

Vorlesung 12a

Schätzen mit Verlass

1. 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 der Variabilität dieser Zufallsvariablen.

Die Stichprobenziehung wird als p -Münzwurf gedeutet.

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

$$\hat{p} := H := \frac{K}{n},$$

mit $K :=$ Anzahl der “Erfolge”.

H ist die (relative) Häufigkeit der Erfolge.

$$\sigma_H = ?$$

K ist $\text{Bin}(n, p)$ -verteilt. Also:

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

Der Zentrale Grenzwertsatz besagt:

H ist approximativ normalverteilt

mit Erwartungswert $\mu_H = p$ und Standardabweichung σ_H .

Also insbesondere:

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

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

Das zufällige Intervall $[H - 2\sigma_H, H + 2\sigma_H]$ überdeckt den Parameter p mit Wahrscheinlichkeit ≈ 0.95 .

In der Praxis ist auch σ_H
(aus der *einen* vorliegenden Stichprobe) zu schätzen.

Als Schätzer für $\sigma_H = \frac{1}{\sqrt{n}}\sqrt{p(1-p)}$ bietet sich an:

$$\widehat{\sigma}_H := \frac{1}{\sqrt{n}}\sqrt{H(1-H)}$$

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

überträgt sich auf

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

Das zufällige Intervall

$$I := [H - 2\widehat{\sigma}_H, H + 2\widehat{\sigma}_H]$$

ist ein

Konfidenzintervall für p

mit approximativer Überdeckungswahrscheinlichkeit 0.95

oder kurz ein

approximatives 95%-Konfidenzintervall für p .

In unserem Eingangsbeispiel ($n = 53$, $k = 23$)
hatten die beobachteten Realisierungen von H und $\widehat{\sigma}_H$
die Werte

$$h = 23/53 = 0.43,$$
$$\frac{1}{\sqrt{n}}\sqrt{h(1-h)} = \sqrt{\frac{0.43 \cdot 0.57}{53}} = 0.07.$$

Als Realisierung von $I = [H - 2\widehat{\sigma}_H, H + 2\widehat{\sigma}_H]$ ergab sich
 $[0.43 - 2 \cdot 0.07, 0.43 + 2 \cdot 0.07] = [0.29, 0.57]$.

Man beachte:

Nicht der Parameter p ist zufällig, sondern das Intervall I .

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

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

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 .

Dabei sollte $np(1 - p)$ “nicht zu klein” sein.

Eine Faustregel für die Anwendbarkeit ist: $nh \geq 9$ und $n(1 - h) \geq 9$.

2. 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 .

\bar{X} ist ein Schätzer für μ .

\bar{X} hat Erwartungswert μ und Standardabweichung σ/\sqrt{n} .

Ein Schätzer für σ^2 ist

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

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

2a. Großer Stichprobenumfang n

Der Zentrale Grenzwertsatz besagt:

$\bar{X} - \mu$ ist approximativ $N(0, \frac{\sigma^2}{n})$ -verteilt.

Bei bekanntem σ 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 μ .

In der Praxis hat man auch σ 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 μ

Dieses Intervall enthält keine unbekannt Parameter.

Der *Standardfehler*

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

ist ein Schätzwert für die

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

Die Realisierung von J ist somit $|\bar{x} - 2f, \bar{x} + 2f|$.

2a. Kleiner Stichprobenumfang n

Für kleine n (etwa: $n \leq 10$) und (exakt bzw. annähernd)
normalverteilte X_i

macht man sich zunutze, dass die Verteilung von

$T := \frac{\sqrt{n}(\bar{X} - \mu)}{S}$ (exakt bzw. annähernd) so verteilt ist 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} .

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

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

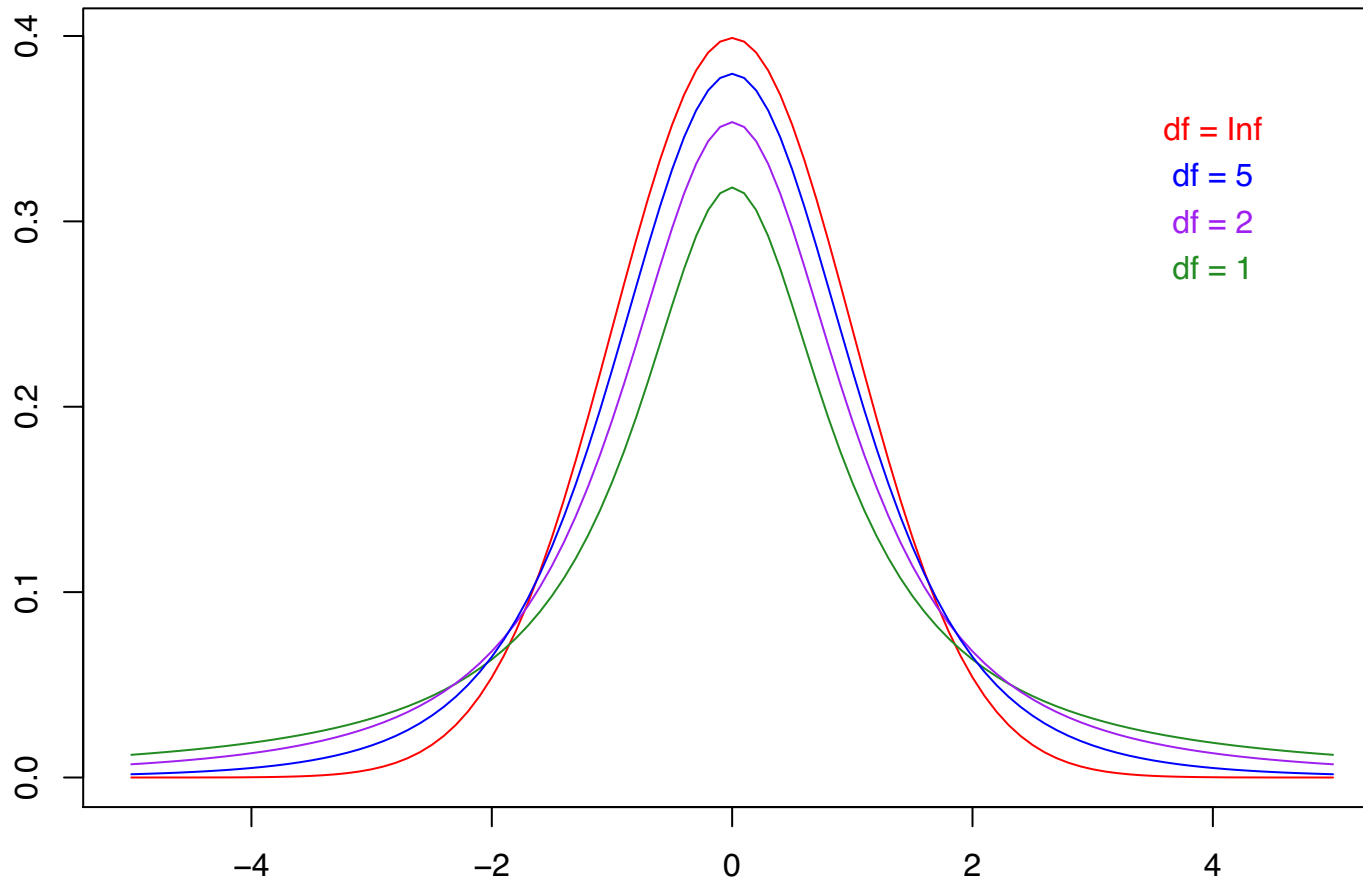
dann ist $T := \frac{\sqrt{n}(\bar{X} - \mu)}{S}$ so verteilt wie

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

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.

Student's t: Dichtefunktionen



“df” steht hier für “degrees of freedom” , d.h. “Freiheitsgrade”

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

Für $n \rightarrow \infty$ ist T_{n-1} asymptotisch $N(0, 1)$ -verteilt
(Gesetz der großen Zahlen für den 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 $n = 6$: $\mathbf{P}(|T_5| \leq 2.57) = 0.95$.

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

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

dann ist $T := \frac{\sqrt{n}(\bar{X} - \mu)}{S}$ 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} .

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

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

Für ein 95%-Konfidenzintervall bestimme c so, dass sich hier 0.95 ergibt.

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

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.

3. Ein Konfidenzintervall für den Median

Seien X_1, X_2, \dots, X_n unabhängig, mit Verteilung ρ .

Es gibt Situationen, in denen die Schätzung der “Lage” von ρ über den Stichprobenmittelwert problematisch ist – etwa wenn ρ so viel Masse weit draußen hat, dass die Varianz von ρ sehr groß ist (oder sogar der Erwartungswert von ρ gar nicht existiert)

In diesem Fall ist es günstig, einen “robusteren” Lageschätzer zu verwenden:

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.

Wenn es nur eine Zahl ν gibt mit $\rho((-\infty, \nu]) = 1/2$
(also z. B. wenn $\rho([\ell, r]) = 1$ gilt
und ρ eine strikt positive Dichtefunktion besitzt),
dann ist ν **der** Median von ρ .

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.

Wenn es mehrere Zahlen ν gibt mit $\rho((-\infty, \nu]) = 1/2$,
dann ist jede dieser Zahlen **ein** Median von ρ .

Bsp: Für die uniforme Verteilung auf $[0, 1] \cup [2, 3]$
ist jedes $\nu \in [1, 2]$ ein Median.

Wie schätzt man den Median?

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:

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

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

Bsp: $n = 6$

$$\mathbf{P}_\rho(\nu \in [X_{(1)}, X_{(6)}]) \geq 1 - \frac{1}{32} = 0.97 :$$

Das Konfidenzintervall $[X_{(1)}, X_{(6)}]$ für den Median hält die Überdeckungswahrscheinlichkeit 0.97 ein.