Grundlagen und Begriffe

Grundgesamtheit	G / GG	
Merkmal	M metrisch (CHF/Monat) / ordinal (schlecht – sehr gut) / nominal (0/1)	
Arithmetisches Mittel von M	μ → bei m. Merkm. μ_x, μ_y, μ_z	$= E(X_1)$
Anteilswert bzw. Proportion (bei nominalem Merkmal)	π → bei nominalem Merkmal anstelle von μ	
Varianz von M	σ^2 → bei m. Merkm. σ_x^2 … usw.	$= Var(X_1)$
Standardabweichung von M	σ → bei m. Merkm. σ_x… usw.	
Grösse der Grundgesamtheit (bekannt / unbekannt)	N	
Gemessene Daten des Merkmals M in der Stichprobe	$x_1, x_2, x_3, \ldots x_n$ (x als Kleinbuchstabe)	
Wahrscheinlichkeit	p	
Stichprobe	**S**	
Arithmetische Mittel der Daten	$\bar{x}$	
Anteilswert bzw. Proportion (bei nominalem Merkmal)	$\hat{p}$ → bei nominalem Merkmal anstelle von $\bar{x}$	
Empirische Varianz der Daten	s^2 → bei m. Merkm. $s^2_{x\ldots}$ usw.	
Empirische Standardabweichung der Daten	s → bei m. Merkm. $s_{x\ldots}$ usw.	
Grösse der Stichprobe	n	
Zufallsvariable	X diskret (bestimmte Werte – Münze) / stetig (beliebige reele Zahl)	
Erwartungswert der Zufallsvariable X_1	$E(X_1)$	$= \mu$
Varianz der Zufallsvariable X_1	$Var(X_1)$	$= \sigma^2$
Standardabweichung der Zufallsvariable X_1	Wurzel von $Var(X_1)$	$= \sigma$

Erwartungswert μ

$$\mu = \sum_{i=1}^{k} x_i \cdot p_i$$

Standardabweichung σ

$$\sigma = \sqrt{\sum_{i=1}^{k} (x_i - \mu)^2 \cdot p_i}$$

Arithmetisches Mittel – Summe der Merkmalswerte mal ihre relativen Gewichte

$$\bar{x} = \sum_{i=1}^{k} x_i \cdot \frac{W_i}{n} = \sum_{i=1}^{k} x_i \cdot w_i$$

Standardabweichung s

$$= \sqrt{\sum_{i=1}^{k} (x_i - \bar{x})^2 \cdot \frac{W_i}{n-1}}$$

Kombinatorik

Permutation (ohne Wiederholung)	Variation (ohne Wiederholung)	Kombination (ohne Wiederholung)
Werte können nur einmal vorkommen (z.B. Lottozahlen) Anzahl Permutationen = n! (Anzahl Möglichkeiten der Zusammensetzung) $n! = 1 \cdot 2 \cdot 3 \cdot \ldots \cdot n$ → **TR: Zahl !**	Reihenfolge entscheidend – eine Auswahl von r Elementen aus n $Anzahl\ Variationen = \frac{n!}{(n-r)!}$ **TR: n nPr r**	Auswahl – Reihenfolge nicht entscheidend - geschrieben $\binom{n}{r}$ $Anzahl\ Kombinationen = \frac{n!}{(n-r)! \cdot r!}$ **TR: n nCr r**

Beispiele

20 Gäste, jeder prostet jedem? (20 nCr 2)

$$\binom{20}{2} = \frac{20 \cdot 19}{1 \cdot 2} = 190$$

Wie viele Möglichkeiten mit 4 Richtigen beim Lotto „6 aus 45“?

4 aus 6 Richtigen: $\binom{6}{4} = 15$ 2 aus 39 Falschen: $\binom{39}{2} = 741$ somit $\binom{6}{4} \cdot \binom{39}{2} = 15 \cdot 741 = 11'115$

Wahrscheinlichkeitsverteilung

Binomialverteilung X ~ B(n, p)

X = r = Anzahl Ausführungen (von n), bei denen Ereignis E eintritt
n = Versuche
p = Wahrscheinlichkeit („Ereignis E“ tritt im Einzelfall ein)

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

X hat den Erwartungswert	X hat die Standardabweichung	X hat die Varianz
$\mu = n \cdot p$	$\sigma = \sqrt{n \cdot p \cdot (1-p)}$	$\sigma^2 = n \cdot p \cdot (1-p)$

TR: Wahrscheinlichkeit, dass E genau r mal eintritt	TR: Wahrscheinlichkeit, dass E höchstens r mal eintritt
P(X=r) = Binomial*pdf* mit n, p, r/X	P(X<=r) = Binomialcdf mit n, p, r/X

Beispiele

Ein holländisches Versandhaus preist Blumenzwiebeln in Packungen zu zwölf Stück an. Auf der Packung wird garantiert, dass mindestens 10 Zwiebeln keimen.
Der Lieferant der Blumenzwiebeln weiss auf Grund langjähriger Erfahrung, dass 6% der Zwiebeln dieser Sorte nicht keimen.

a) Formulieren Sie ein Wahrscheinlichkeitsmodell.
b) Wie gross ist die Wahrscheinlichkeit, dass eine Packung das Garantieversprechen nicht erfüllt?
c) Wie viele Packungen, welche die Garantie nicht erfüllen, sind auf eine Absatzmenge von 20'000 Packungen zu erwarten?

a) Wahrscheinlichkeitsmodell:
X = Anzahl Blumenzwiebeln, die keimen
X ist binomialverteilt mit n = 12, p = 0.94
b) $P(\text{Packung erfüllt Garantie nicht}) = P(X \le 9) = 0.03157 \approx 3.2\%$
`TR: Binomialcdf` mit `12, 0.94, 9`
c) $20\,000 \cdot 0.03157 = 631$ Packungen

a) Wahrscheinlichkeitsmodell:
X = Anzahl Blumenzwiebeln, die *nicht* keimen
X ist binomialverteilt mit n = 12, p = 0.06
b) $P(\text{Packung erfüllt Garantie nicht}) = P(X \ge 3) = P(X > 2) = 1 - P(X \le 2) = 0.03157 \approx 3.2\%$
`TR: Binomialcdf` mit `12, 0.06, 2` speichern unter `y`; dann `1-y`
c) $20\,000 \cdot 0.03157 = 631$ Packungen

Fluggesellschaften überbuchen üblicherweise ihre Flüge, da sie davon ausgehen, dass nicht alle Fluggäste den Flug auch tatsächlich antreten. Langfristige Beobachtungen zeigen, dass rund 4% aller Passagiere vom reservierten Flug zurücktreten.
Für einen Flug werden 385 Buchungen entgegengenommen, obwohl für das betreffende Flugzeug die maximale Auslastung 376 Passagiere beträgt. Mit welcher Wahrscheinlichkeit müssen Passagiere abgewiesen (bzw. auf andere Flüge umgebucht) werden?

Wahrscheinlichkeitsmodell:
X = Anzahl Personen, die den Flug *nicht* antreten
X ist binomialverteilt mit n = 385, p = 0.04
$P(X \le 8) = 0.028 = 2.8\%$
`TR: Binomialcdf` mit `385, 0.04, 8`

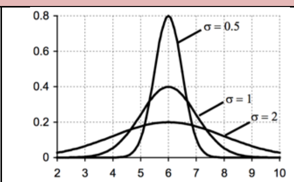
X = Anzahl der zum Flug antretenden Passagiere
X ist binomialverteilt mit n = 385, p = 0.96
$P(X \ge 377) = P(X > 376) = 1 - P(X \le 376) = 1 - 0.9720 = 0.028 = 2.8\%$
`TR: Binomialcdf` mit `385, 0.96, 376` speichern unter `y`; dann `1-y`

Normalverteilung X ~ N(μ, σ^2) → bei vielen verschiedenen Einflüssen

μ = Erwartungswert = Gipfel der Glockenkurve
σ = Standardabweichung = Streuung um den Erwartungswert μ
X = normalverteilte Zufallsvariable
Z = Standardnormalverteilte Zufallsvariable

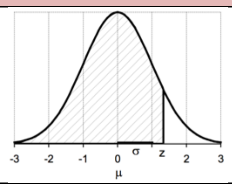
$$Z = \frac{X - \mu}{\sigma}$$

Im Intervall von μ-σ bis μ+σ liegen 68.27% der Fläche
Im Intervall von μ-2σ bis μ+2σ liegen 95.45% der Fläche
Im Intervall von μ-3σ bis μ+3σ liegen 99.73% der Fläche



Beispiel für Normalvert.
μ = 6 und σ = 0.5
μ = 6 und σ = 1
μ = 6 und σ = 2

Standardnormalverteilung



Standardnormalvert.
Immer mit den Werten
μ = 0 und σ = 1

TR: Berechnung von Wahrscheinlichkeit		**TR: Berechnung von z oder x** (Wahrscheinlichkeit gegeben)		
Standardnormalverteilung (μ = 0, σ = 1)	P(Z<z) = Normalcdf mit 0, 1, -1E99, z	Standardnormalverteilung (μ = 0, σ = 1)	P(Z<z) = p	Z = invNormal mit p, 0, 1
Normalverteilung mit μ und σ	P(X<x) = Normalcdf mit μ, σ, -1E99, x	Normalverteilung mit μ und σ	P(X<x) = p	Z = invNormal mit p, μ, σ
Normalverteilung mit μ und σ	P(a<X>b) = Normalcdf mit μ, σ, a, b			

Beispiele

Abfüllgewicht in Gramm / μ = 502.5, σ = 2 → a) wie viele % erfüllen nicht 500 b) zwischen 500 und 504 c) 502.5 und mehr d) 500.5 bis 504.5 e) 96% müssen min. 500 erfüllen

a) $P(X < 500) = 0.10565 = 10.6\%$
`TR: Normalcdf` mit `502.5, 2, -1E99 (oder 0), 500`
oder mit Standardisierung: $z = \frac{500 - 502.5}{2} = -1.25$
`TR: Normalcdf` mit `0, 1, -1E99, -1.25`
b) $P(500 \le X \le 504) = 0.6677 = 66.8\%$
`TR: Normalcdf` mit `502.5, 2, 500, 504`
c) 50%
d) Die Fläche von μ–σ bis μ+σ beträgt 0.6827, somit 68.3%.
e) Gesucht ist z so, dass $P(z) = 1 - 0.96 = 0.04$
$z = -1.75$ (`TR: invNormal` mit `0.04, 0, 1`)
$z = \frac{x - \mu}{\sigma} = -1.75$
also $\mu = x + 1.75 \cdot \sigma = 503.5$ Gramm

Betriebsstunden μA = 360, σA = 40 / μB = 370, σB = 28 a) je Teil nach 290 noch in Ordnung b) zusammen nach 290 nicht mehr i.O.

a) Teil A noch in Ordnung: $P(X_A \ge 290) = 0.95994$
`TR: Normalcdf` mit `360, 40, 290, 1E99`
Teil B noch in Ordnung: $P(X_B \ge 290) = 0.99786$
`TR: Normalcdf` mit `370, 28, 290, 1E99`
b) $P(\text{Maschine funktioniert immer noch}) = P(X_A \ge 290 \text{ und } X_B \ge 290)$
Da die beiden Ereignisse unabhängig sind, gilt:
$P(X_A \ge 290 \text{ und } X_B \ge 290) = P(X_A \ge 290) \cdot P(X_B \ge 290)$
$= 0.95994 \cdot 0.99786 = 0.95789$
$P(\text{Ausfall}) = 1 - 0.95789 = 0.04211$
Die Wahrscheinlichkeit für den Ausfall beträgt 4.2%.

Rechenregeln für Erwartungswert und Varianz

1. $E(X \pm Y) = E(X) \pm E(Y)$
2. $Var(X + Y) = Var(X) + Var(Y)$
   $Var(X - Y) = Var(X) + Var(Y)$
3. $E(a \cdot X) = a \cdot E(X)$
4. $Var(a \cdot X) = a^2 \cdot Var(X)$
5. $Var(X) = E(X^2) - (E(X))^2$

Der zentrlae Grenzwertsatz (zentrale Bedeutung der Normalverteilung)

Die Zufallsvariablen $X_1, X_2, \ldots, X_n$ seien unabhängig und identisch verteilt („independent and identically distributed“),
wobei $E(X_i) = \mu$ und $Var(X_i) = \sigma^2$.
Dann gilt für die Zufallsvariablen $T = X_1 + X_2 + \ldots + X_n$ sowie
$\bar{X} = \frac{T}{n} = \frac{X_1 + X_2 + \ldots + X_n}{n}$:

$E(T) = n \cdot \mu$ $\quad E(\bar{X}) = \mu$
$Var(T) = n \cdot \sigma^2$ $\quad Var(\bar{X}) = \frac{\sigma^2}{n}$

Zentraler Grenzwertsatz:

Sind $X_1, X_2, \ldots, X_n$ unabhängige Zufallsvariable, die alle dieselbe Verteilungsfunktion (und somit dasselbe μ und σ^2) besitzen, so lässt sich die Zufallsvariable

$$\bar{X} = \frac{X_1 + X_2 + \ldots + X_n}{n}$$

mit wachsendem n immer besser durch eine **normalverteilte** Zufallsvariable mit **Erwartungswert** μ und **Varianz (des Durchschnitts)** $\frac{\sigma^2}{n}$ approximieren.

Die entsprechende ***Standardabweichung des Durchschnitts*** $\frac{\sigma}{\sqrt{n}}$ wird als ***Standardfehler*** bezeichnet.

Konfidenzintervalle (die unbekannten Grössen der GG Mittelwert μ, Varianz σ^2, Anteilswert π aus den Stichprobenwerten Stichprobenmittelwert $\bar{x}$, Stichprobenvarianz s^2, relative Häufigkeit $\hat{p}$ schätzen)

Je grösser die Sicherheit, desto breiter der Vertrauensbereich
Je grösser die Standardabweichung, desto breiter der Vertrauensbereich
Je grösser die Stichprobe, desto schmaler der Vertrauensbereich

Vertrauensbereiche für arithmetische Mittelwerte (quantitative Merkmale)

(Punktschätzung: z.B $\bar{x}$ = μ → keine Hinweise darauf wie „genau" die Schätzung ist → daher wird Intervallschätzung verwendet)

z-Verteilung (σ bekannt → Grundgesamtheit)

Fehlerrisiko	α
Konfidenzniveau	1- α
Standardfehler	$\frac{\sigma}{\sqrt{n}}$

α	1–α	$z_{\alpha/2}$ = invNormal mit $1-\frac{\alpha}{2}$, 0, 1
0.317	0.683	1
0.05	0.95	**1.96**
0.046	0.954	2
0.01	0.99	**2.58**

Mit der Wahrscheinlichkeit 1–α enthält das Intervall von $\bar{x} - z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}$ bis $\bar{x} + z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}$ das arithmetische Mittel μ der Grundgesamtheit.*

Bei (relativ gesehen) grossen Stichproben (Faustregel : n > 5% von N) wird der Standardfehler $\frac{\sigma}{\sqrt{n}}$ durch $\frac{\sigma}{\sqrt{n}} \cdot \sqrt{\frac{N-n}{N-1}}$ ersetzt.

Das Intervall $\bar{x} \pm 1.96 \cdot \frac{\sigma}{\sqrt{n}}$ enthält μ mit einer Wahrscheinlichkeit von 95%.

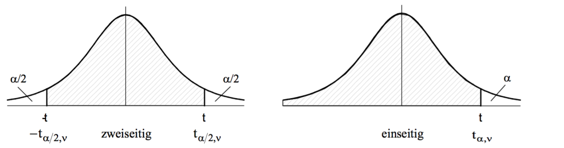
t-Verteilung (s bekannt → Stichprobe)

Anzahl Freiheitsgrade: v = n - 1

$$\bar{x} \pm t_{\alpha/2,v} \cdot \frac{s}{\sqrt{n}} \cdot \sqrt{\frac{N-n}{N-1}}$$

Korrekturfaktor nur wenn wenn n>5%

Quantile der t-Verteilung



t oder z? (t = z? wann?)

t ≈ z falls	n ≥ 100
t falls	n ≤ 100
t oder z falls	30 ≤ n ≥ 100
t falls σ unbekannt	

Beispiel

a) 95% Vertrauensintervall mit $\bar{x}$ = 959 und s = 414

a) 95% Vertrauensintervall: $t_{0.025,1199} \approx z_{0.025} = 1.96$

$$\left[959 - 1.96 \cdot \frac{414}{\sqrt{1200}}, 959 + 1.96 \cdot \frac{414}{\sqrt{1200}}\right] = [936, 982]$$

b) gewichtetes arithmetisches Mittel nach Schichtung $\bar{x}g$

b) $\bar{X}_g = 0.329 \cdot \bar{X}_1 + 0.328 \cdot \bar{X}_2 + ... + 0.078 \cdot \bar{X}_{5+}$

$E(\bar{X}_g) = E(0.329 \cdot \bar{X}_1 + 0.328 \cdot \bar{X}_2 + ... + 0.078 \cdot \bar{X}_{5+})$ gemäss Rechenregeln (Skrip

$E(\bar{X}_g) = 0.329 \cdot E(\bar{X}_1) + 0.328 \cdot E(\bar{X}_2) + ... + 0.078 \cdot E(\bar{X}_{5+})$

$E(\bar{X}_i)$ schätzen wir nun mit $\bar{x}_i$, d.h.

Schätzung von $E(\bar{X}_g)$: $\bar{x}_g = 0.329 \cdot 543 + ... + 0.078 \cdot 1494 = 895.69 \approx 896$ Franken

Haushaltsgrösse (Anzahl Personen)	Stichprobengrösse n_i	ar. Mittel $\bar{x}_i$	Standardabweichung s_i	prozentualer Anteil
1	323	543	130.9	32.9%
2	332	897	236.7	32.8%
3	187	1035	288.7	12.7%
4	221	1267	332.3	13.8%
5+	137	1494	413.4	7.8%

d) 95% Vertrauensintervall für Schätzung in b)

d) $\bar{X}_g = 0.329 \cdot \bar{X}_1 + 0.328 \cdot \bar{X}_2 + ... + 0.078 \cdot \bar{X}_{5+}$

$Var(\bar{X}_g) = Var(0.329 \cdot \bar{X}_1 + 0.328 \cdot \bar{X}_2 + ... + 0.078 \cdot \bar{X}_{5+})$

Nach den Regeln für das Rechnen mit Varianzen ergibt sich:
$Var(\bar{X}_g) = 0.329^2 \cdot Var(\bar{X}_1) + 0.328^2 \cdot Var(\bar{X}_2) + ... + 0.078^2 \cdot Var(\bar{X}_{5+})$

$Var(\bar{X}_1) = \frac{130.9^2}{323}; Var(\bar{X}_2) = \frac{236.7^2}{332}; ... ; Var(\bar{X}_5) = \frac{413.4^2}{137}$

$Var(\bar{X}_g) = 48.19$ und damit $\sqrt{Var(\bar{X}_g)} = 6.94$ (Schätzung des Standardfehlers von $\bar{X}_g$)

d.h. $\bar{X}_g \sim N(895.69; 48.19) \Rightarrow$

95% Vertrauensintervall: $[895.69 - 1.96 \cdot 6.94; 895.69 + 1.96 \cdot 6.94] = [882; 909]$

Breite des Intervalls in a) 982 – 936 = 46
Breite des Intervalls in d) 909 – 882 = 27 (Verkleinerung um 41% !)

Vertrauensbereiche für Anteilswerte (qualitative Merkmale)

Bedingung

$$n > \frac{9}{\hat{p} \cdot (1-\hat{p})}$$

Mit der Wahrscheinlichkeit 1–α enthält das Intervall von $\hat{p} - z_{\alpha/2} \cdot \sqrt{\frac{\hat{p} \cdot (1-\hat{p})}{n}} \cdot \sqrt{\frac{N-n}{N-1}}$ bis $\hat{p} + z_{\alpha/2} \cdot \sqrt{\frac{\hat{p} \cdot (1-\hat{p})}{n}} \cdot \sqrt{\frac{N-n}{N-1}}$ den Anteil π der Grundgesamtheit. (Korrekturfaktor $\sqrt{\frac{N-n}{N-1}}$ nötig , falls $\frac{n}{N} > 0.05$)

$\hat{p}$ = x/n

Stichprobengrösse

Quantitativ ($\sigma^2 = Sx^2$)	Qualitativ
$n \geq \frac{z_{\alpha/2}^2 \cdot \sigma^2}{F^2}$	$n \geq \frac{z_{\alpha/2}^2 \cdot \pi \cdot (1-\pi)}{F^2}$
σ ist entweder durch Studie bekannt, sonst 3-Sigma Regel: Max- Min = Spannweite → Spannweite / 6 = σ	π ist entweder durch Studie bekannt, sonst: π · (1− π) durch 0.25 ersetzen (worst case ist 0.25)
Max. Breite = B = F*2	Max. Fehler = F = B/2

Beispiel

450 Befragungen, 0.74 Ja-Anteil
a) 99% Vertrauensintervall

a) $n \cdot \hat{p} \cdot (1-\hat{p}) = 450 \cdot 0.74 \cdot 0.26 = 86.58 > 9$

$0.74 \pm 2.58 \cdot \sqrt{\frac{0.74 \cdot 0.26}{450}}$

[0.6867, 0.7933]

b) Welche Sicherheit gehört zum Vertrauensbereich 74 + / - 3%?

Normalcdf mit 0.74, $\sqrt{\frac{0.74 \cdot 0.26}{450}}$ = 0.0207, und 0.71, 0.77

0.853 = 85.3% Sicherheit

c) Wie viele Befragungen um bei einer Sicherheit von 95% ein Vertrauensintervall nicht breiter als 5 Prozentpunkte

c) 95%ige Sicherheit: $z_{\alpha/2} = 1.96$

maximaler Fehler $F = \frac{\text{Intervallbreite}}{2} = \frac{0.05}{2} = 0.025$

Die Umfrage liefert eine Schätzung für $\pi \cdot (1-\pi)$:

$\pi \cdot (1-\pi) \approx 0.74 \cdot 0.26$

$n \geq \frac{1.96^2 \cdot 0.74 \cdot 0.26}{0.025^2} = 1'182.6$

minimaler Stichprobenumfang: 1'183 Personen

Grundlagen und Begriffe

Skalierung

Nominal	Codierung (rationale Datenerfass.)	Ausprägung keine Reihenfolge	Nichtmetrisch qualitativ
Ordinal	+ Reihenfolge	Reihenfolge, keine Aussage Abstand	
Intervall	+ Abstände (+/- rechnen)	in Masseinheit gemessen	Metrisches quantitativ
Verhältnis	+ absoluter Nullpunkt	→ nur bei %-Veränderung sinnvoll	

diskret = Merkmal kann *nur bestimmte Werte* annehmen

stetig = Merkmal kann auf Skale *jeden Wert* annehmen

Gliederungstabellen (in jedem Tabellenfach eine Zahl o. Symbol)

-	anstelle einer Zahl bedeutet null (nichts).
0 als	Zahlenwert beträgt zwar mehr als "nichts", doch ist sein Wert kleiner die Hälfte der verwendeten Zähleinheit.
.	anstelle einer Zahl bedeutet, dass die Zahlenangabe nicht möglich ist, weil die begrifflichen Voraussetzungen fehlen.
...	Daten sind nicht erhältlich oder sind ohne Bedeutung oder sind aus anderen Gründen weggelassen.

Mittelwerte

Klassen cm von	bis	Klassen-mitte	Anz. Absolute Häufigkeit W_i	Klassen-breite	Rel. Häufig. w_i	rel. Häufig. kumuliert	Dichte
155	165	160	18	10	13.1 %	13.1 %	1.31 %
165	170	167.5	22	5	16.1 %	29.2 %	3.21 %
170	175	172.5	38	5	27.7 %	56.9 %	5.55 %
175	180	177.5	22	5	16.1 %	73.0 %	3.21 %
180	190	185	34	5	24.8 %	97.8 %	2.48 %
190	200	195	3	10	2.2 %	100 %	0.22 %

$$relative\ Häufigkeit\ (w_i) = \frac{absolute\ Häufikgkeit\ W_i}{gesamte\ Anzahl\ Werte}$$

rechtsschiefe Verteilung: Modus < Median < arithmetisches Mittel.
linksschiefen Verteilung: arithmetisches Mittel < Median < Modus
symmetrischen Verteilung: arithmetisches Mittel = Median = Modus

arithm. Mittel: 174.5 cm
Median: 173.75 cm (Data → L1: 170, 175 | L2: 29.2, 56.9 → LinReg → f(50) → 173.75cm)
1. Quartil: 168.7 cm (Data → L1: 165, 170 | L2: 13.1, 29.2 → LinReg → f(25) → 168.7 cm)
3. Quartil: 180.8 cm (Data → L1: 180, 190 | L2: 73, 97.8 → LinReg → f(75) → 180.8 cm
Quartilsabstand: 12.1 cm (3. Quartil – 1. Quartils)
Modalklasse: Modalklasse 170 bis 175cm (grösste Dichte)
Standardabweichung: 8.7 cm (1-Vart Stats)

Varianz: σ^2 →**Standardabweichung:** σ
Gesch. V: s^2 →**gesch. Stand.abwei.:** s
(1-Var Stats: Sx^2 / Sx / σx^2 / σx)
falls rel. Häuf. In L2 → nur σ

$$s = \sqrt{\frac{n}{n-1}} * \sigma$$

arithmetisches Mittel (gewichteter Durchschnitt)

arithmetische Mittel ist die Summe der Messwerte x_i dividiert durch die Anz. Messwerte

$$\bar{x} = \frac{Summe\ Messwerte\ x_i}{Anzahl\ der\ Messwerte\ (n)}$$

TR: Data → 1-Var Stats, L1 (Klassenmitte), L2 (abs. Häufigkeit W_i) → $\bar{x}$

Achtung bei %-Werten → Faktoren → evtl. GM

Median (Zentralwert) - Quantil

Merkmalswert, teilt der Grösse nach geordnete Merkmalswerte in zwei Hälften. links und rechts des Medians je 50% der der Grösse nach geordneten Werte.

$$Quantil = \text{Untergrenze der Q} - \text{Klasse} + \frac{X - rel.häuf.davor}{rel.häuf.Q - Klasse} * Klassenbreite$$

TR: Data → LinReg mit **L2** (relative Häufigkeiten kumuliert; letzter Wert unter X% und erster Wert über X%),
L1 (Klassenobergrenzen; die entsprechenden beiden Werte), ONE, YES → dann f(X) oder f(X/100)
1. Quartil: X = 25 / Median: X= 50 / 90%-Quantil: X = 90

Modus

*Merkmalsausprägung, welche in einer Messreihe **am häufigsten vorkommt**. Liegen die Messwerte in Klasseneinteilung oder Dichtefunktion vor, so tritt an die Stelle des häufigsten Wertes die Klasse mit der grössten Dichte.*

$$Dichte = \frac{relative\ Häufigkeit\ w_i}{Klassenbreite}$$

Testen von Hypothesen

Verwerfungsbereich allgemein: TR L3 = x / L1 = n-L3 Binomialcdf mit Option List, n, 1-π und L1 unter L2 speichern Verwerfungsbereich startet dort wo der Wert kleiner als α	π = 0.02 / n = 400 400*0.02 = 8 α = 5%	TR L3 = 9, 10, 11, 12...usw. / L1 = 400- L3 Binomialcdf mit Option List, 400, 0.98 und L1 unter L2 speichern Verwerfungsbereich startet bei Wert ≤ 0.05

Tests für Anteilswerte (Qualitatives Merkmal)

Einstichprobentest

Nullhypothese H_0 (X = Anzahl „Erfolge" in n Versuchen)

H_0: $X \sim B(n, \pi_0)$ oder kürzer H_0: $\pi = \pi_0$

Alternativhypothese H_A

$\pi \neq \pi_0$ zweiseitiger Test

$\pi > \pi_0$ bzw. $\pi < \pi_0$ einseitiger Test

Bedingung Stichprobe:

$$n > \frac{9}{\pi_0 \cdot (1 - \pi_0)}$$

Berechnung Testgrösse z:

$\hat{p} = \frac{x}{n}$ Anteil des Merkmals in der Stichprobe

$$z = \frac{\hat{p} - \pi_0}{\sqrt{\pi_0 \cdot (1 - \pi_0)/n}}$$

Verwerfungsbereich:

„Sicherheit" $(1-\alpha)$	Konfidenzintervalle und *zweiseitige* Tests	*ein*seitige Tests
95%	$z_{2.5\%} = 1.96$	$z_{5\%} = 1.64$
99%	$z_{0.5\%} = 2.58$	$z_{1\%} = 2.33$

Zweistichprobentest

Nullhypothese H_0

X_1 = Anzahl „Erfolge" in n_1 Versuchen, $X_1 \sim B(n_1, \pi_1)$

X_2 = Anzahl „Erfolge" in n_2 Versuchen, $X_2 \sim B(n_2, \pi_2)$

X_1 und X_2 voneinander unabhängig

H_0: $\pi_1 = \pi_2 \ (= \pi)$

Alternativhypothese H_A

$\pi_1 \neq \pi_2$ zweiseitiger Test

$\pi_1 > \pi_2$ bzw. $\pi_1 < \pi_2$ einseitiger Test

Stichprobenumfänge n_1, n_2

$\hat{p}_1 = \frac{x_1}{n_1}$ Anteil des Merkmals in der 1. Stichprobe

$\hat{p}_2 = \frac{x_2}{n_2}$ Anteil des Merkmals in der 2. Stichprobe

$\hat{p} = \frac{x_1 + x_2}{n_1 + n_2} = \frac{\hat{p}_1 \cdot n_1 + \hat{p}_2 \cdot n_2}{n_1 + n_2}$ Schätzung für π

Bedingung Stichprobe:

$(n_1 + n_2) \cdot \hat{p} \cdot (1 - \hat{p}) > 9$ sowie $n_1 \geq 50, n_2 \geq 50$

Berechnung Testgrösse Z und Realisation z:

$$Z = \frac{\frac{X_1}{n_1} - \frac{X_2}{n_2}}{\sqrt{\pi \cdot (1-\pi) \cdot \left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} \qquad z = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\hat{p} \cdot (1-\hat{p}) \cdot \left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

Verwerfungsbereich gleich wie oben

Beispiel Anteilswert Test (Einstichprobe und einseitig)

Postchef sagt, 92% werden innerhalb von 7 Minuten bedient / Stichprobe n = 200 / 22 warteten länger als 7 Minuten / Stimmt Aussage Postchef?

Nullhypothese H_0: Der Anteil π der Kunden, die am Schalter länger als 7 Minuten warten müssen, beträgt 8%. $\pi = 0.08$

Alternativhypothese H_A: Der Anteil p der Kunden, die am Schalter länger als 7 Minuten warten müssen, ist grösser als 8%. $\pi > 0.08$ einseitiger Test (Abweichung nach oben wird vermutet)

Bedingung für Approximation mit Normalverteilung prüfen:

$$200 \cdot 0.08 \cdot 0.92 = 14.72 > 9 \checkmark$$

Anteil in Stichprobe: $\hat{p} = \frac{22}{200} = 0.11$

$$z = \frac{(\hat{p} - \pi_0) \cdot \sqrt{n}}{\sqrt{\pi_0 \cdot (1 - \pi_0)}} = \frac{(0.11 - 0.08) \cdot \sqrt{200}}{\sqrt{0.08 \cdot 0.92}} = 1.56$$

einseitiger Verwerfungsbereich $\alpha = 5\%$ $z_{5\%} = 1.64$

Feststellung: $1.56 < 1.64$ H_0 kann nicht verworfen werden. Die Aussage des Postchefs wird durch das beobachtete Ergebnis statistisch nicht widerlegt.

Beispiel Anteilswert Test (Zweistichprobe und zweiseitig)

Stimmen Zeitpunkt t_1 = 156 n_1 = 360

Stimmen Zeitpunkt t_2 = 199 n_2 = 510

Nullhypothese H_0: Anzahl Stimmen Zeitpunkt 1: $X_1 \sim B(n_1, \pi_1)$

Anzahl Stimmen Zeitpunkt 2: $X_2 \sim B(n_2, \pi_2)$

X_1 und X_2 sind voneinander unabhängig.

Die Wähleranteile sind in beiden Zeitpunkten gleich gross.

$\pi_1 = \pi_2$

Alternativhypothese H_A: Die Wähleranteile sind in beiden Zeitpunkten unterschiedlich.

$\pi_1 \neq \pi_2$

zweiseitiger Test

Stichproben $n_1 = 360$ $\hat{p}_1 = \frac{156}{360} = 0.43333$

$n_2 = 510$ $\hat{p}_2 = \frac{199}{510} = 0.39020$

$$\hat{p} = \frac{\hat{p}_1 \cdot n_1 + \hat{p}_2 \cdot n_2}{n_1 + n_2} = \frac{0.43333 \cdot 360 + 0.3902 \cdot 510}{360 + 510} = \frac{156 + 199}{870} = 0.40805$$

Approximation durch Normalverteilung zulässig, da

$(360 + 510) \cdot 0.40805 \cdot (1 - 0.40805) = 210.1 > 9 \ \wedge \ n_1 \geq 50, n_2 \geq 50 \checkmark$

$$z = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\hat{p} \cdot (1-\hat{p}) \cdot \left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} = \frac{0.43333 - 0.3902}{\sqrt{0.40805 \cdot 0.59195 \cdot \left(\frac{1}{360} + \frac{1}{510}\right)}} = 1.28$$

zweiseitiger Verwerfungsbereich $\alpha = 5\%$: $z_{2.5\%} = 1.96$

Feststellung: $1.28 < 1.96$

H_0 kann nicht verworfen werden. Man wird weiterhin davon ausgehen, dass die Wähleranteile zu den Zeitpunkten 1 und 2 gleich gewesen sind.

Tests für arithmetische Mittelwerte (Quantitatives Merkmal)

Einstichprobentest → σ unbekannt (t-Test)

Nullhypothese H_0 (arit. Mittel μ signifikant verschieden von vermutetem Wert μ_0)

H_0: $\mu = \mu_0$

Alternativhypothese H_A

$\mu \neq \mu_0$ zweiseitiger Test

$\mu > \mu_0$ bzw. $\mu < \mu_0$ einseitiger Test

Grösse n

arithmetisches Mittel $\bar{x}$

Standardabweichung s (σ_0 durch s geschätzt)

Berechnung Testgrösse t:

$$z \approx t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}} \quad (z \approx t \text{ gilt ab } n \geq 100)$$

Verwerfungsbereich: siehe Liste Quantile der t-Verteilung

Gepaarter Stichprobentest (Vorher-Nachher-Vergleich)

Das gleiche quantitate Merkmal wird zweimal in einer Beobachtungseinheit gemessen im Zustand A und B der Beobachtungseinheit.

Weicht der Erwartungswert des Merkmals im Zustand A signifikant vom Erwartungswert des Merkmals im Zustand B ab?

Differenzen berechnen, Einstichprobentest anwenden, testen ob das arit. Mittel signifikant von null abweicht

Beispiel arit. Mittel Test (Einstichprobe und einseitig)

Erwartung 2000h, Durchschnitt 2028h, s = 187h, n=100, α = 5%

Nullhypothese H_0: $\mu = 2000$ gleiche durchschnittliche Brenndauer

Alternativhypothese H_A: $\mu > 2000$ einseitiger Test (Abweichung nach oben wird angestrebt)

t-Test mit $n = 100$, $\bar{x} = 2028$, $s = 187$ liefert einen t-Wert von $t = \frac{2028 - 2000}{187/\sqrt{100}} = 1.497$.

Der kritische t-Wert ist $t_{0.05,99} = 1.660$ (s. S. 48).

$t < t_{0.05,99}$: Die Nullhypothese kann nicht verworfen werden.
Eine Verlängerung der durchschnittlichen Brenndauer lässt sich nicht signifikant nachweisen.

Variante mit der Standardnormalverteilung:

$$z \approx t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}} = \frac{2\,028 - 2\,000}{187/\sqrt{100}} = 1.50$$

$z_{5\%} = 1.64$

$z < z_{5\%}$: Die Nullhypothese kann nicht verworfen werden.
Eine Verlängerung der durchschnittlichen Brenndauer lässt sich nicht nachweisen.

Beispiel arit. Mittel Test (gepaarter Stichprobentest)

n = 10, Reifen mit Profil A oder B, α = 5%

Profil A [m]	44.5	55.0	52.5	50.2	45.3	46.1	52.1	50.5	50.6	49.2
Profil B [m]	44.9	54.8	55.6	55.2	55.6	47.7	53.0	49.1	52.3	50.7
Differenz: A–B [m]	–0.4	+0.2	–3.1	–5.0	–10.3	–1.6	–0.9	+1.4	–1.7	–1.5

Nullhypothese H_0: Differenzen $D_i \sim N(0, \sigma^2)$, σ^2 unbekannt, somit $\mu = E[\bar{D}] = 0$

Alternativhypothese H_A: $\mu \neq 0$ (zweiseitiger Test)

t-Test mit $n = 10$, $\bar{d} = -2.29$, $s = 3.3154$ liefert $t = \frac{(-2.29 - 0) \cdot \sqrt{10}}{3.3154} = -2.184$

Der kritische t-Wert liegt für $\nu = 10 - 1 = 9$ bei $t_{0.025,9} = 2.262$ (s. Tabelle auf S. 48).

Damit ist $|t| < t_{0.025,9}$, also kann die Nullhypothese nicht verworfen werden.

Die Reifen unterscheiden sich bezüglich Bremswirkung nicht signifikant.

Bemerkung:
einseitig signifikant (A ist besser als B), da $t_{0.05,9} = 1.833$ (s. Tabelle auf S. 48).

Chi-Quadrat Unabhängigkeitstest (wird benutzt um zu testen, ob zwei qualitative Merkmale voneinander unabhängig sind)

v = Anzahl Freiheitsgrade = (Anzahl Zeilen – 1) · (Anzahl Spalten – 1) Wahrscheinlichkeiten p_i und p_J / n = Stichprobengrösse	f_{ij} = empirisch beobachtete Häufigkeiten f'_{ij} = erwartete Häufigkeiten / $f'_{ij} = p_i \cdot p_J \cdot n$
Bedingungen: n ≥ 50, v > 1	$f'_{ij} \geq 5$

Vorgehen:
- drei Bedingungen prüfen (oben auf dieser Seite)
- Null- und Alternativhypothese ausformulieren
- theoretische Verteilung berechnen (ab Seite 70)
- berechneten Chi-Quadrat-Wert mit dem kritischen Wert auf Seite 69 vergleichen und entscheiden (kann die Nullhypothese verworfen werden oder nicht?)

Die Nullhypothese wird nun geprüft mit der Testgrösse

$$\chi^2 = \sum_{i=1}^{k} \sum_{j=1}^{m} \frac{(f_{ij} - f'_{ij})^2}{f'_{ij}}$$

Beispiel

Abhängigkeit sportliche Betätigung und Berufsgruppe mit Signifikanzniveau 1% (Tabelle 1 rot = gegeben, Rest wird berechnet)

Nullhypothese H_0: Die sportliche Betätigung ist von der Berufsgruppe unabhängig.

Alternativhypothese H_A: Die sportliche Betätigung ist von der Berufsgruppe abhängig.

Signifikanzniveau: $\alpha = 1\%$

Berufsgruppe	sportliche Betätigung: nie	gelegentl.	regelm.	Total	rel. Häuf. = p_i
selbstständig	150	50	100	300	0.30
angestellt	400	150	150	700	0.70
Total	550	200	250	1000	
rel. Häuf. = q_j	0.55	0.20	0.25		

Interpretiert man die relativen Häufigkeiten als Schätzwerte für die Wahrscheinlichkeiten p_i und q_j, so ergeben sich im Fall der Unabhängigkeit der beiden Merkmale nach dem Multiplikationssatz für unabhängige Ereignisse die folgenden Wahrscheinlichkeiten.

Tabelle 2 berechnen mit $p_i \cdot p_J$

Berufsgruppe	sportliche Betätigung: nie	gelegentl.	regelm.
Arbeiter	0.165	0.060	0.075
Angestellter	0.385	0.140	0.175

Tabelle 3 berechnen mit Werten aus Tabelle 2 mal n

Mit Hilfe dieser Wahrscheinlichkeiten lassen sich nun die theoretischen (erwarteten) absoluten Häufigkeiten f'_{ij} berechnen.

Berufsgruppe	sportliche Betätigung: nie	gelegentl.	regelm.
Arbeiter	165	60	75
Angestellter	385	140	175

Die drei Bedingungen sind erfüllt: $n \geq 50$, $f'_{ij} \geq 5$, $\nu > 1$

$$\chi^2 = \frac{(150-165)^2}{165} + \frac{(50-60)^2}{60} + \frac{(100-75)^2}{75} + \frac{(400-385)^2}{385} + \frac{(150-140)^2}{140} + \frac{(150-175)^2}{175} = 16.23$$

$\alpha = 1\%$ $\quad \nu = (2-1)\cdot(3-1) = 1 \cdot 2 = 2$

χ^2-Tabelle: $\chi^2_{1\%} = 9.21$ Feststellung: $16.23 > 9.21$

Interpretation: H_0 verwerfen.

Der Grad der sportlichen Betätigung ist hochsignifikant von der Berufsgruppe abhängig.

TR:
Werte aus Tabelle 1 in L1 (150, 50, 100, 400, 150, 150)
Werte aus Tabelle 3 in L2 (165, 60, 75, 385, 140, 175)
L3 = $(L1 – L2)^2$ / L2
1-Var Stats (L3, One) → ∑x = 16.23

Vorzeichentest

Wertepaare x_i, y_i / Unterschied Wahrscheinlichkeit dass die Differenz $x_i – y_i$ positiv bzw. negativ ist?

Nullhypothese H_0: Wahrscheinlichkeiten sind gleich (je 0.5)	Alternativhypothese H_A: Wahrscheinlichkeiten sind unterschiedlich

Testgrösse
k = Anzahl positive Differenzen
if n > 36, Testgrösse = z und binomialverteilung
if n ≤ 36, Testgrösse = k und normalverteilung

$$z = \frac{k - \frac{1}{2}n}{\frac{1}{2}\sqrt{n}}$$

Signifikanzniveau α wählen. H_0 verwerfen, wenn $|z| > z_{kritisch}$ bzw. wenn k in den, mit Hilfe der binomischen Verteilung für p = 0.5 und n berechneten Verwerfungsbereich fällt.

Beispiel

H_A: Der Median der Differenzen „nachher – vorher" ist positiv. (einseitiger Test)

H_0: Der Median der Differenzen ist null (gleich viele positive wie negative Differenzen). $\pi = 0.5$ (Wahrscheinlichkeit einer positiven Differenz)

n = 16 ≤ 36 → X = Anzahl positiver Differenzen ist binomial verteilt, kurz X ~ B(16,0.5)

Testgrösse k = Anzahl positiver Differenzen = 13

$\alpha = 5\%$

Verwerfungsbereich V: $P(V) = P(X \geq x) < 5\,\%$

$\Leftrightarrow P(X \leq n–x) < 5\%$ mit X ~ B(16, 0.5)

x	n–x	Binomialcdf mit 16, 0.5, n–x
10	6	0.22725
11	5	0.10506
12	4	0.03841

Also V = {12, 13, … ,16}

k = 13 ∈ V: H_0 verwerfen, der Median der Differenzen ist signifikant grösser null.
Die Ladenwerbung hat eine signifikante Wirkung gezeigt.

Bemerkungen:
(1) Es ist viel einfacher, anstatt aufwändig den Verwerfungsbereich nur den p-Wert zu bestimmen: P(X≥13) = Binomialcdf mit 16, 0.5, 16–13: 1.1% < 5%
(2) Wäre z.B. n=100 und k=60 gewesen, würde man mit dem Z-Test für Anteilswerte rechnen:

$z = \frac{0.6-0.5}{\sqrt{0.5\cdot 0.5/100}} = 2$ $\quad 2 > 1.64$ $\quad H_0$ verwerfen

oder

p = Normalcdf mit 0, 1, 2, 1E99 = 0.023 = 2.3%
2.3% < 5% $\quad H_0$ verwerfen

Streuungsmasse

Spannweite	Maximum – Minimum (einfachste Messung einer Streuung, V: Einfachheit, N:es werden nur die beiden Extremwerte berücksichtigt)
Quartilsabstand	3. Quartil – 1. Quartil (Quartilsabstand misst also die Spannweite der mittleren 50% der Messwerte.)
Standardabweichung	durchschnittliche Entfernung aller gemessenen Ausprägungen eines Merkmals vom Durchschnitt →Wurzel von Varianz ist Standardabweichung
Varianz	durchschnittliche quadratische Abweichung
Variationskoeffizient	Will man Streuungen von versch. Merkmalen mit untersch. Grössenordnungen vergleichen, kann man die Streuung im Verhältnis zum Mittelwert angeben

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

$\frac{Standardabweichung}{Arritmetische\ Mittel} = \frac{s}{\bar{x}} = v_\sigma$ Lösungssatz: Grösse Variation des Körpergewichts als der Körpergrösse (KGrösse-v: 9.11%, KGweicht-v: 4.43%)

Korrelation

x-Werte in L1, y-Werte in L2:

Korrelationskoeffizient r $\quad$ LinReg [ax+b] L1, L2, ONE (ggf. L3), YES
r ablesen

r = 0.62	Es besteht ein mittlerer positiver Zusammenhang (resp. ein positiver Zusammenhang mittlerer Stärke) zwischen der Anzahl Kundenkontakte und der Anzahl Abschlüsse.
r = 0.83	Der Zusammenhang kann bestätigt werden
r = – 0.11	Es besteht scheinbar kein (resp. nur ein sehr schwacher) Zusammenhang zwischen Baustärke und Stabilität.

Beurteilung Korrelation

r (lin. Zusammenhang)	*Stärke des Zusammenhang*	R^2
Bis 0.25	kein nachweisbarer funkt. Zusammenhang	5%
Bis 0.5	Schwach	25%
Bis 0.75	Mittel	50%
1	stark	100%

Regression

Zusammenhang zwischen zwei oder mehr Variablen an. Bei der Regressionsanalyse wird *vorausgesetzt*, dass es einen *gerichteten linearen Zusammenhang* gibt, das heißt, es existieren eine *abhängige Variable* und mindestens eine *unabhängige Variable*. Mit Hilfe der Regressionsanalyse kann eine *Regressionsfunktion* errechnet werden, welche die *Anhängigkeit der beiden Variablen mit einer Geraden beschreibt*. Die ermittelte Regressionsgerade erlaubt es, *Prognosen für die abhängige Variable zu treffen*, wenn ein Wert für die unabhängige Variable eingesetzt wird. Umgekehrte Rückschlüsse sind nicht zulässig.

Methode der kleinsten Quadrate

Lineares Modell	$y = ax + b$	LinReg ax+b (4)
Polynom-Model	$y = ax^2 + bx + c$	QuadraticReg (5)
	$y = ax^3 + bx^2 + cx + d$	CubicReg (6)
Logarithmisches Modell	$y = a + b * \ln(x)$	LnReg a+blnx (7)
Potenz-Modell	$y = a * x^b$	PwrReg ax^b (8)
Exponentielles Modell	$y = a * b^x$	ExpReg ab^x(9)

Bestimmtheitsmass R^2

→ erklärten Anteil der Variabilität (Varianz) einer abhängigen Variablen durch ein statistisches Modell
erklärbare Varianz = R^2; nicht erklärbare Varianz = $1-R^2$;
totale Varianz =erklärbare Varianz + nicht erklärbare Varianz
TR: „LinReg" (bzw. passende Gleichung) → r^2 ablesen
$R^2 = 1$: Der vermutete Zusammenhang trifft absolut zu.
$R^2 = 0$: Der vermutete Zusammenhang trifft absolut *nicht* zu.