Datum Name

Schliessende Statistik

Grundgesamtheit

Schluss von der auf die | Stichprobe Grundgesamtheit

Stichprobe

Schätzen & Testen

Problemstellung
Stichprobenverteilung
Schätzen von Mittelwerten und Anteilswerten
Testen von Mittelwerten und Anteilswerten
Relevanz vs Signifikanz

Ein Zufallsversuch „produziert“ Ergebnisse.

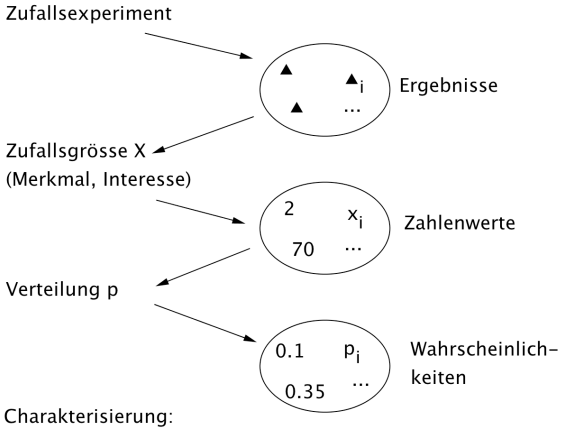
Einem solchen Ergebnis kann man eine Zahl zuordnen
(zB einer zufällig ausgewählten Person ihr Gewicht).

Dieser Zuordnung sagen wir Zufallsgrösse, weil:
Je nach Ergebnis nimmt sie einen anderen Wert an.
Weil das Ergebnis zufällig ist, ist auch der Wert zufällig.

Jeder Wert wird mit einer bestimmten Wahrscheinlichkeit angenommen.
Die Verteilung „verteilt“ die Wahrscheinlichkeiten auf die einzelnen Werte.

Jede Verteilung X wird „charakterisiert“ durch

- den Erwartungswert $\mu(X)$
- die Standardabweichung $\sigma(X)$



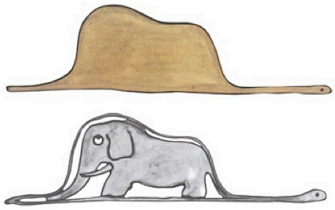
Charakterisierung:
→ Erwartungswert μ (Mittelwert)
→ Standardabweichung σ (Streuung)

Es gibt Verteilungen, die sehr häufig auftreten und deshalb einen eigenen Namen tragen.

Merke

Eine **Verteilung** sagt,

- **welche Zahlen** (sie stehen auf der X-Achse)
- **wie oft vorkommen** (Häufigkeit steht auf der y-Achse)



Verteilungen werden oft als „Kurven“ dargestellt.
Sie enthalten die gesamte Information.

Die **Normalverteilung** ist die wichtigste Verteilung, weil:

- Häufig ist eine Situation abhängig von vielen verschiedenen Einflüssen, welche sich „im Normalfall“ gegenseitig ausgleichen und damit zur Mitte tendieren.
  Solche Situationen werden von der Normalverteilung modelliert (vgl. auch Anhang 1).
- Die Binomialverteilung lässt sich mit Hilfe der Normalverteilung approximieren (Stetigkeitskorrektur!).

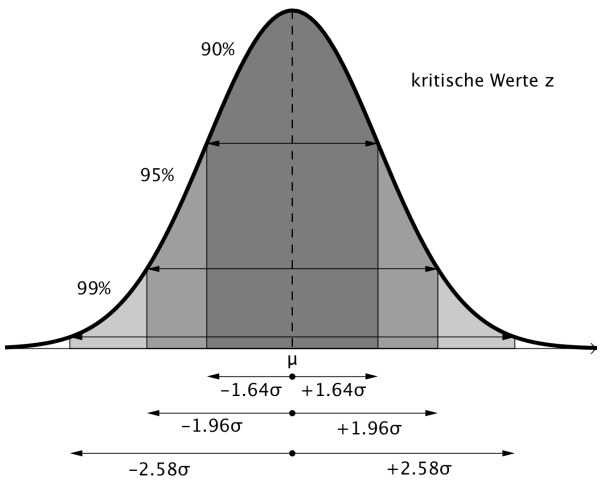
Bei einer Normalverteilung lassen sich symmetrische Bereiche um den Mittelwert μ angeben, in die ein Wert mit einer gewissen Wahrscheinlichkeit „fällt“.

$$P(\mu - z\cdot\sigma \le X \le \mu + z\cdot\sigma) = \ldots\ \%$$

Oft verwendete sogenannte **kritische Werte** sind

- $z = 1.64$ $P(\mu - 1.64\cdot\sigma \le X \le \mu + 1.64\cdot\sigma) = 90\%$
- $z = 1.96$ $P(\mu - 1.96\cdot\sigma \le X \le \mu + 1.96\cdot\sigma) = 95\%$
- $z = 2.58$ $P(\mu - 2.58\cdot\sigma \le X \le \mu + 2.58\cdot\sigma) = 99\%$

Je grösser z, desto „breiter“ das Intervall, desto grösser die dazugehörige Wahrscheinlichkeit.



Merke

„quick-and-dirty“-Regel: P(Bereich $\mu \pm 2\sigma$) ≈ **95%**

Inhalt

1 Einstieg

Das eigentliche Problem – Stichprobe vs Grundgesamtheit

Warum kann eine solche Frage bzw. deren Antwort von Interesse sein?

Die Statistik versucht auf Fragen wie die folgende eine Antwort zu geben:

Wie gross ist ein 18-jähriger Jugendlicher im Schnitt?

Was für eine Vorgehensweise schlagen Sie vor?

Beispiel 1 **Wir „vereinfachen" das Problem**

Wie gross ist dieser Mittelwert in *Ihrer Klasse* ?

Begriffe
- Grundgesamtheit GG
- Stichprobe SP
- Stichprobengrösse/Umfang n

a) Wählen Sie eine Stichprobe, werten Sie sie aus und geben Sie eine Antwort.

b) Richtig oder falsch? Kreuzen Sie die richtigen Aussagen an.

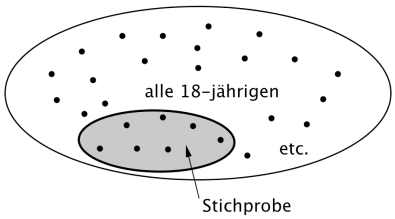
- ❒ Mit einer SP kann ich den genauen Mittelwert der GG nicht berechnen. Ich kann ihn nur *schätzen*.
- ❒ Ich will den Mittelwert der GG wissen, deshalb nützt mir eine SP gar nichts.
- ❒ Zwei verschiedene SP ergeben im Normalfall auch zwei verschiedene Mittelwerte.
- ❒ Gibt es in der GG 20 Werte, dann kann ich genau 4 Stichproben mit Umfang n = 5 ziehen.

c) Wie viele verschiedene Stichproben vom Umfang n sind in Ihrer Klasse möglich?

d) Die GG ist hier überschaubar. Wir können eine „Vollerhebung" durchführen.
Wie weit lag Ihre Stichprobe daneben?

Hinweis *TR / data / STAT–REG / 2: 1–Var Stats*

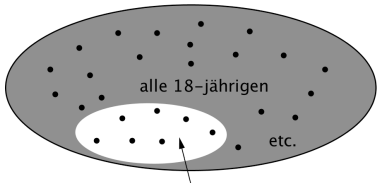
Wir kehren zurück zum ursprünglichen Problem. Die GG sind wieder *alle* 18-Jährigen.
Der Mittelwert ist unbekannt – wir kennen die GG nicht, sie ist viel zu gross.
Wir können den Mittelwert aber *schätzen* durch das Erheben einer Stichprobe.
Annahme: Stichprobe = Ihre Klasse. Wie lautet die Schätzung?



Unser Problem

- Wir haben <u>*nur eine*</u> einzige SP (**sample**) und wollen damit einen Wert in der GG (**population**) schätzen...
- Wir wissen weiter, dass jede Stichprobe wohl eine andere Schätzung ergibt...
- Wie wollen wir wissen, wie „brauchbar" <u>*unsere*</u> Schätzung ist?

Grundgesamtheit unbekannt ("verborgen")



Im nächsten Abschnitt lernen wir ein (geniales!) Werkzeug kennen, das eine Brücke schlägt zwischen der Stichprobe und der Grundgesamtheit.

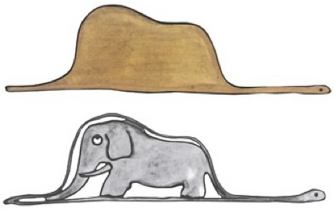
Stichprobenverteilung – Bindeglied zwischen Stichprobe und Grundgesamtheit

... mit welcher *Wahrscheinlichkeit* hatten wir mit unserer Stichprobe „Glück“?

Schätzungen und Schlussfolgerungen aus Stichproben sind nie ganz sicher. Ist etwas nicht ganz sicher, sondern nur wahrscheinlich, dann kommt die Wahrscheinlichkeitsrechnung ins Spiel und mit ihr die Verteilungen.

Eine **Verteilung** gibt an, **„welche Zahlen wie oft vorkommen“**.

- Die Zahlenwerte stehen auf der x-Achse.
- Die Häufigkeiten (absolut, relativ) stehen auf der y-Achse.

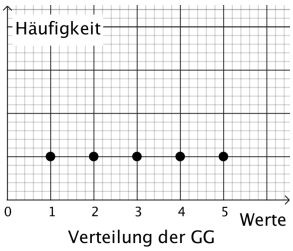


Verteilungen werden oft als „Kurven“ dargestellt. Sie enthalten die gesamte Information.

Beispiel 2 Gedankenexperiment im Kleinen

Zuerst lösen wir ein sehr einfaches Beispiel, an dem wir aber vieles „sehen“ können. Sei GG = $\{1,2,3,4,5\}$. Von dieser (als unbekannt angenommen) GG möchte man den Mittelwert bestimmen... Man „zieht“ eine Stichprobe vom Umfang n = 2.

- Wie viele verschiedene SP sind möglich? *(mit/ohne ZL, Reihenfolge...)*
- Bilden Sie die *Verteilung aller möglichen Stichprobenmittelwerte*. Skizzieren Sie! Was fällt auf?

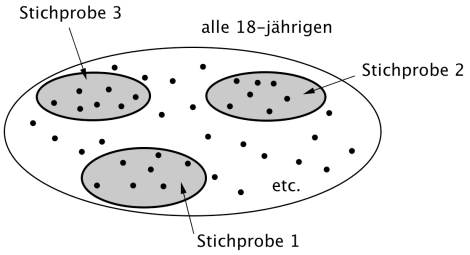


Verteilung der GG

Beispiel 3 Gedankenexperiment im Grossen

a) Wir denken nun „grösser“. Sei GG = $\{$alle 18-Jährigen$\}$.

- Skizzieren Sie eine mögliche Verteilung der Körpergrössen von 18-jährigen.
- Für welchen Wert interessieren wir uns? Zeichnen Sie ihn ein.



b) Wir haben den Mittelwert schon geschätzt – mit Ihrer Klasse als Stichprobe. Eine andere Stichprobe hätte wohl eine andere Schätzung ergeben. Um beurteilen zu können, wie (un-) wahrscheinlich unsere Schätzung ist, machen wir ein **theoretisches – geniales! – Gedankenexperiment**.

Wir stellen uns vor, wir ziehen ...

- ... eine weitere Strichprobe und berechnen deren Mittelwert.
- ... und noch eine und berechnen deren Mittelwert...
- ... und noch eine und berechnen deren Mittelwert...

Wir haben nun ganz viele Mittelwerte!
Wie könnte die *Verteilung dieser Mittelwerte* aussehen? Skizzieren Sie!

Hinweis *Beispiel 2.*
Dazu: in welchem Bereich liegen „viele“ Werte, wo „wenige“?

c*) Wir haben nun zwei Verteilungen:

- die Verteilung der GG (**population distribution**)
- die Stichprobenverteilung (**sampling distribution**)

Worin unterscheiden sie sich? Gibt es auch Gemeinsamkeiten?

Wir tun so, als ob wir alle möglichen Stichproben ziehen könnten – und berechnen deren Mittelwerte. Alle Stichproben sollen denselben Umfang haben.

Beachte

Der Mittelwert der Population heisst auch **„theoretischer Mittelwert“**.
Nicht, weil er nicht existiert, sondern weil er unbekannt ist.
Er ist **„fix“** und wird mit dem **griechischen Buchstaben μ** abgekürzt.

Der Mittelwert der Stichprobe heisst auch **„empirischer Mittelwert“**.
Er wird aus der Stichprobe berechnet. Er ist **„variabel“**, weil jede Stichprobe einen anderen Mittelwert generiert und wird mit dem **lateinischen Buchstaben** $\bar{x}$ abgekürzt.

Beispiel 4 Stichprobenverteilung des Mittelwertes

a) Wir könnten ein anderes „Merkmal“ von 18-Jährigen untersuchen.

X = Körpergewicht

- Skizzieren Sie eine mögliche Verteilung.
- Für welchen Wert interessieren wir uns? Zeichnen Sie ihn ein.
- Skizzieren Sie gerade unterhalb die Verteilung der Mittelwerte (**Stichprobenverteilung des Mittelwertes)**

b) Wie a) für

X = Anzahl Geschwister

c) Während die Verteilung in der GG je nach Merkmal eine *beliebige* Form annehmen kann, scheint die Stichprobenverteilung stets eine *bestimmte* Form zu besitzen. Welche? Haben Sie dafür eine Erklärung?

d) Wir haben *zwei* zusammengehörende Verteilungen und damit

- zwei Mittelwerte μ_{GG} und μ_{SV}. Welchen Zusammenhang erkennen Sie zwischen diesen?
- zwei Standardabweichungen σ_{GG} und σ_{SV}. Welchen Zusammenhang erkennen sie zwischen diesen?

e) Die Stichprobenverteilung hängt ab von der Grösse n der Stichprobe. Es gibt also zu jeder Stichprobengrösse n eine (leicht andere) Stichprobenverteilung. Hat die Stichprobengrösse n einen Einfluss auf

- den Mittelwert der Stichprobenverteilung μ_{SV}?
- die Standardabweichung der Stichprobenverteilung σ_{SV}?

f) Welcher Verteilungstyp scheint die SV zu besitzen?
Geben Sie weiter mögliche Formeln an für

- den Mittelwert μ_{SV}
- die Standardabweichung σ_{SV}.

Je grösser n, umso schmaler wird die Stichprobenverteilung...
Das leuchtet ein! Warum?

Zusammenfassung

Die Gesetzmässigkeit, dass die Stichprobenverteilung – unabhängig von der zugrundeliegenden Verteilung der GG – sich einer Normalverteilung nähert, lässt sich auch mathematisch beweisen.
Weil diese Aussage so wichtig ist, trägt sie einen eigenen Namen. Sie heisst **zentraler Grenzwertsatz**.

- Die Stichprobenverteilung ist angenähert normalverteilt.
  Damit gelten für sie aber alle Eigenschaften (σ-Regeln!),
  die wir von der Normalverteilung her kennen.
- Weiter ist: μ_{SV} = σ_{SV} =

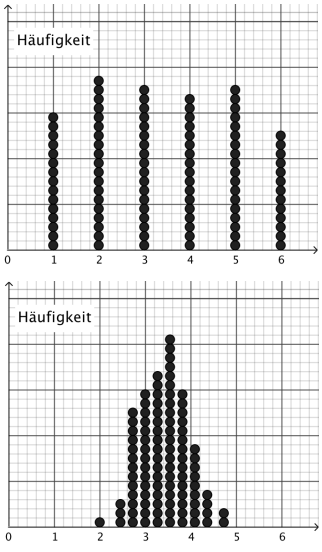
Beispiel *Zentraler Grenzwertsatz würfeln*

Oberes Bild: Ein Würfel wird 100-mal geworfen.
Es ergibt sich eine (fast) gleichverteilte Grundgesamtheit.

Simulation!

Unteres Bild: Aus der GG wurden viele Stichproben mit dem Umfang n = 10 gezogen und jeweils der Mittelwert gebildet.
Es ergibt sich eine glockenförmige Kurve, die Stichprobenverteilung.

Sie sehen: Mittelwert und Streuung verhalten sich gemäss Gesetzmässigkeit!



2 Schätzen

Wir wollen einen unbekannten Wert der GG mit Hilfe einer Stichprobe *schätzen*.
Wir können dies tun mit Hilfe einer

- Punktschätzung
- Intervallschätzung

Die Punktschätzung gibt *einen* Schätzwert für den unbekannten Wert an. Die Intervallschätzung gibt ein *ganzes Intervall* an, in dem der unbekannte Wert mit einer gewissen Wahrscheinlichkeit liegt.

Punktschätzer

Beispiel 5 **Punktschätzung des Mittelwertes**

a) Wie gross ist ein 18-Jähriger im Schnitt?
Eine Stichprobe mit n = 30 ergab die folgenden Werte (X = Körpergrösse in cm):

```
157 160 162 166 168 170 172 174 175 178 179 180 181 184 185 190
    160     166 168 170     174 175         180 181
            166     170     174 175
            166     170
```

Berechnen und interpretieren Sie den Punktschätzer (**point estimation**).

b) Die Interpretation des Punktschätzers als „beste Wahl" ist intuitiv klar, lässt sich aber auch begründen. Sehen wir uns dazu unser einfaches Beispiel GG = $\{1,2,3,4,5\}$ an.

- Listen Sie alle möglichen Stichproben mit n = 2 auf und berechnen Sie deren Mittelwert.
- Berechnen Sie den Mittelwert dieser Mittelwerte. Was erhalten Sie?

Beachte

- Der Mittelwert aller Stichprobenmittelwerte ist gleich dem Mittelwert μ der GG. Das heisst: der Populationsmittelwert μ wird durch das Stichprobenmittel $\bar{x}$ weder systematisch unter- noch überschätzt. Man spricht deshalb von einer unverzerrten oder *erwartungstreuen Schätzung*.
- Wir haben nun an einem Beispiel durchgerechnet, dass tatsächlich gilt: $\mu_{SV} = \mu_{GG}$.

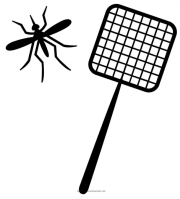
Beispiel 6 **Punktschätzung der Streuung**

Mit einer Stichprobe lässt sich nicht nur der Mittelwert der GG schätzen, sondern auch die *Streuung* der GG.
Wie wird man das wohl machen?
Schätzen Sie in Beispiel 5a) die Streuung der GG.

Hinweis *Der TR gibt „zwei" Antworten: σ_X und s_X ? Und jetzt? Welches ist die „beste" Schätzung? Es ist s_X. Warum? Der Anhang 1 gibt darauf die Antwort.*

Konfidenzintervall

Der Punktschätzer gaukelt eine „Pseudogenauigkeit“ vor:
Wir schätzen den unbekannten Wert „punktgenau“, obwohl wir ihn sehr selten exakt treffen werden.
Wir können eine Punktschätzung mit dem Versuch vergleichen, eine winzige Fliege (den wahren, aber unbekannten Wert) mit einer Stecknadel (dem Punktschätzer) zu treffen.



Erfahrene Fliegenfänger*innen würden aber der Stecknadel wohl eine Fliegenklatsche vorziehen.
Die Entsprechung in der Statistik ist ein Intervall. In diesem Intervall sollte dann der unbekannte Wert liegen. Darauf *vertrauen* wir – zumindest mit einer gewissen Wahrscheinlichkeit.
Ein solches Intervall heisst deshalb *Konfidenzintervall* (**confidence interval**).

Das wirklich Gute: Die grosse Vorarbeit haben wir mit der Stichprobenverteilung bereits geleistet ☺.

Beispiel 7 **Herleitung Konfidenzintervall – der „Umkehrtrick“**

Lesen Sie zuerst folgende Erklärung langsam durch – verstehen Sie *alles*.

1 Der unbekannte Wert μ_{GG} einer GG soll geschätzt werden.
Dazu wird der Wert einer Stichprobe erhoben und ausgewertet: Wir erhalten den sogenannten ***Punkt*schätzer** $\bar{x}$.

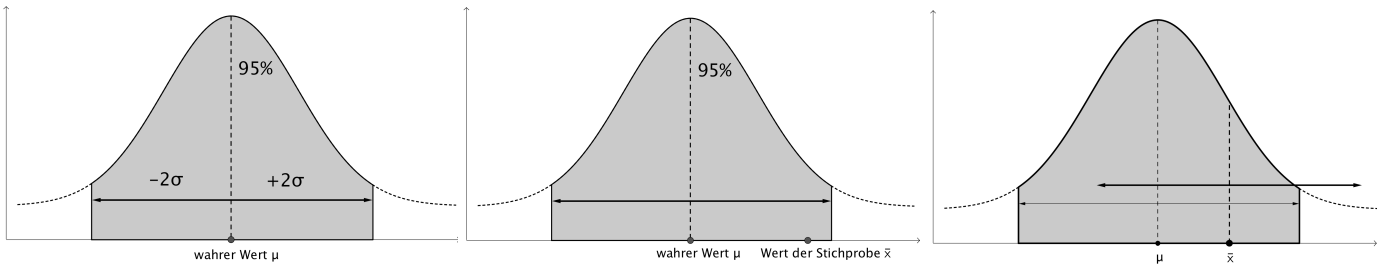
2 Wir wissen von früher: Wir können die Strichprobenverteilung bilden. Für diese gilt:

- normalverteilt (damit gelten die σ-Regeln für Intervalle)
- $\mu_{SV} = \mu_{GG}$; $\sigma_{SV} = \frac{\sigma_{GG}}{\sqrt{n}}$

Das heisst: Mit einer Wahrscheinlichkeit von ≈ 95% liegt ein aus der Stichprobe ermittelter Wert $\bar{x}$ höchstens 2 Standardabweichungen vom wahren Wert μ weg („quick and dirty“ Regel).

3 Das heisst aber umgekehrt **(Umkehrtrick)**:
Der wahre Wert μ liegt höchstens zwei Standardabweichungen vom aus der Stichprobe ermittelten Wert $\bar{x}$ weg!

Die folgende Bildabfolge soll dies verdeutlichen:



a) Begründen Sie folgende Formel für das Konfidenzintervall des Mittelwertes:

μ liegt mit 95 % Wahrscheinlichkeit im Intervall $[\bar{x} - 2\cdot\sigma_{SV}; \bar{x} + 2\cdot\sigma_{SV}] = [\bar{x} - 2\cdot\frac{\sigma_{GG}}{\sqrt{n}}; \bar{x} + 2\cdot\frac{\sigma_{GG}}{\sqrt{n}}]$

bzw. allgemeiner: **μ liegt mit ... % Wahrscheinlichkeit im Intervall $[\bar{x} \pm z\cdot\sigma_{SV}] = [\bar{x} \pm z\cdot\frac{\sigma_{GG}}{\sqrt{n}}]$.**

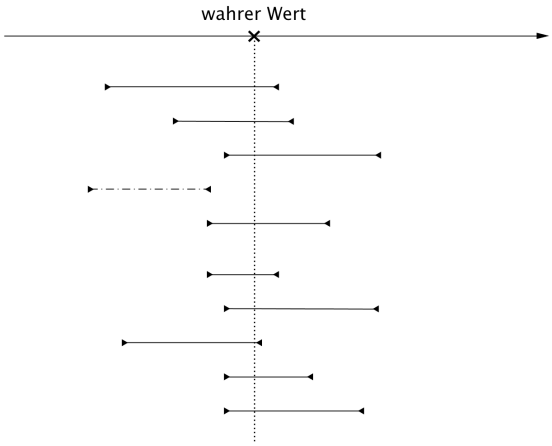
z-Werte
z = 1.65 ≙ 90%
z = 1.96 ≙ 95%
z = 2.58 ≙ 99%

b) *Berechnen und interpretieren* Sie ein 95%-Konfidenzintervall zu den Daten aus Beispiel 5a).

Hinweis *σ_{GG} ist zwar unbekannt. Aber wir schätzen es mit Hilfe der Stichprobe: $\sigma_{GG} \approx s$.*

Beachte

Bei der **Interpretation** eines Konfidenzintervalls ist **Vorsicht** geboten.



Es ist im Grunde falsch zu sagen:
„Mit einer Wahrscheinlichkeit von zB. 95% liegt der wahre Wert im Konfidenzintervall."

Die Wahrscheinlichkeit bezieht sich nämlich nicht auf die Lage des unbekannten Wertes (der liegt nämlich *fix* an seiner Stelle), sondern auf die Wahl der Stichprobe und damit auf das Konfidenzintervall.

Jede Stichprobe erzeugt ein *anderes* Konfidenzintervall mit einer eigenen Lage – und einer eigenen Breite (abhängig von s).
In der Abbildung überdeckt von den 10 berechneten Konfidenzintervallen eines den wahren Wert nicht.

Richtigerweise müsste man formulieren (*Häufigkeitsinterpretation*):
„In 95% aller Fälle, in denen ein 95%-Konfidenzintervall berechnet wird, liegt der wahre Wert der Grundgesamtheit tatsächlich in dem errechneten Intervall."

Nur ist es eben so, dass diese Formulierung ziemlich umständlich ist ☺. Who cares.

Beispiel 8 **Stadianer am Handy**

Sie wollen herausfinden, wie viele Stunden ein Stadianer im Schnitt täglich am Handy ist.
Die entsprechende Umfrage bei 30 Schülern ergab die nebenstehenden Resultate.

1.5	2	2.5	3	3.5	4	4.5	5	5.5
		2.5	3	3.5	4	4.5	5	
		2.5	3	3.5	4	4.5	5	
			3	3.5	4	4.5		
				3.5	4			
				3.5	4			
				3.5				

Berechnen und interpretieren Sie möglichst genau in Worten zu dieser Stichprobe

- den Punktschätzer.
- das 90%-Konfidenzintervall.

Punktschätzer und Konfidenzintervall für Anteilswerte

Sie wollen wissen:

Wie gross ist der Anteil unter den 18-Jährigen, welche blaue Augen haben?

Beispiel 9 **Punktschätzer**

a) Machen Sie einen Punktschätzer!

Hinweis *Als natürliche Stichprobe wählen Sie ...*

b) Als Statistikprofi ist dies einem aber nicht genug. Ein Konfidenzintervall wäre nett...

Hat er blaue Augen?

Beachte

Wir haben bisher immer mit *Mittelwerten* gerechnet. Dazu haben wir *gemessen* (Grösse, Gewicht, Dauer, ...).
Jetzt haben wir es mit etwas anderem zu tun. Wir *zählen* und erhalten einen *Anteilswert* in % (blaue Augen, Raucher, ...).

Beispiel 10 **Mittelwert und Anteilswert, Formel für das Konfidenzintervall**

Lesen Sie zuerst folgende Erklärung langsam durch – verstehen Sie *alles*.

a) Mittelwert und Anteilswert sind das Gleiche

Wir wollen uns das an einem Beispiel verdeutlichen. Lesen Sie!

Wir nehmen dazu eine Gruppe von Personen, sagen wir 10.
Es ist klar, wie wir die Durchschnittsgrösse ausrechnen:
Wir „messen" alle 10 Personen und teilen dann durch 10.

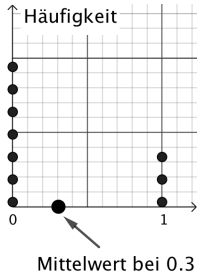
Genau das gleiche machen wir aber auch beim Anteilswert:
Wir „messen", ob eine Person blaue Augen hat oder nicht.
Hat sie blaue Augen, bekommt sie eine 1, sonst eine 0.
(Wir sagen „messen" und meinen „zählen". Dies wird erreicht mit einer 0/1 Gewichtung.)

	Grösse	blaue Augen
Hans	180	0
Nena	160	1
Jakob	175	0
Louise	155	0
Louis	182	0
Lisa	169	0
Lola	173	1
Dario	173	1
Fabio	190	0
Furs	154	0

Mit nebenstehender Tabelle ergäbe sich:

- Mittelwert $\bar{x} = \frac{180+160+\ldots+154}{10} = 171$ cm
- Anteilswert $h = \frac{1+0+\ldots+1}{10} = 0.3$

Sie sehen: Mittelwert und Anteilswert verwenden nur einen anderen „Massstab"!



Nebenan ist die zugehörige Verteilung abgebildet.
Sie ist gewöhnungsbedürftig, weil nur die Werte 0 und 1 vorkommen.
Sie soll aber verdeutlichen, wie eng Anteilswert und Mittelwert verwandt sind.

Konfidenzintervall für den Anteilswert

Nun hindert uns nichts mehr daran, ein Konfidenzintervall auf genau dieselbe Art und Weise wie beim Mittelwert anzugeben ☺.

Für den *wahren* Anteilswert verwenden wir den (lateinischen!) Buchstaben p.

p liegt mit ... % Wahrscheinlichkeit im Intervall $[h \pm z \cdot \sigma_{SV}] = [h \pm z \cdot \frac{\sigma_{GG}}{\sqrt{n}}]$

b) *Berechnen und interpretieren* Sie ein 95%-Konfidenzintervall zu den Daten aus Beispiel 9a).

Hinweis zu σ_{GG}

Es ist: $\sigma_{GG} = \sqrt{p \cdot (1-p)}$ (Nachrechnen!). Nur ist p eben unbekannt.

Dieses Problem – nämlich, dass wir σ_{GG} schätzen müssen – ist uns hingegen bekannt vom „normalen" Mittelwert.

Dazu ersetzen wir p durch den Anteilswert h aus der Stichprobe und bekommen: $\sigma_{GG} = \sqrt{p \cdot (1-p)} \approx \sqrt{h \cdot (1-h)}$.

Beispiel 11 **Gamesucht**

** Es ist nicht Aufgabe der Mathematik zu definieren, was „Gamesucht" bedeutet.*

Wie viele Jugendliche sind „gamesüchtig*"?
Oder: wie gross ist der („wahre") Anteilswert p an gamesüchtigen Jugendlichen?
Es wird eine Studie mit 100 Jugendlichen durchgeführt.
Man kommt zum Schluss: 28 dieser Jugendlichen sind „gamesüchtig".
Berechnen und interpretieren Sie möglichst genau in Worten zu dieser Stichprobe

- den Punktschätzer.
- das 90%-Konfidenzintervall. Man sagt auch: ein Konfidenzintervall mit dem **Konfidenzniveau** 90%.

Bemerkung

Die Formel für das Konfidenzintervall lässt sich auch mit Hilfe der Binomialverteilung herleiten (vgl. dazu Anhang 1).

Notwendiger Stichprobenumfang

Die Schätzung

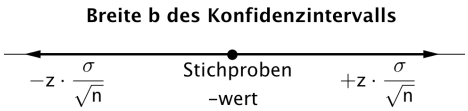
„Der Anteil gamesüchtiger Jugendlicher liegt zwischen 19% und 37%."

ist nicht sehr präzis. Die „Breite" des Konfidenzintervalls ist viel zu gross und macht die Aussage nahezu unbrauchbar. Wir werden uns deshalb noch mit der **Breite des Konfidenzintervalls** befassen.

Beispiel 12 **Breite des Konfidenzintervalls – notweniger Stichprobenumfang**

Offensichtlich gilt für die Breite b eines Konfidenzintervalls:

$$b = 2 \cdot z \cdot \frac{\sigma_{GG}}{\sqrt{n}}$$



a) Die Breite b hängt also von drei Grössen ab. Von welchen?

b) Natürlich möchte man ein möglichst enges Konfidenzintervall (warum?).
Offensichtlich lässt sich die Grösse der Streuung nicht beeinflussen, sie ist eine Eigenschaft der Grundgesamtheit.
Ebenso ist der z-Wert (das Konfidenzniveau) vorgegeben, etwa durch den Auftraggeber der Untersuchung.
Somit ist der Umfang der Stichprobe die einzige Möglichkeit, etwas an der Breite und damit an der Präzision der Schätzung zu ändern.

Soll etwa das Konfidenzintervall für einen *Anteilswert* (höchstens) die Breite b haben, dann muss der dazu *notwendige Stichprobenumfang* n mindestens

- $n > 4z^2 \cdot \frac{h(1-h)}{b^2}$ bzw. im Extremfall
- $n > \frac{z^2}{b^2}$ betragen.

Hinweis *Für welchen Wert h mit $0 \le h \le 1$ ist das Produkt $h(1 - h)$ am grössten?*

c) Wie gross muss der Stichprobenumfang n im Beispiel mit den gamesüchtigen Jugendlichen sein, wenn die Breite b nur 0.04 bzw. 4% betragen soll? (Man spricht dann auch von einer *Schätzung auf ±2 Prozentpunkte*.)

3 Testen

Sie lesen:

18-Jährige sind im Schnitt 175 cm gross.

Glauben Sie immer, was Sie lesen?

Hier liegt gewissermassen die „gegenteilige“ Situation vor wie bis anhin.
Der Wert der GG muss nicht mit Hilfe einer Stichprobe geschätzt werden. Er wird *als gegeben* vorausgesetzt!
Das heisst: wir können *nicht schätzen*, *aber testen* – nämlich, ob dieser Wert tatsächlich stimmt!

Dazu erheben Statistiker*innen sofort eine Stichprobe ☺ und testen, ob der behauptete Wert „vereinbar“ ist mit den Daten in der Stichprobe.

Schätzen oder Testen?

Beispiel 13 ein erster Test

Testen Sie die Hypothese

18-Jährige sind im Schnitt 175 cm gross.

Hinweis *Beispiel 5*
Eine Möglichkeit wäre zu überprüfen, ob der behauptete Wert in unserem Konfidenzintervall liegt. Wir können aber auch anders herum vorgehen: Wir nehmen an, dass der behauptete Wert stimmt und testen, ob der Stichprobenmittelwert nur wenig oder sehr viel von diesem Wert abweicht.

Immer, wenn wir mit Stichproben zu tun haben, läuft alles über die *Stichprobenverteilung*.

Beachte

Schätzen und Testen sind eng miteinander verwandt – zum Verwechseln ähnlich.
Sie haben sich aber als eigenständige statistische Verfahren etabliert und bieten beide die Möglichkeit, *von der Stichprobe auf die Grundgesamtheit zu schliessen*. Wir halten dazu fest:

- **Schätzen** *Welche Werte der GG* sind vereinbar mit den Stichprobendaten?
- **Testen** Sind die Stichprobendaten mit *einem bestimmten Wert der GG* vereinbar?

Beim Testen wird eine Hypothese überprüft und aufgrund der Stichprobe entschieden, ob man die Hypothese beibehält oder verwirft. Es kann zu Fehlentscheidungen kommen.
Beim Schätzen kann man nicht in dem Sinne „Fehler“ machen – es wird nur geschätzt, nicht getestet und entschieden.

Beispiel 14 Akku Laufzeit

Es soll getestet werden, ob die mittlere Laufzeit von Handy-Akkus möglicherweise von den vom Hersteller angegeben 18 Stunden abweicht. Wie könnte man dies tun?

Hinweis *Was für Fehlentscheidungen (Fehlbeurteilungen) könnte man treffen?*

Grundbegriffe der Testtheorie

Das Kennen des unbekannten Wertes der GG ist stets das Ziel.
Der Weg dorthin ist hindernisreich – es gibt stets wieder neue Hypothesen und Begründungen.

Eingebürgert hat sich folgende Namensgebung:

- **Nullhypothese H_0**:
  ist die momentan behauptete bzw. eine vorgegebene Annahme über den unbekannten Wert.

- **Alternativhypothese H_1**:
  ist die Alternative – das Gegenteil – zur Nullhypothese. Sie ist eine „neue" Vermutung.
  Sie sagt, dass die Nullhypothese *nicht (mehr) zutrifft*.

Das Ziel ist, die Alternativhypothese zu begründen, indem man zeigt, dass die Nullhypothese falsch ist.
Wir werden bald sehen, was damit gemeint ist.

Falsifikationsprinzip (Karl Popper)
Eine Person geht öfter an Gewässern spazieren und beobachtet Schwäne. Diese sind alle weiss, daher schliesst die Person vom Besonderen (ihrer Beobachtung) auf die allgemeine Hypothese: Alle Schwäne sind weiss. Die Frage hierbei ist, ab wie vielen Beobachtungen von weissen Schwänen die Person mit Sicherheit sagen kann, dass alle Schwäne weiss sind. Hierauf kann man keine Antwort geben, denn jeder weitere weisse Schwan bringt nur wenig Erkenntnisgewinn. Jedoch reicht nur ein einziger schwarzer Schwan, um die Theorie zu widerlegen.
Anstatt nach weiteren weissen Schwänen zu suchen, ist es daher sinnvoller, nach schwarzen (oder andersfarbigen) Schwänen zu suchen. Bleibt dies erfolglos, so kann daraus geschlossen werden, dass wohl alle Schwäne weiss sind – dies ist das Falsifikationsprinzip.
Daher werden wir im Folgenden bei allen Arten von Hypothesen-Tests immer das **Ziel verfolgen die Nullhypothese** (also das Gegenteil von dem, was wir eigentlich als These aufstellen) **zu widerlegen**. Sollte uns das gelingen, nehmen wir an, dass unsere ursprüngliche These (Alternativhypothese) wahr ist. Dieses „um die Ecke denken" ist nicht ganz leicht.

Beispiel 15 Hypothesen, Testniveau, Verwerfungsbereich, Entscheidung, Fehler

Die Firma „RIGHT TO BEAUTY" behauptet, mit ihrer neuen Diät wäre ein Gewichtsverlust von 5 kg garantiert.
Die Konsumentenschutzorganisation bezweifelt dies und will diese Angabe testen.

a) Wie lauten die *Hypothesen*?

b) Die Konsumentenorganisation gibt sich ein *Testniveau* von 5% vor.
Eine Umfrage unter 50 Proband*innen ergab einen Mittelwert von $\bar{x} = 3.7$ kg und $s = 3$ kg.

- Wie lautet der *Verwerfungsbereich*?
- Welche *Entscheidung* wird die Konsumentenorganisation treffen?

Hinweis *Skizze machen!*

c) Welche *Fehler* kann die Organisation bei ihrer Entscheidung begehen?

- Was bedeutet ein *Fehler 1.Art* und wie gross ist die *Wahrscheinlichkeit* α ihn zu begehen?
- Was bedeutet ein *Fehler 2.Art* und wie gross ist die *Wahrscheinlichkeit* β ihn zu begehen?

Hinweis *Skizze machen!*

d) Annahme: der mittlere Gewichtsverlust beträgt 4.8 kg. Berechnen Sie β.

e*) Wie gross kann die Wahrscheinlichkeit für einen Fehler 2.Art maximal werden?

Testniveau erinnert an Konfidenzniveau...
Testen ist wie Schätzen – einfach umgekehrt!

Salopp formuliert:
Testen ist wie Schätzen einfach ohne Umkehrtrick. Der unbekannte Wert wird durch die Nullhypothese postuliert!

Testen von Mittelwerten und Anteilswerten


Ich werf' 300-mal...

Beispiel 16 **Würfel**

Bei einem Spiel wird der Verdacht geäussert, dass der Würfel nicht in Ordnung ist.
Um dies zu testen, soll der Würfel 300-mal geworfen und dabei *gezählt* werden,
wie oft die „6" auftritt.

Hinweis *Veranschaulichen Sie, was Sie rechnen. Skizze!*

a) Was möchte man zeigen? Wie lauten demnach die Nullhypothese und die Alternativhypothese?

b) Wählen Sie ein Testniveau. Wie lautet der Verwerfungsbereich?

c) Der Test wird durchgeführt: es wird 300-mal gewürfelt. Wie entscheiden Sie sich, falls

- 60-mal die „6" fällt? Was sagt dieses Resultat aus?
- 70-mal die „6" fällt? Was sagt dieses Resultat aus?

d) Bei einem Entscheid sind Fehler möglich. Was bedeutet

- ein Fehler 1.Art und wie gross ist die Wahrscheinlichkeit α ihn zu begehen?
- ein Fehler 2.Art und wie gross ist die Wahrscheinlichkeit β ihn zu begehen?

e) Wie gross ist die Wahrscheinlichkeit β ihn zu begehen, wenn der Würfel tatsächlich gefälscht ist, mit

- $p(6) = \frac{1}{4}$
- $p(6) = \frac{1}{5}$ Für welches p(6) wird die Wahrscheinlichkeit für einen Fehler 2.Art maximal?

Beachte

- **Wahl der Nullhypothese**

  Als Nullhypothese ist immer die Aussage zu wählen, der eine Wahrscheinlichkeit zugeordnet werden kann. Nur bei *bekannter* Verteilung ist es nämlich möglich zu sagen, ob etwas (un–)wahrscheinlich ist und nur dann ist eine begründete Aussage möglich.
  Dies führt oft dazu, dass die Nullhypothese das *Gegenteil* dessen ist, was man vermutet.

- **Interpretation des Tests (Was sagt der Test aus und was sagt er nicht aus.)**

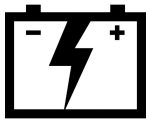
  Wenn man die Nullhypothese *verwirft*, sagt das *viel* aus über deren Ungültigkeit.
  Die Fehlerwahrscheinlichkeit α = p(Fehler 1.Art), die man bei einem solchen Entscheid machen könnte, kann nämlich eine vorgegebene Schranke – z.B. 5% – nicht überschreiten.

  Wenn man jedoch die Nullhypothese *nicht verwirft* – nicht verwerfen kann! – , so sagt das *wenig* über deren Gültigkeit aus, weil: die Fehlerwahrscheinlichkeit β = p(Fehler 2.Art), die man bei einem solchen Entscheid machen könnte, lässt sich nicht kontrollieren und kann sehr gross sein*.

* Der Wert von β hängt ab vom „wahren" Wert von p.

Beispiel 17 Testen

a) eines Mittelwerts



Es soll getestet werden, ob die mittlere Laufzeit von Notebook-Akkus möglicherweise von den vom Hersteller angegeben 4.5 Stunden abweicht.

- Bei einer Stichprobe vom Umfang n = 40 ergab sich ein Stichprobenmittelwert von $\bar{x}$ = 4.38 Stunden und eine Stichprobenstandardabweichung von s = 0.31 Stunden.
  Geben Sie dazu die Hypothesen an. Bestimmen Sie den Verwerfungsbereich zum Testniveau 5%.
  Sind an der Herstellerangabe Zweifel angebracht?
- Was bedeutet bei diesem Test ein Fehler 2.Art?
  Lässt er sich berechnen? Wenn ja, berechnen Sie ihn – wenn nein, was müsste man dazu wissen?

b) eines Anteilwertes

Ein Medikament hilft in 70% aller Fälle. Aufgrund einer neuen chemischen Zusammensetzung des Medikamentes könnte sich die Wirksamkeit verändert haben. Es wird an 100 Patienten getestet.

- Geben Sie die Hypothesen an. Wählen Sie ein Testniveau und bestimmen Sie den Verwerfungsbereich.
- Der Test wird durchgeführt. 58 Patienten werden gesund. Wie ist zu entscheiden?
- Welche Fehler bei der Beurteilung sind möglich und wie gross sind deren Wahrscheinlichkeiten?

c) vermutete Eigenschaft

Eine Biologin vermutet, dass neugeborene Küken schon Körner erkennen können und dies nicht erst durch Erfahrung lernen müssen. Sie möchte ihre Vermutung wissenschaftlich absichern. Sie legt dazu einem Küken „Körner“ aus Papier vor, je zur Hälfte Kreise und Dreiecke und will das Küken 50-mal picken lassen. Eine hohe Anzahl gepickter Kreise spräche für ihre Vermutung, eine irrtümliche Folgerung ist allerdings auch nicht ausgeschlossen.

- Geben Sie die Hypothesen an, wählen Sie ein Testniveau, bestimmen Sie den Verwerfungsbereich.
- Was bedeutet hier ein Fehler 2.Art? Wie gross wäre die zugehörige Wahrscheinlichkeit β, wenn die Küken bereits eine angeborene Fähigkeit hätten und eigentlich in 60% nach den Kreisen pickten?

Beispiel 18 Begründung!

Richtig oder falsch? Kreuzen Sie die richtigen Aussagen an.

- ❒ Um die Alternativhypothese „statisch zu beweisen“ zeigt man, dass die Nullhypothese wegen eines Testergebnisses sehr unwahrscheinlich ist.
- ❒ Das Testniveau gibt die Wahrscheinlichkeit an, dass die Nullhypothese falsch ist.
- ❒ Mit dem Test findet man heraus, ob die Nullhypothese stimmt.
- ❒ Die Wahrscheinlichkeit α für einen Fehler 1.Art gibt an, wie wahrscheinlich es ist, die Nullhypothese beizubehalten, obwohl sie falsch ist.
- ❒ Beim Testen handelt es sich um ein *indirektes* Vorgehen:
  Das Beste, was ich von einem Test erwarten kann, ist *nicht die Bestätigung meiner Hypothese*, *sondern bestenfalls das Verwerfen der Nullhypothese*.
- ❒ Die Fehlerwahrscheinlichkeiten addieren sich zu 1. Es gilt stets: $\alpha + \beta = 1$.

4 Zusammenfassung

Eine **statistische Schätzung** dient dazu, mit Hilfe einer Stichprobe eine Aussage über einen *unbekannten* Wert der Grundgesamtheit zu treffen.

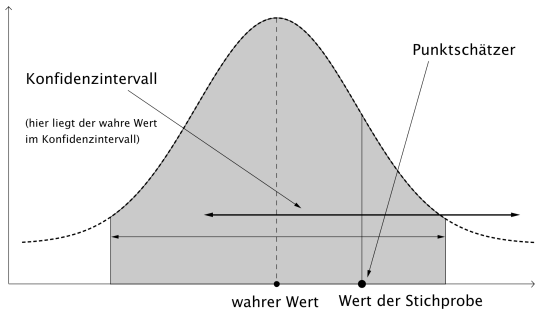
Es handelt sich um ein *direktes* Vorgehen:
Aus den Stichprobendaten wird ein **Punktschätzer** berechnet und um diesen ein Bereich (**Konfidenzintervall**) bestimmt, in dem sich der Wert mit einer bestimmten Wahrscheinlichkeit befindet.

Ein **Konfidenzintervall** wird berechnet durch die Vorgabe eines **Konfidenzniveaus**.

Die **Formeln** lauten:

- **Mittelwert:** $\bar{x} \pm z \cdot \frac{s}{\sqrt{n}}$
  ($\bar{x}$ = Punktschätzer der Stichprobe,
  s = aus Stichprobe geschätzte Standardabweichung)

- **Anteilswert:** $h \pm z \cdot \frac{\sqrt{h(1-h)}}{\sqrt{n}}$
  (h = Punktschätzer der Stichprobe)



Die Begründung dieser Formeln liefern die **Stichprobenverteilung** und der **Umkehrtrick**.
Die Ähnlichkeit der Formeln ist kein Zufall. Anteilswerte sind „spezielle Mittelwerte“.

Ein **statistischer Test** dient der Entscheidung, ob man aufgrund einer Stichprobe eine bestimmte *Annahme über einen Wert* (**Nullhypothese**) der Grundgesamtheit *widerlegen* kann oder nicht.

Es handelt sich um ein *indirektes* Vorgehen:
Die **eigentliche Vermutung (die Alternativhypothese)** wird bestätigt, indem man zeigt, dass ihr Gegenteil (die Nullhypothese) nicht mehr zu halten ist – sprich: sehr unwahrscheinlich ist.

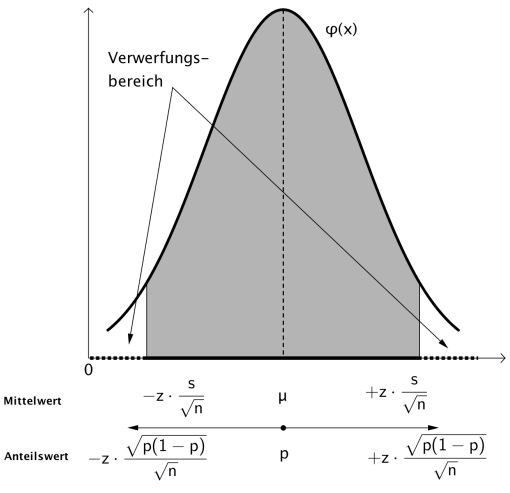
Die Nullhypothese muss dabei statistisch ausgewertet werden können: nur bei *bekannter* Verteilung ist es möglich zu sagen, wie (un–)wahrscheinlich das Testresultat ist und damit ein Entscheid begründbar.

Der **Verwerfungsbereich** wird berechnet aus der Vorgabe eines **Testniveaus**.

Die **Formeln** lauten:

- **Mittelwert:** $\mu \pm z \cdot \frac{s}{\sqrt{n}}$
  (s = aus Stichprobe geschätzte Standardabweichung)

- **Anteilswert:** $p \pm z \cdot \frac{\sqrt{p(1-p)}}{\sqrt{n}}$
  (Formel mit relativer Häufigkeit)



Liegt ein Stichprobenresultat (nicht) im Verwerfungsbereich, dann wird die Nullhypothese (nicht) verworfen. Beim **Entscheid** können Fehler passieren. Man unterscheidet **Fehler 1. und 2.Art**.

Testen und Schätzen sind eng miteinander verwandt.

Beide Verfahren haben zum Ziel, *von der Stichprobe auf die Grundgesamtheit zu schliessen*, um dann eine Aussage über die Grundgesamtheit zu treffen, deren Güte durch eine Wahrscheinlichkeit quantifiziert wird.

Allerdings gelangen sie auf verschiedenen Wegen zum Ziel. Wir merken uns:

„Wenn ich herausfinden will...

- ... welcher Parameter am besten zu meinen Beobachtungen der Stichprobe passt, dann berechne ich einen **Punktschätzer**."
- ... welche Parameterwerte am besten mit meinen Beobachtungen der Stichprobe vereinbar sind, dann berechne ich ein **Konfidenzintervall**."
- ... ob meine Beobachtungen der Stichprobe mit einem bestimmten Wert vereinbar sind, dann führe ich einen **Test** durch."

Beispiel Basketball

Jemand wirft 50-mal einen Freiwurf und trifft dabei 38-mal...

Schätzen Sie?

Oder testen Sie? Und wenn ja: was?

Das Testen hat sich zwar als statistische Vorgehensweise etabliert, gerät jedoch zunehmend in die Kritik, insbesondere in Bezug auf die ***Relevanz*** (vgl. Anhang 4) einer Aussage bzw. eines Entscheides.

Ein neueres statistisches Verfahren ist das sogenannte **Bootstrapping**.
Es bietet eine Möglichkeit, mit schierer Rechenleistung, dafür fast ganz ohne Mathematik, Konfidenzintervalle zu bestimmen. Dabei werden aus *einer* bestehenden Stichprobe *sehr viele* neue Stichproben mit Hilfe eines Computers „erzeugt" – im Wesentlichen durch Ziehen mit Zurücklegen.
Mit diesen vielen „neuen" Stichproben wird dann eine „echte" Stichprobenverteilung generiert.

Anhang 1 mathematische Ab-und Hintergründe

Zentraler Grenzwertsatz

... ist ein mathematisches Resultat und zentral für die gesamte Statistik. Vereinfacht lautet er wie folgt:

Eine Grundgesamtheit mit Erwartungswert μ und Standardabweichung σ sei gegeben. Wir definieren die Zufallsgrössen X_i ($1 \le i \le n$), wie folgt

X_i = i-te Messung.

Jede einzelne Zufallsvariable X_i hat also den Erwartungswert μ und die Standardabweichung σ.

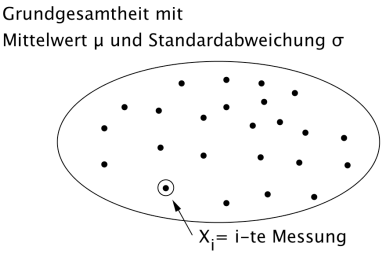
Sind die Zufallsgrössen unabhängig voneinander (beeinflussen sich also nicht gegenseitig), dann ist die ***neu gebildete* Zufallsgrösse**

$$\bar{X} = \frac{X_1 + X_2 + \ldots + X_n}{n}$$

approximativ **normalverteilt** mit

$$\mu_{\bar{x}} = \mu \quad ; \quad \sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$$

sofern n genügend gross ist.
Dieses Resultat gilt unabhängig von der zugrundeliegenden Verteilung!



Der zentrale Grenzwertsatz liefert auch die Begründung, warum wir – bei genügend grossem n – die Binomialverteilung durch die Normalverteilung approximieren dürfen. Die einzelnen X_i sind dabei 0/1-binomialverteilt.

Stichprobenverteilung für einen Anteilswert – „relative" Binomialverteilung

Die Binomialverteilung „zählt" die Anzahl Erfolge.
Wir rechnen also mit *absoluten* Häufigkeiten.

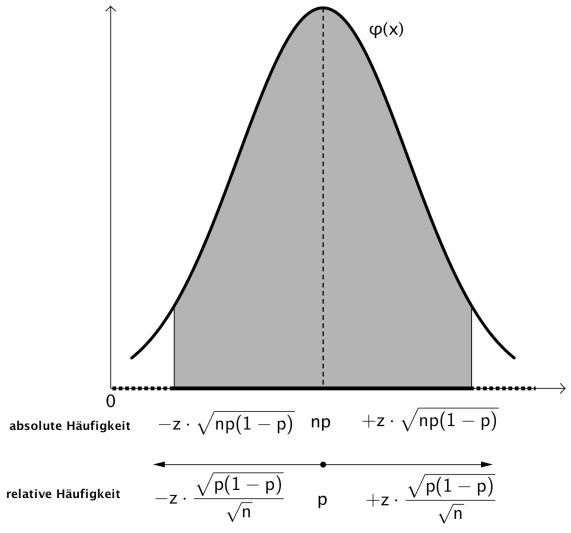
Ein Anteilswert ist immer eine Zahl zwischen 0 und 1.
Wir rechnen also mit *relativen* Häufigkeiten.

Das ist kein Problem: wir dividieren einfach durch n.

Et voilà.

Die Binomialverteilung *ist* eben gerade die Stichprobenverteilung des Anteils bzw. umgekehrt.

Darüber darf man nachdenken.
Dies ist eine neue Blickrichtung ist auf die Binomialverteilung.



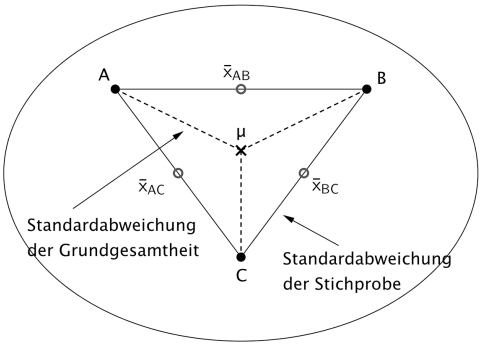
Korrigierte Standardabweichung

Man hat nur *eine* Stichprobe mit n Werten und will damit die Standardabweichung in der Grundgesamtheit schätzen. Man könnte nun einfach die Standardabweichung in der Stichprobe berechnen und – wie beim Mittelwert! – hoffen, dass man einen guten „Punktschätzer" für die Grundgesamtheit hat. Das ist aber nicht so!

Es zeigt sich, dass die „normale Formel"

$$\sigma = \sqrt{\frac{1}{n}\sum_{i=1}^{k}(x_i - \bar{x})^2 \cdot H(x_i)}$$

die „wahre" Standardabweichung *systematisch* unterschätzt.



Dies passiert deshalb, weil beim Berechnen „der Abweichungen vom Mittelwert" bereits der Mittelwert nur ein geschätzter ist (vgl. Abbildung).
Wir müssen sie also grösser machen!

Das machen wir, indem wir die quadratischen Abweichungen nicht durch n teilen, sondern bloss durch n – 1:

$$s = \sqrt{\frac{1}{n-1}\sum_{i=1}^{k}(x_i - \bar{x})^2 \cdot H(x_i)}$$

$\sigma \neq s$

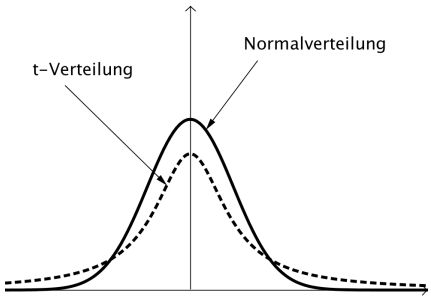
Es lässt sich mathematisch zeigen, dass die so berechnete, sogenannte „korrigierte" Standardabweichung eine bessere Schätzung der „wahren" Standardabweichung ergibt.

t-Verteilung

Die t-Verteilung ist die „breite" Schwester der Normalverteilung.

Warum braucht es die?

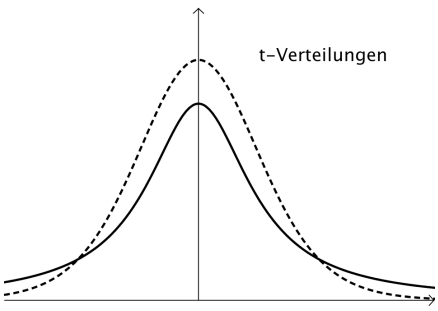
Der zentrale Grenzwertsatz sagt, dass die Stichprobenverteilung angenähert normalverteilt mit einer Standardabweichung, welche abhängt von der „wahren" Standardabweichung σ der Grundgesamtheit!
Diese „wahre" Standardabweichung σ ist aber (fast) immer unbekannt und muss aus der Stichprobe heraus geschätzt werden.



Diese Schätzung ist unsicher, wenn der Stichprobenumfang n klein ist.
Die t-Verteilung berücksichtigt dies und ist deshalb „breiter" als die Normalverteilung.
Die üblichen „kritischen" Werte der t-Verteilung sind deshalb (abhängig von n!) ein bisschen grösser als die, welche wir jeweils für die Normalverteilung brauchen.

Beispiel
Zu 95% gehört der (normalverteilte) „z-Wert" 1.96 und der (t-verteilte) „t-Wert" 2.09 (für n = 20).

Zu jedem n gibt es eine t-Verteilung.
Für grosses n nähert sich die t-Verteilung jedoch immer mehr der Normalverteilung an. Deshalb darf man bei genügend grossem n auch die Normalverteilung anstelle der t-Verteilung ansetzen.
Üblicherweise tut man dies, falls **n ≥ 30**.



Abgebildet sind zwei t-Verteilungen:
Eine zum Wert n = 10, eine zum Wert n = 50.
Welche ist welche?

Beim „Testen von Mittelwerten" ist häufig von einem **t-Test** die Rede.
Sie sollten nun ein Gefühl haben, was damit gemeint ist.

Anhang 2 Formeln der Stichprobenverteilung – sinnvoll? Sinnvoll!

Wir überlegen uns nochmals, dass die angegebenen „Parameter der Stichprobenverteilung“ in der *Art ihres Aufbaus sinnvoll* sind.

Einleuchten sollte dann, dass

- für den Erwartungswert gilt: $\mu_{SV} = \mu_{GG}$ (μ_{SV} ist der„Mittelwert aller Mittelwerte“)
- im Zähler der Standardabweichung σ_{sv} die Standardabweichung σ_{GG} der Grundgesamtheit steht und im Nenner der Stichprobenumfang n auftritt. („Überprüfe“ mit n = 1 und n $\rightarrow +\infty$.)

Überlegung 1, betrifft den Mittelwert der Stichprobenverteilung

Stellen Sie sich vor, Sie ziehen zufällig 30 Werte aus der Grundgesamtheit. Vermutlich würden Sie einige Werte oberhalb und andere Werte unterhalb des „wahren“ Mittelwertes μ_{GG} ziehen; es könnte auch sein, dass Sie gerade die 30 kleinsten oder 30 grössten Werte erwischen würden, aber das ist eher unwahrscheinlich (aber nicht ausgeschlossen).
Der Mittelwert dieser 30 Werte wird irgendwo zwischen den kleinsten und grössten Werten der Originaldaten, wahrscheinlich aber eher in der Gegend des wirklichen Mittelwertes liegen. Das heisst: obwohl sie nichts über den wirklichen Mittelwert wissen, können Sie annehmen, dass der Mittelwert der 30 zufällig gezogenen Zahlen irgendwo in der Gegend des „wahren“ Mittelwertes μ_{GG} liegt.
Das Überraschende ist also: obwohl die wirkliche Verteilung „irgendwie“ aussehen kann, liegen die Mittelwerte $\bar{x}$ der zufällig gezogenen Werte „glockenförmig“ (sprich: normalverteilt) um den tatsächlichen Mittelwert μ_{GG} herum!

Überlegung 2, betrifft die Standardabweichung der Stichprobenverteilung

Die 1.Überlegung sagte uns: wenn wir sehr oft 30 solcher Werte ziehen, dann werden nur selten alle 30 Werte am unteren oder oberen Rand der (wirklichen) Verteilung liegen, und deshalb folgen die Mittelwerte $\bar{x}$ der zufällig „gezogenen“ Werte einer Normalverteilung um den tatsächlichen Mittelwert μ_{GG}.

Es hängt nun von *zwei* Faktoren ab, „wie dicht“ die Mittelwerte der zufällig gezogenen Zahlen beim wirklichen Mittelwert liegen:

Erster Faktor: wenn die Originalwerte stark streuen (d.h., eine große Standardabweichung haben), dann werden die zufällig gezogenen Daten auch stärker streuen und damit auch die aus ihnen berechneten Mittelwerte, als wenn die Originaldaten alle dicht beim wirklichen Mittelwert liegen (denn dann können die zufällig gezogenen Daten auch nicht weit abweichen).
Das heisst: ist σ_{GG} gross, dann ist auch σ_{SV} gross.

Der zweite Faktor ist die Anzahl der zufällig gezogenen Zahlen: wenn es sehr viele sind (z.B. 100 oder 1000), dann wird deren Mittelwert dichter am „wahren“ Mittelwert der Originaldaten liegen, weil es unwahrscheinlicher ist, dass alle 100 (oder 1000) Werte unterhalb oder oberhalb des wirklichen Mittelwertes liegen. Wenn man dagegen nur zufällig drei Werte zieht, dann wird es häufiger vorkommen, dass der Mittelwert der gezogenen Zahlen weiter vom wirklichen Mittelwert abweicht.
Das heisst: je grösser n, umso kleiner σ_{SV}.

Merke
Die Mittelwerte schwanken zwar zufällig (wir ziehen ja eine Stichprobe nach dem Zufallsprinzip), aber sie tun dies immerhin um den wahren Mittelwert herum.
Weiter schwanken sie in berechenbarer Weise und umso weniger, je grösser die Stichprobenumfang ist.

Anhang 3 Testen und Gericht – ein Vergleich

Hypothesentests führt man immer dann durch, wenn man irgendetwas mit Hilfe von erhobenen Daten nachweisen möchte, zum Beispiel, dass es mehr Knabengeburten gibt als Mädchengeburten.
Der Grundsatz bei allen statistischen Tests ist hierbei, dass wir das Gegenteil widerlegen müssen – wir müssen also widerlegen, dass es gleichviele Knabengeburten wie Mädchengeburten gibt.

Um den Grund für dieses Vorgehen zu verstehen, kann man sich eine *Gerichtsverhandlung* vorstellen, und Parallelen zum Ablauf eines Tests ziehen:
Man geht davon aus, dass die Angeklagte unschuldig ist (ohne es genau zu wissen). Bevor man von der Schuld der Angeklagten überzeugt sein kann, muss man ausreichend Beweise gesammelt haben, um zweifelsfrei an die Schuldigkeit glauben zu können. Falls das nicht der Fall ist, muss man davon ausgehen, dass sie unschuldig ist. Man könnte diesen Sachverhalt auch in statistischen Hypothesen formulieren:

- Die Angeklagte ist unschuldig
- Die Angeklagte ist schuldig

Nur: welche der beiden Hypothesen ist nun die Nullhypothese H_0?

Wir wissen: es ist nicht das Gleiche, H_0 beizubehalten und H_0 zu beweisen.

Also: wenn ich beweisen möchte, dass der Angeklagte schuldig ist, *muss* ich meine Hypothesen so herum formulieren:

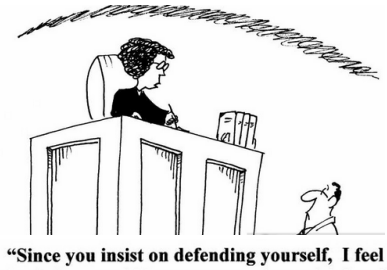
- H_0: Die Angeklagte ist unschuldig
- H_1: Die Angeklagte ist schuldig

Wenn ich nun „Daten erhebe", also in der Verhandlung Beweise gesammelt werden, dann tritt einer der folgenden zwei Fälle ein:

- Es gibt genug Beweise für die Schuld der Angeklagten. Dann kann ich H_0 ablehnen und habe H_1 **nachgewiesen**, d.h. die Angeklagte ist (ziemlich sicher) schuldig. Die Antwort in diesem Fall lautet: „H_1 ist wahr" (natürlich nur zu dem gewählten Testniveau).

- Man hat keine (oder nicht genug) Beweise für die Schuld der Angeklagten gefunden. Damit habe ich aber H_0 (also die Unschuld) **nicht bewiesen**!
  Nur weil keine Beweise für die Schuld gefunden wurden, können wir nicht sagen „wir haben bewiesen dass der Angeklagte unschuldig ist".
  Die Antwort in dieser Situation lautet stattdessen: **„Wir wissen es nicht"**.
  In einer statistischen Auswertung sagt man dann zum Beispiel:
  „Es konnten keine Hinweise auf die Gültigkeit der Alternativhypothese gefunden werden."

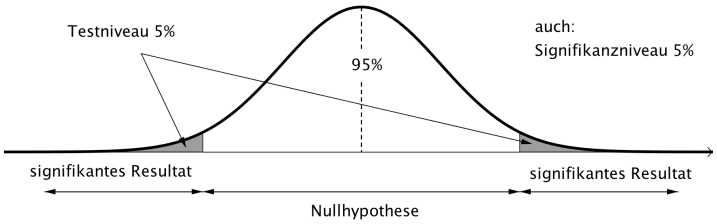
Man kann also H_0 nie (statistisch!) „untermauern", sondern nur H_1.
Aus diesem Grund ist es so wichtig, dass man die Hypothesen richtig herum formuliert:
Der Fall, den man nachweisen möchte, kommt in die Alternativhypothese H_1, das „Gegenteil" in die Nullhypothese H_0.
Die Metapher mit der Gerichtsverhandlung ist eine Eselsbrücke, um sich an dieses Vorgehen zu erinnern.



"Since you insist on defending yourself, I feel April first would be an appropriate trial date."

Anhang 4 Signifikanz vs Relevanz

Ein Begriff, der bei statistischen Aussagen immer auftaucht, ist derjenige der **Signifikanz**.



Liegt ein Stichprobenergebnis im Verwerfungsbereich, spricht man von einem *signifikanten* Ergebnis.
Es liegt also ein „Zeichen" vor, dass man die Nullhypothese verwerfen kann (lat. signum = Zeichen).

Folgende Sprachregelungen haben sich etabliert:

- liegt ein Stichprobenergebnis **im Verwerfungsbereich zum Testniveau 5%**, spricht man von einem **signifikanten** Ergebnis
- liegt ein Stichprobenergebnis **im Verwerfungsbereich zum Testniveau 1%**, spricht man von einem **hochsignifikanten** Ergebnis
- Statt von einem Testniveau spricht man dann von einem „**Signifikanzniveau**" von 5% (bzw. 1%)".
  Statt von „Testen einer Hypothese" spricht man deshalb auch von einem „**Signifikanztest"**.

Zeigen Sie an einem **Beispiel** den Unterschied zwischen signifikantem und hochsignifikantem Stichprobenergebnis.

Der Begriff Signifikanz ist aber nur ein Begriff. Also auch hier: **Vorsicht bei der Interpretation**!

Beispiel **signifikant!, aber...**

a) Bisher schneiden 60% aller SchülerInnen im Fach Mathematik „genügend" ab.
Eine ExpertInnengruppe bewirbt die Schulen und behauptet, mit ihrem Intensivprogramm sei statistisch nachgewiesen worden, dass sich dieser Anteil *signifikant* verbessert habe (auf dem 5%-Niveau).
Es wird die Frage diskutiert, ob diese zwar kostspielige, aber scheinbar erfolgreiche Methode in den Schulen Verbreitung finden soll...

SIE müssen entscheiden. Dazu erinnern Sie sich, dass die Stichprobengrösse n einen Einfluss hat und fragen nach. Die Gruppe antwortet

i) n = 100 **ii)** n = 10'000

Bei welcher Antwort würden Sie sich (eher) für die neuartige Methode entscheiden?

b) Erklären Sie folgende Aussage und erläutern Sie, welche Probleme dies mit sich bringt.

„Bei genügend grossem Stichprobenumfang n wird jede noch so kleine Abweichung von der Nullhypothese als signifikant ausgewiesen."

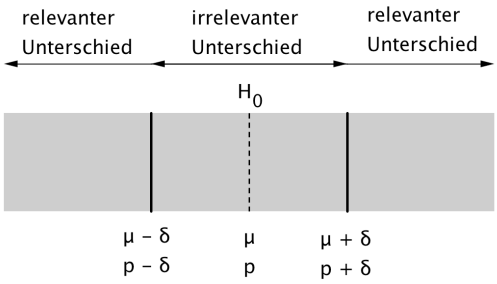
Beachte

Im normalen Sprachgebrauch wird das Wort *signifikant* oft als Synonym für „deutlich" gebraucht.
Eine *statistisch signifikante* Änderung muss allerdings *nicht notwendigerweise deutlich* sein, sondern nur *eindeutig*. Es kann sich also durchaus um eine geringfügige Änderung handeln, die eindeutig gemessen wurde.

Bei genügend hoher Anzahl an Messungen wird jeder (existierende) Unterschied als statistisch signifikant gemessen werden, so klein und unbedeutend er auch sein mag.

In der Praxis ist die wichtigere Frage oft: haben wir ein *relevantes* (und nicht nur ein signifikantes) Resultat?

Wir müssen aber wissen, was **Relevanz** bedeutet...



Was ein relevanter Unterschied ist, hängt ab vom jeweiligen Fachgebiet und ist nicht Aufgabe der Statistik.

Es muss eine „Toleranz" δ angegeben werden, ab der man sagt, dass ein Unterschied relevant ist für die entsprechende Anwendung.

In Bezug auf die Frage Signifikanz vs Relevanz zeigt sich ein Vorteil des Konfidenzintervalles. Konfidenzintervalle erleichtern die Interpretation.

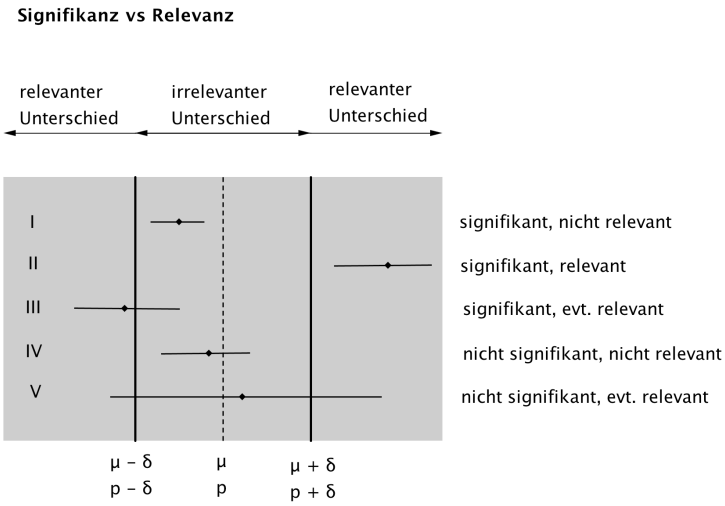
Basierend auf unseren Daten berechnen wir dann ein **Konfidenzintervall** für den Parameter von Interesse. Die Idee besteht nun darin, dass man schaut, wo das Konfidenzintervall bezüglich der obigen Bereiche liegt.

Liegt ein Konfidenzintervall ganz im „relevanten" Bereich, so spricht man von einem relevanten Effekt. Ist zwar der Test signifikant (d.h. das Konfidenzintervall enthält den Parameter nicht), aber das Konfidenzintervall liegt im „irrelevanten Bereich", so hat man zwar ein signifikantes, aber kein relevantes Resultat, siehe Abbildung:

Beispiel

H_0: $\mu = 80$; Toleranz $\delta = 5$

Erhalten wir das Konfidenzintervall [70;78], so müssten wir potenziell beunruhigt sein.



Merke

Ein Konfidenzintervall erleichtert die Interpretation eines Ergebnisses bezüglich der Relevanz.

Beispiel* **Teststärke**

Für die „Enden" des Toleranzbereiches könnte man die Wahrscheinlichkeit β für einen Fehler 2.Art berechnen, also p(Fehler 2.Art/p ± δ). Was würde das bringen?


YES, IT'S REMOTELY POSSIBLE THAT EARTH'S GRAVITY HAS INCREASED, BUT I THINK THERE'S A MUCH SIMPLER EXPLANATION.