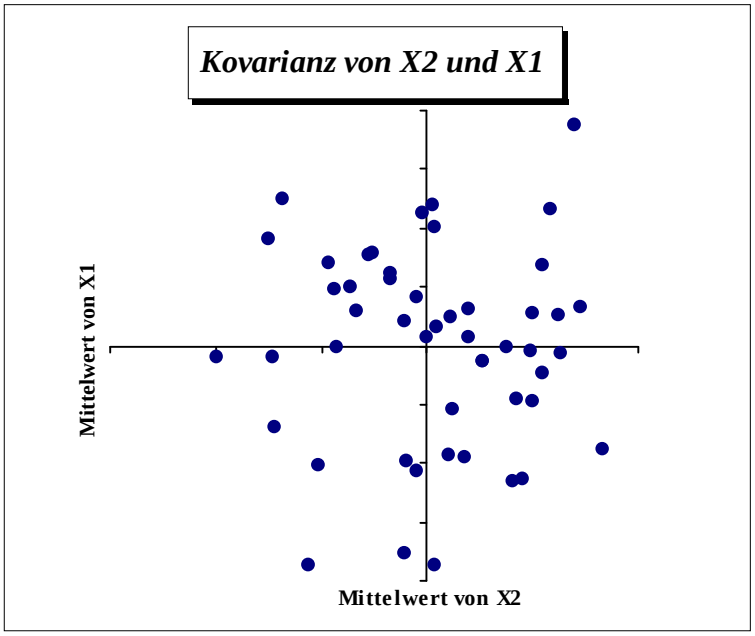
DIE KOVARIANZ

Bei der Berechnung der Varianz hat man das durchschnittliche Abweichungsquadrat der Werte einer Variablen von ihrem Mittelwert berechnet. Die Kovarianz ist nun ein statistischer Kennwert für das Ausmaß an gemeinsamer Variation (Kovariation) *zweier Variablen* — sprich, den linearen „je-desto"–Zusammenhang zweier Variablen. Sie berechnet sich als das gemittelte Produkt der Abweichungen der X_1 und X_2-Werte von ihrem jeweiligen Mittelwert $\overline{X}_1$ bzw. M_1 und $\overline{X}_2$ bzw. M_2:

$$S_{1.2} = \frac{1}{N}\sum_{i=1}^{N}\left(X_1 - M_1\right)\left(X_2 - M_2\right)$$

Wie bei der Varianz erhält man auch hier durch leichte Umformung eine Rohwerteformel:

$$\begin{aligned} S_{1.2} &= \frac{1}{N}\sum_{i=1}^{N}\left(X_1X_2 - X_1M_2 - X_2M_1 + M_1M_2\right) = \\ &= \frac{1}{N}\sum_i X_1X_2 - M_2\overbrace{\frac{1}{N}\sum_i X_1}^{M_1} - M_1\overbrace{\frac{1}{N}\sum_i X_2}^{M_2} + \frac{1}{N}N\cdot M_1M_2 = \\ &= \frac{1}{N}\sum_i X_1X_2 - 2M_1M_2 + M_1M_2 = \\ &= \frac{1}{N}\sum_i X_1X_2 - M_1M_2 \text{ (Rohwerteformel)} \end{aligned}$$



Wie man sieht, bekommt die Kovarianz genau dann einen positiven Wert, wenn die Mehrzahl der X_1 und X_2-Werte gleichzeitig entweder über- oder unterdurchschnittlich ist; in der Abbildung bedeutet das viele Werte im rechten oberen (I.) und linken unteren (III.) Quadranten. Man spricht dann von einem *direkten Zusammenhang*.

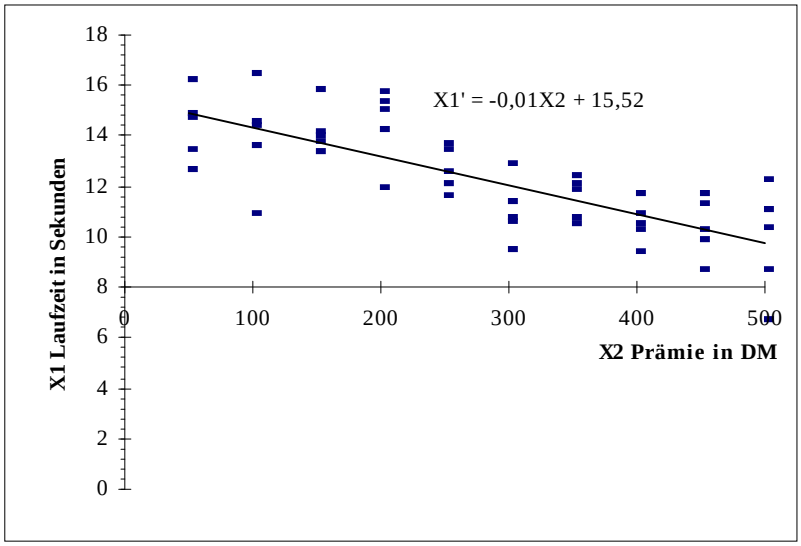
Ein *indirekter oder inverser Zusammenhang* wird durch entsprechend viele Werte im linken oberen (II.) und rechten unteren (IV.) Quadranten ersichtlich. Hier hat immer eine Abweichung ein negatives Vorzeichen, so daß das Produkt ebenfalls negativ wird.

Die Kovarianz ist jedoch kein sehr anschauliches bzw. vergleichbares Zusammenhangsmaß, da die Streuungen der beiden Variablen X_1 und X_2 in ihre Berechnung einfließen. Deshalb gibt es nocht andere Maße, die eben diese Vergleichbarkeit besser ermöglichen, so z.B. der *Produkt-Moment-Korrelationskoeffizient*, oder kurz, die *Korrelation* (Wertebereich [-1;1]).

$$r_{1.2} = \frac{S_{1.2}}{S_1 \cdot S_2}$$

REGRESSION UND KORRELATION

Will man die Werte einer abhängigen Variablen, eines sog. *Kriteriums* auf die Werte einer unabhängigen Variablen, des sog. *Prädiktors* zurückführen, kann man eine Regression machen. Als Beispiel wäre denkbar, daß man am Ende einer 100m Laufbahn in mehreren Durchgängen verschiedene Geldbeträge deponiert (unabh. Var.) und dann mißt, wie schnell ein Sprinter läuft, um diese Strecke zu überwinden (abh. Var.). Die Ergebnisse könnten sich in etwa folgendermaßen darstellen:



Man kann also die Kriteriumswerte (Laufzeiten, X_1) als eine lineare Funktion der Prädiktorwerte (Prämien, X_2), eine Gerade, annähern:

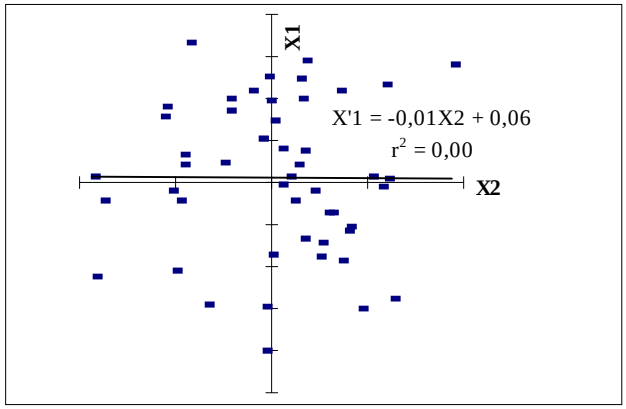
$$X'_1 = b_{1.2} X_2 + a_{1.2}$$
(Regressionsgerade)

Mit Hilfe dieser *Regressionsgeraden* sind nun wiederum Laufzeiten für bestimmte Prämienbeträge zu *schätzen* (X'_1). Allerdings ist der Zusammenhang — wie man in der Graphik schon sehen kann — natürlich kein perfekter. Die tatsächlichen Werte von X_1 weichen von der Schätzung aus X_2 ab; diese Abweichungen nennt man Fehler (X_{1-2}); sie stellen also den aus X_2 nicht schätzbaren bzw. von X_2 unabhängigen Teil in X_1 dar.

Doch wie müssen die Parameter der Geradengleichung (Steigung $b_{1.2}$, Achsenabschnitt $a_{1.2}$) beschaffen sein, sodaß die Schätzgerade wirklich die *beste* Schätzung des Kriteriums darstellt? (siehe Extrablatt mit der Herleitung)

$$b_{1.2} = \frac{S_{1.2}}{S_2^2} \text{ und}$$
$$a_{1.2} = M_1 - b_{1.2} M_2$$

Diese Regressionskoeffizienten bestimmen die beste Schätzgerade. Es kann aber trotzdem sein, daß gar kein bzw. nur ein sehr kleiner (oder nichtlinearer) Zusammenhang zwischen den Variablen besteht. In so einem Fall sind trotz „guter Schätzung“ die Kriteriumswerte relativ weit von der Regressionsgeraden entfernt:



In diesem (simulierten) Fall sind die beiden Variablen voneinander unabhängig, was sich in der wenig elliptischen und mehr kugelförmigen Punktewolke darstellt.

Werden Zwei Variablen gleichzeitig *erhoben* und nicht die Werte einer unabhängigen Variable *a priori zugeteilt*, spricht man nicht mehr von einem *regressionsanalytischen*, sondern von einem *korrelationsanalytischen* Modell, bei dem es gleichgültig ist, welche Variable auf welche regrediert wird.

In beiden Fällen wird meist neben den Regressionskoeffizienten der Korrelationskoeffizient angegeben, der bei optimalem inversen Zusammenhang -1, bei optimalem direkten 1 und bei keinem Zusammenhang zwischen den Variablen den Wert 0 hat. Er hat außerdem den bereits angesprochenen Vorteil der Unabhängigkeit von der spezifischen Streuung der beiden Variablen; er ist — wie man zeigen kann — veränderungsresistent gegenüber linearen Transformationen (bleibt konstant).

Natürlich sind $b_{1.2}$ und $r_{1.2}$ (und $S_{1.2}$) ineinander umzurechnen:

$$r_{1.2} = \frac{S_{1.2}}{S_1 \cdot S_2} = \frac{S_{1.2}}{S_2 \cdot S_2} \cdot \frac{S_2}{S_1} = \frac{S_{1.2}}{S_2^2} \cdot \frac{S_2}{S_1} = b_{1.2} \cdot \frac{S_2}{S_1}$$

Wegen diesem Zusammenhang sind $r_{1.2}$, $b_{1.2}$ und $S_{1.2}$ immer gleichzeitig 0, wenn einer der Kennwerte den Wert 0 hat.

Auch das Quadrat des Korrelationskoeffizienten, das *Bestimmtheitsmaß B*, ist als Anteil gemeinsamer (wechselseitig erklärbarer) Varianz der beiden Variablen ein gebräuchlicher Kennwert:

$$\begin{aligned} S_{1'}^2 &= r_{1.2}^2 \cdot S_1^2 \\ r_{1.2}^2 &= \frac{S_{1'}^2}{S_1^2} \\ &= B \end{aligned}$$

Die Varianz des Fehlers ergibt sich wegen

$$S_1^2 = S_{1'}^2 + S_{1-2}^2$$

als Kriteriumsvarianz (S_1^2) mal Anteil nicht vorhersagbarer Varianz (1 – B):

$$\begin{aligned} S_1^2 &= S_{1'}^2 + S_{1-2}^2 \\ S_{1-2}^2 &= S_1^2 - S_{1'}^2 = S_1^2 - \left(S_1^2 \cdot r_{1.2}^2\right) \\ S_{1-2}^2 &= S_1^2 \cdot \left(1 - r_{1.2}^2\right) \end{aligned}$$

Bei standardisierten Variablen, die eine Varianz und Standardabweichung von 1 haben, ist die Korrelation $r_{1.2}$, die Kovarianz $S_{1.2}$ und der Steigungskoeffizient $b_{1.2}$ der Regressionsgeraden gleich:

$$S_{1.2} = \frac{S_{1.2}}{1 \cdot 1} = \frac{S_{1.2}}{1^2}$$

Der Achsenabschnitt $a_{1.2}$ ist dann genau 0, weil:

$$\begin{aligned} &a_{1.2} = M_1 - b_{1.2} M_2 \wedge M_1 = M_2 = 0 \text{ (stand. Var.)} \\ &a_{1.2} = 0 \end{aligned}$$

MATHEMATISCHE HERLEITUNG DER REGRESSIONSKOEFFIZIENTEN

Man fordert, daß die quadrierten Abweichungen minimal sein sollen (Minimum-Quadrat-Prinzip):

$$\sum_i \left(X_{1i} - X'_{1i}\right)^2 = \min.$$

$$\sum_i \left[X_{1i} - \left(bX_{2i} + a\right)\right]^2 = \min.$$

Diese von b und a abhängige Funktion kann mun minimiert werden, indem man jeweils die Nullstellen der partiellen Ableitungen nach a und b sucht:

$$\begin{aligned} f(a,b) &= \sum_i \left[X_1 - \left(bX_2 + a\right)\right]^2 \\ &= \sum_i \left(X_1^2 - 2aX_1 - 2bX_2X_1 + b^2X_2^2 + 2abX_2 + a^2\right) \\ &= \sum_i X_1^2 - 2a\sum_i X_1 - 2b\sum_i X_2X_1 + b^2\sum_i X_2^2 + 2ab\sum_i X_2 + Na^2 \end{aligned}$$

Ableitung nach a:

$$\frac{df(a,b)}{da} = -2\sum_i X_1 + 2b\sum_i X_2 + 2Na = 0$$

$$a = \frac{1}{N}\sum_i X_1 - b\cdot\frac{1}{N}\sum_i X_2 = M_1 - bM_2$$

Ableitung nach b:

$$\frac{df(a,b)}{db} = -2\sum_i X_2X_1 + 2b\sum_i X_2^2 + 2a\sum_i X_2 = 0$$

$$= -2\sum_i X_2X_1 + 2b\sum_i X_2^2 + 2\cdot\left(\frac{1}{N}\sum_i X_1 - b\cdot\frac{1}{N}\sum_i X_2\right)\cdot\sum_i X_2 = 0$$

$$2b\sum_i X_2^2 - 2b\sum_i X_2\cdot\frac{1}{N}\sum_i X_2 = 2\sum_i X_2X_1 - 2\sum_i X_2\frac{1}{N}\sum_i X_1 \Bigg| \cdot\frac{1}{2}$$

$$b = \frac{\sum_i X_2X_1 - \sum_i X_2\frac{1}{N}\sum_i X_1}{\sum_i X_2^2 - \frac{1}{N}\left(\sum_i X_2\right)^2} \Bigg| \cdot\frac{\frac{1}{N}}{\frac{1}{N}}$$

$$= \frac{\frac{1}{N}\sum_i X_2X_1 - \overbrace{\frac{1}{N}\sum_i X_2\frac{1}{N}\sum_i X_1}^{M_2\cdot M_1}}{\frac{1}{N}\sum_i X_2^2 - \underbrace{\frac{1}{N^2}\left(\sum_i X_2\right)^2}_{M_2^2}} =$$

$$= \frac{S_{1.2}}{S_2^2}$$

$b_{1.2}$ ist also $\frac{S_{1.2}}{S_2^2}$ und $a_{1.2}$ ist $M_1 - b_{1.2}M_2$

POPULATIONSMODELLE FÜR REGRESSIONS- UND KORRELATIONSANALYSE

Das Regressionsanalytische Modell

Untersuchungstyp: Experiment

Vom Untersucher festgelegte Werte des Prädiktors X_2 werden den Probanden zugeteilt. Z.B.: Reaktionszeit nach 1, 2 oder 3 Liter Bier

Zwischen dem einer Person i zugeteilten Wert X_{2i} des Prädiktors und dem im Kriterium X_{1i} beobachteten Wert besteht die Beziehung:

$$X_{1i} = \beta_{1.2} X_{2i} + \alpha + X_{ei}$$

und gilt nur für die apriori festgelegten Werte von X_2.

Das Korrelationsanalytische Modell

Untersuchungstyp: Erhebung

An den Probanden werden die beiden Merkmale X_1 und X_2 simultan erhoben.

Z.B.: Feststellen den Alkoholspiegels und Messung der Reaktionszeit bei Oktoberfestbesuchern

Zwischen den beiden erhobenen Variablen X_1 und X_2 gilt für jede beliebige Person i die Beziehung

$$X_{1i} = \beta_{1.2} X_{2i} + \alpha + X_{ei}$$

welche für jeden Wert von X_2 gelten muß.

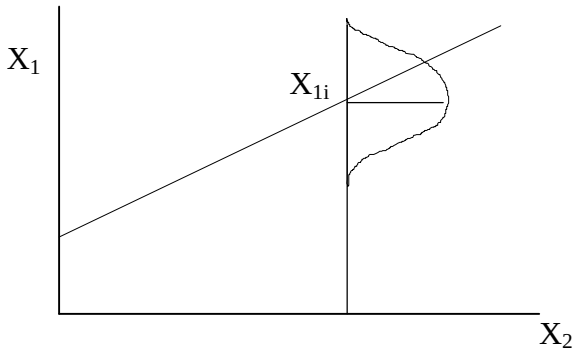
Der Kriteriumswert X_{1i} zerfällt in zwei Komponenten: 1. die Schätzung $X_{1i} = \beta_{1.2} X_{2i} + \alpha$, 2. in die Fehlervariable X_{ei} (e = error)

Für jedes zugewiesene X_2 ist die Schätzung gleich und wird deshalb als der vorhersagbare Teil der Kriteriumvariablen bezeichnet.

Bei zugeteiltem Prädiktor X_2 muß gelten:

Für Personen, die den gleichen Wert in der Variablen X_2 haben, ist die Schätzung gleich.

Bei zufällig gleichen Werten muß gelten:



Die Fehlervariable X_e ist normalverteilt mit dem Mittelwert $\mu_e = 0$ und konstanter Varianz δ_e^2.

Die Stichprobe von Personen, an denen das Experiment vorgenommen wird bzw. an denen die Untersuchung vorgenommen wird, muß eine *Zufallsstichprobe* aus der Gesamtpopulation sein. Alle Elemente müssen also die gleiche Chance haben, in die Stichprobe aufgenommen zu werden.

Die Fehlerwerte müssen voneinander unabhängig sein (es dürfen also keine systematischen Fehler auftreten).

$$\text{Cov}(X_{ei}, X_{eh}) = 0,\ i \neq h;\ i, h = 1, 2, ..., N.\ (\text{Cov ist die Kovarianz})$$

Treten systematische Fehler auf, handelt es sich wahrscheinlich um keine Zufallsstichprobe.

DER SIGNIFIKANZTEST ZUR ÜBERPRÜFUNG EINES LINEAREN ZUSAMMENHANGS

1. Man legt apriori (vor der Erhebung/ dem Experiment) ein Signifikanzniveau fest. Standardmäßig ist das 5%, 1% oder 0,1%.
2. Man errechnet den F-Wert nach der Formel: $F = \frac{(N-2)\cdot r_{1.2}^2}{1-r_{1.2}^2}$.
3. Man bestimmt die Freiheitsgrade: Zählerfreiheitsgrade (FG_Z) = Anzahl der Variablen minus 1; Nennerfreiheitsgrade (FG_N) = Anzahl der Personen minus Anzahl der Variablen.
4. Man schaut in einer F-Werte-Tabelle nach und sucht sich den sog. *kritischen F-Wert* für die ermittelten Zähler- und Nennerfreiheitsgrade und das gewählte Signifikanzniveau.
5. Man vergleicht den errechneten F-Wert mit dem kritischen F-Wert (für welchen mit allen noch größeren F-Werten unter H_o: $\beta = 0$ die Wahrscheinlichkeit $\leq \alpha\%$ ist) aus der Tabelle, und entscheidet über die Nullhypothese H_0: Ist $F_{berechnet} < F_{kritisch}$ ⇨ H_0 beibehalten, $F_{berechnet} \geq F_{kritisch}$ ⇨ H_0 verwerfen und H_1 annehmen.

Rechenbeispiel:

Man *erhebt* bei einer Zufallsstichprobe aus der Population der Psycho-Studenten die Anzahl der zu Freizeitzwecken benützten Zeit pro Tag und die Note in der Schubö-Statistik-Klausur. Die Nullhypothese lautet, daß es da keinen Zusammenhang gibt (H_0: $\beta = 0$, $\rho = 0$); man will jetzt aber testen, ob sich die Note (X_1) nicht doch teilweise auf die Freizeitkapazität (X_2) zurückführen läßt (die Voraussetzungen für das Populationsmodell seien erfüllt!).

	Freizeitstunden/ Tag		Klausurnote		
	X_{2i}	X_{2i}^2	X_{1i}	X_{1i}^2	$X_{1i}X_{2i}$
1	2		1		
2	8		3		
3	7		3		
4	10		5		
5	4		2		
6	4		4		
$\sum$ =					

M =

S_2 = , S_2^2 = , S_1 = , S_1^2 = $S_{1.2}$ =

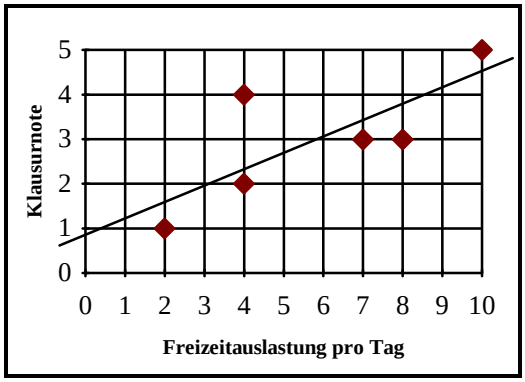
$b_{1.2}$ = , $a_{1.2}$ = Regressionsgleichung: X_1' = ... X_2 + ...

$r_{1.2}$ = , $r_{1.2}^2$ =

$FG_{Zähler}$ = , FG_{Nenner} =

$F_{errechnet}$ = , $F_{krit,\ 5\%}$ = 7.71

☐ H_0 beibehalten *oder* ☐ H_0 verwerfen ?



FORMELN, PARAMETER UND DEREN BEZIEHUNGEN

Mittelwert: $M_1 = \frac{1}{N}\sum_i X_{1i}$ mit N = Anzahl der Beobachtungen

Varianz: $S_1^2 =$

Kovarianz: $S_{1.2} = \frac{1}{N}\sum_i (X_{1i} - M_1)(X_{2i} - M_2) = \frac{1}{N}\sum_i X_{1i}X_{2i} - M_1M_2$

Regressionskoeffizient: $b_{1.2} = \frac{S_{12}}{S_2^2}$; $b_{1.2} = \frac{S_1}{S_2} \cdot r_{1.2}$

Achsenabschnitt: $a = -b_{1.2}\, M_2 + M_1$

Korrelationskoeffizient: $r_{1.2} = \frac{S_{12}}{S_1 \cdot S_2}$; $r_{1.2} = \frac{S_2}{S_1} \cdot b_{1.2}$

Bestimmtheitsmaß: $B = r_{1.2}^2$

Anteil gemeinsamer (= wechselseitig erklärbarer) Varianz: $\frac{S_{1'}^2}{S_1^2} = r_{1.2}^2 = B$

Zusammensetzung der Varianz des Kriteriums: $S_1^2 = S_{1'}^2 + S_{1\text{-}2}^2$

Anteil nicht vorhersagbarer Varianz: $1 - B = 1 - r_{1.2}^2$

Mittelwerte und Varianzen[1]

	Kriterium X_1	Prädiktor X_2	Schätzung $X_{1'}$	Fehler $X_{1\text{-}2}$
Mittelwert	M_1	M_2	M_1	0
Varianz	S_1^2	S_2^2	$r_{1.2}^2 \cdot S_1^2$	$(1 - r_{1.2}^2) \cdot S_1^2$

Korrelationen[2]

	Kriterium X_1	Prädiktor X_2	Schätzung $X_{1'}$	Fehler $X_{1\text{-}2}$
Kriterium X_1	1	$r_{1.2}$	$\lvert r_{1.2} \rvert$	$+\sqrt{1 - r_{1.2}^2}$
Prädiktor X_2		1	± 1	0
Schätzung $X_{1'}$			1	0
Fehler $X_{1\text{-}2}$				1

---

1 Da $M_{1\text{-}2} = 0$ (Voraussetzung bester Schätzung) und $M_1 = M_{1'} + M_{1\text{-}2}$ folgt, daß $M_{1'} = M_1$

2 $r_{1.1'}$ ist vom Betrag $r_{1.2}$ da $X_{1'}$ eine lineare Transformation von X_2 ist, welche die Korrelation nicht ändert. Sie kann nur nie neg. sein, da $X_{1'}$ und X_1 stets ist die gleiche Richtung (also positiv) kovariieren, auch wenn Prädiktor und Kriterium das nicht immer tun (bei r kleiner 0).

$$r_{1.(1\text{-}2)} = \frac{S_{1.(1-2)}}{S_1 \cdot S_{1-2}} = \frac{\frac{1}{N}\sum X_1 X_{1-2} - \overbrace{M_1 M_{1-2}}^{=0}}{S_1 \cdot \sqrt{S_1^2\left(1 - r_{1.2}^2\right)}} = \frac{\frac{1}{N}\sum \left(X_{1'} + X_{1-2}\right) \cdot X_{1-2}}{S_1^2\sqrt{1 - r_{1.2}^2}} = \frac{\underbrace{\frac{1}{N}\sum X_{1'}X_{1-2} - \overbrace{M_1 M_{1-2}}^{=0}}_{=S_{1'.(1-2)}=0} + \overbrace{\underbrace{\frac{1}{N}\sum X_{1-2}X_{1-2} - \underbrace{M_{1-2}M_{1-2}}_{=0}}_{=S_{1-2}^2 = S_1^2(1-r_{1.2}^2)}}^{=S_{1-2}^2=S_1^2(1-r_{1.2}^2)}}{S_1^2\sqrt{1 - r_{1.2}^2}} =$$

$$= \frac{S_1^2\left(1 - r_{1.2}^2\right)}{S_1^2\sqrt{1 - r_{1.2}^2}} = \frac{\left(1 - r_{1.2}^2\right) \cdot \sqrt{1 - r_{1.2}^2}}{\sqrt{1 - r_{1.2}^2} \cdot \sqrt{1 - r_{1.2}^2}} = \frac{\left(1 - r_{1.2}^2\right) \cdot \sqrt{1 - r_{1.2}^2}}{\left(1 - r_{1.2}^2\right)} = +\sqrt{1 - r_{1.2}^2} \text{ q.e.d.}$$

PUNKTSCHÄTZUNGEN UND VERTRAUENSINTERVALLE

Man kann aus einer Stichprobe die Pupulationsparameter der Population schätzen. Erwartungstreu nennt man dabei diejenigen Schätzungen, deren Mittelwert bei einer sehr großen Anzahl von Stichproben mit dem zu schätzenden Parameter der Population übereinstimmt. Man sagt auch, er hat keinen „bias“, keine systematische Abweichung.

Punktschätzungen

Stichprobe	Populationsschätzung
$b_{1.2}$	$\beta_{1.2}$ entspr. vom Wert $b_{1.2}$
a	α entspr. vom Wert a
$S_{1\text{-}2}{}^2$	$\hat{\sigma}_e^2 = \frac{N}{N-2} S_{1-2}^2$
$S_1{}^2$	$\hat{\sigma}_1^2 = \frac{N}{N-1} S_1^2$
$r_{1.2}$	$r_{1.2}$ für $\rho_{1.2}$ hat bias von der Größe 1/N, ist also ab N = 50 vernachlässigbar
$S_{1\text{-}2}$	$\sqrt{\hat{\sigma}_e^2}$ hat einen bias

Wahrscheinlichkeitsbereiche und Vertrauensintervalle

Mit dem Begriff „Wahrscheinlichkeitsbereich“ meint man einen Zahlenbereich $\omega_{1\text{-}\alpha}$, der einen Populationsparameter zum Mittelpunkt hat und innerhalb dessen sich ein Stichprobenkennwert, der diesen Parameter schätzt, mit einer gewissen Wahrscheinlichkeit (1 - α) liegt. Der Bereich wäre dann:

$$\beta_{1.2} - \omega_{1\text{-}\alpha} \le b_{1.2} \le \beta_{1.2} + \omega_{1\text{-}\alpha}$$

Die umgekehrte Fragestellung ist jedoch in der Inferenzstatistik wichtiger: Der Bereich um eine aus der Stichprobe bestimmte Schätzung, der mit vorgegebener Wahrscheinlichkeit (1-α) den Populationsparameter (bzw. Populationswert einer einzelnen Person) enthält. Dieser Bereich heißt das „Vertrauensintervall“.

(Obige Gleichung - $\beta_{1.2}$)	(Multiplikation mit -1)	(Addition von $b_{1.2}$)
$-\omega_{1\text{-}\alpha} \le b_{1.2} - \beta_{1.2} \le +\omega_{1\text{-}\alpha}$	$-\omega_{1\text{-}\alpha} \le \beta_{1.2} - b_{1.2} \le +\omega_{1\text{-}\alpha}$	$b_{1.2} - \omega_{1\text{-}\alpha} \le \beta_{1.2} \le b_{1.2} + \omega_{1\text{-}\alpha}$

Die Vertrauensintervalle um a, $b_{1.2}$ und X'_{1i} (für α, β und X_{1i}) sind:

$$a \pm \sqrt{F_\alpha}\,\frac{1}{N}\sqrt{\frac{\hat{\sigma}_e^2 \sum_i X_{2i}^2}{S_2^2}};$$

$$b_{1.2} \pm \sqrt{F_\alpha \frac{\hat{\sigma}_e^2}{N \cdot S_2^2}};$$

$$X'_{1i} \pm \sqrt{F_\alpha \cdot \hat{\sigma}_e^2 \cdot \left(1 + \frac{1}{N} + \frac{\left(X_{2i} - M_2\right)^2}{N \cdot S_2^2}\right)}$$

DIE REGRESSIONSGLEICHUNG BEI STANDARDISIERTEN VARIABLEN

Bei standardiesierten Variablen vereinfacht sich die Regressionsgleichung folgendermaßen:

$$\underbrace{z_{1i} = \frac{X_{1i} - M_1}{S_1}}_{\text{Standardisierung}} \Rightarrow M_{(z)1} = 0;\ S_{(z)1},\ S^2_{(z)1} = 1$$

Da $a = M_1 - b_{1.2}M_2$, und $M_1, M_2 = 0 \Rightarrow a = 0$;

außerdem gilt: $b_{1.2} = \frac{S_{1.2}}{S^2_{(z)2}}$ und da $S_{(z)1} = S_{(z)2} = 1 \Rightarrow b_{1.2} = r_{1.2}$;

Die Regressionsgleichung wird also nun zu: $z'_1 = r_{1.2}\, z_2$ und $z_{1\text{-}2} = z_1 - r_{1.2}z_2$

KORRELATION UND REGRESSION BEI ALTERNATIVEN UND DICHOTOMEN VARIABLEN

Bei alternativen Variablen mit den Ausprägungen 0 und 1 (z.B. 0 = männlich, 1 = weiblich), oder bei zweigeteilten Variablen (dichotom, wie z.B. Einteilungen in „groß“ = 1 ab 1,80 m Körpergröße und „klein“ = 0 unter 1,80 m) lassen sich ebenfalls Korrelationen bestimmen. Bei der Korrelation einer alternativen mit einer quantitativen Variablen ist jedoch nur die regressionsanalytische Betrachtung bedeutungsvoll, bei der die alternative Variable die Rolle des Prädiktors und die quantitative Variable die Rolle des Kriteriums spielt (andersrum könnten sich Schätzungen ergeben die unsinnig wären, wie z.B. Mit einem Einkommen von X DM ist man 0,83 Mann! im Gegensatz zu: Als Mann/ Frau liegt das geschätzte Einkommen bei X DM).

Bei der Korrelation *zweier* alternativer Variablen hat B (Quadrat des Korrelationskoeffizienten $r_{1.2}$) keine Bedeutung als Anteil erklärter Varianz (da die Werte ja nicht im eigentlichen Sinne mit Zwischenstufen variieren, sondern nur 0 und 1 annehmen). Der Korrelationskoeffizient hat nur noch die Bedeutung eines nicht weiter definierten Zusammenhangsmaßes mit den Grenzen -1 und 1. Bei echten alternativen Variablen kann eine Korrelation vom Wert 1 nur bei gleichen Randverteilungen zustandekommen. Die Korrelation von quantitativen mit alternativen Variablen läßt sich mit dem F-Test auf Überzufälligkeit überprüfen, die Korrelation zweier alternativer Variablen mit dem Chi-Quadrat-Test.

	0	1	
0	22	0	22
1	0	37	37
	22	37	

Beispiel: Zusammenhang zwischen Geschlecht und Note in einer Prüfung.

	Note 1	**Note 2**	**Note 3**	**Note 4**	**Note 5**
männlich (1)	2	5	7	4	1
weiblich (0)	3	5	8	3	0

Jetzt ließe sich wie gewohnt rechnen:

r^2 stellt nun den Anteil der Varianz der Note dar, der aus dem Geschlecht erklärbar ist.

X_2	X_1
1	1
1	1
1	2
1	2
usw.	usw.
0	4

DIE VARIANZANALYTISCHE DARSTELLUNG DER ERGEBNISSE EINER REGRESSIONS- BZW. KORRELATIONSANALYSE

Bei der Varianzanalyse werden Varianzanteile in Abweichungsquadraten (Varianz mal N) einander gegenübergestellt, um die Frage zu klären, ob ein sog. Faktor überzufällige Unterschiede in einem bestimmten Kriterium hervorbringt. Die Ausprägungen dieses Faktors entsprechen den Ausprägungen des Prädiktors der Regressions- bzw. Korrelationsanalyse. Die Unterschiedlichkeit im Kriterium, die sich nicht auf Unterschiede im Prädiktor zurückführen lassen, also die unsystematische Variation *innerhalb* jeweiliger Ausprägungen des Faktors entsprechen dem Fehler in der Regressions- bzw. Korrelationsanalyse. Die übliche Darstellung der Ergebnisse ist die Varianzanalysetabelle, wo man bequem Abweichungsquadrate innerhalb (Fehler) und zwischen (Schätzung) den Ausprägungen des Faktors gegenübergestellt sieht.

Quelle *der Variation*	SAQ *Summe der Abweichungsquadrate*	FG *Freiheitsgrade*	MAQ (= $^{SAQ}/_{FG}$) *Mittleres Abweichungsquadrat*	F $\left(\frac{MAQ_{zwischen}}{MAQ_{innerhalb}}\right)$
Regression *zwischen den Faktoren (Zähler)*	$\sum\left(X'_{1i} - M_1\right)^2 = N \cdot S^2_{1'} = N \cdot r^2_{1.2} \cdot S^2_1$	n - 1	$\frac{N \cdot S^2_1 \cdot r^2_{1.2}}{n-1}$	$\frac{MAQ_{Regression}}{MAQ_{Fehler}} =$
Fehler *innerhalb der Faktoren (Nenner)*	$\sum\left(X_{1i} - X_{1'i}\right)^2 = \sum X^2_{(1-2)i} = NS^2_{1-2} = N \cdot S^2_1\left(1 - r^2_{1.2}\right)$	N - n	$\frac{N \cdot S^2_1 \cdot \left(1 - r^2_{1.2}\right)}{N-n}$	$\frac{(N-n) \cdot r^2_{1.2}}{(n-1) \cdot \left(1 - r^2_{1.2}\right)}$
Total	$\sum\left(X_{1i} - M_1\right)^2 = N \cdot S^2_1$	N-1		

Für die *Multiple Regression* stelle man sich an allen Stellen, an denen $r_{1.2}$ bzw $r^2_{1.2}$ steht $R_{1.23...n}$ bzw. $R^2_{1.23...n}$ vor.

LINEARKOMBINATIONEN

Linearkombinationen sind additive Verknüpfungen mehrerer Geradengleichungen:

$$U = (w_1 X_1 + c_1) + (w_2 X_2 + c_2) + ... + (w_n X_n + c_n)$$

Durch Zusammenfassenn der Konstanten c_i ergibt sich die Form:

$$U = w_1 X_1 + w_2 X_2 + ... + w_n X_n + c \quad \text{mit} \quad c = c_1 + c_2 + ... + c_n$$

Mittelwert einer Linearkombination

$$M_U = w_1 M_1 + w_2 M_2 + ... + w_n M_n + c$$

z.B.: $U_1 = 7X_1 + 8X_2 - 5X_3 - 5$; $M_{U_1} = 7M_1 + 8M_2 - 5M_3 - 5$

Varianz einer Linearkombination

wird bestimmt durch die KKM = Komponenten-Kovarianz-Matrix; sie sieht allgemein folgendermaßen aus:

	$w_1 X_1$	$w_2 X_2$	...	$w_n X_n$
$w_1 X_1$	$w_1^2 S_1^2$	$w_1 w_2 S_{1.2}$	...	$w_1 w_n S_{1.n}$
$w_2 X_2$	$w_1 w_2 S_{1.2}$	$w_2^2 S_2^2$	...	$w_2 w_n S_{2.n}$
...	...	...	...	
$w_n X_n$	$w_1 w_n S_{1.n}$	$w_2 w_n S_{2.n}$	...	$w_n^2 S_n^2$

z.B.:
Linearkombination $4X_1 - 5X_2 + 2X_3$

	$4X_1$	$-5X_2$	$2X_3$
$4X_1$	$16 S_1^2$	$-20 S_{1.2}$	$8 S_{1.3}$
$-5X_2$	$-20 S_{1.2}$	$25 S_2^2$	$-10 S_{2.3}$
$2X_3$	$8 S_{1.3}$	$-10 S_{2.3}$	$4 S_3^2$

Symbol für Matrix: C_{UU}; Summe der Matrix: $S(C_{UU})$
$S_U^2 = S(C_{UU})$ Varianz = Summe der Elemente der Matrix

Kovarianz zweier Linearkombinationen

wird auch über die KKM errechnet: U_1 in die Zeile, U_2 in die Spalte

z.B.: $U = 4X_1 - 3X_2 + 8X_3 - 6$
$V = -2X_4 + 2X_5 + 3$

	$4X_1$	$-3X_2$	$8X_3$
$-2X_4$	$-8 S_{1.4}$	$6 S_{2.4}$	$-16 S_{3.4}$
$2X_5$	$8 S_{1.5}$	$-6 S_{2.5}$	$16 S_{3.5}$

Symbol für die Matrix: C_{UV}; Summe dieser Matrix: $S(C_{UV})$
$S_{U.V} = S(C_{UV})$ Kovarianz der Linearkombinationen = Summe der Elemente der Matrix

Die „totale" KKM

	U	V
U	C_{UU}	C_{UV}
V	C_{UV}	C_{VV}

Vorteil: In einer Matrix können beide Varianzen sowie die Kovarianz bestimmt werden; man benötigt sie für das Berechnen der Korrelation zweier Linearkombinationen:

$$r_{U.V} = \frac{S_{UV}}{S_U \cdot S_V} = \frac{S(C_{UV})}{\sqrt{S(C_{UU})} \cdot \sqrt{S(C_{VV})}}$$

PARTIELLE, SEMIPARTIELLE KORRELATION

Definition: „Partielle Korrelation“ heißt eine korrelationsstatistische Technik, die es erlaubt, die Korrelation zwischen zwei Variablen X_1 und X_2 so zu bestimmen, daß diese Korrelation nicht durch Beziehungen von X_1 und X_2 zu weiteren Variablen X_3, X_4, ..., X_n beeinflußt ist.

Mathematische Herleitung

Für den Fehler gilt: $X_{1\text{-}2} = X_1 - X'_1$ bzw. bei standardisierten Variablen: $z_{1\text{-}2} = z_1 - z'_1$.
Die Schätzung zerlegt sich in: $X'_1 = b_{1.2}\, X_2 + a$ bzw. bei standardisierten Variablen $z'_1 = r_{1.2}\, z_2$.
Als Rest- oder *Partialvariable* bezeichnet man den Teil der Schätzung, der nicht aus dem Prädiktor vorhergesagt werden kann. Der nicht aus z_3 vorhersagbare (bzw. von der Störvariable z_3 unabhängige) Teil von z_1, bzw. z_2 ist also $z_{1\text{-}3}$ und $z_{2\text{-}3}$, für die gilt:

$$z_{1\text{-}3} = z_1 - r_{13}\, z_3 \text{ und } z_{2\text{-}3} = z_2 - r_{23}\, z_3$$

Die Partialvariablen $z_{1\text{-}3}$ und $z_{2\text{-}3}$ sind mit dem Prädiktor z_3 unkorreliert: $r_{3.(1\text{-}3)} = r_{3.(2\text{-}3)} = 0$
Die Korrelation dieser beiden Partialvariablen stellt also den Zusammenhang zwischen z_1 und z_2 dar, der nicht mehr durch z_3 beeinflußt wird.

$$r_{(1\text{-}3)(2\text{-}3)} = \frac{S_{(1-3)(2-3)}}{S_{(1-3)} \cdot S_{(2-3)}} = r_{12\text{-}3} \text{ (andere Schreibweise)}$$

$S_{(1\text{-}3)(2\text{-}3)} = S(C_{(1\text{-}3)(2\text{-}3)})$ = Kovarianz der Linearkombinationen $z_1 - r_{13}\, z_3$ und $z_2 - r_{23}\, z_3$

	z_2	$-r_{23}\, z_3$
z_1	r_{12}	$-r_{23}r_{13}$
$-r_{13}\, z_3$	$-r_{13}r_{23}$	$r_{13}r_{23}$

$S(C_{12\text{-}3}) = r_{12} - r_{13}r_{23} - r_{13}r_{23} + r_{13}r_{23} = r_{12} - r_{13}r_{23} = S_{12\text{-}3} = S_{1(2\text{-}3)}$
$S_{1\text{-}3} = \sqrt{1-r_{13}^2}$ und $S_{2\text{-}3} = \sqrt{1-r_{23}^2}$ (Standardabweichung des Fehlers)
Die Partielle Korrelation der Variablen X_1 mit X_2 ohne den Einfluß von X_3 ist also:

$$r_{12\text{-}3} = \frac{r_{12} - r_{13}r_{23}}{\sqrt{1-r_{13}^2} \cdot \sqrt{1-r_{23}^2}}$$

Da sich die Korrelation bei linearen Transformationen der Variablen nicht verändert, gilt die Formel auch für die nichtstandardisierten Rohwerte.

Wird die Störvariable nur aus einer der Variablen z_1 oder z_2 auspartialisiert, spricht man von der *Semipartiellen Korrelation* $r_{1(2\text{-}3)}$.

$$r_{1(2\text{-}3)} = \frac{S_{1(2-3)}}{S_1 \cdot S_{(2-3)}} = \frac{r_{12} - r_{13}r_{23}}{\sqrt{1-r_{23}^2}}$$

Multiple partielle Korrelation (Auspartialisierung mehrerer Störvariablen)[3]

Ausgangsvariablen: z_1, z_2; z_3, z_4 (die letzten zwei seien Störvariablen)

1. Schritt: Auspartialisieren von z_3 ergibt Partialvariablen 1. Ordnung:
   $z_{1\text{-}3} = z_1 - r_{13}z_3$
   $z_{2\text{-}3} = z_2 - r_{23}z_3$
   $z_{4\text{-}3} = z_4 - r_{34}z_3$ (Auch aus der zweiten Störvariable muß auspartialisiert werden!)

2. Schritt: Auspartialisieren von $z_{4\text{-}3}$ ergibt Partialvariablen 2. Ordnung:
   $z_{1\text{-}34} = z_{1\text{-}3} - b_{1\text{-}3.4\text{-}3}\, z_{4\text{-}3}$
   $z_{2\text{-}34} = z_{2\text{-}3} - b_{2\text{-}3.4\text{-}3}\, z_{4\text{-}3}$

3. Schritt: Die einfache Korrelation dieser beiden Partialvariablen 2. Ordnung ist die gesuchte partielle Korrelation 2. Ordnung:
   $r_{12\text{-}34} = \frac{S_{(1-34)(2-34)}}{S_{1-34} \cdot S_{2-34}}$, allgemein: $r_{12\text{-}34\ldots n} = \frac{S_{12-34\ldots n}}{S_{1-34\ldots n} \cdot S_{2-34\ldots n}}$

Die Reihenfolge der Auspartialisierung ist gleichgültig!

Partieller und semipartieller Regressionskoeffizient

Der *Partielle Regressionskoeffizient* $b_{1\text{-}3.2\text{-}3}$ entsteht durch Regression einer einfachen Partialvariable auf eine andere Partialvariable. Der *Semipartielle Regressionskoeffizient* $b_{1.2\text{-}3}$ entsteht durch Regression einer vollständigen Variablen X_1 auf die Partialvariable $X_{2\text{-}3}$. Es gilt aber:

$$b_{1\text{-}3.2\text{-}3} = b_{1.2\text{-}3} = \frac{S_{1-3}}{S_{2-3}} r_{12\text{-}3} = \frac{S_1 \cdot \left(r_{12} - r_{13}r_{23}\right)}{S_2 \cdot \left(1 - r_{23}^2\right)},$$

$$\text{allgemein: } b_{1.n\text{-}23\ldots(n\text{-}1)} = \frac{S_1}{S_{n-23\ldots(n-1)}} \cdot r_{1(n-23\ldots(n-1))}$$

[3] Achtung! Immer cool bleiben, auch wenn's etwas schwieriger aussieht als es ist!

MULTIPLE REGRESSION UND KORRELATION

Ähnlich wie bei der einfachen Regression geht es um die beste lineare Vorhersage einer Variablen, nur eben aus mehreren Variablen. Es gilt wieder: $X_1 = X'_1 + X_{Rest}$, wobei X'_1 eine Linearkombination aus den Prädiktoren von der Form $X'_1 = w_2X_2 + w_3X_3 + ... + w_nX_n + c$ ist. Das Minimum-Quadrat-Prinzip ist wieder ausschlaggebend (Summe der quadrierten Reste soll minimal sein): $\sum_i X^2_{(Rest)i} = \sum_i (X_{1i} - X'_{1i})^2 = \text{Minimum}$. Diese Bedingung ist dann erfüllt, wenn die Gewichte w_2 bis w_n semipartielle Regressionskoeffizienten sind und die Schätzung also die Form $X'_1 = b_{1.2\text{-}34...n} X_2 + b_{1.3\text{-}24...n} X_3 + ... + b_{1.n\text{-}23...(n\text{-}1)} X_n + a$ hat, wobei $a = M_1 - b_{1.2\text{-}34...n} M_2 - b_{1.3\text{-}24...n} M_3 - ... - b_{1.n\text{-}23...(n\text{-}1)} M_n$ ist.

Zwei andere Formen:

1. $X'_1 = b_{1.2\text{-}34...n} (X_2 - M_2) + b_{1.3\text{-}24...n} (X_3 - M_3) + ... + b_{1.n\text{-}23...(n\text{-}1)} (X_n - M_n) + M_1$ und

2. $X'_1 = M_1 + b_{1.2} (X_2 - M_2) + b_{1.3\text{-}2} X_{3\text{-}2} + ... + b_{1.n\text{-}23...(n\text{-}1)} X_{n\text{-}23...(n\text{-}1)}$

Bedingungen: Zwischen den Prädiktoren darf keine vollständige lineare Beziehung bestehen und die Anzahl N der beobachteten Personen muß größer sein als die Anzahl n der Variablen der Regressionsgleichung.

Restvariable $X_{Rest} = X_1 - X'_1 = X_{1\text{-}23...n}$
Kriteriumsvarianz: $S_1^2 = S_{1'}^2 + S^2_{1\text{-}23...n}$
Kovarianz $S_{1(1')} = S_{1'}^2$

Multiple Korrelation

ist die einfache Korrelation der Schätzung einer multiplen Regression mit dem Kriterium

$$R_{1.23...n} = r_{1(1')} = \frac{S_{1(1')}}{S_1 \cdot S_{1'}} = \frac{S^2_{1'}}{S_1 \cdot S_{1'}} = \frac{S_{1'}}{S_1},$$

$R^2_{1.23...n} = \frac{S^2_{1'}}{S^2_1}$ = multiples Bestimmtheitsmaß = Anteil der im Kriterium aus den Prädiktoren erklärtern Varianz. Varianz des Fehlers: $S^2_{1\text{-}23...n} = S_1^2 (1 - R^2_{1.23...n})$

$$R^2_{1.23...n} = \frac{S_2}{S_1} r_{12} b_{1.2\text{-}34...n} + ... + \frac{S_n}{S_1} r_{1n} b_{1.n\text{-}23...(n\text{-}1)}$$

Bei wechselseitig orthogonalen Prädiktoren ergibt sich die vereinfachte Formel:

$$R^2_{1.23...n} = r^2_{12} + r^2_{13} + ... + r^2_{1n}$$

DER SIGNIFIKANZTEST BEI DER MULTIPLEN REGRESSION UND KORRELATION

H_0: $\beta_2 = \beta_3 = ... = \beta_n = 0$
H_1: $\beta_2 \neq 0$ und/ oder $\beta_3 \neq 0$ ... und/ oder $\beta_n \neq 0$

Zählerfreiheitsgrade: Anzahl der Variablen minus 1: n - 1
Nennerfreiheitsgrade: Anzahl der Personen minus Anzahl der Variablen: N - n

Berechnung des F-Wertes: $$F = \frac{(N-n)\cdot R^2_{1.23...n}}{(n-1)\cdot\left(1-R^2_{1.23...n}\right)}$$

Aufsuchen des kritischen F-Wertes für die entsprechenden Zähler- und Nennerfreiheitsgrade in einer Tabelle
$F \geq F_{krit}$: H_0 verwerfen und H_1 annehmen
$F < F_{krit}$: H_0 beibehalten

Prüfung einzelner Regressionskoeffizienten
(Nur sinnvoll, wenn Zusammenhangshypothese bestätigt wurde)
H_0: $\beta_n = 0$
H_1: $\beta_n \neq 0$
Zählerfreiheitsgrade: 1
Nennerfreiheitsgrade: N - n

Berechnung des F-Wertes: $$F = \frac{(N-n)\cdot\left(R^2_{1.23...n} - R^2_{1.23...(n-1)}\right)}{1-R^2_{1.23...n}}$$

Vergleich mit kritischem F-Wert aus der Tabelle

Beispiel

Variablen Test-Intelligenz (X_1), Ängstlichkeit (X_2), Leistungsmotivation (X_3) und Rigidität (X_4)

	X1	X2	X3	X4
1	3,00	5,00	6,00	4,00
2	2,00	6,00	3,00	8,00
3	2,00	6,00	5,00	6,00
4	0,00	9,00	4,00	8,00
5	4,00	2,00	6,00	2,00
6	2,00	6,00	3,00	8,00
7	5,00	0,00	0,00	7,00
8	2,00	6,00	7,00	4,00
9	0,00	1,00	4,00	0,00
10	1,00	3,00	1,00	8,00
M	2,100	4,400	3,900	5,500
S	1,513	2,653	2,119	2,729

N of Cases = 10

Correlation

	X1	X2	X3	X4
X1	1,000	-,458	-,090	-,012
X2	-,458	1,000	,381	,469
X3	-,090	,381	1,000	-,562
X4	-,012	,469	-,562	1,000

Multiple $R_{1.234}$ 0,87686
$R^2_{1.234}$ 0,76888

Analysis of Variance

	DF	SAQ	MAQ
Regression	3	17,60727	5,86909
Fehler	6	5,29273	0,88212

F = 6,65338 Signif F = ,0245

Semipartieller Regressionskoeffizient

X2 -1,192694 ($b_{1.2\text{-}34}$)
X3 1,305111 ($b_{1.3\text{-}24}$)
X4 1,106949 ($b_{1.4\text{-}23}$)
Constant -3,830301 ($a_{1.234}$)

Multiple $R_{1.24}$ 0,51272
Multiple $R_{1.34}$ 0,11818
Multiple $R_{1.23}$ 0,46718

FAKTORENANALYSEN NACH DEM HAUPTKOMPONENTENMODELL

1. Standardisierte Datenmatrix

Variablen	Personen				
	1	2	3	... i ...	N
z_1	z_{11}	z_{12}	z_{13}	z_{1i}	z_{1N}
z_j	z_{j1}	z_{j2}	z_{j3}	z_{ji}	z_{jN}
z_n	z_{n1}	z_{n2}	z_{n3}	z_{ni}	z_{nN}

2. Korrelationsmatrix

Variablen	Variablen			
	z_1	z_2	... z_j ...	z_n
z_1	1	r_{12}	r_{1j}	r_{1n}
z_2		1	r_{2j}	r_{2n}
z_j			1	r_{jn}
z_n				1

3. Faktor

Ein Faktor eines Variablensatzes ist eine Linearkombination der Variablen
$F_J = w_{1J}z_1 + ... + w_{jJ}z_j + ... + w_{nJ}z_n$
w_{jJ}: Gewicht der Variablen z_j im Faktor F_J

4. Faktorwert

Faktorwert der Person i im Faktor F_J:
$F_{Jj} = w_{1J}z_{1i} + ... + w_{jJ}z_{ji} + ... + w_{nJ}z_{ni}$

5. Faktorwertematrix

Faktoren	Personen			
	1	2	... i ...	N
F_A	F_{A1}	F_{A2}	F_{Ai}	F_{AN}
F_B	F_{B1}	F_{B2}	F_{Bi}	F_{BN}
F_J	F_{J1}	F_{J2}	F_{Ji}	F_{JN}
F_M	F_{M1}	F_{M2}	F_{Mi}	F_{MN}

6. Faktorgewichtematrix

Faktorgewicht: w_{jJ}: Gewicht der Variablen z_j im Faktor F_J

Variablen	Faktoren			
	F_A	F_B	... F_J ...	F_M
z_1	w_{1A}	w_{1B}	w_{1J}	w_{1M}
z_j	w_{jA}	w_{jB}	w_{jJ}	w_{jM}
z_n	w_{nA}	w_{nB}	w_{nJ}	w_{nM}

7. Faktorladung

Die Faktorladung ist die Korrelation einer Variablen des Datensatzes mit einem Faktor des Datensatzes, symbolisiert durch „a“

Ladung der Variablen z_j im Faktor $F_A = a_{jA}$

$$a_{jA} = r_{jA} = \frac{1}{N}\sum_i F_{Ai}z_{ji} = w_{1A}r_{1j} + ... + w_{jA} + ... + w_{nA}r_{jn}$$

(gewichtete Summe der Korrelationen)

8. Faktorladungsmatrix

Variablen	Faktoren			
	F_A	F_B	... F_J ...	F_M
z_1	a_{1A}	a_{1B}	a_{1J}	a_{1M}
z_j	a_{jA}	a_{jB}	a_{jJ}	a_{jM}
z_n	a_{nA}	a_{nB}	a_{nJ}	a_{nM}

9. Kommunalität

... ist der in einer Variablen aus dem Faktor vorhersagbare Varianzanteil, symbolisiert durch „h^2_j“

$h^2_j = r^2_{jA} = a^2_{jA}$

Erweiterter Begriff: Der aus mehreren Faktoren vorhersagbare Teil einer Variablen z_j

$h^2_j = R^2_{j.AB...M} = a^2_{jA} + a^2_{jB} + ... + a^2_{jM}$

10. Varianzanteilmatrix

	F_A	F_B	... F_J ...	F_M
z_1	a^2_{1A}	a^2_{1B}	a^2_{1J}	a^2_{1M}
zj	a^2_{jA}	a^2_{jB}	a^2_{jJ}	a^2_{jM}
z_n	a^2_{nA}	a^2_{nB}	a^2_{nJ}	a^2_{nM}

Zeilensumme: $h^2_j \leq 1$
Kommunalität der Variablen = die durch alle Faktoren erklärte Varianz in einer Variablen.

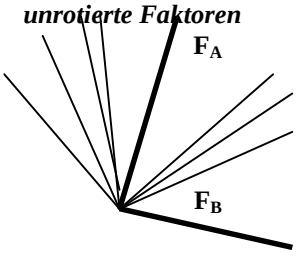
Spaltensumme: Eigenwert des Faktors = Betrag der Varianz, die durch einen Faktor erklärt wird

Zeilensumme = Spaltensumme

Bei n Variablen höchstens n Faktoren extrahierbar. Bei n Faktoren werden die Variablen vollständig erklärt.
Der Eigenwert eines Faktors sollte mindestens so groß sein wie die Varianz, die eine einzige Variable erklärt, nämlich 1 (standardisierte Variablen!).

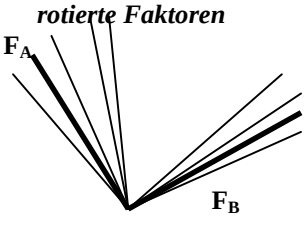
Faktorenrotation

Die Rotation der Faktoren führt dazu, daß die hierarchische Stufung der Faktoren (1. extrahierter Faktor erklärt am meisten Varianz, dann 2., usw.) abgeschwächt wird. Die Eigenwerte unterscheiden sich nicht mehr so sehr wie bei unrotierten Faktoren. Durch die Rotation ergeben sich also neue Faktorladungen, die Kommunalitäten (durch alle Faktoren erklärte Varianz in einer Variablen) bleiben allerdings gleich.



Bei der Einfachstruktur (nach Rotation) lädt jede Variable im Idealfall nur in einem Faktor.

Vorteil: Man hat voneinander unabhängige Faktoren, die jeweils für sich sehr gut eine Gruppe von Variablen repräsentieren.



DIE VARIANZANALYSE: KLASSISCH UND MIT DUMMY-VARIABLEN

Klassische Varianzanalytische Berechnungsweise:

Rohwerte und Berechnung der Mittelwerte

Kontrollgruppe	Experimentalgruppe	
3	4	
1	8	
5	4	
7	8	
$\sum X_K = 16$	$\sum X_E = 24$	$\sum X_G = 40$
$\overline{X}_K = 4$	$\overline{X}_E = 6$	$\overline{X}_G = 5$

Berechnung der Gesamtsumme der Abweichungsquadrate (SAQ_G)

Summe aller quadrierten Abweichungen der Rohwerte vom Gesamtmittelwert:

$(3-5)^2 + (1-5)^2 + (5-5)^2 + (7-5)^2 + (4-5)^2 + (8-5)^2 + (4-5)^2 + (8-5)^2 =$
$= 4 + 16 + 0 + 4 + 1 + 9 + 1 + 9 = 44$

Berechnung der Summe der Abweichungsquadrate innerhalb der Gruppen (SAQ_i)

Summe der quadrierten Abweichungen der Rohwerte von ihrem jeweiligen Gruppenmittelwert:

$(3-4)^2 + (1-4)^2 + (5-4)^2 + (7-4)^2 + (4-6)^2 + (8-6)^2 + (4-6)^2 + (8-6)^2 =$
$= 1 + 9 + 1 + 9 + 4 + 4 + 4 + 4 = 36$

Berechnung der Summe der Abweichungsquadrate zwischen den Gruppen (SAQ_z)

Summe der quadrierten Abweichungen der einzelnen Gruppenmittelwerte vom Gesamtmittelwert mal Anzahl n der Personen in den Gruppen:

$(4-5)^2 + (4-5)^2 + (4-5)^2 + (4-5)^2 + (6-5)^2 + (6-5)^2 + (6-5)^2 + (6-5)^2 = 8 \cdot 1 = 8$

Quelle der Variation	df	SAQ	MAQ	F	Sig.
zwischen den Gruppen	1	8	8	1,33	$p > 0,05$
innerhalb der Gruppen	6	20 + 16 = 36	6		
Gesamt	7	44			

Oder kürzer (Berechnung ohne Abweichungswerte)

	Kontrollgruppe	Experimentalgruppe		
	3	4		
	1	8		
	5	4		
	7	8		
$\sum X$	16	24	$\sum\sum X$	40
n	4	4	N	8
$\overline{X}$	4	6		
$\sum X^2$	84	160	$\sum\sum X^2$	**244**
$(\sum X)^2$	256	576		
$\frac{(\sum X)^2}{n}$	64	144	$\sum_{\text{p Gruppen}} \frac{(\sum X)^2}{n}$	**208**

Korrekturwert K (N mal $\overline{X}_G{}^2$) = $\frac{(\sum\sum X)^2}{N}$ = $\frac{40^2}{8}$ = **200**

Quelle der Variation	df	SAQ	MAQ	F	Sig.
zwischen den Gruppen	1	208 - 200 = 8	8	1,33	p > 0,05
innerhalb der Gruppen	6	244 - 208 bzw. 44 - 8 = 36	6		
Gesamt	7	244 - 200 = 44			

Mit Dummy-Variablen

Übersetzung der p Gruppen in p-1 Dummy-Variablen (0, 1-Variablen):

Abhängige Variable (X_1)	Unabhängige Variable (X_2)
3	0
1	0
5	0
7	0
4	1
8	1
4	1
4	1

$M_1 = 5$ $M_2 = 0,5$

$S_1{}^2 = 5,5$ $S_1{}^2 = 0,25$

Die Regressionsgleichung lautet:

$X_{1'} = 2X_2 + 4$

$r^2 = 0,182$, $F = \frac{(8-2)\cdot 0,182}{1-0,182} = 1.33$

AUFGABEN ZUR VARIANZ- UND FAKTORENANALYSE

1. Bei einer Faktorenanalyse werden aus 60 Variablen 5 Faktoren mit den Eigenwerten 16, 12, 8, 5.5, und 3.5 extrahiert. Welcher prozentuale Anteil der Gesamtvarianz der 60 Variablen wird durch die 5 Faktoren erklärt?

a) 80% b) 75% c) 11,25% d) 78,75% e) 95%

(1992, Nr. 23, b)

2. Die Faktorenanalyse nach der Hauptkomponentenmethode

a) prüft Mittelwertsunterschiede auf Signifikanz
b) rotiert vorher rotierte Faktoren auf Einfachstruktur
c) prüft, ob zwischen einer großen Prädiktorenzahl und einem Kriterium ein signifikanter Zusammenhang besteht
d) transformiert große Datensätze in Rangdaten
e) ist ein reduktionistisches Verfahren, das einen Überblick über eine große Faktorenzahl verschafft, indem wenige Faktoren einen möglichst großen Anteil der Gesamtvarianz der Ausgangsvariablen erklären.

(1990, Nr. 20, e)

3. Die Kommunalität h_{jJ}^2 ist — bei Vorliegen eines Faktors —

1. die quadrierte einfache Korrelation zwischen einer Variablen z_j und einem Faktor F_J
2. gleich h_j
3. der in der Variablen z_j aus dem Faktor F_J vorhersagbare Varianzanteil
4. die Faktorladung zum Quadrat
5. die über alle Personen gerechnete Summe der Produkte aus Variablenwert und Faktorwert

Richtig sind:

a) 1, 2 b) 1, 3, 4 c) 2, 3 d) 1, 4 e) 4, 5

(1990, Nr. 21, b)

4. Das Ziel der Faktorenanalyse ist,

1. die in einer größeren Variablenzahl enthaltene Information auf wenige Faktoren zu verdichten,
2. die Schätzfehler der Faktorwerte zu minimieren,
3. Linearkombinationen der gegebenen Variablen zu finden, die möglichst viel der Varianz der der Variablen erklären,
4. Informationen über Zusammenhänge der Variablen zu erhalten,
5. die Faktorgewichte miteinander zu korrelieren,
6. traumatische frühkindliche Faktoren hermeneutisch zu extrahieren.

Richtig sind:
a) 2, 3 b) 1, 3, 4 c) 5, 6 d) 4, 5, 6 e) 2, 4
(1989, Nr. 19, b)

5. Welche Bedeutung haben die Werte in der Faktorladungsmatrix?

a) Sie geben für alle Personen an, wie stark sie durch die Faktoren beladen sind.
b) Sie geben den Zusammenhang zwischen den verschiedenen Variablen an.
c) Sie geben den Zusammenhang zwischen den Variablen und den Faktoren an.
d) Sie geben den Zusammenhang zwischen den verschiedenen Faktoren an.
e) Sie geben den Zusammenhang zwischen den Personen und den Faktoren an.

(1989, 20, c)

6. Was versteht man unter einem Faktorwert?

a) Der Faktorwert ist der Wert einer Person in einem Faktor.
b) Der Faktorwert ist identisch mit der Kommunaltität.
c) Der Faktorwert ist für standardisierte Ausgangsvariablen mit dem sogenannten kritischen F-Wert der F-Tabelle zu vergleichen.
d) Der Faktorwert ist der Anteil an der Gesamtvarianz, der durch diesen Faktor erklärt wird.
e) Der Faktorwert gibt die Faktorladung an.

(1989, 21, a)

7. Als Faktor wird bezeichnet

1. diejenige Variable, deren Auswirkung varianzanalytisch untersucht werden soll.
2. ein aus mehr als 2 Dummy-Variablen bestehendes Kriterium.
3. eine Linearkombination $F_J = w_{1J}z_1 + w_{2J}z_2 + ... + w_{nJ}z_n$, deren Gewichte durch die Faktorenanalyse bestimmt wurden.
4. jede Variable, die mit dem Kriterium nicht, aber mit mindestens einem der Prädiktoren korreliert.
5. ein signifikanter Mittelwertsunterschied.

Richtig sind:

a) nur 1 b) nur 4 c) 1, 3 d) 2, 5 e) 3, 4, 5

(1987, Nr. 18, c)

8. Folgende Aussagen über die Faktorenanalyse sind zutreffend:

1. Die Kommunaltität einer Variablen kann höchstens den Wert 1 annehmen.
2. Die Summe der Kommunaltitäten der Variablen muß gleich sein der Summe der Eigenwerte aller Faktoren.
3. Der Eigenwert eines Faktors dividiert durch die Anzahl der Variablen gibt den Anteil der Gesamtvarianz an, der durch diesen Faktor erklärt wird.
4. Bei der Faktorenanalyse nach dem Hauptkomponentenverfahren ist der Eigenwert des zuerst extrahierten Faktors am größten.

Richtig sind:

a) alle Aussagen b) 1, 3 c)1, 2, 3 d) 1, 4 e) 1, 2
(1987, Nr. 19, a)

9. Bei einer Faktorenanalyse werden aus 48 Variablen 12 Faktoren extrahiert, wobei die Summe der Eigenwerte 36 beträgt. Welchen Anteil der Gesamtvarianz der 48 Variablen erklären die 12 Faktoren?

a) .25 b) .30 c) .40 d) .75 e) .95

(1987, Nr. 20, d)

10. Gegeben ist folgende Faktorladungsmatrix:

	F_A	F_B	F_C	F_D
z_1	.6	.4	.3	.5
z_2	-.4	.7	0	0
z_3	-.8	0	.6	0
z_4	.5	-.7	0	0
z_5	.4	0	.6	.6

Welche Variable wird durch alle Faktoren am besten erklärt?

a) z_1 b) z_2 c) z_3 d) z_4 e) z_5

(1986, Nr. 19, c)

11. Es wird eine dreifaktorielle Varianzanalyse mit 120 Personen durchgeführt, wobei der erste Faktor 2, der zweite Faktor 3 und der dritte Faktor wieder 2 Ausprägungen hat. Berechnen die $FG_{innerhalb}$ und $FG_{zwischen}$!

a) $FG_i = 12$, $FG_z = 108$ b) $FG_i = 108$ $FG_z = 12$ c) $FG_i = 108$ $FG_z = 11$ d) $FG_i = 11$ $FG_z = 108$ e) $FG_i = 109$ $FG_z = 11$

(Selbst erfunden, c)

12. Es soll mit folgenden Werten eine Varianzanalyse durchgeführt werden:

Kontrollgruppe	**Experimentalgruppe**
3	**1**
6	**0**
5	**3**
2	**2**

Berechnen Sie SAQ_i! SAQ_i =

a) 2,5 b) 15 c) 12,5 d) 6 e) 27,5

Selbst erfunden, b)

13. Berechnen Sie nun MAQ_z! MAQ_z =

a) 7 b) 2,5 c) 6 d) 27,5 e) 12,5

(Selbst erfunden, e)

14. Berechnen Sie nun den F-Wert! F =

a) 0,2 b) 2,5 c) 0,5 d) 5 e) 6

(Selbst erfunden, d)

15. Bei einer Varianzanalyse waren die FG_i = 56 und die FG_z = 3. Wieviele Personen waren beteiligt (N)?

a) 4 b) 60 c) 59 d) 5,5 e) 25

(Selbst erfunden, b)

16. Wieviele Gruppen gab es (bei voriger Analyse)?

a) 4 b) 3 c) 56 d) 9 e) 2

(Selbst erfunden, a)

17. Wieviele Personen waren in jeder der Gruppen (bei voriger Analyse)?

a) 5 b) 56 c) 10 d) 34 e) 15

(Selbst erfunden, e)

18. Bei einem F-Wert = 8, MAQ_z = 16, 4 Gruppen und FG_i = 12 hat SAQ_i welchen Wert?

a) 15 b) keinen Wert c) 24 d) 16 e) 12

(Selbst erfunden, c)

19. Bei der gleichen Analyse waren wieviele Personen in jeder Gruppe?

a) 4 b) 16 c) 5 d) 24 e) 16

(Selbst erfunden, a)

20. Wie groß war SAQ_z?

a) 12 b) 33 c) 45 d) 48 e) 23
(Selbst erfunden, d)

21. Wieviele Dummy-Variablen bräuchte man, wenn man diese Varianzanalyse als Regressionsanalyse rechnen wollte?

a) 1 b) 2 c) 3 d) 4 e) 5
(Selbst erfunden, c)

22. Bei der Statistik-Klausur werde ich

a) cool bleiben b) versagen c) improvisieren d) meditieren e) krank sein

ALLGEMEINES

Die Statistik ist ein Zweig der angewandten Mathematik und umfaßt zwei Hauptgebiete, nämlich die *beschreibende oder deskriptive* und die *schließende oder induktive bzw. Inferenzstatistik*. Das Ziel der Erstgenannten ist eine Ordnung und Reduktion von Datenmaterial in Tabellen, Grafiken und Kennwerten, welche das Datenmaterial gut *darstellt*. Sinn und Zweck der zweiten Disziplin ist es, mittels wahrscheinlichkeitstheoretischer Fundierung von einem Ausschnitt einer zu untersuchenden Grundgesamtheit oder Population — der Zufallsstichprobe — zulässige Aussagen über die Population zu treffen, die selbst meist zu groß ist, in ihrer Gesamtheit untersucht zu werden.

Einige Grundbegriffe und Definitionen

Variable: Eine Variable ist ein Merkmal, das in mindestens zwei Ausprägungen vorkommt. Ihre tatsächlichen Ausprägungen heißen Werte der Variablen.
Es gibt ferner *diskrete, stetige und quasistetige* Variablen. Diskrete Variablen haben nur *feste Ausprägungen* ohne Zwischenwerte (z.B.: Anzahl von Personen in einem Zimmer; es gibt keine 2,74 Personen!); stetige Variablen haben *unendlich viele Zwischenwerte* zwischen beliebigen Intervallen (z.B.: Die Zeit: zwischen 1 und 2 Sekunden gibt es unendlich viele Zwischenzeitpunkte); quasistetige Variablen haben zwar *theoretisch Zwischenwerte* zwischen einzelnen Ausprägungen, sie werden aber *in der Praxis nicht angegeben* (z.B.: Das Alter: Obwohl es zwar rein theoretisch möglich wäre, das exakte Alter einer Person zu einem gegebenen Zeitpunkt anzugeben — in ms etwa — macht man in der Praxis jedoch nur Altersangaben in Jahren). Da es aber sinnvolle Zwischenwerte gibt, darf man quasistetige Variablen wie stetige behandeln.

Grundgesamtheit oder Population: Das sind alle Objekte, die dieselbe Ausprägung eines *örtlichen, sachlichen und zeitlichen Identifikationsmerkmals* besitzen. Z.B.: Die Psychologiestudenten (sachl. I.) des Erstsemesters WS 1996/97 (zeitl. I.) an der LMU München (örtl. I.).

Diejenigen Merkmale, welchen an ihnen erhoben werden, nennt man *Erhebungsmerkmale*.

Teilmengen der Population heißen *Stichproben*. Hatten alle Elemente einer Population dieselbe Wahrscheinlichkeit in die Stichprobe aufgenommen zu werden (z.B. durch ein Losverfahren) nennt man die Teilmenge *Zufallsstichprobe*.

DESKRIPTIVE STATISTIK

Erhobene Daten können zunächst ungeordnet in einer Liste festgehalten werden; so eine Liste nennt man *Urliste*. Jede Person wird dabei einfach mit einer Zahl eines laufenden Index' festgehalten:

Beispiel:

Population = Die Psycho-Studenten des WS 96/97 an der LMU-München;
Stichprobe = Alle Personen, die jetzt in diesem Raum am Fenster sitzen (oder vielleicht auch halb liegen);
Erhebungsmerkmal = Schuhgröße

Person (i)	Schuhgröße (X_i)
1	
2	
3	
4	
5	
6	
7	
8	
9	
10	

Jetzt können wir z.B. schon mal das Arithmetische Mittel, oder kurz den Mittelwert, berechnen. Die Formel lautet:

$$\overline{X} = \frac{1}{N}\sum_{i=1}^{N} X_i$$

in unserem Beispiel heißt das:

$$\overline{X} = \frac{1}{\quad}(\quad + \quad + \quad + \ldots \qquad\qquad)$$

Wir können aber auch die Werte ordnen; das hat den Vorteil, daß wir Werte, die häufiger vorkommen, nicht mehrmals hinschreiben müssen. Dazu fertigen wir eine sog. Häufigkeitstabelle an, in der die Schuhgrößen bereits angeordnet sind (z.B. steigend) und eingetragen wird, wie oft diese Größe in unserer Stichprobe vorkommt. Der Index lautet jetzt nicht mehr i für die Person, sondern j für die jeweilige Schuhgröße (die ja öfter vorkommen kann); die absolute Häufigkeit einer Schuhgröße wird mit n_j bezeichnet, die relative (relativ zur Gesamtanzahl N) mit f_j wobei

$$f_j = \frac{n_j}{N} \text{ und } N = \sum_{j=1}^{J} n_j .$$

f_j liegt zwischen 0 und 1 ($0 \le f_j \le 1$) bzw. $\sum_{j=1}^{J} f_j = 1$

j	X_j	n_j	f_j	N_j	F_j
1	36				
2	37				
3	38				
4	39				
5	40				
6	41				
7	42				
8	43				

Der Mittelwert läßt sich jetzt auch anders berechnen:

$$\overline{X} = \frac{1}{N}\sum_{j=1}^{J} n_j X_j$$

und für unser Beispiel heißt das:

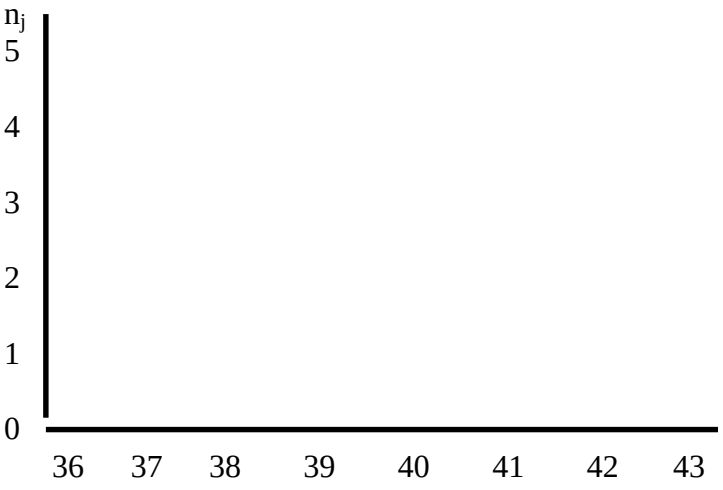
$$\overline{X} = \frac{1}{\quad}(... \cdot 36 + ... \cdot 37 + ... \qquad)$$

Die zwei fehlenden Statistiken N_j und F_j bedeuten folgendes:

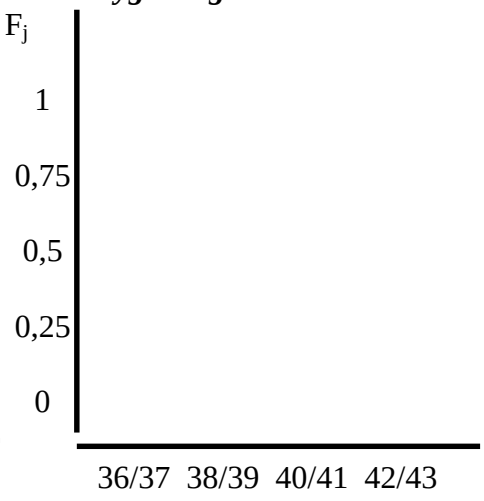
N_j ist die Kummulierte absolute Häufigkeit an Werten, also die Summe der absoluten Häufigkeiten n_j bis zu einem bestimmten j. F_j ist die Kummulierte relative Häufigkeit, also die Summe der relativen Häufigkeiten f_j bis zu einem bestimmten j.

Nun kann man das Ganze auch noch graphisch veranschaulichen mit Stab- und Säulendiagrammen; letztere werden auch Histogramme genannt. Bevor man jedoch Histogramme zeichnen kann, müssen die Werte erst in Klassen zusammengefaßt werden (z.B. von 35,5 - 37,5; 37,5 - 39,5; 39,5 - 41,5; 41,5 - 43,5).

Stabdiagramm



Histogramm bzw. Treppendiagramm bzw. Polygonzug



j	X_j	$\tilde{X}_j$	n_j	f_j	N_j	F_j
1	35,5	37,5				
2	37,5	39,5				
3	39,5	41,5				
4	41,5	43,5				

DIE APPROXIMATIVE VERTEILUNGSFUNKTION

Sind Werte in Klassen eingeteilt, und geht man davon aus, daß alle Werte in den Klassen gleich häufig vorkommen (Gleichverteilung), kann man mit der Approximativen Verteilungsfunktion annähern, welche *Kummulative relative Häufigkeit (F*) bei einem bestimmten Wert (x)* der Variablen erreicht ist (auch wenn die Kummulierten Häufigkeiten nur intervallweise gegeben sind). Die Frage dazu könnte etwa lauten: „Welche Kummulierte relative Häufigkeit ist bei einer Körpergröße von 1,83 m erreicht?“. Nimmt man diesen Wert F* mal 100, so hat man den *Prozentrang*. Er ist quasi die Antwort auf die Frage, wieviel Prozent der Werte bis zu diesem Wert vorkommen (wieviel Prozent der Stichprobenmitglieder haben Schuhgrößen bis 39?)

Der Wert der Approximativen Verteilungsfunktion errechnet sich aus:

$$F^*(x) = F_{e-1} + \frac{x_e - x_e}{\tilde{x}_e - x_e} \cdot \overbrace{\left(F_e - F_{e-1}\right)}^{f_e} \text{ (e steht für Einfallsklasse)}$$

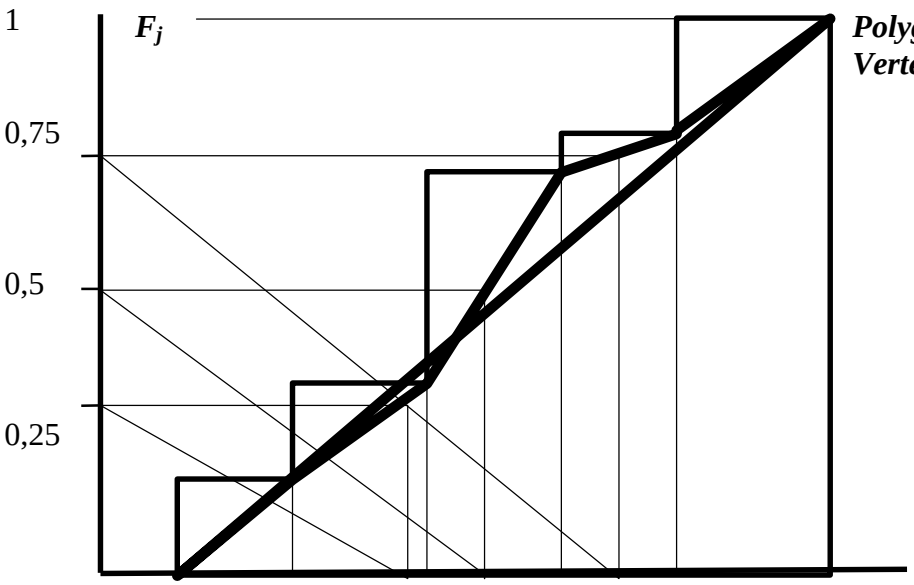
Der Prozentrang PR (x) ist 100 mal F*.

Die Fragestellung in der anderen Richtung würde lauten, unterhalb bzw. bis zu welchem Wert x z.B. 35% der Werte liegen, oder mathematischer, für welchen x-Werte die Approximative Verteilungsfunktion den Wert 0,35 liefert ($F^* = p_x$).

Um das zu ermitteln muß man nur die Approximative Verteilungsfunktion ein wenig umformen und erhält so:

$$x_p = x_e + \frac{p_x - F_{e-1}}{f_e}(\tilde{x}_e - x_e)$$

Der Wert der Variablen x, für den ein gegebener Wert der Verteilungsfunktion p_x zutrifft, heißt *Quantil*. Will man z.B. die Körpergröße wissen, bis zu welcher 37% der Stichprobenmitglieder liegen, sucht man das Quantil für $p_x = 0{,}37$. Die speziellen Quantile für p = 0,25; p = 0.5 und p = 0,75 heißen *1., 2., und 3. Quartil*. Der *Median* — auch ein Maß für zentrale Tendenz wie der Mittelwert — ist das Quantil für p = 0.5 bzw. das 2. Quartil. Die *Perzentile* sind die auf den Prozentrang bezogenen Quantile: Das 30. Perzentil ist also der Wert von x, der den Prozentrang 30% hat.



Polygonzug der Approximativen Verteilungsfunktion (F*)

1. Quartil Median 3. Quartil = Quantil für p = 0,75 = 75. Percentil

STREUUNGSMAßE (SPANNWEITE, VARIANZ, STANDARDABWEICHUNG)

Die Maßzahlen zentraler Tendenz (Mittelwert, Modalwert = häufigster Wert, und Median) liefern keine Information über die Dispersion oder Streuung der Werte in einer Stichprobe. Im Extremfall kann es so z.B. sein, daß sich für einen Schüler mit gleichviel 1er und 3er ein Notendurchschnitt (arithmetisches Mittel) von 2,0 errechnet , obwohl er in keinem Fach die Note 2 hat; also kann es auch wichtig sein, die Streuung von Werten zu kennen. Das einfachste Streuungsmaß ist die Differenz zwischen dem größten und dem kleinsten Wert, die *Spannweite*. Doch sie sagt nichts darüber aus, wie weit die Werte *durchschnittlich* um einen Mittelwert streuen. Diese Information liefern uns die Varianz und deren Wurzel, die Standardabweichung.

Da der Mittelwert tatsächlich die Mitte der Werte bildet, ist die Summe der Differenzen zu ihm gleich Null. So kann man für ein sinnvolles Maß der Streuung um den Mittelwert ganz einfach die Differenzen der Werte zum Mittelwert quadrieren, aufsummieren und durch die Anzahl an Werten teilen; man erhält so die Varianz:

$$s^2 = \frac{1}{N}\sum_{i=1}^{N}\left(x_i - \bar{x}\right)^2 \Rightarrow \text{Abweichungswerteformel}$$

oder durch Umformung:

$$\frac{1}{N}\sum\left(x^2 - 2x\bar{x} + \bar{x}^2\right) = \frac{1}{N}\sum x^2 - 2\bar{x}\underbrace{\frac{1}{N}\sum x}_{\bar{x}} + \underbrace{\frac{1}{N}\sum \bar{x}^2}_{\bar{x}^2} = \frac{1}{N}\sum x^2 - 2\bar{x}^2 + \bar{x}^2 =$$

$$s^2 = \frac{1}{N}\sum_{i=1}^{N} x_i^2 - \bar{x}^2 \Rightarrow \text{Rohwerteformel}$$

Die Quadratwurzel aus der Varianz ist die Standardabweichung:

Sandardabweichung $s = \sqrt{s^2}$; sie stellt praktisch den mittleren Abstand der Werte von ihrem Mittelwert dar.

i	x_i	x_i^2	$(x_i - \bar{x})^2$	$\bar{x}$	$\bar{x}^2$
1	40			40	
2	39				
3	40				
4	39				
5	37				
6	41				
7	42				
8	45				
9	40				
10	37				

LINEARE TRANSFORMATION

Eine Transformation der Werte um $y_i = ax_i + b$ nennt man *lineare Transformation*; wie sich bei einer solchen Transformation der Mittelwert, die Varianz und die Standardabweichung ändert, können wir leicht ableiten:

$$\bar{y} = \frac{1}{N}\sum_{i=1}^{N}(ax_i + b) = a \cdot \underbrace{\frac{1}{N}\sum_{i=1}^{N} x_i}_{\bar{x}} + \underbrace{\frac{1}{N}\sum_{i=1}^{N} b}_{b} = a\bar{x} + b$$ ⇨ der Mittelwert verschiebt sich also wie alle anderen Punkte um den Faktor a und den Summanden b.

Bei der Varianz wirkt sich diese Transformation folgendermaßen aus:

$$s_y^2 = \frac{1}{N}\sum_i\left[(ax_i + b) - (a\bar{x} + b)\right]^2 = \frac{1}{N}\sum_i\left[a\left(x_i - \bar{x} + \frac{b}{a} - \frac{b}{a}\right)\right]^2 = \frac{1}{N}\sum_i a^2(x_i - \bar{x})^2 = a^2\frac{1}{N}\sum_i(x_i - \bar{x})^2 = a^2 \cdot s_x^2$$

Und schließlich bedeutet das für die Standardabweichung eine Transformation um:

$$s_y = |a| \cdot s_x$$

Voraussetzung für die Berechnung von arithmetischem Mittel, Varianz und Standardabweichung sind Daten auf Intervallskalenniveau (siehe eigenes Paper zu den Skalenniveaus). Den Modalwert und die Quartile (z.B. Median) kann man bereits auf Ordinalskalenniveau bestimmen.

DIE SKALENNIVEAUS VON DATEN

Nominalskala

Ausprägungen von Eigenschaften in willkürlicher Reihenfolge bei *qualitativen Variablen*
⇨ Gleichheit oder Ungleichheit von Objekten ist gegeben
Bsp.: Beruf, Geschlecht, Augenfarbe
nicht bestimmbar: arithmetisches Mittel und Median, Varianz und Standardabweichung
bestimmbar: Modalwert

Rang- oder Ordinalskala

Ausprägungen haben eine eindeutige Rangfolge
Elemente sind nach Größe bzw. Intensität zu ordnen
⇨ Rangfolge, Größer-Kleiner-Relation
Frage: Ist das Merkmal unterschiedlich ausgeprägt?
Bsp.: Schulnoten, Reihenfolge der Ankunft von Läufern im Ziel
nicht berechenbar: arithm. Mittel, Varianz und Standardabweichung
bestimmbar: Modalwert und Median

Intervallskala

Die Abstände oder Intervalle zwischen Zahlen sind eindeutig festgelegt; Gleichheit von Intervallen
Frage: Um wieviel mehr ist das Merkmal ausgeprägt?
Bsp: Dioptrien, Celsius (20°C ist energetisch nicht doppelt soviel wie 10°C; doch eine Erwärmung von 10°C auf 20°C bzw. 0°C auf 10°C benötigt jeweils den gleichen Energiebetrag)
berechenbar: arithm. Mittel, Median, Modalwert, Varianz und Standardabweichung

Rationalskala

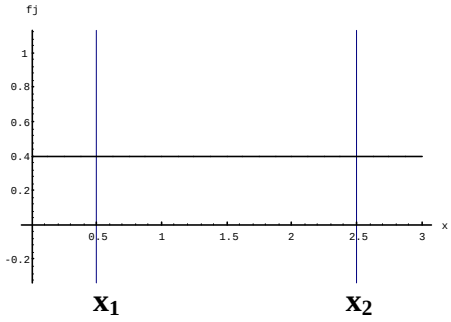
Gleichheit von Verhältnissen
Bsp.: Temparatur in Kelvin, Gewichte, Längenmaße
20 cm ist doppelt so lang wie 10 cm und vierfach so lang wie 5 cm; 20 °K sind vom Energiebetrag doppelt so groß wie 10°K
berechenbar: arithm. Mittel, Median, Modalwert, Varianz, Standardabweichung und absoluter Nullpunkt

In der Psychologie sind die meisten Daten nur auf Intervallskalenniveau zu erheben (es ist z.B. schwer zu beantworten, wo der absolute Nullpunkt der Intelligenz anzusiedeln ist)

VERTEILUNGSFORMEN

Gleichverteilung (Rechteckverteilung)

Die beobachteten Meßwerte verteilen sich in einem bestimmten Intervall völlig gleichmäßig (alle Werte kommen gleich häufig vor).



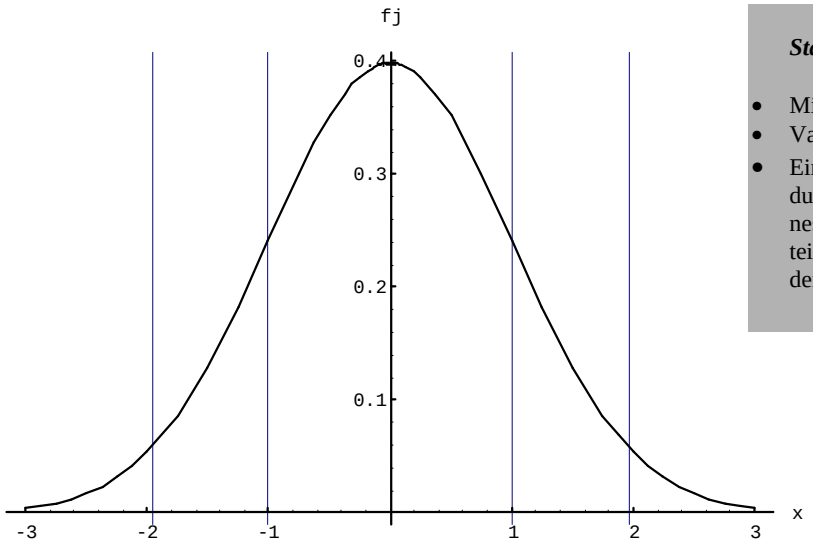
Normalverteilung (GAUSS'sche Verteilung)

Sie wird bestimmt durch 2 Parameter:

- *Varianz* (bzw. Standardabweichung)
- *Mittelwert* (bzw. „Erwartungswert" bei einer Wahrscheinlichkeitsverteilung)

Empirische Verteilungen, die dem Modell der Normalverteilung entsprechen, haben folgende Eigenschaften:

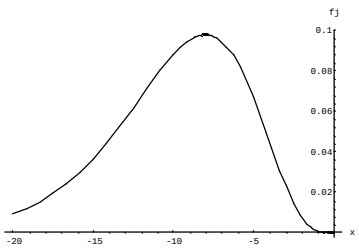
- Im Bereich von einer Standardabweichung unter bis einer Stdabw. über dem Mittelwert $(\bar{x} - s \text{ bis } \bar{x} + s)$ liegen ca. 68% der Werte von x.
- Im Bereich von ca. zwei Standardabweichungen um den Mittelwert liegen 95% der Werte von x.$(\bar{x} - 1{,}96 \cdot s \text{ bis } \bar{x} + 1{,}96 \cdot s)$
- Jenseits der zwei Standardabweichungen unter bzw. über dem Mittelwert liegen je 2,5% der Werte von x.



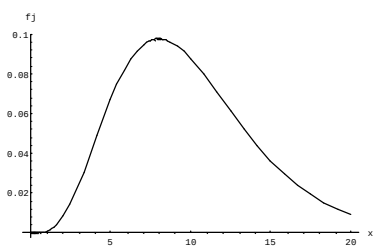
$\mu - 3\sigma$ $\mu - 1{,}96\sigma$ $\mu - \sigma$ μ $\mu + \sigma$ $\mu + 1{,}96\sigma$ $\mu + 3\sigma$

Sonderfall:
Standardnormalverteilung

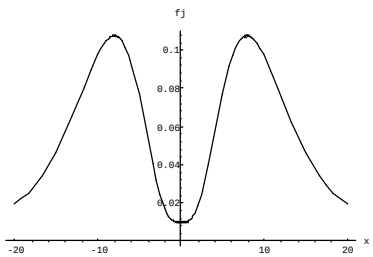
- Mittelwert = 0
- Varianz = Standardabweichung = 1
- Eine normalverteilte Variable x kann durch Standardisierung (siehe eigenes Blatt) in eine standardnormalverteilte Variable z transformiert werden.



Links-schiefe (rechts-steile) Verteilung



Rechts-schiefe (links-steile) Verteilung



Bimodale Verteilung

STANDARDISIERUNG

Eine Standardisierung stellt *im weitesten Sinne* eine Transformation dar, die eine Vereinheitlichung und somit eine bessere Vergleichbarkeit von Daten — z.B. aus verschiedenen Datensätzen oder Stichproben — zur Folge hat. Die Transformation von x-Werten auf Prozentränge wäre eine solche Transformation, die für jede Variable Werte von 0 bis 100 liefert, egal in welcher Einheit die Daten gemessen wurden.

Im *engeren Sinn* versteht man darunter die sog. *z-Transformation*, eine lineare Transformation, die die ursprünglichen x-Werte auf z-Werte so abgebildet, daß sie den Mittelwert 0 und die Standardabweichung und Varianz von 1 haben. Eine solche Transformation führt bei einer normalverteilten Variable zu einer standardnormalverteilten Variable.

Die Formel, welche die x-Werte in z-Werte „übersetzt“, sieht folgendermaßen aus:

$$z_i = \frac{x_i - \overline{x}}{s_x}$$

Von jedem Wert wird also der Mittelwert abgezogen und anschließend Verbleibendes durch die Standardabweichung geteilt. Da der Mittelwert von z-Werten immer 0 ist, weiß man sofort, daß negative Werte unterdurchschnittlich, und positive Werte überduchschnittlich sind. Da man ferner die exakten Wahrscheinlichkeiten für eine Standardnormalverteilung kennt, kann man — wenn die Ausgangsvariable zumindest annähernd normalverteilt ist — Rückschlüsse auf die Auftretenswahrscheinlichkeit von Werten in bestimmten Bereichen machen (siehe auch bei Verteilungsformen).

- Im Bereich von μ -1 bis μ +1 liegen 68% der Werte
- im Bereich von μ - 1,96 bis μ + 1,96 liegen 95% der Werte
- im Bereich unterhalb μ - 1,96 und oberhalb μ + 1,96 liegen je 2,5% der Werte.

(Der Parameter μ aus einer Wahrscheinlichkeitsverteilung entspricht dem Mittelwert in einer empirischen Häufigkeitsverteilung)

Es gibt in der Psychologie noch einige andere gebräuchliche Skalen neben den z-Werten:

IQ-Skala

$IQ_i = 15 \cdot z_i + 100$
⇨ Eigenschaften: Mittelwert $\overline{IQ} = 100$; Standardabweichung $s_{IQ} = 15$

T-Skala

$T_i = 10 \cdot z_i + 50$
⇨ Eigenschaften: $\overline{T} = 50$; $s_T = 10$

Stanines (STAndard NINES)

$SN_i = 1{,}5 \cdot z_i + 5$
⇨ Eigenschaften: $\overline{SN} = 5$; $s_{SN} = 1{,}5$

ÜBUNGSAUFGABEN ZUR DESKRIPTIVEN STATISTIK

1. Welche der folgenden Maßzahlen sind Streuungsmaßzahlen?

1. Spannweite
2. Varianz
3. Interquartilsabstand
4. Standardabweichung
5. Modalwert

Richtig sind:
a) 1, 2, 4, 5 b) alle c) 1, 2, 4 d) 2, 4 e) 1, 2, 3, 4

2. In einem Sportwettkampf erhält eine Teilnehmerin von den Kampfrichtern folgende Wertungen:

8,5 / 8,7 / 8,4 / 8,9 / 8,6 / 9,0 / 8,6
Berechnen Sie arithmetisches Mittel und Median. Als Ergebnis erhalten Sie:
a) arithm. Mittel = 8,67 Median = 8,9
b) arithm. Mittel = 8,67 Median = 8,6
c) arithm. Mittel = 8,70 Median = 8,9
d) arithm. Mittel = 8,65 Median = 8,7
e) arithm. Mittel = 8,80 Median = 8,7

3. Der Schuhverkäufer Herr K. ist Hobbystatistiker und stellt jede Woche heimlich eine Häufigkeitstabelle über die verkauften Paar Schuhe auf, um seinen Chef bei der Bestellung hinsichtlich der verschiedenen Größen besser beraten zu können.

Schuhgröße	**38**	**39**	**40**	**41**	**42**	**43**	**44**	**45**
Häufigkeit	2	10	0	7	13	15	5	9

Wo liegt der Median dieser Verteilung?
a) 42 b) 8,71 c) 13 d) 15 e) 41

4. Was haben Median und Modalwert gemeinsam?

a) Beide sind immer positiv.
b) Beide sind Streuungsmaße.
c) Sie sind beide nicht immer eindeutig bestimmt.
d) Beide können nur für standardisierte Werte berechnet werden.
e) Eines von beiden ist die Wurzel aus dem anderen.

5. Nach vielen Jahren nehmen Sie wieder an den Bundesjugendweltspielen teil. Doch heute laufen Sie nicht 50m, sondern Sie werten die Ergebnisse statistisch aus.
Die Häufigkeit der Sprintergebnisse über 50m sind:

Laufergebnisse	Anzahl
15 - 14 s	3
14 - 13 s	3
13 - 12 s	3
12 - 11 s	1

Der ungefähre Wert des 2. Quartils liegt bei:
a) 12,00 s b) 12,50 s c) 13,33 s d) 13, 67 s e) 12,75 s

6. Welche Aussagen über das arithmetische Mittel sind richtig?

1. Die Berechnung ist nur sinnvoll für Daten, die mindestens Intervallskalenqualität haben.
2. Das arithmetische Mittel ist im Gegensatz zum Median unempfindlich gegen Ausreißer.
3. Wird zu allen x-Werten einer Stichprobe dieselbe Zahl addiert, so vergrößert sich aus das arithmetische Mittel um diese Zahl.
4. Die Summe der Abweichungen aller Werte von ihrem arithmetischen Mittel ist Null.
5. Das arithmetische Mittel kann nicht berechnet werden für bimodale Verteilungen.

Richtig sind:
a) 3, 4 b) 1, 3, 4 c) alle d) 1, 4 e) 1, 2, 5

7. Welche der folgenden Kennwerte sind Maßzahlen der zentralen Tendenz?

1. der Median
2. das 50. Perzentil
3. das arithmetische Mittel
4. der Modalwert
5. das 2. Quartil

Richtig ist/sind:
a) 1, 2, 3, 4 b) alle c) 1, 3 d) 1, 3, 4 e) 1, 2

8. Die angemessene graphische Darstellung der empirischen relativen Häufigkeit für eine klassifizierte Variable ist ein(e)

a) Stabdiagramm
b) Normalverteilung
c) von links nach rechts ansteigende Treppenkurve
d) Histogramm
e) Polygonzug

9. z-standardisierte Meßwerte

a) haben die Varianz 1
b) können nur zwischen 0 und 1 liegen
c) haben den Mittelwert 1
d) werden in der Tabelle der Standardnormalverteilung abgelesen
e) liegen zwischen -1 und +1

10. Welche der Aussagen über die Normalverteilung sind <u>falsch</u>?

a) Die Normalverteilung ist unimodal.
b) Der Erwartungswert, der Median und der Modalwert sind gleich.
c) Die Normalverteilung ist durch zwei Parameter μ und σ bestimmt.
d) Die Wahrscheinlichkeit dafür, daß ein Wert größer als μ auftritt, ist ½.
e) Die Wahrscheinlichkeit dafür, daß ein Wert zwischen μ und $1{,}96\sigma$ auftritt, ist 0.95.

11. Welche Aussage über eine normalverteilte Zufallsvariable ist richtig?

a) Große Werte, ungefähr ab dem Wert 4, sind extrem wenig wahrscheinlich.
b) Sie ist ein gutes Modell der Funktion von Glocken.
c) Der Graph der Dichtefunktion fällt vom höchsten Punkt aus nach beiden Seiten ab, bis er die x-Achse schneidet.
d) Ihre Varianz ist ihrer Standardabweichung gleich.
e) Der Wertebereich darf weder nach oben noch nach unten begrenzt sein.

(Aufgaben des WS 94/95; Lösungen: 1e, 2b, 3a, 4c, 5c, 6b, 7b, 8d, ,9a, 10e, 11e)

1. Die deskriptive Statistik beschreibt

a) Zusammenhänge zwischen Stichprobe und Population
b) die Art der Stichprobenbeziehung
c) die Art des Signifikanztests
d) das Ergebnis des Signifikanztests
e) die Verhältnisse in der Stichprobe.

2. Eine Zufallsstichprobe liegt vor, wenn

a) die Stichprobe normalverteilt ist.
b) die Stichprobe aus der zugehörigen Grundgesamtheit stammt.
c) jedes Element der Grundgesamtheit die gleiche Wahrscheinlichkeit hatte, in die Stichprobe aufgenommen zu werden.
d) die Stichprobe genau die Grundgesamtheit repräsentiert.
e) der Tag der Stichprobenziehung zufällig ausgewählt wurde.

3. Die kumulierte prozentuale Häufigkeit

1. kann aus den relativen Häufigkeiten berechnet werden
2. kann aus den absoluten Häufigkeiten und dem Stichprobenumfang berechnet werden
3. ist immer so groß wie die kumulierte absolute Häufigkeit
4. ist immer so groß wie die kumulierte relative Häufigkeit
5. ist größer oder gleich 0 und kleiner oder gleich 100.

Richtig ist/sind:
a) 1, 2 b) 1, 2, 4 c) 1, 2, 5 d) 1, 4 e) alle Aussagen.

Zu 4./5.: Sie befrage Psychologiestudenten/innen, wie häufig sie trotz des schneearmen Winters beim Skifahren waren. Sie erhalten folgende Daten:

Vp	1	2	3	4	5
Skitage	0	2	1	1	5

4. Berechnen Sie Mittelwert $\overline{x}$ und Varianz s^2 dieser Wertereihe

a) $\overline{x} = 2,5$ $s^2 = 5,2$
b) $\overline{x} = 1,8$ $s^2 = 6,8$
c) $\overline{x} = 1,8$ $s^2 = 2,96$
d) $\overline{x} = 9$ $s^2 = 0$
e) $\overline{x} = 1,8$ $s^2 = 1$

5. Wie groß ist die relative Häufigkeit der Personen der Stichprobe, die höchstens 2 mal beim Skifahren waren?

a) .8
b) .44
c) .6
d) 1
e) .4

6. Ein Stabdiagramm

1. kann die graphische Darstellung der Häufigkeiten einer Variable sein
2. wird nur bei diskreten Variablen angewendet
3. wird nur bei relativen Häufigkeiten angewendet
4. wird auch Säulendiagramm genannt
5. wird bei allen Variablen angewendet

Richtig ist/sind:

a) 1, 3, 5 b) 1, 2, 3 c) 1, 2, 3, 4 d) 1, 5 e) 1, 2

7. Welche Maßzahlen sind Maße zentraler Tendenz?

1. Arithmetisches Mittel
2. Spannweite
3. Median
4. Standardabweichung
5. Interquartilsabstand

Richtig ist/sind:

a) 2, 4 b) 2, 4, 5 c) 1, 3 d) 1, 5 e) 4, 5

8. Sie sind der einzige Nichtraucher in einer 8-Personen-Wohngemeinschaft und wollen ihre Mitbewohner von der Schädlichkeit dieses Lasters überzeugen. Deshalb notieren Sie eines Abends die Anzahl der gerauchten Zigaretten für jeden Mitbewohner. Es ergibt sich folgende Tabelle:

Peter	10
Siggi8	
Gabi	5
Alex	25
Helmut	14
Birgit	12
Anette	3

Wieviele Zigaretten wurden durchschnittlich pro Kopf geraucht?

a) 8 b) 9 c) 10 d) 11 e) 12

9. Was ist der Median des Zigarettenkonsums? (Angaben siehe vorherige Aufgabe)

a) 8
b) 10
c) 11
d) 12
e) 14

Sie sind Mitglied der Jury bei einem Hochsee-Wettfischen und als einziges Mitglied statistisch ausgebildet. Deshalb übernehmen Sie die Aufgabe, aus der Häufigkeitstabelle zu den Längenklassen der gefangenen Haie:

Hai-Längenklasse	Anzahl
150 - 160 cm	2
160 - 170 cm	4
170 - 180 cm	3
180 - 190 cm	1

den ungefähren Wert des 2. Quartils zu bestimmen

a) 167 cm
b) 163 cm
c) 165 cm
d) 170 cm
e) 160 cm

11. Sie führen bei Schulkindern einer Altersstufe einen Schulreifetest durch, der für jeden Schüler zu einen Schulreifewert SRW führt. Ein Schüler hat einen SRW mit einem Prozentrang von 10. Was bedeutet das?

a) Sein SRW ist das 0.1-Quantil
b) Sein SRW ist das 1. Perzentil
c) Etwa 10% der Schüler haben einen höheren SRW
d) Etwa 10% der Schüler haben einen niedrigeren SRW als die Lehrer
e) Der Schüler ist so klug wie 10% der Lehrer

(Aufgaben des WS 92/93; Lösungen: 1e, 2c, 3c, 4c, 5a, 6e, 7c, 8d, 9b, 10a, 11a)

WAHRSCHEINLICHKEITSRECHNUNG

Zufallsexperiment

- Die prinzipiell möglichen Ergebnisse sind bekannt
- Welches Ergebnis vorkommt, ist nicht vorherzusagen
- Die relativen Häufigkeiten für die verschiedenen Ergebnisse stabilisieren sich auf lange Sicht

Ergebnis

Das, was als Resultat bei einem Zufallsexperiment rauskommen kann; mathematisch durch ω (kleines „omega“) repräsentiert.

Ergebnisraum

Menge aller möglichen Ergebnisse; Symbol Ω. Ein Ergebnisraum kann endlich (bei diskreter Zufallsvariable) oder unendlich (bei kontinuierlicher Zufallsvariable) sein.

$$\Omega = \{\omega_1, \omega_2, ..., \omega_n\}$$

Beispiele: Münzwurf Ω ={Zahl, Wappen}, Würfel Ω = {1, 2, 3, 4, 5, 6}

Ereignisse E_i

Mögliche Kombinationen von Ergebnissen; Teilmengen des Ergebnisraums Ω, inklusive der leeren Menge {} (unmögliches Ereignis), aller Elementarereignisse (enthält nur ein Ergebnis $\{\omega_i\}$) und des ganzen Ergebnisraums Ω (sicheres Ereignis); Abkürzung durch Großbuchstaben (A, B, C, ..., Z).

Ereignisse sind *disjunkt*, wenn ihre Schnittmenge leer ist: $A \cap B = \varnothing \Leftrightarrow$ A, B disjunkt; sind zwei Ereignisse disjunkt, gilt, daß die Wahrscheinlichkeit dafür, daß A oder B eintritt, die Summe der Wahrscheinlichkeiten für A und B ist:

$$P(A \cup B) = P(A) + P(B)$$

Ereignisse sind dann komplementär in Bezug auf Ω, wenn ihre Schnittmenge leer und ihre Vereinigungsmenge gleich Ω ist: $A \cap B = \varnothing$, $A \cup B = \Omega \Leftrightarrow$ A, B komplementär; $B = \overline{A}$ (Komplement bezüglich Ω). Beispiel Würfel: A = {1, 3, 5}, $\overline{A}$ = {2, 4, 6}

Zwei Ereignisse heißen dann *unabhängig*, wenn es für das Eintreten von A egal ist, ob das Ereignis B vorher eingetreten ist oder nicht; etwas mathematischer gilt dann:

$$P(A \cap B) = P(A) \cdot P(B);$$ (P ist die Wahrscheinlichkeit, engl. „probability“)

Umgekehrt kann man sagen: Gilt diese Beziehung, sind zwei Ereignisse unabhängig. Demnach sind zwei Ereignisse dann *abhängig*, wenn gilt:

$$P(A \cap B) \neq P(A) \cdot P(B)$$

Ereignisraum von Ω

Menge aller Ereignisse

Potenzmenge Π

ist der Ereignisraum von einem endlichen Ergebnisraum;
Beispiel Münzwurf: $\Pi(\Omega) = \{\{Z, W\}, \{Z\}, \{W\}, \{\}\}$

Mächtigkeit des Ergebnisraums $|\Omega| = n \Rightarrow$ Mächtigkeit der Potenzmenge $|\Pi(\Omega)| = 2^n$

KOLMOGOROWsche Axiome

I. $P(A) \geq 0$
II. $P(\Omega) = 1$
III. $A \cap B = \varnothing \Rightarrow P(A \cup B) = P(A) + P(B)$

Satz von SYLVESTER: A, B beliebig (nicht unbedingt disjunkt);
$P(A \cup B) = P(A) + P(B) - P(A \cap B)$

Das BERNOULLI-Experiment

Ein BERNOULLI-Experiment ist ein Zufallsexperiment mit genau zwei möglichen Ergebnissen, die meist als Treffer und Niete bezeichnet werden.

Eine Kette von n unabhängigen BERNOULLIexperimenten, bei denen die Wahrscheinlichkeit für jeden Treffer konstant p ist, heißt BERNOULLIkette der Länge n mit dem Parameter p.

Bei einer BERNOULLIkette der Länge n mit dem Parameter p gilt:

$$w(y) = P\left(\underbrace{Y}_{\substack{\text{Zufalls-}\\\text{variable}}} = \underbrace{y}_{\substack{\text{Anzahl}\\\text{der "Treffer"}}}\right) = \binom{n}{y} p^y \cdot q^{n-y} \text{ mit } q = 1 - p$$

Die kumulative Wahrscheinlichkeit F(y), daß das Ergebnis Y bis zu y-mal bei n Wiederholungen vorkommt ist:

$$F(y) = P(Y \leq y) = \sum_{k=0}^{y} \binom{n}{k} p^k q^{n-k}$$

Der sog. „Binomialkoeffizient“ $\binom{n}{y}$ läßt sich auf folgende Weisen bestimmen:

1. Mit der Fakultätenformel:

$$\binom{n}{y} = \frac{n!}{y!(n-y)!}; \text{ wobei } n! = 1 \cdot 2 \cdot 3 \cdot \ldots \cdot n \text{ und } 0! := 1$$

2. Mittels additiver Rekursion im PASCALschen Dreieck:

n = 0						1							
1					1		1						
2				1		2		1					
3			1		3		3		1				
4		1		4		6		4		1			
5	1		5		10		10		5		1		
6	1	6		15		20		15		6		1	
	$\binom{6}{0}$	$\binom{6}{1}$		$\binom{6}{2}$		$\binom{6}{3}$		$\binom{6}{4}$		$\binom{6}{5}$		$\binom{6}{6}$	
...						...							

$$\Rightarrow \quad \binom{n}{y} = \binom{n}{n-y}$$

3. Multiplikative Rekursion

$$(a + b)^2 = \underline{1}a^2 + \underline{2}ab + \underline{1}b^2$$
$$(a + b)^5 = \underline{1}a^5 + \underline{5}a^4b + \underline{10}a^3b^2 + \underline{10}a^2b^3 + \underline{5}ab^4 + \underline{1}b^5$$

⇨ Koeffizienten sind wie im PASCALschen Dreieck!

$$\underbrace{(p+q)^n}_{1} = \sum_{y=0}^{n} \binom{n}{y} p^y q^{n-y} = 1 \text{ (Allgemeines Binom)}$$

⇨ Wahrscheinlichkeiten bei einer BERNOULLIkette: $p = 1- q \Leftrightarrow p + q = 1$

Das GALTONsche Nagelbrett

Das GALTONsche Nagelbrett stellt ein Experiment in Form einer BERNOULLIkette mit n Wiederholungen dar, in dem durch die mechanische Anordnung von Nägeln für passierende Kugeln immer die gleiche Wahrscheinlichkeit von p = 0,5 besteht, links bzw. rechts am Nagel vorbeizulaufen. Die Wahrscheinlichkeiten bzw. relativen Häufigkeiten für die einzelnen Ereignisse P(Y = y) bilden sich so mit der Zeit in Auffangrinnen — quasi graphisch — ab.

Beispiel:

Wie groß ist die Wahrscheinlichkeit dafür, daß eine Kugel in einem GALTONschen Nagelbrett mit 6 Reihen a) genau 4mal nach rechts läuft bzw. b) bis zu 4mal nach rechts läuft?

ZUR WIEDERHOLUNG

Was ist

9 über 0? _____

7 über 4? _____

4 über 8? _____

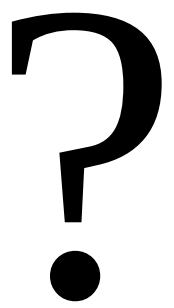
72 über 71? _____

8 über 8? _____

6 über 2? _____

12 über 1? _____

7 über 5? _____



Wie groß ist die Wahrscheinlichkeit, daß ein Marmeladenbrot bei 10 Versuchen 7 mal auf das Gesicht (Marmeladenseite) fällt, wenn p für *„Brot fällt auf's Gesicht“* = 0,6?

Bis zu welcher Anzahl an unglücklichen Fällen auf die Marmeladenseite besteht eine (Gesamt-) Wahrscheinlichkeit (F) kleiner gleich 70%?

Eine (Gesamt-) Wahrscheinlichkeit kleiner gleich 5% besteht ab einem bestimmten, und alle noch größeren y? Welches y ist das?

Hilfstabelle für p = 0.6:

y	n-y	$\binom{n}{y}$	w(y)	F(y)
0	10	1	0.00010	0.00010
1	9	10	0.00157	0.00168
2	8	45	0.01062	0.01229
3	7	120	0.04247	0.05476
4	6	210	0.11148	0.16624
5	5	252	0.20066	0.36690
6	4	210	0.25082	0.61772
7	3	120	0.21499	0.83271
8	2	45	0.12093	0.95364
9	1	10	0.04031	0.99395
10	0	1	0.00605	1.00000

DER SIGNIFIKANZTEST (NACH FISCHER)

1. Oberhypothese(n):

BERNOULLI-Kette

2. 0-Hypothese:

H_0: $p = p_0$ = z.B. 0,5 (ist bei 2 Ergebnissen am naheliegendsten)
(H_1: $p \neq p_0$, d.h. alle von p_0 verschiedenen Werte)

3. Stichprobenumfang n festlegen

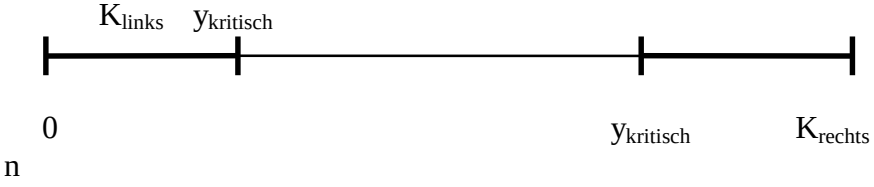
4. Signifikanznivieau α festlegen

α ist per Konvention 0,05 (5%), 0,01 (1%) bzw. 0.001 (1‰)
Alpha wird auch als *Irrtumswahrscheinlichkeit* bezeichnet

5. Kritischen Bereich K bestimmen

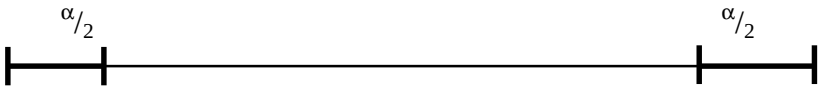
Kritischer Bereich K: $P(Y \in K) \leq \alpha$ unter H_0

a) Einseitiger Kritischer Bereich



K-links:[0, 1, ..., y_{krit}] bzw.
K-rechts [y_{krit},, y_{krit}+1, ..., n-1, n]

b) Zweiseitiger Kritischer Bereich



6. Durchführung des Zufallsexperiments

7. Entscheidung: Falls $Y \in K$: H_0 verworfen, falls nicht, H_0 beibehalten

ÜBUNGSBEISPIEL

Man entnimmt einem Pool von Personen eine Stichprobe von 20 und überprüft die Nullhypothese, die besagt, daß in der zugrunde liegenden Population gleichviele Männer wie Frauen waren. Es sind 14 Frauen und 6 Männer mit der Stichprobe gezogen worden.

Oberhypothese?
Nullhypothese?
Stichprobenumfang n?
Signifikanzniveau (5%, 1% oder 1‰)?
Kritischer Bereich? (ein- oder zweiseitig)?
Realisation?
Entscheidung?

Ausschnitt aus der Hilfstabelle für n = 20 und p = .5:

y	F(y)
0	0,00000
1	0,00002
2	0,00020
3	0,00129
4	0,00591
5	0,02069
6	0,05766
7	0,13159
8	0,25172
9	0,41190
10	0,58810
11	0,74820
12	0,86841
13	0,94234
14	0,97931
15	0,99409
16	0,99871
17	0,99980
18	0,99998
19	1,00000
20	1,00000

DER SIGNIFIKANZTEST (NACH NEYMAN-PEARSON)

Während man sich beim Signifikanztest der FISCHERschen Variante nur darauf festlegte, welchen Wert p unter der Nullhypothese H_0 hat, und p unter der zusammengesetzten Alternativhypothese H_1 nicht über „$p_1 \neq p_0$“ hinaus spezifiziert wurde, legt man sich bei der NEYMAN-PEARSONschen Testvariante auf einen *bestimmten Wert* p unter der *spezifischen Alternativhypothese* H_1’ fest, z.B.: Nullhypothese: H_0: p = 0,5; Alternativhypothese: H_1’: p = 0,75

Es kann nun wieder ein Kritischer Bereich (K) angegeben werden, für den gilt:

$$P(Y \in K|H_0) \leq \alpha$$

Kam beim Fischerschen Signifikanztest ein Y heraus, das ein Element des Kritischen Bereichs war, lehnte man H_0 ab und behielt H_1 bei, wobei natürlich die Möglichkeit mit einer Wahrscheinlichkeit von maximal α (5%, 1%, 1‰) bestand, irrtümlicherweise die H_0 abgelehnt zu haben (deshalb auch *Irrtumswahrscheinlichkeit*), da ein solches Y ja — wenn auch sehr unwahrscheinlich — prinzipiell auch unter H_0 möglich gewesen wäre. Diesen Fehler nennt man *Fehler 1. Art oder α-Fehler*, das Risiko (Wahrscheinlichkeit), ihn zu begehen *α-Risiko*. Bei der NEYMAN-PEARSON-Testvariante kann man nun ferner bestimmen, wie groß die Wahrscheinlichkeit dafür ist, daß man Y-Werte aus diesem Kritischen Bereich unter der Alternativhypothese H_1’ erhält — mit anderen Worten — man unter einer zutreffenden Alternativhypothese auch tatsächlich die Nullhypothese verwerfen kann.

Diese Wahrscheinlichkeit ist:

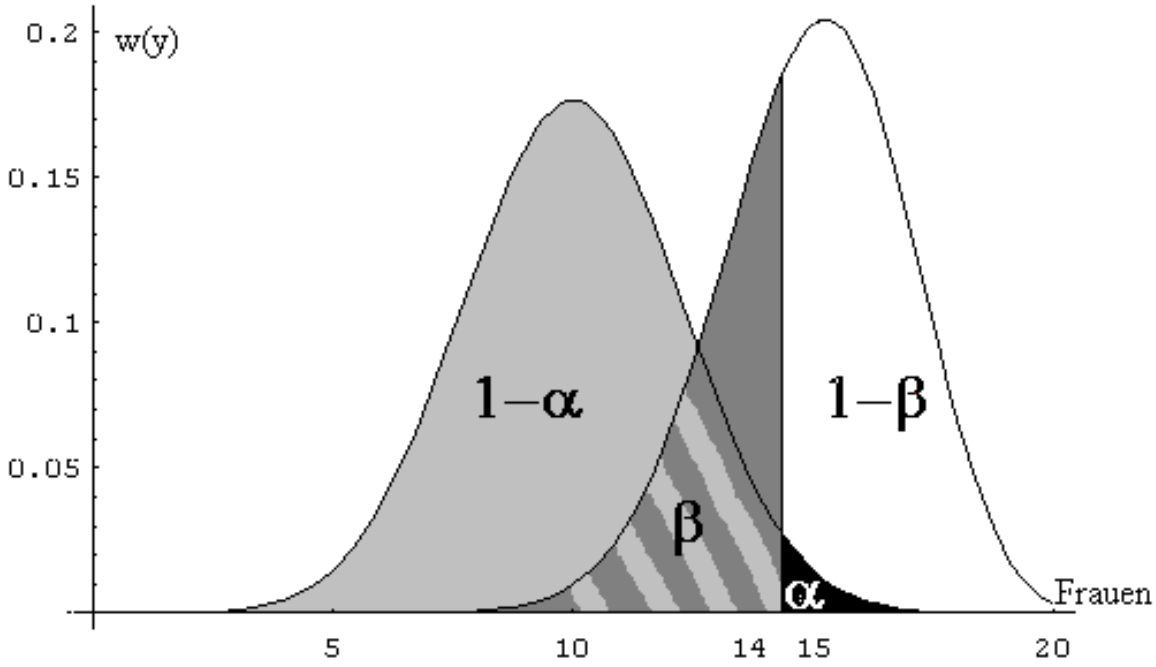
$$P(Y \in K|H_1') = 1 - \beta \text{; sie wird auch } \textit{Teststärke} \text{ genannt.}$$

Demnach ist β die Wahrscheinlichkeit dafür, daß Y unter H_1’ ein Element aus dem Annahmebereich der H_0 ist ($\overline{K}$), man also bei zutreffender H_1’ trotzdem die H_0 beibehalten würde und so ebenfalls einen Fehler beginge:

$P(Y \in \overline{K}|H_1') = \beta$; sie wird *$\beta$-Risiko* genannt; der Fehler heißt *β-Fehler oder Fehler 2. Art*

Zusammenfassend können sich also folgende Fälle ergeben:

man entscheidet sich für:	**zutreffend ist:**	
	H_0	**H_1’**
H_0	**kein Fehler** Wahrscheinlichkeit $\geq 1 - \alpha$	**Fehler 2. Art (β-Fehler)** Wahrscheinlichkeit β (β-Risiko)
H_1’	**Fehler 1. Art (α-Fehler)** Wahrscheinlichkeit (Irrtumswahrscheinlichkeit) $\leq \alpha$ (Signifikanzniveau) 5%, 1%, 1‰ (Alpha-Risiko)	**Teststärke** Wahrscheinlichkeit 1- β



ÜBUNGSBEISPIEL (TEIL 2)

Ausschnitt aus der Hilfstabelle für n = 20 und p = .75:

y	F(y)
0	0.00000
1	0.00000
2	0.00000
3	0.00000
4	0.00000
5	0.00000
6	0.00003
7	0.00018
8	0.00094
9	0.00394
10	0.01386
11	0.04093
12	0.10181
13	0.21422
14	0.38283
15	0.58516
16	0.77484
17	0.90874
18	0.97569
19	0.99683
20	1.00000

Wie groß ist die Wahrscheinlichkeit für einen Fehler 2.Art (β-Fehler) und wie groß die Teststärke (1 - β)? Übrigens: Aus pragmatischen Gründen wird für p_1 oft die relative Häufigkeit f des Ereignisses aus der Stichprobe verwendet (hier wäre z.B. $f_{Frauen} = 0{,}7$).

FORTSETZUNG: SIGNIFIKANZTEST (NEYMAN-PEARSON)

Das β-Risiko wird von folgenden Dingen determiniert:

1) Von der Art der Fragestellung (ein- oder zweiseitig): Bei richtig gewähltem einseitigen Kritischen Bereich wird das β-Risiko kleiner als bei einem zweiseitigen Kritischen Bereich.
2) Von der Größe des α-Risikos: Je größer das α-Risiko ist, desto kleiner wird das β-Risiko, da bei größerem α-Risiko auch der Kritische Bereich größer bzw. der Annahmebereich kleiner wird.
3) Von der Effektstärke, das ist der Abstand zwischen H_0 und H_1. Ist H_0: p = 0,5 und H_1 entweder a) p = 0.6 oder b) p = 0,75, so ist im Fall b) die Effektstärke größer. Es gilt: Je größer die Effektstärke (auch experimenteller Effekt), desto kleiner das β-Risiko.
4) Vom Stichprobenumfang n: Das β-Risiko wird mit wachsendem n kleiner, deshalb: Besser größeren Stichprobenumfang wählen!

Oft ist es nicht so wie in unserem Übungsbeispiel, daß man schon den Stichprobenumfang n und eine bestimmte Effektstärke hat, und nur noch β und 1 - β berechnet, sondern vielmehr überlegt man sich *im vorhinein*, wie groß man den experimentellen Effekt bzw. p_1 einschätzt und bestimmt dazu den nötigen Stichprobenumfang, mit dem man überhaupt ein signifikantes Ergebnis bei *vorgegebener Teststärke* (standardmäßig mindestens 0,8 bzw. 80%) erhalten könnte. Die Formel zur Bestimmung des notwendigen Stichprobenumfangs für ein bestimmtes α und β lautet:

$$\sqrt{n} = \frac{z_{1-\alpha}\sqrt{p_0 q_0} - z_\beta \sqrt{p_1 q_1}}{|p_1 - p_0|}$$

Hierbei ist $z_{1-\alpha}$ bzw. z_β der z-Wert, bei dem die Verteilungsfunktion (F) der Standardnormalverteilung den Wert 1 - α bzw. β hat. Er ist aus Tabellen abzulesen.

Um die H_0: p = 0,5 von der H_1': p = 0,75 signifikant (5%) bei einer Teststärke von 80% unterscheiden zu können, benötigt man ein n von:

$$\sqrt{n} = \frac{z_{1-0.05}\sqrt{0{,}5 \cdot 0{,}5} - z_{0,2}\sqrt{0{,}75 \cdot 0{,}25}}{|0{,}75 - 0{,}5|}; \text{ mit } z_{0,95} = 1{,}645 \text{ und } z_{0,2} = -0{,}84$$

$$\sqrt{n} = \frac{1{,}645\sqrt{0{,}5^2} - (-0{,}84)\sqrt{0{,}1875}}{0{,}25} =$$

$$\sqrt{n} = \frac{0{,}8225 + 0{,}3637}{0{,}25} =$$

$$\sqrt{n} = 4{,}7448 =$$

$$n = 22{,}5 \approx 23,$$ da n nur diskret und mindestens so groß wie das errechnete n sein muß.

Für die Bestimmung des Kritischen y kann man ebenfalls die Annäherung an die Verteilungsfunktion der Standardnormalverteilung verwenden:

$$0{,}95 = P(Z \le z_{krit}) = P\left(Z \le \frac{y + 0{,}5 - np}{\sqrt{npq}}\right) \approx P(Y \le y)$$

mit z_{krit} aus Tabellenwerk ist y_{krit} bestimmbar (nächst größeres y + 1 bei schiefem Wert).

ANDERE ANWENDUNGEN DES SIGNIFIKANZTESTS

Der Mc-Nemar-Test

Anzuwenden für Anordnungen der Art **o**bservation-**t**reatment-**o**bservation (O-T-O), z.B.:
Eine neuartige Therapie wird erprobt:

		nachher		
		gesund	**krank**	
vorher	**gesund**	5	**1**	6
	krank	**20**	4	24
		25	5	30

Man betrachtet nun nur diejenigen Personen, die ihren Zustand *gewechselt* haben; das ist unser n (n = 21). Unser Y sind diejenigen Personen, die nach der Therapie auf *gesund* gewechselt haben; in unserem Fall 20 (Y = 20). Jetzt kann man wie gewohnt berechnen, ab wievielen Wechslern Y (in der positiven Richtung) man nicht mehr davon ausgehen kann, daß das Wechseln von gesund nach krank und umgekehrt gleich wahrscheinlich ist (p = 0,5), was der Nullhypothese entspräche. Man kommt auf ein y_{krit} von 15, der Kritische Bereich K sieht also wie folgt aus: K = {15, 16, ..., 20, 21}. Da in diesem Beispiel Y = 20 ist, würde die H_0 verworfen werden.

Der Nachteil an diesem Test ist, daß man nicht *vorher* den Stichprobenumfang kennt, da man nicht weiß, wieviele Leute wechseln werden.

Die Chi-Quadrat-Näherung

Man kann die Binomialverteilung wie gehabt mit der Standardnormalverteilung annähern:

$$z = \frac{Y - np_0}{\sqrt{np_0q_0}}$$

dabei ist $z_{krit,\ 5\%} = z_{95\%} = 1{,}645$. Bei zweiseitiger Fragestellung bräuchte man $z_{2.5\%}$ und $z_{97{,}5\%}$. Nun kann man durch einfaches Quadrieren von z auf z^2 das gleiche Problem leichter lösen; man hat so keine negativen z-Werte mehr, sondern nur noch positive sog. χ^2-Werte. Die dazugehörige Wahrscheinlichkeitsfunktion heißt Chi-Quadrat-Funktion. Demnach lautet die Chi-Quadrat-Näherung:

$$z^2 = \chi^2 = \left(\frac{Y - np_0}{\sqrt{np_0q_0}}\right)^2 = \frac{(Y - np_0)^2}{np_0} + \frac{((n-Y) - nq_0)^2}{nq_0} = \frac{(O_+ - E_+)^2}{E_+} + \frac{(O_- - E_-)^2}{E_-}$$

mit O = observed
und E = expected;
+ = Treffer,
– = Niete

$$\Rightarrow \text{allgemein: } \chi^2 = \sum_{j=1}^{J} \frac{(O_j - E_j)^2}{E_j} \text{ mit df (Freiheitsgrade)} = J - 1 \text{ (J = Anzahl der Kategorien)}$$

(Chi-Quadrat-Einstichprobentest)

Hier ist zu beachten, daß sich die Intervalle quadrieren: $P(-1 \le z \le 1) \Rightarrow P(0 \le \chi^2 \le 1)$, $P(-2 \le z \le 2) \Rightarrow P(0 \le \chi^2 \le 4)$, und z.B. 95% = $P(-1{,}96 \le z \le 1{,}96) \Rightarrow P(0 \le \chi^2 \le 3{,}84)$. Diesen Test kann man nun auch auf nicht binomialverteilte Variablen anwenden: Man hat z.B. eine Stichprobe von 30 Kugeln in drei Farben (rot, grün, blau). Die einfachste Annahme wäre nun, daß in der Population, aus der die Kugeln kommen, die drei Farben gleich oft vorkommen. Wenn dem so ist, hätte man je 10 Kugeln unterschiedlicher Farbe zu erwarten. Zieht man aber 5 rote, 7 grüne und 18 blaue Kugeln, kommt man auf ein χ^2 von:

$$\chi^2 = \frac{(5-10)^2}{10} + \frac{(7-10)^2}{10} + \frac{(18-10)^2}{10} = \frac{25+9+64}{10} = 9{,}8 > 5{,}99\left(\chi^2_{\text{krit},5\%,\text{df}\,2}\right)$$

$\Rightarrow H_0$ verwerfen!

Anderes Beispiel:

Man will untersuchen, ob ein Schüler in der Verteilung seiner Zeugnisnoten nur zufällig oder signifikant von dem Modell der Schüler-Durchschnitts-Notenwahrscheinlichkeitsverteilung abweicht:

Note	**1**	**2**	**3**	**4**	**5**	**6**	**Summen**
Schüler X	4	2	3	3	0	0	12
Durchschnitts-w'keit	0,15	0,25	0,3	0,15	0,1	0,05	1,0
E_i (12 · W'keit)	1,8	3	3,6	1,8	1,2	0,6	12
O_i - E_i	2,2	-1	-0,6	1,2	-1,2	-0,6	

$$\chi^2 = \frac{2{,}2^2}{1{,}8} + \frac{-1^2}{3} + \frac{-0{,}6^2}{3{,}6} + \frac{1{,}2^2}{1{,}8} + \frac{-1{,}2^2}{1{,}2} + \frac{-0{,}6^2}{0{,}6} = 5{,}722 < 11{,}07\left(\chi^2_{\text{krit},5\%,\text{df}\,5}\right)$$

$\Rightarrow H_0$ beibehalten!

Er unterscheidet sich also nicht signifikant vom Modell.

Die Erweiterung des Chi-Quadrat-Tests (Chi-Quadrat-Mehrstichprobentest)

Man hat nicht mehr nur eine qualitative Variable (wie z.B. Farbe) sondern zwei, von denen man sagt, die Ausprägungen der einen seien durch die andere bestimmt. Eine Frage in der Art wäre etwa, ob Geschlecht und Lieblingsfarbe zusammenhängen:

	rot	**grün**	**gelb**	
Männer	2 0,2933·14 = 4,11	5 0,3467·14 = 4,85	7 0,36·14 = 5,04	14
Frauen	20 0,2933·61 = 17,89	21 0,3467·61 = 21,15	20 0,36·61 = 21,96	61
	22 (29,33%)	26 (34,67%)	27 (36%)	75

$$\chi^2 = \frac{(2-4{,}11)^2}{4{,}11} + \frac{(5-4{,}85)^2}{4{,}85} + \frac{(7-5{,}04)^2}{5{,}04} + \frac{(20-17{,}89)^2}{17{,}89} + \frac{(21-21{,}15)^2}{21{,}15} + \frac{(20-21{,}96)^2}{21{,}96} =$$

$$= 2{,}275 < 5{,}99\left(\chi^2_{\text{krit},5\%,\text{df}\,2}\right), \text{ mit df } = (3-1)\cdot(2-1) = 2$$

H_0 („*Lieblingsfarbe bei Männern und Frauen gleich*“) kann nicht verworfen werden.

CHI-QUADRAT-TEST (NEYMAN-PEARSON VARIANTE)

Ob die beobachteten Werte bei einem Chi-Quadrat-Test von den erwarteten so stark abweichen, daß man zu 95% davon ausgehen kann, daß dies nicht zufällig der Fall ist, hängt natürlich stark vom Stichprobenumfang ab:

Beispiel: Ist die Nullhypothese zu verwerfen, daß sich gleich viele Männer wie Frauen an einem Nähkurs beteiligen? (Geprüft an 10 und an 100 Personen)

	Männer	Frauen	Männer	Frauen
beobachtet	3	7	30	70
erwartet	5	5	50	50
$\chi^2_{krit,\ df1} = 3{,}84$	$\chi^2 = 1{,}600$		$\chi^2 = 15{,}998$	

Man erhielte also bei gleichen relativen Abweichungen von der Erwartung unterschiedliche Testresultate (H_0 beibehalten und H_0 verwerfen) bzw. ist sich nach der Entscheidung über die Teststärke (also die Wahrscheinlichkeit bei tatsächlich unterschiedlicher Verteilung für Männer und Frauen die Nullhypothese auch tatsächlich verwerfen zu können) und das β-Risiko (W'keit die H_0 zu unrecht beibehalten zu haben) noch im Unklaren. Man kann den Chi-Quadrat-Test jedoch auch nach der NEYMAN-PEARSONschen Testtheorie manchen, indem man einfach die erwarteten und beobachteten relativen Häufigkeiten als Wahrscheinlichkeiten auffaßt und mit ihnen einen Wert berechnet, der sich *Effektgröße* nennt:

$$w^2 = \sum_j \frac{\left({}_1p'_j - {}_0p_j\right)^2}{{}_0p_j} = \frac{(0{,}3-0{,}5)^2}{0{,}5} + \frac{(0{,}7-0{,}5)^2}{0{,}5} = 0{,}160$$

allgemein für j Ausprägungen von k Kategorien (Stichproben):

$$w^2 = \sum_k \sum_j \frac{\left({}_1p'_{jk} - {}_0p_{jk}\right)^2}{{}_0p_{jk}} \text{ mit } df = (j-1)\cdot(k-1)$$

Für ${}_1p'_{jk}$ werden oft die relativen Häufigkeiten der Ausprägungen der Merkmale in den einzelnen Stichproben verwendet. Nun kann man in Tabellenwerken schauen, ab welchem Stichprobenumfang die errechnete Effektgröße bei vorgegebenem Signifikanzniveau und vorgegebener Teststärke signifikant wäre.

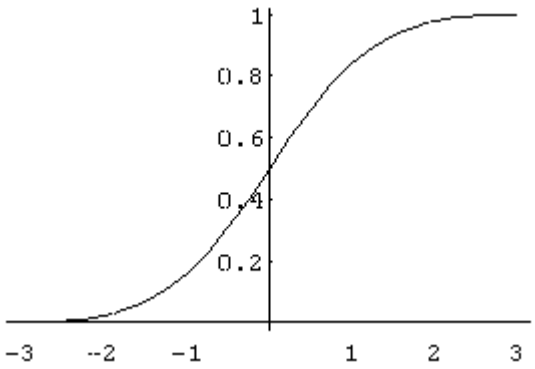
Wenn das Ergebnis signifikant ist, dann wird damit folgende Nullhypothese verworfen:

H_0: ${}_0p_{jk} = {}_0p_{.j} = [{}_0p_{k.}]$ für alle j, k

H_1': $p_{jk} \neq p_{.j}$ für mindestens ein j und k

k =	**j =**	**1**	**2**	**3**	
1		${}_0p_{11}$	${}_0p_{12}$	${}_0p_{13}$	**${}_0p_{1.}$**
2		${}_0p_{21}$	${}_0p_{22}$	${}_0p_{23}$	**${}_0p_{2.}$**
		${}_0p_{.1}$	**${}_0p_{.2}$**	**${}_0p_{.3}$**	

MARKANTE WERTE DER VERTEILUNGSFUNKTION DER STANDARDNORMALVERTEILUNG



Verteilungsfunktion der Standardnormalverteilung

Verteilungsfunktion	**z-Wert**
0,05%	-3,290
0,10%	-3,090
0,50%	-2,576
1,00%	-2,326
2,50%	-1,960
5,00%	-1,645
10,00%	-1,282
15,00%	-1,036
20,00%	-0,842
50,00%	0,000
95,00%	1,645
97,50%	1,960
99,00%	2,326
99,50%	2,576
99,90%	3,090
99,95%	3,290