Parameterschätzung

Induktive Erkenntnisgewinnung

Aus verschiedenen Gründen können nicht immer alle Merkmalsträger einer Grundgesamtheit untersucht werden. Dann entnimmt man repräsentative Stichproben (*Zufallsstichproben*). Mittels der *induktiven Statistik* können dann Rückschlüsse aus den Parametern der Stichprobe auf die Grundgesamtheit gezogen werden.

Schätzfunktion (Schätzer, Schätzstatistik)

Algorithmus zur möglichst treffsicheren Bestimmung eines Parameters $\boldsymbol{\theta}$ (sprich: theta) der Grundgesamtheit aus vorliegenden Stichprobendaten.

Gütekriterien für Schätzer

Erwartungstreue: Es soll gelten $\mathbf{E}(\hat{\theta}) = \theta$, also eine hohe Treffgenauigkeit; ansonsten verzerrter Schätzer, wobei ***Bias*** $\mathrm{B} = \mathrm{E}(\theta) - \theta$.

Variabilität der Schätzung: Standardfehler einer Schätzfunktion SEM = $\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}}$ sollte klein sein. Cave! SEM ist nicht gleich Standardabweichung!

***Entscheidungskriterien**:*
Falls (gleiche) Erwartungstreue, aber unterschiedliche Variabilität: Schätzer mit geringerem Standardfehler verwenden! Falls jedoch auch Verzerrungswerte unterschiedlich: ***MSE*** (mittlerer quadratischer Fehler, mean squared error) berücksichtigt Varianz und Verzerrung. Im Zweifelsfall den Schätzer mit dem kleineren ***MSE*** bevorzugen. Wenn ein Schätzer erwartungstreu ist, stimmen Varianz und MSE überein.

Eigenschaften wichtiger Stichprobenkenngrößen

Stichproben-Mittelwert $\overline{X}$

Er stellt einen ***erwartungstreuen*** Schätzer des Erwartungswertes $\boldsymbol{\mu}$ der Grundgesamtheit dar. Da bei jeder Stichprobenziehung unterschiedliche Werte in die Berechnung eingehen, hat der Stichproben-Mittelwert wiederum den Charakter einer Zufallsgröße. Die Schätzungsqualität erhöht sich mit zunehmendem ***n***.

$$\overline{\mathrm{X}} = \frac{1}{\mathrm{n}} \sum \mathrm{X_i} = \mu$$

Korrigierte Stichprobenvarianz S^{*2}

Erwartungstreu ist die Varianz der Stichprobe nur dann, wenn man (anders als bei der „normalen" Varianz) ***n − 1*** rechnet (ansonsten nur asymptotisch erwartungstreu). *Die* ***korrigierte Stichproben-Standardabweichung S**** erhält man durch Wurzelziehen.

$$S^{*2} = \frac{1}{\mathrm{n}-1} \sum (\mathrm{X_i} - \overline{\mathrm{X}})^2 = \sigma^2 \qquad S^* = \sqrt{S^{*2}}$$

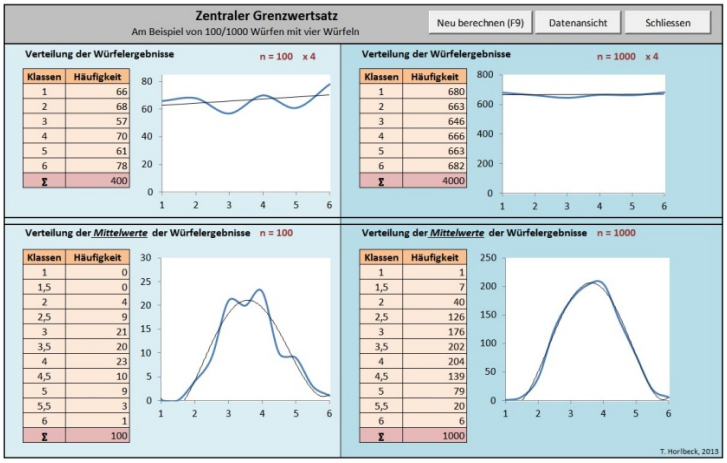
Anteilswert $\hat{p}$

In Analogie zur Bernoulli-Verteilung (Zufallsexperiment mit den zwei Ausgängen A und $\bar{A}$, die mit der Wahrscheinlichkeit $p = P(A)$ bzw. $1 - p = P(\bar{A})$ auftreten) kann man die Bernoulli-Kette als Stichprobenvariablen $\bar{X}_1, \bar{X}_2, \ldots \bar{X}_n$ interpretieren, deren Mittelwert wiederum zur Schätzung des (relativen) Anteilswertes p genutzt werden kann.

$$\hat{\mathrm{p}} = \frac{1}{\mathrm{n}} \sum \overline{\mathrm{X}} = \frac{\mathrm{X}}{\mathrm{n}} = \mathrm{p}$$

Der zentrale Grenzwertsatz

Hilft bei der induktiven Statistik, da er eine sehr günstige Eigenschaft von Stichprobenkenngrößen herausstellt: Die Verteilung der Stichprobenmittelwerte nähert sich mit zunehmendem ***n*** einer Normalverteilung an:

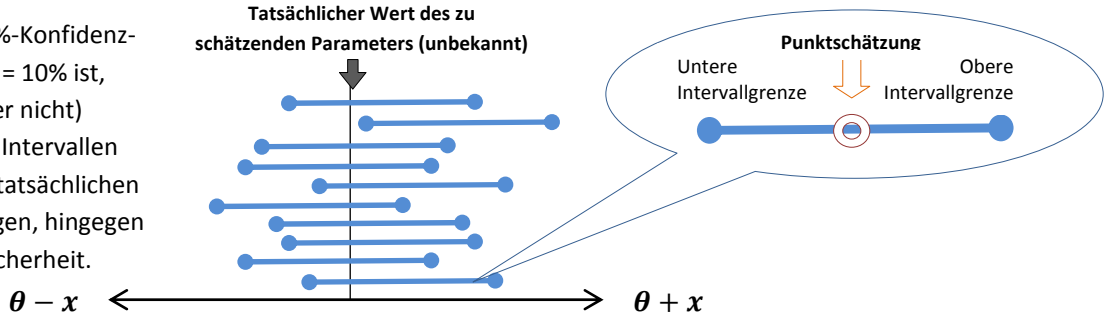


Obwohl diese Eigenschaft hier beispielhaft von einer Gleichverteilung ausgehend ist, gilt sie -bei genügend großem ***n*** - sogar bei unbekannter Verteilung der Grundgesamtheit!

Intervallschätzung

Die Schätzung nur eines Wertes, ***Punktschätzung*** genannt, ist mit einer gewissen Unsicherheit behaftet. Da die entnommenen Stichproben immer zufallsabhängig sind, liefert jede neue Schätzung auch immer leicht streuende Ergebnisse. Mit einem ***Konfidenzintervall*** erhöht man die Aussagekraft einer Schätzung. Dazu bestimmt man ein den Punktschätzungswert umgebendes Intervall, das den zu schätzenden (unbekannten) Parameter der Grundgesamtheit mit einer bestimmten Wahrscheinlichkeit $1 - \alpha$ überdeckt. Der Wert α repräsentiert die (möglichst kleine) ***Irrtumswahrscheinlichkeit***, dessen Komplementärwert $1 - \alpha$ ist die (möglichst große) ***Vertrauens-wahrscheinlichkeit***.

Beispiel am 90%-Konfidenz-intervall. Da α = 10% ist, kann (muss aber nicht) einer von zehn Intervallen außerhalb des tatsächlichen Parameters liegen, hingegen besteht 90% Sicherheit.



Konstruktion des Konfidenzintervalls (zum Konfidenzniveau $1 - \alpha$ für μ ***bei bekannter Varianz/SD***)

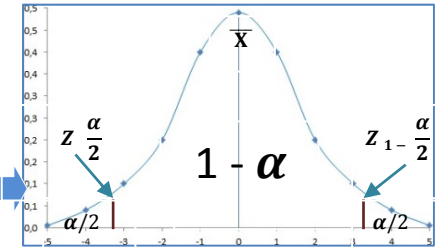
$$KI = \left[\overline{X} - Z_{1-\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}} ; \overline{X} + Z_{1-\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}} \right]$$

= [**Punktschätzwert** - **Kritischer Wert** · **Standardfehler** ; **Punktschätzwert** + **Kritischer Wert** · **Standardfehler**]

(Hier erfolgt Standardisierung und Aufteilung von α zu beiden Seiten)

(Von ***n*** abhängige Variabilität des Schätzers)

Zusammenhang an der Normalverteilungskurve



A. Bestimmung des Konfidenzintervalls für den Mittelwert bei bekannter Varianz und normalverteilter Grundgesamtheit

Gegeben seien 5 Stichprobenwerte mit einem (hier bereits berechneten) Stichprobenmittelwert $\overline{X} = 4{,}02$ *und einer Varianz* $\sigma^2 = 2{,}1$ *bei normalverteilter Grundgesamtheit, und einer Irrtumswahrscheinlichkeit* $\alpha = 10\%$. *Gesucht ist das Intervall* ***KI****, in welchem der (unbekannte) Mittelwert der Grundgesamtheit mit einer Sicherheit 1-* α *(= 90%) enthalten ist.*

1. Mit der Tabelle für Quantile der Standardardnormalverteilung

1.1 ***Formulierung***

$$KI = \left[\overline{X} - z_{1-\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}};\ \overline{X} + z_{1-\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}}\right]$$

$$= \left[4{,}02 - z_{1-\frac{\alpha}{2}} \cdot \frac{\sqrt{2{,}1}}{\sqrt{5}};\ 4{,}02 + z_{1-\frac{\alpha}{2}} \cdot \frac{\sqrt{2{,}1}}{\sqrt{5}}\right]$$

1.2 ***Nachschlagen für*** $z_{1-\frac{\alpha}{2}}$ ***in Tabelle für Quantile der Standardnormalverteilung:***

Vorgehensweise:

a.) Da α hier 10% = 0,1 ist, wird es zur Hälfte auf das linke und das rechte Quantil aufgeteilt. Gesucht ist somit der z_p***-Wert*** von $1 - \frac{\alpha}{2} = 1 - \frac{0{,}1}{2} = 1 - 0{,}05 =$ **0,95**

b.) Das gesuchte $z_{0,95}$ finden:

p	0,5	0,6	0,7	0,8	0,9	0,95	0,975	0,990	0,995	0,999
z_p	0,0000	0,2533	0,5244	0,8416	1,2816	1,6449	1,9600	2,3263	2,5758	3,0902

1.3 ***Einsetzung*** z_p ***in die Formel:***

$$KI = \left[4{,}02 - 1{,}64 \cdot \frac{\sqrt{2{,}1}}{\sqrt{5}};\ 4{,}02 + 1{,}64 \cdot \frac{\sqrt{2{,}1}}{\sqrt{5}}\right] = \left[4{,}02 - 1{,}64 \cdot \frac{\sqrt{2{,}1}}{\sqrt{5}};\ 4{,}02 + 1{,}64 \cdot \frac{\sqrt{2{,}1}}{\sqrt{5}}\right]$$

$$= [4{,}02 - 1.64 \cdot 0{,}648;\ 4{,}02 - 1.64 \cdot 0{,}648] = [4{,}02 - 1{,}06;\ 4{,}02 + 1{,}06] = [\mathbf{2{,}96;\ 5{,}08}]$$

Zur Beachtung: Operatorenrangfolge! Varianz vs. Standardabweichung beachten!
Aufteilung α *auf beide Asymptoten nicht vergessen!*
Quantile < 0,5 ergeben sich aus der Tabelle über die Symmetriebeziehung $z_p = -z_{1-p}$ *!*

1.4 ***Interpretation***

Der gesuchte Mittelwert liegt mit 90-prozentiger Sicherheit (und 10-prozentiger Irrtumswahrscheinlichkeit) innerhalb der Werte **2,96 und 5,08**.

2. Mit Excel (Standardisierung nicht erforderlich, Aufteilung α **/ 2** nicht erforderlich)

KI = [4,02 ± KONFIDENZ(0,1;1,44;5) = [2,96 ; 5,08]

($\overline{X}$ = 4,02; α = 0,1; σ = 1,44; n = 5)

Beinhaltet gesamten Schätzfehlerterm, der vom Stichprobenmittelwert abgesetzt bzw. hinzugefügt wird.

B. Bestimmung des Konfidenzintervalls für den Mittelwert bei unbekannter Varianz und normalverteilter Grundgesamtheit

Gegeben seien 5 Stichprobenwerte {1,9; 3,4; 4,9; 4,4; 5,5} mit einem (hier bereits berechneten) Stichprobenmittelwert $\overline{X} = 4{,}02$ *bei normalverteilter Grundgesamtheit und unbekannter Varianz, und einer Irrtumswahrscheinlichkeit* $\alpha = 10\%$. *Gesucht ist das Intervall* ***KI****, in welchem der (unbekannte) Mittelwert der Grundgesamtheit mit einer Sicherheit 1-* α *(=90%) enthalten ist.*

Da die Varianz der Grundgesamtheit hier unbekannt ist, muss auch diese als ***S**** aus den vorliegenden Stichproben geschätzt werden. Für den Fehlerterm kommt zudem statt der Normalverteilung die ***t-Verteilung*** zum Einsatz, die (jedenfalls bis $n \sim 30$, danach Annäherung an die Normalverteilung) erwartungstreuer schätzt:

$t_{n;\,p}$ (sprich: *p-Quantil* der *t-Verteilung* mit ***n Freiheitsgraden***)

Der kritische Wert wird hier folgendermaßen dargestellt:

$t_{n-1;\,1-\frac{\alpha}{2}} = 1 - \frac{\alpha}{2}$ - ***Quantil der t-Verteilung mit n-1 (= v) Freiheitsgraden***

1. Mit der Tabelle für Quantile der t-Verteilung

Casio fx-991DE plus:
MODE [2 = Statistik]
[1 = 1-Var]
[WERTEEINGABE] AC
SHIFT [1 = Statistik]
[4 = Var] [4 = $\bar{x}$] =

1.1 *Formulierung*

$$KI = \left[\ \bar{X} - t_{n-1;1-\frac{\alpha}{2}} \cdot \frac{S^*}{\sqrt{n}}\ ;\ \bar{X} + t_{n-1;1-\frac{\alpha}{2}} \cdot \frac{S^*}{\sqrt{n}}\ \right]$$

$$= \left[\ \bar{X} - t_{n-1;1-\frac{\alpha}{2}} \cdot \frac{\sqrt{\frac{1}{n-1}\sum(X_i-\bar{X})2}}{\sqrt{n}}\ ;\ \bar{X} + t_{n-1;1-\frac{\alpha}{2}} \cdot \frac{\sqrt{\frac{1}{n-1}\sum(X_i-\bar{X})2}}{\sqrt{n}}\ \right]$$

$$= \left[\ 4{,}02 - t_{n-1;1-\frac{\alpha}{2}} \cdot 0{,}631\ ;\ 4{,}02 + t_{n-1;1-\frac{\alpha}{2}} \cdot 0{,}631\ \right]$$

Korrigierte Sichprobenvarianz **S***

1.2 *Nachschlagen für $t_{n-1;1-\frac{\alpha}{2}}$ in Tabelle für Quantile der t-Verteilung:*

Vorgehensweise:

a.) Da α hier 10% = 0,1 ist, wird es zur Hälfte auf das linke und das rechte Quantil aufgeteilt. Gesucht ist somit der t_p***-Wert*** von $1 - \frac{\alpha}{2} = 1 - \frac{0{,}1}{2} = 1 - 0{,}05 =$ **0,95**

b.) Das gesuchte $t_{0,95}$ für ***n - 1 = 4*** finden:

n	0,800	0,850	0,900	0.950	0,975	0,990	0,995
1	1,376	1,963	3,078	6,314	12,706	31,821	63,657
2	1,061	1,386	1,886	2,920	4, 303	6,965	9,925
3	0,979	1,250	1,638	2,353	3,182	4,541	5,841
4	0,941	1,190	1,533	2,132	2,776	3,747	4,604
5	0,920	1,156	1,476	2.015	2,571	3,365	4,032

1.3 *Einsetzung t_p in die Formel:*

$$KI = [\ 4{,}02 - 2{,}132 \cdot 0{,}631\ ;\ 4{,}02 + 2{,}132 \cdot 0{,}631\ ] = [\mathbf{2{,}67;\ 5{,}37}]$$

Zur Beachtung: Operatorenrangfolge! Varianz vs. Standardabweichung beachten! Beachtung, dass gilt $t_v = t_{n-1}$.

1.4 *Interpretation*

Der gesuchte Mittelwert liegt mit 90-prozentiger Sicherheit (und 10-prozentiger Irrtumswahrscheinlichkeit) innerhalb der Werte **2,67 und 5,37**. Im Vergleich zum Ergebnis mit bekannter Varianz aus A.1. ist das Intervall etwas breiter, da die Genauigkeit der Varianzschätzung aus Stichproben geringer ist und die t-Verteilung einen etwas flacheren Verlauf hat.

2. Mit Excel (Standardisierung nicht erforderlich, Aufteilung $\alpha / 2$ nicht erforderlich)

liefert $t_{n-1;\,1-\alpha/2}$ *liefert korrigierte Standardabweichung*

KI = [4,02 ± TINV(0,1;4) * (WURZEL(VARIANZA(1,9;3,4;4,9;4,4;5,5)) / WURZEL(5))]

$\bar{X}$ α Freiheitsgrade v (=n-1) *gegebene Stichprobenwerte* n

$$KI = [2{,}96\ ;\ 5{,}08\ ]$$