Begriffe	
Statistik	Entwicklung und Anwendung von Methoden zur Erhebung, Aufbereitung, Analyse und Interpretation von Daten
Beschreibende Statistik	Vollständige Kenntnis über das Untersuchungsobjekt
Schliessende Statistik	Für Untersuchung liegend die Daten des zu untersuchenden Objekts nur zum Teil vor.
Hypothese	Eine Hypothese ist eine Aussage deren Gültigkeit man für möglich hält, die aber nicht bewiesen oder verifiziert ist.
Nullhypothese	Die Nullhypothese H0 ist eine Aussage von der angenommen wird, dass sie stimmt.
Alternativhypothese	Die Alternativhypothese H1 beschreibt eine Annahme, sie ist also das Gegenteil der Nullhypothese.
Fehler 1. Art (alpha)	Fehlerhaftes Verwerfen einer Hypothese
Fehler 2. Art (beta)	Fehlerhaftes Annehmen einer Hypothese
Zufällige Fehler	Nicht reproduzierbar
Systematische Fehler	Reproduzierbar (können vermieden werden, unterliegt keinen grossen Schwankungen)
Validierung	Mache ich das Richtige (Überprüfung des Modells)
Verifikation	Mache ich es richtig (Verifiziertes Modell kann nicht valide sein)
Merkmalsträger	Der Gegenstand der statistischen Untersuchung
Abgrenzungsmerkmal (sachlich, räumlich, zeitlich)	Sachlich: wer/was ist unter Merkmalsträger zu verstehen z.B. Wer gilt als „Mitarbeiter" eines Unternehmens
	Räumlich: Räumliche Grenzen, in denen der Merkmalsträger liegen muss z.B. ein Bürogebäude eines Konzerns
	Zeitlich: Zeitpunkt oder Zeitraum, an der ein Merkmalsträger „existieren" muss, um Teil der Grundgesamtheit zu sein ≠ zum Zeitpunkt der Messung/Erhebung!
Grundgesamtheit	Die Menge aller Merkmalsträger die für eine Untersuchung in Frage kommen
Merkmal	Eigenschaften der Merkmalsträger die von Interesse sind
Merkmalswert	Der Wert der Beobachtung / Messung
Primärstatistik	Die Daten wurden genau für diesen Zweck erhoben (teuer)
Sekundärstatistik	Existierende Daten wobei es ungewiss ist, wie die Daten erhoben wurden. (günstig)
Vollerhebung	Befragung aller Merkmalsträger (Kosten und Umfang meist zu gross)
Teilerhebung	Befragung der essentiellen Merkmalsträger (wird meist gemacht)
Diskrete Funktion	Mit Lücken
Stetige Funktion	Ohne Lücken
Formale Abhängigkeit	Zahlenmässig begründete Abhängigkeit
Sachliche Abhängigkeit	Ist der Wert eines Merkmals kausal/ursachlich für den Wert eines zweiten Merkmals abhängig
Menge	Ungeordnet, ohne Redundanzen
Tupel	Geordnet, mit Redundanzen
Zufallsexperiment	Ein Experiment welches beliebig oft durchgeführt werden kann und das Ergebnis komplett vom Zufall abhängig ist
Disjunkt	Keine gemeinsame Teilmenge
Zielgrösse	Beschreiben die Grösse, die man optimieren möchte
Einflussgrösse	Sind Grössen welche die Zielgrösse beeinflussen. Es wird zwischen Streu,- und Störgrössen unterschieden. Man unterscheidet zwischen Steuergrössen und Störgrössen
Steuergrössen	Eine einstellbare Grösse (die man auch für eine gewisse Zeit halten kann)
Störgrössen	Eine Grösse deren Wert man nicht beeinflussen kann
Faktoren	Aus allen Einflussgrössen werden die wesentlichen/relevant Faktoren genannt. Es wird zwischen Quantitativen und Qualitativen Faktoren unterschieden:
Quantitative Faktoren	Quantitative Faktoren: Die Werte sind auf einer Ordinalskala beschrieben
Qualitative Faktoren	Qualitative Faktoren: Die Werte sind auf einer Nominalskala beschrieben
Faktorstufen	Die Werte die ein Faktor in einem Versuch annehmen soll, werden Faktorstufen genannt. Kann ein Faktor nicht genau gemessen werden, so sollte der Abstand der Faktorstufen mindestens 6x die Varianz sein
Komplexität	Hohe Anzahl an Faktoren
Kompliziertheit	Unbekannte oder schwierig zu beschreibende Faktoren

Symbole	
h_i	Absolute Häufigkeit (Anzahl)
f_i	Relative Häufigkeit (Anteil)
H_i	Kumulierte absolute Häufigkeit
F_i	Kumulierte relative Häufigkeit
μ	Mittelwert
σ^2	Varianz
σ	Standardabweichung
$\bar{x}$	Arithmetisches Mittel (Durchschnitt)
H_0	Nullhypothese
H_1	Alternativhypothese
ω	Elementarereignis → Teilmenge der Ergebnismenge
Ω	Ergebnismenge / Ergebnisraum (Menge aller möglicher Ausgänge eines Zufallsexperiments)
σ-Algebra	Siehe Wahrscheinlichkeiten
$\mathcal{A}$	Besteht aus allen möglichen Ergebniskombinationen (Potenzmenge der Ergebnismenge)
n	Anzahl Messungen / Stichprobenumfang
N	Grösse der Grundgesamtheit
Δ	Mittlere absolute Abweichung
R	Spannweite

Ableitungsregeln					
x^a	$a \cdot x^{a-1}$	1	0	$\tan(x)$	$\frac{1}{\cos^2(x)}$
x	1	a^x	$\ln(a) \cdot a^x$	$\tan(x)$	$1+\tan^2(x)$
x^2	$2x$	$\ln(x)$	$\frac{1}{x}$	$\arcsin(x)$	$\frac{1}{\sqrt{1-x^2}}$
$\frac{1}{x}$	$-\frac{1}{x^2}$	$\log_b(x)$	$\frac{1}{\ln(b) \cdot x}$	$\arccos(x)$	$-\frac{1}{\sqrt{1-x^2}}$
$\sqrt{x}$	$\frac{1}{2\sqrt{x}}$	$\sin(x)$	$\cos(x)$	$\arctan(x)$	$\frac{1}{1+x^2}$
e^x	e^x	$\cos(x)$	$-\sin(x)$		

Linearitätsregel	$\frac{d}{dx}(f(x)+g(x)) = f'(x)+g'(x)$
Produktregel	$\frac{d}{dx}(f(x)\cdot g(x)) = f'(x)\cdot g(x) + f(x)\cdot g'(x)$
Produktregel mit Konstante c	$\frac{d}{dx}(c \cdot f(x)) = c \cdot f'(x)$
Kettenregel	$\frac{d}{dx}(f(g(x))) = f'(g(x)) \cdot g'(x)$
Quotientenregel	$\frac{d}{dx}\left(\frac{f(x)}{g(x)}\right) = \frac{f'(x)\cdot g(x) - f(x)\cdot g'(x)}{(g(x))^2}$

Experimente

Zyklischer Prozess des Experimentierens nach Shewhard (Plan → Do → Check → Act)

1. Hypothese aufstellen
2. Experiment durchführen
3. Hypothese überprüfen
4. Hypothese/Modell gegebenenfalls anpassen

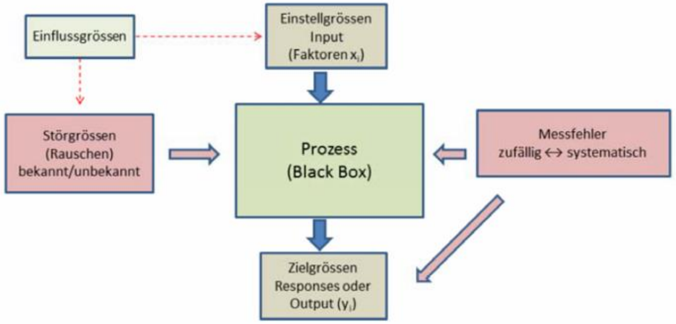
Hindernisse für den Erkenntnisgewinn:

1. Komplexität: Hohe Anzahl an Faktoren
2. Kompliziertheit: Unbekannte oder schwierig zu beschreibende Faktoren
3. Rauschen/Dynamik: Unterschiedliche Ergebnisse bei gleichen Faktoren

Experimente werden immer nach einem bestimmen Schema durchgeführt:

1. Ausgangssituation beschreiben
2. Untersuchungsziele festlegen / Zielgrössen definieren
3. Faktoren auswählen und gewichten
4. Versuchsplanung erstellen
5. Versuche durchführen
6. Ergebnisse auswerten und Vertrauensintervalle bestimmen
7. Ergebnisse interpretieren und Massnahmen ableiten
8. Überprüfen der «Verbesserungen»

Prozessmodell:



DoE: Design of Experiment

Wie sind Experimente zu planen, damit mit möglichst wenigen Einzelexperimenten der Zusammenhang zwischen Einflussfaktoren und Zielgrössen möglichst genau ermittelt werden können.

Vorgehen

1. Ausgangssituation spezifizieren / Problem beschreiben / Ziel definieren
   a. Kunde und dessen Bedürfnisse definieren
   b. Liegen bereits Daten vor
   c. Welche Probleme müssen gelöst werden
   d. Welche Ressourcen (Zeit und Geld) stehen zur Verfügung (Kosten/Nutzen Analyse)
   e. Betroffene Gruppen und deren Beziehung untereinander listen (Wiederstände, Supporter, Wissensträger)
2. Zielgrösse beschreiben: Dabei möglichst alle Grössen sammeln und diese dann auf die wichtigen Reduzieren

Einfluss-Zielgrössen-Matrix

1. Für jede Zielgrösse eine Spalte anlegen
2. In der ersten Spalte alle Einflussgrössen sammeln und in Einflussgrössen und Steuergrössen unterteilen
3. Für jede Einflussgrösse das vorhandenen Wissen über Grösse und Einfluss auf jede Zielgrösse sammeln (z.B. stark, schwach, kein, linear, nicht linear)

Fehlerrechnung:

Die Fehlerrechnung wird benötigt um den Bereich abzuschätzen, in denen der tatsächliche Wert mit einer gewissen Wahrscheinlichkeit liegt.	
Zufällige Fehler	Nicht reproduzierbar
Systematische Fehler	Reproduzierbar
Absoluter Fehler Δt	Bei Summen und Differenzen addieren sich die absoluten Fehler (gleiche Einheit wie Messwert)
	Der absolute Fehler kann mithilfe der relativen Fehlers berechnet werden → Δt = relativer Fehler * Wert
Relativer Fehler	$\frac{\Delta t}{t}$ wobei Δt = absolute Fehler und t = Messwert Bei Produkten und Quotienten addieren sich die relativen Fehler (einheitenlos → %)
	Mit Hilfe des relativen Fehler lässt sich gut Abschätzen, welcher Faktor verbessert werden sollte. (der mit dem grösseren Fehleranteil)
	Bei Potenzen kann der relative Fehler mit dem Exponenten multipliziert werden. Z.B $r^2 * \pi$ → $2 * f_r$ da sich der relative Fehler bei Multiplikationen addiert ($r * r * \pi$ → $f_r + f_r = 2 * f_r$)

Nennwert die Fehlerangabe ergänzen

Mindestens	Letzte Stelle des Messwertes + 1 Stelle (auf halbe gerundet)		
Höchstens	Letzte Stelle des Messwertes (auf 0.3/0.4 gerundet)		
Beispiel	*Gemessener Wert (t)*	***Δ t** von*	***Δ t** bis*
	15.32s	± 0.005s	± 0.04s
	15.3s	± 0.05s	± 0.4s
	15.320s	± 0.0005s	± 0.004s

Beispiele (Masseinheiten beachten)

Schätzung eines Rechtecks

Länge wird abgelesen:	28.15 - 22.35 cm = 5.8 cm
Fehlerschätzung beim Ablesen:	± 0.05 cm
//Subtraktion: absolute Fehler addieren sich (links und rechts)	
Länge des Rechtecks:	5.8 ± 0.1 cm

Berechnung der Fläche

Breite des Rechtecks gegeben mit	0.9 ± 0.1 cm
Berechnung des relativen Fehlers	ΔB = 0.1/0.9 = 11.1%
Relativer Fehler der Länge:	ΔL = 0.1/5.8cm = 1.7%
//Multiplikation: relative Fehler addieren sich	
ΔB + ΔL = 12.8%	
Fläche A = L * B	A = 0.9 * 5.8 = 5.2 ± 0.7 cm2
absoluter Fehler der Fläche	0.128 * 5.22 cm2 = 0.668 cm2

Bei Messgeräten ist der relative Fehler nicht auf den gemessenen Wert, sondern auf Messbereich bezogen.

Tipp für Rechnungen mit Kombinationen von +/- und */:

1. Resultat berechnen ohne beachten der Fehlerangaben
2. Resultat berechnen unter Nutzen der Maximalwerte
3. Fehler Δ ergibt sich durch die Differenz von 1. und 2.

Diagramme

Balkendiagramm	Y-Achse: Häufigkeit und X-Achse: Balken pro Klasse
Histogramm	Der Balken geht über die gesamte Klassenbreite
Polygonzug	Verbinden der Balken mit einer Linie, wobei jeweils der rechte Ecken verbunden wird. Beim Balkendiagramm wird die Mitte genommen.

Skalen

Wir arbeiten hauptsächlich mit metrischen Skala (Intervall und Verhältnis)

Nominalskala (qualitativ)	Sind zwei Einheiten gleich oder ungleich? $= / \neq$ Enthält Namen die gleichgewertet werden
	Geschlecht: {Feminin, Maskulin} Ortsname: {Berlin, Rom, Bern, Paris} Familienstand: {verheiratet, ledig, geschieden, verwitwet}
Ordinalskala / Rangskala (qualitativ)	Es lässt sich zusätzlich eine Ordnung herstellen $= / \neq$ und $> / <$ Die Werte sind nicht mehr gleichgewichtet, sondern intensitätsmässig geordnet (in Klassen)
	Schulnote: {sehr gut, gut, genügend, schlecht} Umfragen: {Trifft zu, Trifft eher zu, Trifft eher nicht zu, Trifft nicht zu} Qualitätsstufe: {Standard, Business, First Class}
Intervallskala (metrische Skala / Kardinalskala) (quantitativ)	Es lässt sich zusätzlich eine Aussage über die Abstände machen $= / \neq$ und $> / <$ und $+/-$ Es kann der einfache Abstand (Intervall) gemessen werden. Hat keinen absoluten Nullpunkt
	Temperatur: {-12, .., 0, .., 42} Uhrzeit: {20:00, 0:00, 10:00}
Verhältnisskala (metrische Skala / Kardinalskala) (quantitativ)	Es lässt sich zusätzlich eine Aussage über das Verhältnis machen $= / \neq$ und $> / <$ und $+/-$ und $\cdot / :$ Hat einen absoluten Nullpunkt, deshalb Vergleich Aussagen möglich. Negative Werte sind nicht möglich. Besitzt das höchste Informationsniveau!
	Umsatz: {0M, 1M, 2M, 3M,.. } Alter: {0,1,..,40,.., gut, gut, genügend, schlecht} Gewicht: {0kg, 50kg, 60kg,..,80kg,120kg }

Häufigkeitsverteilung

- n Gesamtzahl der Merkmalsträger (z.B Glühbirnen)
- v Anzahl verschiedene Merkmalsträger (Klassen)
- h_i absolute Häufigkeit
- H_i kumulierte absolute Häufigkeit
- f_i relative Häufigkeit
- F_i kumulierte relative Häufigkeit
- d_i Klassendichte
- D_i kumulierte Klassendichte
- x Bestimmter Wert innerhalb der Klasse
- $\acute{x}$ Klassenmittelwert
- x^u Untere Klassengrenze
- x^o Obere Klassengrenze
- F_{i-1}/h_{m-1} Häufigkeit der vorherigen Klasse
- F_{i+1}/h_{m+1} Häufigkeit der nächsten Klasse

$$f_i = \frac{h_i}{n}$$

$$H_i = \sum_{a=1}^{i} h_a$$

$$F_i = \sum_{a=1}^{i} f_a$$

$$N = \sum_{i=1}^{v} h_i$$

$$d_i = \frac{h_i}{x^o - x^u}$$

$$D_i = \sum_{a=1}^{i} d_a$$

Dichte:

Wenn die **Klassenbreiten unterschiedlich gross** sind muss mit Dichte d_i gerechnet werden. Ist die Klassenbreite gleich gross bzw. die Häufigkeit unklassifiziert, so ist $d_i = h_i$.

Häufigkeiten berechnen

Häufigkeit für einen bestimmten Wert (z.B. 45)	$F(x) = F_{i-1} + \frac{x - x^u}{x^o - x^u} \cdot (F_i - F_{i-1})$
Arithmetisches Mittel der Gesamtheit	**Unklassifiziert:** $\bar{x} = \frac{1}{n} \sum_{i=1}^{v} x_i \cdot h_i = \sum_{i=1}^{v} x_i \cdot f_i$ **Klassifiziert:** $\bar{x} = \frac{1}{n} \sum_{i=1}^{v} \acute{x}_i \cdot d_i$
Modus (häufigster Wert)	Der Modus ist immer in der Klasse mit der höchsten Dichte. $Mo = x^u + \frac{d_i - d_{i-1}}{(d_i - d_{i-1}) + (d_i - d_{i+1})} \cdot (x^o - x^u)$
Median, 1. Quartil, etc.	$Me/Q = x^u + \frac{\frac{n}{t} - H_{i-1}}{h_i} \cdot (x^o - x^u)$ 1. Bestimmen von t a. t = In wie viele Teile die Gesamtheit unterteilt ist (2=Median, 4=1.Quartil, etc.) 2. Klasse finden, in welcher der Median/Quartil liegt a. (n/t) < H_i → i = Klasse

Beispiel für klassifizierte Häufigkeit

Welcher Anteil der Mitarbeiter ist < 45 Jahre alt?

J	$x_j^u \leq x_i < x_j^o$	h_i	H_i	f_i	F_i
1	0 bis 40	10	10	0.2	0.2
2	40 bis 50	15	25	0.3	0.5
3	50 bis 65	25	50	0.5	1

$$f = \frac{45 - 40}{50 - 40} * (0.5 - 0.2) = 0.15$$

$$F(x < 45) = 0.2 + 0.15 = 0.35 \ bzw. 35\%$$

Anteil < 45 Jahre = 35%

Anteil > 45 Jahre = 100% - 35% = 75%

Rechnung mit Dichte:

d1 = 10/40 = 0.25 d2 = 15/10 = 1.5 d3 = 25/15 = 1.7

Modusklasse ist also Klasse 3, da sie die grösste Dichte hat.

$$M = 50 + \frac{25 - 15}{(25 - 15) + (25 - 0)} * (65 - 50) = 54.3$$

Lagemasse / Lageparameter

Arithmetisches Mittel / Mittelwert

Der Klassische Durchschnitt: Man addiert alle Messwerte und dividiert durch die Anzahl Messwerte

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

Harmonisches Mittel

Ist zur Berechnung des Durchschnitts einzusetzen wenn das Merkmal aus einem Bruch hervorgeht.

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

Bsp. Auf einer Strecke von 2 Kilomer benötigt ein Fahrzeug auf der Hinfahrt 10km/h und auf der Rückfahrt 30km/h

$$\overline{MH} = \frac{\sum_{i=1}^{v} h_i}{\sum_{i=1}^{v} \frac{h_i}{x_i}} = \frac{(2+2)\ km}{\frac{2\ km}{10\ km/h} + \frac{2\ km}{30\ km/h}} = 15\ \frac{km}{h}$$

Geometrisches Mittel

Ist die n-te Wurzel aus dem Produkt aller beobachteten Merkmalswerte

$$\bar{x}_g = \sqrt[n]{\prod_{i=1}^{n} x_i}$$

Verwendete man immer dann, wenn man Mittelwerte aus aufeinander aufbauenden Wachstumsfaktoren berechnen will. Wichtig beim Geometrischen Mittel ist, dass man nicht den Prozentsatz selbst sondern die einzelnen Faktoren (Brüche) $\frac{x}{100}$ einsetzt.

Modus	Gibt den Wert an, der am häufigsten vorkommt
Median	Die Mitte in einem geordneten Datensatz. Gibt es eine gerade Anzahl Elemente wird einfach der Schnitt der beiden in der Mitte liegenden Werte genommen.
Quantil	Unterteilt die Gesamtheit in 2gleich grosse Teile
Quartil	Unterteilt die Gesamtheit in 4 gleich grosse Teile
Dezil	Unterteilt die Gesamtheit in 10 gleich grosse Teile
Perzentil	Unterteilt die Gesamtheit in 100 gleich grosse Teile

Streumasse / Streuparameter

Die Varianz und Standardabweichung werden in der Praxis für die Streuung eingesetzt.

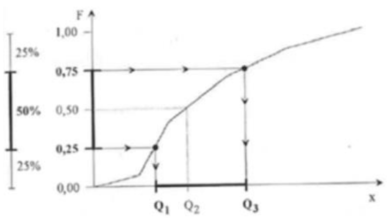
Spannweite

Die Differenz zwischen dem grössten und kleinsten beobachteten Merkmal

Zentraler Quartilsabstand / Interquartilsabstand

Die Differenz zwischen dem ersten und dritten Quartil

$$Q_3 - Q_1$$

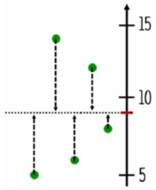


80% Dezilabstand = $D_9 - D_1$

Mittlere absolute Abweichung

Der Durchschnitt der Summe aller Differenzen zum Mittelwert

$$\bar{d} = \frac{1}{n}\sum_{i=1}^{n} |x_i - \bar{x}| \cdot h_i$$



Die Betragsstriche der mittleren absoluten Abweichung ist unvorteilhaft (Fallunterscheidung). Deshalb arbeitet man viel öfter mit der Varianz

Varianz

Durch das Quadrieren wird der Varianzwert sehr gross, weshalb man eher mit der Standardabweichung rechnet.

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

Rsp. vereinfacht:

$$\sigma^2 = \sum_{i=1}^{n} x_i^2 \cdot h_i - \bar{x}^2$$

oder

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

Standardabweichung

Arithmetisches Mittel der Abweichung vom Mittelwert der Gesamtheit $\bar{x}$

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

x_j'	$(x_j' - \bar{x})^2$	$(x_j' - \bar{x})^2 * h_j$
20	689.06	6890.63
45	1.56	23.44
57.5	126.56	3164.06

$$\bar{x} = (20 * 10 + 45 * 15 + 57.5 * 25) * \frac{1}{50} = 46.25$$

$$\sigma^2 = \frac{1}{50} * (6890.63 + 23.44 + 3164.06) = 201.56 \rightarrow \sigma = 14.2$$

Variationskoeffizient (%)

Die Standardabweichung im Verhältnis zum arithmetischen Mittel

$$v = \frac{\sigma}{\bar{x}} \cdot 100$$

1. Standardabweichung: CHF 0.85
2. Durchschnittlicher Preis für einen Espresso: CHF 4.25

$$v = \frac{\sigma}{\bar{x}} = \frac{0.85}{4.25} = 0.2 \cdot 100 = 20\%$$

Vergleich der Streuung der Leistungen eines Weitspringer W und eines Langstreckenläufer L (Achtung: **unterschiedliche Dimension** daher ist die Verwendung der **absoluten Streuung nicht erlaubt**)

$\bar{x}_W = 7.20\ m, \quad \sigma_W = 0.24\ m$

$\bar{x}_L = 29.4\ min, \quad \sigma_L = 0.89\ min$

$$VK_W = \frac{0.24}{7.20} \cdot 100 = 3.3\ \%,$$

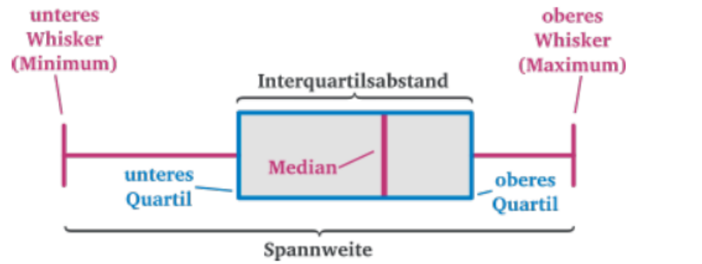
$$VK_B = \frac{0.89}{29.40} \cdot 100 = 3.0\%,$$

Der Langstreckenläufer und der Weitspringer erbringen, relativ gesehen eine nahezu gleichmässige Leistung

Boxplot

Der Boxplot vermittelt einen schnellen Eindruck, in welchem Bereich die Daten liegen und wie sie sich in diesem Bereich aufteilen.

1. Der geordnete Datensatz wird in 4 Abschnitte aufgeteilt, die etwa gleich viele Werte umfassen
   a. Minimum
   b. Maximum
   c. Median
   d. Beim unteren Quartil (Min 25% aller Messwerte kleiner/gleich und Max 75% aller Messwerte (grösser/gleich)
   e. Beim oberen Quartil Max 25% aller Messwerte kleiner/gleich und Min 75% aller Messwerte (grösser / gleich)
2. Die oberen und unteren Enden der Quartile mit Strichen verbinden = B
3. Verbindungslinie zwischen Min und unterem Quartil sowie eine Verbindungslinie zwischen Max und oberem Quartil = Whisker



Zeitreihen

X-Achse = Zeit / Y-Achse = Merkmalswerte → Punktdiagramm

Gleitender Mittelwert

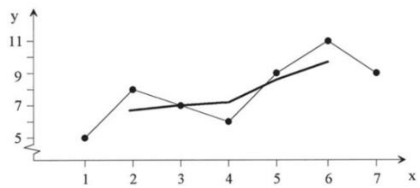
Ziel: Glättung der Zeitreihe/Kurve, in dem die hohen und niedrigen Werte gegeneinander Abgeglichen werden.

Man berechnet immer das arithmetische Mittel über eine Auswahl aller Messwerte und verschiebt diese Auswahl kontinuierlich nach vorne. Aus den neuen Messwerten (arithemtische Mittel) wird anschliessend eine neue Zeitreieh erstellt.

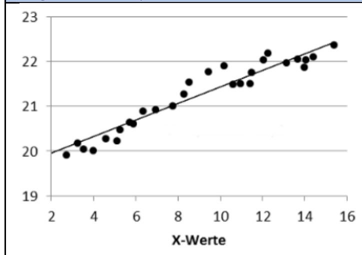
x_i	1	2	3	4	5	6	7
y_i	5	8	7	6	9	11	9
$\bar{y}_i$	-	6,67	7,00	7,33	8,67	9,67	-

$\boxed{5 \quad 8 \quad 7} \quad 6 \quad 9 \quad 11 \quad 9 \quad \rightarrow \quad \bar{y}_2 = \frac{5+8+7}{3} = 6{,}67$

$5 \quad \boxed{8 \quad 7 \quad 6} \quad 9 \quad 11 \quad 9 \quad \rightarrow \quad \bar{y}_3 = \frac{8+7+6}{3} = 7{,}00$



Regressionsanalyse



Die Regressionsanalyse ist das Bindeglied zwischen diskreten Messungen (mit Lücken) und für die Analysen wichtige Stetigkeit (ohne Lücken).

Das Ziel dabei ist die Abweichung zwischen den beiden Funktion so klein wie möglich zu halten

Man ermittelt damit den Zusammenhang zwischen X und Y Achse (linear, exponentiell, etc.)

Regressionsfunktion für lineare Zusammenhänge

Der gesuchte Wert sollte auf der Y-Achse, der gegeben auf der X-Achse liegen.

Regressionsgerade $\hat{y}$	$\hat{y} = a_1 + b_1 * x$ Beschreibt den Zusammenhang zwischen dem unabhängigen Merkmal X und dem abhängigen Merkmal Y.
Regressionsparameter	$a_1 = \bar{y} - b_1 * \bar{x}$ Gibt den tendenziellen Wert des Merkmals Y an, wenn der Wert des Merkmalswert x gleich Null ist.
Regressionsparameter	$b_1 = \frac{\sum(x_i y_i) - n\bar{x}\bar{y}}{\sum(x_i^2) - n\bar{x}^2}$ Gibt als Steigungsmass an, um wie viele Einheiten sich der Wert des Merkmals Y tendenziell änder, wenn der Wert des Merkmals X um eine Einheit erhöht wird.
	$\bar{x} = \frac{\sum x_i}{n}$
	$\bar{y} = \frac{\sum y_i}{n}$
Beispiel	12 Studenten gingen im letzten Semester neben dem Studium einer Erwerbstätigkeit nach. In der nachfolgenden Tabelle sind der zeitliche Aufwand (Std./Woche) für die Erwerbstätigkeit X und der zeitliche Aufwand (Std./Woche) für das Studium Y angegeben.

Student	A	B	C	D	E	F	G	H	I	J	K	L
Erwerbstätigkeit	1	2	2	3	3	4	5	6	8	12	15	23
Studium	39	37	36	40	36	37	34	36	33	33	32	27

Ein Student der 6 Stunden pro Woche erwerbstätig ist, will anhand der vorliegenden Daten ermitteln, wieviel Zeit er für sein Studium aufbringen kann. Es besteht ein linearer Zusammenhang zwischen Erwerbstätigkeit und Studium (Siehe Regressionsanalyse)

Arbeitstabelle für die Regressionsgeraden $\hat{y}$ und $\hat{x}$

Student	x_i	y_i	$x_i y_i$	x_i^2	y_i^2
A	1	39	39	1	1.521
B	2	37	74	4	1.369
C	2	36	72	4	1.296
D	3	40	120	9	1.600
E	3	36	108	9	1.296
F	4	37	148	16	1.369
G	5	34	170	25	1.156
H	6	36	216	36	1.296
I	8	33	264	64	1.089
J	12	33	396	144	1.089
K	15	32	480	225	1.024
L	23	27	621	529	729
Summe	84	420	2.708	1.066	14.834

$\bar{x} = \frac{\sum x_i}{n} = \frac{84}{12} = 7$; $\bar{y} = \frac{\sum y_i}{n} = \frac{420}{12} = 35$

$b1 = \frac{\sum x_i y_i - n\bar{x}\bar{y}}{\sum x_i^2 - n\bar{x}^2} = \frac{2'708 - 12*7*35}{1'066 - 12*7*7} = -0.49$

$a1 = \bar{y} - b1 * \bar{x} = 35 - (-0.49)*7 = 38.43$

Resultat für $\hat{y}$

$\hat{y}$: -0.49x + 38.43

Auf den Fall 6 Studen Erwerbstätigkeit angewendet folgt:

-0.49*6 + 38.43 = 35.49 → der Student kann tendenziell von einem Aufwand von 35.49 Stunden pro Woche für sein Studium ausgehen. Der tatsächliche Aufwand wird jedoch davon abweichen, da noch weitere Faktoren als die Erwerbstätigkeit einen Einfluss auf die Höhe der Studiendauer haben.

$b2 = \frac{\sum x_i y_i - n\bar{x}\bar{y}}{\sum y_i^2 - n\bar{y}^2} = \frac{2'708 - 12*7*35}{14'834 - 12*35*35} = -1.73$

$a2 = \bar{y} - b2 * \bar{y} = 7 - (-1.73)*35 = 67.55$

Resultat für $\hat{x}$

-1.73y + 67.55

Die Regressionsgerade beschreibt die Tendenz des Zusammenhangs zwischen dem Zeitaufwand für da Studium und dem Zeitaufwand für die Erwerbstätigkeit.
Es kann somit der tendenziell anfallende Zeitaufwand für die Erwerbstätigkeit bestimmt werden.

Beispiel für den Zusammenhang mit Häufigkeitsverteilung

Ein Unternehmen bezahlt seine 200 Beschäftigen (n) im Produktionsbereich nach Tarifgruppen aus. Aus der nachstehenden Tabelle kann die Verteilung der 124 weiblichen und der 76 männlichen Beschäftigten auf die Tarifgruppen ersehen werden.

Tarifgruppe	1	2	3	4	Summe
Weiblich	43	32	36	13	124
Männlich	19	18	23	16	76
Summe	62	50	59	29	200

Beschreiben sie den Zusammenhang zwischen den Merkmalen Geschlecht (X) und Tarifgruppenzugehörigkeit:

Resultat

Im Unternehmen arbeiten deutlich mehr Frauen als Männer. Um eine Aussage über die Häufigkeitsverteilung der Geschlechter in den verschiedenen Tarifgruppen zu machen, müssen die relativen Häufigkeiten mit folgender Formel berechnet werden.

$$f_i = \frac{H_i(Tarifgruppe) * H_i(Geschlecht)}{n}$$

Bsp.

$f_i = \frac{62*124}{200} = \underline{38.43}$ → relative Häufigkeit, weiblich in G1

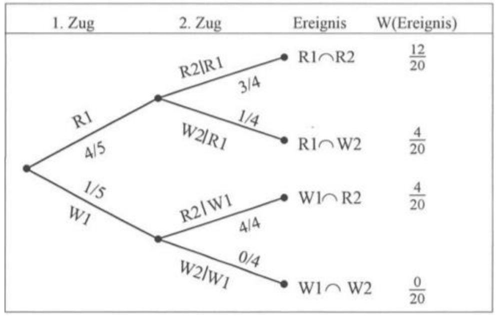
Tarifgruppe	1	2	3	4	Summe
Weiblich	38.44 (43)	31 (32)	36.58 (36)	17.98 (13)	124
Männlich	23.56 (19)	19 (18)	22.42 (23)	11.02 (16)	76
Summe	62	50	59	29	200

Mit der relativen Häufigkeitswerten kann man nun feststellen, dass es in den Tarifgruppen 1 und 4 zu einer Verschiebung kommt, jedoch die Gruppen 2 und 3 geschlechtsunabhängig sind.

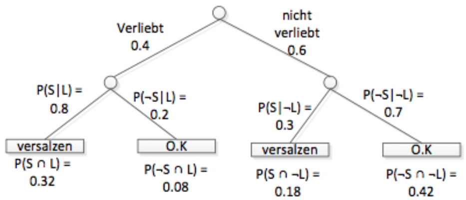
Wahrscheinlichkeitsrechnung / Stochastik

Menge	Ungeordnet, ohne Redundanzen
Tupel	Geordnet, mit Redundanzen
Zufallsexperiment	Ein Experiment welches beliebig oft durchgeführt werden kann und das Ergebnis komplett vom Zufall abhängig ist (z.B Werfen eines Würfels)
Elementarereignis ω	Ist ein möglicher Ausgang des Zufallsexperiments, wobei zwei Elementarereignisse sich immer gegenseitig ausschliessen.
Ergebnismenge Ω	Umfasst alle möglichen Elementarereignisse eines Zufallsexperiments. Z.B {1,2,3,4,5,6}
Ereignis	Eine Teilmenge der Ergebnismenge. Z.B {2,4,6}
System der Ereignisse $\mathcal{A}$	Bei einem Zufallsvorgang gemessene Ereignisse, bilden zusammen ein System der Ereignissen. Dieses weist Eigenschaften auf, welche es ermöglichen Relation (Durchschnitt, Vereinigung, etc.) zu bilden.
Unmögliches Ereignis $\emptyset$	Die leere Menge
Disjunkte Ereignisse	A und B sind disjunkt, wenn sie keine gemeinsame Teilmenge besitzen.
Laplace Experiment (gut Fälle / alle Fälle)	Ein Experiment bei dem jedes Ergebnis <u>dieselbe</u> Wahrscheinlichkeit hat und die Ergebnismenge endlich/abzählbar ist. $P(A) = \frac{\lvert A\rvert}{\lvert\Omega\rvert} = \frac{\text{Anzahl der für das Ereignis A günstigen Ergebnisse}}{\text{Anzahl aller möglichten Ergebnisse}}$
	Wie gross ist die Wahrscheinlichkeit im Lotto (6 aus 49) genau drei Richtige anzukreuzen? $P\{3\ richtige\} = \frac{\binom{6}{3} * \binom{43}{3}}{\binom{49}{6}} = 0.0176$
Unabhängige Ereignisse	Zwei Ereignisse A und B sind voneinander unabhängig wenn gilt: $W(A) = W(A\vert B)$ bzw. $W(A) = W(A\vert \bar{B})$ bzw. $W(A\vert B) = W(A\vert \bar{B})$
	Beispiel: Es wurden folgende Wahrscheinlichkeiten berechnet: P(A) = 0,65; P(A|B) = 0,75; ➔ Da P(A) ≠ P(A|B) sind die beiden Ereignisse A und B abhängig.

Additionssatz	Die Wahrscheinlichkeit das A **oder** B eintritt 1. A und B sind vereinbar $P(A \cup B) = P(A) + P(B) - P(A \cap B)$ 2. A und B sind unvereinbar $P(A \cup B) = P(A) + P(B)$
Multiplikationssatz	Die Wahrscheinlichkeit, dass A **und** B eintritt 1. Sind A und B abhängig $P(A \cap B) = P(A) \cdot P(B\|A)$ oder $P(A \cap B) = P(B) \cdot P(A\|B)$ 2. Sind A und B unabhängig $P(A \cap B) = P(A) \cdot P(B)$
	Es gibt 5 Kugel, 4 rote und 1 weise. Wie gross ist die Wahrscheinlichkeit, dass man ohne zurücklegen zwei rote zieht? → Da man nicht zurücklegt, ist die Wahrscheinlichkeit der weiteren Kugeln abhängig von der von den vorherig gezogenen.



Bedingte Wahrscheinlichkeit	Die Wahrscheinlichkeit von A unter der Voraussetzung von B $P(A\|B) = \frac{P(A \cap B)}{P(B)}$



	S	¬s	Total
l	0.32	0.08	*0.4*
¬l	0.18	0.42	*0.6*
Total	*0.5*	*0.5*	*1*

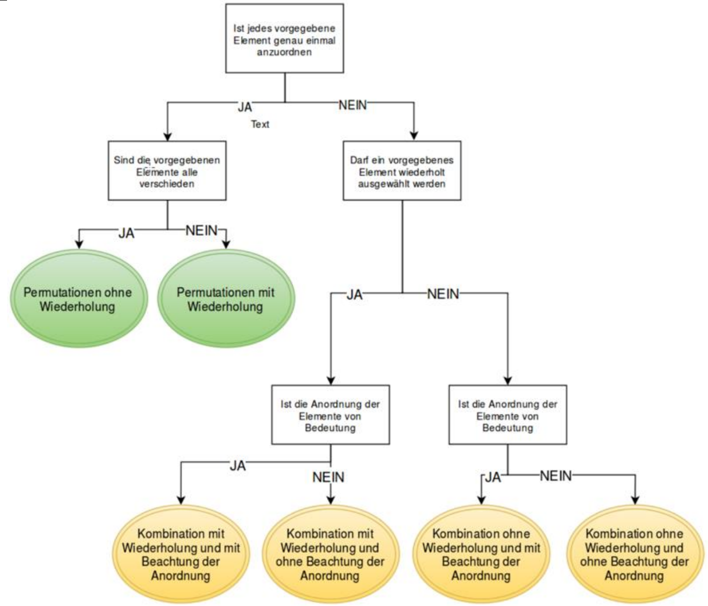
Komplementäre Wahrscheinlichkeit	Die Wahrscheinlichkeit, dass A nicht eintritt $P(\bar{A}) = 1 - P(A)$

Korrelation

Die Korrelation ist eine Kennzahl für den Zusammenhang zwischen mehreren Streudiagrammen

Formale Abhängigkeit	Zahlenmässig begründete Abhängigkeit
Sachliche Abhängigkeit	Ist der Wert eines Merkmals kausal/ursachlich für den Wert eines zweiten Merkmals abhängig
Kovarianz	$\sigma_{xy} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})$ Merkmalswertkombinationen (x_i, y_i)
Kovarianz bei Stichproben	Bei Stichproben verwendet man eine korrigierte Varianz, wobei man nicht nur n sondern durch n – 1 teilt
Korrelationskoeffizient Lineare Abhängigkeit	$r_{xy} = \frac{Kovarianz}{Standardabweichung_x \cdot Standardabweichung_y} = \frac{\sigma_{xy}}{\sigma_x \sigma_y}$ Es resultiert immer ein Wert r zwischen -1 und 1: a) Je mehr der Wert bei -1 liegt, desto mehr ähneln die Punkte im Streudiagramm einer Gerade mit negativer Steigung (stark linear abhängig) b) Je mehr der Wert bei 0 liegt, desto grösser ist die Streuung der Punkte (linear unabhängig) c) Je mehr der Wert bei +1 liegt, desto mehr ähneln die Punkte im Streudiagramm einer Gerade mit positiver Steigung (stark linear abhängig)

Permutationen und Kombinatorik



Permutationen

Permutationen ohne Wiederholung/Zurücklegen (Merke: 0!=1)	Anzahl Möglichkeiten n Objekte anzuordnen = Fakultät $p(n) = n! = n \cdot (n-1) \cdot (n-2) \cdot \ldots \cdot 2 \cdot 1$
	Eine Maschine muss vier Aufträge A, B, C, D nacheinander abarbeiten. Wie viel Anordnungen sind möglich: 4!=24 → für den ersten Platz gibt es 4 Möglichkeiten, für den zweiten 3, usw. → n!
Permutationen mit Wiederholung/Zurücklegen	Bei **identischen** Elementen werden diese in Klassen zusammengefasst. Es gibt dabei k Klassen mit jeweils n_k identischen Elementen
	$p(n_k) = \frac{n!}{n_1! \cdot n_2! \cdot \ldots \cdot n_k!}$
Beispiel:	Von einem 6-stelligen Zahlenschloss weiss man, dass es sich mit einer bestimmten Folge der Ziffern 1, 1, 4, 4, 4 und 8 öffnen lässt. Wie viele Versuche sind maximal notwendig um das Zahlenschloss zu öffnen? Gegeben sind n=6 Ziffern, die in k=3 Klassen von untereinander gleich Ziffern zerfallen. Die Klasse «1» enthält n1 = 2 Elemente, die Klasse «4» n2 = 3 und die Klasse «8» n3 = 1 Element. $p_{2,3,1}(6) = \frac{6!}{2! \cdot 3! \cdot 1!} = \frac{720}{12} = 60 \text{ Permutationen}$

Bei der **Kombinatorik** geht es darum, aus n Elementen, k auszuwählen und anschliessend in eine Ordnung zu bringen.

Binomialkoeffizient	Aus n Optionen, k auswählen $\binom{n}{k} = \frac{n!}{(n-k)! \cdot k!}$ Im Rechner ist das die Funktion nCr(n,k)
Kombinationen ohne Wiederholung/Zurücklegen (Mit Beachtung der Anordnung)	$\text{Anzahl Möglichkeiten} = \frac{n!}{(n-k)!}$ Im Rechner ist das die nPr(n,k) Funktion
	Aus 5 Bewerber soll eine Rangliste der ersten 3 Plätze gemacht werden. Wie viele verschiedene Listen sind möglich? $V_3(5) = \frac{5!}{(5-3)!} = \frac{5!}{2!} = 60$
Kombinationen ohne Wiederholung/Zurücklegen (ohne Beachtung der Anordnung)	Ist gleich dem Binomialkoeffizienten $\text{Anzahl Möglichkeiten} = \binom{n}{k} = \frac{n!}{(n-k)! \cdot k!}$ Im Rechner ist das die Funktion nCr(n,k)
	Bsp. Lotto Beim Lotto müssen aus 49 Zahlen 6 Zahlen ausgewählt werden. Wie viel Tipps sind möglich? $\binom{49}{6} = \frac{49!}{(49-6)! * 6!} = 13'983'816$
Kombinationen mit Wiederholung/Zurücklegen (Mit Beachtung der Anordnung)	$\text{Anzahl Möglichkeiten} = n^k$
	In einem Einkaufsladen gibt es unterschiedlich bemalte Vasen zu kaufen. Der Kunde möchte 3 Vasen für seinen Garten kaufen. Wie viele Möglichkeiten hat er, die Vasen in seinem Garten auf 4 Plätzen anzuordnen. $V_3^W(4) = 4^3 = 64$
	Bei einem Ziffernschloss muss man eine 5-stellige Zahl einstellen, die aus den Ziffern 0-9 gebildet wird. Wie viele Kombinationen gibt es? $10^5 = 100'000$
Kombinationen mit Wiederholung/Zurücklegen (ohne Beachtung der Anordnung)	$\text{Anzahl Möglichkeiten} = \frac{(n+k-1)!}{k! \cdot (n-1)!} = \binom{k+n-1}{k}$
	In einem Rat werden 3 Sitze neu vergeben, es bewerben sich 6 Verbände darauf. Die wiederholte Auswahl eines Verbandes ist möglich. Wie viele mögliche Sitzverteilungen gibt es? k= 3 n = 6 $K_3^W(6) = \binom{6+3-1}{3} = \binom{8}{3} = \frac{8!}{(8-3)! * 3!} = 56$

Zufallsvariablen

Zusammenhang Zufallsvariable und Merkmal

Zufallsvariable X	Merkmal X
Realisation x	Merkmalswert x
Wahrscheinlichkeit	relative Häufigkeit
Wahrscheinlichkeitsfunktion	einfache relative Häufigkeitsverteilung
Verteilungsfunktion	kumulierte relative Häufigkeitsverteilung
Erwartungswert	arithmetisches Mittel
Varianz	Varianz

Realisation	Wert der Zufallsvariable für ein Ereignis z.B. Im Monopoly ist die Summe der Augenzahlen zweier Würfel entscheidend, wie weit ein Spieler vorrücken darf: Zufallsvariable = Augensumme Realisationen = {2,3,4, ... , 12}

Eine Zufallsvariable hat für jedes Ereignis eine bestimmte Wahrscheinlichkeit, dass das Ereignis eintreffen kann.

$$f(x) = P(X = x) = \begin{cases} Wahrscheinlichkeit_1 & \text{für x=Ereignis} \\ Wahrscheinlichkeit_2 & \text{für x=Ereignis} \end{cases}$$

Diskrete Massenfunktion	- Kann mit einem Stabdiagramm veranschaulicht werden (Ordinate (Y) = Wahrscheinlichkeit, Abszisse (X) = Ereigniswerte) - Wahrscheinlichkeit kann direkt abgelesen werden - Hat Lücken und nur positive Werte - Die Summe aller Einzelwahrscheinlichkeiten entspricht 1 = Fläche unter dem Graphen
Stetige Dichtefunktion / Kontinuierliche Verteilungsfunktion	- Auf der X-Achse sind unendliche viele Werte - Die Summe aller Einzelwahrscheinlichkeiten entspricht 1 = Fläche unter dem Graphen
Erwartungswert	$E(X) = \sum_{i=1}^{n} x_i \cdot P(x_i)$
Varianz	$VAR(X) = \sum_{i=1}^{n} (x_i - E(X))^2 \cdot P(x_i)$

Beispiele:

Ein Zufallsvorgang besteht im dreimaligen Werfen einer Münze. Entscheidend ist die Anzahl an Wappen.
Geben Sie die Wahrscheinlichkeits- und Verteilungsfunktion an.
Zufallsvariable: Anzahl Wappen
Realisation: 0,1,2,3

xi	f(xi)	F(xi)
0	0.125	0.125
1	0.375	0.500
2	0.375	0.875
3	0.125	1.000

a) Berechnen Sie den Erwartungswert.

$0 * 0.125 + 1 * 0.375 + 2 * 0.375 + 3 * 0.12 = 1.5\ Wappen$

b) Berechnen Sie die Varianz und die Standardabweichung

$\sigma^2 = (0-1.5)^2 * 0.125 + (1-1.5)^2 * 0.375 + (2-1.5)^2 * 0.375 + (3-1.5)^2 * 0.125 = 0.75$

$\sigma = \sqrt{0.75} = 0.866$

Stichproben

Zufallsstichprobe	Aus der Grundgesamtheit werden Elemente zufällig ausgewählt
Einfache Stichprobe	Die Elemente der Stichprobe haben alle die gleiche Wahrscheinlichkeit
Geschichtete Stichprobe	Ist es möglich, Elemente mit gleichen Eigenschaften in Gruppen einzuteilen ist es sinnvoller, Teilstichproben pro Gruppe/Schicht zu nehmen, um genauer Aussagen über die Gesamtheit zu machen

Schätzverfahren

Ist der Mittelwert, Standardabweichung und die Verteilungsfunktion nicht bekannt müssen diese mit Hilfe von Schätzfunktionen geschätzt werden. Ziel dabei ist es, von einer Stichprobe auf die Grundgesamtheit zu schliessen und dabei den Fehler einer falschen Schätzung zu minimieren.

Punktschätzung

Schätzfunktion für den Mittelwert	$\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i$
Schätzfunktion für Varianz und Standardabweichung	$s^2 = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2$ $s = \sqrt{\frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2}$

Stichprobe \ Varianz σ^2		bekannt	unbekannt
mit Zurücklegen		$\sigma^2_{\bar{X}} = \frac{\sigma^2}{n}$	$\hat{\sigma}^2_{\bar{X}} = \frac{s^2}{n}$
ohne Zurücklegen	$\frac{n}{N} < 0{,}05$	$\sigma^2_{\bar{X}} \approx \frac{\sigma^2}{n}$	$\hat{\sigma}^2_{\bar{X}} \approx \frac{s^2}{n}$
	$\frac{n}{N} \geq 0{,}05$	$\sigma^2_{\bar{X}} = \frac{\sigma^2}{n} \cdot \frac{N-n}{N-1}$	$\hat{\sigma}^2_{\bar{X}} = \frac{s^2}{n} \cdot \frac{N-n}{N}$

ACHTUNG: Evtl. Wurzel ziehen! Wir arbeiten meist mit der Standardabweichung

Stichprobe \ Varianz σ^2		bekannt	unbekannt
mit Zurücklegen		$\sigma^2_P = \frac{\Theta \cdot (1-\Theta)}{n}$	$\hat{\sigma}^2_P = \frac{P \cdot (1-P)}{n}$
ohne Zurücklegen	$\frac{n}{N} < 0{,}05$	$\sigma^2_P \approx \frac{\Theta \cdot (1-\Theta)}{n}$	$\hat{\sigma}^2_P \approx \frac{P \cdot (1-P)}{n}$
	$\frac{n}{N} \geq 0{,}05$	$\sigma^2_P = \frac{\Theta \cdot (1-\Theta)}{n} \cdot \frac{N-n}{N-1}$	$\hat{\sigma}^2_P = \frac{P \cdot (1-P)}{n} \cdot \frac{N-n}{N}$

Verteilung des Merkmals X \ Varianz σ^2	bekannt	unbekannt
bekannt und normalverteilt	$\bar{X}$ ist normalverteilt	$\bar{X}$ ist t-verteilt mit k = n - 1 Freiheitsgraden Wenn n > 30: $\bar{X}$ ist approximativ normalverteilt
bekannt und nicht normalverteilt (n > 30)	$\bar{X}$ ist approximativ normalverteilt	
unbekannt (n > 30)		

$\bar{X} = Stichprobenfunktion$

P = Wahrscheinlichkeit (Anteilswerte)

Intervallschätzung

Konfidenzintervall für den Mittelwert

Beispiel: Bekanntheitsgrad (unbekannte Varianz)
Ein Chemieunternehmen möchte den Bekanntheitsgrad eines von ihm hergestellten Waschmittels in Erfahrung bringen. Dazu werden 400 Personen zufällig ausgewählt und befragt. Das Waschmittel war 30 % der Befragten zumindest namentlich bekannt.
Erstellung des zentralen 95%-Konfidenzintervalls für Θ.

Schritt 1: Festlegung der Verteilungsform von P
Wie: n * P * (1 - P) > 9 = 400 * 0.3 * 0.7 = 84 > 9 → wahr, also approximativ normalverteilt

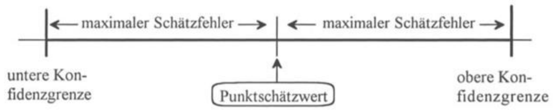
Schritt 2: Festlegung der Varianz / Standartabweichung von P
Resultat: $\hat{\sigma}_P = \sqrt{\frac{P*(1-P)}{n}} = \sqrt{\frac{0.3*0.7}{400}} = 0.02$

Schritt 3: Ermittlung des Quantilswertes z → Gemäss geg. Konfidenzintervall (Unterscheidung einseitig/Beidseitig(zentral))
Resultat aus Tabelle

Schritt 4: Berechnung des maximalen Schätzfehlers
Resultat: z * $\hat{\sigma}_P$ = 1.96 * 0.02 = 0.04

Schritt 5: Ermittlung der Konfidenzgrenze
Resultat: W(0.30 – 0.04 ≤ Θ ≤ 0.30 + 0.04) = 0,95
W(0.26 ≤ Θ ≤ 0.34) = 0.95

Der Bekanntheitsgrad in der Grundgesamtheit wird mit einer Wahrscheinlichkeit von 95% vom Intervall [26%; 34%] überdeckt.



Genauigkeit erhöhen:
- Konfidenzgrenze behalten, Umfang n erhöhen
- Umfang n behalten, Konzidenzniveau senken

Konfidenzintervall für beidseitig begrenzt:

$$W(\mu - z \cdot \sigma_{\overline{X}} \leq \overline{X} \leq \mu + z \cdot \sigma_{\overline{X}}) = 1 - \alpha$$

$\sigma_{\bar{x}}$ => findet sich mit Tabelle links!
1- α gibt die Wahrscheinlichkeit an, dass sich die Zufallsvariable/Stichprobenfunktion innerhalb des Intervalls befindet. (Konfidenzintervall)

Der tägliche Kaffeekonsum in einem Büro:

xi	f(xi)
1	20
2	30
3	40
4	10

Wie hoch ist die Wahrscheinlichkeit, dass der Stichprobenmittelwert bei n=100 im Intervall (2.3;2.5) liegt?

$$\bar{x} = \frac{1*20+2*30+3*40+4*10}{100} = 2.4$$

Varianz-Berechnung siehe "Streuparameter" → 0.84

$$\sigma_{\overline{X}} = \sqrt{\frac{0{,}84}{100}} = 0{,}0917$$

$$z = \frac{\overline{x} - \mu}{\sigma_{\overline{X}}} = \frac{2{,}5 - 2{,}4}{0{,}0917} = 1{,}0905$$

Intervall von z=-1,09 bis +1,09 => 0.8621-0.1379 = 0.7242 → 72.42%

notwendiger Stichprobenumfang n beim Konfidenzintervall für das arithmetische Mittel

In diesem Fall wird gefordert, dass die Schätzung ein vorgegebenes Mindestmass an Genauigkeit e besitzt und dass diese Mindestgenauigkeit mit einer vorgegebenen Konfidenz bzw. Sicherheit erzielt wird.
gegeben: Konfidenz, Genauigkeit (e) → e = x̄ -μ
gesucht: Stichprobenumfang n

Beispiel: Zuckerabfüllung
gegeben: Konfidenz z = 1.96   Genauigkeit e = 0.2 g   Standartabweichung σ = 1.2
$n \geq \frac{z^2 * \sigma^2}{e^2} = \frac{1.96^2 * 1.2^2}{0.2^2}$ = 138.3
Es müssen 139 Packungen entnommen werden, um die gewünschte Genauigkeit zu erzielen (wegen n > 30 ist im Falle einer beliebig verteilten Grundgesamtheit die Approximation durch die Normalverteilung zulässig).

Beispiel: Eine Molkerei liefert an eine Lebensmittelkette 40'000 Flaschen Milch mit 1000ml Soll-Füllmenge. Die Stichproben haben eine durchschnittliche Füllmenge von 1000.25ml. Aufgrund von zahlreichen Kontrollen weiss man, dass die Ist-Füllmenge normalverteilt mit einer Streuung von σ = 1.2ml ist. Wie viele Flaschen Milch müssen der Lieferung entnommen werden, wenn folgende Dinge gegeben sind.
gegeben: z= 1.96 (=95% Konf. Int beidseitig)   e=0.25ml ($\bar{x} - \mu$)
z = $\frac{\bar{x}-\mu}{\frac{\sigma}{\sqrt{n}}} = \frac{e}{\frac{\sigma}{\sqrt{n}}}$ → nach n auflösen = 88.51 → 89 Stichproben

Konfidenzintervall für die Varianz

Es wird die folgende Schätzfunktion verwendet: $s^2 = \frac{1}{n-1} * \sum(x_i - \bar{x})^2$

Voraussetzungen für eine erwartungstreue Schätzung: das Merkmal X ist in der Grundgesamtheit normalverteilt und die Entnahme erfolgt mit Zurücklegen.

-Zweiseitiges Konfidenzintervall

Formel:
$W\left(\frac{(n-1)*s^2}{y_{1-\frac{\alpha}{2}}} \leq \sigma^2 \leq \frac{(n-1)*s^2}{y_{\frac{\alpha}{2}}}\right) = 1 - \alpha$ (Konfidenzintervall)

α = Irrtumswahrscheinlichkeit
n =Stichprobenumfang/Freiheitsgrade
y = Mit k = n-1 und α z.B. 95% Konf. Int → $\alpha = 0.975$ ($y_{1-\frac{\alpha}{2}}$,) und $\alpha = 0.025$ ($y_{\frac{\alpha}{2}}$,), in der Chi² Tabelle → y herausfinden

-Einseitiges Konfidenzintervall (nach oben begrenzt)

Formel:
$W\left(\sigma^2 \leq \frac{(n-1)*s^2}{y_{\alpha,\ k=n-1}}\right) = 1 - \alpha$ (Konfidenzintervall)
α = Irrtumswahrscheinlichkeit
n = Stichprobenumfang/Freiheitsgrade
y = Mit

Testverfahren

Signifikanzniveau	Das Signifikanzniveau wird meist bei 5% angesetzt. Ist der Wert kleiner wie 5% wird angenommen, dass ein Ergebnis signifikant ist.
Hypothese / Nullhypothese H0	Die Nullhypothese H0 ist eine Aussage von der angenommen wird, dass sie stimmt.
Alternativhypothese H1	Die Alternativhypothese H1 beschreibt eine Annahme, sie ist also das Gegenteil der Nullhypothese.
Fehler 1. Art (alpha)	Fehlerhaftes Verwerfen einer Hypothese
Fehler 2. Art (beta)	Fehlerhaftes Annehmen einer Hypothese
Parametertest	Man möchte wissen, ob der angegebene Benzinverbrauch eines Autos eingehalten wird. μ=10l/100km, σ=1l/100km Es werden 25 Autos getestet. Dabei kommt der Mittelwert 10.2L/100km zustande. Liegt das Ergebnis im Bereich statistischer Schwankungen, wenn 1-α=0.95 Annahme: Normalverteilter Strichprobenmittelwert $\mu_0 - z\frac{\sigma}{\sqrt{n}} \leq \bar{x} \leq \mu_0 + z\frac{\sigma}{\sqrt{n}}$ z für zweiseitige Tests: 1.96 Intervall: $10 - 1.96\frac{1}{\sqrt{25}} \leq \bar{x} \leq 10 - 1.96\frac{1}{\sqrt{25}}$ 9.61≤ x̄ ≤ 10.39 → Nullhypothese annehmen!
Anteilswert (unbekannte Wahrscheinlichkeit)	1. Schritt: Wähle die Signifikanzzahl α und bestimme daraus die Werte für Z aus d Tabell (1-α) 2. Schritt: Berechne die Annahmegrenzen zu 1. $c_u = p_0 - z\sqrt{\frac{p_0(1-p_0)}{n}}$ oder $c_u = p_0 - z\sqrt{\frac{p_0(1-p_0)}{n}}\sqrt{\frac{N-n}{N-1}}$ und 2. $c_o = p_0 + z\sqrt{\frac{p_0(1-p_0)}{n}}$ oder $c_o = p_0 + z\sqrt{\frac{p_0(1-p_0)}{n}}\sqrt{\frac{N-n}{N-1}}$ 3. Man berechne den Anteil $\overline{p} = \frac{k}{n}$ 4. Fällt $\overline{p}$ in den Annahmebereich: $c_u \leq \overline{p} \leq c_o$ wird die Hypothese angenommen, sonst abgelehnt

Verteilungen

Zentraler Grenzwertsatz

Der zentrale Grenzwertsatz besagt, dass sich mit grösserem Stichprobenumfang n, alle Verteilungen der Normalverteilung approximieren. Als Faustregel gilt, dass bei einem n über 30 Stichproben die Normalverteilung genommen werden kann.

Stetige Verteilungen

Stetige Verteilungen sind überabzählbar. Das heisst, sie beinhalten so viele Werte, dass diese nicht einfach gezählt werden können.

Normalverteilung

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

Standardnormalverteilung

$\mu = 0, \ \sigma = 1$

Die Standardnormalverteilung ist der einfachste Fall der Normalverteilung, wenn der Mittelwert = 0 und die Varianz = 1 ist.

$$f(x) = \phi = \frac{1}{\sqrt{2\pi}} \cdot e^{-\frac{1}{2}\cdot x^2}$$

Z-Wert berechnen

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

Kleiner: Z Wert direkt ablesen

Grösser: 1 – Z Wert aus der Tabelle

Beidseitig: Min/Max Z-Werte herauslesen und Differenz bilden

In einer Fabrik wird Zucker abgefüllt. Der Mindestinhalt jeder Tüte soll 1000g beinhalten. Gegeben: μ=1002g, σ=1,5g
Wie hoch ist die Wahrscheinlichkeit, dass eine Tüte das Sollgewicht unterschreitet?
$z = \frac{1000-1002}{1{,}5} = -1{,}33$ →Tabelle→0.0918→<u>9.18%</u>

Chi-Quadrat-Verteilung (Varianz!)

Voraussetzungen
- Zufallsvariablen sind unabhängig und normalverteilt

Wird aus der Normalverteilung abgeleitet

Anwendung:
- Schätzung von Verteilungsparametern (z.B. Varianz)

$$\chi_n^2 = Z_1^2 + \ldots + Z_n^2$$

Erwartungswert	Varianz
$\mathrm{E}(X) = n$	$\mathrm{Var}(X) = 2n$.

T-Verteilung

Voraussetzungen

Freiheitsgrade: k=n-1

Das Füllgewicht von Leberwürsten ist normalverteilt. Das Soll Mindestgewicht ist 125g.Aus den täglich produzierten 600 Würsten werden 26 gewogen:

128,4	123,8	123,5	126,9	125,5	123,1	124,9
123,1	126,6	121,9	125,3	123,4	122,1	124,0
123,3	123,2	123,2	124,0	122,8	127,1	125,7
127,1	125,8	123,7	125,9	124,9		

Erstellen Sie das zentrale 95% Konfidenzintervall.

1. berechnen der Stichprobenparameter:
   $\bar{x} = 124.58g, s = 1.72g$
2. Festlegen der Verteilungsform von $\bar{x}$
   X normalverteilt und σ^2 unbekannt →t-verteilt mit k=n-1 Freiheitsgraden
3. Festlegen der Standardabweichung von $\bar{x}$
   Varianz unbekannt, ohne Zurücklegen n/N<0.05 → $\hat{\sigma} = \frac{s}{\sqrt{n}} = 0.34g$
4. Festlegen von t
   1-α =0.95 (zweiseitiges Intervall!) und k=25 → Tabelle: t=2.060
5. Berechnen der Maximalen Schätzfehlers:
   $t^*\sigma_{\bar{x}}$ = 2.060 * 0.34 = 0.70 g
6. Berechnen der Konfidenzgrenzen
   W(124.58 -0.70 ≤ μ ≤ 124.58 + 0.70)=0.95
   W(123.88 ≤ μ ≤ 125.28)=0.95

Exponentialverteilung

Anwendungen:
- Zeitspanne zwischen zwei Anrufen in einer Telefonzentrale
- Dauer eines Telefongesprächs.
- Lebensdauer eines Geräts, wenn Defekte durch äussere Einflüsse und nicht durch Verschleiß verursacht werden.

Wahrscheinlichkeitsdichte: λ = Durchschn. Eintritt

$f(t) = \lambda * e^{-\lambda t}$

Verteilungsfunktion: (entspricht der aufsummierten Wahrsch.)

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

Erwartungswert	Varianz
$E(x) = \frac{1}{\lambda}$	$VAR(x) = \frac{1}{\lambda^2}$

Ein Geschäft wird täglich zwischen 10.00 und 11.00 Uhr von durchschn. 3,5 Kunden besucht. Wie gross ist die Wahrscheinlichkeit, dass der Abstand zwischen dem Eintreffen zweier Kunden höchstens 0.2 Stunden beträgt?

$F_E(0{,}2|3{,}5) = 1 - e^{-3{,}5*0{,}2} = 0{,}503$

Weibull Verteilung

Beschreibt die Lebensdauer von Geräten oder Materialen mit Abnutzungserscheinung

Diskrete Verteilungen

Bernoulli Experiment

Ist ein Zufallsexperiment mit <u>genau zwei</u> möglichen Ergebnissen (Treffer oder Niete).

Bernoulli Kette (binomial Verteilt)

$$P(Ereignis) = \binom{n}{k} \cdot p^k \cdot q^{n-k}$$

Bsp: Wie gross ist die Wahrscheinlichkeit mit 5 Würfel, 2x eine Sechs zu würfeln
n = 5
k = 2
p = 1/6
q = 5/6

$$\binom{5}{2}(\frac{1}{6}^2)(1-\frac{1}{6})^{5-2} = 0.161 = 16\%$$

20% der von einer Maschine produzierten Bolzen sind unbrauchbar. Wie hoch ist die Wahrscheinlichkeit, dass von 4 zufällig ausgewählten Bolzen <u>höchstens</u> zwei unbrauchbar sind?
→ Genau zwei Möglichiche Ausgänge: defekt / i.O

$$P(höchstens\ 2) = \sum_{x=0}^{2} \binom{4}{x} * 0.2^x * 0.8^{4-x}$$

= Summe aus P(0 unbrauchbar) + P(1 unbr.) + P(2 unbr.)

Binomialverteilung

Voraussetzungen
- Die Experimente sind voneinander unabhängig
- Es gibt nur <u>zwei</u> Ausgangsmöglichkeiten
- Anzahl der Versuche ist fix
- Das Experiment wird immer identisch durchgeführt
- Mit grösserem n, nähert sich die Binomialverteilung, ähnlich der Dichtefunktion einer Normalverteilung an

Wahrscheinlichkeitsfunktion (Bernoulli)

$$P(Ereignis) = \binom{n}{k} \cdot p^k \cdot q^{n-k}$$

Erwartungswert μ (relativ zum Nullpunkt)

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

Varianz

$$\sigma^2 = n \cdot p \cdot (1-p)$$

Gegen eine Krankheit wurde ein neues Medikament entwickelt. Die Heilungschance liegt bei 90%. Wie gross ist die Wahrscheinlichkeit, dass bei 5 zufällig gewählten Patienten mindestens 4 geheilt werden?

$P(4) = \binom{5}{4} * 0{,}9^4 * (1-0{,}9)^{5-4} = 0{,}328$

$P(5) = 0{,}590 \quad P(4+5) = 0{,}9185$

Poissonverteilung

Voraussetzungen
- Wie hoch ist die Wahrscheinlichkeit, dass das Ereignis in einem Intervall genau oder höchstens x-Mal eintritt, wenn bekannt ist, dass in diesem Intervall das Ereignis im Mittel μ -Mal auftritt.
- Hängt stark vom Mittelwert (μ) ab
- typische Beispiele: Druckfehler/Seite, Arbeitsunfälle/Tag
- Die Ereignisse treten unabhängig voneinander auf. (z.B Telefonanrufe)

Seltene Ereignisse häufen sich (z.B. Bitfehler bei Flugzeugabstürzen) → Verteilung der seltenen Ereignisse

Ankunftsrate: Eintreffende Ereignisse / Zeit
z.B. 24 Kunden in 8 Stunden → 24/8=3

BEACHTE: Die Werte in der Verteilung Tabelle sind auf kumuliert → Ist ein einzelner Wert gesucht, muss die Differenz zum vorherigen Wert berechnet werden vorherige Tabellenwert abgezogen oder einfach die Formel verwendet werden.

Wahrscheinlichkeitsfunktion

$$f(x) = \frac{\mu^x}{x!} \cdot e^{-\mu}$$

Erwartungswert: $E(X) = \sigma^2 = \mu$

Durchschnittlich 1 Telefonanruf/Minute. Wie gross ist die Wahrscheinlichkeit, dass 2 Anrufe pro Minute eingehen?

$$f_P(2|1) = \frac{1^2 * e^{-1}}{2!} = 0{,}18$$

Achtung: Falls <u>bis zu</u> 2 Anrufe gefragt sind, müssen diese auf kumuliert werden.

$$\sum_{k=0}^{2} \frac{1^k * e^{-1}}{k!}$$

Bei <u>mehr als</u> 2:

$$1 - \sum_{k=0}^{1} \frac{1^k * e^{-1}}{k!}$$

Rechteckverteilung

Alle Realisationen in einem bestimmten Intervall [a, b] sind gleich wahrscheinlich