Rangbasierte Verfahren

12

In den Kap. 9 und 10 haben wir statistische Tests zum Vergleich von zwei Gruppen (t-Tests) oder mehreren Gruppen (ANOVA) kennengelernt. Die Konstruktion der statistischen Tests basierte auf Normalverteilungsannahmen in den zugrunde liegenden Modellen. Bei nicht notwendigerweise glockenförmig verteilten Daten, beispielsweise schiefen Verteilungen, bieten sogenannte *rangbasierte Verfahren* entsprechende Alternativen. Die Grundidee ist, nicht etwa die Rohdaten auszuwerten, sondern zunächst zu ihren *Rängen* überzugehen. Das bedeutet, dass wir den metrischen Informationsgehalt verwerfen und die Beobachtungen im Kontext ihrer Ordnungsstatistik beurteilen. Dieser Übergang zu den Rängen wird die Herleitung der Verteilung entsprechender Teststatistiken unter ganz allgemeinen Verteilungsannahmen ermöglichen.

Als Analogon zum Zweistichproben-t-Test lernen wir in Abschn. 12.1 den *Wilcoxon-Rangsummentest* kennen. In Abschn. 12.2 diskutieren wir den *Kruskal-Wallis-Test,* der als Erweiterung zum Vergleich von zwei oder mehr Gruppen zu sehen ist. Schließlich führen wir den *Wilcoxon-Vorzeichenrangtest* als Analogon des gepaarten t-Tests in Abschn. 12.3 ein. Es liegt jeweils ein nichtparametrisches Modell zugrunde, und wir sprechen daher auch von nichtparametrischen Tests.

12.1 Der Wilcoxon-Rangsummentest

Motivation und Beispiel Im Börsenwesen von Mainhatten kursiert das fiese Bakterium *Bazillus negativus Kursus.* Bei $n = 3$ Börsianerinnen und $m = 7$ Börsianern wurde nach Ansteckung mit dem Bakterium die Zeit bis zum Auftreten der ersten Symptome – Nervosität, Herzrasen, Kontrollverlust – gemessen. Die Inkubationszeiten (in Stunden) seien mit $x_1, \ldots, x_n$ und $y_1, \ldots, y_m$ bezeichnet und waren wie folgt:

M. Messer und G. Schneider, *Statistik,* https://doi.org/10.1007/978-3-662-59339-4_12

Börsianerinnen:	0.5	1.2	2.7				
Börsianer:	2.6	6.1	12.3	13.4	15.3	80.8	112.7

Wir beobachten, dass die Inkubationszeiten bei den Damen tendentiell kürzer sind als bei den Herren. Ist das leicht durch Zufall zu erklären, wenn sich die Inkubationszeiten zwischen Damen und Herren in der Population aller Börsianerinnen und Börsianer eigentlich gar nicht unterscheidet?

Man könnte auf die Idee kommen, den beobachteten Unterschied anhand des Zweistichproben-t-Tests zu quantifizieren. Dazu bemerken wir, dass alle Inkubationszeiten positiv sind. Insbesondere erkennen wir zudem bei den Herren, dass viele Inkubationszeiten um die zehn herum liegen, aber auch zwei sehr hohe Inkubationszeiten von über 80 auftreten. Die Verteilung der Beobachtungen ist also nicht glockenförmig, sondern eher schief, und daher wäre die Modellannahme der Normalverteilung nur schwer zu rechtfertigen. Für dieses Szenario von zwei Stichproben diskutieren wir daher im Folgenden den Rangsummentest von Wilcoxon, der ohne die Annahme der Normalverteilung auskommt.

1. *Wahl eines statistischen Modells:* Für $n, m \in \mathbb{N}\backslash\{0\}$ sei ein statistisches Modell gegeben durch einen Zufallsvektor $\mathfrak{Z} = (X_1, \ldots, X_n, Y_1, \ldots, Y_m)^t$ mit unabhängigen Komponenten, und dabei seien $X_1, \ldots, X_n$ identisch verteilt mit $X_1 \sim \nu_{\vartheta_x}$ und $Y_1, \ldots, Y_m$ identisch verteilt mit $Y_1 \sim \nu_{\vartheta_y}$, und ν_{ϑ_x}, und ν_{ϑ_y} seien Mitglieder der Familie $(\nu_\vartheta)_{\vartheta\in\Theta}$ aller reellwertigen Verteilungen mit stetiger Verteilungsfunktion.
2. *Formulierung der Nullhypothese:* Wir formulieren die Nullhypothese, dass alle Beobachtungen aus derselben Verteilung stammen

$$H_0 : (\vartheta_x, \vartheta_y) \in \{(\vartheta_x, \vartheta_y) \in \Theta \times \Theta \mid \vartheta_x = \vartheta_y\}.$$

3. *Wahl einer Teststatistik:* Die Teststatistik soll erstens die Diskrepanz der Inkubationszeiten zwischen den Gruppen quantifizieren. Zweitens möchten wir ihre Verteilung unter der Nullhypothese berechnen können, und diese Verteilung sollte unter allen Elementen der Nullhypothese identisch sein. Dies leistet die *(Wilcoxon-)Rangsummenstatistik* (Wilcoxon 1945):
   Seien $\mathbf{z} = (x_1, \ldots, x_n, y_1, \ldots, y_m)^t \in \mathbb{R}^{n+m}$. Die Statistik basiert auf der Betrachtung der *Ränge* sämtlicher x_i unter allen Beobachtungen $\mathbf{z}$. Dazu bestimme man den *Rang* (die ‚Position') $R_i = R_i(\mathbf{z}) \in \{1, \ldots, n+m\}$ von x_i unter allen Komponenten von $\mathbf{z}$ via

$$R_i := \sum_{j=1}^{n} \mathbb{1}_{\{x_j \le x_i\}} + \sum_{j=1}^{m} \mathbb{1}_{\{y_j \le x_i\}}.$$

   Beispielsweise finden wir bei $n = m = 2$ mit $x_2 < y_1 < x_1 < y_2$, dass $R_1 = 3$, $R_2 = 1$ etc. Die Ränge R_i sind genau dann paarweise verschieden, wenn die Rohwerte z_i paarweise verschieden sind.
   Als Wilcoxon-Teststatistik wird dann die Summe der Ränge der x_i unter allen $n + m$ Beobachtungen gewählt:

$$S_{n,m}(\mathbf{z}) = \sum_{i=1}^{n} R_i. \tag{12.1}$$

Der Wertebereich von $S_{n,m}$ ist

$$\left\{ \sum_{i=1}^{n} i, \ldots, \sum_{i=m+1}^{n+m} i \right\},$$

denn beispielsweise ergibt sich die kleinstmögliche Rangsumme als $1 + 2 + \cdots + n = (n(n+1))/2$, wenn alle x_i kleiner als alle y_j sind.

Die Idee der Rangsummenstatistik ist folgende: Wenn die x_i vergleichsweise klein (bzw. groß) sind im Vergleich zu den y_j, dann schlägt sich dies in kleinen (bzw. großen) Rängen R_i nieder und führt zu einer vergleichsweise kleinen (bzw. großen) Rangsumme. Extrem kleine oder extrem große Rangsummen sprechen also gegen die Nullhypothese. Damit ist der Rangsummentest sensitiv für die Alternativhypothese, dass die Verteilungen ν_{ϑ_x} und ν_{ϑ_y} gegeneinander verschoben sind. Wenn andererseits ν_{ϑ_x} und ν_{ϑ_y} beispielsweise zwei symmetrische Verteilungen mit gleichem Erwartungswert und unterschiedlicher Standardabweichung beschreiben, dann wird der Wilcoxon-Test typischerweise die Nullhypothese nicht ablehnen.

Übrigens: Manchmal wird anstatt der Rangsummenstatistik die äquivalente *Mann-Whitney*-Statistik betrachtet (Mann und Whitney 1946). Dazu bezeichne V_i den Rang von y_i unter allen $\mathbf{z}$. Dann schreibt sich $S_{n,m}$ als

$$\begin{aligned} S_{n,m}(\mathbf{z}) &= \sum_{i=1}^{n} R_i = \sum_{i=1}^{n} \left[\sum_{k=1}^{n} \mathbb{1}_{\{R_k \le R_i\}} + \sum_{k=1}^{m} \mathbb{1}_{\{V_k \le R_i\}} \right] \\ &= \frac{n(n+1)}{2} + \underbrace{\sum_{i=1}^{n} \sum_{k=1}^{m} \mathbb{1}_{\{V_k \le R_i\}}}_{=:U_{n,m}}. \end{aligned}$$

Die sogenannte Mann-Whitney-Statistik $U_{n,m}$ unterscheidet sich also von $S_{n,m}$ nur um die Konstante $n(n+1)/2$.

Für die Berechnung der Teststatistik im Beispiel des Virus werden den Inkubationszeiten $\mathbf{z} = (x_1, x_2, x_3, y_1, \ldots, y_7)^t$ zunächst wie folgt Ränge zugeordnet:

Rang:	1	2	3	4	5	6	7	8	9	10
Beobachtung:	0.5	1.2	2.6	2.7	6.1	12.3	13.4	15.3	80.8	112.7
Guppe (x oder y):	x	x	y	x	y	y	y	y	y	y

Damit ergibt sich die Rangsumme als

$$S_{n,m}(\mathbf{z}) = 1 + 2 + 4 = 7.$$

Spricht dieser Wert gegen die Nullhypothese? Auf den ersten Blick schon, denn der Wertebereich von $S_{n,m}$ ist $\{6, \ldots, 27\}$, und Werte nahe der Ränder des Wertebereichs sprechen gegen die Nullhypothese.

Um die Verteilung der Teststatistik unter der Nullhypothese zu bestimmen, brauchen wir durch den Übergang von den Beobachtungen z_i zu den Rängen R_i keinerlei Annahmen über die Verteilungen der Daten. Wir geben zwei Möglichkeiten an, eine kombinatorische für feste n, m und eine asymptotische Methode ($n, m \to \infty$).

Kombinatorisch Wir stellen fest, dass unter H_0 der Vektor $(R_1, \ldots, R_n)^t$ verteilt ist wie ein Zufallsvektor im $\mathbb{R}^n$, dessen Komponenten durch n-maliges zufälliges Ziehen ohne Zurücklegen aus $\{1, \ldots, n+m\}$ hervorgehen. Denn unter der Nullhypothese stammen alle Beobachtungen aus der gleichen Verteilung, und somit ist für X_i kein Rang bevorzugt – unabhängig von der zugrunde liegenden Verteilung!

Insgesamt gibt es $\binom{10}{3} = 120$ Möglichkeiten, drei Elemente (die Ränge der x_i) aus zehn Elementen zu wählen. Wir erkennen auch die Symmetrie des Problems: Wenn es genau eine Möglichkeit für die kleinste Rangsumme 6 gibt, dann gibt es auch genau eine Möglichkeit für die größte Rangsumme 27 etc. (vgl. Abb. 12.1).

Ist unsere Rangsumme also extrem, wenn die Nullhypothese stimmt? Ja! Denn aus den 120 Möglichkeiten, drei Ränge zu wählen, gibt es nur vier Kandidaten, die eine mindestens so extreme Rangsumme generieren: Für die Rangsummen 6, 7, 26 und 27 gibt es jeweils nur eine Möglichkeit. Wir erhalten also einen P-Wert von

$$P(7) = \mathbb{P}_{H_0}(\{S_{n,m}(3) \leq 7\} \cup \{S_{n,m}(3) \geq 26\}) = 4/120 < 0.05$$

und können daher die Nullhypothese auf dem 5 %-Niveau ablehnen. Interpretation: Wenn sich in Wirklichkeit die Verteilungen der Inkubationszeiten zwischen Börsianerinnen und Börsianern nicht unterscheiden, dann ist in den obigen Beobachtungen eine recht unwahrscheinliche Konstellation aufgetreten.

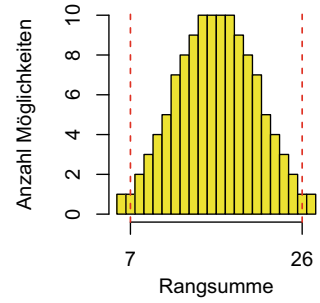


Abb. 12.1 Verteilung von $S_{n,m}(3)$ unter H_0

Wir bemerken, dass für die Herleitung der Verteilung implizit angenommen wurde, dass keine zwei Ränge gleich sind. Das geht mit der Modellannahme einher, dass die zugrunde liegende Verteilungsfunktion stetig ist und somit zwei gleiche Werte mit Wahrscheinlichkeit 1 nicht auftreten.

Asymptotisch Die kombinatorische Herleitung der Verteilung der Rangsummenstatistik $S_{n,m}(\mathfrak{Z})$ unter der Nullhypothese kann rechentechnisch aufwendig werden, wenn n und m groß werden. In diesen Fällen kann man aber folgende Normalapproximation der Verteilung von $S_{n,m}(\mathfrak{Z})$ verwenden. Wir bemerken, dass wir ja schon in Abb. 12.1 eine annähernd glockenförmige Verteilung erkennen.

Satz 12.1 (Asymptotische Normalität der Rangsummenstatistik)
Seien $X, X_2, \ldots, Y_1, Y_2, \ldots$ unabhängige und identisch verteilte Zufallsvariable mit stetiger Verteilungsfunktion. Für $n, m = 1, 2, \ldots$ sei $\mathfrak{Z}_{n,m} := (X_1, \ldots, X_n, Y_1, \ldots, Y_m)^t$. Dann gilt für die Wilcoxon-Statistik $S_{n,m}$ (12.1)

$$\frac{S_{n,m}(\mathfrak{Z}_{n,m}) - \mu_{n.m}}{\sigma_{n,m}} \xrightarrow{d} N(0,1) \quad \textit{für} \quad n, m \to \infty, \quad \textit{mit} \tag{12.2}$$

$$\mu_{n,m} := \frac{n(n+m+1)}{2} = \mathbb{E}[S_{n,m}(\mathfrak{Z}_{n,m})] \quad \textit{und}$$

$$\sigma^2_{n,m} := \frac{nm(n+m+1)}{12} = \mathbb{V}ar(S_{n,m}(\mathfrak{Z}_{n,m})).$$

Heuristik zu Satz 12.1 Wir berechnen zunächst Erwartungswert und Varianz und geben dann eine Heuristik zur Konvergenz von (12.2) an.

Zum Erwartungswert: Da der Rang von X_1 uniform auf $\{1, \ldots, n+m\}$ verteilt ist, gilt

$$\mathbb{E}[R_1(\mathfrak{Z}_{n,m})] = \sum_{k=1}^{n+m} k \cdot \frac{1}{n+m} = \frac{1}{n+m}\left(\frac{(n+m)(n+m+1)}{2}\right) = \frac{n+m+1}{2}.$$

Wegen der Linearität des Erwartungswerts folgt

$$\mathbb{E}[S_{n,m}(\mathfrak{Z}_{n,m})] = n \cdot \mathbb{E}[R_i(\mathfrak{Z}_{n,m})] = \frac{n(n+m+1)}{2}.$$

Zur Berechnung der Varianz von $S_{n,m}$ zerlegen wir

$$\begin{aligned}\mathbb{V}ar(S_{n,m}(\mathfrak{Z}_{n,m})) &= \mathbb{V}ar\left(\sum_{i=1}^{n}\sum_{k=1}^{m}\mathbb{1}_{\{V_k\le R_i\}}\right)\\ &= \sum_{i=1}^{n}\sum_{k=1}^{m}\mathbb{V}ar\left(\mathbb{1}_{\{V_k\le R_i\}}\right)\\ &\quad + \sum_{(k_1,i_1)\neq(k_2,i_2)}\mathbb{C}ov\left(\mathbb{1}_{\{V_{k_1}\le R_{i_1}\}},\mathbb{1}_{\{V_{k_2}\le R_{i_2}\}}\right).\end{aligned}$$

Im Folgenden berechnen wir vier Terme. Term 1 deckt den ersten Summanden ab, und für den zweiten Summanden unterscheiden wir drei Fälle.

1. Es ist

$$\mathbb{V}ar(\mathbb{1}_{\{V_k\le R_i\}}) = \mathbb{E}[\mathbb{1}_{\{V_k\le R_i\}}] - \mathbb{E}[\mathbb{1}_{\{V_k\le R_i\}}]^2 = \frac{1}{2}-\frac{1}{4}=\frac{1}{4},$$

   denn $1^2 = 1$ und $0^2 = 0$, und der Erwartungswert der Indikatorvariable ist gleich der Wahrscheinlichkeit des Ereignisses. Zudem ist die Wahrscheinlichkeit, dass eine Beobachtung kleiner ist als eine andere, gerade 1/2.
2. Für $i_1 \neq i_2$ und $k_1 \neq k_2$ sind $\mathbb{1}_{\{V_{k_1}\le R_{i_1}\}}$ und $\mathbb{1}_{\{V_{k_2}\le R_{i_2}\}}$ unabhängig. Daher verschwindet die Kovarianz.
3. Für $i_1 \neq i_2$ und $k_1 = k_2$ gilt nach dem Verschiebungssatz für die Kovarianz

$$\begin{aligned}&\mathbb{C}ov\left(\mathbb{1}_{\{V_{k_1}\le R_{i_1}\}},\mathbb{1}_{\{V_{k_2}\le R_{i_2}\}}\right)\\ &= \mathbb{E}\left[\mathbb{1}_{\{V_{k_1}\le R_{i_1}\}}\cdot\mathbb{1}_{\{V_{k_2}\le R_{i_2}\}}\right] - \left[\mathbb{1}_{\{V_{k_1}\le R_{i_1}\}}\right]\mathbb{E}\left[\mathbb{1}_{\{V_{k_2}\le R_{i_2}\}}\right]\\ &= \mathbb{P}(\{V_{k_1}\le R_{i_1}\}\cap\{V_{k_2}\le R_{i_2}\}) - \frac{1}{4}\\ &\overset{(*)}{=} \frac{1}{3}-\frac{1}{4}=\frac{1}{12},\end{aligned}$$

   wobei in $(*)$ steckt, dass V_{k_1} mit Wahrscheinlichkeit von 1/3 die kleinste der drei Zufallsvariablen V_{k_1}, R_{i_1} und R_{i_2} ist. Wir bemerken, dass es hier $n(n-1)m$ Summanden gibt.

4. Aus Symmetriegründen findet sich die Kovarianz für $i_1 = i_2$ und $k_1 \neq k_2$ auch als 1/12. Hier gibt es $m(m-1)n$ Summanden.

Insgesamt folgt daher

$$\begin{aligned}\mathbb{V}ar(S_{n,m}(\mathfrak{Z}_{n,m})) &= \frac{1}{4}nm + \frac{1}{12}(n(n-1)m) + \frac{1}{12}(m(m-1)n)\\ &= \frac{1}{12}(3mn + n^2m - nm + nm^2 - nm)\\ &= \frac{1}{12}(nm + n^2m + nm^2)\\ &= \frac{1}{12}(nm(1+n+m)).\end{aligned}$$

Zum Beweis der Konvergenz (12.2) siehe Georgii (2009). Die Idee basiert darauf, die Rangsummenstatistik, die ja eine Summe von nicht unabhängigen Zufallsvariablen ist, durch eine Summe unabhängiger Zufallsvariablen zu approximieren, für die der Zentrale Grenzwertsatz gilt, sodass dann auch die Konvergenz der Rangsummenstatistik anhand eines Slutsky-Argumentes gefolgert werden kann. Dazu bemerken wir noch, dass bei obiger Fallunterscheidung Kovarianzen nur in den Summanden 1., 3. und 4. auftreten. Unter allen n^2m^2 Summanden sind das aber nur wenige, denn

$$nm + n(n-1)m + m(m-1)n < n^2m^2\left(\frac{1}{nm} + \frac{1}{n} + \frac{1}{m}\right) = o(n^2m^2),$$

d. h., die Korrelationen treten nur bei asymptotisch vernachlässigbar vielen Summanden auf.

Die asymptotische Normalität aus Satz 12.1 nutzen wir schließlich zur Formulierung eines asymptotischen Tests.

Lemma 12.2 (Wilcoxon-Rangsummentest)
Seien $X_1, X_2, \ldots, Y_1, Y_2, \ldots$ unabhängige Zufallsvariable, alle X_i seien identisch verteilt gemäß ν_{ϑ_x} und alle Y_j identisch verteilt nach ν_{ϑ_y}, und es seien ν_{ϑ_x} und ν_{ϑ_y} Mitglieder der Familie $(\nu_\vartheta)_{\vartheta\in\Theta}$ aller reellwertigen Verteilungen mit stetiger Verteilungsfunktion. Für $n, m \geq 1$ ist dann ein Modell mit dem Vektor $\mathfrak{Z}_{n,m} := (X_1, \ldots, X_n, Y_1, \ldots, Y_m)^t$ assoziiert. Weiter sei eine Nullhypothese gegeben durch

$$H_0 : (\vartheta_x, \vartheta_y) \in \{(\vartheta_x, \vartheta_y) \in \Theta \times \Theta \mid \vartheta_x = \vartheta_y\}.$$

Zudem sei $\alpha \in (0,1)$, sowie q_α das α-Quantil der $N(0,1)$-Verteilung, $S_{n,m}$ wie in (12.1) und $\mu_{n,m}$ und $\sigma^2_{n,m}$ wie in Satz 12.1. Dann ist die Folge $(T_{n,m})_{n,m}$ via

$$T_{n,m}(z_{n,m}) := \frac{S_{n,m}(z_{n,m}) - \mu_{n,m}}{\sigma_{n,m}}$$

eine Folge von Teststatistiken für einen asymptotischen Test ($n, m \to \infty$) der Nullhypothese H_0 zum Niveau α mit Ablehnungsbereich $\mathcal{R}(\alpha) = (-\infty, q_{\alpha/2}] \cup [q_{1-\alpha/2}, \infty)$.

Die Ausdehnung dieser Idee auf den Vergleich von mehr als zwei Stichproben ist Gegenstand von Abschn. 12.2.

12.2 Der Kruskal-Wallis-Test

Der Kruskal-Wallis-Test ist als nichtparametrische Erweiterung des Wilcoxon-Tests auf zwei oder mehr Gruppen zu sehen – ähnlich der Erweiterung des t-Tests auf die ANOVA. Er arbeitet wieder mit Rangsummen und ist eine Alternative zur ANOVA, wenn man nicht davon ausgehen kann, dass die Beobachtungen in den Gruppen näherungsweise glockenförmig verteilt sind. Wegen der Analogie zum Rangsummentest verzichten wir hier auf die ausführliche Formulierung aller Testschritte.

Es sei $\mathbf{x} = (x_{1,1}, \ldots, x_{1,n_1}, x_{2,1}, \ldots, x_{2,n_2}, \ldots, x_{k,1}, \ldots, x_{k,n_k})^t \in \mathbb{R}^n$ ein Datenvektor von $n = n_1 + \cdots + n_k$ Beobachtungen, stammend aus k Gruppen mit n_i Beobachtungen $x_{i,1}, \ldots, x_{i,n_i}$ in Gruppe i. Wir untersuchen die Nullhypothese, dass die beobachteten Daten Realisierungen aus den gleichen Verteilungen sind.

Den Unterschied zwischen den Gruppen messen wir anhand der Kruskal-Wallis-Statistik H, die die mittleren Ränge der Gruppen jeweils mit dem mittleren Rang aller Beobachtungen vergleicht. Die Denkweise ist analog zur ANOVA, vgl. (10.3), nur dass hier mit den Rängen und nicht mit den Rohdaten argumentiert wird. Es bezeichne wieder $R_{i,j}(\mathbf{x})$ den Rang von $x_{i,j}$ unter allen Komponenten von $\mathbf{x}$, d. h.

$$R_{i,j}(\mathbf{x}) := \sum_{\ell=1}^{k} \sum_{m=1}^{n_\ell} \mathbb{1}_{\{x_{\ell,m} \le x_{i,j}\}}.$$

Der mittlere Rang $\bar{R}_i$ der i-ten Gruppe ist dann

$$\bar{R}_i(\mathbf{x}) := \frac{1}{n_i} \sum_{j=1}^{n_i} R_{i,j}(\mathbf{x}), \tag{12.3}$$

und der mittlere Rang $\bar{R}$ aller Beobachtungen $\mathbf{x}$ ist

$$\bar{R}(\mathbf{x}) := \frac{1}{n} \sum_{i,j} R_{i,j}(\mathbf{x}) = \frac{1}{n} \frac{n(n+1)}{2} = \frac{n+1}{2}. \tag{12.4}$$

Sinnvollerweise hängt $\bar{R}(\mathbf{x})$ nur über die Anzahl n von $\mathbf{x}$ ab. Stammen nun alle Beobachtungen aus der gleichen Verteilung, so wird der mittleren Rang $\bar{R}_i(\mathbf{x})$ von Gruppe i typischerweise nicht allzu stark vom globalen mittleren Rang $\bar{R}(\mathbf{x})$ abweichen. Die Kruskal-Wallis-Statistik H quantifiziert diese Abweichung nun bezüglich aller Gruppen via

$$H(\mathbf{x}) := \frac{12}{n(n+1)} \sum_{i=1}^{k} n_i \left(\bar{R}_i(\mathbf{x}) - \bar{R}(\mathbf{x})\right)^2. \tag{12.5}$$

Sind die Gruppenmittelwerte gegeneinander verschoben, so nimmt $H(\mathbf{x})$ einen großen Wert an. Ein großer Wert $H(\mathbf{x})$ spricht gegen die Nullhypothese.

Eine häufig verwendete äquivalente Schreibweise für H ist

$$H(\mathbf{x}) := \frac{12}{n(n+1)} \left[\sum_{i=1}^{k} n_i \bar{R}_i^2(\mathbf{x})\right] + 3(n+1),$$

welche sich unmittelbar durch Auflösen des Quadrats in (12.6) und unter Ausnutzung von (12.4) ergibt. Wir formulieren folgendes Lemma:

Lemma 12.3 (Kruskal-Wallis-Test)
Seien $(X_{i,j})_{i=1,\ldots,k,\, j=1,2,\ldots}$ unabhängige Zufallsvariable, für jedes $i = 1, \ldots, k$ seien $(X_{i,j})_{j=1,2,\ldots}$ identisch verteilt gemäß ν_{ϑ_i}, und jedes ν_{ϑ_i} sei Mitglied der Familie $(\nu_\vartheta)_{\vartheta\in\Theta}$ aller reellwertigen Verteilungen mit stetiger Verteilungsfunktion. Für $n_1, \ldots, n_k \geq 1$ ist dann ein statistisches Modell mit dem Vektor $\mathfrak{X} = (X_{1,1}, \ldots, X_{1,n_1}, X_{2,1}, \ldots, X_{2,n_2}, \ldots, X_{k,1}, \ldots, X_{k,n_k})^t$ assoziiert. Weiter sei eine Nullhypothese gegeben durch

$$H_0 : (\vartheta_1, \ldots, \vartheta_k) \in \{(\vartheta_1, \ldots, \vartheta_k) \in \Theta^k \mid \vartheta_1 = \cdots = \vartheta_k\}.$$

Zudem sei $\alpha \in (0, 1)$, sowie q_α das α-Quantil der $\chi^2(k-1)$-Verteilung. Dann ist die Folge $(H_{n_1,\ldots,n_k})_{n_1,\ldots,n_k}$ via

$$H_{n_1,\ldots,n_k}(\boldsymbol{x}) = \frac{12}{n(n+1)} \sum_{i=1}^{k} n_i \left(\bar{R}_i(\boldsymbol{x}) - \bar{R}(\boldsymbol{x})\right)^2. \tag{12.6}$$

eine Folge von Teststatistiken für einen asymptotischen Test ($n_1, \ldots, n_k \to \infty$) der Nullhypothese H_0 zum Niveau α mit Ablehnungsbereich $\mathcal{R}(\alpha) = [q_{1-\alpha}, \infty)$.

Das Lemma besagt also, dass $H(\mathfrak{X})$ unter der Nullhypothese, dass alle Beobachtungen aus derselben Verteilung stammen, asymptotisch $\chi^2(k-1)$-verteilt ist. Die Nullhypothese wird verworfen, wenn die Auswertung $H(\mathbf{x})$ große Werte annimmt.

Die Annahme der Stetigkeit der Verteilungen stellt sicher, dass alle Ränge $R_{i,j}(\mathbf{x})$ mit Wahrscheinlichkeit 1 paarweise verschieden sind. Neben der asymptotischen Betrachtungsweise des Lemmas könnten wir bei festgehaltenen Gruppengrößen auch wieder auf kombinatorischem Wege die Verteilung der Teststatistik $H(\mathfrak{X})$ unter der Nullhypothese herleiten,

um damit den Test zu konstruieren. Das funktioniert mit den gleichen Argumenten wie beim Rangsummentest und ist aufgrund der hohen Komplexität ebenfalls wieder nur für nicht zu große Stichproben praktikabel.

Heuristik zu Satz 12.3 Wir betrachten hier nur den Fall von $k = 2$ Gruppen und argumentieren, dass $H(\mathfrak{X})$ unter der Nullhypothese asymptotisch $\chi^2(1)$-verteilt ist. Ganz ähnlich wie beim Übergang vom t-Test zur ANOVA in 10.2, in dem das Quadrat der t-Statistik gerade die F-Statistik war, ist auch das Quadrat der reskalierten Wilcoxon-Statistik aus Lemma 12.2 gerade die Kruskal-Wallis-Statistik. Da die reskalierte Wilcoxon-Statistik unter der Nullhypothese asymptotisch $N(0, 1)$-verteilt ist, ist dessen Quadrat asymptotisch $\chi^2(1)$-verteilt.

Mit der Schreibweise der mittleren Rangsummen schreibt sich die Statistik T aus Lemma 12.2 als

$$T(\mathbf{x}) = \frac{n_1 \bar{R}_1(\mathbf{x}) - n_1(n+1)/2}{[n_1 n_2 (n+1)/12]^{1/2}} = \sqrt{\frac{12}{n+1}} \sqrt{\frac{n_1}{n_2}} \left(\bar{R}_1(\mathbf{x}) - \bar{R}(\mathbf{x})\right).$$

Quadrieren liefert

$$\begin{aligned} T^2(\mathbf{x}) &= \frac{12}{n(n+1)} \frac{n}{n_2} n_1 (\bar{R}_1(\mathbf{x}) - \bar{R}(\mathbf{x}))^2 \\ &= \frac{12}{n(n+1)} \left[n_1 (\bar{R}_1(\mathbf{x}) - \bar{R}(\mathbf{x}))^2 + \frac{n_1^2}{n_2} (\bar{R}_1(\mathbf{x}) - \bar{R}(\mathbf{x}))^2 \right] \\ &= \frac{12}{n(n+1)} \left[n_1 (\bar{R}_1(\mathbf{x}) - \bar{R}(\mathbf{x}))^2 + n_2 (\bar{R}_2(\mathbf{x}) - \bar{R}(\mathbf{x}))^2 \right] = H(\mathbf{x}), \end{aligned}$$

wobei wir in der zweiten Gleichung ausgenutzt haben, dass $n = n_1 + n_2$ und in der dritten, dass $n_1 \bar{R}_1(\mathbf{x}) + n_2 \bar{R}_2(\mathbf{x}) = (n_1 + n_2) \bar{R}(\mathbf{x})$ gilt.

Der allgemeine Fall von k Gruppen lässt sich ähnlich angehen. Hier würde man zunächst die gemeinsame asymptotische Verteilung von $k - 1$ vielen Rangsummen $S_i(\mathbf{x}) = n_i \bar{R}_i(\mathbf{x})$ herleiten und dann H wieder durch diese ausdrücken. Wir bemerken, dass wir bei $k - 1$ Freiheitsgraden landen, weil die letzte Rangsumme durch die Randbedingung $\sum_{i=1}^{k} S_i(\mathbf{x}) = n(n+1)/2$ festgelegt ist. Genaueres dazu ist zum Beispiel in der Originalarbeit von Kruskal und Wallis (1952) zu finden.

12.3 Der Wilcoxon-Vorzeichenrangtest

Im Kontext gepaarter Beobachtungen $(x_1, y_1), \ldots, (x_n, y_n)$ haben wir in Abschn. 9.3.1 den gepaarten t-Test diskutiert. Dabei sind wir zu den Differenzen $d_i := x_i - y_i$ übergegangen. Im dazugehörigen statistischen Modell wurden die Differenzen als unabhängige und normalverteilte Zufallsvariablen $D_1, \ldots, D_n$ modelliert, und der gepaarte t-Test untersuchte die Nullhypothese, dass die D_i einen bestimmten Erwartungswert (zum Beispiel null) haben.

Wie auch beim gepaarten t-Test interessieren wir uns nun dafür, ob die Lage der x_i gegenüber den y_i systematisch verschoben ist. Allerdings möchten wir hier auf die Normalverteilungsannahme verzichten. Das gelingt wieder durch Betrachtung der Ränge.

Beispiel Ein Gitarrenbauer entwickelt eine neue Art von Mechanik, die die Stimmstabilität von Gitarren verbessern soll. Um ihren Nutzen zu bewerten, vergleicht er sie mit einer etablierten Mechanik eines Marktführers an $n = 200$ seiner Instrumente. Jedes Instrument wird zunächst perfekt auf 440 Hertz (Hz) gestimmt und dann einem Belastungstest ausgesetzt, nach dem er misst, um wie viel sich die a-Saite verstimmt hat (Absolutbetrag der Tonhöhenänderung in Hz). Diese Prozedur führt er an jedem Instrument zweimal durch, einmal mit der etablierten, einmal mit seiner neuen Mechanik. Für das i-te Instrument sei die gemessene Verstimmung mit der etablierten Mechanik mit x_i und mit der neuen Mechanik mit y_i bezeichnet. Da die Prozedur für jedes Instrument unter beiden Mechaniken durchgeführt wurde, liegen gepaarte Daten $(x_i, y_i)_{i=1,\dots,n}$ vor (Abb. 12.2a). Fast alle Punkte liegen unterhalb der Diagonalen $\{(x, y)|x = y\}$, die Verstimmung mit der neuen Mechanik ist also typischerweise – zur Freude des Gitarrenbauers – tatsächlich geringer als mit der etablierten Mechanik.

Wir testen die Nullhypothese, dass die Stimmstabilität unter beiden Mechaniken die gleiche ist. Dazu betrachten wir analog zum gepaarten t-Test wieder die n Differenzen $d_i := y_i - x_i$ (Abb. 12.2b). Ihre empirische Verteilung ist zweigipflig, sodass die Modellannahme normalverteilter Differenzen leider nur schwer zu rechtfertigen wäre. Wir führen daher hier Wilcoxon's Vorzeichenrangtest ein (Wilcoxon 1945), der auf explizite Verteilungsannahmen verzichtet.

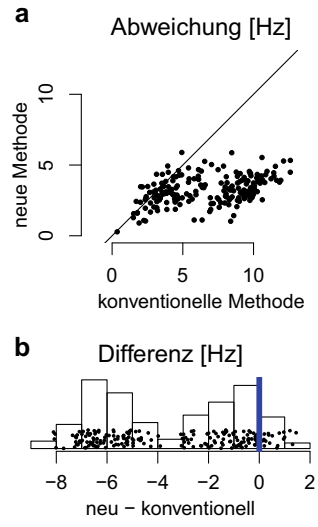


Abb. 12.2 Gepaarte Beobachtungen, **a**: Darstellung im Streudiagramm, jedes Beobachtungspaar ist ein Punkt, **b**: Darstellung der Differenzen

1. *Wahl eines statistischen Modells:* Es sei ein statistisches Modell gegeben durch einen Zufallsvektor $\mathfrak{D} = (D_1, \ldots, D_n)^t$ mit unabhängigen und identisch verteilten Komponenten mit $D_1 \sim \nu_\vartheta$, und ν_ϑ sei ein Mitglied der Familie $(\nu_\vartheta)_{\vartheta\in\Theta}$ aller symmetrischen Verteilungen mit stetiger Verteilungsfunktion. Symmetrisch bedeutet dabei, dass ein $m_\vartheta \in \mathbb{R}$ existiert, sodass $\mathbb{P}_\vartheta(D_1 - m_\vartheta > z) = \mathbb{P}_\vartheta(D_1 - m_\nu < -z)$ für alle $z \in \mathbb{R}$. Dieses Symmetriezentrum m_ϑ ist dann der Median der Verteilung ν_ϑ.
2. *Formulierung der Nullhypothese:* Für $d^{(0)} \in \mathbb{R}$ formulieren wir eine Nullhypothese als

$$H_0 : \vartheta \in \{\vartheta \in \Theta \mid m_\vartheta = d^{(0)}\},$$

   d. h., wir testen die Nullhypothese, dass die Verteilung der Differenzen ein vorgegebenes Symmetriezentrum $d^{(0)}$ besitzt. In der Praxis ist $d^{(0)}$ häufig null, sodass die systematische Verschiebung beider Gruppen verschwindet.
3. *Wahl einer Teststatistik:* Es seien $\mathbf{d} = (d_1, \ldots, d_n)^t \in \mathbb{R}^n$. Wir setzen $y_i := |d_i - d^{(0)}|$ und bezeichnen mit $R_i = R_i(y_i)$ den Rang von y_i unter allen $y_1, \ldots, y_n$. Weicht die Differenz d_i also vergleichsweise stark von $d^{(0)}$ ab, so wird y_i ein großer Rang R_i zugeordnet. Wir definieren die Vorzeichenrangstatistik $W = W_n$ via

$$W(\mathbf{d}) = \sum_{i=1}^{n} R_i \cdot \mathbb{1}_{[d^{(0)},\infty)}(d_i), \tag{12.7}$$

   d. h., wir bilden die Summe aller Ränge derjenigen y_i, für die die Differenz d_i rechts von $d^{(0)}$ liegt. Für den Wertebereich von W finden wir

$$W(\mathbf{d}) \in \left\{0, 1, \ldots, \frac{n(n+1)}{2}\right\}.$$

   Die Idee ist, dass aufgrund der Symmetrie der Verteilung, welche unter der Nullhypothese gilt, jeder Rang mit gleicher Wahrscheinlichkeit rechts wie links vom Symmetriezentrum liegt. Damit erwarten wir unter der Nullhypothese Werte von $W(\mathbf{d})$ in der Mitte des Wertebereichs. Liegen die d_i eher rechts von $d^{(0)}$ und auch noch weit von $d^{(0)}$ entfernt, so wird $W(\mathbf{d})$ groß. Liegen sie links von $d^{(0)}$ und weit von $d^{(0)}$ entfernt, so wird $W(\mathbf{d})$ klein. Beides spricht gegen die Nullhypothese.
   In Abb. 12.3 ist die Berechnung der Statistik $W(\mathbf{d})$ für ein Beispiel mit $n = 7$ dargestellt. Nur eine Beobachtung ist kleiner als $d^{(0)}$, und zwar diejenige, die den zweitkleinsten Abstand zu $d^{(0)}$ hat. Daher geht der Rang $R_i = 2$ nicht in die Berechnung von $W(\mathbf{d})$ ein, und wir finden

$$W(\mathbf{d}) = 1 + 3 + 4 + 5 + 6 + 7 = 26.$$

Um den Test zu konstruieren, bestimmen wir die Verteilung der Teststatistik $W(\mathfrak{D})$ unter der Nullhypothese. Wieder geben wir dafür eine kombinatorische und eine asymptotische Methode an.

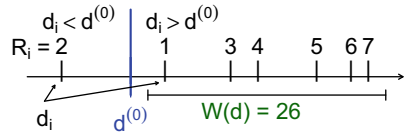


Abb. 12.3 Konstruktion der Teststatistik W

Kombinatorisch Wir diskutieren die kombinatorische Möglichkeit anhand der $n = 7$ Differenzen d_i aus Abb. 12.3. Unter der Nullhypothese ist $W(\mathfrak{D})$ so verteilt wie

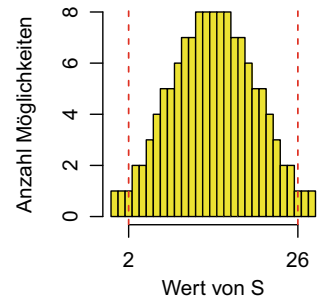
$$S = \sum_{i=1}^{n} i \cdot V_i,$$

wobei $V_1, \ldots, V_n$ unabhängige und $ber(1/2)$-verteilte Zufallsvariablen beschreiben (‚faire Münzwurffolge'). Denn unter der Nullhypothese sind alle D_i unabhängig und identisch verteilt gemäß einer symmetrischen Verteilung mit Symmetriezentrum $d^{(0)}$. Wegen der Symmetrie liegt daher jede Beobachtung unabhängig mit Wahrscheinlichkeit 1/2 rechts wie links von $d^{(0)}$. In der Summe S repräsentiert also i diejenige Differenz, welche den i-größten Abstand zu $d^{(0)}$ besitzt, und V_i indiziert, ob diese Differenz rechts ($V_i = 1$) oder links ($V_i = 0$) von $d^{(0)}$ liegt. Da alle 2^n möglichen Ausgänge der Münzwurffolge $V_1, \ldots, V_n$ die gleiche Wahrscheinlichkeit $1/2^n$ besitzen, müssen wir nur noch für jeden möglichen Ausgang von $S \in \{0, 1, \ldots, 28\}$ lediglich die Anzahl der Münzwurffolgen $V_1, \ldots, V_n$ zählen, die ihn zustande gebracht haben könnten. Durch Vertauschung von 0 und 1 erkennen wir auch die Symmetrie der Verteilung von S, siehe auch Abb. 12.4.

S	Anzahl Mgl.	V_1	V_2	V_3	V_4	V_5	V_6	V_7
0	1	0	0	0	0	0	0	0
1	1	1	0	0	0	0	0	0
2	1	0	1	0	0	0	0	0
3	2	0	0	1	0	0	0	0
		1	1	0	0	0	0	0
4	2	0	0	0	1	0	0	0
		1	0	1	0	0	0	0
⋮								
28	1	1	1	1	1	1	1	1

Wie viele Ausgänge von S sind nun mindestens so extrem wie unsere Beobachtung $W(\mathbf{d}) = 26$? Das sind die sechs Ausgänge $\{0, 1, 2, 26, 27, 28\}$, für die es jeweils eine Möglichkeit gibt. Damit ergibt sich der P-Wert als

Abb. 12.4 Verteilung von S



$$\begin{aligned} P(\mathbf{d}) &= \mathbb{P}_{H_0}(\{W(\mathfrak{D}) \geq 26\} \cup \{W(\mathfrak{D}) \leq 2\}) \\ &= \frac{6}{2^7} \approx 0.047. \end{aligned}$$

Wir können hier also die Nullhypothese auf dem 5 %-Niveau verwerfen.

Asymptotisch Wie im Fall des Rangsummentests kann die kombinatorische Herleitung der Verteilung bei großen Stichproben aufwendig werden. Da sich S aber als Summe unabhängiger Zufallsvariablen schreibt, gibt es auch hier für große n ein Resultat zur asymptotischen Normalität, wie vielleicht auch schon Abb. 12.4 erahnen lässt.

Satz 12.4 (Asymptotische Normalität der Vorzeichenrangstatistik)
Es sei $V_1, V_2, \ldots$ eine Folge unabhängiger und identisch verteilter quadratintegrierbarer Zufallsvariablen mit Erwartungswert $\tilde{\mu} := \mathbb{E}[V_1]$ und positiver Varianz $\tilde{\sigma}^2 = \mathbb{V}ar(V_1)$. Dann gilt für die gewichtete Summe $S_n := \sum_{i=1}^n i V_i$ für $n \to \infty$, dass

$$\frac{S_n - \mu_n}{\sigma_n} \xrightarrow{d} N(0,1), \quad mit \tag{12.8}$$

$$\mu_n := \frac{n(n+1)}{2}\tilde{\mu} = \mathbb{E}[S_n] \quad und \quad \sigma_n^2 := \frac{n(n+1)(2n+1)}{6}\tilde{\sigma}^2 = \mathbb{V}ar(S_n).$$

Heuristik zu Satz 12.4 Wir berechnen aufgrund der Linearität des Erwartungswertes und der Unabhängigkeit der Summanden

$$\mu_n = \sum_{i=1}^n i\tilde{\mu} = \frac{n(n+1)}{2}\tilde{\mu} \quad \text{und} \quad \sigma_n^2 = \sum_{i=1}^n i^2\tilde{\sigma}^2 = \frac{n(n+1)(2n+1)}{6}\tilde{\sigma}^2.$$

Die asymptotische Normalität folgt nach einer Version des Zentralen Grenzwertsatzes nach Lindeberg und Feller, siehe Feller (1971). Dazu bemerken wir, dass die Summanden von S_n zwar unabhängig, aber nicht identisch verteilt sind. Allerdings hat keine der Varianzen eines jeden Summanden zu großen Einfluss auf die Gesamtvarianz, denn $\mathbb{V}ar(iV_i) = O(n^2)$, und andererseits ist die Gesamtvarianz σ_n^2 von der Größenordnung n^3, woraus sich die asymptotische Normalität folgern lässt.

Abschließend formulieren wir den asymptotischen Vorzeichenrangtest, welcher direkt aus dem vorherigen Satz bei Wahl von $V_1 \sim ber(1/2)$ folgt.

Lemma 12.5 (Wilcoxon-Vorzeichenrangtest)
Es seien $D_1, D_2, \ldots$ unabhängige und identisch verteilte Zufallsvariable mit $D_1 \sim \nu_\vartheta$, und ν_ϑ sei ein Mitglied der Familie $(\nu_\vartheta)_{\vartheta\in\Theta}$ aller symmetrischen Verteilungen mit stetiger Verteilungsfunktion. Es sei m_ϑ der eindeutige Median der Verteilung ν_ϑ. Für $n = 1, 2, \ldots$ ist dann ein statistisches Modell assoziiert mit dem Vektor $\mathfrak{D}_n = (D_1, \ldots, D_n)^t$. Es sei eine Nullhypothese gegeben durch

$$H_0 : \vartheta \in \{\vartheta \in \Theta \mid m_\vartheta = d^{(0)}\},$$

und weiter sei $\alpha \in (0, 1)$, sowie q_α das α-Quantil der $N(0, 1)$-Verteilung, W_n wie in (12.7) und $\mu_n = n(n + 1)/4$ und $\sigma_n^2 = n(n + 1)(2n + 1)/24$. Dann ist die Folge $(T_n)_{n=1,2,\ldots}$ via

$$T_n(\boldsymbol{d}_n) := \frac{W_n(\boldsymbol{d}_n) - \mu_n}{\sigma_n}$$

eine Folge von Teststatistiken für einen asymptotischen Test ($n \to \infty$) der Nullhypothese H_0 zum Niveau α mit Ablehnungsbereich $\mathcal{R}(\alpha) = (-\infty, q_{\alpha/2}] \cup [q_{1-\alpha/2}, \infty)$.

Im Beispiel des Gitarrenbauers berechnet sich die auf den Differenzen $\mathbf{d}$ aus Abb. 12.2b basierende Vorzeichenrangstatistik als $W(\mathbf{d}) = 620$. Bei $n = 200$ Beobachtungen erhalten wir $\mu_n = 10050$ und $\sigma_n \approx 820$. Die Teststatistik $T(\mathbf{d}) \approx 11.5$ ist extrem in dem Sinne, dass sie weit in der linken Flanke der Standardnormalverteilung liegt. Der P-Wert ist etwa $P(\mathbf{d}) \approx 10^{-30}$, also winzig klein. Folglich können wir die Nullhypothese zum Beispiel auf dem 1 %-Niveau ablehnen. Die in den Daten beobachtete Verbesserung der Stimmstabilität ist nur extrem schwer durch Zufall zu erklären, wenn sie in Wirklichkeit gar nicht vorhanden ist. Beim Gitarrenbauer kommt erneut Freude auf!