Gütekriterien für Schätzer

5

Nachdem wir im vorherigen Abschnitt erste Schätzer kennengelernt haben, wollen wir nun der Frage nachgehen, nach welchen Kriterien man einen Schätzer beurteilen und mit anderen Schätzern vergleichen kann.

5.1 Gütekriterien für Schätzer

Im Kontext von Beispiel 4.4 haben wir erste Schätzer diskutiert und bemerkt, dass wir prinzipiell verschiedene Schätzer nutzen können, denn ein Schätzer ist ja lediglich eine Abbildung vom Bildraum des Beobachtungsvektors, d. h. eine Auswertung $S(\mathfrak{X})$ des Zufallsvektors $\mathfrak{X}$. Der Schätzer unterliegt damit zufälliger Variabilität und wird i. Allg. die unbekannte zu schätzende Kenngröße verfehlen. Wir brauchen daher statistische Kriterien, die die Güte von Schätzern im Rahmen des zugrunde liegenden statistischen Modells beurteilen. Dazu zählen Eigenschaften, bei denen ein Schätzer den wahren Parameter zumindest in Erwartung trifft *(Erwartungstreue)* oder ihm mit wachsender Wahrscheinlichkeit näherkommt, wenn mehr und mehr Beobachtungen herangezogen werden *(Konsistenz)*.

Definition 5.1 (Erwartungstreue)
Es sei ein statistisches Modell gegeben durch einen Zufallsvektor $\mathfrak{X} = (X_1, \ldots, X_n)^t$ *und eine Verteilungsfamilie* $(\nu_\vartheta)_{\vartheta \in \Theta}$. *Ein unter allen* ν_ϑ *integrierbarer Schätzer* S *heißt erwartungstreu für die abgeleitete Kenngröße* $\tau(\vartheta)$, *falls für alle* $\vartheta \in \Theta$ *gilt*

$$\mathbb{E}_\vartheta[S(\mathfrak{X})] = \tau(\vartheta).$$

M. Messer und G. Schneider, *Statistik*, https://doi.org/10.1007/978-3-662-59339-4_5

Einen erwartungstreuen Schätzer nennen wir auch *unverzerrt* (engl. *unbiased*). Die Erwartungstreue von S bedeutet: Falls ν_ϑ die wahre Verteilung von $\mathfrak{X}$ ist, dann trifft der Schätzer $S(\mathfrak{X})$ die wahre zu schätzende Größe in Erwartung.

Beispiel 5.2 (Erwartungstreue Schätzer)

i. Seien $X_1, \ldots, X_n$ unabhängige und identisch verteilte Zufallsvariable, $X_1 \sim ber(p)$ mit $p \in \Theta := [0,1]$. Dann ist $\hat{p}(\mathfrak{X}) = \bar{X}_n$ erwartungstreu für p. Denn aufgrund der Linearität des Erwartungswertes folgt für alle $p \in \Theta$, dass $\mathbb{E}_p[\bar{X}_n] = (1/n)\sum_{i=1}^n \mathbb{E}_p[X_i] = (1/n)n\mathbb{E}_p[X_1] = \mathbb{E}_p[X_1] = p$.

ii. Analog gilt allgemeiner: Seien $X_1, \ldots, X_n$ unabhängige und identisch verteilte Zufallsvariable mit $X_1 \sim \nu_\vartheta$, und ν_ϑ ist Mitglied der Familie $(\nu_\vartheta)_{\vartheta \in \Theta}$ aller integrierbarer Verteilungen. Dann ist der Mittelwert $\bar{X}_n$ erwartungstreu für den Erwartungswert $\mathbb{E}_\vartheta[X_1]$. Dies folgt analog zu i).

iii. Ersetzt man in ii) die Familie $(\nu_\vartheta)_{\vartheta \in \Theta}$ durch die Familie aller quadratintegrierbaren Verteilungen, dann ist die korrigierte empirische Stichprobenvarianz s^2 aus (3.3) erwartungstreu für die Varianz $\mathbb{V}ar_\vartheta[X_1]$. Für die Berechnung von $\mathbb{E}_\vartheta[s^2(\mathfrak{X})]$ nutze man zum einen wieder die Linearität des Erwartungswertes, und zum anderen die Unabhängigkeit der Beobachtungen, genauer, dass $\mathbb{E}_\vartheta[X_i X_j] = \mathbb{E}_\vartheta[X_i]\mathbb{E}_\vartheta[X_j]$ gilt, falls $i \neq j$. Hier zeigt sich der Grund der Korrektur in s^2 durch Skalierung der Summe mit $n-1$, denn die korrigierte Sichtprobenvarianz ist unverzerrt.

iv. Aus iii) folgt i. Allg. nicht, dass s erwartungstreu für $(\mathbb{V}ar_\vartheta(X_1))^{1/2}$ ist. Denn für $\mathbb{V}ar_\vartheta(X_1) > 0$ gilt nach der Jensen-Ungleichung (vgl. Lemma 2.8)

$$\mathbb{E}_\vartheta[s(\mathfrak{X})] = \mathbb{E}_\vartheta\left[\sqrt{s^2(\mathfrak{X})}\right] < \sqrt{\mathbb{E}_\vartheta\left[s^2(\mathfrak{X})\right]} = \sqrt{\mathbb{V}ar_\vartheta(X_1)},$$

d. h., der Erwartungswert des Schätzers s ist kleiner als der zu schätzende Wert. Wir bemerken, dass $\sqrt{x}$ konkav ist, und wenden die Ungleichung auf $-\sqrt{x}$ konvex an.

Beim zweiten Gütekriterium geht es darum, dass sich ein Schätzer einer Kenngröße der Verteilung ‚annähert', wenn der Stichprobenumfang n groß wird.

Definition 5.3 (Konsistenz)

Es sei ein statistisches Modell gegeben durch einen Zufallsvektor $\mathfrak{X}_\infty = (X_1, X_2, \ldots)^t$ und eine Verteilungsfamilie $(\nu_\vartheta)_{\vartheta \in \Theta}$. Für $n = 1, 2 \ldots$ sei im Modell der Restriktion auf die ersten n Komponenten ($\mathfrak{X}_n = (X_1, \ldots, X_n)^t$) ein Schätzer S_n gegeben. Die Folge $(S_n)_{n=1,2,\ldots}$ heißt konsistent für eine abgeleitete Kenngröße der Verteilung $\tau(\vartheta)$, falls für alle $\vartheta \in \Theta$ gilt, dass

$$\forall\, \varepsilon > 0: \qquad \mathbb{P}_\vartheta\left(|S_n(\mathfrak{X}_n) - \tau(\vartheta)| > \varepsilon\right) \longrightarrow 0 \quad \text{für } n \to \infty.$$

Weiter heißt $(S_n)_{n=1,2,\ldots}$ *stark konsistent für* $\tau(\vartheta)$*, falls für alle* $\vartheta \in \Theta$ *gilt, dass*

$$\mathbb{P}_\vartheta\left(\lim_{n\to\infty} S_n(\mathfrak{X}_n) = \tau(\vartheta)\right) = 1.$$

Wir sprechen abkürzend von der (starken) Konsistenz des Schätzers $S_n(\mathfrak{X}_n)$. Ein für $\tau(\vartheta)$ konsistenter Schätzer $S_n(\mathfrak{X}_n)$ konvergiert also stochastisch gegen $\tau(\vartheta)$, falls ν_ϑ die wahre zugrunde liegende Verteilung ist. Analog bedeutet die starke Konsistenz eines Schätzers, dass er mit Wahrscheinlichkeit 1 gegen die zu schätzende Kenngröße $\tau(\vartheta)$ konvergiert. Starke Konsistenz impliziert Konsistenz, vgl. Lemma 2.14.

Im folgenden Beispiel wird die starke Konsistenz des Mittelwerts und der empirischen Varianz formuliert, welche unmittelbar aus dem Starken Gesetz der großen Zahlen folgen, vgl. Satz 2.10.

Beispiel 5.4 (Konsistente Schätzer)

i. Seien $X_1, X_2 \ldots$ *unabhängige und identisch verteilte Zufallsvariable, mit* $X_1 \sim \nu_\vartheta$*, und* ν_ϑ *ist Mitglied der Familie* $(\nu_\vartheta)_{\vartheta\in\Theta}$ *aller integrierbaren Verteilungen. Dann ist der Mittelwert* $\bar{X}_n$ *stark konsistent für den Erwartungswert* $\mathbb{E}_\vartheta[X_1]$.
Sind die Mitglieder von $(\nu_\vartheta)_{\vartheta\in\Theta}$ *sogar quadratintegrierbar, dann ist die empirische Varianz* s^2 *aus (3.3) stark konsistent für* $\sigma^2_\vartheta := \mathbb{V}ar_\vartheta(X_1)$*. Um dies zu sehen, zerlege man die empirischen Varianz* $s^2(\mathfrak{X})$ *erstens in eine Summe aus Quadraten* X_i^2 *und zweitens in eine Summe aus den Einzeltermen* X_i *und wende dann jeweils das Starke Gesetz der großen Zahlen an. Damit ist auch die empirische Standardabweichung* s *stark konsistent für* σ_ϑ.

ii. Seien $X_1, X_2, \ldots$ *unabhängige und identisch verteilte Zufallsvariable,* $X_1 \sim ber(p)$ *mit* $p \in \Theta := [0,1]$*. Dann ist die relative Häufigkeit* $\hat{p}$ *stark konsistent für* p *und die empirische Varianz* s^2 *stark konsistent für* $p(1-p)$*. Dies gilt analog zu i), denn* $\mathbb{E}_p[X_1] = p$ *und* $\mathbb{V}ar_p(X_1) = p(1-p)$.

5.2 Der mittlere quadratische Fehler

Erwartungstreue eines Schätzers $S(\mathfrak{X})$ sagt uns, dass er im *Mittel,* d. h. in Erwartung, den wahren Parameter trifft. Sie sagt aber nichts über die *Variabilität* des Schätzers aus. Schön wäre es, wenn der Schätzer neben der Erwartungstreue auch eine niedrige Varianz aufwiese – wenn also seine Schwankung gering wäre. Unter Umständen mag man sogar

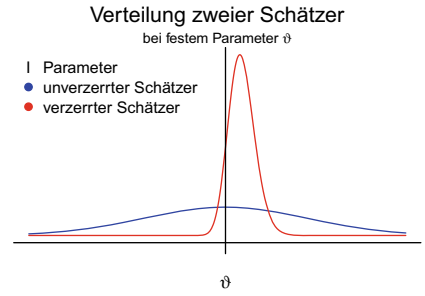

Abb. 5.1 Dichte zweier Schätzer – erwartungstreu (blau) vs. verzerrt (rot) – bei festem Parameter ϑ. Der erwartungstreue Schätzer weist eine vergleichsweise große Varianz auf

weiter gehen und eine Verzerrung in Kauf nehmen, d. h. die Erwartungstreue systematisch aufgeben, um die Variabilität des Schätzers zu reduzieren. Denn bei riesiger Varianz ist Erwartungstreue vergleichsweise wenig wert (vgl. Abb. 5.1). Ein Maß, das simultan sowohl die Verzerrung als auch die Schwankung eines Schätzers bewertet, ist der *mittlere quadratische Fehler.*

Definition 5.5 (Mittlerer quadratischer Fehler)
Es sei ein statistisches Modell gegeben durch einen Zufallsvektor $\mathfrak{X} = (X_1, \ldots, X_n)^t$ *und eine Verteilungsfamilie* $(\nu_\vartheta)_{\vartheta \in \Theta}$. *Weiter sei* S *ein unter allen* ν_ϑ *quadratintegrierbarer Schätzer für die abgeleitete Kenngröße* $\tau(\vartheta)$. *Für alle* $\vartheta \in \Theta$ *ist der mittlere quadratische Fehler (engl. mean squared error, kurz MSE), von* S *bezüglich* $\tau(\vartheta)$ *definiert durch*

$$\mathbb{MSE}_\vartheta(S(\mathfrak{X}), \tau(\vartheta)) := \mathbb{E}_\vartheta\left[(S(\mathfrak{X}) - \tau(\vartheta))^2\right].$$

Interpretation: Der MSE beschreibt die erwartete quadratische Abweichung des Schätzers $S(\mathfrak{X})$ zum abgeleiteten Parameter $\tau(\vartheta)$ unter der Annahme, dass ν_ϑ die wahre zugrunde liegende Verteilung ist. Schätzer mit kleinerem MSE zeigen in Erwartung eine kleinere quadratische Abweichung von dem zu schätzenden Wert und sind daher im Sinne des MSE zu bevorzugen.

Aus der Definition folgt direkt: Ist ein Schätzer S erwartungstreu für $\tau(\vartheta)$ (d. h. $\mathbb{E}_\vartheta[S(\mathfrak{X})] = \tau(\vartheta)$ für alle $\vartheta \in \Theta$), so entspricht der MSE gerade der Varianz des Schätzers, $\mathbb{MSE}_\vartheta(S(\mathfrak{X}), \tau(\vartheta)) = \mathbb{V}ar_\vartheta(S(\mathfrak{X}))$, für alle $\vartheta \in \Theta$. Ist der Schätzer nicht erwartungstreu, so lässt sich der MSE zerlegen in die Varianz des Schätzers plus ein sogenanntes *Verzerrungsquadrat* (siehe Lemma 5.7). Wir benötigen

Definition 5.6 (Bias)
Es sei ein statistisches Modell gegeben durch einen Zufallsvektor $\mathfrak{X} = (X_1, \ldots, X_n)^t$ und eine Verteilungsfamilie $(\nu_\vartheta)_{\vartheta \in \Theta}$. Sei S ein unter allen ν_ϑ integrierbarer Schätzer für eine abgeleitete Kenngröße der Verteilung $\tau(\vartheta)$. Für alle $\vartheta \in \Theta$ ist die Verzerrung (engl. bias, schreibe $\mathbb{B}ias$), von S bezüglich $\tau(\vartheta)$ definiert durch

$$\mathbb{B}ias_\vartheta(S(\mathfrak{X}), \tau(\vartheta)) := \mathbb{E}_\vartheta[S(\mathfrak{X})] - \tau(\vartheta).$$

Ist S erwartungstreu für $\tau(\vartheta)$, so ist $\mathbb{B}ias_\vartheta(S(\mathfrak{X}), \tau(\vartheta)) = 0$, für alle $\vartheta \in \Theta$.

Lemma 5.7 (Zerlegung des mittleren quadratischen Fehlers)
Es sei ein statistisches Modell gegeben durch einen Zufallsvektor $\mathfrak{X} = (X_1, \ldots, X_n)^t$ und eine Verteilungsfamilie $(\nu_\vartheta)_{\vartheta \in \Theta}$. Für den MSE eines Schätzers S bezüglich einer abgeleiteten Kenngröße $\tau(\vartheta)$ gilt für alle $\vartheta \in \Theta$

$$\mathbb{M}\mathrm{SE}_\vartheta(S(\mathfrak{X}), \tau(\vartheta)) = \mathbb{V}ar_\vartheta(S(\mathfrak{X})) + \mathbb{B}ias^2_\vartheta(S(\mathfrak{X}), \tau(\vartheta)).$$

Interpretation: Der mittlere quadratische Fehler beurteilt die Varianz und den Bias eines Schätzers simultan. Der Fehler ist klein, wenn der Schätzer wenig schwankt und im Mittel nahe beim wahren abgeleiteten Parameter liegt.

Beweis Wir berechnen für alle $\vartheta \in \Theta$

$$\begin{aligned}
\mathbb{M}\mathrm{SE}_\vartheta(S(\mathfrak{X}), \tau(\vartheta)) &= \mathbb{E}_\vartheta\left[(S(\mathfrak{X}) - \tau(\vartheta))^2\right] \\
&\overset{(*)}{=} \mathbb{E}_\vartheta\left[([S(\mathfrak{X}) - \mathbb{E}_\vartheta[S(\mathfrak{X})]] + [\mathbb{E}_\vartheta[S(\mathfrak{X})] - \tau(\vartheta)])^2\right] \\
&\overset{(**)}{=} \mathbb{E}_\vartheta[(S(\mathfrak{X}) - \mathbb{E}_\vartheta[S(\mathfrak{X})])^2] + 2(\mathbb{E}_\vartheta[S(\mathfrak{X})] \\
&\quad - \tau(\vartheta))\mathbb{E}_\vartheta[(S(\mathfrak{X}) - \mathbb{E}_\vartheta[S(\mathfrak{X})])] + (\mathbb{E}_\vartheta[S(\mathfrak{X})] - \tau(\vartheta))^2 \\
&= \mathbb{V}ar_\vartheta(S(\mathfrak{X})) + \mathbb{B}ias^2_\vartheta(S(\mathfrak{X}), \tau(\vartheta)).
\end{aligned}$$

In $(*)$ haben wir lediglich ‚die Null addiert', und in Zeile $(**)$ verschwindet der mittlere Summand, da $\mathbb{E}_\vartheta[(S(\mathfrak{X}) - \mathbb{E}_\vartheta[S(\mathfrak{X})])] = 0$.

Beispiel 5.8 (Mittlerer quadratischer Fehler)
Seien $X_1, \ldots, X_n$ unabhängige und identisch verteilte Zufallsvariable mit $X_1 \sim \nu_\vartheta$, und ν_ϑ sei Mitglied der Familie $(\nu_\vartheta)_{\vartheta \in \Theta}$ aller quadratintegrierbaren Verteilungen. Wir interpretieren den Mittelwert $\bar{X}_n$ als einen Schätzer für den Erwartungswert von X_1. Dann ist der MSE von $\bar{X}_n$ bzgl. $\mu_\vartheta := \mathbb{E}_\vartheta[X_1]$

$$\mathbb{MSE}_\vartheta(\bar{X}_n, \mu_\vartheta) = \mathbb{E}_\vartheta[(\bar{X}_n - \mu_\vartheta)^2] \overset{(*)}{=} \mathbb{V}ar_\vartheta(\bar{X}_n) = \frac{1}{n}\mathbb{V}ar_\vartheta(X_1),$$

wobei $(*)$ aufgrund der Erwartungstreue von $\bar{X}_n$ bezüglich μ_ϑ gilt.

Beispiel 5.9 (MSE bei der uniformen Verteilung)
Seien $X_1, \ldots, X_n$ unabhängige und identisch verteilte Zufallsvariable und $X_1 \sim U(0, b]$ mit $b \in \Theta = (0, \infty)$. Wir betrachten zwei Schätzer $\hat{b}_1$ und $\hat{b}_2$ für b.

i. Es sei $\hat{b}_1(\mathfrak{X}) := 2\bar{X}_n$. Dieser Schätzer ist erwartungstreu, denn

$$\mathbb{E}_b\left[\hat{b}_1(\mathfrak{X})\right] = 2\mathbb{E}_b[X_1] = 2 \cdot \frac{b}{2} = b.$$

Daher erhalten wir für den MSE von $\hat{b}_1$ bzgl. b (vgl. Beispiel 2.6iv)

$$\begin{aligned}\mathbb{MSE}_b\left(\hat{b}_1(\mathfrak{X}), b\right) &= \mathbb{E}_\ell\left[\left(\hat{b}_1(\mathfrak{X}) - b\right)^2\right] \\ &= \mathbb{V}ar_b(2\bar{X}_n) = \frac{4}{n}\mathbb{V}ar_b(X_1) = \frac{4}{n}\frac{b^2}{12} = \frac{1}{3n}b^2.\end{aligned}$$

ii. Der zweite Schätzer $\hat{b}_2$ sei gegeben durch

$$\hat{b}_2(\mathfrak{X}) := \frac{n+1}{n}\max\{X_1, \ldots, X_n\}.$$

Er ist ebenfalls erwartungstreu für b: Die Verteilungsfunktion von $X_{(n)} = \max\{X_1, \ldots, X_n\}$ ist gegeben durch

$$F_b(x) = \mathbb{P}_b(X_{(n)} \leq x) = \prod_{i=1}^{n}\mathbb{P}_b(X_i \leq x) = \left(\frac{x}{b}\right)^n,$$

für $x \in (0, b)$; sowie $F_b(x) = 1$ für $x \geq b$. Auf $(0, b)$ ist F_b differenzierbar, und es gilt $F_b'(x) = (n/b^n)x^{n-1}$. Daher ist die Dichte von $X_{(n)}$ gegeben durch $f_b(x) = (n/b^n)x^{n-1}\mathbb{1}_{(0,b]}(x)$. Daraus folgt

$$\mathbb{E}_b[X_{(n)}] = \frac{n}{b^n}\int_0^b x^n dx = \frac{n}{n+1}b,$$

und damit $\mathbb{E}_b[\hat{b}_2(\mathfrak{X})] = b$. *Intuition: Man betrachte den maximalen Wert der* X_i. *Das wahre* b *wird immer etwas größer sein. In Erwartung wird das Intervall von den* n *Beobachtungen in* $(n+1)$ *gleich große Intervalle geteilt, damit liegt der größte von* n *Werten in Erwartung bei* $(n/(n+1)) \cdot b$.
Für den MSE berechnen wir zunächst

$$\mathbb{E}_b\left[X_{(n)}^2\right] = \int_0^b x^2 \frac{n}{b^n} x^{n-1} dx = \frac{n}{n+2} b^2.$$

Aufgrund der Erwartungstreue von $\hat{b}_2$ *für* b *ergibt sich schließlich*

$$\begin{aligned}\mathbb{MSE}_b\left(\hat{b}_2(\mathfrak{X}), b\right) = \mathbb{V}ar_b\left(\frac{n+1}{n} X_{(n)}\right) &= \frac{(n+1)^2}{n^2}\mathbb{E}_b[X_{(n)}^2] - \frac{(n+1)^2}{n^2}\mathbb{E}_b[X_{(n)}]^2 \\ &= \frac{(n+1)^2}{n^2}\frac{n}{n+2} b^2 - b^2 \\ &= \frac{1}{n(n+2)} b^2.\end{aligned}$$

Wir bemerken, dass $\mathbb{MSE}_b(\hat{b}_2(\mathfrak{X}), b) \leq \mathbb{MSE}_b(\hat{b}_1(\mathfrak{X}), b)$ *für alle* n, *mit Gleichheit nur bei* $n = 1$. *Weiter gilt sogar, dass* $\mathbb{MSE}_b(\hat{b}_1(\mathfrak{X}), b) = O(1/n)$ *und* $\mathbb{MSE}_b(\hat{b}_2(\mathfrak{X}), b) = O(1/n^2)$, *für* $n \to \infty$. *Bedeutung: Der* $\mathbb{MSE}_b(\hat{b}_2(\mathfrak{X}), b)$ *fällt bei wachsendem Stichprobenumfang schneller als der* $\mathbb{MSE}_b(\hat{b}_1(\mathfrak{X}), b)$. *Im Sinne des MSE ist der Schätzer* $\hat{b}_2$ *dem Schätzer* $\hat{b}_1$ *– insbesondere für ‚große' n – vorzuziehen.*

Das nächste Lemma besagt, dass ein Schätzer S sogar konsistent für $\tau(\vartheta)$ ist, wenn sein MSE bezüglich $\tau(\vartheta)$ bei wachsender Stichprobengröße verschwindet.

Lemma 5.10 (Verschwinden des MSE impliziert Konsistenz)
Es sei ein statistisches Modell gegeben durch einen Zufallsvektor $\mathfrak{X}_\infty = (X_1, X_2, \ldots)^t$ *und eine Verteilungsfamilie* $(\nu_\vartheta)_{\vartheta \in \Theta}$. *Für* $n = 1, 2 \ldots$ *sei im Modell der Restriktion auf die ersten* n *Komponenten* $(\mathfrak{X}_n = (X_1, \ldots, X_n)^t)$ *ein Schätzer* S_n *gegeben.*
Es bezeichne $\tau(\vartheta)$ *eine abgeleitete Kenngröße der Verteilung. Gilt für alle* $\vartheta \in \Theta$, *dass*

$$\mathbb{MSE}_\vartheta(S_n(\mathfrak{X}_n), \tau(\vartheta)) \longrightarrow 0$$

für $n \to \infty$, *so ist die Folge* $(S_n)_{n=1,2,\ldots}$ *konsistent für* $\tau(\vartheta)$.

Beweis Die Aussage folgt aus Lemma 2.14, da $\mathscr{L}^2$-Konvergenz die stochastische Konvergenz impliziert.

5.3 Suffizienz und Verkleinerung des MSE

Ziel dieses Kapitels ist es, in einer Situation, in der ein Schätzer bereits vorliegt, diesen gewissermaßen zu verbessern, genauer würden wir gerne den MSE verkleinern. Eine Möglichkeit läuft über die Betrachtung sogenannter suffizienter Statistiken. Dazu benötigen wir den Begriff der *Suffizienz,* eingeführt von Fisher (1922), und die bedingte Verteilung sowie den bedingten Erwartungswert.

Wir beschränken uns in diesem Abschnitt auf diskrete Modelle, d. h., dass der Bildraum des Zufallsvektors diskret ist. Eine Kandidatenverteilung ν_ϑ kann dann durch ihre Gewichte g_ϑ beschrieben werden. Die Resultate lassen sich aber verallgemeinern.

5.3.1 Suffizienz

Wir motivieren zunächst den Begriff der Suffizienz an einem Beispiel. Es sei das Bernoulli-Modell aus Beispiel 4.1 gegeben durch $X_1, \ldots, X_n$ unabhängige und identisch verteilte Zufallsvariable mit $X_1 \sim ber(p)$ und $p \in \Theta := [0, 1]$. Dann enthalten die Beobachtungen $\mathbf{x} = (x_1, \ldots, x_n)^t \in \{0, 1\}^n$ als Realisierung von $\mathfrak{X}$ gewisse Information über den unbekannten Parameter. Wir können etwa aus $\mathbf{x} = (0, 1, 0, \ldots)^t$ schließen, dass $p \notin \{0, 1\}$. Andererseits ist eine Statistik T eine Funktion der Beobachtungen $\mathbf{x}$ und damit i. Allg. eine Reduktion der Information von $\mathbf{x}$ über p. Betrachte man etwa die Statistik $T(\mathbf{x}) := x_1$, so könnten wir anhand der Auswertung $T(\mathbf{x}) = 0$ nur schließen, dass $p \neq 1$. Wir haben also weniger Information über p als anhand der Beobachtungen $\mathbf{x}$ selbst, denn wir werfen schließlich alle Beobachtungen $x_2, \ldots, x_n$ einfach weg. Interessant sind für uns nun jene Statistiken S, für welche die Auswertung $S(\mathbf{x})$ in gewisser Hinsicht nicht weniger Information über den zu schätzenden Parameter enthält als die gesamten Beobachtungen $\mathbf{x}$.

Dies führt zum Begriff der Suffizienz, und dafür benötigen wir die Definition der bedingten Verteilung. Zur Erinnerung: Für eine diskrete, integrierbare Zufallsvariable X und ein Ereignis A mit $\mathbb{P}(A) > 0$ ist die *bedingte Verteilung ν_A von X gegeben A* definiert durch

$$\nu_A(\cdot) := \mathbb{P}(X \in \cdot \mid A) := \frac{\mathbb{P}(\{X \in \cdot\} \cap A)}{\mathbb{P}(A)},$$

welche eine Verteilung auf dem Bildraum Γ von X beschreibt.

Definition 5.11 (Suffiziente Statistik)
Es sei ein diskretes Modell gegeben durch einen Zufallsvektor $\mathfrak{X} = (X_1, \ldots, X_n)^t$ mit Bildraum $\mathcal{X}$ und eine Familie von Gewichten $(g_\vartheta)_{\vartheta \in \Theta}$.
Eine Statistik S heißt suffizient für ϑ, falls für alle $\vartheta \in \Theta$ die bedingte Verteilung von $\mathfrak{X}$ gegeben $\{S(\mathfrak{X}) = S(\mathbf{x})\}$ wohldefiniert ist und nicht von ϑ abhängt, für alle $\mathbf{x} \in \mathcal{X}$.

Interpretation: Sämtliche Information von $\mathfrak{X}$ über ϑ steckt auch in $S(\mathfrak{X})$. Insbesondere sehen wir sofort, dass die Beobachtungen $\mathfrak{X}$ selbst suffizient für ϑ sind. Denn offenbar gilt $\mathbb{P}_\vartheta(\mathfrak{X} = \mathbf{x}|\mathfrak{X} = \mathbf{y}) = \mathbb{1}_{\{\mathbf{x}\}}(\mathbf{y})$.

Wir bemerken, dass die Wohldefiniertheit in der Definition bedeutet, dass sämtliche bedingten Verteilungen überhaupt erst existieren müssen, damit der Begriff der Suffizienz gegeben ist.

Beispiel 5.12 (Suffiziente Statistik im Bernoulli-Modell)
Wir betrachten das Bernoulli-Modell aus Beispiel 4.1, schränken aber den Parameterraum Θ aus Gründen der Wohldefiniertheit auf das offene Intervall $(0, 1)$ ein, siehe auch den Kommentar am Ende dieses Beispiels. Dann ist $S(\mathbf{x}) := \sum_{i=1}^n x_i$ suffizient für p, denn für $\mathbf{y} = (y_1, \ldots, y_n)^t \in \{0, 1\}^n$ ist das Gewicht

$$\begin{aligned}
\mathbb{P}_p(\mathfrak{X} = \mathbf{y}|S(\mathfrak{X}) = S(\mathbf{x})) &= \frac{\mathbb{P}_p(\{\mathfrak{X} = \mathbf{y}\} \cap \{S(\mathfrak{X}) = S(\mathbf{x})\})}{\mathbb{P}_p(S(\mathfrak{X}) = S(\mathbf{x}))} \\
&\overset{(*)}{=} \frac{\mathbb{P}_p((X_1, \ldots, X_n)^t = (y_1, \ldots, y_n)^t)}{\mathbb{P}_p(S(\mathfrak{X}) = S(\mathbf{x}))} \cdot \mathbb{1}_{\{S(\mathbf{y})\}}(S(\mathbf{x})) \\
&= \frac{p^{S(\mathbf{x})}(1-p)^{n-S(\mathbf{x})}}{\binom{n}{S(\mathbf{x})} p^{S(\mathbf{x})}(1-p)^{n-S(\mathbf{x})}} \cdot \mathbb{1}_{\{S(\mathbf{y})\}}(S(\mathbf{x})) \\
&= \binom{n}{S(\mathbf{x})}^{-1} \cdot \mathbb{1}_{\{S(\mathbf{y})\}}(S(\mathbf{x})),
\end{aligned}$$

nicht mehr abhängig von p. Intuitiv mag das klar sein: Allein die Anzahl der Erfolge $S(\mathbf{x})$ bei n Münzwürfen enthält Information über die Erfolgswahrscheinlichkeit p. Die Zeitpunkte, an denen die Erfolge auftreten, spielen keine Rolle. Insbesondere haben wir in $(*)$ ausgenutzt, dass $\{\mathfrak{X} = \mathbf{y}\} \subseteq \{S(\mathfrak{X}) = S(\mathbf{x})\}$, wenn $S(\mathbf{x}) = S(\mathbf{y})$, sowie $\{\mathfrak{X} = \mathbf{y}\} \cap \{S(\mathfrak{X}) = S(\mathbf{x})\} = \emptyset$, falls $S(\mathbf{x}) \neq S(\mathbf{y})$. Analog zeigt man, dass auch der Mittelwert $\bar{x}_n$ suffizient für p ist.

Die Ränder $p \in \{0, 1\}$ wurden ausgeschlossen, damit alle betrachteten bedingten Verteilungen wohldefiniert sind. Beispielsweise hätte für $p = 1$ die Beobachtung $\mathbf{x} = (0, \ldots, 0)^t$ Gewicht null, sodass auch $\mathbb{P}_1(S(\mathfrak{X}) = 0) = 0$. Davon abgesehen gilt bei Hinzunahme der Werte $p \in \{0, 1\}$ analog, dass $\mathbb{P}_p(\mathfrak{X} = \mathbf{y}|S(\mathfrak{X}) = S(\mathbf{x}))$ unter allen p, unter denen es wohldefiniert ist, nicht von p abhängt.

In Beispiel 5.12 sind suffiziente Statistiken ‚vom Himmel gefallen', und wir haben dann ihre Suffizienz für einen bestimmen Parameter nachgewiesen. Suffiziente Statistiken lassen sich anhand des Satzes von Neyman und Fisher erschließen.

Satz 5.13 (Satz von Neyman und Fisher)
Es sei ein diskretes Modell gegeben durch einen Zufallsvektor $\mathfrak{X} = (X_1, \ldots, X_n)^t$ mit Bildraum $\mathcal{X}$ und einer Familie von Gewichten $(g_\vartheta)_{\vartheta \in \Theta}$. Dann sind folgende Aussagen äquivalent

1. *S ist eine suffiziente Statistik für ϑ.*
2. *Unter allen $\vartheta \in \Theta$ faktorisieren die Gewichte*

$$g_\vartheta(\mathbf{x}) = r(S(\mathbf{x}), \vartheta) \cdot h(\mathbf{x})$$

für alle $\mathbf{x} \in \mathcal{X}$, wobei der erste Faktor r eine Funktion der suffizienten Statistik $S(\mathbf{x})$ und von ϑ ist, und der zweite Faktor h nicht von ϑ abhängt.

Beweis Seien $\mathbf{x}$ und $\mathbf{y}$ aus $\mathcal{X}$.
1. $\Rightarrow$ 2.:

$$g_\vartheta(\mathbf{y}) = \mathbb{P}_\vartheta(\{\mathfrak{X} = \mathbf{y}\} \cap \overbrace{\{S(\mathfrak{X}) = S(\mathbf{y})\}}^{\supset \{\mathfrak{X} = \mathbf{y}\}}) = \underbrace{\mathbb{P}_\vartheta(S(\mathfrak{X}) = S(\mathbf{y}))}_{=: r_\vartheta(S(\mathbf{y}))} \cdot \underbrace{P_\vartheta(\mathfrak{X} = \mathbf{y} | S(\mathfrak{X}) = S(\mathbf{y}))}_{=: h(\mathbf{y})},$$

wobei h aufgrund der Suffizienz von S nicht von ϑ abhängt.
2. $\Rightarrow$ 1.: Es ist zu zeigen, dass

$$\mathbb{P}_\vartheta(\mathfrak{X} = \mathbf{y} \mid S(\mathfrak{X}) = S(\mathbf{x})) = \frac{\mathbb{P}_\vartheta(\mathfrak{X} = \mathbf{y})}{\mathbb{P}_\vartheta(S(\mathfrak{X}) = S(\mathbf{x}))} \mathbb{1}_{\{S(\mathbf{y})\}}(S(\mathbf{x})) \tag{5.1}$$

nicht von ϑ abhängt. Für den Zähler folgt nach Voraussetzung

$$\mathbb{P}_\vartheta(\mathfrak{X} = \mathbf{y}) = g_\vartheta(\mathbf{y}) = r_\vartheta(S(\mathbf{y})) \cdot h(\mathbf{y}).$$

Für den Nenner finden wir durch Übergang zum Urbildraum $\mathcal{X}$

$$\mathbb{P}_\vartheta(S(\mathfrak{X}) = S(\mathbf{x})) = \sum_{\{\mathbf{z} \in \mathcal{X} | S(\mathbf{x}) = S(\mathbf{z})\}} \overbrace{\mathbb{P}_\vartheta(\mathfrak{X} = \mathbf{z})}^{r_\vartheta(S(\mathbf{z})) \cdot h(\mathbf{z})} \overset{(*)}{=} r_\vartheta(S(\mathbf{x})) \sum_{\{\mathbf{z} \in \mathcal{X} | S(\mathbf{x}) = S(\mathbf{z})\}} h(\mathbf{z}),$$

wobei wir in $(*)$ ausgenutzt haben, dass wir uns auf die Menge $\{\mathbf{z} \in \mathcal{X} | S(\mathbf{x}) = S(\mathbf{z})\}$ einschränken, sodass der Faktor $r_\vartheta(S(\mathbf{z})) = r_\vartheta(S(\mathbf{x}))$ aus der Summe gezogen werden darf. Da nun aber in (5.1) noch der Indikator $\mathbb{1}_{\{S(\mathbf{y})\}}(S(\mathbf{x}))$ auftaucht, fällt dort auch r_ϑ weg, genauer

$$\mathbb{P}_\vartheta(\mathfrak{X} = \mathbf{y} \mid S(\mathfrak{X}) = S(\mathbf{x})) = \frac{h(\mathbf{y})}{\sum_{\{\mathbf{z}\in\mathcal{X}\mid S(\mathbf{y})=S(\mathbf{z})\}} h(\mathbf{z})} \mathbb{1}_{\{S(\mathbf{y})\}}(S(\mathbf{x})),$$

was nicht von ϑ abhängt.

Beispiel 5.14 (Anwendung des Lemmas von Neyman und Fisher)
Es sei das vorherige Bernoullimodell aus Beispiel 5.12 gegeben. Wir wissen schon, dass $S(\mathbf{x}) = \sum_{i=1}^n x_i$ *eine suffiziente Statistik für* p *ist. Dies folgt auch leicht aus dem Lemma von Neyman und Fisher: Sei* $\mathbf{x} = (x_1, \ldots, x_n)^t \in \{0,1\}^n$*, dann ist*

$$g_\vartheta(\mathbf{x}) = \prod_{i=1}^n p^{x_i}(1-p)^{1-x_i} = p^{S(\mathbf{x})}(1-p)^{n-S(\mathbf{x})} =: r_p(S(\mathbf{x}))$$

und $h(\mathbf{x}) := 1$*. Wir sehen auch direkt, dass der Mittelwert* $\bar{x}_n$ *suffizient für* p *ist, wenn wir im Exponenten mit* n *erweitern.*

Suffiziente Statistiken können zur Konstruktion von Schätzern mit kleinerem MSE dienen, wie wir in Abschn. 5.3.2 sehen.

5.3.2 Der bedingte Erwartungswert und der Satz von Rao-Blackwell

Wir erinnern zunächst an den Begriff des bedingten Erwartungswerts. Sei X diskrete, integrierbare Zufallsvariable mit Bildraum Γ und A ein Ereignis mit $\mathbb{P}(A) > 0$. Der *bedingte Erwartungswert von* X *gegeben* A ist definiert als

$$\mathbb{E}[X|A] := \frac{\mathbb{E}[\mathbb{1}_A X]}{\mathbb{P}(A)} = \frac{\sum_{x\in\Gamma} x\mathbb{P}(\{X = x\} \cap A)}{\mathbb{P}(A)} = \sum_{x\in\Gamma} x\mathbb{P}(X = x|A).$$

Dies ist eine deterministische Größe. Interpretation: $\mathbb{E}[X|A]$ beschreibt unsere Erwartung von X, wenn wir wissen, dass A eingetreten ist. Analog schreibt sich für eine diskrete Zufallsvariable Z mit positiven Gewichten und eine Funktion $h : \Gamma \to \mathbb{R}$ mit $h(X)$ integrierbar

$$\mathbb{E}[h(X)|Z = z] = \frac{\mathbb{E}[\mathbb{1}_{\{z\}}(Z)h(X)]}{\mathbb{P}(Z = z)} = \sum_{x\in\Gamma} h(x)\mathbb{P}(X = x|Z = z). \tag{5.2}$$

Andererseits ist der *bedingte Erwartungswert von* $h(X)$ *gegeben* Z eine Zufallsvariable. Es bezeichne Δ den Bildraum der Zufallsvariablen Z. Dann definieren wir

$$\mathbb{E}[h(X)|Z] := \sum_{z\in\Delta} \mathbb{1}_{\{z\}}(Z)\mathbb{E}[h(X)|Z = z].$$

Häufig denken wir an $h = id$, d. h. an den bedingten Erwartungswert von X gegeben Z.

Beispiel 5.15 (Bedingter Erwartungswert)
Seien $Z \sim unif\{1, \ldots, 10\}$ und $X \sim b(Z, 1/2)$. Es beschreibt also X die Anzahl der Erfolge bei einer Münzwurffolge mit zufälliger Länge Z. Dann berechnet sich $\mathbb{E}[X|Z = z] = z/2$, und für den bedingten Erwartungswert von X gegeben Z folgt $\mathbb{E}[X|Z] = Z/2$.

Wir führen drei einfache Eigenschaften des bedingten Erwartungswertes an, die wir später nutzen werden.

Lemma 5.16 (Eigenschaften des bedingten Erwartungswertes)
Seien $a, b \in \mathbb{R}$ und X, Y, Z diskrete Zufallsvariable, mit X, Y integrierbar und Z habe positive Gewichte. Dann gilt:

1. *Linearität: $\mathbb{E}[aX + bY|Z] = a\mathbb{E}[X|Z] + b\mathbb{E}[Y|Z]$.*
2. *Totaler Erwartungswert: $\mathbb{E}[\mathbb{E}[X|Z]] = \mathbb{E}[X]$.*
3. *Verschiebungssatz der bedingten Varianz: Ist $\mathbb{E}[|X|^2] < \infty$, so gilt*

$$\mathbb{V}ar(X|Z) := \mathbb{E}[(X - \mathbb{E}[X|Z])^2|Z] = \mathbb{E}[X^2|Z] - (\mathbb{E}[X|Z])^2 \geq 0.$$

Beweis Die Aussagen können durch Rückführung auf die Definition gefolgert werden. Bezeichnet Δ den Bildraum von Z, so folgt beispielsweise 2. via

$$\mathbb{E}[\mathbb{E}[X|Z]] = \mathbb{E}\left[\sum_{z\in\Delta} \mathbb{1}_{\{Z=z\}}\mathbb{E}[X|Z = z]\right] \overset{(*)}{=} \sum_{z\in\Delta} \mathbb{P}(Z = z)\mathbb{E}[X|Z = z]$$
$$\overset{(5.2)}{=} \sum_{z\in\Delta} \mathbb{E}[\mathbb{1}_{\{z\}}(Z)X] \overset{(**)}{=} \mathbb{E}[X].$$

In $(*)$ haben wir die Summe sowie auch $\mathbb{E}[X|Z = z]$ (konstant) vor den Erwartungswert gezogen, und in $(**)$ wurde die Summe in den Erwartungswert gezogen.

Hilfslemma 5.17 (Bedingter Erwartungswert gegeben eine suffiziente Statistik)
Es sei ein diskretes Modell gegeben durch einen Zufallsvektor $\mathfrak{X} = (X_1, \ldots, X_n)^t$ mit Bildraum $\mathcal{X}$ und Gewichten $(g_\vartheta)_{\vartheta\in\Theta}$. Sei weiter S eine suffiziente Statistik für ϑ und T eine unter allen Gewichten g_ϑ integrierbare Statistik. Dann hängt $\mathbb{E}_\vartheta[T(\mathfrak{X}) \mid S(\mathfrak{X}) = s(\mathbf{x})]$ für sämtliche $\mathbf{x} \in \mathcal{X}$ und damit auch $\mathbb{E}_\vartheta[T(\mathfrak{X})|S(\mathfrak{X})]$ nicht von ϑ ab.

Beweis Nach Definition des bedingten Erwartungswertes ist

$$\mathbb{E}_\vartheta[T(\mathfrak{X})|S(\mathfrak{X}) = s(\mathbf{x})] = \sum_{\mathbf{y}\in\mathcal{X}} T(\mathbf{y})\mathbb{P}_\vartheta(\mathfrak{X} = \mathbf{y} \mid S(\mathfrak{X}) = S(\mathbf{x})),$$

was aufgrund der Suffizienz von S nicht von ϑ abhängt. Dies pflanzt sich offenbar auch auf $\mathbb{E}_\vartheta[T(\mathfrak{X})|S(\mathfrak{X})] = \sum_{\mathbf{x}\in\mathcal{X}} \mathbb{1}_{\{S(\mathbf{x})\}}(S(\mathfrak{X}))\mathbb{E}_\vartheta[T(\mathfrak{X})|S(\mathfrak{X}) = S(\mathbf{x})]$ fort.

Die Bedeutung der Unabhängigkeit von dem Parameter ϑ liegt darin, dass die Abbildung $T^* : \mathcal{X} \to \mathbb{R}$ vermöge

$$T^*(\mathbf{x}) := \mathbb{E}_\vartheta[T(\mathfrak{X}) \mid S(\mathfrak{X}) = S(\mathbf{x})]$$

eine Statistik beschreibt, d. h. eine Abbildung, welche nur von den Beobachtungen und nicht vom unbekannten Parameter ϑ abhängt. In diesem Fall unterdrücken wir den Parameter ϑ und schreiben etwa $T^*(\mathbf{x}) = \mathbb{E}[T(\mathfrak{X}) \mid S(\mathfrak{X}) = S(\mathbf{x})]$.

Mithilfe des Begriffs der suffizienten Statistik formulieren wir nun den

Satz 5.18 (Satz von Rao und Blackwell)
Es sei ein diskretes Modell gegeben durch einen Zufallsvektor $\mathfrak{X} = (X_1, \ldots, X_n)^t$ und eine Familie von Gewichten $(g_\vartheta)_{\vartheta\in\Theta}$.
Sei zudem $\hat{\vartheta}$ ein unter allen Gewichten g_ϑ quadratintegrierbarer Schätzer für einen abgeleiteten Parameter $\tau(\vartheta)$, und es sei S eine suffiziente Statistik für ϑ. Dann heißt ein Schätzer $\hat{\vartheta}^*$ gegeben durch

$$\hat{\vartheta}^*(\mathbf{x}) := \mathbb{E}\left[\hat{\vartheta}(\mathfrak{X}) \mid S(\mathfrak{X}) = S(\mathbf{x})\right]$$

die Rao-Blackwellisierung von $\hat{\vartheta}$ gegeben S, und es gilt für alle $\vartheta \in \Theta$, dass

1. $\mathbb{E}_\vartheta[\hat{\vartheta}^*(\mathfrak{X})] = \mathbb{E}_\vartheta[\hat{\vartheta}(\mathfrak{X})]$,
2. $\mathbb{MSE}_\vartheta(\hat{\vartheta}^*(\mathfrak{X}), \tau(\vartheta)) \le \mathbb{MSE}_\vartheta(\hat{\vartheta}(\mathfrak{X}), \tau(\vartheta))$.

Beweis Die Aussagen folgen direkt aus den Eigenschaften des bedingten Erwartungswertes. Der totale Erwartungswert liefert

$$\mathbb{E}_\vartheta\left[\hat{\vartheta}^*(\mathfrak{X})\right] = \mathbb{E}_\vartheta\left[\mathbb{E}_\vartheta[\hat{\vartheta}(\mathfrak{X})|S(\mathfrak{X})]\right] = \mathbb{E}_\vartheta\left[\hat{\vartheta}(\mathfrak{X})\right],$$

und die zweite Aussage folgt via

$$\begin{aligned}\mathbb{E}_\vartheta\left[(\hat{\vartheta}^*(\mathfrak{X})-\tau(\vartheta))^2\right] &= \mathbb{E}_\vartheta\left[(\mathbb{E}_\vartheta[\hat{\vartheta}(\mathfrak{X})-\tau(\vartheta)|S(\mathfrak{X})])^2\right]\\ &\le \mathbb{E}_\vartheta\left[\mathbb{E}_\vartheta[(\hat{\vartheta}(\mathfrak{X})-\tau(\vartheta))^2|S(\mathfrak{X})]\right]\\ &= \mathbb{E}_\vartheta\left[(\hat{\vartheta}(\mathfrak{X})-\tau(\vartheta))^2\right],\end{aligned}$$

wobei in der Ungleichung der Verschiebungssatz der bedingten Varianz angewendet wurde. In der letzten Gleichung wurde dann nochmal die Turmeigenschaft genutzt.

Die Bedeutung der Rao-Blackwellisierung liegt darin, dass sie den gleichen $\mathbb{B}ias$ wie der Ausgangsschätzer $\hat{\vartheta}$ besitzt und zusätzlich einen MSE, der höchstens so groß ist wie der von $\hat{\vartheta}$. Eine Aussage über Optimalität, d. h. Minimierung des MSE, macht der Satz von Rao-Blackwell aber leider nicht. Der interessierte Leser möge dafür etwa den Satz von Lehmann-Scheffé studieren, siehe zum Beispiel Lehmann und Casella (2006). Zum Satz von Rao und Blackwell seien hier auch Blackwell (1947) und Rao (1992) erwähnt.

Beispiel 5.19 (Rao-Blackwellisierung)
Im Bernoullimodell betrachten wir den Schätzer $\hat{\vartheta}(\mathbf{x}) := x_1 \cdot x_2$ *für* $\tau(p) = p^2$. *Nach Beispiel 5.12 ist* $S(\mathbf{x}) = \sum_{i=1}^n x_i$ *eine suffiziente Statistik für p. Die Rao-Blackwellisierung von* $\hat{\vartheta}(\mathbf{x})$ *bezüglich S ergibt sich als*

$$\begin{aligned}\hat{\vartheta}^*(\mathbf{x}) = \mathbb{E}_p[X_1X_2 \mid S(\mathfrak{X}) = S(\mathbf{x})] &= \mathbb{P}_p(\{X_1=1\}\cap\{X_2=1\} \mid S(\mathfrak{X}) = S(\mathbf{x}))\\ &= \frac{S(\mathbf{x})(S(\mathbf{x})-1)}{n(n-1)}\mathbb{1}_{\{2,3,\ldots,n\}}(S(\mathbf{x})).\end{aligned}$$

Bedeutung der letzten Gleichung: Von $S(\mathbf{x})$ *Erfolgen ist bei n Münzwürfen der erste Münzwurf ein Erfolg mit Wahrscheinlichkeit* $S(\mathbf{x})/n$ *und der zweite ein Erfolg mit Wahrscheinlichkeit* $(S(\mathbf{x})-1)/(n-1)$. *Genauer berechnen wir für* $s := S(\mathbf{x}) \in \{2, 3\ldots, n\}$

$$\begin{aligned}&\mathbb{P}_p(\{X_1=1\}\cap\{X_2=1\} \mid S(\mathfrak{X})=s)\\ &\quad= \frac{\mathbb{P}_p(S(\mathfrak{X})=s \mid \{X_1=1\}\cap\{X_2=1\})}{\mathbb{P}_p(S(\mathfrak{X})=s)}\cdot \mathbb{P}_p(X_1=1)\mathbb{P}_p(X_2=1)\\ &\quad\overset{(*)}{=} \frac{\binom{n-2}{s-2}p^{s-2}(1-p)^{n-s}}{\binom{n}{s}p^s(1-p)^{n-s}}\cdot p^2\\ &\quad= \frac{\binom{n-2}{s-2}}{\binom{n}{s}} = \frac{(n-2)!}{(s-2)!(n-s)!}\cdot\frac{s!(n-s)!}{n!} = \frac{s(s-1)}{n(n-1)},\end{aligned}$$

wobei wir in $(*)$ noch $s-2$ der übrigen Erfolge auf $n-2$ Plätze verteilen.