

Grundlagen der Inferenzstatistik: Was Ihnen nicht erspart bleibt!

- 1 Einführung**
- 2 Wahrscheinlichkeiten kurz gefasst**
- 3 Zufallsvariablen und Verteilungen**
- 4 Theoretische Verteilungen (Wahrscheinlichkeitsfunktion)**
- 5 Von der Grundgesamtheit zur Stichprobe**
- 6 Stichproben(kennwert)verteilungen**
- 7 Der Einfluss der Stichprobengröße**
- 8 Zusammenfassung und Ausblick**

1 Einführung

Grundlegendes Problem

- Daten einer Stichprobe können nicht unmittelbar auf die Grundgesamtheit übertragen werden.

Wichtigste Anwendungen

- Punkt- und Intervallschätzung (einzelne Kennwerte und Konfidenzintervalle)
- Signifikanztests (Testen von Hypothesen)

Voraussetzung

- Daten stammen aus einer einfachen (!) Zufallsstichprobe

2 Wahrscheinlichkeiten kurz gefasst

Unterscheidung in

- Subjektiv (intuitive)
- Klassische
- Empirische (frequentistische)

Generell variieren Wahrscheinlichkeiten zwischen $p = 0$ und $p = 1$ (p steht für *probability*) oder zwischen 0 % und 100 %.

Subjektive Wahrscheinlichkeit

p = Wert (Schätzung, Intuition, Wissen)

Klassische Wahrscheinlichkeit

$$p(A) = \frac{\text{Anzahl günstiger Fälle}}{\text{Anzahl möglicher Fälle}}$$

(A = Ereignis)

Beispiel Würfel

Will man eine „1“ würfeln, ist die Anzahl der günstigen Fälle gleich 1 und die Anzahl möglicher Fälle = 6.

Problem der klassischen Wahrscheinlichkeit: Das Vorgehen ist eingeschränkt, da die Whk. der Einzelereignisse eindeutig aus den Eigenschaften der betreffenden Gegenstände bestimmbar sein muss.

Empirische oder auch frequentistische Wahrscheinlichkeit

$$p(A) = \frac{f(A)}{f(A) + f(\neg A)}$$

(f steht für *frequency*)

A = „passende“ Beobachtung

$\neg A$ = „nicht passende“ Beobachtung

Schätzungen auf der Basis relativen Häufigkeiten, Grundlage sind Beobachtungen keine theoretischen Überlegungen

Beispiel Raucher

In einer Zufallsstichprobe von 20 Personen sind 6 Raucher und 14 Nichtraucher. Die Wahrscheinlichkeit, dass eine

zufällig ausgewählte Person der Grundgesamtheit raucht ist $p = \frac{6}{6+14} = 0,3$

Bedingte Wahrscheinlichkeiten

Beispiel (Sedlmeier 2008)

Ein Schüler hat im Abschlusszeugnis in Mathematik die Note Vier erhalten. Welche der folgenden Aussagen trifft für ihn mit größerer Wahrscheinlichkeit zu? Es gibt 30 Schüler, die eine Vier im Abschlusszeugnis stehen haben. Von denen 8 eine Sechs im Halbjahr hatten und 12 Nachhilfe bekommen haben:

- Er hatte eine Sechs in Mathe im Halbjahreszeugnis
- Er hatte eine Sechs in Mathe im Halbjahreszeugnis, hat aber im zweiten Halbjahr Nachhilfe in Mathe erhalten.

	x	x			
	x	x			
	x	x			
		x			
		x			

Rot = Nachhilfe ($n = 12$)

x = eine Sechs ($n = 8$)

- Wahrscheinlichkeit einer Sechs auf dem Halbjahreszeugnis $8/30$
- Wahrscheinlichkeit einer Nachhilfe $12/30$

Bedingte Wahrscheinlichkeit

Wie groß ist die Wahrscheinlichkeit, dass ein Schüler eine Sechs geschrieben hat, wenn er Nachhilfe bekommen hat? = Anteil der Schüler, auf die beides zutrifft (Sechs und Nachhilfe) an den Schülern die Nachhilfe hatten

$$p(A | B) = \frac{p(A \wedge B)}{p(B)}$$

| = Bedingung

\wedge = „und“

Eine bedingte Wahrscheinlichkeit für A gegen B erhält man, wenn man die Wahrscheinlichkeit der Konjunktion von A und B durch die Wahrscheinlichkeit der Bedingung B teilt!

$$p(\text{Sechs} | \text{Nachhilfe}) = \frac{p(\text{Sechs} \wedge \text{Nachhilfe})}{p(\text{Nachhilfe})} = \frac{5}{12}$$

Mann kann diese Wahrscheinlichkeit auch über die geschätzten Wahrscheinlichkeiten für Nenner und Zähler einzeln berechnen:

$$p(\text{Sechs} \wedge \text{Nachhilfe}) = \frac{f(\text{Sechs} \wedge \text{Nachhilfe})}{f(\text{alle Schüler})} = \frac{5}{30}$$

$$p(\text{Nachhilfe}) = \frac{f(\text{Nachhilfe})}{f(\text{alle Schüler})} = \frac{12}{30}$$

$$p(\text{Sechs} \mid \text{Nachhilfe}) = \frac{p(\text{Sechs} \wedge \text{Nachhilfe})}{p(\text{Nachhilfe})} = \frac{\frac{5}{30}}{\frac{12}{30}} = \frac{5}{12}$$

Beispiel

- a) Eine Hausfrau hat promoviert
- b) Eine promovierte Frau ist Hausfrau

Frage: Welche bedingte Wahrscheinlichkeit ist größer?

- a) $p(\text{promoviert}|\text{Hausfrau})$
- b) $p(\text{Hausfrau}|\text{promoviert})$

Lösung

- Die Wahrscheinlichkeit der Konjunktion (Hausfrau und promoviert) ist in beiden Fällen gleich
- Das eine Frau Hausfrau ist, ist viel wahrscheinlicher als das sie promoviert ist $p(\text{Hausfrau}) > p(\text{promoviert})$

- Die bedingte Wahrscheinlichkeit der Aussage a) $p(\text{promoviert} \mid \text{Hausfrau}) = \frac{p(\text{promoviert} \wedge \text{Hausfrau})}{p(\text{Hausfrau})}$

ist viel kleiner als b)

$p(A/B)$ Wahrscheinlichkeit für das Eintreten des Ereignisses A, wenn B wahr ist. („Wahrscheinlichkeit für A gegeben B“)

Grundlage der Aussage zur Wahrscheinlichkeit eines Signifikanztests: $p(D/H_0)$ ($D = \text{Daten}$)

3 Zufallsvariablen und Verteilungen

- Zufallsvariablen sind Variablen, deren mögliche Ergebnisse zwar bekannt sind, bei denen das aktuelle Ergebnis jedoch vom Zufall (meist dem Ergebnis einer Zufallsziehung einfache Zufallsauswahl) abhängt.
- **Definition:** Eine Zufallsvariable ist eine Funktion, die jedem möglichen Ergebnisse einer Zufallsziehung (z.B. einer Person) einen numerischen Wert zuweist (z.B. 1 = „männlich“ und 2 = „weiblich“). Die möglichen Werte , die eine Zufallsvariable annehmen kann, und deren Wahrscheinlichkeiten können als Verteilungen der Zufallsvariable dargestellt werden, Von solchen Zufallsvariablen kann man nun Erwartungswerte und Varianzen berechnen.
- Es gibt *diskrete* (endlich, abzählbar) und *kontinuierliche* (in einem bestimmten Bereich jede reelle Zahl) Zufallsvariablen.
- Zufallsvariablen werden mit Großbuchstaben bezeichnet.

Beispiel 1

Zufallsvariable „Häufigkeit des Wappens bei zweimaligem Werfen einer Münze“ X . X hat die Ausprägungen $x_1 = 0$, $x_2 = 1$ und $x_3 = 2$.

Der Ereignisraum (Ω) ist dann: $\Omega = \{0,1,2\}$.

Beispiel 2

Befragung zufällig ausgewählter Personen: Die Zufallsvariable Geschlecht kann die Ausprägungen „männlich“ und „weiblich“ haben. $\Omega = \{\text{männlich}, \text{weiblich}\}$.

4 Theoretische Verteilungen

- Für **diskrete** Zufallsvariablen, z.B. Gleichverteilung, Binomialverteilung, Poisson-Verteilung
- Für **metrische** Variablen, z.B. Normalverteilung, Exponentialverteilung

Normalverteilung

Wichtigste Verteilung ist die Normalverteilung.

Eigenschaften der Normalverteilung:

- stetige Verteilung
- symmetrische Dichtefunktion
- S-förmige Verteilungsfunktion
- Erwartungswert: $E(X) = \mu$
- Varianz: $\text{Var}(X) = \sigma^2$

Vorteil

- Approximation durch Normalverteilung: Mit wachsendem n nähern sich viele theoretische Verteilungen der Normalverteilung
- Empirische Verteilungen lassen sich ebenfalls oft durch die Normalverteilung annähern.

Standardnormalverteilung

Parameter:

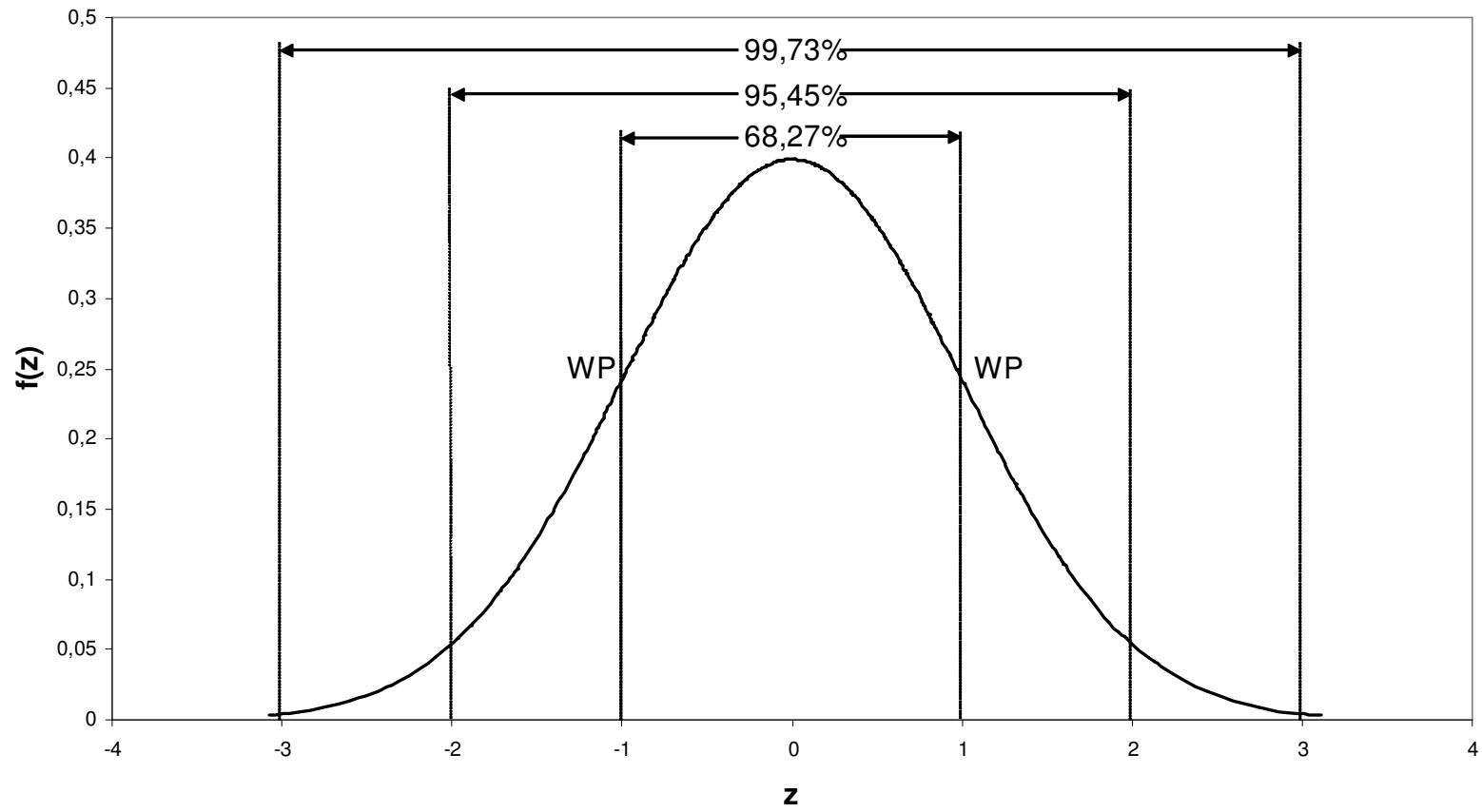
- $\mu = 0$
- $\sigma^2 = 1$
- Ist durch z-Transformation aus der Normalverteilung ableitbar.

$$\circ z_i = \frac{x_i - \mu}{\delta}$$

Vorteil

- Während es unendlich viele Normalverteilungen gibt, existiert nur eine Standardnormalverteilung ($\mu = 0, \sigma^2 = 1$)
- Die Standardabweichungen können genau interpretiert werden.
- Zwischen je einer Standardabweichung links und rechts vom Mittelwert liegen 68 % aller Werte.
- Der Zusammenhang zwischen Flächeninhalt und z-Wert ist in nahezu jedem Statistikbuch tabelliert zu finden.

Standardnormalverteilung



5 Von der Grundgesamtheit zur Stichprobe

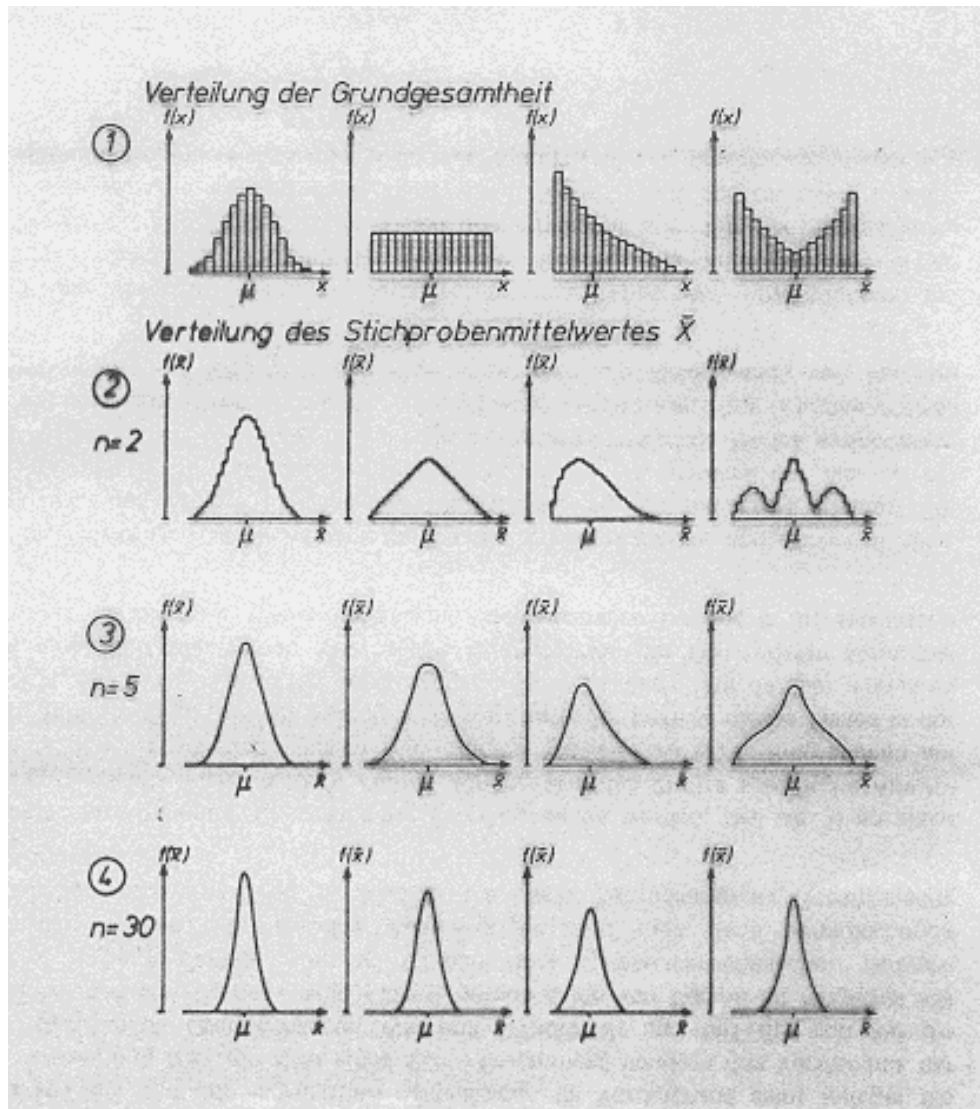
Maßzahlen	Parameter der Grundgesamtheit	Statistiken (Kennwerte) der Stichprobe	Schätzer der Parameter
Arithmetisches Mittel	μ	\bar{x}	$\hat{\mu}$
Standardabweichung	δ	s	$\hat{\delta}$
Varianz	δ^2	s^2	$\hat{\delta}^2$
Fallzahl	N	n	

6 Theoretische Stichproben(kennwert)verteilungen

- Nicht nur die einzelnen Untersuchungseinheiten können als Zufallsvariablen angesehen werden!
- Die Stichprobenkennwerte sind ebenfalls Zufallsvariablen. Sie folgen Stichproben(kennwert)verteilungen!
- Stichprobenverteilung:
 - Verteilung des arithmetischen Mittels der Stichprobe
 - So zusagen: Zusammenfassungen von hypothetischen Ergebnissen (n Stichproben)
 - Oder: Realisierung einer Stichprobenfunktion

Wie entstehen Stichprobenfunktionen?

Trainingsprogramm Sedlmeier/Köhlers (2001): Simulation, Stichprobenverteilung empirisch



Bleymüller et al. (1979)

Das Problem

- Wir haben nur eine Stichprobe!
- Hätte man einem gegebenen Stichprobenkennwert (z.B. Mittelwert) könnte man die Frage stellen: Wie sieht die dazugehörige Grundgesamtheit aus, die diesen Wert hervorgebracht hat!
- Wie kommen wir an eine Verteilung und Kennwerte die uns auf die Parameter der Grundgesamtheit schließen lassen?

Punkt und Intervallschätzung

- Einzelne Kennwerte, die für Stichprobendaten berechnet werden (z.B. Mittelwert), heißen Punktschätzer und schätzen den unbekannt Parameter der Grundgesamtheit.
- Oft gelten die Kennwerte, die die Stichprobe kennzeichnen, auch als beste Schätzung für die Grundgesamtheit (z.B. Mittelwert der Stichprobenverteilung)
- In manchen Fällen müssen aber die Kennwerte, die die Stichprobe charakterisieren, modifiziert werden, wenn sie als Schätzer für die Werte der Grundgesamtheit aufgefasst werden (z.B. **Varianz der Stichprobenverteilung**)
- Grundsätzlich ist aber nie bekannt, wie nahe der Schätzer aus der Stichprobe am wahren Wert der Grundgesamtheit liegt.
- Es kann aber unter bestimmten Annahmen ein Intervall angegeben werden, das mit einer bestimmten (festzulegenden) Wahrscheinlichkeit den wahren Wert überdeckt (enthält) **Konfidenzintervall**

Warum interessieren uns diese Verteilungen?

Weil uns die Frage interessiert:

- Bei einem gegebenen Stichprobenkennwert (z.B. dem Mittelwert für das Alter aller Befragten in einer Stichprobe): **Wie kann die Grundgesamtheit(mit einiger Wahrscheinlichkeit) aussehen, die diesen Wert hervorgebracht hat.**

Lage und Streuungsmaße von Stichprobenverteilungen: Punktschätzung

- Lagemaß: **Erwartungswert** (Ergibt sich als Mittelwert der Ergebnisse des oftmals Wiederholten zugrundeliegenden Experiments)
- Streuungsmaß: Varianz bzw. Wurzel daraus = Standardabweichung für Stichprobenverteilungen:
Standardfehler

Erwartungswert kontinuierlicher Zufallsvariablen

Der Erwartungswert für den Mittelwert ist der erwartungswert einer Zufallsvariable der wiederum dem Populationsmittelwert entspricht:

$$E(\bar{X}) = E(X) = \mu$$

Will man also den Erwartungswert einer Stichprobenverteilung für Mittelwerte bestimmen, nimmt man entweder den Stichprobenmittelwert oder benutzt eine theoretische Vorgabe.

Varianz kontinuierlicher Zufallsvariablen

Die Varianz einer Stichprobenverteilung muss aus der Stichprobenvarianz geschätzt werden.

$$\hat{\delta}^2 = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2$$

n = Anzahl der Werte

Aber: Dieser Schätzer ist nicht erwartungstreu. Er unterschätzt die Populationsvarianz tendenziell. (mathematisch ableitbar)

Deshalb gibt es eine Korrektur (mathematisch ableitbar)

$$\hat{\delta}^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2$$

Die Varianz der Stichprobenverteilung beim Schätzen des Mittelwerts ist dann

$$\hat{\delta}_{\bar{x}}^2 = \frac{\hat{\delta}^2}{n}$$

Für Konkrete Berechnungen, wie z.B. die Konfidenzintervalle wird allerdings meist auf den Standardfehler zurückgegriffen, da dieser in den Originaleinheiten ausgedrückt werden kann

$$\hat{\delta}_{\bar{x}} = \frac{\hat{\delta}}{\sqrt{n}} \text{ (bei unbekannter Varianz in der Grundgesamtheit)}$$

Der Standardfehler wird kleiner mit steigender Stichprobengröße und sinkender Varianz!

Erwartungswert und Varianz diskreter Zufallsvariablen

Werden selten gebraucht und sind deshalb nicht so wichtig.

Der Erwartungswert einer diskreten Zufallsvariable X ist der Populationsmittelwert μ . (mathematisch ableitbar)

Die Varianz einer Binomialverteilung $\sigma_x^2 = \sum_{i=1}^n (x_i - \mu)^2 p_i$ (mathematisch ableitbar)

7 Der Einfluss der Stichprobengröße

Gesetz der großen Zahlen

- Das Gesetz der Großen Zahlen sichert in vielen Fällen zu, dass der Stichprobenmittelwert bei wachsender Stichprobengröße gegen den Erwartungswert konvergiert.
- Der Mittelwert konzentriert sich mit wachsendem n immer mehr um den gemeinsamen Erwartungswert μ der X_i .

Beispiel

Trainingsprogramm Sedlmeier/Köhlers (2001): Simulation, Stichprobenverteilung theoretisch

Zentraler Grenzwertsatz

Beschreibt die mathematische Gesetzmäßigkeit, die Auswirkungen auf die Form von Stichprobenverteilungen hat. Jede gebräuchliche Stichprobenverteilung (also jede Verteilung von Summen, Mittelwerten oder Zufallsvariablen) nähert sich mit steigender Stichprobengröße ($n \rightarrow \infty$) immer mehr der so genannten Normalverteilung an. Oder in standardisierter Form der Standardnormalverteilung. Auch eine Binomialverteilung kann durch eine Normalverteilung ersetzt werden, wenn die Faustregel $np(1-p) > 9$ gilt.

- Es folgt: Die Verteilung des arithmetischen Mittels von n unabhängig identisch verteilten Zufallsvariablen X_i strebt mit wachsendem Stichprobenumfang n gegen eine Normalverteilung mit Erwartungswert μ und Varianz σ^2/n .
- Das arithmetische Mittel ist „asymptotisch normalverteilt“.
- Faustregel: wenn $n > 30$, dann ist die Normalverteilung eine gute Näherung für die Stichprobenverteilung.

8 Zusammenfassung und Ausblick

- Sowohl die Untersuchungseinheiten als auch die Kennwerte einer Stichprobe können als Zufallsvariablen betrachtet werden. Kennwerte einer Stichprobe können also selbst als Stichprobe betrachtet werden.
- Diese Stichprobenverteilungen sind der Schlüssel zu den Werten der Grundgesamtheit.
- Sind Stichproben groß genug ($n > 30$, besser $n > 100$) kann man für die Stichprobenverteilungen in der Regel annehmen, dass sich diese einer Normalverteilung annähern.
- Bei kleinen n können andere zentrale Verteilungen (t-, F- und Chi-Quadrat) angewendet werden. Diese haben aufgrund der theoretischen Verteilung ebenfalls solche Integrationsmöglichkeiten.