Zur statistischen Auswertung von Erkennergebnissen

Matthias Wolff
Brandenburgische Technische Universität

17. August 2014

Kurzfassung

Dieser Bericht erläutert die Berechnung von Konfidenzintervallen für Erkennquoten von Vektor- und Folgenklassifikatoren sowie Verfahren zum Nachweis statistisch signifikanter Unterschiede zwischen Teststichproben bzw. Erkennalgorithmen. Nach der Darstellung der mathematischen Grundlagen werden praktische Berechnungsmethoden sowie Rechenbeispiele angegeben.

1 Die Beta-Verteilung

Bei der statistischen Auswertung von Erkennquoten spielt die Beta-Verteilung mit der Dichtefunktion

$$p_{Beta}(x;\alpha,\beta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\, x^{\alpha-1}(1-x)^{\beta-1} \qquad x \in \mathbb{R}^{[0,1]} \tag{1}$$

eine zentrale Rolle. α und β sind zwei positive reelle Formparameter. Γ bezeichnet die Gammafunktion, eine Erweiterung der Fakultätsfunktion auf reelle und komplexe Argumente, mit der Eigenschaft

$$\Gamma(n+1) = n! \qquad \text{für } n \in \mathbb{Z}^{\geq 0}. \tag{2}$$

Der Vorfaktor dient der Normierung und ist der Kehrwert der EULERschen Beta-Funktion

$$B(\alpha,\beta) = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)} = \int_0^1 \xi^{\alpha-1}\,(1-\xi)^{\beta-1}\,\mathrm{d}\xi \tag{3}$$

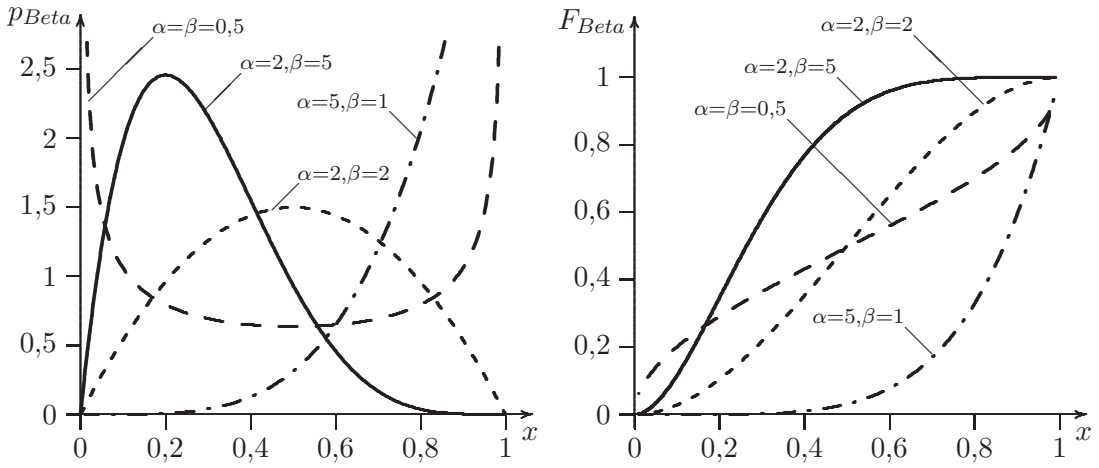
Falls α und β natürliche Zahlen sind, kann die Dichtefunktion der Beta-Verteilung mit (2) auch als

$$p_{Beta}(x;\alpha,\beta) = \frac{(\alpha+\beta-1)!}{(\alpha-1)!\,(\beta-1)!}\, x^{\alpha-1}(1-x)^{\beta-1} \qquad x \in \mathbb{R}^{[0,1]}; \alpha,\beta \in \mathbb{N} \tag{4}$$

geschrieben werden. Bild 1a zeigt die Dichtefunktion für ausgewählte Parameter.

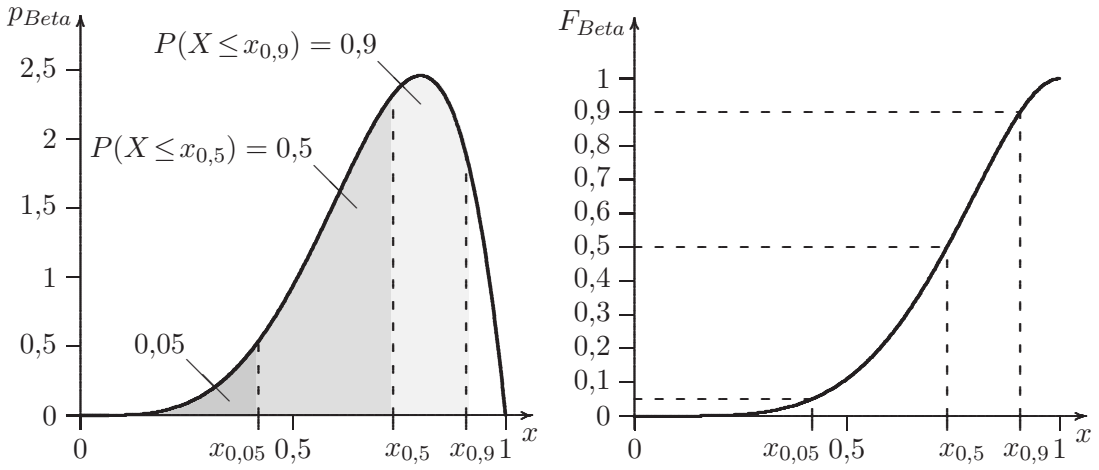
Die Verteilungsfunktion der Beta-Verteilung ist die sogenannte regularisierte unvollständige Beta-Funktion (vgl. 3)

$$\begin{aligned} F_{Beta}(x;\alpha,\beta) &= P(X \leq x) = \int_0^x p_{Beta}(\xi;\alpha,\beta)\,\mathrm{d}\xi \\ &= \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \int_0^x \xi^{\alpha-1}(1-\xi)^{\beta-1}\,\mathrm{d}\xi \qquad x \in \mathbb{R}^{[0,1]}. \end{aligned} \tag{5}$$



(a) Dichtefunktionen (b) Verteilungsfunktionen

Bild 1: Beispiele für die Beta-Verteilung.



(a) Dichtefunktion (b) Verteilungsfunktion

Bild 2: Beispiele für Quantile einer Beta-Verteilung mit den Parametern $\alpha = 5$ und $\beta = 2$.

Bild 1b zeigt einige Beispiele. Die Umkehrfunktion

$$x_P = Q_{Beta}(P;\alpha,\beta) = F^{-1}_{Beta}(P;\alpha,\beta) \tag{6}$$

der Verteilungsfunktion heißt P-Quantil (Beispiele siehe Bild 2). Die stochastische Randbedingung für die Beta-Verteilung lautet

$$F_{Beta}(1;\alpha,\beta) = \int_0^1 p_{Beta}(\xi;\alpha,\beta)\,\mathrm{d}\xi = 1 \tag{7}$$

(man beachte die Integrationsgrenzen!), wie man mit Hilfe von (3) leicht zeigen kann:

$$\begin{aligned}\int_0^1 p_{Beta}(\xi;\alpha,\beta)\,\mathrm{d}\xi &= \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\int_0^1 \xi^{\alpha-1}(1-\xi)^{\beta-1}\,\mathrm{d}\xi \\ &= \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\cdot\frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)} \\ &= 1.\end{aligned}$$

Wichtige Momente und Kenngrößen der Beta-Verteilung sind:

$$\text{Erwartungswert:} \qquad \mathrm{E}(X) = \frac{\alpha}{\alpha+\beta}, \tag{8}$$

$$\text{Varianz:} \qquad \mathrm{Var}(X) = \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)} \quad \text{und} \tag{9}$$

$$\text{Modalwert:} \qquad x_D = \frac{\alpha-1}{\alpha+\beta-2} \qquad \text{für } \alpha,\beta > 1.$$

1.1 Numerische Berechnung

Die Beta-Verteilung ist aufgrund der schnell wachsenden Gamma-Funktion numerisch problematisch. Zur Berechnung der Dichte kann Abhilfe leicht durch Logarithmierung geschaffen werden:

$$p_{Beta}(x;\alpha,\beta) = \exp\Big(\ln\Gamma(\alpha+\beta) - \ln\Gamma(\alpha) - \ln\Gamma(\beta) + (\alpha-1)\ln x + (\beta-1)\ln(1-x)\Big)$$

Die Funktion $\ln\Gamma(x)$ steht in C und dLabPro als `lgamma(x)` direkt zur Verfügung. dLabPro bietet außerdem die Funktion `betadens(x,alpha,beta)` zur Berechnung der Dichtewerte.

Die Werte der Verteilungsfunktion $F_{Beta}(x;\alpha,\beta)$ berechnet man am einfachsten näherungsweise durch numerische Integration der Dichtefunktion.

Auch das P-Quantil $Q_{Beta}(P;\alpha,\beta)$, welches zur Bestimmung des Konfidenzintervalls von Erkennquoten benötigt wird, ist nicht analytisch berechenbar. Eine numerische Näherung wird aber in der Regel von Statistik- und Tabellenkalkulationsprogrammen zur Verfügung gestellt. So enthält beispielsweise Excel die Funktion `BETA.INV(p;alpha;beta)` und dLabPro die Funktion `betaquant(P,alpha,beta)`.

1.2 Parameterschätzung

Die Parameter α und β können mit Hilfe der Maximum-Likelihood-Methode aus einer Stichprobe geschätzt werden. Diese Rechnung ist allerdings sehr kompliziert. Eine wesentlich einfachere Möglichkeit bietet die sogenannte Momentenmethode. Wir schätzen aus der Datenstichprobe nicht direkt die Parameter α und β, sondern Momente. In unserem Fall sind das

$$m = \frac{1}{K}\sum_{k=1}^{K} x_k \qquad \text{Stichprobenmittelwert und}$$

$$s = \frac{1}{K-1}\sum_{k=1}^{K}(x_k-m)^2 \qquad \text{Stichprobenvarianz.}$$

Da der Zusammenhang zwischen den Parametern α und β sowie dem Erwartungswert und der Varianz nach (8) und (9) bekannt ist, können wir die Parameter auf dem Umweg über die Momente ermitteln. Es gilt:

$$m = \frac{\hat\alpha}{\hat\alpha+\hat\beta} \qquad \text{und} \tag{10}$$

$$s = \frac{\hat\alpha\hat\beta}{(\hat\alpha+\hat\beta)^2(\hat\alpha+\hat\beta+1)}, \tag{11}$$

wobei $\hat\alpha$ und $\hat\beta$ Schätzwerte für die Parameter der Beta-Verteilung bezeichnen. Auflösen von (10) nach $\hat\beta$ ergibt zunächst

$$\hat\beta = \frac{\hat\alpha(1-m)}{m}. \tag{12}$$

Wir setzen diesen Ausdruck nun in (11) ein und erhalten

$$s = \frac{\hat{\alpha}^2 \left(\frac{1-m}{m}\right)}{\left(\hat{\alpha} + \frac{\hat{\alpha}(1-m)}{m}\right)^2 \left(\hat{\alpha} + \frac{\hat{\alpha}(1-m)}{m} + 1\right)} = \frac{\hat{\alpha}^2 \left(\frac{1-m}{m}\right)}{\frac{\hat{\alpha}^2}{m^2}\left(\frac{\hat{\alpha}}{m} + 1\right)} = \frac{m(1-m)}{\frac{\hat{\alpha}}{m} + 1}$$

Umstellen nach $\hat{\alpha}$ führt zu

$$\hat{\alpha} = m\left(\frac{m(1-m)}{s} - 1\right)$$

und Einsetzen dieses Ausdrucks in (12) schließlich zu

$$\hat{\beta} = (1-m)\left(\frac{m(1-m)}{s} - 1\right).$$

Beide Schätzformeln gelten nur unter der Bedingung $s < m(1-m)$ (ohne Beweis).

2 Verteilung von Erkennquoten

2.1 Bernoulli- und Binomialverteilung

Das Klassifikationsergebnis von Vektor- und Vektorfolgenklassifikatoren ist eine Bernoulli-verteilte Zufallsgröße S:

$$P_{Berno}(s;x) = \begin{cases} 1-x & \text{für } s=0 \text{ (falsch erkannt)} \\ x & \text{für } s=1 \text{ (richtig erkannt)} \end{cases} \qquad s \in \{0,1\}, x \in [0,1] \subset \mathbb{R}$$

mit dem Definitionsbereich $\{0,1\}$, wobei der Parameter x für die (tatsächliche) Erkennquote steht. $P_{Berno}(s;x)$ bezeichnet also die Wahrscheinlichkeit, dass der Klassifikator ein falsches ($s=0$) oder richtiges ($s=1$) Erkennergebnis liefert unter der Bedingung, dass die tatsächliche Erkennquote x ist. Bild 3 zeigt ein Beispiel für eine Bernoulli-Verteilung.

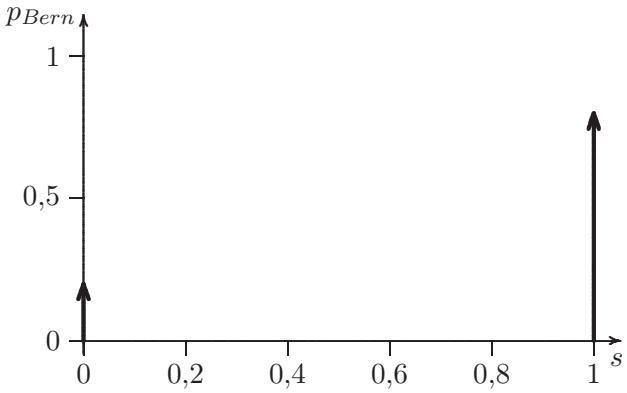


Bild 3: Bernoulli-Verteilung für $x = 0{,}8$.

Klassifiziert man nun eine Teststichprobe mit K Elementen, können $n \in \{0,1,...,K\}$ Einzelklassifikationen korrekt sein. Ist die tatsächliche Erkennquote nach wie vor x, ergibt sich für das Gesamt-Klassifikationsergebnis der Teststichprobe eine binomialverteilte Zufallsgröße N mit

$$\begin{aligned} P_{Bino}(n;x,K) &= \binom{K}{n} x^n (1-x)^{K-n} \\ &= \frac{K!}{n! \cdot (K-n)!} x^n (1-x)^{K-n} \qquad n \in \{0,1,\ldots,K\} \end{aligned} \tag{13}$$

und dem Definitionsbereich $n \in \{0, 1, ..., K\}$. $P_{Bino}(n; x, K)$ bezeichnet die Wahrscheinlichkeit, n korrekte Klassifikationsergebnisse zu erhalten unter den Bedingungen, dass die tatsächliche Erkennquote x ist, dass die Teststichprobe K Elemente enthält und dass die Einzelerkennungen statistisch unabhängig sind.

2.2 Zusammenhang zur Beta-Verteilung

Dichtefunktion

Die Binomalverteilung kann als abgetastete Beta-Verteilungsdichte interpretiert werden. Um dies zu zeigen, führen wir in (13) die Variablen $\alpha = n + 1$ und $\beta = K - n + 1$ ein:

$$P_{Bino}(n; x, K) = \frac{K!}{n! \cdot (K-n)!} \cdot x^{\overbrace{n}^{:=\alpha-1}} \cdot (1-x)^{\overbrace{K-n}^{:=\beta-1}}.$$

und drücken die Fakultätsfunktionen im Nenner nach Gleichung (2) durch die Gamma-Funktion aus:

$$P_{Bino}(n; x, K) = \frac{K!}{\Gamma(\alpha)\Gamma(\beta)}\, x^{\alpha-1}\,(1-x)^{\beta-1}.$$

Nun ersetzen wir nun noch den Zähler des Vorfaktors durch

$$\frac{\Gamma(\alpha+\beta)}{K+1} = \frac{\Gamma\big(n+1+K-n+1\big)}{K+1} = \frac{\Gamma(K+2)}{K+1} = \frac{(K+1)!}{K+1} = K! \qquad K \in \mathbb{N}$$

und erhalten damit schließlich bis auf einen Vorfaktor die Beta-Verteilung nach (1)

$$\begin{aligned} P_{Bino}(n; x, K) &= \frac{1}{K+1}\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\, x^{\alpha-1}\,(1-x)^{\beta-1} \\ &= \frac{1}{K+1}\, p_{Beta}(x; \alpha, \beta) \\ &= \frac{1}{K+1}\, p_{Beta}\big(x; n{+}1, K{-}n{+}1\big) \\ &= \frac{1}{K+1}\frac{\Gamma(K+2)}{\Gamma(n{+}1)\Gamma(K{-}n{+}1)}\, x^{n}\,(1-x)^{K-n} \\ &= p(n; x, K). \end{aligned} \tag{14}$$

Im Gegensatz zur Binomialverteilung kann die „Trefferanzahl“ n bei der Betaverteilung nunmehr reellwertig sein, was wir in der Bezeichnung $p(n; x, K)$ ausgedrückt haben. Bild 4 veranschaulicht den Zusammenhang zwischen Binomial- und Betaverteilung. Man beachte, dass die freie Variable n der Binomialverteilung in die Formparameter der Beta-Verteilung eingegangen ist. Die freie Variable der Betaverteilung ist die tatsächliche Erkennquote x!

Verteilungsfunktion

Auch die Verteilungsfunktion $F_{Bino}(n; x, K)$ der Binomialverteilung kann auf die Beta-Verteilung zurückgeführt werden. Es gilt:

$$\begin{aligned} F_{Bino}(n; x, K) &= F_{Beta}(1{-}x; K{-}n, n{+}1) \\ \sum_{i=0}^{n} \frac{K!}{(K{-}i)! \cdot i!} \cdot x^{i} \cdot (1{-}x)^{K-i} &= \frac{K!}{(K{-}n{-}1)! \cdot n!} \int_{0}^{1-x} \xi^{K-n-1}\,(1-\xi)^{n}\,\mathrm{d}\xi. \end{aligned} \tag{15}$$

Man beachte, dass auch hier links über die „Trefferanzahl“ i summiert, jedoch rechts über die tatsächliche Erkennquote ξ integriert wird. Der Beweis der Identität erfolgt in zwei Schritten:

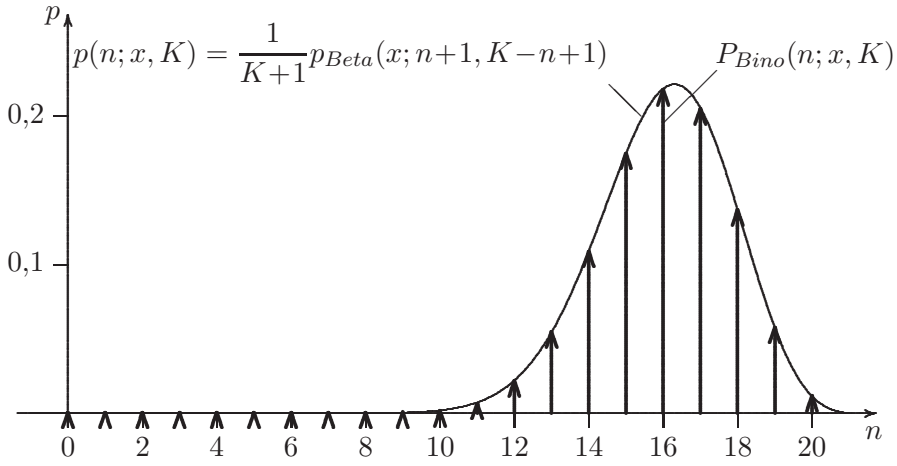


Bild 4: Binomialverteilung nach (13) als Dichtefunktion (DIRAC-Impulse) und angepasste Beta-Verteilungsdichte nach (14) für $K = 20$ und $x = 0{,}8$.

1. Die Anfangswerte beider Seiten bezüglich x (also die Funktionswerte an der Stelle $x = 0$) sind gleich:

$$F_{Bino}(n; 0, K) = \sum_{i=0}^{n} \frac{K!}{(K-i)! \cdot i!} \cdot 0^i \cdot 1^{K-i} = 1 + 0 + 0 + \ldots = 1,$$

$$F_{Beta}(1-0; K-n, n+1) = 1 \quad \text{(stochastische Randbedingung).}$$

2. Die Ableitungen beider Seiten nach x sind gleich.
   - Linke Seite:

$$\begin{aligned}
&\frac{\mathrm{d}}{\mathrm{d}x}\left[\sum_{i=0}^{n} \frac{K!}{(K-i)!\,i!} \cdot x^i \cdot (1-x)^{K-i}\right] \\
&= \sum_{i=0}^{n} \frac{K!}{(K-i)!\,i!}\left[i\,x^{i-1}\,(1-x)^{K-i} - (K-i)\,x^i\,(1-x)^{K-i-1}\right] \\
&= \sum_{i=1}^{n} \frac{K!}{(K-i)!(i-1)!}\,x^{i-1}\,(1-x)^{K-i} - \sum_{i=0}^{n} \frac{K!}{(K-i-1)!i!}\,x^i\,(1-x)^{K-i-1} \\
&= \sum_{j=0}^{n-1} \frac{K!}{(K-j-1)!j!}\,x^j\,(1-x)^{K-j-1} - \sum_{j=0}^{n} \frac{K!}{(K-j-1)!j!}\,x^j\,(1-x)^{K-j-1} \\
&= -\frac{K!}{(K-n-1)! \cdot n!}\,x^n\,(1-x)^{K-n-1}
\end{aligned}$$

   - Rechte Seite: Wir verwenden die aus der Analysis bekannten Beziehungen

$$\frac{\mathrm{d}}{\mathrm{d}x} \int_c^{\varphi(x)} f(\xi)\,\mathrm{d}\xi = \frac{\mathrm{d}}{\mathrm{d}\varphi}\left(\int_c^{\varphi} f(\xi)\,\mathrm{d}\xi\right) \cdot \frac{\mathrm{d}\varphi}{\mathrm{d}x}$$

     (Kettenregel) und

$$\frac{\mathrm{d}}{\mathrm{d}\varphi} \int_c^{\varphi} f(\xi)\,\mathrm{d}\xi = \frac{\mathrm{d}}{\mathrm{d}\varphi}\big[F(\varphi) - F(c)\big] = f(\varphi),$$

(Fundamentalsatz der Analysis). Ineinander eingesetzt ergibt sich:

$$\frac{\mathrm{d}}{\mathrm{d}x}\int_c^{\varphi(x)} f(\xi)\,\mathrm{d}\xi = f(\varphi)\cdot\frac{\mathrm{d}\varphi}{\mathrm{d}x}.$$

Mit $\varphi(x) = 1 - x$ und

$$f(\xi) = \frac{K!}{(K-n-1)!\cdot n!}\,\xi^{K-n-1}\,(1-\xi)^n$$

erhalten wir schließlich

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}x}\left[\frac{K!}{(K-n-1)!\cdot n!}\int_0^{1-x}\xi^{K-n-1}\,(1-\xi)^n\,\mathrm{d}\xi\right]\\ = -\frac{K!}{(K-n-1)!\cdot n!}\,(1-x)^{K-n-1}\,x^n,\end{aligned}$$

also dasselbe Ergebnis wie für die linke Seite.

3 Konfidenzintervall von Erkennquoten

3.1 Clopper-Pearson-Intervall

Erhält man auf einer Teststichprobe mit K Elementen n richtige Klassifikationsergebnisse, kann man eine empirische Erkennquote

$$\hat{x} := \frac{n}{K} \qquad \text{mit } \hat{x} \in \left\{0, \frac{1}{K}, \frac{2}{K}, \ldots, 1\right\} \tag{16}$$

ermitteln. Es stellt sich natürlich die Frage, inwieweit diese Angabe aussagekräftig für die *tatsächliche* Erkennquote x des Klassifikators ist. Zumindest wird die Schätzung augenscheinlich um so unsicherer sein, je kleiner K ist. Um zu einer vernünftigen Einschätzung zu gelangen, betrachten wir die folgenden Fragen:

1. Bei welcher *kleinsten* tatsächlichen Erkennquote x_u kann man mit einer gewissen Wahrscheinlichkeit $P = P_e/2$ noch *mindestens* n Treffer erwarten (siehe Bild 5 oben)?

$$\begin{aligned}x_u: \quad \frac{P_e}{2} = P(N \geq n) &= \sum_{\nu=n}^{K} P_{Bino}(\nu; x_u, K)\\ \rightsquigarrow 1 - \frac{P_e}{2} = P(N \leq n-1) &= \sum_{\nu=0}^{n-1} P_{Bino}(\nu; x_u, K) = F_{Bino}(n-1; x_u, K)\end{aligned}$$

2. Bei welcher *größten* tatsächlichen Erkennquote x_o kann man mit einer gewissen Wahrscheinlichkeit $P = P_e/2$ nur *höchstens* n Treffer erwarten (siehe Bild 5 unten)?

$$x_o: \quad \frac{P_e}{2} = P(N \leq n) = \sum_{\nu=0}^{n} P_{Bino}(\nu; x_o, K) = F_{Bino}(n; x_o, K)$$

Das Intervall $[x_u, x_o]$ heißt Konfidenzintervall (bei dieser speziellen Berechnung auch Clopper-Pearson-Intervall) zum Konfidenzniveau $1 - P_e$. Die Wahrscheinlichkeit P_e selbst heißt Irrtumswahrscheinlichkeit.[1]

[1] In unserem Fall verwenden wir eine sogenannte *zweiseitige* Irrtumswahrscheinlichkeit, die zu gleichen Teilen ($P_e/2$) auf beide Grenzen des Konfidenzintervalls verteilt wird.

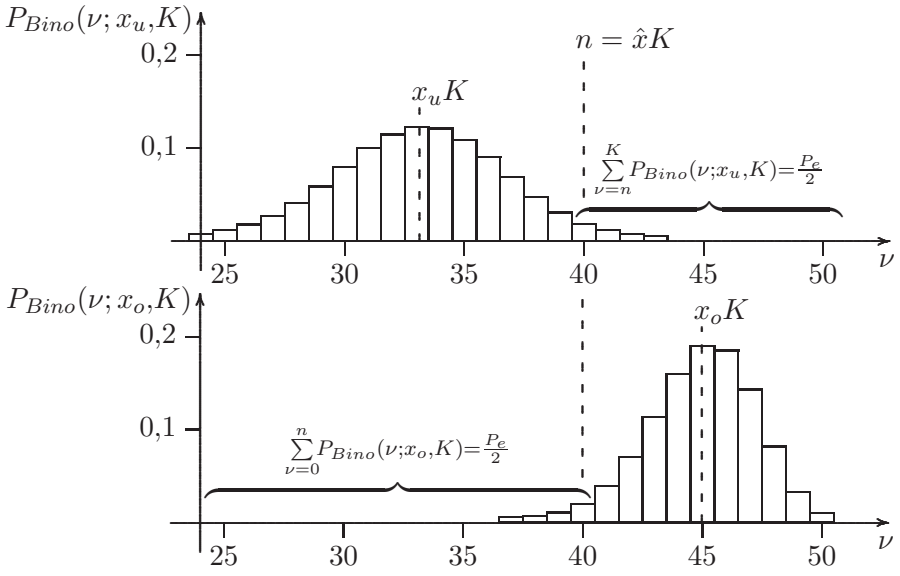


Bild 5: Zur Berechnung der Grenzen x_u und x_o des CLOPPER-PEARSON-Intervalls einer empirischen Erkennquote $\hat{x}$ (im Beispiel: $\hat{x} = 0{,}8$). Weitere Zahlenwerte: Stichprobengröße $K = 50$, Anzahl korrekte Erkennungen $n = \hat{x}K = 40$, Irrtumswahrscheinlichkeit $P_e = 0{,}05$ (d. h. Konfidenzniveau 95 %).

Wegen des Zusammenhangs (15) zur Beta-Verteilung gelten für x_u und x_o auch die folgenden Beziehungen:

$$x_u: \quad 1 - \frac{P_e}{2} = F_{Bino}(n-1; x_u, K) \overset{(15)}{=} F_{Beta}(1-x_u; K-n+1, n),$$
$$x_o: \quad \frac{P_e}{2} = F_{Bino}(n; x_o, K) \quad \overset{(15)}{=} F_{Beta}(1-x_o; K-n, n+1).$$

Die Werte selbst erhält man schließlich mit Hilfe des Quantils der Beta-Verteilung (6)

$$x_u = 1 - Q_{Beta}\Big(1-\frac{P_e}{2}; K-n+1, n\Big), \tag{17}$$
$$x_o = 1 - Q_{Beta}\Big(\frac{P_e}{2}; K-n, n+1\Big). \tag{18}$$

Beispiel

Beim Test eines Erkenners mit einer Teststichprobe der Größe $K = 50$ wurden $n = 40$ korrekte Erkennungsergebnisse erhalten. Mit (16) ist die empirische Erkennquote also

$$\hat{x} = \frac{n}{K} = \frac{40}{50} = 0{,}8.$$

Beim einem Konfidenzniveau von $c = 0{,}95$ beträgt die zulässige Irrtumwahrscheinlichkeit

$$P_e = 1 - c = 1 - 0{,}95 = 0{,}05.$$

Die Grenzen des CLOPPER-PEARSON-Intervalls ergeben sich dann mit (17) und (18) zu:

$$x_u = 1 - Q_{Beta}(0{,}975; 11; 40) \approx 0{,}66 \quad \text{und}$$
$$x_o = 1 - Q_{Beta}(0{,}025; 10; 41) \approx 0{,}90.$$

Damit erhalten wir eine empirischen Erkennquote von

$$\hat{x}^{+(x_o-\hat{x})}_{-(\hat{x}-x_u)} \approx 80^{+10}_{-14}\ \%.$$

beim Konfidenzniveau von 95 %.[2] Zur praktischen Berechnung von $Q_{Beta}(P;\alpha,\beta)$ mit Standardwerkzeugen siehe Abschnitt 1.1.

3.2 Schätzung unter Annahme einer Normalverteilung

Wir erinnern uns, dass ein einzelnes Erkennergebnis eine BENOULLI-verteilte Zufallsgröße S ist (siehe Abschnitt 2.1). Der Schätzer für deren Parameter x ist der Stichproben-Erwartungswert

$$\hat{x} = \mathrm{E}(S) = n/K.$$

Wir berechnen außerdem die Stichproben-Standardabweichung

$$s = \sqrt{\mathrm{Var}(S)} = \sqrt{\frac{K}{K-1}(\mathrm{E}(S^2) - \mathrm{E}(S)^2)} = \sqrt{\frac{K}{K-1}(\hat{x} - \hat{x}^2)}. \tag{19}$$

Dass $\mathrm{E}(S^2) = \mathrm{E}(S) = \hat{x}$ gilt, wird klar, wenn man bedenkt, dass ein einzelnes Erkennergebnis nur entweder falsch oder richtig sein und die Zufallsvariable S somit nur die Werte 0 und 1 annehmen kann. Unterstellt man nun (fälschlicherweise!), dass S normalverteilt sei, kann man die bekannte Faustformel

$$c_{95} \approx 2\frac{s}{\sqrt{K}}.$$

für die halbe Breite des 95 %-Konfidenzintervalls anwenden. Mit (19) wird daraus für unseren Fall

$$c_{95} \approx 2\sqrt{\frac{\hat{x} - \hat{x}^2}{K-1}}. \tag{20}$$

Das Konfidenzintervall lautet dann

$$[x_u, x_o] = [\hat{x} - c_{95}, \hat{x} + c_{95}].$$

Die Genauigkeit dieser Schätzung ist akzeptabel für Anzahlen korrekter Erkennungen von $n > 50$ und $n < K-50$, also *nicht* bei kleinen Teststichproben oder großen Erkennquoten!

Beispiel

Wir kommen noch einmal auf das Beispiel aus Abschnitt 3.1 mit einer Teststichprobengröße von $K = 50$ und einer Anzahl $n = 40$ von korrekten Erkennungen zurück. Die halbe Breite des 95 %-Konfidenzintervalls wird mit Hilfe von (20) zu

$$c_{95} = 2\sqrt{\frac{0{,}8 - 0{,}8^2}{49}} \approx 0{,}114$$

geschätzt. Die empirische Erkennquote beim Konfidenzniveau von 95 % ist dann

$$\hat{x} \pm c_{95} \approx (80 \pm 11)\ \%.$$

Man beachte die relativ deutliche Abweichung zum oben berechneten exakten Ergebnis $\hat{x} \approx 80^{+10}_{-14}$ %.

[2] Die explizite Nennung des Konfidenzniveaus ist notwendig für die Aussagekraft der Angabe!

4 Vergleich von Erkennquoten

Oftmals möchte man empirisch prüfen, ob eine andere Teststichprobe oder irgendeine Änderung an den Erkennalgorithmen zu „besseren“ oder „schlechteren“ Erkennquoten führt. Wir nennen den Referenzversuch (engl. *baseline*) B und den geänderten Versuch A.

Da sowohl die Anzahlen n_A und n_B der korrekten Erkennungen als auch die daraus ermittelten empirischen Erkennquoten $\hat{x}_A$ und $\hat{x}_B$ zufällige Werte sind, ist ein einfacher Vergleich nicht ausreichend. Ermittelt man beispielsweise die Erkennquoten $\hat{x}_A = 92\ \%$ und $\hat{x}_B = 90\ \%$, so ist damit eben *nicht* ohne Weiteres gesagt, dass der Versuch A bessere Ergebnisse liefert als der Versuch B. Eine statistisch gesicherte Aussage erhält man mit Hilfe eines sogenannten Hypothesentests. Man formuliert dazu eine Nullhypothese $H0$ und untersucht, ob diese aufgrund einer Teststatistik bei einer maximal zulässigen Irrtumswahrscheinlichkeit P_e (auch Signifikanzniveau) abgeleht werden kann. Beim Vergleich von Erkennquoten lautet die Nullhypothese

$$\begin{aligned} H0: &\quad \text{Die Erkennquoten von } A \text{ und } B \text{ sind gleich,} \\ H1: &\quad \text{Die Erkennquoten von } A \text{ und } B \text{ sind ungleich.} \end{aligned}$$

$H1$ ist wird – aus offensichtlichen Gründen – als Gegen- oder Alternativhypothese bezeichnet.

4.1 Exakter Test nach Fisher

Mit Hilfe des sogenannten exakten Tests nach Fisher kann der statistische Zusammenhang zwischen zwei binären Kategorisierungen von Daten untersucht werden. In unserem Fall sind die Daten die einzelnen Erkennergebnisse *zweier* zu vergleichender Erkennversuche A und B. Die binären Kategorisierungen sind:

1. Ist das Erkennergebnis richtig oder falsch?
2. Gehört das Erkennergebnis zu Versuch A oder zu Versuch B?

Man trägt die Kategorien und die Anzahlen der jeweils zutreffenden Erkennergebnisse nun in eine Vierfeldertafel ein:

	A	B	Summe
korrekte Erkennungen	n_A	n_B	$n = n_A + n_B$
falsche Erkennungen	$K_A - n_A$	$K_B - n_B$	$K - n$
Summe	K_A	K_B	$K = K_A + K_B$

(21)

Die Anzahlen der korrekten Erkennergebnisse n_A und n_B in beiden Versuchen sowie die Größe der Teststichproben K_A und K_B sind bekannt. Die Gesamtanzahl von Erkennungen ist damit $K = K_A + K_B$ mit insgesamt $n = n_A + n_B$ Erfolgen und $K - n$ Misserfolgen.

Die Nullhypothese des exakten Tests nach Fisher ist, dass *kein* statistischer Zusammenhang zwischen den Kategorien besteht. Sie lautet also in unserem Fall: „Die zu erwartende Anzahl korrekter/falscher Erkennungen ist statistisch unabhängig vom Versuch,“[3] beziehungsweise gleichbedeutend: „Die tatsächlichen Erkennquoten von A und B sind gleich“ (obwohl die experimentell ermittelten sehrwohl verschieden sein können!)

Zur Prüfung der Hypothese stellen wir eine Teststatistik $T(\nu|H_0)$ über die (zufällige) Anzahl ν korrekter Erkennungen in Versuch A auf.[4] Die Teststatistik gilt nur unter

[3] Sie ist aber natürlich abhängig von der Größe der jeweiligen Teststichprobe K_A bzw. K_B!

[4] Die Wahl ist willkürlich. Es könnten auch Versuch B oder die Anzahl von Fehlerkennung betrachtet werden.

der Annahme, dass die Nullhypothese H_0 korrekt ist, also die tatsächlichen Erkennquoten der Versuche A und B gleich sind. Die experimentell ermittelte Anzahl n_A wird als Prüfgröße bezüglich der Teststatistik T bezeichnet. Nimmt man die Gesamtanzahl n korrekter Erkennungen in beiden Versuchen A und B als gegeben an,[5] folgt ν nach Fisher der hypergeometrischen Verteilung

$$T(\nu|H_0) = P_{hgeo}(\nu; n, K_A, K) = \frac{\binom{n}{\nu}\binom{K-n}{K_A-\nu}}{\binom{K}{K_A}} \qquad \nu \in \{0, 1, \ldots, K_A\}. \tag{22}$$

Bild 6 zeigt ein Beispiel.

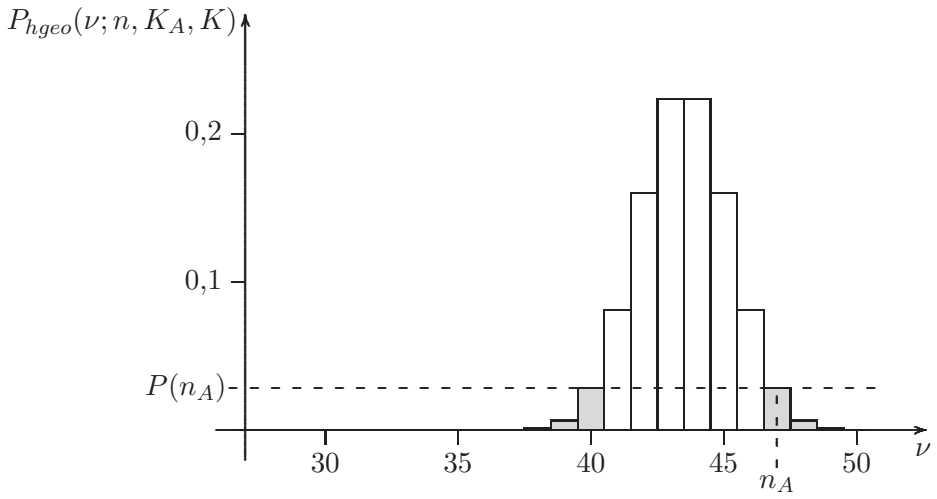


Bild 6: Beispiel für eine Teststatistik (hypergeometrische Verteilung) zum Vergleich zweier Erkennquoten $\hat{x}_A = n_A/K_A$ und $\hat{x}_B = n_B/K_B$. Zahlenwerte: $n_A = 47$, $n_B = 40$, $K_A = K_B = 50$ und damit $n = n_A + n_B = 87$ und $K = K_A + K_B = 100$. Für die grau markierten Wahrscheinlichkeiten gilt $P(\nu) \leq P(n_A)$. Deren Summe ist der P-Wert (im Beispiel $P \approx 0{,}07$).

Man ermittelt nun anhand der Teststatistik die Gesamtwahrscheinlichkeit, dass die Anzahl ν korrekter Erkennungen in Versuch A *zufällig* einen mindestens so extremen Wert annimmt wie die Prüfgröße n_A. Betrachtet werden also alle ν, deren Wahrscheinlichkeit $P(\nu) = T(\nu|H_0)$ höchstens so groß ist, wie die Wahrscheinlichkeit der tatsächlich ermittelten Anzahl $P(n_A) = T(n_A|H_0)$ (grau markierte Balken in Bild 6):

$$P = \sum_{\nu: P(\nu) \leq P(n_A)} P(\nu) \tag{23}$$

Die mit (23) ermittelte Wahrscheinlichkeit wird als P-Wert bezeichnet. Liegt dieser unter dem zuvor festgelegten Signifikanzniveau P_e, so kann die Nullhypothese abgelehnt werden. Folgende Bezeichnungen für Signifikanzniveaus sind etabliert:

$$P < \begin{cases} P_e = 5\ \% & \text{signifikant,} \\ P_e = 1\ \% & \text{sehr signifikant und} \\ P_e = 0{,}1\ \% & \text{hoch signifikant.} \end{cases}$$

[5] Fisher fordert, dass die Randsummen der Vierfeldertafel fest sind. Da dies für K_A und K_B offensichtlich der Fall ist, bleibt hier nur noch die Forderung an n.

Generell sprechen die experimentellen Daten umsomehr gegen die Nullhypothese, je kleiner der ermittelte P-Wert ist. Es sein aber an dieser Stelle ausdrücklich vor einer – leider nicht unüblichen – Überinterpretion des P-Werts gewarnt. Er erlaubt lediglich, eine als zutreffend angenommene Nullhypothese – d. h. $P(H_0) := 1$ – aufgrund einer Teststichprobe abzulehnen. Er ist weder eine Irrtumswahrscheinlichkeit für die Ablehnung der Nullhypothese noch erlaubt er Aussage über die Wahrscheinlichkeit, dass die Nullhypothese zutrifft [4].[6] Der Vergleich mit Abschnitt 4.2 verdeutlicht außerdem die selbstverständliche Tatsache, dass die Wahl der Teststatistik einen entscheidenden Einfluss auf die erhaltene Aussage haben kann.

Beispiel 1

Nach der algorithmischen Optimierung eines Erkenners wurden auf einer Teststichprobe mit $K_A = K_B = 50$ Elemeten $n_A = 47$ korrekte Erkennungen erhalten, vor der Optimierung $n_B = 40$ korrekte Erkennungen. Es soll geprüft werden, ob aufgrund dieser Befunde von einer signifikanten Verbesserung des Erkenners ausgegangen werden kann.

Da ein „signifikantes“ Ergebnis gefordert ist, beträgt die maximal zulässige Irrtumswahrscheinlichkeit $P_e = 5$ %. Insgesamt liegen $K = K_A + K_B = 100$ Erkennergebnisse vor. Davon sind $n = n_A + n_B = 87$ korrekt. Die Teststatistik $T(\nu|H_0)$ nach (22) über die Anzahl der korrekten Erkennungen nach der Optimierung lautet also:

$$T(\nu|H_0) = \frac{\binom{87}{\nu}\binom{100-87}{50-\nu}}{\binom{100}{50}}$$

Es ergibt sich die in Bild 6 dargestellte Verteilungsfunktion. Die Wahrscheinlichkeit, zufällig genau $n_A = 47$ korrekte Erkennungen zu erhalten, ist

$$P(47) = T(47|H_0) = \frac{\binom{87}{47}\binom{13}{3}}{\binom{100}{50}} \approx 0{,}028.$$

Höchstens ebenso groß sind die Wahrscheinlichkeiten für $0 \leq \nu \leq 40$ und $47 \leq \nu \leq 50$. Der P-Wert berechnet sich damit zu

$$P = \sum_{\nu=0}^{40} T(\nu|H_0) + \sum_{\nu=47}^{50} T(\nu|H_0) \approx 0{,}071 > P_e.$$

Da dieser größer als die maximal zulässige Irrtumswahrscheinlichkeit $P_e = 05$ % ist, kann die Nullhypothese: „Die tatsächlichen Erkennquoten vor und nach der Optimierung sind gleich“, *nicht* abgelehnt werden. Obwohl also die empirische Erkennquote von $\hat{x}_B = 80$ % auf $\hat{x}_A = 94$ % (!) gestiegen ist, kann *keine* signifikante Verbesserung festgestellt werden! Das Ergebnis ist auch durch zufällige Einflüsse plausibel erklärbar.

Beispiel 2

Wir verdoppeln nun die Größe der Teststichprobe auf $K_A = K_B = 100$ Elemente und ein erneutes Experiment ergibt $n_B = 80$ korrekte Erkennungen vor der Optimierung und

[6] Beides würde die Kenntnis der A-priori-Wahrscheinlichkeit der Nullhypothese erfordern. Liegt diese beispielsweise bei $P(H_0) = 50$ %, beträgt die A-posteriori-Wahrscheinlichkeit bei einem P-Wert von 5 % lediglich etwa 71 % [4].

$n_A = 94$ danach, also die selben empirischen Erkennquoten wie in Beispiel 1. Für den P-Wert erhalten wir mit der gleichen Rechnung wie oben nun

$$P \approx 0{,}005 < P_e,$$

was sogar eine sehr signifikante Ablehnung der Nullhypothese erlauben würde. Im zweiten Experiment kann also – lediglich aufgrund der größeren Teststichprobe bei ansonsten gleichen Verhältnissen – von einer signifikanten Verbesserung ausgegangen werden. Das Ergebnis ist nicht mehr plausibel durch bloßen Zufall erklärbar.

Numerische Berechnung der hypergeometrischen Verteilung

Auch die hypergeometrische Verteilung nach (22) ist aufgrund der potenziell großen Werte der Binomialkoeffizienten numerisch problematisch. Zur praktischen Berechnung verwenden wir die Beziehung:

$$\frac{\binom{a+b}{a}\binom{c+d}{c}}{\binom{a+b+c+d}{a+c}} = \frac{(a+b)!\,(c+d)!\,(a+c)!\,(b+d)!}{a!\,b!\,c!\,d!\,(a+b+c+d)!}$$

(ohne Beweis). Zur Berechnung von (22) setzen wir

$$\begin{aligned} a &= n_A \\ b &= n_B \qquad\quad = n - n_A \\ c &= K_A - n_A \\ d &= K_B - n_B = K + n_A - K_A - n \end{aligned}$$

und erhalten so

$$\begin{aligned} &P_{hgeo}(n_A; n, K_A, K) \\ &\quad = \frac{n!\,(K-n)!\,K_A!\,(K-K_A)!}{n_A!\,(n-n_A)!\,(K_A-n_A)!\,(K+n_A-K_A-n)!\,K!} \\ &\quad = \frac{\Gamma(n+1)\,\Gamma(K-n+1)\,\Gamma(K_A+1)\,\Gamma(K-K_A+1)}{\Gamma(n_A+1)\,\Gamma(n-n_A+1)\,\Gamma(K_A-n_A+1)\,\Gamma(K+n_A-K_A-n+1)\,\Gamma(K+1)}, \end{aligned}$$

wobei wir in der letzten Zeile die Gamma-Funktion nach (2) eingesetzt haben. Die numerische Berechnung geschieht dann mit Hilfe der logarithmischen Gammafunktion `lgamma(x)` (siehe Abschnitt 1.1):

$$\begin{aligned} &\ln P_{hgeo}(n_A; n, K_A, K) \\ &\quad = \ln\Gamma(n+1) + \ln\Gamma(K-n+1) + \ln\Gamma(K_A+1) + \ln\Gamma(K-K_A+1) \\ &\qquad - \ln\Gamma(n_A+1) - \ln\Gamma(n-n_A+1) - \ln\Gamma(K_A-n_A+1) \\ &\qquad - \ln\Gamma(K+n_A-K_A-n+1) - \ln\Gamma(K+1). \end{aligned}$$

Bei sehr großen Argumenten kann zur Erhöhung der Genauigkeit zusätzlich eine geschickte Umordnung der Summanden hilfreich sein.

4.2 Schätzung unter Annahme einer χ^2-Verteilung

Da die Durchführung des exakten Tests nach FISHER zumindest manuell etwas umständlich ist, kann unter Umständen auf eine Nährung zurückgegriffen werden. Dafür stellt man aus den Einträgen der Vierfeldertafel (21) die Prüfgröße

$$\widehat{\chi^2} = \frac{(K_A+K_B)\Big(n_A(K_B-n_B) - n_B(K_A-n_A)\Big)^2}{K_A\,K_B\,(n_A+n_B)\,(K_A+K_B-n_A-n_B)} \tag{24}$$

auf, die näherungsweise einer χ^2-Verteilung mit einem Freiheitsgrad folgt. Diese Verteilung (und damit die Teststatistik) hat die Dichtefunktion

$$T(x|H_0) = p_{\chi^2,1}(x) = \begin{cases} \dfrac{\mathrm{e}^{-x/2}}{\sqrt{2\pi x}} & x > 0 \\ 0 & x \le 0 \end{cases} \qquad x \in \mathbb{R}$$

(siehe Bild 7) und die Verteilungsfunktion

$$F_{\chi^2,1}(x) = \int_0^x p_{\chi^2,1}(\xi)\,\mathrm{d}\xi = \operatorname{erf}\left(\sqrt{\frac{x}{2}}\right),$$

wobei erf die Fehlerfunktion bezeichnet.

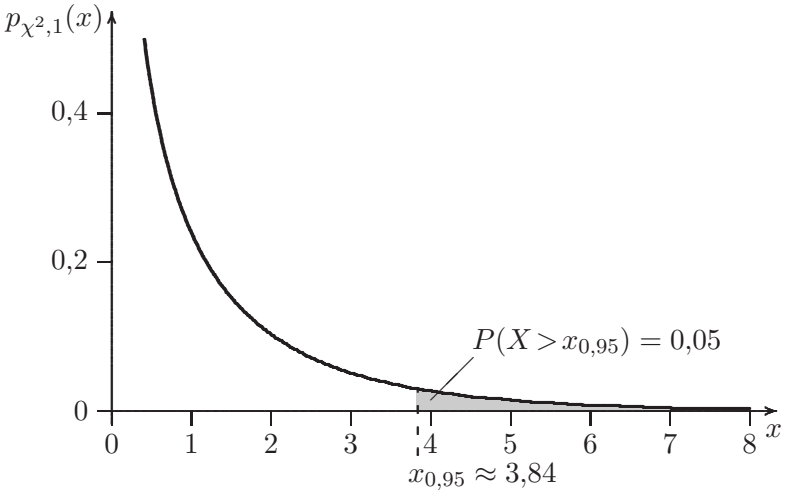


Bild 7: χ^2-Verteilungsdichte mit einem Freiheitsgrad. Die graue Fläche bezeichnet das rechtsseitige 5 %-Quantil.

Der P-Wert der Prüfgröße lautet analog zu (23)

$$P = \int\limits_{x:p(x)\le p\left(\widehat{\chi^2}\right)} p_{\chi^2,1}(x)\,\mathrm{d}x.$$

Die Berechnung ist relativ einfach, da die χ^2-Dichte mit einem Freiheitsgrad streng monoton fällt. Die Werte, deren Wahrscheinlichkeitsdichte kleiner oder gleich $p\left(\widehat{\chi^2}\right)$ ist, sind

also genau diejenigen rechts von $\widehat{\chi^2}$ (siehe Bild 7). Damit gilt:

$$P = \int_{\widehat{\chi^2}}^{\infty} p_{\chi^2,1}(\xi)\,\mathrm{d}\xi \;=\; 1 - \int_{0}^{\widehat{\chi^2}} p_{\chi^2,1}(\xi)\,\mathrm{d}\xi$$

$$= 1 - \operatorname{erf}\left(\sqrt{\frac{\widehat{\chi^2}}{2}}\right). \tag{25}$$

Hinsichtlich der Auswertung des so ermittelten P-Werts gilt das oben gesagte: Er kann lediglich zur Ablehnung der als zutreffend angenommenen Nullhypothese auf einem bestimmten Signifikanzniveau herangezogen werden. Die Näherung gilt als akzeptabel, wenn die Werte n_A, $K_A - n_A$, n_B und $K_B - n_B$ alle größer 5 sind, also *nicht*, wenn mindestens eine der beiden Erkennquoten sehr groß oder die Teststichprobe sehr klein ist.

Beispiel

Wir betrachten noch einmal das erste Beispiel aus Abschnitt 4.1 mit $n_A = 47$, $n_B = 40$ und $K_A = K_B = 50$. Die Prüfgröße (24) ergibt sich damit zu

$$\widehat{\chi^2} = \frac{100 \cdot (47 \cdot 10 - 40 \cdot 3)^2}{50 \cdot 50 \cdot 87 \cdot 13} \approx 4{,}33.$$

Daraus ermitteln wir mit Hilfe der Beziehung (25) einen P-Wert von

$$P = 1 - \operatorname{erf}\left(\sqrt{\frac{4{,}33}{2}}\right) \approx 0{,}037 < P_e$$

Aufgrund dieses P-Werts hätte man die Nullhypothese – im Gegensatz zum exakten Test nach Fisher – auf einem Signifikanzniveau von $P_e = 5\ \%$ also abgeleht und hätte eine signifikante Verbesserung der Erkennalgorithmen vermutet.

5 Weitere Überlegungen

5.1 Beta-Verteilung der empirischen Erkennquote

Wir betrachten noch einmal die empirische Erkennquote

$$\hat{x} := \frac{n}{K} \qquad \text{mit } \hat{x} \in \left(0, \frac{1}{K}, \frac{2}{K}, \ldots, 1\right),$$

welche gleichzeitig ein Schätzer für die tatsächliche Erkennquote x ist. Wegen (14) gilt

$$1 \;=\; \int_{-\infty}^{\infty} p(n; x, K)\,\mathrm{d}n$$

$$\;=\; \int_{-\infty}^{\infty} \frac{1}{K+1} \frac{\Gamma(K+2)}{\Gamma(n+1)\Gamma(K-n+1)}\, x^n\,(1-x)^{K-n}\,\mathrm{d}n \tag{26}$$

Setzen wir nun $n = \hat{x}K$ ein und berücksichtigen, dass $\mathrm{d}n = K\mathrm{d}\hat{x}$ ist, so erhalten wir

$$\begin{aligned} 1 &= \int_{-\infty}^{\infty} \frac{K}{K+1} \frac{\Gamma(K+2)}{\Gamma(\hat{x}K+1)\Gamma\big(K(1-\hat{x})+1\big)}\, x^{\hat{x}K}\,(1-x)^{K(1-\hat{x})}\,\mathrm{d}\hat{x} \\ &= \int_{-\infty}^{\infty} \frac{K}{K+1} p_{Beta}\big(x;\hat{x}K+1, K(1-\hat{x})+1\big)\,\mathrm{d}\hat{x}. \end{aligned}$$

$$p(\hat{x};x,K) = \frac{K}{K+1} p_{Beta}\big(x;\hat{x}K+1, K(1-\hat{x})+1\big) \tag{27}$$

ist dann die Wahrscheinlichkeitsdichte, eine (nun als kontinuierlich angenommene!) empirische Erkennquote von $\hat{x}$ zu ermitteln unter den Bedingungen, dass die tatsächliche Erkennquote x ist und dass die Teststichprobe K Elemente enthält.

$$p_{Beta}(x;\hat{x},K) = p_{Beta}\big(x;\hat{x}K+1, K(1-\hat{x})+1\big) \tag{28}$$

bezeichnet die Wahrscheinlichkeitsdichte der tatsächlichen Erkennquote x unter der Bedingung, dass mit Hilfe einer Teststichprobe mit K Elementen eine empirische Erkennquote von $\hat{x}$ ermittelt wurde. Bild 8 zeigt ein Beispiel für die beiden Dichtefunktionen.

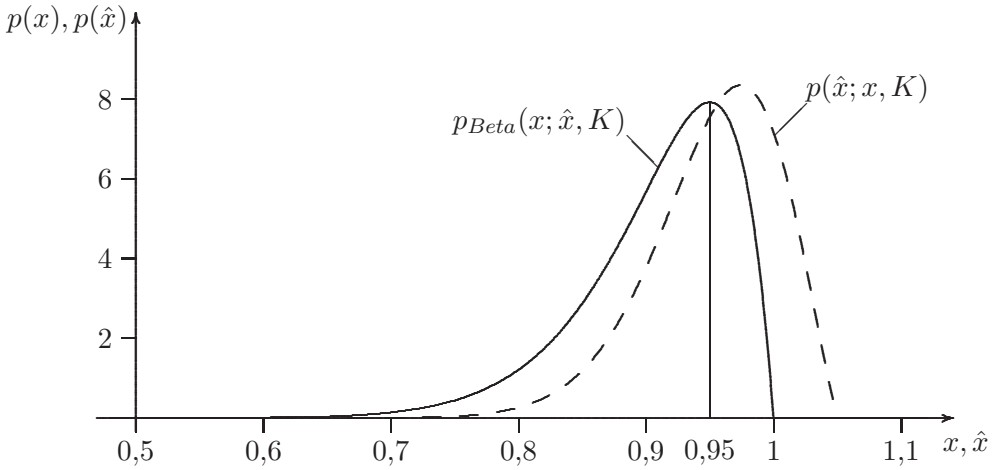


Bild 8: Wahrscheinlichkeitsdichte $p_{Beta}(x;\hat{x},K)$, dass die tatsächliche Erkennquote x ist, falls eine empirische Erkennquote von $\hat{x} = 0{,}95$ ermittelt wurde (durchgezogen, Gl. 28), und dass die empirische Erkennquote $\hat{x}$ ermittelt wird, falls die tatsächliche Erkennquote $x = 0{,}95$ ist (gestrichelt, Gl. 27). Die Größe der Teststichprobe ist jeweils $K = 20$.

Es wäre zu klären, inwieweit (28) zur Bestimmung des Konfizenzintervalls einer empirischen Erkennquote nützlich ist.

5.2 Genauigkeit von Klassenfolgenklassifikatoren

Die Überlegungen in diesem Bericht beziehen sich auf Vektor- und Folgenklassifikatoren, jedoch nicht auf Klassenfolgenklassifikatoren. Deren Genauigkeit ACC kann Werte im Bereich $(-\infty, 1]$ annehmen. Es bleibt zu klären, ob auch diese einer (für beliebige Intervalle erweiterten) Beta-Verteilung folgen. In diesem Fall könnten Konfidenzintervalle und P-Werte analog zur obigen Darstellung ermittelt werden. Abgesehen davon dürfte bei Klassenfolgenklassifikatoren die Unterstellung von Normal- und χ^2-Verteilungen aufgrund der größeren Menge an Basisdaten (viele Einfügungen, Auslassungen, Ersetzungen und

Treffer für *jedes* Element der Teststichprobe) hinreichend genaue Konfidenzintervalle und P-Werte liefern. Auch das wäre jedoch noch genauer zu prüfen.

Literatur

[1] Brandt, S. Datenanalyse für Naturwissenschaftler und Ingenieure. Springer Spektrum. 2013.

[2] Kühlmeyer, M. Statistische Auswertungsmethoden für Ingenieure. Springer. 2001.

[3] Langsrud, Ø.: Fisher's Exact Test, Online-Rechner: `http://www.langsrud.com/stat/fisher.htm`. Abgerufen am 15.05.2014.

[4] Sellke, T.; Bayarri, M. J.; Berger, J.: Calibration of P-values for testing precise null hypotheses. The American Statistician, Jg. 55, H. 1, S. 62–71.

[5] Stahel, W. A.: Statistische Datenanalyse: Eine Einführung für Naturwissenschaftler. Vieweg+Teubner Verlag. 2008.