Experimentieren & Evaluieren | Ex&Ev

Zusammenfassung

INHALTSVERZEICHNIS

Diese Zusammenfassung wurde komplett von Jannis Tschan verfasst, ich habe sie lediglich redigiert. Der Dank gebührt also ihm :)

— eure Nina

Hinweis: *In der Zusammenfassung werden Anweisungen zur Taschenrechnerbenutzung gegeben. Diese beziehen sich auf den TI nSpire CX II-T und das Skript auf https://github.com/takotori/StudentenportalSpick.*

1. GRUNDLAGEN

Durch Experimente soll ein besseres Verständnis der realen Welt und deren Zusammenhänge erhalten werden, mit dem Ziel Verbesserungen zu erreichen.

1.1. MODELL NACH SHEWHART-CYCLE

Hypothese formulieren → Daten im Experiment gewinnen → Hypothese prüfen → ggf. Modell anpassen

1.2. VERSUCHSPLANUNG

Hindernisse im Erkentnisgewinn: Komplexität, Kompliziertheit, Rauschen/Dynamik

1.3. PROZESSMODELL

Einflussgrössen: Einstellgrössen, Störgrössen → Prozess/Experiment → Zielgrössen
Praktisches Problem → Statistisches Problem → Statistische Lösung → Praktische Lösung

- ***Systematischer Fehler:*** durch Versuchsaufbau gegeben, reproduzierbar *(falscher Aufbau, Denkfehler bei Experimentdesign)*
- ***Zufälliger Fehler:*** nicht kontrollierbar, schwankendes Ergebnis *(Messfehler)*

1.4. MODELL NACH DESIGN OF EXPERIMENTS (DOE)

Methodik zur systematischen Planung und statistischen Auswertung von Experimenten

- ***Zielgrössen:*** Ergebnis eines Versuchs, Messwerte oder daraus berechnete Grössen, mehrere pro Versuch möglich
- ***Einflussgrössen:*** Können Zielgrössen in Experiment beeinflussen
  - ***Steuergrössen:*** Können beeinflusst werden *(z.B. Temperatur in Schmelzofen, RAM-Grösse etc.)*
  - ***Störgrössen:*** Können nicht (direkt) beeinflusst werden *(z.B. Ausfallrate, Verunreinigungen)*
- ***Faktoren:*** Wesentliche Einflussgrössen, die das Experiment stark beeinflussen
- ***Faktorstufen:*** Werte der Faktoren
  - ***Quantitative Faktoren:*** Zahlenwerte auf Messskala *(z.B. CPU-Last, Speicherverbrauch)*
  - ***Qualitative Faktoren:*** Namen, Bezeichnungen *(z.B. Ort der Serverfarm, Strassenname)*
- ***Simulation:*** Durchführen von Experimenten am Modell anstatt am realen System

1.4.1. Vorgehensweise

1. Ausgangssituation und Problem beschreiben
2. Untersuchungsszie festlegen
3. Zielgrössen und Faktoren festlegen
4. Entscheidung treffen, ob das Problem ***analytisch-mathematisch*** oder ***experimentell*** gelöst werden soll

Falls das Problem ***experimentell*** gelöst wird:

5. Versuchsplan aufstellen
6. Blockbildung *(Aufteilung der Versuchsobjekte anhand eines Merkmals in Blöcke, vermindert Fehlervarianz)*
7. Aufwandsabschätzung
8. ***Durchführungsplanung***, falls Experiment ***am realen Objekt***, ***Modell erstellen***, falls Experiment am ***Modell***
9. Versuche durchführen
10. Versuchsergebnisse auswerten
11. Ergebnisse interpretieren und Massnahmen ableiten
12. Absicherung, Dokumentation und weiteres Vorgehen

Ausgangssituation und Problem beschreiben (1)

- Wer ist Kunde? Wer ist Konkurrenz? Was braucht der Kunde? Was ist die Verbesserung?
- Was ist die (langfristige) Zielsetzung?
- Welches (Teil-)Problem soll durch die geplante Untersuchung gelöst werden?
- Wie viel Zeit und Geld stehen maximal zur Verfügung? *(Kosten-Nutzen-Analyse)*
- Wer ist von der geplanten Untersuchung betroffen? Ist mit Widerstand zu rechnen?
- Wer ist für was verantwortlich?
- Was ist bereits bekannt? Liegen (aktuelle) Daten vor?
- ***Ist das Problem wirklich verstanden?***

Untersuchungsziel festlegen (2)

- ***Optimale Lage des Mittelwerts:*** Das Prozessergebnis oder ein Produktparameter sollen einen bestimmten Wert annehmen
- ***Reduzierung der Streuung/Robustheit:*** Oft ist weniger die Lage des Mittelwerts des Prozessergebnisses problematisch als dessen Streuung
- ***Erkennen der wichtigsten Störgrössen:*** Durch systematische Beobachtungen (der Fertigung) und einfache Versuche herausfinden
- ***Gleichzeitig machen und lernen:*** Durch systematische Veränderung der Prozessparameter im laufenden Prozess können Verbesserungsmöglichkeiten erkannt werden
- ***Funktion und Zuverlässigkeit nachweisen***

Zielgrössen festlegen (3)

- ***Kundenorientiert und Relevant:*** Zielgrössen müssen die Probleme des Kunden abbilden
- ***Quantifizierung:*** Zielgrössen müssen messbare Grössen sein
- ***Vollständigkeit:*** Alle wesentlichen Prozessergebnisse bzw. Produkteigenschaften müssen als Zielgrössen erfasst werden
- ***Verschiedenheit:*** Möglichst wenige und möglichst unterschiedliche Zielgrössen definieren

Auswahl der Zielgrössen: Auswahl erfolgt in zwei Schritten: Zuerst möglichst viele Einflussgrössen mittels Brainstorming sammeln und anschliessend auf eine handhabbare Anzahl reduzieren.
Beim Brainstorming helfen Prozessablaufdiagramme, Ursache-Wirkungs-Diagramme, Einflussgrössen-Zielgrössen-Matrizen etc.

1.4.2. Beispiel

Zur Fertigung einer bestimmten Chemikalie werden mehrere Ausgangsstoffe einschliesslich Katalysator in einem Reaktionsgefäss vermischt. Die Mischung wird anschliessend über längere Zeit unter Rühren erhitzt, dabei erfolgt die Reaktion. Dann wird das Reaktionsprodukt abgetrennt.

Ziel ist eine ***Erhöhung der Ausbeute bei möglichst geringen Kosten***. Unten ein Auszug aus der Liste der Einflussgrössen, mit den Einschätzungen des Teams zur Grösse ihres Einflusses auf die beiden Zielgrössen «Ausbeute» und «Kosten». Aus der Fertigungsüberwachung ist bekannt, dass die Standardabweichung der Ausbeute aufgrund der Zufallsstreuung $\sigma = 1\%$ beträgt, wenn ein Ausbeute-Unterschied von 2% relevant ist.

Einflussgrösse	*Art*	*Ausbeute*	*Kosten*
Temperatur	Steuergrösse	Stark	Gering
Reaktionszeit	Steuergrösse	Stark	Stark
Katalysatormenge	Steuergrösse	Stark	Gering
Rührrate	Steuergrösse	Gering	Kein
Materialcharge	Störgrösse	Gering	Kein
Bediener	Störgrösse	Gering	Kein
Verunreinigungen	Störgrösse	Stark	Gering

Nach Abschluss der ***Ideenfindung*** werden die Ideen ***bewertet*** und ***gewichtet***. Anschliessend wird eine handhabbare Anzahl von Faktoren ***ausgewählt*** *(3-6)*. Diese müssen ***unabhängig*** voneinander ***veränderbar*** sein.

Im Beispiel werden die obersten 3, ***Temperatur***, ***Reaktionszeit*** und ***Katalysatormenge*** ausgewählt.

Faktorstufen

- ***Kleiner Abstand*** zwischen den Stufen ergibt ein ***kleiner Unterschied der Ergebnisse***. Ein ***grosser Abstand*** führt zu ***Abweichungen der Ergebnisse*** von der Linearität.
- Kann ein Faktor ***nicht genau gemessen*** werden, sollte der Abstand der Faktorstufe ***mindestens*** 6σ sein.
- Eine ***Extrapolation*** der Ergebnisse über den untersuchten Bereich hinaus ist nicht zulässig, daher sollte die Untersuchung ***den interessanten Bereich*** enthalten.

2. FEHLERFORTPFLANZUNG

Fehler können nur ***geschätzt***, nicht berechnet werden. Um den Bereich abzuschätzen, in denen der tatsächliche Wert maximal oder mit einer gewissen Wahrscheinlichkeit liegt, verwendet man ***Fehlerrechnung*** bzw. Fehlerfortpflanzung. Der Fehler wird jeweils mit Δ gekennzeichnet.
Man unterscheidet zwischen dem ***maximalen*** und dem ***wahrscheinlichen*** Fehler.

2.1. IMPLIZITER FEHLER

Oft wird der Fehler nicht explizit angegeben. Dann basiert die Genauigkeit auf der Anzahl Nachkommastellen: Mindestens 0.5 Einheiten ***nach der letzten Stelle*** bis maximal 2-4 Einheiten ***der letzten Stelle*** nach dem Punkt.

Beispiele:
- $t = 15.32s \Rightarrow \Delta t = [\pm 0.005s; \pm 0.04s]$
- $t = 15.3s \Rightarrow \Delta t = [\pm 0.05s; \pm 0.4s]$
- $t = 15.320s \Rightarrow \Delta t = [\pm 0.0005s; \pm 0.004s]$

2.2. MAXIMALER FEHLER

Der maximale Fehler in der Fehlerfortpflanzung bezeichnet den ***grössten möglichen Fehler***, der bei der Berechnung eines Ergebnisses von mehreren Messwerten entstehen kann. Will man diesen ***Fehler reduzieren***, verbessert man möglichst diejenigen Messwerte, welche in der Berechnung am grössten sind. *(z.B. in einer Formel a/b^2 sollte man die Genauigkeit von b verbessern, da es durch das Quadrieren einen grösseren Einfluss auf das Resultat als a hat)*

2.2.1. Absoluter Fehler

Beim absoluten Fehler wird die Abweichung in ***derselben Masseinheit*** angegeben wie der Wert. Wird meist bei ***Messergebnissen*** angegeben. Wird auch als m_{abs} bezeichnet.

Beispiel:

$$\overbrace{L = 5.8\text{cm}}^{\text{Messwert}}, \quad \overbrace{\Delta L = 0.1\text{cm}}^{\text{absoluter Fehler}} \quad \Rightarrow \quad 5.8\text{cm} \pm 0.1\text{cm} = [5.7, 5.9]\text{cm}$$

2.2.2. Relativer Fehler

Der relative Fehler ist ein ***Prozentwert***, also ***einheitslos***. Durch den relativen Fehler kann bestimmt werden, welcher Wert den ***grössten Anteil am Gesamtfehler*** hat, also am ehesten optimiert werden sollte. Er wird sowohl verwendet, um ein ***Gefühl für die Messgenauigkeit*** zu bekommen als auch bei ***Messgeräten*** als ***Angabe*** auf den ***eingestellten Messbereich***. Wird auch als m_{rel} bezeichnet.

Absoluten in relativen Fehler umwandeln:

$$\overbrace{L = 5.8\text{cm}}^{\text{Messwert}}, \quad \overbrace{\Delta L = 0.1\text{cm}}^{\text{absoluter Fehler}} \quad \Rightarrow \quad \frac{\text{absoluter Fehler}}{\text{Messwert}} = \frac{\Delta L}{L} = \frac{0.1\text{cm}}{5.8\text{cm}} = 0.017 = \underline{1.7\%}$$

Relativen in absoluten Fehler umwandeln:

$$\overbrace{L = 5.8\text{cm}}^{\text{Messwert}}, \quad \overbrace{\Delta L = 1.7\%}^{\text{relativer Fehler}} \quad \Rightarrow \quad \text{Messwert} \cdot \text{relativer Fehler} = 5.8\text{cm} \cdot 0.017 = \underline{0.1\text{cm}}$$

2.2.3. Partielle Differentiation

Partielle Differentiation ($\frac{\delta f}{\delta x}$) ist eine ***Ableitung*** eines Terms mit ***mehreren Variablen***. Dabei wird nach jeder Variable in einer eigenen Rechnung abgeleitet, wobei die anderen Variablen wie Konstanten behandelt werden. Mithilfe dieser Methode kann der Gesamtfehler eines Terms mit mehreren Variablen bestimmt werden.
TR: Wie normale Ableitung (Menü → 4 → 1), aber es müssen explizit Mal-Zeichen gesetzt werden.

Anmerkung: Im Skript wird die Notation $\frac{\delta x}{\delta y}$ bzw. δx für eine partielle Ableitung verwendet. $\frac{\delta f}{\delta x} \equiv \frac{\mathrm{d}}{\mathrm{d}x}(f)$

2.2.4. Berechnung

Bei ***Summen & Differenzen addieren*** sich die ***absoluten Fehler***. Beinhaltet die Formel eine ***Multiplikation/Division mit mehreren Messwerten***, darf der Absolute Fehler nicht berechnet werden. Stattdessen den ***relativen Fehler*** berechnen und in den absoluten Fehler ***umwandeln***. Bei ***Produkten & Quotienten addieren*** sich die ***relativen Fehler***.

Der Maximale Fehler Δf ***für absolute Fehler in einer Formel*** f wird über die partielle Differentiation berechnet:

$$\Delta f = \sum_{i=1}^{n} \left| \frac{\delta f}{\delta x_i} \cdot \Delta x_i \right|$$

x_i: Steuergrösse(n) ohne absolutem Fehler
f: Formel, welche Multiplikation/Division mit Steuergrössen enthält
$\delta f / \delta x_i$: Partielle Ableitung von f mit x_i
Δx_i: Absoluter Fehler der Steuergrösse x_i

Beispiel Summen und Differenzen: Was ist der gesamte absolute Fehler?

Absoluter Fehler für die Berechnung verwenden

Gegeben: $W = \frac{m}{2} - \frac{D}{8}, \quad m = 0.300\text{kg} \pm 0.002\text{kg}, \quad D = 10.0\text{kg} \pm 0.5\text{kg}$

1. Fehler in Formel einsetzen, ausrechnen.

$$W = \frac{m_{\text{abs}}}{2} - \frac{D_{\text{abs}}}{8} = \frac{0.002}{2} + \frac{0.5}{8} = 0.001 + 0.0625 = \underline{0.0635\text{kg}}$$

Beispiel Multiplikation und Division: Was ist der gesamte relative Fehler?

Relativer Fehler für die Berechnung verwenden

Ausgangslage: Es gibt meist eine Formel mit Multiplikation/Division sowie mehrere Steuergrössen mit einem absoluten Fehler.

Gegeben: Formel $g = \frac{4\pi^2 l}{t^2}, \quad l = 784\text{mm} \pm 2\text{mm}, \quad t = 17.7s \pm 0.1s$

1. Partielle Ableitung der Formel nach Steuergrössen (l, t) durchführen *(siehe Kapitel «Partielle Differentiation» (Seite 5))*

$$\frac{\delta g}{\delta t} = \frac{-8\pi^2 l}{t^3}, \quad \frac{\delta g}{\delta l} = \frac{4\pi^2}{t^2}$$

2. Formel anwenden, um den absoluten Fehler Δg der Formel g zu erhalten.

$$\Delta g = \left| \frac{-8\pi^2 l}{t^3} \cdot \Delta t \right| + \left| \frac{4\pi^2}{t^2} \cdot \Delta l \right| = \left| \frac{-8\pi^2 \cdot 784\text{mm}}{17.7s^3} \cdot 0.1s \right| + \left| \frac{4\pi^2}{17.7^2} \cdot 2\text{mm} \right| = 1.368 \text{ mm}/s^2$$

3. Relative Fehler der gewünschten Formel ausrechnen

$$\frac{\Delta g}{g} = 1.368 / \frac{4\pi^2 \cdot 784}{17.7^2} = 0.01385 = \underline{1.38\%}$$

TR: Seite 2.3 - Formel und Werte eingeben, gibt den relativen Fehler zurück

Beispiel Multiplikation und Division: Was ist der gesamte absolute Fehler?

Relativer Fehler für die Berechnung verwenden

Gegeben: Formel $g = \frac{4\pi^2 l}{t^2}, \quad g_{\text{rel}} = 1.383\%, \quad l = 784\text{mm} \pm 2\text{mm}, \quad t = 17.7s \pm 0.1s$

1. Werte ohne Verwendung des Fehlers in Formel einsetzen

$$g = \frac{4\pi^2 \cdot 784}{(17.7)^2} = 98.794 \text{ mm}/s^2$$

2. Resultat durch 100 teilen und mit relativem Fehler multiplizieren

$$g_{\text{abs}} = 98.794/100 \cdot 1.383\% = 1.366 \Rightarrow \underline{98.789 \text{ mm}/s^2 \pm 1.366 \text{ mm}/s^2}$$

2.3. FEHLERGRENZEN

Die Fehlergrenzen bestimmen die ***maximal mögliche Abweichung*** von einem Wert innerhalb einer ***festen Intervallgrenze***. Sie können für ***relative & absolute Fehler*** bestimmt werden.
Aus dem ***absoluten Fehler*** können die Intervallgrenzen gebildet werden, indem der absolute Fehler auf den Messwert angewendet wird und der Maximal- & Minimalwert des Messwertes ***als Intervall*** angegeben wird.

Zur Bestimmung gibt es ***2 Methoden***: Minimal-/Maximalwert und partielle Differentiation. Ersteres wird nicht empfohlen, da es bei mehreren Variablen schnell komplex wird.

Beispiel mit partieller Differentiation:
Bei einer Messung werden folgende Werte gemessen. W stellt eine Beziehung der gemessenen Werte dar.

$$m = 0.300\text{kg} \pm 0.002\text{kg},\ \ h = 0.3m \pm 0.002m,$$
$$D = 10.0\frac{\text{kg}}{s^2} \pm 0.5\frac{\text{kg}}{s^2},\quad g = 9.81\frac{m}{s^2} \pm 0.01\frac{m}{s^2},\quad W = \frac{m \cdot g \cdot h}{2} + \frac{D \cdot h^2}{8}$$

1. Nach jeder Variable ableiten

$$\frac{\delta W}{\delta m} = \frac{\mathrm{d}}{\mathrm{d}m}\left(\frac{m \cdot g \cdot h}{2} + \frac{D \cdot h^2}{8}\right) = \frac{g \cdot h}{2}$$
$$\frac{\delta W}{\delta g} = \frac{\mathrm{d}}{\mathrm{d}g}\left(\frac{m \cdot g \cdot h}{2} + \frac{D \cdot h^2}{8}\right) = \frac{m \cdot h}{2}$$
$$\frac{\delta W}{\delta h} = \frac{\mathrm{d}}{\mathrm{d}h}\left(\frac{m \cdot g \cdot h}{2} + \frac{D \cdot h^2}{8}\right) = \frac{m \cdot g}{2} + \frac{D \cdot 2h}{8} = \frac{m \cdot g}{2} + \frac{D \cdot h}{4}$$
$$\frac{\delta W}{\delta D} = \frac{\mathrm{d}}{\mathrm{d}D}\left(\frac{m \cdot g \cdot h}{2} + \frac{D \cdot h^2}{8}\right) = \frac{h^2}{8}$$

2. Ableitungen und Fehler in Formel einsetzen

$$\begin{aligned}\Delta W &= \left|\frac{g \cdot h}{2} \cdot \Delta m\right| + \left|\left(\frac{m \cdot g}{2} + \frac{D \cdot h}{4}\right) \cdot \Delta h\right| + \left|\frac{h^2}{8} \cdot \Delta D\right| + \left|\frac{m \cdot h}{2} \cdot \Delta g\right| \\ &= \left|\frac{9.81 \cdot 0.3}{2} \cdot 0.002\right| + \left|\left(\frac{0.3 \cdot 9.81}{2} + \frac{10 \cdot 0.3}{4}\right) \cdot 0.002\right| + \left|\frac{0.3^2}{8} \cdot 0.5\right| + \left|\frac{0.3 \cdot 0.3}{2} \cdot 0.01\right| \\ &= 0.00294 + 0.00444 + 0.005626 + 0.00045 = \pm 0.0135\,\frac{\text{km} \cdot m^2}{s^2}\end{aligned}$$
$$W = (0.554 \pm 0.0135)\frac{\text{km} \cdot m^2}{s^2} = [0.541, 0.568]\frac{\text{km} \cdot m^2}{s^2}$$

2.4. WAHRSCHEINLICHER FEHLER

Der wahrscheinliche Fehler gibt ähnlich wie der maximale Fehler ein Fehlerintervall zu einem Messwert an. Jedoch befindet sich der Messwert ***nur sehr wahrscheinlich*** in diesem Intervall, im Gegensatz zum absoluten Fehler, dessen Wert sich definitiv in diesem Intervall befindet.

3. STATISTIK-GRUNDLAGEN & UNTERSUCHUNG

3.1. GRUNDBEGRIFFE

- ***Statistik:*** Entwicklung & Anwendung von Methoden zur Erhebung, Aufbereitung, Analyse & Interpretation von Daten
- ***Beschreibende Statistik:*** Vollständige Kenntnis über Untersuchungsobjekt *(z.B. CPU-Load für alle Server der Firma bekannt)*
- ***Schliessende Statistik:*** Untersuchungsdaten liegen nur teilweise vor *(z.B. repräsentative Umfragen)*
- ***Merkmalsträger:*** Gegenstand einer statistischen Untersuchung. Besitzt Merkmale.
  (Beispiele: Server, Werkmaschine, Personengruppe)
- ***Merkmal:*** Eine Eigenschaft, die bei einer statistischen Untersuchung eines Merkmalträgers von Bedeutung ist
  (z.B. Ausfallzeit, Servicedauer, Kosten, Latenz, Geschlecht, Bildungsgrad…)
- ***Merkmalswert:*** Der Wert eines Merkmals, der aufgrund Beobachtung, Messung oder Befragung festgestellt wurde. Ist oft Median mehrerer Messungen. *(z.B. CPU-Last = 45%, Kosten pro Jahr = 4'500 Fr., Latenz = 25ms)*
- ***Grundgesamtheit:*** Menge aller Merkmalsträger, die gemeinsame Abgrenzungsmerkmale besitzen
- ***Abgrenzungsmerkmale:*** Merkmale, mit denen Merkmalsträger gruppiert werden können
  - ***Räumlich:*** Örtliche Gemeinsamkeit *(z.B. alle Server einer Serverfarm, alle Personen an der OST, …)*
  - ***Zeitlich:*** Ergebnisse innerhalb eines Zeitraums *(z.B. alle Messergebnisse eines Tages, Produktion pro Monat, …)*
  - ***Sachlich:*** Merkmalträger gleichen Typus *(z.B. alle Maschinen eines Herstellers, alle Schüler einer Klasse, …)*
- ***Urliste:*** Unsortierte, nicht aufbereitete Daten, die direkt von der Messung stammen.
- ***Primärstatistik:*** Erheben neuer Daten für die Untersuchung. Die erhobenen Daten sind optimal auf die Untersuchungsfrage zugeschnitten, aber die Erhebung ist teuer & aufwändig.
- ***Sekundärstatistik:*** Verwenden bereits vorhandener Daten für die Untersuchung. Günstiger, aber die Daten sind nicht auf die Fragestellung ausgerichtet und evtl. veraltet.
- ***Empirische Daten:*** Am realen Objekt gemessene Daten *(durch Experiment, Messungen etc.)*
- ***Theoretische Daten:*** Daten aus Modell oder Theorie, der eine theoretische Verteilungsfunktion zugrunde liegt
- ***Polygonzug:*** Liniendiagramm *(will de Rinkel ums verrecke eloquent sii muess)*

3.2. SKALEN

Die Merkmalswerte können nach einem bestimmten Ordnungsprinzip als ***Werte*** in eine ***Skala*** eingetragen werden.

Skalentyp	*Erklärung*	*Vergleich*	*Beispiel*
Nominalskala	Werte sind qualitativ gleich, haben keine Wertigkeit	Können nur durch Häufigkeit/Anzahl verglichen werden	Familienstand *(Ledig, verheiratet, …)* Strassen *(Paradeplatz, Seestrasse, …)*
Ordinalskala	Werte haben eine relative Rangordnung	Können nach Intensität geordnet werden	Flugklassen *(Economy, Business Class, First Class)* kalt, warm, heiss
Intervallskala	reelle Zahlen, Nullpunkt ist willkürlich, negative Werte möglich	Der absolute Abstand zwischen Werten kann gemessen werden	Temperatur in °C *-12, 0, 25,…* Uhrzeit *20:00, 0:00, 09:35, …* *8:00 ist nicht doppelt so spät wie 4:00*
Verhältnisskala	reelle Zahlen, Nullpunkt ist absolut Null, keine negativen Werte	Der verhältnismässige Abstand (vielfaches) kann gemessen werden	Gewicht in kg *0, 25, 98, …* Temperatur in °K: *0, 1, 300, …* *50kg ist doppelt so schwer wie 25kg*

Die Intervall- & Verhältnisskala werden unter dem Oberbegriff ***Kardinalskala (metrische Skala)*** zusammengefasst.

ablesbare Merkmalswerte	*Nominalskala*	*Ordinalskala*	*Intervallskala*	*Verhältnisskala*
Verschiedenartigkeit (=, ≠)	✓	✓	✓	✓
Rangordnung (<, >)	×	✓	✓	✓
einfache Abstände / Intervall (+, -)	×	×	✓	✓
verhältnismässige Abstände (*, /)	×	×	×	✓

3.3. ABLAUF DER UNTERSUCHUNG

Die ***Experimentplanung*** besteht aus 3 Phasen: ***Datenerhebung***, ***Datenaufbereitung & -darstellung*** und ***Datenanalyse & -interpretation***. In der Planungsphase muss festgelegt werden, welche ***Merkmale*** bei welchen Merkmalsträgern mit welcher Technik zu erheben sind, welche ***Aufbereitungsverfahren*** einzusetzen sind, welche ***Daten-Darstellungsformen*** zu verwenden ist und welche ***statistischen Analyseverfahren*** angewendet werden.
Werden die Daten ***gezielt und periodisch*** durch Experimente erfasst & gepflegt, spricht man von ***Data Farming***.

3.4. EINFACHE HÄUFIGKEITSVERTEILUNG

Gibt an, ***wie häufig*** ein Merkmalswert x_i aufgetreten ist.

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

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

$$\sum_{i=1}^{n} f_i = 1$$

h_i: Absolute einfache Häufigkeit, Anzahl Merkmalsträger mit Wert x_i
f_i: Relative einfache Häufigkeit, Prozentanteil der Merkmalsträger mit Wert x_i
n: Gesamtanzahl aller Merkmalsträger
v: Anzahl verschiedener Merkmalswerte

Beispiel: Verteilung der Alterskategorien der Mitarbeiter eines Spitals mit $n = 50$ Personen:
- $\boldsymbol{h_1}$: 10 Mitarbeiter in der Altersklasse 30 $\Rightarrow$ $\boldsymbol{f_1} = h_1/n = 10/50 = 20\%$
- $\boldsymbol{h_2}$: 15 Mitarbeiter in der Altersklasse 40 $\Rightarrow$ $\boldsymbol{f_2} = h_2/n = 15/50 = 30\%$
- $\boldsymbol{h_3}$: 25 Mitarbeiter in der Altersklasse 50 $\Rightarrow$ $\boldsymbol{f_3} = h_3/n = 25/50 = 50\%$

3.5. KUMULIERTE HÄUFIGKEITSVERTEILUNG

Auch ***Summenhäufigkeit*** genannt. Misst die Häufigkeit über verschiedene Messungen hinweg, indem sie alle bisherigen Messungen (die einfache Häufigkeit h_i bzw. f_i) aufsummiert.

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

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

$$\sum_{i=1}^{n} f_i = 1$$

H_i: Absolute kumulierte Häufigkeit, Anzahl Messungen mit Merkmalswert $\leq x_i$
F_i: Relative kumulierte Häufigkeit, Prozentanteil der Messungen mit Wert i
n: Gesamtanzahl aller Messungen
k: Anzahl verschiedener Merkmalswerte

Beispiel: Gleiche Aufgabenstellung wie bei der einfachen Häufigkeit

$$\boldsymbol{H_1} = h_1 = 10 \quad \Rightarrow \quad \boldsymbol{F_1} = H_1/n = 10/50 = 20\%$$
$$\boldsymbol{H_2} = h_1 + h_2 = 25 \quad \Rightarrow \quad \boldsymbol{F_2} = H_2/n = 25/50 = 50\%$$
$$\boldsymbol{H_3} = h_1 + h_2 + h_3 = 50 \quad \Rightarrow \quad \boldsymbol{F_3} = H_3/n = 50/50 = 100\%$$

3.6. KLASSIFIZIERTE HÄUFIGKEITSVERTEILUNG

Um Verteilungen mit mehr als 10 Merkmalswerten übersichtlich darstellen zu können, werden Merkmalswerte zu ***Klassen*** zusammengefasst. Die Differenz zwischen der Ober- und Untergrenze einer Klasse ist die ***Klassenbreite***.

Klassenbreite nach Sturges (m aufrunden auf Ganzzahl)	***Faustregel für Anzahl Klassen***
$m \approx 1 + 3.32 \cdot \log(n)$ $K_b = \frac{x_{\max} - x_{\min}}{m}$	$j_{\max} = \sqrt{n}$

j: Anzahl Klassen, K_b: Klassenbreite, x_j^u: untere Klassengrenze, x_j^o: obere Klassengrenze.

Beispiel: Alter der Bewohner einer Nachbarschaft

j	***Rechnungsbetrag*** $x_j^u \leq x_i \leq x_j^o$	***Absolute einfache Häufigkeit*** h_j	***Absolute kumul. Häufigkeit*** H_j	***Relative einfache Häufigkeit*** f_j	***Relative kumul. Häufigkeit*** F_j
1	0 bis 20	8	8	0.05	0.05
2	20 bis 40	40	48	0.25	0.3
3	40 bis 60	80	128	0.5	0.8
4	60 bis 80	32	160	0.2	1
Σ		160			

3.6.1. Klassenrechnungen

Um den Anteil der Klassen ***unterhalb*** eines Wertes g zu bestimmen, kann die ***lineare Interpolation*** zwischen Klassengrenzen genutzt werden:

$$F(x < g) = F_j + \frac{F_{j+1} - F_j}{x_{j+1}^o - x_{j+1}^u} \cdot \left(x - x_{j+1}^u\right)$$

g: Grenzwert
j: Klasse unterhalb des Grenzwertes
$j+1$: Klasse, in welcher sich der Grenzwert befindet

Beispiel: Anteil Bewohner unter 50 Jahre in der obigen Tabelle: $g = 50,\ j = 2,\ j+1 = 3$

$$F(x < 50) = F_2 + \frac{F_3 - F_2}{x_3^o - x_3^u} \cdot (x - x_3^u) = 0.3 + \frac{0.8 - 0.3}{60 - 40} \cdot (50 - 40) = 0.55 = \underline{55\%}$$

Für den Anteil der Klassen ***überhalb*** eines Wertes wird zuerst die lineare Interpolation durchgeführt und dann 1 minus dieses Resultat gerechnet.

$F(x > 50) = 1 - F(x < 50) = 1 - 0.55 = 0.45 = \underline{45\%}$

Liegen ***unterschiedliche Klassenbreiten*** vor, kann nicht mehr mit der absoluten Häufigkeit h_i gearbeitet werden. Je breiter die Klasse, desto mehr Elemente könnten sich darin befinden. Hier muss auf die ***Häufigkeitsdichte*** d_i ausgewichen werden.

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

h_i: Absolute Häufigkeit der Klasse
x_j^u: untere Klassengrenze
x_j^o: obere Klassengrenze

4. HÄUFIGKEITSVERTEILUNGEN & PARAMETER

Typische Eigenschaften der Häufigkeitsverteilung können mit Hilfe von Kenngrössen, den sogenannten ***Parametern***, beschrieben werden. Dabei werden viele Einzelinformationen zu wenigen, aussagekräftigen Grössen verdichtet.

4.1. MITTELWERTE & LAGEPARAMETER

Lageparameter sind ***Kennzahlen***, welche eine zentrale Tendenz der Messwerte aufzeigen, z.B. das Zentrum der Werte.

4.1.1. Modus

Derjenige Merkmalswert, der am häufigsten beobachtet wird. Sind das mehrere, gibt es mehrere Modi (***Multimodi***). Kann für jede Verteilungsart bzw. jedes Skalenniveau bestimmt werden. Ist von Ausreissern unbeeinflusst. Um bei einer klassifizierten Häufigkeit einen einzigen Wert zu erhalten *(und nicht nur die häufigste Klasse)*, kann der Modus mit dieser Formel geschätzt werden:

$$\text{Mo} = x_m^u + \frac{h_m - h_{m-1}}{(h_m - h_{m-1}) + (h_m - h_{m+1})} \cdot (x_m^o - x_m^u)$$

m: Klassennummer der häufigsten Klasse
h_m: Modusklasse (h_j Wert der häufigsten Klasse)
x_m^o / x_m^u: Obere/Untere Klassengrenze

Sind die Klassen unterschiedlich breit, muss zuerst noch die Klasse mit der höchsten Dichte d_i gefunden werden:

$$d_i = \frac{h_i}{x_i^o - x_i^u} \quad \Rightarrow \quad \max(d_i) \quad \Rightarrow \quad \text{Mo} = x_i^u + \frac{h_i - h_{i-1}}{(h_i - h_{i-1}) + (h_i - h_{i+1})} \cdot (x_i^o - x_i^u)$$

Beispiele ohne Klasse:

$$[\text{Löwe}, \text{Giraffe}, \text{Affe}, \text{Löwe}] \Rightarrow \text{Mo} = \underline{\text{Löwe}}, \quad [1,2,6,5,3,4,3] \Rightarrow \text{Mo} = \underline{3}, \quad [1,2,6,3,4,3,2] \Rightarrow \text{Mo} = \underline{[2,3]}$$

Beispiel mit Klassen:
Daten siehe Beispiel-Tabelle in «Klassifizierte Häufigkeitsverteilung» (Seite 10)

1. Klasse auswählen, die am meisten auftaucht (grösster h_j-Wert)

$$m = 3, \quad h_m = h_3 = 80$$

2. Grenzen (x_m^u / x_m^o) und Modusklassen (h_j) aus Tabelle ablesen

$$x_m^u = x_3^u = 40, \quad x_m^o = x_3^o = 60, \quad h_{m-1} = h_2 = 40, \quad h_{m+1} = h_4 = 32$$

3. Werte in Formel einsetzen

$$\text{Mo} = 40 + \frac{80 - 40}{(80 - 40) + (80 - 32)} \cdot (60 - 40) = \underline{49.\overline{09}}$$

4.1.2. Median

Merkmalswert, der in der Rangordnung die mittlere Position einnimmt *(Wert mit Index $n/2$)*. Um den Median berechnen zu können, müssen die Merkmale zuerst ***sortiert*** werden, deswegen müssen sie mindestens ordinalskaliert sein *(keine Nominalwerte)*. Der Median ist ***unbeeinflusst von Ausreissern***, darum gut geeignet für ***schiefe Verteilungen*** *(asymmetrische/ungerade Graphen)*.

Es wird unterschieden zwischen einer ***geraden*** und ***ungeraden*** Anzahl Werte. Ergibt die Berechnung des mittleren Index eine Kommazahl, müssen die Werte darunter und darüber addiert und durch 2 geteilt werden.

ungerade Anzahl Werte	***gerade Anzahl Werte***
$\text{Me} = x_{(n+1)/2}$	$\text{Me} = \frac{x_{n/2} + x_{(n/2)+1}}{2}$

Wie beim Modus wird der Median bei der klassifizierten Häufigkeit wieder durch eine Formel abgeschätzt:

$$\text{Me} = x_m^u + \frac{\frac{n}{2} - H_{m-1}}{H_m - H_{m-1}} \cdot (x_m^o - x_m^u)$$

Beispiele ohne Klassen:

Auf TR: Menü → 6 → 3 → 4: Median. Werte in geschwungene Klammer packen (CTRL + «) »). Müssen nicht sortiert sein.

Ungerade:

$$\{44\%, 42\%, 40\%, 43.5\%, 41.5\%\} \Rightarrow \text{sortieren} \Rightarrow \{40\%, 41.5\%, 42\%, 43.5\%, 44\%\}, \quad n = 5$$
$$\text{Me} = x_{(n+1)/2} = x_{(5+1)/2} = x_3 = \underline{42\%}$$

Gerade:

$$\{44\%, 42\%, 43.5\%, 41.5\%\} \Rightarrow \text{sortieren} \Rightarrow \{41.5\%, 42\%, 43.5\%, 44\%\}, \quad n = 4$$
$$\text{Me} = \frac{x_{n/2} + x_{(n/2)+1}}{2} = \frac{x_{4/2} + x_{(4/2)+1}}{2} = \frac{x_2 + x_3}{2} = \frac{42\% + 43.5\%}{2} = \underline{42.75\%}$$

Beispiel mit Klassen:

Daten siehe Beispiel-Tabelle in «Klassifizierte Häufigkeitsverteilung» (Seite 10).

1. Klasse ermitteln, die sich von der Anzahl Merkmalsträger in der Mitte befindet

$$\frac{n}{2} = \frac{h_n}{2} = \frac{H_4}{2} = \frac{160}{2} = 80 \;\Rightarrow\; 48 < 80 < 128 \;\Rightarrow\; m = 3$$

2. Grenzen & absolute kumulierte Häufigkeiten (H_j) aus Tabelle ablesen

$$x_m^u = x_3^u = 40, \quad x_m^o = x_3^o = 60, \quad H_{m-1} = H_2 = 48, \quad H_m = H_3 = 128$$

3. Werte in Formel einsetzen

$$\text{Me} = 40 + \frac{160/2 - 48}{128 - 48} \cdot (60 - 40) = \underline{48}$$

4.1.3. Quantile & Quartile

Quantile zerlegen Merkmalswerte in eine ***gewisse Anzahl Teile***, ***Quartile*** in ***vier Teile***. Ebenfalls geläufig sind ***Dezile*** *(10 Teile)* und ***Perzentile*** *(100 Teile)*. Das 1. Quantil wird auch als 25% Quantil/Perzentil bezeichnet. Die Berechnung erfolgt analog des Medians, anstatt $n/2$ wird einfach $n/4$ gerechnet. Soll also zb. das 3. Quartil (75% Quantil) ausgerechnet werden, muss $(3n)/4$ gerechnet werden.

ungerade	***gerade***	***klassifizierte Häufigkeit***
$Q_1 = x_{(n+1)/4}$ *falls $(n+1)/4$ keine Ganzzahl ist, muss das arithmetische Mittel von den zwei angrenzenden Werten verwendet werden.*	$Q_1 = \frac{x_{n/4} + x_{(n/4)+1}}{2}$	$Q_1 = x_m^u + \frac{\frac{n}{4} - H_{m-1}}{H_m - H_{m-1}} \cdot (x_m^o - x_m^u)$
$Q_3 = x_{((n+1)\cdot 3)/4}$	$Q_3 = \frac{x_{3n/4} + x_{(3n/4)+1}}{2}$	$Q_3 = x_m^u + \frac{\frac{3n}{4} - H_{m-1}}{H_m - H_{m-1}} \cdot (x_m^o - x_m^u)$

Berechnung des zentralen 80%-Dezilabstand:
Dazu muss das 9. Dezentil minus das 1. Dezentil gerechnet werden *(beim 60%-Dezilabstand wären es 8. Dezentil minus 2. Dezentil)*. Das erste Dezil liegt in der 2. Klasse ($\mathbf{400}/10 = 40$), das letzte in der 5. Klasse ($(9 \cdot \mathbf{400})/10 = 360$).

i	von (x_i^u)	bis (x_i^o)	h_i	d_i	H_i
1	4	10	20	3.3	20
2	10	20	160	16.0	180
3	20	30	80	8.0	260
4	30	40	40	4.0	300
5	40	80	88	2.2	388
6	80	120	12	0.3	400
			$n = \mathbf{400}$		

$$D1 = x_2^u + \frac{(1n/10) - H_1}{h_2} \cdot (x_2^o - x_2^u) = 10 + \frac{40 - 20}{160} \cdot (20 - 10) = 11.25$$

$$D9 = x_5^u + \frac{(9n/10) - H_4}{h_5} \cdot (x_5^o - x_5^u) = 40 + \frac{360 - 300}{88} \cdot (80 - 40) = 67.27$$

$$I_{80} = D9 - D1 = 67.27 - 11.25 = \underline{56.02}$$

4.1.4. Arithmetisches Mittel (Durchschnitt)

De Durchschnitt halt, de söttsch kenne Kolleg. Häufig benutzt, kann aber zu starken Verzerrungen und Fehlschlüssen führen, da Aussreisser den Durchschnitt stark verändern können. Darum nicht geeignet für schiefe Verteilungen, sondern nur für eingipflige, symmetrische Verteilungen *(z.B. Bell Curve, Normalverteilung)*

Auf TR: Menü → 6 → 3 → 3: Mittelwert. Werte in geschwungene Klammer schreiben (CTRL + «) »)

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

$\bar{x}$: Arithmetisches Mittel
n: Anzahl der Werte
x_i: Datenwerte
h_i: Häufigkeit der Datenwerte *(1, wenn nichts angegeben)*

Klassifiziertes Arithmetisches Mittel
Auch hier muss die ***klassifizierte Häufigkeit*** wieder speziell behandelt werden: Zuerst muss die ***Klassenmitte*** $\acute{x}_i$ für jede Klasse berechnet werden, dann die Summe von allen $\acute{x}_i \cdot h_i$ mit $\frac{1}{n}$ multiplizieren.

$$\bar{x} = \frac{1}{n} \sum_{i=1}^{v} \acute{x}_i \cdot h_i$$

$$\acute{x}_i = \frac{x_i^o + x_i^u}{2}$$

$\bar{x}$: Arithmetisches Mittel
n: Anzahl der Werte
$\acute{x}_i$: Klassenmitte von Klasse i
h_i: Häufigkeit von Klasse i

Beispiel mit Klassen:
Daten siehe Beispiel-Tabelle in «Klassifizierte Häufigkeitsverteilung» (Seite 10)

1. Klassenmitten berechnen

$$\acute{x}_i = \left(\frac{x_i^u + x_i^o}{2}\right) \Rightarrow \left\{\frac{0+20}{2} = 10, \quad \frac{20+40}{2} = 30, \quad \frac{40+60}{2} = 50, \quad \frac{60+80}{2} = 70\right\}$$

2. Produkte $\acute{x}_i \cdot h_i$ für alle Klassen berechnen

$$\acute{x}_i \cdot h_i \Rightarrow \{10 \cdot 8 = 80, \quad 30 \cdot 40 = 1'200, \quad 50 \cdot 80 = 4'000, \quad 70 \cdot 32 = 2'240\}$$

3. Summen von allen h_i und allen $\acute{x}_i \cdot h_i$ bestimmen

$$\sum_{i=1}^{n} h_i = 160 \quad \sum_{i=1}^{n} \acute{x}_i \cdot h_i = 7520$$

4. Werte in Formel einsetzen

$$\bar{x} = \frac{1}{160} \cdot 7'520 = \underline{47.023}$$

4.1.5. Harmonisches Mittel

Um den ***Durchschnitt verhältnisskalierter Zahlen*** zu berechnen, verwendet man das harmonische Mittel. Die Merkmalswerte dürfen untereinander keine gemischten Vorzeichen haben. Das einfache Harmonische Mittel wird bei einheitslosen oder beziehungslosen Zahlen verwendet *(siehe «Gewichtetes Harmonisches Mittel» (Seite 14))*.

$$\overline{\text{MH}} = \frac{n}{\sum_{i=1}^{n} \frac{1}{x_i}}$$

In den allermeisten Fällen verwendet man jedoch das gewichtete harmonische Mittel.

Gewichtetes Harmonisches Mittel

Ein Sonderfall sind Brüche, die Beziehungen widerspiegeln; sie werden meist mit ***«x pro y»*** bezeichnet *(z.B. Kilometer pro Stunde, Preis pro Liter, Personen pro Quadratmeter)*. Bei diesen muss das ***gewichtete Harmonische Mittel*** verwendet werden:

$$\overline{\text{MH}} = \frac{\sum_{i=1}^{v} h_i}{\sum_{i=1}^{v} h_i / x_i}$$

Beispiele für gewichtetes HM:

a) *Ein Kipplaster fährt auf dem* 5km *Hinweg* $10\ \text{km/h}$ *und auf dem Rückweg* $30\ \text{km/h}$. *Was ist die mittlere Geschwindigkeit*?

$$h_1 = h_2 = 5\text{km}, \quad x_1 = \frac{5\text{km}}{10\text{km/h}}, \quad x_2 = \frac{5\text{km}}{30\text{km/h}}$$

$$\overline{\text{MH}} = \frac{\sum_{i=1}^{v} h_i}{\sum_{i=1}^{v} h_i / x_i} = \frac{5+5}{(5/10)+(5/30)} = \underline{15\text{km/h}}$$

b) *Eine moderne Abfüllanlage füllt* $50'000$ *Flaschen pro Stunde ab, eine ältere Anlage nur* $30'000$ *Flaschen pro Stunde. Wie viele Flaschen werden durchschnittlich pro Stunde abgefüllt, wenn auf der modernen Anlage* $300'000$ *Flaschen und auf der älteren* $150'000$ *Flaschen abgefüllt werden?*

1) Aus Text Klassenmittelwerte $\acute{x_i}$ suchen *(Beziehungszahl, also Wert pro h)*

$$\acute{x_1} = 30'000\ \text{Fl/h}, \quad \acute{x_2} = 50'000\ \text{Fl/h}$$

2) Aus Text Anzahl Merkmalsträger h_i suchen *(erreichte Anzahl X)*

$$h_1 = 150'000\ \text{Fl} \quad h_2 = 50'000\ \text{Fl}$$

3) Werte in Formel einsetzen

$$\overline{\text{MH}} = \frac{30'000 + 50'000}{300'000/50'000 + 150'000/30'000} = \underline{40'909\ \text{Fl/h}}$$

4.1.6. Geometrisches Mittel

Die n-te Wurzel aus dem Produkt aller beobachteten Merkmalswerte. Es zeigt die ***durchschnittliche Veränderung*** zwischen Merkmalswerten auf: Sie sind Quotienten aus zeitlich benachbarten Grössen *(z.B. Wachstum)*. Die Werte müssen wegen der Division > 0 sein. Funktioniert nur für mindestens verhältnisskalierte Merkmalswerte. Nicht sinnvoll für klassifizierte Häufigkeiten.

Formel für durchschnittlichen Merkmalswert	*Formel für durchschnittliche Veränderung*
$\text{MG} = \sqrt[n]{\prod_{i=1}^{n} x_i}$	$\text{MG} = \sqrt[n-1]{\frac{\text{Endwert}}{\text{Anfangswert}}}$

Beispiel für durchschnittlichen Merkmalswert

$x_i = (3.2, 3.1, 3.4, 3.6, 3.4, 3.1, 3.3, 1.9, 2.0)$

1. Anzahl Werte zählen
   $n = 9$
2. Werte multiplizieren
   $\prod_{i=1}^{n} x_i = 16048.38$
3. Geometrisches Mittel berechnen
   $\sqrt[9]{16048.38} = 2.93$

Beispiel für durchschnittliche Veränderung

$[40] \overrightarrow{\cdot 1.2}\ [48] \overrightarrow{\cdot 1.25}\ [60] \overrightarrow{\cdot 0.95}\ [57]$

1. Anzahl Werte zählen
   $n = 4$
2. Anfangs- und Endwert in Formel einfügen
   $\sqrt[3]{\frac{57}{40}} = \underline{1.125}$

4.2. STREUMASSE

Streumasse sind Kennzahlen, die die ***Verteilung*** der Messwerte ***um ein Zentrum*** angeben.

4.2.1. Spannweite

Differenz aus dem grössten und kleinsten Merkmalswert. Ist als Streumass geeignet, wenn allein die Länge des Streubereiches interessiert, da es keine Information über die Streuung an sich liefert. Äusserst empfindlich auf Aussreisser. Merkmale müssen mindestens ***Intervallskaliert*** sein. *(v = Grösster Wert der Klasse)*

Normale Berechnung	*Klassifizierte Häufigkeit*
$R = \text{grösster Wert} - \text{kleinster Wert}$	$R = x_v^o - x_1^u$

Beispiel ohne Klassen:

$$x_i = (3.2, 3.1, 3.4, 3.6, 3.4, 3.1, 3.3, 1.9, 2.0) \overset{\text{sortieren}}{\Rightarrow} (1.9, 2.0, 3.1, 3.1, 3.2, 3.3, 3.4, 3.4, 3.6)$$

$$x_{\min} = 1.9, \quad x_{\max} = 3.6 \quad R = 3.6 - 1.9 = 1.7$$

Beispiel mit Klassen:

Daten siehe Beispiel-Tabelle in «Klassifizierte Häufigkeitsverteilung» (Seite 10)

$$x_1^u = 0, \quad x_v^o = 80, \quad R = 80 - 0 = \underline{80}$$

4.2.2. Zentraler Quartilsabstand (ZQA)

Abstand zwischen dem 1. & 3. Quartil. Zeigt, wie nahe die Werte um den Median herum zusammenliegen. Dadurch sind Ausreisser kein Problem. Ist geeignet, um den Kernbereich einer Häufigkeitsverteilung darzustellen. Werte müssen mindestens Intervallskaliert sein.

$$\text{ZQA} = Q_3 - Q_1$$

Beispiel Ausfallzeiten (in h) von 20 Maschinen:
Frage: Wie streut die Ausfallzeit der Maschinen in Abhängigkeit von der Wahrscheinlichkeit um das Mittel 0.5?

$$Q_1 = x_{[1/4 \cdot n]} = x_{[1/4 \cdot 20]} = x_{[5]} = 2h$$
$$Q_3 = x_{[3/4 \cdot n]} = x_{[3/4 \cdot 20]} = x_{[15]} = 11h$$
$$\text{ZQA} = 11h - 2h = \underline{9h}$$

Ausfallzeit (h)	0	2	5	6	7	11	12	14
h_i	4	2	2	2	4	3	2	1
H_i	4	**6**	8	10	14	**17**	19	20

4.2.3. Mittlere absolute Abweichung

Durchschnittliche Entfernung der Merkmalswerte vom arithmetischen Mittel. Ist Einheitenbehaftet. Wird häufig bei der Beschreibung der erfassten Datenmenge mit angegeben. Besser geeignet als Varianz/Standardabweichung. Ausreisser werden erfasst, Gefahr einer verzerrten Beschreibung entsteht.

$$\delta = \frac{1}{n} \sum_{i=1}^{v} |x_i - \bar{x}|$$

n: Anzahl der Merkmalsträger
v: Anzahl der verschiedenen Merkmalswerte
x_i: Merkmalswert mit Index i
h_i: absolute einfache Häufigkeit beim Merkmalswert x_i
$\bar{x}$: arithmetisches Mittel der Merkmalswerte

Bei der Klassifizierten Häufigkeit wird die Klassenmitte $\acute{x}$ verwendet und zusätzlich noch die absolute einfache Häufigkeit h_i dazugerechnet:

$$\delta = \frac{1}{n} \sum_{i=1}^{v} |\acute{x}_i - \bar{x}| \cdot h_i$$

n: Anzahl der Merkmalsträger
v: Anzahl der verschiedenen Merkmalswerte
h_i: absolute einfache Häufigkeit beim Merkmalswert x_i
$\bar{x}$: arithmetisches Mittel der Merkmalswerte
$\acute{x}_i$: Klassenmitte

Beispiel ohne Klassen:
Urliste: $x_i = (3.2, 3.1, 3.4, 3.1, 1.9, 2.0)$
1. Arithmetisches Mittel berechnen *(Berechnung des Werts siehe «Arithmetisches Mittel (Durchschnitt)» (Seite 13))*

$$n = 6, \quad \bar{x} = 2.78$$

2. Werte in Formel einsetzen

$$\delta = \frac{1}{6} \cdot |3.2 - 2.78| + |3.1 - 2.78| + |3.4 - 2.78| + |3.1 - 2.78| + |1.9 - 2.78| + |2.0 - 2.78|$$
$$= \frac{1}{6} \cdot 3.34 = \underline{0.556}$$

Beispiel mit Klassen:
Daten siehe Beispiel-Tabelle in «Klassifizierte Häufigkeitsverteilung» (Seite 10):
1. Arithmetisches Mittel berechnen *(Berechnung des Werts siehe «Arithmetisches Mittel (Durchschnitt)» (Seite 13))*:

$$\bar{x} = 47.023$$

2. Für jede Klasse $|\acute{x}_i - \bar{x}|$ berechnen

$$|\acute{x}_i - \bar{x}| \Rightarrow \{|10 - 47.023| = 37.023, |30 - 47.023| = 17.023, |50 - 47.023| = 2.977, |70 - 47.023| = 22.977\}$$

3. Für jede Klasse $|\acute{x}_i - \bar{x}| \cdot h_i$ berechnen

$$|\acute{x}_i - \bar{x}| \cdot h_i \Rightarrow \{8 \cdot 37.023 = 296.184, 40 \cdot 17.023 = 680.92, 80 \cdot 2.977 = 239.76, 32 \cdot 22.977 = 735.264\}$$

4. Summieren und in Formel einfügen

$$\delta = \frac{1}{4} \cdot (296.184 + 680.92 + 239.76 + 735.264) = \frac{1}{4} \cdot 1'952.128 = \underline{488.032}$$

4.2.4. Varianz & Standardabweichung

Die Varianz (σ^2 bzw. $\mathrm{var}(x)$) ist die Verteilung der Werte um das arithmetische Mittel. Die Varianz ist aufgrund der Quadrierung in einer anderen Einheit als die Messwerte und kann darum häufig nicht für konkrete Aussagen verwendet werden *(z.B. Jahre^2)*.

Die Standardabweichung σ ist die durchschnittliche Entfernung aller Werte vom arithmetischen Mittel und die Quadratwurzel der Varianz. Dadurch ist die Standardabweichung wieder in derselben Einheit wie die Messwerte und kann konkrete Angaben lieferen *(z.B. die Länge weicht durchschnittlich 5.2cm vom Mittelwert 25cm ab.)*

Beide Kennzahlen sind ***nur Vergleichswerte*** und liefern nur Informationen über mehr/weniger Streuung. Sind empfindlich auf Ausreisser.

Normale Varianz-Berechnung	*Klassifizierte Häufigkeit-Varianz*	*Standardabweichung*
$\sigma^2 = \frac{1}{n}\sum_{i=1}^{v}(x_i - \bar{x})^2$	$\sigma^2 = \frac{1}{n}\sum_{i=1}^{v}(x_i - \bar{x})^2 \cdot h_i$	$\sigma = \sqrt{\sigma^2}$

Beispiel ohne Klassen:
Urliste: $x_i = (3.2, 3.1, 3.4, 3.1, 1.9, 2.0)$

1. Arithmetisches Mittel berechnen

$$n = 6 \quad \bar{x} = 2.78$$

2. Varianz berechnen

$$\sigma^2 = \frac{1}{6} \cdot ((3.2-2.78)^2 + (3.1-2.78)^2 + (3.4-2.78)^2 + (3.1-2.78)^2 + (1.9-2.78)^2 + (2.0-2.78)^2)$$
$$= \frac{1}{6} \cdot 2.1484 = \underline{0.358}$$

3. Standardabweichung berechnen

$$\sigma = \sqrt{0.358} = \underline{0.598}$$

Beispiel mit Klassen:
Daten siehe Beispiel-Tabelle in «Klassifizierte Häufigkeitsverteilung» (Seite 10)

1. Arithmetisches Mittel berechnen *Berechnung des Werts siehe «Arithmetisches Mittel (Durchschnitt)» (Seite 13)*

$$\bar{x} = 47.023$$

2. Für jede Klasse $(\acute{x_i} - \bar{x})^2$ berechnen

$$(\acute{x_i} - \bar{x})^2 \Rightarrow \{(10-47.023)^2 = 1'370, (30-47.023)^2 = 289, (50-47.023)^2 = 8.86, (70-47.023)^2 = 528\}$$

3. Für jede Klasse $(\acute{x_i} - \bar{x})^2 \cdot h_i$ berechnen

$$(\acute{x_i} - \bar{x})^2 \cdot h_i \Rightarrow \{1'370 \cdot 8 = 10'960, 289 \cdot 40 = 11'560, 8.86 \cdot 80 = 708.8, 528 \cdot 32 = 16'896\}$$

4. h_i und $(\acute{x_i} - \bar{x})^2 \cdot h_i$ summieren und in Formel einfügen

$$\sigma^2 = \frac{1}{4} \cdot (10'960 + 11'560 + 708.8 + 16'896) = \frac{1}{4} \cdot 40124.8 = \underline{10'031.2}$$

5. Standardabweichung berechnen

$$\sigma = \sqrt{10'031.2} = \underline{100.15}$$

4.2.5. Variationskoeffizient

Misst die relative Streuung in Relation zur Lage der Häufigkeitsverteilung. Wird aus der Standardabweichung und dem arithmetischem Mittel berechnet. Dient aufgrunddessen ebenfalls nur als ***Vergleichswert***. Kann zum Vergleich unterschiedlicher Mittelwerte und Dimensionen verwendet werden.

$$\mathrm{VK} = \frac{\sigma}{\bar{x}} \cdot 100$$

σ: Standardabweichung
$\bar{x}$: arithmetisches Mittel der Merkmalswerte

Beispiel:
In welchem Laden ist die Streuung der Preise für ein Produkt geringer?
Gegeben: $\bar{x}_A = 7\ \mathrm{CHF}, \quad \sigma_A = 2.80\ \mathrm{CHF}, \quad \bar{x}_B = 750\ \mathrm{CHF}, \quad \sigma_B = 20.40\ \mathrm{CHF}$

$$\mathrm{VK}_A = \frac{2.80}{7} \cdot 100 = 40\%, \quad \mathrm{VK}_B = \frac{20.40}{750} \cdot 100 = 2.72\%$$

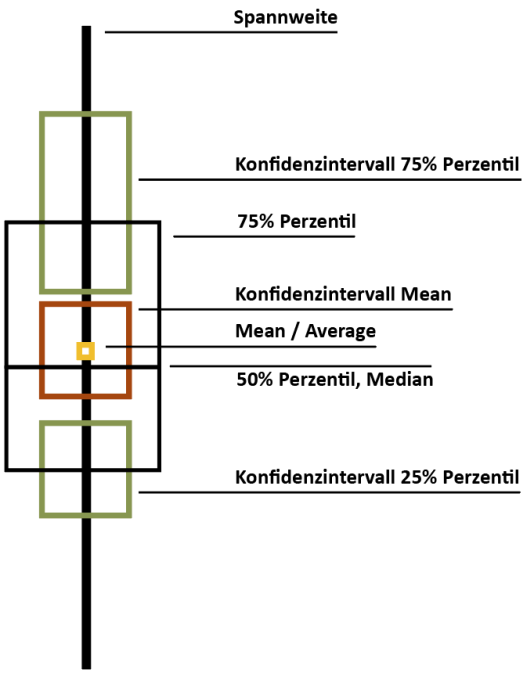
Schlussfolgerung: Die Streuung der Preise für B ist relativ geringer als für A.

4.3. ÜBERSICHT LAGEPARAMETER & STREUMASSE

Parameter	*Nominalskala*	*Ordinalskala*	*Intervallskala*	*Verhältnisskala*
Modus	✓	✓	✓	✓
Median	×	✓	✓	✓
Perzentil	×	✓	✓	✓
Arithmetisches Mittel	×	×	✓	✓
Harmonisches Mittel	×	×	×	✓
Geometrisches Mittel	×	×	×	✓
Spannweite	×	×	✓	✓
Zentraler Quartilsabstand	×	×	✓	✓
Mittlere absolute Abweichung	×	×	✓	✓
Varianz & Standardabweichung	×	×	✓	✓
Variationskoeffizient	×	×	×	✓

4.4. BOXPLOT

Fasst verschiedene ***Streuungs- und Lagemasse*** in ***einer Darstellung*** zusammen und vermittelt einen ***schnellen Eindruck***, in welchem Bereich sich die Daten befinden und wie sie sich über diesen Bereich verteilen.



5. ZEITREIHEN, REGRESSION & KORRELATION

Bei der ***Untersuchung des Zusammenhangs*** zwischen zwei Merkmalen X und Y interessieren folgende Fragen:
- Besteht ein ***Zusammenhang*** zwischen X und Y? *(Feststellung der Abhängigkeit)*
- Von welcher ***Form*** ist der Zusammenhang? *(Regressionsanalyse)*
- Von welcher ***Stärke / Intensität*** ist der Zusammenhang? *(Korrelationsanalyse)*

5.1. ZEITREIHEN

Eine Zeitreihe ist eine ***zeitlich geordnete Folge*** von Merkmalswerten. Es besteht ein Zusammenhang zwischen dem Merkmalsträger x und den diskreten Zeitpunkten t_i. Alle Datensätze, die als x-Achse die Zeit besitzen, sind Zeitreihen. Mithilfe der ***Zeitreihenanalyse*** können Stukturen und Gesetzmässigkeiten einer Zeitreihe erkannt werden.

Ein ***Trend*** beschreibt die langfristige Grundrichtung einer Zeitreihe, er erlaubt einen *(vorsichtigen)* Blick in die Zukunft. Trends können linear, exponentiell, polynomial oder logistisch sein.

5.1.1. Gleitender Mittelwert

Um zu verhindern, dass besondere Ereignisse einen grundlegenden Verlauf verschleiern, werden Werte ***geglättet***. Mit dem ***gleitenden Mittelwert*** wird aus einer bestimmten Anzahl Vergangenheitswerten sowie dem Gegenwartswert ein Mittelwert gebildet. Dieser dient als Prognose bzw. Trenderkennung für die kommende Periode.

$$M_t = \frac{1}{t} \cdot \left(\sum_{T=0}^{t-1} x_T \right) + x_t$$

t: Fenstergrösse
x_T: Vergangenheitswerte
x_t: Gegenwartswert

Beispiel zur Glättung:
Mit einer Fenstergrösse von $t = 3$ wird $y_i = (851, 863, 878, 792, 589, 851, 863)$ geglättet:

$$M_3 = \frac{1}{3} \cdot (851 + 863 + 878) = 864 \quad M_4 = \frac{1}{3} \cdot (863 + 878 + 792) = 844 \quad \ldots$$

$$y_{\text{geglättet}} = (864, 844, 753, 744, 768)$$

Beispiel zur Prognose:
Eine Firma hat in den letzten 6 Jahren diese Anzahl Maschinen verkauft. Welcher Absatz ist für 2024 zu erwarten?

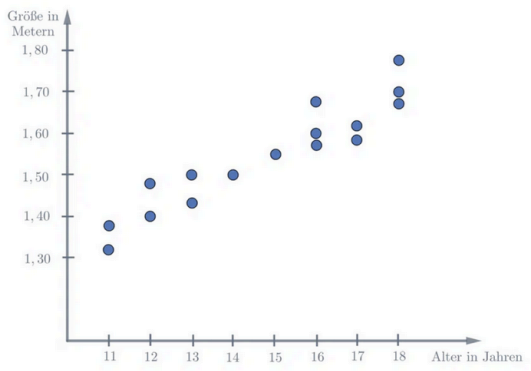
2018	2019	2020	2021	2022	2023	2024
3	7	14	8	3	1	?

$$\begin{aligned} M_7 &= \tfrac{1}{6} \cdot \textstyle\sum_{T=1}^{5} x_T + x_6 \\ &= \tfrac{1}{6} \cdot (3 + 7 + 14 + 8 + 3 + 1) = \tfrac{36}{6} = \underline{6} \end{aligned}$$

Hier zeigt sich die Schwachstelle des gleitenden Mittelwerts zur Prognose: Obgleich fallender Verkaufszahlen liegt der prognostizierte Wert darüber. Darum verwendet man in solchen Fällen besser die Regression.

5.2. REGRESSION

Beim Experimentieren entstehen ***Zahlentupel***: Ein ***Faktorwert*** x_i *(Input, «control»)* und ein ***Ergebniswert*** y_i *(Output, «response»)*. Ziel ist es, die Form/Tendenz des Zusammenhangs durch eine mathematische Funktion zu beschreiben, die ***Regressionsfunktion***. Mit dieser ist auch eine Interpolation möglich. Für Regressionen müssen die Merkmalswerte mindestens ***intervallskaliert*** sein.



Durch eine ***Regressionsanalyse*** kann der funktionale Zusammenhang zwischen einer abhängigen und einer unabhängigen Grösse anhand von Einzelmessungen modelliert werden. Werden die Wertekombinationen (x, y) der Merkmalsträger in ein Koordinatensystem eingetragen, ergibt sich ein ***Streuungsdiagramm*** (Punktewolke).

5.2.1. Regressfunktion

Je nach Abhängigkeit der Parameter voneinander wird für die Regression eine andere ***Regressfunktion*** verwendet.

Bei ***einseitiger*** Beeinflussung *(y wird nur durch Parameter x bestimmt)* gilt diese Regressfunktion:

$$\hat{y} = a_1 + b_1 \cdot x_i, \qquad a_1 = \bar{y} - b_1 \cdot \bar{x}, \qquad b_1 = \frac{\sum_{i=1}^{n} x_i \cdot y_i - n \cdot \bar{x} \cdot \bar{y}}{\sum_{i=1}^{n} x_i^2 - n \cdot \bar{x}^2}$$

Bei ***wechselseitiger*** Beeinflussung *(x und y beeinflussen sich gegenseitig)* oder unbekannter Abhängigkeit gilt:

$$\hat{x} = a_2 + b_2 \cdot y_i, \qquad a_2 = \bar{x} - b_2 \cdot \bar{y}, \qquad b_2 = \frac{\sum_{i=1}^{n} x_i \cdot y_i - n \cdot \bar{x} \cdot \bar{y}}{\sum_{i=1}^{n} y_i^2 - n \cdot \bar{y}^2}$$

Regressvariable	***Definition***
Regressionsgerade $\hat{y}$	Beschreibt den Zusammenhang zwischen dem unabhängigen Merkmal X und dem abhängigen Merkmal Y. Je stärker die Abweichung, desto stärker der Zusammenhang.
Regressionsparameter b_1	Steigungsmass, um wie viele Einheiten sich Y tendentiell ändert, wenn X um eine Einheit grösser wird. Entspricht bei der linearen Regression der Geradensteigung.
Regressionsparameter a_1	Tendenzieller Wert des Merkmals Y, wenn der Wert von $X = 0$

5.2.2. Lineare Regression

Bei der linearen Regression wird ein Modell anhand einer Geraden erstellt. Dieses kann dann mithilfe der Methode der ***kleinsten Quadrate (Mean Squared Error)*** bewertet werden.

Beispiel mit einseitiger Beeinflussung:

Gegeben sind x_i und y_i, die restlichen Tabellenwerte müssen berechnet werden.

1. y_i, $x_i \cdot y_i$ und x_i^2 ausrechnen *(siehe Tabelle)*
2. Arithmetisches Mittel für x und y berechnen.

$$n = 7, \quad \bar{x} = \frac{\sum_1^n x_i}{n} = \frac{28}{7} = 4, \quad \bar{y} = \frac{\sum_1^n y_i}{n} = \frac{60.9}{7} = 8.7$$

3. a und b berechnen

$$b = \frac{\sum_{i=1}^{n} x_i \cdot y_i - n \cdot \bar{x} \cdot \bar{y}}{\sum_{i=1}^{n} x_i^2 - n \cdot \bar{x}^2} = \frac{290.6 - 7 \cdot 4 \cdot 8.7}{140 - 7 \cdot 4^2} = 1.68$$

$$a = \bar{y} - b \cdot \bar{x} = 8.7 - 1.68 \cdot 4 = 1.98$$

4. a und b in lineare Regressionsformel eintragen

$$f(x) = a + b \cdot x = \underline{1.98 + 16.8 \cdot x}$$

x_i	y_i	$x_i \cdot y_i$	x_i^2
1	3.2	3.2	1
2	4.2	8.4	4
3	9	27	9
4	8	32	16
5	12	60	25
6	11.5	69	36
7	13	91	49
Σ 28	60.9	290.6	140

5.2.3. Exponentielle Regression

Folgt der Trend des Streuungsdiagramm einem exponentiellen Verlauf, kann ein exponentielles Modell erstellt werden. Mit Anwendung des natürlichen Logarithmus kann die exponentielle Verteilung linearisiert werden:

$$\underbrace{y_i' = \ln(y_i)}_{\text{So kommt man auf } y_i'} \qquad \underbrace{y_i(x_i) = e^{a+b\cdot x_i}}_{\text{Das ist die exponentielle Regression}} \qquad \underbrace{f(x) = \ln(y_i(x_i)) = a + b \cdot x_i}_{\text{So wird sie in lineare Regression umgewandelt}}$$

Beispiel:

1. $y_i' = \ln(y_i)$, $x_i \cdot y_i'$ und x_i^2 ausrechnen *(siehe Tabelle)*
2. Arithmetisches Mittel von x und y' ausrechnen.

$$n = 7, \quad \bar{x} = \frac{\sum_1^n x_i}{n} = \frac{28}{7} = 4, \quad \bar{y}' = \frac{\sum_1^n y_i}{n} = \frac{2.46}{7} = 0.35$$

3. a und b mit linearer Regressionsformel berechnen

$$b = \frac{\sum_{i=1}^n x_i \cdot y_i' - n \cdot \bar{x} \cdot \bar{y}'}{\sum_{i=1}^n x_i^2 - n \cdot \bar{x}^2} = \frac{27.91 - 7 \cdot 4 \cdot 0.35}{140 - 7 \cdot 4^2} = 0.65$$

$$a = \bar{y}' - b \cdot \bar{x} = 0.35 - 0.65 \cdot 4 = -2.25$$

4. a und b in die exponentielle Polynomformel einsetzen

$$f(x) = e^{a+b\cdot x} = \underline{e^{-2.25+0.65\cdot x}}$$

x_i	y_i	y_i'	$x_i \cdot y_i'$	x_i^2
1	0.20	−1.61	−1.61	1
2	0.40	−0.92	−1.84	4
3	0.80	−0.22	−0.66	9
4	1.44	0.36	1.44	16
5	2.40	0.88	4.40	25
6	5.00	1.16	9.66	36
7	10.60	2.36	16.52	49
Σ 28	20.84	2.46	27.91	140

Logarhitmische Regression

Die Logarhitmische Regression funktioniert vom Vorgehen ähnlich wie die exponentielle, hat jedoch eine andere Zielformel:

$$\underbrace{y_i' = e^{y_i}}_{\text{So kommt man auf } y_i'} \qquad \underbrace{y_i(x_i) = \ln(a + b \cdot x)}_{\text{Das ist die logarithmische Regression}} \qquad \underbrace{f(x) = e^{a+b\cdot x}}_{\text{So wird sie in lineare Regression umgewandelt}}$$

5.2.4. Polynomiale Regression

Ist kein ***linearer*** oder ***exponentieller*** Verlauf sichtbar, kann eine ***Polynomiale Regression*** durchgeführt werden. Der Grad des Polynoms kann ebenfalls festgelegt werden, allerdings wird in der Praxis selten höher als der 3. Grad berechnet. Für die polynomiale Regression ist die Berechnung des Korrelationskoeffizienten nicht sinnvoll.

$$f(x) = a_0 + a_1(x - x_1) + a_2(x - x_1)(x - x_2) + ... + a_n(x - 1)...(x - x_n)$$

Um die Werte für die polynomiale Regression zu bestimmen, wird der ***Newton-Algorhitmus*** angewendet. Dieser berechnet rekursiv die Koeffizienten a des Polynoms. In jedem Schritt werden die x und y-Werte des momentanen und des vorherigen Wertes durch die Spanne der x-Werte dividiert.

$$D_{i,i-1} = \frac{y_i - y_{i-1}}{x_i - x_{i-1}} \quad \Rightarrow \quad D_{2,1} = \frac{y_2 - y_1}{x_2 - x_1} \quad \Rightarrow \quad D_{3,2,1} = \frac{D_{2,3} - D_{2,1}}{x_3 - x_1}$$

Beispiel:

1. Alle notwendigen $D_{i,\dots,i-n}$ berechnen

x_i	y_i	$D_{i,i-1}$	$D_{i,\dots,i-2}$	$D_{i,\dots,i-3}$	$D_{i,\dots,i-4}$
1	52.5	–	–	–	–
2	34.0	$D_{2,1} = \frac{34-52.5}{2-1} = -18.5$	–	–	–
3	13.5	$D_{3,2} = \frac{13.5-34}{3-2} = -20.5$	$D_{3,2,1} = \frac{-20.5-(-18.5)}{3-1} = -1$	–	–
4	0	$D_{4,3} = \frac{0-13.5}{4-3} = -13.5$	$D_{4,3,2} = \frac{-13.5-(-20.5)}{4-2} = 3.5$	$D_{4,3,2,1} = \frac{3.5-(-1)}{4-1} = 1.5$	–
5	2.5	$D_{5,4} = \frac{2.5-0}{5-4} = 2.5$	$D_{5,4,3} = \frac{2.5-(-13.5)}{5-3} = 8$	$D_{5,4,3,2} = \frac{8-3.5}{5-2} = 1.5$	$D_{5,4,3,2,1} = 0$
6	30	$D_{6,5} = 27.5$	$D_{6,5,4} = 12.5$	$D_{6,5,4,3} = 1.5$	$D_{6,5,4,3,2} = 0$
7	91.5	$D_{7,6} = 61.5$	$D_{7,6,5} = 17$	$D_{7,6,5,4} = 1.5$	$D_{7,6,5,4,3} = 0$

($D_{i,\dots,i-5}$ und $D_{i,\dots,i-6}$ wurden aus Platzgründen weggelassen, haben aber alle den Wert 0.)

2. Die hintersten $D_{i,\dots,i-n}$ in einer Zeile als a_i einsetzen.

$$a_0 = y_1 = 52.5 \quad a_1 = D_{2,1} = -18.5 \quad a_2 = D_{3,2,1} = -1 \quad a_3 = D_{4,3,2,1} = 1.5 \quad a_4 = a_5 = a_6 = 0$$

3. Die a's in die Polynomformel einsetzen

$$f(x) = 53.5 - 18.5(x - 1) - 1(x - 1)(x - 2) + 1.5(x - 1)(x - 2)(x - 3)$$

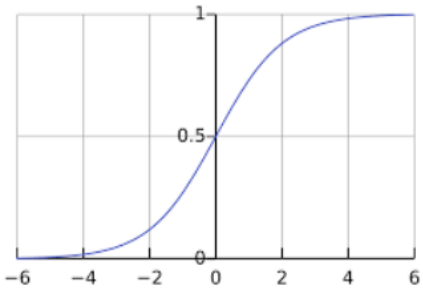
4. Das Polynom ausmultiplizieren

$$\underline{f(x) = 1.5x^3 - 10x^2 + x + 60}$$

(Die Werte sind hier ab a_4 alle 0, daher ist das Polynom vom Grad 3. Die Werte sind in diesem Fall exakt 0, da sie direkt von einem Polynom 3. Grades stammen)

5.2.5. Logistische/Qualitative Regression

Die Logistische Regression unterscheidet sich von den bisherigen Typen, da sie prüft, ob eine ***Abhängigkeit*** zwischen einer binären Variable und einer oder mehreren unabhängigen Variablen besteht. D.h. Sie ***beantwortet `true/false`-Fragen***.



Die Output-Werte liegen immer ***zwischen 0 und 1***. Dazwischen muss ein Wert für die ***Entscheidungsschwelle*** festgelegt werden, also die Grenze, bei der zwischen `true/false` gewechselt wird.

Sie wird nicht mit der kleinsten-Quadrate-Methode *(Mean Squared Error)*, sondern mit der Maximum-Likelihood-Methode oder der Sigmoid-Funktion berechnet. Ist die Grundlage für Klassifizierungs-Algorhitmen und neuronale Netze.

Logistische Regressionsformel	***Sigmoid-Funktion***
$P(Y_k = 1) = \frac{\exp\left(\beta_0 + \sum_{i=1}^{n} \beta_i \cdot X_{k \cdot i}\right)}{1 + \exp\left(\beta_0 + \sum_{i=1}^{n} \beta_i \cdot X_{k \cdot i}\right)}$	$f(x) = \frac{1}{1 + e^{-(x)}}$

5.3. KORRELATION

Eine Korrelation beschreibt eine Beziehung zwischen zwei oder mehreren Merkmalen, Zuständen oder Funktionen. ***Die Beziehung muss keine kausale Beziehung sein***: manche Elemente eines Systems beeinflussen sich gegenseitig nicht, oder es besteht eine stochastische, also vom Zufall beeinflusste Beziehung zwischen ihnen. Korrelation ist nur ein ***Indiz*** für einen kausalen Zusammenhang. Um dies festzustellen, ist die Zusammenarbeit mit Fachkundigen nötig.

Die ***Autokorrelation*** ist die Korrelation eines Merkmals mit sich selbst zu einem früheren Zeitpunkt *(anstatt mit einem anderen Merkmal)*. Damit können Vergleiche zwischen aktuellen und vergangenen Daten aufgestellt werden.

Die ***Korrelationsanalyse*** hat die Aufgabe, die Intensität des rechnerischen *(nicht des kausalen)* Zusammenhangs festzustellen und aufzuzeigen, wie gross der Einfluss des einen Merkmals auf das andere ist. Dafür werden ***Kenngrössen*** benötigt. Welches Verfahren im speziellen Fall eingesetzt werden darf, hängt von der ***Skalierung*** der Merkmale ab. Da wir aber immer davon ausgehen, dass beide Merkmale mindestens intervallskaliert sind, wird in ExEv immer der Korrelationskoeffizient von Bravais Pearson (siehe «Korrelationskoeffizient von Pearson» (Seite 24)) angewendet.

5.3.1. Zusammenhang

Zwei Merkmale sind ***statistisch unabhängig***, wenn der Wert des einen Merkmals nicht vom Wert des anderen Merkmals abhängt. Zu unterscheiden sind eine ***formale Abhängigkeit*** *(eine zahlenmässig begründete Abhängigkeit zwischen den Merkmalen)* und ***Sachliche Abhängigkeit*** *(verursacht der Wert des einen Merkmals den Wert des anderen → Kausalität)*.

5.3.2. Korrelationskoeffizient von Pearson

Der Korrelationskoeffizient von Bravais Person ***zeigt die Stärke des linearen Zusammenhangs an***. Bei einem Wert gegen $+1$ ist ein starker ***positiver/gleichläufiger*** Zusammenhang vorhanden *(wird x grösser, wird y auch grösser)*, bei Werten gegen -1 ein starker ***negativer/gegenläufiger*** Zusammenhang *(wird x grösser, wird y kleiner)*. Ist das Resultat nahe 0, besteht ***kein*** linearer Zusammenhang. Bei entgegengesetzter Steigung der Variablen ist keine Bestimmung möglich. Nur bei linearer Regression sinnvoll.

Die ***Kovarianz*** ist ein Mass für die Streuung der Merkmalsträger bzw. deren Kombinationen (x_i, y_i) um das arithmetische Mittel $\bar{x}, \bar{y}$. Sie berechnet die Varianz in Bezug auf die Abhängigkeit zweier Variablen. Die Kovarianz wird dann auf den Wertebereich $[-1, 1]$ normiert; so entsteht der ***Korrelationskoeffizient***.

Kovarianz	***Korrelationskoeffizient***
$\sigma_{XY} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x}) \cdot (y_i - \bar{y})$	$r = \frac{\sigma_{XY}}{\sigma_X \cdot \sigma_Y} = \frac{\sum_{i=1}^{n}(x_i - \bar{x}) \cdot (y_i - \bar{y})}{\sqrt{\left(\sum_{i=1}^{n}(x_i - \bar{x})^2\right) \cdot \left(\sum_{i=1}^{n}(y_i - \bar{y})^2\right)}}$

Beispiel: Werte aus 1. Aufgabe von «Lineare Regression» (Seite 20)

x_i	y_i	$x_i \cdot y_i$	x_i^2	$(x_i - \bar{x})^2$	$(y_i - \bar{y})^2$	$(x_i - \bar{x}) \cdot (y_i - \bar{y})$
1	3.2	3.2	1	9	30.25	16.5
2	4.2	8.4	4	4	20.25	9
3	9	27	9	1	0.09	-0.3
4	8	32	16	0	0.49	0
5	12	60	25	1	10.89	3.3
6	11.5	69	36	4	7.84	5.6
7	13	91	49	9	18.49	12.9
Σ 28	60.9	290.6	140	28	88.3	47

1. $(x_i - \bar{x})^2$, $(y_i - \bar{y})^2$ und $(x_i - \bar{x}) \cdot (y_i - \bar{y})$ ausrechnen (siehe Tabelle)
2. Kovarianz berechnen

$$\begin{aligned}\sigma_{xy} &= \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x}) \cdot (y_i - \bar{y}) \\ &= \frac{47}{7} = 6.71\end{aligned}$$

3. Standardabweichungen σ_x und σ_y berechnen

$$\sigma_x = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n}} = \sqrt{\frac{28}{7}} = 2$$

$$\sigma_y = \sqrt{\frac{\sum_{i=1}^{n}(y_i - \bar{y})^2}{n}} = \sqrt{\frac{88.3}{7}} = 3.55$$

4. Kovarianzkoeffizient berechnen

$$r = \frac{\sigma_{xy}}{\sigma_x \cdot \sigma_y} = \frac{6.71}{2 \cdot 3.55} = \underline{0.95}$$

$\Rightarrow$ starker positiver Zusammenhang

Achtung: *Wird das wrstat TR Skript «linreg» verwendet, werden beim Schritt drei nicht die Standardabweichungen, sondern die Varianzen angegeben (also ohne $\sqrt{\cdot}$)*

6. WAHRSCHEINLICHKEIT

Die Wahrscheinlichkeitsrechnung befasst sich mit ***Zufallsexperimenten.*** Dabei ist der Ausgang nicht (exakt) vorhersagbar. Wir erhalten unter «***gleichen*** Versuchsbedingungen» jeweils ***verschiedene*** Ergebnisse. Aufgabe der Wahrscheinlichkeitsrechnung ist es, die Wahrscheinlichkeit für den (Nicht-)Eintritt eines Ereignisses zu bestimmen.

6.1. ZUFALLSEXPERIMENTE

- ***Zufallsexperiment:*** Vorgang, der als beliebig oft wiederholbar angesehen werden kann und dessen Ergebnis vom Zufall abhängt. *(z.B. Werfen eines Würfels)*
- ***Zufallsvorgang:*** Vorgang, dessen Ausgang aufgrund von Unkenntnis oder Unwissenheit nicht vorhergesagt werden kann.
- ***Ergebnis:*** Die einzelnen, sich gegenseitig ausschliessenden möglichen Ausgänge eines Zufallsexperiments, also der tatsächlich eingetretene Fall/Messwert *(z.B. Werfen einer 5 mit einem Würfel)*
- ***Ereignis:*** Ergebnisse können zu Ereignissen zusammengefasst werden. Diese Teilmengen werden mit Grossbuchstaben gekennzeichnet. *(z.B Würfeln einer geraden Augenzahl $A = \{2, 4, 6\}$)*
- ***Elementarereignis*** $\{\omega\}$***:*** Ein Ereignis, welches nur ein Ergebnis beinhaltet. Teilmenge von Ω.
- ***Ergebnismenge*** Ω: Alle möglichen Ergebnisse eines Zufallsexperiments. Sie kann endlich *(Werfen eines Würfels $\Omega = \{1, ..., 6\}$)*, unendlich *(Lebensdauer eines Geräts $\Omega = \{\omega \in \mathbb{R} \mid \omega \geq 0\}$)* oder binär sein *(Gerät funktionstüchtig $\Omega = \{0, 1\}$)*.

6.2. WAHRSCHEINLICHKEITSRAUM

Ein ***Wahrscheinlichkeitsraum*** $(\Omega, \mathcal{A}, \mathrm{P})$ wird zur mathematischen Beschreibung von Zufallsexperimenten verwendet. Ein Wahrscheinlichkeitsraum wird mittels drei Elementen beschrieben:

- Der ***Ergebnismenge*** Ω mit allen möglichen Ergebnissen des Zufallsexperiments.
- Dem ***System der Ereignisse*** $\mathcal{A}$, eine Menge von Teilmengen von Ω.
- Dem ***Wahrscheinlichkeitsmass*** $\mathrm{P} : \mathcal{A} \to [0, 1]$, welchem jedem Ereignis eine Wahrscheinlichkeit zwischen 0 und 1 zuordnet.

6.2.1. System der Ereignisse

Alle Ereignisse eines Vorgangs zusammengefasst bilden ein ***System der Ereignisse***, bezeichnet mit kalligrafischen Grossbuchstaben $\mathcal{A}, \mathcal{B}, \mathcal{C}$.... Damit können Relationen gebildet werden *(Vereinigungen, Durchschnitt etc.)*.

Jedes System der Ereignisse enthält folgende Ereigniskombinationen:

- $A \cup B$ ***- Oder:*** Eines der beiden Ergebnisse tritt ein
- $A \cap B$ ***- Und:*** Beide Ereignisse treffen ein
- ***Komplementärereignis*** $\bar{A}/A^{\complement}$: Das Gegenereignis tritt auf *(«nicht A»)*
- ***Sicheres Ereignis:*** Die Ergebnismenge Ω *(alle Ergebnisse zusammen)* tritt immer ein
- ***Unmögliches Ereignis*** $\emptyset$***:*** Ereignis tritt nie ein, beschrieben durch die leere Menge. Immer in Ω enthalten. *(0 würfeln)*
- ***Unvereinbares/Disjunktes Ereignis:*** Ereignisvereinigung kann nie gleichzeitig eintreten: $A \cap B = \emptyset$ *(1 Würfel zeigt nie 3 und 5 gleichzeitig)*

6.2.2. Wahrscheinlichkeitsmass

Das ***Wahrscheinlichkeitsmass*** P ordnet jedem Ereignis A eine Wahrscheinlichkeit $\mathrm{P}(A)$ zwischen 0 und 1 zu. Die Summe aller Wahrscheinlichkeiten in einem Wahrscheinlichkeitsraum (also $\mathrm{P}(\Omega)$) ist immer 1.

Rechenregel	***Formel***
Gegenwahrscheinlichkeit	$\mathrm{P}(A^{\complement}) = 1 - \mathrm{P}(A)$
Unmögliches Ereignis	$\mathrm{P}(\emptyset) = 0$
Monotonieeigenschaft	$A \subset B \Rightarrow \mathrm{P}(A) \leq \mathrm{P}(B)$ *(wenn A in B enthalten, muss B mind. gleich gross wie A sein)*
Additionssatz *(mit Überlappung)*	$\mathrm{P}(A \cup B) = \mathrm{P}(A) + \mathrm{P}(B) - \mathrm{P}(A \cap B)$ *(Schnittmenge entfernen, sonst doppelt)*
Additionssatz *(ohne Überlappung)*	$\mathrm{P}(A \cup B) = \mathrm{P}(A) + \mathrm{P}(B)$
Multiplikationssatz *(ohne Bedingung)*	$\mathrm{P}(A \cap B) = \mathrm{P}(A) \cdot \mathrm{P}(B)$
Multiplikationssatz *(mit Bedingung)*	$\mathrm{P}(A \cap B) = \mathrm{P}(A) \cdot \mathrm{P}(B\|A) = \mathrm{P}(B) \cdot \mathrm{P}(A\|B)$

6.3. LAPLACE-EXPERIMENT

Ein Laplace-Experiment ist eine spezielle Form eines Zufallsexperiments, bei dem es nur ***Elementarereignisse*** $\{\omega\}$ gibt und diese alle die ***gleiche Wahrscheinlichkeit*** haben. Dazu muss die Ergebnismenge Ω ***endlich*** sein. Laplace-Experimente dürfen aber auch Ereignisse haben, die aus mehreren Elementarereignissen bestehen *(siehe Beispiel unten)*.

Wahrscheinlichkeit eines Elementarereignisses	***Wahrscheinlichkeit eines beliebigen Ereignis***
$\mathrm{P}(\{\omega_i\}) = \frac{1}{n}$	$\mathrm{P}(A) = \frac{\lvert A\rvert}{\lvert\Omega\rvert} = \frac{\text{Anz. Ereignisse für A}}{\text{Anz. aller Elementarereignisse}}$

Beispiel:
Das Werfen eines Würfels ist ein Laplace-Experiment, da es nur Elementarereignisse gibt: $\Omega = \{1,2,3,4,5,6\}$
- Wahrscheinlichkeit, eine 2 zu würfeln *(Elementarereignis)*: $\mathrm{P}(\{\omega_i\}) = 1/n = \underline{1/6}$
- Wahrscheinlichkeit für das Ereignis «gerade Zahl»: $A = \{2,4,6\}, \quad \mathrm{P}(A) = \lvert A\rvert/\lvert\Omega\rvert = 3/6 = \underline{1/2}$

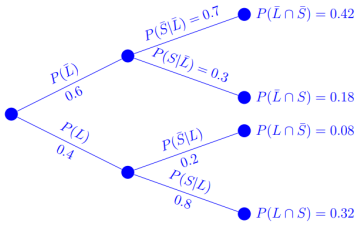
6.4. BEDINGTE WAHRSCHEINLICHKEIT

Will man die Wahrscheinlichkeit eines Ereignis berechnen, das von einem anderen abhängig ist, wendet man die bedingte Wahrscheinlichkeit an. Als Faustregel kann auch gesagt werden ***«Wahrscheinlichkeit von A, wenn B»***.

$$\mathrm{P}(\text{was wir wissen wollen}|\text{was wir wissen}) = \mathrm{P}(A|B) = \frac{\mathrm{P}(A \cap B)}{\mathrm{P}(B)}$$

Sie tritt vor allem bei mehrstufigen Experimenten auf, wenn nach einer Stufe jeweils andere Ereignisse eintreffen können *(z.B. mehrere Münzwürfe hintereinander stellen einen Kopf/Zahl Binary Tree auf)*. Aus der Formel oben schliesst sich die ***bedingte Pfadregel***:

$$\mathrm{P}(A \cap B) = \mathrm{P}(B) \cdot \mathrm{P}(A|B)$$
$$\mathrm{P}(B) \to \mathrm{P}(A|B) \to \mathrm{P}(A \cap B)$$



Aus dieser lässt sich der ***Satz von Bayes*** schliessen. Damit kann die «Bedingung» umgekehrt werden: Wissen wir $\mathrm{P}(A|B)$, können wir auch $\mathrm{P}(B|A)$ ausrechnen.

$$\mathrm{P}(A|B) \cdot \mathrm{P}(B) = \mathrm{P}(A \cap B) = \mathrm{P}(B|A) \cdot \mathrm{P}(A) \quad \Rightarrow \quad \mathrm{P}(A|B) = \mathrm{P}(B|A) \cdot \frac{\mathrm{P}(A)}{\mathrm{P}(B)}$$

Beispiel

	Erkrankt (B)	*Nicht erkrankt $\bar{B}$*	*Summe Σ*
Geimpft (A)	117	389	506
Nicht Geimpft ($\bar{A}$)	289	165	454
Summe Σ	406	554	960

Wahrscheinlichkeit, an Grippe zu erkranken:

$$\mathrm{P}(B) = \frac{|B|}{|\Omega|} = \frac{406}{960} = 42.3\%$$

Wahrscheinlichkeit, dass Person nicht geimpft ist:

$$\mathrm{P}\left(\bar{A}\right) = \frac{\left|\bar{A}\right|}{|\Omega|} = \frac{454}{960} = 47.3\%$$

Wahrscheinlichkeit, trotz Impfung zu erkranken *(Laplace)*:

$$\mathrm{P}(B|A) = \frac{\mathrm{P}(A \cap B)}{\mathrm{P}(A)} = \frac{|A \cap B|/\Omega}{|A|/|\Omega|} = \frac{|A \cap B|}{|A|} = \frac{117}{506} = 32.1\%$$

6.4.1. Unabhängige Ereignisse

Wird das Ereignis A nicht vom Eintreten des Ereignisses B beeinflusst, gilt $\mathrm{P}(A|B) = \mathrm{P}(A)$. Die Ereignisse sind also voneinander ***stochastisch unabhängig***. Bei diesen gelten auch die jeweiligen Komplemente A^{C}, B^{C} und $A^{\mathrm{C}} \cap B^{\mathrm{C}}$ als unabhängig. Die ***unabhängige Pfadregel*** lautet also

$$\mathrm{P}(A \cap B) = \mathrm{P}(B) \cdot \mathrm{P}(A)$$

Durch ***Umformen*** dieser Formel kann man die Zahlenwerte für die Ereignisse berechnen. Dies ist nützlich, um herauszufinden, ob Variablen ***voneinander unabhängig*** sind oder nicht. Weichen die Werte erheblich von den ursprünglichen Wahrscheinlichkeitswerten ab, sind die Variablen voneinander abhängig.

$$|X \cap Y| = \frac{|X| \cdot |Y|}{|\Omega|}$$

Sind alle Ereignisse innerhalb einer Menge voneinander unabhängig, spricht man von ***stochastischer/vollständiger Unabhängigkeit***. Dann kann bei einer fehlender Wahrscheinlichkeit diese aus den Schnittmengen der anderen mithilfe des ***Multiplikationssatzes*** berechnet werden. Gesucht wird $\mathrm{P}(A)$, gegeben sind $\mathrm{P}(B)$ und $\mathrm{P}(C)$.

$$\mathrm{P}(A) = \mathrm{P}(B \cap A) + \mathrm{P}(C \cap A) \quad \Rightarrow \quad \mathrm{P}(B) \cdot \mathrm{P}(A|B) + \mathrm{P}(C) \cdot P(A|C)$$

Der Begriff «unabhängig» wird manchmal verwechselt mit dem Begriff «disjunkt». Zwei disjunkte Ereignisse A und B, also mit $\mathrm{P}(A \cdot B) = \emptyset$, können aber nur dann unabhängig sein, wenn eines der beiden Ereignisse die Wahrscheinlichkeit 0 hat. Nur dann ist $\mathrm{P}(A) \cdot \mathrm{P}(B) = 0 = \mathrm{P}(\emptyset) = \mathrm{P}(A \cdot B)$.

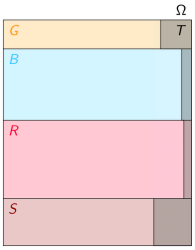
6.4.2. Totale/Vollständige Wahrscheinlichkeit

Hat man für ein Ereignis A mehrere Bedingungen B_i *(z.B. mehrere Fälle oder Ursachen für A)*, kann man die Wahrscheinlichkeit für A berechnen, wenn man die bedingten Wahrscheinlichkeiten für B_i zusammenzählt *(Aus Einzelfällen lässt sich die Gesamtsituation zusammenstellen)*.

$$\mathrm{P}(A) = \sum_{i=1}^{n} \mathrm{P}(A|B_i) \cdot \mathrm{P}(B_i)$$

Beispiel:
Wie gross ist die Wahrscheinlichkeit $\mathrm{P}(T)$, auf der Enterprise umzukommen?

Wahrscheinlichkeitsfallunterscheidung: Aus den bedingten Wahrscheinlichkeiten die Bedingung herausrechnen, um die Totale Wahrscheinlichkeit zu erhalten.



$$\begin{aligned} \mathrm{P}(T \cap G) &= \mathrm{P}(T|G) \cdot \mathrm{P}(G) \\ +\mathrm{P}(T \cap B) &= \mathrm{P}(T|B) \cdot \mathrm{P}(B) \\ +\mathrm{P}(T \cap R) &= \mathrm{P}(T|R) \cdot \mathrm{P}(R) \\ \hline \mathrm{P}(T) & \end{aligned}$$

Beispiel:
Aus einem Jasskartenset (36 Karten) werden 2 Karten gezogen. Wie gross ist die Wahrscheinlichkeit, beim zweiten Zug ein Ass zu ziehen?

$$\begin{aligned} \mathrm{P}(A) &= \mathrm{P}(A|B_1) \cdot P(B_1) + \mathrm{P}(A|B_2) \cdot \mathrm{P}(B_2) \\ &= \frac{4}{35} \cdot \frac{32}{36} + \frac{3}{35} \cdot \frac{4}{36} = \underline{\frac{1}{9}} \end{aligned}$$

$\mathrm{P}(A|B_i)$: Ass im zweiten Zug (nur noch 35 Karten)
$\mathrm{P}(B_1)$: Kein Ass im ersten Zug $= 32/36$
$\mathrm{P}(B_2)$: Ass im ersten Zug $= 4/36$

Beispiel:
In einem Behälter liegen 3 Arten von Batterien im Verhältnis 20:30:50. B_1 hält zu 70% länger als 100 Stunden, B_2 zu 40% und B_3 zu 30%. Wie hoch ist die Wahrscheinlichkeit, dass eine zufällig gewählte Batterie länger als 100 Stunden hält?

$$\begin{aligned} \mathrm{P}(A) &= \sum_{i=1}^{3} \mathrm{P}(A|B_i) \cdot \mathrm{P}(B_i) \\ &= 0.7 \cdot 0.2 + 0.4 \cdot 0.3 + 0.3 \cdot 0.5 = \underline{0.41} \end{aligned}$$

n: Anzahl Arten von Batterien
$\mathrm{P}(A)$: Gewählte Batterie hält länger als 100h
$\mathrm{P}(A|B_i)$: Wahrsch., dass B_i länger als 100h hält
$\mathrm{P}(B_i)$: Wahrscheinlichkeit, dass B_i gewählt wird

7. KOMBINATORIK

Die Kombinatorik liefert mathematische Modelle zur ***Bestimmung der Anzahl möglicher Anordnungen*** von Elementen. Dabei kann mit oder ohne Berücksichtigung der Elementreihenfolge und -wiederholung gearbeitet werden.

- ***Menge:*** Liste, die jedes Element nur einmal enthält, die Reihenfolge der Elemente ist irrelevant.
- ***Tupel:*** Liste, die Elemente mehrfach enthalten kann. Die Reihenfolge der Elemente ist relevant.
- ***n-Tupel:*** Tupel mit n Elementen. Hat an der n-ten Stelle jeweils k_n mögliche Permutationen. Es gibt für die Besetzung jeweils $k_1 \cdot k_2 \cdot ... \cdot k_n$ verschiedene n-Tupel.

Man unterscheidet zwei Arten der Auswahlverfahren einer Reihenfolge:
- ***Mit Wiederholung/Zurücklegen:*** Die Elemente sind nicht einzigartig und dürfen sich in der Anordnung wiederholen *(z.B. Zahlen in Zahlenschloss, Bälle mit verschiedenen Farben)*
- ***Ohne Wiederholung/Zurücklegen:*** Die Elemente sind alle einzigartig und dürfen nicht mehrmals vorkommen. *(z.B. Personen an Tisch, Lotto-Zahlen)*

Es gibt drei verschiedene Techniken, um Elemente anzuordnen:

7.1. PERMUTATION

«Auf wieviele Arten lassen sich n verschiedene Objekte anordnen?»

Bei der Permutation wird die ganze Menge angeordnet, jedes Element der Menge wird also genau einmal in die Anordnung gelegt.

ohne Wiederholung	***mit Wiederholung***
$$P(n) = n!$$	$$P_{n_1,\dots,n_k}(n) = \frac{n!}{n_1! \cdot n_2! \cdot \dots \cdot n_k!}$$
Für das erste Objekt stehen n Plätze zur Verfügung. Für das zweite Objekt muss einer der $n-1$ verbleibenden Plätze gewählt werden. Bisher sind nun $n \cdot (n-1)$ Möglichkeiten gefunden. Führt man diese Reihenfolge fort, ergeben sich $n!$ Möglichkeiten.	Sind von den Elementen mindestens zwei identisch, werden diese zu Klassen zusammengefasst. Die n zu füllenden Plätze werden durch das Produkt der Fakultäten der Häufigkeiten aller Klassen n_i geteilt.

Beispiel:

Eine Maschine muss vier Aufträge nacheinander abarbeiten. Wie viele Anordnungen sind möglich?

$$\mathrm{P}(n) = 4! = \underline{24}$$

Beispiel:

a) *Ein Zahlenschloss hat eine Kombination mit den Ziffern «$1, 1, 4, 4, 4, 8$». Wie viele Codes gibt es?*

$$n_1(1) = 2,\ n_2(4) = 3,\ n_3(8) = 1, \quad \Rightarrow \quad n = 2 + 3 + 1 = 6$$

$$\mathrm{P}_{2,3,1} = \frac{6!}{2! \cdot 3! \cdot 1!} = \frac{720}{12} = \underline{60}$$

b) *Wie viele Kombinationen beginnen mit 4?*
Da der erste Platz nun fix mit einer 4 besetzt ist, haben wir eine neue Fragestellung mit den Ziffern «$1, 1, 4, 4, 8$»:

$$n_1(1) = 2,\ n_2(4) = 2, n_3(8) = 1 \quad \Rightarrow \quad n = 2 + 2 + 1 = 5$$

$$\mathrm{P}_{2,2,1} = \frac{5!}{2! \cdot 2! \cdot 1!} = \frac{120}{4} = \underline{30}$$

7.2. KOMBINATION

«Auf wie viele Arten kann man k Objekte aus n auswählen?»

Es wird eine Teilmenge aller Elemente angeordnet. Die Reihenfolge der Elemente ist hier nicht relevant, man spricht auch von einer ***ungeordneten (Stich-)probe***. Die Permutation ist ein Spezialfall der Kombination, bei welcher $n = k$.

ohne Wiederholung	***mit Wiederholung***
$$K_k(n) = \frac{\prod_n^{n-k+1} n}{k!} = \frac{n!}{k! \cdot (n-k)!} = \binom{n}{k}$$	$$K_k^W(n) = \frac{(n+k-1)!}{k! \cdot (n-k)!} = \binom{n+k-1}{k}$$
Es ist zuerst k mal eine Auswahl zu treffen. Für die erste Auswahl stehen n Objekte zur Verfügung. Danach muss noch $k-1$ mal eine Auswahl getroffen werden, es stehen noch $n-1$ Alternativen zur Verfügung. So lassen sich insgesamt $n \cdot (n-1) \cdot (n-2) \dots (n-k+1)$ Möglichkeiten finden.	Die Wiederholung addiert noch $k-1$ zum oberen Term der Matrix, da wir die Elemente wieder zur Verfügung haben. Die Matrixschreibweise entspricht dem Binomialkoeffizienten und kann auch als C_k^n geschrieben werden.

Es gibt mehrere Schreibweisen für die Kombinationsformeln. Für das Lösen der Aufgaben kannst du diejenige verwenden, die dir am besten liegt. Die Beispiele zeigen absichtlich alle Schreibweisen an.

TR: Menü → 5: Wahrscheinlichkeit → 3: Kombinationen → $\mathrm{nCr}(\text{obere Zahl}, \text{untere Zahl})$

Beispiel:
Bei einem Pokémon-Turnier soll jeder der 25 Teilnehmer einmal gegen jeden spielen. Wie viele Spiele werden ausgetragen?

$$K_2(25) = \frac{25 \cdot 24}{2!} = \frac{25!}{2! \cdot (25-2)!} = \binom{25}{2} = \underline{300}$$

Beispiel:
Im Nationalrat werden 3 Sitze neu vergeben. Es stellen sich 6 Parteien dafür auf. Eine Partei kann mehr als einen Sitz erhalten. Die Reihenfolge ist irrelevant, da es keinen Unterschied macht, ob ABC oder CBA *(egal ob Partei A 1. oder 3. wird, sie erhalten auf beide Arten 1 Sitz)*. Wie viele Kombinationen sind möglich?

$$K_3^W(6) = \binom{6+3-1}{3} = \binom{8}{3} = \frac{8 \cdot 7 \cdot 6}{3!} = \frac{8!}{3! \cdot (8-3)!} = \underline{56}$$

Beispiel:
Aus 50 Glühbirnen wird eine Stichprobe von 5 entnommen. Wie gross ist die Wahrscheinlichkeit, dass aus diesen 5 genau 2 defekt sind, wenn total 10 defekt sind?

Das gewünschte Ergebnis kann in einem Laplace-Experiment mit zwei Fällen dargestellt werden:
- 2 defekte Glühbirnen aus total 10 defekten ziehen
- $(5-2) = 8$ funktionsfähige Glühbirnen aus total $(50-10) = 40$ Funktionsfähigen ziehen

$$\mathrm{P}\{3 \text{ Defekte}\} = \frac{\binom{10}{2} \cdot \binom{50-10}{5-2}}{\binom{50}{5}} = \frac{(10 \cdot 9)/2 \cdot (40 \cdot 39 \cdot 38)/(2 \cdot 3)}{2'118'760} = \underline{0.21}$$

7.3. VARIATION

«Auf wie viele Arten kann man k mal unter n verschiedenen Objekten auswählen?»
Es wird eine Teilmenge, bei welcher die Reihenfolge relevant ist, angeordnet. Diese Teilmenge wird als ***geordnete (Stich-)probe*** bezeichnet.

ohne Wiederholung	*mit Wiederholung*
$V_k(n) = \frac{n!}{(n-k)!} = \prod_{n}^{n-k+1} n$	$V_k^W(n) = n^k$

Beispiel:
Auf wie viele Arten kann man eine Perlenkette der Länge $k = 10$ aus $n = 4$ Farben von Perlen herstellen?

$$V_{10}(4) = 4^{10} = \underline{1'048'576 \text{ Möglichkeiten}}$$

Beispiel:
Wie viele Kombinationen gibt es bei einem 5-stelligen Zahlenschloss mit jeweils 10 Ziffern pro Stelle?

$$10^5 = \underline{100'000 \text{ Kombinationen}}$$

Beispiel:
Aus 18 Teilnehmer eines Rennens müssen die ersten 3 in der richtigen Reihenfolge getippt werden. Wie viele Möglichkeiten gibt es?

$$k = 3, \qquad n = 18$$
$$V_3^W(18) = \prod_{18}^{18-3+1} n = \prod_{18}^{16} n = 18 \cdot 17 \cdot 16 = \underline{4'896}$$

7.4. BESTIMMUNG DER KOMBINATORIK-FORMEL

Wird Frage mit «Ja» beantwortet: ✓ folgen, wird Frage mit «Nein» beantwortet: × folgen

1. **Ist jedes vorgegebene Element genau einmal anzuordnen?**
   (Ganze Menge verwenden, keine Stichprobe)
   - ✓ Schritt 2
   - × Schritt 3
2. **Sind die vorgegebenen Elemente alle verschieden?**
   - ✓ Permutation ohne Wiederholung
   - × Permutation mit Wiederholung
3. **Darf ein vorgegebenes Element wiederholt ausgewählt werden?**
   (Ist Grundelement nicht nur einmal vorhanden?)
   - ✓ Schritt 5
   - × Schritt 4
4. **Ist die Anordnung der Elemente von Bedeutung?**
   - ✓ Variation ohne Wiederholung
   - × Kombination ohne Wiederholung
5. **Ist die Anordnung der Elemente von Bedeutung?**
   - ✓ Variation mit Wiederholung
   - × Kombination mit Wiederholung

8. DISKRETE VERTEILUNGEN

Mit Hilfe einer Wahrscheinlichkeitsverteilung lassen sich zufallsbehaftete Ereignisse oder Variablen *(sogenannte Zufallsvariablen)* modellieren. ***Diskrete Verteilungen*** stellen die Ergebnisse von Experimenten dar, welche eine feste Anzahl von Ereignissen haben *(z.B. 6 Ereignisse eines Würfels, Anzahl Studenten mit einer bestimmten Note)*

8.1. ZUFALLSVARIABLE

Eine Zufallsvariable $X(\omega)$ ist eine Funktion, welche jedem möglichen Ergebnis ω eines Zufallsexperiments ***eine reelle Zahl x zuordnet***. Dabei steht in der Fallunterscheidung links die Realisierung der Zufallsvariable *(konkreter Wert)* und rechts das Ereignis *($\omega \in \{...\}$)*.

Beispiel: Beim Roulette auf erstes Dutzend $(1-12)$ setzen. Ergebnisse ω: 2 Fr. Gewinn, 1 Fr. Verlust (des Einsatzes)

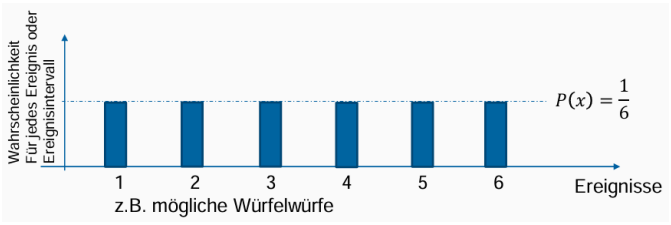
$$X : \{0, 1, ..., 36\} \to \mathbb{R}, \quad X(\omega) = \begin{cases} 2 & \text{für } \omega \in \{1, 2, ..., 12\} \\ -1 & \text{für } \omega \in \{0, 13, 14, ..., 36\} \end{cases}$$

8.2. DISKRETE WAHRSCHEINLICHKEITSFUNKTION

Eine ***Wahrscheinlichkeitsfunktion $f(x)$*** weist den Werten einer Zufallsvariable die Wahrscheinlichkeit ihres Auftretens p zu und bildet damit ein Tupel $[x_i, p_i]$. Dieses wird häufig als ***Graph*** dargestellt. Jede Verteilung hat ihre eigene Wahrscheinlichkeitsfunktion. Eine Wahrscheinlichkeitsfunktion wird als Fallunterscheidung dargestellt mit einem sonst-Fall mit Wahrscheinlichkeit 0 *(default switch case)*.

$$f(x) = \begin{cases} P(X = x_i) = p_i, & x = x_i \in \{x_1, ..., x_k\} \\ 0, & \text{sonst.} \end{cases}$$

Die ***diskrete Wahrscheinlichkeitsfunktion*** $f(x_i) = \mathrm{P}(X = x)$ wird als ***Balkendiagramm*** dargestellt. Die Wahrscheinlichkeit $\mathrm{P}(x_i)$ eines Ereignisses kann unmittelbar an ihrem Balken abgelesen werden und die Summe aller Wahrscheinlichkeiten ist 1.



- ***Verteilung der Zufallsvariable:*** Die Gesamtheit der Tupel der Wahrscheinlichkeitsfunktion. Wird normalerweise als Graph dargestellt.
- ***Realisation:*** Wert, den die Zufallsvariable in einem konkreten Experimentdurchlauf annimmt.
- ***Erwartungswert $E(x)$:*** Wert, der die Zufallsvariable im Mittelwert annimmt. Die Wahrscheinlichkeit sollte um diesen Wert am grössten sein. Pro Verteilungstyp unterschiedlich.
- ***Erfolgswahrscheinlichkeit p:*** Wahrscheinlichkeit, dass ein Ereignis eintritt
- ***Misserfolgswahrscheinlichkeit q:*** Wahrscheinlichkeit, dass ein Ereignis nicht eintritt *(häufig $1-p$)*

Die Konzepte der Zufallsvariable lassen sich auf diese der ***beschreibenden Statistik*** mappen:

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

Beispiel:
Bestimme die Wahrscheinlichkeiten der Zahlen der Urliste

$$x_i = \{1,4,2,2,4,3,1,2,2,2,3,4,5,1,1,2,3,4,5,5\}, \quad n = 20$$

i	x_i	h_i	f_i	F_i	$x_i f_i$	$x_i^2 f_i$
1	1	4	0.20	0.20	0.20	0.20
2	2	6	0.30	0.50	0.60	1.20
3	3	3	0.15	0.65	0.45	1.35
4	4	4	0.20	0.85	0.80	3.20
5	5	3	0.15	1.00	0.75	3.75
Σ		20			2.80	9.70

1. Tabelle mit abs. & relat. Häufigkeiten h_i, f_i erstellen
2. Wahrscheinlichkeitsfunktion mit f_i erstellen

$$f(x) = \begin{cases} 0.2 & x = 1 \\ 0.3 & x = 2 \\ 0.15 & x = 3 \\ 0.2 & x = 4 \\ 0.15 & x = 5 \\ 0 & \text{sonst.} \end{cases}$$

8.3. DISKRETE VERTEILUNGSFUNKTION

Eine ***Verteilungsfunktion*** $F(x) = \mathbf{P}(X \leq x)$ stellt die Summe aller Wahrscheinlichkeiten dar, bei denen sich die Realisierungen unterhalb eines Wertes x befinden. Wird auch als ***summierte/kumulierte Wahrscheinlichkeit*** bezeichnet.

Viele Verteilungsfunktionen haben eigene ***Verteilungsparameter***, welche die genaue Grösse der Verteilung bestimmen. Während ein ***Skalenparameter*** die Streuung und somit die Breite einer Verteilung bestimmt, beeinflusst ein ***Formparameter*** die Form einer Verteilungsfunktion; er bewirkt mehr als nur eine Skalierung oder Verschiebung.

Beispiel:
Bestimme die Verteilungsfunktion der Urliste vom Beispiel der Diskreten Wahrscheinlichkeitsfunktion

1. Tabelle mit relativen kumulierten Häufigkeit F_i erstellen *(siehe oben)*
2. Verteilungsfunktion mit F_i erstellen. ***Die Fälle ausserhalb des Wertebereichs nicht vergessen!***

$$F(x) = \begin{cases} 0 & x < 1 \\ 0.2 & x = 1 \\ 0.5 & x = 2 \\ 0.65 & x = 3 \\ 0.85 & x = 4 \\ 1 & x \geq 5 \end{cases}$$

8.3.1. Empirische Verteilungsfunktion

Die empirische Verteilungsfunktion zeigt die Wahrscheinlichkeit an, dass ein Messwert ***höchstens eine bestimmte Grösse*** hat. Sie wird angewendet, wenn konkrete Messwerte vorliegen, also beispielsweise berechnet sie, in welchem Anteil bei 20 Würfen mit 2 Würfeln höchstens eine 5 gefallen ist.

$$F(x_i) = \sum_{j=1}^{i} f(x_j) = \sum_{j=1}^{i} = \frac{n_j}{n}$$

$F(x_i)$: Verteilungswert des Messwerts x_i
$\sum_{i=1}^{i}$: Summiere für alle Werte bis zum Messwert x_i
$f(x_j)$: Relative Häufigkeit des Messwerts j
n_j: Absolute Häufigkeit des Messwerts j
n: Gesamtanzahl der Messwerte der Stichprobe

Erwartungswert (Mittelwert)	*Varianz*
$E(X) = \mu = \sum_i x_i \cdot f(x_i) = \sum_i x_i \cdot \frac{n_j}{n}$	$\begin{aligned}\mathrm{var}(X) &= \sum_{i=1}^{v} (x_i - \bar{x})^2 \cdot f(x_i) \\ &= E(x^2) - (E(x))^2\end{aligned}$

Beispiel:
Bestimme Erwartungswert & Varianz der Urliste vom Beispiel der Diskreten Wahrscheinlichkeitsfunktion «Diskrete Wahrscheinlichkeitsfunktion» (Seite 31).

1. Tabelle mit relativer Häufigkeit f_i und Erwartungswert $x_i \cdot n_j/n$ erstellen *(siehe Beispiel oben)*
2. Summe bilden, um Gesamt-Erwartungswert zu erhalten

$$E(x) = \sum_i x_i \cdot f(x_i) = \underline{2.8}$$

3. Tabelle um Spalte $x_i^2 \cdot f(x_i)$ erweitern und Summe bilden
4. Varianz berechnen mit der Summe der neuen Spalte

$$\mathrm{var}(x) = E(x^2) - (E(x))^2 = \sum_{i=1}^{v} x_i^2 \cdot f(x_i) - (E(x))^2 = 9.7 - 2.8^2 = \underline{1.86}$$

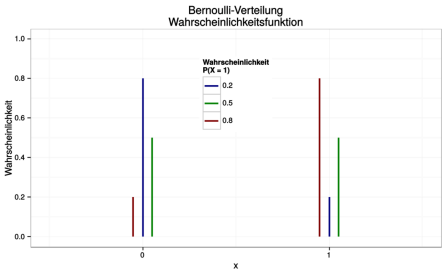
Beispiel:
An der letzten Ex&Ev-Prüfung hatten die 20 Studierenden folgenden Notenspiegel. Wie wahrscheinlich ist es, dass eine Person eine 4 oder besser erreicht hat?

Note	6	5	4	3	2	1
Häufigkeit	4	5	7	2	1	1

$$\mathrm{P}(x_i \leq 4 \leq x_{20}) = \sum_{j=i}^{20} f(x_j) = \sum_{j=i}^{20} \frac{n_j}{20} = \frac{4}{20} + \frac{5}{20} + \frac{7}{20} = \frac{16}{20} = 0.8 = \underline{80\%}$$

8.4. BERNOULLI-VERTEILUNG

Die ***Bernoulli-Verteilung*** $\mathrm{Ber}(p)$ hat nur 0 und 1 als mögliche Ereignisse, eignet sich also um festzustellen, ob ein Ereignis eintritt oder nicht. Ist von Wiederholungen unabhängig, die (Miss-)Erfolgswahrscheinlichkeiten p und $q = 1 - p$ bleiben also gleich.



Die Wahrscheinlichkeitsfunktion der Bernoulli-Verteilung ist:

$$f(x) = P(X = x) = \begin{cases} 1 - p & , \text{falls } x = 0 \\ p & , \text{falls } x = 1 \\ 0 & , \text{sonst} \end{cases}$$

Erwartungswert	***Varianz***
$E(X) = p$	$\mathrm{var}(X) = p \cdot q$

Beispiel:
In einem Experiment mit Bernoulli-Verteilung auf $\{0, 1\}$ ist das Ereignis 1 viermal so wahrscheinlich wie das Ereignis 0. Bestimme Wahrscheinlichkeitsfunktion, Verteilungsfunktion, Erwartungswert & Varianz.

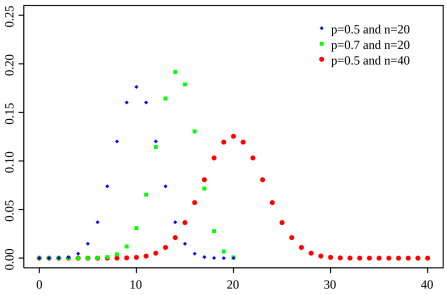
Die Summe der Wahrscheinlichkeit ist 1, also folgt: $1 = p + (1 - p) \quad \Rightarrow \quad p + 4p = 5p \rightarrow p = 0.2$

Daraus lassen sich alle Werte bilden:

Wahrscheinlichk.funktion	***Verteilungsfunktion***	***Erwartungswert***	***Varianz***
$f(x) = \begin{cases} 0.2 & x = 0 \\ 0.8 & x = 1 \\ 0 & \text{sonst} \end{cases}$	$F(X) = \begin{cases} 0 & x < 0 \\ 0.2 & x = 0 \\ 1 & x \geq 1 \end{cases}$	$E(x) = p = 0.2$	$\sigma^2 = p \cdot (1 - p) = 0.16$

8.5. BINOMIAL-VERTEILUNG

Die ***Binomial-Verteilung*** $\mathrm{Bin}(n,p)$ hat ebenfalls nur 0 und 1 als mögliche Ergebnisse. Sie zeigt aber die Wahrscheinlichkeit auf, ob ein Ereignis wahrscheinlich/genau/höchstens/mindestens x-mal auftritt – also die ***Anzahl Erfolge in einer Serie von zufälligen Ereignissen***. Die Bernoulli-Verteilung ist ein Spezialfall der Binomial-Verteilung, bei welcher $n = 1$.



Die Wahrscheinlichkeitsfunktion der Binomialverteilung ist:

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

x: Anzahl Ereignisse
n: Anzahl Realisationen, in denen Ereignis eintritt
p: Wahrscheinlichkeit, dass Ereignis eintritt

Erwartungswert	***Varianz***
$E(X) = n \cdot p$	$\mathrm{var}(X) = n \cdot p \cdot q = n \cdot p \cdot (1-p)$

Beispiel:
Auf eine Reise sind 10 Personen angemeldet. Die Wahrscheinlichkeit einer Absage ist 5%. Wie hoch ist die Chance, dass genau 2 Personen absagen?

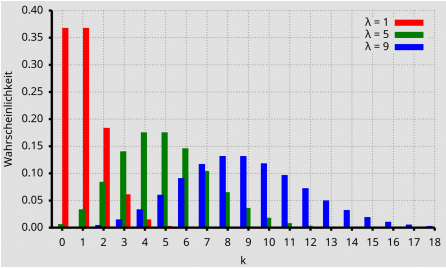
$$\mathrm{P}(X = 2) = \binom{n}{k} \cdot p^k \cdot (1-p)^{n-k} = \binom{10}{2} \cdot 0.05^2 \cdot (1-0.05)^{10-2} = \binom{10}{2} \cdot 0.05^2 \cdot 0.95^8 = \underline{0.075}$$

Wie hoch ist die Wahrscheinlichkeit, dass mindestens 3 Gäste absagen? Dazu muss 1 minus die Wahrscheinlichkeit von 0 bis 2 Absagen gerechnet werden.

$$\mathrm{P} = 1 - \mathrm{P}(X \leq x = 2) = 1 - \sum_{i=0}^{x} \binom{n}{x_i} \cdot p^{x_i} \cdot (1-p)^{n-x_i} == 1 - \sum_{i=0}^{2} \binom{10}{i} \cdot 0.05^i \cdot 0.95^{10-i} = \underline{0.011}$$

8.6. POISSON-VERTEILUNG

Die ***Poisson-Verteilung*** $\mathrm{Poi}(\mu)$ zeigt, wie hoch die Wahrscheinlichkeit ist, dass ein Ereignis in einem Intervall genau oder höchstens x-mal eintritt, wenn bekannt ist, dass in diesem Intervall das Ereignis durchschnittlich μ-mal auftritt. Durch sie können Wahrscheinlichkeiten ***seltener Ereignisse*** bestimmt werden.



Die Wahrscheinlichkeitsfunktion der Poisson-Verteilung ist:

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

x oder λ: Anzahl Ereignisse
μ: Durchschnittliches Auftreten des Ereignis im Intervall

TR Skript: $\mathrm{p\_atleast\_k}$ *beinhaltet Grenze* k ***nicht***

Erwartungswert	***Varianz***
$E(X) = \mu$	$\mathrm{var}(X) = \mu$

Beispiel:
An einer Hotline rufen in einer Stunde durchschnittlich 5 Kunden an. Sie kann 9 Anfrufe gleichzeitig bearbeiten.
(Nicht diesselbe Aufgabe wie bei der Exponential-Verteilung, es hängt von der Fragestellung ab!)

Wie gross ist die Wahrscheinlichkeit, dass genau 3 Kunden anrufen?

$$\mathrm{P}(3) = \frac{\mu^x}{x!} \cdot e^{-\mu} = \frac{5^3}{3!} \cdot e^{-5} = \underline{0.14}$$

Wie gross ist die Wahrscheinlichkeit, dass die Hotline mit mehr als 9 Anrufen gleichzeitig zu kämpfen hat und somit überlastet ist?
Dazu muss 1 minus die Wahrscheinlichkeit von 0 bis 9 Anrufen gerechnet werden.

$$\mathrm{P} = 1 - \mathrm{P}(X \leq x = 9) = 1 - \sum_{i=0}^{x} \frac{\mu^x}{x!} \cdot e^{-\mu} = 1 - \sum_{i=0}^{9} \frac{5^i}{i!} \cdot e^{-5} = 1 - 0.9682 = \underline{0.0318}$$

8.6.1. Approximation der Binomialverteilung

Mit der Poissonverteilung kann die Binomialverteilung approximiert werden, wenn die Datengrundlage genügend klein ist *(Faustregel: Bei $np \leq 10$ und $n \geq 1500p$ kann die Approximation gemacht werden)*. Die Rechnung funktioniert dann ganz normal wie bei der Poisson-Verteilung.

9. STETIGE VERTEILUNGEN

Häufig arbeitet man nicht mit Werten, die exakt gleich sind *(Würfelaugen, Anzahl Personen in Schlange, Ausfallzeit)*, sondern nur annähernd gleich *(Länge von 2 Gegenständen, Füllstand von 2 Behältern)* oder mit Systemen, bei denen sich der Zustand über die Zeit kontinuierlich ändert *(Wartezeit in Schlange, Stromverbrauch eines Gebäude, Alterungsprozesse)*. Solche Prozesse können durch ***stetige Zufallsvariablen*** beschrieben werden.

9.1. WAHRSCHEINLICHKEITSDICHTEFUNKTION

Anstatt der Wahrscheinlichkeitsfunktion bei den diskreten Verteilungen gibt es bei den stetigen Verteilungen die ***Wahrscheinlichkeitsdichte*** bzw. ***Dichtefunktion*** $f(x)$.

$$f(x) = \mathrm{P}(a \leq X \leq b) = \int_a^b f(x)\,\mathrm{d}x$$

Mithilfe der Dichtefunktion kann die ***Wahrscheinlichkeit*** ermittelt werden, dass ein Wert ***realisiert*** wird, der ***innerhalb*** eines vorab definierten ***Intervalls*** liegt. Im Gegensatz zu Wahrscheinlichkeiten können Wahrscheinlichkeitsdichtefunktionen auch Werte ***über 1*** annehmen.

9.1.1. Stetige Verteilungsfunktion

Die stetige Verteilungsfunktion $F(x)$ wird aus der ***Integration*** der Dichtefunktion $f(x)$ gebildet. Anders herum wird durch ***Ableiten*** der Stetigen Verteilungsfunktion die Dichtefunktion berechnet.

$$F(X) = \int_{-\infty}^{\infty} f(t)\,\mathrm{d}t, \qquad f(x) = F'(x)$$

9.2. RECHTECK-VERTEILUNG/GLEICHVERTEILUNG

Die Rechteck-Verteilung *(auch als Gleichverteilung bezeichnet)* beschreibt Vorgänge, bei denen die Ereignisse nur Zahlen im Intervall $[a, b]$ sein können. Alle Ergebnisse in $[a, b]$ sind ***gleich wahrscheinlich***. Sie wird beispielsweise angenommen, wenn Fehler- oder Temperaturgrenzen angegeben sind.

Die ***Dichte- und Verteilungsfunktion*** der Rechteck-Verteilung sind:

$$f(x|a;b) = \frac{1}{b-a} \qquad F(x|a;b) = \frac{x-a}{b-a}$$

Erwartungswert	***Varianz***
$E(X) = \frac{a+b}{2}$	$\mathrm{var}(X) = \frac{1}{12} \cdot (b-a)^2$

Beispiel:

Eine Person trifft zu einer zufälligen Zeit an einer Bushaltestelle ein, bei der alle 10min ein Bus fährt.

a) Wie gross ist die Wahrscheinlichkeit, dass sie 3 Minuten auf den Bus warten muss?

1) a und b aus dem Text bestimmen *(Wie oft geschieht etwas?)*
   Wartezeit zwischen 0 und 10 Minuten: $[a, b] = [0, 10]$
2) x aus Text bestimmen
   $x = 3$
3) Wahrscheinlichkeit ist ein konkreter Wert, in Dichte-Formel einsetzen

$$f(3|0;10) = \frac{1}{10-0} = 0.1 = \underline{10\%}$$

b) Wie hoch ist die Wahrscheinlichkeit, dass sie höchstens 3 Minuten warten muss?

1) Wahrscheinlichkeit ist ein Intervall, deswegen Verteilungsfunktion verwenden

$$F(x \leq x|\ 0;10) = \frac{3-0}{10-0} = 0.3 = \underline{30\%}$$

9.3. DREIECKS-VERTEILUNG

Die Dreiecks-Verteilung hat zusätzlich zu den ***Maximalwerten*** $[a, b]$ noch den ***wahrscheinlichsten Wert*** c. Alle Werte sammeln sich also ***um*** c an, befinden sich aber immer ***zwischen*** a und b. Bei vielen praxisnahen Anwendungen sind nur ***spärlich Daten*** vorhanden, um eine konkrete Verteilung der Grundgesamtheit zu schätzen. Sind Werte wie Min, Max und Modus bekannt, nimmt man oft die Dreiecksverteilung.

Die ***Dichte- und Verteilungsfunktion*** der Dreieck-Verteilung sind:

$$f(x) = \begin{cases} \frac{2\cdot(x-a)}{(b-a)\cdot(c-a)}, & a \le x \le c \\ \frac{2}{b-a}, & x = c \\ \frac{2\cdot(b-x)}{(b-a)\cdot(b-c)}, & c < x \le b \end{cases} \qquad F(x) = \begin{cases} \frac{(x-a)^2}{(b-a)\cdot(c-a)}, & a \le x \le c \\ \frac{c-a}{b-a}, & x = c \\ 1 - \frac{(b-x)^2}{(b-a)\cdot(b-c)}, & c < x \le b \end{cases}$$

Erwartungswert	***Varianz***
$E(X) = \frac{a+b+c}{3}$	$\mathrm{var}(X) = \frac{(a-b)^2 + (b-c)^2 + (a-c)^2}{36}$

9.4. EXPONENTIAL-VERTEILUNG

Die Exponential-Verteilung ist der ***Kehrwert*** der ***Poisson-Verteilung***. Sie hat den Parameter λ, der die ***Zahl eines erwarteten Ereignis A pro Einheitsintervall*** festlegt. Damit wird die Wahrscheinlichkeit berechnet, dass der Abstand zwischen zwei aufeinanderfolgenden Ereignissen A höchstens das x-Fache der gegebenen Zeit oder Strecke beträgt. Ein häufiger Einsatzzweck ist die ***Berechnung der Länge von zufälligen Zeitintervallen*** *(z.B. Zeit zwischen 2 Anrufen, Lebensdauer von Atomen beim radioaktiven Zerfall, Lebensdauer von Maschinen ohne Berücksichtigung von Verschleiss).*

Die ***Dichte- und Verteilungsfunktion*** der Exponential-Verteilung sind:

$$f(t) = \begin{cases} \lambda \cdot e^{-\lambda\cdot x} & \text{, für } x \ge 0 \\ 0 & \text{, sonst} \end{cases} \qquad F(x) = \begin{cases} 1 - e^{-\lambda\cdot x} & \text{, für } x \ge 0 \\ 0 & \text{, für } x < 0 \end{cases}$$

Erwartungswert	***Varianz***	***Median***
$E(X) = \frac{1}{\lambda}$	$\mathrm{var}(X) = \frac{1}{\lambda^2}$	$\mathrm{Me} = \frac{\ln(2)}{\lambda}$

Beispiel:

Bei einer Hotline rufen pro Stunde durchschnittlich 5 *Kunden an.*

(Nicht die gleiche Aufgabe wie bei der Poisson-Verteilung, es hängt von der Fragestellung ab!)

a) Wie viele Minuten vergehen durchschnittlich zwischen Anrufen?
   1) λ bestimmen: $\lambda = 5$
   2) Aufgabenstellung beinhaltet «Durchschnittlich» $\Rightarrow$ gesucht ist der **Erwartungswert**:

$$E(x) = \mu = \frac{1}{\lambda} = \frac{1}{5} = 0.2 \quad \Rightarrow \quad 60 \cdot 0.2 = \underline{12 \text{ Minuten}}$$

b) Wie gross ist die Wahrscheinlichkeit, dass höchstens 6 Minuten zwischen 2 Anrufen vergehen?
   1) Wahrscheinlichkeit ist ein Intervall, also **Verteilungsfunktion**

$$6 \text{ Minuten}/60 = 0.1 \quad \Rightarrow \quad F(x \le 0.1) = 1 - e^{5\,\cdot\,0.1} = 0.39 = \underline{39\%}$$

c) Wie gross ist die Wahrscheinlichkeit, dass zwischen zwei Anrufen 6 bis 15 Minuten vergehen?

$$F(0.1 \le x \le 0.25) = F(0.25) - F(0.1) = 0.71 - 0.39 = 0.32 = \underline{32\%}$$

9.5. WEIBULL-VERTEILUNG

Die Weibull-Verteilung kann je nach ***Einstellung ihrer Parameter*** der ***Exponential-*** oder der ***Normalverteilung*** ähneln. Sie kann wie die Exponential-Verteilung eine zufällige Lebensdauer abschätzen, jedoch kann sie die Vorgeschichte eines Merkmals berücksichtigen.

Ein häufiger ***Einsatzbereich*** ist die Bestimmung von Wahrscheinlichkeiten für ***Lebenszeiten von Maschinen oder Bauteilen***, wobei anders als bei der Exponentialverteilung ***Alter*** und ***Nutzungsintensität*** mit in die Berechnungen eingehen können. Besonders im Falle von kostenintensiven Anlagen ist diese aufwendige Differenzierung von Bedeutung. Empirische Untersuchungen zeigen nämlich bei vielen Anlagen:

1. In der ***ersten Phase*** eine zunächst hohe, dann sinkende Ausfallwahrscheinlichkeit, etwa bis die optimale Einrichtung und Einstellung erfolgt ist
2. In der ***zweiten Phase*** eine gleichleibend niedrige Ausfallrate
3. Mit zunehmendem Alter in der ***dritten Phase*** altersbedingt eine ansteigende Ausfallrate.

Die Verteilung hat zwei Parameter: Den ***Skalenparameter*** λ und den ***Formparameter*** k.

λ gibt die mittlere Ausfallwahrscheinlichkeit pro Intervall an.

Beim Formparameter k der Weibull-Verteilung können wir verschiedene Interpretationen ablesen:
– $k < 1$: Ausfallrate nimmt mit der Zeit ab *(Ausfälle finden frühzeitig statt)*
– $k = 0$: Ausfallrate ist konstant *(zufällige äussere Einflüsse verursachen Ausfall)*
– $k > 1$: Ausfallrate nimmt mit der Zeit zu *(Alterungsprozesse verursachen Ausfälle)*

Da die Verteilung bei $k = 0$ wie die Exponentialverteilung auch eine ***konstante Ausfallrate*** annimmt, stellt diese also einen Spezialfall der Weibull-Verteilung dar.

Die ***Dichte- und Verteilungsfunktion*** der Weibull-Verteilung sind:

$$f(t) = \begin{cases} \lambda \cdot k \cdot t^{k-1} \cdot e^{-\lambda \cdot t^k} & \text{für } t \geq 0 \\ 0 & \text{sonst} \end{cases} \qquad F(x) = \begin{cases} 1 - e^{-\lambda \cdot x^k} & \text{für } x \geq 0 \\ 0 & \text{für } x < 0 \end{cases}$$

Erwartungswert	***Varianz***
$E(X) = \frac{1}{\lambda} \cdot \Gamma\left(1 + \frac{1}{k}\right) \quad \Rightarrow \quad \frac{1}{\lambda} \cdot \left(1 + \frac{1}{k}\right)!$	$\text{var}(X) = \frac{1}{\lambda^2} \cdot \left(\Gamma\left(1 + \frac{2}{k}\right) - \Gamma^2\left(1 + \frac{1}{k}\right)\right)$

9.5.1. Gammafunktion $\Gamma()$

Die beiden Kenngrössen der Weibull-Verteilung und die Gamma-Verteilung verwenden die Gammafunktion $\Gamma()$. Sie definiert, wie die ***Fakultät*** für positive reelle Zahlen gehandhabt wird.

$$\Gamma(n) = \begin{cases} \Gamma(n+1) = n! & ,n \in \mathbb{N} \\ \Gamma(n) = \int_0^\infty (t^{n-1} \cdot e^{-t})\, \mathrm{d}t & ,n \in \mathbb{R}^+ \end{cases}$$

In der Prüfung kann tendenziell einfach die Fakultät für die Gamma-Funktion eingesetzt werden.

9.6. GAMMA-VERTEILUNG

Die ***Gamma-Verteilung*** $G(\theta, k)$ wird häufig verwendet, um Warteschlangen zu modellieren *(z.B. Bedien- oder Reparaturzeiten)*. Sie hat den ***Skalenparameter*** θ und den ***Formparameter*** k. Für $k = 1$ erhält man die ***Exponentialverteilung.***

Die ***Dichtefunktion*** der Gamma-Verteilung. Die Verteilungsfunktion ist zu kompliziert für die Prüfung.

$$f(t) = \frac{x^{k-1} \cdot e^{-t/\theta}}{\theta^k \cdot \Gamma(k)} \quad \text{, für } t \geq 0$$

Erwartungswert	***Varianz***
$E(X) = k \cdot \theta$	$\text{var}(X) = k \cdot \theta^2$

9.7. NORMALVERTEILUNG

Die ***Normalverteilung*** $\mathcal{N}(\mu, \sigma^2)$ ist eine der wichtigsten Verteilungen, sie bildet eine ***Gauss-Glockenkurve***. Die Werte sammeln sich symmetrisch um den ***höchsten Punkt*** μ der Glockenkurve an. Die Wendepunkte der Kurven sind $\pm\sigma$ Einheiten von μ entfernt. Die ***Varianz*** σ^2 bestimmt, wie eng die Werte um μ liegen, bei einer hohen Varianz ist die Kurve entsprechend breiter.

Der ***zentrale Grenzwertsatz besagt***, dass sich der Mittelwert und die Summe unabhängig und identisch verteilter Zufallsvariablen bei einer beliebigen Verteilung mit zunehmenden Stichprobenumfang der Normalverteilung annähern.
Oder anders gesagt: Viele kleine unabhängige Zufallseffekte summieren sich ungefähr zu einer Normalverteilung. Dadurch sind z.B. Mittelwerte von Stichproben normalverteilt.

Ist also die Verteilung unbekannt, kann die Wahrscheinlichkeit approximativ mit der Normalverteilung berechnet werden. Dazu sollte aber eine ***Stichprobe*** $n > 30$ vorliegen, wenn nicht, sollte man besser die Student-t-Verteilung verwenden.

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

$$F(x) = \int_{-\infty}^{x} \left(\frac{1}{\sigma \cdot \sqrt{2 \cdot \pi}} \cdot e^{-\frac{1}{2} \cdot \left(\frac{t-\mu}{\sigma}\right)} \right) \text{d}t, \sigma > 0$$

μ: Erwartungswert *(Mittelwert)*
σ: Standardabweichung

Da dieses Integral nicht integrierbar ist, müssen die Verteilungswerte aus der Standardnormalverteilungstabelle abgelesen werden → Standardisierung erforderlich!

Erwartungswert	***Varianz***
$E(X) = \mu$	$\text{var}(X) = \sigma^2$

9.7.1. Standardnormalverteilung

Für die Verteilungsfunktion der Normalverteilung kann keine Stammfunktion gefunden werden, deswegen müssen die Wahrscheinlichkeitswerte aus einer Verteilungstabelle abgelesen werden. Es gibt aber unendlich viele Normalverteilungen, weil die Parameter μ und σ unendlich viele Werte annehmen können. Um dieses Problem zu lösen und für jede Normalverteilung einen Wahrscheinlichkeitswert zu erhalten, ***standardisiert*** man die Normalverteilung.

Eine ***standardisierte Normalverteilung*** $\mathcal{N}(0, 1)$ ist gegeben, wenn $\mu = 0$ und $\sigma^2 = 1$. Um eine ***standardisierte Zufallsvariable*** Z zu erhalten, gilt folgende Formel. Mit diesem Wert kann die ***Wahrscheinlichkeit*** $\Phi(Z)$ *(manchmal auch $F(Z)$)* aus einer Standardnormalverteilungstabelle abgelesen werden.
Achtung: Ist $Z < 0$, muss 1 minus Φ vom positiven Wert gerechnet werden!

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

$$\Phi(Z) = \begin{cases} \Phi(Z) & ,Z \geq 0 \\ 1 - \Phi(|Z|) & ,Z < 0 \end{cases}$$

Z: standardnormalverteilte Zufallsvariable
X: nicht-standardisierte normalverteilte Zufallsvariable
μ, σ^2: Parameter der nicht-standardisierten Normalverteilung
Φ: Verteilungsfunktion, Wert aus Tabelle ablesen

Taschenrechner: Menü-5-5-2 $\text{normCdf}(-\infty, x, 0, 1)$

Ebenfalls gibt es eine Umkehrfunktion um von der Wahrscheinlichkeit q das q-Quantil z_q zu erhalten. Dazu schlägt man in der Quantile der Standartnormalverteilungstabelle nach.

$$\Phi(Z_q) = q \text{ (aus Quantil-Tabelle oder TR normCdf)}$$

Beispiel:
Bier wird in Dosen mit einer durchschnittlichen Füllmenge von $753\ \text{ml}$ *mit einer Standardabweichung von* 2ml *abgefüllt.*

a) Wie gross ist die Wahrscheinlichkeit, dass die Sollfüllmenge von 750ml unterschritten wird?
   1) Transformieren zu Standardnormalverteilung

$$Z = \frac{x - \mu}{\sigma} = \frac{750 - 753}{2} = -1.5$$

   2) Wert von Z in Standardnormalverteilungstabelle nachschlagen

$$\text{P}(Z \leq -1.5) = \Phi(Z) = 1 - \Phi(|Z|) = 1 - \Phi(1.5) = 1 - 0.9332 = 0.0668 = \underline{6.68\%}$$

b) Wie gross ist die Wahrscheinlichkeit, dass in einer Dose mindestens 757ml enthalten sind?
   1) Transformieren zu Standardnormalverteilung

$$Z = \frac{x - \mu}{\sigma} = \frac{757 - 753}{2} = 2$$

   2) Grösser als etwas $\Rightarrow$ 1 - kleiner als etwas

$$\text{P}(Z > 2) = 1 - \text{P}(Z < 2) = 1 - \Phi(Z) = 1 - \Phi(2) = 1 - 0.9773 = 0.0227 = \underline{2.28\%}$$

Bei zwei Grenzen (x zwischen y und z), müssen die Wahrscheinlichkeiten einfach voneinander subtrahiert werden. Um aus der Wahrscheinlichkeit die Anzahl zu erhalten, Gesamtanzahl mal Wahrscheinlichkeit rechnen.

9.8. ÜBERBLICK STETIGE VERTEILUNGEN

Name	***Anwendung***
Rechteck-Verteilung / Gleichverteilung	Zahlen sind in einem Intervall und alle gleich wahrscheinlich *(z.B. Fehler- oder Temperaturgrenzen)*
Dreiecks-Verteilung	Zahlen sind in einem Intervall und haben einen wahrscheinlichsten Wert *(wenn Minimum, Maximum und Modus bekannt sind)*
Exponential-Verteilung	Wenn Zahlen exponentialverteilt sind *(Radioaktiver Zerfall, Lebensdauer)*
Weibull-Verteilung	Ähnelt der Exponential- oder der Normalverteilung *(z.B. für Lebensdauer, wenn Vorgeschichte relevant)*
Gamma-Verteilung	Wird häufig für Warteschlangenmodellation verwendet
Normalverteilung	Viele kleine unabhängige Zufallseffekte. Ist die Verteilung unbekannt, kann sie mit der Normalverteilung approximiert werden.

9.9. ERZEUGEN VON PSEUDO-ZUFALLSZAHLEN

Um Simulationsexperimente durchzuführen, müssen Zufallszahlen erzeugt werden.

- Diese müssen ***einer bestimmten Verteilungsfunktion folgen*** *(muss auf die Experimentdaten passen)*
- Diese müssen ***eindeutig in ihrer Reihenfolge wiederholbar sein***
  (damit Experiment unter gleichen Bedingungen wiederholbar ist)
- Es müssen ***beliebig viele Zufallszahlenfolgen erzeugbar sein***
  (um verschiedenen Prozessen verschiedenenes Verhalten zu geben und unterschiedliche Experimente durchzuführen)

«Echte» Zufallszahlen sind aufgrund von fehlender Reproduzierbarkeit und schlechten statistischen Eigenschaften ungeeignet, deshalb verwendet man ***Pseudo-Zufallszahlen***.

Ein ***Random Number Generator (RNG)*** kann gleichverteilte Pseudo-Zufallszahlen z.B. durch ***linear congruential generators*** realisieren. Diese haben einen Initialwert, den ***Seed***, welcher immer die gleiche Folge an Pseudo-Zufallszahlen generiert. Ein Durchlauf dauert so lange, bis wieder der Seed ausgegeben wird. Es ist nicht garantiert, dass jeder Wert im Intervall $[0, m]$ ausgegeben wird. Je nach Parameter kann ein Durchlauf eine sehr kurze/lange Zyklenlänge haben.

$$X_{n+1} = (a \cdot X_n + c) \bmod (m)$$

m: Modulus, die Zufallszahl wird im Intervall $[0, m]$ sein, meist Primzahl
a: Multiplier, die letzte Zahl wird damit multipliziert
c: Increment, die Multiplikation wird um diesen Wert verschoben
x_0: Seed, Initialwert

Beispiel mit $m = 9, a = 2, c = 0, x_0 = 1$

$$[1] \xrightarrow{(2 \cdot 1 + 0) \bmod 9} [2] \xrightarrow{(2 \cdot 2 + 0) \bmod 9} [4] \xrightarrow{(2 \cdot 4 + 0) \bmod 9} [8] \xrightarrow{(2 \cdot 8 + 0) \bmod 9} [7] \rightarrow [5] \rightarrow [1]$$

Beispiel mit $m = 9, a = 4, c = 1, x_0 = 0$

$$[0] \xrightarrow{(4 \cdot 0 + 1) \bmod 9} [1] \xrightarrow{(4 \cdot 1 + 1) \bmod 9} [5] \xrightarrow{(4 \cdot 5 + 1) \bmod 9} [3] \rightarrow [4] \rightarrow [8] \rightarrow [6] \rightarrow [7] \rightarrow [2] \rightarrow [0]$$

9.9.1. Inversive Transformationsmethode

Durch die inversive Transformationsmethode, auch ***Simulationslemma*** genannt, kann aus gleichverteilten Zufallszahlen, welche durch eine Zufallsfunktion $U = Fx(X)$ erzeugt wurden, mithilfe der Umkehrfunktion $X = F^{-1}x(U)$ eine andere Verteilungsfunktion generiert werden.

Beispiel:
Gegeben sind die zwischen $[0, 1]$ gleichverteilten Zufallszzahlen:

$$\text{Intervall} = [0, 1], \quad u_i = (0.71, 0.11, 0.98, 0.64)$$

a) Transformiere die Zahlen so, dass sie einer Gleichverteilung mit $a = 2, b = 7$ folgen.
   1) Verteilungsfunktion der Gleichverteilung aufschreiben und mit u gleichsetzen

$$F(x) = u = \frac{x - a}{b - a}$$

   2) Verteilungsfunktion invertieren (nach x auflösen) und Werte von u_i in Formel einfügen

$$F^{-1}(u) = x = u \cdot (b - a) + a = u_i \cdot (7 - 2) + 2 \quad \Rightarrow \quad \underline{x_i = (5.55, 2.55, 6.90, 5.20)}$$

b) Transformiere die Zahlen so, dass sie einer Exponentialverteilung mit $\lambda = 0.5$ folgen.

$$F(x) = u = 1 - e^{-\lambda \cdot x} \quad \Rightarrow \quad F^{-1}(u) = x = \frac{1}{\lambda} \cdot \ln(1 - r) \quad \Rightarrow \quad \underline{x_i = (2.48, 0.23, 7.82, 2.04)}$$

10. SCHLIESSENDE STATISTIK

Die schliessende Statistik versucht auf der ***Basis statistischer Modelle*** und ***Daten aus Stichproben*** zu allgemeinen Aussagen über eine Grundgesamtheit zu gelangen. Es werden Hypothesen über Verteilungen von Merkmalen getestet, indem die angenommenen Verteilungen mit gemessenen Werten *(Stichproben)* verglichen werden.

Liegt nur eine ***Stichprobe*** der Grundmenge vor *(z.B. eine Messreihe, diese ist auch nur eine Stichprobe)*, ***ohne*** dass der ***Mittelwert*** μ oder die ***Varianz*** σ^2 bekannt sind, können mithilfe der Schliessenden Statistik Testverfahren bzw. -verteilungen bestimmt werden, die Aussagen über die Genauigkeit der Schätzungen und die Verteilfunktion erlauben.

10.1. KONFIDENZINTERVALL

Das Konfidenzintervall gibt den Wertebereich an, in welcher sich ein Datenpunkt einer Verteilung mit einer bestimmten Wahrscheinlichkeit, dem ***Konfidenzniveau*** α, befinden sollte.
Beliebte Konfidenzniveaus sind 90%, 95%, 99% *(verwendet als 0.9, 0.95, 0.99)*.

Haben wir ***eine Stichprobe*** *(gemessene Daten)* $\{x_1, ..., x_n\}$, ist der Mittelwert dieser:

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

Hat man ***mehrere Stichproben***, ist der Mittelwert aller Stichproben die ***Stichprobenfunktion***:

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

Der Konfidenzintervall ist selbst wieder eine Zufallsvariable und streut um den Mittelwert der Grundgesamtheit.

$$\mathrm{P}\left(\bar{x}_u \leq \bar{X} \leq \bar{x}_o\right) = 1 - \alpha$$

$\bar{x}_u, \bar{x}_o$: Untere/Obere Grenze des Konfidenzintervalls
$\bar{X}$: Wert der Realisation
α: Konfidenzniveau als Kommazahl in $[0, 1]$

Satz 1: Wenn der Stichprobenmittelwert diesselbe Verteilung wie X hat, gilt

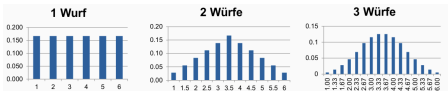
- Die Summe der Zufallsvariablen *(einzelne Mittelwerte)* ist wieder eine Zufallsvariable
- Der Erwartungswert der Stichprobe ist gleich dem Erwartungswert der Grundgesamtheit

$$\bar{X} = \mu_{\bar{x}} = \frac{1}{n}\sum_{i=1}^{n} X_i = \mu$$

- Die Varianz der Stichprobe wird immer kleiner, je grösser der Stichprobenumfang n wird *(die gemessenen Mittelwerte nähern sich dem tatsächlichen Mittelwert)*.

$$\sigma^2_{\bar{X}} = \frac{\sigma^2}{n}$$

Beispiel Würfel:
Je öfters ein Würfel geworfen wird, desto kleiner wird die Varianz, während der Mittelwert gleich bleibt.



	1. Wurf	*2. Wurf*	*3. Wurf*
Mittelwert μ	3.5	3.5	3.5
Varianz σ^2	2.29	$2.29/2 = 1.46$	$2.29/3 = 0.97$
Standardabweichung σ	1.71	1.21	0.89

Satz 2: Ist die Zufallsvariable X zusätzlich noch normalverteilt, ist auch der Stichprobenmittelwert normalverteilt

$$\bar{X} \Rightarrow N\left(\mu; \frac{\sigma}{\sqrt{n}}\right)$$

σ, μ: Standardabweichung & Mittelwert der Grundmenge
N: Verteilungsfunktion der Normalverteilung
n: Anzahl Stichproben
$\bar{X}$: Zufallsvariable der Stichprobe

10.2. TESTVERFAHREN DER NORMALVERTEILUNG

Viele Ergebnisse sind annähernd normalverteilt, je grösser die Anzahl Stichproben wird.

(siehe «Normalverteilung» (Seite 40))

Faustregel: Ab $n \geq 30$ ist die Normalverteilung unabhängig von der tatsächlichen Verteilung als Approximation geeignet.

Standardisierte Stichproben-Zufallsvariable

$$Z = \frac{X - \mu}{\sigma / \sqrt{n}}$$

n: Anzahl Durchführungen
X: Gesuchter Wert
σ, μ: Standardabweichung & Mittelwert

Beispiel mit unterer Schranke:

Eine Maschine produziert im normalverteilten Mittel 10 Stück pro Sekunde mit einer Standardabweichung von 1.5 pro Sekunde. Es werden 50 Messungen durchgeführt.

Mit welcher Wahrscheinlichkeit liegt die mittlere Produktion unter 9.5 Stück?

1. Variablen aus dem Text bestimmen

$$n = 50, \quad \mu = 10, \quad X = 9.5, \quad \sigma = 1.5$$

2. Werte in Formel einsetzen

$$Z = \frac{9.5 - 10}{1.5/\sqrt{50}} = -2.36$$

3. Berechneten Wert in die Tabelle der Standardnormalverteilung einsetzen

$$\Phi(Z) = \Phi(-2.36) = 0.0091$$

$\Rightarrow$ Der Mittelwert liegt mit einer Wahrscheinlichkeit von 0.91% im Intervall $[-\infty, 5]$

Beispiel: mit Konfidenzintervall (erlaubte Toleranz) die Wahrscheinlichkeit berechnen

Eine Maschine produziert im normalverteilten Mittel 10 Stück pro Sekunde mit einer Standardabweichung von 1 pro Sekunde. Es werden 25 Messungen durchgeführt. Mit welcher Wahrscheinlichkeit liegt die mittlere Produktion zwischen 9.8 und 10.2 Stück pro Sekunde?

1. Wahrscheinlichkeitsformel aufstellen

$$\mathrm{P}(9.8 \leq \bar{x} \leq 10.2) = 1 - \alpha$$

2. Die Konfidenzintervallschranken standardisieren *(in standardisierte Zufallsvariable konvertieren)*

$$\mathrm{P}\left(\frac{9.8 - 10}{1/\sqrt{25}} \leq \bar{Z} \leq \frac{10.2 - 10}{1/\sqrt{25}}\right) = \mathrm{P}\left(-\frac{0.2}{0.2} \leq \bar{Z} \leq \frac{0.2}{0.2}\right) = \mathrm{P}\big(-1 \leq \bar{Z} \leq 1\big)$$

3. Mit den Schranken die Wahrscheinlichkeit in der Standardnormalverteilungstabelle ablesen

$$\Phi(1) = \underbrace{\mathrm{normCdf}(-\infty, 1, 0, 1)}_{5\text{–}5\text{–}2 \text{ im TR}} = 0.8413, \quad \Phi(-1) = 1 - \Phi(1) = 1 - 0.8413 = 0.1587$$

4. Die Differenz der Wahrscheinlichkeit der Schranken berechnen

$$0.8413 - 0.1587 = 0.6843 = \underline{68.43\%}$$

Beispiel: mit Wahrscheinlichkeit das Konfidenzintervall berechnen
In welchem Konfidenzintervall liegt der Mittelwert bei der obenstehenden Aufgabe, wenn das Konfidenzniveau *(vorgegebene Wahrscheinlichkeit)* 95% ist?

1. Variablen aus dem Text bestimmen
$$n = 50, \quad \alpha = 0.95, \quad \sigma = 1.5$$
2. Z-Wert aus der Standardnormalverteilungs-Quantiltabelle auslesen
(Da die Abweichung beidseitig sein kann, muss der Wert 0.975 anstelle von 0.95 verwendet werden.)
$$\Phi^{-1}(Z_q) = \Phi(0.975) = \underbrace{\text{invNorm}(0.975, 0, 1)}_{5\text{–}5\text{–}3 \text{ im TR}} = \underline{1.96}$$

10.3. STUDENTISCHE T-VERTEILUNG

Ist die Anzahl Stichproben < 30, ist die ***t-Verteilung*** oft besser geeignet als die Normalverteilung. Sie flacht an den Enden weniger stark ab als die Normalverteilung, ist aber dafür um den Mittelwert weniger hoch. Auch wird bei der t-Verteilung die ***Standardabweichung nicht benötigt***. Der Vertrauensfaktor t wird aus der t-Verteilungstabelle abgelesen und ist von der Anzahl Freiheitsgrade abhängig.

10.3.1. Freiheitsgrad

Der ***Freiheitsgrad*** r/ν bestimmt bei einer Gleichung, wie viele Parameter «frei» wählbar sind, wenn das Resultat y bekannt ist. Bei den Verteilungstests ist vor allem der Mittelwert wichtig. Ist dieser bekannt, können $n-1$ Parameter zufällig gewählt werden. Der letzte aber muss so gewählt werden, dass die Summe aller Parameter μ ergibt.

Beispiel:
Bei $n = 5, \mu = 7$ können die ersten 4 Werte frei gewählt werden, der $5.$ aber so, dass die Werte in der Summe 7 ergeben. Die Anzahl Freiheitsgrade ist also 4.

Dichtefunktion der t-Verteilung:

$$f(t) = \frac{1}{\sqrt{r} \cdot B\left(\frac{1}{2}, \frac{r}{2}\right)} \cdot \left(1 + \frac{t^2}{r}\right)^{-\frac{1}{2}\cdot(r+1)}$$
$$B(x, y) = \int_0^1 \left(t^{x-1} \cdot (1-t)^{y-1}\right) \mathrm{d}t$$
$$t = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}}$$

t: Vertrauensfaktor *(aus der t-Verteilungstabelle ablesen)*
r: Anzahl Freiheitsgrade
$B()$**:** Betafunktion

TR: Menü-5-5-5 $\rightarrow \text{tCdf}(-\infty, x, r)$

Erwartungswert	***Varianz***
$E(X) = 0, \text{für } r > 1$	$\text{var}(X) = \frac{r}{r-2}, \text{für } r > 2$

Beispiel: *Eine Zufallsvariable X ist t-verteilt mit $\nu = 10$ Freiheitsgraden.*

a) Wie hoch ist die Wahrscheinlichkeit, dass x zwischen -1.4 und 1.8 liegt?

1) Vertrauensfaktoren aus der t-Verteilungstabelle bei entsprechendem k ablesen. Dabei sollten die Werte in der Tabelle die oben genannten Grenzwerte vollständig umfassen

$$F(x < 1.4) \approx 0.9, \quad F(x < 1.8) \approx 0.95$$

2) Wahrscheinlichkeit ausrechnen, dabei 1 minus negativer Wert rechnen

$$\begin{aligned} \mathrm{P}(-1.4 < x < 1.8) &= F(x < 1.8) - F(x < -1.4) \\ &= F(x < 1.8) - (1 - F(x < 1.4)) = 0.95 - (1 - 0.9) = \underline{0.85} \end{aligned}$$

Einfacher geht es mit dem Taschenrechner: Menü-5-5-5 $\mathrm{tCdf}(x_u, x_o, r) \rightarrow \mathrm{tCdf}(-1.4, 1.8, 10) = 0.853091$

b) In welchem mittlerem Bereich liegen die Realisationen mit einer Wahrscheinlichkeit von 99%?

1) Wert aus t-Verteilungstabelle bei entsprechendem k ablesen

$$0.995 = \mathrm{P}(x > 3.1693)$$

2) Intervall bilden, in welchem sich der Prozentsatz der Werte befindet

$$x = [-3.1693, +3.1693]$$

10.4. CHI-QUADRAT-VERTEILUNG

Hat man Zufallsvariablen, die ***unabhängig*** und ***standardnormalverteilt*** sind, ist die Chi-Quadratverteilung χ^2 die Verteilung der Summe der quadrierten Zufallsvariablen. Solche Summen quadrierter Zufallsvariablen treten auf bei: ***Varianz σ^2 einer Stichprobe***, ***Hypothesentest über die Verteilungsform***, ***Unabhängigkeitstest***. Wie die t-Verteilung hat sie die ***Anzahl Freiheitsgrade*** als Parameter. Nähert sich mit zunehmender Anzahl Freiheitsgrade der Normalverteilung an.

Dichtefunktion der χ^2_r-Verteilung:

$$f(t) = \begin{cases} \frac{1}{2^{r/2} \cdot \Gamma(r/2)} \cdot t^{(r/2)-1} \cdot e^{-(t/2)} & , \text{für } t > 0 \\ 0 & , \text{für } t \leq 0 \end{cases}$$

t: Vertrauensfaktor, aus der χ^2-Tabelle ablesen
r: Anzahl Freiheitsgrade
$\Gamma()$: Gammafunktion
(siehe Kapitel «Gammafunktion $\Gamma()$» (Seite 39))

TR: Menü-5-5-8 $\rightarrow \chi^2\mathrm{norm}(0, x, r)$

Erwartungswert	***Varianz***
$E(X) = r$	$\mathrm{var}(X) = 2 \cdot r$

Beispiel: *Eine Zufallsvariable X ist χ^2-verteilt mit $\nu = 10$ Freiheitsgraden.*

Wie hoch ist die Wahrscheinlichkeit, dass x zwischen 15 und 20 liegt?

a) Werte aus der χ^2-Verteilungstabelle ablesen *(ungenaue Werte!)*

$$F(x < 15) \approx 0.9, \quad F(x < 20) \approx 0.975$$

b) Wahrscheinlichkeit ausrechnen

$$\mathrm{P}(15 < x < 20) = F(x < 20) - F(x < 15) = 0.975 - 0.9 = \underline{0.075}$$

Einfacher geht es mit dem Taschenrechner: Menü-5-5-8 $\rightarrow \chi^2\mathrm{norm}(x_u, x_o, r)$

$$\chi^2\mathrm{norm}(15, 20, 10) = 0.102809 \text{ (genauer als mit der Tabelle)}$$

11. SCHÄTZVERFAHREN

Häufig sind μ, σ ***und die Verteilungsfunktion bekannt***, wenn sehr viele Messungen/Experimente durchgeführt werden. Diese sind jedoch teuer und sollen deshalb auf ein Minimum reduziert werden. Deswegen werden die Parameter oft geschätzt.

- ***Schätzverfahren:*** Schätzt unbekannte Parameter der Verteilung eines Merkmals in der Grundgesamtheit anhand einer Stichprobe
  - ***Punktschätzung:*** Schätzung durch Angabe eines einzigen Wertes
  - ***Intervallschätzung:*** Schätzung durch Angabe eines Intervalls
- ***Schätzfunktion:*** Ordnet einer Stichprobe einen Wert zu und lässt damit von der Stichprobe auf die Grundgesamtheit schliessen. Damit lässt sich der Fehler einer falschen Schätzung bestimmen/minimieren
- ***Schätzwert $\hat{T}$:*** Ergebnis der Schätzung von T, dem Parameter der Grundgesamtheit
- ***Anteilswert $p = k/n$***: Die Wahrscheinlichkeit eines Ereignisses. Ist ein Laplace-Experiment.
- ***Zufallsstichprobe:*** Es werden zufällig Elemente aus der Grundgesamtheit für eine Stichprobe gewählt.

11.1. PUNKTSCHÄTZUNG

Mit der Punktschätzung wird ein oder mehrere Parameter so gut wie möglich durch einen einzelnen Wert angenähert. Dies geschieht durch die ***quadratische Abweichung***:

$$E\left[\left(\hat{T}-T\right)^2\right] = \operatorname{var}\left(\hat{T}\right) + \left[E\left(\hat{T}-T\right)^2\right]$$

Ist $E\left(\hat{T}-T\right) = 0$, ist die Schätzung ***erwartungstreu***, hat also keine Abweichung vom tatsächlichen Wert.

Mittelwert-Schätzfunktion	***Varianz-Schätzfunktion***
$\mu = \bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$	$\sigma^2 = s^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(x_i - \bar{X}\right)^2$

Die durch obenstehende Schätzfunktionen erzeugten Parameter sind normalverteilt (μ) bzw. Chi-Quadrat-verteilt (σ)

11.2. INTERVALLSCHÄTZUNG DES ERWARTUNGSWERTES

Die Intervallschätzung zielt darauf ab, einen Bereich anzugeben, der mit einer gewissen *(selbstgewählten)* Wahrscheinlichkeit den wahren Wert enthält. Dieser Bereich wird auch ***Konfidenzintervall*** genannt. Häufige Konfidenzintervalle sind 90%, 95% und 99%.

Es gibt ***5 Schritte zur Erstellung eines Konfidenzintervalls*** für den Mittelwert $\bar{X}$ der Stichprobe X:

1. Feststellung der Verteilungsform

Zuerst muss festgestellt werden, welche Verteilung das Stichprobenmittel $\bar{X}$ besitzt:

Verteilung des Merkmals $\bar{X}$	***Varianz σ^2 bekannt***	***Varianz σ^2 unbekannt***
Normalverteilt	$\bar{X}$ ist normalverteilt	Bei $n \leq 30$: $\bar{X}$ ist t-verteilt mit $k = n-1$ Freiheitsgraden Bei $n > 30$: $\bar{X}$ ist approximativ normalverteilt
Nicht normalverteilt	$\bar{X}$ ist approximativ normalverteilt	
Unbekannt ($n > 30$)		

2. Feststellung der Varianz

Als zweites wird die ***Varianz*** für das Stichprobenmittel ***bestimmt***, wobei gilt:

$$N = \text{Grösse der Grundmenge}, \quad n = \text{Grösse der Stichprobe}, \quad s^2 = \text{Varianz-Schätzfunktion}$$

Für die Bestimmung von mit/ohne Zurücklegen/Wiederholung siehe Kapitel «Bestimmung der Kombinatorik-Formel» (Seite 31).

Stichprobenart	*Varianz σ^2 bekannt*	*Varianz σ^2 unbekannt*
Mit Zurücklegen *(unendliche Grundgesamtheit)*	$\sigma^2_{\bar{X}} = \frac{\sigma^2}{n}$	$\hat{\sigma}^2_{\bar{X}} = \frac{s^2}{n}$
Ohne Zurücklegen und $\frac{n}{N} < 0.05$	$\sigma^2_{\bar{X}} \approx \frac{\sigma^2}{n}$	$\hat{\sigma}^2_{\bar{X}} \approx \frac{s^2}{n}$
Ohne Zurücklegen und $\frac{n}{N} \geq 0.05$	$\sigma^2_{\bar{X}} = \frac{\sigma^2}{n} \cdot \frac{N-n}{N-1}$	$\sigma^2_{\bar{X}} = \frac{s^2}{n} \cdot \frac{N-n}{N}$

3. Bestimmen des Quantilwerts Z
Mittels Tabelle oder Rechner.

4. Berechnen des maximalen Schätzfehlers
Der maximale Schätzfehler ist das Produkt aus Quantilwerts und Standardabweichung von X.

5. Ermitteln der Konfidenzintervallgrenzen
Die Konfidenzgrenzen ergeben sich durch Addition/Subtraktion des max. Schätzfehlers vom Stichprobenmittel $\bar{X}$.

11.3. INTERVALLSCHÄTZUNG EINES ANTEILWERTS

Funktioniert ähnlich wie die Schätzung des Erwartungswerts

1. Bestimmen der Verteilungsform von P
Ist $n \cdot P \cdot (1-P) > 9$, ist die Schätzfunktion approximativ normalverteilt.

2. Bestimmen der Varianz von P
Für die Bestimmung von mit/ohne Zurücklegen/Wiederholung siehe Kapitel «Bestimmung der Kombinatorik-Formel» (Seite 31).

Stichprobenart	*Varianz σ^2 bekannt*	*Varianz σ^2 unbekannt*
Mit Zurücklegen *(unendliche Grundgesamtheit)*	$\hat{\sigma}^2_P = \frac{P \cdot (1-P)}{n}$	
Ohne Zurücklegen und $\frac{n}{N} < 0.05$	$\hat{\sigma}^2_P \approx \frac{P \cdot (1-P)}{n}$	
Ohne Zurücklegen und $\frac{n}{N} \geq 0.05$	$\hat{\sigma}^2_{\bar{X}} = \frac{P \cdot (1-P)}{n} \cdot \frac{N-n}{N-1}$	$\hat{\sigma}^2_{\bar{X}} = \frac{P \cdot (1-P)}{n} \cdot \frac{N-n}{N}$

3. Bestimmen des Quantilwerts Z
Mittels Tabelle oder Rechner.

4. Berechnung des maximalen Schätzfehlers
Der maximale Schätzfehler ist das Produkt aus Quantilwerts und Standardabweichung von P.

5. Ermitteln der Konfidenzintervallgrenzen
Die Konfidenzgrenzen ergeben sich durch Addition/Subtraktion des max. Schätzfehlers vom Stichprobenmittel $\bar{P}$.

11.4. STICHPROBENUMFANGBERECHNUNG

Für die Intervallschätzung muss manchmal die ***Stichprobengrösse*** bestummen werden, also wie viele Proben mindestens in der Stichprobe enthalten sein müssen, um eine bestimmte ***Genauigkeit*** α erreichen zu können. Für die Bestimmung des Z-Wertes aus der Standardnormalverteilung muss die Wahrscheinlichkeit $1-\alpha$ angewendet werden.

Mit Zurücklegen/Wiederholung

$$n \geq \frac{Z^2 \cdot \sigma^2}{e^2}$$

Ohne Zurücklegen/Wiederholung

$$n \geq \frac{Z^2 \cdot N \cdot \sigma^2}{e^2 \cdot (N-1) + Z^2 \cdot \sigma^2}$$

N: Anzahl Elemente in der Grundgesamtheit
σ^2: Varianz der Verteilung
e: Absolute Abweichung bzw. Fehler vom Mittelwert
Z: standardnormalverteilte Zufallsvariable
(siehe «Standardnormalverteilung» (Seite 40))

Beispiel ohne Zurücklegen

Eine Lieferung von $1'000$ Paketen Zucker ist mit einer 95% Konfidenz zu untersuchen, ob der garantierte Mittelwert eingehalten wird. Der Fehler $e = 0.2g$ und die Standardabweichung $\sigma = 1.2g$ sind bekannt.

a) Wie viele Pakete müssen mindestens entnommen werden?

1) Z aus Quantil-Standardnormalverteilungstabelle $\Phi(Z_q)$ bestimmen

$$95\% \Rightarrow 2.5\% \text{ links \& rechts der Normalverteilung} = 0.975 \quad \Rightarrow \quad Z = \Phi(0.975) = 1.96$$

2) Werte in Formel einsetzen

$$n \geq \frac{1.96^2 \cdot 1'000 \cdot 1.2^2}{0.2^2 \cdot (1'000-1) + 1.96^2 \cdot 1.2^2} = 121.6 \quad \Rightarrow \quad \underline{n > 122}$$

11.5. INTERVALLSCHÄTZUNG DER VARIANZ

$$\mathrm{P}\left(\frac{(n-1)\cdot s^2}{y_{1-\frac{\alpha}{2};r=n-1}} \leq \sigma^2 \leq \frac{(n-1)\cdot s^2}{y_{\frac{\alpha}{2};r=n-1}}\right)$$

r: Anzahl Freiheitsgrade
s: Varianzschätzfunktion
y: Verteilungsfunktion für neue, χ^2-verteilte Zufallsvariable

12. TESTVERFAHREN

Durch Erstellen und Experimentieren an einem Modell eines realen Produktionssystems möchte man ein verbessertes Modell erhalten, welches sich dann auf die realen Systeme anwenden lässt. In diesem Prozess werden zwei Hypothesen *(Annahmen)* aufgestellt:

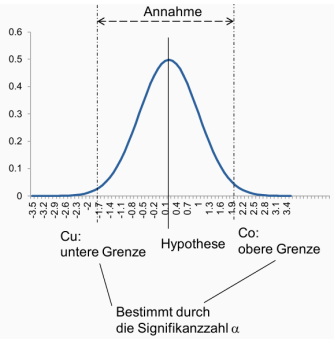
- H_0: Das Modell verhält sich bezüglich der zu untersuchenden Fragestellung wie das reale System *(Keine Veränderung)*
- H_1: Das verbesserte Modell verhält sich signifikant leistungsstärker als das ursprüngliche System *(Veränderung)*

Diese beiden Hypothesen müssen überprüft werden, bzw. die Wahrscheinlichkeit des Fehlers soll bestimmt werden.

12.1. FEHLERARTEN

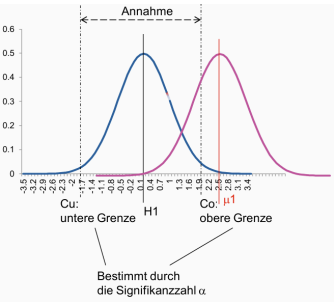
Bei der Überprüfung einer Hypothese kann es zu zwei verschiedenen Arten von Fehlern kommen.

12.1.1. Fehler erster Art (false negative)



Der Fehler erster Art *(auch Produzentenrisiko genannt)* ist ***die Ablehnung einer korrekten Hypothese:*** Wenn der Wert ausserhalb des Konfidenzintervalls liegt, wird die Hypothese verworfen. Hat man beispielsweise einen 95%-Konfidenzintervall und der erhaltene Wert liegt aber in den anderen 5%, so wird die Hypothese als «unwahrscheinlich» verworfen, obwohl sie eigentlich korrekt ist.

12.1.2. Fehler zweiter Art (false positive)



Der Fehler zweiter Art *(auch Konsumentenrisiko genannt)* ist ***das Annehmen einer inkorrekten Hypothese:*** Der erhaltene Wert bzw. die erhaltene Verteilung hat beispielsweise einen anderen Mittelwert als die Hypothese angenommen hat. Durch den gesetzten Konfidenzintervall liegt dieser aber immer noch in einem wahrscheinlichen Bereich. Die Hypothese wird somit angenommen, obwohl sie nicht der tatsächlichen Verteilung entspricht.

12.2. PARAMETERTEST

Der ***Parametertest*** prüft anhand einer Stichprobe, ob eine Hypothese zu einem bestimmten Parameter *(Mittelwert oder Varianz)* zutrifft. Um diesen durchführen zu können, wird jeweils eine Annahme über die Verteilung der ***Grundgesamtheit*** getroffen.

Bevor ein Parametertest durchgeführt werden kann, müssen zunächst verschiedene Werte festgelegt werden:

- ***Nullhypothese H_0:*** Fragestellung, welche aussagt, dass die geprüften Daten **keinen Zusammenhang** haben, das Experiment hat nicht den gewünschten Effekt
- ***Alternativhypothese H_1:*** Fragestellung, welche aussagt, **dass sich etwas ändert** – das Experiment hat den gewünschten Effekt. Sie ist immer die Negation *(das Gegenteil)* der Nullhypothese *(vgl. indirekter Beweis in der Mathematik)*
- ***Signifikanzzahl α:*** Die Irrtumswahrscheinlichkeit der Nullhypothese, also um wie viel Prozent die Messung abweichen darf, ohne dass sie verworfen wird. Häufig auch in der Form $1 - \alpha$ angegeben.
- ***Kritischer Wert:*** Wird dieser Wert mit Wahrscheinlichkeit α überschritten, ist die Nullhypothese widerlegt.
- ***Annahmebereich:*** Befindet sich das Parametertest-Resultat innerhalb des Annahmebereichs, ist die Nullhypothese angenommen. Analog wird beim ***Ablehnungsbereich*** die Nullhypothese verworfen. Krit. Wert ist die Grenze.

Das ***Ziel eines Parametertests*** ist es, aufzuzeigen, ob die Messung den kritischen Wert über-/unterschreitet und damit die Nullhypothese widerlegt. Tut sie das, hat das Experiment ein ***signifikantes Ergebnis***.

Es kann viele Möglichkeiten für die Widerlegung der Nullhypothese geben. Scheitert ein Test, heisst das nicht, dass die Nullhypothese wahr ist, sondern nur, dass sie noch nicht widerlegt wurde.

Beispiel zur Aufstellung der Hypothesen

Fragestellung: Hilft das neue Medikament dem Patienten, schneller gesund zu werden?

- ***Nullhypothese H_0:*** Es ist **kein** Unterschied in der Genesungszeit zwischen Medikament und Placebo feststellbar.
- ***Alternativhypothese H_1:*** Es ist **ein** Unterschied in der Genesungszeit zwischen Medikament und Placebo feststellbar.

Beispiel:
Die Zugstärke eines Rasenmäher-Modells hat einen Mittelwert von $1500N$ und eine Standardabweichung von $50N$. Man behauptet, mit einem neuen Verfahren die Zugstärke erhöhen zu können. Ein Test wird mit 60 Rasenmähern durchgeführt und ergibt eine mittlere Zugstärke von $1550N$.

a) Kann man an dieser Behauptung mit einer Irrtumswahrscheinlichkeit von 0.05 festhalten?
   1) Werte aus Text herausschreiben

$$X = 1550, \quad \text{Mittelwert } \mu = 1500, \quad \text{Standardabweichung } \sigma = 50, \quad n = 60, \quad \alpha = 0.05$$

   2) Hypothesen definieren:
      – H_0: Das neue Verfahren hat keine Veränderung der Zugstärke ergeben, $\mu = 1500N$
      – H_1: Es hat eine Verbesserung der Zugstärke gegeben, $\mu > 1500N$
   3) p-Wert und kritischer Z-Wert ausrechnen:

$$p = 1 - \alpha = 0.95, \quad z = \text{invNorm}(0.95, 0, 1) = 1.645$$

   3) Effektiver Z-Wert ausrechnen *(Formel siehe «Testverfahren der Normalverteilung» (Seite 44))*:

$$Z = \frac{X - \mu}{\sigma/\sqrt{n}} = \frac{1550 - 1500}{50/\sqrt{60}} = 7.74$$

   4) Effektiver Z-Wert mit Z-krit vergleichen. Ist $Z > Z$-krit: Nullhypothese verwerfen.
      $7.74 > 1.645$, das heisst, die Null-Hypothese muss verworfen und die Alternativ-Hypothese angenommen werden.

12.3. DIFFERENZTESTS FÜR MITTELWERTE VON NORMALVERTEILUNGEN

Manchmal will man mithilfe von Stichproben untersuchen, ob zwei Mittelwerte $\mu_1 = \mu_2$ gleich sind oder signifikant voneinander abweichen. z.B. ob ein angepasstes Modell im Durchschnitt besser ist als das reale System.
– ***Abhängige Stichproben:*** Die Messwerte haben eine Beziehung zueinander
– ***Unabhängige Stichproben:*** Die Messwerte haben keine Beziehung zueinander

Beispiel:
Vergleich Lohn im Alter von 30 Jahren vs. 50 Jahren: Befragt man 30- und 50-jährige Personen, ist die Stichprobe ***unabhängig***. Befragt man 50-jährige nach ihrem derzeitigen Lohn und ihrem Lohn als sie 30 Jahre alt waren, ist die Stichprobe ***abhängig***.

Annahmebereich für unabhängige Stichproben

$$c = \pm z \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

Z: Aus Tabelle Normalverteilung
σ_1^2, σ_2^2: Varianz der Stichproben 1 und 2
n_1, n_2: Gesamtanzahlen der Stichproben 1 und 2

12.3.1. Abhängige Stichproben

1. Bilden der Nullhypothese H_0: $\mu_1 = \mu_2$ *(Trifft sie ein, müsste die Differenz gegen 0 gehen)*
2. Signifikanzzahl α festlegen
3. Annahmegrenzen mithilfe Normalverteilungstabelle festlegen

12.3.2. Unabhängige Stichproben

Bei einer unabhängigen Stichprobe muss die Varianz für jede Probe einzeln geschätzt werden.
1. Bilden der Nullhypothese H_0: $\mu_1 = \mu_2$
2. $\pm Z$ mit Normalverteilungstabelle festlegen
3. Annahmebereich mit Formel oben berechnen
4. Mittelwerte der Stichproben μ_1 und μ_2 berechnen
5. Nullhypothese wird angenommen, wenn Differenz $d_i = \mu_1 - \mu_2$ in den Annahmebereich fällt

12.4. CHI-QUADRAT-TEST

Mit ***Verteilungstests*** können Hypotesen über die Wahrscheinlichkeitsverteilung einer Stichprobe überprüft werden. Wir verwenden in unserem Fall den ***Chi-Quadrat-Test***: Bei ihm werden die Häufigkeiten von empirisch ermittelten Verteilungen *(Messwerte)* mit der theoretischen Verteilung verglichen. Dabei werden die Differenzen der Häufigkeitswerte quadriert, normiert und aufaddiert.

Beispiel:
Test, ob ein Würfel eine Gleichverteilung produziert und damit fair ist. Ein Würfel wurde 60-mal geworfen mit folgenden Häufigkeiten:

$$h^e = \{\boxdot \mapsto 7,\ \boxdot \mapsto 8,\ \boxdot \mapsto 13,\ \boxdot \mapsto 8,\ \boxdot \mapsto 9,\ \boxdot \mapsto 15\}$$

1. Werte aus Text herausschreiben

$$n = 6, \quad p = \frac{1}{6}, \quad h^{\text{th}} = n \cdot p = 10, \quad \text{Annahme: } \alpha = 0.05$$

2. Hypothesen definieren
   - H_0: Die Würfel produzieren eine Gleichverteilung, sind also fair
   - H_1: Die Würfel produzieren keine Gleichverteilung, sind also nicht fair
3. Differenz zwischen der empirischen h^e und theoretischen h^{th} Verteilung bilden

$$h^e - h^{\text{th}} = \{\boxdot \mapsto 7 - 10 = -3,\ \boxdot \mapsto -2,\ \boxdot \mapsto 3,\ \boxdot \mapsto -2,\ \boxdot \mapsto 1,\ \boxdot \mapsto 5\}$$

4. Normieren, um negative Werte zu entfernen

$$\frac{\left(h_i^e - h_i^{\text{th}}\right)^2}{h_i^{\text{th}}} \Rightarrow \left\{1 \mapsto \frac{(7-10)^2}{10} = 0.9, 2 \mapsto \frac{(8-10)^2}{10} = 0.4, 3 \mapsto 0.9, 4 \mapsto 0.4, 5 \mapsto 0.1, 6 \mapsto 2.5\right\}$$

5. Summe der normierten Werte bilden, um D zu erhalten

$$D = 0.9 + 0.4 + 0.9 + 0.4 + 0.1 + 2.5 = \underline{5.2}$$

6. D_{krit} aus χ^2-Tabelle ablesen oder von TR und mit D vergleichen. Ist $D > D_{\text{krit}}$: Nullhypothese verwerfen.
   $5.2 < 11.0705$, das heisst, die Null-Hypothese kann nicht verworfen werden.