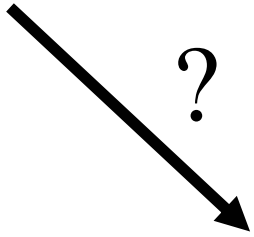
# Inferenzstatistik (=schließende Statistik)

- Grundproblem der Inferenzstatistik:
  - Wie kann man von einer Stichprobe einen gültigen Schluß auf di Grundgesamtheit ziehen
  - Bzw.: Wie groß sind die Fehler, die man dabei macht
- Stichprobenparameter: $\bar{x}, s, s^2$




- Parameter der Grundgesamtheit: $\mu, \sigma, \sigma^2$
- Weitere Parameter: Anteilswerte, Zusammenhangsmaße, Regressionskoeffizienten etc.

# Punktschätzungen

- Sind bereits bekannt:
- Punktschätzung für $\hat{\mu}$ : $\bar{x}$
- Punktschätzung für $\hat{\sigma}^2$ : $s^2 \times \frac{n}{n-1}$
- Punktschätzung für $\hat{\sigma}$ : $\sqrt{\hat{\sigma}^2}$
- Problem:
  - Schätzung beruht auf (Zufalls-)Stichproben
  - Stichprobenwerte sind Ausprägungen einer Zufallsvariable
  - Deshalb entsprechen sie praktisch niemals exakt dem Wert in der Grundgesamtheit

# Wiederholung: Repräsentativität

- „Die Repräsentativitätslüge. Meinungsumfragen nennen sich oft ‚repräsentativ', tatsächlich aber werden die Befragten meist nach dem Zufallsprinzip ausgewählt" (aus „Vorwärts", 10/1994, S. 23)
- Für uns sind Stichproben repräsentativ wenn gilt:
  - Jedes Element der Grundgesamtheit hat die gleiche
  - oder eine angebbare Chance in die Stichprobe zu gelangen

# Wiederholung: Zufallsexperiment

- Zufallsexperimente
  - können theoretisch beliebig oft wiederholt werden
  - Einzelergebnisse hängen vom Zufall ab, Verteilung der Ergebnisse ist aber bekannt
  - Bei häufiger Wiederholung nähert sich die empirische Verteilung der theoretischen Verteilung an
- Stichprobenziehung ist ein Zufallsexperiment:
  - Wer in die Stichprobe kommt, hängt nur vom Zufall ab
  - Stichprobenparameter (z. B. Mittelwert) sind Zufallsvariablen, die von vielen zufälligen Größen (den Ausprägungen für die einzelnen Fälle) abhängig sind

# Zufallsvariablen

- Im Einzelfall weiß man nicht, welchen Wert die Variable annimmt
- Aber: Ausprägungen von Zufallsvariablen sind nicht willkürlich, sondern höchst regelmäßig verteilt
- Die *Verteilung* der Werte einer Zufallsvariablen ist in der Regel bekannt
- Zufallsvariablen (und ihre Verteilungen) können diskret oder stetig sein

# Einfaches Zufallsexperiment mit n=1

- Ein nicht-präparierter Würfel wird einmal (n=1) geworfen
- Das Ergebnis wird notiert
- Dieses Zufallsexperiment wird sehr oft (>10000) wiederholt
- Die relativen Häufigkeiten (=Prozentwerte) für die einzelnen Augenzahlen nähern sich dabei den erwarteten Häufigkeiten an
- Für die diskrete Zufallsvariable ergibt sich eine diskrete Gleichverteilung (sechs Ausprägungen á 16,667%)

# Additives Zusammenwirken von Zufallsgrößen (n>1)

- Das Zufallsexperiment wird mit *zwei* Würfeln wiederholt
- Die Zufallsvariable wird nun gebildet, indem die *Summe* beider Augenzahlen gebildet wird
- Für die Zufallsvariable ergibt sich eine flache unimodale Verteilung mit elf diskreten Ausprägungen
- Noch mehr Würfel:
  - die Zahl der Ausprägungen nimmt immer mehr zu
  - die Form der Verteilung nähert sich einer Glocke an (Normalverteilung)

# Normalverteilung

- Ist ein *Modell* für die Verteilung von Zufallsvariablen
- Ist dann geeignet, wenn diese ihrerseits auf das additive Zusammenwirken (Beispiel: arithmetisches Mittel in einer Stichprobe) von Zufallsgrößen zurückgeführt werden können und n>30

$$y = f(x) = \frac{1}{\sigma \times \sqrt{2 \times \pi}} \times e^{-\frac{(x-\mu)^2}{2 \times \sigma^2}}$$

- Ist eine stetige Verteilung mit zwei Parametern:
  - D.h.: für einen einzelnen Wert ist die Wahrscheinlichkeit = 0
  - Für alle Werte zwischen -∞ und + ∞ ist die Wahrscheinlichkeit=1. Dieser Wahrscheinlichkeit entsprich die gesamte Fläche unter der Kurve
  - Für einen bestimmten Wertebereich ist die Wahrscheinlichkeit gleich der Fläche unter der Kurve
- Die Normalverteilung ist symmetrisch und unimodal

# Zentraler Grenzwertsatz

- Ein Stichprobenmittelwert kann als Ausprägung einer Zufallsvariable verstanden werden, die ihrerseits auf das additive Zusammenwirken von Zufallsvariablen zurückgeht
- Die Mittelwerte von (theoretische unendlich vielen) Zufallsstichproben, die aus einer Grundgesamtheit gezoge werden, verteilen sich in Gestalt einer Normalverteilung um den Mittelwert der Grundgesamtheit (den „wahren Mittelwert“)
- Dabei spielt es keine Rolle, wie die Ausgangsvariable in der Grundgesamtheit und in den Stichproben verteilt ist

# ZGWS II

| Variable | Verteilung |
| --- | --- |
| Ausgangsvariable in der GG, z.B. Alter | z.B. linkssteil |
| ⬇ | ⬇ |
| Zufallsvariable: Alter eines einzelnen Befragten | wie in GG, d.h. linkssteil |
| ⬇ | ⬇ |
| additive Zufallsvariable: Arithmetisches Mittel des Alters in einer Stichprobe | normalverteilt |

# ZGWS III

- Für die Normalverteilung, der die Stichprobenmittelwerte folgen, gilt:
  - Der Mittelwert der Normalverteilung ist gleich dem Mittelwert in der GG (wahrer Mittelwert)
  - Die Streuung (Standardabweichung) der Normalverteilung wird als Standardfehler (des Mittelwertes) bezeichnet
- Der Standardfehler hängt ab:
  - von der Streuung in der GG (die in der Regel geschätzt werden muß)
  - von der Stichprobengröße

# Formel für den Standardfehler

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} = \sqrt{\frac{\sigma^2}{n}}$$

Der Standardfehler (Streuung der Stichprobenmittelwerte um den wahren Mittelwert) ist um so größer, je größer die Streuung in der GG

Der Standardfehler ist um so größer, je kleiner die Stichprobe

Der Zusammenhang zwischen Stichprobenumfang und Standardfehler ist umgekehrt quadratisch: Um den Schätzfehler zu halbieren, muß der Stichprobenumfang vervierfacht werden

# Eigenschaften der Normalverteilung

- Es existiert eine ganze Familie von Normalverteilungen: Breite und Lage einer Normalverteilung hängen von Mittelwert und Streuung a
- Für *alle* Normalverteilungen gilt:
  - 68% der Fläche bzw. der Fälle liegen in einem Intervall von einer Standardabweichung um den Mittelwert
  - 95% der Fläche liegen in einem Intervall von 1,96 Standardabweichungen um den Mittelwert
- Die Verteilung der Standardnormalverteilung (Mittelwert=0, Standardabweichung=1, auch: z-Verteilung) ist tabelliert
- Die Werte dieser Tabelle sind für *alle* Normalverteilungen anwendbar (z-Transformation und deren Umkehrung)
- Viele theoretischer Verteilungen gehen bei steigender Fallzahl in die Normalverteilung über (t-Verteilung, Binomialverteilung etc.)

# Konfidenzintervalle

- Aus dem ZGWS und den Eigenschaften der Normalverteilung ergib sich:
- Bei wiederholter Stichprobenziehung werden 95% der Stichprobenmittelwerte in einem Intervall von ±1,96 Standardfehler vom wahren Mittelwert der GG liegen
- Daraus ergibt sich umgekehrt: Ein Intervall von ±1,96 Standardfehlern um den Stichprobenmittelwert wird in 95% aller Fälle den wahren Mittelwert einschließen
- D.h., es ist jetzt möglich, aufgrund der Stichprobenwerte ein Konfidenzintervall anzugeben, das mit einer Wahrscheinlichkeit vor 95% den wahren Wert beinhaltet (Intervallschätzung)

# Konfidenzintervalle II

- Zur Berechnung von Konfidenzintervallen müssen noch zwei kleinere Probleme gelöst werden:
  - Die Stichprobenmittelwerte sind nicht standardnormal-, sondern normalverteilt ➔ Umkehrung z-Transformation, damit man mit den „kritischen Werten“ aus der Tabelle rechnen kann
  - Schätzung der Streuung in der Grundgesamtheit nach der schon bekannten Formel: $\hat{\sigma}^2 = s^2 \times \frac{n}{n-1} = \frac{SAQ}{n} \times \frac{n}{n-1} = \frac{SAQ}{n-1}$
  - Wegen dieser zusätzlichen Unsicherheit müßte eigentlich die bei gleicher Grundform breitere t-Verteilung verwendet werden (breitere Intervalle)
  - Bei großen Fallzahlen (n > 1000) geht die t-Verteilung aber in di Normalverteilung über

# Konfidenzintervalle III

- Die Wahrscheinlichkeit, daß das Intervall den wahren Wert *nicht* umfaßt (Irrtumswahrscheinlichkeit), wird mit $\alpha$ bezeichnet.
- Die Vertrauenswahrscheinlichkeit (Wahrscheinlichkeit, daß das Intervall den wahren Wert einschließt) beträgt 1- $\alpha$ und wird manchmal mit $\gamma$ bezeichnet
- In der Forschungspraxis werden üblicherweise Vertrauenswahrscheinlichkeiten von 95% und von 99% verwendet, das entspricht kritischen z-Werten von ±1,96 bzw. ±2,58

$$\bar{x} - z_{(1-\frac{\alpha}{2})} \times \frac{\hat{\sigma}}{\sqrt{n}} \leq \mu \leq \bar{x} + z_{(1-\frac{\alpha}{2})} \times \frac{\hat{\sigma}}{\sqrt{n}}$$

# Konfidenzintervalle IV

- Konfidenzintervalle können auch für andere Stichprobenparameter (z.B. Anteilswerte) berechnet werden
- Die Verteilung von Anteilswerten aus Stichproben folgt einer Binomialverteilung, die durch die Normalverteilung approximiert werden kann, wenn die Bedingung $n*p*(1-p) \geq 9$ erfüllt ist
- Als Streuung eines Anteilswertes gilt $p*(p-1)$
  - wenn zwei nominale Merkmale gleichhäufig sind, sind die zugehörigen Objekte sehr heterogen
  - je geringer der Anteil eines Merkmals, desto homogener sind die Objekte
- Durch Einsetzen ergibt sich

$$p - z_{(1-\frac{\alpha}{2})} \times \sqrt{\frac{p \times (1-p)}{n}} \leq \Theta \leq p + z_{(1-\frac{\alpha}{2})} \times \sqrt{\frac{p \times (1-p)}{n}}$$

# Warum z-Transformation?

- Verteilung der Stichprobenmittelwerte um den wahren Mittelwert = Normalverteilung
  - Mittelwert = wahrer Mittelwert z.B. 37,5
  - Streuung = Standardabweichung = Standardfehler des Mittelwertes z.B. 0,8
  - Problem: In welchem Bereich liegen 95% der Fläche???
- Mathematisches Modell = *Standard*normalverteilung
  - Mittelwert = 0
  - Streuung = Standardabweichung = 1
  - 95 % der Fläche im Bereich 1,96 -1,96

# Warum z-Transformation?

- Lösung: Für alle Normalverteilungen gilt, daß 95 Prozent der Fläche im Bereich von 1,96 Standardabweichungen, 99 Prozent im Bereich von 2,58 Standardbweichungen um den Mittelwert liegen
- Um die „kritischen Werte“ aus der tabellierten Standardabweichung verwenden zu können, müssen diese mit der tatsächlichen Standarbweichung der Normalverteilung multipliziert werden (Umkehrung der z-Transformation)
- Anschließend muß noch der tatsächliche Mittelwert addiert werden, weil sich die Intervalle beziehen sich auf diesen und nicht auf den MW der Standardnormalverteilung von 0 beziehen

# Warum z-Transformation

- Z-Transformation x➔z
- Vom x-Wert Mittelwert substrahieren
- Dann durch Std.abweichung teilen
- Bsp. (35,9-37,5)/0,8=-1,96; (39,1-37,5)/0,08=1,96
- Nachschlagen in der tabellierten Std.normalverteilung zeigt, daß zwischen 35,9 und 39,1 95% der Fläche (95% aller Stichprobenmittelwerte) liegen

- Umkehrung z➔x
- Z-Wert mit Standardabweichung multiplizieren
- Und Mittelwert addieren
- Bsp. (-1,96*0,8)+37,5=35,9; (1,96*0,8)+37,5=39,1
- 95% der Fläche (und der Stichprobe liegen zwischen diesen beiden Werten. Aus den kritischen Werten der Standardnormalverteilung könne kritische Werte für eine beliebige Normalverteilung ermittelt werden