3. Intervallschätzungen

3.1. Zufallsstichproben und Stichprobenfunktionen

3.1.1 Zufallsstichproben

Sei X eine Zufallsvariable und seien $X_1, X_2, \ldots, X_n$ Zufallsvariablen mit gemeinsamer Verteilung,

- dann heißen $X_1, X_2, \ldots, X_n$ **Zufallsstichprobe** aus X, wenn jedes X_i wie X verteilt ist. n ist **Stichprobenumfang** und die Werte $x_1, x_2, \ldots, x_n$

der Zufallsvariablen $X_1, X_2, \ldots, X_n$ heißen **Realisierung** oder **Wert der Stichprobe**.

- Eine Zufallsstichprobe $X_1, X_2, \ldots, X_n$ heißt **einfache Stichprobe** aus X, wenn $X_1, X_2, \ldots, X_n$ außerdem stochastisch unabhängig sind.

Eine Zahl, die von der Verteilung von X abhängt, nennt man einen **Verteilungsparameter**. Die wichtigsten Parameter sind der Erwartungswert, die Varianz, die Quantile sowie die Interwallwahrscheinlichkeiten.

<u>Beispiel:</u>

Die untersuchte Grundgesamtheit G besteht aus allen Haushalten in Deutschland im Jahr 2001. Untersuchungsmerkmal ist das verfügbare monatliche Haushaltseinkommen. Da die Grundgesamtheit sehr groß ist, soll nur ein Teil der Einkommen statistisch erhoben werden. Durch ein Experiment wird ein Haushalt zufällig ausgewählt (zufällig heißt, jeder Haushalt besitzt die gleiche Chance, ausgewählt zu werden). Dann werden durch erneute Zufallsauswahl aus der Grundgesamtheit G unabhängig voneinander weitere Haushalte gezogen und entsprechende Einkommen festgestellt.

Durch X_i (verfügbares monatliches Einkommen des Haushalts i) ist eine Zufallsvariable X gegeben. Die Einkommenswerte $x_1, x_2, \ldots, x_n$ können als Realisierungen einer Stichprobe aus $X_1, X_2, \ldots, X_n$ aus X aufgefasst werden.

3.1.2. Stichprobenfunktionen

Statistische Verfahren beruhen auf Zufallsstichproben aus einer oder mehreren Zufallsvariablen. Eine Stichprobe $X_1, X_2, \ldots, X_n$ aus X geht jedoch meistens nicht direkt, sondern über eine Funktion

$$g(X_1, X_2, \ldots, X_n)$$

in das Verfahren ein. Die Funktion $g(X_1, X_2, \ldots, X_n)$ ist selbst eine Zufallsvariable und wird **Stichprobenfunktion** genannt.

Wichtige Stichprobenfunktionen sind das Stichprobenmittel und die Stichprobenvarianz.

Das arithmetische Mittel der Beobachtungen X_i heißt **Stichprobenmittel**:

$$\overline{X} = \frac{1}{n} \sum_{i=1}^{n} X_i$$

Stichprobenvarianz:

$$S^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2 = \frac{1}{n}\sum_{i=1}^{n} X_i^2 - \overline{X}^2$$

3.2 Stichprobenverteilungen

Die Stichprobenparameter wie Mittelwert, Anteilswert, Varianz und andere sind Realisationen von Zufallsvariablen. Ihre Wahrscheinlichkeitsverteilungen nennt man **Stichprobenverteilungen**.

Verteilung des Stichprobenmittelwertes

Sei $X \sim N(\mu,\sigma)$ und sei $X_1, X_2, \ldots, X_n$ eine Stichprobe aus X (normal-verteilte Stichprobe). Dann gilt für das **Stichprobenmittel**:

$$\bar{X} \sim N(\mu, \frac{\sigma}{\sqrt{n}})$$

Denn bei großem Stichprobenumfang n gilt für die Verteilung des Stich-probenmittelwertes $\bar{X}$:

$$E(\bar{X}) = \mu \quad und \quad \sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}}$$

$\overline{X}$ ist (nach dem zentralen Grenzwertsatz) **annähernd normalverteilt**.

Dies bedeutet, dass die ZV, für die der beobachtete Stichprobenmittelwert $\overline{x}$ eine Realisation darstellt, asymptotisch normalverteilt ist mit den Parametern μ und $\frac{\sigma}{\sqrt{n}}$ (gilt bei beliebiger Verteilung von X und Unabhängigkeit (Stichprobe mit Zurücklegen)).

Man geht davon aus, dass bei einem Stichprobenumfang von $n > 30$ die Ausgangsverteilung keine Rolle spielt.

Aber:

nicht nur Stichprobenumfang, sondern auch Verteilung des Merkmals in der GG ist relevant

Durch Standardisieren der ZV $\bar{X}$ erhält man den **standardisierten Stichprobenmittel**:

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

ist **annähernd standardnormalverteilt**.

Hieraus folgt die Wahrscheinlichkeitsaussage

$$P(\mu - z \cdot \sigma_{\overline{X}} < \overline{X} < \mu + z \cdot \sigma_{\overline{X}}) = D(z)$$ **(direkter Schluss)**

Direkter Schluss = Schluss von der Grundgesamtheit auf die Stichprobe: „Mit welcher Wahrscheinlichkeit fällt ein Stichprobenmittelwert in ein vorher bestimmtes Intervall oder umgekehrt.“ Dabei beschreibt $D(z)$ die über dem symmetrischen Intervall [-*z*, +*z*] liegende Wahrscheinlichkeit.

Verteilung des Stichprobenanteilswertes

Entsprechendes gilt für den Anteilswert H einer großen Stichprobe:

$$E(H) = p \quad und \quad \sigma_H = \sqrt{\frac{p(1-p)}{n}}$$

H ist **annähernd normalverteilt** bzw. mit $q := 1 - p$,

$$\frac{H-p}{\sqrt{pq/n}} \sim N(0,1)$$

annähernd standardnormalverteilt

3.3 Intervallschätzungen mit großen Stichproben

Trotz der guten Eigenschaften der Punktschätzungen, weiß man nicht, welches Vertrauen man in sie legen kann. Anders ist das bei den Intervallschätzungen. Bei Intervallschätzungen kann man die Wahrscheinlichkeiten angeben, mit denen ein Schätzer in ein bestimmtes Intervall fällt.

Intervallschätzung basiert wie der direkte Schluss auf der Wahrscheinlichkeitsaussage. Man erhält nun ein Intervall um μ.

= Umkehrschluss oder Rückschluss

$$P(\bar{X} - z\sigma_{\bar{X}} \leq \mu \leq \bar{X} + z\sigma_{\bar{X}}) = D(z) = 1 - \alpha$$

Begriff des Konfidenzintervalls:

Gilt für das Zufallsintervall [Q_1;Q_2] und einen bestimmten Parameter q $P(Q_1 \leq q \leq Q_2) = 1 - \alpha$, d.h. überdeckt das Intervall den unbekannten Parameter q mit einer vorgegebenen Wahrscheinlichkeit 1-α, dann heißt [Q_1;Q_2] Konfidenzintervall für q zum Konfidenzniveau 1-α.

- 1-α = Konfidenzwahrscheinlichkeit. Sie gibt an, inwieweit man darauf vertrauen kann, dass der unbekannte Wert q im Konfidenzintervall liegt.

- α = Irrtumswahrscheinlichkeit. Sie gibt an, wie oft man sich im Mittel irrt, wenn man Konfidenzintervalle aufstellt.

- Q_1; Q_2 = Konfidenz- oder Vertrauensgrenzen.

Konfidenzintervalle für Mittelwerte

Die Intervallschätzung stellt die Umkehrung des direkten Schlusses (s.o.) dar und heißt auch **Umkehrschluss** oder **Rückschluss (indirekter Schluss)**.

Indirekter Schluss: Schluss von der Stichprobe auf die unbekannte Grundgesamtheit.

Konfidenzintervall um den Mittelwert μ:

$$KONF\ (\bar{x} - z\sigma_{\bar{X}} \leq \mu \leq \bar{x} + z\sigma_{\bar{X}}) = 1 - \alpha$$

In den meisten Fällen ist aber auch die Varianz des Merkmals in der Grundgesamtheit unbekannt, so dass man sie schätzen muss,

$$\hat{\sigma}_{\bar{X}} = \frac{\hat{\sigma}}{\sqrt{n}}$$

um das Konfidenzintervall zu bestimmen:

$$KONF\ (\bar{x} - z\hat{\sigma}_{\bar{X}} \leq \mu \leq \bar{x} + z\hat{\sigma}_{\bar{X}}) = 1 - \alpha$$

Dadurch entsteht eine zusätzliche Ungenauigkeit.

Konfidenzintervalle für Anteilswerte:

Bei großen Stichproben werden die Intervalle für Anteilswerte analog geschätzt Normalverteilung als Näherungsverteilung):

$$KONF\left(H - z\sqrt{\frac{pq}{n}} \leq p \leq H + z\sqrt{\frac{pq}{n}}\right) = D(z) = 1 - \alpha \qquad \text{(Rückschluss)}$$

Da zu schätzende p wird durch den Stichprobenanteilswert h ersetzt und das Konfidenzintervall für Anteilswerte lautet:

$$KONF\left(h - z\sqrt{\frac{h(1-h)}{n}} \leq p \leq h + z\sqrt{\frac{h(1-h)}{n}}\right) = 1-\alpha$$

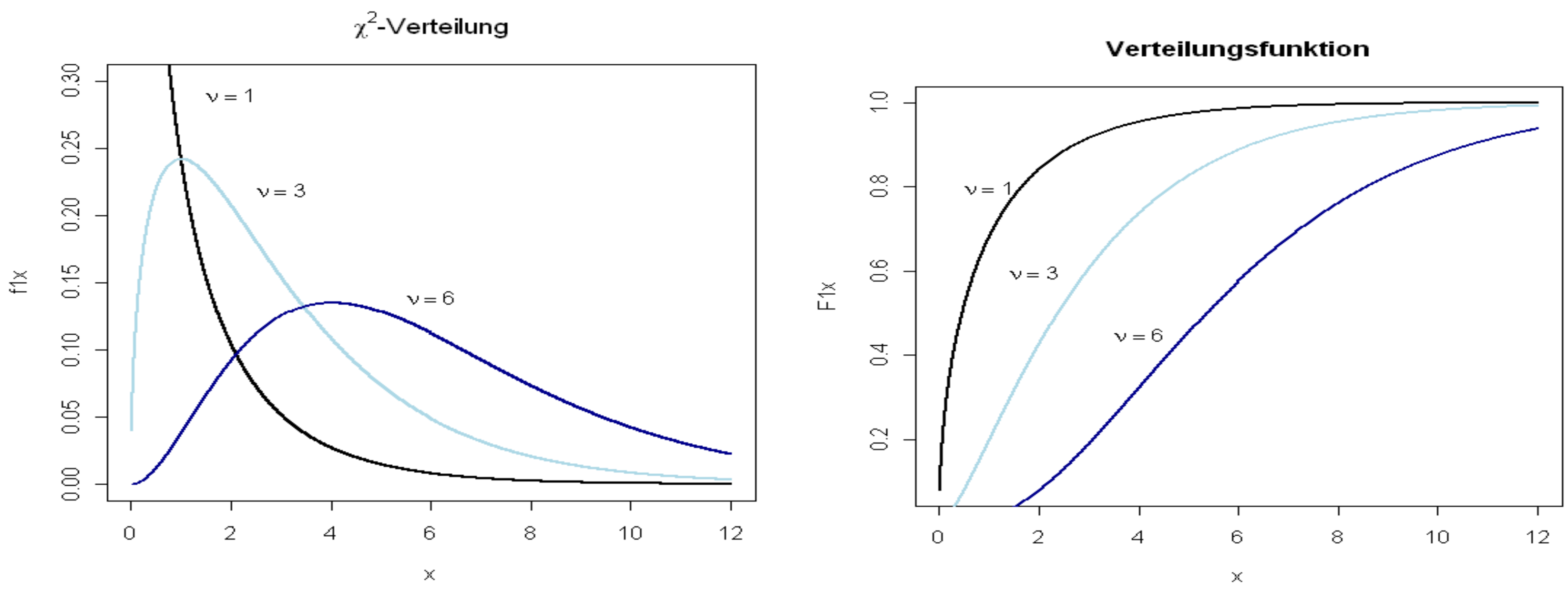
3.4 Chi-Quadrat-Verteilung

Seien $U_1, U_2, \ldots, U_\nu$ unabhängige *N*(0,1)-verteilte Zufallsvariablen. Dann besitzt

$$Q = \sum_{i=1}^{\nu} U_i^2 \sim \chi_\nu^2$$

eine Chi-Quadrat-Verteilung (ν = Anzahl der Freiheitsgrade).

Chi-Quadrat-Verteilung ist eine stetige Verteilung auf [0,∞] (positive Wahrscheinlichkeitsdichte). Diese Verteilung ist asymmetrisch und rechtsschief:



Der Erwartungswert und die Varianz dieser Verteilung sind:

$$E[Q] = \nu \quad und \quad V[Q] = 2\nu$$

Sei $X_1, X_2, \ldots, X_n$ eine Stichprobe aus $X \sim N(\mu, \sigma)$. Aus der Definition der Chi-Quadrat-Verteilung folgt, dass

$$\sum_{i=1}^{n} \left(\frac{X_i - \mu}{\sigma} \right)^2 \sim \chi_n^2$$

Es gilt außerdem:

$$\frac{nS^2}{\sigma^2} = \frac{1}{\sigma^2}\sum_{i=1}^{n}(X_i - \bar{X})^2 = \sum_{i=1}^{n}\left(\frac{X_i - \bar{X}}{\sigma}\right)^2 \sim \chi^2_{n-1}$$

Die Stichprobenvarianz S^2 ist somit verteilt wie das $\frac{\sigma^2}{n}$-fache einer χ^2_{n-1}-verteilten Zufallsvariable.

3.5 *t*-Verteilung

Seien U und Q stochastisch unabhängig, wobei U standardnormal verteilt $U \sim N(0,1)$ und Q chi-Quadrat-verteilt $U \sim \chi^2_\nu$ ist.

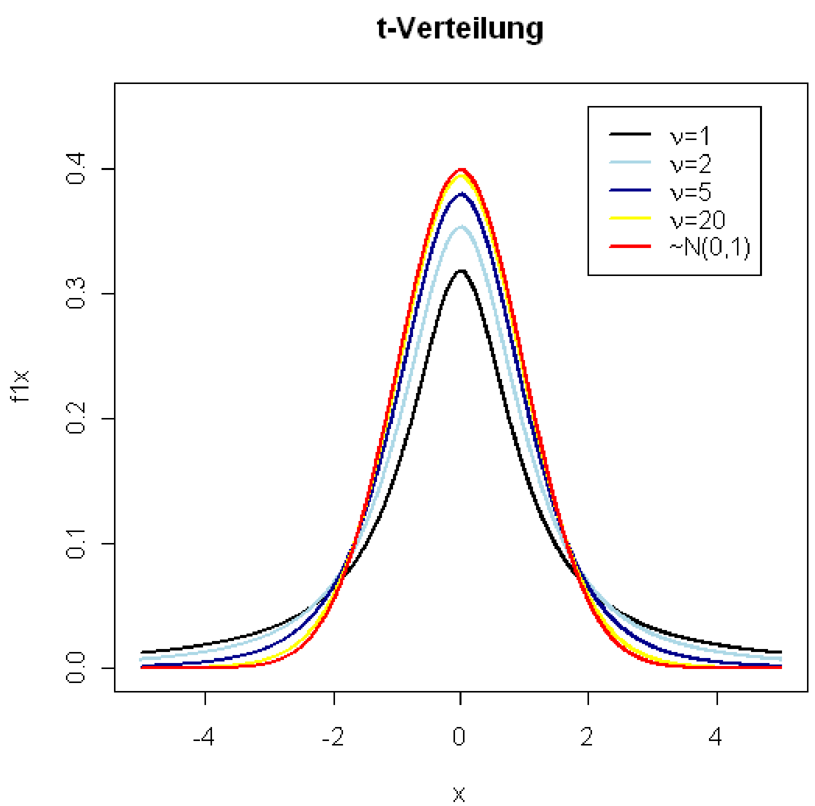
Die Verteilung von $W = \frac{U}{\sqrt{Q}}\sqrt{\nu}$ heißt *t*-Verteilung (auch Student-Verteilung genannt) mit Parameter ν:

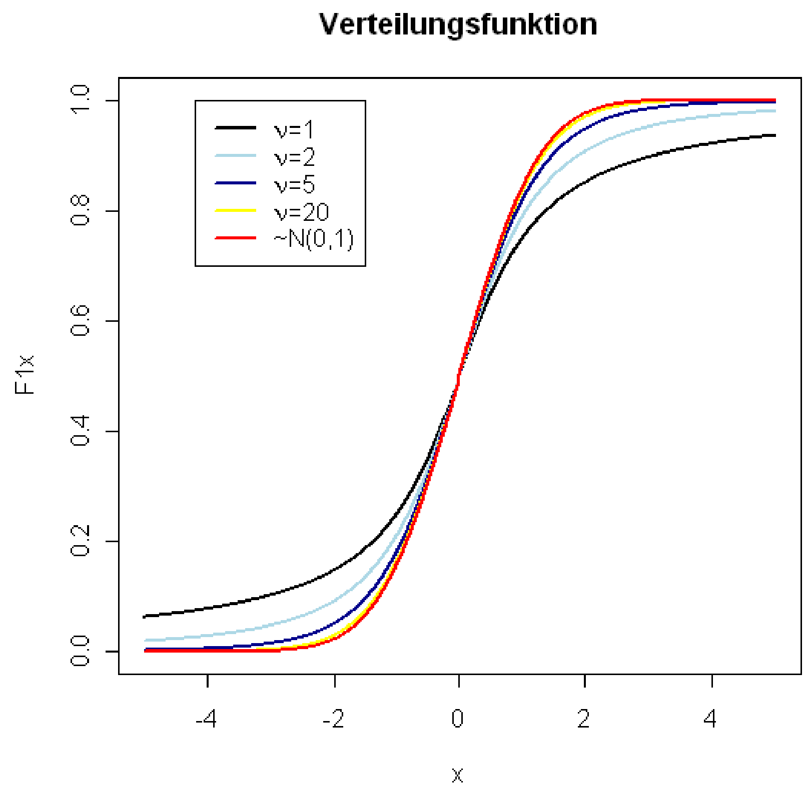
$$W = \frac{U}{\sqrt{Q}}\sqrt{\nu} \sim t_\nu$$ (mit ν = Anzahl der Freiheitsgrade).

Für diese Verteilung gilt:

$$E[W] = 0, \text{ falls } \nu \geq 2 \qquad \text{und} \qquad V[W] = \frac{\nu}{\nu - 2} > 1, \text{ falls } \nu \geq 3.$$

Die Dichtefunktion der *t*-Verteilung ähnelt der Dichte der Standardnormalverteilung. Sie ist symmetrisch um Null. An den Flanken weist jedoch die Dichte der *t*-Verteilung mehr Masse als die *N*(0,1)-Dichte auf. Mit steigender Anzahl der Freiheitsgrade (für $\nu \geq 40$) konvergiert die *t*-Verteilung gegen die Standardnormalverteilung.


t-Verteilung
f1x
x
0.0
0.1
0.2
0.3
0.4
-4
-2
0
2
4
ν=1
ν=2
ν=5
ν=20
~N(0,1)


Verteilungsfunktion
F1x
x
0.0
0.2
0.4
0.6
0.8
1.0
-4
-2
0
2
4
ν=1
ν=2
ν=5
ν=20
~N(0,1)

3.6 Intervallschätzung mit kleinen Stichproben

Bei *großen Stichproben* lassen sich die Konfidenzintervalle leicht konstruieren, da nach dem zentralen Grenzwertsatz die Normalverteilung eine gute Näherung für die wahre Stichprobenverteilung darstellt.

Bei *kleinen Stichproben* ist es jedoch nicht gegeben. Nur in den Situationen, wenn das Merkmal bereits *in der Grundgesamtheit* normalverteilt ist, wird die Konstruktion von Konfidenzintervallen wieder einfach. Wegen der *Reproduktionseigenschaft der Normalverteilung* ist der Stichprobenmittelwert auch normalverteilt.

Konfidenzintervalle für Mittelwerte

Wenn die Stichprobe aus einer normalverteilten Grundgesamtheit stammt, ist der Quotient $\frac{\bar{X}-\mu}{\hat{\sigma}_{\bar{X}}}$ auch für kleine Stichproben Student-*t*-verteilt mit *n*-1 Freiheitsgraden.

Hieraus folgt die Wahrscheinlichkeitsaussage:

$$P(-t \leq \frac{\bar{X}-\mu}{\hat{\sigma}_{\bar{X}}} \leq +t) = F_T(t) - F_T(-t)$$

Das Konfidenzintervall für μ lautet:

$$KONF(\bar{x} - t_{n-1}\hat{\sigma}_{\bar{X}} \leq \mu \leq \bar{x} + t_{n-1}\hat{\sigma}_{\bar{X}}) = 1 - \alpha$$

mit dem (1- α/2)-Quantil: $t_{n-1}[1-\alpha/2] = -t_{n-1}[\alpha/2]$

Konfidenzintervalle für Varianzen

Die Varianz S^2 einer Stichprobe ist eine Zufallsvariable. Ihre Verteilung lässt sich berechnen für den Fall, dass das Merkmal in der Grundgesamtheit annähernd normalverteilt ist und die einzelnen Stichproben unabhängig gezogen werden.

Wenn das Merkmal in der Grundgesamtheit normalverteilt ist, so ist der Quotient $n\frac{S^2}{\sigma^2} = \chi^2_{n-1}$ chi-Quadrat-verteilt mit n–1 Freiheitsgraden.

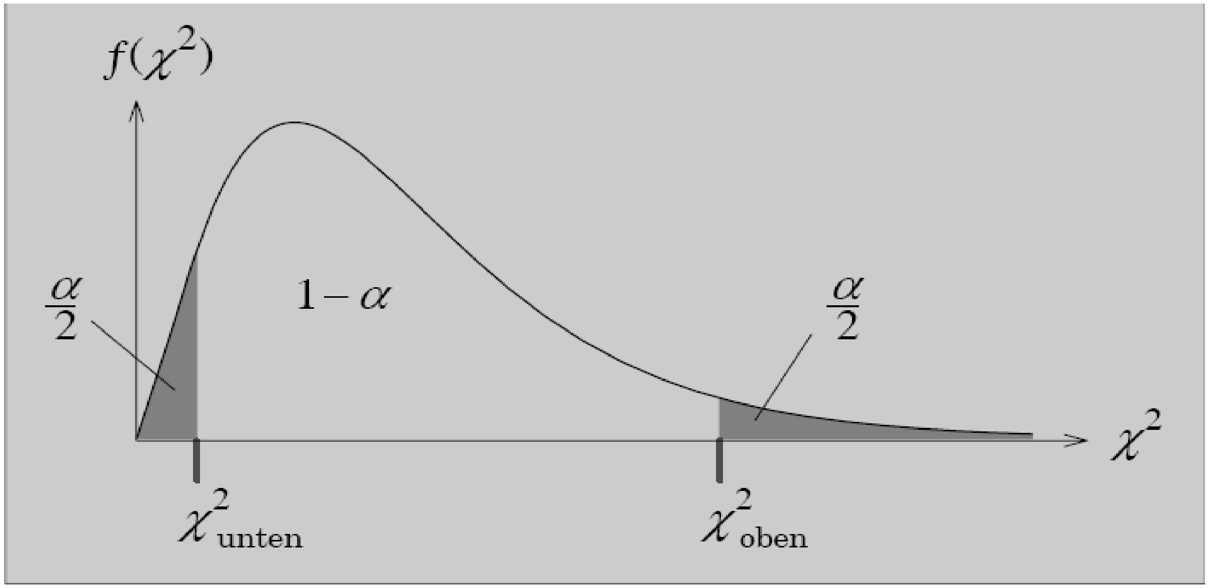
$$KONF\left(\frac{n \cdot s^2}{\chi^2_{n-1}[1-\alpha/2]} \leq \sigma^2 \leq \frac{n \cdot s^2}{\chi^2_{n-1}[\alpha/2]}\right) = 1-\alpha$$

Chi-Quadrat-Verteilung ist keine symmetrische Verteilung,

$$F(\chi^2_{unten}) = \alpha / 2$$

$$F(\chi^2_{oben}) = 1 - \alpha / 2$$

beide Stellen sind positiv.



Quelle: J. Schira (2003), S. 461.

3.7 Übersicht Varianzen

- Varianz in der Stichprobe

$$s^2 = \frac{1}{n}\sum_{i=1}^{n}(x_j - \bar{x})^2$$ mit n = Stichprobenumfang und $$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_j$$

- Varianz in der Grundgesamtheit

$$\sigma^2 = \frac{1}{N}\sum_{i=1}^{n}(x_j - \mu)^2$$ mit N = Stichprobenumfang und $$\mu = \frac{1}{N}\sum_{i=1}^{N} x_j$$

- Geschätzte Varianz in der Grundgesamtheit

$$\hat{\sigma}^2 = \frac{n}{n-1} s^2$$ mit *n-1* = Anzahl der Freiheitsgrade

- Varianz des Stichprobenmittelwertes

$$V(\bar{X}) = \sigma_{\bar{X}}^2 = \frac{\sigma^2}{n}$$

- Geschätzt Varianz des Stichprobenmittelwertes

$$\hat{V}(\bar{X}) = \hat{\sigma}_{\bar{X}}^2 = \frac{\hat{\sigma}^2}{n}$$