

Vorlesung 11a

Schätzen mit Verlass

Schätzen von Anteilen

Große Population (♀ und ♂)
mit unbekanntem Weibchenanteil p

In einer Stichprobe vom Umfang $n = 53$
gab es 23 Weibchen.

Wie zuverlässig ist der Schätzwert $\frac{23}{53}$?

Goldene Idee der Statistik:

In einem idealisiert gedachten Szenario interpretiert man den Schätzwert als Realisierung einer Zufallsvariable und rechnet mit deren Variabilität.

Deutung der Stichprobenziehung als p -Münzwurf

$$\hat{p} = \frac{B}{n}, \quad \text{mit } B := \text{Anzahl der "Erfolge"}$$

$$\sigma_{\hat{p}} = \frac{1}{\sqrt{n}} \sqrt{p(1-p)}$$

(B ist $\text{Bin}(n, p)$ -verteilt.)

Ein hübsche Illustration dieser Formel

(und ein Weg, wie man sie experimentell entdecken kann)

findet sich auf

<http://math.uni-frankfurt.de/ferebee/explorativ/Woche5.html>

Der Zentrale Grenzwertsatz besagt:

\hat{p} ist approximativ normalverteilt
mit Erwartungswert p und Standardabweichung $\sigma_{\hat{p}}$.

Also insbesondere:

$$\mathbf{P}_p(|p - \hat{p}| \leq 2\sigma_{\hat{p}}) \approx 0.95$$

$$\mathbf{P}_p(p \in [\hat{p} - 2\sigma_{\hat{p}}, \hat{p} + 2\sigma_{\hat{p}}]) \approx 0.95$$

In der Praxis ist auch $\sigma_{\hat{p}}$
(aus der *einen* vorliegenden Stichprobe) zu schätzen.

$$\sigma_{\hat{p}} = \frac{1}{\sqrt{n}} \sqrt{p(1-p)}$$

wird geschätzt durch

$$g := \frac{1}{\sqrt{n}} \sqrt{\hat{p}(1-\hat{p})}$$

$$\mathbf{P}_p(p \in [\hat{p} - 2\sigma_{\hat{p}}, \hat{p} + 2\sigma_{\hat{p}}]) \approx 0.95$$

überträgt sich auf

$$\mathbf{P}_p(p \in [\hat{p} - 2g, \hat{p} + 2g]) \approx 0.95$$

Das zufällige Intervall

$$I := \left[\hat{p} - \frac{2}{\sqrt{n}} \sqrt{\hat{p}(1 - \hat{p})}, \hat{p} + \frac{2}{\sqrt{n}} \sqrt{\hat{p}(1 - \hat{p})} \right]$$

ist ein

Konfidenzintervall für p

mit approximativer Überdeckungswahrscheinlichkeit 0.95

oder kurz ein

approximatives 95%-Konfidenzintervall für p .

Faustregel für die Anwendbarkeit: $n\hat{p} \geq 9$ und $n(1 - \hat{p}) \geq 9$.

R-Programme und Illustrationen zur exakten Überdeckungswahrscheinlichkeit von I findet man auf <http://math.uni-frankfurt.de/ferabee/explorativ/Woche6.html>

Schätzung des Erwartungswertes einer Verteilung auf \mathbb{R} (Lageschätzung)

$$m := \frac{1}{n}(x_1 + \cdots + x_n)$$

wird gedacht als eine Realisierung der Zufallsvariablen

$$\bar{X} := \frac{1}{n}(X_1 + \cdots + X_n)$$

mit X_1, \dots, X_n unabhängig, identisch verteilt

mit Erwartungswert μ und Standardabweichung σ .

Anders als bei der Anteilsschätzung ist hier σ i.a. keine Funktion von μ .

Der Zentrale Grenzwertsatz besagt:

$\frac{\bar{X} - \mu}{\sigma/\sqrt{n}}$ ist approximativ $N(0, 1)$ -verteilt.

Bei bekanntem σ ist also

$$\left[\bar{X} - \frac{2\sigma}{\sqrt{n}}, \bar{X} + \frac{2\sigma}{\sqrt{n}} \right]$$

ein approximatives 95%-Konfidenzintervall für μ .

In der Praxis hat man auch σ aus den Daten zu schätzen.

Wir wissen schon:

$$s^2 := \frac{1}{n-1} \left((X_1 - \bar{X})^2 + \dots + (X_n - \bar{X})^2 \right)$$

ist ein "erwartungstreuer" Schätzer für σ^2

Satz (W. Gosset (alias “Student”, 1908), R. Fisher (1924))

Sind X_1, \dots, X_n unabhängig und $N(\mu, \sigma^2)$ -verteilt,

dann ist $T := \frac{\bar{X} - \mu}{s/\sqrt{n}}$ so verteilt wie

$$T_{n-1} := \frac{N_0}{\sqrt{\frac{1}{n-1} (N_1^2 + \dots + N_{n-1}^2)}}$$

mit unabhängigen und $N(0, 1)$ verteilten N_0, \dots, N_{n-1} .

Bezeichnung: Die Verteilung von T_{n-1} heißt ***t*-Verteilung**
(oder ***Student-Verteilung***) mit $n - 1$ Freiheitsgraden.

Folgerung: Sind X_1, \dots, X_n unabhängig und $N(\mu, \sigma^2)$ -verteilt, dann ist für jedes $c > 0$:

$$\mathbf{P}(|T_{n-1}| \leq c) = \mathbf{P}\left(\left|\frac{\bar{X} - \mu}{s/\sqrt{n}}\right| \leq c\right) = \mathbf{P}\left(\mu \in \left[\bar{X} - \frac{cs}{\sqrt{n}}, \bar{X} + \frac{cs}{\sqrt{n}}\right]\right)$$

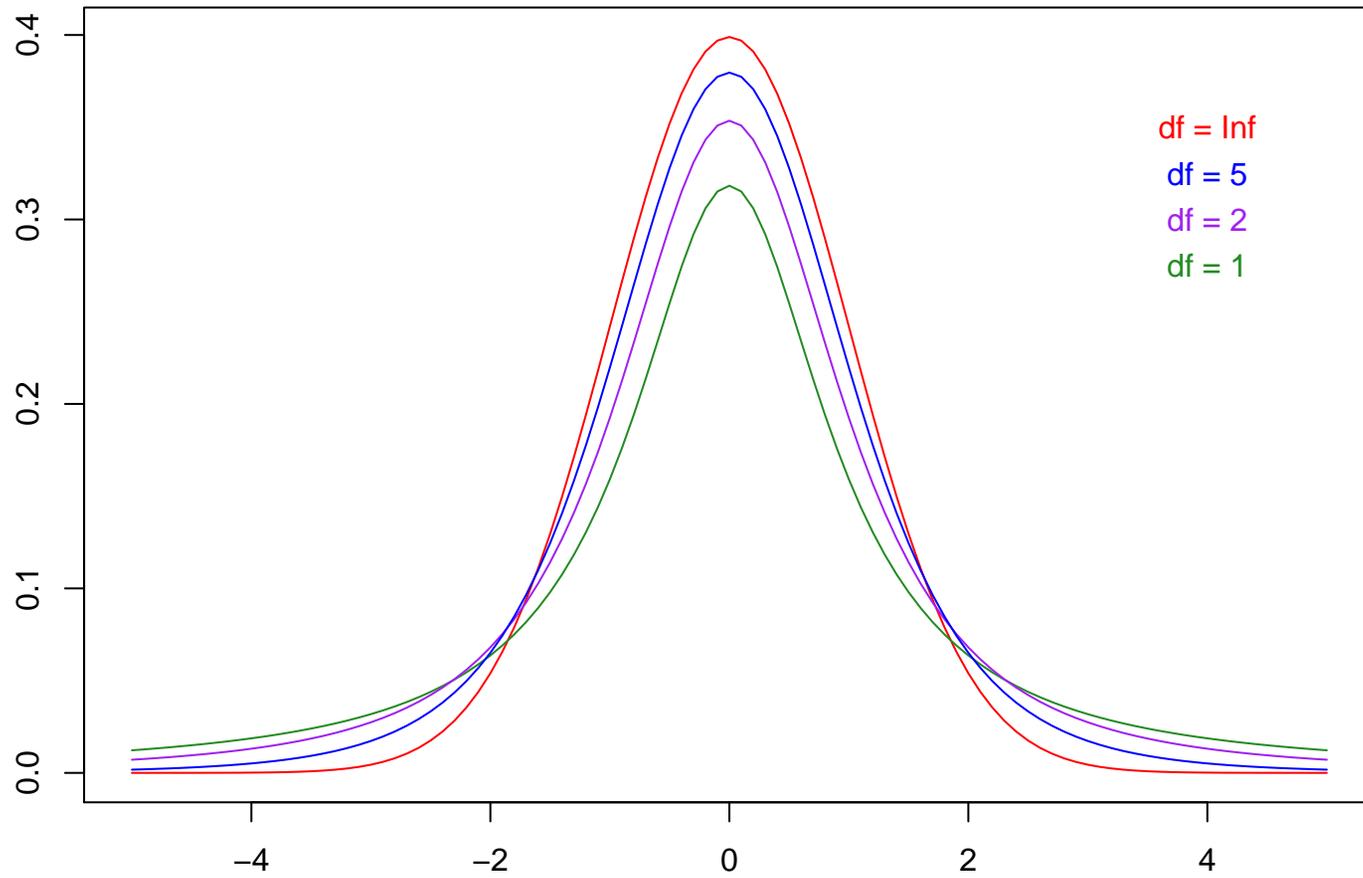
Wie groß muss c sein, damit sich hier 0.95 ergibt?

Für $n \rightarrow \infty$ ist T_{n-1} asymptotisch $N(0, 1)$ -verteilt (Gesetz der großen Zahlen im Nenner von T_{n-1}).

Je kleiner n , um so *breitschultriger* ist die Verteilung von T_{n-1}

Z.B. für $n = 6$: $\mathbf{P}(|T_5| \leq 2.57) = 0.95$.

Student's t: Dichtefunktionen



Beweis des Satzes von Gosset und Fisher: X_1, \dots, X_n ist von der Form

$$X_1 = \mu + \sigma Z_1, \dots, X_n = \mu + \sigma Z_n$$

mit Z_1, \dots, Z_n unabhängig und standard-normalverteilt. Also:

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

$$s = \sigma \sqrt{\frac{1}{n-1} (Z_1 - \bar{Z})^2 + \dots + (Z_n - \bar{Z})^2}$$

$$\frac{\bar{X} - \mu}{s/\sqrt{n}} = \frac{\sqrt{n}\bar{Z}}{\sqrt{\frac{1}{n-1} ((Z_1 - \bar{Z})^2 + \dots + (Z_n - \bar{Z})^2)}}$$

$\sqrt{n}\bar{Z}$ ist die Koordinate von $\vec{Z} := (Z_1, \dots, Z_n)$ bzgl. des Einheitsvektors in Richtung $(1, \dots, 1)$,

$(Z_1 - \bar{Z})^2 + \dots + (Z_n - \bar{Z})^2$ ist das Längenquadrat der Projektion orthogonal dazu.

Die Behauptung des Satzes folgt dann aus der Rotationssymmetrie der Verteilung von \vec{Z} wie in Aufgabe 36 (siehe auch Seite 138 im Buch).

Ein Konfidenzintervall für den Median

Eine Zahl ν heißt *Median* der Verteilung ρ auf \mathbb{R} ,
wenn $\rho((-\infty, \nu]) \geq 1/2$ **und** $\rho([\nu, \infty)) \geq 1/2$ gilt.

Die *Ordnungsstatistiken* $X_{(1)} \leq \dots \leq X_{(n)}$
sind die aufsteigend geordneten X_1, \dots, X_n

Ein Kandidat für ein **Konfidenzintervall für den Median** ist

$$[X_{(1+j)}, X_{(n-j)}]$$

mit $0 \leq j < n/2$.

$$\mathbf{P}_\rho(\nu \notin [X_{(1)}, X_{(n)}]) = \mathbf{P}_\rho(X_{(1)} > \nu) + \mathbf{P}_\rho(X_{(n)} < \nu) .$$

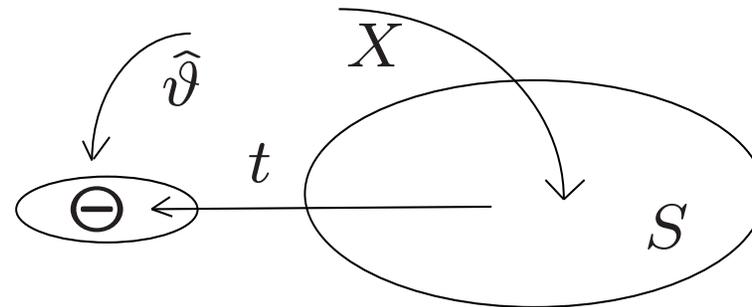
$$\mathbf{P}_\rho(X_{(1)} > \nu) \leq 2^{-n}$$

$$\mathbf{P}_\rho(X_{(n)} < \nu) \leq 2^{-n} .$$

Also:

$$\mathbf{P}_\rho(\nu \in [X_{(1)}, X_{(n)}]) \geq 1 - \frac{1}{2^{n-1}} .$$

Ein Logo der Statistik:



$$\mathbf{P}_{\vartheta}(X \in da) = \rho_{\vartheta}(da), \quad \vartheta \in \Theta$$

Θ ... *Parameterraum*

S ... *Beobachtungsraum*

$\hat{\vartheta} := t(X)$... *Schätzer für den Parameter ϑ*

Sei $m(\vartheta)$ ein reelles Parametermerkmal
und $I = I(X)$ ein aus den Daten konstruiertes Intervall.

Gilt für ein $\alpha \in (0, 1)$

$$\mathbf{P}_{\vartheta}(m(\vartheta) \in I) \geq 1 - \alpha \quad \text{für jedes } \vartheta \in \Theta$$

dann sagt man:

I ist ein **Konfidenzintervall** für $m(\vartheta)$ mit Niveau $1 - \alpha$,
es hält die *Überdeckungswahrscheinlichkeit* $1 - \alpha$ ein.