# Vorlesung 14a

# 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 $\frac{23}{53}$ als Schätzung für $p$ ?

Goldene Idee der Statistik:

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

Deutung der Stichprobenziehung als $p$-Münzwurf

Als *Schätzer* für $p$ betrachten wir die *Zufallsvariable*

$$\hat{p} := H := \frac{K}{n}, \quad \text{mit} \quad K := \text{Anzahl der “Erfolge”}.$$

$H$ ist die (relative) Häufigkeit der Erfolge.

$$\sigma_H = ?$$

$K$ ist Bin$(n, p)$-verteilt. Also:

$$\boxed{\sigma_H = \frac{1}{\sqrt{n}}\sqrt{p(1-p)}}$$

Der Zentrale Grenzwertsatz besagt:

$\widehat{H}$ ist approximativ normalverteilt

mit Erwartungswert $H$ und Standardabweichung $\sigma_H$.

Also insbesondere:

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

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

In der Praxis ist auch $\sigma_H$

(aus der *einen* vorliegenden Stichprobe) zu schätzen.

Ein Schätzer für $\sigma_H = \frac{1}{\sqrt{n}}\sqrt{p(1-p)}$ ist

$$G := \frac{1}{\sqrt{n}}\sqrt{H(1-H)}$$

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

überträgt sich auf

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

Das zufällige Intervall

$$I := [H - 2G,\, H + 2G]$$

ist ein

Konfidenzintervall für $p$

mit approximativer Überdeckungswahrscheinlichkeit 0.95

oder kurz ein

approximatives 95%-Konfidenzintervall für $p$.

In unserem Eingangsbeispiel ($n = 53,\ k = 23$)
waren die Realisierungen von $H$ und $G$ gleich

$$h = 23/53 = 0.43, \quad g = \sqrt{\frac{0.43 \cdot 0.57}{53}} = 0.07.$$

Als Realisierung von $I = [H - 2G,\ H + 2G]$ ergibt sich

$$[h - 2g, h + 2g] = [0.36, 0.57].$$

Man beachte:
Zufällig ist nicht der Parameter $p$, sondern das Intervall $I$.

Im Jargon der Statistik wird oft
sowohl der Schätzer (die Zufallvariable) $H$
als auch der Schätzwert (die Zahl) $h$
mit dem Symbol $\widehat{p}$ bezeichnet.

Auf der nächsten Folie,
die das Obige zsammenfasst,
steht $\widehat{p}$ für die Zufallsvariable $H$
(vgl. Buch S. 122):

Das zufällige Intervall

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

ist ein

Konfidenzintervall für $p$

mit approximativer Überdeckungswahrscheinlichkeit 0.95

oder kurz ein

approximatives 95%-Konfidenzintervall für $p$.

Dabei sollte $np(1-p)$ “groß” sein.
Eine Faustregel für die Anwendbarkeit ist: $nh \geq 9$ und $n(1-h) \geq 9$.

## 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 $X_1, \ldots, X_n$ unabhängig, identisch verteilt

mit Erwartungswert $\mu$ und Standardabweichung $\sigma$.

Anders als bei der Anteilschätzung ist hier $\sigma$ i.a. keine Funktion von $\mu$.

$\bar{X}$ ist ein Schätzer für $\mu$.

$\bar{X}$ hat Erwartungswert $\mu$ und Standardabweichung $\sigma/\sqrt{n}$.

Ein Schätzer für $\sigma^2$ ist

$$S^2 := \tfrac{1}{n-1}\left((X_1 - \bar{X})^2 + \cdots + (X_n - \bar{X})^2\right)$$

Im Jargon der Statistik schreibt man meist $s^2$ statt $S^2$ und verwendet $s^2$ auch zur Bezeichung von Schätzwerten (Realisierungen).

Der Zentrale Grenzwertsatz besagt:

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

Bei bekanntem $\sigma$ ist also für große $n$

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

ein approximatives 95%-Konfidenzintervall für $\mu$.

In der Praxis hat man auch $\sigma$ aus den Daten zu schätzen.

Für große $n$ ist auch

$$J := \left[\bar{X} - 2\frac{S}{\sqrt{n}}, \bar{X} + 2\frac{S}{\sqrt{n}}\right]$$

ein approximatives 95%-Konfidenzintervall für $\mu$

Dieses Intervall enthält keine unbekannten Parameter.

Seine Realisierung ist $|\bar{x} - 2f, \bar{x} + 2f]$.

Der *Standardfehler*

$$f := \frac{s}{\sqrt{n}}$$

ist ein Schätzwert für die

Standardabweichung $\sigma/\sqrt{n}$ des Stichprobenmittelwertes $\bar{X}$.

Der Zentrale Grenzwertsatz besagt:

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

Bei bekanntem $\sigma$ ist also

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

ein approximatives 95%-Konfidenzintervall für $\mu$.

In der Praxis hat man auch $\sigma$ aus den Daten zu schätzen.

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

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

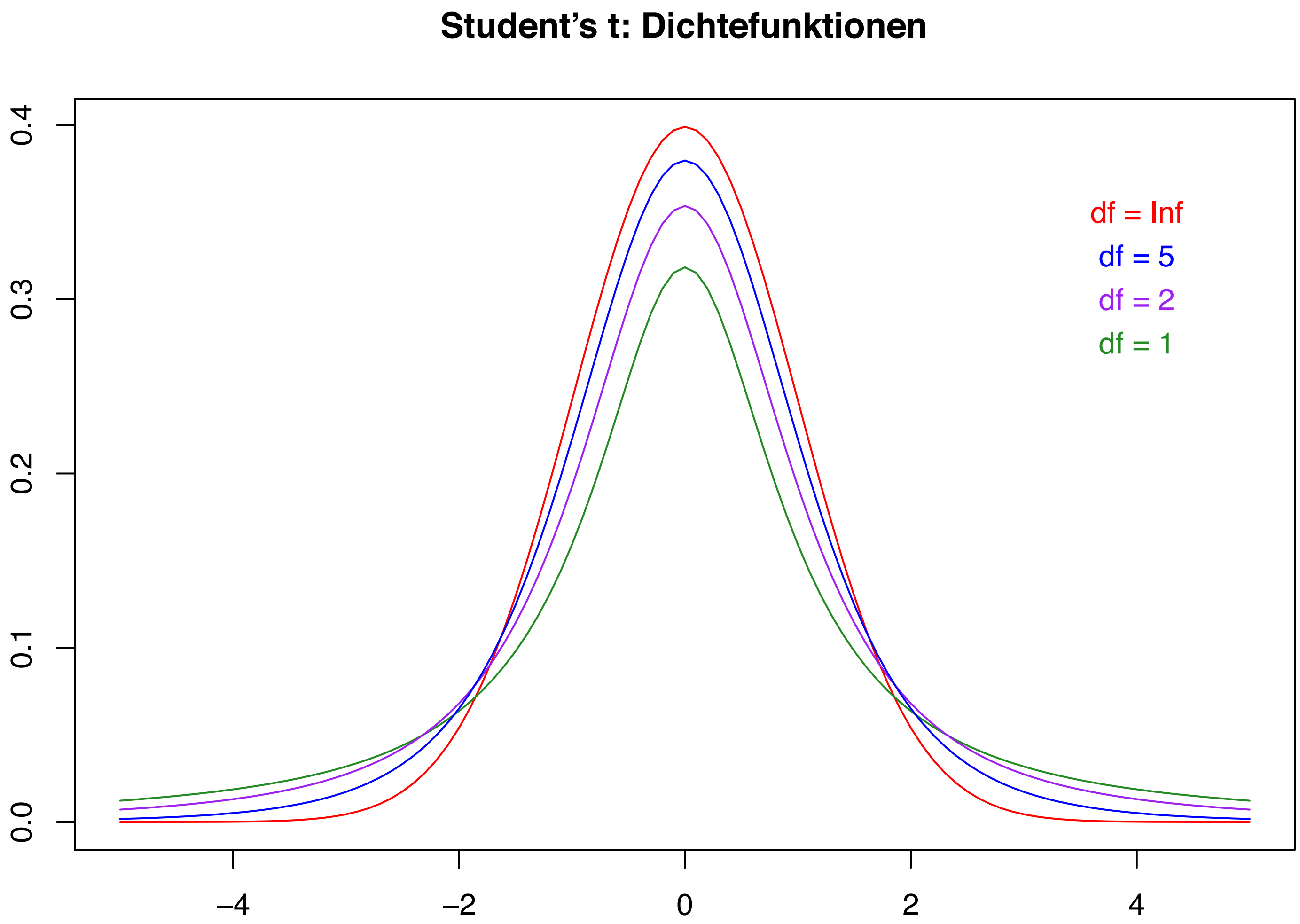
dann ist $\quad T := \dfrac{\bar{X} - \mu}{s/\sqrt{n}} \quad$ so verteilt wie

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

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.


Student's t: Dichtefunktionen
df = Inf
df = 5
df = 2
df = 1
0.0
0.1
0.2
0.3
0.4
−4
−2
0
2
4

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

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

Je kleiner $n$, um so mehr schwankt der Nenner, und um so *breitschultriger* ist die Verteilung von $T_{n-1}$

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

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

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

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

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

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

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

dann ist für jedes $c > 0$:

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

Für ein $95\%$-Konfidenzintervall bestimme $c$ so,

dass sich hier $0.95$ ergibt.

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

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

dann ist für jedes $c > 0$:

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

Für ein $95\%$-Konfidenzintervall bestimme $c$ so,

dass sich hier $0.95$ ergibt.

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

Der passende R-Befehl ist `qt(0.975, 5)`,

mit der Ausgabe 2.57

Denn: $\mathbf{P}(T_5 \leq 2.57) = 0.975$.

Man sagt: Das $0.975$-Quantil der $t(5)$-Verteilung ist $2.57$.

**Beweisskizze** des Satzes von Gosset und Fisher:

$X_1, \ldots, X_n$ ist von der Form $X_1 = \mu + \sigma Z_1, , \ldots, X_n = \mu + \sigma Z_n$ mit $Z_1, \ldots, Z_n$ unabhängig und standard-normalverteilt. Also:

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

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

**Beweisskizze** 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:

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

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

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

**Beweisskizze** des Satzes von Gosset und Fisher:

$X_1, \ldots, X_n$ ist von der Form $X_1 = \mu + \sigma Z_1, , \ldots, X_n = \mu + \sigma Z_n$ mit $Z_1, \ldots, Z_n$ unabhängig und standard-normalverteilt. Also:

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

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

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

$\sqrt{n}\bar{Z}$ ist die Koordinate von $\vec{Z} := (Z_1, \ldots, Z_n)$ bzgl. des Einheitsvektors in Richtung $(1, \ldots, 1)$,
$(Z_1 - \bar{Z})^2 + \cdots + (Z_n - \bar{Z})^2$ ist das Längenquadrat der Projektion orthogonal dazu.

Die Behauptung des Satzes folgt dann aus der Rotationssymmetrie der Verteilung von $\vec{Z}$; vgl. Seite 138 im Buch).

# Ein Konfidenzintervall für den Median

Eine Zahl $\nu$ heißt *Median* der Verteilung $\rho$ auf $\mathbb{R}$,

wenn $\rho((-\infty, \nu]) \geq 1/2$ und $\rho([\nu, \infty)) \geq 1/2$ gilt.

Die *Ordnungsstatistiken* $X_{(1)} \leq \cdots \leq X_{(n)}$

sind die aufsteigend geordneten $X_1, \ldots, X_n$

Ein Kandidat für ein Konfidenzintervall für den Median ist

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

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

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

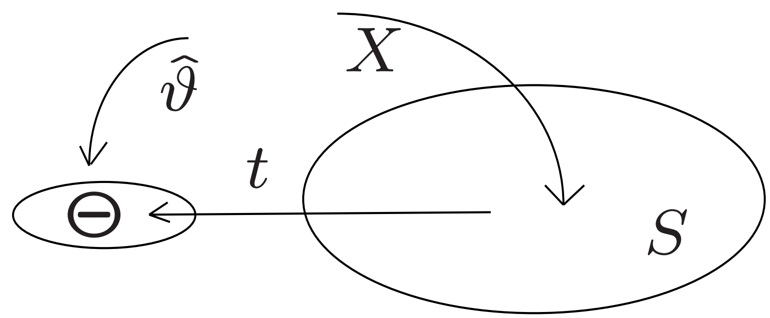
$$\mathbf{P}_\rho(X_{(1)} > \nu) \leq 2^{-n}$$

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

Also:

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

Ein Logo der Statistik:




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

$\vartheta$ ... *Parameter* (z.B. $\vartheta = (\mu, \sigma^2)$ bei der Normalvert.)

$\Theta$ ... *Raum der Parameter*

$S$ ... *Beobachtungsraum*

$\widehat{\vartheta} := h(X)$ ... *Schätzer* für den Parameter $\vartheta$

Sei $m(\vartheta)$ ein reelles Parametermerkmal (z.B. $m(\vartheta) = \mu$)

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.