

Lösungen einiger Verständnisfragen und -aufgaben zu Kap. 25: Schätzfunktionen

Verständnisfrage:

Zu 25.1: Wie kommt man von der Verteilung von X auf die für $\hat{\theta}$?

Es gilt $\hat{\theta} = X/n$, wobei die Zufallsvariable X den Erwartungswert $E(X) = n\theta$ und die Varianz $V(X) = n\theta(1-\theta)$ besitzt. Mit der Formel für die Linearkombination $Y = a + bX$ von Zufallsvariablen,

$$E(Y) = a + bE(X), \quad V(Y) = b^2V(X)$$

bekommt man direkt

$$E(\hat{\theta}) = \theta, \quad V(\hat{\theta}) = \frac{\theta(1-\theta)}{n}.$$

Verständnisfrage:

Zu 25.2(b):

- (i) Finden Sie ein Beispiel eines konsistenten, aber nicht erwartungstreuen Schätzers
- (ii) Finden Sie ein Beispiel eines erwartungstreuen, aber nicht konsistenten Schätzers
- (iii) Zeigen Sie, dass \bar{X} bzw. f als Schätzer des Erwartungswertes $E(X) = \mu$ bzw. des Anteilswertes θ konsistent, erwartungstreu und effizient sind.

(i) z.B. folgender Schätzer für den Erwartungswert für $n > 8$:

$$\bar{X}'_n = \frac{1}{n} \sum_5^{n-4} X_i.$$

Da für unabhängige Zufallsvariable mit $E(X) = \mu$, $V(X) = \sigma^2$:

$$E(\bar{X}'_n) = \frac{n-8}{n}E(X) \neq E(X), \quad V(\bar{X}'_n) = \frac{\sigma^2}{n-8}$$

ist \bar{X}'_n er nicht erwartungstreu, aber wegen $\lim_{n \rightarrow \infty} V(\bar{X}'_n) = 0$ konsistent. Der Varianz-Schätzer $\frac{1}{n} \sum_{i=1}^n (X_i - E(X))^2$ ist übrigens auch konsistent, aber nicht erwartungstreu.

(ii) z.B. folgender Schätzer für den Erwartungswert:

$$\bar{X}'_n = \frac{X_1 + X_2 + X_3}{3}$$

Es gilt $E(X'_n) = E(X)$. Da aber die Varianz wegen der fehlenden Ausnutzung aller Stichprobenelemente > 3 für $n \rightarrow \infty$ nicht gegen Null geht, ist dieser Schätzer nicht konsistent.

(iii) Es gilt für unabhängige Zufallsvariable mit $E(X) = \mu$, $V(X) = \sigma^2$:

$$E(\bar{X}) = \mu, \quad E(\bar{X} - \mu)^2 = \frac{\sigma_x^2}{n}$$

Dies ist erwartungstreu wegen $E(\bar{X}) = \mu$ und konsistent wegen $\lim_{n \rightarrow \infty} E(\bar{X} - \mu)^2 \rightarrow 0$.
Da darüberhinaus $c = \bar{X}$ die Fehlerquadratsumme

$$F(c) = \sum_{i=1}^n (X_i - c)^2$$

minimiert (vgl. Statistik 1), hat \bar{X} bei endlichem n von allen möglichen Konstanten die geringste Schwankungsbreite, ist also effizient. Ein nicht effizienter, aber erwartungstreuere und konsistenter Schätzer wäre z.B.

$$\bar{X}' = \frac{2}{n} \sum_{i=1}^{n/2} X_{2i}.$$

Verständnisaufgabe zu Kap. 25.3:

Konkretisieren Sie die Formel für die Gauß-Statistik für den Anteilswert ($n \geq 30$, Auswahlsatz der Stichprobe $< 5\%$).

Am einfachsten definiert man die "Pseudovariablen" X so, dass $X = 1$, wenn ein Element der Stichprobe zum Anteil gehört, und $X = 0$ sonst. Dann ist $X \sim B(1, \theta)$ mit

- $\mu = E(X) = \theta$,
- $\sigma^2 = V(X) = \theta(1 - \theta)$
- und der Mittelwert-Schätzer gleich dem Anteilswert: $\bar{X} = f$.

Für $n > 30$ bzw. $n\theta(1 - \theta) > 9$ sowie Unabhängigkeit der Stichprobenelemente (dafür ist ein Auswahlsatz kleiner 5% notwendig) ist f annähernd gaußverteilt, so dass für ihn die Gauß-Statistik gilt. Setzt man nun in den im Skript angegebenen Ausdruck die obigen Spezialisierungen für den Anteilswert ein, erhält man direkt

$$Z = (f - \theta) \sqrt{\frac{n}{\theta(1 - \theta)}}.$$

Beispielaufgabe zu Kap. 25.3:

Mittels einer Stichprobe soll der Anteil der Studenten in Sachsen ermittelt werden, die ein eigenes Kfz besitzen.

- Geben Sie die allgemeine Formel für den notwendigen Stichprobenumfang als Funktion des tatsächlichen Anteilswertes θ und der zulässigen Standardabweichung σ_f seines Stichprobenschätzers f an.
- Erwartet wird ein Anteil von 35%. Wie groß muss die Stichprobe sein, damit der Stichprobenfehler (Standardabweichung von f) weniger als 2 % beträgt?
- Mit welcher Wahrscheinlichkeit ist bei einem tatsächlichen Anteilswert von $\theta = 35\%$ und $n = 569$ Studenten der Fehler oberhalb 3%, die relative Stichprobenhäufigkeit also oberhalb 38%?

- (a) Da beim Anteilswert die zugeordnete (Pseudo-) Zufallsvariable $X_i \sim B(1, \theta)$ binomialverteilt ist mit der Varianz $V(X_i) = \theta(1 - \theta)$ (vgl. die vorige Aufgabe), ist nach den Rechenregeln für die Varianz und mit $\bar{X} = f$:

$$\sigma_f^2 = V(f) = V\left(\frac{1}{n} \sum X_i\right) = \frac{1}{n^2} n V(X_i) = \frac{\theta(1 - \theta)}{n}.$$

(Hier wurde die Summenregel für Varianzen angewandt, welche nur bei unabhängigen Stichprobenelementen, also für einen Auswahlsatz $< 5\%$ gilt. Da die Grundgesamtheit nach Aufgabenstellung die Studenten in ganz Sachsen sind, ist dies bei sinnvollen Stichprobengrößen immer erfüllt.) Diese Stichprobenvarianz muss man lediglich nach n umstellen:

$$n = \frac{\theta(1 - \theta)}{\sigma_f^2}.$$

(b) Es ist

$$\theta = 0.35, \quad \sigma_f^2 = 0.02^2$$

und damit

$$n \geq \underline{\underline{569.}}$$

(c) Die Wahrscheinlichkeit dafür, dass der Anteilswert-Schätzer f bei einem tatsächlichen Anteilswert $\theta=35\%$ den Wert 38% überschreitet, ergibt sich aus dem entsprechenden Wert der Gauß-Statistik, also der Standardnormalverteilung $\Phi(z)$. Mit $E(f) = \theta$, $V(f) = \theta(1 - \theta)/n$ sowie einer Gaußverteilung für f ist

$$Z = \frac{f - E(f)}{\sqrt{V(f)}} = (f - \theta) \sqrt{\frac{n}{\theta(1 - \theta)}}$$

standardnormalverteilt (vgl. die vorige Aufgabe). Es gilt also

$$f = \theta + \sqrt{\frac{\theta(1 - \theta)}{n}} Z$$

und damit für die Wahrscheinlichkeit:

$$\begin{aligned} P(f > 0.38) &= P\left(\theta + \sqrt{\frac{\theta(1 - \theta)}{n}} Z > 0.38\right) \\ &= P\left(Z > (0.38 - \theta) \sqrt{\frac{n}{\theta(1 - \theta)}}\right) \\ &= P(Z > 1.5) \\ &= 1 - P(Z < 1.5) \\ &= 1 - \Phi(1.5) \\ &\stackrel{\text{Tabelle}}{=} 1 - 0.9332 = \underline{\underline{0.0668}}. \end{aligned}$$

Herleitungs-Aufgabe zu Kap. 25.4:



Warum steht bei der Stichprobenvarianz $(n - 1)$ und nicht n im Nenner?

Siehe das Dokument [exkurse17.pdf](#).

Rechen-Aufgabe zu Kap. 25.4:

Eine Abfüllmaschine für Nutella füllt im Mittel $\mu = 402$ g Nutella in die 400g-Gläser, wobei die tatsächliche Streuung $\sigma = \sqrt{\sigma^2} = 2$ g beträgt. Ferner seien die Füllmengen gaußverteilt (z.B. der Fall, wenn es viele kleine, unabhängige Ursachen für die Schwankungen gibt!)

Stiftung Warentest kauft

- (i) 2 Gläser,
- (ii) 10 Gläser,
- (iii) 100 Gläser

und bestimmt jeweils Stichprobenmittel \bar{X} und Stichprobenvarianz S^2 bzw. Stichprobenstreuung $S = \sqrt{S^2}$.

Nutella ist “technisch durchgefallen”,

- wenn $\bar{X} < 400$ g beträgt,
- oder wenn die ermittelte Stichprobenstreuung des Füllgewichts $S > 3$ g ist.

Berechnen Sie jeweils für diese beiden Einzelereignisse die Wahrscheinlichkeit.

Noch aufzuschreiben (Übungen?)

Verständnisfrage:

Zu 25.5 Philosophische” Frage: Für einen Schätzer (für das Mittel) gibt es sowohl die Gauß-Statistik als auch die T-Statistik. Gibt es also zwei Verteilungen?

Nein. Falls die Zufallsvariable X gaußverteilt ist, so ist \bar{X} immer ebenfalls gaußverteilt, unabhängig davon, ob man die Standardabweichung kennt oder nicht. Im letzteren Fall hat man allerdings das Problem, dass man nicht nur den Mittelwert abschätzen muss, sondern auch die Verteilung des *Schätzers*, da man sonst keine Wahrscheinlichkeitsaussagen machen kann. Man muss also Eigenschaften des *Schätzers ebenfalls abschätzen*, um Aussagen über den Erwartungswert machen zu können. Diese zweistufige Schätzung führt letztendlich zu einer im Vergleich zur Gauß-Statistik “breiteren” Verteilung, der T -Statistik.

Aufgabe zu Kap. 25.6: Regressionsparameter



Berechnen Sie Mittelwert und Varianz der aus einer Stichprobe gemäß Abschnitt 25.6 gewonnenen empirischen Regressionsfunktion $\hat{Y}(x)$!

Wir setzen zunächst die Ausdrücke für die Schätzer \hat{A} und \hat{B} der Regressionskoeffizienten in die empirische Regressionsfunktion ein:

$$\begin{aligned}\hat{Y}(x) &= \hat{A} + \hat{B}x \\ &= \bar{Y} + \hat{B}(x - \bar{x}) \\ &= \frac{1}{n} \sum_{i=1}^n Y_i + \frac{(x - \bar{x})}{ns_x^2} \sum_{i=1}^n (Y_i - \bar{Y})(x_i - \bar{x}) \\ &= \frac{1}{n} \sum_{i=1}^n Y_i \left(1 + \frac{(x_i - \bar{x})(x - \bar{x})}{s_x^2} \right).\end{aligned}$$

Dabei wurde die übliche Definition $\sum_{i=1}^n (x_i - \bar{x})^2 = ns_x^2$ verwendet (wobei s_x^2 hier die Streubreite der x -Werte, aber *keine* Varianz im statistischen Sinne darstellt!), sowie die Identität $\sum_i (x_i - \bar{x})\bar{Y} = 0$.

Damit ist, für beliebige Werte von x , $\hat{Y}(x)$ als Linearkombination der Zufallsvariablen Y_i dargestellt. (die x_i sowie die unabhängige Variable sind ja im statistischem Sinne feste Werte!)

Der *Erwartungswert* ergibt sich wie üblich durch den Erwartungswert einer Linearkombination:

$$\begin{aligned}E(\hat{Y}(x)) &= \frac{1}{n} \sum_{i=1}^n E(Y_i) \left(1 + \frac{(x_i - \bar{x})(x - \bar{x})}{s_x^2} \right) \\ &= \frac{1}{n} \sum_{i=1}^n (a + bx_i) \left(1 + \frac{(x_i - \bar{x})(x - \bar{x})}{s_x^2} \right) \\ &= a + b\bar{x} + \frac{b}{n} \sum_{i=1}^n \left(\frac{x_i(x_i - \bar{x})(x - \bar{x})}{s_x^2} \right) \\ &= \underline{\underline{a + bx.}}\end{aligned}$$

Hier wurden die Identitäten $\sum_i x_i = n\bar{x}$, $\sum_i (x_i - \bar{x})(x - \bar{x}) = 0$ sowie

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

ausgenutzt.

Mit der Unabhängigkeit der Schwankungen der Y_i bekommt man schließlich für die Varianz der Regressionsfunktion

$$V(\hat{Y}(x)) = \frac{1}{n^2} \sum_{i=1}^n V(Y_i) \left(1 + \frac{(x_i - \bar{x})(x - \bar{x})}{s_x^2} \right)^2$$

$$\begin{aligned}
&= \frac{1}{n^2} \sum_{i=1}^n \sigma_R^2 \left(1 + \frac{(x_i - \bar{x})(x - \bar{x})}{s_x^2} \right)^2 \\
&= \frac{\sigma_R^2}{n^2} \left(n + 2(x - \bar{x}) \sum_{i=1}^n \frac{(x_i - \bar{x})}{s_x^2} + (x - \bar{x})^2 \sum_{i=1}^n \frac{(x_i - \bar{x})^2}{s_x^4} \right) \\
V(\hat{Y}(x)) &= \underline{\underline{\frac{\sigma_R^2}{n} \left(1 + \frac{(x - \bar{x})^2}{s_x^2} \right)}}.
\end{aligned}$$

Die Zone innerhalb der Standardabweichung wird also beschrieben durch eine Hyperbel mit der Taille (minimale Breite) σ_R/\sqrt{n} bei $x = \bar{x}$ und der Steigung $\sigma_R/(\sqrt{n}s_x)$ der Hyperbeläste.

Schreibt man die Schätzfunktion als

$$\hat{Y}(x) = \bar{Y} + \bar{B}(x - \bar{x}),$$

und vergleicht mit der Varianzformel, sieht man, dass \bar{Y} und $\bar{B}(x - \bar{x})$ unabhängig voneinander sind, da die Varianzformel einen konstanten Anteil und einen proportional zu $(x - \bar{x})^2$ hat, aber keinen proportional zu $(x - \bar{x})$. Da x und \bar{x} feste Größen sind, bedeutet dies, dass \bar{Y} und \hat{B} voneinander unabhängig sind. Wendet man die Regeln für die Varianz einer Linearkombination von unabhängigen Zufallsvariablen an, erhält man

$$V(\hat{Y}(x)) = V(\bar{Y}) + (x - \bar{x})^2 V(\bar{B})$$

Vergleich mit der allgemeinen Varianzformel liefert so die Varianz des Schätzers des Regressionsparameters b :

$$V(\hat{B}) = \frac{\sigma_R^2}{ns_x^2} \quad (1)$$

Verständnisfrage zu Kap. 25.6(c)

Machen Sie sich anhand einer Stichprobe vom Umfang $n = 2$ klar, dass obiger Schätzer mit $n - 1$ oder n im Nenner nicht erwartungstreu sein kann.

Mit zwei Wertepaaren erhält man sonst immer als Schätzer $\hat{\sigma}_R^2 = 0$ (durch zwei Punkte kann man immer exakt eine Regressionsgerade legen), was nicht erwartungstreu sein kann.