Über den Zusammenhang von Stichprobengröße und Stichprobenfehler

Empirische Sozialforschung & Statistik

1. **Stichprobenfehler** – in welcher Hinsicht ‚Fehler'?

2. **Weshalb kommt es in Stichproben zu Stichprobenfehlern?**

3. Das $\sqrt{n}$**-Gesetz**

Stichprobenfehler – in welcher Hinsicht ‚Fehler'?

- **Kennwert aus Stichprobe** nicht exakt gleich dem entsprechenden **Parameter in der Population**
- Also Fehler hinsichtlich **Repräsentativität**
  - Repräsentativität: Verteilung von Variable X in Stichprobe = Verteilung X in Population ⇔ Kennwert = Parameter
- **Stichprobenfehler**…
  - …sind **zufällige Fehler**
  - …treten also **in Zufallsstichproben** auf
  - Der Zusammenhang von Stichprobengröße und Stichprobenfehler lässt sich lediglich für Zufallsstichproben rekonstruieren

Weshalb kommt es in Stichproben zu Stichprobenfehlern?

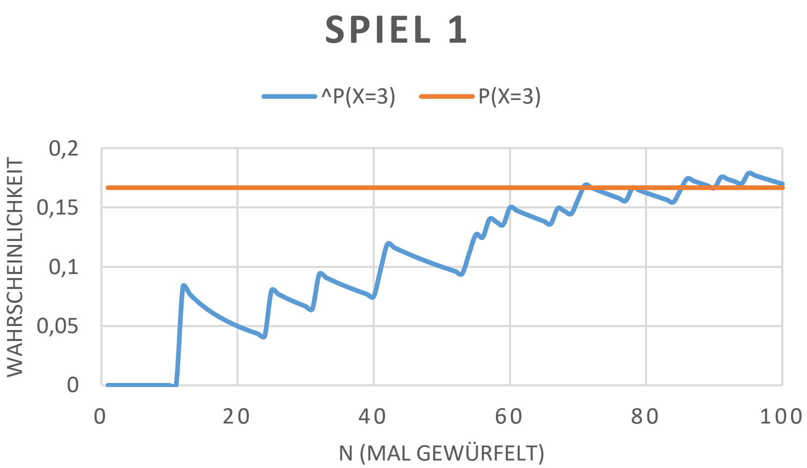
- **Sampling** (in Zufallsstichproben) entspricht **Realisierung eines Zufallsexperiments**
- Sampling-Wahrscheinlichkeit p_i eines jeden Elements i aus Population ist $0 < p < 1$
- Realisierung der Stichprobe ist also **kontingent**: die Stichprobe hätte sich auch aus anderen Elementen zusammensetzen können
- Die Frage hier also: wieso nähert sich mit zunehmendem n die Verteilung von X in der Stichprobe der Verteilung von X in Population an?
- Antwort: es gilt das **Gesetz der großen Zahl**
  - Sei A eine Teilmenge aller möglichen Ausprägungen von X, dann ist die (statistische) Wahrscheinlichkeit für $\mathrm{x} = A$

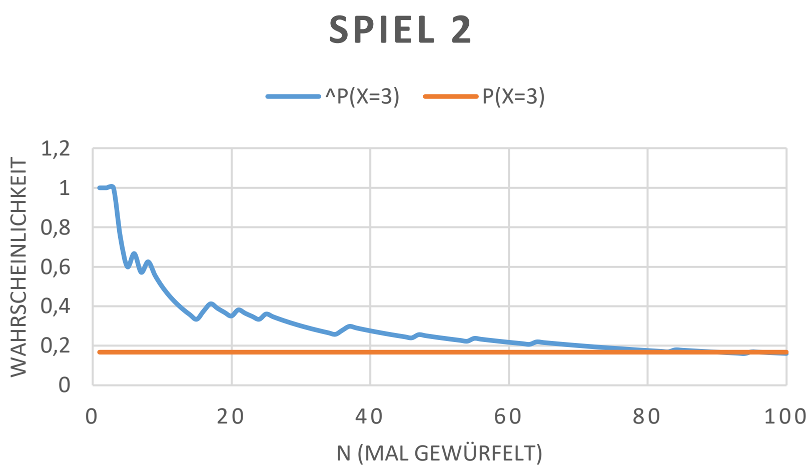
$$P(x = A) = \lim_{n \to \infty} p(x = A) = \lim_{n \to \infty} \frac{f(x = A)}{n}$$

Weshalb kommt es in Stichproben zu Stichprobenfehlern?

Gesetz der großen Zahl

- Würfelspiel
  - Würfeln und schauen, wie groß der Anteil von ‚Augenzahl 3' ist
    - Kennwert $\hat{P}(X=3)=\frac{f(x=3)}{n}$, Stichprobe n-mal würfeln
    - Parameter $P(X=3)=\frac{1}{6}$, Population ∞-mal würfeln (P theoretisch ermittelt: ‚Augenzahl 3' auf 1 von 6 Seiten)





Das $\sqrt{n}$-Gesetz

- Stichprobenfehler nicht in jeder Stichprobe der Größe n gleich
  - Siehe Würfelspiele: Abweichung $\hat{P}(X=3) - P(X=3)$ in Spiel 1 verschieden von entsprechender Abweichung in Spiel 2
  - Aber die **Wahrscheinlichkeitsverteilung des Stichprobenfehlers** ist gleich und **abhängig von $\boldsymbol{n}$**: je größer n, desto kleiner die Wahrscheinlichkeit großer Stichprobenfehler
- Stichprobenziehung als Realisierung eines Zufallsexperiments ➔ Stichprobenfehler als Zufallsvariable; **Standardfehler ($\boldsymbol{SE}$)** = Standardabweichung der Stichprobenfehler
- **Schätzgenauigkeit** und Stichprobengröße
  - Für die Parameterschätzung kann eine **Fehlerwahrscheinlichkeit $\boldsymbol{\alpha}$** vorgegeben werden
  - Ermittelt wird dann, wie groß der Stichprobenfehler mit der vorgegebenen Wahrscheinlichkeit maximal ist
  - $\sqrt{n}$-Gesetz präzisiert diesen Zusammenhang zwischen Schätzgenauigkeit und Stichprobengröße (zumindest für nicht zu kleine Stichproben)

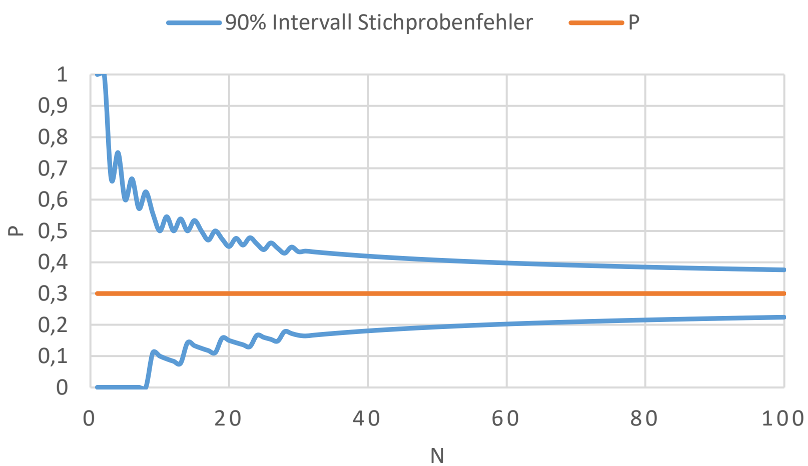
Das $\sqrt{n}$-Gesetz

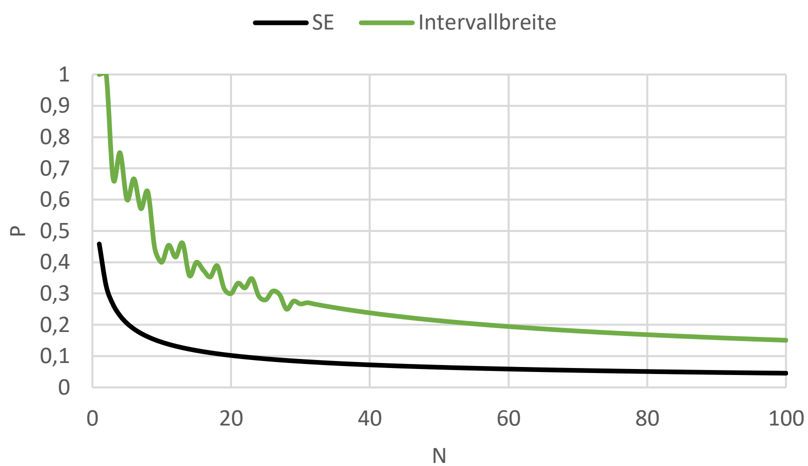
- Jedoch für kleine Stichproben das Prinzip der Schätzgenauigkeit und Zusammenhang von mit Stichprobengröße gut darstellbar
  - Wenn Parameter Anteil P der Ausprägung einer Variable ist,
  - dann ist die Binomialverteilung als Wahrscheinlichkeitsverteilung des Stichprobenfehlers adäquat
  - (und die Binomialverteilung geht unmittelbar aus der grundlegenden Wahrscheinlichkeitsrechnung hervor)
  - Für binomialverteilte Variable: $SE = \sqrt{\frac{P(1-P)}{n}} = \frac{\sqrt{P(1-P)}}{\sqrt{n}}$ *(für normalverteilte Zufallsvariablen: $SE = \frac{s}{\sqrt{n}}$)*
  - Wahrscheinlichkeitsverteilung $\hat{P}$ (und mithin Wahrscheinlichkeitsverteilung $\hat{P} - P$) also abhängig von SE und damit abhängig von n
  - Beispiel mit $P = 0.3$ und $\alpha = 0.1$:

n	Kennwert mit Wahrscheinlichkeit $1-\alpha$ im Intervall
10	$[0.1\,, 0.5]$
30	$[0.167\,, 0.433]$
60	$[0.2\,, 0.4]$
100	$[0.23\,, 0.38]$

Das $\sqrt{n}$-Gesetz

- Für größere Stichproben kann Binomialverteilung durch Normalverteilung approximiert werden
    - Beispiel Wahrscheinlichkeitsverteilung $\hat{P}$ mit $P = 0.3$ und $\alpha = 0.1$
    - für $n \leq 30$ Binomialverteilung, für $n > 30$ approximiert durch Normalverteilung





Das ‚Gezacke' im Bereich $n = [1\,, 30]$ kommt Zustande, weil die Binomialverteilung diskret ist

Bedeutung für die Praxis der statistischen Datenauswertung

- So ein Schwankungsintervall sagt uns ja: wenn wir theoretisch unendlich viele Stichproben der Größe n ziehen, aus einer Population mit Parameter P, dann wird der Kennwert $\hat{P}$ bei einem Anteil von $1-\alpha$ der Stichproben innerhalb des entsprechend ermittelten Intervalls liegen
- Je enger das Intervall ist, desto genauer ist die Schätzung
- Nun könnte man ja einfach sehr viele Stichproben mit gleichem Umfang aus der gleichen Population ziehen, um diese Wahrscheinlichkeit (=empirischer Anteil) – d.h. also: die Schätzgenauigkeit – zu ermitteln
- In der Praxis ziehen wir aber eben nur *eine* Stichprobe
- Wenn wir dann trotzdem etwas über die Schätzgenauigkeit sagen wollen, dann müssen wir uns eben an die hier diskutierten Aspekte der Wahrscheinlichkeitstheorie erinnern…
    - (…in der Praxis ist es dann doch noch ein klein wenig anders, weil wir i.d.R. den Wert des Parameters nicht kennen…aber die statistischen Grundlagen sind dieselben, die Anpassung an die Praxis geringfügig…)