Konfidenzintervalle

Hanno Gottschalk
Institute of Mathematics, TU Berlin

Stochastik für Informatik | SoSe2023, TU Berlin | 27. Juni 2023

9. Konfidenzintervalle

- Konsistente Schätzer sind in der Praxis nur Approximationen. Aber wie gut sind diese?

9. Konfidenzintervalle

- Konsistente Schätzer sind in der Praxis nur Approximationen. Aber wie gut sind diese?
- Wie groß muss eine Stichprobe sein, dass diese Approximation hinreichend gut wird?

9. Konfidenzintervalle

- Konsistente Schätzer sind in der Praxis nur Approximationen. Aber wie gut sind diese?
- Wie groß muss eine Stichprobe sein, dass diese Approximation hinreichend gut wird?
- Mit welchem Schätzfehler muss man bei gegebener Stichprobengröße rechnen?

Technische Universität Berlin

9. Konfidenzintervalle

- Konsistente Schätzer sind in der Praxis nur Approximationen. Aber wie gut sind diese?
- Wie groß muss eine Stichprobe sein, dass diese Approximation hinreichend gut wird?
- Mit welchem Schätzfehler muss man bei gegebener Stichprobengröße rechnen?

Wir geben uns eine W.-keit $\alpha \in (0,1)$ vor und konstruieren ein Intervall um einen Schätzwert, sodass der wahre Wert mit W.-keit $1-\alpha$ in diesem Intervall liegt.

Definition

Sei $(\Omega, \{P_\theta\}_{\theta\in\Theta}, \{X_j\}_{j\in\mathbb{N}})$ ein statistisches Modell und θ_j die j-te Komponente von $\theta = (\theta_1, \dots, \theta_q)$.

Definition

Sei $(\Omega, \{P_\theta\}_{\theta\in\Theta}, \{X_j\}_{j\in\mathbb{N}})$ ein statistisches Modell und θ_j die j-te Komponente von $\theta = (\theta_1, \ldots, \theta_q)$.

(i) Es seien S_-, S_+ zwei Schätzer und $\alpha \in (0,1)$ eine (kleine) Irrtumswahrscheinlichkeit. Falls $S_- \leq S_+$ und

$$
\begin{aligned}
&P_\theta(\theta_j \in (S_-, S_+]) \\
&= P_\theta\left(\{S_-(X_1, ..., X_n) < \theta_j \leq S_+(X_1, ..., X_n)\}\right) \\
&= 1 - \alpha
\end{aligned}
$$

für alle $\theta \in \Theta$

Definition

Sei $(\Omega, \{P_\theta\}_{\theta\in\Theta}, \{X_j\}_{j\in\mathbb{N}})$ ein statistisches Modell und θ_j die j-te Komponente von $\theta = (\theta_1, \dots, \theta_q)$.

(i) Es seien S_-, S_+ zwei Schätzer und $\alpha \in (0,1)$ eine (kleine) Irrtumswahrscheinlichkeit. Falls $S_- \leq S_+$ und

$$\begin{aligned} &P_\theta(\theta_j \in (S_-, S_+]) \\ &= P_\theta\left(\{S_-(X_1, ..., X_n) < \theta_j \leq S_+(X_1, ..., X_n)\}\right) \\ &= 1 - \alpha \end{aligned}$$

für alle $\theta \in \Theta$, dann heißt das paar S_-, S_+ ein **Intervallschätzer** für θ_j zum Konfidenzniveau $1-\alpha$.

Technische Universität Berlin

Definition

(ii) Das zufallsabhängige Intervall

$$\big(S_-(X_1,\dots,X_n),S_+(X_1,\dots,X_n)\big]$$

nennen wir auch **Konfidenzintervall** für θ_j zum Konfidenzniveau $(1-\alpha)$.

Einfacher Fall:

Satz

Seien X_j i.i.d. mit $X_j \sim N(\mu, \sigma^2)$ mit bekannter Varianz σ^2 und sei $\theta = \mu \in \mathbb{R}$ der zu schätzenden Parameter.

Einfacher Fall:

Satz

Seien X_j i.i.d. mit $X_j \sim N(\mu, \sigma^2)$ mit bekannter Varianz σ^2 und sei $\theta = \mu \in \mathbb{R}$ der zu schätzenden Parameter. Mit dem p-Quantil z_p der von $N(0,1)$, erhält man

(i) *Das symmetrische, beidseitige Konfidenzintervall zum Konfidenzniveau $1-\alpha$ für den Erwartungswert μ mit gegebener Varianz σ^2 ist definiert durch*

$$S_{\pm}(X_1, \dots, X_n) = \bar{X} \pm z_{1-\alpha/2} \frac{\sigma}{\sqrt{n}}.$$

Satz

(ii) *Das linksoffene, einseitige Konfidenzintervall zum Konfidenzniveau* $1-\alpha$ *für den Erwartungswert* μ *mit gegebener Varianz* σ^2 *ist definiert durch*

$$S_- = -\infty, \quad S_+(X_1, \ldots, X_n) = \bar{X} + z_{1-\alpha}\frac{\sigma}{\sqrt{n}}.$$

Satz

(ii) *Das linksoffene, einseitige Konfidenzintervall zum Konfidenzniveau $1-\alpha$ für den Erwartungswert μ mit gegebener Varianz σ^2 ist definiert durch*

$$S_- = -\infty, \quad S_+(X_1, \ldots, X_n) = \bar{X} + z_{1-\alpha}\frac{\sigma}{\sqrt{n}}.$$

Gleiches gilt für das rechtsoffene, einseitige Konfidenzintervall

$$S_+ = \infty, \quad S_-(X_1, \ldots, X_n) = \bar{X} - z_{1-\alpha}\frac{\sigma}{\sqrt{n}} \ .$$

Beweis: (i):

- Für unabhängige Z.V. $X \sim N(\mu_X, \sigma_X^2)$, $Y \sim N(\mu_Y, \sigma_Y^2)$ gilt

$$X + Y \sim N(\mu_X + \mu_Y, \sigma_X^2 + \sigma_Y^2)$$

Beweis: (i):

- Für unabhängige Z.V. $X \sim N(\mu_X, \sigma_X^2)$, $Y \sim N(\mu_Y, \sigma_Y^2)$ gilt
- Induktiv folgt

$$X + Y \sim N(\mu_X + \mu_Y, \sigma_X^2 + \sigma_Y^2)$$
$$n\bar{X} \sim N(n\mu, n\sigma^2)$$

Beweis: (i):

- Für unabhängige Z.V. $X \sim N(\mu_X, \sigma_X^2)$, $Y \sim N(\mu_Y, \sigma_Y^2)$ gilt
- Induktiv folgt

$$X + Y \sim N(\mu_X + \mu_Y, \sigma_X^2 + \sigma_Y^2)$$

$$n\bar{X} \sim N(n\mu, n\sigma^2), \quad \text{also} \quad \bar{X} \sim N(\mu, \frac{1}{n}\sigma^2)$$

Beweis: (i):

- Für unabhängige Z.V. $X \sim N(\mu_X, \sigma_X^2)$, $Y \sim N(\mu_Y, \sigma_Y^2)$ gilt
$$X + Y \sim N(\mu_X + \mu_Y, \sigma_X^2 + \sigma_Y^2)$$
- Induktiv folgt
$$n\bar{X} \sim N(n\mu, n\sigma^2), \quad \text{also} \quad \bar{X} \sim N(\mu, \frac{1}{n}\sigma^2)$$

Damit

$$P\left(\bar{X} - z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}} \le \mu < \bar{X} + z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}\right)$$
$$= P\left(Z - z_{1-\alpha/2} \le 0 < Z + z_{1-\alpha/2},\right)$$

wobei $\mu = \mathbb{E}[X_1]$ den wahren Erwartungswert bezeichne (Standartisierung auf $Z \sim N(0,1)$).

- Wegen $-z_{1-\alpha/2} = z_{\alpha/2}$ erhalten wir

$$\begin{aligned} P\big(Z - z_{1-\alpha/2} \leq 0 < Z + z_{1-\alpha/2}\big) \\ &= F(z_{1-\alpha/2}) - F(-z_{1-\alpha/2}) \\ &= 1 - \frac{\alpha}{2} - \frac{\alpha}{2} \\ &= 1 - \alpha \end{aligned}$$

mit der Verteilung F von $N(0,1)$.

(ii)

- Wieder Standartisieren:

$$
\begin{aligned}
P\left(\mu < \bar{X} + z_{1-\alpha}\frac{\sigma}{\sqrt{n}}\right) &= P(0 < Z + z_{1-\alpha}) \\
&= P(-z_{1-\alpha} < Z) \\
&= P(z_\alpha < Z) \\
&= 1 - P(Z \leq z_\alpha) \\
&= 1 - \alpha
\end{aligned}
$$

(ii)

- Wieder Standartisieren:

$$
\begin{aligned}
P\left(\mu < \bar{X} + z_{1-\alpha}\frac{\sigma}{\sqrt{n}}\right) &= P(0 < Z + z_{1-\alpha}) \\
&= P(-z_{1-\alpha} < Z) \\
&= P(z_\alpha < Z) \\
&= 1 - P(Z \leq z_\alpha) \\
&= 1 - \alpha
\end{aligned}
$$

- Der linksoffene Fall geht analog. $\square$

- Beidseitige Konfidenzintervalle geben einen Vertrauensbereich, indem man mit sehr großer W.-keit $(1-\alpha)$ den wahren Wert findet. Man kann ihre (halbierte) Breite als typischen Schätzfehler auffassen.

- Beidseitige Konfidenzintervalle geben einen Vertrauensbereich, indem man mit sehr großer W.-keit $(1-\alpha)$ den wahren Wert findet. Man kann ihre (halbierte) Breite als typischen Schätzfehler auffassen.
- Einseitige Konfidenzintervalle geben eine obere/untere Schranke, die man mit sehr großer W.-keit $(1-\alpha)$ nicht über-/unterschreitet.

- Beidseitige Konfidenzintervalle geben einen Vertrauensbereich, indem man mit sehr großer W.-keit $(1-\alpha)$ den wahren Wert findet. Man kann ihre (halbierte) Breite als typischen Schätzfehler auffassen.
- Einseitige Konfidenzintervalle geben eine obere/untere Schranke, die man mit sehr großer W.-keit $(1-\alpha)$ nicht über-/unterschreitet.
- Typische Werte: $\alpha = 10\%, 5\%, 1\%$ oder 0.1%

Beispiel

- Eine Abfüllanlage füllt Limonade in Flaschen mit
  - einer Ungenauigkeit $(= \sigma)$ von 0.01 l und
  - einem Füllmengen-Sollwert von 0.75 l

Beispiel

- Eine Abfüllanlage füllt Limonade in Flaschen mit
  - einer Ungenauigkeit $(= \sigma)$ von 0.01 l und
  - einem Füllmengen-Sollwert von 0.75 l
- Fragestellung: Liegt die tatsächliche durchschnittliche Füllmenge unter 0.751 l?

Beispiel

- Eine Abfüllanlage füllt Limonade in Flaschen mit
  - einer Ungenauigkeit $(= \sigma)$ von 0.01 l und
  - einem Füllmengen-Sollwert von 0.75 l
- Fragestellung: Liegt die tatsächliche durchschnittliche Füllmenge unter 0.751 l?
- Stichprobe mit 10 Flaschen liefert durchschnittliche Füllmenge von $\bar{x} =$0.749 l.

Beispiel

- Eine Abfüllanlage füllt Limonade in Flaschen mit
  - einer Ungenauigkeit ($= \sigma$) von 0.01 l und
  - einem Füllmengen-Sollwert von 0.75 l
- Fragestellung: Liegt die tatsächliche durchschnittliche Füllmenge unter 0.751 l?
- Stichprobe mit 10 Flaschen liefert durchschnittliche Füllmenge von $\bar{x} =$0.749 l.
- Kann man mit 99%-iger Sicherheit sagen, dass der tatsächliche Durchschnittswert unter 0.751 l liegt?

Lösung:

- Wir nehmen an, die Füllmenge X ist normalverteilt mit bekanntem $\sigma_X =$0.01 und unbekanntem μ_X.

Lösung:

- Wir nehmen an, die Füllmenge X ist normalverteilt mit bekanntem $\sigma_X =$0.01 und unbekanntem μ_X.
- Wir berechnen das linksoffene Konfidenzintervall:

Technische Universität Berlin

Lösung:

- Wir nehmen an, die Füllmenge X ist normalverteilt mit bekanntem σ_X =0.01 und unbekanntem μ_X.
- Wir berechnen das linksoffene Konfidenzintervall:
  mit $z_{0.99} = 2.33$ wissen wir, dass zu 99% der wahre Wert μ_X kleiner ist als

$$\bar{x} + z_{0.99}\frac{\sigma_X}{\sqrt{n}} = 0.749 + 2.33\frac{0.01}{\sqrt{10}} = 0.756 \ .$$

Lösung:

- Wir nehmen an, die Füllmenge X ist normalverteilt mit bekanntem σ_X =0.01 und unbekanntem μ_X.
- Wir berechnen das linksoffene Konfidenzintervall:
  mit $z_{0.99} = 2.33$ wissen wir, dass zu 99% der wahre Wert μ_X kleiner ist als
  $$\bar{x} + z_{0.99}\frac{\sigma_X}{\sqrt{n}} = 0.749 + 2.33\frac{0.01}{\sqrt{10}} = 0.756 \ .$$
- Das 1%-Konfidenzintervall $(-\infty, 0.756)$ enthält 0.751

Lösung:

- Wir nehmen an, die Füllmenge X ist normalverteilt mit bekanntem σ_X =0.01 und unbekanntem μ_X.
- Wir berechnen das linksoffene Konfidenzintervall:
  mit $z_{0.99} = 2.33$ wissen wir, dass zu 99% der wahre Wert μ_X kleiner ist als

$$\bar{x} + z_{0.99}\frac{\sigma_X}{\sqrt{n}} = 0.749 + 2.33\frac{0.01}{\sqrt{10}} = 0.756 \ .$$

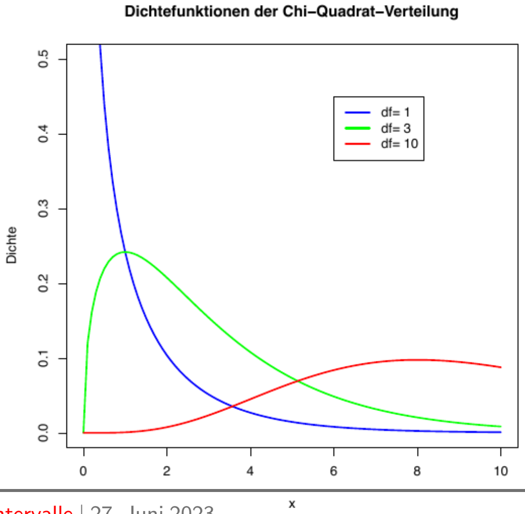
- Das 1%-Konfidenzintervall $(-\infty, 0.756)$ enthält 0.751
- Wir können also nicht mit 99%-iger Sicherheit sagen, dass der Wert kleiner 0.751 l, da die statistische Streuung bei einer derart kleinen Stichprobe zu groß ist.

Definition

Seien $X_1, \ldots, X_n$ i.i.d. mit $X_j \sim N(0,1)$. Sei weiter

$$K^2 = K^2(X_1, \ldots, X_n) = \sum_{j=1}^{n} X_j^2 \ .$$

Dann ist die χ^2-**Verteilung mit** n **Freiheitsgraden** definiert als $\mathrm{Vert}(K^2)$.


Dichtefunktionen der Chi–Quadrat–Verteilung
Dichte
x
0.0
0.1
0.2
0.3
0.4
0.5
0
2
4
6
8
10
df= 1
df= 3
df= 10

Satz

Im Gauß-Produktmodell $P_\theta^{(n)} = N(\mu, \sigma^2)^{\otimes n}$ mit zu schätzenden Parametern $\theta = (\mu, \sigma^2) \in \mathbb{R} \times \mathbb{R}_+ = \Theta$ sind das arithmetische Mittel $\bar{X}$ und die empirische Varianz σ^2 unabhängig. Es gilt

$$Z = \sqrt{n}\frac{\bar{X} - \mu}{\sigma} \sim N(0,1)$$

und

$$(n-1)\frac{\hat{\sigma}^2}{\sigma^2} = (n-1)\frac{\hat{\sigma}^2(X_1, \dots, X_n)}{\sigma^2} \sim \chi(n-1).$$

Definition

Seien X, Y ZV'n mit $X \sim N(0,1)$ und $Y \sim \chi^2(n)$. Dann definiert die Verteilung der Zufallsvariable

$$T = \sqrt{n}\frac{X}{\sqrt{Y}} \sim t(n)$$

die **Student-t-Verteilung** zu n **Freiheitsgraden**.

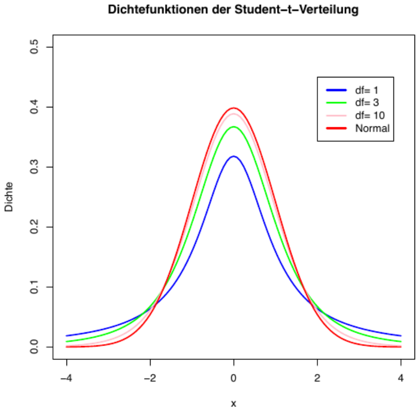
Definition

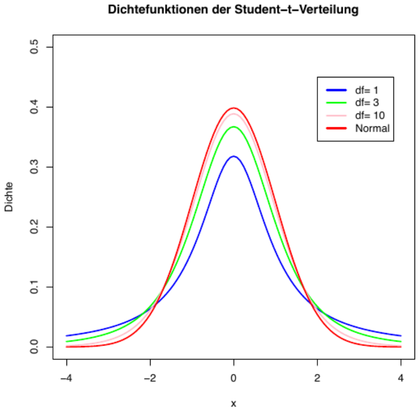
Seien X, Y ZV'n mit $X \sim N(0,1)$ und $Y \sim \chi^2(n)$. Dann definiert die Verteilung der Zufallsvariable

$$T = \sqrt{n}\frac{X}{\sqrt{Y}} \sim t(n)$$

die **Student-t-Verteilung** zu n **Freiheitsgraden**.

Diese wird die Rolle von $N(0,1)$ beim Schätzen der Varianz übernehmen.


Dichtefunktionen der Student–t–Verteilung
Dichte
0.0
0.1
0.2
0.3
0.4
0.5
–4
–2
0
2
4
x
df= 1
df= 3
df= 10
Normal



Symmetrisch: für $p < 0.5$ gilt wieder $-t_p(n) = t_{1-p}(n)$

Satz

Es seien wie oben $X_1, \ldots, X_n$ i.i.d. Zufallsvariablen mit $X_j \sim N(\mu, \sigma^2)$. Dann ist $T = T(X_1, \ldots, X_n)$

$$T = \sqrt{n}\frac{\bar{X} - \mu}{\hat{\sigma}}$$

Student-t-verteilt mit $n-1$ Freiheitsgraden, $T \sim t(n-1)$.

Satz

Es seien wie oben $X_1, \ldots, X_n$ i.i.d. Zufallsvariablen mit $X_j \sim N(\mu, \sigma^2)$. Dann ist $T = T(X_1, \ldots, X_n)$

$$T = \sqrt{n}\frac{\bar{X} - \mu}{\hat{\sigma}}$$

Student-t-verteilt mit $n-1$ Freiheitsgraden, $T \sim t(n-1)$.

Beweis: Wir rechnen

$$T = \sqrt{n}\frac{\bar{X} - \mu}{\hat{\sigma}} = \sqrt{n-1}\frac{\sqrt{n}\frac{\bar{X}-\mu}{\sigma}}{\sqrt{(n-1)\frac{\hat{\sigma}^2}{\sigma^2}}} = \sqrt{n-1}\frac{Z}{\sqrt{K^2}} \; .$$

Technische Universität Berlin

Satz

Seien, für $j = 1, \dots, n$, X_j normalverteilte i.i.d. Zufallsvariablen $X_j \sim N(\mu, \sigma^2)$ und $\theta = (\mu, \sigma^2) \in \mathbb{R} \times \mathbb{R}_+ = \Theta$ die im Produktmodell zu schätzenden Parameter.
Mit $t_p(n-1)$, dem p-Quantil der Student-t-Verteilung mit $n-1$ Parametern, erhält man

(i) *Das symmetrische, beidseitige Konfidenzintervall zum Konfidenzniveau $1-\alpha$ für den Erwartungswert μ mit geschätzter Varianz ist definiert durch*

$$S_\pm(X_1, \dots, X_n) = \bar{X} \pm t_{1-\alpha/2}(n-1)\frac{\hat{\sigma}}{\sqrt{n}}.$$

Satz

(ii) *Das linksoffene, einseitige Konfidenzintervall zum Konfidenzniveau $1-\alpha$ mit geschätzter Varianz ist definiert durch $S_- = -\infty$ und*

$$S_+(X_1,\dots,X_n) = \bar{X} + t_{1-\alpha}(n-1)\frac{\hat{\sigma}}{\sqrt{n}}.$$

und es ist ebenfalls ein Konfidenzintervall.
Gleiches gilt für das rechtsoffene, einseitige Konfidenzintervall mit $S_+ = \infty$ und

$$S_-(X_1,\dots,X_n) = \bar{X} - t_{1-\alpha}(n-1)\frac{\hat{\sigma}}{\sqrt{n}}\ .$$

Der Beweis ist analog zum Beweis für die Konfidenzintervalle im Gauß-Fall.

- $\sigma \quad \to \quad \hat{\sigma}$,
- $Z \sim N(0,1) \quad \to \quad T \sim t(n-1)$ und
- Quantile von $N(0,1)$
  $\to$ entsprechende Quantile von $t(n-1)$

Beispiel

- Mittlerer Ertrag von Maisanbauflächen (in Einheiten von ha) wird untersucht:
  - Stichprobengröße: $n = 30$
  - Mittlerer Ertrag: $\hat{\mu} = 25\,\mathrm{t}$
  - empirische Standardabweichung: $\hat{\sigma} = 5\,\mathrm{t}$

Beispiel

- Mittlerer Ertrag von Maisanbauflächen (in Einheiten von ha) wird untersucht:
  - Stichprobengröße: $n = 30$
  - Mittlerer Ertrag: $\hat{\mu} = 25\,\mathrm{t}$
  - empirische Standardabweichung: $\hat{\sigma} = 5\,\mathrm{t}$
- Gibt es ein Konfidenzintervall, sodass der wahre Wert μ für den zu erwartenden Maisertrag mit 95% W.-keit in diesem Intervall liegt?

Beispiel

- Mittlerer Ertrag von Maisanbauflächen (in Einheiten von ha) wird untersucht:
  - Stichprobengröße: $n = 30$
  - Mittlerer Ertrag: $\hat{\mu} = 25\,\mathrm{t}$
  - empirische Standardabweichung: $\hat{\sigma} = 5\,\mathrm{t}$
- Gibt es ein Konfidenzintervall, sodass der wahre Wert μ für den zu erwartenden Maisertrag mit 95% W.-keit in diesem Intervall liegt?

Lösung:

- Mit dem 2.5%-Quantil $t_{1-\alpha/2}(n-1) = t_{0.975}(n-1) = 2.045$ der $t(n-1) = t(29)$-Verteilung erhalten wir das 95%-Konfidenzintervall

$$\left[\bar{x} - t_{1-\alpha/2}\frac{\hat{\sigma}}{\sqrt{n}}, \bar{x} - t_{1-\alpha/2}\frac{\hat{\sigma}}{\sqrt{n}}\right) = [23.1\,,\ 26.9)\ .$$