Exkurs: Exakte Konfidenzintervalle für den Erfolgsparameter in der Binomialverteilung

27. Januar 2014

Unter n unabhängigen Versuchen seien x Erfolge beobachtet worden,
wir fassen x als Realisierung einer $\mathrm{Bin}_{n,\vartheta}$-verteilten ZV auf und wollen anhand der Beobachtung auf ϑ schließen.

Wir hatten in der Vorlesung das auf asymptotischer Normalität fußende (approximative) Konfidenzintervall für ϑ zum Niveau $1-\alpha$ kennen gelernt:

$$\left[\widehat{\vartheta} - q\frac{\widehat{\sigma}}{\sqrt{n}}, \widehat{\vartheta} + q\frac{\widehat{\sigma}}{\sqrt{n}}\right]$$

mit $\widehat{\vartheta} = \dfrac{x}{n}$, $\widehat{\sigma} = \sqrt{\widehat{\vartheta}(1-\widehat{\vartheta})}$, q das $1-\frac{\alpha}{2}$-Quantil von $\mathcal{N}_{0,1}$

Wie könnten wir vorgehen, wenn wir uns nicht auf die Asymptotik verlassen möchten?

Beobachte $X \sim P_\vartheta := \mathrm{Bin}_{n,\vartheta}$,
Konfidenzintervall für $\vartheta \in \Theta = [0,1]$?

Idee: Zu $\vartheta \in \Theta := [0,1]$ wähle $c_\vartheta \in (0,1)$, so dass für

$$C_\vartheta := \big\{x \in \{0,1,\dots,n\} : \mathrm{Bin}_{n,\vartheta}(\{x\}) \geq c_\vartheta\}$$

gilt $\mathrm{Bin}_{n,\vartheta}(C_\vartheta) \geq 1-\alpha$
(und c_ϑ möglichst groß, so dass C_ϑ möglichst klein).

Setze $C(x) := \{\vartheta \in \Theta : x \in C_\vartheta\}$ für $x \in \mathcal{X} := \{0,1,\dots,n\}$,
dann gilt

$$\forall\, \vartheta \in \Theta\ :\ P_\vartheta(\vartheta \in C(X)) = P_\vartheta(X \in C_\vartheta) \geq 1-\alpha$$

nach Konstruktion.

Es gilt

1. Für $\vartheta \in (0,1)$ ist $\{0,\dots,n\} \ni x \mapsto \mathrm{Bin}_{n,\vartheta}(\{x\})$ strikt wachsend auf $\{0,1,\dots,\lceil (n+1)\vartheta - 1\rceil\}$, strikt fallend auf $\{\lfloor (n+1)\vartheta\rfloor,\dots,n\}$, also maximal auf $x = \lfloor (n+1)\vartheta \rfloor$ (und auf $(n+1)\vartheta - 1$, wenn $(n+1)\vartheta \in \mathbb{Z}$).
2. Für $x \in \{1,\dots,n\}$ ist $[0,1] \ni \vartheta \mapsto Bin_{n,\vartheta}(\{x, x+1,\dots,n\})$ stetig, strikt monoton wachsend mit

   $$\mathrm{Bin}_{n,\vartheta}(\{x, x+1,\dots,n\}) = \mathrm{Beta}_{x,n-x+1}([0,\vartheta]),$$

   wo $\mathrm{Beta}_{a,b}$ die Dichte

   $$f_{\mathrm{Beta}_{a,b}}(u) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} u^{a-1}(1-u)^{b-1}$$

   auf $(0,1)$ hat.

denn:

1.

$$\begin{aligned}\frac{\mathrm{Bin}_{n,\vartheta}(\{x\})}{\mathrm{Bin}_{n,\vartheta}(\{x-1\})} &= \frac{\binom{n}{x}\vartheta^x(1-\vartheta)^{n-x}}{\binom{n}{x-1}\vartheta^{x-1}(1-\vartheta)^{n-x+1}} \\ &= \frac{(n-x+1)\vartheta}{x(1-\vartheta)} \\ &> 1 \\ \iff x &< (n+1)\vartheta\end{aligned}$$

2. $U_1, \ldots, U_n$ unabhängig und uniform auf $[0,1]$,

$$S_\vartheta := \sum_{i=1}^{n} \mathbf{1}_{[0,\vartheta]}(U_i)$$

ist $\mathrm{Bin}_{n,\vartheta}$-verteilt.

Sei $U_{(1)} < U_{(2)} < \ldots < U_{(n)}$ die „Ordnungsstatistik“.

$$
\begin{aligned}
\mathrm{Bin}_{n,\vartheta}(\{x,\ldots,n\}) &= P(S_\vartheta \ge x) = P(U_{(x)} \le \vartheta)\\
&= \sum_{k=1}^{n} \sum_{\substack{B\subseteq\{1,\ldots,n\}\setminus\{k\}\\ |B|=x-1}} \underbrace{P\begin{pmatrix} U_k \le \vartheta, U_m \le U_k \text{ für } m\in B,\\ U_l > U_k \text{ für } l \in \{1,\ldots,n\}\setminus(\{k\}\cup B)\end{pmatrix}}_{=\int_0^\vartheta u^{|B|}(1-u)^{n-|B|-1}\,du=\int_0^\vartheta u^{x-1}(1-u)^{n-x}\,du}\\
&= \underbrace{n\binom{n-1}{x-1}}_{=\frac{n!}{(x-1)!(n-x)!}=\frac{\Gamma(n+1)}{\Gamma(x)\Gamma(n-x+1)}} \int_0^\vartheta u^{x-1}(1-u)^{n-x}\,du
\end{aligned}
$$

Wähle $C_\vartheta := \{x_-(\vartheta), x_-(\vartheta)+1, \ldots, x_+(\vartheta)\}$ mit
$x_-(\vartheta) = \max\{x : \mathrm{Bin}_{n,\vartheta}(\{0,\ldots,x-1\}) \leq \frac{\alpha}{2}\}$ und
$x_+(\vartheta) = \min\{x : \mathrm{Bin}_{n,\vartheta}(\{x+1,\ldots,n\}) \leq \frac{\alpha}{2}\}$.

Es gilt:

- $x \leq x_+(\vartheta) \iff$
  $Bin_{n,\vartheta}(\{x,\ldots,n\}) = Beta_{x,n-x+1}([0,\vartheta]) > \frac{\alpha}{2}$
  $\iff \vartheta > p_-(x) := \frac{\alpha}{2}$-Quantil von $\mathrm{Beta}_{x,n-x+1}$.
- $x \geq x_+(\vartheta) \iff \mathrm{Bin}_{n,\vartheta}(\{0,\ldots,x\}) =$
  $1 - \mathrm{Bin}(\{x+1,\ldots,n\}) = \mathrm{Beta}_{x+1,n-x}([\vartheta,1]) \geq \frac{\alpha}{2}$.
  $\iff \vartheta < p_+(x) := 1 - \frac{\alpha}{2}$-Quantil von $\mathrm{Beta}_{x+1,n-x}$.

Satz (exaktes Konfidenzintervall im Binomialmodell)

$$p_-(x) := \tfrac{\alpha}{2}\text{-Quantil von } \mathrm{Beta}_{x,n-x+1},$$
$$p_+(x) := 1 - \tfrac{\alpha}{2}\text{-Quantil von } \mathrm{Beta}_{x+1,n-x}$$

$x \mapsto [p_-(x), p_+(x)]$ ist ein Konfidenzintervall für ϑ zum Sicherheitsniveau $1-\alpha$.

Satz (exaktes Konfidenzintervall im Binomialmodell)

$$p_-(x) := \tfrac{\alpha}{2}\text{-Quantil von } \mathrm{Beta}_{x,n-x+1},$$
$$p_+(x) := 1 - \tfrac{\alpha}{2}\text{-Quantil von } \mathrm{Beta}_{x+1,n-x}$$

$x \mapsto [p_-(x), p_+(x)]$ ist ein Konfidenzintervall für ϑ zum Sicherheitsniveau $1 - \alpha$.

Bem.:

- Quantile der Beta-Verteilungen sind tabelliert, gelegentlich kann beim Nachschlagen in Tabellen die Symmetrieeigenschaft

  $$\mathrm{Beta}_{a,b}([0, x]) = \mathrm{Beta}_{b,a}([1-x, 1]) = 1-\mathrm{Beta}_{b,a}([0, 1-x])$$

  nützlich sein.
- `R` kennt die Beta-Verteilungen, ihre Verteilungsfunktionen `pbeta(x, a, b)` und ihre Quantile `qbeta(p, a, b)`

Beispiel: $n = 53$, $x = 23$, wähle $\alpha = 0{,}05$

$\widehat{\vartheta} = \frac{23}{53} \approx 0{,}434$, $\widehat{\sigma} \approx 0{,}496$, $q_{0{,}975} \approx 1{,}96$

$\left[\widehat{\vartheta} \pm q\frac{\widehat{\sigma}}{\sqrt{53}}\right] \approx [0{,}30, 0{,}57]$

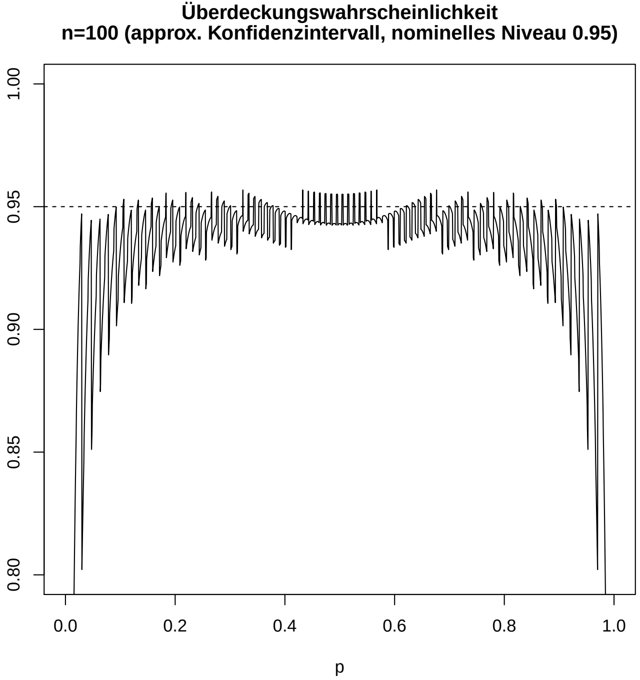
$p_-(23) = 0{,}025$-Quantil von $\mathrm{Beta}_{23,31} \approx 0{,}30$,

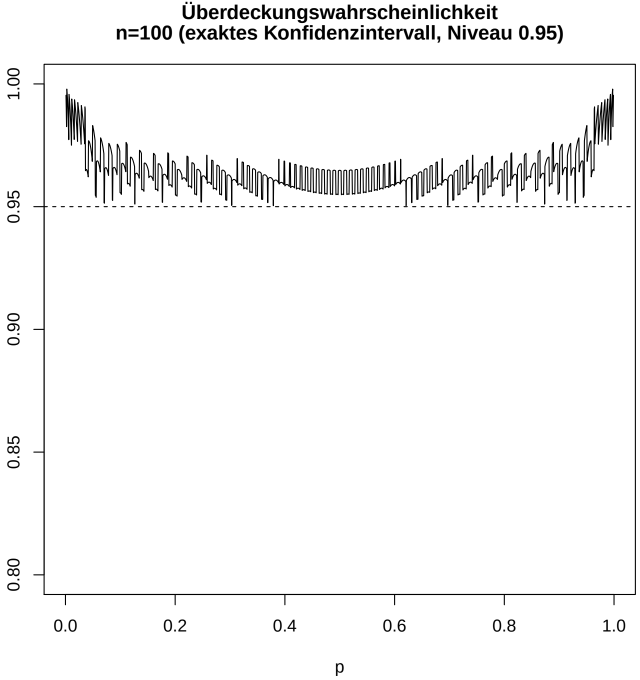
$p_+(23) = 0{,}975$-Quantil von $\mathrm{Beta}_{24,30} \approx 0{,}57$

Absurd präzise Werte wären:

$\left[\widehat{\vartheta} \pm q\frac{\widehat{\sigma}}{\sqrt{53}}\right] \approx [0{,}3005306, 0{,}5673939]$

$[p_-(23), p_+(23)] \approx [0{,}2983921, 0{,}5771742]$


Überdeckungswahrscheinlichkeit
n=100 (approx. Konfidenzintervall, nominelles Niveau 0.95)
1.00
0.95
0.90
0.85
0.80
0.0
0.2
0.4
0.6
0.8
1.0
p


Überdeckungswahrscheinlichkeit
n=100 (exaktes Konfidenzintervall, Niveau 0.95)
1.00
0.95
0.90
0.85
0.80
0.0
0.2
0.4
0.6
0.8
1.0
p

Manchmal betrachtet man auch den Schätzer

$$\widetilde{\vartheta} = \frac{x+1}{n+2}$$

für ϑ und bildet als (approximatives) Konfidenzintervall zum Niveau $1-\alpha$

$$\left[\widetilde{\vartheta} - q\frac{\widetilde{\sigma}}{\sqrt{n}}, \widehat{\vartheta} + q\frac{\widetilde{\sigma}}{\sqrt{n}}\right]$$

mit $\widetilde{\sigma} = \sqrt{\widetilde{\vartheta}(1-\widetilde{\vartheta})}$, $\quad q$ das $1-\frac{\alpha}{2}$-Quantil von $\mathcal{N}_{0,1}$

[$\widetilde{\vartheta}$ wirkt vielleicht zunächst „ad hoc“ gewählt.
Die Form von $\widetilde{\vartheta}$ ist natürlich im Kontext eines Bayesschen Ansatzes, den wir in dieser Vorlesung nicht weiter diskutieren werden.]

Überdeckungswahrscheinlichkeit
n=100 (modifiziertes approx. Konfidenzint., nominelles Niveau 0.95)