Karlsruher Institut für Technologie

Institut für Volkswirtschaftslehre (ECON)

Lehrstuhl für Ökonometrie und Statistik

# Kapitel VII - Intervallschätzungen

## Induktive Statistik

Prof. Dr. W.-D. Heller

Hartwig Senska

Carlo Siebenschuh

**Gegeben:** Zufallsvariable $Y$, deren Verteilung von einem Parameter $\gamma \in \Gamma \subset \mathbb{R}$ abhängt.

**Gesucht:** Schätzbereich für $\gamma$ aus einer einfachen Stichprobe mit Zurücklegen zu $Y$.

**Sinnvolle Schätzbereiche:** Intervalle

**Damit:**
Aktionenraum$=$Menge der nichtleeren Intervalle in $\mathbb{R}$

Ein Intervall $[\gamma_1, \gamma_2]$ in $\mathbb{R}$ ist festgelegt durch die Intervallgrenzen $\gamma_1$ und $\gamma_2$ mit $\gamma_1 < \gamma_2$.

**Bemerkung:** Anstelle von abgeschlossenen Intervallen können auch offene oder halboffene Intervalle betrachtet werden. Als Intervallgrenzen können auch $-\infty$ und $+\infty$ zugelassen werden.

Der Aktionenraum kann damit auch durch

$$I_\Gamma = \{(\gamma_1, \gamma_2) \in \mathbb{R}^2 | \gamma_1 < \gamma_2\}$$

beschrieben werden. $(\gamma_1, \gamma_2)$ Paar von Intervallgrenzen

Intervallschätzfunktion: $\delta : \mathcal{X} \rightarrow I_\Gamma$

$\delta(x_1, x_2, \ldots, x_n) \in \mathbb{R}^2$ Paar der Intervallgrenzen

$\delta(x_1, x_2, \ldots, x_n) = (\delta_1(x_1, x_2, \ldots, x_n), \delta_2(x_1, x_2, \ldots, x_n))$

$\delta_1(x_1, x_2, \ldots, x_n)$ Intervalluntergrenze

$\delta_2(x_1, x_2, \ldots, x_n)$ Intervallobergrenze

mit $\delta_1(x_1, x_2, \ldots, x_n) < \delta_2(x_1, x_2, \ldots, x_n)$.

Beurteilung von Intervallschätzfunktionen mit Hilfe von Schadensfunktionen:

Der Schaden bei einem Schätzintervall $[\gamma_1, \gamma_2]$ und dem wahren Parameter $\gamma$ sollte berücksichtigen,

- ob der Schätzbereich "trifft", d.h. $\gamma \in [\gamma_1, \gamma_2]$ ist,
- ob die Aussage "präzise"ist, d.h. wie schmal das Intervall ist, also die Intervallbreite $\gamma_2 - \gamma_1$.

Mögliche Ansätze für Schadensfunktionen:

a. Konstante Schadensfunktion

$$S([\gamma_1, \gamma_2], \gamma) = s_l(\gamma_2 - \gamma_1) + \begin{cases} s_o & \gamma > \gamma_2 \\ 0 & \gamma \in [\gamma_1, \gamma_2] \\ s_u & \gamma < \gamma_1 \end{cases}$$

$s_o$ Schaden, falls der wahre Parameterwert oberhalb des Schätzintervalls (Unterschätzung)

$s_u$ Schaden, falls der wahre Parameterwert unterhalb des Schätzintervalls (Überschätzung)

b. Lineare Schadensfunktion:

$$S([\gamma_1,\gamma_2],\gamma) = s_l(\gamma_2-\gamma_1) + \begin{cases} s_o(\gamma-\gamma_2) & \gamma > \gamma_2 \\ 0 & \gamma \in [\gamma_1,\gamma_2] \\ s_u(\gamma_1-\gamma) & \gamma < \gamma_1 \end{cases}$$

$s_o$ Schaden pro Einheit bei Unterschätzung

$s_u$ Schaden pro Einheit bei Überschätzung

c. quadratische Schadensfunktion:

$$S([\gamma_1, \gamma_2], \gamma) = s_l(\gamma_2 - \gamma_1) + \begin{cases} s_o(\gamma - \gamma_2)^2 & \gamma > \gamma_2 \\ 0 & \gamma \in [\gamma_1, \gamma_2] \\ s_u(\gamma_1 - \gamma)^2 & \gamma < \gamma_1 \end{cases}$$

Im folgenden betrachten wir die konstante Schadensfunktion.

Der erwartete Schaden wird durch die Risikofunktion angegeben.

$\delta(X) = (\delta_1(X), \delta_2(X))$ Intervallschätzfunktion

Risikofunktion bei konstanter Schadensfunktion:

$$
\begin{aligned}
R(\delta, \gamma) &= E_\gamma[S([\delta_1(X), \delta_2(X)], \gamma)] \\
&= E_\gamma[s_l(\delta_2(X) - \delta_1(X))] + E_\gamma[S_2([\delta_1(X), \delta_2(X)], \gamma)] \\
&= s_l(E_\gamma(\delta_2(X)) - E_\gamma(\delta_1(X))) + s_o P_\gamma(\delta_2(X) < \gamma) \\
&\quad +0 \cdot P_\gamma(\delta_1(X) \leq \gamma \leq \delta_2(X)) + s_u P_\gamma(\delta_1(X) > \gamma)
\end{aligned}
$$

Neben der erwarteten Intervallbreite

$$E_\gamma(\delta_2(X)) - E_\gamma(\delta_1(X))$$

gehen die Wahrscheinlichkeiten für die Fehler

$$\delta_2(X) < \gamma$$

und

$$\gamma < \delta_1(X)$$

ein.

Die Konstellation der Koeffizienten $s_l, s_o$ und $s_u$ berücksichtigt die (unterschiedliche) Wichtigkeit der Zielsetzungen:

- Präzision (ausgedrückt durch die erwartete Intervallbreite)
- geringe Wahrscheinlichkeit einer Unterschätzung
- geringe Wahrscheinlichkeit einer Überschätzung

Ansatz ohne Festlegung der Koeffizienten $s_l, s_o$ und $s_u$:
1. Lege eine obere Schranke $\alpha$ für die Wahrscheinlichkeit für einen Fehler fest.
Forderung:

$$P_\gamma(\gamma < \delta_1(X)) + P_\gamma(\delta_2(X) < \gamma) \leq \alpha$$

2. Minimiere erwartete Intervallbreite unter Einhaltung von 1.

Definition: **Konfidenzintervallschätzung**
Eine Intervallschätzfunktion $[\delta_1, \delta_2]$ heißt **Konfidenzintervallschätzung**, wenn

$$P_\gamma(\gamma \in [\delta_1(X), \delta_2(X)]) = 1 - \alpha \text{ für alle } \gamma \in \Gamma \text{ gilt.}$$

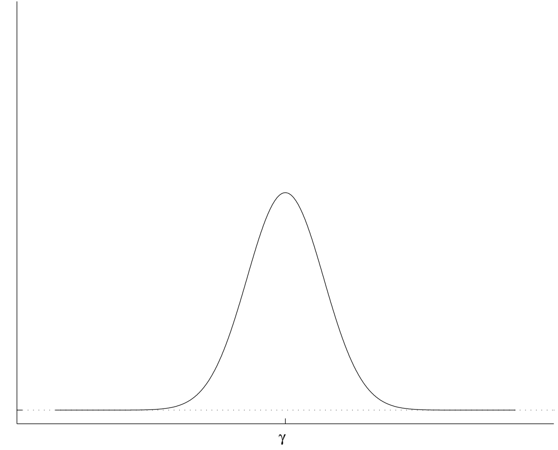
$1 - \alpha$ heißt Konfidenzniveau **(Vertrauenswahrscheinlichkeit)**,
die Wahrscheinlichkeit für einen Fehler beträgt $\alpha$ **(Irrtumswahrscheinlichkeit)**.

Bei einer symmetrischen Verteilung bewirkt die Minimierung der erwarteten Intervallbreite, dass die Irrtumswahrscheinlichkeit zu gleichen Teilen auf die beiden Fehler aufgeteilt wird, wie man sich an der Abbildung klar machen kann. In diesem Fall spricht man von einer

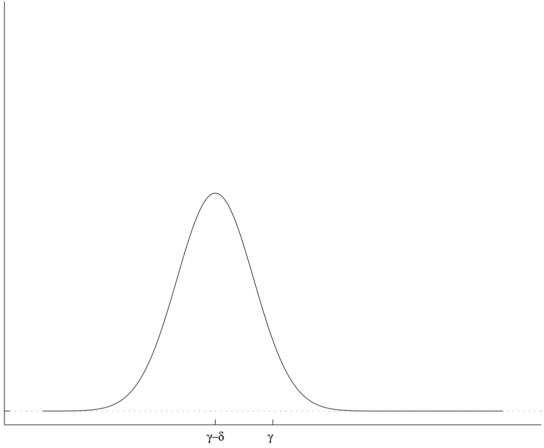
**symmetrischen Konfidenzintervallschätzung**.

Je kleiner die Intervallbreite bei gegebener Fehlerwahrscheinlichkeit, desto nützlicher die Schätzung. Dies wird im folgenden verdeutlicht.

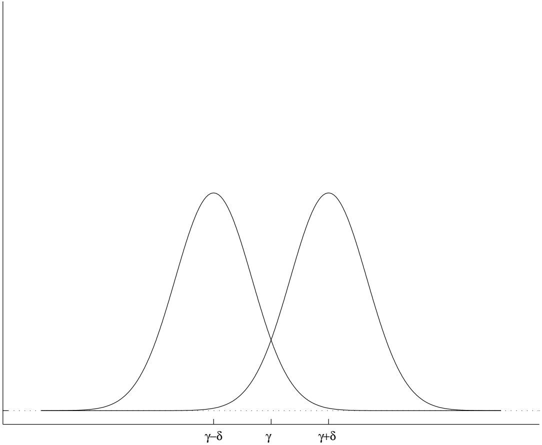
1.Dichtefunktion von $\overline{X}$



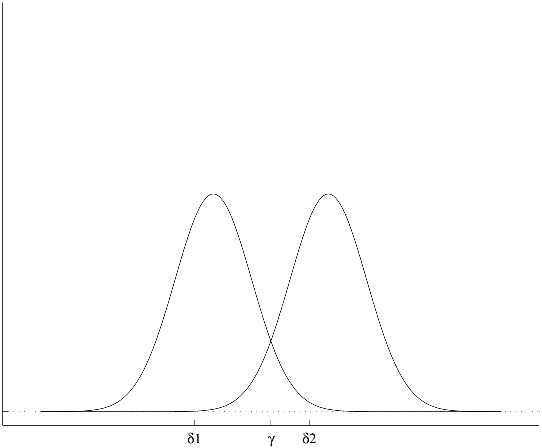
2.Schätzintervall $[\overline{x} - d, \overline{x} + d]$
Dichtefunktion von $\overline{X} - d$



2.Schätzintervall $[\overline{x} - d, \overline{x} + d]$
Dichtefunktion von $\overline{X} - d$ und $\overline{X} + d$



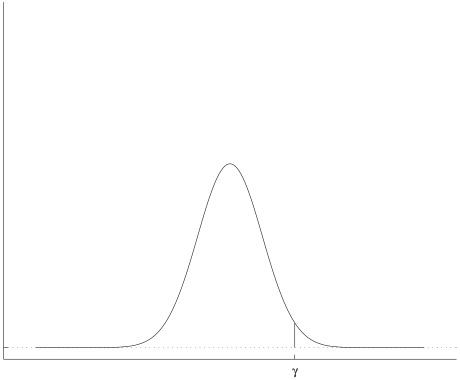
$\delta_1(x) = \overline{x} - d$ und $\delta_2(x) = \overline{x} + d$ gehören zusammen



3.Was kann falsch laufen?

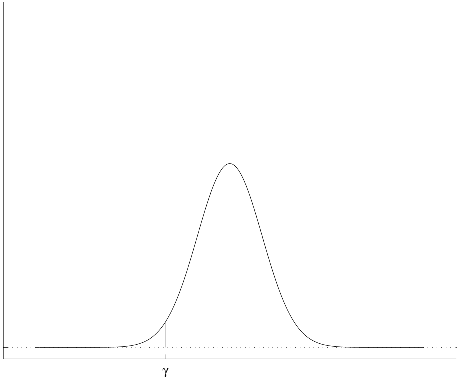
a) $\gamma < \delta_1(x)$

Wahrscheinlichkeit dafür: anhand Dichtefunktion von $\delta_1(X)$

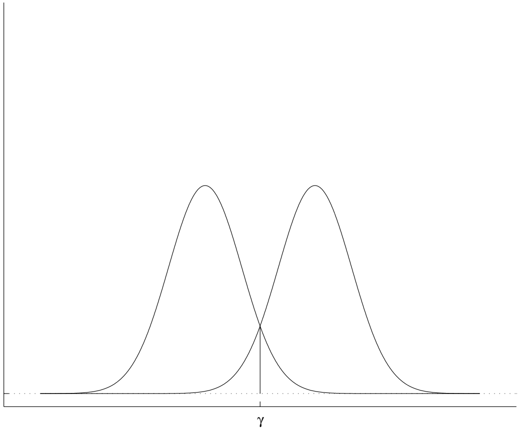


b) $\delta_2(x) < \gamma$

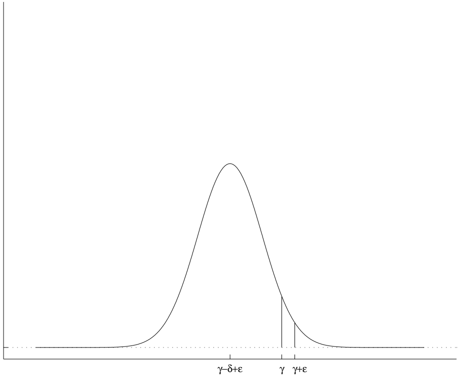
Wahrscheinlichkeit dafür: anhand Dichtefunktion von $\delta_2(X)$



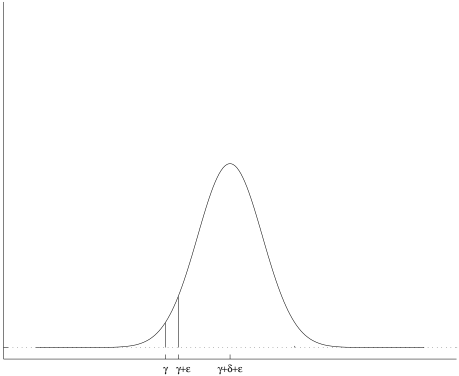
Fehlerwahrscheinlichkeit insgesamt



4. Schätzintervall nicht symmetrisch um $\overline{x}$ bei gleicher Breite $2d$ z.B nach rechts verschoben $[\overline{x} - d + \epsilon, \overline{x} + d + \epsilon], \epsilon > 0$.
$\Rightarrow$ Vergrößerung der Wahrscheinlichkeit von $\gamma < \delta_1(X)$



4. Schätzintervall nicht symmetrisch um $\overline{x}$ bei gleicher Breite $2d$ z.B nach rechts verschoben $[\overline{x} - d + \epsilon, \overline{x} + d + \epsilon], \epsilon > 0$
$\Rightarrow$ Verkleinerung der Wahrscheinlichkeit von $\delta_2(X) < \gamma$



4. Schätzintervall nicht symmetrisch um $\overline{x}$ bei gleicher Breite $2d$ z.B nach rechts verschoben $[\overline{x} - d + \epsilon, \overline{x} + d + \epsilon], \epsilon > 0$
Vergrößerung $>$ Verkleinerung
$\Rightarrow$ Fehlerwahrscheinlichkeit insgesamt größer

Bemerkungen: 1. Das Gleichheitszeichen in

$$P_\gamma(\gamma \in [\delta_1(X), \delta_2(X)]) = 1 - \alpha \text{ für alle } \gamma \in \Gamma$$

soll bewirken, dass das Intervall schmal bleibt, da die Schranke $\alpha$ für den Fehler exakt eingehalten wird.
Implizit bedeutet dies aber, dass $P_\gamma(\gamma \in [\delta_1(X), \delta_2(X)])$ unabhängig von $\gamma$ ist.
Dies ist häufig nicht zu erreichen. Deshalb fordert man

$$P_\gamma(\gamma \in [\delta_1(X), \delta_2(X)]) \geq 1 - \alpha \text{ für alle } \gamma \in \Gamma$$

mit dem Gleichheitszeichen für mindestens ein $\gamma$.

2. Die Formulierung “die Wahrscheinlichkeit, dass $\gamma$ im Intervall liegt, betrage mindestens $1-\alpha$“ sollte man vermeiden, da sie wegen der Ähnlichkeit mit der Bezeichnung bei Zufallsvariablen zu der Fehlerinterpretation führen könnte, dass $\gamma$ die Zufallsvariable ist. Zufällig ist aber das Schätzintervall, während $\gamma$ eine feste Größe ist.

Problem bei der Konfidenzintervallschätzung:

Die Wahrscheinlichkeit

$$P_\gamma(\gamma \in [\delta_1(X), \delta_2(X)])$$

bei der $\gamma$ als Parameter der Verteilung und bei der Formulierung des Ereignisses vorkommt, ist ohne Kenntnis von $\gamma$ zu berechnen bzw. abzuschätzen.

In Spezialfällen wie der Normalverteilung ist dies aber möglich.

Beispiel: Normalverteilung

$Y$ $N(\mu, \sigma^2)$-verteilt mit bekannter Varianz $\sigma^2$

Es ist ein $(1-\alpha)$-Konfidenzintervall für den Mittelwert mit einer Stichprobe mit Zurücklegen vom Umfang $n$ zu schätzen.

Eine erwartungstreue Schätzfunktion für den Erwartungswert $\mu$ ist

$$\bar{X} = \frac{1}{n} \sum_{i=1}^{n} X_i \text{ mit den Realisationen}$$

$$\hat{\mu} = \bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i .$$

Die zugehörige Zufallsvariable ist $N(\mu, \frac{1}{n}\sigma^2)$-verteilt.

Daher ist

$$\hat{\mu}_{stand}(X) = \frac{\mu - \frac{1}{n}\sum\limits_{i=1}^{n} X_i}{\frac{\sigma}{\sqrt{n}}}$$

standardnormalverteilt und damit unbhängig von $\mu$ und $\sigma^2$.
Damit gilt

$$P(-u_{1-\frac{\alpha}{2}} \leq \hat{\mu}_{stand}(X) \leq u_{1-\frac{\alpha}{2}}) = 1 - \alpha,$$

wobei $u_{1-\frac{\alpha}{2}}$ das $(1 - \frac{\alpha}{2})$-Quantil der Standardnormalverteilung ist: $\Phi(u_{1-\frac{\alpha}{2}}) = 1 - \frac{\alpha}{2}$

Daraus ergibt sich:

$$P(\frac{1}{n}\sum_{i=1}^{n} X_i - \frac{\sigma}{\sqrt{n}} u_{1-\frac{\alpha}{2}} \leq \mu \leq \frac{1}{n}\sum_{i=1}^{n} X_i + \frac{\sigma}{\sqrt{n}} u_{1-\frac{\alpha}{2}}) = 1 - \alpha$$

und das $(1-\alpha)$-Konfidenzintervall

$$\left[\frac{1}{n}\sum_{i=1}^{n} X_i - \frac{\sigma}{\sqrt{n}} u_{1-\frac{\alpha}{2}}, \frac{1}{n}\sum_{i=1}^{n} X_i + \frac{\sigma}{\sqrt{n}} u_{1-\frac{\alpha}{2}}\right]$$

Das $(1-\alpha)$-Konfidenzintervall ist ein Zufallsintervall, daher muss zweimal $x_i$ durch $X_i$ ersetzt werden.

Beispiel: $\sigma = 0,02$ bekannt
Zu den Messwerten
10.00, 10.01, 10.07, 10.05, 10.02, 9.99, 9.96, 10.02, 10.04, 10.08
und der Vertauenswahrscheinlichkeit 0.95 ist

$$\text{mit } u_{1-\frac{\alpha}{2}} = u_{0.975} = 1.96$$

und einer Standardabweichung von 0.02

$$[10.024 - \frac{0.02}{\sqrt{10}} 1.96, 10.024 + \frac{0.02}{\sqrt{10}} 1.96] = [10.012, 10.036]$$

eine Realisation des 95%-Konfidenzintervall $[\bar{X} \pm \frac{0.02}{\sqrt{10}} 1.96]$.

Bei einer anderen Verteilung der Irrtumswahrscheinlichkeit $\alpha$ auf die beiden Seiten erhalten wir ein breiteres Intervall:
Fordern wir beispielsweise
$P(\mu < \delta_1(X)) = \frac{1}{4}\alpha$ und $P(\mu > \delta_2(X)) = \frac{3}{4}\alpha$

so gilt

$$P(\frac{1}{n}\sum_{i=1}^{n} X_i - \frac{\sigma}{\sqrt{n}} u_{1-\frac{\alpha}{4}} \leq \mu \leq \frac{1}{n}\sum_{i=1}^{n} X_i + \frac{\sigma}{\sqrt{n}} u_{1-\frac{3\alpha}{4}}) = 1 - \alpha$$

und das Intervall

$$\left[\tfrac{1}{n}\sum_{i=1}^{n} X_i - \tfrac{\sigma}{\sqrt{n}} u_{1-\frac{\alpha}{4}}, \tfrac{1}{n}\sum_{i=1}^{n} X_i + \tfrac{\sigma}{\sqrt{n}} u_{1-\frac{3\alpha}{4}}\right]$$

hat die Überdeckungswahrscheinlichkeit $\alpha$.

Für $\alpha = 0.05$ ist

$u_{1-\frac{3\alpha}{4}} = u_{0.9625} = 1.78$ und
$u_{\frac{\alpha}{4}} = -u_{1-\frac{\alpha}{4}} = -u_{0.9875} = -2.24$

und die Intervallbreite ist mit

$$\odot 4.02 \frac{\sigma}{\sqrt{n}} \text{ gegenüber } \oplus 3.92 \frac{\sigma}{\sqrt{n}}$$

größer als bei symmetrischer Aufteilung.

$\odot$ KI nicht symmetrisch
$\oplus$ KI symmetrisch

Ist $\sigma^2$ nicht bekannt, so können wir einen Schätze für $\sigma^2$ bzw. $\sigma$ verwenden, also z.B. die korrigierte Stichprobenstandardabweichung

$$S^*(X) = \sqrt{\tfrac{1}{n-1} \sum_{i=1}^{n} (X_i - \bar{X})^2}.$$

Diese Schätzfunktion ist **nicht erwartungstreu** für $\sigma$, was im Moment nicht erforderlich ist. Damit ist jetzt

$$\hat{\mu}_t(X) = \frac{\mu - \frac{1}{n} \sum_{i=1}^{n} X_i}{\frac{S^*(X)}{\sqrt{n}}}$$

nicht mehr standardnormalverteilt, hat aber den Erwartungswert 0, falls $n \geq 3$, und die Varianz $\frac{n-1}{n-3}$, falls $n \geq 4$.

Die Verteilung dieser Zufallsvariaben heißt $t$-Verteilung

und ihre Quantile sind tabelliert.

Mit den Quantilen

$$t(n-1)_{1-\frac{\alpha}{2}} \quad \text{und} \quad t(n-1)_{\frac{\alpha}{2}} = -t(n-1)_{1-\frac{\alpha}{2}}$$

gilt

$$P(-t(n-1)_{1-\frac{\alpha}{2}} \leq \hat{\mu}_t(X) \leq t(n-1)_{1-\frac{\alpha}{2}}) = 1-\alpha$$

Das $(1-\alpha)$-Konfidenzintervall ist damit

$$\left[\frac{1}{n}\sum_{i=1}^{n} X_i - \frac{S^*(X)}{\sqrt{n}} t(n-1)_{1-\frac{\alpha}{2}}, \frac{1}{n}\sum_{i=1}^{n} X_i + \frac{S^*(X)}{\sqrt{n}} t(n-1)_{1-\frac{\alpha}{2}}\right] \blacksquare$$

Das $(1-\alpha)$-Konfidenzintervall ist ein Zufallsintervall, daher muss zweimal $x_i$ durch $X_i$ und zweimal $S^*(x)$ durch $S^*(X)$ ersetzt werden. Im Zahlenbeispiel erhalten wir mit $S^*(X) = 0.0369$, für $\alpha = 0.05$

$$t(n-1)_{1-\frac{\alpha}{2}} = t(9)_{0.975} = 2.2622$$

und damit als Realisation des 95%-Konfidenzintervall [9.998,10.050], also ein geringfügig größeres Intervall (vgl.: [10,012; 10,036]), da sozusagen die Unsicherheit bei der Varianz berücksichtigt wird.

**Allgemeine Vorgehensweise**

$Y$ Zufallsvariable mit unbekanntem Parameter $\gamma$

$X$ Stichprobenvektor zu $Y$

$\hat{\gamma}(X)$ (erwartungstreue) Schätzfunktion für $\gamma$

$E_\gamma(\hat{\gamma}(X)) = \mu(\gamma)$ Funktion von $\gamma$
($\mu(\gamma) = \gamma$, falls $\hat{\gamma}(X)$ erwartungstreu)

$Var_\gamma(\hat{\gamma}(X)) = \sigma^2(\gamma)$ Funktion von $\gamma$

Im Beispiel der Normalverteilung:

$\hat{\gamma}(x) = \bar{x}$, also $\mu(\hat{\gamma}) = \gamma$

$Var_{\gamma}(\hat{\gamma}(X)) = \frac{\sigma^2}{n}$, also $\sigma^2(\gamma) = \frac{\sigma^2}{n}$ konstant bezüglich $\gamma = \mu$

Dann ist

$$\hat{\gamma}_{Stand.}(X) = \frac{\mu(\gamma) - \hat{\gamma}(X)}{\sigma(\gamma)}$$

eine Zufallsvariable mit Erwartungswert 0 und Varianz 1.

"Wenn man Glück hat" (s. Hartung(1993), S.130), ist die Verteilung dieser Zufallsvariable unabhägig von $\gamma$ und bekannt.

Sei $F$ die Verteilungsfunktion der Zufallsvariablen und

$$t_{\frac{\alpha}{2}} \text{ das } \tfrac{\alpha}{2}\text{-Quantil, d.h. } F(t_{\frac{\alpha}{2}}) = \frac{\alpha}{2}$$

$$t_{1-\frac{\alpha}{2}} \text{ das } (1-\tfrac{\alpha}{2})\text{-Quantil, d.h. } F(t_{1-\frac{\alpha}{2}}) = 1-\frac{\alpha}{2}$$

so gilt

$$\begin{aligned} 1-\alpha &= F(t_{1-\frac{\alpha}{2}}) - F(t_{\frac{\alpha}{2}}) \\ &= P(t_{\frac{\alpha}{2}} \leq \hat{\gamma}_{Stand}(X) \leq t_{1-\frac{\alpha}{2}}) \\ &= P(t_{\frac{\alpha}{2}} \leq \tfrac{\mu(\gamma)-\hat{\gamma}(X)}{\sigma(\gamma)} \leq t_{1-\frac{\alpha}{2}}) \end{aligned}$$

Die Ungleichung in der Klammer muss jetzt noch in eine Ungleichung für $\gamma$ umgeformt werden.

**Beispiel:** Bernoulli-Verteilung

**Gesucht:** Konfidenzintervall für den Parameter $p$ einer Bernoulli-Verteilung $P(Y=1)=p$

$\hat{p}=\bar{x}=\hat{\gamma}(x)$, $\mu(p)=E_p(\hat{\gamma}(X))=E_p(\bar{X})=p$

$Var_\gamma(\hat{\gamma}(X))=Var_p(\bar{X})=\frac{p(1-p)}{n}=\sigma^2(\gamma)$

Damit lautet die Ungleichung

$$t_{\frac{\alpha}{2}} \leq \frac{p-\bar{X}}{\sqrt{\frac{p(1-p)}{n}}} \leq t_{1-\frac{\alpha}{2}},$$

die in zwei Ungleichungen für $p$ umgeformt werden kann.

Falls der Erwartungswert zu schätzen ist, kann auf der Grundlage des Zentralen Grenzwertsatzes eine Näherungslösung bestimmt werden:
$n$ hinreichend groß, dann

$$\frac{1}{n}\sum_{i=1}^{n} X_i \text{ näherungsweise normalverteilt}$$

$$\text{mit } \mu = E(Y), \sigma^2 = \frac{1}{n} Var(Y)$$

Dementsprechend liefert das Konfidenzintervall für den Mittelwert der Normalverteilung eine **Näherungslösung** des Konfidenzintervalls für den Erwartungswert von $Y$.

Die Güte der Approximation hängt von der Verteilung von $Y$ ab.

**Beispiel:** Bernoulli-Verteilung

$Y$ Bernoulli-verteilt mit Parameter $p: \quad P(Y=1)=p$

$X=(X_1, X_2, \ldots, X_n)$ Stichprobenvektor m.Z. zu $Y$

$(x_1, x_2, \ldots, x_n)$ Stichprobenergebnis;

$a = x_1 + x_2 + \ldots + x_n$ suffiziente und vollständige Statistik,

$A = X_1 + X_2 + \ldots + X_n$ binomialverteilt mit Parameter $p$.

Zahlenbeispiel: $n = 1000, a = 19$

1. Approximation durch die Normalverteilung

$$\hat{\mu}_{Stand}(X) = \frac{\mu - \frac{1}{n}\sum_{i=1}^{n} X_i}{\frac{\sigma}{\sqrt{n}}} \text{ mit}$$

$$\mu = E(Y) = p, \text{ und } \sigma = \sqrt{Var(Y)} = \sqrt{p(1-p)}$$

ist näherungsweise standardnormalverteilt.

a) Ersetzen der Standardabweichung durch Schätzwert

Die Standardabweichung von $Y$ kann durch die korrigierte Stichprobenstandardabweichung oder durch Einsetzen eines Schätzwerts für $p$ geschätzt werden:

$$\left.\begin{aligned} \hat{\sigma} &= \sqrt{\bar{x}(1-\bar{x})} &&= \quad 0.1365 \\ \hat{\sigma} &= \sqrt{\tfrac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})^2} &&= \quad 0.1366 \end{aligned}\right\}\text{kein wesentlicher Unterschied}$$

$(1-\alpha)$-Konfidenzintervall:

$$[\frac{1}{n}\sum_{i=1}^{n} X_i - \frac{\hat{\sigma}(X)}{\sqrt{n}} u_{1-\frac{\alpha}{2}}, \frac{1}{n}\sum_{i=1}^{n} X_i + \frac{\hat{\sigma}(X)}{\sqrt{n}} u_{1-\frac{\alpha}{2}}]$$

95%-Konfidenzintervall für $p$: $\sqrt{n} = \sqrt{1000} = 31.623$

$[0.019 - 31.623^{-1} \cdot 0.1365 \cdot 1.96, 0.019 + 31.623^{-1} \cdot 0.1365 \cdot 1.96] = [0.011, 0.027]$.
Das $(1-\alpha)$-Konfidenzintervall ist ein Zufallsintervall, daher muss zweimal $x_i$ durch $X_i$ und zweimal $\hat{\sigma}$ durch $\hat{\sigma}(X)$ ersetzt werden.

b) besser: unter Verwendung von $Var(\overline{X}) = \frac{p(1-p)}{n}$

$\frac{p-\bar{X}}{\sqrt{\frac{p(1-p)}{n}}}$ ist näherungsweise standardnormalverteilt

Daraus ergibt sich mit den Quantilen der Standardnormalverteilung:

$$-u_{1-\frac{\alpha}{2}} \leq \frac{p-\bar{X}}{\sqrt{\frac{p(1-p)}{n}}} \leq u_{1-\frac{\alpha}{2}}$$

Umformen in eine Gleichung der Form

$$\delta_1(a) \leq p \leq \delta_2(a) \text{ mit } a = n\bar{x} \text{ ergibt}$$

$$\delta_1(a),\ \delta_2(a) = \frac{2a + u_{1-\frac{\alpha}{2}}^2 \pm u_{1-\frac{\alpha}{2}} \sqrt{u_{1-\frac{\alpha}{2}}^2 + 4a(1-\frac{a}{n})}}{2(n + u_{1-\frac{\alpha}{2}}^2)}$$

Mit $a = 19$, $n = 1000$ erhalten wir das realisierte 95%-Konfidenzintervall

$[0.012, 0.029]$.

2. Exakte Konfidenzintervallschätzung

   Gesucht: Funktionen $\delta_1, \delta_2$ mit

$$P_p(\delta_1(A) < p < \delta_2(A)) \geq 1 - \alpha$$

   bzw. $P_p(\delta_2(A) \leq p) \leq \frac{\alpha}{2}$ und $P_p(p \leq \delta_1(A)) \leq \frac{\alpha}{2}$

   Verteilung von $A$: Binomialverteilung $B(n, p)$

$$L_{n,c}(p) := P_p(A \leq c) = \sum_{k=0}^{c} \binom{n}{k} p^k (1-p)^{n-k}$$

   Damit erhält man: (Herleitung, Buch S.108)

$$\delta_1(a) = L_{n,a-1}^{-1}(1 - \tfrac{\alpha}{2}) \text{ und } \delta_2(a) = L_{n,a}^{-1}(\tfrac{\alpha}{2})$$

$$\delta_1(a) = L^{-1}_{n,a-1}(1 - \tfrac{\alpha}{2}) \text{ und } \delta_2(a) = L^{-1}_{n,a}(\tfrac{\alpha}{2})$$

Diese Werte können mit Hilfe der F-Verteilung bestimmt werden.

Es gilt:

$$L_{n,c}(p) = 1 - F_{2(c+1),2(n-c)}\left(\frac{n-c}{c+1}\frac{p}{1-p}\right)$$

$$= F_{2(n-c),2(c+1)}\left(\frac{c+1}{n-c}\frac{1-p}{p}\right)$$

Mit etwas Rechnung folgt daraus:

$$\begin{aligned} L_{n,c}(p) = \alpha \Leftrightarrow p &= \frac{(c+1)F^{-1}_{2(c+1),2(n-c)}(1-\alpha)}{n-c+(c+1)F_{2(c+1),2(n-c)}(1-\alpha)} \\ &= \frac{c+1}{c+1+(n-c)F^{-1}_{2(n-c),2(c+1)}(\alpha)} \end{aligned}$$

Wobei für $\alpha \leq 0.5$ die obere und für $\alpha \geq 0.5$ die untere Formel Verwendung findet, da die meisten Tabellen die $\alpha$-Quantile der $F$-Verteilung nur für $\alpha \geq 0.5$ enthalten. Bei unseren Zahlenwerten ergibt sich

$$\delta_1(a) = \frac{19}{19+982F^{-1}_{1964,38}(0.975)} = 0.012$$

$$\delta_2(a) = \frac{20F^{-1}_{40,1962}(0.975)}{981+20F^{-1}_{40,962}(0.975)} = 0.029$$

Konfidenzintervallschätzung für die Varianz:
Varianz von $Y$=Erwartungswert von $(Y - E(Y))^2$
Damit:
Setze $Z = (Y - E(Y))^2$ und bestimme
ein Konfidenzintervall für $E(Z)$

Speziell: Normalverteilung

a) $\mu$ bekannt

$$\sum_{i=1}^{n} (x_i - \mu)^2 \text{ ist suffizient und vollständig für } \sigma^2$$

$$\frac{\sum_{i=1}^{n} (X_i - \mu)^2}{\sigma^2} \text{ ist } \chi^2(n)\text{-verteilt}$$

Damit ist

$$\begin{aligned}
1-\alpha &= P\left(\chi^2(n)_{\frac{\alpha}{2}} \leq \frac{\sum\limits_{i=1}^{n}(X_i-\mu)^2}{\sigma^2} \leq \chi^2(n)_{1-\frac{\alpha}{2}}\right) \\
&= P\left(\frac{1}{\chi^2(n)_{1-\frac{\alpha}{2}}} \leq \frac{\sigma^2}{\sum\limits_{i=1}^{n}(X_i-\mu)^2} \leq \frac{1}{\chi^2(n)_{\frac{\alpha}{2}}}\right) \\
&= P\left(\frac{\sum\limits_{i=1}^{n}(X_i-\mu)^2}{\chi^2(n)_{1-\frac{\alpha}{2}}} \leq \sigma^2 \leq \frac{\sum\limits_{i=1}^{n}(X_i-\mu)^2}{\chi^2(n)_{\frac{\alpha}{2}}}\right)
\end{aligned}$$

$(1-\alpha)$-Kofidenzintervall:

$$\left[\frac{\sum_{i=1}^{n}(X_i-\mu)^2}{\chi^2(n)_{1-\frac{\alpha}{2}}} \quad , \quad \frac{\sum_{i=1}^{n}(X_i-\mu)^2}{\chi^2(n)_{\frac{\alpha}{2}}}\right]$$

Das $(1-\alpha)$-Konfidenzintervall ist ein Zufallsintervall, daher muss zweimal $x_i$ durch $X_i$ ersetzt werden.

b) $\mu$ unbekannt

$S^{*^2}(x) = \frac{1}{n-1} \sum\limits_{i=1}^{n} (x_i - \bar{x})^2$ ist erwartungstreue Schätzfunktion

für $\sigma^2$

$\frac{(n-1)S^{*^2}(X)}{\sigma^2}$ ist $\chi^2(n-1)$-verteilt

Damit lautet das $(1-\alpha)$-Konfidenzintervall:

$$\left[ \frac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{\chi^2(n-1)_{1-\frac{\alpha}{2}}} \quad , \quad \frac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{\chi^2(n-1)_{\frac{\alpha}{2}}} \right]$$

Das $(1-\alpha)$-Konfidenzintervall ist ein Zufallsintervall, daher muss zweimal $x_i$ durch $X_i$ und zweimal $\bar{x}$ durch $\bar{X}$ ersetzt werden.