

Hochschule München
Fakultät
Wirtschaftsingenieurwesen



Datenanalyse

Prof. Dr. Volker Abel

Version 2.0

Inhaltsverzeichnis

1. Auswertung und Modellierung von Zähldaten.....	1
1.1 Auswertung von prozentualen Häufigkeiten.....	1
1.2 Auswertung einer mittleren Anzahl	4
1.3 Modellierung von Zähldaten	5
1.3.1 Modell Hypergeometrische Verteilung	5
1.3.2 Modell Binomial – Verteilung.....	8
1.3.3 Modell Poisson – Verteilung.....	10
1.4 Modellwahl	11
1.5 Der χ^2 -Anpassungstest	11
2. Auswertung von metrischen Daten bei einer Stichprobe	16
2.1 Maßzahlen für die Lage	16
2.1.1 Arithmetisches Mittel	16
2.1.2 Median (Zentralwert).....	16
2.1.3 α -Quantil	17
2.2 Maßzahlen für die Streuung.....	17
2.2.1 Spannweite	17
2.2.2 Inter-Quartil-Spannweite	17
2.2.3 Varianz	17
2.2.4 Standardabweichung	18
2.2.5 Mittlere absolute Abweichung vom Mittelwert	18
2.2.6 Mittlere absolute Abweichung vom Median.....	18
2.2.7 Median absolute Abweichung	18
2.3 Maßzahlen für die Schiefe.....	19
2.4 Dot Plot.....	20
2.5 Histogramm.....	20
2.6 Stamm- und Blatt- Darstellung	21
2.7 Box-and-Whisker-Plot	22
2.8 Summenhäufigkeitsfunktion / empirische Verteilungsfunktion.....	24
2.9 Eine Zufallsstichprobe	25
2.10 Statistische Anteilsbereiche.....	26
2.10.1 Zweiseitiger Anteilsbereich.....	26
2.10.2 Einseitiger Anteilsbereich	27
3. Auswertung und Vergleich von metrischen Daten bei zwei oder mehreren Stichproben.....	29

3.1	Kolmogorov-Smirnov-Test	29
3.2	Tücken von Signifikanztests (Hypothesentests)	32
3.3	Zusammenfassung	33
4.	Modellierung von metrischen Daten	34
4.1	Allgemeine Vorbemerkungen	34
4.2	Exponentialverteilung	35
4.2.1	Das Modell	35
4.2.2	Erkennen der Verteilung und Schätzungen des Parameters	36
4.3	Normalverteilung	38
4.3.1	Das Modell	38
4.3.2	Erkennen der Verteilung und Schätzen der Parameter	43
4.4	Approximationen statistischer Verteilungen durch die Normalverteilung	47
4.5	Die Kosinusverteilung	48
4.5.1	Das Modell	48
4.5.2	Schätzung der Parameter und Vergleich mit der Normalverteilung	50
4.6	Weibull-Verteilung	51
4.6.1	Das Modell	51
4.6.2	Erkennen der Verteilung und Schätzen der Parameter	52
5.	Industrial Statistics	53
5.1	Statistical Process Control (SPC)	53
5.1.1	Introduction	53
5.1.2	Die wichtigsten Regelkarten	58
5.1.3	Normalverteilung und SPC	59
5.2	Capability Indices	60
5.2.1	Introduction	60
5.2.2	Prozessbeherrschung und Prozessfähigkeit	62
5.2.3	Normalverteilung und Fähigkeitsindizes	65
5.2.4	Kosinusverteilung und Fähigkeitsindizes	67
5.3	Reliability Engineering	67
5.3.1	Ausfallrate / failure rate $h(x)$	67
5.3.2	$h(x)$ für die Exponentialverteilung	68
5.3.3	$h(x)$ für die Weibullverteilung	68
6.	Korrelation und Regression	69
6.1	Korrelation	69
6.2	Regression	71
6.2.1	Einfache lineare Regression	71
6.2.2	Einfache nicht lineare Regression	76
6.2.3	Mehrfache (Multiple) lineare Regression	79

7.	Etwas Allerlei zum guten (?) Schluss	86
7.1	Das Geburtstagsproblem.....	86
7.2	Das Ziegenproblem	87
7.3	Das Gesetz von Benford	88

1. Auswertung und Modellierung von Zähldaten

1.1 Auswertung von prozentualen Häufigkeiten

Ist bei einer Zufallsstichprobe vom Umfang n

- das Ergebnis A m mal eingetreten, oder
- die Frage A m mal bejaht worden,

so ist $p = \frac{m}{n} (\cdot 100\%)$ die prozentuale (relative) Häufigkeit von A in der Stichprobe (m ist die absolute Häufigkeit).

Eine Stichprobe ist ein Teil einer umfassenderen Menge, der Grundgesamtheit oder Population. Wie groß ist, mit Hilfe des Stichprobenergebnisses „geschätzt“, die prozentuale Häufigkeit von A in der Grundgesamtheit?

Exakt weiß man das natürlich nicht.

$\hat{p} = \frac{m}{n}$ ist ein sogenannter „Punktschätzer“.

Sein Vorteil: Nur ein Wert

Sein Nachteil: Fast sicher falsch, denn warum sollten prozentuale Häufigkeit in der Stichprobe und in der Grundgesamtheit identisch sein?

Statistiker bevorzugen in dieser Situation eine Intervallschätzung (**Konfidenzintervall**).

Ihr Nachteil: Ein Intervall ist nicht so „scharf“ wie ein Punkt.

Ihr Vorteil: Man kann angeben, welches Vertrauen man in dieses Intervall haben kann. D.h. man kennt die Wahrscheinlichkeit, mit der das Intervall die unbekannte Häufigkeit in der Grundgesamtheit überdeckt.

Beispiel

Von $n = 300$ zufällig ausgewählten Personen (in Deutschland lebend, über 18 Jahre) wissen $m = 176$ Personen, wer als neuer SPD-Vorsitzender vorgeschlagen ist.

$p = \frac{176}{300} = 0,586\bar{6} = 58,76\%$ ist die prozentuale Häufigkeit in der Stichprobe. Wie sicher

(vertrauenswürdig) ist dieser Wert für die Grundgesamtheit?

➔ **Konfidenzintervall für die prozentuale Häufigkeit in der Grundgesamtheit**

(Voraussetzung: $n \cdot p \cdot (1 - p) \geq 9$)

$$(1.1) \quad p_u = \frac{2m + z^2 - z \cdot \sqrt{z^2 + 4 \cdot m \cdot \left(1 - \frac{m}{n}\right)}}{2 \cdot (n + z^2)}$$

$$(1.2) \quad p_o = \frac{2m + z^2 + z \cdot \sqrt{z^2 + 4 \cdot m \cdot \left(1 - \frac{m}{n}\right)}}{2 \cdot (n + z^2)}$$

Dabei ist z (die Werte stammen aus der sog. Normalverteilung, siehe S.42)

Vertrauen	90%	95%	99%
z	1,645	1,960	2,578

Bemerkungen

1. Ist $n \cdot p \cdot (1 - p) < 9$, so sind noch komplizierte Formeln von CLOPPER und PEARSON zu verwenden (Fachliteratur).
2. Was geschieht mit den Intervallgrenzen p_u und p_o , wenn das Ereignis A
 - a) gar nicht b) immer eintritt?

→Tafelanschrieb

Das Intervall $[p_u, p_o]$ heißt zweiseitig. Manchmal interessiert man sich für ein einseitiges Konfidenzintervall $[0, p_o]$ bzw. $[p_u, 1]$.

Es gelten dieselben Formeln (1-1 / 1-2) mit dem Unterschied, dass das Vertrauen in der z-Tabelle nicht mehr 90%, 95%, 99% ist, sondern **95%, 97,5%, 99,5%**.

In vielen Lehrbüchern findet man noch einfachere, aber etwas schlechtere Formeln:

$$(1.3) \quad p_u = \frac{m}{n} - z \cdot \sqrt{\frac{\frac{m}{n} \cdot \left(1 - \frac{m}{n}\right)}{n}}$$

$$(1.4) \quad p_o = \frac{m}{n} + z \cdot \sqrt{\frac{\frac{m}{n} \cdot \left(1 - \frac{m}{n}\right)}{n}}$$

(Diese Formeln wollen wir hier nicht verwenden)

Aber: Mit dieser Formel kann man die Frage nach dem notwendigen Stichprobenumfang n beantworten. Die Antwort hängt ab vom

- Vertrauen und der
- Intervallbreite ($p_o - p_u$).

$$p_o - p_u = \frac{m}{n} + z \cdot \sqrt{\frac{\frac{m}{n} \cdot \left(1 - \frac{m}{n}\right)}{n}} - \left(\frac{m}{n} - z \cdot \sqrt{\frac{\frac{m}{n} \cdot \left(1 - \frac{m}{n}\right)}{n}} \right) = 2 \cdot z \cdot \sqrt{\frac{\frac{m}{n} \cdot \left(1 - \frac{m}{n}\right)}{n}}$$

$$= 2 \cdot z \cdot \sqrt{\frac{p \cdot (1-p)}{n}} \leq d$$

$$(1.5) \quad n \geq \left(\frac{2 \cdot z}{d} \right)^2 \cdot p \cdot (1-p)$$

Problem: p und damit $p \cdot (1-p)$ sind unbekannt.

Lösung:

a) Wir nehmen den schlimmsten Fall:

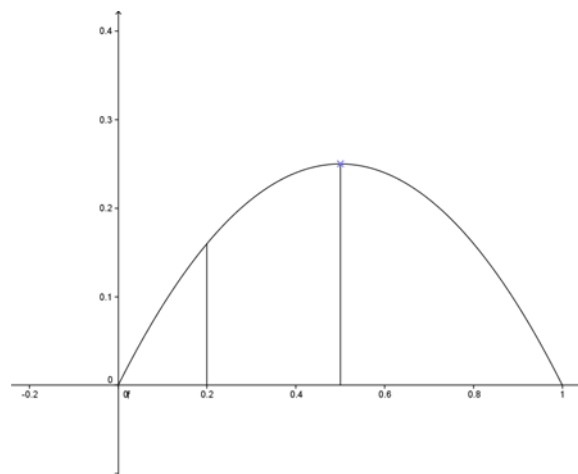
$$f(p) = p \cdot (1-p) \text{ Maximal?}$$

$$= p - p^2$$

$$f'(p) = 1 - 2p = 0 \quad \Leftrightarrow p = 0,5$$

$$f''(p) = -2 \quad \Rightarrow \text{also Maximum}$$

$$\Rightarrow f(0,5) = 0,5(1-0,5) = 0,25 = \frac{1}{4} \text{ ist Maximalwert}$$



$$n \geq \left(\frac{2 \cdot z}{d} \right)^2 \cdot \frac{1}{4} = \left(\frac{2 \cdot z}{2 \cdot d} \right)^2$$

$$n \geq \left(\frac{z}{d} \right)^2$$

b) Es liegt Erfahrungswissen vor

z.B. Grüne $p \leq 0,20$

$$f(p) = p \cdot (1-p) \text{ maximal wenn } p = 0,2$$

$$= 0,2 \cdot 0,8 = 0,16$$

$$n \geq \left(\frac{2 \cdot z}{d} \right)^2 \cdot 0,16$$

Bisher wurde vorausgesetzt, dass der Umfang N gleich der Grundgesamtheit ∞ ist.

Jetzt soll N eine endliche Zahl sein. Dann sind die Grenzen des Konfidenzintervalls gegeben durch:

$$(1.6) \quad p_u' = \frac{m-0,5}{n} - \left(\frac{m-0,5}{n} - p_u \right) \cdot \sqrt{\frac{N-n}{N-1}}$$

$$(1.7) \quad p_o' = \frac{m-\frac{m}{n}}{n} + \left(p_o - \frac{m-\frac{m}{n}}{n} \right) \cdot \sqrt{\frac{N-n}{N-1}}$$

wobei p_u und p_o die aus den Formeln 1-1 und 1-2 berechneten Werte sind.

Faustregel:

Ist $\frac{n}{N} \leq 0,05$, kann auf die Endlichkeitskorrekter $\left(\sqrt{\frac{N-n}{N-1}} \right)$ in den Formeln 1-6 und 1-7 verzichtet

werden.

→ Beweis Tafelanschrieb

1.2 Auswertung einer mittleren Anzahl

Ausgangspunkt

ist das von dem deutsch-russischen Mathematiker L. von Borkiewicz 1898 veröffentlichte Standardbeispiel, welches in vielen Lehrbüchern auf der ganzen Welt benutzt wird.

In 10 Regimentern der preußischen Armee gab es in einem Zeitraum von 20 Jahren folgende (absolute) Häufigkeiten von durch Huftritt getöteten Soldaten.

Anzahl der Todessfälle pro Jahr & Regiment	Absolute Häufigkeit in 20 Jahren bei 10 Regimentern
0	109
1	65
2	22
3	3
4	1
≥ 5	0
	200 = 10 · 20 = n

Wie groß ist die mittlere Anzahl der durch Huftritt getöteten Soldaten pro Jahr und Regiment?

$$\lambda = \frac{0 \cdot 109 + 1 \cdot 65 + 2 \cdot 22 + 3 \cdot 3 + 4 \cdot 1 + 5 \cdot 0}{200} = \frac{122}{200} = 0,61$$

Sieht man dies als eine Zufallsstichprobe an, stellt sich die Frage nach der mittleren Anzahl in der Grundgesamtheit (alle Regimenter, längere Zeit). Es stellt sich also die Frage nach einem

Konfidenzintervall für λ .

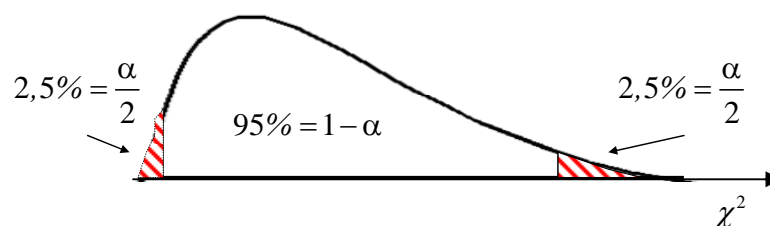
$$1 - \alpha = \text{Vertrauen}$$

Notation: n = Umfang der Stichprobe

x = Gesamtzahl von 'A' in der Stichprobe

$$(1.8) \quad \lambda_u = \frac{\chi^2_{2x; \frac{\alpha}{2}}}{2 \cdot n}$$

$$(1.9) \quad \lambda_o = \frac{\chi^2_{2x+2; 1-\frac{\alpha}{2}}}{2 \cdot n}$$



Eine Tabelle der sogenannten χ^2 -Verteilung findet sich auf der Seite 15.

1.3 Modellierung von Zähldaten

X = Zahl der „A“ in der Zufallsstichprobe

1.3.1 Modell Hypergeometrische Verteilung

Die Grundgesamtheit besteht aus N Elementen. Davon haben M Elemente die Eigenschaft A (und der Rest $N - M$ die Eigenschaft \bar{A} (nicht A)).

Aus der Gesamtheit werden n Elemente zufällig und ohne Zurücklegen ausgewählt.

Dann ist die Wahrscheinlichkeit, dass davon k Elemente die Eigenschaft A (und $n - k$ die Eigenschaft \bar{A}) haben:

$$(1.10) \quad P(X = k) = \frac{\binom{M}{k} \cdot \binom{N-M}{n-k}}{\binom{N}{n}}$$

wobei der Binomialkoeffizient definiert ist als

$$\binom{a}{b} = \frac{a!}{b!(a-b)!} \quad \begin{array}{l} a! = 1 \cdot 2 \cdot 3 \cdot \dots \cdot a \\ 0! = 1 \end{array}$$

X heißt dann hypergeometrisch verteilt und hat den

Erwartungswert

$$(1.11) \quad E(X) = n \cdot \frac{M}{N}$$

und die Varianz

$$(1.12) \quad \text{var}(X) = n \cdot \frac{N-n}{N-1} \cdot \frac{M}{N} \cdot \left(1 - \frac{M}{N}\right)$$

und

$$p = \frac{M}{N} \text{ als den Anteil in der Gesamtheit mit Eigenschaft A.}$$

(nicht zu verwechseln mit $p = \frac{m}{n}$ in Kapitel 1.1)

Beispiel

$N = 19$ Studenten $M = 7$ Frauen $N - M = 12$ Männer
 $n = 5$ "Ziehungen" $X = \text{Zahl der Frauen}$

Wie hoch ist die Wahrscheinlichkeit, dass man aus 19 Studenten, von denen 7 Frauen sind, mit 5 „Ziehungen“ ohne Zurücklegen 4 Frauen auswählt?

$$P(X = 4) = \frac{\binom{7}{4} \cdot \binom{12}{1}}{\binom{19}{5}} = \frac{35 \cdot 12}{11628} = 0,036 = 3,6\%$$

Beispiel

Ein Kunde erhält eine Lieferung von 500 Antriebswellen. Aus Erfahrung muss der Lieferant der annehmen, dass 2% zu beanstanden sind.

Die Eingangskontrolle des Kunden wählt 50 Wellen zufällig aus.

a) Wie viele schlechte Wellen werden wohl gefunden?

kommt drauf an: $\underbrace{0}_{\text{minimal}}, 1, 2, \dots, \underbrace{10}_{\text{maximal}}$

zu erwarten sind: $E(X) = n \cdot \frac{M}{N} = n \cdot p = 50 \cdot 0,02 = 1$

b) Wie groß ist die Wahrscheinlichkeit, dass

- o keine
- o höchstens 1
- o mindestens 2

schlechte Wellen gefunden werden?

$$\begin{aligned} N &= 500 & M &= 10 \\ N - M &= 490 & (p &= 2\%) \\ n &= 50 \end{aligned}$$

$$P(X=0) = \frac{\binom{10}{0} \cdot \binom{490}{50}}{\binom{500}{50}} = 0,345 = 34,5\%$$

$$P(X=1) = \frac{\binom{10}{1} \cdot \binom{490}{49}}{\binom{500}{50}} = 0,391 = 39,1\%$$

$$P(X \leq 1) = 0,345 + 0,391 = 0,736 = 73,6\%$$

$$P(X \geq 2) = 1 - P(X < 2) = 1 - \underbrace{P(X \leq 1)}_{\text{diskrete Verteilung}} = 1 - 0,736 = 0,264 = 26,4\%$$

$$\begin{aligned} P(X \geq 3) &= 1 - P(X \leq 2) \\ &= 1 - [P(X=0) + P(X=1) + P(X=2)] \end{aligned}$$

Die Parameter N und M werden in den Lehrbüchern immer als bekannt vorausgesetzt.

In der Realität sind sie meist unbekannt und müssen mittels einer Stichprobe geschätzt werden.

- o N unbekannt: capture / recapture-Modelle
s. Übungsaufgabe
- o M unbekannt: also auch $p = \frac{M}{N}$ unbekannt

Ist in der Stichprobe m die Anzahl der Elemente mit Eigenschaft A , so ist der Anteil in der

$$\text{Stichprobe } \hat{p} = \frac{m}{n}$$

Setzen wir für das unbekannte p den Punktschätzer \hat{p} , also $\frac{M}{N} = \frac{m}{n}$, so ist $\hat{M} = \frac{m}{n} \cdot N$ der

Punktschätzer für M .

1.3.2 Modell Binomial – Verteilung

Ein Experiment wird n mal unabhängig durchgeführt. Die Zufallsvariable X zählt, wie oft dabei das Ergebnis A eintritt. Setzt man $p = P(A)$, so gilt:

$$(1.13) \quad P(X = k) = \binom{n}{k} \cdot p^k \cdot (1-p)^{n-k}$$

für $k = 0, 1, 2, \dots, n$

X heißt dann binomialverteilt mit dem Erwartungswert:

$$(1.14) \quad E(X) = n \cdot p$$

und der Varianz:

$$(1.15) \quad \text{var}(X) = n \cdot p \cdot (1-p).$$

Beispiel

Sie würfeln 5-mal und interessieren sich bei jedem Wurf, ob eine Primzahl eintritt.

X = Anzahl der Primzahlen bei fünf Würfeln

$$A = \{1, 2, 3, 5\} \quad P(A) = \frac{4}{6} = \frac{2}{3} = p \quad n = 5$$

$$P(X = 0) = \binom{5}{0} \cdot \left(\frac{2}{3}\right)^0 \cdot \left(\frac{1}{3}\right)^5 = 0,004$$

$$P(X = 1) = \binom{5}{1} \cdot \left(\frac{2}{3}\right)^1 \cdot \left(\frac{1}{3}\right)^4 = 0,041$$

$$P(X = 2) = \binom{5}{2} \cdot \left(\frac{2}{3}\right)^2 \cdot \left(\frac{1}{3}\right)^3 = 0,165$$

$$P(X = 3) = \binom{5}{3} \cdot \left(\frac{2}{3}\right)^3 \cdot \left(\frac{1}{3}\right)^2 = 0,329$$

$$P(X = 4) = \binom{5}{4} \cdot \left(\frac{2}{3}\right)^4 \cdot \left(\frac{1}{3}\right)^1 = 0,329$$

$$P(X = 5) = \binom{5}{5} \cdot \left(\frac{2}{3}\right)^5 \cdot \left(\frac{1}{3}\right)^0 = 0,132$$

$$\sum = 1,000$$

$$E(X) = n \cdot p = 5 \cdot \frac{2}{3} = 3 \cdot \frac{1}{3}$$

Setzt man $\frac{M}{N} = p$, so haben die hypergeometrische Verteilung (Ziehen ohne Zurücklegen) und die

Binomial-Verteilung (Ziehen mit Zurücklegen) den gleichen Erwartungswert, aber verschiedene Varianzen:

Varianz der hypergeometrischen Verteilung

$$\begin{aligned} &= n \cdot \frac{M}{N} \left(1 - \frac{M}{N}\right) \cdot \frac{N-n}{N-1} \\ &= n \cdot p \cdot (1-p) \cdot \frac{N-n}{N-1} \\ &\approx n \cdot p \cdot (1-p) \cdot \frac{N-n}{N} \\ &= n \cdot p \cdot (1-p) \cdot \left(1 - \frac{n}{N}\right) \\ &\approx \underbrace{n \cdot p \cdot (1-p)}_{\text{=Varianz der Binomialverteilung}} \quad \text{wenn } \frac{n}{N} \text{ "klein"} \end{aligned}$$

Faustregel für „klein“: $\frac{n}{N} \leq 0,05$ (d.h. Auswahlatz $\leq 5\%$).

Hat man Ziehen ohne Zurücklegen und $\frac{n}{N} \leq 0,05$, darf stattdessen die Binomial-Verteilung (Ziehen mit Zurücklegen) verwendet werden.

Punktschätzung von p

$p = P(A)$ wird durch die relative Häufigkeit von A in der Stichprobe geschätzt.

Der Wertebereich der binomialverteilten Zufallsvariablen X ist $0, 1, 2, \dots, n$. Eine solche Obergrenze entfällt bei der POISSON Verteilung.

1.3.3 Modell Poisson – Verteilung

Eine Zufallsvariable X heißt Poissonverteilt, wenn

$$(1.16) \quad P(X = k) = e^{-\lambda} \cdot \frac{\lambda^k}{k!}$$

für $k = 0, 1, 2, \dots, \infty$.

Es ist
der Erwartungswert:

$$(1.17) \quad E(X) = \lambda$$

und die Varianz:

$$(1.18) \quad \text{var}(X) = \lambda.$$

Punktschätzung von λ

Da $\lambda = E(X)$, wird λ durch den Mittelwert \bar{x} der Stichprobe geschätzt. (\bar{x} ist das λ von Kapitel 1.2)

$$\hat{\lambda} = \bar{x} \quad \text{Punktschätzer für } \lambda$$

Beispiel

In Europa stürzen im langjährigen Mittel 0,23 Verkehrsflugzeuge pro Jahr ab. Wie groß ist die Wahrscheinlichkeit, dass in einem Jahr 0,1,2,3... Verkehrsflugzeuge in Europa abstürzen?

$$\lambda = 0,23$$

$$P(X = 0) = e^{-0,23} \cdot \frac{0,23^0}{0!} = 0,7945$$

$$P(X = 1) = e^{-0,23} \cdot \frac{0,23^1}{1!} = 0,1827$$

$$P(X = 2) = e^{-0,23} \cdot \frac{0,23^2}{2!} = 0,0210$$

$$\overline{\sum 0,9982}$$

$$P(X > 2) \approx 0$$

1.4 Modellwahl

$0 \leq X \leq n$:

Ziehen ohne Zurücklegen: Hypergeometrische Verteilung mit N, M, n .

Ziehen „mit“ Zurücklegen: Binomial-Verteilung mit n, p .

Ziehen ohne Zurücklegen und $\frac{n}{N} \leq 0,05$: Binomial-Verteilung mit n, p .

$0 \leq X$ (keine Obergrenze)
Poissonverteilung.

Ist bei der hypergeometrischen Verteilung

- N unbekannt, so wird es mit einem capture / recapture-Verfahren geschätzt (vgl. Übungsaufgabe).
- M unbekannt, so wird gesetzt $\frac{m}{n} = \frac{M}{N}$, also $\hat{M} = \frac{m}{n} \cdot N = p \cdot N$.

Ist bei einer Binomial-Verteilung p unbekannt, so wird es durch die prozentuale Häufigkeit $\frac{m}{n}$ der Stichprobe geschätzt.

Für alle anderen Verteilungen von Zähldaten (inkl. der Poissonverteilung) ist das Modell zunächst nur eine Vermutung (Hypothese), die anhand der Stichprobe durch einen Anpassungstest überprüft werden muss.

1.5 Der χ^2 -Anpassungstest

$\chi = CHI$

Soll die Hypothese einer vermuteten Wahrscheinlichkeitsverteilung getestet werden, so ist der χ^2 -Anpassungstest der bekannteste Test.

$H_0 : (p_1, p_2, \dots, p_m)$ ist die Wahrscheinlichkeitsverteilung von m „Klassen“.

Um H_0 zu testen, wird eine Stichprobe vom Umfang n gezogen.

O_1, O_2, \dots, O_m sind die beobachteten („observed“) absoluten Häufigkeiten.

$$(1.19) \quad \sum_{i=1}^m O_i = n$$

Die O_i werden mit den, unter H_0 zu erwartenden („expected“) Häufigkeiten

$$(1.20) \quad E_i = n \cdot p_i$$

in folgendem Ausdruck verglichen:

$$(1.21) \quad T = \sum_{i=1}^m \frac{(O_i - E_i)^2}{E_i}$$

Sind alle $n \cdot p_i \geq 5$, so folgt T näherungsweise der so genannten χ^2 -Verteilung mit

$$(1.22) \quad f = m - 1$$

Freiheitsgraden.

Tabelle der χ^2 -Verteilung auf Seite 15. Ablesebeispiel:

$$f = 8 ; 1 - \alpha = 0,975$$

$$\chi_{8;0,975}^2 = 17,53$$

Zum Signifikanzniveau α wird H_0 abgelehnt, wenn

$$T > \chi_{f;1-\alpha}^2 \text{ (aus Tabelle).}$$

Andernfalls wird H_0 nicht abgelehnt.

α = Wahrscheinlichkeit, den Fehler 1.Art zu begehen (H_0 abzulehnen, obwohl H_0 wahr ist)

Beispiel

Bei einem Spaltungsversuch in der Biologie werden 3 Phänotypen im Verhältnis 1 : 2 : 1 erwartet.

Experimentell ermittelt wurde das Verhältnis 14 : 50 : 16 .

Ist damit die Vermutung widerlegt?

$$H_0 : \text{Biologische Theorie} \quad p_1 = \frac{1}{4} \quad p_2 = \frac{2}{4} \quad p_3 = \frac{1}{4}$$

Bei $n = 80$ Versuchen (14 + 50 + 16) wären die Häufigkeiten

$$E_1 = 20 \quad E_2 = 40 \quad E_3 = 20$$

zu erwarten. Widersprechen die Beobachtungen

$$O_1 = 14 \quad O_2 = 50 \quad O_3 = 16$$

auf signifikante Weise der Vermutung?

$$\begin{aligned} T &= \sum_{i=1}^3 \frac{(O_i - E_i)^2}{E_i} \\ &= \frac{(14 - 20)^2}{20} + \frac{(50 - 40)^2}{40} + \frac{(16 - 20)^2}{20} \end{aligned}$$

$$= \frac{36}{20} + \frac{100}{40} + \frac{16}{20} = 5,1$$

$$\alpha = 10\% \quad \chi_{2;0,90}^2 = 4,61 \Rightarrow H_0 \text{ ablehnen}$$

1:2:1 falsch

$$\alpha = 5\% \quad \chi_{2;0,95}^2 = 5,99 \Rightarrow H_0 \text{ nicht ablehnen}$$

1:2:1 könnte richtig sein

Beispiel

Ein homogener Teil eines Sees wird in 120 Parzellen unterteilt. Es wird gezählt, wie viele Teichmuscheln es pro Parzelle gibt.

Die Daten sind:

Teichmuscheln	0	1	2	3	≥ 4
Anzahl der Parzellen	57	39	13	8	3



zu lesen: In 57 von 120 Parzellen findet der Taucher keine Teichmuschel.

Ist die Zahl der Teichmuscheln pro Parzelle POISSONverteilt mit Mittelwert 1?

$$P(X=0) = e^{-1} \cdot \frac{1^0}{0!} = 0,368 \quad \Rightarrow E_1 = 120 \cdot 0,368 = 44,16$$

$$P(X=1) = e^{-1} \cdot \frac{1^1}{1!} = 0,368 \quad \Rightarrow E_2 = 120 \cdot 0,368 = 44,16$$

$$P(X=2) = e^{-1} \cdot \frac{1^2}{2!} = 0,184 \quad \Rightarrow E_3 = 120 \cdot 0,184 = 22,08$$

$$P(X=3) = e^{-1} \cdot \frac{1^3}{3!} = 0,061 \quad \Rightarrow E_4 = 120 \cdot 0,061 = 7,32$$

$$\underline{\underline{\sum = 0,981}}$$

$$P(X \geq 4) = 1 - 0,981 = 0,019 \quad \Rightarrow E_5 = 120 \cdot 0,019 = 2,28$$

< 5!!!

Neue Einteilung

Teichmuscheln	0	1	2	≥ 3
Anzahl der Parzellen (O_i)	57	39	13	11
neue E_i	44,16	44,16	22,08	9,60

H_0 : Zahl der Teichmuscheln pro Parzelle ist POISSONverteilt mit $\lambda = 1$.

$$T = \frac{(57 - 44,16)^2}{44,16} + \frac{(39 - 44,16)^2}{44,16} + \frac{(13 - 22,08)^2}{22,08} + \frac{(11 - 9,60)^2}{9,60} = 8,274$$

$$f = m - 1 = 4 - 1 = 3$$

$$\alpha = 0,05 \quad \Rightarrow \chi_{3;0,95}^2 = 7,81$$

Da $8,274 > 7,81$ ist, ist H_0 abzulehnen.

$$\alpha = 0,025 \quad \Rightarrow \chi_{3;0,975}^2 = 9,35$$

Da $8,274 < 9,35$ ist, ist H_0 nicht abzulehnen.

Müssen bei der Wahrscheinlichkeitsverteilung r unbekannte Parameter geschätzt werden, so ist die Anzahl der Freiheitsgrade nur noch:

$$(1.23) \quad f = m - 1 - r$$

Beispiel

H_0 : Zahl der Teichmuscheln pro Parzelle ist POISSONverteilt.

Da λ in H_0 nicht explizit aufgeführt wird, muss es geschätzt werden.

Punktschätzung von λ :

$$\hat{\lambda} = \bar{x} = \frac{0 \cdot 57 + 1 \cdot 39 + 2 \cdot 13 + 3 \cdot 8 + 4 \cdot 3}{120} = 0,84$$

Da λ geschätzt wird, „kostet“ es uns $r = 1$ Freiheitsgrade.

Teichmuscheln pro Parzelle	0	1	2	≥ 3
O_i	57	39	13	11
$p_i = e^{-0,84} \cdot \frac{0,84^i}{i!}$	0,43	0,36	0,15	0,06
$E_i = 120 \cdot p_i$	51,6	43,2	18	7,2

$$T = \frac{(57 - 51,6)^2}{51,6} + \frac{(39 - 43,2)^2}{43,2} + \frac{(13 - 18)^2}{18} + \frac{(11 - 7,2)^2}{7,2} = 4,36$$

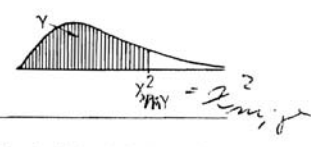
$$f = m - 1 - r = 4 - 1 - 1 = 2$$

$\alpha = 0,10 \Rightarrow \chi^2_{2;0,90} = 4,61$

Da $4,36 < 4,61$ wird H_0 nicht abgelehnt.

Für $\alpha < 0,10$ wird H_0 erst recht nicht abgelehnt.

➔ Zahl der Teichmuscheln pro Parzelle könnte POISSONverteilt sein.



Tab. 4:¹ Quantile $\chi^2_{m, \gamma}$ der χ^2 -Verteilung

$m \backslash \gamma$	0,995	0,990	0,975	0,950	0,900	0,750	0,500	0,250	0,100	0,050	0,025	0,010	0,005
1	7,879	6,635	5,034	3,841	2,706	1,323	0,455	0,102	0,0158	0,00393	0,000982	0,0004157	0,0002393
2	10,60	9,210	7,378	5,991	4,605	2,773	1,386	0,575	0,211	0,103	0,0506	0,0201	0,0100
3	12,84	11,34	9,348	7,815	6,251	4,108	2,366	1,213	0,584	0,352	0,216	0,115	0,0717
4	14,86	13,28	11,14	9,488	7,779	5,385	3,357	1,923	1,064	0,711	0,484	0,297	0,207
5	16,75	15,09	12,83	11,07	9,236	6,626	4,351	2,675	1,610	1,145	0,381	0,554	0,412
6	18,55	16,81	14,45	12,59	10,64	7,841	5,348	3,455	2,204	1,635	1,237	0,872	0,676
7	20,28	18,48	16,01	14,07	12,02	9,037	6,346	4,255	2,833	2,167	1,690	1,239	0,989
8	21,96	20,09	17,53	15,51	13,36	10,22	7,344	5,071	3,490	2,733	2,180	1,647	1,344
9	23,59	21,67	19,02	16,92	14,68	11,39	8,343	5,899	4,168	3,325	2,700	2,088	1,735
10	25,19	23,21	20,48	18,31	15,99	12,55	9,342	6,737	4,865	3,940	3,247	2,558	2,156
11	26,76	24,73	21,92	19,68	17,28	13,70	10,34	7,584	5,578	4,575	3,816	3,053	2,603
12	28,30	26,22	23,34	21,03	18,55	14,85	11,34	8,438	6,304	5,226	4,404	3,571	3,074
13	29,82	27,69	24,74	22,36	19,81	15,98	12,34	9,299	7,042	5,892	5,009	4,107	3,565
14	31,32	29,14	26,12	23,68	21,06	17,12	13,34	10,17	7,790	6,571	5,629	4,660	4,075
15	32,80	30,58	27,49	25,00	22,31	18,25	14,34	11,04	8,547	7,261	6,262	5,229	4,601
16	34,27	32,00	28,85	26,30	23,54	19,37	15,34	11,91	9,312	7,962	6,908	5,812	5,142
17	35,72	33,41	30,19	27,59	24,77	20,49	16,34	12,79	10,09	8,672	7,564	6,408	5,697
18	37,16	34,81	31,53	28,87	25,99	21,60	17,34	13,68	10,86	9,390	8,231	7,015	6,265
19	38,58	36,19	32,85	30,14	27,20	22,72	18,34	14,56	11,65	10,12	8,907	7,633	6,844
20	40,00	37,57	34,17	31,41	28,41	23,83	19,34	15,45	12,44	10,85	9,591	8,260	7,434
21	41,40	38,93	35,48	32,67	29,62	24,93	20,34	16,34	13,24	11,59	10,28	8,897	8,034
22	42,80	40,29	36,78	33,92	30,81	26,04	21,34	17,24	14,04	12,34	10,98	9,542	8,643
23	44,18	41,64	38,08	35,17	32,01	27,14	22,34	18,14	14,85	13,09	11,69	10,20	9,260
24	45,56	42,98	39,36	36,42	33,20	28,24	23,34	19,04	15,66	13,85	12,40	10,86	9,886
25	46,93	44,31	40,65	37,65	34,38	29,34	24,34	19,94	16,47	14,61	13,12	11,52	10,52
26	48,29	45,64	41,92	38,89	35,56	30,43	25,34	20,84	17,29	15,38	13,84	12,20	11,16
27	49,64	46,96	43,19	40,11	36,74	31,53	26,34	21,75	18,11	16,15	14,57	12,88	11,81
28	50,99	48,28	44,46	41,34	37,92	32,62	27,34	22,66	19,94	16,93	15,31	13,56	12,46
29	52,34	49,59	45,72	42,56	39,09	33,71	28,34	23,57	19,77	17,71	16,05	14,26	13,12
30	53,67	50,89	46,98	43,77	40,26	34,80	29,34	24,48	20,60	18,49	16,79	14,95	13,79
40	66,77	63,69	59,34	55,76	51,81	45,62	39,34	33,66	29,05	26,51	24,43	22,16	20,71
50	79,49	76,15	71,42	67,50	63,17	56,33	49,33	42,94	37,69	34,76	32,36	29,71	27,99
60	91,95	88,38	83,30	79,08	74,40	66,98	59,33	52,29	46,46	43,19	40,48	37,48	35,53
70	104,2	100,4	95,02	90,53	85,53	77,58	69,33	61,70	55,33	51,74	48,76	45,44	43,28
80	116,3	112,3	106,6	101,9	96,58	88,13	79,33	71,14	64,28	60,39	57,15	53,54	51,17
90	128,3	124,1	118,1	113,1	107,6	98,65	89,33	80,62	73,29	69,13	65,65	61,75	59,20
100	140,2	135,8	129,6	124,3	118,5	109,1	99,33	90,13	82,36	77,93	74,22	70,06	67,33
150	198,4	193,2	185,8	179,6	172,6	161,3	149,3	138,0	128,3	122,7	118,0	112,7	109,1
200	255,3	249,4	241,1	234,0	226,0	213,1	199,3	186,2	174,8	168,3	162,7	156,4	152,2
250	311,3	304,9	295,7	287,9	279,1	264,7	249,3	234,6	221,8	214,4	208,1	200,9	196,2
300	366,8	359,9	349,9	341,4	331,8	316,1	299,3	283,1	269,1	260,9	253,9	246,0	240,7
400	476,6	468,7	457,3	447,6	436,6	418,7	399,3	380,6	364,2	354,6	346,5	337,2	330,9
600	693,0	683,5	669,8	658,1	644,8	623,0	599,3	576,3	556,1	544,2	534,0	522,4	514,5
800	906,8	896,0	880,3	866,9	851,7	826,6	799,3	772,7	749,2	735,4	723,5	709,9	700,7
1000	1119,	1107,	1090,	1075,	1058,	1030,	999,3	969,5	943,1	927,6	914,3	898,9	888,6

Tabelle 1-1 Chi²-Verteilung

2. Auswertung von metrischen Daten bei einer Stichprobe

Daten: x_1, x_2, \dots, x_n in ungeordneter (chronologischer) Reihenfolge (**Urliste**)

$x_{(1)}, x_{(2)}, \dots, x_{(n)}$ in geordneter Reihenfolge: $x_{(1)} \leq x_{(2)} \leq \dots \leq x_{(n)}$ (**sortiert nach Größe**)

Beispiel „Steel ball bearings“

First line					
	1.18	1.42	0.69	0.88	1.62
	1.09	1.53	1.02	1.19	1.32

2.1 Maßzahlen für die Lage

2.1.1 Arithmetisches Mittel (Average, Mean)

$$(2.1) \quad \bar{x} = \frac{x_1 + x_2 + \dots + x_n}{n} = \frac{1}{n} \sum_{i=1}^n x_i = \frac{\sum_{i=1}^n x_i}{n}$$

$$\text{Bemerkung: } \sum_{i=1}^n (x_i - \bar{x}) = \sum_{i=1}^n x_i - n \cdot \bar{x} = \sum_{i=1}^n x_i - \sum_{i=1}^n x_i = 0$$

Beispiel „Steel ball bearings“

$$\bar{x} = \frac{11,94}{10} = 1,194$$

2.1.2 Median (Zentralwert)

$$(2.2) \quad n \text{ ungerade} : \tilde{x} = x_{\left(\frac{n+1}{2}\right)}$$

$$(2.3) \quad n \text{ gerade} : \tilde{x} = \frac{x_{\left(\frac{n}{2}\right)} + x_{\left(\frac{n}{2}+1\right)}}{2}$$

Der Median gibt den Zentralwert einer Stichprobe an. Ist die Anzahl der Messungen (**n**) **ungerade**, entspricht der Zentralwert genau dem mittleren Element. D.h.: $n = 7 \rightarrow \tilde{x} = x_{(4)}$

Ist die Anzahl der Messungen (**n**) **gerade**, entspricht der Zentralwert dem Mittelwert der beiden in der Mitte platzierten Messungen. D.h.: $n = 8 \rightarrow \tilde{x} = \frac{x_{(4)} + x_{(5)}}{2}$

Beispiel „Steel ball bearings“

$$\tilde{x} = \frac{x_{(5)} + x_{(6)}}{2} = \frac{1,18 + 1,19}{2} = 1,185$$

2.1.3 α -Quantil (Quantile, Percentile)

$n \cdot \alpha$ ganzzahlig:

$$(2.4) \quad Q(\alpha) = \frac{x_{(n \cdot \alpha)} + x_{(n \cdot \alpha + 1)}}{2} \quad \text{mit } 0 < \alpha < 1$$

$n \cdot \alpha$ nicht ganzzahlig:

$$(2.5) \quad Q(\alpha) = x_{(k)}$$

wobei k die auf $n \cdot \alpha$ **nächst** folgende ganze Zahl ist

Besondere Quantile:	$Q(0,50)$ = Median (Zentralwert)	= 50%-Punkt
	$Q(0,25)$ = unteres Quartil	= 25%-Punkt
	$Q(0,75)$ = oberes Quartil	= 75%-Punkt

2.2 Maßzahlen für die Streuung

2.2.1 Spannweite (Range)

$$(2.6) \quad R = \text{maximum} - \text{minimum} = x_{(n)} - x_{(1)}$$

Beispiel „Steel ball bearings“

$$R = x_{(10)} - x_{(1)} = 1,62 - 0,69 = 0,93$$

2.2.2 Inter-Quartil-Spannweite

$$(2.7) \quad IQR = Q(0,75) - Q(0,25)$$

2.2.3 Varianz (Variance)

$$(2.8) \quad s^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2 = \frac{\sum_{i=1}^n x_i^2 - n \cdot (\bar{x})^2}{n-1}$$

Bemerkung: $n-1$ = Zahl der Freiheitsgrade

Beispiel „Steel ball bearings“

$$s^2 = \frac{1}{10-1} \sum_{i=1}^{10} (x_i - 1,194)^2 \approx 0,084$$

2.2.4 Standardabweichung (Standard Deviation)

$$(2.9) \quad s = +\sqrt{s^2} = \sqrt{\frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2}$$

Beispiel „Steel ball bearings“

$$s = \sqrt{0,084} \approx 0,29$$

2.2.5 Mittlere absolute Abweichung vom Mittelwert

$$(2.10) \quad MA_x = \frac{\sum_{i=1}^n |x_i - \bar{x}|}{n}$$

Beispiel „Steel ball bearings“

$$MA_x = \frac{0,504 + 0,314 + 0,174 + 0,104 + 0,014 + 0,004 + 0,126 + 0,226 + 0,336 + 0,426}{10} = 0,2228$$

2.2.6 Mittlere absolute Abweichung vom Median

$$(2.11) \quad MA_x = \frac{\sum_{i=1}^n |x_i - \tilde{x}|}{n}$$

Beispiel „Steel ball bearings“

$$MA_x = \frac{0,495 + 0,305 + 0,165 + 0,095 + 0,005 + 0,005 + 0,135 + 0,235 + 0,345 + 0,435}{10} = 0,222$$

2.2.7 Median absolute Abweichung (Median Deviation)

$$(2.12) \quad MAD = \text{Median} \left\{ |x_i - \tilde{x}|; i = 1, \dots, n \right\}$$

Beispiel „Steel ball bearings“

$|x_i - \tilde{x}|$ nach der Größe sortiert

0,005

0,005

0,095

0,135

0,165 ← Median

0,235

0,305

0,345

0,435

0,495

$$D = \frac{0,165 + 0,235}{2} = 0,2$$

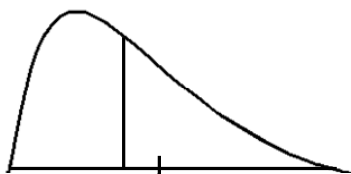
Anmerkung: Der Median geht nicht auf ungewöhnliche „Ausreißer“ ein.

2.3 Maßzahlen für die Schiefe

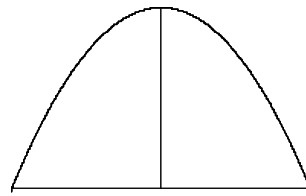
(Skewness)

$$(2.13) \quad \text{Schiefe } I = \frac{\bar{x} - \tilde{x}}{s}$$

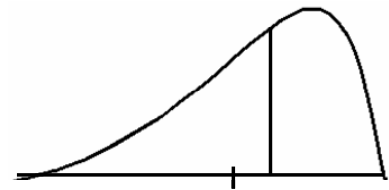
$$-1 \leq \text{Schiefe } I \leq +1$$



\tilde{x} \bar{x}
linkssteil (rechtsschief)
Schiefe $I > 0$



$\bar{x} = \tilde{x}$
Schiefe $I = 0$



\bar{x} \tilde{x}
rechtssteil (linksschief)
Schiefe $I < 0$

$$(2.14) \quad \text{Schiefe IIa} = \frac{Q(0,75) + Q(0,25) - 2 \cdot Q(0,50)}{Q(0,75) - Q(0,25)}$$

$$(2.15) \quad \text{Schiefe IIb} = \frac{Q(0,90) + Q(0,10) - 2 \cdot Q(0,50)}{Q(0,90) - Q(0,10)}$$

$$(2.16) \quad \text{Schiefe III} = \frac{\frac{1}{n} \cdot \sum_{i=1}^n (x_i - \bar{x})^3}{s^3}$$

Beispiel aus D. Drain (Intel Corporation) Widerstandsmessungen (Daten auf der nächsten Seite)

$$\bar{x} = 13,1, \tilde{x} = 12,0; s = 3,8; Q(0,75) = 13,3; Q(0,25) = 11,1; Q(0,90) = x_{(30)} = 15,6; Q(0,10) = x_{(4)} = 10,3$$

$$\text{Schiefe I} = \frac{13,1 - 12}{3,8} = 0,29$$

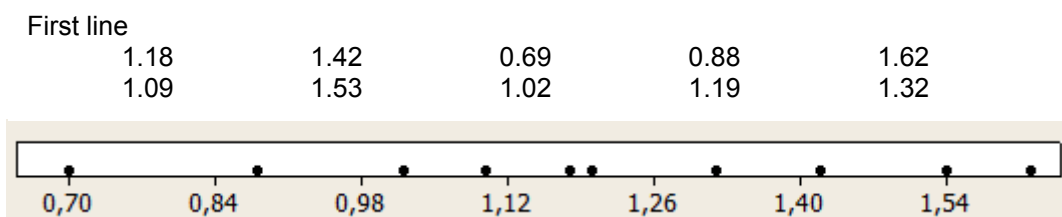
$$\text{Schiefe IIa} = \frac{13,3 + 11,1 - 2 \cdot 12}{13,3 - 11,1} = 0,18$$

$$\text{Schiefe IIb} = \frac{15,6 + 10,3 - 2 \cdot 12}{15,6 - 10,3} = 0,36$$

2.4 Dot Plot

Auf der x-Achse wird zu jedem Wert der Stichprobe ein Symbol (z.B. Punkt) geplottet. Kommt ein Wert mehrfach vor, werden die Symbole übereinander gestapelt.

Beispiel „Steel ball bearings“



2.5 Histogramm

Der Messwertbereich wird in eine gewisse Anzahl von gleichlangen Intervallen (Klassen) unterteilt und über jede Klasse ein Rechteck gezeichnet, dessen Höhe proportional zur (relativen) Häufigkeit der Werte ist, die in dieser Klasse liegen.

Probleme: Je nach Wahl der **Klassenbreite** und **Klassenlagen (Klassenmitte)** können sehr unterschiedliche Darstellungen entstehen.

Richtige Wahl der Klassenbreite: Es gibt Faustregeln (u.a. vom DIN), aus der Steinzeit der Statistik, aber auch wissenschaftliche Lösungen z.B. von Freedman u. Diaconis:

$$(2.17) \quad \text{optimale Klassenbreite} = 2 \cdot \frac{(Q(0,75) - Q(0,25))}{\sqrt[3]{n}}$$

mit n Zahl der Messwerte

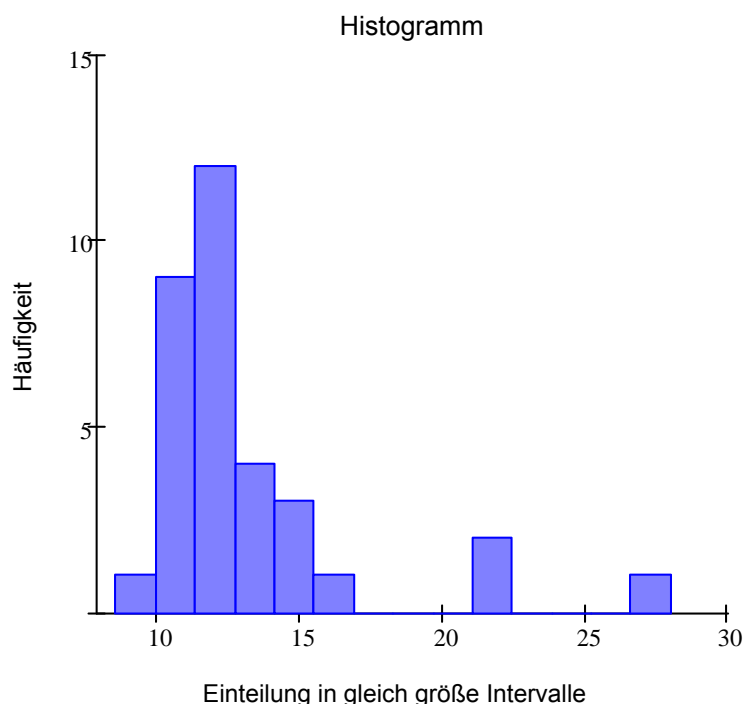
Richtige Wahl der Klassenlagen (Klassenmitte): Es gibt weder eine Faustregel noch wissenschaftliche Ansätze.

Beispiel aus D. Drain (Intel Corporation) Widerstandsmessungen

Order i	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
RHO	8,6	10,2	10,3	10,3	10,6	10,8	10,8	10,9	11,1	11,3	11,5	11,5	11,7	11,7	11,8	11,8
$\frac{i}{31}(\%)$	3,2	6,5	9,7	12,9	16,1	19,4	22,6	25,8	29,0	32,2	35,5	38,7	41,9	45,2	48,4	51,6
17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33
12,0	12,4	12,5	12,5	12,5	12,6	13,0	13,1	13,3	13,7	14,4	14,6	15,5	15,6	21,2	21,6	28,0
54,8	58,1	61,3	64,5	67,7	71,0	74,2	77,4	80,6	83,9	87,1	90,3	93,5	96,8	-	-	-

$$\bar{x} = 13,1 \quad Q(0,50) = 12,0 \quad Q(0,25) = 11,1 \quad Q(0,75) = 13,3 \quad R = 19,4 \quad s = 3,8$$

$$\text{Klassenbreite} = 2 \cdot \frac{(Q(0,75) - Q(0,25))}{\sqrt[3]{n}} = 2 \cdot \frac{(13,3 - 11,1)}{\sqrt[3]{33}} \approx 1,4$$



2.6 Stamm- und Blatt- Darstellung

(Stem-and-Leaf-Plot)

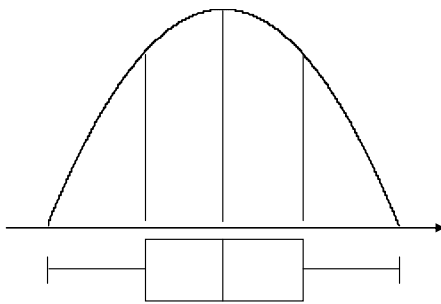
Die führenden Ziffern bilden den Stamm, die letzte Ziffer das Blatt. Die Blätter eines Stammes werden in aufsteigender Folge aufgereiht.

Interpretation:

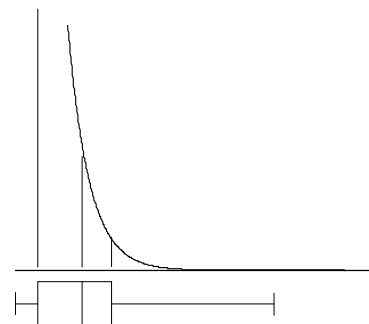
- | | |
|------------------------|---|
| a) Mittelwert | $Q(0,50) = \text{Median}$ |
| b) Streuung | Die Kastenlänge umfasst die zentralen 50%;
Box & Whisker die Streubreite des ganzen,
ausreißerbereinigten Datensatzes.
Die beiden Whisker und Kastenteile vierteln
(ausreißerbereinigt) den ganzen Datensatz. |
| c) Schiefe / Symmetrie | Perfekte Symmetrie liegt vor, wenn beiden Kastenteile gleich
breit und Whisker gleich lang sind. Je mehr davon
abgewichen wird, desto „schiefer“ ist die Verteilung. |
| d) Ausreißer | Werte, die „nicht normal“ sind. |

Achtung: Aus dem Box-Plot geht nicht hervor, ob die Verteilung ein- oder mehrgipflig ist. Dazu vielleicht ein Histogramm...

zu c) Schiefe / Symmetrie



Symmetrisch
gleich große Whisker
gleich große Kastenteile



unsymmetrisch
verschieden lange Whisker
verschieden große Kastenteile

Beispiel aus D. Drain (Intel Corporation) Widerstandsmessungen

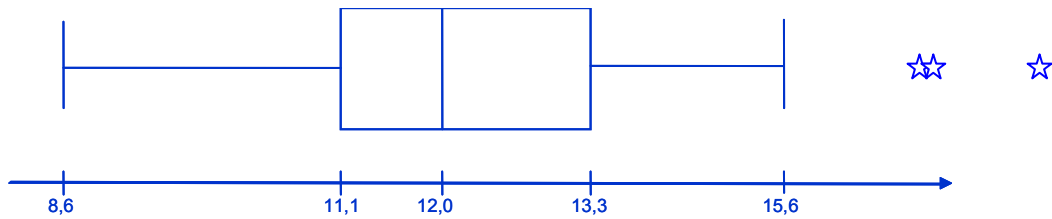
$$Q(0,50) = 12,0 \quad Q(0,25) = 11,1 \quad Q(0,75) = 13,3$$

$$\text{Unterer Whisker: } Q(0,25) - 1,5 \cdot (Q(0,75) - Q(0,25)) = 11,1 - 1,5 \cdot (13,3 - 11,1) = 7,8$$

→ Whisker geht bis 8,6 (der kleinste Wert der Stichprobe, welcher über 7,8 liegt)

$$\text{Obere Whisker: } Q(0,75) + 1,5 \cdot (Q(0,75) - Q(0,25)) = 13,3 + 1,5 \cdot (13,3 - 11,1) = 16,6$$

→ Whisker geht bis 15,6 (der größte Wert der Stichprobe, welcher nicht über 16,6 liegt)



2.8 Summenhäufigkeitsfunktion / empirische Verteilungsfunktion

Die Summenhäufigkeitsfunktion oder die empirische Verteilungsfunktion ist definiert durch:

$$(2.18) \quad \hat{F}_n(x) = \frac{1}{n} \cdot \text{Anzahl} \{i \mid x_{(i)} \leq x\}$$

Die grafische Darstellung ergibt eine TREPPE, die für $x \in [x_{(k)}; x_{(k+1)}[$ konstant ist und an den

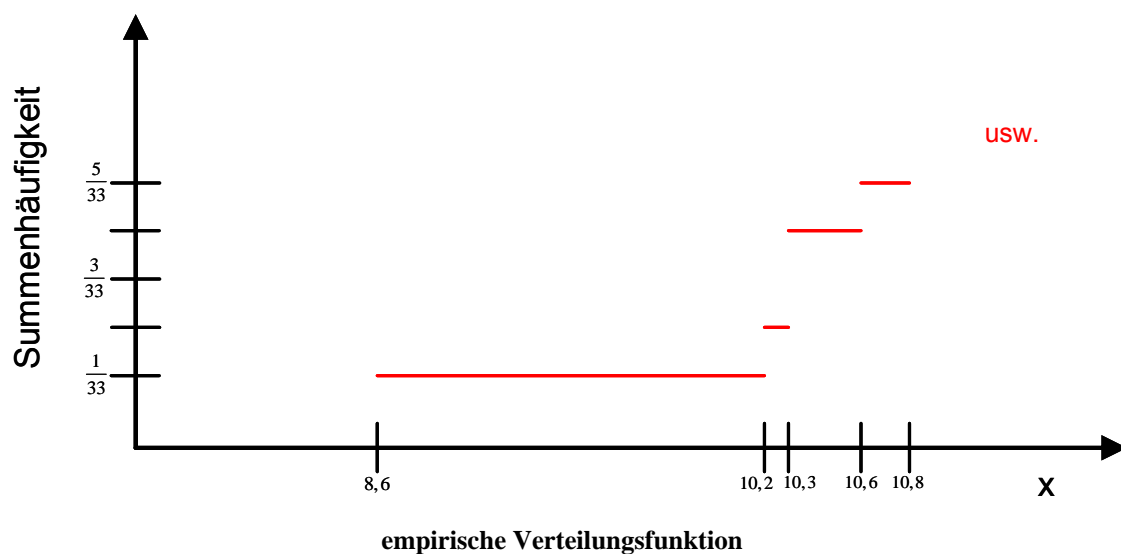
Stellen $x_{(k+1)}$ um deren Häufigkeit in die Höhe springt: $\hat{F}_n(x) = \frac{k}{n}$ für $x \in [x_{(k)}; x_{(k+1)}[$

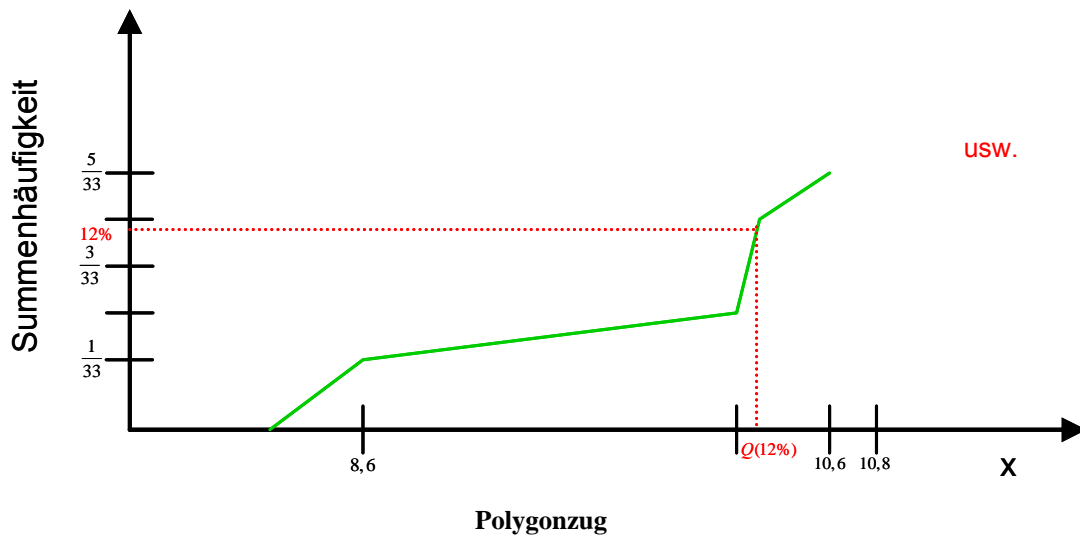
(Manchmal werden die „Stufenanfänge“ durch einen **Polygonzug** miteinander verbunden)

→ sieht gut aus, ist aber in der Anwendung nicht korrekt!!!

$\hat{F}_n(x)$ spielt in der mathematischen Statistik eine große Rolle.

Auch Ingenieure verwenden sie gerne, allerdings in der Polygonversion, um damit Quantile zu bestimmen.





Das Ergebnis ist aber lediglich eine Näherung an die Quantile $Q(\alpha)$, wie wir sie definiert haben.

2.9 Eine Zufallsstichprobe

Eine **Zufallsstichprobe** (random sample) ist eine **rein zufällig**, aus einer festen Gesamtheit (population) oder einem festen Prozess, entnommene Stichprobe.

Positives Beispiel: Ziehen der Lottozahlen („7 aus 49“)

Negatives Beispiel: Die anwesenden 44 Studenten sind keine Zufallsstichprobe aus der Gesamtheit der immatrikulierten Studenten.

Die Entnahme einer kleinen Milchprobe an der Oberfläche einer Milchkanne ist keine Zufallsstichprobe.

Mit einer Zufallsstichprobe will man die Repräsentativität einer Stichprobe für die zugehörige Gesamtheit erreichen.

„Zufall ist blind“

Wie gewährleistet man eine Zufallsauswahl?

- Gesamtheit **endlich** (auch große)
Jedes Element der Gesamtheit muss die gleiche Chance haben, in die Stichprobe zu gelangen.
→ Hilfsmittel: Zufallsgenerator
- Gesamtheit **unendlich** / Prozeß

Die Messwerte sollen statistisch unabhängig sein, d.h. der Wert einer Messung beeinflusst in kein(st)er Weise den Wert einer anderen Messung.

→ Hilfsmittel: Fachwissen des Messenden

Die etwas anspruchsvolleren mathematischen Verfahren setzen eine Zufallsstichprobe voraus.

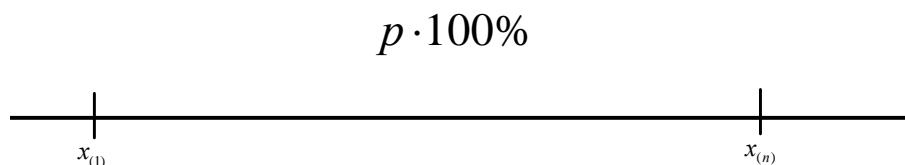
In der Praxis sind die hehren Prinzipien nicht leicht, manchmal gar nicht realisierbar.

Auch wenn keine Zufallsstichprobe vorliegt, tut man so, „als ob“ bzw. man ist sich der Tatsache gar nicht bewusst, dass eine Zufallsstichprobe bei der Anwendung des Verfahrens verlangt wird.

2.10 Statistische Anteilsbereiche

2.10.1 Zweiseitiger Anteilsbereich

In einer Zufallsstichprobe liegen alle (d.h. 100%) Werte mit 100%-iger Sicherheit zwischen dem kleinsten Wert $x_{(1)}$ und dem größten Wert $x_{(n)}$.



Von der Gesamtheit / dem Prozess werden es aber nur $p \cdot 100\%$ sein ($0 \leq p \leq 1$).

Da es sich um eine zufällige Stichprobe handelt, kann man das aber nicht mit 100%-iger Sicherheit sagen, sondern nur mit dem Vertrauen (confidence) von $\gamma \cdot 100\%$ ($0 \leq \gamma \leq 1$).

Liegt eine Zufallsstichprobe (!) vom Umfang n vor, so liegt zwischen dem kleinsten und größten Stichprobenwert $x_{(1)}$ bzw. $x_{(n)}$ mindestens ein Anteil p der Gesamtheit mit einem Vertrauen von γ , gemäß der Formel von **WILKS**.

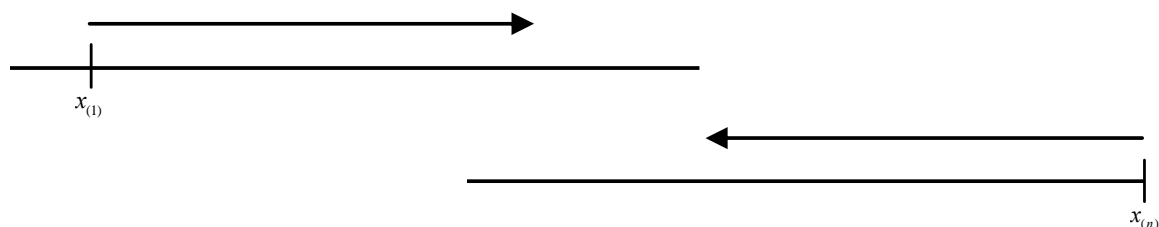
$$(2.19) \quad n \cdot p^{n-1} - (n-1) \cdot p^n = 1 - \gamma$$

Auflösen nach p oder n ist nicht möglich!

Folgende kleine Tabelle soll helfen, eine Vorstellung von möglichen Werten zu erhalten. Weitere Werte mit dem Taschenrechner **ausprobieren**.

$p \backslash \gamma$	0,95	0,90	0,70	0,50
0,99	473	388	244	168
0,95	93	77	49	34
0,90	46	38	24	17
0,85	30	25	16	11

2.10.2 Einseitiger Anteilbereich



Es liegt mindestens ein Anteil $p \cdot 100\%$ ($0 \leq p \leq 1$) der Gesamtheit mit einem Vertrauen von $\gamma \cdot 100\%$ ($0 \leq \gamma \leq 1$) unter dem Stichprobenwert $x_{(n)}$.

Es liegt mindestens ein Anteil $p \cdot 100\%$ ($0 \leq p \leq 1$) der Gesamtheit mit einem Vertrauen von $\gamma \cdot 100\%$ ($0 \leq \gamma \leq 1$) über dem Stichprobenwert $x_{(1)}$.

→ gemäß der Formel von Wilks.

$$(2.20) \quad p^n = 1 - \gamma \quad \Leftrightarrow \quad n = \frac{\ln(1 - \gamma)}{\ln(p)} \quad \Leftrightarrow \quad p = \sqrt[n]{1 - \gamma}$$

Beispiel aus D. Drain (Intel Corporation) Widerstandsmessungen

a) Einseitiger Anteilbereich

Angenommen, die 33 Widerstandsmessungen sind eine Zufallsstichprobe; Welches Vertrauen kann man haben, dass mindestens 90% aller Werte über dem kleinsten Stichprobenwert 8,6 liegen?

$$p^n = 1 - \gamma$$

$$\gamma = 1 - p^n = 1 - 0,9^{33} = 0,969 \approx 97\%$$

b) Zweiseitiger Anteilbereich

Wie groß muss der Umfang einer Zufallsstichprobe sein, damit mindestens 92% der Werte der Gesamtheit mit einem Vertrauen von 90% zwischen dem kleinsten & größten Wert der Stichprobe liegen werden?

$$n \cdot p^{n-1} - (n-1) \cdot p^n = 1 - \gamma$$

$$p = 0,92 \quad \gamma = 0,9 \quad n = ?$$

$$\text{Tabelle: } p = 0,9 \quad \gamma = 0,9 \Rightarrow n = 38$$

Durch probieren ein genaueres Ergebnis erzielen:

$$n = 50$$

$$50 \cdot 0,92^{49} - 49 \cdot 0,92^{50} = 0,0827$$

$$n = 45$$

$$45 \cdot 0,92^{44} - 44 \cdot 0,92^{45} = 0,1153$$

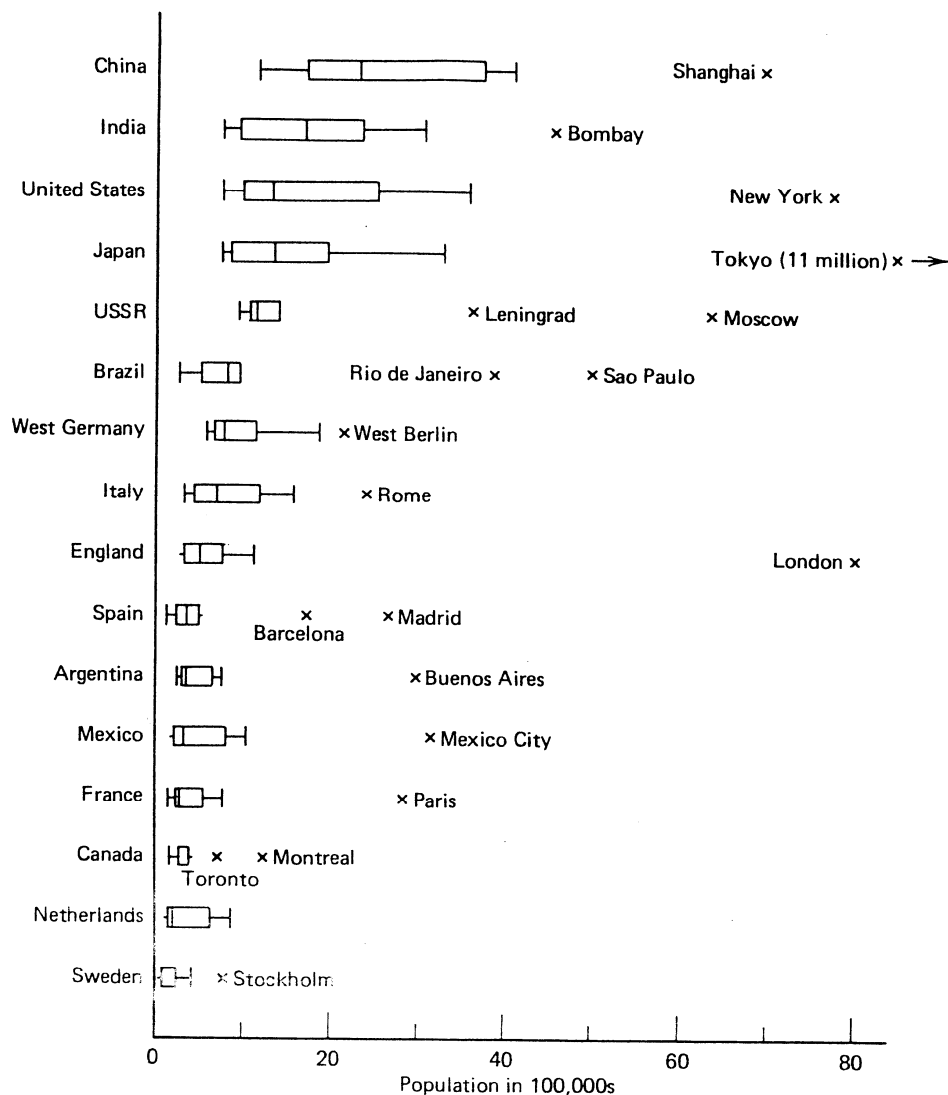
$$n = 47$$

$$47 \cdot 0,92^{46} - 46 \cdot 0,92^{47} = \mathbf{0,1010}$$

3. Auswertung und Vergleich von metrischen Daten bei zwei oder mehreren Stichproben

Man stellt die Stichproben **grafisch** gegenüber und stellt Vergleiche an.

Beispiel: 16 Staaten, deren 10 größten Städte.



Zusätzlich (!) kann man auch noch Maßzahlen angeben.

3.1 Kolmogorov-Smirnov-Test

Der Zwei-Stichproben-Test von Kolmogorov und Smirnov prüft, ob zwei vorliegende Zufallsstichproben aus derselben Grundgesamtheit stammen.

(Allgemein sind Signifikanz- oder Hypothesentests sehr beliebt in der statistischen Literatur. Am Beispiel des K-S-Tests soll auf ihre großen Tücken hingewiesen werden.)

Gegeben: Zwei Zufallsstichproben

$$(x_{11}, x_{12}, x_{13}, \dots, x_{1n}), \quad (\text{aus Grundgesamtheit 1})$$

$$(x_{21}, x_{22}, x_{23}, \dots, x_{2m}) \quad (\text{aus Grundgesamtheit 2})$$

die **unabhängig** sind.

Überprüft werden soll die Nullhypothese.

H_0 : Identische statistische Verteilung in Grundgesamtheit 1 und Grundgesamtheit 2.

Anhand der Stichprobendaten werden wir H_0 ablehnen, wenn die Daten deutlich („signifikant“) gegen die Hypothese sprechen; oder die Nullhypothese nicht ablehnen, wenn sie mit den Daten kompatibel ist.

Für jede Stichprobe wird ihre empirische Verteilungsfunktion („Treppe“) gezeichnet (vgl. Seite 24),

$$\hat{F}_n(x) \text{ bzw. } \hat{F}_m(x)$$

und dann die maximale absolute Differenz ermittelt.

$$(3.1) \quad \hat{D} = \max_x \left| \hat{F}_n(x) - \hat{F}_m(x) \right|$$

Ist \hat{D} „klein“, so sind sich die beiden Verteilungsfunktionen sehr nahe; es besteht kein Grund zu behaupten, dass die beiden dahinter stehenden Grundgesamtheiten verschieden sind.

Damit ist H_0 zwar nicht bewiesen, wir können es aber als Arbeitshypothese gelten lassen.

Ist \hat{D} „groß“, so besteht ein signifikanter Widerspruch zur Behauptung H_0 , dass die beiden Grundgesamtheiten die gleiche statistische Verteilung haben.

Wir sehen also verschiedene statistische Verteilungen.

Was ist aber die Schranke, welche „klein“ bzw. „groß“ festlegt? Statistiker legen hierzu ein

Signifikanzniveau α

zu Grunde, meist $\alpha = 0,05 = 5\%$.

α ist das Risiko, H_0 abzulehnen, obwohl H_0 richtig ist.

$1 - \alpha$ ist dann das Vertrauen / die Sicherheit, H_0 nicht abzulehnen, wenn H_0 richtig ist.

Den russischen Mathematikern Kolmogorov und Smirnov ist es 1939 gelungen, solche Schranken in Abhängigkeit von α , n und m zu finden.

α	0,20	0,15	0,10	0,05	0,01	0,001
$K(\alpha)$	1,07	1,14	1,22	1,36	1,63	1,95

und damit die Schranken:

$$(3.2) \quad D(\alpha) = K(\alpha) \cdot \sqrt{\frac{n+m}{n \cdot m}}$$

Ist $\hat{D} > D(\alpha)$, wird H_0 abgelehnt

Ist $\hat{D} \leq D(\alpha)$, wird H_0 nicht abgelehnt

Beispiel zweier unabhängiger Zufallsstichproben

Stichprobe 1	2,1	3,0	1,2	2,9	0,6	2,8	1,6	1,7	3,2	1,7
Stichprobe 2	3,2	3,8	2,1	7,2	2,3	3,5	3,0	3,1	4,6	3,2

Gefragt ist, ob die zwei Stichproben aus derselben Grundgesamtheit stammen könnten

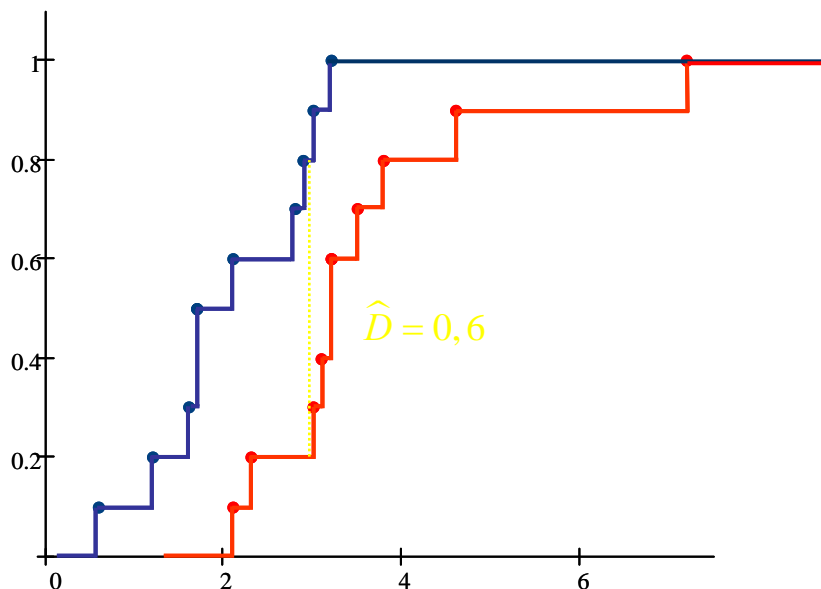
→ K-S-Test

Vorgehensweise

a) Daten ordnen

Stichprobe 1	0,6	1,2	1,6	1,7	1,7	2,1	2,8	2,9	3,0	3,2
Stichprobe 2	2,1	2,3	3,0	3,1	3,2	3,2	3,5	3,8	4,6	7,2

b) Verteilungsfunktion ermitteln



c) Bestimmung der maximalen vertikalen Differenz

$$\hat{D} = 0,6$$

d) Über H_0 entscheiden

$$\alpha = 10\% \quad K(\alpha) = 1,22 \quad D(\alpha) = 1,22 \cdot \sqrt{\frac{10+10}{100}} = 0,55$$

Da $0,6 > 0,55$, wird H_0 abgelehnt.

$$\alpha = 5\% \quad K(\alpha) = 1,36 \quad D(\alpha) = 1,36 \cdot \sqrt{\frac{10+10}{100}} = 0,61$$

Da $0,6 < 0,61$, wird H_0 nicht abgelehnt.

→ Grundgesamtheiten könnten also statistisch gleich sein.

$$\alpha = 1\% \quad K(\alpha) = 1,63 \quad D(\alpha) = 1,63 \cdot \sqrt{\frac{10+10}{100}} = 0,73$$

Da $0,6 < 0,73$, wird H_0 nicht abgelehnt.

3.2 Tücken von Signifikanztests (Hypothesentests)

- a) Wenn H_0 nicht abgelehnt wird, sagen manche, H_0 wird angenommen oder gar H_0 ist bewiesen.

Ein grandioser Irrtum! Die Richtigkeit von H_0 wird doch vorausgesetzt, wie kann dann H_0 akzeptiert oder bewiesen werden?

- b) Das Testergebnis hängt von der Wahl des Signifikanzniveaus α ab.

Es besteht die Versuchung, es so zu wählen, dass das Ergebnis passt.

- c) Manche sagen, dass sie mit 95 %iger Sicherheit die richtige Entscheidung getroffen haben (wenn $\alpha = 5\%$ gewählt wurde).

Absolut falsch! Die Entscheidung ist entweder zu 100 % richtig oder zu 100 % falsch.

α ist die Wahrscheinlichkeit, eine richtige Nullhypothese abzulehnen (Fehler 1. Art). Es gibt aber auch einen Fehler 2. Art, nämlich eine falsche Nullhypothese nicht abzulehnen. Über dieses Risiko wird geschwiegen (man kann es auch schlecht berechnen).

Wareneingangskontrolle: H_0 : Die Ware ist in Ordnung

	H_0 richtig	H_0 falsch
Ware akzeptiert	okay	Fehler 2. Art Konsumenterrisiko
Ware zurück gewiesen	Fehler 1. Art Produzenterrisiko	okay

- d) Tests werden durchgeführt, ohne zu berücksichtigen, dass die Stichproben zufällig und unabhängig sein müssen.

- e) Für große Stichprobenumfänge tendieren statistische Tests klar zu Ablehnung, für kleine zur Nichtablehnung.

Testentscheidung nur eine Frage von n !?

- f) Tests werden häufig unkritisch und mechanistisch durchgeführt. Gefahr einer Alibi- oder Pseudowissenschaft.

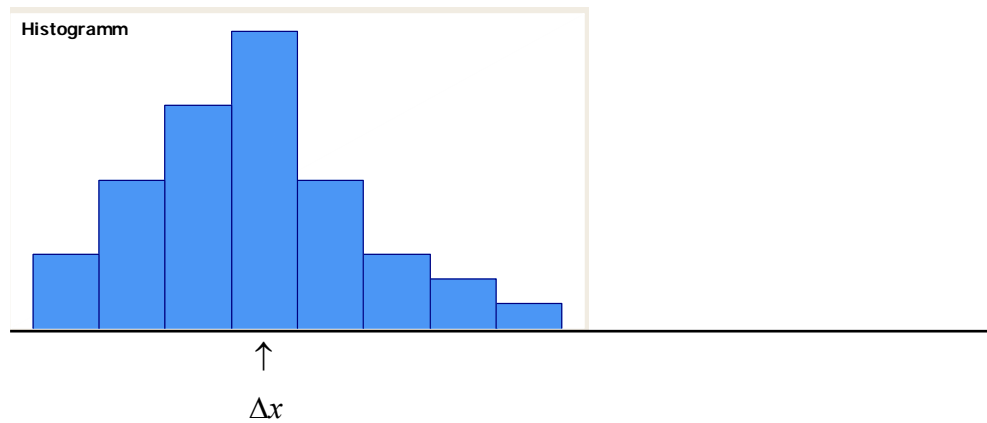
3.3 Zusammenfassung

Für viele, auch für Statistiker, scheinen Signifikanztests das „non-plus-ultra“ zu sein. → Man schaue in die Lehrbücher.

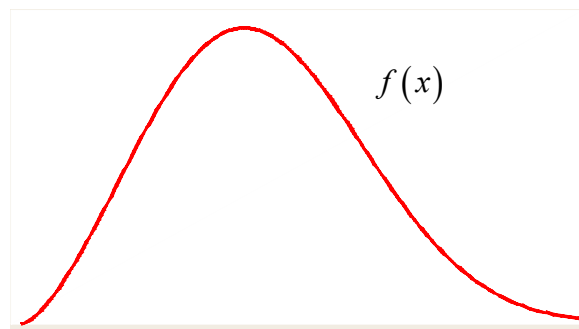
Mathematisch sind diese Tests wunderbar, ihr Nutzen für die Wirklichkeit und ihr Verstandenwerden seitens der Anwender kann angezweifelt werden.

4. Modellierung von metrischen Daten

4.1 Allgemeine Vorbemerkungen

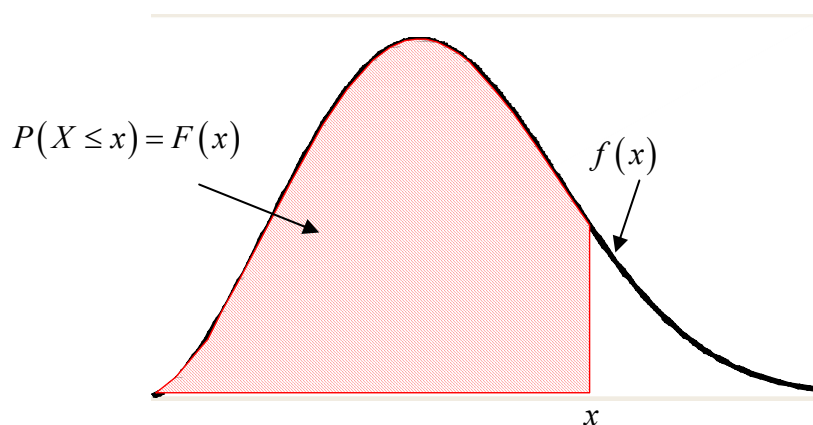


$$\left. \begin{array}{l} n \rightarrow \infty \\ \Delta x \rightarrow 0 \end{array} \right\} \Rightarrow$$



Dichtefunktion $f(x)$ der Zufallsvariablen X hat folgende Eigenschaften:

1. $f(x) \geq 0$
2. $\int_{-\infty}^{+\infty} f(x) dx = 1$



Verteilungsfunktion (Summenhäufigkeit $\hat{F}(x)$):

$$(4.1) \quad F(x) = \int_{-\infty}^x f(t) dt$$

$$(4.2) \quad P(a \leq X \leq b) = F(b) - F(a)$$

$$(4.3) \quad P(X \leq c) = F(c)$$

$$(4.4) \quad P(X \geq c) = 1 - F(c)$$

Erwartungswert (häufiges Symbol μ (Mittelwert)):

$$(4.5) \quad E(X) = \int_{-\infty}^{+\infty} x \cdot f(x) dx$$

Varianz (häufiges Symbol σ^2):

$$(4.6) \quad Var(X) = \int_{-\infty}^{+\infty} (x - \mu)^2 \cdot f(x) dx$$

Standardabweichung:

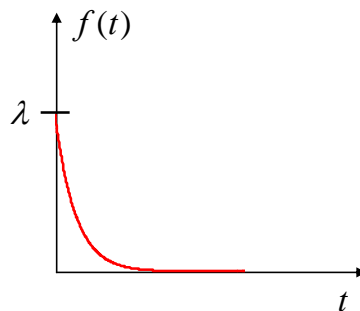
$$(4.7) \quad \sigma = +\sqrt{Var(X)}$$

4.2 Exponentialverteilung

4.2.1 Das Modell

Eine Zufallsvariable T heißt exponentialverteilt, wenn sie die Dichtefunktion hat:

$$(4.8) \quad f(t) = \lambda \cdot e^{-\lambda t} \quad \text{für } \begin{matrix} t \geq 0 \\ \lambda \in \mathbb{R} \end{matrix}$$



$$(4.9) \quad F(t) = 1 - e^{-\lambda t} \quad \text{für } \begin{matrix} t \geq 0 \\ \lambda \in \mathbb{R} \end{matrix}$$

$$(4.10) \quad E(T) = \frac{1}{\lambda}$$

$$(4.11) \quad Var(T) = \frac{1}{\lambda^2}$$

Median von $T = \tilde{x}$

$$P(T \leq \tilde{x}) = 0,5$$

$$F(\tilde{x}) = 0,5$$

$$1 - e^{-\lambda \cdot \tilde{x}} = 0,5$$

$$e^{-\lambda \cdot \tilde{x}} = 0,5$$

$$-\lambda \cdot \tilde{x} = \ln(0,5) = -0,693$$

$$(4.12) \quad \tilde{x} = \frac{0,693}{\lambda} \quad \text{"Halbwertszeit"}$$

Einsatz

- Physik (Atomzerfall radioaktiver Stoffe)
- Wartezeiten
- Bedienzeiten / Sprechzeiten
- Lebensdauer

4.2.2 Erkennen der Verteilung und Schätzungen des Parameters

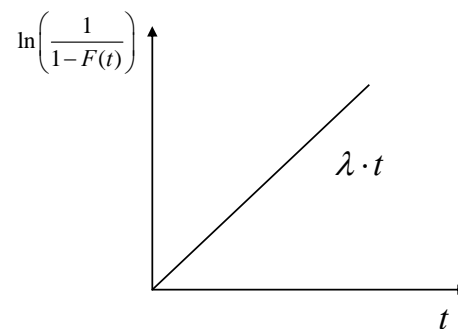
$$F(t) = 1 - e^{-\lambda \cdot t}$$

$$e^{-\lambda \cdot t} = 1 - F(t)$$

$$-\lambda \cdot t = \ln(1 - F(t))$$

$$\lambda \cdot t = -\ln(1 - F(t))$$

$$\lambda \cdot t = \ln\left(\frac{1}{1 - F(t)}\right)$$



$$\text{Mit } F(t_{(i)}) = \frac{i}{n+1}$$

$$\text{ist } \frac{1}{1 - F(t_{(i)})} = \frac{1}{1 - \frac{i}{n+1}} = \frac{1}{\frac{n+1-i}{n+1}} = \frac{n+1}{n+1-i}$$

$$\text{Plotte: } \left(t_{(i)}; \ln\left(\frac{n+1}{n+1-i}\right) \right)$$

Liegen diese Punkte „gut“ an einer Geraden, so liegt eine Exponentialverteilung vor mit λ = Steigung der Geraden.

Punktschätzung

Stichprobe t_1, t_2, \dots, t_n

Da $E(t) = \frac{1}{\lambda}$ und damit $\lambda = \frac{1}{E(t)}$ schätzen wir

$$(4.13) \quad \hat{\lambda} = \frac{1}{\frac{\sum_{i=1}^n t_i}{n}} = \frac{n}{\sum_{i=1}^n t_i}$$

Konfidenzintervall zum Vertrauen $1 - \alpha$

$$(4.14) \quad \left[\frac{\chi_{2n; \frac{\alpha}{2}}^2}{2 \cdot \sum_{i=1}^n t_i}; \frac{\chi_{2n; 1 - \frac{\alpha}{2}}^2}{2 \cdot \sum_{i=1}^n t_i} \right]$$

Beispiel

1. Reperaturzeiten von Autos in Minuten

10, 5, 40, 60, 30

$$\bar{x} = \frac{145}{5} = 29 \text{ min}$$

$$\text{Schätze } \hat{\lambda} = \frac{1}{29 \text{ min}} = \frac{2,06}{h}$$

$$\text{„Halbwertszeit“ (Median)} = \frac{0,693}{\hat{\lambda}} = \frac{0,693}{2,06} h = 0,336 = 20 \text{ min}$$

2. Operationen im Krankenhaus

$$\lambda = \frac{1}{4h} = \frac{0,25}{h} \quad (\bar{x} = 4h)$$

Wie viel Prozent der Operationen dauern länger als 7 Stunden?

$$\begin{aligned} P(T \geq 7) &= 1 - P(T \leq 7) \\ &= 1 - F(7) \\ &= 1 - \left(1 - e^{-\frac{1}{4} \cdot 7} \right) = e^{-1,75} \\ &= 0,17 = 17\% \end{aligned}$$

Und wie viel Prozent der Operationen dauern weniger als eine halbe Stunde?

$$\begin{aligned} P(T \leq 0,5) &= F(0,5) \\ &= 1 - e^{-\frac{1}{4} \cdot 0,5} \\ &= 0,12 = 12\% \end{aligned}$$

4.3 Normalverteilung

4.3.1 Das Modell



Eine „Zufallsvariable“ X heißt normalverteilt (Gauß-verteilt), wenn sie die Dichtefunktion hat:

$$(4.15) \quad f(x) = \frac{1}{\sqrt{2\pi} \cdot \sigma} \cdot e^{-\frac{1}{2} \left(\frac{x-\mu}{\sigma} \right)^2}$$

$$= \frac{1}{\sqrt{2\pi} \cdot \sigma} \cdot \exp \left(-\frac{1}{2} \left(\frac{x-\mu}{\sigma} \right)^2 \right)$$

$$-\infty < x < +\infty$$

$$\left. \begin{array}{l} \mu \in \mathbb{R} \\ \sigma > 0 \end{array} \right\} \text{konstante Parameter}$$

kleine Kurvendiskussion von $y = f(x)$

- Maximum bei $x = \mu$
- $x = \mu$ Symmetrieachse
- $D = \mathbb{R}$
- $\lim_{x \rightarrow \pm\infty} f(x) = 0$
- $\int_{-\infty}^{+\infty} f(x) dx = 1 = 100\%$
- Erwartungswert („Mittelwert“)

$$\int_{-\infty}^{+\infty} x \cdot f(x) dx = \mu$$

- Varianz

$$\int_{-\infty}^{+\infty} (x - \mu)^2 \cdot f(x) dx = \sigma^2$$

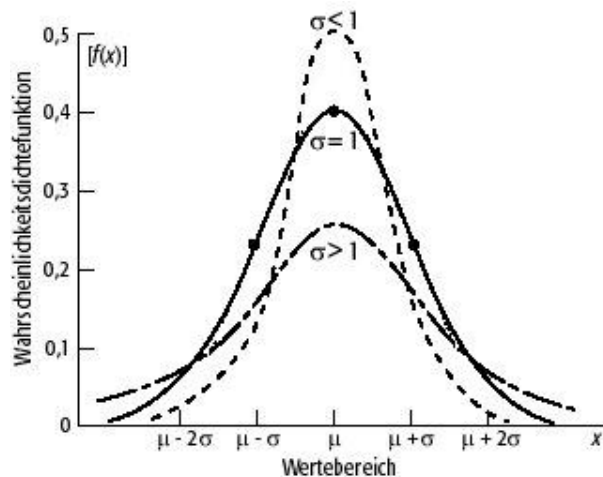
- Standardabweichung = σ

- Mittlere Abweichung vom Mittelwert

$$\rho = \int_{-\infty}^{+\infty} |x - \mu| \cdot f(x) dx = \sqrt{\frac{2}{\pi}} \cdot \sigma$$

$$\rho = 0,7979 \cdot \sigma \approx 0,8 \cdot \sigma$$

$$\sigma = 1,253 \cdot \rho$$



**kleines Sigma: kleine Streuung
großes Sigma: große Streuung**

Die Verteilungsfunktion $F(x)$ (Summenhäufigkeitsfunktion)

$$F(x) = \int_{-\infty}^x \frac{1}{\sqrt{2\pi} \cdot \sigma} \cdot \exp\left(-\frac{1}{2} \left(\frac{t - \mu}{\sigma}\right)^2\right) dt$$

gibt es nicht, da die Stammfunktion $F(x)$ nicht existiert. Dadurch wird der Umgang mit der Normalverteilung etwas umständlich.

Die Normalverteilung mit $\mu = 0$ und $\sigma = 1$ heißt Standardnormalverteilung, für die es eine Tabelle gibt. Jede Normalverteilung mit beliebigen μ und σ lässt sich in eine Standardnormalverteilung durch die Transformationsgleichung

$$(4.16) \quad Z = \frac{X - \mu}{\sigma}$$

überführen und auch wieder rücktransformieren:

$$(4.17) \quad X = Z \cdot \sigma + \mu$$

Es gilt:

$$(4.18) \quad \begin{aligned} F(x) &= P(X \leq x) = P\left(\frac{X - \mu}{\sigma} \leq \frac{x - \mu}{\sigma}\right) = P(Z \leq z) \\ &= \frac{1}{\sqrt{2\pi}} \cdot \int_{-\infty}^z e^{-\frac{1}{2}t^2} dt = \Phi(z) \quad \text{mit} \quad z = \frac{x - \mu}{\sigma} \end{aligned}$$

$\Phi(z)$ ist tabelliert.

Beispiele:

$$\Phi(0,23) = 0,59095$$

$$\Phi(2,57) = 0,99492$$

$$z > 4,09 \implies \Phi(z) \approx 1$$

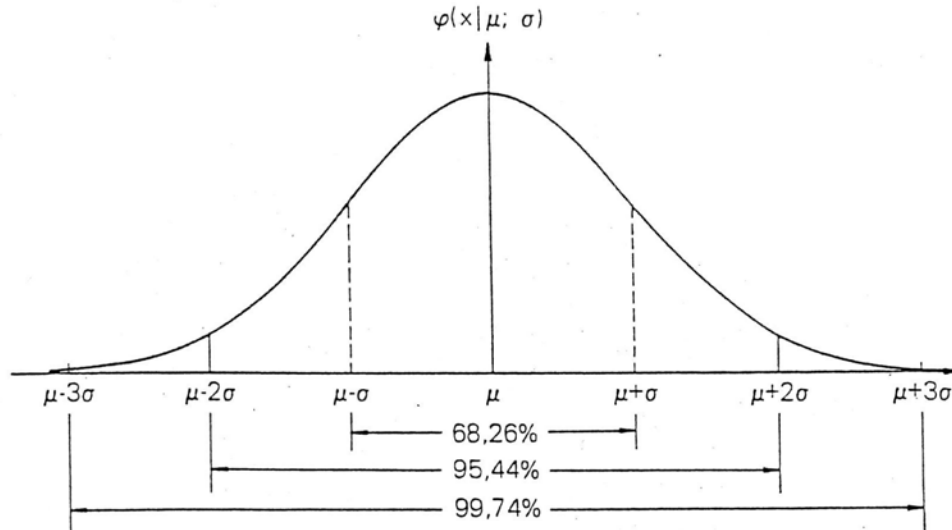
$$(4.19) \quad \Phi(-z) = 1 - \Phi(z)$$

Tabelle zur Standardnormalverteilung auf der nächsten Seite!

4. Modellierung von metrischen Daten

$\Phi(z) \rightarrow$

$z \setminus *$	0	1	2	3	4	5	6	7	8	9
0,0*	0,50000	0,50399	0,50798	0,51197	0,51595	0,51994	0,52392	0,52790	0,53188	0,53586
0,1*	0,53983	0,54380	0,54776	0,55172	0,55567	0,55962	0,56356	0,56749	0,57142	0,57535
0,2*	0,57926	0,58317	0,58706	0,59095	0,59483	0,59871	0,60257	0,60642	0,61026	0,61409
0,3*	0,61791	0,62172	0,62552	0,62930	0,63307	0,63683	0,64058	0,64431	0,64803	0,65173
0,4*	0,65542	0,65910	0,66276	0,66640	0,67003	0,67364	0,67724	0,68082	0,68439	0,68793
0,5*	0,69146	0,69497	0,69847	0,70194	0,70540	0,70884	0,71226	0,71566	0,71904	0,72240
0,6*	0,72575	0,72907	0,73237	0,73565	0,73891	0,74215	0,74537	0,74857	0,75175	0,75490
0,7*	0,75804	0,76115	0,76424	0,76730	0,77035	0,77337	0,77637	0,77935	0,78230	0,78524
0,8*	0,78814	0,79103	0,79389	0,79673	0,79955	0,80234	0,80511	0,80785	0,81057	0,81327
0,9*	0,81594	0,81859	0,82121	0,82381	0,82639	0,82894	0,83147	0,83398	0,83646	0,83891
1,0*	0,84134	0,84375	0,84614	0,84849	0,85083	0,85314	0,85543	0,85769	0,85993	0,86214
1,1*	0,86433	0,86650	0,86864	0,87076	0,87286	0,87493	0,87698	0,87900	0,88100	0,88298
1,2*	0,88493	0,88686	0,88877	0,89065	0,89251	0,89435	0,89617	0,89796	0,89973	0,90147
1,3*	0,90320	0,90490	0,90658	0,90824	0,90988	0,91149	0,91309	0,91466	0,91621	0,91774
1,4*	0,91924	0,92073	0,92220	0,92364	0,92507	0,92647	0,92785	0,92922	0,93056	0,93189
1,5*	0,93319	0,93448	0,93574	0,93699	0,93822	0,93943	0,94062	0,94179	0,94295	0,94408
1,6*	0,94520	0,94630	0,94738	0,94845	0,94950	0,95053	0,95154	0,95254	0,95352	0,95449
1,7*	0,95543	0,95637	0,95728	0,95818	0,95907	0,95994	0,96080	0,96164	0,96246	0,96327
1,8*	0,96407	0,96485	0,96562	0,96638	0,96712	0,96784	0,96856	0,96926	0,96995	0,97062
1,9*	0,97128	0,97193	0,97257	0,97320	0,97381	0,97441	0,97500	0,97558	0,97615	0,97670
2,0*	0,97725	0,97778	0,97831	0,97882	0,97932	0,97982	0,98030	0,98077	0,98124	0,98169
2,1*	0,98214	0,98257	0,98300	0,98341	0,98382	0,98422	0,98461	0,98500	0,98537	0,98574
2,2*	0,98610	0,98645	0,98679	0,98713	0,98745	0,98778	0,98809	0,98840	0,98870	0,98899
2,3*	0,98928	0,98956	0,98983	0,99010	0,99036	0,99061	0,99086	0,99111	0,99134	0,99158
2,4*	0,99180	0,99202	0,99224	0,99245	0,99266	0,99286	0,99305	0,99324	0,99343	0,99361
2,5*	0,99379	0,99396	0,99413	0,99430	0,99446	0,99461	0,99477	0,99492	0,99506	0,99520
2,6*	0,99534	0,99547	0,99560	0,99573	0,99585	0,99598	0,99609	0,99621	0,99632	0,99643
2,7*	0,99653	0,99664	0,99674	0,99683	0,99693	0,99702	0,99711	0,99720	0,99728	0,99736
2,8*	0,99744	0,99752	0,99760	0,99767	0,99774	0,99781	0,99788	0,99795	0,99801	0,99807
2,9*	0,99813	0,99819	0,99825	0,99831	0,99836	0,99841	0,99846	0,99851	0,99856	0,99861
3,0*	0,99865	0,99869	0,99874	0,99878	0,99882	0,99886	0,99889	0,99893	0,99896	0,99900
3,1*	0,99903	0,99906	0,99910	0,99913	0,99916	0,99918	0,99921	0,99924	0,99926	0,99929
3,2*	0,99931	0,99934	0,99936	0,99938	0,99940	0,99942	0,99944	0,99946	0,99948	0,99950
3,3*	0,99952	0,99953	0,99955	0,99957	0,99958	0,99960	0,99961	0,99962	0,99964	0,99965
3,4*	0,99966	0,99968	0,99969	0,99970	0,99971	0,99972	0,99973	0,99974	0,99975	0,99976
3,5*	0,99977	0,99978	0,99978	0,99979	0,99980	0,99981	0,99981	0,99982	0,99983	0,99983
3,6*	0,99984	0,99985	0,99985	0,99986	0,99986	0,99987	0,99987	0,99988	0,99988	0,99989
3,7*	0,99989	0,99990	0,99990	0,99990	0,99991	0,99991	0,99992	0,99992	0,99992	0,99992
3,8*	0,99993	0,99993	0,99993	0,99994	0,99994	0,99994	0,99994	0,99995	0,99995	0,99995
3,9*	0,99995	0,99995	0,99996	0,99996	0,99996	0,99996	0,99996	0,99996	0,99997	0,99997
4,0*	0,99997	0,99997	0,99997	0,99997	0,99997	0,99997	0,99998	0,99998	0,99998	0,99998



Für die Glockenkurve (und nur diese!) gilt:

Bereich	Anteil Werte innerhalb	Anteil Werte außerhalb
$\mu \pm 1 \times \sigma$	68,26 %	31,74 %
$\mu \pm 2 \times \sigma$	95,44 %	4,56 %
$\mu \pm 3 \times \sigma$	99,74 %	0,26 %
$\mu \pm 4 \times \sigma$	99,9937 %	63 ppm
$\mu \pm 5 \times \sigma$	99,999943 %	0,57 ppm
$\mu \pm 6 \times \sigma$	99,9999998 %	0,002 ppm = 2 ppb

Achtung: Die Glockenkurve geht nur asymptotisch (wie die e-Kurve) gegen die x-Achse. Unrealistisch!

$$P(\mu - k \cdot \sigma \leq X \leq \mu + k \cdot \sigma)$$

$$k \in \mathbb{R}$$

$$= P(-k \cdot \sigma \leq X - \mu \leq k \cdot \sigma)$$

$$= P\left(-k \leq \frac{X - \mu}{\sigma} \leq +k\right)$$

$$Z = \frac{X - \mu}{\sigma}$$

Standard-Normalverteilung

$$= \Phi(k) - \Phi(-k)$$

$$= \Phi(k) - [1 - \Phi(k)] = 2 \cdot \Phi(k) - 1$$

$$(4.20) \quad P(\mu - k \cdot \sigma \leq X \leq \mu + k \cdot \sigma) = 2 \cdot \Phi(k) - 1$$

Beispiel

$$k = 2$$

$$P(\mu - 2 \cdot \sigma \leq X \leq \mu + 2 \cdot \sigma) = 2 \cdot \Phi(2) - 1 = 2 \cdot 0,97725 - 1 = 0,9545 = 95,45\%$$

Andersrum: k gesucht

Was ist das 90 % Intervall?

$$P(\underbrace{\mu - k \cdot \sigma \leq X \leq \mu + k \cdot \sigma}_{2 \cdot \Phi(k) - 1}) = 0,9$$

$$2 \cdot \Phi(k) - 1 = 0,9$$

$$\Phi(k) = 0,95$$

$$k = 1,645 \text{ (vgl. Tabelle)}$$

„Die Mathematiker denken, die Physiker haben die Gültigkeit der Normalverteilung empirisch nachgewiesen. Die Physiker denken, die Mathematiker haben die Gültigkeit der Normalverteilung theoretisch nachgewiesen.“

Es gibt Signifikanztests für:

$$H_0 : \text{Es liegt eine Normalverteilung vor}$$

Die bekanntesten Tests sind

- Anderson-Darling
- d'Agostino
- Kolmogorov-Smirnov
- Shapiro-Wilk

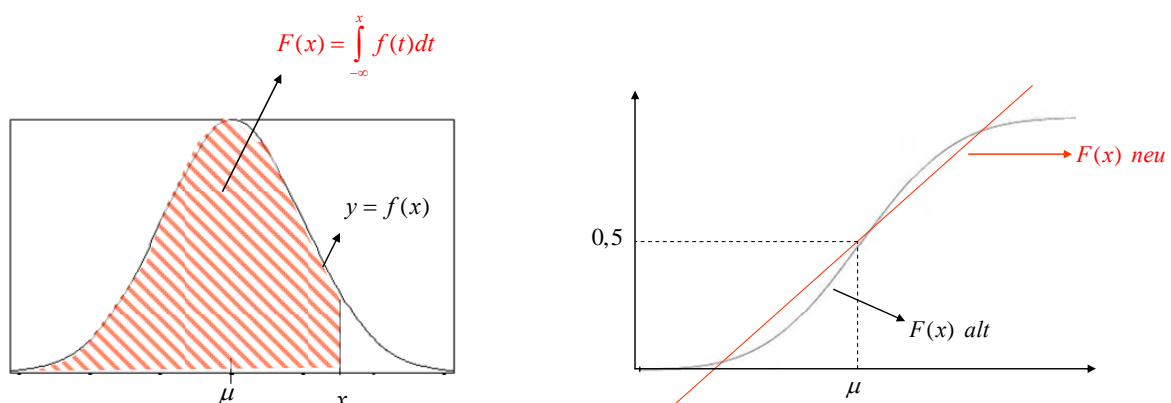
→ der letzte gilt als der Beste.

Allen diesen Tests gemeinsam ist das Problem, dass es bei kleinen Stichproben fast unmöglich ist, H_0 abzulehnen und dass bei großen Stichproben H_0 fast immer abgelehnt wird.

Deshalb beschäftigen wir uns nicht mit solchen Tests.

4.3.2 Erkennen der Verteilung und Schätzen der Parameter

Interessant für die Praxis ist hingegen das **Wahrscheinlichkeitsnetz** (-papier), *englisch: normal probability plot*. Damit lässt sich eine fundierte Einsicht gewinnen. (kein Test!)



4. Modellierung von metrischen Daten

Das Wahrscheinlichkeitsnetz ist ein formatiertes Blatt Papier mit unveränderter x-Achse und einer so veränderten y-Achse, dass $F(x)$ darin als Gerade erscheint.

Die Stammfunktion

$$F(x) = \int_{-\infty}^x f(t)dt$$

entspricht der empirischen Verteilungsfunktion

$$\hat{F}_n(x)$$

$\hat{F}_n(x)$ wird nicht als Treppe, sondern nur in Punktform gezeichnet: die Punkte

$$\left(x_{(i)}, \frac{i}{n+1} \right)$$

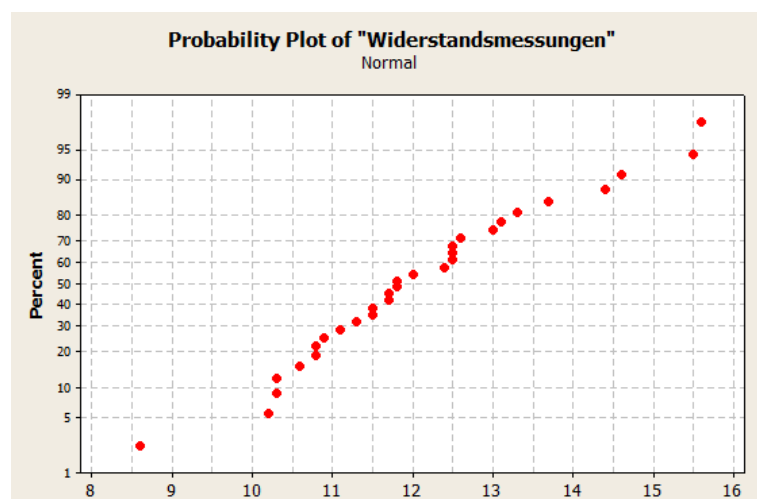
Lässt sich im Wahrscheinlichkeitsnetz durch diese „Punktwolke“ auf befriedigende Art und Weise eine Gerade legen, so ist die Normalverteilung eine plausible Arbeitshypothese, andernfalls nicht.

Beispiel

33 Widerstandsmessungen (3 Ausreißer, die wir rausnehmen) von Kapitel 2, S.21

Sind die anderen 30 Daten normalverteilt?

Order i	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
RHO	8,6	10,2	10,3	10,3	10,6	10,8	10,8	10,9	11,1	11,3	11,5	11,5	11,7	11,7	11,8	11,8
$\frac{i}{31}(\%)$	3,2	6,5	9,7	12,9	16,1	19,4	22,6	25,8	29,0	32,2	35,5	38,7	41,9	45,2	48,4	51,6
17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33
12,0	12,4	12,5	12,5	12,5	12,6	13,0	13,1	13,3	13,7	14,4	14,6	15,5	15,6	21,2	21,6	28,0
54,8	58,1	61,3	64,5	67,7	71,0	74,2	77,4	80,6	83,9	87,1	90,3	93,5	96,8	-	-	-



Selbst wenn man eine Normalverteilung nachweisen könnte (was prinzipiell unmöglich ist), wäre noch die Frage nach μ und σ .

Gegeben eine Zufallsstichprobe x_1, \dots, x_n mit dem Mittelwert:

$$(4.21) \quad \bar{x} = \frac{\sum_{i=1}^n x_i}{n}$$

und der Standardabweichung:

$$(4.22) \quad s = \sqrt{\frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n-1}}$$

<u>Punktschätzer</u> für μ	$\hat{\mu} = \bar{x}$
für σ^2	$\hat{\sigma}^2 = s^2$
für σ	$\hat{\sigma} = \frac{s}{a_n}$

n	a _n
2	0,798
3	0,887
4	0,922
5	0,940
6	0,951
7	0,960
8	0,965
9	0,969
10	0,973
≥11	≈1,000

Konfidenzintervalle zum Vertrauensniveau $1 - \alpha$ sind

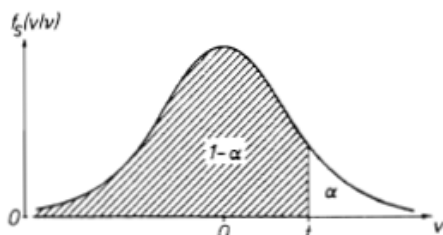
$$(4.23) \quad \text{für } \mu: \quad \bar{x} \pm t_{n-1; 1-\frac{\alpha}{2}} \cdot \frac{s}{\sqrt{n}}$$

t_{n-1} : Student-t-Verteilung mit $n-1$ Freiheitsgraden

$$(4.24) \quad \text{für } \sigma^2: \quad \left[\frac{(n-1) \cdot s^2}{\chi_{n-1; 1-\frac{\alpha}{2}}^2}, \frac{(n-1) \cdot s^2}{\chi_{n-1; \frac{\alpha}{2}}^2} \right]$$

(4.25) für σ :
$$\left[\sqrt{\frac{(n-1) \cdot s^2}{\chi^2_{n-1; 1-\frac{\alpha}{2}}}}; \sqrt{\frac{(n-1) \cdot s^2}{\chi^2_{n-1; \frac{\alpha}{2}}}} \right]$$

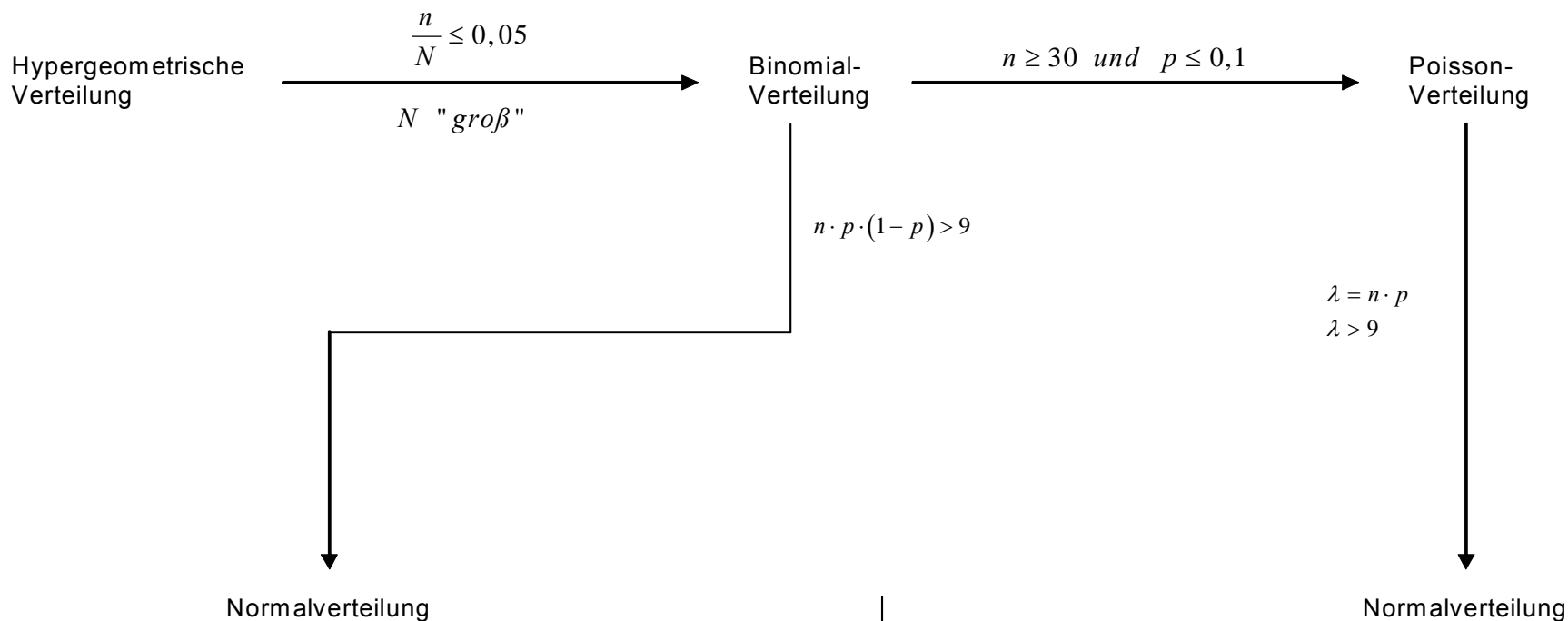
Studentverteilung – Werte von t zu gegebenen Werten der Verteilungsfunktion



Tabelliert sind die Werte t , für die $W(-\infty < T \leq t) = F_S(t/\nu) = 1 - \alpha$ gilt.

ν	$1-\alpha$									
	0.600	0.700	0.750	0.800	0.900	0.950	0.975	0.990	0.995	0.999
1	0.325	0.727	1.000	1.376	3.078	6.314	12.71	31.82	63.66	318.3
2	0.289	0.617	0.816	1.061	1.886	2.920	4.303	6.965	9.925	22.33
3	0.277	0.584	0.765	0.978	1.638	2.353	3.182	4.541	5.841	10.21
4	0.271	0.569	0.741	0.941	1.533	2.132	2.776	3.747	4.604	7.173
5	0.267	0.559	0.727	0.920	1.476	2.015	2.571	3.365	4.032	5.893
6	0.265	0.553	0.718	0.906	1.440	1.943	2.447	3.143	3.707	5.208
7	0.263	0.549	0.711	0.896	1.415	1.895	2.365	2.998	3.499	4.785
8	0.262	0.546	0.706	0.889	1.397	1.860	2.306	2.896	3.355	4.501
9	0.261	0.543	0.703	0.883	1.383	1.833	2.262	2.821	3.250	4.297
10	0.260	0.542	0.700	0.879	1.372	1.812	2.228	2.764	3.169	4.144
11	0.260	0.540	0.697	0.876	1.363	1.796	2.201	2.718	3.106	4.025
12	0.259	0.539	0.695	0.873	1.356	1.782	2.179	2.681	3.055	3.930
13	0.259	0.538	0.694	0.870	1.350	1.771	2.160	2.650	3.012	3.852
14	0.258	0.537	0.692	0.868	1.345	1.761	2.145	2.624	2.977	3.787
15	0.258	0.536	0.691	0.866	1.341	1.753	2.131	2.602	2.947	3.733
16	0.258	0.535	0.690	0.865	1.337	1.746	2.120	2.583	2.921	3.686
17	0.257	0.534	0.689	0.863	1.333	1.740	2.110	2.567	2.898	3.646
18	0.257	0.534	0.688	0.862	1.330	1.734	2.101	2.552	2.878	3.610
19	0.257	0.533	0.688	0.861	1.328	1.729	2.093	2.539	2.861	3.579
20	0.257	0.533	0.687	0.860	1.325	1.725	2.086	2.528	2.845	3.552
21	0.257	0.532	0.686	0.859	1.323	1.721	2.080	2.518	2.831	3.527
22	0.256	0.532	0.686	0.858	1.321	1.717	2.074	2.508	2.819	3.505
23	0.256	0.532	0.685	0.858	1.319	1.714	2.069	2.500	2.807	3.485
24	0.256	0.531	0.685	0.857	1.318	1.711	2.064	2.492	2.797	3.467
25	0.256	0.531	0.684	0.856	1.316	1.708	2.060	2.485	2.787	3.450
26	0.256	0.531	0.684	0.856	1.315	1.706	2.056	2.479	2.779	3.435
27	0.256	0.531	0.684	0.855	1.314	1.703	2.052	2.473	2.771	3.421
28	0.256	0.530	0.683	0.855	1.313	1.701	2.048	2.467	2.763	3.408
29	0.256	0.530	0.683	0.854	1.311	1.699	2.045	2.462	2.756	3.396
30	0.256	0.530	0.683	0.854	1.310	1.697	2.042	2.457	2.750	3.385
40	0.255	0.529	0.681	0.851	1.303	1.684	2.021	2.423	2.704	3.307
50	0.255	0.528	0.679	0.849	1.299	1.676	2.009	2.403	2.678	3.261
100	0.254	0.526	0.677	0.845	1.290	1.660	1.984	2.364	2.626	3.174
150	0.254	0.526	0.676	0.844	1.287	1.655	1.976	2.352	2.609	3.146
∞	0.253	0.524	0.674	0.842	1.282	1.645	1.960	2.326	2.576	3.090

4.4 Approximationen statistischer Verteilungen durch die Normalverteilung



$$P(X \leq k) \approx \Phi \left(\frac{k + 0,5 - n \cdot p}{\sqrt{n \cdot p \cdot (1-p)}} \right)$$

$$P(a \leq X \leq b) \approx \Phi \left(\frac{b + 0,5 - n \cdot p}{\sqrt{n \cdot p \cdot (1-p)}} \right) - \Phi \left(\frac{a - 0,5 - n \cdot p}{\sqrt{n \cdot p \cdot (1-p)}} \right)$$

$$P(X \leq k) \approx \Phi \left(\frac{k + 0,5 - \lambda}{\sqrt{\lambda}} \right)$$

$$P(a \leq X \leq b) \approx \Phi \left(\frac{b + 0,5 - \lambda}{\sqrt{\lambda}} \right) - \Phi \left(\frac{a - 0,5 - \lambda}{\sqrt{\lambda}} \right)$$

Achtung: es muss sein $a \leq X \leq b$!

4.5 Die Kosinusverteilung

4.5.1 Das Modell

Viele Leute glauben an die Gaußsche Normalverteilung blindlings. Die Mathematiker („Zentraler Grenzwertsatz“) können beweisen, dass unter gewissen Voraussetzungen eine Zufallsvariable näherungsweise normal verteilt sein muss.

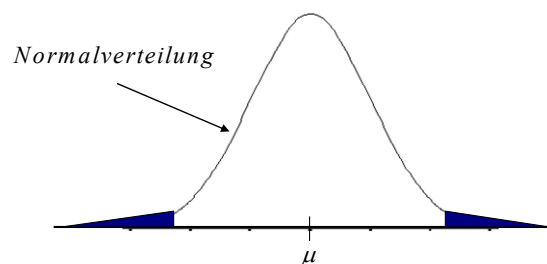
Das dient vielen als Rechtfertigung der Normalverteilung. Nun sind diese Voraussetzungen in der Realität kaum zu überprüfen.

Und was heißt „näherungsweise“, wie gut ist die Näherung?

Wir präsentieren eine andere Verteilung als Rivalen.

Auch sie hat die Form einer Glocke, ist mathematisch einfacher (ein Taschenrechner genügt, keine Tabelle) und hat einen **endlichen** Definitionsbereich.

Bei der Normalverteilung ist der Definitionsbereich die ganze Menge der reellen Zahlen (\mathbb{R}).



Beliebig große und auch negative Werte haben eine positive Wahrscheinlichkeit (blaue Flächen).

Das sollte stutzig machen.

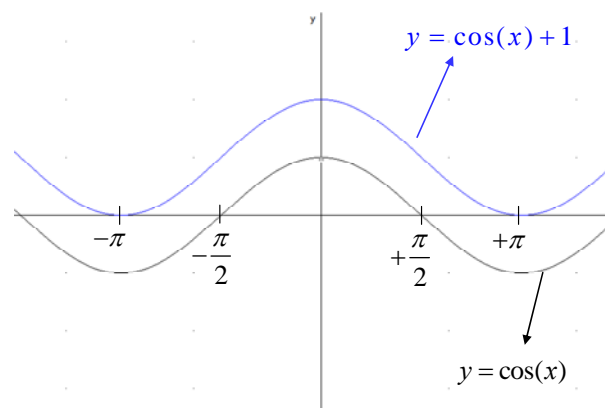
Die Rivalin ist die so genannte

„angehobene“ Kosinusverteilung,

eine in der Fachliteratur praktisch totgeschwiegene Verteilung (hat man etwas zu verbergen?).

Motivation

Zwischen $-\pi$ und $+\pi$ hat die Kosinusfunktion die Form einer Glocke, insbesondere wenn man sie um +1 anhebt.



Die Fläche unter dieser Kurve ist

$$\int_{-\pi}^{+\pi} (1 + \cos(x)) dx = \left[x + \sin(x) \right]_{-\pi}^{+\pi} = (\pi + \sin(\pi) - (-\pi) - \sin(-\pi)) = 2\pi$$

sie muss aber **gleich 1** sein, damit sie eine **statistische** Verteilung ist.

Damit haben wir die Kosinus-Verteilung

$$(4.26) \quad f(x) = \frac{1}{2\pi} (1 + \cos(x)) \quad -\pi \leq x \leq +\pi$$

als Wettbewerber zur Normalverteilung.

Im Gegensatz zur Normalverteilung ist sie auf ein endliches Intervall beschränkt und hat eine Verteilungsfunktion, die integrierbar ist.

$$F(x) = \int_{-\pi}^x \frac{1}{2\pi} (1 + \cos(t)) dt = \frac{1}{2\pi} [t + \sin(t)]_{-\pi}^x = \frac{1}{2\pi} \left(x + \sin(x) - \underbrace{(-\pi) - \sin(-\pi)}_0 \right)$$

$$(4.27) \quad F(x) = \frac{1}{2\pi} (\pi + x + \sin(x)) = 0,5 + \frac{x + \sin(x)}{2\pi} \quad -\pi \leq x \leq +\pi$$

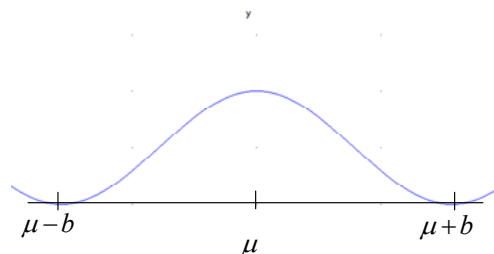
Dafür reicht ein Taschenrechner; Tabelle ist überflüssig.

Die Verteilung hat den Erwartungswert (Mittelwert) 0 und die Spannweite 2π .

Allgemein (ohne Beweis)

Die Kosinusverteilung mit dem Mittelwert μ und der Spannweite $2b$ hat die Funktion

$$(4.28) \quad f(x) = \frac{1}{2b} \left(1 + \cos \left(\frac{x - \mu}{b} \cdot \pi \right) \right) \quad \mu - b \leq x \leq \mu + b$$



und die Verteilungsfunktion

$$(4.29) \quad F(x) = \frac{1}{2} \left(1 + \frac{x - \mu}{b} + \frac{1}{\pi} \cdot \sin \left(\frac{x - \mu}{b} \cdot \pi \right) \right)$$

Diese allgemeine Kosinusverteilung hat den Mittelwert μ (Median) und die Spannweite $2b$.

$$\text{Varianz: } \sigma^2 = b^2 \left(\frac{1}{3} - \frac{2}{\pi^2} \right) \approx 0,1307 \cdot b^2$$

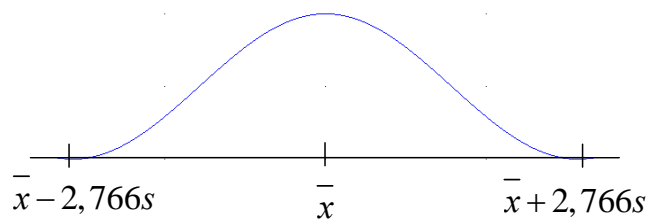
Ist eine Glockenkurve!!!

4.5.2 Schätzung der Parameter und Vergleich mit der Normalverteilung

Mit der Transformation $Z = b \cdot (X - \mu) \Rightarrow X = \mu + \frac{1}{b} \cdot Z$ kann die allgemeine Kosinusverteilung (4.28) bzw. (4.29) in die Standard Kosinusverteilung (4.26) bzw. (4.27) gebracht werden (auch wieder zurück).

Wie werden μ und b geschätzt (Punktschätzer)?

$$\hat{\mu} = \bar{x}$$
$$\hat{b} = 2,766s$$



Bei der Kosinusverteilung liegen zwischen

$$\mu \pm b = \bar{x} \pm 2,766s \quad 100 \% \text{ der Werte.}$$

Bei der Gauß-Verteilung liegen zwischen

$$\mu \pm 3\sigma = \bar{x} \pm 3s \quad 99,73 \% \text{ der Werte}$$

und zwischen

$$\mu \pm \infty \quad 100 \% \text{ der Werte}$$

Was ist plausibler?

4.6 Weibull-Verteilung

4.6.1 Das Modell

Der schwedische Ingenieur Waloddi Weibull schlug 1939 folgende Verteilung zur Modellierung von Materialermüdung – und festigkeit vor.

$$(4.30) \quad f(x) = \frac{b}{a} \cdot \left(\frac{x}{a}\right)^{b-1} \cdot e^{-\left(\frac{x}{a}\right)^b} \quad \text{für } x \geq 0$$

$$(4.31) \quad F(x) = 1 - e^{-\left(\frac{x}{a}\right)^b} \quad \text{für } x \geq 0$$

$E(X)$ und $\text{var}(X)$ sind bekannt, aber nicht von Bedeutung.

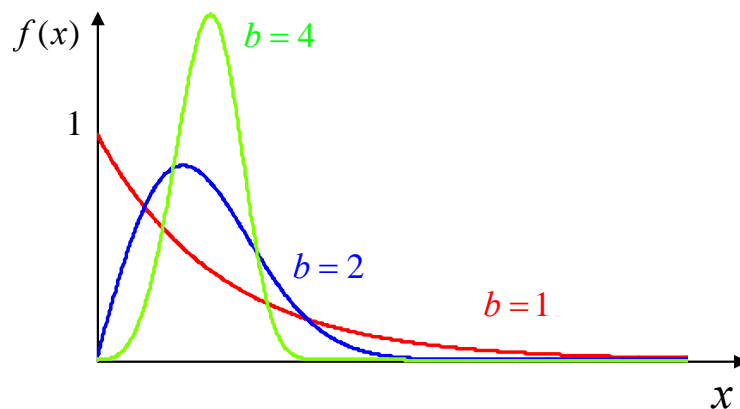
$$(4.32) \quad \text{Median} = a \cdot \sqrt[b]{\ln(2)}$$

Lustiges: man setze $x = a$

$$(4.33) \quad F(a) = 1 - e^{-\left(\frac{a}{a}\right)^b} = 1 - e^{-1} = 0,632$$

→ unabhängig von b.

a = charakteristische Lebensdauer.



Für $b = 1$

$$F(x) = 1 - e^{-\left(\frac{x}{a}\right)^1}$$

$$= 1 - e^{-\frac{1}{a}x}$$

hat man die Exponentialfunktion mit $\lambda = \frac{1}{a}$.

Für $b = 3, 4$ hat man ungefähr die Normalverteilung.

4.6.2 Erkennen der Verteilung und Schätzen der Parameter

Es ist

$$1 - F(x) = e^{-\left(\frac{x}{a}\right)^b}$$

$$\frac{1}{1 - F(x)} = e^{\left(\frac{x}{a}\right)^b}$$

$$\ln\left(\frac{1}{1 - F(x)}\right) = \left(\frac{x}{a}\right)^b$$

$$\ln\left(\ln\left(\frac{1}{1 - F(x)}\right)\right) = b \cdot \ln\left(\frac{x}{a}\right) = b \cdot \ln(x) - b \cdot \ln(a)$$

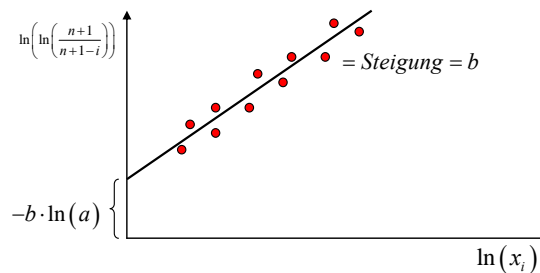
GERADE!!!

Plotte $\left(\ln(x), \ln\left(\ln\left(\frac{1}{1 - F(x)}\right)\right)\right)$ d.h. $\left(\ln(x_{(i)}), \ln\left(\ln\left(\frac{n+1}{n+1-i}\right)\right)\right)$

Liegen diese Punkte an einer Geraden, so kann man von einer Weibull-Verteilung ausgehen.

Punktschätzung von a und b

Weibull-Plot



$$\hat{b} = \text{Steigung}$$

$$\hat{a} \text{ aus Achsenabschnitt} = -\hat{b} \cdot \ln(\hat{a})$$

Beispiel IV Lines Connection Time

$$UCL_x = \bar{x} + E_2 \cdot \bar{R}_m = 325,77 + 2,659 \cdot 119,6 = 645$$

beim einzelnen Patienten

$$LCL_x = \bar{x} - E_2 \cdot \bar{R}_m = 325,77 - 2,659 \cdot 119,6 = 7$$

$$UCL_{R_m} = D_4 \cdot \bar{R}_m = 3,267 \cdot 119,6 = 392$$

Änderungen zum vorherigen Patienten

$$LCL_{R_m} = D_3 \cdot \bar{R}_m = 0 \cdot 119,6 = 0$$

Pro Patient wurde nur eine Messung durchgeführt!

„Sample Size“ = 1; aber die Spannweite (Range R_m) wird immer aus 2 Werten berechnet.

Für die Werte E_2, D_4, D_3 werden Tabellenwerte für $n = 2$ (Sample Size) genommen.

Bewertung der errechneten Werte: $UCL_x = 645 \rightarrow$ Ein Messwert liegt über dem Upper Control Limit (690) Wieso? Aufgabe des Unternehmens: Ursache finden \rightarrow Ursache abstellen!

5.1.2 Die wichtigsten Regelkarten

Die wichtigsten Regelkarten sind:

1. Mittelwert kombiniert mit der Spannweite (\bar{x} und R).
2. Mittelwert kombiniert mit der Standardabweichung (\bar{x} und s).
3. Einzelwert kombiniert mit gleitender Spannweite (x und R_m).
4. Ausschussanteil p (Fraction Defective)
5. Zahl der Ausfälle $c = count$.

In der High-Tech-Industrie zeigen sich diese Karten oft als untauglich.

- \rightarrow Bei der Mittelwertkarte wird statt der durchschnittlichen Standardabweichung \bar{s} (Variation **within** the sample) die Standardabweichung der Gruppenmittelwerte (Variation **between** the groups) genommen.

$$(5.1) \quad s_x = \sqrt{\frac{1}{k-1} \cdot \sum_{i=1}^k (\bar{x}_i - \bar{x})^2}$$

$k =$ Anzahl Gruppen

$\bar{x}_i =$ Mittelwert der Gruppe i

$\bar{x} =$ Mittelwert aller Gruppen

Als Eingriffsgrenzen werden folgende gewählt (Abel, Stark, Drain):

$$(5.2) \quad UCL_x : \bar{x} + 3 \cdot s_x$$

$$(5.3) \quad LCL_x : \bar{x} - 3 \cdot s_x$$

Analog dazu wird bei der p-Karte die Standardabweichung der Ausschussanteile

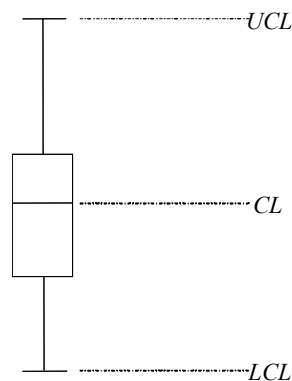
$$(5.4) \quad s_p = \sqrt{\frac{1}{k-1} \cdot \sum_{i=1}^k (p_i - \bar{p})^2}$$

genommen und als Eingriffsgrenzen werden folgende gewählt (Abel, Heimann):

$$(5.5) \quad UCL_p : \bar{p} + 3 \cdot s_p$$

$$(5.6) \quad LCL_p : \bar{p} - 3 \cdot s_p$$

Manche nehmen als Eingriffsgrenzen die Enden der Whisker des Box-Whiskers-Plots.



Für viele Prozesse braucht man diese Flexibilität und Freiheit. Wichtig ist nicht so sehr die statistische Formel, sondern das Ziel:

- a) „common causes“ von „special causes“ zu trennen
- b) „special causes“ zu entfernen und damit
- c) die Prozessstreuung zu verringern d.h., den Prozess zu verbessern.

5.1.3 Normalverteilung und SPC

In Büchern und Schulungsunterlagen wird das gerne thematisiert. Es ist richtig, dass bei einer Normalverteilung 99,73% der Werte zwischen $\mu - 3\sigma$ und $\mu + 3\sigma$ liegen.

Dass eine Normalverteilung vorliegt und dass $\mu = \bar{x}$ und $\sigma = s$ ist, ist nichts als Wunschdenken.

Der Vater der Regelkarten, der amerikanische Physiker Walter SHEWHART, hat in seinem 1931 erschienenen Buch

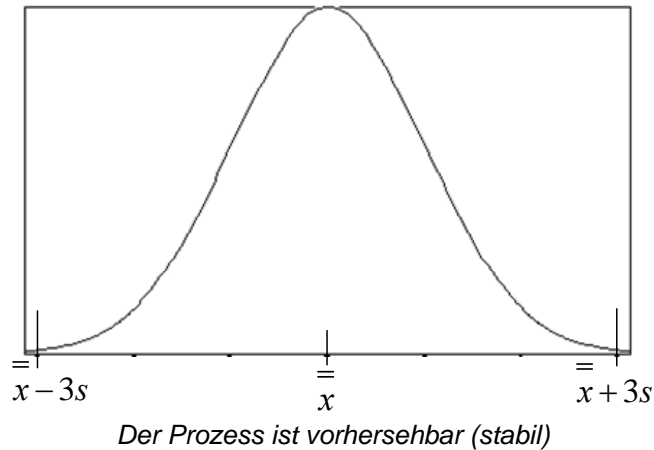
„Economic Control of Quality of Manufactured Product“

die Gültigkeit der Regelkarten in einem allgemeinen mathematischen Rahmen nachgewiesen, die Normalverteilung eher am Rande betrachtet und die Wirtschaftlichkeit des SPC hervorgehoben.

5.2.2 Prozessbeherrschung und Prozessfähigkeit

Beherrschter Prozeß: process in control

keine „special causes“, nur zufällige Schwankungen. (Normalverteilung)



Fähiger Prozess: capable process

Ein beherrschter (!) Prozess hält auch noch die von außen vorgegeben Spezifikationsgrenzen (Toleranzgrenzen) ein.

(→ technische Notwendigkeit)

USL: Upper Specification Limit

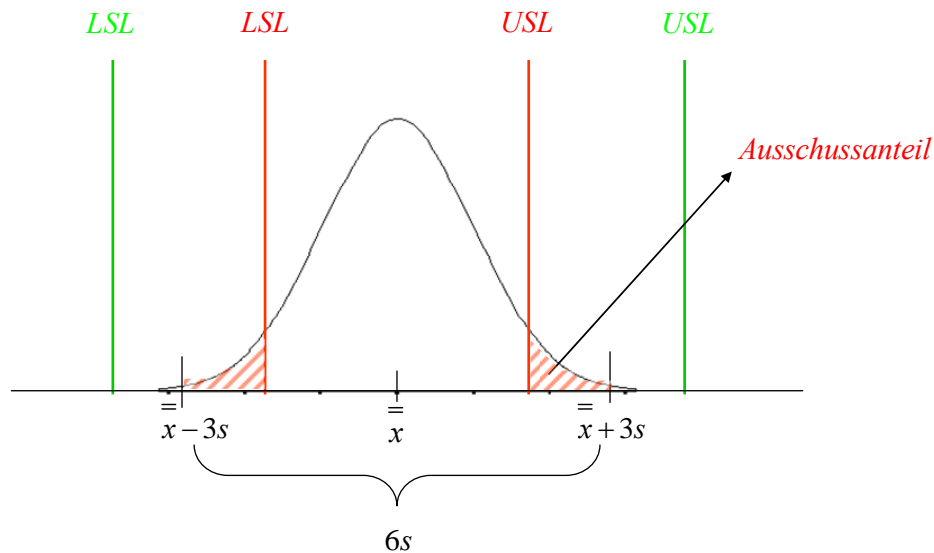
LSL: Lower Specification Limit

Annahme: Ein Prozess ist beherrscht innerhalb der Grenzen

$$(5.7) \quad \bar{x} \pm 3 \cdot s$$

Wichtig ist nun das Verhältnis!

$$(5.8) \quad \frac{\text{erlaubte Streuung}}{\text{tatsächliche Streuung}} = \frac{USL - LSL}{6s}$$



Liegen die Grenzen LSL und USL wie in grün dargestellt, ist das Verhältnis $\frac{\text{erlaubte Streuung}}{\text{tatsächliche Streuung}}$ **größer als 1: SUPER!**

Liegen die Grenzen LSL und USL nun wie in rot dargestellt, ist das Verhältnis $\frac{\text{erlaubte Streuung}}{\text{tatsächliche Streuung}}$ **kleiner als 1: SCHLECHT!** Es liegt ein Ausschussanteil vor, der beseitigt werden muss.

Legt man die durchschnittliche Standardabweichung \bar{s} zur Grunde (**Kurz-Zeit-Streuung**), so erhält man den Prozessfähigkeitsindex:

$$(5.9) \quad C_p = \frac{USL - LSL}{6 \cdot \bar{s}}$$

Verwendet man dagegen die Standardabweichung aller Daten (**Lang-Zeit-Streuung**):

$$(5.10) \quad s_g = \sqrt{\frac{1}{k \cdot n - 1} \cdot \sum_{i,j} (x_{ij} - \bar{x})^2}$$

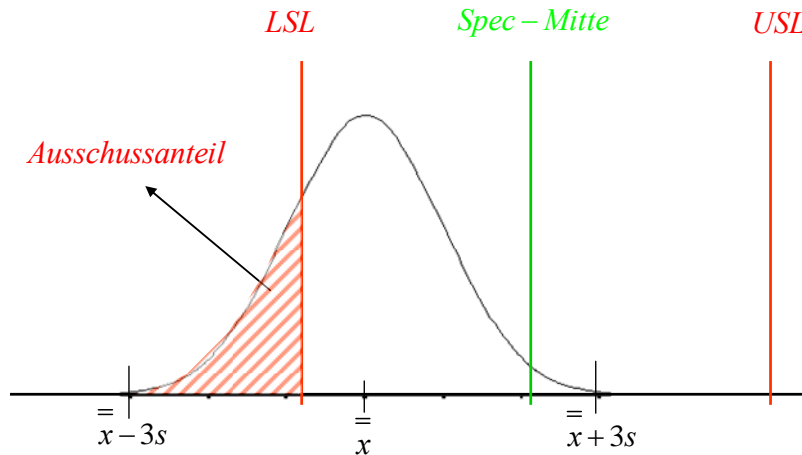
$g = \text{gesamt}$

so erhält man den Prozessperformanceindex

$$(5.11) \quad P_p = \frac{USL - LSL}{6 \cdot s_g}$$

Bemerkung: $s_g \geq \bar{s} \text{ also } C_p \geq P_p$

C_p und P_p sind nur aussagefähig, wenn der Prozess perfekt zentriert ist, d.h. der Ist-Mittelwert \bar{x} mit dem Sollmittelwert $\frac{USL + LSL}{2}$ übereinstimmt. Sonst ist C_p unsinnig.

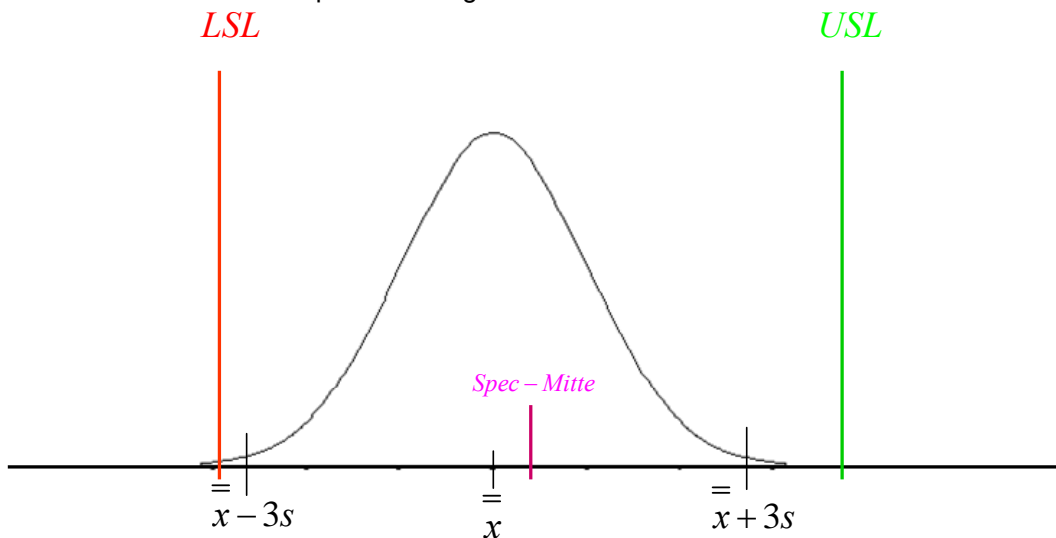


$C_p = 1,9$ ist nicht aussagefähig, da sehr viele Messungen um „Ausschussanteil“ liegen (ca. 40%).

Dieses Problem umgeht man mit

$$C_{p_k} \text{ bzw. } P_{p_k}$$

Zunächst Indizes mit nur einer Spezifikationsgrenze



grüner Fall:
Prozess soll $\leq USL$ sein

$$C_{p_u} = \frac{USL - \bar{x}}{3 \cdot \bar{s}} \quad P_{p_u} = \frac{USL - \bar{x}}{3 \cdot s_g}$$

roter Fall:
Prozess soll $\geq LSL$ sein

$$C_{p_l} = \frac{\bar{x} - LSL}{3 \cdot \bar{s}} \quad P_{p_l} = \frac{\bar{x} - LSL}{3 \cdot s_g}$$

Soll der Prozess $\leq USL$ und $\geq LSL$ sein (zweiseitiger Fall) so führt eine „worst-case“ Betrachtung zu:

$$(5.12) \quad C_{p_k} = \min\{C_{p_u}, C_{p_l}\}$$

bzw.

$$(5.13) \quad P_{p_k} = \min\{P_{p_u}, P_{p_l}\}$$

Beispiel:

$$C_{p_u} = 1,7 \quad C_{p_l} = 1,05$$

$$\Rightarrow C_{p_k} = 1,05$$

C_{p_k} bzw. P_{p_k} ist der „richtige“ Wert.

Warum dann C_p ?

Die Differenz $C_p - C_{p_k}$ sagt aus, wie gut oder schlecht der Prozess zentriert ist. C_{p_k} könnte maximal gleich C_p sein, wenn der Prozess perfekt zentriert wäre.

Es gilt:

$$C_{p_k} \leq C_p$$

und

$$C_{p_k} = C_p \cdot (1 - k),$$

wobei

$$(5.14) \quad k = \frac{\left(\frac{USL + LSL}{2} - \bar{x} \right)}{\left(\frac{USL - LSL}{2} \right)}$$

Ein Prozess mit	$C_{p_k} = 1$	geht gerade noch
	$C_{p_k} = 1,33$	gut
	$C_{p_k} = 1,67$	sehr gut, hervorragend

5.2.3 Normalverteilung und Fähigkeitsindizes

In Büchern und Schulungsunterlagen findet man häufig Umrechnungen von C_{p_k} in Ausschussanteile wie folgt:

$C_p = C_{p_k}$	0,33	0,67	1,00	1,33	1,67	2,00
Anteil außerhalb der Spec	31,7% =317000ppm	4,56% =45500ppm	0,27% =2700ppm	63ppm	0,57ppm	2ppb

Beispiel:

$$C_p = C_{pk} = 0,67 = \frac{USL - LSL}{6 \cdot \sigma} = \frac{2}{3}$$

$$USL - LSL = 4 \cdot \sigma = 0 \pm 2 \cdot \sigma$$

$$P(-2 \leq Z \leq +2) = 2 \cdot \Phi(2) - 1 = 2 \cdot 0,97725 - 1$$

$$= 0,95450 = 95,45\% \text{ (innerhalb der Spec)}$$

$$P(Z \leq -2 \text{ oder } Z \geq +2) = 1 - 0,95450 = 0,0455 = 4,55\% \text{ (außerhalb der Spec)}$$

Viele Leute nehmen solche Rechnungen für bare Münze (Mathematisch sind sie korrekt!).

Richtig ist

Wenn wir eine perfekte Normalverteilung haben und wenn μ und σ genau bekannt sind, dann bedeutet z.B.: $C_p = C_{pk} = 0,67$, dass man 456 von 100.000 Werten außerhalb der Spec hat (Ausschuss).

Da wir nicht wissen können, zu welchem Grad die beiden „Wenns“ erfüllt sind, ist diese Aussage Spekulation.

Das zweite „Wenn“, kann man wie folgt in den Griff bekommen:

Wenn die Zufallsstichprobe normalverteilt ist, so liefert

$$(5.15) \quad C_{pk} \pm 1,96 \sqrt{\frac{C_{pk}^2}{2(N-1)} + \frac{1}{9N}}$$

ein 95% Konfidenzintervall für den wahren Wert von C_{pk} .

Man kann sich zu 95% sicher sein, dass dieses Intervall den wahren C_{pk} -Wert überdeckt (enthält).

Dabei ist N die Gesamtzahl der Stichprobenwerte, also $N = k \cdot n$.

Beispiel

Vorausgesetzt eine perfekte Normalverteilung und eine Zufallsstichprobe von $N = k \cdot n = 20 \cdot 5 = 100$ Werten.

Ein aus einer Stichprobe gerechneter Wert von $C_{pk} = 1,33$ bedeutet, dass der wahre C_{pk} -Wert im Intervall (mit 95% Vertrauen)

$$1,33 \pm 1,96 \sqrt{\frac{1,33^2}{198} + \frac{1}{900}} = 1,33 \pm 0,2$$

liegt, also irgendwo zwischen 1,13 und 1,53.

5.2.4 Kosinusverteilung und Fähigkeitsindizes

Beispiel

$$USL=10 \quad LSL=4 \quad s=1,25$$

Normalverteilung

$$C_p = \frac{10-4}{6 \cdot 1,25} = \frac{6}{7,5} = 0,80$$

Kosinusverteilung

$$b = 2,766 \cdot s = 3,4575$$

$$\text{tatsächliche Streuung} = 2b = 6,915$$

$$\Rightarrow C_p = \frac{6}{6,915} = 0,87$$

Die Kosinusverteilung ergibt den höheren C_p -Wert. (Abel will das patentieren lassen)

5.3 Reliability Engineering

5.3.1 Ausfallrate / failure rate h(x)

$X = \text{Variable}$ $x = \text{konkrete Zahl}$

$F(x) = P(X \leq x)$ = Wahrscheinlichkeit, dass der Ausfallzeitpunkt (Variable X) vor der Zeit x geschieht.

$1 - F(x) = P(X > x)$ = Wahrscheinlichkeit, dass der Ausfallzeitpunkt (Variable X) nach der Zeit x geschieht.

$$(5.16) \quad h(x) = \frac{f(x)}{1 - F(x)}$$

↗ „Wahrscheinlichkeit des Ausfalls“

Ausfallrate

= Wahrscheinlichkeit, zur Zeit x auszufallen, wenn die Zeit x erreicht wurde

= momentane Ausfallwahrscheinlichkeit

h=hazard

5.3.2 h(x) für die Exponentialverteilung

$$(5.17) \quad h(x) = \frac{\lambda \cdot e^{-\lambda \cdot x}}{1 - (1 - e^{-\lambda \cdot x})} = \lambda \quad \text{konstante Ausfallrate}$$

5.3.3 h(x) für die Weibullverteilung

$$h(x) = \frac{\left(\frac{b}{a}\right) \cdot \left(\frac{x}{a}\right)^{b-1} \cdot e^{-\left(\frac{x}{a}\right)^b}}{e^{-\left(\frac{x}{a}\right)^b}}$$
$$(5.18) \quad h(x) = \left(\frac{b}{a}\right) \cdot \left(\frac{x}{a}\right)^{b-1}$$

Für $b < 1$: Fallende Ausfallrate
Für $b = 1$: konstante Ausfallrate
Für $b > 1$: Steigende Ausfallrate

Graph Badewannenkurve!!!

BURN IN

Normalverteilung, logarithmische Normalverteilung, Exponentialverteilung und Weibull-Verteilung gehören zur Klasse der **stetigen** Verteilungen.

Eine stetige Zufallsvariable X kann prinzipiell jede reelle Zahl annehmen.

6. Korrelation und Regression

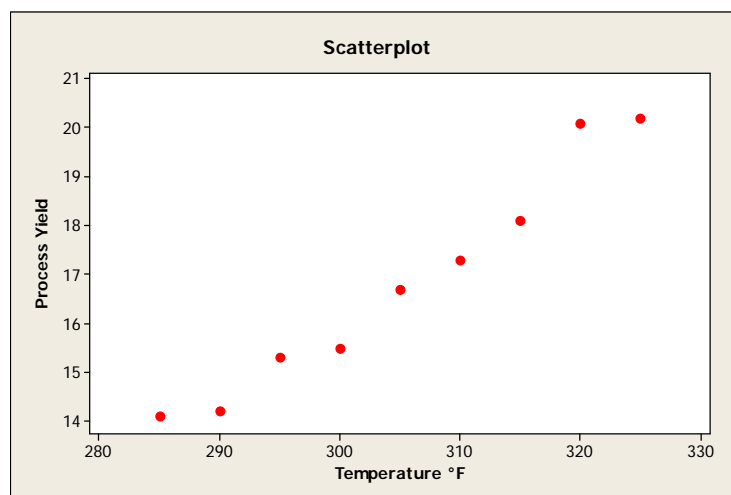
6.1 Korrelation

Bisher betrachteten wir eine Größe X und hatten dazu eine Stichprobe x_1, x_2, \dots, x_n .

Nun werden zwei Größen X und Y gleichzeitig betrachtet. Dazu liegt eine Stichprobe mit den Wertepaaren $(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$ vor.

Beispiel: Prozessausbeute bei bestimmten Temperaturen

Temperature X, °F	Process Yield Y	XY	X ²
310	17,3	5363,0	96100
290	14,2	4118,0	84100
300	15,5	4650,0	90000
315	18,1	5701,5	99225
320	20,1	6432,0	102400
305	16,7	5093,5	93025
325	20,2	6565,0	105625
295	15,3	4513,5	87025
285	14,1	4018,5	81225
2745	151,5	46455,0	838725



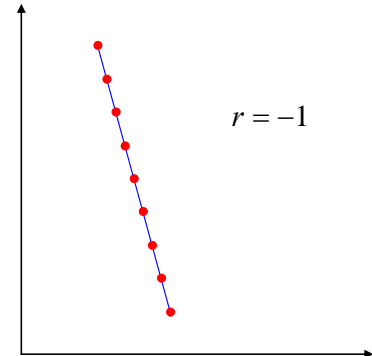
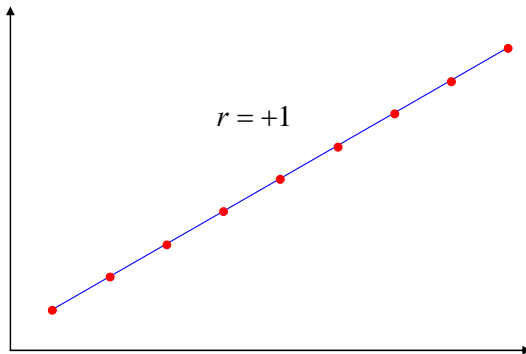
Eine zentrale Frage ist, wie stark oder schwach hängen X und Y zusammen. Der Korrelationskoeffizient des englischen Mathematikers Karl Pearson von 1896

$$(6.1) \quad r = \frac{\sum_{i=1}^n (x_i - \bar{x}) \cdot (y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2 \cdot \sum_{i=1}^n (y_i - \bar{y})^2}}$$

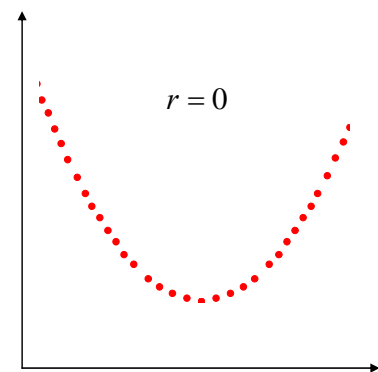
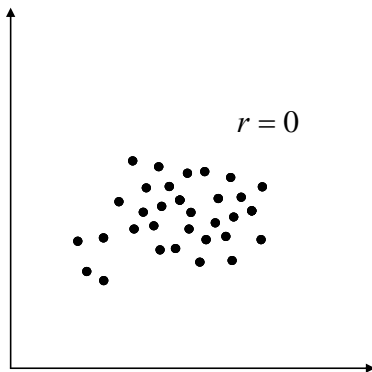
misst die Stärke des linearen Zusammenhangs $y = a + bx$ bzw. $x = c + dy$ und liegt immer zwischen

$$-1 \leq r \leq +1.$$

Liegen die Wertepaare (x_i, y_i) $i = 1, 2, \dots, n$ perfekt auf einer Geraden mit positiver Steigung, so hat man den Grenzfall $r = +1$. Liegen die Punkte dagegen auf einer Geraden mit negativer Steigung, so liegt der Grenzfall $r = -1$.



Je kleiner $|r|$ ist, desto schlechter ist der lineare Zusammenhang.



(r misst nur den **linearen** Zusammenhang, nicht den quadratischen!!!)

Achtung!

Ein Streudiagramm der Paare (x_i, y_i) ist unverzichtbar, da es

- anschaulich ist
- vermeidet, dass man in die Falle tappt, bei kleinem $|r|$ zu sagen, dass es keinen Zusammenhang gäbe, obwohl es eventuell einen nicht linearen Zusammenhang (z.B. Parabel) gibt.

r misst nicht einen generellen Zusammenhang, sondern nur den linearen.

r lässt sich in eine Formel umformen, welche zum praktischen Rechnen besser geeignet ist:

$$(6.2) \quad r = \frac{\sum_{i=1}^n x_i \cdot y_i - \frac{1}{n} \cdot \left(\sum_{i=1}^n x_i \right) \cdot \left(\sum_{i=1}^n y_i \right)}{\sqrt{\left(\sum_{i=1}^n x_i^2 - \frac{1}{n} \left(\sum_{i=1}^n x_i \right)^2 \right) \cdot \left(\sum_{i=1}^n y_i^2 - \frac{1}{n} \left(\sum_{i=1}^n y_i \right)^2 \right)}}$$

oder

$$(6.3) \quad r = \frac{\sum_{i=1}^n x_i \cdot y_i - n \cdot \bar{x} \cdot \bar{y}}{\sqrt{\left(\sum_{i=1}^n x_i^2 - n \cdot \bar{x}^2\right) \cdot \left(\sum_{i=1}^n y_i^2 - n \cdot \bar{y}^2\right)}}$$

Beispiel: Prozessausbeute bei bestimmten Temperaturen

$$r = \frac{46455,0 - \frac{1}{9} \cdot 2745 \cdot 151,5}{\sqrt{\left(838725 - \frac{1}{9} \cdot 2745^2\right) \cdot \left(2592,63 - \frac{1}{9} \cdot 151,5^2\right)}} = 0,9968$$

d.h. es liegt eine fast perfekte positive Korrelation

zwischen X = Temperatur
und Y = Ausbeute
vor.

Der Korrelationskoeffizient ist bei Praktikern sehr beliebt und weit verbreitet.

Ein Problem besteht darin, dass man nicht allgemein sagen kann, ab welchem Wert der Korrelationskoeffizient „signifikant“ ist, zumal dies auch noch von n abhängig sein müsste.

Selbst eine hohe **Korrelation** heißt nicht **Kausalität!**

Die Anzahl der Störche im Land korreliert sehr gut mit der Anzahl der Babies. Aber das lag wohl mehr an der Industrialisierung.

6.2 Regression

6.2.1 Einfache lineare Regression

Während bei der Korrelation die Größe X und Y „gleichberechtigt“ sind, wird bei der Regression, einem von dem englischen Biologen Francis GALTON (Cousin von Charles DARWIN) Ende des 19. Jahrhunderts eingeführten Begriffs, eine Variable (Y) durch die andere Variable (X) bestimmt, also

$$y = f(x)$$

Im einfachsten Fall geschieht das in Form einer Geraden:

$$y = b_0 + b_1 \cdot x$$

(einfache lineare Regression)

Die n Wertepaare (x_i, y_i) werden uns nicht den Gefallen tun (dann wäre die Bestimmung von b_0 und b_1 kinderleicht), dass sie auf einer Geraden liegen, also in der Regel wird gelten,

$$y_i \neq b_0 + b_1 \cdot x_i$$

Die n Abweichungen

$$y_i - (b_0 + b_1 \cdot x_i)$$

heißen RESIDUEN. Sie können negativ, null oder positiv sein.

$$(6.4) \quad e_i = y_i - (b_0 + b_1 \cdot x_i)$$

Hätten wir schon den Achsenabschnitt b_0 und die Steigung b_1 , so könnten wir mittels

$$b_0 + b_1 \cdot x_i$$

den zu x_i gehörenden theoretischen Wert „vorhersagen“. Wir wollen diesen \hat{y}_i nennen, also:

$$(6.5) \quad \hat{y}_i = b_0 + b_1 \cdot x_i$$

und damit

$$(6.6) \quad e_i = y_i - \hat{y}_i$$

e_i ist die Differenz zwischen dem gemessenen Wert y_i und dem vorhergesagten Wert \hat{y}_i .

Der deutsche Mathematiker Carl Friedrich Gauß hat 1809 erforscht, welche Eigenschaften ein Verfahren zur Bestimmung von b_0 und b_1 hat, das man **Methode der kleinste Quadrate** nennt. b_0 und b_1 sind die Lösungen des folgenden Minimierungsproblems:

$$\begin{aligned} \sum_{i=1}^n e_i^2 &\rightarrow \text{Minimum} \\ \sum_{i=1}^n (y_i - \hat{y}_i)^2 &\rightarrow \text{Minimum} \\ \sum_{i=1}^n (y_i - (b_0 + b_1 \cdot x_i))^2 &\rightarrow \text{Minimum} \end{aligned}$$

$\sum_{i=1}^n (y_i - b_0 - b_1 \cdot x_i)^2$ ist eine Funktion von b_0 und b_1 .

$$f(b_0, b_1) = \sum_{i=1}^n (y_i - b_0 - b_1 \cdot x_i)^2$$

$$\frac{\delta f}{\delta b_0} = \sum_{i=1}^n 2 \cdot (y_i - b_0 - b_1 \cdot x_i) \cdot (-1) = 0$$

$$\frac{\delta f}{\delta b_1} = \sum_{i=1}^n 2 \cdot (y_i - b_0 - b_1 \cdot x_i) \cdot (-x_i) = 0$$

Die beiden Gleichungen haben die Lösungen:

$$(6.7) \quad b_1 = \frac{\sum_{i=1}^n x_i y_i - \frac{1}{n} \cdot \left(\sum_{i=1}^n x_i \right) \cdot \left(\sum_{i=1}^n y_i \right)}{\sum_{i=1}^n x_i^2 - \frac{1}{n} \cdot \left(\sum_{i=1}^n x_i \right)^2}$$

$$(6.8) \quad b_0 = \bar{y} - b_1 \cdot \bar{x}$$

Zwischen b_1 und dem Korrelationskoeffizienten r besteht die folgende Beziehung:

$$r > 0 \leftrightarrow b_1 > 0$$

$$r < 0 \leftrightarrow b_1 < 0$$

Beispiel: Prozessausbeute bei bestimmten Temperaturen

$$b_1 = \frac{46455 - \frac{1}{9} \cdot 2745 \cdot 151,5}{838725 - \frac{1}{9} \cdot 2745^2} = 0,1650$$

$$b_0 = \frac{151,5}{9} - 0,1650 \cdot \frac{2745}{9} = -33,49$$

Nach diesem Verfahren lassen sich b_0 und b_1 immer bestimmen, egal wie gut oder schlecht die Gerade „passt“.

Eine Maßzahl für die Güte der Anpassung der Stichprobe $(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$ an die Gerade $y = b_0 + b_1 \cdot x$ ist das Bestimmtheitsmaß (coefficient of determination).

$$(6.9) \quad B = \frac{\sum_{i=1}^n (\hat{y}_i - \bar{y})^2}{\sum_{i=1}^n (y_i - \bar{y})^2}$$

Erläuterung

Es gilt die Gleichung

$$\begin{aligned} \sum_{i=1}^n (y_i - \bar{y})^2 &= \sum_{i=1}^n \underbrace{(y_i - \hat{y}_i)}_a + \underbrace{\hat{y}_i - \bar{y}}_b \\ (1) \quad &= \sum_{i=1}^n (y_i - \hat{y}_i)^2 + \sum_{i=1}^n (\hat{y}_i - \bar{y})^2 \end{aligned}$$

Wenn alle Abweichungen $e_i = y_i - \hat{y}_i$ gleich 0 sind, haben wir eine perfekte Gerade und die obige Gleichung (1) reduziert sich auf

$$\sum_{i=1}^n (y_i - \bar{y})^2 = \sum_{i=1}^n (\hat{y}_i - \bar{y})^2$$

und damit $B = 1$.

Andererseits kann B nie negativ werden, also $0 \leq B \leq 1$.

Je größer B ist, desto besser ist die Anpassung der Punktwolke an die Gerade.

Inhaltliche Interpretation

Im **Nenner** von B steht die linke Seite der Gleichung (1), das ist die gesamte Varianz.

Im **Zähler** steht die Varianz der \hat{y}_i -Werte, also die durch die Gerade $y = b_0 + b_1 \cdot x$ erklärte Varianz.

Also ist B der Anteil der durch die Gerade erklärten Varianz an der gesamten Varianz.

Sprechweise (nicht ganz korrekt)

Ist z.B. $B=0,83$, so sagt man, y werde zu 83 % durch $y = b_0 + b_1 \cdot x$ erklärt. (und zu 17 % durch andere Faktoren)

Es lässt sich beweisen, dass B gerade der quadrierte Korrelationskoeffizient ist

$$\text{also } B = r^2$$

(R-square)

Warum dann B? Der Wert von B lässt sich interpretieren (siehe oben), der Wert von r nicht. Allerdings weiß man durch r, ob die Korrelation positiv oder negativ ist, was B nicht tut. Man hat aber sowieso das b_1 und damit auch das Vorzeichen von r.

→ B statt r!!!

Beispiel: Prozessausbeute bei bestimmten Temperaturen

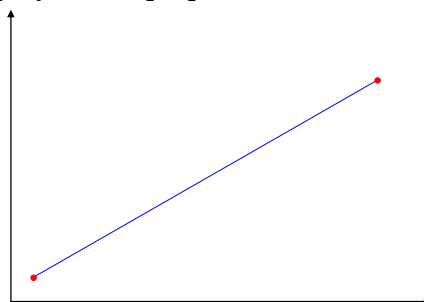
$$R - sq = B = 0,9936$$

sehr gut

$$r = +\sqrt{0,9936} = +0,9968$$

Achtung

Wenn man nur 2 Wertepaare (x_1, y_1) und (x_2, y_2) hat, dann ist B immer gleich 1.



Der Faule ist der Gewinner ($B=1$).

Der Fleißige ($n>2$) ist der Verlierer ($B<1$).

In der Formel für B muss noch die Zahl der Wertepaare, also n , berücksichtigt werden. Dies führt zum korrigierten Bestimmtheitsmaß (adjusted R-square):

$$(6.10) \quad R_{adj}^2 = R^2 - \frac{1}{n-2} \cdot (1 - R^2)$$

mit $R^2 = B$

Es gilt immer

$$R_{adj}^2 \leq R^2$$

$$n \rightarrow \infty \quad R_{adj}^2 \rightarrow R^2$$

$$n = 2 \quad R_{adj}^2 = \text{"ERROR"}$$

$$R^2 = 1$$

Kommentar

Es ist geistlos zu fragen, wie gut sich eine Gerade an zwei Punkte anpasst!

Beispiel: Prozessausbeute bei bestimmten Temperaturen

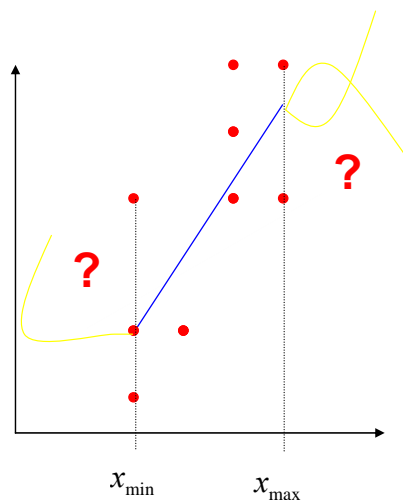
$$R_{adj}^2 = 0,9936 - \frac{1}{9-2} \cdot (1 - 0,9936) = 0,9927$$

sehr gut

Der ermittelte Zusammenhang $y = b_0 + b_1 \cdot x$ ist nur im untersuchten Bereich

$$x_{\min} \leq x \leq x_{\max}$$

gesichert. Außerhalb dieses Bereiches ist Spekulation.

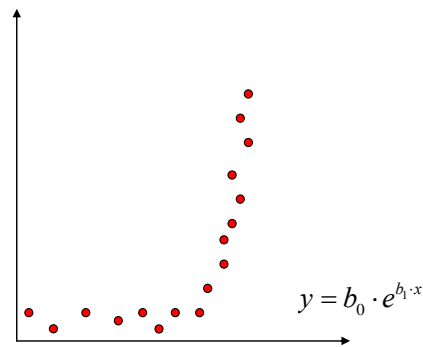


6.2.2 Einfache nicht lineare Regression

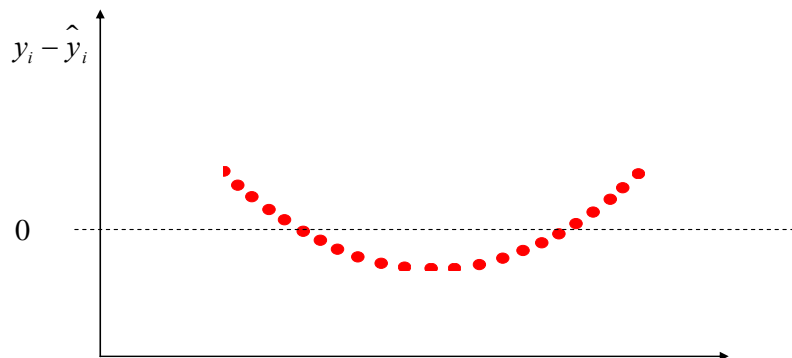
Manchmal sagt einem die Theorie, dass zwischen x und y kein linearer Zusammenhang besteht.

Beispiel: $s(t) = \frac{1}{2} \cdot g \cdot t^2$

Oder es gibt zwar keine Theorie, aber die Punkte des Streudiagramms weisen ganz klar auf eine nicht lineare Funktion hin.



Oder der Residuen-Plot weist ein merkwürdiges Muster auf



was vermuten lässt, dass eine nicht lineare Funktion besser geeignet ist als $y = b_0 + b_1 \cdot x$.

Beispiel

Exponentielle Modellgleichung: $y = b_0 \cdot e^{b_1 \cdot x}$

Bestimmung von b_0 und b_1 nach der Methode der kleinsten Quadrate.

$$f(b_0, b_1) = \sum_{i=1}^n (y_i - \hat{y}_i)^2 = \sum_{i=1}^n (y_i - b_0 \cdot e^{b_1 \cdot x_i})^2$$

$\rightarrow \text{Minimum}$

$$\frac{\delta f}{\delta b_0} = \sum_{i=1}^n 2 \cdot (y_i - b_0 \cdot e^{b_1 \cdot x_i}) \cdot (-e^{b_1 \cdot x_i}) = 0$$

$$\frac{\delta f}{\delta b_1} = \sum_{i=1}^n 2 \cdot (y_i - b_0 \cdot e^{b_1 \cdot x_i}) \cdot (-b_0 \cdot e^{b_1 \cdot x_i}) \cdot (x_i) = 0$$

Dieses Gleichungssystem ist nicht exakt zu lösen, sondern nur näherungsweise mit Computerprogrammen.

Ein für uns gangbarer Weg besteht darin, die nicht lineare Funktion linear zu machen.

$$\begin{array}{ll} y = b_0 \cdot e^{b_1 \cdot x} & \text{nicht linear} \\ \ln(y) = \ln(b_0) + b_1 \cdot x & \text{linear} \end{array}$$

So dann lassen sich die Resultate des linearen Modells verwenden:

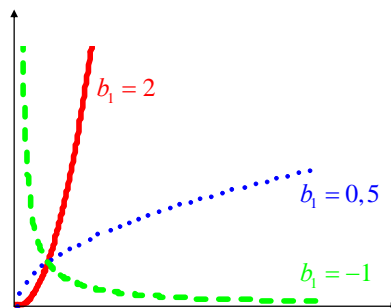
$$(6.11) \quad b_1 = \frac{\sum_{i=1}^n x_i \cdot \ln(y_i) - \frac{1}{n} \cdot \left(\sum_{i=1}^n x_i \right) \cdot \left(\sum_{i=1}^n \ln(y_i) \right)}{\sum_{i=1}^n x_i^2 - \frac{1}{n} \cdot \left(\sum_{i=1}^n x_i \right)^2}$$

$$\ln(b_0) = \frac{\sum_{i=1}^n \ln(y_i)}{n} - b_1 \cdot \frac{\sum_{i=1}^n x_i}{n}$$

$$(6.12) \quad b_0 = e^{\ln(b_0)}$$

Weiteres Beispiel

Potenzfunktion: $y = b_0 \cdot x^{b_1}$



$$\ln(y) = \ln(b_0) + b_1 \cdot \ln(x)$$

$$(6.13) \quad b_1 = \frac{\sum_{i=1}^n \ln(x_i) \cdot \ln(y_i) - \frac{1}{n} \cdot \left(\sum_{i=1}^n \ln(x_i) \right) \cdot \left(\sum_{i=1}^n \ln(y_i) \right)}{\sum_{i=1}^n \ln(x_i)^2 - \frac{1}{n} \cdot \left(\sum_{i=1}^n \ln(x_i) \right)^2}$$

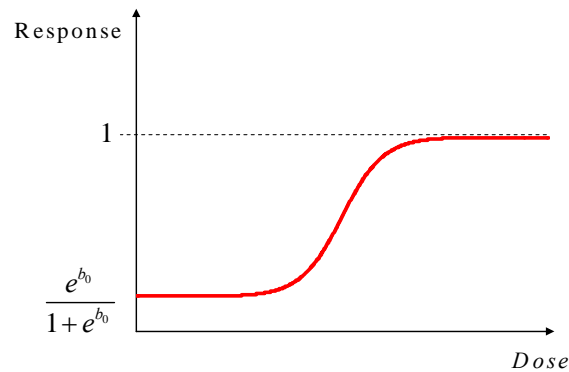
$$\ln(b_0) = \frac{\sum_{i=1}^n \ln(y_i)}{n} - b_1 \cdot \frac{\sum_{i=1}^n \ln(x_i)}{n}$$

$$(6.14) \quad b_0 = e^{\ln(b_0)}$$

Weiteres Beispiel

Sättigungsfunktion: $y = \frac{e^{b_0 + b_1 \cdot x}}{1 + e^{b_0 + b_1 \cdot x}}$

$$y(0) = \frac{e^{b_0}}{1 + e^{b_0}} \quad y(\infty) = 1$$



$$\ln\left(\frac{y}{1-y}\right) = \dots = b_0 + b_1 \cdot x$$

Setze $y^* = \ln\left(\frac{y}{1-y}\right)$

$$(6.15) \quad b_1 = \frac{\sum_{i=1}^n x_i \cdot y_i^* - \frac{1}{n} \cdot \sum_{i=1}^n x_i \cdot \sum_{i=1}^n y_i^*}{\sum_{i=1}^n x_i^2 - \frac{1}{n} \cdot \left(\sum_{i=1}^n x_i\right)^2}$$

$$(6.16) \quad b_0 = \frac{\sum_{i=1}^n y_i^*}{n} - b_1 \cdot \frac{\sum_{i=1}^n x_i}{n}$$

Der Fantasie sind keine Grenzen gesetzt

z.B.: $\log_{10} y = b_0 + b_1 \cdot \frac{1}{x}$

lässt sich folgendermaßen lösen:

statt $x_i \rightarrow \frac{1}{x_i}$

statt $y_i \rightarrow \log_{10} y_i$

dann die Formeln für b_1 und b_0

mit $\frac{1}{x_i}$ (statt x_i)

mit $\log_{10} y_i$ (statt y_i)

Das Gleiche gilt für die rein quadratische Gleichung

$$y = b_0 + b_1 \cdot x^2$$

(x_i^2 statt x_i in den Formeln für b_1 und b_0)

Will man allerdings ein echtes Polynom modellieren, also

$$y = b_0 + b_1 \cdot x + b_2 \cdot x^2 \text{ oder}$$

$$y = b_0 + b_1 \cdot x + b_2 \cdot x^2 + b_3 \cdot x^3$$

ist man mit den bisherigen Formeln am Ende, da jetzt nicht zwei sondern drei bzw. vier Koeffizienten zu bestimmen sind.

→ Mehrfache (Multiple) Lineare Regression

6.2.3 Mehrfache (Multiple) lineare Regression

Möchte man die abhängige Variable y nicht durch eine, sondern mehrere unabhängige Variablen x_1, x_2, \dots, x_k modellieren d.h.:

$$y = b_0 + b_1 \cdot x_1 + b_2 \cdot x_2 + \dots + b_k \cdot x_k$$

so sind die gesuchten Koeffizienten b_0, b_1, \dots, b_k wieder die Lösung der Optimierungsaufgabe (Methode der kleinsten Quadrate).

$$\sum_{i=1}^n (y_i - \hat{y}_i)^2 = \sum_{i=1}^n (y_i - (b_0 + b_1 \cdot x_{1i} + b_2 \cdot x_{2i} + \dots + b_k \cdot x_{ki}))^2$$

$$\Rightarrow \text{MINIMUM}$$

Das führt zu $k + 1$ partiellen Ableitungen, die gleich Null zu setzen sind. Ohne eine entsprechende Software ist dieses Gleichungssystem praktisch nicht zu lösen.

Die Formel für b_0, b_1, \dots, b_k können nicht angegeben werden; das Bestimmtheitsmaß ist wieder

$$(6.17) \quad B = R^2 = \frac{\sum_{i=1}^n (\hat{y}_i - \bar{y})^2}{\sum_{i=1}^n (y_i - \bar{y})^2}$$

und das korrigierte Bestimmtheitsmaß ist nun

$$(6.18) \quad R_{adj}^2 = R^2 - \frac{k}{n - (k + 1)} \cdot (1 - R^2).$$

Bei festen k gilt

$$\lim_{n \rightarrow \infty} R_{adj}^2 = R^2 \quad \text{und} \quad R_{adj}^2 \leq R^2.$$

William Occam (Ockham), geb. ca. 1280 in Occam / England, Franziskaner, gestorben 1349 in München. (Zuflucht bei Kaiser Ludwig der Bayer)

„Eine Theorie sollte nicht unnötig kompliziert gemacht werden.“

Occam's razor

Bei k unabhängigen Variablen gibt es $2^k - 1$ mögliche Modellgleichungen (Teilmengen).

z.B. für $k = 3$ gibt es die Teilmengen

$\{x_1\}, \{x_2\}, \{x_3\}; \{x_1, x_2\}, \{x_1, x_3\}, \{x_2, x_3\}; \{x_1, x_2, x_3\}$

Wie findet man die beste Teilmenge?

Das „volle“ Modell

$$y = b_0 + b_1 \cdot x_1 + \dots + b_k \cdot x_k$$

hat von allen $2^k - 1$ möglichen Gleichungen den höchsten R^2 – Wert (wäre also der Beste), aber nicht unbedingt den höchsten R_{adj}^2 – Wert, auf den es schließlich ankommt.

Welche Strategien gibt es, die beste der $2^k - 1$ Gleichungen zu finden? (Alle Rechner unterstützt).

1. „All subsets regression“

Von allen $2^k - 1$ möglichen Gleichungen wählt man jene mit dem höchsten R_{adj}^2 – Wert aus. Dieses, nur an Zahlenwerte orientierte Vorgehen gilt bei Eingeweihten als etwas kindisch, zumal wenn sich die R_{adj}^2 – Werte eventuell nur um Hundertstel oder Tausendstel unterscheiden.

2. „Backward Selection“

Man startet mit dem vollen Modell und entfernt Schritt für Schritt jene Variable, deren Elimination die geringste Reduktion im R_{adj}^2 – Wert mit sich bringt. Solange bis man einerseits eine kurze Gleichung hat (Occam's razor) und andererseits noch einen akzeptablen R_{adj}^2 – Wert.

3. „Forward Selection“

Man startet mit dem „Nullmodell“ ($y = b_0$) und fügt Schritt für Schritt jene Variable hinzu, die den größten Zuwachs beim R_{adj}^2 – Wert mit sich bringt. Solange bis man einerseits eine kurze Gleichung hat (Occam's razor) und andererseits noch einen akzeptablen R_{adj}^2 – Wert.

4. Mallows C_p

(nicht zu verwechseln mit dem Prozessfähigkeitsindex C_p)

Collin Mallows, englisch-amerikanischer Statistiker erfand 1964 folgende Formel

$$(6.19) \quad C_p = \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{\frac{1}{n-k-1} \cdot \sum_{i=1}^n (y_i - \hat{y}_i)^2} + 2 \cdot p - n$$

wobei im Zähler

$$\hat{y}_i = b_0 + b_1 \cdot x_{1i} + b_2 \cdot x_{2i} + \dots + b_{p-1} \cdot x_{p-1,i}$$

also $p = 1$ (für Konstante, oder „Nullmodell“) + $p - 1$ (für die $p - 1$ Variable)

= Zahl der Koeffizienten im Teilmodell

im Nenner

$$\hat{y}_i = b_0 + b_1 \cdot x_{1i} + b_2 \cdot x_{2i} + \dots + b_k \cdot x_{ki}$$

k = maximale Zahl der unabhängigen Variablen (volles Modell)

Es ist C_p für $p = 2, \dots, k + 1$ zu berechnen.

Für gute Gleichungen hat zu gelten $C_p \approx p$

Wir wählen diejenige Teilmenge aus, für die p klein und $C_p \approx p$ ist.

Anmerkung

$$C_{k+1} = \frac{1}{\frac{1}{n-k-1}} + 2 \cdot \underbrace{(k+1)}_p - n$$

$$= n - k - 1 + 2k + 2 - n = k + 1$$

5. Da die obigen Verfahren zu keinem eindeutigen Ergebnis führen, sollte der Anwender den Mut aufbringen, selbst bei der Auswahl der Variablen aufgrund seines Fachwissens einzugreifen. Welche Variablen müssen unbedingt, welche könnten auch aus der Sicht der Fachdisziplin (Physik, BWL, Psychologie) in der Gleichung sein?

7. Etwas Allerlei zum guten (?) Schluss

7.1 Das Geburtstagsproblem

x Personen treffen sich zufällig. Wie groß ist die Wahrscheinlichkeit, dass mindestens 2 von ihnen am gleichen Tag Geburtstag haben?

Lösung

Wahrscheinlichkeit, dass die x Personen alle an verschiedenen Tagen Geburtstag haben:

$$= \frac{365}{365} \cdot \frac{364}{365} \cdot \frac{363}{365} \cdot \dots \cdot \frac{365-x+1}{365}$$

Wahrscheinlichkeit, dass von x Personen mindesten zwei am gleichen Tag Geburtstag haben:

$$(*) = 1 - \frac{365 \cdot 364 \cdot 363 \cdot \dots \cdot (365 - x + 1)}{365^x}$$

Die Wahrscheinlichkeit beträgt in Abhängigkeit von x:

x	10	15	20	25	30	35	40
Wahrscheinlichkeit	11,7%	25,3%	41,1%	56,9%	70,6%	81,4%	89,1%

Voraussetzungen für die Gültigkeit der Formel (*):

- x Personen stellen eine Zufallsstichprobe dar
- jeder Tag als Geburtstag gleich wahrscheinlich

Realität

Im Sommer werden mehr Kinder geboren als im Winter. Am seltensten im November & Dezember. Am häufigsten im Juli & August.

Nicht gleichmäßig über die Tage des Jahres verteilte Geburtstage erhöhen die Wahrscheinlichkeit (*) sogar.

Schaltjahr: 1 „extra Tag“ in vier Jahren ergibt eine ungleiche Verteilung. Effekt: siehe oben.

Ein anderes Geburtstagsproblem

Außer Ihnen sind noch n Personen im Raum. Wie groß ist die Wahrscheinlichkeit, dass mindestens eine weitere Person am gleichen Tag Geburtstag hat wie Sie?

→ Ziehen ohne Zurücklegen, aber N ist sehr groß.

$$\frac{n}{N} \leq 0,05 \rightarrow \text{Binomialverteilung}$$

$$p = P(\text{Geburtstag}) = \frac{1}{365}$$

$$\begin{aligned} P(X \geq 1) &= 1 - P(X = 0) = 1 - \binom{n}{0} \cdot p^0 \cdot (1-p)^{n-0} \\ &= 1 - \left(\frac{364}{365}\right)^n \end{aligned}$$

n	10	20	30	50	70
$P(X \geq 1)$	0,027	0,053	0,075	0,128	0,175

„Break Even“

$$\begin{aligned} 1 - \left(\frac{364}{365}\right)^n &= 0,5 \\ \left(\frac{364}{365}\right)^n &= 0,5 \\ n &= \frac{\ln(0,5)}{\ln\left(\frac{364}{365}\right)} = 252,65 \\ n &= 253 \end{aligned}$$

7.2 Das Ziegenproblem

Bei einer Spielshow soll der Kandidat eines von drei aufgebauten Toren auswählen.

Hinter einem verbirgt sich der Gewinn, ein Auto, hinter den beiden anderen jeweils eine Ziege, also Nieten oder Trostpreise.

Folgender Spielablauf ist immer gleich und den Kandidaten vorab bekannt.

1. Der Kandidat wählt ein Tor aus, das aber vorerst verschlossen bleibt.
2. Daraufhin öffnet der Moderator, der die Position des Gewinnes kennt, eines der beiden, nicht vom Kandidaten ausgewählten Tore, hinter dem sich eine Ziege befindet.
Im Spiel befinden sich also jetzt noch ein Gewinn und eine Niete.
3. Der Moderator bietet dem Kandidaten an, seine Entscheidung zu überdenken und das andere Tor zu wählen.

Wie soll sich der Kandidat entscheiden, um seine Gewinnchancen zu maximieren?

Wechseln bringt Erfolg, wenn der Kandidat zunächst das falsche Tor gewählt hat. Wahrscheinlichkeit, dass das falsche Tor gewählt wurde ist $\frac{2}{3}$.