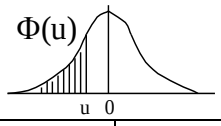
Biomedizinische Statistik

Version 03 / 2024 (Skript 1295)

Prof. Dr. Stefan von Weber, HS Furtwangen University,
Fakultät MME

Inhaltsverzeichnis

0. Tabelle der Sicherheitspunkte der t-, χ^2-, F- und $\Phi(u)$-Verteilung für α=0.05 (5%)

FG	t		χ^2	F (einseitig rechts)							FG1 / FG2	$\Phi(u)$ 	
	eins.	zweis		1	2	3	4	5	10	20		u	$\Phi(u)$
1	6,31	12,71	3,84	161	200	216	225	230	242	248	1		
2	2,92	4,30	5,99	18,5	19,0	19,2	19,2	19,3	19,4	19,4	2	-0,1	0,4602
3	2,35	3,18	7,81	10,1	9,55	9,28	9,12	9,01	8,79	8,66	3	-0,2	0,4207
4	2,13	2,78	9,49	7,71	6,94	6,59	6,39	6,26	5,96	5,80	4	-0,3	0,3821
5	2,02	2,57	11,07	6,61	5,79	5,41	5,19	5,05	4,74	4,56	5	-0,4	0,3446
6	1,94	2,45	12,59	5,99	5,14	4,76	4,53	4,39	4,06	3,87	6	-0,5	0,3085
7	1,89	2,36	14,07	5,59	4,74	4,35	4,12	3,97	3,64	3,44	7	-0,6	0,2742
8	1,86	2,31	15,51	5,32	4,46	4,07	3,84	3,69	3,35	3,15	8	-0,7	0,2420
9	1,83	2,26	16,92	5,12	4,26	3,06	3,63	3,48	3,14	2,93	9	-0,8	0,2119
10	1,81	2,23	18,31	4,96	4,10	3,71	3,48	3,33	2,98	2,77	10	-0,9	0,1841
11	1,80	2,20	19,68	4,84	3,98	3,59	3,36	3,20	2,85	2,65	11	-1,0	0.1587
12	1,78	2,18	21,03	4,75	3,89	3,49	3,26	3,11	2,75	2,54	12	-1,1	0,1357
13	1,77	2,16	22,36	4,67	3,81	3,41	3,18	3,03	2,67	2,46	13	-1,2	0,1151
14	1,76	2,14	23,68	4,60	3,74	3,34	3,11	2,96	2,60	2,39	14	-1,3	0,0968
15	1,75	2,13	25,00	4,54	3,68	3,29	3,06	2,90	2,54	2,33	15	-1,4	0,0808
16	1,75	2,12	26,30	4,49	3,63	3,24	3,01	2,85	2,49	2,28	16	-1,5	0,0668
17	1,74	2,11	27,59	4,45	3,59	3,20	2,96	2,81	2,45	2,23	17	-1,6	0,0548
18	1,73	2,10	28,87	4,41	3,55	3,16	2,93	2,77	2,41	2,19	18	-1,7	0,0446
19	1,73	2,09	30,14	4,38	3,52	3,13	2,90	2,74	2,38	2,15	19	-1,8	0,0359
20	1,72	2,09	31,41	4,35	3,49	3,10	2,87	2,71	2,35	2,12	20	-1,9	0,0287
21	1,72	2,08	32,67	4,32	3,47	3,07	2,84	2,68	2,32	2,09	21	-2,0	0,0227
22	1,72	2,07	33,92	4,30	3,44	3,05	2,82	2,66	2,30	2,07	22	-2,1	0,0179
23	1,71	2,07	35,17	4,28	3,42	3,03	2,80	2,64	2,27	2,04	23	-2,2	0,0139
24	1,71	2,06	36,42	4,26	3,40	3,01	2,78	2,62	2,25	2,02	24	-2,3	0,01072
25	1,71	2,06	37,65	4,24	3,39	2,99	2,76	2,60	2,24	2,00	25	-2,4	0,00820
26	1,71	2,06	38,89	4,23	3,37	2,98	2,74	2,59	2,22	1,99	26	-2,5	0,00621
27	1,70	2,06	40,11	4,21	3,35	2,96	2,73	2,57	2,20	1,97	27	-2,6	0,00466
28	1,70	2,05	41,34	4,20	3,34	2,95	2,71	2,56	2,19	1,96	28	-2,7	0,00347
29	1,70	2,05	42,56	4,18	3,33	2,93	2,70	2,55	2,18	1,94	29	-2,8	0,00255
30	1,70	2,04	43,77	4,17	3,32	2,92	2,69	2,53	2,16	1,93	30	-2,9	0,00187
34	1,69	2,03	48,60	4,13	3,28	2,88	2,65	2,49	2,12	1,89	34	-3,0	0,001350
40	1,68	2,02	55,76	4,08	3,23	2,84	2,61	2,45	2,08	1,84	40	-3,1	0,000967
44	1,68	2,02	60,48	4,06	3,21	2,82	2,58	2,43	2,05	1,81	44	-3,2	0,000688
50	1,68	2,01	67,50	4,03	3,18	2,79	2,56	2,40	2,03	1,78	50	-3,3	0,000484
60	1,67	2,00	79,08	4,00	3,15	2,76	2,53	2,37	1,99	1,75	60	-3,4	0,000337
70	1,67	1,99	90,53	3,98	3,13	2,74	2,50	2,35	1,97	1,72	70	-3,5	0,000233
80	1,66	1,99	101,88	3,96	3,11	2,72	2,49	2,33	1,95	1,70	80	-3,6	0,000159
90	1,66	1,99	113,15	3,95	3,10	2,71	2,47	2,32	1,94	1,69	90	-3,7	0,0001080
100	1,66	1,98	124,34	3,94	3,09	2,70	2,46	2,31	1,93	1,68	100	-3,8	0,0000723
150	1,66	1,98	179,58	3,90	3,06	2,66	2,43	2,27	1,89	1,64	150	-3,9	0,0000480
200	1,65	1,97	233,99	3,89	3,04	2,65	2,42	2,26	1,88	1,62	200	-4,0	0,0000317
∞	1,65	1,96	∞	3,84	3,00	2,60	2,37	2,21	1,83	1,57	∞		

N=22 systolische Blutdruckwerte in [mmHg] zu Beispiel 3 in der Vorlesung:

122	127	131	104	138	139	148	115	137	144
113	132	134	142	119	127	133	111	141	134
129	131								

1. Einführung

Statistik ist deskriptive oder beschreibende	konfirmatorische oder hypothesenprüfende
Mittelwerte, Standardabweichungen, Regressionskoeffizienten, Korrelationskoeffizienten, Wahrscheinlichkeitsschätzungen	Hypothese → Stichprobe → Test → Aussage zur Population einschließlich Irrtumswahrscheinlichkeit
Beispiel Umfrage in München: Würden Sie gern eine Diät machen? 23 von 100 Probanden antworten mit "JA" → Wahrscheinlichkeit p = 23/100 = 0.23 in der Stichprobe, d.h. unter unter den 100 Befragten. Das ist lediglich eine Schätzung des p-Wertes aller Münchner.	Hypothese (Beispiel): Weniger als 20% aller Münchner wollen eine Diät machen → Umfrage siehe links → asymptotischer Binomialtest 0.23 gegen 0.2 bei n=100 → u=0.75 → Hypothese abgeschmettert, d.h. keine signifikante Abweichung vom Wert 20% gefunden

Die ***klassische konfirmatorische Statistik***, auch *frequentistische Statistik* genannt, setzt voraus, dass man theoretisch unendlich viele Stichproben ziehen kann, und dass dann die aus den Stichproben berechnete Prüf- oder Testgröße unter der Nullhypothese H_0 eine bestimmte Verteilung annimmt. Meistens sind die Testgrößen so konstruiert, dass bei Ziehung der Stichproben aus immer derselben Grundgesamtheit (es gilt die Nullhypothese H_0) eine Verteilung der Testgröße um den Wert null herum entsteht, z.B. in Form einer Glockenkurve, d.h., kleine Werte überwiegen. Große Werte der Prüf- oder Testgröße kommen mit geringer Wahrscheinlichkeit vor und signalisieren einen möglichen Ausnahmefall. Statt anzunehmen, dass einer der seltenen Fälle einer großen Prüfgröße eingetroffen ist, nimmt man lieber an, dass sich die Grundgesamtheiten unterscheiden (Alternativhypothese H_A).

Die ***Bayes'sche Statistik*** setzt keine unendlich oft durchführbare Zufallsexperimente voraus. Sie ist deswegen oft schon bei kleinen Datenvolumen im Vorteil, ist aber auch rechnerisch anspruchsvoller, d.h., nur mit Computerprogrammen zu bewältigen. Durch stufenweise Annäherung, auch unter Verwendung von Vorwissen bzw. Expertenwissen, gelangt man zu einer Wahrscheinlichkeitsaussage für eine bestimmte Hypothese. Wir verwenden in diesem Skript von der Bayes'schen Statistik nur den ***Satz von Bayes*** zur Berchnung einer bedingten Wahrscheinlichkeit.

Statistik heißt Komprimierung, Visualisierung und Analyse von Daten. Ziele der **deskriptiven** Statistik sind Information und Vorhersage künftiger Daten, die Ziele der **konfirmatorischen** Statistik sind die Prüfung von Hypothesen mittels Stichproben. Aus den Daten einer **Stichprobe** zieht man Schlussfolgerungen für die gesamte **Population (**oder **Grundgesamtheit)**. Beispiel: Aus einer Studie mit 15 Patienten zieht man Schlüsse, die für alle Patienten mit dieser Krankheit Gültigkeit haben sollen, mit Angabe der **Irrtumswahrscheinlichkeit.** Eine **Stichprobe** sind z.B. 10 zufällig ausgewählte Bäume aus einem Waldstück. Die **Population** ist das Waldstück. Personen heißen **Patient, Proband, Fall**, Objekte **Fall, Punkt, Messpunkt.**

Arten von Studien

- Therapiestudien (Querschnitts-/Längsschnittstudien, retrospektive / prospektive Studien, retrospektive mit Archivmaterial (Fall-Kontroll-Studie), prospektive (Kohorten-Studien))
- Beobachtungsstudien
- Physiologische und neurophysiologische Untersuchungen
- Experimente mit Tieren, Mikroorganismen oder Technik

Klinische Therapie-Studien: Nachweis der Wirksamkeit einer Therapie. Wichtig ist die Placebogruppe, da auch Placebos eine (psychologische) Wirkung haben. Die Farbe des Placebo ist von Bedeutung (Wirkung steigt von gelb-grün-blau-rot). Wichtig nach Gasser&Seifert:

- Randomisierung der Patienten (Zufallsgruppeneinteilung)
- nach Störgrößen schichten (Alter, Schweregrad,, ...) und diese erfassen (als Kovariable)
- einfach blind (Patient) oder doppelt blind (Arzt und Patient ahnungslos, ob Placebo)
- standardisiertes Vorgehen (replizierbar): Z.B. wann wie viele Pillen in welchen Gefäßen ...
- Einschluss- und Ausschlusskriterien (Übergewicht, Untergewicht, Altersgrenze, ...)
- Zustimmung der Probanden/Patienten nach Aufklärung (informed consent).

Beobachtungsstudien: keine Therapie oder anderweitige Beeinflussung, sondern schätzt Wahrscheinlichkeiten. Beispiele sind:
Bestimmung der Prävalenz = Bestand an Fällen einer bestimmten Krankheit
Bestimmung von Risikofaktoren, z.B. für Herzinfarkt

Versuchsplanung: Versuchsplanung heißt:

- repräsentative Stichproben auswählen
- mit möglichst wenig Kosten ein signifikantes Ergebnis erzielen
- Störfaktoren entweder ausschließen oder als Kovariable messen

Vorversuche: Oft sind Vorversuche nötig, um die auftretenden Varianzen zu kennen.
Beispiel: Schafft Bakterienstamm 2 die Hürde von **d**=3% Verbesserung bei der Insulinproduktion gegenüber Standard-Bakterienstamm 1? Aus Vorversuchen ist bekannt, daß die Insulin-Messungen mit Standardabweichung s=4.3 streuen.

Statistische Aufgabe: Es soll bei s = 4.3 der Insulinmessungen die **notwendige Anzahl n von Messungen** in jeder Gruppe bestimmt werden, damit der gewünschte Effekt **d =** $\mu_2 - \mu_1 \geq 3.0$ in den Mittelwerten auf dem α=5%-Niveau gesichert werden kann.

Lösung hier z.B.: Man nimmt die Teststatistik aus Abschnitt *Mittelwertvergleich zweier normalverteilter Grundgesamtheiten* her, $t=(d/s)*((n*n)/(n+n))^{0.5}$ bzw. $t=(d/s)*(n/2)^{0.5}$ mit FG=2n–2, und probiert so lange, bis man das kleinste n gefunden hat, das ein $t>t_a$ liefert. Gruppenumfang n=19 liefert $t=2.15 > t_\alpha=2.11$, also Signifikanz. Empfohlene Gruppenstärke: Je 19 Messungen mit Stamm 1 und mit Stamm 2.

2. Wahrscheinlichkeitsrechnung

Wozu? Qualitätskontrolle, Chancen berechnen, Im PC stochastische Modelle simulieren, Grundlage für einige Testverteilungen

Die möglichen Ausgänge eines **Zufallsexperiments** heißen **Elementarereignisse** (z.B. eine 4 beim Würfeln). Ihre Menge heißt **Ereignisraum R** (1-6 beim Würfel). Das sichere Ereignis (eine Zahl $1 \leq x \leq 6$) trifft immer ein, das unmögliche Ereignis (z.B. eine 0 oder 7) nie. Die **Wahrscheinlichkeit** P eines Ereignisses ist eine Zahl $0 \leq P \leq 1$ bzw. $0 \leq P\% \leq 100\%$.

Der **Erwartungswert** E der **Häufigkeit**, mit der ein Ereignis eintrifft ist $E = N \cdot P$
N=Zahl der Ziehungen insgesamt, P=Wahrscheinlichkeit für das Eintreffen des Ereignisses

Beispiel Tablettenfehler R={1,2}, N=1.000.000

Elementarereignis	N_i	$P_i = N_i / N$	$P_i\% = P_i \cdot 100$

A_1 (untergewichtig)	632	0,000632	0,0632
A_2 (übergewichtig)	869	0,000869	0,0869

Multiplikationssatz: Die Wahrscheinlichkeit $P[A \cap B]$ für das gemeinsame Eintreffen stochastisch unabhängiger Ereignisse **A und B**: $P[A \cap B]$= P(A)·P(B). Die Wahrscheinlichkeit mit zwei Würfeln A und B zwei Sechser zu würfeln ist $P[6 \cap 6]$= (1/6) · (1/6) = (1/36). **Stochastische Unabhängigkeit** heißt, dass das Eintreffen von A_i nicht von A_j abhängt, A_j nicht von A_i, und es auch keine versteckte Abhängigkeit zwischen den Ereignissen A und B gibt.

Additionssatz: Die Wahrscheinlichkeit P(A∨B) für das Eintreffen entweder des Ereignisses A **oder aber** des Ereignisses B. A und B sind disjunkt, d.h., sie schließen sich gegenseitig aus: P(A∨B) = P(A) + P(B). Auf das Tablettenbeispiel bezogen ist $P(A_1 \vee A_2)$ =0,000632+0,000869= 0,001501, dh., wir finden mit dieser Wahrscheinlichkeit eine unter- oder übergewichtige Tablette. Beides kann eine Tablette nicht sein - unter- und übergewichtig.

Bedingte Wahrscheinlichkeit (oder Satz von Bayes): P[A|B] ist die Wahrscheinlichkeit, dass das Ereignis A eintritt, vorausgesetzt, dass vorher das Ereignis B eingetreten ist. Beispiel: P(A) sei die Wahrscheinlichkeit, bei einem beliebigen Patienten Diabetes zu diagnostizieren. P(B) sei die Wahrscheinlichkeit, dass ein beliebiger Patient übergewichtig ist. $P[A \cap B]$ ist die Wahrscheinlichkeit, dass die Ereignisse A und B zusammen eintreffen, d.h., die Wahrscheinlichkeit, dass Übergewicht und Diabetes zusammen auftreten. P[B|A] ist dann die ***bedingte Wahrscheinlichkeit***, dass eine übergewichtige Person Diabetes bekommt.

$$P[B \mid A] = \frac{P[A \cap B]}{P[A]}.$$

P[A|B] ist dann die ***bedingte Wahrscheinlichkeit***, dass eine Person mit Diabetes übergewichtig ist.

$$P[A|B] = \frac{P[A \cap B]}{P[B]}.$$

Zahlenbeispiel: Diabetes haben im Schnitt 9,3% aller Patienten über 50 Jahre, d.h. P(A)=0,093.
Übergewichtig sind im Schnitt 37,1% aller Patienten über 50 Jahre, d.h. P(B)=0,371.
$P[A \cap B]$=0,069 (6,9%) ist die Wahrscheinlichkeit, dass die Ereignisse A und B zusammen eintreffen, d.h., die Wahrscheinlichkeit, dass Übergewicht und Diabetes zusammen auftreten. Dann ist P[B|A]=0,069/0,093=0,742 (74,2%) die ***bedingte Wahrscheinlichkeit***, dass eine übergewichtige Person Diabetes bekommt, und P[A|B]=0,069/0,371=0,186 (18,6%) die ***bedingte Wahrscheinlichkeit***, dass eine Person mit Diabetes übergewichtig ist.

Wahrscheinlichkeit eines metrischen Messwertes: Die Mittagstemperatur exakt 20°C hat als mathematischer Punkt Wahrscheinlichkeit 0. Wir müssen ein Intervall vorgeben, z.B. P[19.5<x<20.5], um eine endliche Wahrscheinlichkeit zu erreichen.

Permutationen: Die Zahl möglicher Anordnungen von N unterscheidbaren Objekten ist N_P=N! (sprich N-Fakultät) mit 0!=1, 1!=1, 2!=1·2=2, 3!=1·2·3=6, N!=1·2·3⋯N.

Beispiel: 3 Aufträge A, B, C lassen sich in die 6 Reihenfolgen ABC, ACB, BAC, BCA, CAB, CBA bringen. Besteht die Menge der N Objekte aus k Gruppen (mit den Gruppenumfängen N_1, N_2, ..., N_k,) innerhalb derer sich die Objekte nicht unterscheiden, dann ist die Anzahl der unterscheidbaren Anordnungen $N_P = \frac{N!}{N_1! \cdot N_2! \cdots N_K!}$. Z.B. ein weibliches und ein männliches Zwillingspaar ergibt die 6 unterscheidbaren Anordnungen WWmm, WmWm, ..., mmWW.

3. Grundlagen der Statistik

3.1 Fehlerarten, Zufallszahlen, Skalen :

Systematische Fehler entstehen z.B. durch falsche Versuchspläne (z.B. sehr ungleiche Gruppengrößen, großer Altersunterschied, ...), falsch kalibrierte Messinstrumente, nicht operationalisierte Kriterien bei unterschiedlichen Ärzten. Man kann systematische Fehler vermeiden oder teilweise korrigieren, wenn man sich an die →Empfehlungen der Versuchsplanung hält.

Zufallsfehler / Zufallszahlen: Messwerte oder beobachtete Größen werden in der Statistik als Zufallszahlen aufgefasst. Die mittlere Mittagstemperatur im Sommer ist 21,3 °C. Abweichungen vom Mittel werden nicht durch Wetterforschungen erklärt, sondern als Zufall aufgefasst. Eine Frau hat im Schnitt 1,4 Kinder. Die tatsächliche Kinderzahl einer Frau wird nicht durch die Lebensumstände erklärt, sondern als Zufall. Eine **Zufallsvariable** ist eine Funktion, die dem Ausgang eines Zufallsexperiments eine reelle Zahl zuordnet. Ein Zufallswert x der Zufallsvariablen X heißt **Realisierung** oder **Ausprägung**.

Diskrete Zufallszahlen können nur bestimmte, meist ganzzahlige Werte (Realisierung, Ausprägung, Symptom, Kategorie) annehmen (Kinderzahl 1, 2, ..., oder Weinflaschenvolumen 0.5, 0.75, 1.0, 1.5, 2.0, 3.0, 5.0 usw.). **Kontinuierliche Zufallszahlen** können im Definitionsbereich beliebige Werte (Realisierungen) annehmen (z.B. T=21.3 °C oder T= 21.297 °C).

Nominale (qualitative) Daten sind immer diskret und dienen nur zur Sortierung und Gruppeneinteilung. Z.B. ist die Postleitzahl in Patientenadressen eine nominale Größe. Summen oder Mittelwerte aus nominalen Daten sind Unsinn. **Kategoriale Daten** werden wie nominale behandelt, wenn sie nicht **ordinal** sind, d.h. keine Rangordnung dahinter versteckt ist (z.B. "weiße", "rote", "schwarze" Pilzkolonien, auch wenn sie mit 1,2,3 im Rechner codiert sind.).

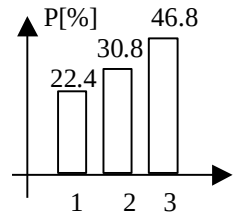
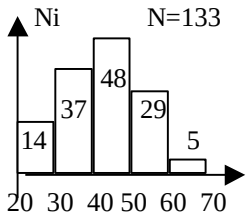
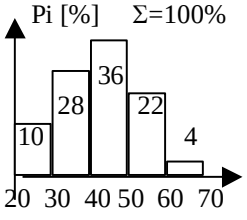
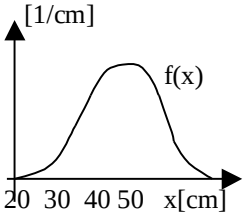
Metrische (quantitative, stetige) Daten lassen sich auf einer Zahlengeraden anordnen. Es besteht zwischen zwei Werten immer eine der Beziehungen "<", "=" oder ">". Mit metrischen Daten darf man rechnen (Summen, Mittelwerte, ...). **Ranggeordnete kategoriale Daten** werden oft wie metrische Daten behandelt, z.B. Zensuren 1,2,...,5 oder Wagenklassen 1="klein", 2="mittel", 3="groß". **Binärdaten** (mit nur zwei Ausprägungen) sind alles: Nominal, kategorial und sie können ebenfalls wie metrische Daten behandelt werden (z.B. weiblich=1, männlich=2 oder gesund=0, krank=1). Oft erzeugt man Binärdaten aus metrischen Daten durch eine **Dichotomisierung** am Median, d.h nimm z. B. Binärzahl 0, wenn Zahl<Median ist, und nimm die 1, wenn Zahl ≥Median ist.

Prozentzahlen sind ein heikles Thema, und es wird kontrovers diskutiert, ob man mit ihnen Statistik treiben darf.. Beispiel: Ein Patient aus einer Gruppe von zwei ergibt 50%, oder ein Patient aus einer Gruppe von drei ergibt 33,3%. Die Prozentzahlen verbergen die tatsächlichen Häufigkeiten, und führen so leicht zu falschen Schlussfolgerungen. Vorschlag: Benutze Prozentzahlen nur, wenn sie aus genügend großen Häufigkeiten berechnet werden

Ipsative Skalierung: Die Skalen unterscheiden sich von Individuum zu **Individuum. Ipsativierung** ist der Ausgleich der am Individuum gemessenen Werte am individuellen Mittelwert. Ipsative Daten entstehen z.B. durch ***forced-choice***-Befragungen, d.h., man darf nur *eine* Auswahlmöglichkeit von zwei möglichen zu einer Frage ankreuzen. Ein ipsativer Persönlichkeitstest zeigt die Rangordnung von Merkmalen *in einer Person* auf. Der Vergleich von Personen untereinander wird dadurch fragwürdig. Dazu bedarf es ***normativer Daten***. Hier kann man zu einer Frage angeben, z.B. auf einer Skala von 0-10, wie sehr man die Frage bejahen möchte. Die normative Befragung ermöglicht den Vergleich, wie jemand *in Bezug auf eine Normgruppe* abschneidet. In der **Zeitreihenanalyse** bezeichnet die ipsative Skalierung, dass man die Daten z.B. auf Jahresmittel zentriert (es gibt gute und schlechte Jahre), in der **Varianzanalyse**, dass man Daten auf das Mittel der Faktorstufen zentriert.

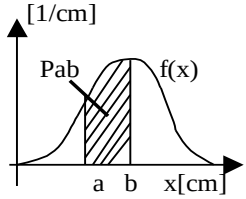
3.2 Verteilungen

Die **Verteilungsfunktion** gibt Auskunft, wie viele Daten mit welcher Abweichung vom Mittelwert erwartet werden. Die Darstellung der Verteilung diskreter Zufallszahlen erfolgt mit dem **Balken- oder Tortendiagramm**. Jeder Balken entspricht einer Ausprägung der Zufallszahl. Die Darstellung der Verteilung kontinuierlicher Zufallszahlen erfolgt bei **beobachteten Daten** mit dem Balkendiagramm (**Histogramm** der absoluten oder relativen Häufigkeiten), bei theoretischen Verteilungen mit dem **Liniendiagramm**. Die Festlegung der Klassenanzahl und damit der Klassenbreite (z.B. 10 cm bei den Stammdurchmessern) richtet sich nach der Gesamtzahl N und erfordert einiges Probieren. Großes N $\Rightarrow$ viele Klassen, kleines N $\Rightarrow$ wenig Klassen. Es gibt keine Vorschrift. Das **kumulative Histogramm** beobachteter Daten ist eine Treppenfunktion, die aufsteigend die Werte von 0 bis 1 annimmt. (Siehe Summenverteilung)

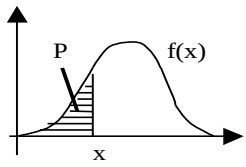
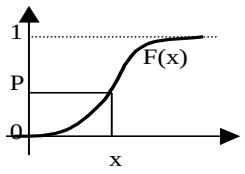
Diskrete Verteilung (Tablettenfehler A_1, A_2, A_3)	Histogramm absoluter Häufigkeiten (Stammdurchmesser)	Histogramm relativer Häufigkeiten (Stammdurchmesser)	Liniendiagramm einer Dichte-Normalverteilung
			

- Bei einer diskreten Verteilung ist $\Sigma P_i=1$ bzw. $\Sigma P_i\%=100\%$
- Beim Histogramm der absoluten Häufigkeiten ist $N=\Sigma N_i$ (N = Gesamtzahl der Objekte)
- Beim Histogramm der relativen Häufigkeiten ist $\Sigma P_i=1$ bzw. $\Sigma P_i\%=100\%$
- Bei einer Dichteverteilung ist die Gesamtfläche unter der Dichtekurve f(x) immer gleich 1.

Theoretische Verteilungen folgern aus einem **Modellprozess**. Die **Dichtefunktion** f(x) gibt mit ihrer Fläche über dem Intervall [a,b] die Wahrscheinlichkeit P_{ab} an, dass ein x-Wert aus dem Intervall [a,b] auftritt. X ist eine kontinuierliche Zufallsvariable.

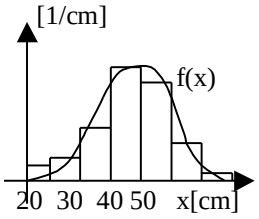
	Wahrscheinlichkeit $P_{ab} = \int_a^b f(x)dx$	Normierung $\int_{-\infty}^{+\infty} f(x)dx = 1$

Die **Verteilungsfunktion** (Summenverteilung) F(x) gibt mit ihrem Funktionswert F(x) die Wahrscheinlichkeit P an, mit der ein Zufallswert aus dem Intervall [−∞,x] auftritt.

Dichteverteilung	Verteilungsfunktion	Formel Verteilungsfunktion
		$F(x) = \int_{-\infty}^{x} f(u)du$

Wann man die Dichtefunktion verwendet oder die Verteilungsfunktion, dafür gibt es keine Vorschriften. Die Verteilungsinformation steckt in beiden Kurven. Eine Verteilung kann durch die **Momente** μ_i charakterisiert werden, ohne dass man das genaue Bild der Funktion vorliegen hat. Das entspricht in etwa der Taylorreihenentwicklung der Dichtefunktion. Die Momente μ_1–μ_4 haben die Namen Mittelwert, Varianz, Schiefe und Exzess. Die höheren Momente (ab μ_2) werden auf das arithmetische Mittel bezogen berechnet (x–E).

1. Moment: Erwartungswert (Mittelwert, arithmetisches Mittel, Schwerpunkt)	2. Moment: Varianz Bei Normalverteilung ist $\mu_2=\sigma^2/2$ mit σ = Standardabw.	3. Moment: Schiefe $\mu_3>0$: Gipfel links von E $\mu_3<0$: Gipfel rechts von E
$\mu_1 = E(x) = \int_{-\infty}^{+\infty} x \cdot f(x)dx$	$\mu_2 = \int_{-\infty}^{+\infty} (x-E)^2 f(x)dx$	$\mu_3 = \int_{-\infty}^{+\infty} (x-E)^3 f(x)dx$

Theoretische Datenverteilungen werden beobachteten oder gemessenen Daten **unterstellt.** Man sagt z.B., die Daten seien normal verteilt oder sie seien binomial verteilt. Einen Beweis, dass die Daten tatsächlich so verteilt sind, gibt es nicht. Mit dem χ^2-Anpassungstest oder dem Kolmogorov-Smirnov-Test kann man jedoch Abweichungen zwischen beobachteter Verteilung und unterstellter theoretischer Verteilung statistisch bewerten, und zu einer Aussage z.B. der Form kommen: "Es gibt keine signifikante Abweichung von der Normalverteilung".	

Wichtige theoretische Verteilungen für diskrete Zufallszahlen sind die Poisson-Verteilung, die Binomialverteilung, multinomiale Verteilung und hypergeometrische Verteilung. Alle vier Verteilungen werden auch als **Prüfverteilungen** zur Prüfung von Hypothesen benutzt, wenn auch seltener, als die u-, t-, χ^2- und F-Verteilung..

Die **Poisson-Verteilung** ist eine diskrete Verteilung mit Parameter λ=Np >0. N ist die Zahl der Ziehungen mit $N \to \infty$ und Wahrscheinlichkeit $p \to 0$. Sie liefert die Wahrscheinlichkeit, dass ein seltenes Ereignis (z.B. Todesfall	$P_{N.k} = \frac{\lambda^k}{k!} e^{-\lambda}$

durch Meteoriteneinschlag bei N Meteoriten pro Jahr, oder der Zerfall eines radioaktiven Materials bei N Atomen und im Zeitraum t Sekunden) genau k mal eintritt. Die Verteilung hat Erwartungswert $\lambda=Np$ und Varianz $\sigma^2=\lambda=Np$	

Die Binomialverteilung hat als Modell eine Urne mit Anteil p an schwarzen und Anteil q=1−p an weißen Kugeln. $P_{n.k}$ ist die Wahrscheinlichkeit, bei n Ziehungen mit Zurücklegen genau k schwarze Kugeln zu ziehen. p heißt Parameter der Binomialverteilung. Erwartungswert der Binomialverteilung ist E= n p, Varianz ist σ^2 = pq n.

$$P_{n.k} = \binom{n}{k} p^k q^{n-k} \quad \text{mit} \quad \binom{n}{0} = 1, \quad \binom{n}{k} = \frac{n(n-1)\ldots(n-k+1)}{1\cdot 2\cdot \ldots \cdot k} \quad \text{(sprich "n über k")}$$

Beispiel: Ein Prozess gerät mit Wahrscheinlichkeit p=0,068 außer Kontrolle (Erfahrungswert). Wie hoch ist die Wahrscheinlichkeit, dass von den 10 Chargen einer Woche 3 versaut sind?

$$P_{10.3} = \binom{10}{3} 0.068^3 \cdot 0.932^7 = \frac{10\cdot 9\cdot 8}{1\cdot 2\cdot 3} \cdot 0.00031 \cdot 0.611 = 0.023 \quad \text{oder 2,3\%}$$

Man rechnet also etwa jede 40. Woche mit 3 versauten Chargen. Die Summe P der 11 Wahrscheinlichkeiten P = $P_{10.0}$+ $P_{10.1}$+...+ $P_{10.10}$ ist exakt P=1.

Die hypergeometrische Verteilung folgt demselben Urnenschema, wie die Binomialverteilung. Der Unterschied ist, dass ohne Zurücklegen gezogen wird, d.h. jede Ziehung ändert p und q, sofern die Zahl N der Kugeln in der Urne nicht allzu groß ist. Auch hier wird die Wahrscheinlichkeit φ berechnet, mit n Ziehungen k schwarze Kugeln von den änfänglich M schwarzen Kugeln zu ziehen. (p=M/N). Für große N geht die hypergeometrische Verteilung in die Binomialverteilung über	$$\varphi = \frac{\binom{n}{k}\binom{N-M}{n-k}}{\binom{N}{n}}$$

Die wichtigsten kontinuierlichen theoretischen Verteilungen sind die Normalverteilung (auch u- oder Gauß-Verteilung), die lognormale Verteilung, die t- oder Student-Verteilung, die χ^2-Verteilung (Chi-Quadrat-Verteilung) und die F-Verteilung (Fisher-Verteilung). Die Normalverteilung (u-Verteilung) und die lognormale Verteilung treten häufig als Datenverteilung auf. Die t-, χ^2- und F-Verteilung sind seltener Datenverteilungen, sondern werden weit häufiger als **Prüfverteilungen** zum Testen von Hypothesen benutzt. Die Gaußverteilung (u-Verteilung) ist beides - Datenverteilung und Prüfverteilung.

Dichtefunktion der Normalverteilung: μ (Erwartungswert) und σ^2 (Varianz) heißen Parameter der Normalverteilung. Man schätzt sie durch eine Stichprobe, indem man für μ den *Mittelwert* und für σ^2 die *Varianz* σ^2_{n-1} einsetzt. Normalverteilte Zufallszahlen entstehen, wenn sich viele Zufallseinflüsse addieren.	$$f(x) = \frac{1}{\sqrt{2\pi}\,\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

Bei angenommener Normalverteilung einer Population und Schätzung ihrer Parameter μ und σ^2 der Population aus einer Stichprobe gilt:

Stichprobenstatistik		Schätzwert		Parameter der Population
Mittel	$\bar{x} = \Sigma x_i / n$	$\rightarrow$	$\hat{\mu}$	$\rightarrow$ μ

Varianz $\sigma_{n-1}^2 = \frac{\sum (x_i - \bar{x})^2}{n-1}$ $\rightarrow$ $\hat{\sigma}^2$ $\rightarrow$ σ^2

Beispiel n=10 Stammdurchmesser: 36 41 39 52 48 53 55 61 54 49 cm. Das Mittel 48.8 cm ist Schätzwert für das unbekannte Populationsmittel μ. Die Standardabweichung σ_{n-1}=7,91 cm ist Schätzwert der unbekannten Standardabweichung σ der Population. Die **wahren Parameter** μ und σ^2 der Population kann man nur für $n \rightarrow \infty$ erhalten. Die Schätzwerte sind fehlerbehaftet.

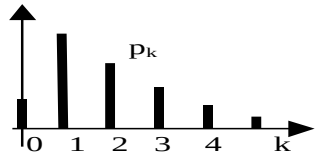
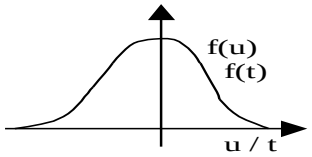
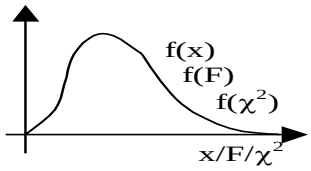
Normalverteilung mit Mittelwert μ und **Varianz** σ^2 wird mit **$N(\mu ; \sigma^2)$** abgekürzt. N(0;1) ist die **Standard-Normalverteilung** mit Mittel 0 und Varianz 1. Die Verteilungsfunktion (Summenkurve) **$\Phi(u)$** zur Normalverteilung f(x) wird auch **Gaußsches Fehlerintegral** genannt und ist in vielen Büchern tabelliert. $\Phi(u)$ und Umkehrfunktion **$u(\Phi)$** sind wichtige Prüfverteilungen. Die Normalverteilung ist wichtig wegen des **zentralen Grenzwertsatzes**: *Die Verteilung der Summe beliebig verteilter Zufallszahlen* **z** *nähert sich für wachsende Zahl an Summanden der Normalverteilung*, d.h. in der Praxis ist die Größe $S=z_1+z_2+...+z_n$ schon ab n=5 recht gut normalverteilt. Darunter fällt z.B. jedes Stichprobenmittel mit Stichprobenumfang $n \geq 5$.

Dichtefunktion der lognormalen Verteilung: M (Erwartungswert) und S^2 (Varianz) heißen Parameter. Man berechnet aus den logarithmierten Daten Mittelwert und Varianz und setzt diese gleich M und S. Lognormale Zufallszahlen entstehen, wenn sich Zufallseinflüsse multiplizieren. Die Verteilung ist unsymmetrisch.	$f(x) = \frac{1}{S \cdot x\sqrt{2\pi}} e^{-\frac{(\ln(x)-M)^2}{2S^2}}$
Die t-Verteilung (auch Student-Verteilung nach dem Pseudonym *Student* von W. P. Gosset) ist die Verteilung des Quotienten $t = u / \chi$. Dabei ist u N(0;1)-verteilt und χ^2 ist χ^2-verteilt mit k Freiheitsgraden. Die Verteilung ist symmetrisch.	$t = \frac{u}{\chi}\sqrt{k}$
Die χ^2-Verteilung (Chi-Quadrat-Verteilung von F.R. Helmert und K. Pearson) ist die Verteilung der Summe $\chi^2 = u_1^2+...u_k^2$. Die u_i sind N(0;1)-normalverteilt und stochastisch unabhängig. Freiheitsgrad FG der Verteilung ist k. Unsymmetrische Verteilung.	$\chi^2 = u_1^2 + \ldots + u_k^2$ mit k Freiheitsgraden
Die F-Verteilung von R. A. Fisher ist die Verteilung des Quotienten $F= \chi^2_1 / \chi^2_2$. Dabei ist χ^2_1 mit FG_1 Freiheitsgraden und χ^2_2 mit FG_2 Freiheitsgraden verteilt.	$F= \chi^2_1 / \chi^2_2$ Mit FG_1und FG_2 Freiheitsgraden

Die F-Verteilung ist insofern interessant, da sie die t- und die χ^2-Verteilung quasi enthält.
Es gilt $t^2(FG) = F$ mit $FG_1=1$ und $FG_2=FG$.
Es gilt $\chi^2(FG)= FG_2\, F$ mit $FG_1 \rightarrow \infty$ und $FG_2=FG$.

Der **Freiheitsgrad FG** ist die Zahl der „freien Datenpunkte“, die zur Berechnung einer Streuung herangezogen werden können. Beispiel Abweichung der Punkte von einer Ausgleichsgeraden. Bei zwei Punkten geht die Gerade exakt durch beide Punkte. Kein Punkt ist frei (FG=0).
Bei drei Punkten ist einer überzählig (FG=1). Allgemein im Fall der Geraden ist $FG = n - 2$.

Die folgenden drei Abbildungen zeigen das typische Aussehen der folgenden Verteilungen:

Poisson-, Binomial-, Hypergeometrische Verteilung	Normalverteilung, t-Verteilung	lognormale, χ^2-, F-Verteilung
		

3.3 Schätzung von Verteilungsparametern

Ein Schätzwert (oder Schätzer) ist eine nach einer bestimmten Formel berechnete Zahl, die dem gesuchten Parameter einer Population, z.B. dem Mittelwert, möglichst nahe kommt. Es gibt gute, sehr gute und den **besten Schätzwert**. Eine allgemeine Methode zum Aufspüren des besten Schätzers heißt **Maximum Likelihood**. Die beobachteten Daten haben *höchste Wahrscheinlichkeit*, wenn gerade die besten Schätzwerte als Parameter der **angenommenen Datenverteilung** benutzt werden. Bei Annahme der Normalverteilung sind *Maximum Likelihood* und die *Methode der kleinsten Quadrate* asymptotisch (d.h. für $n\to\infty$) identisch.

Punktschätzung und Konfidenzintervall: Punktschätzung heißt die Berechnung eines einzelnen Wertes aus einer Stichprobe, z.B. des Stichprobenmittels als Punktschätzung für das unbekannte Populationsmittel. In der deskriptiven (beschreibenden) Statistik haben Punktschätzungen einen festen Platz. In der konfirmatorischen (hypothesenprüfenden) Statistik werden Punktschätzungen nur berechnet als Grundlage für die Konstruktion der Konfidenzintervalle. **Konfidenzintervalle**: Bei oftmaliger Wiederholung einer Studie würden wir ähnliche, aber andere Schätzwerte für einen gesuchten Parameter θ erhalten. Das ist der Zufallseffekt - andere Patienten, andere Jahreszeit usw. Wo liegt jetzt der wirkliche Wert unserer gesuchten Zahl θ? Hier hilft das Konfidenzintervall weiter:

> Ein $(1-\alpha)$–Konfidenzintervall $[\theta_U, \theta_O]$ für den Parameter θ ist ein zufälliges Intervall, das mit Wahrscheinlichkeit $(1-\alpha)$ den gesuchten Wert θ enthält.

Einige Konfidenzintervalle als Beispiele sind:

Konfidenzintervall für μ bei bekanntem σ_0^2, aus n Werten geschätztem Stichprobenmittel $\bar{x}$ und vorausgesetzter Normalverteilung	$\bar{x} \pm \frac{\sigma_0}{\sqrt{n}} u(1-\alpha/2)$
Konfidenzintervall für μ. Sowohl $\bar{x}$ und σ^2_{n-1} werden bei vorausgesetzter Normalverteilung aus einer Stichprobe des Umfangs n geschätzt	$\bar{x} \pm \frac{\sigma_{n-1}}{\sqrt{n}} t(\alpha, FG = n-1, zweis.)$
Exaktes Konfidenzintervall $[p_U,p_O]$ für relative Häufigkeit $\hat{p} = k/n$. F_1 ist die F-Verteilung mit den Freiheitsgraden $FG_{11}=2(k+1)$, $FG_{12}=2(n-k)$, F_2 ist die F-Verteilung mit den Freiheitsgraden $FG_{21}=2(n-k+1)$, $FG_{22}=2k$.	$p_U = \frac{k}{k + (n-k+1)\cdot F_1(\alpha/2, FG_{11}, FG_{12})}$ $p_O = \frac{(k+1)\cdot F_2(\alpha/2, FG_{21}, FG_{22})}{n-k+(k+1)\cdot F_2(\alpha/2, FG_{21}, FG_{22})}$
Approximatives ($n\to\infty$) Konfidenzintervall $[p_U,p_O]$ für relative Häufigkeit $\hat{p} = k/n$.	$\hat{p} \pm u(1-\alpha/2)\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$

4. Datenerhebung

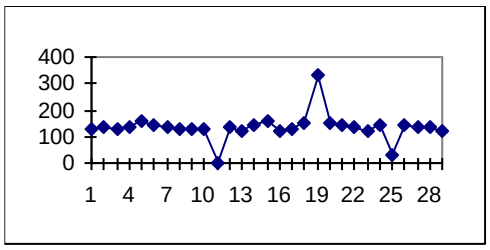
4.1 Dateneingabe, Kontrolle

Daten werden als Tabellen eingegeben. Am besten schreibt man sie mit einem einfachen Texteditor als sogenanntes ASCII-File oder als EXCEL-Tabelle. Schon bei der Benutzung von Umlauten oder des Tabulators sollte man vorsichtig sein. Manche Statistikprogramme mögen sie nicht. Buchstaben als Datenwerte vermeiden, sondern als Zahl codieren, z.B. weiblich w=1, männlich m=2 oder Blutgruppe 0=0, A=1, B=2, AB=3. Fehlende Werte werden von EXCEL nicht immer toleriert, manche andere Programme akzeptieren sie. Notfalls muss man per Hand alle Zeilen mit fehlenden Werten streichen, bevor man mit EXCEL arbeitet. Auf keinen Fall eine Null einsetzen!

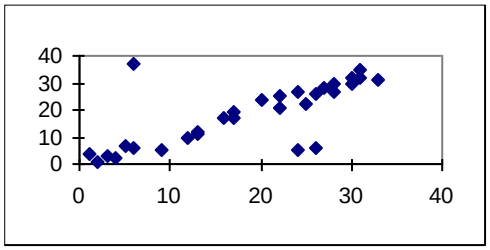
Datensichtung ist ein wichtiger Schritt vor der eigentlichen Arbeit mit den Daten:
Kommafehler beim Eintippen, z.B. Gewichte von Patienten 84.3 77.1 59.0 ... **820.** ...
Zahlendreher, z.B. Alter von Schülern 12 17 14 ... **71** ...
Null als Fehlwert, z.B. Gewichte von Patienten 84.3 77.1 59.0 ... **0** ...
Falsche Gruppenzuordnung, z.B. 2=krank statt 1=gesund 1 1 1 1 ... **2** ...

Eine **erste Kontrolle** erfolgt mit den gewöhnlichen statistischen Maßzahlen, wie Mittelwert, Standardabweichung, Maximum, Minimum, Anzahl der gültigen Werte, die es in jedem Statistikpaket und in EXCEL gibt.

Visuelle Ausreißerkontrolle im X-N-Plot (Excel)
Die Daten stehen in einer Spalte oder Zeile → Markieren der Daten → *Einfügen* → *Diagramm* → *Auf dieses Blatt* → ein Rechteck für die Graphik ziehen → *weiter* → *Linien* → *weiter* → *1* → *weiter* → *weiter* → *ende*



Visuelle Ausreißerkontrolle im X-Y-Plot (Excel)
Die Daten stehen in zwei Spalten oder Zeilen → Markieren der Daten → *Einfügen* → *Diagramm* → *Auf dieses Blatt* → ein Rechteck für die Graphik ziehen → *weiter* → *Punkte* → *weiter* → *1* → *weiter* → *weiter* → *ende*



Ausreißerkontrolle mit der 3-Sigma-Regel: Ist der Wert $u = (|X - \bar{X}|)/\sigma_{n-1}$ zu einem Datenwert X größer als 3, dann ist X als Ausreißer verdächtig. $\bar{X}$ ist der Mittelwert und σ_{n-1} die Standardabweichung der Daten. Entfernt man den Ausreißer aus den Daten, dann ist eine neue Kontrolle nötig, da sich σ_{n-1} und $\bar{X}$ damit ändern.

4.2 Transformation von Daten

Die meisten Datentransformationen haben das Ziel, normal verteilte Daten zu erzeugen. Warum? Die besten statistischen Verfahren setzen normal verteilte Daten voraus. Viele Verteilungen sind jedoch **rechtsschief**, d.h. der Modalwert (höchster Punkt der Dichteverteilung) und der Median (dieser teilt die Fläche zu je 50%) liegen links vom Mittelwert.

Die Ziele einer Datentransformation sind:

- Erreichen einer quasi Normalverteilung

- Varianzstabilisierung: Bei der χ^2-Verteilung z.B. ist $\sigma^2=2\mu$, d.h. die Varianz ändert sich mit dem Mittelwert, bei Normalverteilung sind μ und σ^2 unabhängig voneinander. Beim Gruppenvergleich z.B. stören solche Varianzeffekte.

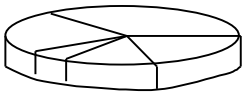
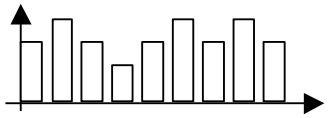
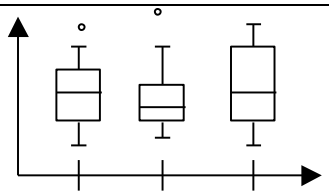
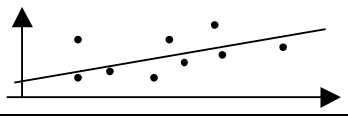
Folgende Transformationen sind u. a. gebräuchlich: $\sqrt{x}, \sqrt[3]{x}, \ln(x), \log(x), -\frac{1}{\sqrt{x}}, -\frac{1}{x}$

Davon ist die erste die schwächste, die letzte, -1/x, die stärkste. ln(x) und log(x) unterscheiden sich nur durch einen konstanten Faktor. Empfehlungen:

Körpergewicht	Transformation $-1/\sqrt{x}$ (Gasser&Seifert)
Prozentwerte, relative Häufigkeiten, z.B. x=h/n	Transformation $\arcsin\left(\sqrt{x/n}\right)$
absolute Häufigkeiten, Zählwerte	Transformationen $\sqrt{x}, \sqrt{x+3/8}$
Verweildauer	Transformation $-1/x$

Indexierung einer Zahlenreihe $X_1, X_2, ..., X_n$ auf Start bei 100%: Bei der Darstellung sehr unterschiedlich hoher Kurven (z.B. Umsatzvergleiche) sieht die untere Kurve oft miserabel aus. Hier hilft die Indexierung. Jede Kurve startet bei 100% und verändert sich nur relativ zu diesem Startpunkt. Die Formel ist $X'_i = (X_i \cdot 100) / X_1$
Man dividiert jeden Wert durch den ersten Wert der Zahlenreihe und multipliziert mit 100.

5. Diagramme

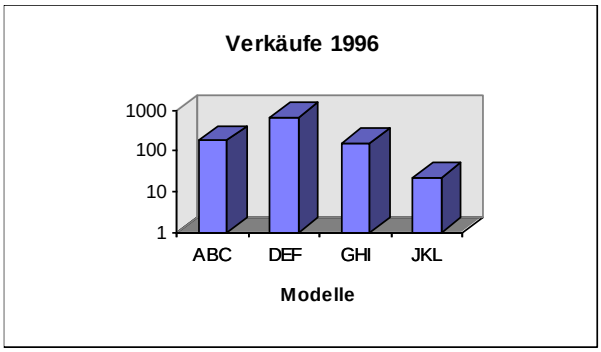
Tortendiagramm bei der Aufteilung eines Kuchens (100%) unter verschiedenen Klassen, z.B. Marktanteile an der Europäischen Vitaminproduktion	
Balkendiagramm bei der Darstellung von Summendaten (Histogramme, Vergleich der Quartalssummen, Vergleich von Gruppenmitteln, Vergleich der Umsätze in den Jahren 2000 - 2005, ...)	
Liniendiagramm bei der Darstellung von Punkdaten (Verlauf der Tagesmitteltemperaturen, Gewichtsverlauf, Fieberkurve,...)	
Boxplots zeigen auf einen Blick die Verteilung von Daten. Die eigentliche Box gibt den Bereich vom 25%- bis zum 75%-Perzentil an mit dem Median als Teilung. Die "whiskers" an den Enden geben das 10% und das 90%-Perzentil an. Manche Boxplots zeigen als Punkte oder Kreise noch die extremen Werte an. Beispiel: 3 Gruppen im Vergleich.	
Scatterplots (x-y-Diagramme) zeigen die Messwerte als Punkte in einem Koordinatensystem, oft mit einem Liniendiagramm gekoppelt.	

Koordinatenachsen haben einen **Maßstab**. Dieser hat
- einen **Bereich** von **Anfang** bis **Ende,** der Anfang muss nicht immer Null sein
- eine **Teilung**, die fein oder grob sein kann,

- eine **Skala**, die **linear** oder **logarithmisch** sein kann

Beispiel 3-D-Säulendiagramm mit logarithmischer Skala in EXCEL 5.0

	A	B
Zeile 1	ABC	177
Zeile 2	DEF	672
Zeile 3	GHI	154
Zeile 4	JKL	22



Markiere die Zellen A1 bis B4 → *Einfügen* → *Diagramm* → *Auf demselben Blatt* → Rahmen ziehen → *weiter* → *3-D-Säulen* → *weiter* → *1* → *weiter* → *weiter* → Legende nein → Titel, x-Achsenbeschriftung → *Ende* → irgendeine Zelle anklicken → Doppelklick im Diagramm → Doppelklick Verkaufzahlenachse → *Skalierung* → *logarithmisch*

6. Statistische Maßzahlen

Die wichtigsten Maßzahlen sind Mittelwerte, Streumaße, Anzahlen (Häufigkeiten).

- Die wichtigsten Mittelwerte sind das arithmetische Mittel, der Median und der Modalwert. Weniger wichtige Unterarten des arithmetischen Mittels sind das gewogene arithmetische Mittel, das geometrische Mittel, das harmonische Mittel.
- Die wichtigsten Streumaße sind die geschätzte Standardabweichung der Grundgesamtheit σ_{n-1}, die Standardabweichung der Stichprobe σ_n, der Fehler des Mittelwerts $\sigma_{\bar{x}}$, der Interquartilabstand $IQR = Q_{0,75} - Q_{0,25}$, und Spannweite $R = x_{max} - x_{min}$.
- Die wichtigste Anzahl ist der Stichprobenumfang n.

Geschätzte Klassenwahrscheinlichkeit $\hat{p}_i = h_i / n$

h_i = Zahl der in Klasse i ausgezählten Werte, n = Stichprobenumfang

arithmetisches Mittel

x_i = Der i-te Wert einer Stichprobe
n = Stichprobenumfang

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

gewogenes arithmetisches Mittel

g_i = Gewicht zum Wert x_i
Die Gewichte müssen positiv (>0) sein.

$$\bar{x} = \left(\sum_{i=1}^{n} g_i x_i\right) / \left(\sum_{i=1}^{n} g_i\right)$$

Beispiel: Gegeben sind die Klassenmitten und Frequenzen (Zahl der Stämme in der Klasse) von 7 Durchmesserklassen von Fichten. Klasse 1 sind z.B. Stämme von 25-30 cm Durchmesser.

Klassenmitte x_i:	27.5	32.5	37.5	42.5	47.5	52.5	57.5
Klassenumfang g_i:	41	84	207	213	156	47	9

$G=\Sigma g_i=757$, $\Sigma g_i x_i=31067.5$, gewichtetes Mittel = 31067.5/757 = 41.04 cm

geometrisches Mittel als n-te Wurzel des Produktes der Einzelwerte

$$\bar{x}_G = \sqrt[n]{\prod x_i} = \sqrt[n]{x_1 \cdot x_2 \cdot \ldots \cdot x_n}$$

$$\bar{x}_G = e^{\left(\sum \ln(x_i)\right)/n}$$

als alternative Formel bei großem n mit ln(x) als natürlichem Logarithmus und e^x als Exponentialfunktion

Beispiel: Ein Aktienfond veränderte sich in den letzten Jahren von einem Jahr zum anderen um +3,6%, – 7,2%, +1.6%, +13.4%. Wegen des Minuszeichens müssen wir auf die absoluten Prozentwerte gehen: 103,6%, 92,8%, 101,6%, 113,4%. Das geometrische Mittel der absoluten Prozentzahlen ist $\sqrt[4]{103.6 \cdot 92.8 \cdot 101.6 \cdot 113.4}$ =102.59. Gehen wir wieder zu relativen Wachstumsraten über, erhalten wir einen jährlichen Zuwachs von 2,59% gemittelt über die 4 Jahre.

Median: Zuerst die Stichprobe sortieren. Bei ungeradem n ist der Wert in der Mitte der Median, bei geradem n ist das arithmetische Mittel der beiden mittleren Werte der Median.
Beispiel: Sortiert man die 10 Durchmesser 54 46 61 47 43 59 38 44 49 41, erhält man die Folge 38 41 43 44 **46 47** 49 54 59 61. Das Mittel der 2 mittleren Werte, 46.5, ist der Median.

Modalwert: Der am häufigsten auftretende Messwert in einer Messreihe sehr großen Umfangs mit unimodaler (eingipfliger) Verteilung. Der Modalwert wird seltener benutzt.

Wann nimmt man welchen Mittelwert?

- Den Median, wenn entweder der *typische Wert* die beste Aussage macht, oder aber ein gegen Datenausreißer robuster Mittelwert gesucht wird. Ein Millionär und 100 arme Schlucker im Dorf haben ein Gesamteinkommen von 1.000.000 +100 x 1000 Euro. Mittelwert 11.000. Typisch für das Dorf sind aber 1000 Euro.
- Das arithmetische Mittel, wenn es um Bilanzen geht. Ein Bach mit 1000 Gramm Schmutzfracht pro m^3 und 100 Gewässer mit 1 Gramm pro m^3 verschmutzen den Bodensee im Mittel mit 11 Gramm pro m^3. Das arithmetische Mittel ist empfindlich für Datenausreißer.
- Den Modalwert, um eine Verteilung zusätzlich zu charakterisieren
- Das gewichtete arithmetische Mittel, um bereits vorverdichtete Zahlen zu mitteln (z.B. möchte man aus Klassenmitteln das Gesamtmittel berechnen, weil die Originaldaten fehlen)

Standardabweichung in der Stichprobe, d.h. nur für die n Daten der Stichprobe

$$\sigma_n = \sqrt{\frac{\sum (x_i - \bar{x})^2}{n}} \qquad \sigma_n = \sqrt{\frac{(\sum x_i^2) - n \cdot \bar{x}^2}{n}}$$

Standardabweichung der Grundgesamtheit geschätzt aus einer Stichprobe des Umfangs n

$$\sigma_{n-1} = \sqrt{\frac{\sum (x_i - \bar{x})^2}{n-1}} \qquad \sigma_{n-1} = \sqrt{\frac{\left(\sum x_i^2\right) - n \cdot \bar{x}^2}{n-1}}$$

Varianz: Das Quadrat einer Standardabweichung heißt Varianz, z.B. σ^2, $\hat{\sigma}^2_{n-1}$
Wird nicht spezifiziert oder nur von σ^2 gesprochen, dann ist immer σ^2_{n-1} gemeint.

Kovarianz der Grundgesamtheit geschätzt aus einer Stichprobe des Umfangs n

$$\mathrm{cov}(x,y) = \frac{\sum (x_i - \bar{x}) \cdot (y_i - \bar{y})}{n-1} \qquad \mathrm{cov}(x,y) = \frac{\left(\sum x_i y_i\right) - n\bar{x}\bar{y}}{n-1}$$

Fehler des Mittelwerts $\sigma_{\bar{x}}$: Ziehen wir aus der Population immer wieder neue Stichproben des Umfangs n, dann streuen die berechneten Mittelwerte um das das unbekannte Mittel μ. Der
Fehler des Mittelwerts schätzt die Ungenauigkeit bei der Bestimmung des wahren Mittelwertes μ (Erwartungswert) einer Grundgesamtheit aus einer Stichprobe des Umfangs n. Ein Mittelwert aus n Einzelmessungen berechnet hat demnach die Genauigkeit oder Standardabweichung $\sigma_{\bar{x}}$, d.h. es gilt $\bar{x} \pm \sigma_{\bar{x}}$.

$$\sigma_{\bar{x}} = \frac{\sigma_{n-1}}{\sqrt{n}}$$

Wann nimmt man welche Fehlerangabe bzw. Streuungsmaß?

- σ_{n-1} (s, Standardabweichung, SD, Standard Deviation) bei allen Angaben, wo man die Variabilität der gemessenen Daten angeben möchte, z.B. die Größe 12-jähriger Knaben ist in Deutschland 143 ± 6 cm. Die Größe schwankt um das Mittel mit durchschnittlich 6 cm.
- Den Interquartilabstand IQR statt σ_{n-1} bei sehr schief verteilten Daten (75%–25%-Quartil)
- $\sigma_{\bar{x}}$ (SE, Standard Error of Mean) wenn man die Genauigkeit einer Schätzung dokumentieren möchte, z.B. aus einer repräsentativen Stichprobe mit 1600 deutschen 12-jährigen Knaben wurde die mittlere Größe deutscher 12-jähriger Knaben zu 143.6 ± 0.15 cm bestimmt. Die Genauigkeit der Schätzung des unbekannten Populationsmittels ist 0.15 cm.
- σ_n in den extrem seltenen Fällen, wo man die Standardabweichung der Stichprobe selbst dokumentieren möchte, z.B. *unsere Testgruppe aus 12-jährigen Knaben hatte eine mittlere Größe von 147.8 ± 3.6 cm.* Hier bezieht sich die Standardabweichung **nur** auf die Personen der Testgruppe, ist also keine Schätzung für die Grundgesamtheit.

95%-Konfidenzintervall, in dem sich der wahre Mittelwert μ einer Grundgesamtheit mit 95% Wahrscheinlichkeit aufhält bei Schätzung aus einer Stichprobe des Umfangs n. t(FG,zws,α) ist der zweiseitige Sicherheitspunkt der t-Verteilung mit α=0.05.

$$\bar{\bar{x}} \pm \sigma_{\bar{x}} \cdot t(FG, zws., \alpha)$$

FG = n-1

Beispiel: 11 Drahtdicken in mm gemessen: 0,141 0,138 0,143 0,142 0,145
0,141 0,142 0,144 0,143 0,139 0,144

$\bar{x}$ =0,1420 mm	Arithmetisches Mittel, Schätzwert für μ in der Population
σ_{n-1}=0,00214 mm	Standardabweichung, Schätzwert für σ in der Population
$\sigma_{\bar{x}}$ =0,000645 mm	Fehler des Mittelwertes (bei n=11 Messungen)
0,1420 ± 2.23*0.000645	95%-Konfidenzintervall für das wahre Mittel μ mit t_α=2.23, zweiseitig und Freiheitsgrad FG=10.

Quantile oder Perzentile: Als Quantil X_P zur Wahrscheinlichkeit P bezeichnet man eine Zahl x auf der x-Achse, für die gilt, dass genau der Anteil P (bzw. P%) der Population kleinere Werte als X_P aufweist. Gibt man die Wahrscheinlichkeit in % an, spricht man von Perzentilen. Mit welcher Wahrscheinlichkeit P sind z.B. Zufallszahlen x kleiner als Quantil X_P, wenn x eine normalverteilte Zufallszahl mit Mittelwert $\bar{x}$ und Standardabweichung σ_{n-1} ist?
Berechne $u = (X_p - \bar{x})/\sigma_{n-1}$ und bestimme aus der Tafel Φ(u) das P.

Welches Quantil X_P gehört zu den P% unteren normalverteilten Werten einer Population?
P ist gegeben, suche in Φ(u) dazu den u-Wert. $X_P = \bar{x} + u \cdot \sigma_{n-1}$

Beachte, dass die Tafel Φ(u) manchmal nur für negative u vorliegt. Positive u ergeben Wahrscheinlichkeiten P>0,5. Wegen der Symmetrie der Normalverteilung gilt Φ(u)=1–Φ(–u)

Der **Median** einer Verteilung ist das 50%-Perzentil. Das 25%- und 75%-Perzentil werden auch **Quartile** genannt. Ihr Abstand auf der x-Achse heißt **Interquartilabstand**.

Indexzahlen: P sind Preise, g sind Gewichte (Stückzahlen z.B.) 0 indiziert das Basisjahr (Bezugsjahr), 1 indiziert das aktuelle Jahr, n ist die Anzahl der Produkte im Warenkorb.

Preisindex nach Paasche	**Mengenindex nach Laspeyres**
$I_P = \left(\sum_{i=1}^{n}(g_{1i}P_{1i})\right) / \left(\sum_{i=1}^{n}(g_{1i}P_{0i})\right)$	$I_Q = \left(\sum_{i=1}^{n}(g_{0i}P_{1i})\right) / \left(\sum_{i=1}^{n}(g_{0i}P_{0i})\right)$

7. Test von Hypothesen

Eine **wissenschaftliche Hypothese** ist eine Aussage über eine Grundgesamtheit. Beispiel: *Nach Einnahme unseres neu entwickelten ACE-Hemmers sinkt der Blutdruck von Hochdruckpatienten.* Gemeint ist die Grundgesamtheit aller Hochdruckpatienten. Überprüfen können wir eine solche Hypothese nur mit einer Stichprobe. Wir verallgemeinern die Ergebnisse einer (meist kleinen) Stichprobe auf die (zumeist große) Grundgesamtheit. Dabei können uns **Zufallsfehler** einen Schabernack spielen. Sie gaukeln uns eine Blutdruckabnahme vor, weil wir zufällig mehr Patienten ausgewählt hatten, bei denen unser neuer ACE-Hemmer eine Blutdruckerniedrigung bewirkt, als solche, bei denen nichts oder gar das Gegenteil eintritt. Um solche Fehler bewerten zu können, legt die konfirmatorische Statistik eine zulässige **Irrtumswahrscheinlichkeit** fest und prüft, ob sie durch die Ergebnisse der Stichprobe nicht überschritten wird. Fast alle Hypothesentests lassen sich in folgendes **Schema** pressen:

Schritt 0: **Festlegung einer Nullhypothese H_0**, die Effekte als zufällig abtut (z.B. nur Zufallsschwankungen des Blutdrucks). Dagegen steht die **Alternativhypothese H_A**, die einen signifikanten Effekt (z.B. eine Blutdrucksenkung) postuliert. Bezeichnen wir z.B. die Differenz $P_1 - P_2$ der Blutdruckmessungen vor und nach der Behandlung mit **d**, dann lauten die beiden Hypothesen bei **zweiseitiger Fragestellung** H_0: **d**=0 und H_A: **d**≠0. Der **Fehler 1. Art, α,** gibt die Wahrscheinlichkeit an, mit der wir eine richtige Nullhypothese H_0 ablehnen, d.h. uns für die falsche Hypothese H_A entscheiden. Übliche Wertevorgaben für α sind 0,05 bzw. 0,01 (5% bzw. 1%). Der selten kontrollierte **Fehler 2. Art, β,** gibt die Wahrscheinlichkeit, dass wir eine richtige Alternativhypothese H_A ablehnen. Hier ist man mit Werten von β=10-30% schon zufrieden.

Schritt 1: Auswahl einer geeigneten **Teststatistik (Methode).** Eine Teststatistik ist eine Größe, <u>die bei Gültigkeit von H_0</u> einer bekannten Verteilung folgt, z.B. der **u-, t-, χ^2- oder F-Verteilung.** (Es gibt weitere Testverteilungen.). Schritt 1 erfordert viel Erfahrung und Einfühlungsvermögen in die Statistik. Im Zweifelsfall konsultiere man einen erfahrenen Statistiker.

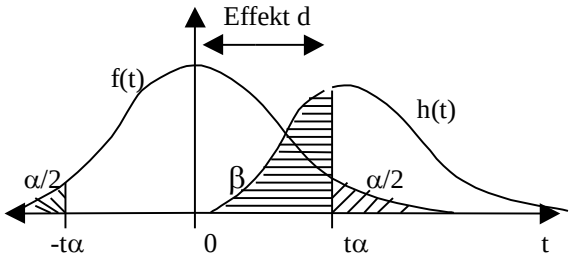
Schritt 2: **Berechne den Wert der Teststatistik**, z.B. **t** und eventuelle Freiheitsgrade aus den Daten der Stichprobe. Dabei kann es sein, dass man vorher eine Datentransformation ausführen muß, wenn die Datenverteilung nicht den Testanforderungen genügt, z.B. definitiv keine Normalverteilung der Daten vorliegt, wie sie der Test vielleicht verlangt.

Schritt 3: Entscheidung über die **Annahme oder Ablehnung von H_0**. Ist der absolute Wert der Teststatistik, z.B. $|\mathbf{t}|$, so groß, dass die Wahrscheinlichkeit seines Auftretens bei Annahme der Gültigkeit von H_0 auf den Wert α sinkt oder gar darunter, dann zweifelt man an der

Gültigkeit von H_0 und entscheidet sich für H_A. Extrem große Werte der Teststatistik sind nämlich bei Annahme von H_A sehr viel wahrscheinlicher als bei Annahme von H_0. Die zum α gehörigen **Sicherheitspunkte** der Testverteilungen, z.B. t_α, sind in Tabellen verfügbar. Sie markieren die Punkte z.B. auf der t-Achse, ab der die t-Werte bei Annahme von H_0 nur noch mit kleiner Wahrscheinlichkeit auftreten. So verläuft bei der zweiseitigen Fragestellung die Entscheidungsfindung auf den simplen Vergleich von z.B. $|\mathbf{t}|$ und t_α hinaus:

Nimm H_0, wenn $|\mathbf{t}| < t_\alpha$ ist,
Nimm H_A, wenn $|\mathbf{t}| \geq t_\alpha$ ist.

Den **Zusammenhang zwischen dem Fehler 1. Art, α, und dem Fehler 2. Art, β,** zeigt die folgende Graphik an einem Beispiel: Kurve f(t) ist die von Gosset gefundene t-Verteilung bei Gültigkeit von H_0. Kurve h(t) ist ein Beispiel für irgendeine meist unbekannt bleibende Verteilung der t-Werte bei Gültigkeit von H_A. (Diese Verteilung interessiert nicht wirklich.) Der Sicherheitspunkt bei zweiseitiger Fragestellung für f(t) ist t_α. Er tritt symmetrisch auf als $+t_\alpha$ und $-t_\alpha$. Jeder Zwickel der f(t)-Kurve hat Wahrscheinlichkeit $\alpha/2$, zusammen α.
Fall1: H_0 sei gültig, d.h. Kurve h(t) existiert nicht. f(t) ist die gültige Verteilung der t-Werte. Für einen aus der Stichprobe berechneten t-Wert mit $|\mathbf{t}| < t_\alpha$ nehmen wir H_0 zu recht an.
Fall2: H_0 sei gültig. Für einen aus der Stichprobe berechneten t-Wert mit $|\mathbf{t}| \geq t_\alpha$ lehnen wir H_0 zu unrecht ab. Wir realisieren den Fehler 1. Art, α.
Fall3: H_A sei gültig, d.h. Kurve h(t) ist jetzt die gültige Verteilung der t-Werte, die man aus Stichproben berechnet. Für einen berechneten t-Wert mit $|\mathbf{t}| < t_\alpha$ nehmen wir H_0 zu unrecht an. Wir realisieren den Fehler 2. Art, β.
Fall4: H_A sei gültig. Für einen berechneten t-Wert mit $|\mathbf{t}| > t_\alpha$ nehmen wir H_A zu recht an.



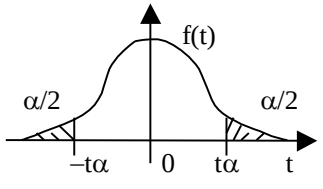
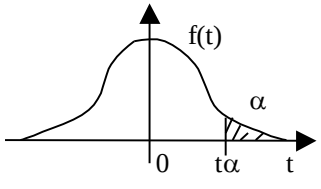
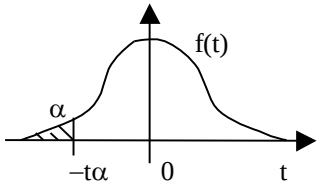
Ist der **Effekt** *d* klein, dann überlappen sich die beiden Verteilungen sehr stark und der Fehler 2. Art, β, wird immer größer. Man kann einen Effekt statistisch nur sichern, wenn er genügend groß ist. Allgemein gilt jedoch: **Großes α $\longleftrightarrow$ kleines β** und umgekehrt. Man muss den Kompromiss finden, was oft eine finanzielle Optimierungsaufgabe ist ($\rightarrow$ Versuchsplanung).

p-Wert (p-Value) eines Tests ist die Wahrscheinlichkeit für das Auftreten der Testgröße bzw. noch größerer Werte, alles unter der Annahme, dass H_0 gültig ist. Ein p-Wert ≤ 0.05 bedeutet Signifikanz auf dem 5%-Niveau, ein p-Wert ≤ 0.01 bedeutet Signifikanz auf dem 1%-Niveau, usw. **Trennschärfe (Power, Macht) eines Tests** ist definiert als **1-β,** d.h. die Wahrscheinlichkeit, eine richtige Alternativhypothese statistisch zu sichern. **Optimale Tests** haben maximale Trennschärfe, wenn die Voraussetzungen erfüllt sind (richtige Datenverteilung, ..., usw.).

- Die Trennschärfe steigt mit $\sqrt{n}$. Über das Stichproben-n kann bei festem α das β beliebig heruntergedrückt werden, falls genug Geld und Zeit da ist und tatsächlich ein Effekt existiert.
- Die Trennschärfe sinkt, wenn α heruntergesetzt wird, d.h., man sollte mit dem höchsten zulässigen α arbeiten (5% in der Biologie und Medizinforschung, 1% oder manchmal sogar 0.1% bei der Zulassung von Medikamenten).

- Die Trennschärfe steigt mit besserer Messmethodik (kleineren Varianzen in den Gruppen).
- Die Trennschärfe ist bei einseitiger Fragestellung besser (aber Vorsicht! Sie müssen die einseitige Hypothese gut begründen).

Zweiseitige und einseitige Fragestellung: Weiß man nichts über die Richtung des Effekts, dann ist immer die zweiseitige Fragestellung angebracht. Hat man jedoch **Vorwissen** aus früheren wissenschaftlichen Untersuchungen oder schreibt die Logik zwingend einen positiven oder einen negativen Effekt vor, dann darf man die Hypothesen einseitig aufstellen. Man wird durch kleinere Werte der Sicherheitspunkte belohnt, d.h. man erreicht leichter (mit weniger Probanden) eine signifikante Aussage.

Zweiseitige Fragestellung H_0: $\mu_1=\mu 2$, H_A: $\mu 1\neq\mu 2$	einseitig positive Fragestellung H_0: $\mu_1\leq\mu 2$, H_A: $\mu 1>\mu 2$	einseitig neg. Fragestellung H_0: $\mu_1\geq\mu 2$, H_A: $\mu 1<\mu 2$
		
α verteilt sich auf den linken und den rechten Zwickel. Entsprechend weit sind die Sicherheitspunkte t_α von der 0 entfernt	Der gesamte Fehler 1. Art, α, ist im rechten Zwickel zu finden. Entsprechend liegt der einseitige Sicherheitspunkt t_α näher an der Null	Der gesamte Fehler 1. Art, α, ist im linken Zwickel zu finden. Entsprechend liegt der einseitige Sicherheitspunkt $-t_\alpha$ näher an der Null

Nehmen Sie aber nur die einseitige Fragestellung, wenn Sie sie auch gut begründen können!

Freiheitsgrad: Der Begriff stammt aus der Mechanik und gibt dort die Zahl der möglichen Translations- und Rotationsbewegungen einer Ansammlung von Objekten an. In der Statistik ist es die Anzahl unabhängiger Werte, die in einer Quadratsumme stecken. Die Anzahl unabhängiger Werte ist $FG=N-N_p$. Dabei ist N die Anzahl der quadrierten Werte, N_p ist die Anzahl unabhängiger Stichprobenparameter, die in den quadrierten Daten stecken. Stichprobenparameter sind Parameter, die aus den Stichprobendaten selbst berechnet werden.

Beispiel Gesamt-χ^2 einer 4x2-Kontingenztafel	Beispiel Varianz aus n Messwerten
$\chi^2_{ij}=\frac{\left(n_{ij}-\hat{e}_{ij}\right)^2}{\hat{e}_{ij}}$, $\hat{e}_{ij}=\frac{n_{i.}n_{.j}}{n}$, $\chi^2=\sum_{i=1}^{I}\sum_{j=1}^{J}\chi^2_{ij}$	$\sigma_{n-1}=\sqrt{\frac{\sum\left(x_i-\bar{x}\right)^2}{n-1}}$
8 beobachtete unabhängige Häufigkeiten n_{ij}. Es gibt 5 benutzte unabhängige Parameter zur Berechnung der Erwartungswerte e_{ij} : Gesamtzahl n, die Zeilensumme $n_{1.}$ und die 3 Spaltensummen $n_{.1}$, $n_{.2}$, $n_{.3.}$.	n unabhängige Messwerte x_i liegen vor. Es gibt nur einen benutzten Parameter, der aus den Daten berechnet wird: $\bar{x}$
FG = 8–5 = 3	FG = n–1

Multiples Testen und **Alpha-Adjustierung**

Führt man an einer Stichprobe mehrere Tests durch bzw. berechnet mehrere Konfidenzintervalle oder Schätzwerte, von denen jeder die Irrtumswahrscheinlichkeit α hat, z.B. α=5%, dann findet man bei n=100 Tests etwa 5 signifikante Alternativen, auch wenn in Wirklichkeit überall die Nullhypothese gültig ist. Man spricht von ***Inkonsistenz***, wenn z.B. bei den 3 paarweisen Testen $\mu_1 = \mu_2$, $\mu_1 = \mu_3$, $\mu_2 = \mu_3$ immer die Nullhypothesen nicht abgelehnt wird, wohl aber beim Test $\mu_1 = \mu_2 = \mu_3$. Man spricht von einer ***Inflation des α-Fehlers***, wenn, wie im obigen Beispiel, das α des Einzeltest geringerr ist, als das α der Gesamthypothese bestehend aus den n Einzelhypothesen.

Wie geht man mit dem Problem um? Wir klammern hier in diesem Skript das Problem der Inkosistenz aus und betrachten nur die Inflation. Es gibt zwei Möglichkeiten:

1. Wir stellen nur unabhängige Einzelhypothesen auf. Es macht uns nichts aus, wenn einige falsch bewertet werden. Die große Masse ist richtig bewertet.
2. Wir fordern, die Gesamtheit aller unserer n Hypothesen wird als eine multiple Hypothese aufgefaßt und darf nur mit Irrtumswahrscheinlichkeit α falsch sein, d.h., selbst bei n=100 Hypothesentests darf die Wahrscheinlichkeit, dass auch nur eine Nullhypothese fälschlich abgelehnt wurde, nicht größer als α sein. Das bedeutet, wir müssen das α der Einzeltests anpassen (adjustieren).

Die **Bonferroni-Adjustierung** dividiert α durch die Hypothesenzahl *n*, d.h. $\alpha^* = \alpha / n$, und testet bei den Einzelhypothesen mit α^* statt mit α. **Holms sequentielle Prozedur** (Bonferroni-Holm-Adjustierung) berechnet zuerst die p-Werte für alle *n* Einzeltest, ordnet die p-Werte aufsteigend nach der Größe, vergleicht den kleinsten p-Wert mit $\alpha_1 = \alpha / n$, den nächstgrößeren mit $\alpha_2 = \alpha / (n-1)$, usw. bis zum größten p-Wert, der mit α verglichen wird. Ist jedoch ein p-Wert größer als sein α_i, dann ist dieser Test und alle nachfolgenden Tests nicht signifikant. Die Bonferroni-Adjustierung ist einfacher durchzuführen, liefert eventuell aber weniger Signifikanzen, als Holms Prozedur. Noch genauer testet man mit der **Sidak-Adjustierung**. Hier berechnet sich das adjustierte α^* nach der Formel $\alpha^* = 1-(1-\alpha)^{1/n}$. Es gibt noch weitere Möglichkeiten, mit dem Problem der Inflation umzugehen, wie z.B. die Tukey T-Methode, die Dunnett-Prozedur oder die Benjamini-Hochberg-Prozedur.

8. Test von Häufigkeitszahlen

8.1 Vergleich einer beobachteten relativen Häufigkeit mit einer Konstanten

Vergleich einer beobachteten relativen Häufigkeit $\hat{p}$ mit einer vorgegebenen konstanten Wahrscheinlichkeit p_o. p sei die „unbekannte" Wahrscheinlichkeit der Grundgesamtheit.

Schritt 0: Hypothese $H_o: p = p_o$ $H_A: p \neq p_o$ (zweiseitiger Test) α=0.05 (5%)
oder z.B. $H_A: p > p_o$(einseitiger >Test)

Schritt1 : Methode asymptotischer Binomial-Test: **u** ist unter H_0 asymptotisch normal verteilt

Schritt 2: Berechne $\hat{p}$ = h / n
h=Zahl der JA-Antworten,
n=Antworten insgesamt

$$u = \frac{\hat{p} - p_0}{\sqrt{p_o(1-p_0)}}\sqrt{n}$$

Schritt 3: Aussage: Die Sicherheitspunkte für $u(\alpha)$ sind identisch mit denen von $t(\alpha, FG \to \infty)$ bzw. mit denen der Standardnormalverteilung $\Phi(u)$.
Bei zweiseitigem Test und α=0.05 ist $u(\alpha)$=1,96, bei einseitigem Test ist $u(\alpha)$=1,65.

2-seitiger Test: Wenn $u > u(\alpha)$, dann ist signifikant $p > p_o$ → nimm H_A
Wenn $u < -u(\alpha)$, dann ist signifikant $p < p_o$ → nimm H_A
1-seitiger Test: Wenn $u > u(\alpha)$, dann ist signifikant $p > p_o$ → nimm H_A
In allen anderen Fällen wird H_o: $p = p_o$ angenommen.

Zahlenbeispiel: Die Biofirma *Laktozar* will in München eine Kampagne starten, wenn der Anteil von 20% Diätfreunden signifikant überschritten wird. Eine Umfrage unter 100 Personen ergab 23 JA-Stimmen für eine neue Diät.
$p=23/100=0.23$, $p_o=0.2$, $u= ((0.23-0.2)/(0.2*0.8)) * 100^{0.5} = 0.75$
$0.75 < 1.96$, d.h. wir akzeptieren H_0. Keine signifikante Abweichung vom Wert 20% wurde gefunden. Die Kampagne findet nicht statt.

8.2 Vergleich zweier relativer beobachteter Häufigkeiten

(genauer der Vergleich der geschätzten Wahrscheinlichkeiten **p_1 und p_2** in zwei Grundgesamtheiten). Gegeben sind 2 Stichproben mit Umfang n_1 bzw. n_2 und h_1 bzw. h_2 „JA-Antworten".
Schritt 0: Hypothese H_o: $p_1 = p_2$ H_A: $p_1 \neq p_2$ (zweiseitiger Test) $\alpha=0.05$ (5%)
oder z.B. H_A: $p_1 > p_2$ (einseitiger >Test)
Schritt 1 : Methode t-Test für Häufigkeiten

Schritt 2 : Berechne

$\hat{p}_1 = h_1 / n_1$	$\hat{p}_2 = h_2 / n_2$	FG = $n_1 + n_2 - 2$
$p = \frac{h_1 + h_2}{n_1 + n_2}$	q = 1 − p	$t = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{pq}} \sqrt{\frac{n_1 n_2}{n_1 + n_2}}$

Schritt3 : Aussage : Suche Sicherheitspunkt $t(\alpha, FG)$ aus der Tafel (beachte 1- oder 2-seitig)
2-seitiger Test: Wenn $t < -t(\alpha, FG)$, dann ist signifikant $p_1 < p_2$ → H_A
Wenn $t > t(\alpha, FG)$, dann ist signifikant $p_1 > p_2$ → H_A
1-seitiger Test : Wenn $t > t(\alpha, FG)$, dann ist signifikant $p_1 > p_2$ → H_A
in allen anderen Fällen H_o: $p_1 = p_2$ annehmen (kein signifikanter Unterschied)

Zahlenbeispiel: Die Biofirma *Laktozar* will ihre Kampagne für die neue Diät in Deutschland starten, wenn Frankreich nicht signifikant mehr Diätfreunde hat. Es wurden zwei Umfragen erhoben, eine in Deutschland, eine in Frankreich.

D: $h_1=127$ von $n_1=500$ Probanden waren für eine neue Diät
F: $h_2=64$ von $n_2=300$ Probanden waren für eine neue Diät
H_0: $p_1=p_2$, H_A:$p_1 \neq p_2$ (zweiseitige Fragestellung), $\alpha=0.05$, d.h. $t_\alpha=1.96$

$\hat{p}_1 = \frac{127}{500} = 0.254$	$\hat{p}_2 = \frac{64}{300} = 0.213$	FG=500+300-2=798
$p = \frac{127 + 64}{500 + 300} = 0.239$	q = 1–0.239 = 0.761	$t = \frac{0.254 - 0{,}213}{\sqrt{0.239 \cdot 0.761}} \sqrt{\frac{500 \cdot 300}{500 + 300}} = 1.316$

Wegen $t < t_\alpha$, d.h., $1.316 < 1.96$ akzeptieren wir H_0. Es besteht kein signifikanter Unterschied in der Zahl der Diätfreunde zwischen Deutschland und Frankreich. Die Kampagne wird in Deutschland gestartet.

9. Kontingenztafeln

Kontingenztafeln entstehen beim Auszählen von kategorialen Merkmalen. Die Merkmalszahl bestimmt die Dimension der Tafel (2 Merkmale ergeben z.B. eine Matrix aus Zeilen und Spalten, 3 Merkmale ein 3-dimensionales Zahlenfeld usw.). Beispiel: Befragung von 100 Probanden nach ihren Rauchgewohnheiten. Merkmal Geschlecht hat zwei Kategorien: weiblich / männlich. Merkmal Rauchgewohnheit hatte hier 3 Kategorien: nie / mäßig / stark. Die einzelne Häufigkeit n_{ij} (auch Frequenz genannt) heißt Konfiguration oder Zelle. Zellen werden durch die Indizes i, j, ... bezeichnet.

	rauche nie	mäßig	stark
w	$n_{11}=22$	$n_{12}=17$	$n_{13}=11$
m	$n_{21}=26$	$n_{22}=16$	$n_{23}=8$

Kontingenztafeln können zumeist als fertige Tafeln gelesen werden oder aus gelesenen Daten ausgezählt werden. Nur kategoriale Merkmale sind zur Auszählung geeignet. Man kann jedoch metrische Merkmale in kategoriale umwandeln (z.B. Transformation in ein dichotomes 0/1-Merkmal). Die Kategorien eines Merkmals müssen als Nummern 1, 2, 3, .. kodiert sein.

Was leistet die Kontingenztafelanalyse u.a.?

- Kontingenztest (Chi-Quadrat-Test auf Unabhängigkeit von kategorialen Merkmalen)
- Typensuche mit Konfigurationsfrequenzanalyse nach G. A. Lienert und N. Victor
- Chi-Quadrat-Zerlegung nach LANCASTER (Abhängigkeitsstruktur erkennen)
- Merkmalsselektion durch stufenweise Reduktion einer n-dimensionalen Tafel
- Analyse von 2x2-Tafeln (Zusammenhangs- bzw. Assoziationsmasse, Typensuche nach dem Zero-Order-Modell von A.v.Eye)

9.1 Kontingenztest oder Homogenitätstest

auf Zusammenhang oder Unabhängigkeit zweier kategorialer Merkmale. Gegeben ist eine **Kontingenztafel** für k≥2 kategoriale Merkmale. H_o beim Globaltest: Die Merkmale sind unabhängig - kein Zusammenhang. H_A beim Globaltest: Die Merkmale sind abhängig - es gibt einen Zusammenhang. Der Test erfolgt mit Chi-Quadrat und testet einseitig auf Überschreitung des oberen Sicherheitspunktes der Chi-Quadrat-Verteilung mit FG Freiheitsgraden.

Schritt 0: Hypothese H_o: „Kein Zusammenhang“, H_A: „Signifikanter Zusammenhang“
α=0,05 (5%)

Schritt 1: Methode Globaler χ^2-Test in Kontingenztafeln

Schritt 2: n_{ij} = Häufigkeit der Kategorienkombination (i,j) (Beispiel k=2)
$n_{i.}$ = Zeilensumme i $n_{.j}$ = Spaltensumme j
I = Zeilenzahl der Tafel J = Spaltenzahl der Tafel
n = Gesamtzahl aller Probanden (Fälle)
Berechne Freiheitsgrad, Erwartungswerte, χ^2-Komponenten und Gesamt-χ^2

FG = I J – (I–1) – (J–1) –1 $\qquad \hat{e}_{ij} = \frac{n_{i.} n_{.j}}{n}$

$$\chi_{ij}^2 = \frac{\left(n_{ij} - \hat{e}_{ij}\right)^2}{\hat{e}_{ij}} \qquad \chi^2 = \sum_{i=1}^{I}\sum_{j=1}^{J}\chi_{ij}^2$$

Schritt 3: Suche den Sicherheitspunkt $\chi^2(\alpha,FG)$. Wenn $\chi^2 \geq \chi^2(\alpha,FG)$, dann nimm einen signifikanten Zusammenhang der Merkmale (bzw. Kontingenz) an, sonst akzeptiere H_o: „Kein signifikanter Zusammenhang (bzw. keine Kontingenz)".

Beispiel: Aus einer Fragebogenaktion zum Trinkverhalten von Patienten ergab sich u. a. die Frage: Gibt es Unterschiede zwischen Männern und Frauen bezüglich der Wichtigkeit von Alkohol, Säften, Heißgetränken? (Trinktyp). Die Kontingenztafel lautet:

		Trinktyp		
		Alkohol	Säfte	Heißgetränke
Geschlecht	m	84	23	42
	w	27	82	54

Wir testen auf einen signifikanten Zusammenhang zwischen den Merkmalen Geschlecht und Trinktyp (Hypothesen, e_{ij}, χ_{ij}^2, χ_{Gesamt}^2, Hypothese wählen, Antwortsatz)

Lösung: e_{11}=53,0 e_{12}=50,1 …. χ^2_{11}=18,13 χ^2_{12}=14,66 … χ^2_{ges}=63,3

FG=2 χ^2_{alfa}=5,99 H_A Es besteht ein signifikanter Zusammenhang zwischen den Merkmalen Geschlecht und Trinktyp

9.2 Konfigurationsfrequenzanalyse (KFA) nach Lienert und Victor

Kontingenztafeln (oder Kreuzklassifikationstabellen) sind ein Mittel, um Beziehungen zwischen kategorialen Merkmalen zu analysieren. Die Konfigurationsfrequenzanalyse (KFA) nach G. A. Lienert (1969) hat sich hierbei als universelle Analysemethode fuer Kontingenztafeln bewährt.

Vereinfachter Test auf Kontingenztyp bzw. Antityp nach Krauth und Lienert
Gegeben ist eine Kontingenztafel (siehe oben). Machen Sie den Kontingenztest! Wenn er Signifikanz liefert, dann ist ein Kontingenztyp vorhanden, sonst nicht. Dieser kann jedoch Typ oder Antityp sein. Ein ***Typ ist signifikant überfrequentiert*** bezogen auf seinen Erwartungswert. Ein ***Antityp ist signifikant unterfrequentiert*** bezogen auf seinen Erwartungswert. Zum Auffinden des Kontingenztyps/-antityps im Falle einer signifikanten Kontingenz suchen Sie unter den Zellen die Zelle mit der größten χ^2-Komponente heraus. Diese Zelle ist der gesuchte Kontingenztyp bzw. Antityp nach Krauth und Lienert. Die χ^2-Komponenten werden für den Kontingenztest eh schon berechnet.

Zahlenbeispiele Typensuche: Per Videoauswertung in einer Großapotheke soll der Lienertsche Kontingenztyp bzw. Antityp gesucht werden, d.h. der Typ Kunde, dessen Auftreten überdurchschnittlich weit über oder unter dem Erwartungswert liegt. Bei der Auswertung wurden die beiden Merkmale Geschlecht (w, m) und Einkaufszeit (vorm., mitt., nachm., abends) erfasst. Die Kontingenztafel mit Zeilen- und Spaltensummen und den bereits berechneten Erwartungswerten $\hat{e}_{ij}$ und den berechneten χ^2-Werten ist:

	vorm.	mittags	nachm.	abends	Summe
w	n_{11}=132 $\hat{e}_{11}$=119.4 χ^2_{11}=**1. 32**	n_{12}=92 $\hat{e}_{12}$=105.6 χ^2_{12}=1.75	n_{13}=74 $\hat{e}_{13}$=58.2 χ^2_{13}=**4. 29**	n_{14}=179 $\hat{e}_{14}$=193.8 χ^2_{14}=1.13	$n_{1.}$=477
m	n_{21}=67 $\hat{e}_{21}$=79.6 χ^2_{21}=1.99	n_{22}=84 $\hat{e}_{22}$=70.4 χ^2_{22}=**2. 63**	n_{23}=23 $\hat{e}_{23}$=38.8 χ^2_{23}=6.43	n_{24}=144 $\hat{e}_{24}$=129.2 χ^2_{24}=**1.70**	$n_{2.}$=318
Summe	$n_{.1}$=199	$n_{.2}$=176	$n_{.3}$=97	$n_{.4}$=323	n=795

Z.B. ist $n_{1.}$ =132+92+74+179=477 oder $\hat{e}_{11}$ = 477*199 / 795 = 119.4
Bei den Zellen, bei denen die beobachtete Frequenz $n_{ij} > \hat{e}_{ij}$ ist, wurde der χ^2-Wert fett gedruckt. Diese Zellen sind Typkandidaten. Die anderen Zellen sind Antitypkandidaten.
Summe der χ^2-Werte ist: $\chi^2 = \chi^2_{11} + \chi^2_{12} + ... + \chi^2_{24}$ = 1.32 + 1.75 + ... +1.70 = 21.24
Zahl der Freiheitsgrade ist FG=2*4–1–3–1 = 3
Der Sicherheitspunkt für α=0.05 und 3 Freiheitsgrade ist χ^2_α = 7.81
Wir lehnen H_0 ab, da $\chi^2 > \chi^2_\alpha$ ist, d.h., 21.24>7.81.
Es existiert mindestens ein signifikanter Typ bzw. Antityp auf dem α=5%-Niveau.
Der höchste χ^2-Wert tritt mit 6.43 in Zelle (2,3) auf, einer Antityp-Zelle.
Resultat: Männliche Kunden kaufen überdurchschnittlich selten am Nachmittag ein, ein Grund, in dieser Zeit in dieser Apotheke Reklamevideos speziell für Frauen abzuspielen.

Vollständige Typensuche mittels KFA
Typ/Antityp nach G. A. Lienert: Ist die beobachtete Zellfrequenz n_{ijk} signifikant größer, als der Erwartungswert e_{ijk}, dann liegt ein Kontingenz-Typ vor. Bei signifikanter Unterschreitung ($n_{ijk} < e_{ijk}$) sprechen einige Autoren von einem Antityp. Die Definition von Antitypen ist jedoch umstritten. Typen nach Victor: Ist die beobachtete Zellfrequenz n_{ijk} signifikant größer, als ein speziell berechneter Erwartungswert V_{ijk} (Victor-Erwartungswert), dann liegt ein Typ nach Victor vor. Das V_{ijk} wird zwar auch aus den Randsummen geschätzt, aber vermindert um den Häufigkeitsüberhang der Typzellen. Das "Zuviel" an Häufigkeit bei einer Typzelle soll nicht in die Berechnung des Erwartungswertes eingehen. Das Problem ist nur iterativ zu lösen, da die Typzellen a priori nicht bekannt sind.

H_o beim lokalen Zelltest (Einzeltest): Die Zelle ist kein Typ - Abweichungen vom Erwartungswert unter der Unabhängigkeitshypothese sind zufällig. H_A beim Einzeltest: Die Zelle ist Typ oder Antityp - Abweichungen sind ursächlich und reproduzierbar vorhanden. Die Absicherung der multiplen Hypothese erfolgt immer mit Alpha-Adjustierung. Lokale Zelltests bewerten durch eine Testgöße (Teststatistik) den Abstand $n_{ijk} - e_{ijk}$ einer jeden Zelle. Z.B. berechnet der Chi-Quadrat-Komponententest die Testgrößen $\chi^2_{ijk}=(n_{ijk} - e_{ijk})^2 / e_{ijk}$ für jede Zelle (i,j,k). Der Freiheitsgrad der einzelnen Komponente ist (nach Perli u.a.) festgelegt auf FG=1. Ist der Abstand signifikant, liegt ein Typ (oder Antityp) vor.

Einseitiger Test: Man testet einseitig auf Typen, wenn $n_{ijk} > e_{ijk}$ (bzw. nach Victor $n_{ijk} > V_{ijk}$) ist. Man testet einseitig auf Antitypen, wenn $n_{ijk} < e_{ijk}$ bzw. $n_{ijk} < V_{ijk}$ ist. Die gesamte vorgegebene Irrtumswahrscheinlichkeit α wird einseitig angenommen und verringert so die erforderliche Testgröße, die einen signifikanten Typ anzeigt. Mögliche Begründung: Eine deutliche Abweichung der Frequenz n_{ijk} von ihrem Erwartungswert in die eine oder andere Richtung ist im Bayes'schen Sinne eine Vorinformation, die die den einseitigen Test rechtfertigt**.**
Zweiseitiger Test: Man läßt für jede Zelle beide alternativen Hypothesen (Typ oder Antityp) offen. α wird zu gleichen Teilen auf Typ und Antityp verteilt. Die Signifikanzschwelle liegt höher, als beim einseitigen Test. Der zweiseitige Test bedarf keiner Begründung.

Stetigkeitskorrektur: Kleine Zellwahrscheinlichkeiten (e_{ijk} <5) führen leicht auf antikonservative Ergebnisse. Man reduziert die Testgröße gezielt (was sich bei kleinen Frequenzen besonders auswirkt) und beugt so Irrtümern vor. Nach einer Studie aus dem Jahre 2001 (Lautsch / Weber) zeigt jeder Test entweder antikonservatives (liefert zu viele Typen) oder konservatives Verhalten (liefert zu wenig Typen). Eine für die vorliegende Tafel zugeschnittene Korrekturkonstante ***K*** sorgt z. B. in DASY dafür, dass asymptotisch das vorgegebene α eingehalten, aber auch ausgeschöpft wird. Nur so kann auch das β minimiert werden.

Nach einer Untersuchung von v.Eye, Lautsch und v.Weber (2004) werden folgende lokale Zelltests in dieser Reihenfolge empfohlen:

1. Combinatoric Search nach Dunkl, von Eye, Lautsch, Victor, von Weber
2. Gradientenverfahren von Lautsch und von Weber, falls die Rechenzeiten der Combinatoric Search zu lang werden.
3. Chi-Quadrat-Test nach Lienert (bei 2^d-Tafeln und FG>20)
4. Alternativ zu 3. der asymptotische Test nach Perli et al.. (bei 2^d-Tafeln und FG>20)

Zahlenbeispiel KFA: Die berühmten LSD-Daten von G.A.Lienert aus dem Jahre 1970 zeigen das psychotoxische Syndrom, das Leuner 1962 bereits beschrieben hat. 65 Studenten nahmen freiwillig Lysergsäurediethylamid (LSD) ein und unterzogen sich, soweit noch fähig, verschiedenen Tests. Das Leuner'sche Syndrom ist eine Kombination aus

M01 = Bewusstseinseinschränkung (clouded consciouness)
M02 = Denkstörung (disturbed thinking)
M03 = Affektivitätsbeeinflussung (altered affectivity)

```
.
Die Typensuche mit der Combinatoric Search:
65 Probanden,    8 Zellen,    mBl=    8.13 mittlere Belegung
     37.92 Chi-Quadrat-Gesamt mit FG=4
 6.346E-06 (***) einseitige Irrtumswahrscheinlichkeit
      6.00 geschaetztes maximales Typgewicht
Test:  Combinatoric Search (Weber et al.)     Zweiseitig
Geschaetzter Korrekturwert K =  -1.11
Geschaetztes Beta            =  34.03 %
Sie arbeiten mit Alpha       =   0.05
.
Nr.  i j k l m   Nijk     Eijk      Vijk      koTw    KIW    T/AT Signif
.
001  1 1 1 . .     20     12.51      0.69     4.47   0.00000   1   ***
002  1 1 2 . .      1      6.85      2.12    -0.26   0.39743   0
003  1 2 1 . .      4     11.40      3.65     0.08   0.47006   0
004  1 2 2 . .     12      6.24     11.24     0.10   0.45887   0
005  2 1 1 . .      3      9.46      2.92     0.02   0.49232   0
006  2 1 2 . .     10      5.18      8.97     0.15   0.43883   0
007  2 2 1 . .     15      8.63     15.44    -0.05   0.47935   0
008  2 2 2 . .      0      4.73     47.51    -3.24   0.00060  -1   ***
```

```
ijklm sind Zellindizes
Eijk  Unabhaengigkeits-Erwartungswerte aus den Randsummen berechnet
Vijk  VICTOR-Erwartungswerte bei der kombinatorischen Suche und
      beim Gradientenverfahren. Sonst ist Eijk=Vijk gesetzt.
      Ein Eijk bzw. Vijk kleiner 3 wird im Test auf 3 hochgesetzt
koTw  Testwerte mit Stetigkeitskorrektur nach Lautsch und v. Weber
KIW   Einseitige Irrtumswahrscheinlichkeiten zum Testwert koTw
A/AT  Eine 1 bedeutet Typ, eine -1 Antityp, eine 0 weder/noch.
      * bedeutet ein KIW ≤ 0.05, ** ≤ 0.01, *** ≤ 0.001
```

Die Combinatoric Search findet den Typen (1,1,1) und den Antitypen (2,2,2). Überraschend groß ist der Victorerwartungswert V_{ijk}=47.51 zum Antitypen (2,2,2) und der kleine Wert Vijk=0.69 zum Typen (1,1,1). Die Summe der V_{ijk} muss nicht die Probandenzahl N=65 ergeben, wie wir von der Summe der E_{ijk} gewohnt sind. Die 6 Zellen 002-007 definieren ein mittleres LSD-Wirkungsniveau. Man sieht, dass die Victorerwartungswerte sehr genau den gefundenen Frequenzen n_{ijk} entsprechen. Zelle 001 ist ein Ausreißer in dem Sinne, dass sich die Wirkungen der Droge bei diesem Probandentyp extrem verstärken, so dass keine normale Reaktion mehr erkennbar ist. Zelle 008 ist ein Ausreißer in dem Sinne, dass eigentlich viel mehr Probanden mit völlig unbeeinflussten Reaktionen erwartet werden. Die beiden Ausreißer zeigen, dass die LSD-Wirkung keinem log-linearem Modell folgt.

9.3 χ^2-Zerlegung nach Lancaster

Krauth und Lienert schlugen 1973 im Zusammenhang mit der Assoziations- und Kontingenzstrukturanalyse fuer 2^d-Tafeln vor, die Chi-Quadrat-Zerlegung nach Lancaster (1951) für den Nachweis von **Wechselwirkungen** zwischen den Merkmalen (Symptomen) zu nutzen. Das Chi-Quadrat einer Zweier-KFA wird als Maß der Wechselwirkung (Interaktion) 1. Ordnung angenommen, d.h. als Maß für die Abweichung vom Grundmodell der KFA (Wechselwirkung 0. Ordnung). Indizieren wir die Chi-Quadrat-Anteile mit kleinen Buchstaben, dann ergeben sich für die 3 Variablen A, B, C und Wechselwirkung 1. Ordnung die formalen Identitäten

$$\chi^2\ ab = \chi^2\ (A,B) \quad \chi^2\ ac = \chi^2\ (A,C) \qquad \chi^2\ bc = \chi^2\ (B,C)$$

Für 4-Felder-Tafeln mit den Feldern $\begin{bmatrix} a & b \\ c & d \end{bmatrix}$ gibt es eine **kompakte Näherungsformel** zur Berechnung des χ^2, die man gut für die 2-er-Beziehungen (Wechselwirkungen 1. Ordnung) zwischen zwei Merkmalen verwenden kann.

$$\chi^2 = \frac{(a\,d - b\,c)^{\,2}\,n}{(a+b)(c+d)(a+c)(b+d)} \quad \text{mit} \quad FG=1.$$

Bei fehlender Wechselwirkung 2. Ordnung lässt sich nach Lancaster das Gesamt- χ^2 der Dreier-KFA aus den Wechselwirkungen 1. Ordnung erklären. Verbleibt ein Residuum, wird es als Wechselwirkung 2. Ordnung gedeutet (Zusammenwirken von 3 Merkmalen):

$$\chi^2\ abc = \chi^2\ (A,B,C) - \chi^2\ ab - \chi^2\ ac - \chi^2\ bc.$$

Jedem χ^2 -Anteil ordnen wir einen Freiheitsgrad zu. Liegt als Ausgangstafel keine 2^d-Tafel vor, dann sind mit den Mitteln der KFA die interessanten Konfigurationen (Typen) zu ermitteln. Die Zusammenhangsstruktur eines speziellen Typs kann dann mittels einer kollabierten 2^d-Tafel in der oben genannten Weise zerlegt werden.

Zahlenbeispiel: Die berühmten LSD-Daten von G. A. Lienert aus dem Jahre 1970 (siehe Tabelle unten) zeigen das psychotoxische Syndrom, das Leuner 1962 bereits beschrieben hat. 65 Studenten nahmen freiwillig Lysergsäurediethylamid (LSD) ein und unterzogen sich, soweit noch fähig, verschiedenen Tests. Das Leuner'sche Syndrom ist eine Kombination aus

i = Bewußtseinseinschränkung (B) (clouded consciouness) mit den Ausprägungen 1 und 2
j = Denkstörung (D) (disturbed thinking) mit den Ausprägungen 1 und 2
k = Affektivitätsbeeinflussung (A) (altered affectivity) mit den Ausprägungen 1 und 2

Ausprägung 1 heißt hier stark ausgeprägt, Ausprägung 2 heißt nicht oder schwach ausgeprägt.

Zelle	i	j	k	n_{ijk}
001	1	1	1	20
002	1	1	2	1
003	1	2	1	4
004	1	2	2	12
005	2	1	1	3
006	2	1	2	10
007	2	2	1	15
008	2	2	2	0

Zuerst bilden wir die 6 Randsummen, die paarweise immer die Gesamtanzahl n = 65 ergeben:

$n_{1..}$ = 20 +1+4+12 = 37		(Addiere alle n_{ijk} mit Ausprägung i=1)
$n_{2..}$ = 3 + 10+15+ 0 = 28		(Addiere alle n_{ijk} mit Ausprägung i=2)
$n_{.1.}$ = 20 +1 +3+10 = 34		(Addiere alle n_{ijk} mit Ausprägung j=1)
$n_{.2.}$ = 4 +12 +15+0 = 31		(Addiere alle n_{ijk} mit Ausprägung j=2)
$n_{..1}$ = 20 + 4+ 3+ 15 = 42		(Addiere alle n_{ijk} mit Ausprägung k=1)
$n_{..2}$ = 1+ 12+ 10+ 0 = 23		(Addiere alle n_{ijk} mit Ausprägung k=1)

Wir vervollständigen die Tabelle der LSD-Daten mit den Erwartungswerten e_{ijk} und den χ^2-Komponenten χ^2_{ijk}. Die Formeln für die e_{ijk} und die χ^2_{ijk} sind:

$$\hat{e}_{ijk} = \frac{n_{i..}\, n_{.j.}\, n_{..k}}{n^2} \qquad \text{z.B.} \quad \hat{e}_{111} = \frac{n_{1..}\, n_{.1.}\, n_{..1}}{n^2} = \frac{37 \cdot 34 \cdot 42}{65^2} = 12{,}51$$

$$\chi^2_{ijk} = \frac{\left(n_{ijk} - \hat{e}_{ijk}\right)^2}{\hat{e}_{ijk}} \qquad \text{z.B.} \quad \chi^2_{111} = \frac{\left(n_{111} - \hat{e}_{111}\right)^2}{\hat{e}_{111}} = \frac{\left(20 - 12{,}51\right)^2}{12{,}51} = 4{,}48$$

Zelle	i	j	k	n_{ijk}	e_{ijk}	χ^2_{ijk}
001	1	1	1	20	12,51	4,48
002	1	1	2	1	6,85	5,00
003	1	2	1	4	11,40	4,80
004	1	2	2	12	6,24	5,32
005	2	1	1	3	9,46	4,41
006	2	1	2	10	5,18	4,49
007	2	2	1	15	8,63	4,70

```
008     2 2 2       0       4,73       4,73
```

Die Summe der χ^2_{ijk} ergibt das χ^2 (B,D,A) = $\chi^2_{Gesamt} = 37{,}93$ mit FG = 2·2·2 -1 -1 -1 -1 = 4.

Wir bilden die drei 2x2-Untertafeln (i, j), (i, k) und (j, k) und bestimmen jeweils das χ^2 jeder dieser 4-Felder-Tafeln als Maß für die Wechselwirkung 1. Ordnung. Eine Untertafel, z.B. (i, j), entsteht, wenn wir über den fehlenden Index summieren, z.B. Feld a=n_{11} ergibt sich aus allen Frequenzen n_{ijk} mit Index i=1 und j=1, Feld b=n_{12} aus allen Frequenzen mit Index i=1 und j=2 usw. Die Gesamtanzahl n mit n=65 bleibt für alle Untertafeln dieselbe wie bei der originalen 2x2x2-Tafel.

Untertafel (i, j) bzw. (B,D), d.h. die Tafel (Bewußtseinseintrübung / Denkstörung) mit χ^2 (B,D) = 0,682 nach der Näherungsformel und FG=1:

	j = 1	j = 2
i = 1	a = n_{11}= 20+1 = **21**	b = n_{12} = 4+12= **16**
i = 2	c = n_{21}= 3+10= **13**	d = n_{22} = 15+0= **15**

Untertafel (i, k) bzw. (B,A), d.h. die Tafel (Bewußtseinseintrübung / Affektivitätsbeeinflussung) mit χ^2 (B,A) = 0,002 nach der Näherungsformel und FG=1:

	k = 1	k = 2
i = 1	a = n_{11}= 20+4 = **24**	b = n_{12} = 1+12= **13**
i = 2	c = n_{21}= 3+15= **18**	d = n_{22} = 10+0= **10**

Untertafel (j, k) bzw. (D,A), d.h. die Tafel (Denkstörung / Affektivitätsbeeinflussung) mit χ^2 (D,A) = 0,287 nach der Näherungsformel und FG=1:

	k = 1	k = 2
j = 1	a = n_{11}= 20+3 = **23**	b = n_{12} = 1+10= **11**
j = 2	c = n_{21}= 4+15= **19**	d = n_{22} = 12+0= **12**

Berechnung der Wechselwirkung 2. Ordnung χ^2 bda:

χ^2 bda = χ^2 (B,D,A) - χ^2 (B,D) - χ^2 (B,A) - χ^2 (D,A) = 37,93 – 0,69 – 0,00 – 0,29 = **36,95** mit FG =4 -1 -1 -1 = 1.

Bestimmung der Signifikanzen der einzelnen Wechselwirkungen (Assoziationen):
Sicherheitspunkt der χ^2-Verteilung für FG=1 und α=5% ist χ^2_α = 3,84.

Ho: Es besteht keine signifikante Assoziation
H_A: Es besteht eine signifikante Assoziation.
Wir akzeptieren Ho, wenn $\chi^2 < \chi^2_\alpha$ ist.
Wir akzeptieren H_A, wenn $\chi^2 \geq \chi^2_\alpha$ ist.

Für die LSD-Daten von Krauth und Lienert mit den Merkmalen B= Bewußtseinseintrübung, D=Denkstörung, A=Affektivitätsbeeinflussung, ergeben sich zusammengefasst folgende χ^2, Freiheitsgrade und χ^2 -Anteile, d.h. Wechselwirkungsmaße, und Signifikanzen:

χ^2-Tafel bzw. -Untertafel	FG	χ^2 –Anteil	FG	Signifikanz des χ^2 –Anteils
χ^2 (B,D,A) = 37,93	4	χ^2 bda = 36,95	1	***
χ^2 (B,D) = 0,682	1	χ^2 bd = 0,68	1	---
χ^2 (B,A) = 0,002	1	χ^2 ba = 0,00	1	---

χ^2 (D,A) = 0,287 1 χ^2 da = 0,29 1 ---

Wechselwirkung χ^2 bda ist sogar auf dem 0,1%-Niveau (α=0,001) signifikant, nicht nur auf dem Niveau α=0,05, so dass Wechselwirkungen 2. Ordnung der Form BDA mit großer Sicherheit angenommen werden. Alle 2-er-Beziehungen sind hier nicht signifikant.

9.4 Merkmalsselektion - Suche der signifikantesten Tafeln

Hat man Rohdaten mit mehr als 5 kategorialen Merkmalen, dann ist oft eine Merkmalsauswahl notwendig, um z.B. lokale Tests oder eine Chi-Quadrat-Zerlegung überhaupt ausführen zu können. Wie findet man informationsträchtige Merkmalsgruppen? Eine Möglichkeit ist die Bewertung durch das Chi-Quadrat der Kontingenztafel. Es wird probeweise abwechselnd immer ein Merkmal entfernt und mit den verbliebenen Merkmalen die Kontingenztafel, Chi-Quadrat-Gesamt, Freiheitsgrad und die Wahrscheinlichkeit P berechnet. Dasjenige Merkmal wird selektiert und aus der Menge entfernt, dessen Entfernung die Wahrscheinlichkeit P der reduzierten Tafel am wenigsten erhöht. Ziel ist eine Merkmalsmenge, die auf eine Kontingenztafel mit möglichst kleiner Wahrscheinlichkeit führt. Die Selektionsschritte werden solange wiederholt, bis nur noch zwei Merkmale in der Menge verblieben sind.

9. 5 2x2-Tafeln: Zusammenhangsmaße, Typensuche

	Nichtraucher	Raucher	Zeilensumme
weiblich	a= 37	b= 13	a+b= 50
männlich	c= 29	d= 21	c+d= 50
Spaltensumme	a+c= 66	b+d= 34	Gesamt N=100

2x2-Tafeln entstehen bei der Auswertung zweier binärer Merkmale. Zumeist wird die Frage nach der Unabhängigkeit der Merkmale oder ihrem Zusammenhang gestellt, selten die Frage nach einem Typ. Beispiel: Merkmal M1 ist die Rauchgewohnheit (Nichtraucher/Raucher), Merkmal M2 ist das Geschlecht (weiblich /männlich)
Die vier Zahlen a=N_{11}=37, b=N_{12}=13, c=N_{21}=29, d=N_{22}=21 heißen Häufigkeit oder Zellfrequenz. Die Randsummen Ni. und N.j und das Gesamt-N werden daraus berechnet. Aus den Randsummen und Gesamt-N werden die Zellerwartungswerte E_{ij} berechnet. Je nach Modell sind die Schätzungen unterschiedlich: Das Kontingenzmodell nimmt Formel E_{ij} = Ni.N.j / N, dagegen A. v. Eyes Clustermodell (KFA 0. Ordnung) nimmt $E_{ij}=P_i P_j$ (P_i, P_j sind hier gegebene Wahrscheinlichkeiten). Das χ^2 obiger 4-Felder-Tafel ist übrigens χ^2=2,852 mit FG=1, d.h. auf dem 5%-Niveau liegt wegen Sicherheitspunkt $\chi^2_{\alpha=5\%,\ FG=1}$ = 3,84 keine Signifikanz vor.

Nach Untersuchungen aus dem Jahre 2003 (A. v. Eye, Lautsch und v. Weber gilt: Es gibt 6 besonders empfehlenswerte Zusammenhangsmasse (Kontingenzmaße):

Friedrich Vogel's Zv (F. Vogel, Bamberg)	Normalapproximiertes Z
log-odds ratio θ Teta	log-linear interaction λ (Lambda)
Goodman's Lambda λ_{Good}	Binomialtest Kr (J. Krauth, Düsseldorf)

Zusammenhangsmaße haben zwei Aufgaben:

1. Den Zusammenhang zwischen zwei Merkmalen messbar und vergleichbar zu machen, wie es der Korrelationskoeffizient bei metrischen Merkmalen leistet. Leider erfüllen die

Maße diese Bedingung nur teilweise, da sie nicht auf das Intervall [−1; +1] normiert sind oder aber dieses Intervall nicht ausschöpfen können.

2. Als Teststatistik einen Test auf "Unabhängigkeit der Merkmale" zu gestatten. Diese Aufgabe erfüllen alle oben genannten Maße recht gut.

Rechenbeispiel zu Vogel's Zv: Eines der besten Kontingenzmaße (Assoziationsmaße).

$$Zv = \frac{d(U,G)}{d(U,M)} \quad \text{mit} \quad d(U,G) = \sum_{ij} \left| \hat{e}_{ij} - n_{ij} \right| \quad \text{und} \quad d(U,M) = \sum_{ij} \left| \hat{e}_{ij} - m_{ij} \right|$$

Dabei sind:
n_{ij} = beobachtete Frequenzen
m_{ij} = maximal-minimal-Werte der Tafel bei Einhaltung der Randsummen
Das Vorzeichen von Zv wird durch die Determinante der 2x2-Tafel festgelegt. Ist D=ad-bc>0, dann ist auch Zv>0.

Gegeben ist eine beobachtete 4-Felder-Tafel a, b, c, d :

	Raucher	Nichtraucher	Summe
Hochdruck	a = 49	b = 10	59
Kein Hochdr.	c = 20	d =39	59
Summe	69	49	118

Wir bilden die maximal-minimale Tafel. Der kleinste Wert wird 0 gesetzt (hier wird der Wert 10 auf 0 gesetzt). Unter Einhaltung der Randsummen ergibt sich dann automatisch:

	Raucher	Nichtraucher	Summe
Hochdruck	59	0	59
Kein Hochdr.	10	49	59
Summe	69	49	118

Mit den Erwartungswerten e_{ij}: z.B. $e_{11} = 59 \cdot 69/118 = 34{,}5$ oder $e_{12} = 59 \cdot 49/118 = 24{,}5$ ist:

	Raucher	Nichtraucher
Hochdruck	34,5	24,5
Kein Hochdr.	34,5	24,5

Damit wird d(U,G)= | 49-34,5 | + | 10-24,5 | +…= 58
Und d(U,M)= | 59-34,5 | + | 0-24,5 | +…= 98
Und der Determinante D =ad−bc= 59·49 − 10·20 = 2691 > 0.

Damit wird Zv = 58/98 = +0,59.

Die Sicherheitspunkte für Zv lassen sich nur durch Simulation vieler Tafeln mit diesen Randsummen ermitteln. Das ist umständlich und nur mit einem PC und entsprechender Software möglich. Deshalb benutzen wir den Chi-Quadrat-Kontingenztest für die Originaltafel. Die Testgröße χ^2 ist bei einer 4-Felder-Tafel

$$\chi^2 = \frac{(ad-bc)^2 n}{(a+b)(c+d)(a+c)(b+d)} \text{ mit FG=1.}$$

Die Tafelwerte a, b, c, d und n = a+b+c+d eingesetzt in die Formel ergeben $\chi^2 = 29{,}35$. Der Sicherheitspunkt der χ^2-Verteilung für FG=1 und α=5% ist $\chi^2_\alpha = 3{,}84$.

Ho: Es besteht keine signifikante Assoziation zwischen Blutdruck und Rauchgewohnheit.
H_A: Es besteht eine signifikante Assoziation zwischen Blutdruck und Rauchgewohnheit.

Wir akzeptieren Ho, wenn $\chi^2 < \chi^2_\alpha$ ist.
Wir akzeptieren H_A, wenn $\chi^2 \geq \chi^2_\alpha$ ist.

Hier ist $\chi^2 \geq \chi^2_\alpha$. Wir akzeptieren H_A und postulieren eine signifikante Assoziation zwischen Blutdruck und Rauchgewohnheit.

Odds-Ratio (kurz **OR)**, auch Chancenverhältnis, Quotenverhältnis, oder relative Chance ist gleichzeitig ein Assoziationsmaß und ein Chancenverhältnis. Das Chancenverhältnis wird in der Medizin benutzt, um zu beziffern, wie stark ein Risikofaktor mit einer bestimmten Erkrankung zusammenhängt. Gegeben sind die 4 Felder a, b, c, d einer 4-Feldertafel. Dann ist ***OR= (a/c) / (b/d) = ad / (bc).***

Beispiel: Risiko eines Herzinfarkts bei belasteten / unbelasteten Diabetikern (fiktive Zahlen)

	Raucher und Adipositas	Nichtraucher und Normalgewicht
Herzinfarkt	127	87
kein Herzinfarkt	1433	6423

OR=127·6423 / (87·1433) = 6,54.
Die Chance als übergewichtiger rauchender Diabetiker einen Herzinfarkt zu bekommen, ist 6,54 mal höher, als bei einem normalgewichtigen nichtrauchenden Diabetiker.

Nach George UdnyYule lässt sich das OR auf das Intervall [-1, +1] abbilden und erfüllt so eine Bedingung eines guten Assoziationsmaßes. Es gilt ***Q=(OR-1)/(OR+1).***
Bei unserem obigen Beispiel ergibt sich Q=0,735. Man kann diesen Wert als positiven, schon recht ausgeprägten Zusammenhang zwischen dem Merkmale *Herzinfarkt (Ja/Nein)* und dem Merkmal *Rauchgewohnheit und Körpergewicht* ansehen. Einen Signifikanztest macht man jedoch besser mit dem Chiquadrat gemäß Formel

$$\chi^2 = \frac{(ad-bc)^2 n}{(a+b)(c+d)(a+c)(b+d)} \text{ mit FG=1.}$$

Bei unserem Beispiel wäre das Ergebnis χ^2=225,72 mit einem p-value von p<0,000001, d.h. einem sehr signifikantem Zusammenhang.

Typensuche in 2x2-Tafeln: Die Suche von Kontingenztypen in 2x2-Tafeln ist sinnlos. Ein Fehler 1. Art Alpha < 0.5 (Alpha<50%) ist nicht realisierbar. Beim Kontingenzmodell werden die Zellwahrscheinlichkeiten aus den Randsummen geschätzt. Man verliert dadurch 2 Freiheitsgrade und hat nur noch eine unabhängige Hypothese. Das sieht beim "Zero-Order-

Modell" von A. von Eye, auch Clustermodell genannt oder **KFA 0. Ordnung**, anders aus. Das Zero-Order-Modell setzt feste Wahrscheinlichkeiten von allen 4 Zellen als Basismodell an (z.B. Pij=0.25 für alle 4 Zellen bei Gleichverteilung beider Merkmale). Man hat bei Festlegung des Gesamt-N drei Freiheitsgrade, also 2 mehr als beim Kontingenz- oder Unabhängigkeitsmodell, und damit bei Vorhandensein nur einer Typzelle mehr "redundante" Information, die eine bessere Identifikation der einen Typzelle erlaubt.

Beispiel: Eine Studie über das Rauchverhalten der Geschlechter setze Gleichverteilung der Geschlechter im beobachteten Erhebungsgebiet (50% weiblich, 50% männlich) und ein Verhältnis von 40% Raucher zu 60% Nichtraucher an. Diese Zahlen (40%, 60%) könnten z.B. aus einer anderen Erhebung für das gesamte Land stammen, während die aktuelle Kontingenztafel als Erhebungsgebiet etwa eine Firma, eine Hochschule, eine Region umfasst. Die 4 Zellwahrscheinlichkeiten sind dann die Produkte der vorgegebenen Randwahrscheinlichkeiten, d.h. 0.5*0.4, 0.5*0.6. Was man testet, ist eine Abweichung vom Landesdurchschnitt, d.h., ein
gefundener Typ hat lokale Bedeutung, etwa im Sinne "Hier an der FH rauchen mehr Männer als im Landesdurchschnitt". Lokaler Cluster-Typ wäre demnach der "rauchende männliche FH-Student an der FH XYZ". Der Anwender kann unter 3 Typentests wählen:

1. Kleingruppentest nach von Eye / Dunkl
2. Chi-Komponententest nach G. A. Lienert
3. Der exakte Binomialtest nach J. Krauth

10. χ^2-Anpassungstest für eine Verteilung

Parametrische Analyseverfahren (wie Varianzanalyse, t-Test oder Regression) setzen voraus, dass die Vorhersagefehler (Residuen) normalverteilt sind. Ein **Anpassungstest** (*goodness-of-fit test*) ist ein nichtparametrischer Hypothesentest, der die Wahrscheinlichkeitsverteilung einer Zufallsvariablen auf eine bestimmte Verteilung prüfen soll. Der Test liefert jedoch **keinen Beweis für die Verteilung**, sondern erhärtet nur eine Vermutung, welche Verteilung zutreffen könnte. Der Stichprobenumfang *n* spielt hier eine wichtige Rolle.

Die bekanntesten Anpassungstests (eine Auswahl aus ca. 40 Testverfahren) sind:

- Chi-Quadrat-Anpassungstest (kleiner bis mittlerer Aufwand, ab etwa n>20)
- Kolmogorow-Smirnow-Anpassungstest (kleiner bis mittlerer Aufwand, ab etwa n>10)
- Shapiro-Wilk-Test (auch als Ryan-Joiner-Test bekannt) (großer mathematischer Aufwand, ab n>2 durchführbar, ist in vielen Softwarepaketen vorhanden)
- Anderson-Darling-Anpassungstest (kleiner bis mittlerer Aufwand, ab n>7)
- Jarque-Bera-Test (kleiner bis mittlerer Aufwand, ab etwa n>10)

Wir benutzen hier eine spezielle Form des Chi-Quadrat-Anpassungstest auf Normalverteilung, die nur bei wesentlichen Abweichungen von der Normalverteilung die Nullhypothese ablehnt.

Test auf Normalverteilung in der Grundgesamtheit: Gegeben ist eine Stichprobe mit den *n* Werten $x_1, x_2, \ldots, x_n$. Die Wertezahl *n* sollte mindestens 20 betragen, da ansonsten dieser Test fragwürdig wird. Man will prüfen, ob die Daten als normal verteilt angesehen werden können.

Schritt 0: Hypothese H_o: „Es spricht nichts gegen eine Normalverteilung der Grundgesamtheit“

Hypothese H_A: „Die Daten der Grundgesamtheit sind signif. nicht normal verteilt“
Schritt 1: Methode χ^2-Test von Klassenhäufigkeiten mit 5 gleichwahrscheinlichen Klassen.
Schritt 2: Berechne $\bar{x}$, σ_{n-1}, n der Stichprobe, bilde k=5 gleichwahrscheinliche Klassen symmetrisch zum Mittelwert gemäß der unten stehenden Tabelle, zähle die Häufigkeiten h_i der Werte in den Klassen aus, berechne die Klassenerwartungswerte $\hat{e}_i$ und die χ^2-Komponenten χ^2_i der Klassen. (σ_{n-1} wird in der Tabelle unten kurz σ genannt).

Klasse	uG_i	oG_i	h_i	$\hat{e}_i$	χ^2_i
1	$-\infty$	$\bar{x}$ -0,84 σ	h_1	n/5	χ^2_1
2	$\bar{x}$ -0,84 σ	$\bar{x}$ -0,25 σ	h_2	n/5	χ^2_2
3	$\bar{x}$ -0,25 σ	$\bar{x}$ +0,25 σ	h_3	n/5	χ^2_3
4	$\bar{x}$ +0,25 σ	$\bar{x}$ +0,84 σ	h_4	n/5	χ^2_4
5	$\bar{x}$ +0,84 σ	$+\infty$	h_5	n/5	χ^2_5
					$\chi^2=\Sigma\chi^2_i$

Ein Wert *x* fällt bei 5 Klassen mit 20% Wahrscheinlichkeit in die Klasse i, wenn $uG_i \le x < oG_i$ ist. (Die Zahlen 0,25 und 0,84 ergeben sich aus der Summenkurve der Normalverteilung, denn es gilt $\Phi(-0{,}8416)=0{,}2$, $\Phi(-0{,}2533)=0{,}4$, usw.). Häufigkeit h_i ist die Zahl der Werte, die in Klasse i ausgezählt wurden (i=1,2,...,k). Der Faktor 1/5 in der Spalte $\hat{e}_i$ der Erwartungswerte ergibt sich aus der benutzten Klassenzahl k=5. Die Chi-Quadrat-Komponenten berechnen sich zu $\chi^2_i = (h_i - \hat{e}_i)^2 / \hat{e}_i$. Das χ^2 hat bei 5 Klassen zwei Freiheitsgrade (FG = 2).

Schritt 3: Aussage: Sicherheitspunkt ist $\chi^2_\alpha = \chi^2(\alpha=0.05, FG=2)=5{,}99$. Wenn $\chi^2 < 5{,}99$, dann akzeptiere H_o, d.h. Normalverteilung, ansonsten H_A, d.h. signifikant keine Normalverteilung.

11. Mittelwertvergleiche

11.1 Einstichproben-t-Test (Test Messreihenmittel gegen Konstante)

Gegeben ist eine Messreihe $x_1, x_2, ..., x_n$. Der Mittelwert μ der Grundgesamtheit, aus der die Messreihe stammt, soll gegen einen Konstanten Wert μ_0 getestet werden. μ_0 kann eine vom Gesetzgeber festgelegte Norm sein, ein Literaturwert ohne Fehlerangabe oder eine sonstwie theoretisch begründete Zahl.
Schritt 0: Hypothese H_o: $\mu=\mu_0$ H_A: $\mu \neq \mu_0$ (zweiseitiger Test) $\alpha=0.05$ (5%)
oder z.B. H_A: $\mu > \mu_0$ (einseitiger >Test)
Schritt1 : Methode t-Test
Schritt 2: $t = \frac{\bar{x} - \mu_0}{\sigma_{n-1}} \sqrt{n}$, FG=n−1
Schritt3 : Aussage : Suche Sicherheitspunkt $t(\alpha, FG)$ aus der Tafel (beachte 1- oder 2-seitig)
2-seitiger Test: Wenn $t < -t(\alpha, FG, zweis.)$, dann ist signifikant $\mu < \mu_0 \rightarrow H_A$
Wenn $t > t(\alpha, FG, zweis.)$, dann ist signifikant $\mu > \mu_0 \rightarrow H_A$
1-seitiger Test : Wenn $t > t(\alpha, FG, eins.)$, dann ist signifikant $\mu > \mu_0 \rightarrow H_A$
in allen anderen Fällen H_o: $\mu = \mu_0$ annehmen (kein signifikanter Unterschied)

11. 2 Mittelwertvergleich zweier normalverteilter Grundgesamtheiten

Gegeben sind zwei unabhängige Stichproben (Messungen, Beobachtungen) $x_{11}, x_{12}, ..., x_{1n1}$ und $x_{21}, x_{22}, ..., x_{2n2}$ mit Umfang n_1 und n_2. Der erste Index bezeichnet die Stichprobe 1 oder

2, der zweite Index nummeriert die Beobachtungen innerhalb der Messreihe mit 1, 2, 3,... Sie wollen prüfen, ob die Mittelwertunterschiede zwischen den Populationen signifikant sind.

Schritt 0: Hypothese H_o: $\mu_1=\mu_2$ H_A: $\mu_1 \neq \mu_2$ (zweiseitiger Test) α=0.05 (5%)
oder z.B. H_A: $\mu_1 > \mu_2$ (einseitiger >Test)
Schritt1 : Methode t-Test mit gemittelter Standardabweichung
Schritt2 : Berechne für jede Stichprobe $\bar{x}_i$, $SAQ_i = n_i \sigma^2_{in} = (\Sigma x_{ij}^2) - n_i (\bar{x}_i)^2$, i=1,2
Berechne
FG = $n_1 + n_2 - 2$

$$\bar{\sigma} = \sqrt{\frac{SAQ_1 + SAQ_2}{n_1 + n_2 - 2}} \qquad t = \frac{\bar{x}_1 - \bar{x}_2}{\bar{\sigma}} \sqrt{\frac{n_1 \cdot n_2}{n_1 + n_2}}$$

Schritt3 : Aussage : Suche Sicherheitspunkt t(α, FG) aus der Tafel (beachte 1- oder 2-seitig)
2-seitiger Test: Wenn $t < -t(\alpha, FG, zweis.)$, dann ist signifikant $\mu_1 < \mu_2$
Wenn $t > t(\alpha, FG, zweis.)$, dann ist signifikant $\mu_1 > \mu_2$
1-seitiger Test : z.B. H_A: $\mu_1 > \mu_2$
Wenn $t > t(\alpha, FG, eins.)$, dann ist signifikant $\mu_1 > \mu_2$
in allen anderen Fällen H_o: $\mu_1 = \mu_2$ annehmen (kein signifikanter Unterschied)

Zahlenbeispiel Mittelwertvergleich unabhängiger Stichproben: Ein Photoreaktor baut mittels UV-Strahlung organische Substanzen im Wasser ab. Seine Leistung wird in [mg/KWh] für eine standardisierte Testsubstanz gemessen. Es liegen zwei unterschiedlich lange Messreihen für zwei Lampen (Lampe A und Lampe B) vor.

Abbauleistung mit Lampe A in mg/KWh	3.6	2.9	3.0	4.1	---
Abbauleistung mit Lampe B in mg/KWh	3.9	4.4	3.2	3.8	4.3

Frage: *Besteht ein signifikanter Unterschied in der Abbauleistung von Lampen des Typs A gegenüber Lampen des Typs B?*

Zuerst das Rechenschema für die Summen:

	Lampe A=x_1	x_1^2		Lampe B=x_2	x_2^2
1	3.6	12.96		3.9	15.21
2	2.9	8.41		4.4	19.36
3	3.0	9.00		3.2	10.24
4	4.1	16.81		3.8	14.44
5	---	---		4.3	18.49
Σ	13.6	47.18		19.6	77.74

n_1=4, $\bar{x}_1$=3.40 [mg/l], $SAQ_1 = n_1\sigma_{1,n}^2 = (47.18 - 4*3.40^2) = 0.940$ Anzahl, Mittel, SAQ
n_2=5, $\bar{x}_1$=3.92 [mg/l], $SAQ_2 = n_1\sigma_{1,n}^2 = (77.74 - 5*3.92^2) = 0.908$ Anzahl, Mittel, SAQ

Hypothese H_o: $\mu_1=\mu_2$, H_A: $\mu_1 \neq \mu_2$ (zweiseitiger Test), α=0.05 (5%) Hypothesenpaar
$\bar{\sigma} = \sqrt{(0.940 + 0.908)/(4 + 5 - 2)} = 0.5138$ [mg/l] Gemitteltes σ
FG=4+5–2=7, $t_\alpha = t(\alpha=0.05, FG=7, zweis.) = 2.36$ Freiheitsgrad, Sicherheitsp.
$t = ((3.40 - 3.92)/0.5138) \cdot \sqrt{(4 \cdot 5)/(4+5)} = -1.509$ t-Statistik
Da $|t| < t_\alpha$ akzeptieren wir H_0 Hypothesenauswahl
Es besteht kein signifikanter Unterschied in den Abbauleistungen von Lampen des Typs A gegenüber Lampen des Typs B.

Man nimmt im 2-Stichproben-t-Test für unabhängige Stichproben die gemittelte Standardabweichung bei angenommener Gleichheit der Varianzen (homoscedasticity). Bei unterschiedli-

chen Varianzen (heteroscedasticity) ist der Fakt der Ungleichheit unerheblich, wenn die Stichprobenumfänge $n_1>30$ und $n_2>30$ sind. Ist das jedoch nicht der Fall, dann nimmt man den Welch-Test bzw. einen ähnlich aufgebauten Test. Der Welch-Test führt auf nichtganzzahlige Freiheitsgrade, die dann zu runden sind.

Mittelwertvergleich zweier normalverteilter Grundgesamtheiten bei ungleichen Varianzen und entweder $n_1 \leq 30$ oder $n_2 \leq 30$ oder beide $n \leq 30$ (Welch-Test).

Hypothese H_o: $\mu_1=\mu_2$ H_A: $\mu_1 \neq \mu_2$ (zweiseitiger Test) $\alpha=0.05$ (5%)
Methode: Welch-Test mit gemittelter Standardabweichung und adjustierten Freiheitsgraden.

Berechne für jede Stichprobe Mittelwert $\bar{x}_i$, Standardabweichung $\sigma_{i,n-1}$ für i=1,2

$$\bar{\sigma} = \sqrt{\frac{\sigma_{1,n-1}^2}{n_1} + \frac{\sigma_{2,n-1}^2}{n_2}}$$

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\bar{\sigma}}$$

mit Freiheitsgraden $FG = \dfrac{(g_1+g_2)^2}{\dfrac{g_1^2}{n_1-1}+\dfrac{g_2^2}{n_2-1}}$ mit $g_1 = \dfrac{\sigma_{1,n-1}^2}{n_1}$ und $g_2 = \dfrac{\sigma_{2,n-1}^2}{n_2}$.

Berechne mit EXCEL-Funktion =TINV(…) Sicherheitspunkt t(α, FG, zweiseitig)
2-seitiger Test: Wenn $t \leq -t(\alpha, FG, zweis.)$, dann ist signifikant $\mu_1 < \mu_2 \rightarrow H_A$
Wenn $t \geq t(\alpha, FG, zweis.)$, dann ist signifikant $\mu_1 > \mu_2 \rightarrow H_A$
in allen anderen Fällen H_o: $\mu_1 = \mu_2$ annehmen (kein signifikanter Unterschied).

F-Test zur Entscheidung, ob gleiche oder signifikant ungleiche Varianzen in den Grundgesamtheiten vorliegen.

Sind $\sigma_{1,n-1}^2$ und $\sigma_{2,n-1}^2$ die Varianzschätzungen aus den beiden Stichproben $x_{11}, x_{12}, ..., x_{1n1}$ und $x_{21}, x_{22}, ..., x_{2n2}$ mit Umfang n_1 und n_2, dann dann ist die Testgröße

$F = \dfrac{\sigma_{1,n-1}^2}{\sigma_{2,n-1}^2}$ unter H_o F-verteilt mit $FG_1=n_1-1$ und $FG_2=n_2-1$ Freiheitsgraden.

H_o: $\sigma_1^2 = \sigma_2^2$ Gleichheit der Varianzen (homoscedasticity) in den Grundgesamtheiten.
H_A: $\sigma_1^2 \neq \sigma_2^2$ Ungleichheit der Varianzen (heteroscedasticity).

Wir akzeptieren H_A, wenn $F \geq F(\alpha, FG_1, FG_2)$ ist (Sicherheitspunkt der F-Verteilung).
Wir akzeptieren H_o, wenn $F < F(\alpha, FG_1, FG_2)$ ist.

Ist F<1, dann bildet man den Kehrwert 1/F und testet mit diesem, statt mit F. Dabei vertauschen sich die Freiheitsgrade. Es wird $FG_1=n_2-1$ und $FG_2=n_1-1$.

Die Tafel der Sicherheitspunkte der F-Verteilung auf Seite 2 gibt die einseitigen oberen Sicherheitspunkte für eine Irrtumswahrscheinlichkeit α=5%.

Wir nehmen den einfachen t-Test im Falle gleicher Varianzen (Hypothese H_o).
Wir nehmen den Welch-Test im Falle ungleicher Varianzen (Hypothese H_A).

Zahlenbeispiel F-Test und anschließender Welch-Test im Falle ungleicher Varianzen

Gegeben sind zwei unabhängige Stichproben. Zwei Patientengruppen mit leichter Demenz unterzogen sich IQ-Tests. Gruppe 1 ohne Medikament, Gruppe 2 mit Medikament. Frage: Beeinflusst das Medikament signifikant den IQ?

Gruppe 1	102	89	97	88	94	100	91	97	105	102	95	93	99	90	95
Gruppe 2	82	116	104	87	98	74	114	79	98	84	113	117	114	123	

$\sigma_{1,n-1} = 5{,}12974519$ $\sigma_{2,n-1} = 16{,}4652481$ $n_1=15$ (FG=14) $n_2=14$.(FG=13)

Da $\sigma_2 > \sigma_1$ ist, vertauschen wir Zähler und Nenner, d.h. es ist $F = \sigma^2_2 / \sigma^2_1$.

$F = (16{,}4652481)^2 / (5{,}12974519)^2 = 10{,}30255575$ mit $FG_1=14$ und $FG_2=13$.

Den Sicherheitspunkt $F(\alpha=5\%, FG_1=14, FG_2=13) = 2{,}554$ finden wir

a) Durch Interpolation in unserer F-Tafel Seite 2
b) Oder mit der Excelfunktion =FINV(0,05 ; 14 ; 13)

Wir akzeptieren H_A, weil $F \geq F(\alpha, FG_1, FG_2)$ ist. Es besteht ein signifikanter Unterschied in den Varianzen der beiden Grundgesamtheiten.
Wir empfehlen den Welch-Test.

$$\bar{\sigma} = \sqrt{\frac{\sigma^2_{1,n-1}}{n_1} + \frac{\sigma^2_{2,n-1}}{n_2}} = 4{,}5955 \quad \text{und}$$

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\bar{\sigma}} = (95{,}8 - 100{,}214) / 4{,}5955 = -0{,}9606$$

$$g_1 = \frac{\sigma^2_{1,n-1}}{n_1} = 1{,}754 \quad , \quad g_2 = \frac{\sigma^2_{2,n-1}}{n_2} = 19{,}365 \quad , \quad FG = \frac{(g_1 + g_2)^2}{\frac{g_1^2}{n_1 - 1} + \frac{g_2^2}{n_2 - 1}} = 15{,}345.$$

Wir runden FG=15. Sicherheitspunkt ist $t(\alpha=5\%, FG=15, \text{zweiseitig}) = 2{,}13$.
Da $|t| < t(\alpha=5\%, FG=15, \text{zweiseitig}) = 2{,}13$ ist, akzeptieren wir Hypothese H_o.
Das Medikament hat keinen signifikanten Einfluss auf den IQ.

11. 3 Mann-Whitney-Test (Vergleich zweier Mittelwerte, Rangtest)

Gegeben sind zwei unabhängige Stichproben (Messungen, Beobachtungen) $x_1, x_2, ..., x_n$ und $y_1, y_2, ..., y_m$ mit Umfang *n* bzw. *m*. Sie wollen prüfen, ob die Mittelwertunterschiede signifikant sind. Es lässt sich definitiv **keine Normalverteilung** der zwei Messreihen herstellen, oder aber man will dieser Diskussion aus dem Wege gehen. Der Mann-Whitney-Test ist ein Rangtest:

Schritt 0: Hypothese H_o: $\mu_1=\mu_2$ H_A: $\mu_1 \neq \mu_2$ (zweiseitiger Test) $\alpha=0.05$ (5%)
oder z.B. H_A: $\mu_1 > \mu_2$ (einseitiger >Test)

Schritt 1: Rangtest von Mann-Whitney für beliebig verteilte Daten

Schritt 2: Man sortiert die vermischten Daten in eine aufsteigende Folge und vergibt Rangzahlen 1, 2, 3, ... Treten gleiche Messwerte auf, dann vergibt man bei gerader Zahl gleicher Messwerte allen denselben Rang, z.B., .25, **27, 27, 27**, 29 ..., wie der mittlere Wert erhalten würde. Bei einer geraden Anzahl gleicher Messwerte nimmt man den Durchschnitt der betroffenen Ränge, z.B. ...,25, **26.5, 26.5**, 28, ., Anschließend bildet die Rangsummen R_x der x-Messwerte, und ebenso R_y, daraus U_X, U_Y:

$$U_X = nm + \frac{n(n+1)}{2} - R_X\,, \qquad U_Y = nm + \frac{m(m+1)}{2} - R_Y\,,$$

Ist n≤10 oder m≤10, dann berechnet man U=Min(U_X, U_Y) und ist fertig,

sonst berechnet man aus U das $u = \dfrac{U - (nm/2)}{\sqrt{nm(n+m+1)/12}}$

Schritt 3: Sicherheitspunkte $U_\alpha = U(\alpha, n, m)$ finden wir z.B. in E. Weber, Tab. 19 ff.

2-seitiger Test: Wenn $U = U_X > U_\alpha$, dann ist signifikant $\mu_X > \mu_Y$
Wenn $U = U_Y > U_\alpha$, dann ist signifikant $\mu_Y > \mu_X$

1-seitiger Test : Wenn $U = U_X > U_{\alpha/2}$, dann ist signifikant $\mu_X > \mu_Y$

in allen anderen Fällen H_o: $\mu_X = \mu_Y$ annehmen (kein signifikanter Unterschied)

Bei n>10 und m>10 vergleichen wir **u** mit dem zweiseitigem Sicherheitspunkt der Normalverteilung, $u_\alpha = 1.96$, einseitig $u_\alpha = 1.65$: (gültig bei $\alpha = 0.05$)

11. 4 Gepaarter t-Test

(Mittelwertvergleich einer normalverteilten korrelierten Stichprobe) Gegeben ist eine korrelierte Stichproben aus n Wertepaaren (y_1, x_1), (y_2, x_2), ..., (y_n, x_n). Korreliert heißt, die Messwertpaare (y_i, x_i) sind am selben Objekt gewonnen, z.B. ist y der Blutdruck vor Medikamentengabe, x der Blutdruck nach Medikamentengabe. Welche Größe mit x und welche mit y bezeichnet wird, ist egal. Man muß nur das Vorzeichen des Effekts **d**=y–x beachten.

Schritt 0: Hypothese H_o: $\mathbf{d} = 0$ H_A: $\mathbf{d} \neq 0$ (zweiseitiger Test) α=0.05 (5%)
oder z.B. H_A: $\mathbf{d} > 0$ (einseitiger >Test)

Schritt 1 : Methode t-Test für das Differenzenmittel

Schritt 2 : Berechne alle Differenzen $d_i = y_i - x_i$, daraus Mittelwert und Standardabweichung.

$\bar{d} = \dfrac{\sum d_i}{n}$	$s_d = \sqrt{\dfrac{\sum (d_i - \bar{d})^2}{n-1}} = \sqrt{\dfrac{\left(\sum d_i^2\right) - n \cdot \bar{d}^{\,2}}{n-1}}$
$t = (\bar{d} / s_d) \cdot \sqrt{n}$	FG=n–1

Die linke s_d-Formel ist genauer, die rechte einfacher zu berechnen.

Schritt3 : Aussage : Suche Sicherheitspunkt t(α, FG) aus der Tafel (beachte 1- oder 2-seitig)

2-seitiger Test: Wenn $t < -t(\alpha, FG)$, dann ist signifikant $\mu_y < \mu_x$ bzw. $\mathbf{d} < 0$
Wenn $t > t(\alpha, FG)$, dann ist signifikant $\mu_y > \mu_x$ bzw. $\mathbf{d} > 0$

1-seitiger Test : z.B. H_A: $\mu_y > \mu_x$ bzw. H_A: $\mathbf{d} > 0$
Wenn $t > t(\alpha, FG)$, dann ist signifikant $\mu_y > \mu_x$ bzw. $d > 0$

in allen anderen Fällen H_o: $\mu_y = \mu_x$ annehmen (kein signifikanter Unterschied)

11. 5 Gepaarter Mittelwert-Rangtest von Wilcoxon

(Matched-pairs signed-ranks test) Mittelwertvergleich einer **nicht normalverteilten** korrelierten Stichprobe. Gegeben ist dieselbe Datenanordnung wie beim gepaarten t-Test.

Schritt 0: Hypothese H_o: $\mathbf{d} = 0$ H_A: $\mathbf{d} \neq 0$ (zweiseitiger Test) α=0.05 (5%)
oder z.B. H_A: $\mathbf{d} > 0$ (einseitiger >Test)

Schritt 1: Wilcoxon-Test (Rangtest)

Schritt 2: Die Differenzen $d_i=y_i-x_i$ werden ohne Rücksicht auf das Vorzeichen mit Rangzahlen versehen. Sind zwei oder mehr d_i von gleicher absoluter Größe, so erhalten sie das Rangmittel der ihnen zustehenden Ränge. Differenzen $d_i=0$ werden entfernt und das **n** entsprechend erniedrigt. Jetzt werden die Rangzahlen mit dem Vorzeichen ihres d_i versehen und getrennt addiert. R_N ist die Summe der negativen, R_P die Summe der positiven Rangzahlen.

Bei $n\leq 25$ berechnet man $U=\text{Min}(R_N, R_P)$

Bei $n>25$ berechnet man $u=\dfrac{U-(n(n+1)/4)}{\sqrt{n(n+1)(2n+1)/24}}$

Schritt 3: Sicherheitspunkte $U_\alpha=U(\alpha, n)$ finden wir z.B. in E. Weber, Tab. 25.

2-seitiger Test: Wenn $U=R_P < U_\alpha$, dann ist signifikant $\mu_Y > \mu_X \rightarrow H_A$

Wenn $U= R_N < U_\alpha$, dann ist signifikant $\mu_X > \mu_Y \rightarrow H_A$

1-seitiger Test : Wenn $U= R_P < U_{\alpha/2}$, dann ist signifikant $\mu_Y > \mu_X \rightarrow H_A$

in allen anderen Fällen H_o: $\mu_X = \mu_Y$ annehmen (kein signifikanter Unterschied)

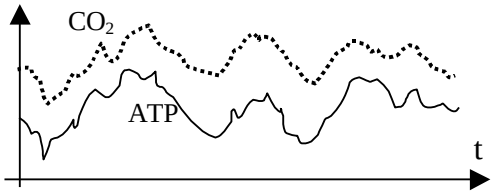
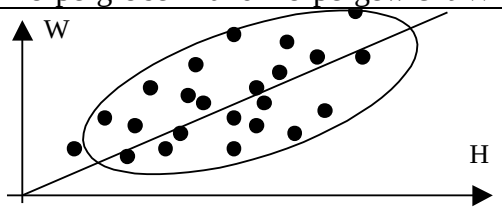
Bei $n>25$ vergleichen wir **u** mit dem zweiseitigem Sicherheitspunkt der Normalverteilung, $u_\alpha=1.96$, einseitig $u_\alpha=1.65$: (gültig bei $\alpha=0.05$)

12. Korrelation und Regression

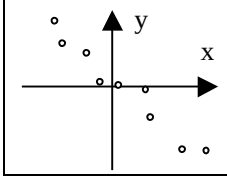
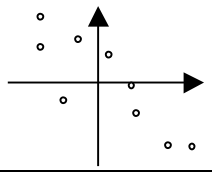
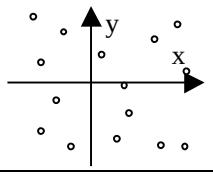
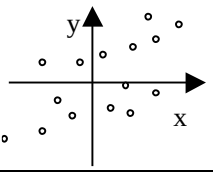
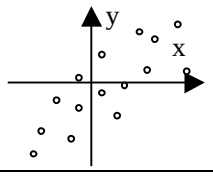
Wichtige Analysen bei einem metrischen Merkmalspaar sind die Korrelationsanalyse und die einfache Regressionsanalyse. **Korrelationsanalyse** ist angebracht, wenn zwei beobachtete oder gemessene Merkmale in Beziehung stehen und sich keines der beiden Merkmale als Einflussgröße oder Zielgröße qualifizieren lässt. Beispiel: Blutdruck P_1 am Puls gemessen und Blutdruck P_2 an der Aorta gemessen. Man kann nicht sagen, dass eines der Merkmale vom anderen abhängt, sondern beide Merkmale hängen vermutlich von einer oder sogar mehreren dritten Größen (**Faktoren**) ab. Einfache **Regressionsanalyse** ist angebracht, wenn definitiv eine Zielgröße von einer Einflussgröße abhängt. Beispiel: Der Blutdruck P_1 am Puls gemessen hängt von der Dosis ***x*** eines blutdrucksenkenden Präparats ab.

12.1 Korrelation nach Bravais-Pearson

Gleichlaufendes oder ähnliches Verhalten zweier Merkmale wird als Korrelation bezeichnet, wobei eine direkte Abhängigkeit des einen Merkmals vom anderen nicht Voraussetzung ist. Zeitliche Korrelation ist sogar ohne jeden Zusammenhang denkbar, wenn man an die gesellschaftlichen und kulturellen Entwicklungen voneinander isolierter Kontinente denkt.

Zeitliche Korrelation zwischen ATP- und CO_2-Produktion von Candida saccharomyces	**Produkt-Momenten-Korrelation** zwischen Körpergröße H und Körpergewicht W
	
Die über der Zeit aufgetragenen Werte der gemessenen ATP-Produktion der Hefezellen und der CO_2-Ausstoß des Fermenters haben einen ähnlichen Verlauf. Hohe ATP-Werte z.B. korrelieren mit hohen CO_2 -Werten	Große Probanden wiegen im Schnitt mehr als kleine Probanden, wobei es aber keinen sklavischen Zusammenhang gibt. Die **Korrelationsellipse** ist eine Höhenlinie der 2-dimensionalen Dichteverteilung der Messpunkte

Der Korrelationskoeffizient **r** wird so normiert, dass er nur Werte zwischen –1 und +1 annehmen kann. Ein Wert r = 1 bedeutet, dass ein exakter linearer Zusammenhang der Form y=a+bx oder x=c+dy zwischen den beiden Merkmalen besteht ohne jede Abweichung. r = –1 bedeutet einen ebenso exakten Zusammenhang, aber von der Form y=a–bx bzw. y=c–dy. Die Werte ***a, b, c, d*** sind Konstante. Die Graphiken unten zeigen verschiedene Korrelationswerte und das Streubild der Messpunkte.

r = –0.9	r = –0.5	r = 0	r = 0.5	r = 0.9
				

12.2 Linearer Korrelationskoeffizient r

(Produkt-Momenten-Korrelationskoeffizient nach Bravais und Pearson) zweier metrischer Merkmale in einer Grundgesamtheit. Gegeben ist eine Stichprobe mit Wertepaaren (x_1,y_1), (x_2,y_2), ..., (x_n,y_n), wobei es egal ist, welche der beiden Größen x bzw. y genannt wird. Berechne zuerst die drei Abweichungsprodukt- bzw. Abweichungsquadratsummen SAPxy, SAQxx, SAQyy, wobei die linke Formel genauer, die rechte schneller zu berechnen ist. Folgendes Rechenschema bietet sich an, wenn man lediglich mit einem einfachen Taschenrechner ausgerüstet ist. Man berechnet die 5 Summen und benutzt anschließend den rechten Formelsatz für SAPxy, SAQxx, SAQyy. Vorsicht! Die Mittelwerte nicht zu sehr runden. 6 signifikante Ziffern sollten bleiben.
Es folgt das Rechenschema:

Nr	x_i	y_i	x_i^2	$x_i\, y_i$	y_i^2
1	x_1	y_1	x_1^2	$x_1\, y_1$	y_1^2
2	x_2	y_2	x_2^2	$x_2\, y_2$	y_2^2
...	...	...	...	...	...
n	x_n	y_n	x_n^2	$x_n\, y_n$	y_n^2
	$\Sigma\, x_i$	$\Sigma\, y_i$	$\Sigma\, x_i^2$	$\Sigma\, x_i\, y_i$	$\Sigma\, y_i^2$

$$SAPxy = \sum_{i=1}^{n}\left((x_i - \bar{x})(y_i - \bar{y})\right) \qquad \text{bzw.} \quad SAPxy = \left(\sum_{i=1}^{n} x_i y_i\right) - n \cdot \bar{x}\bar{y}$$

$$SAQxx = \sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2 \qquad \text{bzw.} \quad SAQxx = \left(\sum_{i=1}^{n} x_i^2\right) - n \cdot \bar{x}^2$$

$$SAQyy = \sum_{i=1}^{n}\left(y_i - \bar{y}\right)^2 \qquad \text{bzw.} \quad SAQyy = \left(\sum_{i=1}^{n} y_i^2\right) - n \cdot \bar{y}^2$$

$$\hat{r} = \frac{SAPxy}{\sqrt{SAQxx \cdot SAQyy}} \qquad t = \frac{\hat{r}}{\sqrt{1-\hat{r}^2}}\sqrt{n-2} \qquad \text{FG} = n - 2$$

$\hat{r}$ schätzt den Korrelationskoeffizienten ***r*** der Grundgesamtheit. Die Nullhypothese ist H_0: **r**=0 (keine Korrelation in der Grundgesamtheit), HA: **r**≠0 (signifikante Korrelation in der Grundgesamtheit). Die Korrelation **r** in der Grundgesamtheit ist signifikant von Null verschieden, wenn $|t| \geq t(\alpha,FG)$ für zweiseitigen Test ist. Sonst nimmt man H_o: **r** = 0 an, d.h. "keine signifikante Korrelation in der Grundgesamtheit."

Zahlenbeispiel Korrelationskoeffizient: Der Sauerstoffgehalt y [mg/l] wurde zusammen mit dem Luftstrom x [m^3/h] in einem Fermenter gemessen. Zuerst das Rechenschema für die Summen:

Nr	x	y	x^2	xy	y^2
1	50	1.3	2500	65	1.60
2	110	1.9	12100	209	3.61
3	110	2.1	12100	231	4.41
4	300	3.7	90000	1110	13.69
5	370	5.1	136900	1887	26.01
Σ	940	14.1	253600	3502	49.41

$\bar{x}$ =188, $\bar{y}$ =2.82, SAQxx=253600–5*188^2 =76880, SAQyy=49.41–5*2.82^2=9.648,
SAPxy=3502–5*188*2.82=851.2,

$\hat{r}$ = 851.2 / $(76880*9.648)^{0.5}$ = 0.98834 Korrelationskoeffizient
H_0: **r**=0 , HA: **r**≠0, α=0.05 Hypothesenpaar
t = (0.98834 / $(1–0.98834^{\,2})^{0.5}$) * $3^{\,0.5}$ = 11.22 t-Statistik
FG = 5–2 = 3 Freiheitsgrad
t_α = t (α=0.05, FG, zweiseitig) = 3.18 Sicherheitspunkt der t-Verteilung
Wir akzeptieren H_A, d.h. im Fermenter sind Sauerstoffgehalt und Lufteintrag (hoch) korreliert.

12.3 Einfache lineare Regression, Ausgleichsgerade

Wir haben eine Einflussgröße x, von deren Werten angenommen wird, daß sie fehlerfrei einstellbar sind (Modellannahme) und eine Zielgröße y, die Zufallsfehler e_i enthält und über eine einfache Geradengleichung von der Einflussgröße abhängt. Ist x die Zeit, dann sprechen wir auch von **Trendanalyse**.

Regressionsmodell $y_i = \mathbf{a} + \mathbf{b}\, x_i + e_i$

Gesucht sind Schätzwerte für die Regressionskonstante **a** und den Regressionskoeffizienten **b** in der Grundgesamtheit. Gegeben ist eine Stichprobe mit den n Wertepaaren (x_1,y_1), (x_2,y_2), ..., (x_n,y_n) wie beim Korrelationskoeffizienten. **y** heißt Zielgröße, **x** heißt Einflussgröße. $\mathbf{e_i}$ heißt Residuum (Abweichung, Fehler) im Punkt i. Regressionskonstante **a** ist der Erwartungswert der Zielgröße im Punkt x=0. Regressionskoeffizient **b** heißt auch *Anstieg* der Geraden. Die Koeffizienten **a** und **b** werden nach der "*Kleinsten-Quadrate-Methode*" geschätzt, d.h. so, dass die Summe Σe_i^2=Minimum wird. Berechne zuerst SAPxy, SAQxx, SAQyy wie bei der Korrelation, dann:

$\hat{b} = SAPxy / SAQxx$ schätzt den Regressionskoeffizienten **b**

$\hat{a} = \bar{y} - \hat{b} \cdot \bar{x}$ schätzt die Regressionskonstante **a**

$\hat{y}_i = \hat{a} + \hat{b} \cdot x_i = \bar{y} + \hat{b} \cdot (x_i - \bar{x})$ schätzt **y** im Punkt x_i (Erwartungswert)

$\hat{e}_i = y_i - \hat{y}_i$ schätzt das Residuum $\mathbf{e_i}$ im Punkt x_i

$$\hat{S}_R = \sqrt{\frac{\sum (y_i - \hat{y}_i)^2}{n-2}} = \sqrt{\frac{SAQyy - \hat{b} \cdot SAPxy}{n-2}} = \sqrt{\frac{\sum \hat{e}_i^2}{n-2}}$$

$\hat{S}_R$ schätzt den mittleren Fehler σ_R in der Grundgesamtheit (Reststreuung der Punkte um die Gerade (in y-Richtung gesehen)). Die mittlere Formel ist für den Taschenrechner geeignet.

FG = n-2 Freiheitsgrad der Reststreuung $\hat{S}_R$

$S_b = \hat{S}_R / \sqrt{SAQxx}$ Schätzfehler für Regressionskoeffizienten **b**

Schätzfehler der Regressionskonstanten **a**	Schätzfehler des Erwartungswertes $\hat{y}_i$
$S_a = \hat{S}_R \sqrt{\frac{1}{n} + \frac{\bar{x}^2}{SAQxx}}$	$S_{\hat{y}} = \hat{S}_R \sqrt{\frac{1}{n} + \frac{(x_i - \bar{x})^2}{SAQxx}}$

$t_a = \hat{a} / S_a$ mit FG = n-2 testet H_o: **a**=0 gegen H_A: **a**≠0 (2-seitig)

$t_b = \hat{b} / S_b$ mit FG = n-2 testet H_o: **b**=0 gegen H_A: **b**≠0 (2-seitig)

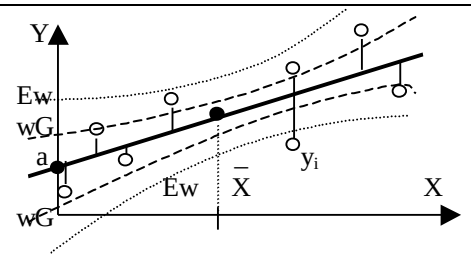
Beide Tests lassen sich bei nachweisbarem Vorwissen auch einseitig anlegen. Ein signifikantes **a**≠0 heißt, dass die Zielgröße y für den Wert x=0 der Einflussgröße einen Wert y≠0 hat. Ein signifikanter Wert **b**≠0 sagt, dass die Einflussgröße x die Zielgröße y tatsächlich beeinflusst, d.h., daß der Anstieg der Geraden nicht Zufall ist.

$$\hat{y}_i \pm t(\alpha, FG, zweis.) \cdot \hat{S}_R \cdot \sqrt{\frac{1}{n} + \frac{(x_i - \bar{x})^2}{SAQxx}}$$ Konfidenzintervall der wahren Regressionsgeraden

$$\hat{y}_i \pm t(\alpha, FG, zweis.) \cdot \hat{S}_R \cdot \sqrt{1 + \frac{1}{n} + \frac{(x_i - \bar{x})^2}{SAQxx}}$$ Konfidenzintervall der Einzelwerte bei Prognose.

Zieht man immer wieder neue Stichproben des Umfangs n und berechnet man aus jeder Stichprobe die Regressionsgerade, dann erwartet man 100-α% der Geraden im Konfidenzintervall der "**wahren Geraden**". Ebenso liegt die wahre (unbekannte) Regressionsgerade der Grundgesamtheit mit 100-α% im Konfidenzintervall. Für Prognosen ist der zu erwartende Fehler der **Einzelbeobachtung** wichtig. 100-α% der Einzelwerte werden im Konfidenzintervall der Einzelwerte erwartet. Wie man sieht, erweitert sich das Konfidenzintervall außerhalb des Messbereichs dramatisch, so dass sich allzu kühne Prognosen z.B. in die Zukunft verbieten.

Die Abbildung rechts zeigt die Regressionsgerade im X-Y-Koordinatensystem. Sie geht durch den Punkt **a** auf der Y-Achse und durch den Punkt $(\bar{x}, \bar{y})$. Die Messwerte y_i sind durch kleine Kreise, die Residuen e_i durch Striche dargestellt. Das Konfidenzintervall der wahren Geraden (wG) ist gestrichelt, das der Einzelwerte (Ew) ist gepunktet dargestellt.



Folgende Bedingungen stellt das Regressionsmodell an die Daten:

1. Das einfache lineare Modell $y_i = a + b\,x_i + e_i$ trifft auf die Grundgesamtheit zu
2. Die Messpunkte streuen überall normalverteilt mit N(μ=0; σ= $\hat{S}_R$) um die Gerade.

Zahlenbeispiel Ausgleichsgerade: Der Sauerstoffgehalt y [mg/l] wurde in Abhängigkeit vom Luftstrom x [m^3/h] in einem Fermenter gemessen. Das Rechenschema für die Summen ist identisch mit dem des Zahlenbeispiels für den Korrelationskoeffizienten.

$\bar{x}$ =188, $\bar{y}$ =2.82, SAQxx=253600–5*188^2 =76880, SAQyy=49.41–5*2.82^2=9.648, SAPxy=3502–5*188*2.82=851.2,

$\hat{b}$ =851.2/76880=0.0110718 [mg/l / m^3/h]	Anstieg der Geraden
$\hat{a}$ =2,82–0.01107*188=0.7388 [mg/l]	Regressionskonstante
$\hat{S}_R$ =((9.648-0.0110718*851.2) / (5-2))$^{0.5}$ = 0.273 [mg/l]	Reststreuung
FG=5–2=3	Freiheitsgrad der Reststreuung
t (α=0.05, FG, zweiseitig) = 3.18	Sicherheitspunkt der t-Verteilung
$\hat{y}_{x=500}$=0.7388+0.01107*500=6.2738 [mg/l]	Erwartungswert für x=500 [m^3/h]
6.27 ± 0.273 * **(** 1/5 + (500-188)^2/76880 **)**$^{0.5}$ * 3.18 =	95%-Konfidenzintervall der
6.27 ± 1.055, gerundet 6.27 ± 1.0	"wahren Geraden" für x=500
6.27 ± 0.273 * **(** 1 + 1/5 + (500-188)^2/76880 **)**$^{0.5}$ * 3.18 =	95%-Konfidenzintervall der Einzel-
6.27 ± 1.368, gerundet 6.27 ± 1.4	werte für x=500
H_o: **b**=0 gegen H_A: **b**≠0 (2-seitig), α=0.05	Hypothesenpaar zum Anstieg **b**
t=0.0110718*(76880)$^{0.5}$/0.273=11.24	t-Statistik zum Anstieg **b**
Da $\lvert t \rvert \geq 3.18$, akzeptieren wir H_A	Hypothesenauswahl H_o oder H_A

Der Anstieg **b** der Geraden in der Grundgesamtheit unterscheidet sich signifikant von 0. Es besteht ein signifikanter Zusammenhang zwischen Sauerstoffkonzentration y [mg/l] und Lufteintrag x [m^3/h].

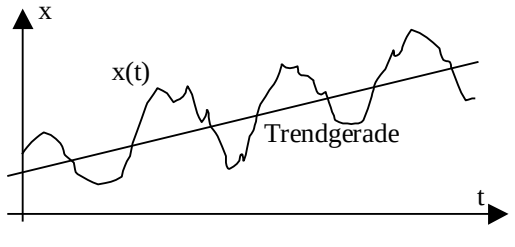
H_o: **a**=0 gegen H_A: **a**≠0 (2-seitig), α=0.05	Hypothesenpaar zur Konstanten **a**
t=0.7388 / (0.274 * (1/5 + 188^2/76880)$^{0.5}$) = 3,320	t-Statistik zur Konstanten **a**
Da $\lvert t \rvert \geq 3.18$, akzeptieren wir H_A	Hypothesenauswahl H_o oder H_A

Die Regressionskonstante **a** der Grundgesamtheit unterscheidet sich signifikant (auf 5%-Niveau) von 0. Auch bei Null Lufteintrag besteht eine Sauerstoffkonzentration ≠0.

Hier sieht man, wie wichtig das Signifikanzniveau sein kann: Das Testergebnis für die Konstante **a** ist Müll! Jeder Biologe weiß, bei Null Lufteintrag ist kein Sauerstoff im Fermenter. Wegen t(α=0.01, FG=3, zweis.)=5.84 ist die Konstante **a** auf dem 1%-Niveau nicht signifikant. Man sollte sie aus diesem Modell entfernen und mit dem Modell $y_i = b\, x_i + e_i$ arbeiten, was einer Geraden durch den Koordinatenursprung entspricht. Man gewinnt einen Freiheitsgrad bei der Reststreuung, da nur noch Anstieg **b** aus den Daten geschätzt werden muss.

12.4 Zeitreihen (Time series)

Die Graphik zeigt die Trendgerade überlagert von einer einfachen periodischen Schwingung, wie sie beim Planktonwachstum z.B. der Einfluss der Gezeiten (Einflüsse von Sonne und Mond) hervorrufen können. Periodische Schwingen lassen sich z.B. durch Sinuswellen mit Phasenverschiebung modellieren.



Ist die Zeit die Einflussgröße, dann spricht man von Zeitreihen. Der Einfluss anderer Variablen ist vorhanden, wird aber nicht direkt modelliert. Zumeist spaltet man im Modell die zeitliche Änderung in einen linearen bzw. nichtlinearen Trend und eine Anzahl periodischer Schwingungen (saisonale Schwankungen) um diese Trendkurve auf. Bei der Modellierung der periodischen Schwingungen, die durch Tages-, Wochen, Monats-, Mond-, Quartals-, Jahres- oder andere Rhythmen bestimmt sein können, unterscheiden sich die Theorien. Diese bilden ein Buch für sich.

In der betrieblichen Praxis berechnet man gern die Abweichungen der Monatsmittel (bzw. Quartalsmittel) vom Langzeittrend für die Vorjahre und erhält so einen Planwert für die Abweichungen der Monatswerte (bzw. Quartalswerte) des laufenden Jahres vom aktuellen Trend. Das entspricht einer ***ipsativen Skalierung*** der Abweichungen.

Beispiel: Gegeben sind die Verkaufssmengen X an Kästen Radler für 2 Jahre (8 Quartale). Es besteht ein langfristiger linearer Aufwärtstrend , indem die Bestellmengen pro Jahr um einen konstanten Betrag steigen. Der Anstieg ***b*** der Trendgeraden ist noch zu ermitteln.

Quartal	**X in 2020**	**X in 2021**
1	27350	28000
2	29650	30500
3	28400	29150
4	26850	27600

Wir berechnen den Anstieg ***b*** der Trendgeraden, indem wir zuerst die Jahresmittel $\boldsymbol{X_{2020}}$ und $\boldsymbol{X_{2021}}$ berechnen.

$\boldsymbol{X_{2020}}$ = 28062,5 Kästen/Quartal

$\boldsymbol{X_{2021}}$ = 28812,5 Kästen/Quartal

Der Einfachheit halber rechnen wir mit 4 Quartalen (zu 365/4= 91,26 Tagen). Mit diesen Werten ist der Anstieg $b = ((\boldsymbol{X_{2021}} - \boldsymbol{X_{2020}}) / 4) = 187{,}5$ Kästen/Quartal. Die Trendgerade X(t) gehe exakt in den Jahresmitten durch die Jahresmittel, d.h., $\boldsymbol{X(t) = X_{2020} + bt}$, wobei ***t*** in Quartalen seit Jahresmitte 2020 anzugeben ist. Der Jahresanfang von 2020 hat dann negative Quartalsnummern. Beispiel: Die Mitte des 1. Quartals 2020 liegt 1,5 Quartale vor der Jahresmitte 2020, d.h. X(1.Quartal 2020)=X(-1,5)= 28062,5+187,5·(-1,5)=27781, d.h., es werden laut Trend im 1. Quartal 27781 Kästen erwartet. Der Trendwert für das 4. Quartal 2021 ist dann X(5,5)=29094.

Wir berechnen die 8 Abweichungen **ΔX** der 8 Quartalswerte vom jeweiligen Quartalstrendwert und bilden dann die 4 Mittelwerte $\boldsymbol{\Delta X_{M,Q}}$ für jedes Quartal nach dem Beispiel für das 1. Quartal

$\boldsymbol{\Delta X_{Q1,2022}} = (\boldsymbol{\Delta X_{Q1,\,2020}} + \boldsymbol{\Delta X_{Q1,\,2021}}) / 2 = (-431 - 531) / 2 = \mathbf{-481}$

Quartal	**ΔX in 2020**	**ΔX in 2021**	**ΔX in 2022 erw.**
1	27350-27781= -431	28000-28531= -531	-481
2	29650-27968= 1681	30500-28719= 1781	1731

3	28400-28156= 244	29150-28906= 244	244
4	26850-28344= -1494	27600-29094= -1494	-1494

Wir berechnen die Trendwerte X(Qi) für die 4 Quartale 2022, addieren dazu die für 2022 erwarteten Abweichungen **ΔX** aus der obigen Tabelle, und erhalten so eine Schätzung der erwarteten Quartalsumsätze für 2022.

Quartal	X(Qi 2022) Trendwert	X(Qi 2022) erwartet
1	X(6,5)= 29281	28800
2	X(7,5)= 29469	31200
3	X(8,5)= 29656	29900
4	X(9,5)= 29844	28350

Natürlich ist bei einem nichtlinearen Trend die entsprechende nichtlineare Formel für X(t) statt der Geradengleichung $X(t)= X_{2020} + bt$ einzusetzen, z.B. $X(t) = X_{2020}\, e^{\alpha(t-to)}$. Dabei ist α der Wachstumskoeffizient der e-Kurve, wie ihn eine nichtlineare Kurvenanpassung an eine e-Kurve bei Verwendung von Quartalen als Zeitintervalle berechnen würde, und t_0 die Zeitangabe, z.B. $t_0=2$, für den Zeitpunkt, für den das Jahresmittel X_{2020} exakt wiedergegeben werden soll (wegen $e^0=1$).

Autokorrelation

Die Autokorrelationsfunktion Ak(LAG) entsteht, wenn man eine Zeitfunktion x(t) mit sich selbst korreliert und die "Kopie" der Kurve nach und nach immer weiter gegen das "Original" verschiebt. Für jede Verschiebung wird der Korrelationskoeffizient ***r*** berechnet und über der Verschiebung aufgetragen. Dabei ist "LAG" die zeitliche Verschiebung zwischen x(t) und der zeitlich verschobenen Kurve x (t - LAG).

Liegt x(t) im Zeitintervall [t1, t2] vor, dann kann LAG maximal t2-t1 sein, da sonst keine Wertepaare mehr aufeinandertreffen. In der Praxis berechnet man Ak(LAG) maximal bis zum Wert (t2-t1)/2. Ak(LAG) ist eine gerade Funktion, d.h. es gilt Ak(LAG)=Ak(-LAG).

Aus der Lage der Maxima der Autokorrelationsfunktion, d.h. aus dem "LAG" eines Maximums, kann man Periodizitäten der Funktion x(t) ablesen. Wiederholt sich ein Kurvenbild in x(t) nach einer Periode, dann steigt der Korrelationskoeffizient bei einer Verschiebung um diese Periode stark an. Das "LAG" eines Maximums ist gleich der Periodendauer τ der gesuchten Schwingung.

Kreuzkorrelation

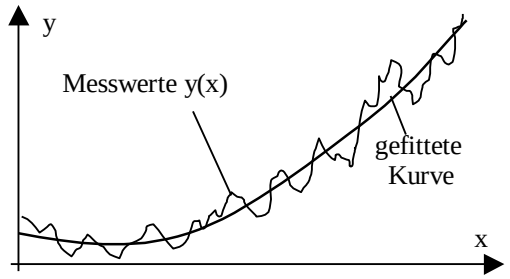
Die Kreuzkorrelationsfunktion Kk(LAG) entsteht wenn man zwei Zeitfunktionen x(t) und y(t) miteinander korreliert und y(t) nach und nach gegenüber x(t) verschiebt. Für jede Verschiebung wird der lineare Korrelationskoeffizient ***r*** berechnet und über der Verschiebung aufgetragen. Dabei ist "LAG" die zeitliche Verschiebung zwischen x(t) und der zeitlich verschobenen Kurve y(t-LAG).

Liegen x(t) und y(t) im Zeitintervall [t1,t2] vor, dann kann LAG maximal t2-t1 (positives LAG oder Linksverschiebung von y(t) gegenüber x(t)) sein bzw. maximal t1-t2 (negatives LAG oder Rechtsverschiebung), da sonst keine Wertepaare mehr aufeinandertreffen. In der Praxis berechnet man Kk(LAG) maximal im Intervall [-(t2-t1)/2 , +(t2-t1)/2]. Kk(LAG) ist im Allgemeinen keine gerade Funktion, d.h. es gilt fast immer Kk(LAG) ungleich Kk(-LAG).

Aus der Lage der Maxima der Kreuzkorrelationsfunktion kk(LAG) lässt sich eine zeitliche Verschiebung zwischen den Kurven x(t) und y(t) ablesen. Wiederholt sich nämlich nach der Zeit dt in y(t) ein Kurvenbild aus x(t), dann steigt für diese Verschiebung die Korrelation an. Das "LAG" eines Maximums ist die gesuchte Verschiebung ***dt***. Ein positives dt (LAG>0) bedeutet, dass y(t) der Kurve x(t) zeitlich nachhinkt. Man muss y(t) nach links verschieben, um die beiden Kurven besser zur Deckung zu bringen. Umgekehrt bedeutet ein negatives *dt* (LAG<0), dass y(t) der Kurve x(t) zeitlich vorauseilt.

Beide Funktionen, die Autokorrelationsfunktion und die Kreuzkorrelationsfunktion, lassen sich relativ leicht in Excel programmieren und dann graphisch darstellen, indem man die gewöhnliche Korrelationsfunktion =KORREL „zieht“.

12.5 Nichtlineare Regression



Die Aufgabe ist, einen Kurventyp zu finden, der die Messwerte ohne ihre zufälligen Schwankungen bestmöglich wiedergibt.
a) Das Glasmodell beruht auf den biologisch-chemisch-physikalischen Grundlagen
b) Die Black-Box verwendet dem Datenverlauf ähnliche Kurven (Parabel, e-Funktion, ...)
c) Eine Kombination aus a) und b)

Bei der Berechnung der Kurvenanpassung unterscheiden wir zwischen linearisierten, quasilinearen und nichtlinearen Modellen.

Beispiel Linearisierung der Exponentialfunktion: Hefewachstum oder Wachstum allgemein ist in seiner Anfangsphase oft durch die Exponentialfunktion $Z(t) = Z_o\, e^{\alpha t}$ darstellbar. Der Wachstumskoeffizient α hat die Dimension $[h^{-1}]$. Z_o ist die Startmenge bei t=0. Logarithmieren der Modellgleichung ergibt $\ln(Z)=\ln(Z_o)+\alpha t$. Durch die Umbenennungen $y=\ln(Z)$, $a= \ln(Z_o)$ und $b=\alpha$ erhalten wir das einfach lineare Regressionsmodell $y=a+bt$. Man schätzt die beiden Koeffizienten a und b und erhält durch die rückwärtigen Ersetzungen $Z_o=e^a$ und $\alpha=b$ die gesuchten Koeffizienten für das nichtlineare Modell. Der Fehler des Anstiegs s_b kann (mit Einschränkungen) direkt als Fehler von α, d.h. als s_α interpretiert werden. Der Fehler der Konstanten s_a aus dem logarithmierten Modell wird zum Multiplikator für den Originalkoeffizienten Z_o, d.h $Z_o+s_Z = Z_o * e^{sa}$ und $Z_o-s_Z = Z_o / e^{sa}$. Man beachte jedoch:

- Die so gefundene Kurve minimiert im Originalplot nicht die Fehlerquadratsumme, sondern nur im logarithmierten Modell
- Die Hypothesenprüfung erfolgt nur am logarithmierten Modell korrekt

Quasilineare Modelle: Man ersetzt x durch eine oder mehrere Funktionen von x. Jede Funktion bildet eine neue Variable, die in ein multiples lineares Regressionsmodell eingesetzt wird:

Das Polynom z.B.	$y = a + b\,t + c\,t^2$	wird ersetzt durch das
quasilineare Modell	$y = b_0 + b_1 X_1 + b_2 X_2$	mit $X_1 = t$ und $X_2 = t^2$

Ein anderes Beispiele für ein quasilineares Modelle ist:

$$y = b_0 + b_1 \sin(c*x) + b_2 \cos(c*x) + b_3 x + b_4 e^{dx}$$

mit $X_1 = \sin(x)$

und $X_2 = \cos(x)$ usw.

und mit den festen Konstanten c und d, deren Werte bekannt sein müssen. Die multiple Regressionsanalyse vermag nur die Werte der sogenannten linearen Koeffizienten b_0, b_1, b_2, ... zu schätzen. Auch hier gilt, dass die Fehlerquadratsumme nicht in allen Fällen minimiert wird.

Vorsicht bei der Verwendung von Polynomen höheren Grades!!! Man kann durch *n* Punkte mit voneinander verschiedenen x-Werten immer exakt ein Polynom (n-1)-ten Grades legen. Eine Gerade durch 2 Punkte, eine Parabel durch 3 Punkte, ein Polynom 9. Grades durch 10 Punkte. Was macht das Polynom zwischen den Punkten? Oft liefert es konfuse Werte! Deshalb die **Empfehlung:** Ist n die Zahl der Messpunkte, dann sollte die Zahl p+1 der benutzten Koeffizienten b_0, b_1, b_2, ..., b_p im Modell immer kleiner als n/2 sein, d.h. doppelt so viele Messpunkte wie Koeffizienten. Benutzt man ein schrittweises Aufbau- bzw. Abbauverfahren oder die CWA-Regression, dann darf die Zahl der Merkmale im Modell beliebig hoch sein. Hier übernimmt das Regressionsprogramm die Auswahl der geeigneten Menge an Merkmalen.

Nichtlineare Modelle: Solver, wie sie in EXCEL z.B. zur Verfügung stehen, können beliebige Kurven an Daten fitten. Hier spielt es keine Rolle, ob die Koeffizienten linear oder nichtlinear in das Modell eingehen. Es kann jedoch sein, dass ein Solver nicht immer eine zulässige Lösung findet. Dann muß man die Startwerte der Koeffizienten ändern. Viele Programme liefern noch die Standardfehler der Koeffizienten, wobei diese Fehlerschätzungen jedoch mit Vorsicht zu genießen sind. Es sind allenfalls Richtwerte für die Fehler der Koeffizienten

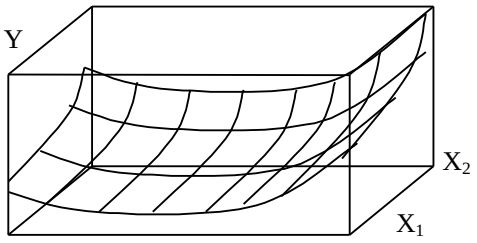
Beispiel eines nichtlinearen Modells ist die Haldane-Kinetik des Bakterienwachstums mit Wachstum μ in Abhängigkeit von der Substratkonzentration c_s und den 3 Koeffizienten μ^*, K_S und K_{SI}. Bei gegebenen Daten μ und c_s berechnet der Solver durch Kurvenfit die 3 Koeffizienten, wobei die Fehlerquadratsumme minimiert wird.	$$\mu(c_s) = \frac{\mu^* \cdot c_s}{K_S + c_s + \frac{c_s^2}{K_{SI}}}$$

12.6 Multiple Regression

Die multiple Regression verknüpft p Einflussgrößen X_1, X_2,..., X_p mit einer Zielgröße Y. Das Modell kann mit oder ohne Regressionskonstante b_o sein:

$$Y = b_o + b_1 * X_1 + b_2 * X_2 + ... + b_p * X_p + e$$

Die geometrische Interpretation ist eine Funktion über dem von X_1, X_2, ... aufgespannten Raum. Die Regressionskoeffizienten b_1, b_2,..., b_p (und b_0) werden nach der Methode der kleinsten Quadrate (Σe^2= Minimum) geschätzt. **e** ist der zufällige Fehler oder Residuum (Abweichung). Beispiel Parameterpreisbildung: Der durchsetzbare Marktpreis eines neu zu entwickelnden Elektromotors soll geschätzt werden. Man benötigt dazu die technischen Kennzahlen des Motors (z.B. Gewicht, Leistung, Drehzahl, Spannung) sowie die verfügbaren Marktpreise und Kenndaten schon verfügbarer ähnlicher Motore. Zielgröße ist der Marktpreis, Einflussgrößen sind die Kenndaten. Das lineare Modell (mit Fehler ***e***) lautet:



$$\text{Marktpreis} = b_o + b_1 * \text{Gewicht} + b_2 * \text{Leistung} + b_3 * \text{Drehzahl} + b_4 * \text{Leistung} + e$$

Die multiple Regression schätzt aus den vorhandenen Preisen und Kennzahlen die Regressionskoeffizienten. Setzt man die "extra Daten" des eigenen geplanten Motors ein, dann erhält man eine Schätzung des Erwartungswertes für dessen Preis.

Zum **quasilinearen Modell** siehe "nichtlineare Regression". Von **gewichteter Regression** spricht man, wenn jedem Datenpunkt ***i*** ein Gewicht G_i zugeordnet wird. Die Zahl der Freiheitsgrade wird dadurch nicht verändert.

Es gibt drei Hauptaufgaben der multiplen Regression:

1. **Prognose** (Vorhersage) von Y-Werten außerhalb des durch die x-Werte vorgegebenen Bereichs und/oder für neue Datenpunkte innerhalb des vorgegebenen X-Bereichs. Beispiele: Parameterpreisbildung, Schätzung der Energiekosten im nächsten Jahr auf der Basis der Produktionszahlen und Verbräuche in den vergangenen Jahren.
2. **Reproduktion** von Y-Werten exakt auf den Punkten des beobachteten X-Bereichs (Stützstellen). Es handelt sich hier um eine reine Datenreduktion (Regressionsparameter statt einzelner Y-Werte). Beispiel: Speicherung der Sicherheitspunkte der t-Verteilung für die Freiheitsgrade FG>5 mit einem Regressionsmodell der Form $t = b_0 + b_1 * (1/FG) + b_2 * (1/FG^2)$
3. **Finden von signifikanten Einflussgrößen** X: Beispiel: Welches sind die Haupteinflussgrößen auf den Ertrag einer neu entwickelten Rapssorte (Temperatur? Regenmenge? Kalk? Stickstoff?, ...)

Entsprechend den Hauptaufgaben sind verschiedene Regressionsalgorithmen zu empfehlen:

- Für Prognose bei hoher Merkmalszahl, wenig vorhandenen Datenpunkten und stark streuenden Zielgrößenwerten Y wird die "CWA-Regression" empfohlen
- Für Prognose (Vorhersage) mit wenigen Einflussgrößen und/oder wenig streuenden Zielgrößenwerten wird das "Schrittweise Aufbauverfahren" oder das "Schrittweise Abbauverfahren" empfohlen

- Für die exakte Reproduktion der Y-Werte an den Stützstellen wird die "Regression mit allen Einflussgrößen" empfohlen.
- Für das Auffinden signifikanter Einflussgrößen wird das "Schrittweise Aufbauverfahren" oder das "Schrittweise Abbauverfahren" empfohlen

Bei der **Prognose** (Vorhersage) interessiert die Genauigkeit der prognostizierten Y-Werte. Nicht die kleinste Reststreuung ist gefragt, sondern der kleinste Prognosefehler. Zur Bestimmung benutzt man **Lachenbruch**-, **Jackknife**- oder **Bootstrap**-Methoden. Die Frage nach der Signifikanz der Einflussgrößen stellt sich nicht bzw. ist untergeordnet.

Die **exakte Reproduktion** (Datenverdichtung) erfordert ein besonders gutes Regressionsmodell, das die Reststreuung bei möglichst wenigen Einflussgrößen minimiert. Bei ungeschickter Wahl des Modells ist es möglich, dass Sie bei X-Werten außerhalb der benutzten Stützstellen Phantasiewerte erhalten, die weit jenseits von gut und böse liegen. Ein Prognosefehler oder die Signifikanz der Einflussgrößen ist von untergeordneter Bedeutung.

Das Auffinden **signifikanter Einflussgrößen** ist oft von wissenschaftlichem oder praktischem Wert. Das Ergebnis kann kausale Zusammenhänge erkennbar machen, oder aber den Erhebungsaufwand fuer Prognosedaten erheblich reduzieren helfen. Probleme beim Auffinden der signifikanten Einflussgrößen sind:

1. Aus einer Gruppe untereinander hoch korrelierter Einflussgrößen wird zumeist nur ein Merkmal mehr oder weniger zufällig ausgewählt. Es ist durchaus möglich, dass eine ganze Reihe von Merkmalen denselben signifikanten Einfluss haben, wie das zufällig gewählte Merkmal. Das kann dann ein Hinweis darauf sein, dass ein versteckter Faktor alle diese Merkmale steuert. Seltener tritt der Faktor selbst als Merkmal auf. Es wäre in diesem Falle sinnvoll, eine Faktoranalyse vorzuschalten und mit den Faktoren als Einflussgrößen zu arbeiten
2. Liegen sehr viele Einflussgrößen vor, dann kann eine Alpha-Adjustierung, z.B. nach Bonferroni vorgenommen werden, da sonst Zufallsmerkmale eine Chance bekommen, als signifikant gemeldet zu werden (bei α=5% würden von 100 Zufallsmerkmalen immerhin 5 akzeptiert werden, wenn keine Alpha-Adjustierung vorgenommen wird.)

"CWA-Regression" ist ein Verfahren, das die Regressionskoeffizienten iterativ nach einem speziellen Abstiegsverfahren zur Minimierung der Reststreuung berechnet (Cierzynski / v.Weber 1989). Die Vorteile sind:

- Hochkorrelierte Merkmale schließen sich nicht gegenseitig aus, sondern werden zu einer Art Faktor gemittelt (man spart die Faktorregression)
- Die Iteration wird abgebrochen, wenn der Prognosefehler wieder ansteigt
- Es entsteht eine "robuste Lösung", die auch bei moderaten Veränderungen in der Datenbasis (X-Werte) noch Bestand hat

"Schrittweises Aufbauverfahren" bzw. "Abbauverfahren": Ein Signifikanztest (t-Test, F-Test) entscheidet über Aufnahme oder Verbleib einer Einflussgröße im Modell. Vorteile sind:

- Nur Einflussgrößen mit einem statistisch gesicherten Einfluss auf die Reduktion der Reststreuung werden in das Modell aufgenommen (Ausnahme: Ist kein Merkmal signifikant, wird das mit dem höchsten t-Wert genommen)
- Eine Gruppe hoch korrelierter Merkmale wird durch ein Merkmal vertreten
- Es entsteht eine robuste Lösung, die auch bei moderaten Veränderungen in der Datenbasis noch Bestand hat

"Regression mit allen Einflussgrößen" ist ein Verfahren, bei dem nur Merkmale aus dem Modell entfernt werden, wenn eine so starke lineare Abhängigkeit der Merkmale diagnostiziert wird, dass numerische Instabilitäten auftreten. Der Vorteil ist: Für die Stützstellen (und nur für diese) lässt sich die Reststreuung maximal minimieren. Es hängt sehr vom Modell ab, ob die Zielgrößenschätzung auch für Werte außerhalb der Stützstellen noch vernünftige Zahlen liefert. Am besten testet man dieses aus, indem man selbst einmal die X-Werte leicht variiert und in das berechnete Modell einsetzt.

Polynomiale Standardmodelle sind: *Einfach Polynomial*: Zu jedem im Modell vorhandenen X_i-Merkmal wird bei Polynomgrad PG=2 ein X_i^2-Merkmal zusätzlich erzeugt, bei Polynomgrad PG=3 ein Merkmalspaar X_i^2und X_i^3 zusätzlich erzeugt usw.
Vollständig Polynomial: Wie einfach Polynomial, aber zusätzlich noch alle Produkte der vorhanden X-Merkmale, z.B. bei PG=2 und X_1, X_2 entstehen zusätzlich X_1^2, X_2^2, X_1*X_2, bei PG=3 und X_1, X_2 entstehen zusätzlich X_1^2, X_1^3, X_2^2, X_2^3, X_1*X_2, $X_1^2*X_2$, $X_1* X_2^2$

Erklärung der von der multiplen Regression benutzten und berechneten Größen:

Y	Das Zielgrößenmerkmal
X j	Einflussgrößenmerkmale (j = 1, 2, 3,...p) mit p= Einflussgrößenzahl
n	Auswertbare Punktzahl (Datensätze ohne Ausfall)
B	Multiples Bestimmtheitsmaß (multiples R^2), ein Maß für die Verbesserung der Vorhersage durch Kenntnis von X_1, X_2, ..., X_p. Es ist $0 \le B \le 1$. $B= SAQ_{Reg} / SAQ_{Rest}$. Dabei ist SAQ_{Reg} die Summe der Abweichungsquadrate aus Erwartungswerten und Mittelwert ($\sum(\hat{y}_i - \bar{y})^2$) und SAQ_{Rest} ist die Σe^2.
F	F-Testwert für R^2 bzw. B Die Nullhypothese ist Ho: B=0 (Kein modellmäßiger Zusammenhang zwischen Y und den Xj nachweisbar) mit F=B(n–k)/(1-B) und mit FG_1=p und FG_2=n–k, k= Koeffizientenzahl einschließlich des b_o.
FG	FG=N-k, Freiheitsgrad der Reststreuung, k= Koeffizientenzahl einschließlich des b_o.
KIW(B)	Die Irrtumswahrscheinlichkeit bei einseitigem Test für die Ablehnung der Nullhypothese Ho:B=0 (H_A:B>0)
b_j	Koeffizient Der Zahlenwert des Regressions-Koeffizienten
s_{bj}	Stdabw. Die geschätzte Standardabweichung des Koeffizienten,
t_j	T-Wert t-verteilte Prüfgroeße zum Test der Nullhypothese Ho: bj =0 (Koeffizient bj in der Grundgesamtheit Null?)
p-Wert	KIW Kritische Irrtumswahrscheinlichkeit bei zweiseitigem Test fuer die Ablehnung der Nullhypothese Ho:bj=0 (HA:bj<>0)
S_R	Reststreuung oder mittleres Residuum (mittlerer Fehler **e**)
Sj	Prognosefehler/Vorhersagefehler **e** nach Jackknife-Methode geschätzt
Sw	Prognosefehler/Vorhersagefehler **e** mit Arbeitsstichprobe (working sample) geschätzt

Multiple lineare Regression: Modell $Y = b + m_1X_1 + m_2X_2 + ... + m_qX_q + e$
Dabei ist Y die Zielgröße, X_1 bis X_q die q Einflussgrößen, e das Residuum (Abweichung), b ist die Regressionskonstante, m_1 bis m_q die Regressionskoeffizienten.

Die rechte Tabelle zeigt einen Ausschnitt aus einer EXCEL-Tabelle mit den Spalten A,B,C,... und den Zeilen 1,2,...

	A	B	C	D
Zeile1	Preis	Drehz.	Spann.	Gewicht
Zeile2	1400	1400	380	240

Wir wollen z.B. die multiple Regression ohne Konstante b berechnen:

$P = m_1D+m_2S+m_3G$

Wir markieren eine Matrix mit immer 5 Zeilen (hier ab Zeile 9) und soviel Spalten, wie Koeffizienten zu berechnen sind (hier 3 Spalten). Konstante b würde bei den Spalten mitzählen, hätten wir sie gewünscht. Wir geben über die Tastatur die Anweisung *=rgp(a2:a7;b2:d7;falsch;wahr)* und die Tasten-Kombination Strg-Shift-Enter. *a2:a7* bezeichnet hier die Zielgröße *Preis*, *b2:d7* bezeichnet hier die drei Einflussgrößen, *falsch* legt fest, dass die Konstante b entfällt, *wahr* legt fest, dass zusätzliche Statistiken (Fehler der Koeffizienten usw. erscheinen.) Siehe auch HELP-Möglichkeit von EXCEL. EXCEL berechnet in Zeile 9 die Koeffizienten in der Reihenfolge m_3,m_2,m_1. In Zeile 10

Zeile3	3800	2000	600	900
Zeile4	1850	2800	380	180
Zeile5	4450	12000	380	95
Zeile6	5900	1200	600	1800
Zeile7	22500	600	15000	3250
Zeile8				
Zeile9	2,796	0,881	0,323	
Zeile10	0,0549	0,0139	0,008	
Zeile11	0,999	112,6		
Zeile12	.....	3		
Zeile13	.....	.....		

stehen die Fehler der Koeffizienten s_{m3}, s_{m2}, s_{m1}. Zeile 11 liefert das Bestimmtheitsmaß r^2 und die Reststreuung (mittlere Abweichung). Zeile 12 enthält den Freiheitsgrad zur Reststreuung und wird für eventuelle t-Tests benötigt, die Sie zu den Koeffizienten durchführen wollen. Der Rest ist hier unwichtig.

13. Varianzanalyse

13.1 Einfache Varianzanalyse

Die Varianzanalyse ist eine statistische Methode zur Beurteilung gruppierter metrischer oder ranggeordneter Daten. Die Gruppierung erfolgt mit Hilfe eines kategorialen oder nominalen Merkmals. Das Gruppierungsmerkmal wird in der Literatur meist "**Faktor**" genannt. Seine Werte werden als "Faktorstufen" bezeichnet und im Computer als ganze Zahlen behandelt (kategoriales Merkmal). Die Werte des metrischen Merkmals zu einer Faktorstufe bilden eine Gruppe. Die Literatur (**EISENHART**) unterscheidet zwei Modelle der Varianzanalyse:

Modell 1 (**festes Modell**): Die Gruppierung der Daten ist durch den Versuchsplan vorgegeben. Hier interessiert, ob Mittelwertunterschiede zwischen den Gruppen existieren. Beispiel: Reißfestigkeit eines textilen Gewebes [N/m] nach 14-tägiger Exposition mit UVB-Bestrahlung in Abhängigkeit von einer Oberflächenbeschichtung. Zielgröße ist die Reißfestigkeit, Faktor die Beschichtung mit den Stufen 1="unbeschichtet", 2="8 g Al/m^2", 3="16 g Al/m^2".

Modell 2 (**zufälliges Modell**): Die Gruppierung wird beobachtet, ist also zufällig. Hier interessiert, ob die Werte der Zielgröße innerhalb der Gruppen stärker oder schwächer streuen, als die Gruppenmittelwerte untereinander. Beispiel: Eine Herde Kühe wird in Gruppen eingeteilt. Gruppierungsmerkmal ist der Vater. Kühe vom gleichen Vater bilden eine Gruppe. Faktor ist Merkmal "Vater" mit den Stufen 1="Anton", 2="Bogumil", 3=... Zielgröße ist die Jahresmilchleistung, die eine Kuh bringt. Streuen diese Werte in den Gruppen weniger, als die Mittelwerte zwischen den Gruppen, dann vermutet man einen genetischen Einfluss des Vaters auf die Milchleistung durch Vererbung eines "Milchleistungsgens".

Globaler Test: Ist die Varianz zwischen den Gruppen signifikant größer, als die Varianz innerhalb der Gruppen, d.h. gibt es einen signifikanten Einfluss des Faktors bzw. signifikante Unterschiede in den Gruppenmitteln? Die Formeln werden hier nur kurz angedeutet:

Xges = Mittelwert aller n beteiligten x-Werte (metrisches Merkmal)
Xi = Mittelwert der x-Werte aus Gruppe i, i=1,..g, g = Gruppenzahl

SAQges = Summe der Abweichungsquadrate der x-Werte von Xges
SAQinn = Summe der Abw.quadrate innerhalb der Gruppen, d.h. über alle
g Gruppen die Abweichungsquadrate der x von ihrem zuständigen Xi
SAQzwi = SAQges-SAQinn, die Summe der Abweichungsquadrate zwischen den Gruppen

MQzwi = SAQzwi/FGzwi Mittleres Quadrat mit Freiheitsgrad FGzwi=g-1
MQinn = SAQinn/FGinn Mittleres Quadrat mit Freiheitsgrad FGinn=n-g

Fgl = MQzw/MQinn globaler Fwert mit Freiheitsgraden (FGzwi, FGinn)

KIWgl = Kritische Irrtumswahrscheinlichkeit, d.h. die Wahrscheinlichkeit, dass dieser oder größere F-Werte auftreten können unter der Nullhypothese (kein Einfluss des Faktors). Der Wert wird mit der vorgegebenen Irrtumswahrscheinlichkeit α verglichen (einseitiger F-Test, da ein negativer Faktoreffekt ohne Natur- oder Datenmanipulation nicht möglich ist).

Die Mittelwertvergleiche erfolgen paarweise, jedes Xi mit jedem Xj, d.h. es werden h=g(g –1)/2 Einzelhypothesen getestet. Der Einzeltest erfolgt mit

$Fij = ((Xi-Xj)^2 * n1*n2) / (MQinn*(n1 + n2))$

Einzel-F-Werte mit den Freiheitsgraden (1, FGinn)

KIWij= Kritische Irrtumswahrscheinlichkeiten. Sie werden mit α^* verglichen (entspricht in dieser Konstellation dem zweiseitigen t-Test). α^* ist zumeist das adjustierte α nach Bonferroni.

Ein Sonderfall: Nach PERLI kann man den ersten Test (Maximales F=F1) auch durch den globalen Test ersetzen. Zeigt der globale F-Test einen signifikanten Faktoreinfluss an, dann wird der am höchsten bewertete Mittelwertunterschied ebenfalls als signifikant erachtet, auch wenn der Einzeltest ihn verwerfen sollte.

13.2 Varianzanalyse mit Kreuzklassifikation und Mittelwertvergleichen

Wir behandeln hier nur die Kreuzklassifikation mit festen Effekten. Die Beobachtungen sind nach 2 Merkmalen klassifiziert. Wir haben in unserem Beispiel r = 4 Zeilen (4 A-Klassen) und s = 3 Spalten (B-Klassen). Es liegt „einfache Besetzung“ vor, d.h. jede Zelle ist nur mit einer Zahl besetzt. Die 4 A-Klassen sind 4 Agarplatten, die mit bacterium subtilis (Heubakterium) beimpft wurden. Die 3 B-Klassen sind 3 Proben an Penicillinlösungen unterschiedlicher Herkunft. Die Messwerte x_{ij} sind die Breiten der Wirkungshöfe um die 3 ausgestanzten Löcher in den Agarplatten, in die die 3 Proben getropft wurden. Studiert werden sollen hier die Platteneffekte und die Effekte der unterschiedlichen Penicillinlösungen.

Zuerst bestimmen wir die Randsummen $S_{i.}$ und $S_{.j}$, dann die Gesamtsumme S..

	Probe 1	Probe 2	Probe 3	Summe	Mittel
Platte 1	**x_{11} = 27**	**x_{12} = 21**	**x_{13} = 34**	**$S_{1.}$ = 82**	27,33
Platte 2	**X_{21} = 26**	**X_{22} = 19**	**X_{23} = 31**	**$S_{2.}$ = 76**	25,33
Platte 3	**X_{31} = 27**	**X_{32} = 18**	**X_{33} = 34**	**$S_{3.}$ = 79**	26,33

Platte 4	**X_{41} = 29**	**X_{42} = 22**	**X_{43} = 33**	**$S_{4.}$ = 84**	28,00
Summe	$S_{.1}$ = 109	$S_{.2}$ = 80	$S_{.3}$ = 132	S.. = 321	**Ges = 26,75**
Quadratsumme	Σx_{i1}^2 = 2975	Σx_{i2}^2 = 1610	Σx_{i3}^2 = 4362	Σx_{ij}^2 = 8947	
Mittel	27,25	20,00	33,00		

Aus den Randsummen lassen sich dann leicht die Zeilenmittel $\mu_{i.}$, die Spaltenmittel $\mu_{.j}$, und das Gesamtmitttel $\mu_{..}$ bestimmen. Die Quadratsummen für die Berechnung der totalen Quadratsumme über alle Werte x_{ij} werden hier mit den Spalten gebildet (mit den Zeilen wäre ebenso möglich).
Das Modell der Varianzanalyse fordert hier, dass die x_{ij} mit $N(\mu_{ij}, \sigma_e^2)$ verteilt sind, d.h. normalverteilt mit Mittelwert μ_{ij} und Varianz σ_e^2. Das μ_{ij} ist der Erwartungswert für die Hemmhofbreite auf Platte i um die Probenlösung j herum. Um die Effekte (hier Unterschiede der mittleren Hofbreiten) zu illustrieren, zerlegt man den Beobachtungswert folgendermaßen:

$$\mu_{ij} = \mu_{..} + (\mu_{i.} - \mu_{..}) + (\mu_{.j} - \mu_{..}) + (\mu_{ij} - \mu_{i.} - \mu_{.j} + \mu_{..})$$

oder

$$\mu_{ij} = \mu_{..} + \alpha_i + \beta_j + e_{ij}$$

Die Summe der Effekte α_i und β_j ist einzeln und auch zusammen immer 0. Die Versuchsfehler e_{ij} sind mit $N(0, \sigma_e^2)$ normalverteilt, d.h. mit Mittelwert 0 und Varianz σ_e^2. Unser Ziel ist es, den Versuchsfehler σ_e zu bestimmen und danach die einfachen linearen Kontraste $\mu_1 - \mu_2$, $\mu_1 - \mu_3$ und $\mu_2 - \mu_3$ auf $\neq 0$ zu testen (Unterscheiden sich die 3 Penicillinproben in ihrer Wirkung signifikant?)

Mathematisch lässt sich eine Summe von Abweichungsquadraten zerlegen, z.B. :

$$\sum_{j=1}^{s}\sum_{i=1}^{r}\left(x_{ij} - \bar{x}_{..}\right)^2 = \sum_i s\left(\bar{x}_{i.} - \bar{x}_{..}\right)^2 + \sum_j r\left(x_{.j} - \bar{x}_{..}\right)^2 + \sum_j\sum_i\left(x_{ij} - \bar{x}_{i.} - \bar{x}_{.j} + \bar{x}_{..}\right)^2$$

oder

$$SAQT = SAQA + SAQB + SAQR$$

mit

$$SAQT = \sum_j\sum_i x_{ij}^2 - \frac{S_{..}^2}{n} \quad mit \quad S_{..} = \sum_j\sum_i x_{ij} \quad und \quad n = r \cdot s \ ,$$

$$SAQA = \sum_{i=1}^{r}\frac{S_{i.}^2}{s} - \frac{S_{..}^2}{n} \quad mit \quad S_{i.} = \sum_j x_{ij} \ ,$$

$$SAQB = \sum_i \frac{S_{.j}^2}{r} - \frac{S_{..}^2}{n} \quad mit \quad S_{.j} = \sum_i x_{ij} \ ,$$

$$SAQR = SAQT - SAQA - SAQB$$

Jetzt können wir die Tafel der Varianzanalyse aufstellen mit den aufgeteilten Summen der Abweichungsquadrate (*SAQ*), den zugehörigen Freiheitsgraden (*FG*), den mittleren Quadraten (*MQ*) und den Erwartungswerten der MQ (*E(MQ)*) :

Variabilitätsursache	SAQ	FG	MQ	E(MQ)
Varianz zwischen A-Klassen	SAQA	FGA = r–1	$MQA = \frac{SAQA}{r-1}$	$\sigma_e^2 + \frac{s}{r-1}\sum_i \alpha_i^2$

Varianz zwischen B-Klassen	SAQB	FGB = s–1	$MQB = \frac{SAQB}{s-1}$	$\sigma_e^2 + \frac{r}{s-1}\sum_j \beta_j^2$
Restvarianz	SAQR	FGR = (r–1)(s–1)	$MQR = \frac{SAQR}{(r-1)(s-1)}$	σ_e^2

Für einen globalen Test auf Homogenität z.B. der B-Mittelwerte (hier der Penicillinproben) kann man den F-Test nehmen:

$$F_{\alpha, FG1=s-1, FG2=(s-1)(r-1)} = \frac{MQB}{MQR}$$

Wir nehmen Homogenität der Mittelwerte an (Hypothese H_o bzw. *keine signifikanten Mittelwertunterschiede zwischen den B-Klassen*), wenn $F < F\alpha$. Wir nehmen die Existenz wenigstens eines signifikanten Mittelwertunterschiedes an, wenn $F \geq F\alpha$ ist.

Als linearen Kontrast bezeichnet man eine Linearkombination aus Mittelwerten μ_i und Konstanten c_i der Form

$$K = c_1\mu_1 + c_2\mu_2 + \ldots + c_k\mu_k \qquad mit \qquad \sum_{i=1}^{k} c_k = 0.$$

Wir betrachten hier nur die einfachsten Kontraste, die möglich sind, nämlich die für k = 2. Sie haben die Form $K_{ij} = \mu_i - \mu_j$ mit $c_i = 1$ und $c_j = -1$. Für die Signifikanzprüfung der einzelnen B-Klassen-Kontraste können wir einen F-Test, aber auch den t-Test nehmen.

$$t_{ij} = \frac{(\bar{x}_i - \bar{x}_j)}{\sqrt{MQR}}\sqrt{\frac{r \cdot r}{r + r}} \qquad \text{mit FGR Freiheitsgraden und } \sqrt{MQR} = \sigma_e .$$

Wollen wir die A-Klassen-Mittelwerte testen, dann heißt die Formel

$$t_{ij} = \frac{(\bar{x}_i - \bar{x}_j)}{\sqrt{MQR}}\sqrt{\frac{s \cdot s}{s + s}} \qquad \text{mit FGR Freiheitsgraden.}$$

Die Tabelle der Varianzanalyse zu unserem Zahlenbeispiel der r = 4 Agarplatten und s = 3 Penicillinproben lautet

SAQT = 360,25	FGT = n –1 = 12 –1 = 11	
SAQA = 12,25	FGA = r – 1 = 4 – 1 = 3	MQA = SAQA/FGA = 4,083
SAQB = 339,50	FGB = s – 1 = 3– 1 = 2	MQB = SAQB/FGB = 169,75
SAQR = 8,50	FGR = (r–1)(s–1) = 6	MQR = SAQR/FGR = 1,4166 = σ_e^2

Der erste t-Test prüft den Kontrast zwischen den Penicillinproben 1 und 2.

$$t_{12} = \frac{(27{,}25 - 20)}{\sqrt{1{,}4166}}\sqrt{\frac{4 \cdot 4}{4 + 4}} = 8{,}61 \qquad \text{mit Freiheitsgrad FG = 6.}$$

Wir akzeptieren Hypothese H_o (kein signifikanter Unterschied), wenn $|t_{ij}| < t_\alpha$.
Wir akzeptieren Hypothese H_A (signifikanter Unterschied), wenn $|t_{ij}| \geq t_\alpha$.
Wegen $t_{\alpha = 5\%, FG = 6, zws.} = 2{,}45$ akzeptieren wir hier Hypothese H_A für den ersten Vergleich. Probe 1 hat mit 27,25 mm Breite des Wirkungsbereichs einen signifikant besseren Effekt als Probe 2 mit nur 20 mm. Ebenso berechnen sich die Vergleiche der Proben 1 und 3

sowie 2 und 3. Die t-Werte sind $t_{13} = -6{,}83$ und $t_{23} = -15{,}44$. Auch diese Unterschiede sind signifikant.

Will man nicht drei Einzelhypothesen testen, sondern eine multiple Hypothese, dann bieten sich die Bonferroni-Adjustierung der Irrtumswahrscheinlichkeit α an, oder aber z.B. Holms sequentielle Prozedur. Für beide Verfahren benötigt man jedoch t-Tafeln, die die Sicherheitspunkte zum adjustierten α enthalten. Bei drei Mittelwerten lautet die Bonferroni-Adjustierung $\alpha^* = \alpha/3 = 0{,}05 / 3 = 0{,}01667$. Für ein solches α ist unsere Tafel am Anfang des Skripts nicht eingerichtet. Excel liefert jedoch mit der Funktion =TINV (Irrtumswahrscheinlichkeit ; Freiheitsgrade) die zweiseitigen Sicherheitspunkte t_α für beliebige α.

Interessieren z.B. die Unterschiede der A-Gruppen selbst nicht sonderlich, kann man durch eine ***ipsative Skalierung*** aus der Kreuzklassifikation eine einfache Klassifikation machen. Man berechnet zuerst, wie im obigen Beispiel, alle 4 A-Mittel , d.h. $\bar{x}_{1.} = 27{,}33$, $\bar{x}_{2.} = 25{,}33$ usw. Dann subtrahiert man die berechneten A-Mittel von allen Werten der jeweiligen Gruppe, z.B. $x'_{1,1} = x_{1,1} - \bar{x}_{1.} = 27\text{-}27{,}33 = -0{,}33$, $x'_{1,2} = 21\text{-}27{,}33 = -6{,}33$, usw. Anschließend macht man mit den x'-Daten eine einfache Varianzanalyse. Man hat jetzt nur noch 3 Gruppen (B_1, B_2, B_3), dafür hat jede Gruppe jetzt 4 Wiederholungen. Freiheitsgrade für die Restvarianz σ_e^2 gewinnt man dadurch nicht, da wir ja die 4 subtrahierten Gruppenmittel aus den Daten selbst berechnet haben.

Der Vorteil der ***ipsativen Skalierung*** liegt hier in der Anschaulichkeit der neu skalierten Daten. Ein Beispiel aus der Landwirtschaft soll das illustrieren: Drei Düngungsarten (Naturdünger fest, Naturdünger flüssig, Industriedünger als Pulver) werden in 8 Jahren auf Parzellen eines Ackers ausgebracht und die Erträge x_{ij} gemessen. Faktor A seien die Jahre i=1, 2, ..., 8. Faktor B die Düngungsarten j=1, 2, 3. Da Jahre sehr unterschiedlich ausfallen (zu nass, zu trocken, zu warm, zu kalt), überdeckt ihr Einfluss den weniger starken Einfluss der Düngungsmethode. Durch die Subtraktion der Jahresmittel treten jedoch die Unterschiede im Ertrag aufgrund der unterschiedlichen Düngung plötzlich stark hervor und fallen sofort ins Auge.

14. Klassifikation

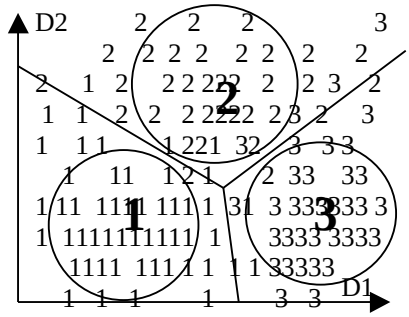
Die Diskriminanzanalyse kann Objekte klassifizieren, wenn Lernobjekte mit bekannter Klasseneinteilung zur Verfügung stehen. Die Clusteranalyse versucht bei völlig ungruppierten Daten eine Klasseneinteilung zu finden.

14.1 Lineare Diskriminanzanalyse

Die Diskriminanzanalyse hat folgende Hauptaufgaben:

1. Klassifikationsregeln für Objekte auf der Grundlage einer Lernstich probe bereits klassifizierter Lernobjekte aufstellen und den zu erwartenden Klassifikationsfehler schätzen.
2. Klassifikation neuer Objekte (Arbeitsobjekte) mit den aufgestellten Klassifikationsregeln durchführen und graphisch oder tabellarisch darstellen
3. Aufsuchen von signifikanten Trennmerkmalen zur Reduktion des Erfassungsaufwandes von Klassifikationsdaten
4. Test auf multivariate Mittelwertunterschiede zwischen Objektklassen.
5. Test auf Isoliertheit von Objektklassen, insbesondere in Verbindung mit der Clusteranalyse.

Das Bild zeigt den Diskriminanzraum, der von den Diskriminanzmerkmalen D1 und D2 aufgespannt wird. Die Klassen 1, 2, 3 haben in dieser Projektion die Form von Kreisen. Trennlinien teilen die Klassengebiete ab. Die Objekte (die kleinen Ziffern) können nicht immer sauber ihrer Klasse zu-geordnet werden. Die Minimierung des Klassifikationsfehlers ist eines der Hauptziele des Anwenders. Der Diskriminanzraum hat die Dimension d=k−1, wenn k die Klassenzahl ist (hier d=2)



Beispiel Klassifikation: Ein Computerprogramm soll lernen, die GC-Kurven (Gas-Chromatographie) von 10 verschiedenen Farbstofflösemitteln sicher zu unterscheiden. Man gibt von jeder GC-Kurve 10 bis 30 Werte aus charakteristischen Bereichen des Chromatogramms als Datensatz ein. Um die Redundanz zu verbessern, nimmt man pro Lösemittel mindestens 10 Chromatogramme unterschiedlicher Proben desselben Lösemittels.

Die lineare Diskriminanzanalyse berechnet aus den 10-30 originalen GC-Merkmalen ein oder mehrere Diskriminanzmerkmale sowie die Klassengrenzen. Die Klassengrenzen trennen im Diskriminanzraum, der von den Diskriminanzmerkmalen aufgespannt wird, die Klassen voneinander ab. Ein neues Chromatogramm ergibt einen Punkt im Diskriminanzraum. Man nimmt die Klasse an, in deren Gebiet der neue Punkt liegt. Auf diese Weise ist das Lösemittel über seine GC-Werte automatisch klassifizierbar.

Beispiel Merkmalsauswahl: Für die Klassifikation von Produktionsfehlern soll die Anzahl der Messpunkte aus Zeitgründen minimiert werden. An einer Stichprobe von Teilen mit bekannten Fehlern werden genügend viele Messungen gemacht, um jeden Fehler eindeutig klassifizieren zu können. Jetzt erfolgt eine automatische Reduktion der Merkmale auf die wesentlichen Diskriminanzmerkmale, d.h. die Merkmalsmenge, die gerade noch eine sichere Klassifikation erlaubt. Das "schrittweise Verfahren" nimmt nur signifikante Trennmerkmale auf. Sind Merkmale hoch korreliert, wird zumeist nur ein Merkmal der unter sich hoch korrelierten Gruppe mehr oder weniger zufällig ausgewählt.

Beispiel multivariater Mittelwertvergleich: Unterscheiden sich Neugeborene aus Großstädten von Neugeborenen aus ländlichen Gebieten. Zu jedem Neugeborenen werden Messdaten erhoben, z.B. Gewicht, Länge, Temperatur usw., aber auch die Herkunft (Großstadt oder ländlicher Raum). Das Programm berechnet den ***Mahalanobisabstand*** der beiden Klassen (Stadt / Land), eine Art gewichteter Mittelwertabstand über **alle gemessenen Merkmale**, und testet diesen Abstand auf Signifikanz.

Entsprechend den Hauptaufgaben sind verschiedene Diskriminanzalgorithmen zu empfehlen.

- Bei hoher Merkmalszahl und wenig Lernobjekten wird immer das **schrittweise Aufbauverfahren** empfohlen. Ebenso bei der Suche nach signifikanten Trennvariablen. Als Alternative gibt es das Abbauverfahren. "Schrittweises Aufbauverfahren" ist ein Verfahren, bei dem ein Signifikanztest (F-Test) über die Aufnahme oder den Verbleib einer Trennvariablen im Modell entscheidet
- Bei wenig Merkmalen und vielen Lernobjekten wird die **Analyse mit allen Merkmalen** empfohlen. (Nur extrem hoch korrelierte Merkmalsgruppen werden ausgedünnt.)

Bei der **Klassifikation** neuer Objekte interessiert die Genauigkeit der prognostizierten Klassifikation. Nicht der kleinste Reklassifikationsfehler ist gefragt, sondern der kleinste Klassifikationsfehler bei neuen, bislang noch nicht klassifizierten Objekten. Zur Bestimmung benutzt man Jack-knife- oder Bootstrap-Methoden bzw. eine Teilung der vorhandenen Daten in einen

Lern- und einen Arbeitsteil. Die Frage nach der Signifikanz der Trennvariablen stellt sich nicht bzw. ist untergeordnet.

Multiple multivariate Mittelwertvergleiche: Es wird ein globaler F-Test ausgeführt (Ahrens/Laeuter S.106, Gl.7.12) Er zeigt an, ob es insgesamt "irgendwelche" multivariate Mittelwertunterschiede gibt. Der Simultanvergleich jeder Klasse i gegen jede andere Klasse j führt auf eine Matrix Fij von F-Werten, die mit einem gemeinsamen Sicherheitspunkt Fsim verglichen werden. Gleichzeitig mit dem Mittelwertvergleich wird auch ein paarweiser Test auf Isoliertheit (Ahrens/Laeuter S.138, Gl. 7.73) der Klassen durchgeführt. Nicht isolierte Klassen lassen sich schlecht trennen. Im Zusammenhang mit der Clusteranalyse sind sie ein Indiz für eine mögliche Klassenzusammenlegung.

Klassifikationsstrategien: ***Ohne Apriori-Wahrscheinlichkeit***: Die Einordnung in eine Klasse ist im Diskriminanzraum nur vom Quadrat k des Euklidischen Abstands des Objektes zum nächstgelegenen Klassenmittel abhängig, abgesehen von einem Faktor Nj/(Nj+1), der sich kaum von 1 unterscheidet. Nj ist der Klassenumfang (Objektzahl) der ausgewählten Lernklasse). ***Mit Apriori-Wahrscheinlichkeit***: Die Einordnung in eine Klasse ist im Diskriminanzraum sowohl vom Quadrat k des Euklidischen Abstands als auch von der Wahrscheinlichkeit Pj der Klasse abhängig. (Ahrens/Laeuter S. 131, Gl. 7.63). Als Apriori-Wahrscheinlichkeit wird die relative Häufigkeit in den Klassen der Lerndaten genommen. Eine große Lernklasse hat automatisch eine größere Wahrscheinlichkeit, dass benachbarte Objekte ihr zugeordnet werden. Wann man ohne oder mit Apriori-Wahrscheinlichkeit arbeitet, dafür gibt es kein Rezept. Richtschnur ist nur die Güte der Klassifikation neuer Objekte, die durch die Fehlerschätzung bewertet wird.

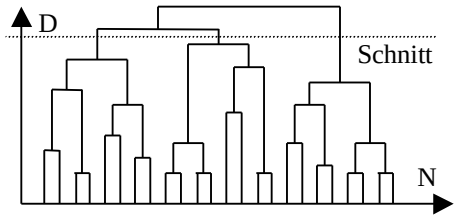
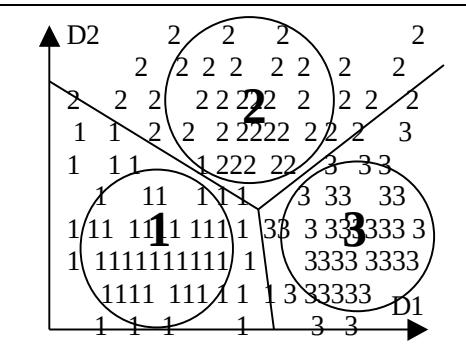
Fehlerschätzung der Klassifikation: Wird ein Objekt einer falschen Klasse zugeordnet, liegt ein Klassifikationsfehler vor. Wir unterscheiden:

- **Reklassifikationsfehler:** Die Objekte der Lernstichprobe werden reklassifiziert, d.h. einer Klasse zugeordnet. Mit steigender Merkmalszahl p nimmt dieser Fehler ab. Man darf sich davon jedoch nicht täuschen lassen. Bei einer Klassifikation von Objekten, die nicht in der Lernstichprobe waren, werden diese desto schlechter klassifiziert, je mehr *unnötige* Merkmale verwendet werden.
- **Jackknife-Fehler:** Die Lernstichprobe wird in viele zufällig ausgewürfelte Teile unterteilt (meist 10). Neun werden als Lernstichprobe benutzt für die Merkmalsauswahl, dann werden die Objekte der 10. Teilstichprobe klassifiziert. Das Ganze wird 10 mal durchgeführt, bis jede Teilstichprobe einmal klassifiziert wurde. Diese Art der Fehlerschätzung ist recht realistisch, was die Fehlerrate bei völlig neuen Objekten betrifft.
- **Working-Sample Fehler:** Hat man sehr viele Daten, kann man die Daten in Lern- und Arbeitsdaten teilen. An der Lernstichprobe wird die Merkmalsauswahl vorgenommen, an der Arbeitsstichprobe wird die richtige Klassifikation überprüft. Diese Art der Fehlerschätzung ist die realistischste, was die Fehlerrate bei völlig neuen Objekten betrifft.

Datenaufbau für eine lineare Diskriminanzanalyse: Sie benötigen eine kategoriale **Zielvariable** Y mit den Klassennummern und eine oder mehrere **Trennvariablen** Xj. Diese können metrisch, binär oder ranggeordnet sein. Es können aus den eingelesenen Trennvariablen Xj durch Potenzieren und/ oder Multiplikation weitere Trennmerkmale gewonnen werden (**polynomiale Modelle**). Ein kategoriales X-Merkmal mit k Kategorien muss durch eine Datentransformation in k-1 binäre Merkmale umcodiert werden (Beispiel Merkmal Haarfarbe mit den 3 Kategorien: K1=schwarz, K2=rot, K3=blond muss in zwei binäre Merkmale umcodiert werden: M1=schwarz/nichtschwarz, M2=rot/nichtrot)

14.2 Clusteranalyse

Hat man keinerlei Vorstellung, wie sich Daten strukturieren, dann versucht man mit der Clusteranalyse eine erste Klassenstruktur zu erzeugen. Es ist wie der Blick in den Sternhimmel, an dem der Mensch "Figuren" zu erkennen sucht. Ob sich so gefundene Klassen später als wertvoll erweisen, muss dann eine nachfolgende Analyse der Eigenschaften der Objekte, die in eine Klasse "geworfen" wurden, klären. Es gibt zwei prinzipiell verschiedene Clusterungsstrategien:

Hierarchische Methoden: Diese erzeugen ein Dendrogramm (Baumstruktur), indem sie die N Objekte nach ihrer Distanz D im mehrdimensionalen Merkmalsraum ordnen. Sich nahestehende Objekte wandern in eine Klasse. Durch einen Schnitt in geeigneter Höhe kann man k Klassen erzeugen (Hier k=3 Klassen)	
Partitionierende Methoden: Man sucht "Kondensationskeime", d.h. Objekte mit vielen anderen Objekten drum herum und baut sie zu Klassen aus. Durch Austausch werden störende Objekte an benachbarte Klassen abgegeben. Ziel ist eine Klasse ohne Ausreißer und etwa von Kugelform. Die graphische Darstellung erfolgt dann mit den Mitteln der Diskriminanzanalyse. Zunächst gibt es keine Fehlzuordnungen, da ja eine Definition der Klassen noch völlig offen ist.	

Allen Methoden gemeinsam ist, dass der Anwender eine gewisse Vorstellung von der Anzahl der Klassen haben sollte, die er erwartet. Weiterhin ist allen Methoden gemeinsam, dass sie immens viel Computerzeit *verbraten*. Die **Eingangsdaten** sind Merkmalsvektoren - je einen pro Objekt. Die Merkmale unterliegen denselben Einschränkungen, wie die Trennmerkmale der Diskriminanzanalyse: Nur metrische, binäre oder rangeordnete kategoriale Merkmale sind zugelassen. Das **Ergebnis** der Clusteranalyse ist eine Klassennummer für jedes Objekt und einige Kennzahlen zu den Klassen (Mittelwert, Klassenumfang usw. Ab hier kann man dann auf die Diskriminanzanalyse zurückgreifen.

14. 3 Logistische Regression

Die logistische Regression teilt die Objekte in genau 2 Klassen (y=0 und y=1). Der Wert 0 oder 1 für y wird ähnlich wie bei der Regression aus p Einflussgrößen x_1, x_2, ..., x_p geschätzt. Beispiel: y= Karies 0/1, x_1=Wasserfluoridierung, x_2=Anteil % Zucker in der Ernährung.

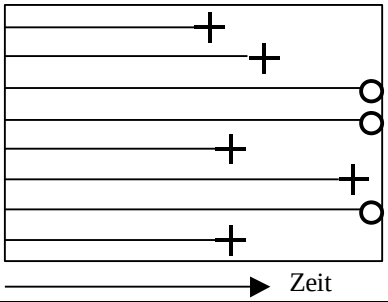
Variable y folgt einer **Bernoulli-Verteilung** mit $P(y=r) = p^{r}(1-p)^{1-r}$ und r = 0 / 1. Erwartungswert ist E(y)=p, Varianz Var(y)= $\sigma^2_y = p(1-p)$. Für die Modellierung der Wahrscheinlichkeit p in Abhängigkeit von x-Variablen benutzt man die **logistische Verteilungsfunktion** p(x). Wegen p(1−p) = exp(	$$p(x) = \frac{\exp(b_0 + b_1 x_1 + \ldots + b_p x_p)}{1 + \exp(b_0 + b_1 x_1 + \ldots + b_p x_p)}$$

$b_0+b_1x_1+...+b_px_p$) ist g(x) = log(p/1−p)) = $b_0+b_1x_1+...+b_px_p$.	$g(x)=\log\left(\frac{p(x)}{1-p(x)}\right)=b_0+b_1x_1+\ldots$

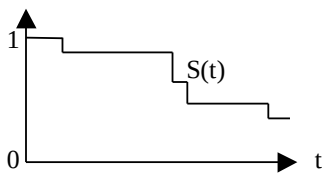
Die Einflussgrößen können metrisch, kategorial ranggeordnet oder binär sein. Man berechnet die Wahrscheinlichkeit p(x), dass y den Wert 1 annimmt. Wird von einem noch nicht klassifiziertem Objekt die Wahrscheinlichkeit p(x) auf Grund der geschätzten Parameter b_0, b_1, ... der Lernstichprobe und seiner eigenen x-Werte geschätzt, dann muss man irgendwo (meist bei p=0.5) die **Klassengrenze** ziehen. Diese sollte so gesetzt werden, dass der Klassifikationsfehler ein Minimum wird. Die Schätzung der b_0, b_1, ... erfolgt iterativ durch rechentechnisch aufwendige Maximierung der **Maximum-Likelihood-Funktion**. Das Verfahren liefert auch die Fehler s_{bj} der Koeffizienten. Mit dem **Wald-Test** (nach Abraham Wald) $W=b_j / s_{bj}$, wobei W approximativ als normal verteilt angenommen wird, prüft man die Signifikanz der Koeffizienten, und damit den Einfluss der x-Merkmale. Der **Likelihood-ratio-Test** ist ein globaler Test, mit dem man unterschiedliche Modellansätze vergleichen kann.

Binäre Einflussgrößen (x_j=0 bzw. x_j=1) führen auf das **Odds-Ratio** OR mit log(OR)=g(1)−g(0)=b_j. OR=e^{bj} ist die Wahrscheinlichkeit, die z.B. x_j =1 zum Krankheitsrisiko beiträgt.	$OR=\frac{p(1)}{1-p(1)}\Big/\frac{p(0)}{1-p(0)}$
Bei kontinuierlichen Einflussgrößen x_j gibt man die Erhöhung der Risikowahrscheinlichkeit an, falls sich x_j um 1 erhöht. Alle anderen Patientenwerte x_1,..., x_{j-1} , x_{j+1},..., x_p bleiben unverändert.	$\frac{p(x_1,\ldots,x_j+1,\ldots,x_p)}{p(x_1,\ldots,x_j,\ldots,x_p)}$

15. Survivalanalyse

Ziel ist die Modellierung der **Überlebenszeit** nach einer Operation oder einer Therapie in Abhängigkeit verschiedener Einflussfaktoren, wie Alter, Geschlecht usw. Während der Studie können Patienten an der untersuchten Krankheit sterben (+) oder aus der Studie ausscheiden (o). Gründe für das Ausscheiden sind u.a. das Ende der Studie, Umzug, Unfalltod usw. Man spricht von **zensierten Daten** (abgehackten Daten).	

Die **Überlebensfunktion** ist S(t)=1−F(t), wobei F(t) die Summenkurve der Überlebenszeiten t ist. S(t) gibt den Prozentsatz an Patienten an, die nach der Zeit t überlebt haben. Die **Kaplan-Meier-Schätzung** für S(t) ist das Produkt der Überlebenswahrscheinlichkeiten p_t von einem Tag t zum Tag t+1.	$S(t)=\prod_t p_t$

Die Überlebenswahrscheinlichkeiten p_t werden so bestimmt (n_0=Startzahl an Patienten): Niemand stirbt/wird zensiert: $p_t = (n_t - 0) / n_t$ und $n_{t+1} = n_t$ Jemand stirbt: $p_t = (n_t - 1) / n_t$ und $n_{t+1} = n_t - 1$ Jemand wird zensiert: $p_t = (n_t - 0) / n_t$ und $n_{t+1} = n_t - 1$	

Der **Median der Überlebenszeit** (50% Überlebende-Zeitpunkt) ist der Zeitpunkt mit S(t)=0.5. Ein approximatives **Vertrauensintervall** der "wahren Überlebensfunktion" ist

$$\hat{S}(t) \pm u_{\alpha,zweis.} \hat{S}(t) \sqrt{(1-\hat{S}(t))/n_t}$$

mit dem zweiseitigen Sicherheitspunkt der Normalverteilung, z.B. u_α=1.96 bei α=0.05.

Der **Vergleich zweier Überlebensfunktionen** $S_1(t)$ und $S_2(t)$ erfolgt z.B. mit dem Log-rank Trend Test. Man bildet man für jeden Ausfallzeitpunkt i=1,2,...,I, egal ob Gruppe 1 oder 2 oder beide betroffen sind, die 2x2-Tafel $a=d_{1i}$, $b=d_{2i}$, $c=n_{1i}$, $d=n_{2i}$, wobei die d_{ji} die Zahl der Ausfälle zu diesem Zeitpunkt in der Gruppe j=1,2 sind (meist 0 in einer der beiden Gruppen), die n_{ji} die Patientenzahlen in den Gruppen vor dem Ausfall. Aus jeder Tafel berechnet man den Erwartungswert E(a)=(a+b)(a+c)/(a+b+c+d) und E(b)=(a+b)(b+d)/(a+b+c+d) und setzt die Werte in die Summen O_1= Σa, O_2= Σb, E_1= ΣE(a), E_2= ΣE(b) ein. Mit $T=(O_1+O_2)-(E_1+E_2)$ und $V_T=(E_1+E_2)^2/(E_1+E_2)= E_1+E_2$ berechnet man die TREND-Test-Statistik $\chi^2_{Trend} = T^2/V_T$, die unter H_o χ^2-verteilt ist mit FG=1. Man vergleicht einseitig mit dem Sicherheitspunkt χ^2_α. Wenn $\chi^2_{Trend} \geq \chi^2_\alpha$, dann wird H_o abgelehnt.

Die **Hazardfunktion** (hazard rate, force of mortality, Ausfallrate (in der Qualitätssicherung), failure rate) h(t) ist die Wahrscheinlichkeit am Tag t zu sterben / als Gerät auszufallen. Die Funktion ist schwierig zu schätzen (Programmpakete BMDP, SPSS, SAS, SPSSX, Stat-View haben Programme für die Survivalanalyse). Die Definition der Hazardfunktion ist h(t) = f(t)/S(t), wobei f(t) die Dichtefunktion der Überlebenszeiten ist. Typische Hazardverläufe zeigen die folgenden 4 Graphiken:

negative aging, z.B. nach einer Operation	no aging (intakter Geräteservice)	positive aging mit steigendem Alter	bathtube curve vom Säugling zum Greis
			

Beim **Modell von Cox** handelt es sich wieder um ein typisches Regressionsmodell, bei dem die Überlebenszeit t, aber auch Einflussgrößen x_1, x_2, ..., x_p eingehen. Das Model heiß "proportional hazards model" und lautet:

$$h(t)= h_0(t)\cdot\exp(b_1 x_1 + b_2 x_2 + ...+ b_p x_p).$$

$h_0(t)$ heißt Baseline Hazard und ist eine nichtparametrische, unbekannte Funktion, die unabhängig von der Regressionsgleichung $b_1 x_1 + b_2 x_2 + ...+ b_p x_p$ berechnet wird. Jeder Term der Form $b_j x_j$ wirkt über die Exponentialfunktion wie ein Vorfaktor für $h_0(t)$.

16. Literatur

allgemeine Biostatistik	Spezialliteratur
Sachs., Lothar. (1997). *Angewandte Statistik: Anwendung statistischer Methoden*. 8th ed., Springer Verlag (Haupt- und Kochbuch des deutschen Naturforschers)	Lautsch, Erwin & Stefan von Weber (1995). *Methoden und Anwebdungen der Konfigurationsfrequenzanalyse (KFA)*, Beltz Psychologische Verlags Union
Gasser, Theo & Burkhardt Seifert (1999). *Biostatistik*, 3. Auflage, Universität Zürich, Abteilung Biostatistik (zu empfehlen, gegen Schutzgebühr bei den Autoren anfordern)	Ahrens, Heinz. & Jürgen Läuter: "Mehrdimensionale Varianzanalyse", Akademieverlag Berlin 1981.
Weber, Erna (1967). *Grundriss der biologi-*	Mucha, Hans-Jo.: Clusteranalyse mit Mikro-

schen Statistik, Gustav Fischer Verlag Jena	computern, Akademieverlag Berlin, 1992
Altmann, D.G. (1991): Practical Statistics for medical research, Chapman and Hall	Josef Puhani, Statistik, Bayrische Verlagsanstalt, Bamberg 1995 (mehr für Betriebswirte)

17. Praktikumsanleitung mit Excel

Zur Vorlesung ist kein Praktikum vorgesehen. Aber für die Berechnungen zur **Heimarbeit** oder wenn sich eine Hörerin oder ein Hörer selbst Praxis im Umgang mit Excel aneignen möchte, dann folgt hier eine kleine Anleitung.

Einige EXCEL-Funktionen

EXCEL hat gute Hilfe-Möglichkeiten. Diese Tabelle kann nur eine Anregung sein.

Funktion und Parameter	Aufrufbeispiel	Was liefert sie?
geomittel(xwerte)	=geomittel(a1:a5);	geometrisches Mittel
häufigkeit(x;klassengrenzen);	=häufigkeit(a2:a35;b7:b8)	Klassenhäufigkeiten
norminv(p;mittelwert;sigma)	=norminv(b5;c1;d1)	Quantil Xp Normalverteilung
normvert(x;mittel;sigma;typ)	=normvert(a8:a12;b1;c1;1)	$\Phi(u)$ mit u=(x-mittel)/sigma
rgp(y;x;konst;zusatzstatistik)	=rgp(a2:a7;b2:d7;1;0)	(multiple) lineare Regression
stabw(xwerte)	=stabw(c1:k1)	Das σ_{n-1} aller Werte
tvert(t;df;s)	=tvert(d8;b9;2)	Irrtumswahrscheinlichkeit α zum Freiheitsgrad, zweiseitig
trend(y;x;x*;k)	=trend(a2:a7;b2:b7;b8:b12;1)	Werte der Ausgleichsgeraden (k=1 mit konstantem Glied)
ttest(g1;g2;s;typ)	=ttest(a2:a9;b2:b14;2;2)	Mittelwertvergleich zweier normal verteilter Populationen
*	=(a1:a5)*(b1:b5)	paarweise Multiplikation
potenz(x;y)	=potenz(((a1:a5)-a6);2)	(Ai – A6)^2 für i=1,...,5

Verschiedene Mittelwerte und Standardabweichung: Starten Sie Excel. Tippen Sie in Zelle A1 irgend eine Spaltenbezeichnung, z.B. „Daten“, darunter 7 Zahlen, die für Sie Sinn machen (Z.B. 7 Umsätze oder 7 Temperaturen oder 7 Zinssätze oder 7 Gewichte).
In Zelle A10 tippen Sie =Mittelwert(**A2:A8**) und geben dann ENTER.
A2:A8 sagt, dass Sie mit der Maus (linke Taste gedrückt) über Ihre 7 Zahlen fahren („Ihre Daten selektieren“) oder aber den Feldbezug A2:A8 selbst eintippen.
Schreiben Sie daneben in Zelle B10 als Erklärung das Wort „Mittelwert“
Auf A11 die Standardabweichung σ_{n-1} mit =Stabw(A2:A8) , in B11 das Wort „Sigma“
Auf A12 den Median mit =Median(A2:A8), in B12 das Wort „Median“
Auf A13 das Geometrische Mittel, =Geomittel(A2:A8) , in B13 das Wort „Geomittel“
Auf A14 das Gestutzte Mittel mit =Gestutztmittel(A2:A8 ; 0,3). In Wort B14 das Wort „Gestutztmittel“. Erforschen Sie mit der Hilfeoption, was das Gestutztmittel genau macht.

Ausreißerkontrolle mit der 3-Sigma-Regel:
Tippen Sie in C1 das Wort „u-Werte“ ein.

Tippen Sie in C2 die Formel =(A2-A10)/A11 ein. Enter. Gehen Sie in die rechte untere Ecke der Zelle und ziehen Sie das Feld bis C8 unten. Die mathematische Formel, die dahinter steckt, ist $u_i = (P_i - \text{Mittelwert}) / \text{Stabw.}$ Die Dollarzeichen verhindern, dass beim Ziehen der Bezug A10 bzw. A11 verändert wird (feste Bezüge).
Tippen Sie in D1 „Betrag u“ ein.
Tippen Sie in D2 die Formel =abs(D2) ein und ziehen das Feld bis D8.
Auf Zelle D10 das maximale u mit Funktion =max(…). Tippen Sie in E10 „Max u“ ein.
Entscheiden Sie, ob ein Ausreißer unter den Daten ist (Max u>3 ?).

Graphische Ausreißerkontrolle: Stellen Sie Ihre Datenspalte graphisch dar und suchen Sie visuell nach Ausreißern:
A1 bis A8 selektieren → Diagrammassistent → Punkte (x,y) → Nur Punkte → Fertigstellen

Quartile berechnen: Berechnen Sie auf A15 mit =Quartile(A2:A8 ; 1) das 1. Quartil (Grenze der unteren 25%) Ihrer Daten, dann auf A16 das 2. Quartil (Grenze der unteren 50%) usw. bis zum 3. Quartil. Vergleichen Sie die Quartile mit dem Median. Was fällt Ihnen auf? Schreiben Sie in die Zellen

Momente der Datenverteilung: Berechnen Sie aus Ihren Daten die ersten 4 Momente.
Auf Zelle A20 das Mittel =Mittelwert(A2:A8) Tippen Sie in B20 „Mittelwert“ ein.
Auf Zelle A21 die Varianz =Varianz(A2:A8) Tippen Sie in B21 „Varianz“ ein.
Auf Zelle A22 die Schiefe =Schiefe(A2:A8) Tippen Sie in B22 „Schiefe“ ein.
Auf Zelle A23 die Kurtosis =Kurt(A2:A8) Tippen Sie in B23 „Kurtosis“ ein.
(Die Kurtosis oder der Excess ist eine Randverdickung gegenüber der Gausskurve.)

Datenverteilung durch ein Histogramm graphisch darstellen: Legen Sie in Mappe 2 (Tabelle 2) auf Excel-Spalte A eine neue Spalte *Daten* an. Tippen Sie auf A1 das Wort „Daten“. Kopieren Sie Ihre Daten aus Tabelle 1 unter das Wort *Daten* und verlängern Sie die Zahlenkolonne mit ausgedachten Zahlen bis A26 (insgesamt 25 Zahlenwerte).

Tippen Sie in Zelle B1 das Wort „Klassengrenzen“. Geben Sie darunter 5 aufsteigend sortierte Zahlen ein als Klassengrenzen für zu bildende Klassen. Die erste Klassengrenze sollte größer sein als Ihr kleinster Datenwert, die 5. Klassengrenze kleiner als Ihr größter Datenwert.
Tippen Sie in C1 das Wort „Häufigkeiten“.

Selektieren Sie das Feld C2 bis C7 mit der Maus. Tippen Sie in die weiß gebliebene Zelle C2 die Formel =Häufigkeit(A2:A26 ; B2:B6) und geben Sie die 3-fach-Taste STRG-UMSCH-ENTER. Die 6 Zellen füllen sich mit den ausgezählten Häufigkeiten. Die erste Häufigkeit ist die Anzahl Ihrer Datenwerte in Klasse 1 (Kleinster Wert bis einschließlich 1. Klassengrenze). Der letzte Häufigkeitswert ist für die Klasse jenseits und einschließlich der 5. Klassengrenze.

Tippen Sie in D1 das Wort „Klasse“. Waren Ihre Klassengrenzen z.B. 10, 20, 30, 40, 50, dann schreiben Sie in D2 den folgenden Text „*bis einschl. 10* “, in D3 „*von 11 bis einschl. 20* “, usw. und in D7 „*ab einschl. 50* “.

C1 bis C7 selektieren → Diagrammassistent → Säule → weiter → Reihe → ein Klick in das Feld rechts von „Beschriftung der Rubrikenachse (x)“ und mit der Maus D2 bis D7 selektieren → Fertigstellen.

Indexierung auf gemeinsamen Startwert 100%: Spielen Sie das Beispiel Indexierung aus der Vorlesung mit eigenen Daten durch. Machen Sie eine Liniengraphik der beiden Datenreihen vor und nach der Indexierung.

Einfache lineare Regression mit Teststatistiken in Excel

Die einfach lineare Regression setzt man z.B. bei folgenden Aufgaben ein:

- Man möchte eine Ausgleichsgerade durch Datenpunkte ziehen
- Man möchte den Anpassungsfehler (die Reststreuung) wissen
- Man möchte testen, ob der Anstieg signifikant ist
- Man möchte testen, ob die Konstante signifikant von Null verschieden ist, oder ob nicht eine Gerade durch den Ursprung die bessere Wahl wäre
- Man möchte die Gerade für eine Prognose verlängern und wissen, wie genau sind die prognostizierten Werte.

Die Funktion **=trend(y-Werte ; x-Werte)** berechnet die Erwartungswerte $\hat{y}_i$ der Ausgleichsgeraden, die durch die y- und x-Werte definiert ist. Die Funktion **=rgp(y-Werte ; x-Werte ; wahr ; wahr)** berechnet die Regressionskoeffizienten, deren Standardabweichungen, die Reststreuung, die Bestimmtheit r^2, deren Freiheitsgrad usw. einer einfachen oder multiplen Regression. Das erste *wahr* steht für ein "Modell **mit Regressionskonstante**", das zweite *wahr* für "außer den Koeffizienten weitere statistische Kennzahlen ausgeben", wie oben genannt. Die Abkürzung **SSE** steht im Schema unten für die 3-fach-Tastenbelegung Strg-Shift-Enter (bzw. Strg-Umsch-Enter). Drücken Sie erst die beiden linken Tasten Strg und ⇑, dann zusätzlich ENTER. Zuerst tippen Sie die Spaltenbezeichnungen x, y, y-Dach als Text ein, dann die x-Zahlenwerte in die Felder A2 bis A7, dann die y-Zahlenwerte in B2 bis B7, dann laut Schema:

	S1=A	S2=B	S3=C
Z1	x	y	y-Dach
Z2	1,7	3,3	
Z3	2,3	4,1	
Z4	2,1	4,5	
Z5	2,4	4,7	
Z6	3,9	8,3	
Z7	1,6	3,3	
Z8			
Z9	2,20	-0,45	b1,bo
Z10	0,18	0,45	sb1,sbo
Z11	0,97	0,34	r2, sR
Z12	144	4	F, df
Z13	16,8	0,47	ssreg,ssres
Z14			
Z15	12	-1.0	t1, t2

Selektiere *C2:C7* und tippe ein:
=trend(*sel. B2:B7* **;** *sel. A2:A7* **)** **SSE**
(y-Werte x-Werte)

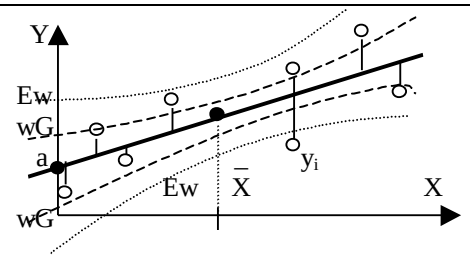
Auf *C2 bis C7* erscheinen die berechneten y-Dach-Werte. Jetzt wollen wir die Koeffizienten und Statistiken berechnen: Selektiere *A9:B13* und tippe ein:
=rgp(*sel.B2:B7***;***sel.A2:A7***;wahr;wahr)** **SSE**
Es erscheinen die Zahlen in Spalte A und B z.B. b1=Anstieg, bo=Regressionskonstante der Geraden y = bo + b1 x

Berechnung der t-Statistiken: Sel. *A15:B15*
= *sel.A9:B9* / *sel. A10:B10* **SSE**
die beiden Teststatistiken erscheinen

In den berechneten Statistiken bedeuten:

sb1, sb0	die Standardabweichungen (Fehler) der beiden Koeffizienten b_1 und bo
r^2	die multiple Bestimmtheit (bei einer einfach linearen Regression ist es das Quadrat des Korrelationskoeffizienten r)
sR	Reststreuung der Messpunkte um die Gerade (mittlere Abweichung)
F	Testgröße (F-Statistik) hier zur Hypothese Ho: $b_1=0$ mit den Freiheitsgraden df1=1 und df2=df. Bei einer einfachen Regression wie hier im Beispiel ist

	$F=t_1^2$, und t_1 die t-Statistik für b_1 mit df Freiheitsgraden.
t_1, t_2	sind die Teststatistiken zu den Koeffizienten b1 und b0. Man testet damit die Hypothesen Ho: b1=0 bzw. Ho: bo=0
ssreg	$=\sum(y_i-\bar{y})^2$, auch Summe der Abweichungsquadrate der y genannt (SAQyy)
ssresid	$=\sum(\hat{y}_i-y_i)^2=\sum e_i^2$, auch Summe der Abweichungsquadrate bzw. Fehlerquadratsumme genannt.

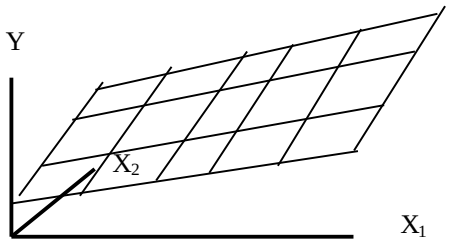
Die Abbildung rechts zeigt die Regressionsgerade im X-Y-Koordinatensystem. Sie geht durch den Punkt **a** auf der Y-Achse und durch den Punkt $(\bar{x}, \bar{y})$. Die Messwerte y_i sind durch kleine Kreise, die Residuen e_i durch Striche dargestellt. Das Konfidenzintervall der wahren Geraden (wG) ist gestrichelt, das der Einzelwerte (Ew) ist gepunktet dargestellt.	

Wir wollen eine Graphik der Regressionsgeraden mit den Messpunkten als gif-Datei oder jpg-Datei in ein WORD-Dokument einfügen, ohne dass WORD später auf EXCEL zugreifen will:

1. Mache die Graphik so groß wie möglich und kopiere sie in das Zeichenprogramm Paint
2. Mache eventuelle zusätzliche Beschriftungen oder andere Änderungen
3. Speichere sie unter Dateityp *.gif oder *.jpg ab

Die so erzeugte Bilddatei Trend.gif können Sie jetzt problemlos an beliebiger Position in ein Word-Dokument oder eine Powerpoint-Präsentation einfügen.

Multiple lineare Regression mit Excel

Die multiple Regression verknüpft p Einflussgrößen X_1, X_2,..., X_p mit einer Zielgröße Y. Das Modell kann mit oder ohne Regressionskonstante b_o sein: $Y = b_o + b_1 * X_1 + b_2 * X_2 + ... + b_p * X_p + e$ Die geometrische Interpretation ist eine Ebene	

über dem von X_1, X_2, ... aufgespannten Raum. Die Regressionskoeffizienten b_1, b_2,..., b_p (und b_0) werden nach der Methode der kleinsten Quadrate (Σe^2= Minimum) geschätzt. **e** ist der zufällige Fehler oder Residuum (Abweichung). Die multiple lineare Regression setzt man z.B. für folgende Aufgaben ein:

- Man möchte eine Ausgleichsebene durch Datenpunkte legen, d.h. den Einfluss mehrerer Einflussgrößen X_1, X_2, … auf eine Zielgröße Y durch eine lineare Formel darstellen. Mit dieser Formel kann man Werte vorhersagen (Prognose) oder zwischen Datenpunkten interpolieren.
- Man möchte wissen, ob die lineare Formel die Zielgröße genau genug wiedergibt. Man kann den Gesamteinfluss aller Einflussgrößen auf die Zielgröße global bewerten.

- Man möchte aus sehr vielen Einflussgrößen diejenigen heraussuchen, die einen signifikanten Einfluss auf die Zielgröße haben, d.h. man bewertet jede Einflussgröße einzeln.

Die multiple Regression schätzt aus p Einflussgrößen X_1, X_2, …, X_p die Werte einer Zielgröße Y. Das am meisten benutzte Regressionsmodell ist die Ebenengleichung

$$Y_i = b_0 + b_1 X_{i1} + b_2 X_{i2} + ... + b_p X_{ip} + e_i$$

Dabei ist Y_i ein beobachteter Wert der Zielgröße, X_{ij} ist der i-te Wert der j-ten Einflussgröße, b_0 ist die Regressionskonstante, b_1 , b_2 , ... , b_p sind Regressionskoeffizienten, e_i ist der Fehler im Datenpunkt i (oder Abweichung bzw. Residuum).

Beispiel Pflanzenwachstum: Der Ertrag in Abhängigkeit unterschiedlicher Parameter wird bestimmt. Die verfügbaren Daten sind in der folgenden Exceltabelle zu sehen.

Zeile	A	B	C	D	E	F
1	Bodenwert	Beregnung	Düngung	Temperatur	Bodendichte	Y = Ertrag
2	2	2	0,10	17	1320	1,1
3	2	3	0,15	19	1410	1,5
4	4	2	0,10	22	1190	1,8
5	3	4	0,20	20	1240	2,0
6	2	1	0	18	1240	0,80
7	1	3	0,10	18	1350	1,20
8	4	4	0	21	1270	1,95
9	2	3	0,20	15	1300	1,15

Wir markieren das Feld von A11:F15 und tippen eine Regressionsanweisung ein, die zuerst die Zielgrößenwerte Y nennt, dann die Einflussgrößenwerte X. Das erste „*wahr*“ legt ein Modell „*mit Konstante*“ fest, das zweite „*wahr*“ legt fest, dass wir außer den Koeffizienten weitere Werte berechnet haben möchten, z.B. die s_{bi}, R^2, s_R, usw. Es sind immer 5 Zeilen, die Sie markieren. Die Spaltenzahl richtet sich jedoch nach der Anzahl der Koeffizienten im Regressionsmodell (bo zählt mit, falls es berechnet werden soll).

=rgp(F2:F9; A2:E9; wahr; wahr) Strg-Shift-Enter

Zeile 11	b_5=0,000129	b_4=0,0995	b_3=1,379	b_2=0,185	b_1=0,137	b_0= -1,597
12	s_{b5}=0,000369	s_{b4}=0,0119	s_{b3}=0,253	s_{b2}=0,0214	s_{b1}=0,0322	s_{b0}=0,514
13	R^2=0,997	s_R =0,0431				
14	F= 147,44	FG= 2				
15	ssreg=1,37	ssresid=0,0037	ssreg	ssresid		
16	Dichte	Temperatur	Düngung	Beregnung	Bodenwert	b_0

Wie man sieht, kehrt Excel die Reihenfolge der Regressionskoeffizienten um (b_5, b_4, ..., b_0). In den berechneten Statistiken in den Zeilen 12 bis 15 bedeuten im Falle der multiplen Regression:

s_{b5},..., s_{b0} die geschätzten Standardfehler der Koeffizienten b_5 ,..., bo

R^2 die multiple Bestimmtheit (bei einer einfach linearen Regression ist es das Quadrat des Korrelationskoeffizienten r. R^2 =0 heißt, dass keinerlei linearer Zusammenhang zwischen der Gesamtheit aller Einflussgrößen mit der Zielgröße besteht. R^2 =1 heißt, dass die Einflussgrößen die gegebenen y-Werte absolut exakt reproduzieren ohne jede Abweichung.)

s_R Reststreuung der Messpunkte um die berechnete Ebene (mittlere Abweichung)

F Testgröße (F-Statistik nach Fisher) zur Bewertung der multiplen Bestimmtheit. Hypothese Ho: „keine Bestimmtheit, ein Wert von $R^2>0$ ist rein zufällig". Hypothese Ha: „Es besteht ein signifikanter Einfluss der Einflussgrößen auf die Zielvariable, ein Wert von $R^2>0$ ist nicht zufällig". Die Irrtumswahrscheinlichkeit p bei Ablehnung von Ho (bzw. Annahme von Ha) berechnet man mit der Funktion FVERT(F ; n – FG – 1 ; FG) wenn b_o mitberechnet wird (mit TRUE ausgewählt), und wird auch als p-Value zum F-Test bezeichnet. Falls b_o nicht berechnet wird (mit FALSE abgewählt), schreiben Sie FVERT(F ; n – FG ; FG).

Die Summen ssreg und ssresid wurden schon bei der einfachen Regression kurz beschrieben. Für die Bewertung der Wichtigkeit der einzelnen Einflussgrößen bzw. der Konstanten für das Regressionsmodell hat man zu jedem Koeffizienten das Hypothesenpaar Ho und Ha. Ho sagt: „Diese Einflussgröße hat keinen linearen Einfluss auf die Zielgröße. Ein Wert bj ≠ 0 eines Koeffizienten ist rein zufällig". Hypothese Ha sagt: „Diese Einflussgroße trägt signifikant zur Erklärung der Zielgröße bei." Praktisch berechnet man zu jedem Koeffizienten eine Teststatistik. Meistens wird die t-Statistik verwendet. Es gilt $t_i = |b_i / s_{bi}|$. Wir dividieren mit einer Excelanweisung gleich alle Koeffizienten und die Konstante durch ihren geschätzten Standardfehler und bilden den Absolutbetrag. Dazu markieren wir die Felder A18:F18 und tippen die nachfolgende Befehlszeile ein:

=ABS(A11:F11/A12:F12) Strg-Shift-Enter

Zeile 18	t_5=0,351	t_4=8,33	t_3=5,43	t_2=8,65	t_1=4,27	t_0=3,10

Die t-Verteilung hat eine ähnliche Gestalt wie die Normalverteilung (Glockenkurve). Die Funktion TVERT berechnet aus einem t-Wert, dem Freiheitsgrad FG von oben und der Zahl 2 die zweiseitige Irrtumswahrscheinlichkeit (p-Value) bei Ablehnung der Hypothese H_0 zum betreffenden Koeffizienten. Dieser p-Value (die Irrtumswahrscheinlichkeit) sollte möglichst klein sein, z.B. <0,05, denn dann bewerten wir die Einflussgröße als wesentlich (signifikant). Zur Berechnung der p-Values markieren wir die Zellen A20:F20 und tippen folgende Anweisung ein:

=TVERT(A18:F18; B14; 2) Strg-Shift-Enter

Zeile 20	p_5 = 0,75	p_4 = 0,014	p_3 = 0,032	p_2 = 0,013	p_1 = 0,0506	p_0 = 0,09

In der Forschung gibt man meist eine zulässige Irrtumswahrscheinlichkeit α = 5% (0,05) vor. D.h. mit 5% Wahrscheinlichkeit wollen wir uns bei der Bewertung einer Einflussgröße irren dürfen. Ist der berechnete p-Value größer α, dann entscheiden wir uns für Hypothese Ho (unwesentliche Einflussgröße). Ist $p \le \alpha$, dann entscheiden wir uns für Hypothese Ha (wesentliche Einflussgröße). Zur Darstellung der Hypothesenwahl markieren wir die Felder A22:F22 und tippen folgende Anweisung ein:

=wenn(A20:F20 > 0,05 ; „Ho" ; „ Ha") Strg-Shift-Enter

Zelle 22	Ho	Ha	Ha	Ha	Ho	Ho
23	Dichte	Temperatur	Düngung	Beregnung	Bodenwert	b_0

Den schlechtesten p-Value (höchste Irrtumswahrscheinlichkeit) hat Einflussgröße X_5=Dichte. Wenn wir unser Regressionsmodell von unwesentlichen Bestandteilen befreien wollen, sollten wir zuerst diese Einflussgröße entfernen (Schrittweiser Abbau). Entfernen Sie jedoch in

jedem Schritt immer nur einen Term, d.h. eine Einflussgröße oder die Konstante bo. Durch Korrelationen zwischen den Einflussgrößen ändern sich die p-Values oft dramatisch bei Wegnahme oder Hinzunahme einer einzelnen Einflussgröße. Die Regressionskonstante bo kann man entfernen, indem man statt des ersten „*wahr*“ in der rgp-Anweisung ein „*falsch*“ schreibt.

Den globalen Test auf einen signifikanten linearen Zusammenhang der Gesamtheit der Einflussgrößen auf die Zielgröße macht man mit dem F-Test (siehe oben bei der Erklärung des F). Das folgende Rechenschema liefert den p-Value und die Hypothese Ho bzw. Ha. Jede eingetippte Anweisung schließen Sie mit ENTER ab.

	A =ANZAHL(A2:A9)	B	C =FVERT(A14; A24-B14-1; B14)	D	E	F =wenn(….
Zeile 24	**8**		**0,0067**			Ha
Zeile 25	Anzahl		p-Value			Hypothese

Die Wenn-Anweisung lautet vollständig: =wenn(C24 > 0,05 ; „Ho“ ; „Ha“)

18. Alte Beispielklausuren

Beispielklausur 1

1. (10 P) **Bedingte Wahrscheinlichkeit, Multiplikation von Wahrscheinlichkeiten**:
   a) Ein Patient ist von leptosomen Körperbau, den 27 % aller Patienten aufweisen. Er gehört zu den 7% Patienten, die leptosom sind und an Skoliose leiden. Berechnen Sie die Wahrscheinlichkeit P_1, mit der ein leptosomer Patient Skoliose hat.
   b) Berechnen Sie Wahrscheinlichkeit P_2 , mit der ein Patient weiblich **und** leptosom ist. Frauen stellen 57% der Patienten.

   Lösung: a) 25,9% b) 15,4%

2. (15 P) **Suchen Sie Ausreißer** mit der 3–σ–Regel in den folgenden n=17 LDL-Cholesterinwerten::

   145 132 178 138 127 152 157 147 163 204 144 153 166 158 149 128 151

   Lösung: $\bar{x}$=152,47 $\sigma_{n-1}=18{,}87$ u_{max}=2,73 u_{min}=-1,34 keine Ausreißer

3. (15 P) **Statistische Maßzahlen**: Berechnen Sie aus den LDL-Cholesterinwerten der Aufgabe 2 $\bar{x}, \sigma_{n-1}, \sigma_{\bar{x}}$ sowie den Median und das 95%-Konfidenzintervall des wahren Mittels. Einen eventuellen Ausreißer lassen Sie in den Daten drin.

   Lösung: $\bar{x}$=152,47 $\sigma_{n-1}=18{,}87$ $\sigma_{\bar{x}}$ = =4,58 Median=151 FG=16
   $t_\alpha = 2{,}12$ Konfidenzinte4rvall 152,47 ± 9,71

4. (20 P) Machen Sie den **χ^2-Homogenitätstest** zu folgender Kontingenztafel, die die

Häufigkeit von Morbus Scheuermann in Abhängigkeit von der Haarfarbe und vom Geschlecht untersucht:

Haarfarbe	blond	braun	schwarz	
weiblich	$n_{11} = 27$	$n_{12} = 43$	$n_{13} = 30$	
männlich	$n_{21} = 13$	$n_{22} = 61$	$n_{23} = 76$	

Beantworten Sie die Frage nach der Unabhängigkeit der Merkmale Haarfarbe und Geschlecht

Lösung: e_{11}=16 e_{12}=41,6 …. χ^2_{11}=7,56 χ^2_{12}=o,o5 … χ^2_{ges}=18,73

FG=2 χ^2_{alfa}=5,99 H_A Haarfarbe und Geschlecht sind Keine unabhängigen Merkmale bezüglich der Häufigkeit von Morbus Scheuermann

5. (20 P) Testen Sie Sie mit dem **Mann-Whitney Test**, ob ein signifikanter Unterschied zwischen den Alpha2-Globulin-Werten der Elektrophorese von Gruppe1-Patienten (rein vegetarische Ernährung) und Gruppe2-Patienten (gemischte Ernährung) vorliegt. Die Messwerte werden als nicht normalverteilt eingestuft. Die Werte sind bereits innerhalb der Gruppen sortiert.

Gruppe 1:	5.3	5.5	5.5	6.1	6.7	6.8	6.8	7.3	7.5	7.9	7.9	8.4
Gruppe 2:	6.1	6.4	6.9	8.3	8.4	8.5	8.9	8.9	9.8	12.3		

Lösung: Rangsumme Gr. 1 ist 101 m=12 ux=97
Rangsumme Gr. 2 ist 152 n=10 uy=22 U=23
u=-2,44 p=0,015 (<0,05) und damit akzeptieren wir H_A
Es besteht ein signifikanter Unterschied beim Alpha2-Globulin

Beispielklausur 2

1. (17 P) Versuchsplanung: Es soll gegenüber den Zulassungsbehörden nachgewiesen werden, dass *Hepuramol* den Blutdruck von Hochdruckpatienten in Abhängigkeit vom Maß des Überdrucks senkt. Aus früheren Untersuchungen weiß man, dass die Reststreuung s_R bei Blutdruckdaten **s_R =5 ,7 mmHg** beträgt. Die Varianz der Blutdruckmesswerte bei Hochdruckpatienten ist $\sigma_x^2 = \left[\sum_{i=1}^{n}(x_i - \bar{x})^2\right] / (n-1)$ und hat hier den Wert **σ_x^2 = 193,6** [mm^{2}Hg]. Gesucht ist die **Mindestzahl n** an Probanden, um eine 10%-ige Senkung des Blutüberdrucks (Regressionskoeffizient **b_1 = – 0,1**) auf dem 5%-Niveau der Irrtumswahrscheinlichkeit nachweisen zu können.

a) Geben Sie die Formel an, wie man aus **σ_x^2** die Summe **SAQxx** der Abweichungsquadrate erhält (Suchen Sie im Skript unter "Korrelationskoeffizient" die Formel für SAQxx).

Lösung: SAQxx=193,6 (n-1) (mit n=70 gemäß Aufgabenteil b wird) $SAQ\ xx = 193{,}6 \cdot 69$

b) Setzen Sie die aus der Aufgabenstellung bekannten Werte in den t-Test für den Regressions-koeffizienten b_1 ein und probieren Sie mit verschiedenen Probandenzahlen n die Signifi-

kanz, bis Sie das kleinste n gefunden haben, das noch auf ein signifikantes t führt. Arbeiten Sie sich von n = 70 rückwärts Richtung n = 60. Notieren Sie die Zwischenergebnisse n, t, t_α .

Lösung: n=70 $t = \frac{0,1}{5,7}\sqrt{193,6 \cdot 69} = 2,03$ t_α =1,99 signifikant

n=68 t=2,00 t_α =1,99 signifikant

n=67 t=1,98 t_α =1,99 nicht signifikant

Man benötigt 68 Probanden

2. (21 P) Test auf Normalverteilung mit dem χ^2-Anpassungstest: Gegeben sind die 25 HDL-Werte

174	157	133	107	148	144	140	154	137	150
116	147	136	156	130	182	167	128	163	162
121	161	202	159	191					

a) Berechnen Sie $\bar{x}$ *und* σ_{n-1} Lösung: $\overline{X}$ =150,6 $\sigma_{n-1} = 22,74$

b) Machen Sie den χ^2-Anpassungstest, d.h., berechnen Sie die Histogrammgrenzen, die Klassenhäufigkeiten, Erwartungswerte, χ^2_i-Werte, χ^2_{gesamt}, Freiheitsgrad, Sicherheitspunkt χ^2_α. Wählen Sie die Hypothese aus und geben Sie in einem Satz eine fachliche Umsetzung.

Lösung:

Klassengrenzen	-∞	131,5	144,9	156,3	169,7	∞	
Häufigkeiten	5	5	5	6	4		
Erwartungswerte	5	5	5	5	5		
χ_i^2	0	0	0	0,2	0,2	χ^2_{ges}=0,4	

FG=2 χ_α^2 =5,99 Ho Normalverteilung der Daten akzeptiert

3. (20 P) Mann-Whitney-Test: Gegeben sind zwei Gruppen von Blutgerinnungszeiten, von denen die Datenverteilung unbekannt ist. Die Gruppe-B-Patienten wurden mit *Fibrilanol* behandelt. Machen Sie den Mann-Whitney-Test auf Mittelwertunterschied:

a)

A:	4,3	4,9	3,7	2,3	1,7	4,3	5,1	2,4	4,4	6,3
B:	0,9	2,2	1,8	2,7	1,8	1,8	1,4	2,3	2,5	3,7

Legen Sie die gemeinsame Rangfolge fest und die Rangzahl jedes einzelnen Messwertes

Lös.: 1 2 3 5 5 5 7 8.5 8.5 10 11 12 13.5 13.5 15.5 15.5 17 18 19 20

b) Geben Sie das Hypothesenpaar und die Rangsummen der beiden Gruppen an

Lös: $Ho: \mu_1 = \mu_2$ $H_A: \mu_1 \neq \mu_2$ Rx=140 Ry=70

c) Testen Sie mit der u-Formel für große n, m unter Benutzung der Φ(u)-Tafel, wählen Sie Ihre Hypothese und setzen Sie das Ergebnis in einem Satz fachlich um.

Lös.: m=10 n=10 ux=15 uy=85 U=15 u=-2,65 u_α =1,96 $p_{zws.}$=2 Φ(u) ~ 0,01

Wegen p<0,05 akzeptieren wir H_A. Die Blutgerinnungszeiten nach Fibrilanolgabe sind signifikant kürzer.

4. (22 P) Mittelwertvergleiche zweier normalverteilter Grundgesamtheiten: Gegeben sind die systolischen Blutdruckwerte von Laborratten vor und kurz nach einer Beschallung mit lauter Musik:

Vor: 114 117 116 121 119 122 118 - - - 126 123
Nach: 122 119 115 124 - - - 123 121 121 129 - - -

(2 Ratten starben an Herzversagen, bei einer rutschte die Messkanüle bei der ersten Messung ab)

a) Machen Sie den t-Test für ungepaarte Werte nach Beispiel 19, d.h., Hypothesenpaar, Mittelwerte, gemitteltes Sigma, t-Test, Ihre Testentscheidung und eine fachliche Umsetzung

Lös.: $\bar{x}_1 = 119{,}55$ $\sigma_{n1}{}^2 = 12{,}69$ $n_1=9$ $\bar{x}_2 = 121{,}75$ $\sigma_{n2}{}^2 = 14{,}1875$ $n_2=8$
$\bar{\sigma} = 3{,}89$ t=-1,162 FG=15 $t_\alpha = 2{,}13$ Wir akzeptieren Ho
Es gibt keinen signifikanten Unterschied der Blutdruckwerte

b) Machen Sie den gepaarten t-Test (Beispiel 21) mit den 7 vollständigen Paaren, d.h. Hypothesenpaar, Differenzen bilden, mittlere Differenz, Sigma der Differenzen, t-Test, Testentscheidung, fachliche Umsetzung

Lös.: n=7 Differenzen: 8 2 -1 3 1 3 3 $\bar{d} = 2{,}714$ $\sigma_{n-1} = 2{,}752$ t=2,609
FG=6 $t_\alpha = 2{,}45$ Wir akzeptieren H_A
Es gibt einen signifikanten Unterschied der Blutdruckwerte

c) Welcher der beiden Tests bringt hier das bessere Ergebnis? Lösung: Der gepaarte t-Test

Beispielklausur 3

1.) (10 P) Aus einer Fragebogenaktion zum Trinkverhalten von Patienten ergab sich u. a. die Frage: Gibt es Unterschiede zwischen Männern und Frauen bezüglich der Wichtigkeit von Alkohol, Säften, Heißgetränken? (Trinktyp). Die Kontingenztafel lautet:

		Trinktyp		
		Alkohol	Säfte	Heißgetränke
Geschlecht	m	84	23	42
	w	27	82	54

Testen Sie auf einen signifikanten Zusammenhang zwischen den Merkmalen Geschlecht und Trinktyp (Hypothesen, e_{ij}, χ^2_{ij}, χ^2_{Gesamt}, Hypothese wählen, Antwortsatz)

Lösung: $e_{11}=53{,}0$ $e_{12}=50{,}1$ …. $\chi^2{}_{11}=18{,}13$ $\chi^2{}_{12}=14{,}66$ … $\chi^2{}_{ges}=63{,}3$
FG=2 $\chi^2{}_{alfa}=5{,}99$ H_A Es besteht ein signifikanter Zusammenhang zwischen den Merkmalen Geschlecht und Trinktyp

2.) (8 P) Machen Sie den Vergleiche von relativen Häufigkeitszahlen für das Zahlenpaar aus der Tabelle von Aufgabe 1, Spalte 1 (Alkohol). Hier ist n_1 die Zeilensumme 1, n_2 ist die Zeilensumme 2 der Tabelle (Hypothesen, p, q, t, Hypothese wählen, Antwortsatz).

Lös.: Ho:p1=p2 H_A:p1≠p2 h1=84 h2=27 n1=149 n2=163
$\hat{p}_1 = 84/149 = 0{,}564$ $\hat{p}_2 = 27/163 = 0{,}166$ $\bar{p} = 111/312 = 0{,}356$ q=0,644
FG=310 t=7,33 $t_\alpha = 1{,}96$ Wir akzeptieren H_A
Der Alkoholkonsum ist bei den Männern signifikant größer als bei den Frauen.

3. (15 P) Gegeben Sind die n=28 β-Globulinwerte von 10 Patientinnen und 18 Patienten

137	162	182	279	191	187	244	143	169	172	336	233	155	175	191
174	183	88	151	306	102	161	206	274	167	173	183	241		

Berechnen Sie aus den gesamten Daten (alle 28 Werte) Mittelwert $\bar{x}$, Standardabweichung σ_{n-1}, den Fehler des Mittelwerts $\sigma_{\bar{x}}$, das 95%-Konfidenzintervall für das wahre Mittel, den Median und die Spannweite (Max−Min). (Beispiele 11, 7, 8)

Lös.: n=28 $\bar{X}$ =191,6 $\sigma_{n-1} = 56{,}8$ $\sigma_{\bar{x}} = 10{,}73$ Median=178,5 FG=27
$t_\alpha = 2{,}06$ Konfidenzinte4rvall 191,6 ± 22,1 Spannweite =336-88 =248

4. (15 P) Machen Sie unter Benutzung von Mittelwert $\bar{x}$, Standardabweichung σ_{n-1} und der Daten aus Aufgabe 3 den Test auf Normalverteilung (Beispiel 17). (Hypothesen, Klassengrenzen, Häufigkeiten, Erwartungswerte, χ_i^2, χ_{Ges}^2, Hypothese wählen, Antwortsatz.)

Lösung: Klassengrenzen -∞ 143,9 177,4 205,8 239,3 ∞
Häufigkeiten 4 10 6 2 6
Erwartungswerte 5,6 5,6 5,6 5,6 5,6
χ_i^2 0,457 3,457 0,029 2,314 0,029 χ^2_{ges}=6,286
FG=2 χ_α^2=5,99 Ha keine Normalverteilung der Daten

5. (10 P) Wieviel von 5.000 Patienten werden schätzungsweise einen β-Globulinwert von x>250 aufweisen, wenn man Mittelwert $\bar{x}$, Standardabweichung σ_{n-1} aus Aufgabe 3 zugrunde legt? Bei welchem β-Globulinwert enden die 25% der „wenig belasteten" Patienten (Quantil X_{25}) (Beispiel 11)

Lös.: a) u=1,028 $\Phi(-u) = 0{,}1587$ E= N p =5000·0,1587= 793 Patienten
b) p=0,25 u=-0,6 x=157,5

6. (15 P) Machen Sie mit den Daten von Aufgabe 3 den Test auf unterschiedliche β-Globulinbelastung bei Frauen und Männern (Beispiel 19) mit dem ungepaarten t-Test für normalverteilte Grundgesamtheiten.

Lös.: $Ho: \mu_1 = \mu_2$ $H_A: \mu_1 \neq \mu_2$
$\bar{x}_1 = 185{,}3$ $\sum x^2$ =380405 n_1=10 SAQ1=37044,1
$\bar{x}_2 = 195{,}1$ $\sum x^2$ =734640 n_1=18 SAQ2=49409,6
$\bar{\sigma} = 57{,}66$ t=-0,43 FG=26 $t_\alpha = 2{,}06$ Wir akzeptieren Ho
Es gibt keinen signifikanten Unterschied bei den β-Globulinwerten

Beispielklausur 4

Gegeben sind die Gewichte von 2 Patientengruppen. Gruppe 1 ist normalgewichtig, die Patienten der Gruppe 2 sind für ihre Größe zu schwer.

Gr. 1	67	72	56	77	71	87	74	94	83	
Gr. 2	93	123	109	98	133	107	103	94	97	109

1.) (10 P) (Beispiel 7 und 8) Berechnen Sie nur zur Gruppe 1 die Anzahl n, Mittelwert $\bar{x}$, Standardabweichung σ_{n-1}, Fehler des Mittelwerts $\sigma_{\bar{x}}$, das 95%-Konfidenzintervall des wahren Mittels, den Median, das Quantil Q_{25} und Wahrscheinlichkeit P, mit der Gewichte x>90 erwartet werden.

Lös.: $n_1 = 9$ $\bar{x}_1 = 75{,}67$ $\sigma_{n-1} = 11{,}29$ $\sigma_{\bar{x}} = 3{,}763$ FG=8 $t_\alpha = 2{,}31$ Median = 74
95%-Intervall $75{,}67 \pm 8{,}7$ $u_{25} = -0{,}7$ $Q_{25} = 75{,}67 - 0{,}7 \cdot 11{,}29 = 67{,}77$
U= (90 – 75,67) / 11,29 = 1,27 P = Φ(-1,27) = 0,1

2.) (6 P) (Beispiel 19) Machen Sie den F-Test auf Varianzhomogenität zwischen Gruppe 1 und Gruppe 2 (Hypothesen, F-Wert, Hypothese wählen, Antwortsatz).

Lös.: Ho: Varianzhomogenität Ha: Varianzinhomogenität
$\bar{x}_2 = 106{,}6$ $\sigma_{n-1} = 12{,}91$ $n_2 = 10$ $F = 12{,}91^2 / 11{,}29^2 = 1{,}31$
FG1 = 9 FG2 = 8 $F_{\alpha=5\%,\ FG1=9,\ FG2=8} = 3{,}35$ (Sicherheitspunkt) Ho: „Homogenität“

3.) (8 P) (Beispiel 19a) Machen Sie den Mittelwertvergleich zweier normalverteilter Grundgesamtheiten (t-Test), um den Gewichtsunterschied zwischen Gruppe 1 und Gruppe zwei auf Signifikanz zu testen (Hypothesen, t-Wert, Hypothese wählen, Antwortsatz).

Lös.: $Ho: \mu_1 = \mu_2 \quad H_A: \mu_1 \neq \mu_2$
$SAQ_1 = (67^2 + 72^2 + \ldots + 83^2) - 9 \cdot 75{,}667^2 = 52549 - 51529 = 1020$
$SAQ_2 = (93^2 + 123^2 + \ldots + 109^2) - 10 \cdot 126{,}6^2 = 115136 - 113636 = 1500$
$\bar{\sigma} = \sqrt{(1020+1500)/17} = 12{,}17$ FG=17 $t_\alpha = 2{,}11$
$t = ((75{,}67 - 106{,}6)/12{,}17) \cdot \sqrt{(9 \cdot 10)/(9+10)} = -5{,}53$ Ha: „Sign. Gewichtsunterschied“

4.) (10 P) Runden Sie die beiden Gruppenmittel jeweils zur nächstgelegenen ganzen Zahl (als Beispiele 86,4 →86 oder 86,5 → 87), bilden Sie die Differenzen $d_{i1} = x_{i1} - \bar{x}_1$ in Gruppe 1 und $d_{i2} = x_{i2} - \bar{x}_2$ in Gruppe 2, fassen Sie alle 19 Differenzen in einer Liste zusammen und machen Sie den Chiquadrat-Anpassungstest auf Normalverteilung für die 19 Differenzen (Hypothesen, Mittelwert $\bar{d}$, Standardabweichung σ_{n-1} der d_i, dann die 5 Klassengrenzenpaare berechnen, das Histogramm auszählen, die χ_i^2, χ^2_{Gesamt}, Hypothese wählen, Antwortsatz.)
(Beispiel 17)

Lös.: $\bar{x}_1^* = 76$ $d_i = (-9 \quad -4 \quad -20 \quad 1 \quad -5 \quad 11 \quad -2 \quad 18 \quad 7)$
$\bar{x}_2^* = 107$ $d_i = (-14 \quad 16 \quad 2 \quad -9 \quad 26 \quad 0 \quad -4 \quad -13 \quad -10 \quad 2)$
Ho: „Normalverteilung“ Ha: „keine Normalverteilung“

Klassengrenzen	-∞	-10,4	-3,33	2,59	9,69	∞
Häufigkeiten		3	6	5	1	4
Erwartungswerte		3,8	3,8	3,8	3,8	3,8
χ_i^2		0,17	1,27	0,38	2,06	0,01

χ^2_{ges}= 3,89

FG=2 χ^2_α=5,99 Ho Normalverteilung der Daten akzeptiert

5. (10 P) Mann-Whitney-Test (Beispiel 20): Gegeben sind zwei Gruppen von Blutgerinnungszeiten. Die Gruppe-B-Patienten wurden mit *Targophagin* behandelt. Machen Sie den Mann-Whitney-Test auf Mittelwertunterschied: Zuerst in jeder Gruppe sortieren.

a) A: 3,3 3,9 3,7 2,3 2,7 4,3 4,1 2,4 4,4 5,3 2,6
B: 0,9 2,2 1,7 2,2 1,8 1,8 1,4 1,3 2,1 2,7
Legen Sie die gemeinsame Rangfolge und die Rangzahlen der einzelnen Werte fest.

b) Geben Sie das Hypothesenpaar und die Rangsummen der beiden Gruppen an

c) Testen Sie mit der u-Formel für große n, m unter Benutzung der □(u)-Tafel, wählen Sie Ihre Hypothese und setzen Sie das Ergebnis in einem Antwortsatz fachlich um.

Lös.: $Ho: \mu_1 = \mu_2 \quad H_A: \mu_1 \neq \mu_2$

Gr 1 sortiert = (2,3 2,4 2,6 2,7 3,3 3,7 3,9 4,1 4,3 4,4 5,3) n = 11

Gr 2 sortiert = (0,9 1,3 1,4 1,7 1,8 1,8 2,1 2,2 2,2 2,7) m = 10

Werte:	0,9	1,3	1,4	1,7	1,8	1,8	2,1	2,2	2,2	2,3	2,4	2,6	2,7	2,7	3,3	3,7	3,9	4,1	4,3	4,4	5,3
Gruppe:	B	B	B	B	B	B	B	B	B	A	A	A	A	B	A	A	A	A	A	A	A
Rang.	1	2	3	4	5,5	5,5	7	8,5	8,5	10	11	12	13,5	13,5	15	16	17	18	19	20	21

Rx= 10+11+12+13.5+15+16+…+21 =172,5 Ux=11·10+11·12/2 – 172,5 = 3,5

Ry= 1+2+3+4+5,5+5,5+7+8,5+8,5+13,5 =58,5 Uy=11·10+10·11/2 – 58,5 = 106,5

U = 3,5

$u = (3{,}5 - 10 \cdot 11/2)/\sqrt{11 \cdot 10 \cdot (11+10+1)/12}$ = -3,63 P = Φ(-3,63) = 0,00016

Wegen P<0,05 akzeptieren wir Ha. „Signifikanter Unterschied der Gerinnungszeiten“

6. (8 P) Versuchsplanung (Beispiel 1): Auf der Grundlage des t-Tests für 2 relative Häufigkeiten **(Beispiel 13)** soll die Anzahl **n** an Frauen und Männern bestimmt werden, die für eine Signifikanzprüfung der Frage „Wer greift bei Kopfschmerzen schneller zur Tablette?“ erforderlich sind. Jede Gruppe hat den gleichen Umfang **n**. Die durchschnittliche Wahrscheinlichkeit, dass eine Frau oder ein Mann zur Tablette greift, sei **p=0,32**. Ein Unterschied von $\hat{p}_1 - \hat{p}_2 = 0{,}032$ soll mit Irrtumswahrscheinlichkeit α=5% gerade noch einen signifikanten t-Wert liefern. Starten Sie Ihre Suche nach dem richtigen **n** bei n=1500 und variieren Sie Ihr **n** dann in 50-er Schritten nach oben bzw. nach unten.

Lös.: $t = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{pq}} \sqrt{\frac{n_1 n_2}{n_1 + n_2}} \qquad t = \frac{0{,}032}{\sqrt{0{,}32 \cdot 0{,}68}} \sqrt{\frac{n^2}{2n}} = \frac{0{,}032}{0{,}466 \cdot \sqrt{2}} \sqrt{n} = 0{,}0486 \cdot \sqrt{n}$

n=1500 FG→∞ t_α = 1,96 $t = 0{,}0486 \cdot \sqrt{1500}$ = 1,88 nicht signif.

n=1550 ……..

……..

n=1650 FG→∞ t_α = 1,96 $t = 0{,}0486 \cdot \sqrt{1650}$ = 1,97 signifikant

n=1650 Frauen und n=1650 Männer werden für die Prüfung benötigt.

19. Liste der Beispiele, die in der Vorlesung gerechnet werden

1. Versuchsplanung 2. Bedingte Wahrscheinlichkeiten 3. Histogramme 4. Momente 5. Poissonverteilung 6. Binomialverteilung 7. Normalverteilung von Daten 8. Konfidenzintervalle 9. 3-Sigma-Regel 10. Indexierung 11. Statistische Maßzahlen 12. Test Häufigkeit gegen Konstante 13. Test zweier relativer Häufigkeiten 14. Kontingenz- oder Homogenitätstest	15. χ^2-Zerlegung nach Lancaster 16. Kontingenzmaße 17. χ^2-Anpassungstest für Verteilung 18. Einstichproben t-Test 19. F-Test auf Varianzhomogenität 19a) t-Test (gleiche Varianzen) 19b) Welch-Test (ungleiche Varianzen) 20. Mann-Whitney-Test 21. Gepaarter t-Test 22. Wilcoxon-Test 23. Produkt-Momenten Korrelation r 24. Lineare Regression (1 und 2) 25. VA Kreuzklassifikation

20. Liste der Aufgaben für die Heimarbeit

Jede(r) Student(in) muss für den 3. ECTS-Punkt bis **zum Ende der Vorlesungszeit** eine **Heimarbeit** aus **einer Aufgabe bestehend** anfertigen (**nur 1-er oder 2-er-Gruppen**). Zur Auswahl steht eine Reihe von Aufgaben, die Sie in den nachfolgenden zwei Tabellen finden. Es finden drei Praktikumstermine statt, an denen ich Hilfestellung gebe. Sie können die fertige Heimarbeit

- an einem der drei Praktikumstermine vorzeigen
- als ungezippte PDF an ***webers@hs-furtwangen.de*** senden
- als Schwarz-weiß-Kopie in mein Postfach legen

Liste der **A-Aufgaben**: Die Zuordnung erfolgt über den **Anfangsbuchstaben des Rufnamens**

Aufgabe	**Anfang Rufname**	**Thema**
A1	A, B, C	Ca. 50% aller Hecken sind Liguster. Die Blätter sind schmal, lanzettförmig und ca. 5-8 cm lang. Sammeln Sie 50 Blätter ein. Messen Sie die Länge (Excel-Tabelle). Berechnen Sie Mittelwert, Standardabweichung, Fehler des Mittelwerts, Median mit **Excel** und machen Sie den Test auf Normalverteilung der Längen laut Skript mit dem **Taschenrechner oder Excel**
A2	D, E, F	Ilex oder Stechpalme wächst in Vorgärten, Parks oder Friedhöfen. Die Blätter sind winterhart, dunkelgrün und haben Stacheln am Rand. Sammeln Sie 25 Blätter der Ilex (Stechpalme) und messen Sie die Längen (Spalte A). Dann bei 25 Blättern der Ilex eines anderen Standorts die Längen messen (Spalte B). Machen Sie mit **Excel** den t-Test auf einen möglichen signifikanten Längenunterschied der beiden Populationen

A3	G, H, I	Ilex oder Stechpalme wächst in Vorgärten, Parks oder Friedhöfen. Die Blätter sind winterhart, dunkelgrün und haben Stacheln am Rand. Sammeln Sie 25 Blätter und zählen Sie die Stacheln eines jeden Blattes, machen Sie eine Häufigkeitstabelle geordnet nach aufsteigender Stachelzahl, und erzeugen Sie mit **EXCEL** ein Säulendiagramm und ein Tortendiagramm zur Verteilung der Stachelanzahl
A4	J, K, L, M, N, O	Ca. 50% aller Hecken sind Liguster. Die Blätter sind schmal, lanzettförmig und ca. 5-8 cm lang. Behauptung: Das "Deutsche Ligusterblatt" ist 66 mm lang. Messen Sie 30 zufällig gewählte Blätter und berechnen Sie mit **Excel** Mittelwert, Standardabweichung und machen Sie den Einstichproben-t-test gegen die Konstante 66
A5	P, Q, R	Alkohol ist Zielwasser. Messen Sie die Abweichungen von jeweils 20 Dartpfeilen vom Zentrum vor und nach einem Bier (Spalte A und B). Machen Sie mit **Excel** den t-Test auf einen möglichen signifikanten mittleren Abstandsunterschied der beiden Messwertpopulationen
A6	S, T	Machen Sie auf ein Stück weißen Papiers einen dicken Zielpunkt. Legen Sie 20 mal die rechte Hand mit dem Kugelschreiber ans Ohr, schließen Sie die Augen und markieren Sie mit geschlossenen Augen einen Punkt in der Nähe des Zielpunktes. Messen Sie den Abstand vom Ziel in mm. Wiederholen Sie das Spiel auf einem neuen Blatt 20 mal für die linke Hand. (Spalte A und B). Machen Sie mit **Excel** den t-Test auf einen möglichen signifikanten Abstandsunterschied der beiden Messwertpopulationen
A7	U, V, W, X, Y, Z	Kaffee ist der Feind des Schützen. Messen Sie die Abweichungen von jeweils 20 Dartpfeilen vom Zentrum vor und nach einem Kaffee bzw. Tee (Spalte A und B). Machen Sie mit **Excel** den t-Test auf einen möglichen signifikanten Abstandsunterschied der beiden Messwertpopulationen

Liste der **B-Aufgaben**: Die Zuordnung über den **Anfangsbuchstaben des Familiennamens**

Aufgabe	**Anfang Familienname**	**Thema**
B1	A, B, C, D, E, F	Wiegen Sie im Labor 20 Mandeln (LIDL, ca. 2,00 € /Tüte) einzeln aus, messen von jeder Mandel Länge L, Breite B und Dicke D und machen Sie mit **EXCEL** die zwei multiplen Regressionen (1) Gewicht= a+ b*L*B (Spalte1=G, Spalte2=L*B) (2) Gewicht= a+ b*L*B*D (Spalte1=G, Spalte2=L*B*D) Die t-Werte ebenfalls mit **Excel** berechnen laut Anleitung im Skript. Welche Regression gibt gemäß Reststreuung das Gewicht besser wieder?
B2	G, H, I,	Wiegen Sie im Labor 20 Haselnüsse (LIDL, ca. 2 € /Tüte) einzeln aus, messen von jeder Nuss max. und min. Durchmesser Dx und

	J, K, L	Dm und machen Sie mit **Excel** die multiple Regression Gewicht= a+b*Dx + c*Dm + d*Dx*Dm (Spalte1=G, Spalte 2=DX, Spalte3=Dm, Spalte4=Dx*Dm) Die t-Werte ebenfalls mit EXCEL berechnen.
B3	M, N, O	Nehmen Sie 1 Seite deutschen und 1 Seite spanischen Text. Zählen Sie alle Buchstaben auf der Seite und die Häufigkeit des Buchstabens "e". Machen Sie laut Skript per **Taschenrechner oder Excel** den t-Test für die zwei relative Häufigkeiten des Buchstabens „e“. Gibt es einen signifikanten Unterschied?
B4	P, Q, R	Nehmen Sie 1 Seite deutschen und 1 Seite englischen Text. Zählen Sie alle Buchstaben auf der Seite und die Häufigkeit des Buchstabens "e". Machen Sie laut Skript per **Taschenrechner oder Excel** den t-Test für die zwei relative Häufigkeiten des Buchstabens „e“. Gibt es einen signifikanten Unterschied?
B5	S, T	Ca. 50% aller Hecken sind Liguster. Die Blätter sind schmal, lanzettförmig und ca. 5-8 cm lang. Wiegen Sie im Labor 10 Ligusterblätter einzeln aus und messen sie von jedem Blatt Länge L und Breite B, und machen Sie mit **Excel** die multiple Regression *Gewicht= a+b*Länge + c*Breite + d*Länge*Breite* (Spalte1=G, Spalte 2=L, Spalte3=B, Spalte4=L*B) Die t-Werte ebenfalls mit EXCEL berechnen (siehe Skript).
B6	U, V, W, X, Y, Z	Erfragen Sie von 10 Kommilitonen Schuhgröße S und Größe G. Dasselbe für 10 Kommilitoninnen. Berechnen Sie mit **Excel** getrennt für die Männer und die Frauen die zwei Korrelationsanalysen *=korrel(S ; G)* und testen Sie den Korrelationskoeffizienten *r* auf Signifikanz. Berechnen Sie bitte auch die t-Werte und prüfen Sie die beiden Korrelationskoeffizienten *r* auf Signifikanz.