Nicht-parametrische statistische Verfahren

von Peter Pfaffelhuber

Version: 7. Dezember 2015

Die statistischen Verfahren, die wir bisher kennengelernt haben, basieren auf statistischen Modellen, die immer eine bestimmte Klasse von Verteilungen voraussetzen; man erinnere sich beispielsweise an die Normalverteilungsannahme bei der linearen Regression. Ist eine solche Annahme nicht gerechtfertigt oder verletzt, so greift man auf nicht-parametrische Verfahren zurück. Das statistische Modell ist hier viel flexibler, so dass unter sehr wenigen Grundannahmen Aussagen getroffen werden können. Formal ist es so: Die Parametermenge $\mathcal{P}$ eines statistischen Modells $(X, \{\mathbb{P}_\theta : \theta \in \mathcal{P}\})$ ist oftmals eine Teilmenge eines $\mathbb{R}^k$, etwa beim Normalverteilungsmodell $(X, \{\mathbb{P}_{\theta=(\mu,\sigma^2)} = N(\mu, \sigma^2)^n : \mu \in \mathbb{R}, \sigma^2 \in \mathbb{R}_+\})$. Ist diese Annahme zu restriktiv, so müssen wir $\mathcal{P}$ als viel größere Menge annehmen, so dass $\mathcal{P}$ keine Teilmenge eines $\mathbb{R}^k$ mehr ist. In genau diesem Fall spricht man von nicht-parametrischer Statistik. Etwa könnte $\mathcal{P} = \{\theta : \mathbb{R} \to \mathbb{R}_+ \text{ Dichte bzgl } \lambda\}$ die Menge der regulären, stetigen Modelle (mit $E = \mathbb{R}$) bezeichnen oder $\mathcal{P} = \{\theta : \mathbb{R} \to \mathbb{R}_+ \text{ Dichte bzgl } \lambda \text{ mit } \theta(m + x) = \theta(m - x) \text{ für ein m}\}$ die Menge der bezüglich $m \in \mathbb{R}$ symmetrischen regulären, stetigen Modelle. Wir wollen in diesem Abschnitt statistische Verfahren mit solchen *großen* Parametermengen $\mathcal{P}$ angeben.

1 Quantil-Tests

Wir beginnen mit dem einfachen Beispiel eines Tests auf ein Quantil. Wir verwenden das statistische Modell $(X, \{\mathbb{P}_\theta^n : \theta \in \mathcal{P}\})$ mit $\mathcal{P} = \{\theta : \mathbb{R} \to \mathbb{R}_+ \text{ Dichte bzgl } \lambda\}$ der stetigen, regulären Modelle. Wir bezeichnen mit $\kappa_{\theta,p}$ das p-Quantil von $\mathbb{P}_\theta$. Laut Definition gilt

$$\mathbb{P}_\theta(X_1 \leq \kappa_{\theta,p}) = p,$$

außerdem ist $\sum_{i=1}^n 1_{X_i \leq \kappa_{\theta,p}} \sim B(n, p)$. Daraus lässt sich bereits ein Test auf ein vorgegebenes Quantil ableiten.

Beispiel 1 (Schlafdauern). Wir erinnern an das Datenbeispiel aus dem t-Test. Ein Medikament wir daraufhin untersucht, ob es den Schlaf von Probanden verlängert. Dazu wird jeweils die Schlafdauerdifferenz bei zehn Patienten notiert. Man erhält

$$1.9, 0.8, 1.1, 0.1, -0.1, 4.4, 5.5, 1.6, 4.6, 3.4.$$

Wir wollen nun testen, ob der Median (das 50%-Quantil) 0 ist oder nicht.

```
> a<-c(1.9, 0.8, 1.1, 0.1, -0.1, 4.4, 5.5, 1.6, 4.6, 3.4)
> length(a)
[1] 10
> sum(a>0)
```

```
[1] 9
> binom.test(c(9,1), 0.5)

Exact binomial test

data:  c(9, 1)
number of successes = 9, number of trials = 10, p-value = 0.02148
alternative hypothesis: true probability of success is not equal to 0.5
95 percent confidence interval:
 0.5549839 0.9974714
sample estimates:
probability of success
                   0.9
```

Vorzeichentest auf ein Quantil

Annahme	$X_1, ..., X_n$ unabhängig, verteilt nach einer Verteilung $\mathbb{P}_\theta = \theta \cdot \lambda$
Hypothese	$H_0 : \kappa_{\theta,p} = \kappa^*$ für ein vorgegebenes κ^* gegen $H_1 : \kappa_{\theta,p} \neq \kappa^*$
Teststatistik	$Q := \sum_{i=1}^{n} 1_{X_i \leq \kappa^*} \sim B(n,p)$ unter H_0
Ablehnungsbereich	$\{0, \ldots, k, l, \ldots, n\}$ mit $B(n,p)(1, ..., k), B(n,p)(l, ..., n) \leq \alpha/2$
p-Wert	$B(n,p)(1, ..., Q' \wedge Q, Q \vee Q', ..., n)$ mit $Q' = 2np - Q$

2 Tests auf Zufälligkeit

In einer Warteschlange stehen 6 Frauen und 5 Männer, etwa in der Reihenfolge `F`, `M`, `M`, `F`, `M`, `M`, `M`, `F`, `F`, `F`, `F`. Ist diese Folge eine *zufällige* Folge?

Um diese Frage zunächst zu formalisieren, sei $E = \{x \in \{0,1\}^n : x_1 + \cdots + x_n = n_1\}$ und $n_0 := n - n_1$. Weiter bezeichne für $x \in E$

$$r(x) := 1 + \sum_{i=2}^{n} 1_{x_i \neq x_{i-1}}$$

die Anzahl der *Runs* in x. Etwa ist $r(0,1,1,0,1,1,1,0,0,0,0) = 5$. Außerdem bezeichne $\mathbb{P}$ die Gleichverteilung auf E.

Theorem 2 (Verteilung der Anzahl der Runs unter Zufälligkeit). *Es gilt für* $X \sim \mathbb{P}$

und $R = r(X)$

$$\mathbb{P}(R=r) = \begin{cases} 2\frac{\binom{n_0-1}{r/2-1}\binom{n_1-1}{r/2-1}}{\binom{n_0+n_1}{n_0}}, & r \text{ gerade,} \\ \frac{\binom{n_0-1}{(r-1)/2-1}\binom{n_1-1}{(r-3)/2-1} + \binom{n_0-1}{(r-3)/2-1}\binom{n_1-1}{(r-1)/2-1}}{\binom{n_0+n_1}{n_0}}, & r \text{ ungerade.} \end{cases}$$

Beweis. Sei zunächst r gerade. Dann gibt es genau $r/2$ Runs mit 0 und $r/2$ runs mit 1. Sehen wir uns zunächst die $r/2$ Runs mit 0 an. Es gibt insgesamt $\binom{n_0-1}{r/2-1}$ Möglichkeiten, die n_0 möglichen 0er auf $r/2$ verschiedene Runs (der Länge ≥ 1) zu verteilen. (Denn: Jede solche Möglichkeit lässt sich als Reihung, etwa $0|000|0|...|0$ mit genau $r/2-1$ mal $|$ und n_0 mal 0 aufschreiben. Da zwischen zwei $|$ mindestens eine 0 stehen muss, gibt es eine Bijektion dieser Reihungen auf die Darstellungen $|00||...|$, bei der zwischen zwei $|$ (und vor der ersten und nach der letzten) eine 0 entfernt wurde. Die Anzahl dieser Möglichkeiten ist nun gegeben, wenn man die Möglichkeiten abzählt, $r/2-1$ mal $|$ auf insgesamt $r/2-1+n_0-r/2 = n_0-1$ Stellen zu verteilen. Dies ist bekanntlich $\binom{n_0-1}{r/2-1}$.) Die gesuchte Wahrscheinlichkeit ergibt sich nun aus dem Quotienten der Anzahl der Möglichkeiten, $r/2$ Runs mit 0 und $r/2$ Runs mit 1 zu erhalten, und der Gesamtzahl an Möglichkeiten, n_0 mal 0 auf insgesamt n_0+n_1 Plätze aufzuteilen. Der Vorfaktor 2 entsteht dadurch, dass entweder mit 0 oder mit 1 begonnen werden kann.

Für r ungerade bemerken wir, dass entweder $(r+1)/2$ Runs mit 0 und $(r-1)/2$ Runs mit 1 oder umgekehrt vorliegen, wobei die Folge immer mit der Ziffer begonnen werden muss, von der mehr Runs vorhanden sind. Dieselben kombinatorischen Überlegungen wie oben führen auf das Ergebnis. Man beachte hierbei $(r+1)/2-1 = (r-1)/2$ und $(r-1)/2-1 = (r-3)/2$. □

Proposition 3 (Erwartungswert und Varianz von R). *Es gilt, falls* $n_0 \to \infty, n_1 \to \infty$ *und so, dass* $n_0/n \to p, n_1/n \to q := 1-p$

$$\mathbb{E}[R] \xrightarrow{n\to\infty} 2pq,$$
$$\frac{1}{n}\mathbb{V}[R] \xrightarrow{n\to\infty} 4p^2q^2.$$

Beweis. Wir berechnen zunächst für $i, j = 2, ..., n$ mit $j > i$

$$\mathbb{E}[1_{X_i \neq X_{i-1}}] = \frac{n_0}{n}\frac{n_1}{n-1} + \frac{n_1}{n}\frac{n_0}{n-1} = 2\frac{n_0}{n}\frac{n_1}{n-1} = 2pq + O(1/n),$$
$$\mathbb{E}[1_{X_i \neq X_{i-1}} 1_{X_j \neq X_{j-1}}] = \begin{cases} \frac{n_0n_1(n_0-1)+n_1n_0(n_1-1)}{n(n-1)(n-2)} = \frac{n_0n_1}{n(n-1)}, & j = i+1, \\ 4\frac{n_0n_1(n_0-1)(n_1-1)}{n(n-1)(n-2)(n-3)}, & j > i+1. \end{cases}$$

Damit sehen wir, dass

$$\mathbb{V}[1_{X_i \neq X_{i-1}}] = \mathbb{E}[1_{X_i \neq X_{i-1}}] - \mathbb{E}[1_{X_i \neq X_{i-1}}]^2 = 2pq(1-2pq) + O(1/n)$$

und für $j = i+1$

$$\begin{aligned} \mathbb{COV}[1_{X_i \neq X_{i-1}}, 1_{X_j \neq X_{j-1}}] &= \frac{n_0n_1}{n(n-1)} - 4\frac{n_0^2n_1^2}{n^2(n-1)^2} \\ &= \frac{n_0n_1}{n(n-1)}\Big(1 - 4\frac{n_0n_1}{n(n-1)}\Big) = pq(1-4pq) + O(1/n) \end{aligned}$$

sowie für $j > i+1$

$$\begin{aligned}
\tfrac{1}{4}\mathbb{COV}[1_{X_i\neq X_{i-1}}, 1_{X_j\neq X_{j-1}}] &= \frac{n_0n_1(n_0-1)(n_1-1)}{n(n-1)(n-2)(n-3)} - \frac{n_0^2n_1^2}{n^2(n-1)^2} \\
&= \frac{n_0n_1}{n(n-1)}\Big(\frac{(n_0-1)(n_1-1)}{(n-2)(n-3)} - \frac{n_0n_1}{n(n-1)}\Big) \\
&= \frac{n_0n_1}{n(n-1)}\frac{n(n-1)(n_0-1)(n_1-1) - n_0n_1(n-2)(n-3)}{n(n-1)(n-2)(n-3)} \\
&= \frac{n_0n_1}{n(n-1)}\frac{-nn_0n_1 - n^2n_0 - n^2n_1 + 5n_0n_1n + O(n^2)}{n(n-1)(n-2)(n-3)} \\
&= \frac{1}{n}pq(4pq-p-q) + O(1/n^2) \\
&= -\frac{1}{n}pq(1-4pq) + O(1/n^2).
\end{aligned}$$

Daraus ergibt sich für die Varianz

$$\begin{aligned}
\mathbb{V}[R] &= n\mathbb{V}[1_{X_2\neq X_1}] + 2n\mathbb{COV}[1_{X_2\neq X_1}, 1_{X_3\neq X_2}] + n^2\mathbb{COV}[1_{X_2\neq X_1}, 1_{X_4\neq X_3}] + O(1) \\
&= n(2pq(1-2pq) + 2pq(1-4pq) - 4pq(1-4pq)) + O(1) = 4np^2q^2 + O(1)
\end{aligned}$$

□

Bemerkung 4 (R approximativ normalverteilt). Zwar sind die Zufallsvariablen $1_{X_i\neq X_{i-1}}, i = 2, ..., n$ nicht unabhängig, jedoch kann man für R doch einen zentralen Grenzwertsatz angeben. Genauer ist (für große n) die Statistik

$$\frac{R - 2npq}{2\sqrt{n}pq}$$

approximativ $N(0,1)$-verteilt.

Tests auf die Anzahl von Runs in einer zufälligen Folge

Annahme	$X_1, ..., X_n \in \{0,1\}$ mit $X_1 + \cdots + X_n = n_1$
Hypothese	$H_0 : X$ rein zufällig gegen $H_1 : X$ nicht rein zufällig
Teststatistik	$R = 1 + \sum_{i=2}^{n} 1_{X_i\neq X_{i-1}}$ unter H_0 verteilt wie in Theorem 2, approximativ wie in Bemerkung 4.
Ablehnungsbereich	ergibt sich aus der Verteilung von R
p-Wert	ergibt sich aus der Verteilung von R

Beispiel 5 (Zufälligkeit von Zufallszahlgeneratoren). Ein linearer Kongruenzgenerator für Pseudo-Zufallszahlen ist bekanntermaßen gegeben durch die Rekursionsvorschrift (mit einem Startwert $x_0 \in \{0, ..., m-1\}$)

$$x_i = ax_{i-1} + b \mod m.$$

Typischerweise ist hier $m = 2^e$ für eine implementierte Wortlänge e. Eine R-Implemetierung könnte also etwa sein (siehe auch POSIX.1-2001)

```
> myrand<-function(n, seed=1) {
    res<-rep(seed,n)
    for(i in 2:n) {
      res[i] = (res[i-1] * 1103515245 + 12345) %% 32768;
    }
    res/32768
}
```

Wir wollen nun sehen, ob eine so generierte Folge dem Test auf Zufälligkeit standhält. Wir laden zunächst das entsprechende R-Paket.

```
> install.package("randtests")
> library("randtests")
```

In einer Stichprobe der Größe 10000 kann die Zufälligkeit nicht verworfen werden.

```
> x<-myrand(10000)
> runs.test(x)

Runs Test

data:  x
statistic = 1.3544, runs = 5067, n1 = 5092, n2 = 4908, n = 10000,
p-value = 0.1756
alternative hypothesis: nonrandomness
```

3 Der Wald-Wolfowitz-Runs-Test

Wir wenden uns nun – im Gegensatz zur Situation in Abschnitt 1 – Tests mit zwei unabhängigen Stichproben zu. Insbesondere geben wir nun eine nicht-parametrische Alternative zum doppelte t-Test an. Hierzu sei $X_1, ..., X_m$ unabhängig und identisch nach $\mathbb{P}_\theta = \theta \cdot \lambda$ und $Y_1, ..., Y_n$ unabhängig und identisch nach $\mathbb{P}_{\theta'} = \theta' \cdot \lambda$ verteilt. Ziel ist es, den Test $H_0 : \theta = \theta'$ zu testen. Seien hierzu $X_{(1)}, ..., X_{(m)}$ und $Y_{(1)}, ..., Y_{(n)}$ die Ordnungsstatistiken von X und Y. Weiter sei $Z = (X, Y)$ und $Z_{(1)}, ..., Z_{(m+n)}$ die Ordnungsstatistiken der gemeinsamen Stichprobe $X_1, ..., X_m, Y_1, ..., Y_n$. Im weiteren verwenden wir den Vektor

$$W := (1_{Z_{(1)} \in \{X_1, ..., X_m\}}, ..., 1_{Z_{(m+n)} \in \{X_1, ..., X_m\}}).$$

Unter H_0 ist W ein rein zufälliger Vektor in $\{0, 1\}^{m+n}$ mit genau n mal 0 und m-mal 1. Die Verteilung der Anzahl von Runs in W haben wir also im letzten Kapitel hergeleitet. Einzig für

die Berechnung des Ablehnungsbereiches bemerken wir, dass H_0 nur dann abgelehnt wird, wenn die Anzahl der Runs zu klein ist. (Etwa seien alle X_i kleiner als alle Y_j. Dann ist $W = 1, ..., 1, 0, ..., 0$ und die Anzahl der Runs ist 2.)

Beispiel 6 (Der Runs-Test mit t-verteilten Daten). Schon beim Überprüfen von Modellannahmen haben wir untersucht, welche t-Verteilungen von einer Normalverteilung zu unterscheiden sind. Dies wollen wir nochmal vertiefen, indem wir den Runs-Test auf einen Datensatz t- und einen Datensatz normalverteilter Daten anwenden. Wir verwenden hier 10 Freiheitsfrage für die t-Verteilung.

```
> set.seed(1)
> x<-rnorm(100)
> y<-rt(100, df=10)
> perm<-sort(c(x,y), index.return=TRUE)$ix
> w<-as.numeric(perm<=100)
> runs.test(w)

Runs Test

data:  w
statistic = -0.8507, runs = 95, n1 = 100, n2 = 100, n = 200, p-value =
0.395
alternative hypothesis: nonrandomness
```

Der Wald-Wolfowitz-Runs-Test

Annahme	$X_1, ..., X_m$ unabhängig, verteilt nach einer Verteilung $\mathbb{P}_\theta = \theta \cdot \lambda$ $Y_1, ..., Y_n$ unabhängig, verteilt nach einer Verteilung $\mathbb{P}_{\theta'} = \theta' \cdot \lambda$
Hypothese	$H_0 : \theta = \theta'$ gegen $H_1 : \theta \neq \theta'$
Teststatistik	$R := r(W)$, unter H_0 verteilt nach Theorem 2, approximativ with in Bemerkung 4 mit $Z = (X, Y)$ und $W := (1_{Z_{(1)} \in \{X_1, ..., X_m\}}, ..., 1_{Z_{(m+n)} \in \{X_1, ..., X_m\}})$.
Ablehnungsbereich	ergibt sich aus der Verteilung von R
p-Wert	ergibt sich aus der Verteilung von R

4 Der Kruskal-Wallis-Test

Nachdem wir nun eine nicht-parametrische Version des doppelten t-Tests kennengelernt haben, kommt nun eine nicht-parametrische Version der einfaktoriellen Varianzanalyse. Wir erinnern daran, dass hierfür Y_{ki} die i-te Messung der k-ten Gruppe ist, wobei wir die Gleichheit

der Verteilungen von p Gruppen testen wollen. Etwas genauer seien hier $Y_{k\bullet} = Y_{k1}, ..., Y_{kn_k}$ unabhängig und nach $\mathbb{P}_{\theta_k} \sim \theta_k \cdot \lambda$ verteilt, $k = 1, ..., p$. Wie im Wald-Wolfowitz-Test definieren wir $Y_{\bullet\bullet} = (Y_{ki})_{k=1,...,p,i=1,...,n_k}$ und $Z = Y_1, ..., Y_n$ die als Vektor geschriebenen Daten $Y_{\bullet\bullet}$. Für die Ordnungsstatistiken $Z_{(1)}, ..., Z_{(n)}$ verwenden wir den Vektor $W = (W_1, ..., W_p)$ mit

$$R_k = \sum_{i=1}^{n} i1(Z_{(i)} \in \{Y_{k1}, ..., Y_{kn_k}\}),$$

d.h. R_k ist die Summe der Ränge der Größen $Y_{k\bullet}$ in Z. Nun ist die Summe aller Ränge immer gleich

$$\sum_{k=1}^{p} R_k = \sum_{i=1}^{n} i \sum_{k=1}^{p} 1(Z_{(i)} \in \{Y_{k1}, ..., Y_{kn_k}\}) = \sum_{i=1}^{n} i = \binom{n+1}{2}.$$

Gilt außerdem $H_0 = \theta_1 = \cdots = \theta_p$, so gilt für die erwartete Summe der Ränge von Gruppe k

$$\mathbb{E}[R_k] = \sum_{i=1}^{n} i\mathbb{P}(Z_{(i)} \in \{Y_{k1}, ..., Y_{kn_k}\}) = \frac{n_k}{n}\binom{n+1}{2} = \frac{n_k(n+1)}{2}.$$

Damit können wir nun den Kruskal-Wallis-Test angeben. Allerdings ist die Verteilung der Teststatistik S (siehe unten) nur für kleine Werte von p einfach anzugeben.

Kruskal-Wallis-Test (nicht-parametrische einfaktorielle Varianzanalyse)

Annahme	$Y_{k1}, ..., Y_{kn_k}$ unabhängig, nach $\mathbb{P}_{\theta_k} = \theta_k \cdot \lambda$ verteilt, $k = 1, ..., p$
Dabei sind	
$Y_{11}, \ldots, Y_{pn_p}$	gegebene Merkmalsausprägungen eines Merkmals gemessen in Levels $1, \ldots, p$
Hypothesen	$H_0 : \theta_1 = \cdots = \theta_p$ gegen $H_1 : \theta_k \neq \theta_\ell$ für ein Paar k, ℓ
Teststatistik	$S = \sum_{k=1}^{p} \left(R_k - \frac{n_k(n+1)}{2}\right)^2$
Ablehnungsbereich	durch Verteilung von S gegeben
p-Wert	durch Verteilung von S gegeben

Beispiel 7. Wir verwenden dieselben normalverteilten Daten X und t-verteilten Daten Y aus dem letzten Beispiel. Nun ergibt sich

```
> a<-list(x,y)
> kruskal.test(a)
```

```
Kruskal-Wallis rank sum test

data:  a
Kruskal-Wallis chi-squared = 0.0539, df = 1, p-value = 0.8164
```

Also wird auch hier die Nullhypothese nicht verworfen. R berechnet hier nicht das S von oben, sondern eine normalisierte Version davon, wodurch die Teststatistik approximativ χ^2-verteilt ist mit $p - 1$ Freiheitsgraden.