5 Nichtparametrische Methoden

Bisher haben wir angenommen, dass eine Familie von Verteilungen gegeben ist. Wenn die möglichen Verteilungen unbekannt sind, so versucht man, aus den Daten strukturelle Zusammenhänge zu finden.

5.1 Empirische Verteilungsfunktion

Wir nehmen an, dass die Daten mindestens ordinales Niveau haben. Also zum Beispiel Klassennoten, Ranglisten oder Körpergrößen.

Wir ziehen eine Stichprobe $x_1, x_2, \ldots, x_n$. Die Voraussetzungen sind:

1. Die Zufallsvariablen $X_1, \ldots, X_n$ sind identisch unabhängig
2. Die unbekannte Verteilungsfunktion $F(x)$ ist stetig.

Aus der Stichprobe ermittelt man zunächst folgende Größen:

1. Die Reihenfolge $x_{(1)} < x_{(2)} < \cdots < x_{(n)}$, wobei Bindungen $x_i = x_j$ ausgeschlossen werden oder durch Werfen einer Münze aufgelöst werden.
2. $x_{(1)}$ =Minimum, $x_{(n)}$ = Maximum, Median m

$$m = \begin{cases} x_{(\frac{n+1}{2})} & n \text{ ungerade} \\ \frac{1}{2}(x_{(\frac{n}{2})} + x_{(\frac{n}{2}+1)}) & n \text{ gerade,} \end{cases}$$

   $d = x_{(n)} - x_{(1)}$ = Spanne.

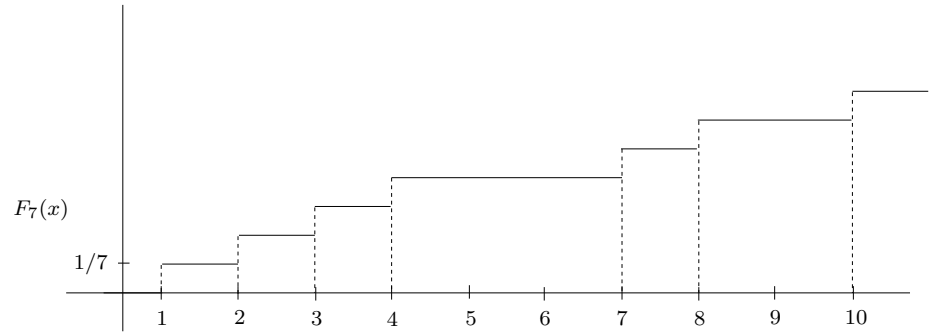
Definition. $F_n(x) = \frac{1}{n}\#\{i : x_i \leq x\}$ heißt die *empirische Verteilungsfunktion.*

$F_n(x)$ ist also eine Treppenfunktion.

Beispiel.

x_1	x_2	x_3	x_4	x_5	x_6	x_7
3	7	1	8	10	2	4

Dann ist $x_{(1)} = 1$, $x_{(7)} = 10$, , $m = 4$, $d = 9$



Die empirische Verteilungsfunktion hat offenbar folgende Eigenschaften:

1. $F_n(x)$ ist monoton steigend,
2. $\lim_{x\to-\infty} F_n(x) = 0$, $\lim_{x\to\infty} F_n(x) = 1$.

Ferner ist $F_n(x)$ diskrete Zufallsvariable mit Werten in $\{\frac{0}{n}, \frac{1}{n}, \cdots, \frac{n}{n}\}$ für jedes x.

Satz 5.1. *Sei $F(x)$ die unbekannte Verteilungsfunktion. Dann gilt*

$$P(F_n(x) = \frac{k}{n}) = \binom{n}{k} F(x)^k (1 - F(x))^{n-k} \ (k = 0, \ldots, n)\,,$$

das heißt die Zufallsvariable $nF_n(x)$ ist binomialverteilt $b(k, n; F(x))$ für jedes x.

Beweis. Wir haben $P(X_i \le x) = F(x)$ für $i = 1, \ldots, n$. Sei x fest und $Y_i(x)$ erklärt durch

$$Y_i(x) = \begin{cases} 1 & X_i \le x \\ 0 & X_i > x\,, \end{cases}$$

dann ist $P(Y_i(x) = 1) = F(x)$. Die Variable $Y_i(x)$ ist also Bernoulli verteilt mit Erfolgswahrscheinlichkeit $p = F(x)$. Aus $nF_n = Y_1(x) + \cdots + Y_n(x)$ folgt daher, dass $nF_n(x)$ binomialverteilt ist mit $b(k, n; F(x))$, und somit

$$P(F_n(x) = \frac{k}{n}) = P(nF_n(x) = k) = \binom{n}{k} F(x)^k (1 - F(x))^{n-k}\,. \qquad \square$$

Folgerung 5.2. *Wir haben für jedes x*

a. $E[F_n(x)] = F(x)$

b. $\mathrm{Var}[F_n(x)] = \frac{F(x)(1-F(x))}{n}$.

Beweis. Aus $E[nF_n(x)] = nF(x)$ folgt $E[F_n(x)] = F(x)$, und aus $\mathrm{Var}[nF_n(x)] = nF(x)(1-F(x))$ folgt $\mathrm{Var}[F_n(x)] = \frac{F(x)(1-F(x))}{n}$. $\square$

Definition. Bei kardinalem Niveau sind der *empirische Erwartungswert* und die *empirische Varianz* gegeben durch

$$\overline{x} = \frac{x_1 + \cdots + x_n}{n},\ s^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2 = \frac{1}{n}\sum_{i=1}^{n} x_i^2 - \overline{x}^2 .$$

Es seien $\overline{X} = \frac{X_1+\cdots+X_n}{n}$, $S^2 = \frac{1}{n}\sum\limits_{i=1}^{n}(X_i - \overline{X})^2 = \frac{1}{2}\sum\limits_{i=1}^{n} X_i^2 - \overline{X}^2$ die entsprechenden Zufallsvariablen.

Satz 5.3. *Es gilt*

a. $E[\overline{X}] = E[X]$, $\mathrm{Var}[\overline{X}] = \frac{1}{n}\mathrm{Var}[X]$,

b. $E[S^2] = \frac{n-1}{n}\mathrm{Var}[X]$.

Beweis. a. Wir haben

$$E[\overline{X}] = \frac{1}{n}E[X_1 + \cdots + X_n] = E[X] ,$$

$$\mathrm{Var}[\overline{X}] = \frac{1}{n^2}\mathrm{Var}[X_1 + \cdots + X_n] = \frac{1}{n}\mathrm{Var}[X] .$$

b. Für S^2 erhalten wir

$$E[S^2] = \frac{1}{n}\sum_{i=1}^{n} E[X_i^2] - E[\overline{X}^2] = E[X^2] - E[\overline{X}^2] .$$

Ferner ist

$$\begin{aligned} E[\overline{X}^2] &= \frac{1}{n^2}E[(X_1 + \cdots + X_n)^2] = \frac{1}{n^2}(\sum_{i=1}^{n} E[X_i^2] + 2\sum_{i<j} E[X_i X_j]) \\ &= \frac{1}{n}E[X^2] + \frac{2}{n^2}\sum_{i<j} E[X_i]E[X_j] \\ &= \frac{1}{n}E[X^2] + \frac{n-1}{n}E[X]^2 , \end{aligned}$$

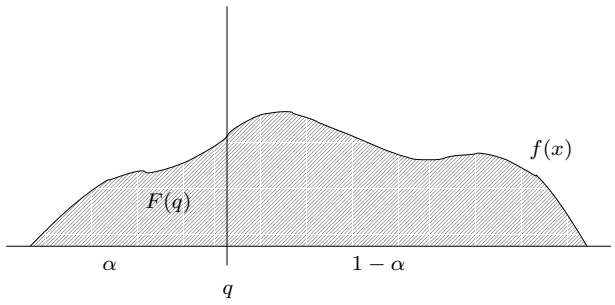
also

$$E[S^2] = \frac{n-1}{n}(E[X^2] - E[X]^2) = \frac{n-1}{n}\mathrm{Var}[X]\,. \qquad \square$$

Definition. Sei $F(x)$ stetige Verteilungsfunktion, $0 < \alpha < 1$. Die Zahl q heißt α-*Quartil*, falls

$$P(X < q) \leq \alpha,\ P(X > q) \leq 1 - \alpha\,,$$

das heißt $F(q) = \alpha$.



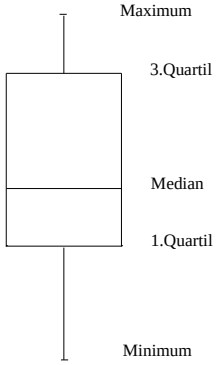
Für $\alpha = \frac{1}{4}$ sprechen wir vom 1. *Quartil*, und für $\alpha = \frac{3}{4}$ vom 3. *Quartil*.

Für die empirische Verteilungsfunktion $F_n(x)$ bedeutet dies

$$\alpha = \frac{1}{4} : F_n(q) \leq \frac{1}{4} \Longrightarrow x_{\lfloor(\frac{n}{4})\rfloor}$$

$$\alpha = \frac{3}{4} : F_n(q) \leq \frac{3}{4} \Longrightarrow x_{\lfloor(\frac{3n}{4})\rfloor}\,.$$

Die übliche graphische Darstellung einer Stichprobe $x_1, \ldots, x_n$ ist das sogenannte Boxplot:



5.2 Verteilung der Ränge

Die Daten seien mindestens auf ordinalem Niveau, $x_1, \ldots, x_n$ die Stichprobe. Die entsprechenden Zufallsvariablen sind identisch unabhängig verteilt.

Definition. Die Variable $R_i = R(X_i)$ ist die Zufallsvariable, die X_i den Rang in der Stichprobe zuteilt, mit Wertebereich $\{1, 2, \ldots, n\}$.

Satz 5.4. *Es gilt*

a. $P(X_1 = r_1 \wedge X_2 = r_2 \wedge \ldots \wedge X_n = r_n) = \frac{1}{n!}$ *für alle Permutationen* $r_1 r_2 \ldots r_n$ *von* $\{1, 2, \ldots, n\}$,

b. $P(R_i = j) = \frac{1}{n}$ *für alle* i *und* j,

c. $P(R_i = k \wedge R_j = \ell) = \frac{1}{n(n-1)}$ *für alle* $i \neq j$, $k \neq \ell$,

d. $E[R_i] = \frac{n+1}{2}$, $\mathrm{Var}[R_i] = \frac{n^2-1}{12}$ *für alle* i,

e. $\mathrm{cov}(R_i, R_j) = -\frac{n+1}{12}$ *für alle* $i \neq j$,

f. $\rho(R_i, R_j) = -\frac{1}{n-1}$ *für alle* $i \neq j$.

Beweis. a. Dies ist wegen der Unabhängigkeit klar.
b. $P(R_i = j) = P(X_i = x_{(j)}) = \frac{(n-1)!}{n!} = \frac{1}{n}$.
c. $P(R_i = k \wedge R_j = \ell) = P(X_i = x_{(k)} \wedge X_j = x_{(\ell)}) = \frac{(n-2)!}{n!} = \frac{1}{n(n-1)}$.
d. $E[R_i] = \sum_{j=1}^{n} jP(R_i = j) = \frac{1}{n}\sum_{j=1}^{n} j = \frac{n+1}{2}$. Ferner ist

$$E[R_i^2] = \sum_{j=1}^{n} j^2 P(R_i = j) = \frac{1}{n}\sum_{j=1}^{n} j^2 = \frac{n(n+1)(2n+1)}{6n} = \frac{(n+1)(2n+1)}{6},$$

also

$$\mathrm{Var}[R_i] = E[R_i^2] - E[R_i]^2 = \frac{(n+1)(2n+1)}{6} - \frac{(n+1)^2}{4} = \frac{n^2-1}{12}.$$

e. Wir haben $\mathrm{cov}(R_i, R_j) = E[R_i R_j] - E[R_i]E[R_j]$, also

$$\begin{aligned}
\mathrm{cov}(R_i, R_j) &= \sum_{k=1}^{n} \sum_{\substack{\ell=1 \\ \ell \neq k}}^{n} k\ell P(R_i = k \wedge R_j = \ell) - (\frac{n+1}{2})^2 \\
&= \frac{1}{n(n-1)} \sum_{k=1}^{n} \sum_{\substack{\ell=1 \\ \ell \neq k}}^{n} k\ell - (\frac{n+1}{2})^2 \\
&= \frac{1}{n(n-1)} \sum_{k=1}^{n} k(\frac{n(n+1)}{2} - k) - (\frac{n+1}{2})^2 \\
&= \frac{1}{n(n-1)} \frac{n^2(n+1)^2}{4} - \frac{1}{n(n-1)} \frac{n(n+1)(2n+1)}{6} - (\frac{n+1}{2})^2 \\
&= -\frac{n+1}{12} \,.
\end{aligned}$$

f. Wir haben

$$\rho(R_i, R_j) = \frac{\mathrm{cov}(R_i, R_j)}{\sqrt{\mathrm{Var}[R_i]\mathrm{Var}[R_j]}} = -\frac{n+1}{12} \cdot \frac{12}{n^2-1} = -\frac{1}{n-1} \,. \qquad \square$$

5.3 Tests mittels geordneter Statistiken

Typische Beispiele aus der Praxis sind:

A. Männer haben mehr Autounfälle als Frauen.

B. Medikament A ist wirksamer als Medikament B.

C. Kindergartenkinder haben in der Schule bessere Noten als Nicht-Kindergartenkinder.

Unsere Vorgangsweise orientiert sich am Hypothesen Testen nach den üblichen 5 Schritten:

1. Wir treffen Annahmen über das Modell und die Daten.

2. Wir stellen die Hypothesen auf.

3. Wir wählen das Irrtumsniveau α.

4. Wir formulieren die Teststatistik und die Entscheidungsregel.

5. Jetzt erst wird das Experiment durchgeführt.

A. Vorzeichentest
Die Daten sind mindestens auf ordinalem Niveau. Sie werden in Paaren (x_i, y_i) erhoben, $i = 1, \ldots, n$. Die Zufallsvariabeln (X_i, Y_i) sind unabhängig und intern unabhängig.

Beispiel. Die Daten (X_i, Y_i) entsprechen Mann/Frau, und es werden die Körpergröße oder Unfallhäufigkeit erhoben.

Die Ereignisse werden folgendermaßen bezeichnet:

$$\begin{aligned} X_i < Y_i: &\quad \text{Ereignis} \quad + \\ X_i = Y_i: &\qquad\qquad\quad 0 \\ X_i > Y_i: &\qquad\qquad\quad - \end{aligned}$$

Der zweiseitige Test betrachtet die Hypothesen

$$\begin{aligned} H_0: &\quad P(+) = P(-) \\ H_1: &\quad P(+) \neq P(-)\,. \end{aligned}$$

Der einseitige Test betrachtet

$$\begin{aligned} H_0: &\quad P(+) \leq P(-) \\ H_1: &\quad P(+) > P(-) \end{aligned} \quad .$$

Unentschieden (Ereignisse 0) werden entfernt. Ist α Irrtumsniveau im einseitigen Test, so wird $\frac{\alpha}{2}$ im zweiseitigen Test verwendet.

Zweiseitiger Test. Wir nehmen als Teststatistik

$$T = \#(+)\,.$$

Unter H_0 haben wir $P(+) = P(-) = \frac{1}{2}$, und T ist binomialverteilt $b(k, n; \frac{1}{2})$. Die Entscheidungsregel ist demnach

$$t < T < n - t \;\Rightarrow\; H_0$$

$$T \leq t \text{ oder } T \geq n - t \;\Rightarrow\; H_1\,,$$

für geeignestes t.

Analyse: Unter H_0 ist der Irrtum 1. Art

$$\sum_{k=0}^{t} \binom{n}{k} \frac{1}{2^n} \leq \frac{\alpha}{2}\,.$$

Für große n und $\alpha = 0,05$ ergibt dies $t \sim \frac{n}{2} - \sqrt{n}\,.$

Einseitiger Test. Unter H_0 ist $X_i > Y_i$ wahrscheinlicher. Wir verwenden als Teststatistik

$$T = \#(-)$$

mit der Entscheidungsregel

$$\begin{aligned} T > t &\Rightarrow H_0 \\ T \leq t &\Rightarrow H_1\,. \end{aligned}$$

Für große n und $\alpha = 0,025$ ist $t \sim \frac{n}{2} - \sqrt{n}\,.$

B. Wilcoxon Rangsummentest
Die Daten sind mindestens auf ordinalem Niveau. Gegeben sind $m+n$ unabhängige Variablen $X_1, \ldots, X_m$ und $Y_1, \ldots, Y_n$ mit den Rängen $1, \ldots, N = m+n$. $F(x)$ sei die (unbekannte) Verteilungsfunktion der identisch verteilten X_i und $G(x)$ jene der identisch verteilten Y_j.

Beispiel. Es sollen 4 Kindergartenkinder $X_1, \ldots, X_4$ gegen 8 Nicht-Kindergartenkinder $Y_1, \ldots, Y_8$ nach ihren schulischen Leistungen verglichen werden. Die X_i haben die Ränge 4,7,9,12 (1 = schlechtester bis 12 = bester Rang).

Die Hypothesen sind

$$\begin{aligned} &H_0 : F(x) = G(x) \quad \text{(kein Unterschied)} \\ &H_1 : F(x) \neq G(x)\,. \end{aligned}$$

Sei $Z_{(1)} < Z_{(2)} < \cdots < Z_{(N)}$ die geordnete Stichprobe, und

$$V_i = \begin{cases} 1 & Z_{(i)} \text{ ist } X\text{-Variable} \\ 0 & Z_{(i)} \text{ ist } Y\text{-Variable.} \end{cases}$$

Die Variablen $V_1, \ldots, V_N$ sind nicht unabhängig, aber unter der Hypothese H_0 haben alle $0,1$-Vektoren mit m 1en und n 0en dieselbe W-keit $\frac{1}{\binom{N}{m}}\,.$

Lemma 5.5. *Unter H_0 gilt*

a. $E[V_i] = \frac{m}{N}$ *für alle* i,

b. $\mathrm{Var}[V_i] = \frac{mn}{N^2}$ *für alle* i,

c. $\mathrm{cov}(V_i, V_j) = -\frac{mn}{N^2(N-1)}$ *für alle* $i \neq j$.

Beweis. a. $P(V_i = 1) = P(Z_{(i)} = X \text{ Variable}) = \frac{m}{N}$. V_i ist Bernoulli Variable mit $p = \frac{m}{N}$, somit gilt $E[V_i] = \frac{m}{N}$.

b. $\mathrm{Var}[V_i] = \frac{m}{N} \cdot \frac{N-m}{N} = \frac{mn}{N^2}$.

c. Wir haben für $i \neq j$

$$E[V_i V_j] = P(V_i = 1 \wedge V_j = 1) = \frac{\binom{m}{2}}{\binom{N}{2}} = \frac{m(m-1)}{N(N-1)}\,,$$

und somit

$$\begin{aligned}\mathrm{cov}(V_i, V_j) &= E[V_i V_j] - E[V_i]E[V_j] \\ &= \frac{m(m-1)}{N(N-1)} - \frac{m^2}{N^2} = -\frac{mn}{N^2(N-1)}\,. \qquad \square\end{aligned}$$

Als Teststatistik verwenden wir

$$W_N = \sum_{i=1}^{N} i V_i = \text{Summe der Ränge der } X_j\,.$$

Unter H_0 ist

$$E[W_N] = \sum_{i=1}^{N} i E[V_i] = \frac{m}{N} \sum_{i=1}^{N} i = \frac{m(N+1)}{2}\,.$$

Ausrechnen ergibt

$$\mathrm{Var}[W_N] = \frac{mn(N+1)}{12}\,.$$

Sei $W_{\min} = \sum_{i=1}^{m} i$, $W_{\max} = \sum_{i=N-m+1}^{N} i$. Die Entscheidungsregel ist demnach aus Symmetriegründen

$$W_N \leq t \;\text{ oder }\; W_N \geq W_{\max} - (t - W_{\min}) \Longrightarrow H_1$$

mit

$$P(W_N \leq t) \leq \frac{\alpha}{2} \text{ unter } H_0 .$$

Beispiel. Analysieren wir das Kindergartenbeispiel. Hier ist $m = 4$, $n = 8$, $N = 12$, $\binom{N}{m} = \binom{12}{4} = 495$, $\alpha = 0,05$. Wir haben $W_{\max} = 9+10+11+12 = 42$, $W_{\min} = 1 + 2 + 3 + 4 = 10$. Wir stellen eine Tabelle auf:

$$\begin{array}{rl} W = 10 & 1,2,3,4 \\ 11 & 1,2,3,5 \\ 12 & 1,2,3,6; 1,2,4,5 \\ 13 & 1,2,3,7; 1,2,4,6; 1,3,4,5 \\ 14 & 1,2,3,8; 1,2,4,7; 1,2,5,6; 1,3,4,6; 2,3,4,5 . \end{array}$$

Also ist $P(T \leq 14) = \frac{12}{495} \sim 0,024 \leq 0,025$

Die Entscheidungsregel lautet demnach

$$W \leq 14 \text{ oder } W \geq 38 \Longrightarrow H_1 .$$

In unserem Beispiel ist $W = 4+7+9+12 = 32$. Die Nullhypothese H_0 kann also nicht verworfen werden.

C. Mediantest

Beispiel. Eine Reifenfirma entwickelt einen Reifentyp und behauptet, dass die Reifen im Median ≥ 33.000 km halten. Wie sollen wir das testen?
Wir nehmen kardinales Niveau an. Die Variablen $X_1, \ldots, X_n$ sind unabhängig identisch verteilt mit stetiger Verteilungsfunktion $F(x)$, die symmetrisch um den Median liegt. M_0 ist vorgegeben.

Zweiseitiger Test:

$$\begin{array}{ll} H_0 : & M = M_0 \\ H_1 : & M \neq M_0 . \end{array}$$

Einseitiger Test:

$$\begin{array}{ll} H_0 : & M \geq M_0 \\ H_1 : & M < M_1 . \end{array}$$

Es sei $Y_i = X_i - M_0$ und $r(|Y_i|)$ der Rang von $|Y_i|$. Nun setzen wir

$$Z_i = \begin{cases} 1 & \text{falls} \quad Y_j > 0 \text{ wobei } r(|Y_j|) = i \\ 0 & \text{falls} \quad Y_j < 0 \text{ wobei } r(|Y_j|) = i . \end{cases}$$

Wir können $Y_j = 0$ wegen $P(Y_j = 0) = 0$ vernachlässigen.

Nun betrachten wir die Teststatistik

$$W^+ = \sum_{i=1}^{n} iZ_i = \text{Summe der Ränge der } \textit{positiven} \text{ Differenzen.}$$

Zweiseitiger Test: Unter H_0 ist $E[Z_i] = \frac{1}{2}$ wegen der symmetrischen Verteilung um M_0, also

$$E[W_i^+] = \sum_{i=1}^{n} iE[Z_i] = \frac{n(n+1)}{4}\,.$$

Alle $0,1$-Vektoren $(z_1, \ldots, z_n)$ haben W-keit $\frac{1}{2^n}$, somit ist

$$P(W^+ = w) = \frac{a(w)}{2^n}\,,$$

wobei $a(w)$ die Anzahl der n-Tupel $(y_1, \ldots, y_n)$ ist, so dass die Summe der Ränge der positiven y_i gleich w ist.

Die Entscheidungsregel ist demnach

$$W^+ \leq t \;\text{ oder }\; W^+ \text{ groß } \Longrightarrow H_1\,,$$

wobei

$$P(W^+ \leq t) \leq \frac{\alpha}{2}\,.$$

Im einseitigen Test wird $P(W^+ \leq t) \leq \alpha$ verwendet.

5.4 Tests auf Korrelation

Wir untersuchen zwei Merkmale X, Y, zum Beispiel

X:	Körpergröße Vater	Y:	Körpergröße Sohn
	Geschlecht		Wahlverhalten
	Gesundheit		Schulerfolg

und stellen uns die Fragen

A. Sind X und Y unabhängig?

B. Sind X und Y positiv (negativ) korreliert?

Die Daten sind mindestens auf ordinalem Niveau.

Beispiel. Sieben Bewerber stellen sich vor. Zwei Personalvertreter stellen jeweils eine Rangliste auf (1 =bester bis 7 = schlechtester) .

Bewerber	1	2	3	4	5	6	7
A	5	7	1	3	4	6	2
B	3	6	1	2	4	7	5

Sind die Ranglisten korreliert?

Es seien $x_1, \ldots, x_n$ und $y_1, \ldots, y_n$ die Stichprobenwerte,

$$\begin{array}{lll} r_1, \ldots, r_n & \text{die Ränge von} & x_1, \ldots, x_n \\ s_1, \ldots, s_n & \text{die Ränge von} & y_1, \ldots, y_n, \end{array}$$

$\{r_1, \ldots, r_n\} = \{s_1, \ldots, s_n\} = \{1, \ldots, n\}$.

Definition. Der *Korrelationskoeffizient von Spearman* ist

$$r = \frac{\sum_{i=1}^n (r_i - \overline{r})(s_i - \overline{s})}{\sqrt{\sum_{i=1}^n (r_i - \overline{r})^2}\sqrt{\sum_{i=1}^n (s_i - \overline{s})^2}},$$

wobei $\overline{r} = \overline{s} = \frac{n+1}{2}$.

Satz 5.6. *Wir haben*

$$r = 1 - \frac{6\sum_{i=1}^n d_i^2}{(n-1)n(n+1)}, d_i = r_i - s_i \ (i = 1, \ldots, n).$$

Beweis. Es ist

$$\begin{aligned} \sum_{i=1}^n (r_i - \overline{r})^2 &= \sum_{i=1}^n (s_i - \overline{s})^2 = \sum_{i=1}^n (i - \frac{n+1}{2})^2 \\ &= \frac{n(n+1)(2n+1)}{6} - (n+1)\frac{n(n+1)}{2} + \frac{n(n+1)^2}{4} \\ &= \frac{n(n+1)}{12}(4n + 2 - 6n - 6 + 3n + 3) \\ &= \frac{(n-1)n(n+1)}{12}. \end{aligned}$$

Daraus folgt

$$r = \frac{12}{(n-1)n(n+1)} \sum_{i=1}^{n} (r_i - \frac{n+1}{2})(s_i - \frac{n+1}{2}) \,.$$

Setzen wir $d_i = r_i - s_i = (r_i - \frac{n+1}{2}) - (s_i - \frac{n+1}{2})$, so ergibt sich

$$\begin{aligned} \sum_{i=1}^{n} d_i^2 &= \sum_{i=1}^{n} (r_i - \frac{n+1}{2})^2 - 2\sum_{i=1}^{n} (r_i - \frac{n+1}{2})(s_i - \frac{n+1}{2}) + \sum_{i=1}^{n} (s_i - \frac{n+1}{2})^2 \\ &= \frac{(n-1)n(n+1)}{6} - 2\sum_{i=1}^{n} (r_i - \overline{r})(s_i - \overline{s}) \,, \end{aligned}$$

also

$$\begin{aligned} r &= \frac{12}{(n-1)n(n+1)} \frac{(n-1)n(n+1) - 6\sum_{i=1}^{n} d_i^2}{12} \\ &= 1 - \frac{6\sum_{i=1}^{n} d_i^2}{(n-1)n(n+1)} \,. \qquad \square \end{aligned}$$

Beispiel. In unserem Beispiel ist

$$d_1^2 = 4, d_2^2 = 1, d_3^2 = 0, d_4^2 = 1, d_5^2 = 0, d_6^2 = 1, d_7^2 = 9 \,,$$

also

$$r = 1 - \frac{6 \cdot 16}{6 \cdot 7 \cdot 8} = \frac{5}{7} = 0,714 \,.$$

Folgerung 5.7. *Für den Korrelationskoeffizienten von Spearman gilt*

a. $-1 \leq r \leq 1$,

b. $r = 1 \Longrightarrow r_i = s_i \ (i = 1, \ldots, n)$,

c. $r = -1 \Longrightarrow r_i = n + 1 - s_i \ (i = 1, \ldots, n)$.

Beweis. Aus Satz 5.6 folgt, dass $-1 \leq r \leq 1$ äquivalent ist zu

$$-1 \leq 1 - \frac{6\sum_{i=1}^{n} d_i^2}{(n-1)n(n+1)} \leq 1$$

also zu

$$0 \leq \sum_{i=1}^{n} d_i^2 \leq \frac{(n-1)n(n+1)}{3}\,,$$

wobei die linke Seite genau $r = 1$ entspricht und die rechte Seite $r = -1$. Die linke Seite $0 \leq \sum_{i=1}^{n} d_i^2$ ist offensichtlich gültig mit $\sum_{i=1}^{n} d_i^2 = \sum_{i=1}^{n}(r_i - s_i)^2 = 0 \Longrightarrow$ $r_i = s_i$ für alle i. Um die rechte Seite zu verifizieren, können wir o.B.d.A. $s_i = i$ setzen, also müssen wir

$$\sum_{i=1}^{n}(r_i - i)^2 \leq \frac{(n-1)n(n+1)}{3}$$

zeigen, mit Gleichheit genau für $r_i = n+1-i$. Betrachten wir eine Permutation $(r_1, \ldots, r_n)$. Falls $r_k < r_{k+1}$ gilt, so gilt für die Permutation $(r'_1, \ldots, r'_k,$ $r'_{k+1}, \ldots, r'_n)$ mit $r'_k = r_{k+1}$, $r'_{k+1} = r_k$, $r'_i = r_i$ $(i \neq k, k+1)$

$$\begin{aligned}
&\sum_{i=1}^{n}(r'_i - i)^2 - \sum_{i=1}^{n}(r_i - i)^2 = (r'_k - k)^2 + (r'_{k+1} - (k+1))^2 \\
&- (r_k - k)^2 - (r_{k+1} - (k+1))^2 \\
&= (r_{k+1} - k)^2 + (r_k - (k+1))^2 - (r_k - k)^2 - (r_{k+1} - (k+1))^2 \\
&= -2kr_{k+1} - 2(k+1)r_k + 2kr_k + 2(k+1)r_{k+1} \\
&= 2r_{k+1} - 2r_k > 0\,.
\end{aligned}$$

Der größte Ausdruck ist daher genau für $r_1 = n, r_2 = n-1, \ldots, r_n = 1$ gegeben und für diesen berechnen wir

$$\begin{aligned}
\sum_{i=1}^{n}(n+1-2i)^2 &= n(n+1)^2 - 4\frac{(n+1)^2 n}{2} + 4\frac{n(n+1)(2n+1)}{6} \\
&= -n(n+1)^2 + \frac{2}{3}n(n+1)(2n+1) \\
&= \frac{n(n+1)}{3}(4n+2-3n-3) = \frac{(n-1)n(n+1)}{3}\,,
\end{aligned}$$

wie gewünscht. $\square$

Damit können wir sagen:

Falls $r \sim 1$ ist,	dann sind X, Y positiv korreliert, große Ränge der x_i entsprechen großen Rängen der y_i, und umgekehrt.
Falls $r \sim -1$ ist,	dann sind X, Y negativ korreliert, kleine Ränge der x_i entsprechen großen Rängen der y_i und umgekehrt.
Falls $r \sim 0$ ist,	dann sind X, Y unkorreliert.

5.5 Korrelationstest von Spearman

Die Variablen $X_1, \ldots, X_n$ sind unabhängig identisch verteilt, ebenso die Variablen $Y_1, \ldots, Y_n$, die Paare $(X_1, Y_1), \ldots, (X_n, Y_n)$ sind unabhängig.

Zweiseitiger Test.

$$\begin{array}{ll} H_0: & X, Y \text{ unabhängig} \\ H_1: & X, Y \text{ korreliert}. \end{array}$$

Einseitiger Test.

$$\begin{array}{ll} H_0: & X, Y \text{ unabhängig} \\ H_1: & X, Y \text{ positiv korreliert}. \end{array}$$

Wir definieren die Rang-Zufallsvariablen

$$R_i = R(X_i),\ S_i = R(Y_i),\ D_i = R_i - S_i\,,$$

$$D = \sum_{i=1}^{n} D_i^2 = \sum_{i=1}^{n} (R_i - S_i)^2\,.$$

O.B.d.A. sei $s_i = i$, also $R(S_i) = i$, und somit

$$D = \sum_{i=1}^{n} (R_i - i)^2 = \sum_{i=1}^{n} R_i^2 - 2\sum_{i=1}^{n} iR_i + \sum_{i=1}^{n} i^2\,.$$

Wir wissen

$$E[R_i] = \frac{n+1}{2},\ E[R_i^2] = \frac{(n+1)(2n+1)}{6},\ \mathrm{Var}[R_i] = \frac{n^2-1}{12}\,.$$

Daraus folgt

$$\begin{aligned} E[D] &= \frac{n(n+1)(2n+1)}{3} - 2\sum_{i=1}^{n} iE[R_i] \\ &= \frac{n(n+1)(2n+1)}{3} - \frac{n(n+1)^2}{2} = \frac{(n-1)n(n+1)}{6}\,. \end{aligned}$$

Ist R die Zufallsvariable für den Spearman Koeffizienten, so folgt aus Satz 5.6

$$E[R] = 1 - \frac{6E[D]}{(n-1)n(n+1)} = 0\,.$$

Wir nehmen als Teststatistik

$$d = \sum_{i=1}^{n} (r_i - s_i)^2\,.$$

Niedrige Werte von d weisen auf positive Korrelation hin, hohe Werte auf negative Korrelation.

Einseitiger Test. Die Entscheidungsregel ist

$$\begin{aligned} d > t_\alpha &\Rightarrow H_0 \\ d \le t_\alpha &\Rightarrow H_1\,, \end{aligned}$$

wobei t_α das α-Quantil der Gleichverteilung über alle Permutationen ist.

Beispiel. In unserem Beispiel haben wir $n = 7$, $d = 16$. Alle Permutationen $(r_1, \ldots, r_7)$ haben dieselbe W-keit $\frac{1}{7!}$. Gesucht ist das α-Quantil $t = t_\alpha$ mit

$$\sum_{w=0}^{t} \frac{a(w)}{7!} \le \alpha, \quad t = \text{maximal groß}\,,$$

wobei $a(w) = \#\{(r_1, \ldots, r_7)\colon d = \sum_{i=1}^{7}(r_i - i)^2 = w\}$. Wir haben

$$\begin{aligned} a(0) &= 1 \qquad 1\,2\,3\,4\,5\,6\,7 \\ a(1) &= 0 \\ a(2) &= 6 \qquad 2\,1\,3\,4\,5\,6\,7, \ldots, 1\,2\,3\,4\,5\,7\,6 \\ a(3) &= 0 \\ &\vdots \end{aligned}$$

Es ergibt sich bei $\alpha = 0,05$ das α-Quantil $t_\alpha = 18$. Die Entscheidungsregel $d \le 18$ besagt also, dass H_0 verworfen wird, die Ranglisten sind positiv korreliert.

Für große Stichproben nimmt man in der Praxis

$$Z_n = \frac{R_n}{\sqrt{n-1}} \xrightarrow{\text{i.V.}} N(0,1) \text{ (wenn } X, Y \text{ unabhängig sind)}\,.$$

Die Nullhypothese wird abgelehnt, wenn $Z \geq z_{1-\alpha}$ mit $\phi(z_{1-\alpha}) = 1-\alpha$.

5.6 Tests für nominale Skalen

Ein typisches Beispiel aus der Wahlforschung vergleicht die Variable X =Einkommen, Y =Parteipräferenz, und stellt die Frage, ob X und Y unabhängig sind oder nicht.

Wir haben zwei Merkmale A und B mit den disjunkten Klassen

$$\begin{aligned} A: &\quad A_1, A_2, \ldots, A_k \\ B: &\quad B_1, B_2, \ldots, B_\ell \end{aligned}$$

Die Daten fasst man in einer sogenannten *Kontingenztafel* zusammen, das heißt in einer Matrix (n_{ij}), $i = 1, \ldots, k$, $j = 1, \ldots, \ell$, wobei $n_{ij} = \#$Merkmale in $A_i \cap B_j$.

Beispiel. A =Einkommen, B =Parteipräferenz

A \ B	CDU	SPD	FDP	Andere	$\sum$
hoch	35 / 22,5	7 / 21,5	5 / 4,5	3 / 1,5	50
mittel	250 / 270	250 / 258	80 / 54	20 / 18	600
niedrig	165 / 157,5	173 / 150,5	5 / 31,5	7 / 10,5	350
	450	430	90	30	1000

In den Kästchen wird n_{ij}, $\widetilde{n}_{ij}$ notiert, wobei

$n_{ij} = \#$ in $A_i \cap B_j$ tatsächlich erhoben
$\widetilde{n}_{ij} = \#$ in $A_i \cap B_j$ wenn die Merkmale unabhängig sind.

Wir schreiben für die *Randhäufigkeiten* kurz

$$n_{i\cdot} = \#A_i = \sum_{j=1}^{\ell} n_{ij}\,,$$

$$n_{\cdot j} = \#B_j = \sum_{i=1}^{k} n_{ij}\,,$$

alsos

$$n = \sum_{i=1}^{k} n_{i\cdot} = \sum_{j=1}^{\ell} n_{\cdot j} = \sum_{i,j} n_{ij}\,.$$

Falls die Merkmale unabhängig sind, so haben wir

$$\frac{\widetilde{n}_{ij}}{n} = P(X \in A_i \wedge Y \in B_j) = P(X \in A_i)P(Y \in B_j) = \frac{n_{i\cdot}}{n} \cdot \frac{n_{\cdot j}}{n}\,,$$

also

$$\widetilde{n}_{ij} = \frac{n_{i\cdot} n_{\cdot j}}{n} \quad (i = 1, \ldots, k, j = 1, \ldots, \ell)\,.$$

Es gilt

$$\sum_{i=1}^{k} \widetilde{n}_{ij} = \frac{n_{\cdot j}}{n} \sum_{i=1}^{k} n_{i\cdot} = n_{\cdot j}\,, \sum_{j=1}^{\ell} \widetilde{n}_{ij} = n_{i\cdot}\,.$$

A. Chi-Quadrat Test auf Unabhängigkeit

Dies ist wohl das bekannteste aller Testverfahren. Wir haben die Merkmale $A_1, \ldots, A_k$ und $B_1, \ldots, B_\ell$ mit den Zufallsvariablen X und Y.

Die zu testenden Hypothesen sind

H_0: Merkmale A und B sind unabhängig
H_1: Merkmale sind abhängig.

Unter H_0 bietet sich als Teststatistik

$$X^2 = \sum_{i=1}^{k} \sum_{j=1}^{\ell} \frac{(n_{ij} - \widetilde{n}_{ij})^2}{\widetilde{n}_{ij}}$$

an. Der Nenner erklärt sich aus folgender Überlegung. Wenn n_{ij} und $\widetilde{n}_{ij}$ klein sind, so fällt $n_{ij} - \widetilde{n}_{ij}$ mehr ins Gewicht, daher die Gewichtung mit $\frac{1}{\widetilde{n}_{ij}}$.

Sei

$$P(X \in A_i) = p_{i\cdot}\,,\ P(Y \in B_j) = p_{\cdot j}\,,\ P(X \in A_i \wedge Y \in B_j) = p_{ij}\,.$$

Unter H_0 gilt dann

$$p_{ij} = p_{i\cdot} p_{\cdot j} \quad \text{für alle } i, j\,.$$

Somit haben wir

$$\begin{array}{lll} H_0\text{:} & p_{ij} = p_{i\cdot} p_{\cdot j} & \text{für alle } i, j \\ H_1\text{:} & p_{ij} \neq p_{i\cdot} p_{\cdot j} & \text{für ein Paar } (i, j)\,. \end{array}$$

Unter H_0 sei N_{ij} die Zufallsvariable

$$N_{ij} = \#\text{ in } A_i \cap B_j\,,$$

somit

$$\begin{aligned} P(N_{11} = n_{11} \wedge \ldots \wedge N_{k\ell} = n_{k\ell}) &= \frac{n!}{n_{11}! \cdots n_{k\ell}!}\, p_{11}^{n_{11}} \ldots p_{k\ell}^{n_{k\ell}} \\ &= \frac{n!}{n_{11}! \ldots n_{k\ell}!}\, (p_{1\cdot} p_{\cdot 1})^{n_{11}} \cdots (p_{k\cdot} p_{\cdot \ell})^{n_{k\ell}}\,. \end{aligned}$$

Die Teststatistik X^2 hängt also von den unbekannten Parametern $p_{i\cdot}$, $p_{\cdot j}$ ab. Wir verwenden nun die Maximum Likelihood Methode zur Schätzung dieser Parameter. Wir haben

$$P(N_{11} = n_{11} \wedge \ldots \wedge N_{k\ell} = n_{k\ell}) = \max\,,$$

das heißt

$$\prod_{i=1}^{k} \prod_{j=1}^{\ell} (p_{i\cdot} p_{\cdot j})^{n_{ij}} = \max$$

unter den Nebenbedingungen $\sum\limits_{i=1}^{k} p_{i\cdot} = 1\,,\ \sum\limits_{j=1}^{\ell} p_{\cdot j} = 1\,.$

Mit der Methode von Lagrange aus der Analysis ergeben sich die Schätzer

$$\widehat{p}_{i\cdot} = \frac{n_{i\cdot}}{n}\,,\ \widehat{p}_{\cdot j} = \frac{n_{\cdot j}}{n}$$

und daraus die Teststatistik

$$X^2 = \sum_{i,j} \frac{(n_{ij} - n\widehat{p}_{i\cdot}\widehat{p}_{\cdot j})^2}{n\widehat{p}_{i\cdot}\widehat{p}_{\cdot j}}\,.$$

Diese Statistik ist χ^2-verteilt mit $k\ell - (k-1) - (\ell-1) - 1 = (k-1)(\ell-1)$ Freiheitsgraden, also

$$X^2 \sim \gamma_{\frac{1}{2},\frac{(k-1)(\ell-1)}{2}} \text{ gamma-verteilt.}$$

Die Entscheidungsregel besagt:

$$X^2 \geq \chi^2_{1-\alpha,(k-1)(\ell-1)} \Longrightarrow H_1\,,$$

wobei $\chi^2_{1-\alpha,(k-1)(\ell-1)}$ das $(1-\alpha)$-Quantil der χ^2-Verteilung ist.

Beispiel. In userem Ausgangsbeispiel berechnet man $X^2 = 59,63$. Beim Niveau $\alpha = 0,05$ ist $\chi^2_{0,95;6} = 12,59$. Die Nullhypothese der Unabhängigkeit wird abgelehnt.

B. Fisher Test bei 2×2-Tafeln

Bei Merkmalen mit jeweils zwei Ausprägungen gibt es einen weiteren sehr bekannten Test auf Unabhängigkeit.

Beispiel. Gegeben sei folgende Kontinggenztafel

A \ B	B_1	B_2	Σ
A_1	2	8	10
A_2	3	7	10
Σ	5	15	

Wiederum soll gestestet werden:

$$\begin{aligned} H_0: &\quad X, Y \text{ unabhängig} \\ H_1: &\quad X, Y \text{ abhängig}. \end{aligned}$$

Die Idee besteht darin, alle 2×2-Tafeln mit den gleichen Randhäufigkeiten zu betrachten. In unserem Beispiel sind dies

$$\begin{array}{|c|c|}\hline 0 & 10 \\ \hline 5 & 5 \\ \hline \end{array} \quad \begin{array}{|c|c|}\hline 1 & 9 \\ \hline 4 & 6 \\ \hline \end{array} \quad \begin{array}{|c|c|}\hline 2 & 8 \\ \hline 3 & 7 \\ \hline \end{array} \quad \begin{array}{|c|c|}\hline 3 & 7 \\ \hline 2 & 8 \\ \hline \end{array} \quad \begin{array}{|c|c|}\hline 4 & 6 \\ \hline 1 & 9 \\ \hline \end{array} \quad \begin{array}{|c|c|}\hline 5 & 5 \\ \hline 0 & 10 \\ \hline \end{array}.$$

Allgemein sei die beobachtete Tafel

	B_1	B_2	$\sum$
A_1	a	b	$a+b$
A_2	c	d	$c+d$
$\sum$	$a+c$	$b+d$	

dann sind alle Tafeln mit gleicher Randhäufigkeit von der Form mit $0 \leq x \leq \min(a+b, a+c)$.

x	$a+b-x$	$a+b$	Ziehung
$a+c-x$	$d-a+x$	$c+d$	
$a+c$	$b+d$		
rot	weiß		

Als Modell stellen wir uns eine Urne vor mit $a+c$ roten Kugeln und $b+d$ weißen Kugeln. Man ziehe $a+b$ Kugeln ohne Zurücklegen. Die Merkmale entsprechen also folgenden Ereignissen.

A: Ziehung der Kugel

- A_1: Kugel in der Ziehung
- A_2: Kugel nicht in der Ziehung

B: Kugelfarbe

B_1: rot
B_2: weiß.

Als Teststatistik wählen wir

$$T = \#\text{rote Kugeln in Stichprobe.}$$

Unter H_0 ist T hypergeometrisch verteilt mit

$$P(T = x) = \frac{\binom{a+c}{x}\binom{b+d}{a+b-x}}{\binom{n}{a+b}} .$$

Die Entscheidungsregel ist demnach

$$T \leq c_{\alpha/2} \text{ oder } T \geq C_{1-\alpha/2} \Longrightarrow H_1 ,$$

wobei $c_{\alpha/2}$ das $\alpha/2$-Quantil der hypergeometrischen Verteilung ist.

Beispiel. In unserem Beispiel ist $T = 2$. Für $\alpha = 0,05$ erhalten wir $c_{\alpha/2} = 0$, $c_{1-\alpha/2} = 5$. H_0 kann also nicht abgelehnt werden.

Literatur

H. Büning, G. Trenkler: Nichtparametrische statistische Methoden 1978, de Gruyter.

W.J. Conover: Practical Nonparametric Statistics, 2nd edition 1971, John Wiley.

W. Feller: Probability Theory and its Applications, vol. I, 1950, John Wiley.

H.-O. Georgii: Stochastik. Einführung in die Wahrscheinlichkeitstheorie und Statistik, 2. Auflage 2007, de Gruyter.

G.R. Grimmett, D.R. Stirzaker: Probability and Random Processes, 2nd edition, 1992, Clarendon Press.

A. Klenke: Wahrscheinlichkeitstheorie, 2. Auflage 2008, Springer.

U. Krengel: Einführung in die Wahrscheinlichkeitstheorie und Statistik, 8. Auflage 2008, Vieweg-Verlag.

H. Toutenburg, C. Heumann: Deskriptive Statistik, 5. Auflage 2006, Springer.