Werkzeuge der Empirischen Forschung

Sommersemester 2020

Wolfgang Kössler

Humboldt-Universität zu Berlin, Institut für Informatik

6. Juli 2020

Übungen

Inhalt (1)

1. Mathematik, Folgen und Reihen, Potenzreihen, Funktionen
2. Mathematik, Differential- und Integralrechnung, Matrizen
3. Zufallsvariablen, Wahrscheinlichkeiten
4. Rechnen mit Erwartungswerten
5. Berechnen von robusten Lage- und Skalenschätzungen

Inhalt (2)

6. Fisher Information, optimale Schätzungen, Dichteschätzung
7. Abhängigkeitsmaße, Korrelationen, Regression
8. Hypothesentest
9. Varianzanalyse
10. Quadratische Formen, Matrizen und Eigenwerte

Inhalt (3)

11. Matrizenrechnung, Anwendung auf χ^2 Statistiken
12. Anpassungstests und Nichtparametrische Tests
13. Unabhängigkeitstest, Regression
14. Matrizenrechnung

Inhalt (4)

- Zahlbereiche: $\mathbb{N}, \mathbb{Z}, \mathbb{R}, (\mathbb{C})$
- Elementare Rechenoperationen: $+, -, *, **$
- Mengen, Rechnen mit Mmengen: $\cap, \cup, \bar{A}, \bigcap_{n=1}^{\infty}, \bigcup_{n=1}^{\infty}$
- Einige endliche Summen
- Folgen reeller Zahlen $(a_n) : \mathbb{N} \to \mathbb{R}$
- Unendliche Reihen=Folgen von Partialsummen $S_n = \sum_{i=1}^{n} a_i$
- Potenzreihen $\sum_{i=1}^{\infty} a_i x^i$ oder $\sum_{i=1}^{\infty} a_i (x - x_0)^i$
- Reelle Funktionen einer Veränderlichen $f : \quad \mathbb{R} \to \mathbb{R}$
- Stetigkeit von Funktionen

Einige endliche Summen

$$
\begin{aligned}
(a+b)^n &= \sum_{j=0}^{n} \binom{n}{j} a^j b^{n-j} \\
\sum_{j=0}^{n} j &= \frac{n(n+1)}{2} \\
\sum_{j=0}^{n} j^2 &= \frac{n(n+1)(2n+1)}{6} \\
\sum_{j=0}^{n} q^n &= \frac{1-q^{n+1}}{1-q}
\end{aligned}
$$

Folgen

monoton wachsend: $a_n \leq a_{n+1} \quad \forall n \in \mathbb{N}$

strikt monoton wachsend: $a_n < a_{n+1} \quad \forall n \in \mathbb{N}$

monoton fallendd: $a_n \geq a_{n+1} \quad \forall n \in \mathbb{N}$

strikt monoton fallend: $a_n > a_{n+1} \quad \forall n \in \mathbb{N}$

Folge (a_n) nach oben beschränkt: $\exists K : \; a_n \leq K$

Folge (a_n) nach unten beschränkt: $\exists K : \; a_n \geq K$

Folge (a_n) beschränkt: $\exists K : \; |a_n| \leq K$

Grenzwert a: $\forall \epsilon > 0 \; \exists n_0 : \quad \forall n \geq n_o : \quad |a_n - a| < \varepsilon$

Cauchy-Folge: $\forall \epsilon > 0 \; \exists n_0 : \quad \forall n, m \geq n_o : \quad |a_n - a_m| < \varepsilon$

Grenzwertsätze

- Eine monoton wachsende nach oben beschränkte Folge konvergiert
- Eine monoton fallende nach unten beschränkte Folge konvergiert
- Seien $(a_n), (b_n)$ konvergent, $\lim_{n\to\infty} a_n = a$, $\lim_{n\to\infty} b_n = b$. Dann gilt:

$$\begin{aligned} \lim(a_n \pm b_n) &= \lim a_n \pm \lim b_n \\ \lim(a_n \cdot b_n) &= \lim a_n \cdot \lim b_n \\ \lim \frac{a_n}{b_b} &= \frac{\lim a_n}{\lim b_n} = \frac{a}{b} \quad \text{falls} \quad b_n \neq 0, b \neq 0 \end{aligned}$$

- Eine Cauchy-Folge (=Fundamentalfolge) konvergiert.

Beispiele

- $a_n = n$ Folge natürlicher Zahlen
- Die Folge $\left((1+\frac{1}{n})^n\right)$ konvergiert. Erinnerung:
  Die Folge $\left((1+\frac{1}{n})^n\right)$ ist monoton wachsend,
  Die Folge $\left((1+\frac{1}{n})^{n+1}\right)$ ist monoton fallend,
  beide Folgen haben denselben Grenzwert (vgl. Grenzwersätze)
  $2 \leq (1+\frac{1}{n})^n < (1+\frac{1}{n})^{n+1} \leq 4$
  Der Grenzwert definiert die Eulersche Zahl e.

Sei $x \in \mathbb{R}, x \neq 0$ fest. Mit Hilfe der Grenzwertsätze bekommen wir

$$\begin{aligned}
\lim_{n\to\infty}\left(1+\frac{x}{n}\right)^n &= \lim_{m\to\infty}\left(1+\frac{1}{m}\right)^{m\cdot x} \qquad m=\frac{n}{x} \\
&= \left(\lim_{m\to\infty}\left(1+\frac{1}{m}\right)^{m}\right)^{x} \quad \text{Stetigkeit der Exponentialfkt.} \\
&= e^x \\
\lim_{n\to\infty}\left(1-\frac{x}{n}\right)^n &= e^{-x} \\
\lim_{n\to\infty}\left(1-\frac{1}{n}\right)^n &= e^{-1}
\end{aligned}$$

Unendliche Reihen, Konvergenzkriterien

- Es gelten die selben Grenzwertsätze wie für Folgen
- Majorantenkriterium: Seien $|a_i| \leq |b_i|$ und

  $\sum_{i=1}^{\infty} |b_i|$ konvergiert $\Rightarrow \sum_{i=1}^{\infty} |a_i|$ konvergiert

  $\sum_{i=1}^{\infty} |a_i|$ divergiert $\Rightarrow \sum_{i=1}^{\infty} |b_i|$ divergiert
- Quotientenkriterium: $\left|\frac{a_{i+1}}{a_i}\right| < q < 1 \Rightarrow \sum_{i=1}^{\infty} |a_i|$ konvergiert
- Wurzelkriterium: $\sqrt{|a_i|} < q < 1 \Rightarrow \sum_{i=1}^{\infty} |a_i|$ konvergiert

Geometrische Reihe, Harmonische Reihe, e, $\ln 2$

$$\sum_{i=0}^{n} q^i = \frac{1-q^{n+1}}{1-q}$$ Beweis: linke Seite mal (1-q) und ausrechnen

$$\sum_{i=0}^{\infty} q^i = \frac{1}{1-q} \quad \text{falls} \quad |q|<1, \qquad \sum_{n=0}^{\infty} \frac{1}{n!} = e, \qquad \sum_{n=1}^{\infty} \frac{(-1)^{n+1}}{n} = \ln 2$$

$$\sum_{n=1}^{\infty} \frac{1}{n} = \infty, \qquad \sum_{n=1}^{\infty} \frac{1}{n^2} = \frac{\pi^2}{6}, \qquad \sum_{n=1}^{\infty} \frac{1}{n^\alpha} \text{ konvergiert für } \alpha > 1$$

Funktionen

- Lineare Funktionen
- Polynome
- Potenzfunktionen
- Exponentialfunktionen
- Trigonometrische Funktionen $\sin, \cos, \tan, \cot, \arcsin, \arccos$
- Logarithmische Funktionen log ist hier der natürliche Logarithmus
- Gammafunktion

Potenzreihen

Sei die Funktion f in einer Umbebung von x_0 unendlich oft differenzierbar. Bezeichne $f^{(n)}(x_0)$ die n-te Ableitung an der Stelle x_0. Dann

$$f(x) = \sum_{n=0}^{\infty} \frac{f^{(n)}(x_0)}{n!}(x - x_0)^n.$$

Beispiele

$$e^x = \sum_{n=0}^{\infty} \frac{x^n}{n!}, \qquad \log(1+x) = \sum_{n=1}^{\infty} (-1)^{n+1} \frac{x^n}{n}$$

$$\sin(x) = \sum_{n=0}^{\infty} (-1)^n \frac{x^{2n+1}}{(2n+1)!}, \qquad \cos(x) = \sum_{n=0}^{\infty} (-1)^n \frac{x^{2n}}{(2n)!}$$

Stetigkeit von Funktionen

Die Funktion f heißt stetig in x_0, falls

$$\forall \varepsilon > 0 \quad \exists \delta > 0 \text{ so dass } \forall x : |x - x_o| < \delta \text{ gilt: } |f(x) - f(x_0)| < \varepsilon$$

$$\Leftrightarrow \quad \lim_{x \to x_0} f(x) = f(x_0)$$

Die Funktion f heißt stetig, falls sie in jedem Punkt x_0 des Definitionsbereichs stetig ist.

Differenzierbarkeit, Ableitung

Die Funktion f heißt differenzierbar in x_0, falls f stetig in x_0 und der Grenzwert

$$\lim_{x \to x_0} \frac{f(x) - f(x_0)}{x - x_0} = f'(x_0)$$

existiert. Der Grenzwert heißt Ableitung von f in x_0.

$f'(x_0)$ ist also der Anstieg der Tangente an f in x_0.

- Bestimmung von Extremwerten.

  Notwendige Bedingung $f'(x_0) = 0$, hinreichende Bedingung: $f'(x_0) > 0$ oder $f'(x_0) < 0$
- Potenzrehenentwicklung von Funktionen
- Approximation von Funktionen durch Polynome (Satz von Taylor)

Rechenregeln

$$
\begin{aligned}
(f \pm g)' &= f' \pm g' \\
(f \cdot g)' &= f' \cdot g + f \cdot g' && \text{Produktregel} \\
\left(\frac{f}{g}\right)' &= \frac{f' \cdot g - f \cdot g'}{g^2} \\
(f(g(x))' &= f'(g(x)) \cdot g'(x) && \text{Kettenregel}
\end{aligned}
$$

Integrierbarkeit, Bestimmtes Integral, Stammfunktion

Sei f eine Funktion, auf dem Intervall $[a, b]$ definiert. f heißt integrierbar in $[a, b]$, falls die Grenzwerte der Ober- und Untersummen existieren (und gleich sind).

Der Grenzwert bez. $\int_a^b f(x)\,dx$ heißt dann bestimmtes Integral (Riemann-Integral)

Die Funktion F heißt unbestimmtes Integral oder Stammfunktion von f, falls F differenzierbar und $F' = f$.

Hauptsatz der Differential- und Integralrechnung

$$\int_a^b f(x)\,dx = F(b) - F(a)$$

Rechenregeln

$$\int (f \pm g) = \int f \pm \int g$$

$$\int f'g = f \cdot g - \int fg' \quad \text{Partielle Integration, vgl. Produktregel}$$

$$F(g(x)) = \int f(g(x)) \cdot g'(x)\,dx \quad \text{Substitutionsregel, Kettenregel}$$

$$\int_a^c f(x)\,dx = \int_a^b f(x)\,dx + \int_b^c f(x)\,dx$$

Matrizen

Seien $\mathbf{A} = (a_{ij})_{i=1,\dots n, j=1,\dots,n}$ und $\mathbf{B} = (b_{jk})j = 1, \dots n, j = 1, \dots, l$ Matrizen. Dann ist das Produkt

$$\mathbf{A} \cdot \mathbf{B} = \left(\sum_{j=1}^{n} a_{ij} b_{jk}\right)_{i=1,\dots n, k=1,\dots,l}$$

Matrizenmultiplikation ist assoziativ, nicht kommutativ.

Rang der Matrix: maximale Anzahl linear unabhängiger Spalten

Sei $\mathbf{A}$ (n, n) Matrix. λ ist Eigenwert von $\mathbf{A}$ falls $\mathbf{Ax} = \lambda\mathbf{x}$, $\mathbf{x}$ heißt Eigenvektor.

Determinante von zweireihiger Matrix $\mathbf{A} = \begin{pmatrix} a & b \\ c & d \end{pmatrix} = a \cdot d - b \cdot c$

Binomialwahrscheinlichkeiten

$X \sim Bi(n,p), \quad p = 0.1, n = 25$

exakt (Binomial):

$$\begin{aligned} P(X \geq 5) &= \sum_{k=5}^{25} \binom{n}{k} p^k (1-p)^{n-k} \\ &= 1 - P(X \geq 4) = 1 - \text{CDF('Binomial',4,p,n)} \\ &= 0.097994 \end{aligned}$$

approximativ (Zentraler Grenzwertsatz, normal)

$$
\begin{aligned}
P(X > 4) &= 1 - P(X \leq 4) \\
&= 1 - P\Big(\frac{X - np}{\sqrt{np(1-p)}} \leq \frac{4 - np}{\sqrt{np(1-p)}}\Big) \\
&\approx 1 - \Phi\Big(\frac{1.5}{5 \cdot 0.3}\Big) = 1 - \Phi(1) = \Phi(-1) \\
&= \text{CDF('Normal',-1,0,1)} = 0.15866 \\
P(X \geq 5) &= 1 - P(X < 5) \\
&\approx 1 - P\Big(\frac{X - np}{\sqrt{np(1-p)}} < \frac{5 - np}{\sqrt{np(1-p)}}\Big) \\
&= 1 - \Phi\Big(\frac{5 - 2.5}{5 \cdot 0.3}\Big) \\
&= 1 - \Phi\Big(\frac{5}{3}\Big) = \Phi\Big(-\frac{5}{3}\Big) = 0.04779
\end{aligned}
$$

approximativ (ZGWS, normal, Stetigkeitskorrektur):

Beachten Sie den zusätzlichen Summanden -0.5 im Zähler.

$$\begin{aligned} P(X > 5) &\approx 1 - P\Big(\frac{X - np - 0.5}{\sqrt{np(1-p)}} \leq \frac{5 - np - 0.5}{\sqrt{np(1-p)}}\Big) \\ &= 1 - \Phi\Big(\frac{4.5 - 2.5}{5 \cdot 0.3}\Big) = 1 - \Phi\Big(\frac{4}{3}\Big) \\ &= \Phi\Big(-\frac{4}{3}\Big) = 0.09121 \end{aligned}$$

Die Approximation der Binomial- durch eine Normalverteilung ist hier nicht so gut (vor allem ist $p = 0.1$ klein).

Hypergeometrische und Binomial Verteilung

Seien

$$f(k|Bi(m,p)) = \binom{m}{k} p^k (1-p)^{m-k} \quad \text{Binomialwahrscheinlichkeit}$$

und

$$f(k|H_{N,n,m}) = \frac{\binom{n}{k} \cdot \binom{N-n}{m-k}}{\binom{N}{m}} \quad \text{Hypergeometrische Wahrscheinlichkeit.}$$

Satz: Es gilt Für $N \to \infty$, $n \to \infty$, $\frac{n}{N} \to p$ gilt:

$$f(k|H_{N,n,m}) \to \binom{m}{k} p^k (1-p)^{m-k} = f(k|Bi(m,p))$$

Hypergeometrische und Binomial Verteilung (2)

Beweis des Satzes Es gilt

$$f(k|H_{N,n,m}) = \frac{\binom{n}{k} \cdot \binom{N-n}{m-k}}{\binom{N}{m}}$$

$$= \frac{n! \cdot (N-n)! \cdot m!(N-m)!}{k!(n-k)! \cdot (N-n-m+k)!(m-k)! \cdot N!}$$

$$= \binom{m}{k} \frac{(N-n)!}{(N-n-m+k)!} \frac{n!}{(n-k)!} \frac{(N-m)!}{N!}$$

$$= \binom{m}{k} (N-n-m+k+1) \cdots (N-n) \cdot \frac{(n-k+1) \cdots n}{(N-m+1) \cdots N}$$

$$
\begin{aligned}
&= \binom{m}{k} \frac{(N-n-m+k+1)\cdots(N-n)}{N^{m-k}} \cdot \\
&\quad \frac{(n-k+1)\cdots n}{N^k} \frac{N^m}{(N-m+1)\cdots N} \\
&= \binom{m}{k} (1-p-\frac{m-k-1}{N})(1-p-\frac{m-k-2}{N})\cdots(1-p)\cdot \\
&\quad \frac{(p-\frac{k-1}{N})\cdots p}{(1-\frac{m-1}{N})(1-\frac{m-2}{N})\cdots(1-\frac{1}{N})\cdot 1} \\
&\to \binom{m}{k}(1-p)^{m-k}p^k = f(k|Bi(m,p))
\end{aligned}
$$

- $X \sim Exp(\lambda)$. Bestimmen Sie das 0.95-Quantil $u_{0.95}$, d.h. $F(u_{0.95}) = 0.95$.

  Bei $\lambda = 1 : u_{0.95} = 2.99573$

  `Quantile('Exponential',0.95,1)`

  Bestimmen Sie die Wahrscheinlichkeiten $P(X \leq \mathbf{E}(X))$ und $P(X \leq \text{med}(X))$

  Bestimmen Sie den Median $\text{med}(X)$, d.h. $P(X \leq \text{med}(X)) = \frac{1}{2}$

- Sei Φ die Verteilungsfunktion der Standardnormal. Berechnen Sie

$$\Phi(1), \quad \Phi(-1), \quad \Phi^{-1}(0.1), \quad \Phi^{-1}(0.9),$$

  `CDF('normal',1,0,1), CDF('normal',-1,0,1),`

  `Quantile('normal',0.1,0,1),`

  `Quantile('normal',0.9,0,1)`

Sei $X \sim Exp(\lambda)$. Es gilt:

$$\mathbf{E}X = \lambda \qquad \text{und} \qquad \text{var}X = \lambda^2.$$

Beweis:

$$\begin{aligned}
\mathbf{E}X &= \int_0^\infty x \cdot \frac{1}{\lambda} e^{-\frac{x}{\lambda}}\, dx = \\
&= x(-e^{-\frac{x}{\lambda}})|_0^\infty - \int_0^\infty (-e^{-\frac{x}{\lambda}})\, dx = \lambda. \\
\mathbf{E}X^2 &= \int_0^\infty x^2 \frac{1}{\lambda} e^{-\frac{x}{\lambda}}\, dx = \\
&= x^2(-e^{-\frac{x}{\lambda}})|_0^\infty - \int_0^\infty 2x(-e^{-\frac{x}{\lambda}})\, dx = 2\lambda^2 \\
\text{var}X &= 2\lambda^2 - \lambda^2 = \lambda^2.
\end{aligned}$$

Eigenschaften des Erwartungswerts

$$
\begin{aligned}
\mathbf{E}(X)+\mathbf{E}(Y) &= \sum_{j,k} x_j p(x_j,y_k) + \sum_{j,k} y_k p(x_j,y_k) \\
&= \sum_{j,k} (x_j+y_k) p(x_j,y_k) \qquad \text{Reihen konvergieren absolut} \\
&= \mathbf{E}(X+Y) \qquad \text{(X, Y diskret)} \\
\mathbf{E}(X)+\mathbf{E}(Y) &= \int\int x f(x,y)\,dx\,dy + \int\int y f(x,y)\,dx\,dy \\
&= \int\int (x+y) f(x,y)\,dx\,dy \qquad \text{(X, Y stetig)} \\
&= \mathbf{E}(X+Y)
\end{aligned}
$$

Im gemischten Fall gilt das auch.

- Zeigen Sie die in der Vorlesung angegebenen Eigenschaften der Varianz.
- Berechnen Sie Erwartungswert und Varianz der auf Folie 154 der Vorlesung angegebenen Verteilungen.
- Berechnen Sie Schiefe und Kurtosis dieser Verteilungen.
- Berechnen Sie die Varianz bei geometrischer Verteilung

χ^2-Verteilung

Def.: Seien $X_i \sim \mathcal{N}(0,1), i = 1, \ldots, n$, und unabhängig. Dann ist

$$Y = \sum_{i=1}^{n} X_i^2 \sim \chi_n^2$$

χ^2-verteilt mit n Freiheitsgraden.

Erwartungswert und Varianz lassen sich leicht ausrechnen:

$$\begin{aligned} \mathbf{E}(Y) &= n\mathbf{E}(X)^2 = var(X) + (\mathbf{E}(X))^2 = n(1+0) = n \\ \text{var}(Y) &= n \cdot \text{var}(X^2) = n(\mathbf{E}X^4 - (\mathbf{E}(X^2))^2) \\ &= n(3 - 1^2) = 2n \end{aligned}$$

Anmerkung: Die Dichte von Y ist gegeben durch $(y > 0)$

$$f_{\chi_n^2}(y) = \frac{1}{2^{n/2}\Gamma(\frac{n}{2})} e^{-\frac{y}{2}} y^{\frac{n}{2}-1}.$$

Seien X und Y unabhängige stetige oder diskrete Zufallsvariablen. Zeigen Sie

$$\mathbf{E}(X \cdot Y) = \mathbf{E}X \cdot \mathbf{E}Y.$$

Beweis: Sei X und Y stetige Zufallsvariablen mit gemeinsamer Dichte $h(x, y)$ und "Randdichten" $f(x)$ und $g(y)$. Da X und Y unabhängig sind, gilt für alle $x, y \in \mathbb{R}$

$$\int_{-\infty}^{x} \int_{-\infty}^{y} h(t, s)\, dt\, ds = P(X \leq x, Y \leq x)$$

$$= P(X \leq x)P(Y \leq y) = \int_{-\infty}^{x} f(s)\, ds \int_{-\infty}^{y} g(t)\, dt$$

Also: $h(x, y) = f(x)g(y)$. Daraus folgt:

Die letzte Gleichung folgt auch schon durch zweimaliges partielles Differenzieren der Unabhängigkeitsgleichung

$$F(x,y) = F(x) \cdot F(y)$$

$$\begin{aligned}
\mathbf{E}(X \cdot Y) &= \int_{\mathbb{R}^2} x \cdot y \cdot h(x,y)\, dx\, dy \\
&= \int_{\mathbb{R}^2} x \cdot y \cdot f(x) \cdot g(y)\, dx\, dy \\
&= \int_{-\infty}^{\infty} x\, f(x)\, dx \cdot \int_{-\infty}^{\infty} y\, g(y)\, dy \\
&= \mathbf{E}X \cdot \mathbf{E}Y.
\end{aligned}$$

Für diskrete Zufallsvariablen analog.

Korrelationskoeffizient bei bivariater Normalverteilung

Die Dichtefunktion der bivariaten Normalverteilung ist gegeben durch

$$f(x,y) = \frac{1}{2\pi\sqrt{1-\rho^2}} e^{-\frac{1}{1-\rho^2}(x^2-2xy\rho+y^2)}$$

Zeigen Sie $\mathbf{E}(X \cdot Y) = \rho$

Pareto-Verteilung

$$f(x) = (\alpha - 1)\frac{1}{x^\alpha}\, dx$$

ist Dichte von zwei Zufallsvariablen X und Y auf $[1, \infty)$ falls $\alpha > 1$ bzw. auf dem Intervall $[0, 1]$ falls $\alpha < 1$:

$$\begin{aligned}
\int_1^\infty f(x)\, dx &= \int_1^\infty (\alpha - 1)\frac{1}{x^\alpha}\, dx = -x^{-\alpha+1}\big|_1^\infty = 1 \quad \text{falls } \alpha > 1 \\
\int_0^1 f(x)\, dx &= \int_0^1 (\alpha - 1)\frac{1}{x^\alpha}\, dx = -x^{-\alpha+1}\big|_0^1 = 1 \quad \text{falls } \alpha < 1 \\
\mathbf{E}(X) &= (\alpha - 1)\int_1^\infty x \cdot \frac{1}{x^\alpha}\, dx = (\alpha - 1)\int_1^\infty x^{1-\alpha}\, dx \\
&= \frac{\alpha - 1}{2 - \alpha} x^{2-\alpha}\Big|_1^\infty = \begin{cases} \frac{\alpha-1}{\alpha-2} & \text{falls } \alpha > 2 \\ \infty & \text{sonst} \end{cases}
\end{aligned}$$

Pareto-Verteilung (2)

- Zeigen Sie

$$\mathbf{E}(X^2) = \begin{cases} \frac{\alpha-1}{\alpha-3} & \text{falls} \quad \alpha > 3 \\ \infty & \text{sonst} \end{cases}$$

  Damit haben Sir ein Beispiel für eine Verteilung für die Erwartungswert oder Varianz nicht existieren.
- Berechnen Sie var(X).
- Berechnen Sie Erwartungswert und Varianz von Y.

Beobachtungen: 1,2,4,8,10,17

- $\overline{X} = \frac{42}{6} = 7$, med(X)$=\frac{4+8}{2} = 6$, lomed=4, highmed=8
- $\overline{X}_{tr,1} = \frac{1}{4} \cdot 24 = 6$, $\overline{X}_{wi,1} = \frac{1}{6} \cdot 36 = 6$
- $s^2 = \frac{1}{5}((1-7)^2+(2-7)^2+(4-7)^2+(8-7)^2+(10-7)^2(17-7)^2)$
  $= \frac{1}{5}(6^2 + 5^2 + 3^2 + 1^2 + 3^2 + 10^2) = \frac{1}{5} \cdot 180 = 36$,
  $s = 6$, $S^2 = \frac{1}{6} \cdot 180 = 30$
- $x_{0.75} = x_{(5)} = 10$, $x_{0.25} = x_{(2)} = 2$, $IR = 10 - 2 = 8$
- MAD$=$med$|x_i - x_{0.5}| =$
  med $(|1-6|, |2-6|, 4-6|, |8-6|, |10-6|, |17-6|)=$
  med(5,4,2,2,4,11)$=4$
- $CV = \frac{s \cdot 100}{\overline{X}} = \frac{600}{7}$
- Gini$= 7.2$ (bitte nachrechnen)

Beobachtungen: 1,2,4,8,10,17, Berechnung von S_n (ohne Faktor)

i	$\lvert x_i - x_j \rvert$	highmed$\lvert x_i - x_j \rvert$
1	0,1,3,7,9,16	7
2	1,0,2,6,8,15	6
3	3,2,0,4,6,13	4
4	7,6,4,0,2,9	6
5	9,8,6,2,0,7	7
6	16,15,13,9,7,0	13
lomed		$S_n = 6$

Zur Berechnung von Q_n ordnen wir die Einträge in der oberen Dreiecksmatrix: 1,2,2,3,4,6,**6**,7,7,8,9,9,13,15,16.

$h = \lfloor \frac{n}{2} \rfloor + 1 = 4, k = \binom{h}{2} = 6$. Der 6.größte Eintrag oben ist 6, also $Q_n = 2.22. \cdot 6.$ (ohne Korrekturfaktor)

Berechnung von S_n und Q_n (ohne Faktor)

	Beobachtungen: 1,2,3,4,5		Beobachtungen: 1,2,3,4,10	
i	$\lvert x_i - x_j \rvert$	highmed$\lvert x_i - x_j \rvert$	$\lvert x_i - x_j \rvert$	highmed$\lvert x_i - x_j \rvert$
1	0,1,2,3,4	2	0,1,2,3,9	2
2	1,0,1,2,3	1	1,0,1,2,8	1
3	2,1,0,1,2	1	2,1,0,1,7	1
4	3,2,1,0,1	1	3,2,1,0,6	1
5	4,3,2,1,0	2	9,8,7,6,0	7
lomed		$S_n = 1.19 \cdot 1$		$S_n = 1.19 \cdot 1$

Zur Berechnung von Q_n ordnen wir die Einträge in der oberen Dreiecksmatrix: 1,1,**1**,1,2,2,2,3,3,4 bzw. 1,1,**1**,2,2,3,6,7,8,9

$h = \lfloor \frac{n}{2} \rfloor + 1 = 3, k = \binom{h}{2} = 3$. Der 3.größte Eintrag oben ist jeweils 1, also $Q_n = 2.22. \cdot 1.$

- Bestimmen Sie Erwartungswert und Varianz von $\overline{X}$.
- Bestimmen Sie Erwartungswert der getrimmten und winsorisierten Mittel
- Zeigen Sie die Erwartungstreue von s^2, d.h. $\mathbf{E}(s^2) = \sigma^2$ (vgl. Stochastik-Vorlesung)
- Berechnen Sie Schiefe und Kurtosis für die Beispiele aus der Vorlesung
- Zeichnen Sie Boxplots (skeletal und schematic) für die fiktiven Daten aus der Übung

- Geben Sie eine theoretische Begründung für die Boxplot-Variante bei der die fences an den Stellen $x_{0.25} - 3 \cdot IR$ und $x_{0.75} + 3 \cdot IR$ definiert sind (anstelle an den Stellen $x_{0.25} - 1.5 \cdot IR$ und $x_{0.75} + 1.5 \cdot IR$ wie in der Vorlesung).
  Hinweise: $\Phi^{-1}(0.9995) = 3.29$ und $3.29 \cdot 0.7434 \approx 2.5$.
- Pferdetritt-Daten
  Wenn Poisson richtig ist dann $\mathbf{E}(X) = \text{var}(X) = \lambda$ und die beiden Schätzungen $\overline{X} = 0.7$ und S^2 für Erwartungswert und Varianz müssten etwa gleich sein.
  Wenn nicht, so kein Poisson oder es liegen Ausreißer vor.

Fisher-Informationen

Bestimmen Sie die Fisher-Information $I(f, \theta)$ in folgenden Modellen

- Exponential, Parameter $\theta = \lambda$, vgl. Vorlesung, Folie 177.
- Doppelexponential, Parameter $\theta = \mu$, vgl. Vorlesung, Folie 178.
- Binomial, Parameter $\theta = p$, $I(f, p) = \frac{1}{p(1-p)}$

Ist das arithmetische Mittel $\overline{X}$ jeweils optimal?

Zeigen Sie (unter der Annahme, dass alle Terme wohldefiniert sind)

$$\begin{aligned} \mathbf{E}\left(\frac{\partial \log f(X_i, \theta)}{\partial \theta}\right) &= 0 \quad \text{und} \\ \operatorname{var}\left(\frac{\partial \log f(X_i, \theta)}{\partial \theta}\right) &= I(f, \theta) \end{aligned}$$

Wartezeiten zwischen zwei Ausbrüchen des Old Faithful Geyser in Yellowstone

Die Datei geyser.dat enthält die Wartezeiten zwischen 300 Ausbrüchen des Old Faithful Geyser, in min.

Laden Sie sich die Datei geyser.dat aus dem Verzeichnis Kursdaten (Vgl. Webseite) herunter und führen Sie nichtparametrische Dichteschätzungen durch. Probieren Sie verschiedene Glättungsparameter.

aus Berliner Zeitung, Juni 2009

Von 200 Personen mit Übergewicht nahmen 100 Personen 9g Salz pro Tag zu sich, und 100 nur 3g Salz pro Tag. Die jeweilige Anzahl der Herzinfarkt- oder Schlaganfall-Toten finden Sie in der Tabelle unten. Bestimmen Sie Abhängigkeitsmaße und Odds Ratio OR.

Salz	Tote	Lebende	
9g	14	86	100
3g	10	90	100
	24	176	200

$$OR = \frac{\frac{14}{86}}{\frac{10}{90}} = \frac{14 \cdot 90}{86 \cdot 10} = \frac{63}{43}$$

$$\Phi^2 = \frac{(14 \cdot 90 - 10 \cdot 86)^2}{24 \cdot 176 \cdot 100^2} = \frac{1}{264}$$

$$\frac{\chi^2}{200} = \frac{(14 - \frac{24 \cdot 100}{200})^2}{24 \cdot 100} + \frac{(86 - \frac{176 \cdot 100}{200})^2}{176 \cdot 100} + \frac{(10 - \frac{24 \cdot 100}{200})^2}{24 \cdot 100} + \frac{(90 - \frac{176 \cdot 100}{200})^2}{176 \cdot 100}$$

$$\chi^2 = \frac{1}{3} + \frac{1}{22} + \frac{1}{3} + \frac{1}{22} = \frac{25}{33}$$

(bitte nachrechnen)

Einfluss von Risikofaktoren auf Krankheiten

Von 760 Personen wurden 120 einem Risiko ausgesetzt, z.B. radioaktiver Strahlung. Von diesen wurden 24 krank, wärend von den übrigen 640 Personen 48 krank wurden. Bestimmen Sie Abhängigkeitsmaße und Odds Ratio OR.

Risiko	krank	gesund	
exponiert	24	96	120
nicht exponiert	48	592	640
	72	144	720

$$OR = \frac{\text{Odds(exponiert)}}{\text{Odds(nicht exponiert)}} = \frac{\frac{24}{96}}{\frac{48}{592}} = 3.083 = \frac{\frac{24}{48}}{\frac{96}{592}} = \frac{\text{Odds(krank)}}{\text{Odds(gesund)}}$$

$\chi^2 = 18.387$ bitte von Hand nachrechnen

Korrelationskoeffizienten

Berechnen Sie empirische Pearson-, Spearman- und Kendall Korrelationskoeffizienten der Variablen (X, Y)

X	-2	-1	0	2
Y	0	1	3	1

Pearson: 0.349

Spearman: 0.632

Kendall: 0.548

bitte von Hand nachrechnen

Führen Sie mit den gegebenen Daten eine lineare Regression durch.

Spearman-Korrelationskoeffizient bei $S_i = n + 1 - R_i \quad \forall i$

Zeigen Sie: $r_S = -1$. (vgl. Folie 270)

Einfaches lineares Regressionsmodell

$$Y_i = \alpha + \beta X_i + \epsilon_i, \quad \epsilon_i \sim (0, \sigma^2)$$

Fall a.: X_i fest, nicht zufällig.

Fall b. X_i vorgegeben, aber fehlerbehaftet gemessen.

Modifikation am Modell

$$\begin{aligned} Y_i &= \alpha + \beta(X_i + \eta_i) + \epsilon_i, \quad \epsilon_i \sim (0, \sigma_\epsilon^2), \eta_i \sim (0, \sigma_\eta^2) \\ &= \alpha + \beta X_i + (\epsilon_i + \beta \eta_i) \end{aligned}$$

X_i, ϵ_i, η_i unkorreliert. $\Rightarrow$ Schätzungen für α und β sind dieselben.

Fall c. X_i und Y_i vertauschbar.

verschiedene Modelle:

$$\begin{aligned} Y_i &= \alpha + \beta X_i + \epsilon_i \\ X_i &= \gamma + \delta Y_i + \kappa_i \end{aligned}$$

Diese Modelle unterscheiden sich!

Kleinste Quadrat-Schätzungen:

$$\begin{aligned} \hat{\beta} &= \frac{s_{XY}}{s_{XX}} \\ \hat{\delta} &= \frac{s_{XY}}{s_{YY}} \end{aligned}$$

und für die anderen Parameter auch i.A. unterschiedliche Schätzungen.

Betrachten den Parameterraum $\Theta \subseteq \mathbb{R}$.

Im zweiseitigen Fall ist $H_0 : \mu = \mu_0, \quad H_A : \mu \neq \mu_0. \quad \Theta = \mathbb{R}.$

Im einseitigen Fall gibt es folgende Varianten

$H_0 : \mu \leq \mu_0, \quad H_A : \mu > \mu_0, \quad \Theta = \mathbb{R}$

$H_0 : \mu = \mu_0$ und $H_A : \mu > \mu_0. \quad \Theta = \{\mu : \mu \geq \mu_0\}$ (ÜA 16)

Beide Fälle werden gleich behandelt.

Wenn t die Realisierung von T (Prüfgröße des t-Tests) ist und $t < 0$ dann wird H_0 ohnehin nicht abgelehnt, der p-Wert ist größer als 0.5.

Der Fall

$H_0 : \mu \geq \mu_0$ und $H_A : \mu < \mu_0$ oder

$H_0 : \mu = \mu_0$ und $H_A : \mu < \mu_0$

ist analog.

- (Einstichprobenfall). Es gilt: Null oder μ_0 im Konfidenzintervall $\Leftrightarrow$ H_0 ($\mu = 0$ oder $\mu = \mu_0$) nicht abgelehnt.
- (Konfidenzintervall- Toleranzintervall)

  Ein $(1 - \alpha)$ Konfidenzintervall ist ein Intervall, in dem ein unbekannter Parameter (z.B. der Erwartungswert μ) mit Wkt. $(1 - \alpha)$ liegt.
  $\alpha = 0.05$ führt also zu einem 95% Konfidenzintervall. α heißt Signifikanzniveau.

  Ein $(1 - \alpha)$ Toleranzintervall ist ein Intervall, in dem eine künftige Beobachtung mit Wkt. $(1 - \alpha)$ liegen wird (dazu müssen aber die Parameter bekannt sein).

T: Prüfgröße (Zufallsvariable)

t: beobachteter Wert der Prüfgröße (der Wert den man erhält wenn man die Werte der Beobachtungen einsetzt)

Der p-Wert ist eine Überschreitungswahrscheinlichkeit,

$P(T > t)$ oder

$P(|T| > t)$ (beim Ein- und Zweistichprobenfall in SAS Standard)

Der p-Wert ist keine Nullhypothesenwahrscheinlichkeit

Kleine p-Werte indizieren Abweichung von H_0

große p-Werte: nichts gegen H_0 einzuwenden.

t-Statistik bei $X_i \sim \mathcal{N}$

Wenn die Beobachtungen normalverteilt sind, dann sind Zähler und Nenner der t-Statistik unabhängig!

Um das zu sehen vgl. Sie die Faktorisierung der Likelihood-Funktion auf Folie 189 im Kapitel Schätzmethoden.

t-Test-Datei

Wir haben $n = 10, \quad \overline{X} = 1.67, \quad s = 1.1304, \quad \alpha = 0.05.$

$(1 - \alpha)$-Konfidenzintervall (wenn Normalverteilung vorliegen würde)

$$\overline{X} \pm \frac{s}{\sqrt{n}} t_{1-\frac{\alpha}{2}, n-1} : \quad [0.8613, 2.4787]$$

Verallgemeinerung von Aufgabe 21, Werte der t-Teststatistik

$x_1 = \cdots = x_m = c = y_1 = \ldots = y_{n-1}; \quad y_n = y \neq c$

$$
\begin{aligned}
\overline{X} &= c, \quad S_X^2 = 0, \quad \overline{Y} = \frac{(n-1)\cdot c + y}{n}, \quad s_Y^2 = \frac{(c-y)^2}{n} : \\
(n-1)s_Y^2 &= (n-1)(c - \frac{(n-1)\cdot c + y}{n})^2 + (y - \frac{(n-1)\cdot c + y}{n})^2 \\
&= (n-1)\frac{(c-y)^2}{n^2} + (\frac{n-1}{n}(y-c))^2 \\
&= (n-1)(\frac{(c-y)^2}{n^2} + \frac{(n-1)^2}{n^2}(c-y)^2) = (n-1)\frac{(c-y)^2}{n} \\
s^2 &= \frac{(m-1)s_X^2 + (n-1)s_Y^2}{n+m-2} = \frac{(n-1)(c-y)^2}{n\cdot(n+m-2)} \\
t &= \frac{\overline{X} - \overline{Y}}{s} = \frac{n\cdot c - (n-1)c - y}{n} \cdot \frac{\sqrt{n(n+m-2)}}{\sqrt{n-1}(c-y)} \\
&= \sqrt{\frac{n+m-2}{n(n-1)}} \qquad \textit{d.h. } c \textit{ und } y \textit{ spielen keine Rolle}
\end{aligned}
$$

Aufgabe 21b, Werte der t-Teststatistik

$x_1 = 1, x_2 = 2, x_3 = 3, y_1 = y_2 = y_3 = 0; \quad m = n = 3,$

$\overline{X} = 2, \overline{Y} = 0, s_X^2 = \frac{1}{2} \cdot (1 + 0 + 1) = 1, s_Y^2 = 0.$

$$t = \frac{2}{\sqrt{\frac{2}{3}}\sqrt{\frac{2S_X^2 + 2S_Y^2}{4}}} = \sqrt{3} \cdot 2 > 2.132 = t_{0.95,4}$$

d.h. H_0 (Lage beider Populationen ist gleich) wird abgelehnt, was ja auch zu erwarten war.

Nehmen wir jetzt eine zusätzliche Beobachtung hinzu, $x_4 = 10$ dann

$\overline{X} = 4, \overline{Y} = 0, s_X^2 = \frac{1}{3} \cdot (9 + 4 + 1 + 36) = \frac{50}{3}, s_Y^2 = 0, m = 4, n = 3.$

$$t = \frac{4}{\sqrt{\frac{1}{4} + \frac{1}{3}}\sqrt{\frac{3S_X^2 + 2S_Y^2}{5}}} = \frac{\sqrt{5}\sqrt{12} \cdot 4}{\sqrt{7}\sqrt{50}} \approx 1.656 < 2.015 = t_{0.95,5}$$

d.h. H_0 (Lage beider Populationen ist gleich) wird NICHT abgelehnt, was aber nicht zu erwarten war.

Vergleich von 2 Vorhersagealgorithmen

Dazu betrachten wir 2 Testmengen, die Beobachtungen $X_1, \ldots, X_n$, $X_i = 1$ falls Vorhersage richtig, $X_i = 0$ sonst, $X_i \sim Bi(1, p_1)$ für Algorithmus 1, die Beobachtungen $Y_1, \ldots, Y_n$, $Y_i = 1$ falls Vorhersage richtig, $Y_i = 0$ sonst, $Y_i \sim Bi(1, p_2)$ für Algorithmus 2.

- Schätzen die Wahrscheinlichkeit einer korrekten Vorhersage durch $\hat{p}_1 = \frac{\sum X_i}{n}$ bzw. $\hat{p}_2 = \frac{\sum Y_i}{m}$.
- Seien n_1 und n_2 die Anzahl der Einsen bzw. Nullen von den X_i.

$$
\begin{aligned}
S_X^2 &= \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2 = \frac{1}{n-1}\Big(n_1(1-\frac{n_1}{n})^2 + n_2(0-\frac{n_1}{n})^2\Big) \\
&= \frac{1}{n-1}\Big(\frac{n_1 n_2^2}{n^2} + \frac{n_1^2 n_2}{n^2}\Big) = \frac{n}{n-1}\frac{n_1 n_2}{n^2} = \frac{n}{n-1}\hat{p}_1(1-\hat{p}_1) \\
S_Y^2 &= \frac{m}{m-1}\hat{p}_2(1-\hat{p}_2) \qquad \text{analog}
\end{aligned}
$$

Vergleich von 2 Vorhersagealgorithmen, Fortsetzung

- Testen $H_0 : p_1 = p_2$ gegen $H_1 : p_1 \neq p_2$ mit dem t-Test für gleiche Varianzen (unter H_0 sind die Varianzen ja gleich)

$$
\begin{aligned}
t &= \frac{\overline{X} - \overline{Y}}{\sqrt{\frac{1}{n} + \frac{1}{m}} \sqrt{\frac{(n-1)S_X^2 + (m-1)S_Y^2}{n+m-2}}} \\
&= \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\frac{1}{n} + \frac{1}{m}} \sqrt{\frac{n\hat{p}_1(1-\hat{p}_1) + m\hat{p}_2(1-\hat{p}_2)}{n+m-2}}} \\
&= \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\frac{n+m}{n+m-2}} \sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{m} + \frac{\hat{p}_2(1-\hat{p}_2)}{n}}} \approx \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\hat{\text{var}}\hat{p}_1 + \hat{\text{var}}\hat{p}_2}}
\end{aligned}
$$

$\approx$ falls m=n. $\quad \text{var}\hat{p}_1 + \text{var}\hat{p}_2 = \text{var}(\hat{p}_1 - \hat{p}_2)$

Vergleich von 2 Vorhersagealgorithmen, Fortsetzung

Die t-Verteilung (mit m+n-2 Freiheitsgraden) gilt nur approximativ, da keine Normalverteilung vorliegt.

Es stellt sich die Frage ob man hier nicht lieber einen χ^2-Unabhängigkeitstest oder Fisher's exakten Test durchführt.

Dazu siehe unten.

Einseitiger Binomialtest, Beispiel 1..3.3.7 aus Reiß et al.

Bei 55-65 jährigen Arbeitern eines Atomkraftwerks gab es 13 Todesfälle von denen 5 auf einen Tumor zurückzuführen waren. Im Jahr 1995 war der Anteil der Todesfälle durch Tumor in Deutschland $\frac{1}{5}$. Frage: Ist die Häufung von 5 Todesfällen von 13 signifikant?

X: zufällige Anzahl der Tumortoten von 13, $X \sim Bi(13, p)$.

Testproblem: $H_0 : p \leq \frac{1}{5}$ gegen $H_1 : p > \frac{1}{5}$.

Wir führen einen Binomialtest durch:

$$P_{\frac{1}{5}}(X \leq 4) = \sum_{i=0}^{4} \binom{13}{i} p^i (1-p)^{13-i} = \sum_{i=0}^{4} \binom{13}{i} (\frac{1}{5})^i (\frac{4}{5})^{13-i} = 0.901$$

also: $P_{\frac{1}{5}}(X \geq 5) = 0.099$ und H_0 wird bei $\alpha = 0.05$ nicht abgelehnt.

Betrachten wir die F-Statistik aus der einfaktoriellen Varianzanalyse

$$F = \frac{MSB}{MSE} = \frac{N-k}{k-1}\frac{SSB}{SSW} = \frac{N-k}{k-1}\frac{\sum_{i=1}^{k} n_i(\overline{Y}_i - \overline{Y})^2}{\sum_{i=1}^{k}\sum_{j=1}^{n_i}(Y_{ij} - \overline{Y}_i)^2}$$

und setzen die Anzahl der Behandlungen auf $k = 2$. Dann wird mit $\overline{Y} = \frac{1}{N}(n_1\overline{Y}_1 + n_2\overline{Y}_2)$, $N = n_1 + n_2$

$$\begin{aligned} F &= \frac{N-2}{2-1}\left(\frac{n_1(\overline{Y}_1 - \frac{1}{N}(n_1\overline{Y}_1 + n_2\overline{Y}_2))^2}{(n_1-1)s_1^2 + (n_2-1)s_2^2} + \right. \\ &\quad \left. \frac{n_2(\overline{Y}_2 - \frac{1}{N}(n_1\overline{Y}_1 + n_2\overline{Y}_2))^2}{(n_1-1)s_1^2 + (n_2-1)s_2^2}\right) \end{aligned}$$

Die Zähler in den rechten Brüchen sind zusammen gleich

$$\begin{aligned} &= n_1((1-\frac{n_1}{N})\overline{Y}_1 - \frac{n_2}{N}\overline{Y}_2)^2 + n_2(-\frac{n_1}{N}\overline{Y}_1 + (1-\frac{n_2}{N})\overline{Y}_2)^2) \\ &= \frac{n_1 n_2^2}{N^2}(\overline{Y}_1 - \overline{Y}_2)^2 + \frac{n_1^2 n_2}{N^2}(\overline{Y}_1 - \overline{Y}_2)^2 \\ &= \frac{n_1 n_2}{N}(\overline{Y}_1 - \overline{Y}_2)^2 \end{aligned}$$

Also:

$$\begin{aligned} F &= \frac{(N-2)n_1 n_2}{N} \frac{(\overline{Y}_1 - \overline{Y}_2)^2}{(n_1-1)s_1^2 + (n_2-1)s_2^2} \\ &= \frac{1}{\frac{1}{n_1}+\frac{1}{n_2}} \frac{(\overline{Y}_1 - \overline{Y}_2)^2}{\frac{(n_1-1)s_1^2+(n_2-1)s_2^2}{N-2}} \end{aligned}$$

Vergleichen Sie das mit der t-Statistik (gemeinsame Varianzen)

$$T = \frac{\overline{X}_1 - \overline{X}_2}{\sqrt{\frac{1}{n}+\frac{1}{m}}\sqrt{\frac{(n-1)S_1^2+(m-1)S_2^2}{n+m-2}}}$$

d.h. F ist dassselbe wie das Quadrat von T.

Bem.: Wir wissen

$$t = \frac{\mathcal{N}(0,1)}{\sqrt{\frac{1}{n+m-2}\chi^2_{n+m-2}}}$$

(so ist die t-Verteilung definiert)

und es ist

$$t^2 = \frac{\mathcal{N}(0,1)^2}{\frac{1}{n+m-2}\chi^2_{n+m-2}} = \frac{\frac{1}{1}\chi^2_1}{\frac{1}{n+m-2}\chi^2_{n+m-2}} = F_{1,n+m-2}$$

(so ist die F-Verteilung bei der der 1. Freiheitsgrad 1 ist, definiert).

Allgemein (Zähler und Nenner seien unabhängig)

$$F = \frac{\frac{1}{df1}\chi^2_{df1}}{\frac{1}{df2}\chi^2_{df2}} \sim F_{df1,df2}$$

Bei der Varianzanalyse haben wir Quotienten von mittleren Quadratsummen.

Wenn Normalverteilung vorliegt (vorliegen sollte) so

- Zähler und Nenner sind unabhängig (siehe folgende Folie)
- Die Quadratsummen sind χ^2-verteilt.

  Die Quadratsummen können als Quadratische Formen $\mathbf{x}'\mathbf{A}\mathbf{x}$ geschrieben werden. Die Freiheitsgrade sind gleich dem Rang der Matrix $\mathbf{A}$. Dazu später.

F-Statistik der Varianzanalyse bei $X_i \sim \mathcal{N}$

Wenn die Beobachtungen normalverteilt sind, und die Nullhypothese (keine Effekte) zutrifft, dann sind Zähler und Nenner der F-Statistik

$$F = \frac{MSB}{MSE} = \frac{\frac{SSB}{k-1}}{\frac{SSW}{N-I}}$$

unabhängig!

Analog zur t-Statistik vergleichen Sie dazu die Faktorisierung der Likelihood-Funktion (gemeinsame Dichte aller Beobachtungen) auf Folie 189 im Kapitel Schätzmethoden. Wegen der Summeneigenschaft der Quadratsummen $SST = SSB + SSW$ faktorisiert sich die Dichte weiter, und wir bekommen 3 Faktoren die jeweils nur von den (suffizienten) Statistiken $\overline{X}$ (Gesamtmittel) und SSB und SSW abhängen.

Dopamin Aktivität von Schizophrenie-Patienten

Der Einfachheit halber betrachten wir nur die jeweils ersten drei Patienten der Gesunden- und Krankheitsgruppe (die erste Beobachtung in der Krankengruppe wurde hier leicht modifiziert, damit wir hier mit der Hand leichter rechnen können, vgl. Datei Dopamin.sas)

Gruppe	Beobachtungen			$\overline{X}$
0:	0.0104	0.0105	0.0112	0.0107
1:	0.0155	0.0204	0.0208	0.0189
	$\lvert X_i - \overline{X}\rvert$			
0:	0.0003	0.0002	0.0005	
1:	0.0034	0.0015	0.0019	
	$\lvert X_i - medX_i\rvert$			
0:	0.0001	0.0	0.0007	
1:	0.0049	0.0	0.0004	

Rechnen Sie mit den modifizierten Beobachtungen jeweils t-Tests und einfaktorielle Varianzanalysen.

Lassen sich evtl. Skalenunterschiede zwischen beiden Gruppen feststellen?

Lassen Sie Programme laufen

- Levene und Brown-Forsythe Test
- t-Test mit den modifizierten Beobachtungen (2. bzw. 3. Block in der vorigen Tabelle)
- Varianzanalyse mit den modifizierten Beobachtungen

Was stellen Sie fest?

Sei $\mathbf{B}$ eine $n \times n$ Matrix und $\mathbf{x} = (x_1, \ldots, x_n)$.

Ein Ausdruck der Form $\mathbf{x}'\mathbf{B}\mathbf{x}$ heißt Quadratische Form.

1. Quadratsummen als Quadratische Formen

$$\sum_{i=1}^{n}(X_i - \overline{X})^2 = \mathbf{X}'\mathbf{A}\mathbf{X}$$

wobei $\mathbf{X}' = (X_1, \ldots, X_n)$ Beobachtungsvektor.

Bezeichen wir mit $\mathbf{I}$ die Einheitsmatrix und mit $\mathbf{1}$ die Matrix die nur aus Einsen besteht, dann können wir schreiben (bitte nachrechnen)

$$\mathbf{A} = \mathbf{I} - \frac{1}{n}\mathbf{1}.$$

Finden Sie entsprechende Quadratische Formen für die Quadratsummen SSB und SSW aus der einfaktoriellen Varianzanalyse!

2. Eigenschaften der Matrix $\mathbf{A}$

- $\mathbf{A}$ ist symmetrisch
- $\mathbf{A}$ ist positiv semidefinit, $\mathbf{x}'\mathbf{A}\mathbf{x} \geq 0$ folgt aus 1.
- $\mathbf{A}$ ist idempotent, d.h. $\mathbf{A}^2 = \mathbf{A}$

$$\mathbf{A}^2 = (\mathbf{I} - \frac{1}{n}\mathbf{1})'(\mathbf{I} - \frac{1}{n}\mathbf{1}) = \mathbf{I} - \frac{2}{n}\mathbf{1} + \frac{1}{n^2}\underbrace{\mathbf{1}'\mathbf{1}}_{n\mathbf{1}} = \mathbf{I} - \frac{1}{n}\mathbf{1} = \mathbf{A}$$

Erinnerung: Eigenwerte und Eigenvektoren

$\lambda \in \mathbb{C}$ heißt Eigenwert einer Matrix $\mathbf{B}$, falls $\exists \mathbf{x}$ so dass $\mathbf{Bx} = \lambda \cdot \mathbf{x}$.

$\mathbf{x}$ heißt zum Eigenwert λ gehöriger Eigenvektor.

3. Eigenwerte von $\mathbf{A} = \mathbf{I} - \frac{1}{n}\mathbf{1}$

- alle Eigenwerte von $\mathbf{A}$ sind reell, da $\mathbf{A}$ symmetrisch
- alle Eigenwerte sind ≥ 0 da $\mathbf{A}$ positiv semidefinit,

  $$0 \leq \mathbf{x}'\mathbf{Ax} = \mathbf{x}'\lambda\mathbf{x} = \lambda \underbrace{\mathbf{x}'\mathbf{x}}_{\geq 0}$$
- alle Eigenwerte sind 0 oder 1, da $\mathbf{A}$ idempotent.

4. Alle Eigenwerte einer idempotenten Matrix $\mathbf{B}$ sind 0 oder 1

Das folgt aus der Spektralzerlegung (Hauptachsentransformation)

$$\mathbf{B} = \mathbf{U}\mathbf{\Lambda}\mathbf{U}'$$

wobei $\mathbf{\Lambda}$ Diagonalmatrix mit Eigenwerten in den Diagonal-Einträgen und $\mathbf{U}$ Orthogonalmatrix mit den entsprechenden Eigenvektoren (die Eigenvektoren sind die Spalten von $\mathbf{U}$)

$$\mathbf{B}^2 = (\mathbf{U}\mathbf{\Lambda}\underbrace{\mathbf{U}'(\mathbf{U}}_{\mathbf{I}}\mathbf{\Lambda}\mathbf{U}') = \mathbf{U}\mathbf{\Lambda}^2\mathbf{U}'$$

Wenn wir die Gleichung $\mathbf{B} = \mathbf{B}^2$ von links mit $\mathbf{U}'$ und von rechts mit $\mathbf{U}$ multiplizieren, erhalten wir wegen $\mathbf{U}'\mathbf{U} = \mathbf{I}$

$$\mathbf{\Lambda} = \mathbf{U}'\mathbf{U}\mathbf{\Lambda}\mathbf{U}'\mathbf{U} = \mathbf{U}'\mathbf{B}\mathbf{U} = \mathbf{U}'\mathbf{B}^2\mathbf{U} = \mathbf{U}'\mathbf{U}\mathbf{\Lambda}^2\mathbf{U}'\mathbf{U} = \mathbf{\Lambda}^2$$

also $\mathbf{\Lambda}^2 = \mathbf{\Lambda}$. Da $\mathbf{\Lambda}$ Diagonalmatrix sind alle Diagonaleinträge (also die Eigenwerte) Null oder Eins.

Erinnerung: Rang einer Matrix

Der Rang $rg(\mathbf{B})$ einer Matrix $\mathbf{B}$ ist die maximale Anzahl linear unabhängiger Spalten. Die Vektoren $\mathbf{x}_1, \ldots, \mathbf{x}_k$ heißen linear unabhängig, falls aus $\sum_{i=1}^{k} \mu_i \mathbf{x}_i = 0$ folgt: $\mu_1 = \ldots = \mu_k = 0$.

5. Der Rang von $\mathbf{A} = \mathbf{I} - \frac{1}{n}\mathbf{1}$ ist $n - 1$.

- $rg(\mathbf{A}) \leq n - 1$. Wenn wir alle Zeilen aufsummieren erhalten wir den Nullvektor, d.h. die Spalten sind nicht linear unabhängig.
- $rg(\mathbf{A}) = n - 1$. Sei $\widetilde{\mathbf{A}}$ die Matrix $\mathbf{A}$ ohne letzte Zeile und Spalte und $\mathbf{x}_1, \ldots, \mathbf{x}_{n-1}$ die Spaltenvektoren von $\widetilde{\mathbf{A}}$. Dann folgt aus

$$0 = \sum_{i=1}^{n-1} \mu_i \mathbf{x}_i = \mu_1 \begin{pmatrix} 1 - \frac{1}{n} \\ -\frac{1}{n} \\ \dots \\ -\frac{1}{n} \end{pmatrix} + \mu_2 \begin{pmatrix} -\frac{1}{n} \\ 1 - \frac{1}{n} \\ \dots \\ -\frac{1}{n} \end{pmatrix} + \dots + \mu_{n-1} \begin{pmatrix} -\frac{1}{n} \\ -\frac{1}{n} \\ \dots \\ 1 - \frac{1}{n} \end{pmatrix}$$

dass $\mu_i = \sum_{j=1}^{n-1} \mu_j \quad \forall i = 1, \dots, n-1$, d.h. alle Koeffizienten μ_i sind gleich, sagen wir μ, und $\mu = \sum_{j=1}^{n-1} \mu = (n-1)\mu$. Daraus folgt: $\mu = 0$. D.h. $n-1$ Spalten sind linear unabhängig. Da der Rang einer $(n \times n)$ Matrix gleich der Anzahl der von Null verschiedenen Eigenwerte ist folgt:

6. **A** hat einfachen Eigenwert Null und $n-1$ fachen Eigenwert Eins.

$rg(A) = n - 1$

Dieses Resultat können Sie auch wie folgt erhalten.

Wenn Sie in der Matrix **A** von den ersten $n-1$ Zeilen jeweils die letzte abziehen, und anschließend die letzte Zeile mit n multiplizieren und die letzte Zeile dann durch die Summe aller Zeilen ersetzen, dann erhalten Sie

$$rg\begin{pmatrix} 1-\frac{1}{n} & -\frac{1}{n} & \cdots & -\frac{1}{n} \\ -\frac{1}{n} & 1-\frac{1}{n} & \cdots & -\frac{1}{n} \\ \cdots & & & \\ -\frac{1}{n} & -\frac{1}{n} & \cdots & 1-\frac{1}{n} \end{pmatrix} = rg\begin{pmatrix} 1 & 0 & \cdots & -1 \\ 0 & 1 & \cdots & -1 \\ & & \cdots & \\ -\frac{1}{n} & -\frac{1}{n} & \cdots & 1-\frac{1}{n} \end{pmatrix}$$

$$= rg\begin{pmatrix} \mathbf{I}_{n-1} & \mathbf{1}_{n-1} \\ \mathbf{0_{n-1}} & 0 \end{pmatrix} = n-1$$

7. $\mathbf{A} = \mathbf{U}\Lambda\mathbf{U}' = \mathbf{T}\mathbf{T}'$ wobei $\mathbf{T}$ gleich $\mathbf{U}$ ohne letzte Spalte

Seien $\mathbf{e}_1, \ldots, \mathbf{e}_{n-1}$ Eigenvektoren zum Eigenwert Eins, und $\mathbf{e}_n$ Eigenvektor zum Eigenwert Null, $\mathbf{U} = (\mathbf{e}_1, \ldots, \mathbf{e}_{n-1}, \mathbf{e}_n)$.

Da die Eigenvektoren orthogonal sind, gilt $\mathbf{U}'\mathbf{U} = \mathbf{I}$. Weiterhin

$$\mathbf{U}\mathbf{U}' = \begin{pmatrix} e_{11} & \ldots & e_{1n} \\ \ldots & & \\ e_{n1} & \ldots & e_{nn} \end{pmatrix} \begin{pmatrix} e_{11} & \ldots & e_{n1} \\ \ldots & & \\ e_{1n} & \ldots & e_{nn} \end{pmatrix} = \left(\sum_{j=1}^{n} e_{ij} e_{kj} \right)_{i,k=1,\ldots,n}$$

$$\sum_{i=1}^{n} \mathbf{e}_i \mathbf{e}_i' = \sum_{i=1}^{n} \begin{pmatrix} e_{1i} \\ \ldots \\ e_{ni} \end{pmatrix} (e_{1i}, \ldots, e_{ni}) = \sum_{i=1}^{n} \begin{pmatrix} e_{1i}e_{1i} & \ldots & e_{1i}e_{ni} \\ \ldots & & \\ e_{ni}e_{1i} & \ldots & e_{ni}e_{ni} \end{pmatrix}$$

und die letzte Summe ist gleich

$$\begin{pmatrix} \sum_{i=1}^n e_{1i}e_{1i} & \dots & \sum_{i=1}^n e_{1i}e_{ni} \\ \dots & & \\ \sum_{i=1}^n e_{ni}e_{1i} & \dots & \sum_{i=1}^n e_{ni}e_{ni} \end{pmatrix} = \mathbf{U}\mathbf{U}'$$

Da $\lambda_1 = \cdots \lambda_{n-1} = 1$ und $\lambda_n = 0$ folgt

$$\mathbf{A} = \mathbf{U}\mathbf{\Lambda}\mathbf{U}' = \sum_{i=1}^{n} \lambda_i \mathbf{e}_i \mathbf{e}_i' = \sum_{i=1}^{n-1} \mathbf{e}_i \mathbf{e}_i' = \mathbf{T}\mathbf{T}'$$

$\mathbf{T}$ ist also eine $(n \times n-1)$ Matrix.

$$\mathbf{X}'\mathbf{A}\mathbf{X} = \mathbf{X}'\mathbf{T}\mathbf{T}'\mathbf{X} = (\mathbf{T}'\mathbf{X})'(\mathbf{T}'\mathbf{X}), \quad \mathbf{T}'\mathbf{X} = \left(\sum_{j=1}^n t_{ij}X_j\right)_{i=1,\dots,n-1}$$

Da $\mathbf{U}$ und $\mathbf{T}$ orthogonal: $\mathbf{U}'\mathbf{U} = \mathbf{I}_n$ und $\mathbf{T}'\mathbf{T} = \mathbf{I}_{n-1}$.

8. Seien jetzt $X_1, \ldots, X_n$ unabhängige Zufallsvariablen

Seien $\mathbf{X} = (X_1, \ldots, X_n)$ und $\boldsymbol{\mu} = \mathbf{E}(\mathbf{X}) = (\mu_1, \ldots, \mu_n)$.

$$\mathbf{E}(\mathbf{T}'\mathbf{X}) = \mathbf{E}\Big(\sum_{j=1}^{n} t_{ij} X_j\Big)_{i=1,\ldots,n} = \Big(\sum_{j=1}^{n} t_{ij}\mathbf{E}X_j\Big)_{i=1,\ldots,n} = \mathbf{T}'\boldsymbol{\mu}$$

$$\mathrm{var}(\mathbf{T}'\mathbf{X}) = \left(\sum_{j=1}^{n} t_{ji}\, t_{jk} \mathrm{var} X_j\right)_{i.k=1,\ldots,n} = \mathbf{T}'\mathbf{T} \qquad \text{falls} \qquad \mathrm{var}X_j = 1$$

9. Seien nun zusätzlich $\mathbf{X} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$

$$\mathbf{T}'\mathbf{X} \sim \mathcal{N}(\mathbf{T}'\boldsymbol{\mu}, \mathbf{T}'\boldsymbol{\Sigma}\mathbf{T})$$

Wenn nun $\boldsymbol{\mu} = \mathbf{0}$ und $\boldsymbol{\Sigma} = \mathbf{I}_n$ dann wegen $\mathbf{T}'\mathbf{T} = I_{n-1}$

$$\mathbf{T}'\mathbf{X} \sim \mathcal{N}(\mathbf{0_{n-1}}, \mathbf{I_{n-1}})$$

d.h. die Komponenten Y_i von $\mathbf{Y} = \mathbf{T}'\mathbf{X}$ sind unkorreliert (und wegen Normalverteilung auch unabhängig).

10. Die Quadratsumme $\sum_{i=1}^{n}(X_i - \overline{X})^2 = \mathbf{X'AX}$ ist χ^2_{n-1} verteilt

$$\sum_{i=1}^{n}(X_i - \overline{X})^2 = \mathbf{X'AX} = (\mathbf{T'X})'(\mathbf{T'X}) = \mathbf{Y'Y} = \sum_{i=1}^{n-1} Y_i^2$$

ist die Summe von $n-1$ unabhängigen, standardnormal verteilten Zufallsvariablen.

Diese Summe ist also χ^2 verteilt mit $n-1$ Freiheitsgraden. (So ist die χ^2-Verteilung definiert.)

Verallgemeinerung auf beliebige quadratische Formen

Wenn die $(n \times n)$ Matrix $\mathbf{B}$ idempotent mit $rg(B) = k$ dann hat $\mathbf{B}$ genau k Eigenwerte Eins und $n-k$ Eigenwerte Null.

wenn also $\mathbf{X} \sim \mathcal{N}(\mathbf{0_n}, \mathbf{I_n})$ so

$$\mathbf{X'BX} \sim \chi^2_{rg(\mathbf{B})}$$

Von der χ^2-Verteilung abgeleitete Verteilungen

- Wenn Zähler Z Standardnormal und der Nenner $D \sim \frac{1}{df}\chi^2$ und Zähler und Nenner unabhängig sind, dann ist

$$\frac{Z}{D} \sim t_{df} \qquad (t - \text{verteilt mit df Freiheitsgraden}).$$

  Das kennen Sie von Einstichproben- und Zweistichproben t-Test. Bei der Regressionsanalyse werden wir weitere Anwendungen des t-Tests sehen.

- Wenn Zähler $Z \sim \frac{1}{df_1}\chi^2_{df_1}$ und der Nenner $D \sim \frac{1}{df_2}\chi^2_{df_2}$ und außerdem Zähler und Nenner unabhängig sind, dann ist

$$\frac{Z}{D} \sim F_{df_1,df_2} \qquad (F - \text{verteilt mit}(df_1, df_2) \text{ Freiheitsgraden}).$$

  Solche F-Tests kennen Sie aus der Varianzanalyse.

Daten: 1.5, 2.7, 2.8, 3.0, 3.1

Testen Sie die fiktiven Beispieldaten auf Normalverteilung mit der Hand. Rechnen Sie den Kolmogorov-Smirnov Test.

Vielleicht schaffen Sie es auch die Prüfgrößen des Cramer-von Mises Test oder des Anderson-Darling Tests zu berechnen (vgl. folgende Seiten). Sie können auch selbst ein Programm schreiben und vergleichen mit dem was die Programmpakete ausgeben?

Wir wollen versuchen, die Prüfgröße $\frac{W^2}{n} = \int (F_n(x) - F(x))^2 \, dF(x)$ des Cramer- von Mises Tests explizit auszurechnen. Das nur um zu veranschaulichen, dass die Prüfgrößen explizit ausgerechnet werden können.

$$F_n(x) = \begin{cases} 0 & \text{falls} \quad x < x_{(1)} \\ \frac{i}{n} & \text{falls} \quad x_{(i)} \leq x < x_{(i+1)} \\ 1 & \text{sonst} \end{cases}$$

$$(F_n(x) - F(x))^2 = \begin{cases} F(x)^2 & \text{falls} \quad x < x_{(1)} \\ \left(\frac{i}{n} - F(x)\right)^2 & \text{falls} \quad x_{(i)} \leq x < x_{(i+1)} \\ (1 - F(x))^2 & \text{sonst} \end{cases}$$

$$\frac{W^2}{n} = \int_{-\infty}^{x_{(1)}} F^2(x) \, dF(x) + \sum_{i=1}^{n} \int_{x_{(i)}}^{x_{(i+1)}} (\frac{i}{n} - F(x))^2 \, dF(x) + \int_{x_{(n)}}^{\infty} (1 - F(x))^2 \, dF(x$$

Der 1. und 3. Summand S_1 und S_3 sind einfach auszurechnen.

$$
\begin{aligned}
S_1 &= \frac{1}{3}F^{-1}(x_{(1)}) =: \frac{1}{3}u_{(1)} \\
S_3 &= \frac{1}{3}(1 - F^{-1}(x_{(n)})) =: \frac{1}{3}(1 - u_{(n)}) \\
S_2 &= \sum_{i=0}^{n} \int_{u_{(i)}}^{u_{(i+1)}} (\frac{i}{n} - u)^2 \, du \qquad u_{(0)} := 0, u_{(n+1)} = 1 \\
&= \sum_{i=0}^{n} \frac{1}{3}(u - \frac{i}{n})^3 \Big|_{u_{(i)}}^{u_{(i+1)}} \qquad a_i := u_{(i)} \\
&= \frac{1}{3}\sum_{i=1}^{n} \Big((a_{i+1} - \frac{i}{n})^3 - (a_i - \frac{i}{n})^3 \Big) \\
&= a_{n+1}^3 - a_0^3 + \sum_{i=1}^{n} (a_{i+1}^2 - a_i^2)\frac{i}{n} + \sum_{i=1}^{n} (a_{i+1} - a_i)\Big(\frac{i}{n}\Big)^2 \\
&= 1 + \frac{1}{n}\Big(\sum_{i=1}^{n} i(a_{i+1}^2 - a_i^2) + \sum_{i=1}^{n} i^2(a_{i+1} - a_i) \Big)
\end{aligned}
$$

1. Summe T_1 und 2. Summe T_2:

$$
\begin{aligned}
T_1 &= \sum_{i=1}^{n} i(a_{i+1}^2 - a_i^2) \\
&= (a_2^2 - a_1^2) + (2a_3^2 - 2a_2^2) + (3a_4^2 - 3a_3^2) + \ldots \\
&\quad +((n-1)a_n^2 - (n-1)a_{n-1}^2) + (na_{n+1}^2 - na_n^2) \\
&= -a_1^2 - a_2^2 - \ldots - a_n^2 + na_{n+1}^2 = n - \sum_{i=1}^{n} a_i^2
\end{aligned}
$$

$$
\begin{aligned}
T_2 &= \sum_{i=1}^{n} i^2(a_{i+1} - a_i) \\
&= (a_2 - a_1) + 2^2(a_3 - a_2) + 3^2(a_4 - a_3) + \ldots \\
&\quad +(n-1)^2(a_n - a_{n-1}) + n^2(a_{n+1} - a_n) \\
&= -a_1 + (1^2 - 2^2)a_2 + (2^2 - 3^2)a_3 + \ldots + ((n-1)^2 - n^2)a_n + n^2 \\
&= n^2 - \sum_{i=1}^{n}(i - 1 + i)a_i = n^2 - \sum_{i=1}^{n}(2i-1)a_i
\end{aligned}
$$

Anpassungstest auf Poisson

Die Anzahl der Anrufe in einem Zeitintervall von 15 Minuten sei wie folgt verteilt

Anzahl	0	1	2	3	4	5	6	7	≥ 8
Häufigkeit	1	0	8	10	6	6	7	1	5

Testen Sie ob diese Anzahlen einer Poisson-Verteilung genügen! Wie wählen Sie den Parameter?

Beachten Sie: Wenn Sie eine Parameterschätzung verwenden, dann verringert sich der Freiheitsgrad der entsprechenden χ^2 Verteilung um Eins. Das Programm weis aber nicht, dass Sie Schätzungen verwenden.

In SAS können Sie die Wahrscheinlichkeiten im TESTP Kommando angeben, diese müssen Sie aber vorher ausrechnen (lassen).

bei $\lambda = 3$ wird Poisson abgelehnt, bei $\lambda = \overline{X} = 3.9$ nicht.

bei $\lambda = 3$ keine Parameterschätzung, also 8 Freiheitsgrade

bei $\lambda = 3.9$ wurde λ geschätzt, also 8-1=7 Freiheitsgrade

Wenn Sie SAS nutzen (oder ein anderes Programmpaket dann ist der p-Wert von Hand (mit der Funktion CDF) auszurechnen.

Anpassungstest auf diskrete Gleichverteilung

Wir werfen einen Spielwürfel 20000 mal Die Ergebnisse des Experiments finden Sie in der Tabelle unten

Augenzahl	1	2	3	4	5	6
Häufigkeit	3407	3631	3176	2916	3448	3422

Ist der Würfel fair? Testen Sie ob diese Anzahlen einer diskreten Gleichverteilung genügen!

Probieren Sie es auch mit dem Taschenrechner oder mit Computeralgebra!

- Zeigen Sie für die Vorzeichen-Wilcoxon Rangstatistik W_n^+, vgl. Folie 438: $\mathrm{var} W_n^+ = \frac{n \cdot (n+1)(2n+1)}{24}$.
- Zeigen Sie für die Wilcoxon Teststatistik S_1, vgl. Folie 454, $\mathbf{E} S_1 = \frac{n(n+m+1)}{2}$ und $\mathrm{var}\ S_1 = \frac{n \cdot m \cdot (n+m+1)}{12}$.
- Berechnen Sie die Prüfgröße für den Kruskal-Wallis Test im Fall von zwei unverbundenen Stichproben. Vergleichen Sie mit der Prüfgröße des Wilcoxon Tests.
- Vergleichen Sie die Prüfgröße KW des Kruskal-Wallis Tests mit der der Prüfgröße F der einfaktoriellen Varianzanalyse! Warum gilt $KW \sim \chi^2$ aber $F \sim F$?
- Rechnen Sie den Kruskal-Wallis Test für die Maschinen-Daten (vgl. Beispieldatei `Anova_Maschinen.sas`) mit der Hand nach.

Varianz von $W_n^+ = \sum_{i=1}^n R_i^+ V_i$ unter H_0

$V_i = 1$ falls $X_i - \mu_0 > 0$, $V_i = 0$ sonst. $\quad \mathbf{E}V_i = \frac{1}{2} = \mathbf{E}V_I^2$

$$\begin{aligned}
\mathbf{E}W_n^{+2} &= \mathbf{E}\Big(\sum_{i=1}^n R_i^+ V_i\Big)^2 = \mathbf{E}\Big(\sum_{i=1}^n R_i^{+2} V_i^2\Big) + \mathbf{E}\Big(\sum_{i\neq j}^n R_i^+ R_j^+ V_j\Big)^2 \\
&= \sum_{i=1}^n i^2 \mathbf{E}V_i^2 + \sum_{i\neq j}^n ij\ \mathbf{E}(V_i)\mathbf{E}(V_j) \\
&= \frac{1}{2}\sum_{i=1}^n i^2 + \frac{1}{4}\left(\Big(\sum_{i=1}^n i\Big)^2 - \sum_{i=1}^n i^2\right) \\
&= \frac{n(n+1)(2n+1)}{6}\Big(\frac{1}{2} - \frac{1}{4}\Big) + \frac{1}{4}\Big(\frac{n(n+1)}{2}\Big)^2 \\
&= \frac{n(n+1)(2n+1)}{24} + \Big(\frac{n(n+1)}{4}\Big)^2 \\
\text{var } W_n^+ &= \mathbf{E}W_n^{+2} - (\mathbf{E}W_n^+)^2 = \frac{n(n+1)(2n+1)}{24}
\end{aligned}$$

Seien W und KW die Prüfgrößen des Wilcoxon bzw. des Kruskal-Wall Tests. $T_i = \frac{1}{n_i}\sum_{j=1}^{n_i} R_{ij}$, $i = 1, 2$, $n_1 = n, n_2 = m$, $N = n + m$. $nT_1 + mT_2 = \sum R_{ij} = \sum_{j=1}^{N} j = \frac{N(N+1)}{2}$. Es gilt

$$
\begin{aligned}
W^2 &= \left(\frac{\sum_{j=1}^{n} R_{1j} - \frac{n(N+1)}{2}}{\frac{mn(N+1)}{12}}\right)^2 = \frac{\left(nT_1 - \frac{n(N+1)}{2}\right)^2}{\frac{mn(N+1)}{12}} \\
&= 12\frac{n^2\left(T_1 - \frac{(N+1)}{2}\right)^2}{nm(N+1)} = \frac{12n}{m(N+1)}\left(T_1 - \frac{(N+1)}{2}\right)^2 \\
KW &= \frac{n\left(T_1 - \frac{(N+1)}{2}\right)^2 + m\left(T_2 - \frac{(N+1)}{2}\right)^2}{\frac{N(N+1)}{12}} = \frac{12}{N(N+1)} \cdot \\
&\quad \left(n\left(T_1 - \frac{(N+1)}{2}\right)^2 + m\left(\frac{1}{m}\left(\frac{N(N+1)}{2} - nT_1\right) - \frac{N+1}{2}\right)^2\right)
\end{aligned}
$$

Der letzte große Klammerausdruck (der 2. Summand, ohne Quadrat) wird zu

$$-\frac{n}{m}T_1 + \frac{N+1}{2}(\frac{N}{m} - 1) = \frac{n}{m}(-T_1 + \frac{N+1}{2})$$

und damit

$$\begin{aligned} KW &= \frac{12}{N(N+1)}\Big(T_1 - \frac{(N+1)}{2}\Big)^2(n + m\frac{n^2}{m^2}) \\ &= \frac{12n}{N(N+1)}\Big(T_1 - \frac{(N+1)}{2}\Big)^2(1 + \frac{n}{m}) \\ &= \frac{12n}{m(N+1)}\Big(T_1 - \frac{(N+1)}{2}\Big)^2 = W^2 \end{aligned}$$

Exakter Test

Beim exakten Test werden möglichen Fälle der Aufteilung der Stichproben betrachtet, und jeweils die Prüfgröße ausgerechnet. Damit bekommt man die exakte Wahrscheinlichkeitverteilung. Die Anzahl der Aufteilungen der Stichproben ist beim

- Wilcoxon-Test: $\binom{n+m}{n}$
- Kruskal-Wallis Test: $\binom{N}{n_1}\binom{N-n_1}{n_2}\cdots\binom{n_{k-1}+n_k}{n_k}$ bei k Gruppen.

Stetigkeitskorrektur am Beispiel des Wilcoxon-Tests

Teststatistik des Wilcoxon-Tests:

$$S = \sum_{i=1}^{n} R_{i1}, \quad \text{wobei } R_{i1} \text{ die Ränge die zur ersten Stichprobe gehören}$$

Diese Summe ist ganzzahlig (wenn keine Bindungen vorliegen).

$$\begin{aligned} P(S < s) &= P\Big(\frac{S - \mathbf{E}S}{\sqrt{\text{var}}(S)} < \frac{s - \mathbf{E}S}{\sqrt{\text{var}}(S)}\Big) \approx \Phi\Big(\frac{s - \mathbf{E}S}{\sqrt{\text{var}}(S)}\Big) \\ P(S \leq s) &= P\Big(\frac{S - \mathbf{E}S}{\sqrt{\text{var}}(S)} \leq \frac{s - \mathbf{E}S}{\sqrt{\text{var}}(S)}\Big) \approx \Phi\Big(\frac{s - \mathbf{E}S}{\sqrt{\text{var}}(S)}\Big) \end{aligned}$$

aber i.A.: $P(S < s) \neq P(S \leq s)$. Deshalb Kompromiss:

$$P(S < s + \frac{1}{2}) = P\Big(\frac{S - \mathbf{E}S}{\sqrt{\text{var}}(S)} < \frac{s - \mathbf{E}S + \frac{1}{2}}{\sqrt{\text{var}}(S)}\Big) \approx \Phi\Big(\frac{s - \mathbf{E}S + \frac{1}{2}}{\sqrt{\text{var}}(S)}\Big)$$

EDF Tests im 2-Stichprobenproblem

Ausgabe bei SAS

$$
\begin{aligned}
KS &= \max_j \sqrt{\frac{1}{n}\sum_{i=1}^{k} n_i (F_i(x_j) - F(x_j))^2}, \qquad F = \sum_{i=1}^{2} \frac{1}{n_i} F_i \\
KSa &= KS \cdot \sqrt{n} \qquad \text{asymptotische KS-Statistik} \\
\text{Display} &= \max_j n_i (F_i(x_j) - F(x_j)) \\
D &= \max_j |F_1(x_j) - F_2(x_j)| \qquad \text{Kolmogorov-Statistik} \\
CM &= \text{Cramer- von Mises Statistik wie oben} \\
CM_a &= CM \cdot n \\
K &= \max_j n_i (F_i(x_j) - F(x_j)) - \min_j n_i (F_i(x_j) - F(x_j)) \\
K_a &= K \cdot \sqrt{\frac{n_1 n_2}{n}} \qquad \text{(asymptotische) Kuiper-Teststatistik}
\end{aligned}
$$

Binomialtest, vgl. Vorlesung, Folie 448

Testproblem: $H_0 : p \leq 0.05 \quad$ gegen $\quad H_1 : p > 0.05.$

Stichprobe vom Umfang 20, davon sind 3 schlecht.

X: zufällige Anzahl der schlechten Stücke

exakter p-Wert$= P(X \geq 3) = 1 - P(X \leq 2) =$

$1 - CDF(\text{`}Binomial', 2, 0.05, 20) = 0.0754$

Gegeben seien die (fiktiven) Datenpaare (X_i, Y_i)

X:	-2	-1	0	3
Y:	0	1	3	0

Berechnen Sie de Pearson-Korrelationskoeffizienten,

Spearman-Korrelationskoeffizienten, und den

Kendall-Konkordanzkoeffizienten mit der Hand. Vergleichen Sie:

$$
\begin{aligned}
r_P &= \frac{-1}{\sqrt{4+1+0+9}\sqrt{1+0+4+1}} = \frac{-1}{\sqrt{84}} \\
r_S &= \frac{1.5 \cdot 1 - 0.5 \cdot 0.5 + 0.5 \cdot 1.5 - 1.5 \cdot 1}{\sqrt{1.5^2+0.5^2+0.5^2+1.5^2}\sqrt{1^2+0.5^2+1.5^2+0.5^2}} \\
&= \frac{0.5}{\sqrt{5}\sqrt{3.75}} = \frac{1}{\sqrt{75}} \\
r_K &= \frac{3-2}{6} = \frac{1}{6} \quad \text{(3 Paare konkorant, 2 diskordant, 1 gebunden)}
\end{aligned}
$$

χ^2 Unabhängigkeitstest

Berliner Zeitung, 5.3.2008

Es stehen 2 Schmerzmittel zur Verfügung, ein billiges und ein teures (B bzw, T). In einer Studie bekommen jeweils 41 Patienten eines der beiden Mittel, und in beiden Gruppen wird gezählt wieviel Patienten schmerzfrei wurden. Ergebnisse sind in der untenstehenden Tabelle

Pille	T	B	Gesamt
keine Schmerzen	33	25	58
Schmerzen	8	16	24
	41	41	82

Die Frage ist, besteht eine statistisch signifikanter Unterschied zwischen der Wirkung der beiden Pillen?

Fisher's exakter Test

Wir berechnen den exakten p-Wert, d.h. die Wahrscheinlichkeit, dass $X \geq 33$, wobei X die zufällige Anzahl der Patienten ist, die die teure Pille nehmen und keine Schmerzen haben.

$$P(X \geq 33) = \sum_{r \geq 33} \frac{\binom{41}{r}\binom{41}{58-r}}{\binom{82}{58}} = 1 - CDF('Hyper', 32, 82, 58, 41)$$

32: Argument

82: Umfang der Population

58: Anzahl der Patienten ohne Schmerzen

41: Anzahl der Patienten mit der teuren Pille

Beim χ^2 Unabhängigkeitstest würden wir bekommen:

$$\sum_{i=1}^{2}\sum_{j=1}^{2}\frac{(n_{ij}-n_{i.}n_{.j}/n)^2}{n_{i.}n_{.j}/n} = \frac{(33-58\cdot 41/82)^2}{41\cdot 58/82}+\frac{(25-58\cdot 41/82)^2}{41\cdot 58/82}+$$
$$\frac{(8-41\cdot 24/82)^2}{41\cdot 24/82}+\frac{(16-24\cdot 41/82)^2}{41\cdot 24/82}$$

Berliner Zeitung, Juni 2009, vgl. S.45

In einer Studie wurden 2 Gruppen von jeweils 100 Übergewichtigen mit 9g Salz pro Tag bzw. 3g Salz pro Tag behandelt. In der ersten Gruppe starben 14 Patienten an Herzinfarkt oder Schlaganfall, in der anderen Gruppe waren es nur 10.

Frage: Ist dieser Unterschied statistisch signifikant?

Dosis	tot	$\sim$ tot	
9g	14	86	100
3g	10	90	100
	24	176	200

$\chi^2 = \frac{25}{33}$: p-Wert $>$0.1.

Daten von E.J.Gumbel, 1922: 4 Jahre politischer Mord, vgl. Artikel von U.Rendtel im DAGStat Bulletin, Juni 2018

Die Tabelle gibt die Anzahl der politischen Morde in den Jahren 1918-1922 an, getrennt nach links- und rechtsstehenden, sowie die Anzahlen der Verurteilungen.

	Politische Morde begonnen von		
	linksstehend	rechtstehend	Gesamt
ungesühnt	4	326	330
teilw. gesühnt	1	27	28
gesühnt	17	1	18
	22	354	376

Was ist die Fragestellung? Welchen Test führen Sie durch? Was ist das Ergebnis?

Punktbiserialer Korrelationskoeffizient

Sei jetzt die Zufallsvariablen X binär, und Y stetig. N die Gesamtanzahl der Boebachtungspaare (X_i, Y_i)

$$X_i = \begin{cases} 1 & n \text{ mal} \\ 0 & m \text{ mal} \end{cases}, \qquad \overline{X} = \frac{n}{N}$$

$$\begin{aligned} s_X^2 &= \frac{1}{N-1}\left(m(-\frac{n}{N})^2 + n(1-\frac{n}{N})^2\right) = \frac{1}{N-1}\frac{mn}{N^2} \cdot N = \frac{mn}{(N-1)N} \\ s_Y^2 &= \frac{1}{N-1}\sum_{i=1}^{N}(Y_i - \overline{Y})^2 = \frac{1}{N-1}\sum_{i=1}^{N} Y_i'^2, \qquad Y_i' = Y_i - \overline{Y} \\ s_{XY} &= \frac{1}{N-1}\sum_{i=1}^{N}(X_i - \frac{n}{N})\, Y_i' \end{aligned}$$

$$
\begin{aligned}
s_{XY} &= \frac{1}{N-1}\Big(m(-\frac{n}{N})\sum_{X_i=0} Y_i' + n(1-\frac{n}{N})\sum_{X_i=1} Y_i'\Big) \\
&= \frac{mn}{N(N-1)}\Big(-\sum_{X_i=0} Y_i' + \sum_{i=1} Y_i'\Big) \\
r_P &= \frac{s_{XY}}{s_X s_Y} = \sqrt{\frac{mn}{N}}\frac{\sum_{X_i=1} Y_i' - \sum_{X_i=0} Y_i'}{\sqrt{\sum_{i=1}^{N} Y_i'^2}}
\end{aligned}
$$

Das ist der punktbiseriale Korrelationskoeffizient, vgl. Hartung, Multivariate Statistik, S.202.

Ich würde an dieser Stelle einen 2-Stichproben t-Test (oder Wilcoxon-Test oder einen anderen Test im 2-Stichprobenproblem) empfehlen.

Sei n die Anzahl der Beobachtungen und m die Anzahl der Parameter.

Lineare Regression: Minimieren $(\mathbf{Y} - \mathbf{X}\boldsymbol{\theta})'(\mathbf{Y} - \mathbf{X}\boldsymbol{\theta})$, vgl. Folie 501

Wir leiten den gegebenen Term ab und setzen ihn dann Null

$$(\mathbf{Y} - \mathbf{X}\boldsymbol{\theta})'(\mathbf{Y} - \mathbf{X}\boldsymbol{\theta}) = \sum_{i=1}^{n}\Big(y_i - \sum_{j=1}^{m} x_{ij}\theta_j\Big)^2$$

$$\begin{aligned}
\frac{\partial}{\partial\theta_k}\sum_{i=1}^{n}\Big(y_i - \sum_{j=1}^{m} x_{ij}\theta_j\Big)^2 &= 2\sum_{i=1}^{n}\Big(y_i - \sum_{j=1}^{m} x_{ij}\theta_j\Big)x_{ik} \\
= 2\Big(\sum_{i=1}^{n} y_i x_{ik} - \sum_{i=1}^{n}\sum_{j=1}^{m} x_{ij}x_{ik}\theta_j\Big) &= 2\Big(\sum_{i=1}^{n} y_i x_{ik} - \sum_{j=1}^{m}\sum_{i=1}^{n} x_{ij}x_{ik}\theta_j\Big) \\
= 2\Big(\sum_{i=1}^{n} y_i x_{ik} - \sum_{j=1}^{n}(\mathbf{X}'\mathbf{X})_{jk}\theta_j\Big) &= 2\Big((\mathbf{Y}'\mathbf{X})_k + (\mathbf{X}'\mathbf{X}\boldsymbol{\theta})_k\Big)
\end{aligned}$$

Varianz der Residuen

Sei $\mathbf{e} = (e_1, \ldots, e_n) = (Y_1 - \hat{Y}_1, \ldots, Y_n - \hat{Y}_n) = (\mathbf{Y} - \hat{\mathbf{Y}})$ der Vektor der Residuen.

$$\mathbf{Y} - \hat{\mathbf{Y}} = \mathbf{Y} - \mathbf{HY} = (\mathbf{I} - \mathbf{X}'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X})\mathbf{Y} = (\mathbf{I} - \mathbf{H})\mathbf{Y}$$

$$\begin{aligned}
\operatorname{var}(e_i) &= \operatorname{var}(Y_i - \hat{Y}_i) = \operatorname{var} \sum_{j=1}^{n} (\delta_{ij} - h_{ij}) y_j, \qquad \delta_{ij} = \begin{cases} 1 & \text{falls } i = \\ 0 & \text{sonst} \end{cases} \\
&= (1 - h_{ii}) \operatorname{var} Y_i - \sum_{j=1}^{n} h_{ij} \operatorname{var} Y_j = (1 - h_{ii})\sigma^2 + \sum_{j=1}^{n} h_{ij}\sigma^2 \\
&= (1 - h_{ii})\sigma^2
\end{aligned}$$

da wir voraussetzen, dass die Y_i unabhängig sind.

(bleibt noch zu zeigen, $\sum_{j=1}^{n} h_{ij} = 0$.)

Output der Prozedur REG

- Anzahl der Beobachtungen, abhängige Variable
- Varianzanalysetabelle

  Quadratsummen, mittlere Quadratsummen, F-Test zum globalen Testproblem
- RootMSE$=\hat{\sigma}$, R^2, adjustiertes R^2, $\overline{Y}$
- Parameterschätzung und Test ob Parameter Null
- Konfidenzbereiche (Optionen CLM, CLI)
- Residuen, studentisierte Residuen, Cook's D (Option R)
- Test auf Autokorrelation der Residuen (Option DW)

  empirische Pearson-Korrelation der Vektoren $\mathbf{e} = (e_1, \ldots, e_n)$ und $\text{Lag}(\mathbf{e}) = (e_2, \ldots, e_{n+1})$.

Sei $A = (a_{ij}), B = (b_{jk})$. Es gilt $(\mathbf{AB})' = \mathbf{B}'\mathbf{A}'$

$$\begin{aligned}(\mathbf{AB})' &= \left(\sum_{j=1}^{n} a_{ij} b_{jk}\right)'_{i,k} = \left(\sum_{j=1}^{n} a_{ij} b_{jk}\right)_{k,i} \\ &= \left(\sum_{j=1}^{n} a'_{ji} b'_{kj}\right)_{i,k} = \left(\sum_{j=1}^{n} b'_{kj} a'_{ji}\right)_{i,k} = \mathbf{B}'\mathbf{A}'\end{aligned}$$

Sei $\mathbf{A}$ regulär. Dann ist $\mathbf{A}^{-1}$ gegeben als Lösung von

$$\mathbf{A}\mathbf{A}^{-1} = \mathbf{I}$$

Verallgemeinerte (Moore-Penrose) Inverse A^+

$$\begin{aligned}\mathbf{A}\mathbf{A}^{+}\mathbf{A} &= \mathbf{A} \\ \mathbf{A}^{+}\mathbf{A}\mathbf{A}^{+} &= \mathbf{A}^{+} \\ \mathbf{A}\mathbf{A}^{+} &= (\mathbf{A}\mathbf{A}^{+})'\end{aligned}$$

Sei $\mathbf{X} = (X_1, \ldots, X_p)$ Zufallsvariable, $\mathbf{b} = (b_1, \ldots, b_p)$.

$\mathbf{Y} = \mathbf{b}'\mathbf{X} \Rightarrow \mathrm{var}\mathbf{Y} = \mathbf{b}'\mathbf{\Sigma}\mathbf{b}$

$$\mathrm{var}(\mathbf{b}'\mathbf{X}) = \mathrm{var}\Big(\sum_{j=1}^{p} b_j X_j\Big) = \sum_{i=1}^{p}\sum_{j=1}^{p} b_i b_j \mathrm{cov}(X_i, X_j) = \mathbf{b}'\mathbf{\Sigma}\mathbf{b}$$

$\mathbf{B}$: $p \times q$ Matrix. $\mathbf{Y} = \mathbf{B}'\mathbf{X} \Rightarrow \mathrm{cov}\mathbf{Y} = \mathbf{B}'\mathbf{\Sigma}\mathbf{B}$

$$\begin{aligned}
\mathrm{cov}(\mathbf{B}'\mathbf{X}) &= \mathrm{cov}\Big(\sum_{j=1}^{p} b_{ji} X_j\Big)_{i=1,\ldots,q} \\
&= \left(\mathrm{cov}\Big(\sum_{j=1}^{p} b_{ji} X_j, \sum_{j'=1}^{p} b_{j'k} X_{j'}\Big)\right)_{i,k=1,\ldots,q} \\
&= \Big(\sum_{j=1}^{p}\sum_{j'=1}^{p} b_{ji} b_{j'k} \mathrm{cov}(X_j, X_{j'})\Big)_{i,k=1,\ldots,q} = \mathbf{B}'\mathbf{\Sigma}\mathbf{B}
\end{aligned}$$

$\mathbf{E}(\mathbf{X}\mathbf{X}') = \mathbf{\Sigma}$

$$\begin{aligned}\mathbf{E}(\mathbf{X}\mathbf{X}') &= \mathbf{E}\Big(X_i X_k\Big)_{i,k=1,\ldots,p} = \Big(\mathbf{E}(X_i X_k)\Big)_{i,k=1,\ldots,p} \\ &= \Big(\mathrm{cov}(X_i, X_k)\Big)_{i,k=1,\ldots,p} = \mathbf{\Sigma}\end{aligned}$$

Sei $\mathbf{\Sigma}$ Korrelationsmatrix

$$n = sp\Big(\mathbf{\Sigma}\Big) = sp\Big(\mathbf{B}\mathbf{\Lambda}\mathbf{B}'\Big) = sp\Big(\underbrace{\mathbf{B}'\mathbf{B}}_{=I}\mathbf{\Lambda}\Big) = \sum_{i=1}^{p}\lambda_i$$

D.h. die Summe der Eigenwerte der Korrelationsmatrix ist n.

Ausgabe der Prozedur PRINCOMP

Die Hauptkomponenten $\mathbf{Y} = \mathbf{B}'\mathbf{X}$ sind gegeben durch die Eigenvektoren $\mathbf{b}$, die spaltenweise in der Matrix $\mathbf{B}$ angeordnet sind. In SAS wird die Matrix $\mathbf{B}$ ausgegeben.

Eigenvektoren der Korrrelationsmatrix im Fall von Dimension $p = 2$

$$\mathbf{\Sigma}\begin{pmatrix}1\\1\end{pmatrix} = \begin{pmatrix}1 & \rho_{12}\\ \rho_{12} & 1\end{pmatrix}\begin{pmatrix}1\\1\end{pmatrix} = \begin{pmatrix}1+\rho_{12}\\1+\rho_{12}\end{pmatrix} = (1+\rho_{12})\begin{pmatrix}1\\1\end{pmatrix}$$

d.h. $(1+\rho_{12})$ ist Eigenwert und $(1,1)'$ ist (nicht normierter) Eigenvektor von $\mathbf{\Sigma}$. Multiplikation mit Faktor $\frac{1}{\sqrt{2}}$ liefert den normierten Eigenvektor $\frac{1}{\sqrt{2}}(1,1)$. Der Vektor $\frac{1}{\sqrt{2}}(1,-1)$ ist dazu senkrechter Eigenvektor.