

5.6 Nichtparametrische Tests

Bis jetzt haben wir entweder angenommen, dass wir große Stichproben haben (so dass der ZGS eingesetzt 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, \dots, D_n unabhängige ZV mit identischer stetigen Verteilung Q , die um 0 symmetrisch ist. d.h. es gelte

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

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

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

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

Es gilt

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

weil jeder Rang $1, \dots, n$ 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 $\{x_1, \dots, x_n\}, \{y_1, \dots, y_m\}$ aus Grundgesamtheiten, die wir vergleichen möchten. Wir kombinieren die Daten, berechnen die Rangstatistiken und die Summen der Rangstatistiken für die zwei Stichproben:

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

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 $\{X_1, \dots, X_n\}$. Wir werden deshalb nur den Fall $X_1 < X_2 < \dots < X_n$ betrachten. Für die zugehörigen Ränge $R_1 < \dots < R_n$ 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

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

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 $k = 0, \dots, mn$ 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 $k_1 \leq k_2 \leq \dots \leq k_n$ aus der Menge $\{0, \dots, m\}$

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.

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

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

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

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

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

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

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

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

Jetzt gilt:

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

$i = k, j = l$ 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.

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

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

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