Statistik I — bis jetzt

K 1 Einführung

K 2 Beschreibende Statistik

K 3 Graphiken

K 4 Schätzen

K 5 Testen

5.1 Tests im allgemeinen
5.2 Binomial Tests
5.3 Normal Test
5.4 Andere Mittelwert-Tests

und jetzt

5.5 Optimalität von statistischen Tests

5.5 Optimalität von statistischen Tests

Einfache und zusammengesetzte Hypothesen

Eine Hypothese wie $\theta = \theta_0$ ist eine einfache Hypothese, weil es die Verteilung vollständig bestimmt. Hypothesen wie $\theta \leq \theta_0$ werden zusammengesetzte Hypothesen genannt.

z.B.

$$H_0: \quad \theta = \theta_0 \qquad H_A: \quad \theta = \theta_1$$

$$H_0: \quad \theta = \theta_0 \qquad H_A: \quad \theta \neq \theta_0$$

$$H_0: \quad \theta \leq \theta_0 \qquad H_A: \quad \theta > \theta_0$$

$$H_0: \quad \theta_1 \leq \theta \leq \theta_2 \qquad H_A: \quad \theta \leq \theta_1 \quad \text{oder} \quad \theta > \theta_2$$

Einseitige und zweiseitige Tests

Wenn die Hypothesen Alternativen in nur einer Richtung berücksichtigen (wie im dritten Beispiel), spricht man von einem einseitigen Test. Das führt dazu, dass ein einseitiger Test vom Niveau α einen anderen Ablehnungsbereich hat als ein zweiseitger Test vom Niveau α. Vorsicht!

Beurteilungskriterien für statistische Tests

Testniveau α
Für einen Test des Niveaus α gilt $\quad g(\theta) \leq \alpha \quad \forall \theta \in H_0$

Unverfälschheit
Für einen unverfälschten Test des Niveaus α gilt

$$g(\theta) \geq \alpha \qquad\qquad \forall \theta \in H_A$$

(Die Wahrscheinlichkeit H_A abzulehnen, wenn diese nicht vorliegt, sollte mindestens so groß sein, als wenn sie vorgelegen hätte.)

Konsistenz
Mit wachsender Stichprobengröße n soll der Test immer besser werden.

$$g(\theta) \to 1 \quad \text{falls } n \to \infty, \qquad \forall \theta \in H_A$$

Optimalität
Sei $g(\theta), g_1(\theta)$ Gütefunktionen von Tests des Niveaus α, dann ist der Test mit Gütefunktion $g(\theta)$ der gleichmäßig beste Test vom Niveau α wenn

$$g(\theta) \geq g_1(\theta) \qquad \forall \theta \in H_A$$

für $g_1(\theta)$ eines beliebigen anderen Tests vom Niveau α.

Optimale Tests
Neyman-Pearson

Neyman-Pearson Lemma für zwei einfache Hypothesen

$$H_0: \qquad \text{Dichte } f_0(\underline{x})$$

$$H_A: \qquad \text{Dichte } f_A(\underline{x})$$

R sei irgendeine Region mit

$$P_0(\underline{x} \in R) \leq \alpha \qquad \text{d.h.} \qquad \int_R f_0(\underline{x})\mathrm{d}\underline{x} \leq \alpha$$

Nehmen wir an, es gebe eine Region

$$R^* = \{\underline{x} : f_A(\underline{x})/f_0(\underline{x}) \geq k\} \text{ mit } P_0(\underline{x} \in R^*) = \alpha$$

wo k so groß wie möglich ist, dann haben wir

$$P_A(\underline{x} \in R^*) \geq P_A(\underline{x} \in R)$$

Dieser Test (H_0 ablehnen, wenn $\underline{x} \in R^*$) ist der optimale Test des Niveaus α und R^* ist die sogenannte kritische Region.

Neyman-Pearson Beweis

Sei $\bar{R}^*$ und $\bar{R}$ die Komplemente von R^* bzw. R.

$$P_{\theta_A}(R^*) - P_{\theta_A}(R) = \int\limits_{R^* \cap \bar{R}} f_A(x)\mathrm{d}x - \int\limits_{\bar{R}^* \cap R} f_A(x)\mathrm{d}x$$

$$\text{Für } R^* \cap \bar{R} \quad f_A(x) \geq k f_0(x)$$

$$\int\limits_{R^* \cap \bar{R}} f_A(x)\mathrm{d}x \geq k \int\limits_{R^* \cap \bar{R}} f_0(x)\mathrm{d}x$$

Ähnlicherweise gilt

$$\int\limits_{\bar{R}^* \cap R} f_A(x)\mathrm{d}x < k \int\limits_{\bar{R}^* \cap R} f_0(x)\mathrm{d}x$$

Deshalb haben wir, dass

$$P_{\theta_A}(R^*) - P_{\theta_A}(R) \geq k \left[\int\limits_{R^* \cap \bar{R}} f_0(x)\mathrm{d}x - \int\limits_{\bar{R}^* \cap R} f_0(x)\mathrm{d}x \right]$$

$$
\begin{aligned}
&= k\left[\int\limits_{R^*} f_0(x)\mathrm{d}x - \int\limits_{R} f_0(x)\mathrm{d}x\right] \\
&= k\left[P_{\theta_0}(R^*) - P_{\theta_0}(R)\right] \\
&\geq 0
\end{aligned}
$$

Eine ähnliche Logik führt bei zusammengesetzten Hypothesen zum Likelihood-Quotienten-Test (Likelihood Ratio LR Test). Man entscheidet sich für die Alternative, wenn

$$
R(x) = \frac{\sup_{\theta \in H_A} f(x,\theta)}{\sup_{\theta \in H_0} f(x,\theta)}
$$

einen Schwellenwert a überschreitet.

LR Test Beispiel

Betrachten wir die Hypothesen:

$$H_0: \quad X \sim N(\mu_0, \sigma_0^2)$$

$$H_1: \quad X \sim N(\mu_1, \sigma_1^2)$$

Wie sollte ein guter Test aussehen?

$$\bar{X} \sim N(\mu, \sigma_i^2/n)$$

$$R = \frac{f_1}{f_0} = \frac{\sigma_0}{\sigma_1} e^{-\frac{n}{2\sigma_1^2}(\bar{x}-\mu_1)^2 + \frac{n}{2\sigma_0^2}(\bar{x}-\mu_0)^2}$$

Für $\sigma_0^2 = \sigma_1^2 = \sigma^2$ und $\mu_1 > \mu_0$

$$R \geq a \Rightarrow \log R \geq \log a$$

$$\begin{aligned} \log R &= -\frac{n}{2\sigma^2}\left[(\bar{x}-\mu_1)^2 - (\bar{x}-\mu_0)^2\right] \\ &= -\frac{n}{2\sigma^2}\left[-2\bar{x}(\mu_1-\mu_0) + (\mu_1^2 - \mu_0^2)\right] \\ &= \frac{n(\mu_1-\mu_0)}{\sigma^2}\left[\bar{x} - \frac{(\mu_1+\mu_0)}{2}\right] \end{aligned}$$

In diesem Fall hat der optimale Test eine erkennbare Form.

Wenn man die zwei Alternativen gleich behandelt, dann würde man nach Maximum Likelihood handeln und die entsprechende Hypothese auswählen.

Für $\sigma_0^2 \neq \sigma_1^2$

$$R \geq a \Rightarrow \log \frac{\sigma_1}{\sigma_0} R \geq \log \frac{\sigma_1}{\sigma_0} a$$

$$\Rightarrow -\frac{1}{\sigma_1^2}(\bar{x} - \mu_1)^2 + \frac{1}{\sigma_0^2}(\bar{x} - \mu_0)^2 \geq \frac{2 \log \frac{\sigma_1}{\sigma_0} a}{n}$$

$$\bar{x}^2 \left(\frac{1}{\sigma_0^2} - \frac{1}{\sigma_1^2} \right) + 2\bar{x} \left(\frac{\mu_1}{\sigma_1^2} - \frac{\mu_0}{\sigma_0^2} \right) \geq \frac{2 \log \frac{\sigma_1}{\sigma_0} a}{n} - \frac{\mu_0^2}{\sigma_0^2} + \frac{\mu_1^2}{\sigma_1^2}$$

Der Ablehnungsbereich eines optimalen Tests wird von der Lösung der quadratischen Gleichung abhängen.

KIs und Tests

Ist C ein Konfidenzbereich zum Niveau $1 - \alpha$, dann ist $\{\theta_0 \in C\}$ der Annahmebereich eines Tests von $H_0 : \theta = \theta_0$ gegen $H_1 : \theta \neq \theta_0$ zum Niveau α

Ist umgekehrt für jedes $\theta_0 \in \Theta$ ein (nichtrandomisierter) Test für $H_0 : \theta = \theta_0$ gegen $H_1 : \theta \neq \theta_0$ zum Niveau α gegeben, so läßt sich daraus ein Konfidenzbereich zum Niveau α gewinnen.

Bayes und Tests

Die ZV X hat die Dichte $f(x, \theta)$ und es muß eine apriori Verteilung für θ geben, $h(\theta)$. Die aposteriori Verteilung

$$h(\theta|x) \propto f(x, \theta)h(\theta)$$

enthält alle Informationen aus den Daten. Wir testen nicht. Statt dessen können wir Wahrscheinlichkeiten angeben:

$$P(\theta_1 < \theta < \theta_2) = \int_{\theta_1}^{\theta_2} h(\theta|x) \mathrm{d}\theta$$

WM Gelbe Karten

Gab es signifikant weniger gelbe Karten an der WM 2010 als an der WM 2006?

Annahmen:

alle m Spiele sind unabhängig von einander

die Anzahl gelber Karten pro Spiel ist Poissonverteilt

Insgesamt gab es 345 gelbe Karten in 64 Spielen.
Im WM 2010 gab es 260 gelbe Karten.

$$S = \sum_{i=1}^{m} X_i \sim P(m\lambda)$$

$$H_0 : \lambda_{2006} = \lambda_{2010}$$

Wie könnte ein statistischer Test aussehen?

5.6 Nichtparametrische Tests

Bis jetzt haben wir entweder angenommen, dass wir große Stichproben haben (so dass der ZGS einsetzt werden kann), oder, dass die ZV normalverteilt sind (so dass, bei gleichen Varianzen, t-Tests durchgeführt werden können). Wie geht man sonst vor?

Bei nichtparametrischen Tests werden Statistiken benutzt, deren Verteilungen von der Grundgesamtheit unabhängig sind, z.B. Rangstatistiken.

5.6.1 Der Vorzeichentest (sign test)

Sei X eine ZV mit stetiger Verteilungsfunktion $F(x;\theta)$, wo der Lageparameter θ der Median ist. Wir möchten

$$H_0 : \theta = \theta_0 \quad \text{gegen} \quad H_1 : \theta > \theta_0$$

testen. Betrachten wir die Teststatistik

$$T = \sum_{i=1}^{n} T_i$$

wo

$$T_i = 1 \quad \text{falls } x_i > \theta_0 \quad \text{und sonst } = 0$$

$$T \sim B(n, 0.5) \quad \text{unter } H_0$$

Gegeben ein Testniveau α, lehnen wir H_0 ab, wenn

$$T \geq t_{1-\alpha}$$

und $t_{1-\alpha}$ wird aus

$$\sum_{j=t_{1-\alpha}}^{n} \binom{n}{j} (0.5)^n \approx \alpha$$

gefunden. Da T auch unter H_1 binomialverteilt ist, läßt sich die Gütefunktion leicht bestimmen:

$$g(p) = \sum_{j=t_{1-\alpha}}^{n} \binom{n}{j} p^j (1-p)^{n-j}$$

Dieser Test benutzt nur die Information, ob $x_i > \theta_0$ ist, nicht den eigentlichen Wert. Deshalb sollte es uns nicht überraschen, dass er weniger gut als ein parametrischer Test ist — wenn die notwendigen Annahmen stimmen.

5.6.2 Wilcoxon Signed-rank Test für gepaarte Daten

Wenn gepaarte Daten vorliegen (z.B. Messungen vor und nach, Bruder und Schwester, links und rechts), können wir uns mit den Unterschieden beschäftigen, statt mit den einzelnen Verteilungen. Unter H_0: die Werte sollen im Durchschnitt gleich sein, wird der Erwartungswert der Unterschiede 0 sein.

Seien $D_1, \ldots, D_n$ unabhängige ZV mit identischer stetigen Verteilung Q, die um 0 symmetrisch ist. d.h. es gelte

$$F_Q(-a) = 1 - F_Q(a) \quad \forall \quad a \in \mathbb{R}$$

Sei $Z_i = 1_{[D_i > 0]}$ und R_i^+ der Rang von $|D_i|$ unter $|D_1|, \ldots |D_n|$

Wir betrachten die Teststatistik

$$W^+ = \sum_{i=1}^{n} Z_i R_i^+$$

die Summe der Ränge mit positiven Zeichen. Da Q stetig ist, können wir annehmen, dass es keine Bindungen ($D_j = D_k$) gibt. Die genaue Verteilung kann man kombinatorisch bestimmen, aber für größere Datensätze ($n > 20$?), wird eine Normalapproximation genommen.

Satz 5.6.2

$$E[W^+] = \frac{n(n+1)}{4}$$

$$V[W^+] = \frac{n(n+1)(2n+1)}{24}$$

Beweis:
Unter H_0 sind die Z_i unabhängige Bernoulli ZV mit $p = 1/2$, so dass

$$E[Z_i] = 1/2 \qquad V[Z_i] = 1/4$$

Es gilt

$$W^+ = \sum_{i=1}^{n} iZ_i$$

weil jeder Rang $1, \ldots, n$ gerade einmal erscheinen wird und weil Z_i und R_i^+ unabhängig sind. Dann haben wir

$$E[W^+] = \frac{1}{2}\frac{n(n+1)}{2}$$

$$V[W^+] = \frac{1}{4}\frac{n(n+1)(2n+1)}{6}$$

Wenn es Bindungen trotzdem gibt, kann man den durchschnittlichen Rang benutzen, so lange es nicht viele gibt. Sonst muß man Modifizierungen vornehmen.

5.6.3
Mann-Whitney U Test, Wilcoxon Rangsummen Test

Gegeben seien zwei Stichproben $\{x_1, \ldots, x_n\}, \{y_1, \ldots, y_m\}$ aus Grundgesamtheiten, die wir vergleichen möchten. Wir kombinieren die Daten, berechnen die Rangstatistiken und die Summen der Rangstatistiken für die zwei Stichproben:

$$W_X = R_1 + \ldots + R_n \quad W_Y = R_{n+1} + \ldots + R_{n+m}$$

wo natürlich

$$W_X + W_Y = \frac{(n+m)(n+m+1)}{2}$$

Man kann W_X bzw. W_Y auch anders darstellen:

$$W = U + \frac{n(n+1)}{2}$$

wo die Statistik U zählt wie oft eine Beobachtung aus der X Gruppe größer ist als eine Beobachtung aus der Y Gruppe.

W und U sind invariant unter Permutationen von $\{X_1, \ldots, X_n\}$. Wir werden deshalb nur den Fall $X_1 < X_2 \ldots < X_n$ betrachten. Für die zugehörigen Ränge $R_1 < \ldots < R_n$ gilt dann

$$R_i = i + \#(Y_j < X_i)$$

Im Mann-Whitney U Test wird U benutzt und im Wilcoxon Rangsummen Test wird W benutzt, aber wir sehen, dass sie zum gleichen Ergebnis führen werden, da

$$\{U < c\} = \{W < c + \frac{n(n+1)}{2}\}$$

Es gibt zwei verschiedene Rechtfertigungen für den Einsatz solcher Tests.

1) (das Permutation Argument) Wir können unsere Statistik beurteilen, indem wir sie mit den entsprechenden Statistiken für alle mögliche

$$\binom{n+m}{n}$$

Zuteilungen der Werte auf die Gruppen X und Y vergleichen.

2) Wie können Verteilungsfunktionen F_X und G_Y annehmen und H_0: $F_X = G_Y$ testen (unter H_0 sind alle Permutationen gleichwahrscheinlich). Es ist wichtig zu bemerken, dass H_0 die Gleichheit der ganzen Verteilungsfunktion testet und nicht die Gleichheit der Mittelwerte oder Mediane.

Die Verteilung von U

Für $k = 0, \ldots, mn$ gilt

$$P(U = k) = N(k; m, n) / \binom{n+m}{m}$$

Hier bezeichnet $N(k; m, n)$ die Anzahl aller Partitionen $\sum_{i=1}^{n} k_i = k$ von k in n aufsteigend geordnete Zahlen $k_1 \leq k_2 \ldots \leq k_n$ aus der Menge $\{0, \ldots, m\}$

Es gibt Tabellen für diese Verteilungen (und sie sind sowieso in modernen Softwarepaketen eingebaut), aber man braucht sie nur für kleine n und m. Hauptsächlich, wenigstens als eine erste Approximation, arbeitet man mit der approximierenden Normalverteilung.

Satz 5.6.3

$$E[U] = \frac{nm}{2}$$

$$V[U] = \frac{nm(m+n+1)}{12}$$

Beweis:

Sei $Z_{ij} = 1$ wenn $X_i > Y_j$ und sonst 0.

$$U = \sum_{i=1}^{n} \sum_{j=1}^{m} Z_{ij}$$

$$E[Z_{ij}] = 1/2 \Rightarrow E[U] = \frac{nm}{2}$$

$$V[U] = Cov[\sum_{i=1}^{n} \sum_{j=1}^{m} Z_{ij}, \sum_{k=1}^{n} \sum_{l=1}^{m} Z_{kl}]$$

$$= \sum_{i=1}^{n} \sum_{j=1}^{m} \sum_{k=1}^{n} \sum_{l=1}^{m} Cov(Z_{ij}, Z_{kl})$$

$$Cov(Z_{ij}, Z_{kl}) = E[Z_{ij} Z_{kl}] - 1/4$$

$$E[Z_{ij} Z_{kl}] = P(X_i > Y_j \text{ und } X_k > Y_l)$$

$$P(X_i > Y_j \text{ und } X_k > Y_l) = \begin{cases} 1/2 & : \quad i = k, j = l \\ 1/4 & : \quad i \neq k, j \neq l \\ 1/3 & : \quad \text{sonst} \end{cases}$$

Das Resultat für den dritten Fall folgt daraus, dass aus drei u.i.v. ZV muß eine die kleinste sein.

Jetzt gilt:

$$Cov(Z_{ij}, Z_{kl}) = \begin{cases} 1/4 & : \quad i = k, j = l \\ 0 & : \quad i \neq k, j \neq l \\ 1/12 & : \quad \text{sonst} \end{cases}$$

$i = k, j = l$ taucht nm mal auf.
$i = k, j \neq l$ taucht $m(m-1)n$ mal auf.
$i \neq k, j = l$ taucht $n(n-1)m$ mal auf.

$$V[U] = \frac{mn}{4} + \frac{n(n-1)m + m(m-1)n}{12}$$

$$= \frac{mn(m+n+1)}{12}$$

Für m, n groß (beide > 10?) soll eine Normalapproximation nicht schlecht sein. Dieses Resultat folgt nicht aus dem ZGS, weil U zwar eine Summe von i.v. ZV ist, aber sie sind nicht unabhängig.

Aus Satz 5.6.3 sehen wir, dass

$$E[W] = \frac{n(m+n+1)}{2}$$

und

$$V[W] = \frac{mn(m+n+1)}{12}$$