Mathe III

Testtheorie

Ralf Herbrich

Überblick

1. Konzept der Hypothesentests
2. Hypothesentests für den Erwartungswert
   - Gauss-Test
   - t-Test
3. Hypothesentests für Modelle
   - *Mann-Whitney-Wilcoxon*-Test
   - *Likelihood Ratio*-Test
4. Irrtumswahrscheinlichkeit und p-Wert
5. Multiples Testen

Überblick

1. Konzept der Hypothesentests
2. Hypothesentests für den Erwartungswert
   - Gauss-Test
   - t-Test
3. **Hypothesentests für Modelle**
   - *Mann-Whitney-Wilcoxon*-Test
   - *Likelihood Ratio*-Test
4. Irrtumswahrscheinlichkeit und p-Wert
5. Multiples Testen

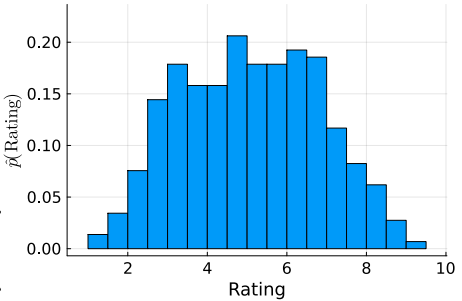
Motivation: Hypothesentests für Modelle

- **Beispiel (Filmbewertungen)**. Wenn wir alle Science-Fiction Filme in Betracht ziehen, kommen die Ratings aus 2021 aus der gleichen Verteilung wie die Ratings aus 2022?
- **Ansatz**: Wir folgen dem Konstruktionsprinzip von Hypothesentests
  1. **Festlegen eines parametrischen Modells** $(\Omega, \mathcal{F}, \{P_\theta \mid \theta \in \Theta\})$

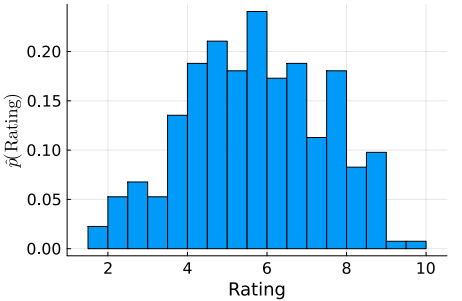
     $$\left([1,10]^{291+266}, \mathcal{B}([1,10]^{291+266}), \left\{(X_1, \dots, X_{291}, Y_1, \dots, Y_{266}) \to \prod_{i=1}^{291} P_X(X_i) \cdot \prod_{j=1}^{266} P_Y(Y_j)\right\}\right)$$

  2. **Formulierung von Hypothesen**
     - H_0: $P_X = P_Y$
     - H_1: $P_X \neq P_Y$
  3. **Wahl eines Irrtumsniveaus** α
     - $\alpha = 0.05 = 5\%$
  4. **Festlegen einer Entscheidungsregel**
     - Welche Teststatistik kann man benutzen, deren Verteilung man unter der Nullhypothese $P_X = P_Y$ kennt?

2021 (291 Filme)



2022 (266 Filme)



Überblick

1. Konzept der Hypothesentests
2. Hypothesentests für den Erwartungswert
   - Gauss-Test
   - t-Test
3. Hypothesentests für Modelle
   - ***Mann-Whitney-Wilcoxon*-Test**
   - *Likelihood Ratio*-Test
4. Irrtumswahrscheinlichkeit und p-Wert
5. Multiples Testen

Nicht-parametrischer *Mann-Whitney-Wilcoxon* Test

- **Idee**: Wenn die beiden Stichproben $X_1, \dots, X_n$ und $Y_1, \dots, Y_m$ aus der gleichen Verteilung kommen, dann muss es gleichwahrscheinlich sein, dass $X_i > Y_j$ oder $Y_j > X_i$!
  - ***Mann-Whitney-*U Statistik**: Führe alle $n \cdot m$ paarweisen Vergleiche von X_i und Y_j durch und zähle, wie häufig $X_i > Y_j$ (bei Gleichheit zählen wir $^1/_2$).

$$U = \sum_{i=1}^{n} \sum_{j=1}^{m} \begin{cases} 1 & \text{wenn } X_i > Y_j \\ \frac{1}{2} & \text{wenn } X_i = Y_j \\ 0 & \text{wenn } X_i < Y_j \end{cases}$$

  - ***Wilcoxon*-Rangsummenstatistik** W: Sortiere alle $X_1, \dots, X_n$ und $Y_1, \dots, Y_m$ gemeinsam und summiere die Ränge $R_1, \dots, R_n$ der $X_1, \dots, X_n$ nach dem Sortieren.

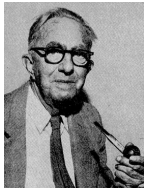
$$W = \sum_{i=1}^{n} R_i$$

- **Satz (*Mann-Whitney-Wilcoxon*)**. Die Teststatistiken U und W unterscheiden sich nur durch eine additive Konstante:

$$U = W - \frac{n \cdot (n+1)}{2}$$

Henry Berthold Mann
(1905 – 2000)

Donald Ransom Whitney
(1915 – 2007)



Frank Wilcoxon
(1892 – 1965)

Verteilung der *Wilcoxon*-Rangsummenstatistik W

- Sei $F(n, m, k)$ die Wahrscheinlichkeit, dass die Summe der Ränge in $X_1, \dots, X_n$ kleiner oder gleich $k \in \mathbb{R}$ sind (die zweite Stichprobe ist $Y_1, \dots, Y_m$).
- Die Verteilungsfunktion der W-Statistik unter der Nullhypothese kann **rekursiv** hergeleitet werden!

1. $F(n, m, k) = 0$ wenn $n < 0$ oder $m < 0$ oder $k < 0$.
2. $F(1,0, k) = 0$ wenn $k = 0$ aber $F(1,0, k) = 1$ wenn $k > 0$.
3. $F(0,1, k) = 1$ wenn $k \geq 0$.
4. $F(n, m, k) = \frac{n}{n+m} \cdot F(n-1, m, k-n-m) + \frac{m}{n+m} \cdot F(n, m-1, k)$

Nur ein Element X_1 und keine Stichprobe Y

Nur ein Element Y_1 und keine Stichprobe X

Wahrscheinlichkeit, dass das größte Element $X_1, \dots, X_n$ und $Y_1, \dots, Y_m$ aus $X_1, \dots, X_n$ kommt

Wahrscheinlichkeit, dass die Summe der Ränge in $X_1, \dots, X_{n-1}$ und $Y_1, \dots, Y_m$ kleiner gleich $k - (n + m)$ ist

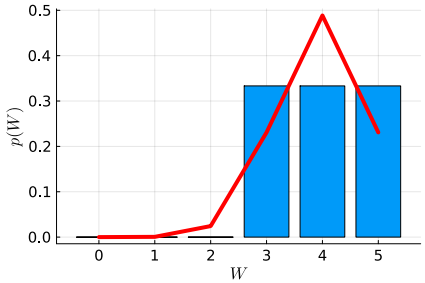
Wahrscheinlichkeit, dass das größte Element $X_1, \dots, X_n$ und $Y_1, \dots, Y_m$ aus $Y_1, \dots, Y_m$ kommt

Wahrscheinlichkeit, dass die Summe der Ränge in $X_1, \dots, X_n$ und $Y_1, \dots, Y_{m-1}$ kleiner gleich k ist

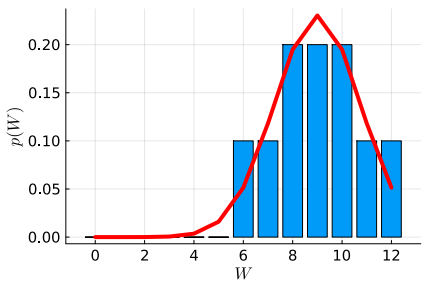
Verteilung der *Wilcoxon*-Rangsummenstatistik W

- Beispiele der Verteilung von W

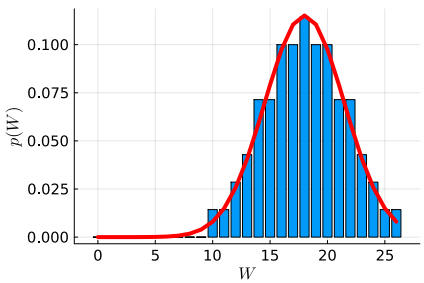
$n = 2, m = 1$



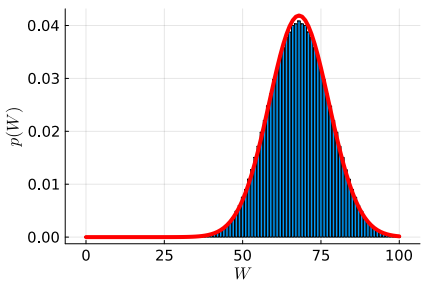
$n = 3, m = 2$



$n = 4, m = 4$



$n = 8, m = 8$



- **Satz (Approximative Verteilung der W-Statistik)**. Die Teststatistik W ist approximativ normalverteilt mit

$$W \sim^d \mathcal{N}\left(\frac{n \cdot (n+m+1)}{2}, \frac{n \cdot m \cdot (n+m+1)}{12}\right)$$

- **Bemerkung (Approximative Verteilung der W-Statistik)**
  - Der Beweis ist kompliziert, da die $n \cdot m$ Zufallsvariablen der paarweisen Vergleiche abhängig sind.

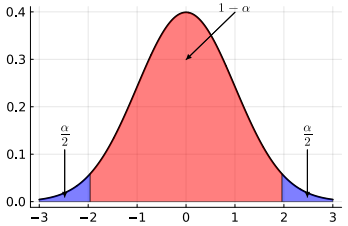
Mann-Whitney-Wilcoxon-Tests

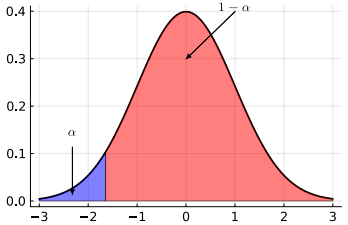
- **Daten**: Zwei Stichproben von identisch und unabhängig verteilten Zufallsvariablen $X_1, \dots, X_n$ und $Y_1, \dots, Y_m$.
- **Nullhypothese** H_0: $P_X = P_Y$
- **Teststatistik**: Wir betrachten die W Statistik, weil unter H_0

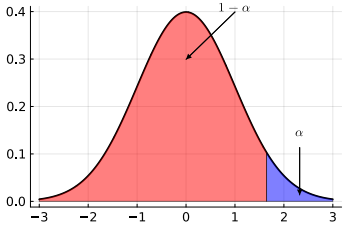
$$W \sim^d \mathcal{N}\left(\frac{n \cdot (n+m+1)}{2}, \frac{n \cdot m \cdot (n+m+1)}{12}\right) \Leftrightarrow \sqrt{3} \cdot \frac{2W - n \cdot (n+m+1)}{\sqrt{n \cdot m \cdot (n+m+1)}} \sim^d \mathcal{N}(0,1)$$

$W_{n,m}$

- **Annahmeregion (rotes Intervall)**:
  1. $|W_{n,m}| \leq z_{1-\frac{\alpha}{2}}$
  2. $W_{n,m} \geq z_{\alpha}$
  3. $W_{n,m} \leq z_{1-\alpha}$







- **Bemerkung (*Mann-Whitney-Wilcoxon*-Tests)**
  - Die Wahl der Annahmeregion hängt davon ab, was wir für eine Verschiebung des Mittelwertes erwarten.

Beispiel: *Mann-Whitney-Wilcoxon*-Tests

- **Beispiel (Filmbewertungen)**. Wenn wir alle Science-Fiction Filme in Betracht ziehen, kommen die Ratings aus 2021 aus der gleichen Verteilung wie die Ratings aus 2022?
  - **Ansatz**: Wir berechnen die W-Statistik für diesen Datensatz

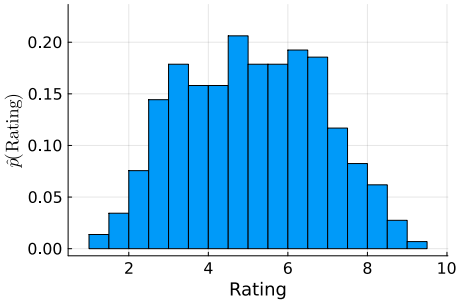
    $$W = 74542.5$$

    Damit ergibt sich

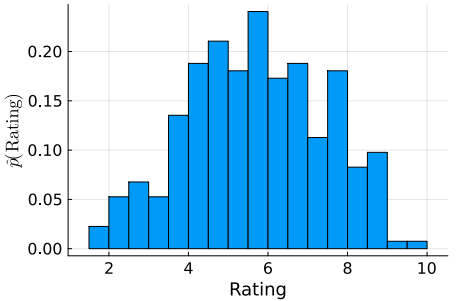
    $$W_{n,m} \approx -4.7$$

    Egal welche Annahmeregion wir benutzen, bei $\alpha = 0.05$ führt das zur Ablehnung der Nullhypothese.

2021 (291 Filme)



2022 (266 Filme)



- **Bemerkungen (*Mann-Whitney-Wilcoxon*-Test in `R`)**
  - Für diesen Test gibt es eine spezielle `R` Funktion: **`wilcox.test(x,y)`**

    ```
    W = 31178, p-value = 1.74e-06

    alternative hypothesis: true location shift is not equal to 0
    ```

  - Beachte, dass `R` hier den Wert der U-Statistik als W ausgibt!

Überblick

1. Konzept der Hypothesentests
2. Hypothesentests für den Erwartungswert
   - Gauss-Test
   - t-Test
3. Hypothesentests für Modelle
   - *Mann-Whitney-Wilcoxon*-Test
   - ***Likelihood Ratio*-Test**
4. Irrtumswahrscheinlichkeit und p-Wert
5. Multiples Testen

Motivation: *Likelihood-Ratio*-Test

- **Beispiel (Filmbewertungen)**. Wenn wir alle Science-Fiction Filme in Betracht ziehen, kommen die Ratings aus 2021 aus der gleichen Normalverteilung wie die Ratings aus 2022 bei angenommener Varianz von 3?
- **Ansatz**: Wir folgen dem Konstruktionsprinzip von Hypothesentests aber nehmen an, wir haben zwei **Mengen** von Modellen (nicht nur zwei Modelle!)

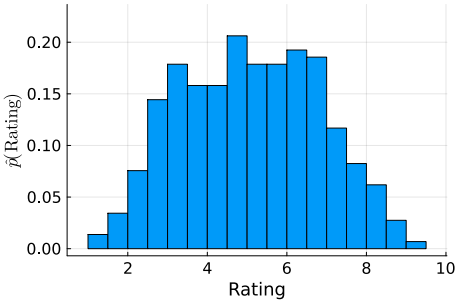
1. **Festlegen eines parametrischen Modells** $(\Omega, \mathcal{F}, \{P_\theta \mid \theta \in \Theta_0 \cup \Theta_1\})$

$$\left([1,10]^{291+266}, \mathcal{B}([1,10]^{291+266}), \left\{(X_1, \dots, X_{291}, Y_1, \dots, Y_{266}) \to \prod_{i=1}^{291} \mathcal{N}(X_i; \mu_X, 3) \cdot \prod_{j=1}^{266} \mathcal{N}(Y_j; \mu_Y, 3)\right\}\right)$$

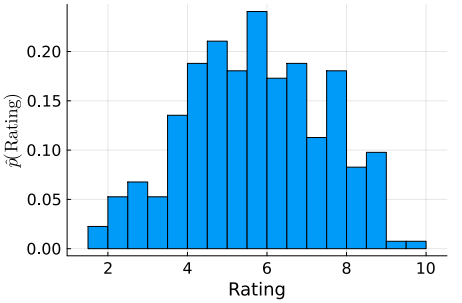
$$\Theta_0 = \{(\mu_X, \mu_Y) \in \mathbb{R}^2 \mid \mu_X = \mu_Y\} \leftarrow \text{Ein-Parameter Modell}$$
$$\Theta_1 = \{(\mu_X, \mu_Y) \in \mathbb{R}^2 \mid \mu_X \neq \mu_Y\} \leftarrow \text{Zwei-Parameter Modell}$$

2. **Formulierung von Hypothesen**
   - $H_0: \theta \in \Theta_0$
   - $H_1: \theta \in \Theta_1$

2021 (291 Filme)



2022 (266 Filme)



Likelihood-Ratio-Test

- Wir vergleichen ganze Mengen von parametrischen Modellen, welche den Daten unterliegen sollen.
- **Frage**: Welcher der Verteilungen in den parametrischen Modellen Θ_0 und Θ_1 benutzen wir, wenn wir einen Datensatz gegeben haben?
  - **Antwort**: Die wahrscheinlichste Verteilung der Daten auch bekannt als die *Maximum-Likelihood* Schätzung in den beiden parametrischen Modellen!
- **Definition (*Likelihood*)**. Ist $(\Omega, \mathcal{F}, \{P_\theta \mid \theta \in \Theta\})$ ein parametrisches Modell und sei $x \in \Omega$ eine Stichprobe, so heißt die Funktion $\mathcal{L}\colon \Theta \times \Omega \to [0, +\infty)$ mit $\mathcal{L}(\theta, x) = p_\theta(x)$ die zugehörige *Likelihood*, wobei p_θ die (Zähl)dichte von P_θ ist.
- **Teststatistik des *Likelihood-Ratio*-Tests** für einen Datensatz x

$$\lambda_{\mathrm{LR}} = -2 \cdot \log\left(\frac{\sup_{\theta \in \Theta_0} \mathcal{L}(\theta, x)}{\sup_{\theta \in \Theta_0 \cup \Theta_1} \mathcal{L}(\theta, x)}\right)$$

Maximale Wahrscheinlichkeit der Daten unter Modell Θ_0

Maximale Wahrscheinlichkeit der Daten unter Modell $\Theta_0 \cup \Theta_1$

- **Bemerkung (*Likelihood-Ratio*-Test)**
  - Unter der Nullhypothese konvergiert λ_{LR} gegen Null mit steigender Stichprobengröße!

Samuel Stanley Wilks (1906 – 1964)

Likelihood-Ratio-Test

- **Satz (von Wilks)**. Für ein parametrisches Modell $(\mathcal{F}^n, \mathcal{B}(\mathcal{F}^n), \{P_\theta \mid \theta \in \Theta_0 \cup \Theta_1\}), \mathcal{F} \subseteq \mathbb{R}$ und $n \to \infty$ folgt unter der Nullhypothese, dass

$$\lambda_{\mathrm{LR}} \sim^d \chi^2(r)$$

  wobei r der Unterschied der frei wählbaren Parameter in $\Theta_0 \cup \Theta_1$ und Θ_0 ist.

- **Annahmeregion (rotes Intervall)**: Da unter H_0 die wahrscheinlichen Werte von λ_{LR} nahe Null sind, wählen wir

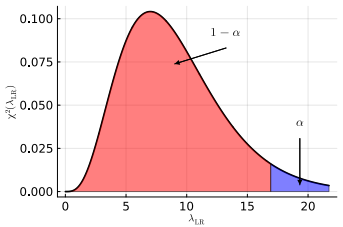
$$\lambda_{\mathrm{LR}} \leq c_{1-\alpha}$$

- **Beispiel (Filmbewertungen)**. Wenn wir alle Science-Fiction Filme in Betracht ziehen, kommen die Ratings aus 2021 aus der gleichen Verteilung wie die Ratings aus 2022?
  - Im Modell Θ_0 ist $\hat{\mu}_X = \hat{\mu}_Y = 5.39$ und $\log\left(\mathcal{L}\left((\hat{\mu}_X, \hat{\mu}_Y), x\right)\right) = -1137.07$
  - Im Modell $\Theta_0 \cup \Theta_1$ ist $\hat{\mu}_X = 5.05$ und $\hat{\mu}_Y = 5.74$ und $\log\left(\mathcal{L}\left((\hat{\mu}_X, \hat{\mu}_Y), x\right)\right) = -1125.76$
  - Daraus folgt, dass $\lambda_{\mathrm{LR}} = 22.62 > 3.84 = c_{0.95}$ für

$$r = 2 - 1 = 1$$

    daher Ablehnung der Nullhypothese.

$r = 9$



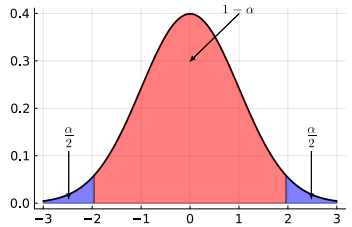
Überblick

1. Konzept der Hypothesentests
2. Hypothesentests für den Erwartungswert
   - Gauss-Test
   - t-Test
3. Hypothesentests für Modelle
   - *Mann-Whitney-Wilcoxon*-Test
   - *Likelihood Ratio*-Test
4. **Irrtumswahrscheinlichkeit und p-Wert**
5. Multiples Testen

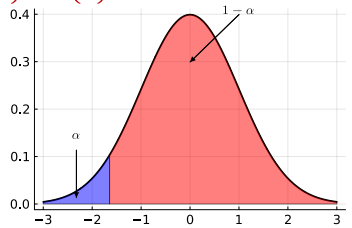
Irrtumswahrscheinlichkeit und p-Wert

- **Wiederholung: Konstruktion eines Hypothesentests in 4 Schritten**
  1. Festlegen eines parametrischen Modells $(\Omega, \mathcal{F}, \{P_\theta \mid \theta \in \Theta\})$
  2. Formulierung von Null- und Alternativhypothese
  3. Wahl eines Irrtumsniveaus α
  4. Konstruktion einer Ablehnungsregion (*rejection region*) unter Nullhypothese
- Es ist **garantiert**, dass der Fehler 1. Art (d.h., $P(\text{Entscheidung für } H_1 \mid H_0)$) maximal so hoch ist, wie das Irrtumsniveau über die wiederholte Anwendung des Hypothesentests!
  - Diese Garantie gibt es, da die Konstruktion **nicht** von der Stichprobe abhängt und damit auch nicht die Entscheidung!
- **Alternative p-Wert**: Wir berechnen aus der Stichprobe, für welches Irrtumsniveau $\alpha(X)$ die Stichprobe in der Ablehnungsregion gewesen wäre!
- **Bemerkungen (*p-Wert*)**
  - Das stichprobenabhängige Irrtumsniveau wird auch p-Wert genannt.
  - Intuitiv: bei welchem Irrtumsniveau lehnt der Test die Nullhypothese ab.

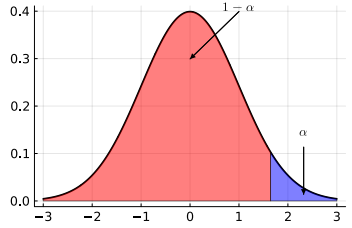
$\alpha(X) = 2 \cdot \Phi(-|X|)$



$\alpha(X) = \Phi(X)$



$\alpha(X) = 1 - \Phi(X)$



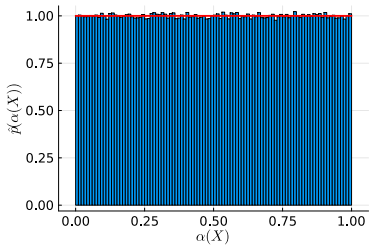
Verteilung von p-Werten

- Da der p-Wert von der Stichprobe abhängt, hat er auch eine Verteilung!
- **Satz (Verteilung von p-Werten)**. Für ein parametrisches Modell $(\mathbb{R}^n, \mathcal{B}(\mathbb{R}^n), \{P_\theta \mid \theta \in \mathbb{R}\})$ und Nullhypothesen H_0 der Form $\theta = \theta_0$ gilt, dass der p-Wert $\alpha(X)$ gleichverteilt auf dem Intervall $[0,1]$ ist, d.h.

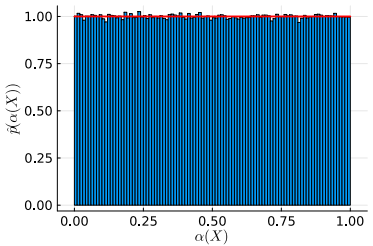
$$\alpha(X) \sim \mathcal{U}([0,1])$$

- **Beispiel (Verteilung von p-Werten)**. Wir simulieren $X \sim \mathcal{N}(0,1)$

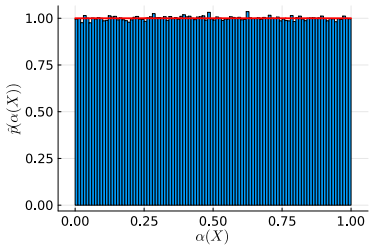
$\alpha(X) = 2 \cdot \Phi(-|X|)$



$\alpha(X) = \Phi(X)$



$\alpha(X) = 1 - \Phi(X)$



- **Bemerkungen (p-*Wert*)**
  - p-Werte sind sinnvoll zur Entscheidungsfindung (R berechnet immer den p-Wert).
  - p-Werte ist **keine** Wahrscheinlichkeit, dass die Nullhypothese wahr ist!

Überblick

1. Konzept der Hypothesentests
2. Hypothesentests für den Erwartungswert
   - Gauss-Test
   - t-Test
3. Hypothesentests für Modelle
   - *Mann-Whitney-Wilcoxon*-Test
   - *Likelihood Ratio*-Test
4. Irrtumswahrscheinlichkeit und p-Wert
5. **Multiples Testen**

Multiple Hypothesentests

- Bis jetzt haben wir nur Hypothesentests betrachtet, die **eine** logische Aussage als Nullhypothese repräsentieren (Elementarhypothesen).
  - **Beispiele (Filmbewertungen)**. Basierend auf IMDb:
    1. H_1: Das durchschnittliche Rating von Science-Fiction Filmen ist größer als 5.6.
    2. H_2: Das durchschnittliche Rating von Science-Fiction Filmen ist kleiner als 6.5.
- Bei einem **multiplen Test** konstruieren wir eine (globale) Nullhypothese H_0 durch Konjunktion von m Elementarhypothesen: $H_0 = H_1 \cap \cdots \cap H_m$.
  - **Beispiele (Filmbewertungen)**. Basierend auf IMDb:
    1. H_0: Das durchschnittliche Rating von Science-Fiction Filmen ist zwischen 5.6 und 6.5.
- **Frage**: Wenn wir die Hypothesentests H_i auf einem Irrtumsniveau von α ausführen, wie hoch ist die Wahrscheinlichkeit eines Typ I Fehlers von H_0?
- **Antwort**:

Teststatistik $T_i(X)$ ist in Ablehnungsregion R_i

$P(A \cup B) = P(A) + P(B) - P(A \cap B) \leq P(A) + P(B)$

Jeder einzelne Test hat ein Irrtumsniveau von α

$$P(\text{Typ I Fehler}) = P_X\left(\bigcup_{i=1}^{m} T_i(X) \in R_i \mid H_0\right) \leq \sum_{i=1}^{m} P_X(T_i(X) \in R_i | H_0) \leq m \cdot \alpha$$

Bonferroni-Korrektur

- **Satz (Bonferroni-Korrektur)**. Für eine Familie von m Nullhypothesen $\{H_1, \dots, H_m\}$ lehnt man jede der Nullhypothesen auf einem Irrtumsniveau von α/m ab. Dann ist garantiert, dass die globale Nullhypothese $H_0 = H_1 \cap \cdots \cap H_m$ das Irrtumsniveau von α einhält.
- **Beweis**: Folgt direkt aus $P(A \cup B) \leq P(A) + P(B)$
- **Beispiel (Gauss-Tests)**. Wir betrachten die folgenden zwei Elementarhypothesen
  1. $H_1: E[X] \geq \mu_0$
  2. $H_2: E[X] \leq \mu_0$

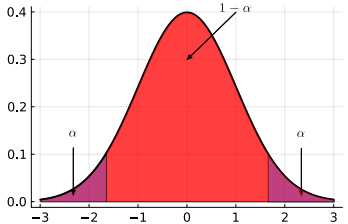
  Dann ist die globale Hypothese $H_0 = H_1 \cap H_2: E[X] = \mu_0$. Damit sind die Ablehnungsregionen:
  1. $\left(-\infty, \mu_0 + z_{\frac{\alpha}{2}} \cdot \frac{\sigma_X}{\sqrt{n}}\right]$
  2. $\left[\mu_0 + z_{1-\frac{\alpha}{2}} \cdot \frac{\sigma_X}{\sqrt{n}}, +\infty\right)$

  Daher entspricht die Annahmeregion des multiplen Tests:

$$\left[\mu_0 + z_{\frac{\alpha}{2}} \cdot \frac{\sigma_X}{\sqrt{n}}, \mu_0 + z_{1-\frac{\alpha}{2}} \cdot \frac{\sigma_X}{\sqrt{n}}\right]$$

Entspricht genau der Annahmeregion des zweiseitigen Gauss-Tests!

Carlo Emilio Bonferroni
(1892 – 1960)



Hasso Plattner Institut
Digital Engineering · Universität Potsdam

Viel Spaß bis zur nächsten Vorlesung!