Nichtparametrische Statistik

Kap 1 Gauss-Test

$H_0: \mu = \mu_0$

Behauptung: $Z = \dfrac{\sqrt{n}(\bar{X} - \mu_0)}{\sigma} \overset{H_0}{\sim} N(0,1)$

Ann: $X_1, \dots, X_n$ unabhängig und identisch verteilt wie $X \sim N(\mu, \sigma^2)$

[Geg: $n = 9$ und $\sigma^2 = 0{,}18^2$ und $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$]

Bew.: (1) Normalverteilung von Z folgt direkt aus der Verteilungsannahme für X

(X ist normalverteilt $\Rightarrow$ $\bar{X}$ ist normalverteilt $\Rightarrow$ Z ist normalv.)

$\sim N(\mu, \sigma^2)$ $\quad \sim N(\mu, \frac{\sigma^2}{n})$ $\quad \overset{H_0}{\sim} N(0,1)$

(2) $E\left[\dfrac{\sqrt{n}(\bar{X} - \mu_0)}{\sigma}\right] = \dfrac{\sqrt{n}}{\sigma} E\left[\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) - \mu_0\right] =$

$= \dfrac{\sqrt{n}}{\sigma}\left[\underbrace{\left(\frac{1}{n}\sum_{i=1}^{n} \underbrace{E(X_i)}_{\mu}\right)}_{=\mu} - \mu_0\right] = \dfrac{\sqrt{n}}{\sigma}[\mu - \mu_0] = 0$

"unter H_0"
(= H_0 wahr $(\mu = \mu_0)$)

(3) $\mathbb{V}\left(\dfrac{\sqrt{n}(\bar{X} - \mu_0)}{\sigma}\right) = \dfrac{n}{\sigma^2} \mathbb{V}\left(\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) - \mu_0\right) =$

$= \dfrac{n}{\sigma^2} \mathbb{V}\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \dfrac{1}{\sigma^2 n} \mathbb{V}\left(\sum_{i=1}^{n} X_i\right) =$

$$= \frac{1}{\sigma^2 n}\left[\sum_{i=1}^{n} \mathbb{V}(X_i)\right] = \frac{1}{\sigma^2 n}\cdot\left[n\cdot\sigma^2\right] = 1$$

$\square$

Unabhängigkeit zwischen X_i und X_j für alle $i \neq j = 1, \dots, n$

---

p-Wert

Zweiseitiger Test:

$$p = P_{H_0}\left(|T| \geq |T_{beob}|\right)$$

Einseitige Versionen

$H_1: \mu > \mu_0$

$$p_{größer} = P_{H_0}\left(T \geq T_{beob}\right)$$

$H_1: \mu < \mu_0$

$$p_{kleiner} = P_{H_0}\left(T \leq T_{beob}\right),$$

wobei

$$p = 2\cdot \min\{p_{kleiner}, p_{größer}\}$$

(zweiseitig)

Die Gütefunktion

$\beta(n, \alpha, \theta, \theta_0) = P$ ("Ablehnung von H_0, falls wahrer Parameterwert $\theta \in \Omega_0 \cup \Omega_1$")

oft unterschlagen

- $\beta(n, \alpha, \theta) \leq \alpha$ für alle $\theta \in \Omega_0$

Achtung!: Falls $\beta(n, \alpha, \theta) > \alpha$ für mind. ein $\theta \in \Omega_0$

$\Rightarrow$ $\boxed{\text{ungültiger Test}}$

Ideal: $\beta(n, \alpha, \theta) = \alpha$ für alle $\theta \in \Omega_0$

(oder zumindest für mind. ein $\theta \in \Omega_0$)

- Falls $\beta(n, \alpha, \theta) < \alpha$ für all $\theta \in \Omega_0$, nennt man den dazugehörigen Test "konservativ".

- Für $\theta \in \Omega_1$ ist $\beta(n, \alpha, \theta)$ die Wahrscheinlichkeit die richtige Entscheidung zu treffen (H_0 ablehnen).

Ideal: $\beta(n, \alpha, \theta) \approx 1$ für alle $\theta \in \Omega_1$

- "unverfälschter Test"

$$\beta(n,\alpha,\theta\in\Omega_1) \geq \beta(n,\alpha,\theta\in\Omega_0)$$

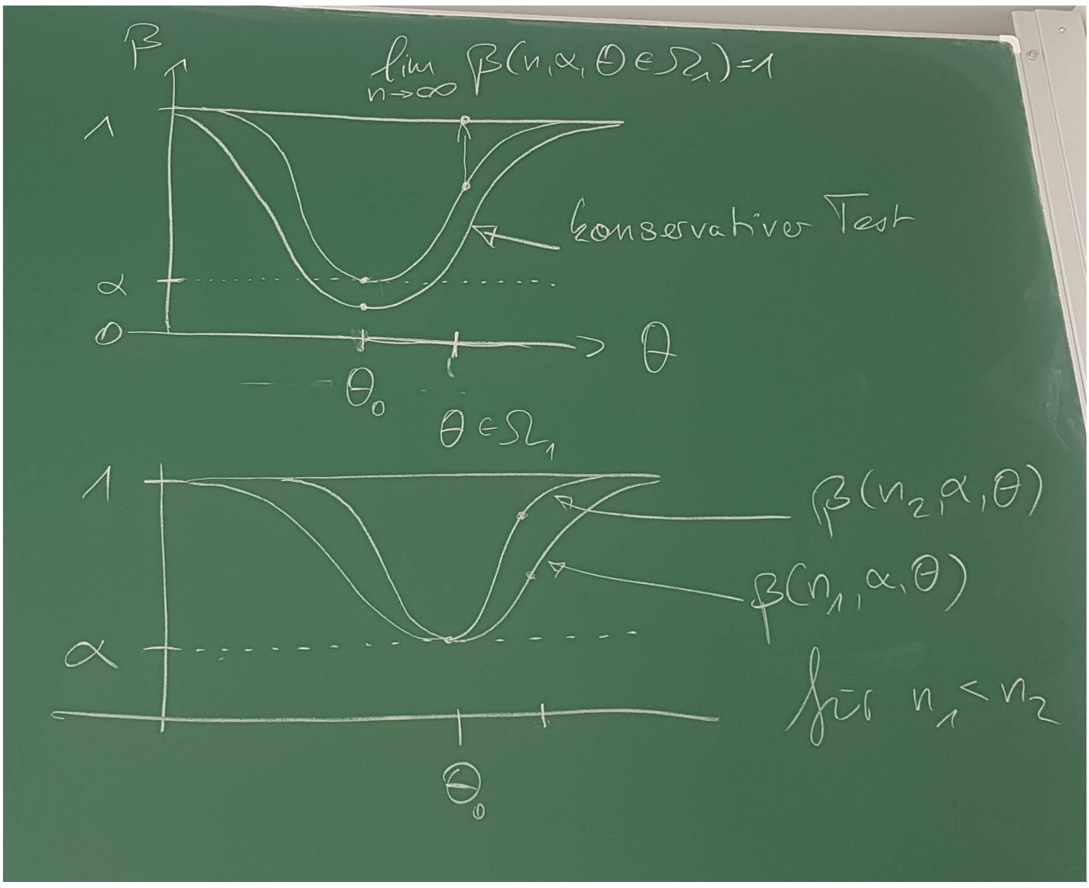
- "konsistenter Test"

$$\beta(n,\alpha,\theta) \xrightarrow[n\to\infty]{} 1 \quad \text{für alle } \theta\in\Omega_1$$

- Seien β und β^* die Gütefunktionen von zwei unterschiedlichen Tests (t-Test und z-Test).

Falls $\beta(n,\alpha,\theta) \geq \beta^*(n,\alpha,\theta)$ für <u>alle n</u> und für alle $\theta\in\Omega_1$,

dann nennt man den zu β gehörigen Test "gleichmäßig besten" Test.



Die Gütefunktion des z-Tests

- $X_i \overset{iid}{\sim} N(\mu, \sigma^2)$ für alle $i = 1, \dots, n$
- $H_0: \mu = \mu_0$ versus $H_1: \mu \neq \mu_0$

$$z = \frac{\sqrt{n}(\bar{X} - \mu_0)}{\sigma} = \frac{\bar{X} - \mu_0}{\sqrt{\underbrace{\frac{\sigma^2}{n}}_{Var(\bar{X})}}} = \frac{\bar{X} - \mu + (\mu - \mu_0)}{\sqrt{\frac{\sigma^2}{n}}}$$

$$= \underbrace{\frac{\bar{X} - \mu}{\sqrt{\frac{\sigma^2}{n}}}}_{\sim N(0,1)} + \frac{\mu - \mu_0}{\sqrt{\frac{\sigma^2}{n}}} \sim N\left(\frac{\mu - \mu_0}{\sqrt{\frac{\sigma^2}{n}}}, 1\right)$$

Verteilung von Z unter H_0 und H_1

$$\beta_Z(n, \alpha, \mu, \mu_0) = P(Z > z_{1-\frac{\alpha}{2}}) + P(Z < z_{\frac{\alpha}{2}})$$

$(1-\frac{\alpha}{2})$ und $(\frac{\alpha}{2})$ Quantil von $N(0,1)$

Simulation einer Gütefunktion (Beispiel t-Test)

Zufallsstichprobe

$(X_1, \dots, X_n)$, wobei: $X_i \overset{iid}{\sim} N(\mu, \sigma^2)$

$$T = T(X_1, \dots, X_n) = \frac{\sqrt{n}(\bar{X} - \mu_0)}{S} \overset{H_0}{\sim} t_{n-1} \quad \left(\overset{z.B.}{10000}\right)$$

Per Zufallszahlengenerator erzeugen wir nun $b = 1, 2, \dots, B$ Realisationen von $(X_1^{(b)}, \dots, X_n^{(b)}) \Rightarrow T^{(b)} = \frac{\sqrt{n}(\bar{X}^{(b)} - \mu_0)}{s^{(b)}}$

$$\beta_t(n, \alpha, \mu, \mu_0) = P\left(\text{„}H_0 \text{ ablehnen falls wahrer Parameterwert } \mu\text{“}\right)$$

$$\approx \frac{1}{B} \sum_{b=1}^{B} 1\!\!1\left(T^{(b)} \in C\right)$$

Starke Gesetz der großen Zahlen:

$$\frac{1}{B} \sum_{b=1}^{B} 1\!\!1\left(T^{(b)} \in C\right) \xrightarrow[a.s.]{B \to \infty} E\left(1\!\!1(T \in C)\right) = \underline{\underline{P(T \in C)}} \cdot 1 + P(T \notin C) \cdot 0$$

unter H_0:

$$P_{H_0}(T \in C) = E\left(1\!\!1(T \in C)\right) = \alpha$$

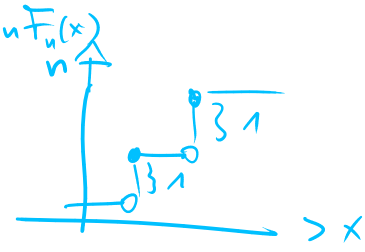
unter H_1:

$$E\left(1\!\!1(T \in C)\right) \geq \alpha$$

Kapitel 2

Lokale Eigenschaften von $F_n(x)$

↑ d.h. für ein "festes" x



$$F_n(x) \in \left\{0, \frac{1}{n}, \frac{2}{n}, \dots, 1\right\}$$

$$\Leftrightarrow \quad n \cdot F_n(x) \in \{0, 1, \dots, n\}$$

Satz: Sei $X_1, \dots, X_n$ iid wie $X \sim F$,

wobei ~~F stetig und~~ $X \in \mathbb{R}$.

Dann gilt für jedes feste $x \in \mathbb{R}$

$$P\left(F_n(x) = \frac{m}{n}\right) = \binom{n}{m} (F(x))^m \, (1 - F(x))^{n-m}$$

$$P(n \cdot F_n(x) = m)$$

d.h. $F_n(x) \sim \text{Binom}(n, p = F(x))$

Setze $\quad Y_i(x) = 1\!\!1(X_i \leq x) \in \{0, 1\}$

$$P(Y_i(x) = 1) = P(X_i \leq x) = F(x)$$
$$P(Y_i(x) = 0) = P(X_i > x) = 1 - F(x)$$

Somit ist $\quad Y_i(x) \in \{0, 1\}$ eine

Bernoulli ZV.

$$Y_i(x) \sim \text{Bern}(p = F(x))$$

Damit:

$$n \, F_n(x) = \sum_{i=1}^{n} 1\!\!1(X_i \leq x) \in \{0, 1, \dots, n\}$$

$\underbrace{\qquad\qquad\qquad}$ Binomialverteilt mit Parametern n und $p = F(x)$

D.h.:

$$P(nF_n(x) = m) = \binom{n}{m} (F(x))^m (1 - F(x))^{n-m}$$

$$P\left(F_n(x) = \frac{m}{n}\right)$$

$\square$

Da $n \cdot F_n(x) \sim \text{Binom}(n, p = F(x))$

1. $E(nF_n(x)) = n \cdot F(x)$

$\Leftrightarrow \quad E(F_n(x)) = F(x)$

D.h. $F_n(x)$ ist erwartungstreuer Schätzer für $F(x)$

2. $Var(nF_n(x)) = n \cdot F(x)(1 - F(x))$

$\Leftrightarrow \quad n^2 Var(F_n(x)) = n \cdot F(x)(1 - F(x))$

$\Leftrightarrow \quad Var(F_n(x)) = \frac{1}{n} F(x)(1 - F(x))$

D.h. $Var(F_n(x)) \xrightarrow[n \to \infty]{} 0$

3. Behauptung: $F_n(x)$ ist (schwach) konsistenter Schätzer für $F(x)$

D.h. $F_n(x) \xrightarrow[n \to \infty]{p} F(x)$

Tschebyschef-Ungleichung:

$$P(|X - E(X)| < k) \geq 1 - \frac{Var(X)}{k^2} \quad \text{für } k > 0$$

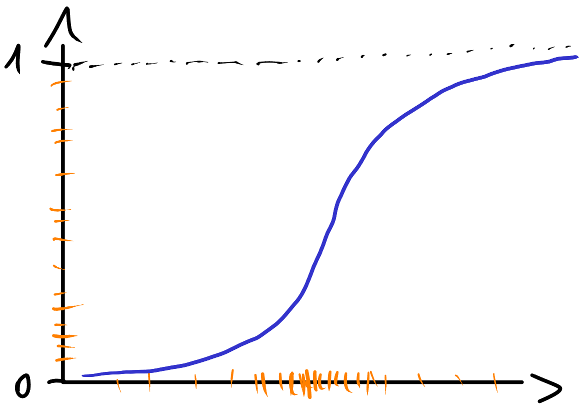
$$\left(\Leftrightarrow \quad P(|X - E(X)| > k) \leq \frac{Var(X)}{k^2}\right)$$

Hier:

$$1 \geq P\left(\left| F_n(x) - \overset{E(F_n(x))}{\widetilde{F(x)}} \right| < b \right) \geq \underbrace{1 - \frac{\frac{1}{n} F(x)(1-F(x))}{b^2}}_{\xrightarrow[n \to \infty]{} 1}$$

Zum Satz, dass $F(X) \sim U[0,1]$, wobei $X \sim F$ und F stetig ist.

in etwa gleichverteilt →



$X \overset{\text{z.B.}}{\sim} N(\mu, \sigma^2)$

Randnotiz

Inverse uniform integral transform

$$X = F^{-1}(U) \qquad U \sim U[0,1]$$

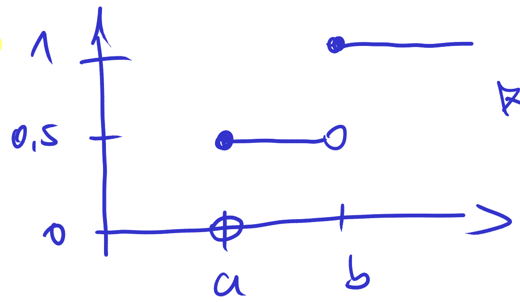
$$\Rightarrow \quad X \sim F$$

Nicht-Robustheit des Mittelwerts

$$\overline{X} = \frac{1}{n} \sum_{i=1}^{n} x_i \longrightarrow \infty$$

für $x_1 \to \infty$ und alle anderen $(x_2, \dots, x_n)$ fest.

Quantile



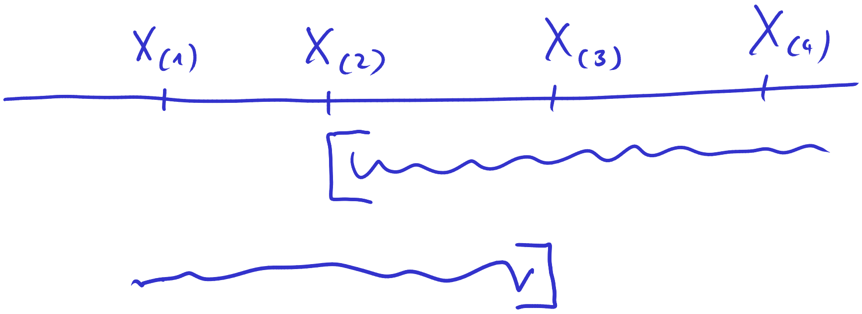
Verteilungsfunktion von X

$p = 0{,}5$

$$P(X < \psi_{0,5}) \leq 0{,}5 \leq P(X \leq \psi_{0,5})$$

gilt für alle $\psi_{0,5} \in [a, b]$

Empirisches Quantil (Bsp. $p = 0{,}5$ und $n = 4$)



$$\frac{\# X_i \leq \psi_{n,p}}{n} \geq 0{,}5$$

$$\frac{\# X_i \geq \psi_{n,p}}{n} \geq 0{,}5$$

$$\Rightarrow \quad \psi_{n,p} = \psi_{4;0,5} \in [X_{(2)}, X_{(3)}]$$

$\psi_{n,p}$ „approximiert" ψ_p

$\psi_{n,p}$ ist ein konsistenter „Punktschätzer" für ψ_p.

Unter gewissen Voraussetzungen (siehe B&T, Kap 3.4.8) gilt, dass $\psi_{n,p} \xrightarrow{P} \psi_p$ für $n \to \infty$

Konfidenzintervall für Quantile

<u>Ziel:</u> Finde $X_{(k)}$ und $X_{(l)}$, wobei $k<l$, sodass

$$P(X_{(k)} < \psi_p < X_{(l)}) =$$

$$= P(\psi_p \in \underbrace{[X_{(k)}, X_{(l)}]}) \geq 1-\alpha$$

$(1-\alpha)\cdot 100\,\%$ Konfidenzintervall für ψ_p.

$F(X_{(k)}) \sim U[0,1]$

$F(X) \sim U[0,1]$

<u>Ann:</u> Die Verteilungsfunktion F von X sei stetig und <u>streng</u> monoton wachsend.

Da F streng monoton wachsend ist, gilt folgende Identität von Zufallsereignissen:

$$X_{(k)} < \psi_p \iff F(X_{(k)}) < \underbrace{F(\psi_p)}_{=p}$$

$$X_{(l)} > \psi_p \iff F(X_{(l)}) > \overbrace{F(\psi_p)}$$

<u>Somit:</u> $P(X_{(k)} < \psi_p < X_{(l)}) =$

$$= P(F(X_{(k)}) < p < F(X_{(l)}))$$

Mögliche Zufallsrealisationen

1. $X_{(k)} < \psi_p < X_{(k+1)} \iff \sum_{i=1}^{n} \mathbb{1}(\overbrace{F(X_{(i)})}^{\sim U[0,1]} < p) = k$

2. $X_{(k+1)} < \psi_p < X_{(k+2)} \iff \sum_{i=1}^{n} \mathbb{1}(F(X_{(i)}) < p) = k+1$

$\vdots$

$(l-k-1)$ $X_{(l-1)} < \psi_p < X_{(l)} \iff \underbrace{\sum_{i=1}^{n} \mathbb{1}(F(X_{(i)}) < p)}_{\sim \text{Binom}(n,p)} = l-1$

Also: $P(X_{(k)} < \psi_p < X_{(l)}) =$

$$= \sum_{j=k}^{l-1} \binom{n}{j} p^j (1-p)^{n-j} \geq 1-\alpha$$

Ziel: suche k und l, sodass diese Ungleichung erfüllt ist.

KS-Test

Annahme A: $X_1, \ldots, X_n \overset{iid}{\sim} F$, wobei F stetig und streng monoton wachsend

(d.h. $F(x_1) < F(x_2)$, für all $x_1 < x_2$)

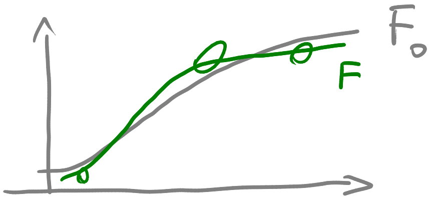
Zweiseitiger Test:

$H_0: F = F_0$ $\quad (H_0: F(x) = F_0(x)$ für alle $x \in \mathbb{R})$

$H_1: F(x) \neq F_0(x)$ für mind. ein $x \in \mathbb{R}$

$$D_n = \sup_{x \in \mathbb{R}} |F_n(x) - F_0(x)|$$

<u>H_1-Szenario:</u>



---

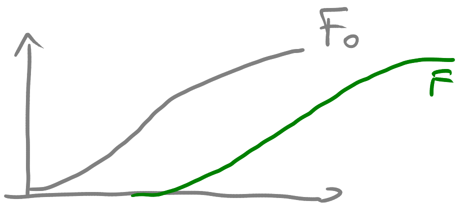
einseitig

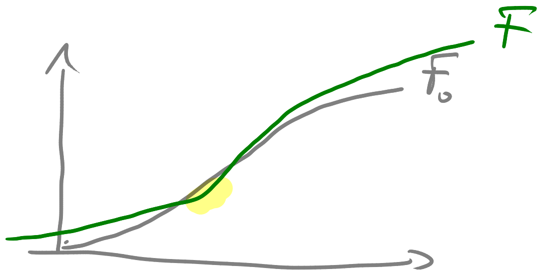
$$D_n^- = \sup_{x \in \mathbb{R}} (F_0(x) - F_n(x))$$

H_0: $F(x) \geq F_0(x)$ für alle $x \in \mathbb{R}$

H_1: $F(x) < F_0(x)$ für mind. ein $x \in \mathbb{R}$

H_1-Szenarien:





$$D_n^+ = \sup_{x \in \mathbb{R}} (F_n(x) - F_0(x))$$

H_0: $F(x) \leq F_0(x)$ für alle $x \in \mathbb{R}$

H_1: $F(x) > F_0(x)$ für mind. ein $x \in \mathbb{R}$.

<u>Beachte:</u> $D_n = \max\{D_n^+, D_n^-\}$

Berechnung von D_n^+, D_n^- und D_n

Unter der Annahme A (F stetig und str. monoton wachsend) gilt, dass $F_n(x) = \frac{i}{n}$ für alle $x \in [x_{(i)}, x_{(i+1)})$

$(x_{(i)} \le x < x_{(i+1)})$

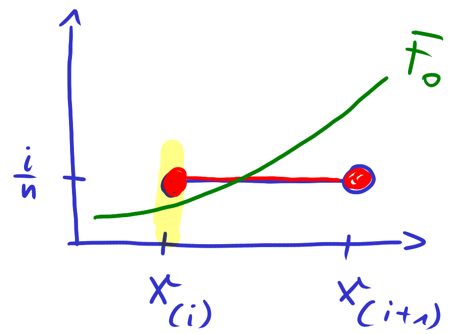
(D.h. nur solche F_0, die Annahme A erfüllen, machen Sinn!)

Sei $x_{(0)} = -\infty$ und $x_{(n+1)} = \infty$

$$D_n^+ = \sup_{x \in \mathbb{R}} \left(F_n(x) - F_0(x) \right)$$

$$= \max_{0 \le i \le n} \left(\sup_{x_{(i)} \le x < x_{(i+1)}} \left(F_n(x) - F_0(x) \right) \right)$$

da F_0 (erfüllt Annahme A) streng monoton wachsend und F_n konstant



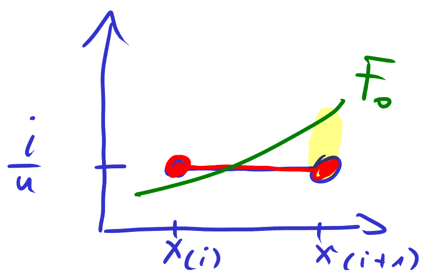
$$= \max_{0 \le i \le n} \left(F_n(x_{(i)}) - F_0(x_{(i)}) \right)$$

$$= \max_{0 \le i \le n} \left(\frac{i}{n} - F_0(x_{(i)}) \right)$$

$$D_n^- = \sup_{x \in \mathbb{R}} \left(F_0(x) - F_n(x) \right)$$

$$= \max_{0 \le i \le n} \left(\sup_{x_{(i)} \le x < x_{(i+1)}} \left(F_0(x) - F_n(x) \right) \right)$$

$$= \max_{0 \le i \le n} \left(F_0(x_{(i+1)}) - F_n(x_{(i)}) \right)$$



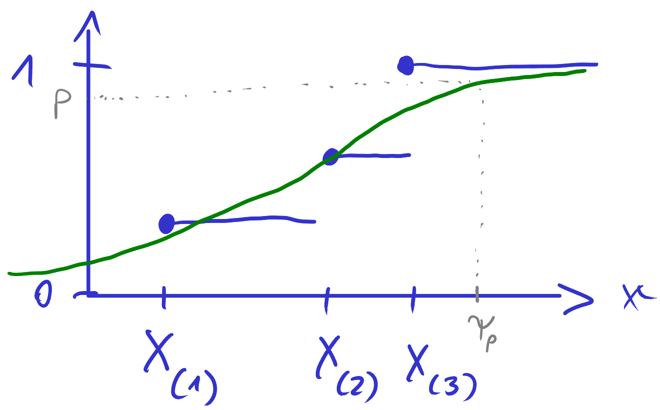
$$= \max_{0 \le i \le n} \left(F_0(X_{(i+1)}) - \frac{i}{n} \right)$$

Für $D_n = \max\{D_n^+, D_n^-\}$

Verteilungsfreiheit von D_n unter $H_0: F_0 = F$

(und im Folgenden unter der Annahme A)

Motivation:



$$F_n(x) = F_n(\overbrace{F^{-1}(p)}^{x})$$

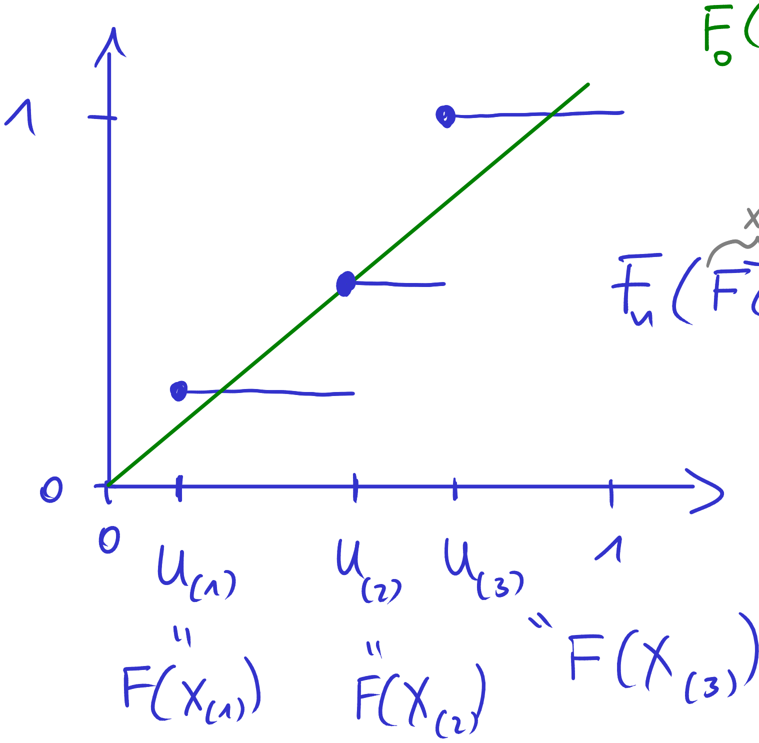
$$F(x) \overset{H_0}{=} F_0(x) = \underbrace{F_0(\overbrace{F^{-1}(p)}^{x})}_{}$$

$$\overset{H_0}{=} p_x$$

unter H_0 Identitäts-Funktion

- Für jedes $x \in \mathbb{R}$ existiert genau ein $p_x \in [0,1]$, sodass $\boxed{\psi_p = x}$, wobei ψ_p das p-Quantil der Verteilung F ist.
- $\psi_p = x \iff F^{-1}(p) = x \overset{H_0}{\iff} F_0^{-1}(p) = x$
- $F(X_{(i)}) = U_{(i)}$ mit $U \sim U[0,1]$

$$\overset{H_0}{\iff} F_0(X_{(i)}) = U_{(i)}$$



$$F_0(\overbrace{F^{-1}(p_x)}^{x}) = \tilde{F}_0(p) \overset{H_0}{=} I(p) = p$$

"Identitätsfunktion"

$$F_n(\overbrace{F^{-1}(p_x)}^{x}) = \tilde{F}_n(p)$$

unter $H_0: F_0 = F$ ist $\tilde{F}_n$ die empirische Verteilungsfunktion basierend $(F(X_1), \dots, F(X_n))$, wobei $F(X_i) \overset{iid}{\sim} U[0,1]$

Herleitung der Verteilungsfreiheit von D_n unter $H_0: F_0 = F$ (und unter Annahme A)

$$D_n = \sup_{x \in \mathbb{R}} |F_0(x) - F_n(x)| = \sup_{p \in [0,1]} |F_0(\psi_p) - F_n(\overbrace{\psi_p}^{F^{-1}(p)})|$$

• Für jedes $x \in \mathbb{R}$ $\overbrace{\text{existiert genau ein}}^{\exists!}$ $p \in [0,1]$, $\overset{\text{"s.t."}}{\text{sodass}}$ $\psi_{p_x} = x$.

[$\forall x \in \mathbb{R}, \exists!\, p \in [0,1]$ s.t. $\psi_p = x$]

"für alle"

- Unter $H_0: F_0 = F$ gilt:

$$F_0(\Psi_p) = F_0(F^{-1}(p)) \overset{H_0}{=} F(F^{-1}(p)) = p$$

$$F^{-1}(p) \overset{H_0}{=} F_0^{-1}(p)$$

$$\Rightarrow \quad D_n \overset{H_0}{=} \sup_{p \in [0,1]} \Big| p - \underbrace{F_n(F_0^{-1}(p))}_{\tilde{F}_n(p)} \Big|$$

Empirische VF basierend auf $(F_0(X_1), \dots, F_0(X_n))$

wobei unter $H_0: F_0 = F$ gilt

$$F(X_i) \overset{iid}{\sim} U[0,1] \text{ für alle } i = 1, \dots, n.$$

Motivation warum Testentscheidungen basierend auf

$$D_n = \sup_{x \in \mathbb{R}} |F_n(x) - F_0(x, \hat{\theta})| \quad (\hat{\theta} \approx \theta_{wahr})$$

und den kritischen Werten von

$$D_n = \sup_{x \in \mathbb{R}} |F_n(x) - F_0(x, \theta_{wahr})|,$$

wobei die Stochastik von $\hat{\theta}$ ignoriert wird, konservativ sind.

Grund:

$$\sup_{x \in \mathbb{R}} |F_n(x) - F_0(x, \hat{\theta})| \leq \sup_{x \in \mathbb{R}} |F_n(x) - F_0(x, \theta_{wahr})|$$

χ^2-Anpassungstest

Beispiel:

$q = 2$

$n = 6 \; (= n_1 + n_2)$

$\boxed{\begin{array}{l} n_1 = 2 \\ n_2 = 4 \end{array}}$

$H_0: \pi_1 = \pi_1^0 = 0{,}5$ und $\pi_2 = \pi_2^0 = 0{,}5$ $\quad (\pi_1^0 + \ldots + \pi_q^0 = 1)$

H_1: „H_0 falsch"

beobachtete Zufallsrealisation

$$Q_{beob} = \sum_{j=1}^{2} \frac{(n_j - n \cdot \pi_j^0)^2}{n \cdot \pi_j^0} = \frac{(2-3)^2}{3} + \frac{(4-3)^2}{3} = \frac{2}{3}$$

Berechnung des p-Wertes

$$\text{p-Wert} = P_{H_0}(Q \geq Q_{beob})$$

$$= 1 - P_{H_0}(Q < \underbrace{Q_{beob}}_{2/3})$$

$$= 1 - 0{,}3125 = 0{,}6875$$

Q ist diskrete Zufallsvariable, da n_1 und n_2 nur diskrete Werte annehmen.

$$P_{H_0}\left(Q < \tfrac{2}{3}\right) = \sum_{\{(m_1, m_2) \mid Q < \frac{2}{3}\}} \boxed{P(n_1 = m_1, \; n_2 = m_2)} = 0{,}3125$$

Binomialverteilung $(n, p = 0{,}5)$

(m_1, m_2)	Q	$P(n_1 = m_1, n_2 = m_2)$
$(3,3)$	$\frac{(3-3)^2}{3} + \frac{(3-3)^2}{3} = 0$	$0{,}3125$ dmultinom(x = c(3,3), size = 6,

			prob = c(0.5, 0.5))
(2,4)	$\frac{2}{3}$	/	
(4,2)	$\frac{2}{3}$	/	

Daumenregel zur Diskretisierung von stetigen Verteilungen (Folie 37)

$$\frac{n}{10} \leq q \leq \frac{n}{5} ,$$

und die A_j - Klassen so dass ungefähr gleich viele Datenpunkte pro Klasse vorhanden sind.

Verteilung der Ränge

Sei $X_1, \dots, X_n$ iid wie $X \sim F$ mit F stetig.

Rang von X_i

$$r(X_i) = \sum_{j=1}^{n} 1\!\!1 (\overbrace{X_i - X_j \geq 0}^{X_i \geq X_j})$$

Anzahl der Stichprobenvariablen, die X_i nicht übertreffen.

- $r(X_i) \in \{1, 2, \dots n\}$

$$P(r(X_i) = a) = \frac{1}{n} \quad \text{für alle } a = 1, 2, \ldots, n$$

- $(r(X_1), \ldots, r(X_n)) \in A = \{(a_1, \ldots, a_n) \in \underbrace{\{1, \ldots, n\}^n}_{\{1,\ldots,n\} \times \{1,\ldots,n\} \times \ldots \times \{1,\ldots,n\}} \mid a_i \neq a_j, 1 \leq i, j \leq n, i \neq j\}$

$\quad\quad a_1 \quad\quad a_2 \quad\quad a_n$

$1, 2, \ldots, n \quad\quad a_1 = 1, 2, \ldots, n$

$3, 1, \ldots, 5 \quad\quad a_2 = 1, 2, \ldots, n$

$2, 7, \ldots, 3$

$$P((r(X_1), \ldots, r(X_n)) = (a_1, \ldots, a_n)) = \frac{1}{n} \cdot \frac{1}{n-1} \cdot \frac{1}{n-2} \cdot \ldots \cdot \frac{1}{1}$$
$$= \frac{1}{n!}$$

für alle $(a_1, \ldots, a_n) \in A$

- $E(r(X_i)) = \sum_{j=1}^{n} j \underbrace{P(r(X_i) = j)}_{\frac{1}{n}} = \frac{1}{n} \sum_{j=1}^{n} j = \frac{1}{n} \frac{n(n+1)}{2}$

$$\Rightarrow E(r(X_i)) = \frac{(n+1)}{2}$$

BSP $n = 5$, $E(r(X_i)) = 3$, $1\ 2\ 3\ 4\ 5$

- $Var(r(X_i)) = E(r(X_i)^2) - (E(r(X_i)))^2$

$$E(r(X_i)^2) = \sum_{j=1}^{n} j^2 P(r(X_i)^2 = j^2)$$
$$= \sum_{j=1}^{n} j^2 \underbrace{P(r(X_i) = j)}_{= \frac{1}{n}}$$

$$= \frac{1}{n} \sum_{j=1}^{n} j^2 = \frac{1}{n} \left(\frac{1}{6} n(n+1)(2n+1) \right)$$

$$\Rightarrow Var(r(X_i)) = \frac{1}{6}(n+1)(2n+1) - \left(\frac{n+1}{2}\right)^2$$

$$= \ldots = \frac{n^2-1}{2}$$

$Cov(r(X_i), r(X_j)) =$ Übungsblatt 2

für $i \neq j$

Lineare Rangstatistiken

Tests auf Lagealternativen

$H_0: M_{med} = M_0$

$H_1: M_{med} \neq M_0$

IQ Beispiel:

$H_0: M_{med} = 110$

$H_1: M_{med} \neq 110$

- Differenzen: $D_i = X_i - M_0$
- Ränge der absoluten Differenzen: $r(|D_i|)$
- Vorzeichen (der Differenzen) $V_i = \begin{cases} 1, \text{ falls } X_i - M_0 > 0 \\ 0, \text{ falls } \underbrace{X_i - M_0}_{D_i} \leq 0 \end{cases}$

Allgemeine lineare Rangstatistik

$$L_n^+ = \sum_{i=1}^{n} g(r(|D_i|)) \cdot V_i$$

Unterschiedliche Funktionen g führen zu unterschiedlichen Tests.

-> "Vorzeichentest": $g(x) = 1$

-> "Wilcoxon Vorzeichen-Rangtest": $g(x) = x$

Vorzeichentest (Binomialtest)

Annahmen:

Verteilungsfunktion

Sei $X_1, \ldots, X_n$ iid wie $X \sim F$ mit F stetig

$H_0: M_{med} = M_0$

$H_1: M_{med} \neq M_0$

Teststatistik:

$$V_n^+ = \sum_{i=1}^{n} V_i \quad, \text{wobei} \quad V_i = \begin{cases} 1, & \text{falls } X_i - M_0 > 0 \\ 0, & \text{falls } X_i - M_0 \leq 0 \end{cases}$$

Beachte:

$$P_{H_0}(\overbrace{X_i - M_0 \leq 0}^{V_i = 0}) = \frac{1}{2}$$

$$P_{H_0}(\underbrace{X_i - M_0 > 0}_{V_i = 1}) = 1 - P_{H_0}(X_i - M_0 \leq 0) = \frac{1}{2}$$

(folgt direkt aus der Definition des Medians)

D.h. $V_i \overset{H_0}{\sim} \text{Bernoulli}(p = \frac{1}{2})$

$\Rightarrow V_n^+ \overset{H_0}{\sim} \text{Binomial}(n = n, p = \frac{1}{2})$

Zweiseitiger Test:

H_0 wird abgelehnt, falls $V^+_{n,beob}$ "zu groß" ist

, d.h. falls $P_{H_0}(V_n^+ \geq V^+_{n,beob}) < \frac{\alpha}{2}$,

oder falls $V^+_{n,beob}$ "zu klein" ist,

d.h. falls $P_{H_0}(V_n^+ \leq V^+_{n,beob}) < \frac{\alpha}{2}$.

Herleitung der Verteilung (unter H_0) durch Auszählen:

Beispiel $n = 5$

$$V_n^+ = V_1 + V_2 + V_3 + V_4 + V_5 \in \{0, 1, \dots, 5\}$$

$$(V_1, V_2, V_3, V_4, V_5) \in \{0,1\}^5$$

$\Rightarrow$ 2^5 Möglichkeiten

$V_n^+ = v$	$(V_1, V_2, V_3, V_4, V_5)$	$P_{H_0}(V_n^+ = v)$	$P_{H_0}(V_n^+ \leq v)$
$v = 0$	$(0,0,0,0,0)$	$\frac{1}{2^5} = \frac{\binom{5}{0}}{2^5}$	$\frac{\binom{5}{0}}{2^5}$
$v = 1$	$(1,0,0,0,0)$ $(0,1,0,0,0)$ $(0,0,1,0,0)$ $(0,0,0,1,0)$ $(0,0,0,0,1)$	$\frac{5}{2^5} = \frac{\binom{5}{1}}{2^5}$	$\sum_{k=0}^{v} \frac{\binom{5}{k}}{2^5}$

⋮

usw.

Allgemein

Wahrscheinlichkeitsfunktion von V_n^+ unter H_0:

$$P_{H_0}(V_n^+ = v) = \frac{\binom{n}{v}}{2^n} \quad \text{für } v \in \{0, 1, \ldots, n\}$$

symmetrisch: $P_{H_0}(V_n^+ = v) = P_{H_0}(V_n^+ = n - v)$

denn $\binom{n}{v} = \binom{n}{n-v}$

vgl. Wahrscheinlichkeitsfunktion von $B \sim \text{Binom}(n,p)$:

$$P(B = k) = \binom{n}{k} p^k (1-p)^{n-k}$$

Verteilungsfunktion unter H_0:

$$P_{H_0}(V_n^+ \leq v) = \sum_{u=0}^{v} \frac{\binom{n}{u}}{2^n} \qquad v \in \{0, \ldots, n\}$$

wegen symmetrie:

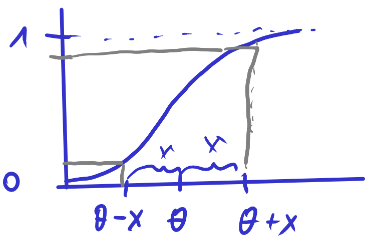
$$P_{H_0}(V_n^+ \leq v) = P_{H_0}(V_n^+ \geq n - v)$$

Wilcoxons Vorzeichen-Rangtest

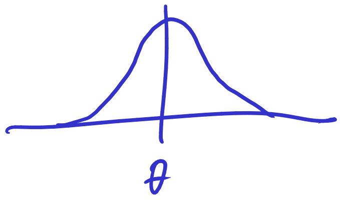
Annahmen:

Sei $X_1, \ldots, X_n$ iid wie $X \sim F$, wobei F stetig und symmetrisch um θ, d.h.

$$F(\theta - x) + F(\theta + x) = 1$$



$$f(\theta - x) = f(\theta + x)$$



$$H_0: M_{med} = M_0$$

$$H_1: M_{med} \neq M_0$$

Teststatistik:

$$W_n^+ = \sum_{i=1}^{n} r(|D_i|)\, V_i$$

$$V_i = \begin{cases} 1, & \text{falls } X_i - M_0 > 0 \\ 0, & \text{falls } X_i - M_0 \leq 0 \end{cases}$$

$$D_i = X_i - M_0$$

Herleitung der Verteilung von W_n^+ unter H_0 durch "Auszählen"

$$W_n^+ = \sum_{i=1}^{n} \underbrace{r(|D_i|)}_{\in\{1,\dots,n\}} \overbrace{V_i}^{\in\{0,1\}} \in \left\{0, 1, 2, \frac{n(n+1)}{2}\right\}$$

W_n^+ ist symmetrisch verteilt mit

$$E_{H_0}(W_n^+) = \frac{n(n+1)}{4} \quad \text{und} \quad Var_{H_0}(W_n^+) = \frac{n(n+1)(2n+1)}{24}$$

Beispiel $n = 5$

$$P_{H_0}(W_n^+ = w) = \frac{a(w)}{2^5}$$

$a(w)$: Anzahl der $\boxed{\text{Rangtupel mit } V_i > 0}$ für welche $W_n^+ = w$

von Vektoren $(r(|D_1|)V_1, \dots, r(|D_n|)V_n)$ mit $V_i > 0$

$$W_{5,max}^+ = \sum_{i=1}^{n} i = \frac{n(n+1)}{2} = 15 \quad (E(W_n^+) = 7{,}5)$$

w	Rangtupel mit $V_i = 1$ und $W_n^+ = w$	$P_{H_0}(W_n^+ = w) = \frac{a(w)}{2^n}$	wegen Symmetrie
15	(1, 2, 3, 4, 5)	$\frac{1}{32}$	$w = 0$ (0, 0, 0, 0, 0)
14	(2, 3, 4, 5)	$\frac{1}{32}$	$w = 1$ (1, 0, 0, 0, 0)
13	(1, 3, 4, 5)	$\frac{1}{32}$	$w = 2$ (0, 2, 0, 0, 0)
12	(3, 4, 5) (1, 2, 4, 5)	$\frac{2}{32}$	$w = 3$ (1, 2, 0, 0, 0) (0, 0, 3, 0, 0)
11	(1, 2, 3, 5) (2, 4, 5)	$\frac{2}{32}$	$\hat{=}\ w = 4$
10	(1, 2, 3, 4) (2, 3, 5) (1, 4, 5)	$\frac{3}{32}$	$\hat{=}\ w = 5$
9	(4, 5); (2, 3, 4); (1, 3, 5)	$\frac{3}{32}$	$\hat{=}\ w = 6$
8	(3, 5); (1, 3, 4); (1, 2, 5)	$\frac{3}{32}$	$\hat{=}\ w = 7$

Zweistichprobenprobleme:

Lineare Rangtests

Allgem. lineare Rangstatistik

$$L_N = \sum_{i=1}^{N} a_i V_i$$

$$V_i = \begin{cases} 1, & \text{falls } i\text{te Ordnungsstatistik eine X-Variable ist} \\ 0, & \text{sonst} \end{cases}$$

$$E(V_i) = 1 \frac{m}{N} + 0 \cdot \frac{n}{N} = \frac{m}{N}$$

$$Var(V_i) = E(V_i^2) - (E(V_i))^2$$

$$= \overbrace{1^2 \frac{m}{N} + 0^2 \frac{n}{N}} - \left(\frac{m}{N}\right)^2 = \frac{m(m+n)}{N^2} - \frac{m^2}{N^2} = \frac{mn}{N^2}$$

$$Cov(V_i, V_j) = E(V_i V_j) - E(V_i) E(V_j)$$

$(i \neq j)$

$$E(V_i V_j) = 1 \cdot P(V_i = 1, V_j = 1) + 0$$

$$= 1 \cdot P(V_i = 1 \mid V_j = 1) \cdot P(V_j = 1)$$

$$= 1 \cdot \frac{m-1}{N-1} \cdot \frac{m}{N}$$

$$\Rightarrow Cov(V_i, V_j) = \frac{(m-1)m}{(N-1)N} - \left(\frac{m}{N}\right)^2 = \ldots = \frac{-nm}{N^2(N-1)}$$

$$E(L_N) = \sum_{i=1}^{N} a_i E(V_i) = \frac{m}{N} \sum_{i=1}^{N} a_i$$

$$Var(L_N) = Var\left(\sum_{i=1}^{N} a_i V_i\right)$$

$$= \sum_{i=1}^{N} a_i^2 Var(V_i) + \sum_{\substack{i=1 \\ (i \neq j)}}^{N} \sum_{j=1}^{N} a_i a_j Cov(V_i, V_j)$$

$$= \frac{mn}{N^2} \sum_{i=1}^{N} a_i^2 - \frac{nm}{N^2(N-1)} \sum_{\substack{i=1 \\ (i \neq j)}}^{N} \sum_{j=1}^{N} a_i a_j$$

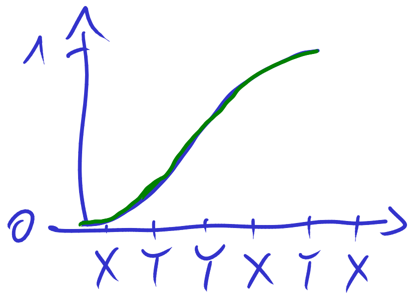
$$= \frac{mn}{N^2(N-1)} \left[(N-1) \sum_{i=1}^{N} a_i^2 - \sum_{\substack{i=1 \\ (i \neq j)}}^{N} \sum_{j=1}^{N} a_i a_j \right]$$

$$= \frac{mn}{N^2(N-1)} \left[N \sum_{i=1}^{N} a_i^2 - \underbrace{\sum_{i=1}^{N} \sum_{j=1}^{N} a_i a_j}_{= \left(\sum_{i=1}^{N} a_i\right)^2} \right]$$

$$= \frac{mn}{N^2(N-1)} \left[N \sum_{i=1}^{N} a_i^2 - \left(\sum_{i=1}^{N} a_i\right)^2 \right]$$

Lagealternativen

$H_0: F_X = F_Y$

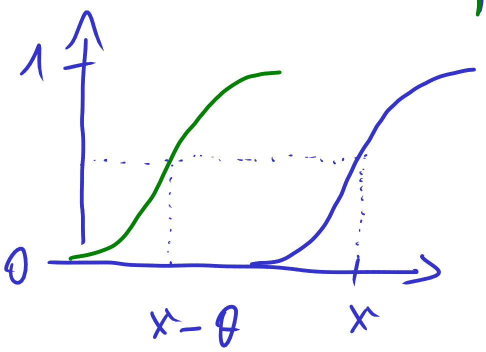


$V_i: \quad 1 \; 0 \; 0 \; 1 \; 0 \; 1$

$a_i: \quad \underline{\underline{1}} \; 2 \; 3 \; \underline{\underline{4}} \; 5 \; \underline{\underline{6}}$

$W_N = \sum_{i=1}^{N} a_i V_i = 11$

$H_1: F_X(x) = F_Y(x-\theta)$, für alle $x \in \mathbb{R}$, mit $\theta \neq 0$



$Y \, Y \, Y \quad X \, X \, X$

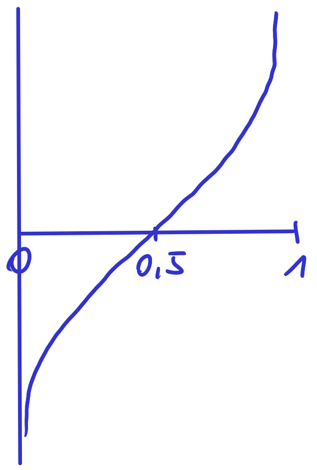
$V_i \quad 0 \; 0 \; 0 \quad 1 \; 1 \; 1$

$a_i \quad 1 \; 2 \; 3 \quad \underline{\underline{4}} \; \underline{\underline{5}} \; \underline{\underline{6}}$

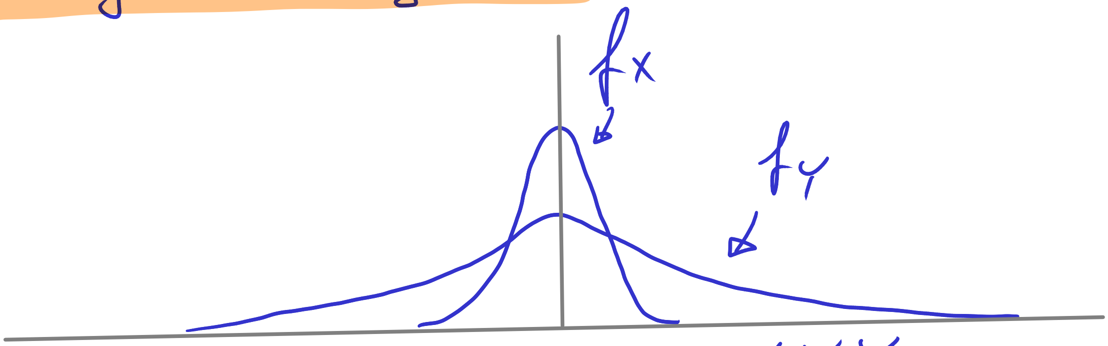
$W_N = \sum_{i=1}^{N} a_i V_i = 15$

a_i beim Test von v.d. Waerden

$$a_i = \Phi^{-1}\left(\frac{i}{N+1}\right) \quad ; \quad i = 1, \dots, N$$



Siegel - Tukey Test



	Y	Y	Y	Y	X	X	Y	X	Y	X	Y	Y	Y
V_i	0	0	0	0	1	1	0	1	0	1	0	0	0
a_i	1	4	5	8	$\underline{9}$	$\underline{12}$	13	$\underline{11}$	10	$\underline{7}$	6	3	2

$$S_N = 9 + 12 + 11 + 7$$

K-S Test im Zweistichprobenproblem

BSP: $m = 2$ und $n = 3$ $\quad (N = m + n = 5)$

$\Rightarrow$ Es gibt $\binom{n+m}{n} = \binom{5}{3} = 10$ z.B. $(X\,Y\,X\,Y\,Y)$

verschiedene Realisationsmöglichkeiten der kombinierten geordneten Stichprobe.

$$P_{H_0}(D_{m,n} = 1) = \frac{2}{10} \quad \left.\begin{cases} (X,X,Y,Y,Y) \\ (Y,Y,Y,X,X) \end{cases}\right\} D_{m,n} = 1$$

$$P_{H_0}\left(D_{m,n} = \tfrac{2}{3}\right) = \frac{4}{10} \quad \left.\begin{cases} (Y,X,X,Y,Y) \\ (Y,Y,X,X,Y) \\ (Y,Y,X,Y,X) \\ (X,Y,X,Y,Y) \end{cases}\right\} D_{m,n} = \frac{2}{3}$$

$$P_{H_0}\left(D_{m,n} = \tfrac{1}{2}\right) = \frac{3}{10} \quad \left.\begin{cases} (X,Y,Y,X,Y) \\ (X,Y,Y,Y,X) \\ (Y,X,Y,Y,X) \end{cases}\right\} D_{m,n} = \frac{1}{2}$$

$$P_{H_0}\left(D_{min}=\frac{1}{3}\right)=\frac{1}{10}\quad\{(Y,X,Y,X,Y)\}\ D_{min}=\frac{1}{3}$$

Übungsblatt #2

Aufg 2 $\quad Cov(r(X_i), r(X_j)) \quad i\neq j$

Einzelne Schritte:

- $P(r(X_i)=k,\ r(X_j)=\ell) =$

$$= \underbrace{P(r(X_i)=k \mid r(X_j)=\ell)}_{\frac{1}{n-1}} \cdot \underbrace{P(r(X_j)=\ell)}_{\frac{1}{n}} = \frac{1}{n(n-1)}$$

- $\displaystyle\sum_{\substack{k=1\\k\neq\ell}}^{n}\sum_{\ell=1}^{n} k\ell = ?$

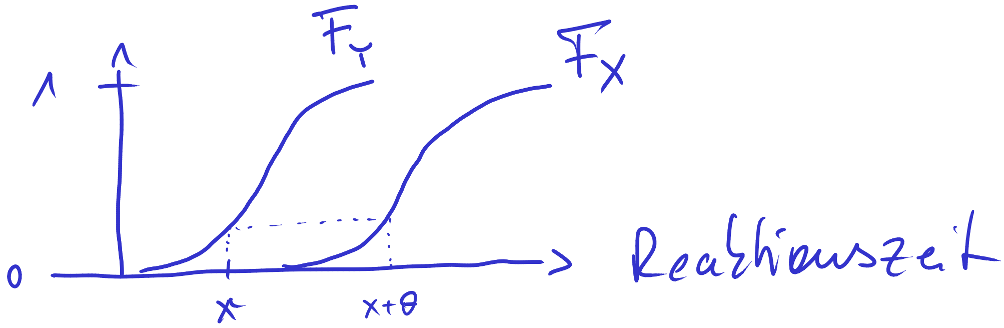
$$\begin{bmatrix} \cancel{11} & 12 & \cdots & 1n \\ 21 & \cancel{22} & \cdots & 2n \\ \vdots & & \ddots & \\ & & & \cancel{nn} \end{bmatrix} \begin{matrix} \rightarrow & 1\cdot\frac{n(n+1)}{2} - 1\cdot 1 \\ \rightarrow & 2\ \frac{n(n+1)}{2} - 2\cdot 2 \\ \\ \rightarrow & n\ \frac{n(n+1)}{2} - n\cdot n \end{matrix}$$

für die kte Zeile: $\quad k\,\frac{n(n+1)}{2} - kk$

Summe über alle Zeilen: $\quad \displaystyle\sum_{k=1}^{n}\left(k\,\frac{n(n+1)}{2} - kk\right) =$

$$= \frac{n(n+1)}{2} \sum_{k=1}^{n} k - \sum_{k=1}^{n} k^2$$

Aufg 3



a) $H_0 : F_Y = F_X$ $\qquad H_1 : F_Y(x) = F_X(x+\theta)$ für $\theta > 0$

$H_0 : F_Y(x) = F_X(x+\theta)$ für $\theta \leq 0$

b) [Notiz: $E(W_N) = E\left(\sum_{i=1}^{m} r(X_i)\right)$

$$= \sum_{i=1}^{m} E(r(X_i))$$

$$= \sum_{i=1}^{m} \left(\sum_{k=1}^{N} \frac{1}{N} k \right)$$

$$= m \left(\frac{1}{N} \sum_{k=1}^{N} k \right)$$

$$= m \, \frac{1}{N} \, \frac{N(N+1)}{2} = \frac{m(N+1)}{2}$$]

Kap 3.

$$\hat{\theta}_n = g(X_1, \dots, X_n)$$

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

$$E(\bar{X}_n) = \frac{1}{n}\sum_{i=1}^{n} E(X_i)$$

Da $X_1, \dots, X_n$ alle iid wie $X \sim F_X$

mit $E(X) = \mu$

$$= \frac{1}{n}\sum_{i=1}^{n} \mu = \frac{n}{n}\mu = \mu$$

Folie 18

$$MSE(\hat{\theta}_n) = E\left[(\hat{\theta}_n - \theta)^2\right] =$$

$$= E\left[\left((\hat{\theta}_n \underbrace{- E(\hat{\theta}_n)) + (E(\hat{\theta}_n)}_{=0} - \theta)\right)^2\right]$$

$$= E\left[(\hat{\theta}_n - E(\hat{\theta}_n))^2 + 2(\hat{\theta}_n - E(\hat{\theta}_n))(E(\hat{\theta}_n) - \theta) + (\underbrace{E(\hat{\theta}_n - \theta)}_{Bias(\hat{\theta}_n) = E(\hat{\theta}_n) - \theta})^2\right]$$

$$= \underbrace{E\left[(\hat{\theta}_n - E(\hat{\theta}_n))^2\right]}_{Var(\hat{\theta}_n)} + \underbrace{2(E(\hat{\theta}_n) - \theta)\underbrace{E\left[\hat{\theta}_n - E(\hat{\theta}_n)\right]}_{E(\hat{\theta}_n) - E(\hat{\theta}_n)}}_{=0} + (Bias(\hat{\theta}_n))^2$$

$$= Var(\hat{\theta}_n) + (Bias(\hat{\theta}_n))^2$$

Konsistenz

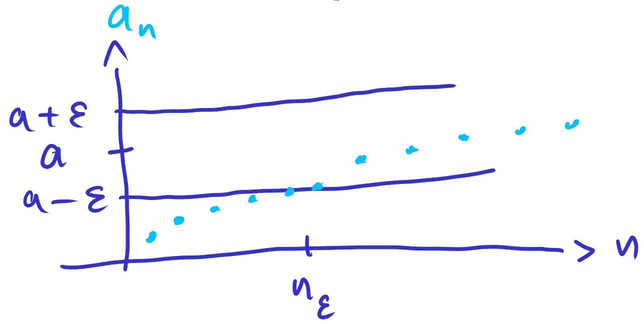
Ausflug zu deterministischen Folgen

$$a_n \longrightarrow a \quad \text{für } n \to \infty$$

"a_n konvergiert zu a"

D.h. Für hinreichend große n, ist und bleibt a_n beliebig nahe bei a.

$\varepsilon > 0$



Für jedes $\varepsilon > 0$ gibt es ein n_ε, sodass $|a_n - a| \leq \varepsilon$ für alle $n \geq n_\varepsilon$

$\iff a_n \longrightarrow a$, für $n \to \infty$

⚠ $\hat{\theta}_n$ ist jedoch eine Zufallsvariable!

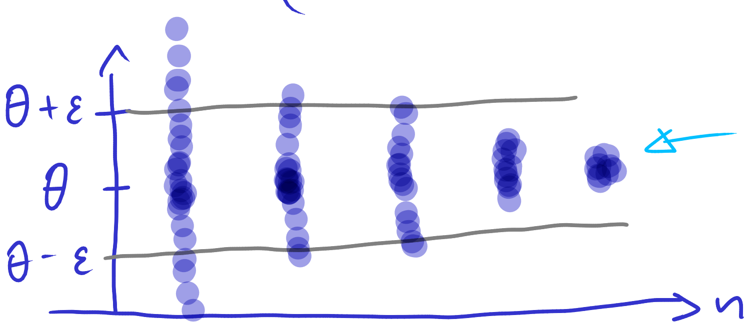
↳ (Eine) Lösung: "Konvergenz in Wahrscheinlichkeit"

Konvergenz in Wahrscheinlichkeit

$$\hat{\theta}_n \xrightarrow[p]{} \theta, \quad n \to \infty$$

$\iff$ D.h. für jedes $\varepsilon > 0$

$$P(|\hat{\theta}_n - \theta| < \varepsilon) \longrightarrow 1 \quad \text{für } n \to \infty$$



Realisationen von $\hat{\theta}_n$ für verschiedene n.

Behauptung $\boxed{\hat{\theta}_n \xrightarrow[MSE]{} \theta} \Rightarrow \boxed{\hat{\theta}_n \xrightarrow[p]{} \theta}$

D.h. $\boxed{\lim_{n\to\infty} E\left[(\hat{\theta}_n - \theta)^2\right] = 0} \Rightarrow \boxed{\lim_{n\to\infty} P\left[|\hat{\theta}_n - \theta| < \varepsilon\right] = 1}$

Beweis:

Tool: Markov-Ungleichung

$$P(X \geq c) \leq \frac{E(X)}{c}$$

mit $X \geq 0$ — Zufallsvariable

und $0 < c < \infty$ — Konstante

$$P\Big[\underbrace{\overbrace{(\hat{\theta}_n - \theta)^2}^{=X} \geq \overbrace{\varepsilon^2}^{c}}_{\Leftrightarrow\ |\hat{\theta}_n - \theta| \geq \varepsilon}\Big] \leq \frac{E\left[(\hat{\theta}_n - \theta)^2\right]}{\varepsilon^2}$$

$$\Leftrightarrow P\left[|\hat{\theta}_n - \theta| \geq \varepsilon\right] \leq \frac{\overbrace{E\left[(\hat{\theta}_n - \theta)^2\right]}^{=MSE(\hat{\theta}_n)}}{\varepsilon^2}$$

Dies zeigt die Behauptung, denn:

1.) $\lim_{n\to\infty} \frac{E\left[(\hat{\theta}_n - \theta)^2\right]}{\varepsilon^2} = 0$

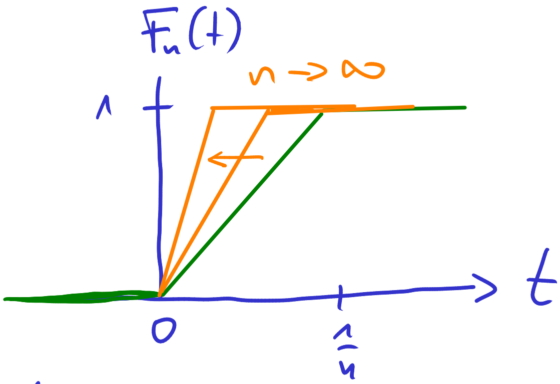
2.) $0 \leq P\left[|\hat{\theta}_n - \theta| \geq \varepsilon\right]$

$\square$

Konvergenz in Verteilung

Def $\lim_{n \to \infty} F_n(t) = F(t)$ „für alle Stetigkeitspunkte t von F."

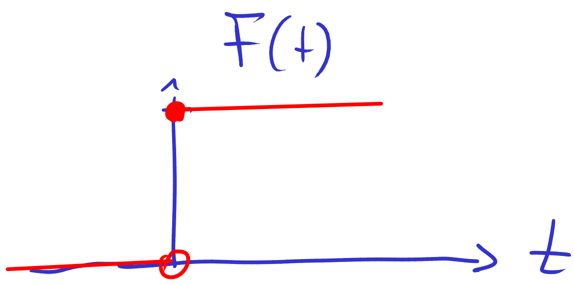
<u>BSP:</u>



$Z_n \sim U[0, \frac{1}{n}]$

$$F_n(t) = \begin{cases} 0, & t < 0 \\ tn, & 0 \le t \le \frac{1}{n} \\ 1, & t > \frac{1}{n} \end{cases}$$

$Z_\infty = 0$



$$(F_\infty(t) =)\ F(t) = \begin{cases} 0, & t < 0 \\ 1, & t \ge 0 \end{cases}$$

D.h. für alle $t > 0$ existiert ein m_t, sodass $F_n(t) = F(t) = 1$, für alle $n \ge m_t$.

Ljapunov

$$\frac{\frac{1}{n}\sum_{i=1}^{n}(X_i-\mu_i)}{\frac{1}{n}\left(\sum_{i=1}^{n}\sigma_i^2\right)^{1/2}} = \frac{\frac{1}{n}\sum_{i=1}^{n}(X_i-\mu_i)}{\frac{1}{\sqrt{n}}\frac{1}{\sqrt{n}}\left(\sum_{i=1}^{n}\sigma_i^2\right)^{1/2}} =$$

$$= \frac{\frac{1}{n}\sum_{i=1}^{n}(X_i-\mu_i)}{\frac{1}{\sqrt{n}}\left(\frac{1}{n}\sum_{i=1}^{n}\sigma_i^2\right)^{1/2}} = \frac{\sqrt{n}\left(\frac{1}{n}\sum_{i=1}^{n}(X_i-\mu_i)\right)}{\left(\frac{1}{n}\sum_{i=1}^{n}\sigma_i^2\right)^{1/2}}$$

ZGS von Lindeberg-Lévy: $E(X_i)=\mu$ für alle $i=1,\dots,n$

$Var(X_i)=\sigma^2$ für alle $i=1,\dots,n$

$$\frac{\sqrt{n}(\bar{X}_n-\mu)}{\sigma} \xrightarrow[L]{} N(0,1), \text{ für } n\to\infty$$

Tschebyscheff

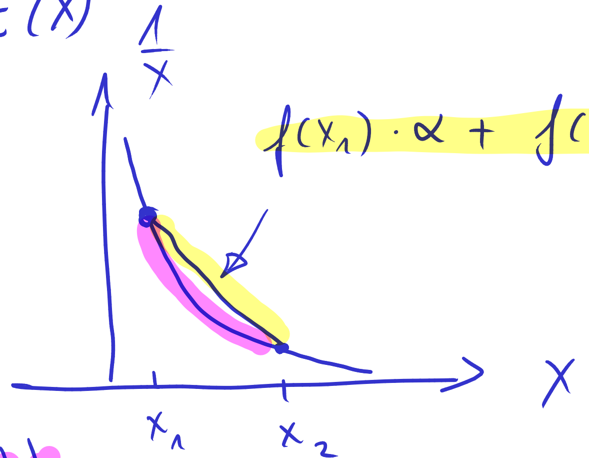
$$P(|X-\mu|>k\sigma_x) \le \frac{1}{k^2} \left(=\left(\frac{1}{k}\right)^2\right), k>0$$

mit $k=\frac{\tilde{k}}{\sigma_x} \iff \tilde{k}=k\cdot\sigma_x$, $\tilde{k}>0$

$$P(|X-\mu|>\tilde{k}) \le \frac{\sigma_x^2}{\tilde{k}^2} \quad \text{mit } \sigma_x^2=Var(X)$$

Jensens Ungleichung

$$E\left(\frac{1}{X}\right) \geq \frac{1}{E(X)}$$



$f(x_1)\cdot\alpha + f(x_2)(1-\alpha)$, $\alpha \in (0,1)$

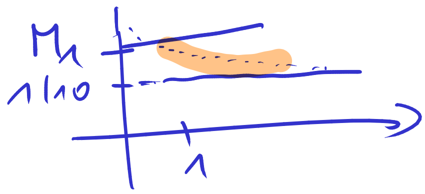
$$f(x_1\alpha + x_2(1-\alpha)) \leq f(x_1)\alpha + f(x_2)(1-\alpha)$$

Deterministische Ordnungssymbole

$z_n = O(1)$

"$|z_n|$ ist beschränkt für alle hinreichend großen n."

$\Leftrightarrow$ "$\exists\, M_m$, sodass $|z_n| \leq M_m$ für alle $n \geq m$."



BSP $\quad z_n = \frac{1}{n} + \frac{1}{10}$

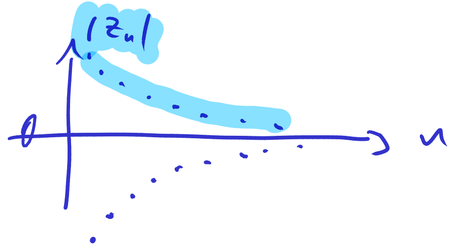
$\Rightarrow \quad |z_n| \leq M_1 = 1{,}1$ für alle $n \geq 1$

$\Rightarrow \quad z_n = O(1)$

$z_n = o(1)$

" $|z_n|$ ist eine "Nullfolge" (d.h. $|z_n| \xrightarrow[n\to\infty]{} 0$)."

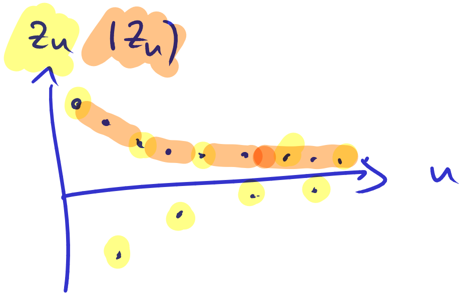
BSP



a) $z_n = -\frac{1}{n}$

$\Rightarrow |z_n| \longrightarrow 0$ für $n \to \infty$

$\Leftrightarrow z_n = o(1)$



b) $z_n = \frac{1}{n}(-1)^n$

$|z_n| \longrightarrow 0, \; n \to \infty$

$\Leftrightarrow z_n = o(1)$

Beachte:

$$z_n = o(1) \quad \Rightarrow \quad z_n = O(1)$$

$$z_n = O(r_n) \quad \Leftrightarrow \quad \frac{z_n}{r_n} = O(1)$$

$\Leftrightarrow$ " $|z_n|$ ist von gleicher oder kleinerer Größenordnung wie/als $|r_n|$."

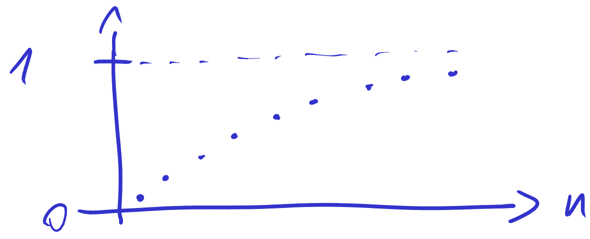
$\Leftrightarrow$ " $\frac{|z_n|}{|r_n|}$ ist beschränkt für hinreichend große n"

$\left(\exists M_m > 0, \text{ sodass } 0 \leq \frac{|z_n|}{|r_n|} \leq M_m \text{ für alle } n \geq m\right)$

BSP:

$$z_n = \frac{1}{100+n^2}$$

$$r_n = n^{-2}$$



$$\frac{|z_n|}{|r_n|} = \frac{n^2}{100+n^2}$$

$$\Rightarrow \quad 0 \leq \frac{|z_n|}{|r_n|} \leq M = 1 \quad \text{für alle } n \geq 1$$

$$\Leftrightarrow \quad \frac{z_n}{r_n} = O(1)$$

$$\Leftrightarrow \quad z_n = O(r_n)$$

$$z_n = o(r_n) \quad \Leftrightarrow \quad \frac{z_n}{r_n} = o(1)$$

$\Leftrightarrow$ „ $|z_n|$ ist von kleinerer Größenordnung als $|r_n|$."

$\Leftrightarrow$ „ $\frac{|z_n|}{|r_n|}$ ist eine Nullfolge."

BSP $\quad z_n = \frac{1}{n^2} \qquad r_n = \frac{1}{n}$

$$\frac{|z_n|}{|r_n|} = \frac{n}{n^2} = \frac{1}{n} \xrightarrow[n \to \infty]{} 0$$

$$\Rightarrow \quad z_n = o(r_n)$$

BSP $\quad z_n = \sum_{i=1}^{n} i = O(n^2)$

$$z_n = \sum_{i=1}^{n} i = \frac{n(n+1)}{2} = \frac{n^2+n}{2}$$

$$r_n = n^2$$

$$\frac{|z_n|}{|r_n|} = \frac{\frac{n^2+n}{2}}{n^2} = \dots = \frac{1}{2} + \frac{1}{2n}$$

$$\Rightarrow \frac{|z_n|}{|r_n|} \leq M = 1 \quad \text{for all } n \geq 1$$

$$\Rightarrow \frac{n^2+n}{2} = O(n^2)$$

BSP $\quad z_n = \sum_{i=1}^{n} i = o(n^3)$

$$z_n = \frac{n^2+n}{2}$$

$$r_n = n^3$$

$$\frac{|z_n|}{|r_n|} = \frac{1}{2} n^{-1} + \frac{1}{2} n^{-2} \xrightarrow[n \to \infty]{} 0$$

$$\Rightarrow \frac{z_n}{r_n} = o(1) \Leftrightarrow z_n = o(r_n)$$

$Z_n = O_p(1)$

$\Leftrightarrow$ " $|Z_n|$ ist stochastisch beschränkt für alle hinreichend großen n."

BSP $\quad Z_n \sim N(0,1)$ für alle n

$$\Rightarrow \quad P\left(|Z_n| > z_{1-\frac{\varepsilon}{2}}\right) = \varepsilon \quad \text{für all } n \geq 1$$

$$\Leftrightarrow \quad Z_n = O_p(1)$$

$Z_n = o_p(1)$

$\Leftrightarrow \quad Z_n \xrightarrow[p]{} 0$ für $n \to \infty$

BSP $\quad \bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$ mit $X_i \sim N(\mu, \sigma^2)$, für all i

$$Z_n = \bar{X}_n - \mu \xrightarrow[p]{n\to\infty} 0$$

$$\Leftrightarrow \quad Z_n = \bar{X}_n - \mu = o_p(1)$$

$$(\bar{X}_n - \mu) = O_p(n^{-1/2})$$

$$\Leftrightarrow \sqrt{n}(\bar{X}_n - \mu) = O_p(1)$$

Laut dem zentralen Grenzwertsatz (angenommen alle Voraussetzungen sind erfüllt)

$$\sqrt{n}(\bar{X}_n - \mu) \xrightarrow[L]{n \to \infty} Z \sim N(0, \sigma^2)$$

$$\Rightarrow \sqrt{n}(\bar{X}_n - \mu) = O_p(1)$$

$$\Leftrightarrow (\bar{X}_n - \mu) = O_p(n^{-1/2})$$

Integration durch Substitution

BSP

$$\int_0^2 x\, e^{x^2}\, dx = \frac{1}{2} \int_0^2 \underbrace{\boxed{2x}}_{g'(x)} \underbrace{\boxed{e^{x^2}}}_{f(g(x))} dx =$$

Substitution

$$\underset{u(x)}{u} = x^2 \quad \Rightarrow \quad \frac{du}{dx} = 2x \quad \Leftrightarrow \quad \frac{1}{2x} du = dx$$

$$u(2) = 2^2 = 4$$
$$u(0) = 0^2 = 0$$

$$= \frac{1}{2} \int_0^4 \cancel{2x}\; e^u\; \frac{1}{\cancel{2x}} du = \frac{1}{2} \int_0^4 e^u du = \frac{1}{2} \left[e^u \right]_0^4$$

$$= \frac{1}{2}(e^4 - 1)$$

Partielle Integration

BSP

$$\int_1^2 \overbrace{\boxed{x^2}}^{g'(x)} \overbrace{\boxed{\ln(x)}}^{f(x)} \, dx = \left[\frac{1}{3}x^3 \cdot \ln(x)\right]_1^2 - \int_1^2 \frac{1}{3}x^3 \, \frac{1}{x} \, dx =$$

$$= \left(\frac{8}{3}\ln(2) - 0\right) - \frac{1}{3}\int_1^2 x^2 \, dx$$

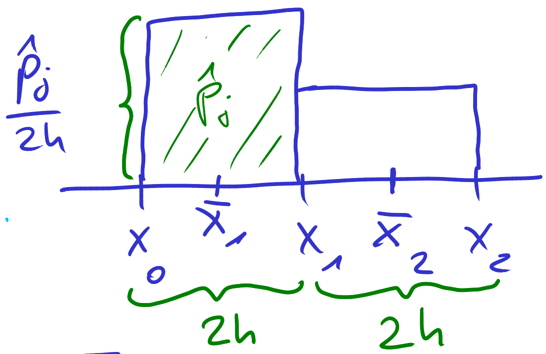
$$= \left(\frac{8}{3}\ln(2)\right) - \frac{1}{3}\left[\frac{1}{3}x^3\right]_1^2 = \frac{8}{3}\ln(2) - \left(\frac{8}{9} - \frac{1}{9}\right)$$

Kap 4 Nichtparametrische Dichteschätzung

Ann: $X_1, \dots, X_n \overset{iid}{\sim} X$ mit $X \sim F$ (bzw. $X \sim f = F'$)

Frage: $f = ?$

Das Histogramm



$\bar{x}_j = \frac{x_{j-1} + x_j}{2}$, wobei $(x_j)_{j=0,1,2,\dots}$ die Balkenränder definiert.

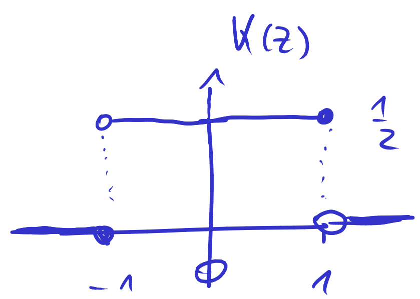
$$\hat{f}_{hist}(\bar{x}_j) = \frac{\# X_i \in (x_{j-1}, x_j]}{2hn}$$

$$= \frac{1}{2h} \cdot \hat{p}_j, \quad \text{wobei } \hat{p}_j = \frac{\# X_i \in (x_{j-1}, x_j]}{n}$$

$$\hat{f}_{hist}(\bar{x}_j) = \frac{1}{2h} \cdot \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}(x_{j-1} < X_i \leq x_j)$$

$$= \frac{1}{nh} \sum_{i=1}^{n} \frac{1}{2} \mathbb{1}(x_{j-1} < X_i \leq x_j)$$

$$= \frac{1}{nh} \sum_{i=1}^{n} K\left(\frac{X_i - \bar{x}_j}{h}\right)$$



für $K(z) = \begin{cases} \frac{1}{2} & \text{für } z \in (-1, 1] \\ 0 & \text{sonst} \end{cases}$

Da $K(z)$ achsensymm.: $K\left(\frac{X_i - \bar{x}_j}{h}\right) = K\left(\frac{\bar{x}_j - X_i}{h}\right)$

Wie gut ist $\frac{F(x+h) - F(x)}{h} \approx F'(x)$?

Behauptung $F'(x) = \frac{F(x+h) - F(x)}{h} + O(h)$ für $h \to 0$

Bew:

Taylorentwicklung von F um x für $h \to 0$:

$$F(x+h) = F(x) + F'(x)\overset{x+h-x}{h} + \frac{1}{2} F''(x) h^2 + o(h^2)$$

$$\Rightarrow \frac{F(x+h) - F(x)}{h} = \frac{F'(x) h + \frac{1}{2} F''(x) h^2 + o(h^2)}{h}$$

$$\left[\begin{array}{l} v_n = O(h^{-1}) \quad \text{und} \quad z_n = o(h^2) \\ \Rightarrow \quad v_n z_n = o(h) \end{array} \right]$$

$$\Rightarrow \frac{F(x+h) - F(x)}{h} = F'(x) + \underbrace{\overbrace{\frac{1}{2} F''(x) h}^{O(h)} + o(h)}_{O(h)}$$

$$\Rightarrow \frac{F(x+h) - F(x)}{h} = \underbrace{F'(x)}_{f(x)} + O(h)$$

Bias des Kerndichteschätzers

$$\hat{f}_{nh}(x) = \frac{1}{nh} \sum_{i=1}^{n} K\left(\frac{x - X_i}{h}\right), \quad x \in \mathbb{R}$$

Ann:
- $X_1, \dots, X_n \overset{iid}{\sim} X$ mit $X \sim f$
- f sei zweimal stetig differenzierbar
- $n \to \infty$, $h = h_n \to 0$, $nh_n \to \infty$
- K wie auf Folie 35

Satz: Unter obigen Annahmen lautet der Bias von $\hat{f}_{nh}(x)$:

$$\operatorname{Bias}(\hat{f}_{nh}(x)) = E(\hat{f}_{nh}(x)) - f(x)$$

$$= \frac{1}{2} f''(x)\, \nu_2(K)\, h^2 + o(h^2),$$

wobei $\nu_2(K) = \int K(y)\, y^2\, dy$

Beweis:

$$E[\hat{f}_{nh}(x)] = \frac{1}{nh} \sum_{i=1}^{n} E\left[K\left(\frac{x - X_i}{h}\right)\right]$$

$$\star \quad = \frac{1}{h} E\left[K\left(\frac{x - X}{h}\right)\right] \qquad \text{iid}$$

$$= \frac{1}{h} E\left[K\left(\frac{X - x}{h}\right)\right] \qquad | \; K \text{ symmetrisch}$$

$$= \frac{1}{h} \int K\left(\frac{z - x}{h}\right) f(z)\, dz$$

Subst $\frac{z - x}{h} = y \Rightarrow \frac{dy}{dz} = \frac{1}{h} \Leftrightarrow dz = h\, dy$

$\Leftrightarrow z = x + hy$

$$= \frac{\cancel{h}}{\cancel{h}} \int K(y)\, f(x+hy)\, dy$$

$$f(x+hy) = f(x) + f'(x)\, hy + \frac{1}{2} f''(x)\, h^2 y^2 + o(h^2)$$

$$= \int K(y) \left[f(x) + f'(x)\, hy + \frac{1}{2} f''(x)\, h^2 y^2 + o(h^2) \right] dy$$

$$= f(x) \underbrace{\int K(y)\, dy}_{=1} + f'(x)\, h \underbrace{\int K(y)\, y\, dy}_{=0} + \frac{1}{2} f''(x)\, h^2 \underbrace{\int K(y)\, y^2\, dy}_{=\nu_2(K)} + o(h^2)$$

$$\Rightarrow E(\hat{f}_{nh}(x)) = f(x) + \frac{1}{2} f''(x)\, \nu_2(K)\, h^2 + o(h^2)$$

$\square$

Asymptotischer Bias:

$$ABias(\hat{f}_{nh}(x)) = \frac{1}{2} f''(x)\, \nu_2(K)\, h^2$$

<u>denn:</u>

$$\frac{Bias(\hat{f}_{nh}(x))}{ABias(\hat{f}_{nh}(x))} = 1 + o(h^2)$$

<u>Satz:</u> Unter obiger Annahme lautet die Varianz von $\hat{f}_{nh}(x)$:

$$Var(\hat{f}_{nh}(x)) = \frac{1}{n} \frac{f(x)\, R(K)}{h} + O\left(\frac{1}{n}\right),$$

wobei $R(K) = \int (K(y))^2\, dy$

Beweis:

$$\mathrm{Var}\left(\hat{f}_{nh}(x)\right) = \frac{1}{n^2 h^2} \sum_{i=1}^{n} \mathrm{Var}\left[K\left(\frac{x-X_i}{h}\right)\right] \quad \text{iid}$$

$$= \frac{1}{n h^2} \mathrm{Var}\left[K\left(\frac{x-X}{h}\right)\right]$$

$$= \frac{1}{nh^2} E\left[\left(K\left(\frac{x-X}{h}\right)\right)^2\right] - \frac{1}{n}\Big\{\underbrace{\frac{1}{h} E\left[K\left(\frac{x-X}{h}\right)\right]}_{E\left[\hat{f}_{nh}(x)\right] \text{ siehe } \star}\Big\}^2$$

$$= \underbrace{\frac{1}{nh^2} E\left[\left(K\left(\frac{x-X}{h}\right)\right)^2\right]}_{\text{NR}} - \underbrace{\frac{1}{n}\Big\{\underbrace{E\left(\hat{f}_{nh}(x)\right)}_{0 \le \text{const} < \infty}\Big\}^2}_{\frac{1}{n}\text{const} = O(n^{-1})}$$

NR:

$$\frac{1}{nh^2} E\left[\left(K\left(\frac{x-X}{h}\right)\right)^2\right] \overset{K \text{ symmetrisch}}{=} \frac{1}{nh^2} E\left[\left(K\left(\frac{X-x}{h}\right)\right)^2\right] =$$

$$= \frac{1}{nh^2} \int \left(K\left(\frac{z-x}{h}\right)\right)^2 f(z)\, dz = \frac{h}{nh^2} \int (K(y))^2 \underbrace{f(x+hy)}_{f(x)+O(h)}\, dy$$

Taylorapprox 1. Ordnung

$$= \frac{1}{nh} f(x) \underbrace{\int (K(y))^2 dy}_{R(K)} + \underbrace{\frac{1}{nh} R(K)\, O(h)}_{= O(\frac{1}{n})}$$

Damit:

$$\mathrm{Var}(\hat{f}_{nh}(x)) = \frac{1}{nh} f(x) R(K) + O(n^{-1}) + O(n^{-1})$$

$$= \frac{1}{nh} f(x) R(K) + \underbrace{O(\tfrac{1}{n})}_{= o(\frac{1}{nh})}$$

Asymptotische Varianz

$$\mathrm{AVar}(\hat{f}_{nh}(x)) = \frac{R(K) f(x)}{nh}$$

denn:

$$\frac{\mathrm{Var}(\hat{f}_{nh}(x))}{\mathrm{AVar}(\hat{f}_{nh}(x))} = 1 + o\left(\frac{1}{nh}\right)$$

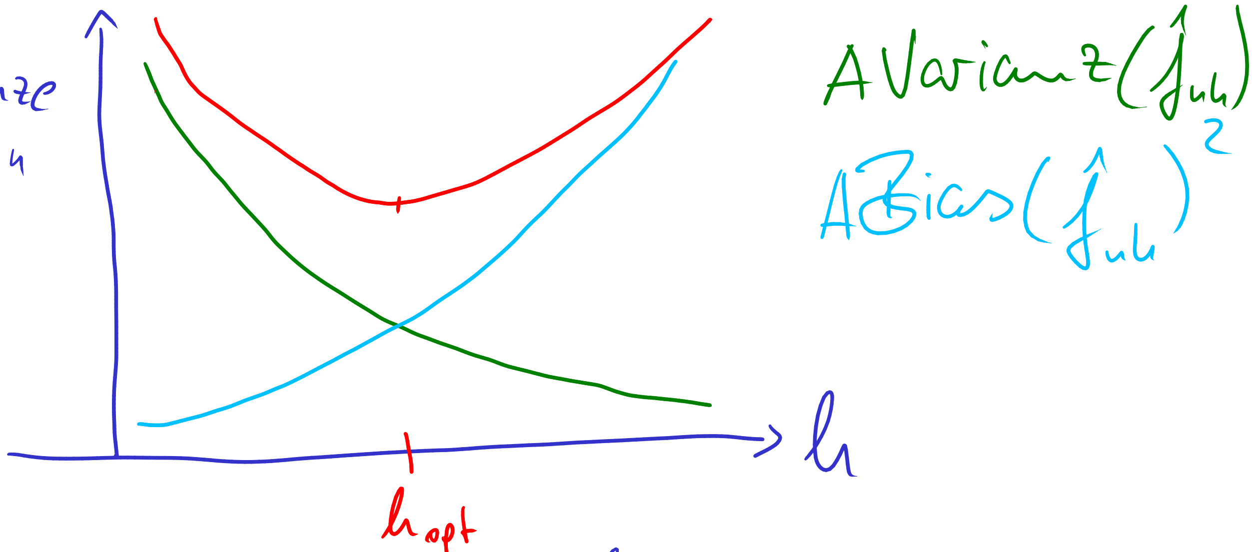
Asymptotischen mittleren quadratischen Fehler von $\hat{f}_{nh}(x)$:

$$\mathrm{AMSE}(\hat{f}_{nh}(x)) = \frac{f(x) R(K)}{nh} + \left(\frac{1}{2} f''(x) \nu_2(K) h^2\right)^2$$

$$\mathrm{AMISE}(\hat{f}_{nh}) = \int \mathrm{AMSE}(\hat{f}_{nh}(x))\, dx$$

$$\mathrm{AMISE}(\hat{f}_{nh}) = \frac{R(K)}{nh} + \frac{1}{4}(\mu_2(K))^2 h^4 \int (f''(x))^2\,dx$$

„Bias-Varianz Trade-off"



$\longrightarrow$ minimieren bzgl. h.

$\Longrightarrow$ h_{opt}.

Kreuzvalidierung (Ziel: h_{ISE})

Behauptung:

$\frac{1}{n}\sum_{i=1}^{n} \hat{f}_{nh,-i}(X_i)$ ist

ein „vernünftiger" Schätzer für $\int \hat{f}_{nh}(x) f(x)\,dx$,

zumindest in dem Sinne, dass

$$\boxed{E\left[\frac{1}{n}\sum_{i=1}^{n}\hat{f}_{nh,-i}(X_i)\right] = E\left[\int \hat{f}_{nh}(x) f(x)\,dx\right]}$$

Beweis zu dieser Behauptung

NR: $\frac{1}{n}\sum_{i=1}^{n} \hat{f}_{nh,-i}(X_i) =$

$$= \frac{1}{n}\sum_{i=1}^{n}\left[\frac{1}{(n-1)h}\sum_{\substack{j=1\\(j\neq i)}}^{n} K\left(\frac{X_j - X_i}{h}\right)\right]$$

$$= \frac{1}{n(n-1)h}\sum_{i=1}^{n}\sum_{\substack{j=1\\(j=i)}}^{n} K\left(\frac{X_j - X_i}{h}\right)$$

$X_1,\dots,X_n \overset{iid}{\sim} X$

$X_j \perp\!\!\!\perp X_i$

"unabhängig von"

$$E\left[\frac{1}{n}\sum_{i=1}^{n}\hat{f}_{nh,-i}(X_i)\right] =$$

wegen iid

$$= E\left[\frac{1}{h}K\left(\frac{X_i - X_j}{h}\right)\right] =$$

mit $i \neq j$

$$= \int\int \frac{1}{h}K\left(\frac{x-u}{h}\right)\underbrace{f(x)\,f(u)}_{\text{wegen Unabhängigkeit zwischen } X_i \text{ und } X_j}\,dx\,du$$

$$E\left[\int \frac{1}{h}K\left(\frac{x-X_j}{h}\right)f(x)\,dx\right] =$$

$$= E\left[\int \frac{1}{hn}\sum_{i=1}^{n} K\left(\frac{x-X_j}{h}\right) f(x)\,dx\right] = \text{wegen iid}$$

$$= E\left[\int \hat{f}_{nh}(x)\, f(x)\,dx\right]$$

$\square$

Übungsblatt 3