§70: PARAMETERSCHÄTZUNG UND KONFIDENZINTERVALLE

70.1. Motivation

Bisher sind wir stets von theoretischen Modellen (z.B. „fairer Würfel") ausgegangen, die erlaubten, Parameter wie Erwartungswert oder Varianz einer Verteilung exakt zu berechnen.

In vielen realen Situation kennt man jedoch nur den Verteilungstyp und muss auf Grund von Stichproben die Parameter schätzen. Wie geht man dabei vor?

Die geschätzten Parameter sind i.A. fehlerhaft. Lässt sich ein Vertrauensintervall angeben, innerhalb dessen ein Parameter mit einer vorgegebenen Sicherheit liegt?

70.2. Def.:

Gegeben seien n Beobachtungswerte $x_1, \dots, x_n$ eines Zufallsexperiments. Dann nennen wir $(x_1, \dots, x_n)^T$ Stichprobe vom Umfang n. Die einzelnen x_i heißen Stichprobenwerte.

70.3. Beispiel

In einer Kiste befinden sich 10 000 Schrauben. Ein Teil davon ist fehlerhaft. Für eine Stichprobe werden 100 Schrauben entnommen. Die Zufallsvariable X_i beschreibt den Zustand der i-ten entnommenen Schraube:

$$X_i(\omega) = \begin{cases} 0 & \text{falls i-te Schraube in Ordnung} \\ 1 & \text{" " " defekt.} \end{cases}$$

Eine konkrete Realisierung des Zufallsvektors $(X_1, \dots, X_{100})^T$ liefere die Stichprobe $(x_1, \dots, x_{100})^T = (0, 1, 0, 0, 1, 0, \dots, 0, 1)^T$.

So wie wir den Zufallsvariablen Parameter wie Erwartungswert und Varianz zugeordnet haben, können wir auch für Stichproben Kenngrößen definieren:

70.4. Def.: Für eine Stichprobe $(x_1, \dots, x_n)$ definiert man

- den <u>Mittelwert</u> durch $\bar{x} := \frac{1}{n}(x_1 + \dots + x_n)$
- die <u>Varianz</u> durch $s^2 := \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2$
- die <u>Standardabweichung</u> durch $s := \sqrt{s^2}$.

70.5. <u>Bemerkungen</u>:

a) Man kann zeigen, dass der Mittelwert $\bar{x}$ und die Varianz s^2 geeignete Approximationen an den Erwartungswert μ und die Varianz σ^2 einer Zufallsvariablen sind.

b) Die Tatsache, dass im Nenner von s^2 die Größe $n-1$ und nicht n steht, hat tiefere theoretische Gründe, auf die wir hier nicht eingehen. (Siehe z.B. Hartmann, Satz 21.9).

c) Ähnlich zum Verschiebungssatz 65.12 gibt es eine häufig benutzte Formel zum Berechnen der Varianz einer Stichprobe:

$$\boxed{s^2 = \frac{1}{n-1}\left(\sum_{i=1}^{n} x_i^2 - n\bar{x}^2\right)}$$

Beweis:

$$\sum_{i=1}^{n} (x_i - \bar{x})^2 = \sum_{i=1}^{n} (x_i^2 - 2x_i\bar{x} + \bar{x}^2) = \sum_{i=1}^{n} x_i^2 - 2\bar{x} \underbrace{\sum_{i=1}^{n} x_i}_{n\bar{x}} + n\bar{x}^2$$

$$= \sum_{i=1}^{n} x_i^2 - n\bar{x}^2. \qquad \square$$

70.6. Beispiel

Bei einer Wahlumfrage geben 400 von 1000 Personen an, die Partei A zu wählen. Das Umfrageinstitut prognostiziert auf Grund dieser Stichprobe einen Wahlausgang mit 40 % aller Stimmen für Partei A.

70.7. Konfidenzintervalle

In Beispiel 70.6 werden verschiedene Stichproben zu leicht unterschiedlichen Resultaten führen, die wiederum i.A. alle vom tatsächlichen Wahlausgang abweichen.
Können wir statt eines einzelnen Werts $p = 0{,}4$ ein Vertrauensintervall (Konfidenzintervall) $[p_u, p_o]$ angeben, innerhalb dessen das Endresultat mit einer vorgegebenen W.S. (Konfidenzniveau) von z.B. 95 % liegt?

70.8. Beispiel: Wahlumfrage aus 70.6

Wir gehen von einer Binomialverteilung aus und schätzen p durch $p = \frac{400}{1000} = 0{,}4$ ab. Sei

$$X_i := \begin{cases} 1 & \text{Befragte/r } i \text{ wählt A} \\ 0 & \text{" " " " nicht} \end{cases}$$

und $X = \sum_{i=1}^{1000} X_i$. Dann gilt nach 67.4. mit $p = 0{,}4$ und $n = 1000$:

$$\mu = E(X) = np = 400$$

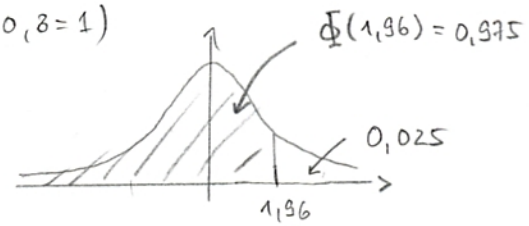
$$\sigma^2 = V(X) = np(1-p) = 240 \quad \Rightarrow \quad \sigma \approx 15{,}49$$

Wegen $np = 400 > 5$ und $n(1-p) = 600 > 5$ können wir für X auch eine Normalverteilung mit $\mu = 400$ und $\sigma = 15{,}49$ annehmen. (vgl. 68.8).

Wir suchen ein Intervall $[\mu - r\sigma\,;\, \mu + r\sigma]$, innerhalb dessen das Integral über die Dichtefunktion den Wert 0,95 annimmt.

Tabelliert ist die Standardnormalverteilung ($\mu = 0, \sigma = 1$)

$$\Phi(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-\frac{t^2}{2}}\, dt$$



Man findet: $\Phi(1{,}96) \approx 0{,}975$.

Somit ist aus Symmetriegründen:

$$\frac{1}{\sqrt{2\pi}} \int_{-1{,}96}^{1{,}96} e^{-\frac{t^2}{2}}\, dt \approx 0{,}95$$

und $r = 1{,}96$.

Damit ergibt sich das Konfidenzintervall

$$[\mu - r\sigma\,;\, \mu + r\sigma] = [400 - 1{,}96 \cdot 15{,}49\,;\, 400 + 1{,}96 \cdot 15{,}49]$$
$$\approx [369{,}6\,;\, 430{,}4]$$

Bei einem Konfidenzniveau von 95% erhält Partei A also zwischen 36,96 und 43,04 % der Stimmen. Möchte man ein kleineres Konfidenzintervall, muss man mehr Personen befragen.

10.9. Beispiel: Überbuchung eines Flugzeugs

Ein Flugzeug hat 200 Sitze. Wie viele Reservierungen dürfen angenommen werden, wenn erfahrungsgemäß 5 % aller Passagiere nicht erscheinen? Die Fluggesellschaft ist bereit, in 1 von 50 Flügen in Verlegenheit zu geraten.

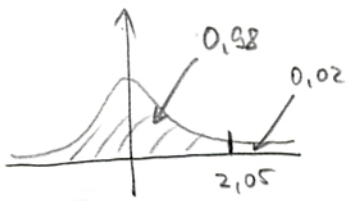
Sei n die Anzahl der Reservierungen und X die Anzahl der tatsächlich erscheinenden Passagiere.
Legt man eine Binomialverteilung zu Grunde mit $p = 0,95$, so gilt:

$$\mu = E(X) = n \cdot p = 0,95 \cdot n$$

$$\sigma^2 = n\,p\,(1-p) = 0,0475\,n \quad \Rightarrow \quad \sigma = 0,2179\sqrt{n}$$

Der Tabelle der Standardnormalverteilung entnimmt man:

$$\Phi(2,05) \approx 0,98 = 1 - \frac{1}{50}$$



Fordert man

$$\mu + 2,05\,\sigma \overset{!}{\leq} 200$$

ergibt sich

$$0,95\,n + 2,05 \cdot 0,2179\sqrt{n} \leq 200$$

Man prüft leicht nach, dass dies für $n \leq 203$ erfüllt ist.
(ausprobieren oder $y := \sqrt{n}$ setzen und quadr. Gleichung lösen)

Bem.: Die Approximation durch die Normalverteilung war gerechtfertigt wegen $np > 5$ und $n(1-p) > 5$.