Binomialverteilung

In Österreich haben rund $p = 41\,\%$ der Personen Blutgruppe A.
Es wird eine Zufallsstichprobe mit $n = 100$ Personen in Österreich durchgeführt.
Ermittle die Wahrscheinlichkeit, dass davon ...

1) genau 41 Personen die Blutgruppe A haben.

2) höchstens 35 Personen die Blutgruppe A haben.

3) mehr als 43 Personen die Blutgruppe A haben.

Anmerkung: Die Stichprobengröße ist im Vergleich zur Gesamtbevölkerung so klein, dass „Ziehen ohne Zurücklegen“ und „Ziehen mit Zurücklegen“ fast die gleichen Wahrscheinlichkeiten liefern.

X ... Anzahl Personen in Stichprobe mit Blutgruppe A
X ist (näherungsweise) binomialverteilt mit $n = 100$ und $p = 0{,}41$.

1) $P(X = 41) = 8{,}09...\,\%$ **2)** $P(X \leq 35) = 13{,}13...\,\%$ **3)** $P(X > 43) = P(X \geq 44) = 30{,}41...\,\%$

Wozu Konfidenzintervalle?

Es soll untersucht werden, welcher relative Anteil p einer Bevölkerung ein bestimmtes Merkmal hat.
Merkmale sind zum Beispiel „hat Blutgruppe A“, „wählt Partei X“, „kann Gitarre spielen“, „hat Antikörper gegen Virus Y“, ...

Es ist zu aufwändig, jede Person in der Bevölkerung zu befragen/untersuchen.

Deshalb wird nur eine repräsentative Stichprobe der Bevölkerung auf dieses Merkmal untersucht.
Das heißt: Es werden n Personen ausgewählt, die die Gesamtbevölkerung möglichst gut abbilden.

Wenn in dieser Stichprobe genau k Personen dieses Merkmal haben, dann ist der relative Anteil

$$h = \frac{k}{n}$$

in der Stichprobe die bestmögliche Schätzung für den relativen Anteil p in der Gesamtbevölkerung.

Wir können aber *nicht* damit rechnen, dass *exakt* $h = p$ gilt.
Je größer die Stichprobe ist, desto wahrscheinlicher ist es, dass h *nahe bei* p liegt.

Mit Konfidenzintervallen wollen wir *exakte* Aussagen treffen, was mit „nahe bei“ gemeint ist.

Konfidenzniveau & Interpretation

Die Sicherheit eines Konfidenzintervalls heißt auch **Konfidenzniveau** (**Vertrauensniveau**) $\boldsymbol{\gamma}$.
Typische Werte in der Praxis sind $\gamma = 95\,\%$, $\gamma = 99\,\%$ und $\gamma = 99{,}9\,\%$.

Gegeben ist eine Stichprobe der Größe n, die Anzahl der „Erfolge“ k und ein Konfidenzniveau γ.
Gesucht ist ein Intervall um $h = \frac{k}{n}$ mit folgender Eigenschaft:

Das Konfidenzintervall enthält den unbekannten relativen Anteil p mit der Wahrscheinlichkeit γ.

Je größer das Konfidenzniveau γ ist, desto verlässlicher soll das Intervall den Wert p enthalten und desto breiter wird also das Konfidenzintervall.

Die Wahrscheinlichkeit $\gamma = 95\,\%$ können wir zum Beispiel folgendermaßen interpretieren:

1) Wir nehmen 1000 Stichproben der Größe n.

2) Wir berechnen daraus die 1000 zugehörigen 95-%-Konfidenzintervalle.

3) Dann erwarten wir, dass rund 950 dieser Intervalle den unbekannten relativen Anteil p enthalten.

In der Praxis können wir natürlich nicht einfach so 1000 Stichproben erzeugen.
Diese frequentistische Deutung von Wahrscheinlichkeit hilft hier aber beim Interpretieren von Konfidenzintervallen.

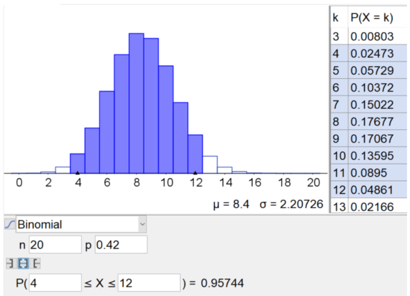
Wie berechnet man aber nun ein Konfidenzintervall für p in Abhängigkeit von n, k und γ?
Tatsächlich gibt es bei der Binomialverteilung verschiedene Methoden und Formeln zur Berechnung eines Konfidenzintervalls.
Mehr dazu erfährst du auf den nächsten Seiten.

Grenzen der Machbarkeit

X ist binomialverteilt mit den Parametern n und p. Der Parameter p ist unbekannt.
Die Zufallsvariable X kann nur die Werte $k = 0, 1, 2, \ldots, n$ annehmen.
Bei gegebenem Konfidenzniveau γ berechnen wir für jedes k mit einer Formel das Konfidenzintervall.

Die Formel ist noch geheim.

Zum Beispiel: $n = 20$, $p = 0{,}42$, $\gamma = 95\,\%$



Wir untersuchen welche der 21 möglichen Konfidenzintervalle den Parameter p enthalten:

- Wenn $k = 4, 5, \ldots, 12$ ist, dann enthält das Konfidenzintervall mit dieser (geheimen) Formel den Wert $p = 0{,}42$.
- Wenn $k = 0, 1, 2, 3, 13, 14, \ldots, 20$ ist, dann ist $p = 0{,}42$ mit dieser Formel *nicht* im Konfidenzintervall enthalten.
  Diese Werte für k sind zu „ungewöhnlich“ bei $n = 20$ und $p = 0{,}42$.

Für die Wahrscheinlichkeit, dass das Konfidenzintervall den Wert $p = 0{,}42$ enthält, gilt hier also:

$$P(4 \leq X \leq 12) = 95{,}74...\,\%$$

Wir können uns bei der Binomialverteilung noch so bemühen eine passende Formel zu finden.
Wir werden es nicht schaffen, dass die Wahrscheinlichkeit stets *exakt* $\gamma = 95\,\%$ ist.
Das ist anders als beim Konfindenzintervall für den Erwartungswert μ einer normalverteilten Zufallsvariable.

Konfidenzintervall für p (Formel I)

In der ehemaligen Formelsammlung (AHS) findest du die folgende Formel für dieses Konfidenzintervall:

Konfidenzintervall

h ... relative Häufigkeit in einer Stichprobe
p ... unbekannter relativer Anteil in der Grundgesamtheit
γ ... Konfidenzniveau (Vertrauensniveau)

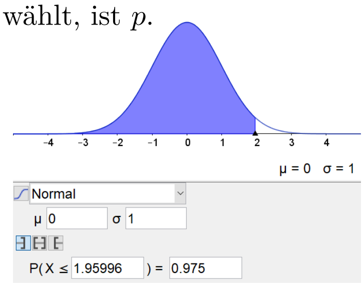
γ-Konfidenzintervall für p (diejenigen Werte p, in deren γ-Schätzbereich der Wert h liegt):

$$\left[h - z \cdot \sqrt{\frac{h \cdot (1-h)}{n}};\ h + z \cdot \sqrt{\frac{h \cdot (1-h)}{n}}\right], \text{ wobei für } z \text{ gilt: } \gamma = 2 \cdot \phi(z) - 1$$

In dieser Formel wird die Binomialverteilung durch die Normalverteilung angenähert.
Die Näherungslösung mit der Normalverteilung wird nochmal gerundet, um diese Formel zu erhalten.

Berechnung eines Konfidenzintervalls (Formel I)

In einer Stichprobe der Größe $n = 1000$ geben $k = 420$ Personen an, die Partei X zu wählen.
Der relative Anteil der Personen in der Gesamtbevölkerung, die Partei X wählt, ist p.
Berechne auf Basis dieser Stichprobe ein 95-%-Konfidenzintervall für p.



1) Relative Häufigkeit in der Stichprobe: $h = \frac{420}{1000} = 0{,}42$

2) $\Phi(z) = \frac{\gamma+1}{2} = \frac{95\,\%+100\,\%}{2} = 97{,}5\,\% \implies z = 1{,}9599...$

Im Bild rechts ermitteln wir dieses 97,5 %-Quantil der Standardnormalverteilung.

3) $h - z \cdot \sqrt{\frac{h \cdot (1-h)}{n}} = 0{,}3894...$ $\qquad h + z \cdot \sqrt{\frac{h \cdot (1-h)}{n}} = 0{,}4505...$

4) 95-%-Konfidenzintervall: $[0{,}3894...; 0{,}4505...]$

Berechnung eines Konfidenzintervalls (Formel I)

Gleiche Aufgabe mit $n = 1000$, $k = 420$ und $\gamma = 95\,\%$ in GeoGebra:

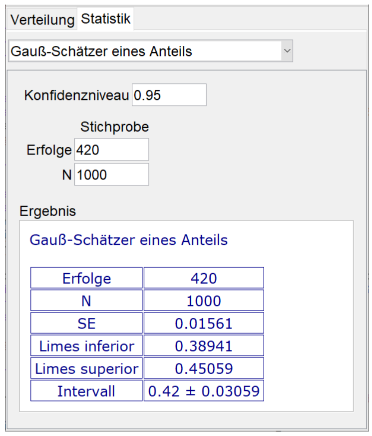
1) Ansicht → Wahrscheinlichkeitsrechner öffnen

2) Statistik auswählen

3) Gauß-Schätzer eines Anteils auswählen

4) Konfidenzniveau $\gamma = 95\,\%$ eingeben
Erfolge $k = 420$ eingeben
Stichprobengröße $n = 1000$ eingeben

5) Konfidenzintervall ablesen: [0,3894...; 0,4505...]



Überdeckungswahrscheinlichkeit (Formel I)

Enthält dieses Konfidenzintervall (KI) den unbekannten Anteil p mit 95 %-iger Wahrscheinlichkeit?

Zum Beispiel: $n = 1000$, $\gamma = 95\,\%$

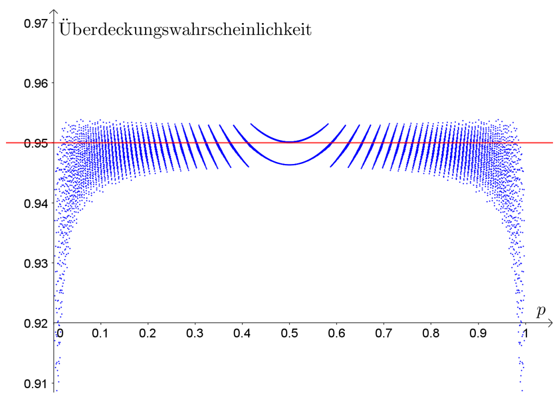
Bei gegebenem p können wir für jedes $k = 0, 1, 2, \ldots, 1000$ berechnen, ob das KI den Wert p enthält. Wir summieren die Wahrscheinlichkeiten aller Werte k, für die p im KI enthalten ist. Damit erhalten wir die **Überdeckungswahrscheinlichkeit** für diesen Wert p.

Rechts ist jeder Wert p mit der zugehörigen Überdeckungswahrscheinlichkeit als Punkt dargestellt.

$p \in [0; 1]$ mit Schrittweite 0,0001

Von diesen 10 001 Werten für p haben 6545 eine Überdeckungswahrscheinlichkeit *unter* 95 %.

Trotz der recht großen Stichprobe kann die Überdeckungswahrscheinlichkeit also deutlich *unter* 95 % liegen.



Überdeckungswahrscheinlichkeit (Formel II)

Wir verwenden wieder die Normalverteilung als Approximation für die Binomialverteilung. Die Näherungslösung runden wir diesmal aber nicht. Die Formel ist dann umfangreicher:

$$\left[\frac{1}{n+z^2}\cdot\left(n\cdot h+\frac{z^2}{2}-z\cdot\sqrt{n\cdot h\cdot(1-h)+\frac{z^2}{4}}\right)\,;\,\frac{1}{n+z^2}\cdot\left(n\cdot h+\frac{z^2}{2}+z\cdot\sqrt{n\cdot h\cdot(1-h)+\frac{z^2}{4}}\right)\right]$$

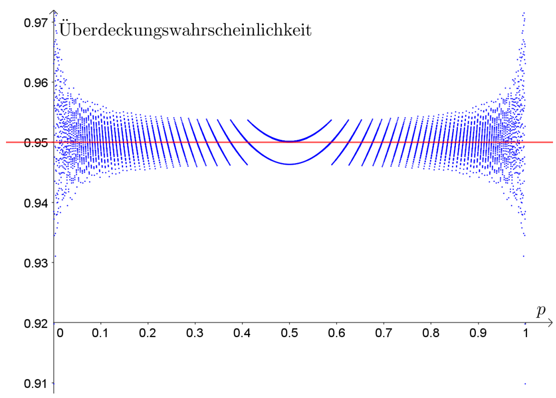
Zum Beispiel: $n = 1000$, $\gamma = 95\,\%$

Rechts ist jeder Wert p mit der zugehörigen Überdeckungswahrscheinlichkeit als Punkt dargestellt.

$p \in [0; 1]$ mit Schrittweite 0,0001

Im Vergleich zur gerundeten Formel verbessert sich das Ergebnis dadurch für Werte p nahe bei 0 oder 1.

Von diesen 10 001 Werten für p haben dennoch 4487 eine Überdeckungswahrscheinlichkeit *unter* 95 %.



Clopper-Pearson-Konfidenzintervall

Tatsächlich ist seit 1934 eine Formel bekannt, mit der wir *zuverlässige* Aussagen zum KI treffen können. Diese Formel von **Clopper** und **Pearson** lautet:

$$\left[b(\tfrac{\alpha}{2}; k; n-k+1)\,;\, b(1-\tfrac{\alpha}{2}; k+1; n-k)\right]$$

Dabei ist n die Stichprobengröße und k die Anzahl der „Erfolge“ in der Stichprobe.

Für das Signifikanzniveau α und das Konfidenzniveau γ gilt: $\alpha = 1-\gamma$

Mit $b(p; a; b)$ wird das p-Quantil einer betaverteilten Zufallsvariable mit Parametern a und b abgekürzt.

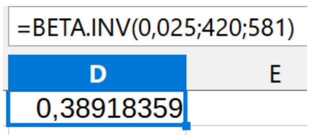
Clopper-Pearson-Konfidenzintervall

Wir berechnen das Clopper-Pearson-Konfidenzintervall mit $n = 1000$, $k = 420$ und $\gamma = 95\,\%$.

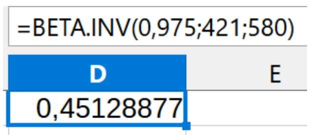
1) $\alpha = 1-\gamma = 5\,\%$

2) Berechnung der Quantile in einer Tabellenkalulation:

$b(0{,}025; 420; 581) = 0{,}3891...$



$b(0{,}975; 421; 580) = 0{,}4512...$



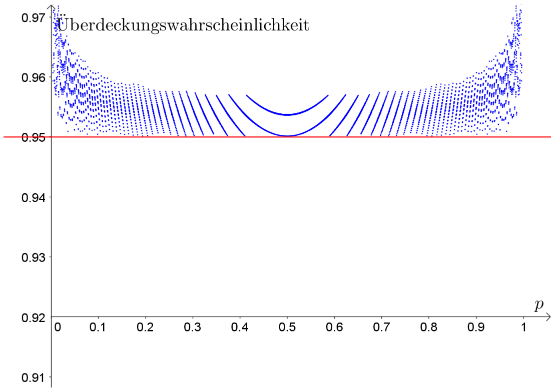
3) Das Clopper-Pearson-KI für p ist bei diesen Werten n, k und γ also $[0{,}3891...; 0{,}4512...]$.

Überdeckungswahrscheinlichkeit (Clopper-Pearson)

Wir untersuchen wieder die Überdeckungswahrscheinlichkeit für $n = 1000$ und $\gamma = 95\,\%$:

Rechts ist jeder Wert p mit der zugehörigen Überdeckungswahrscheinlichkeit als Punkt dargestellt.

$p \in [0; 1]$ mit Schrittweite 0,0001



Für *jede* Zahl p ist die Überdeckungswahrscheinlichkeit *mindestens* 95 %. Das ist kein Zufall.

Tatsächlich liefert die Clopper-Pearson-Formel für jedes n und γ ein *zuverlässiges* Konfidenzintervall. Man kann nämlich folgende Aussage beweisen:

Die Wahrscheinlichkeit, dass das Clopper-Pearson-Konfidenzintervall den unbekannten relativen Anteil p enthält, ist *mindestens* γ.

Vergleich

Das Clopper-Pearson-KI liefert eine zuverlässige Aussage. Es kann dafür etwas breiter sein.
Um wie viel Prozent ist das Clopper-Pearson-KI bei $n = 1000$, $k = 420$ und $\gamma = 95\,\%$ breiter?

Approximation mit Normalverteilung:
$[0{,}389\,409...; 0{,}450\,590...]$

Breite $b_1 = 0{,}061\,181...$

Clopper-Pearson:
$[0{,}389\,183...; 0{,}451\,288...]$

Breite $b_2 = 0{,}062\,105...$

$$\frac{b_2 - b_1}{b_1} = 0{,}0151... \implies \text{Das Clopper-Pearson-KI ist in diesem Beispiel um } 1{,}51...\,\% \text{ breiter.}$$