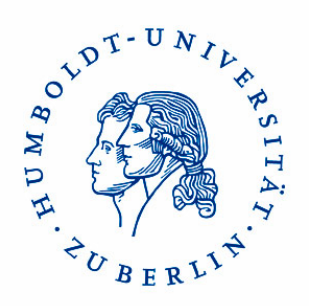
Stochastik–Praktikum
Testtheorie

Thorsten Dickhaus

Humboldt-Universität zu Berlin

11.10.2010



Definition

X: Zufallsgröße mit Werten in Ω,
$(\Omega, \mathcal{F}, (\mathbb{P}_\vartheta)_{\vartheta \in \Theta})$ statistisches Modell

Problem: Teste $H_0 : \vartheta \in \Theta_0$ gegen $H_1 : \vartheta \in \Theta_1 = \Theta \setminus \Theta_0$

(Nicht-randomisierter) Test: $\varphi : \Omega \to \{0, 1\}$, wobei

$$\varphi(x) = 1 \iff: \quad H_0 \text{ verwerfen},$$
$$\varphi(x) = 0 \iff: \quad H_0 \textbf{ nicht } \text{verwerfen}.$$

Randomisierter Test:

$\varphi : \Omega \to [0, 1]$ liefert eine Ablehnwahrscheinlichkeit.

Entscheidungsmuster, Sprechweisen

$\varphi(x)$	H_0 wahr	H_0 falsch
1	Fehler 1. Art	Trennschärfe (power)
0	Signifikanzniveau	Fehler 2. Art

H_0 einfach	$\lvert\Theta_0\rvert = 1$
H_0 zusammengesetzt	$\lvert\Theta_0\rvert > 1$ (z. B. ein Intervall)
φ einseitig	Θ_1 einseitig beschränkt
φ zweiseitig	Θ_1 beidseitig unbeschränkt

Niveau α–Tests, Gütefunktion

$T : \Omega \to \mathbb{R}$ messbare Abbildung, $\alpha \in [0, 1]$ fest vorgegeben, Γ_α so, dass

$$\sup_{\vartheta \in \Theta_0} \mathbb{P}_\vartheta \left(T(X) \in \Gamma_\alpha \right) \leq \alpha.$$

$\varphi(x) = \mathbf{1}_{\Gamma_\alpha}(T(x))$ heißt Test zum Niveau α.

Gütefunktion: $\beta_\varphi(\vartheta) = \mathbb{P}_\vartheta(T(X) \in \Gamma_\alpha)$ für $\vartheta \in \Theta_1$.

Definition

Ein Test φ heißt gleichmäßig bester Test zum Niveau α, falls für jedes $\vartheta_1 \in \Theta_1$ die Güte jedes anderen Niveau α–Tests $\tilde{\varphi}$ kleiner oder gleich der Güte von φ ist, also $\forall \vartheta_1 \in \Theta_1 : \beta_{\tilde{\varphi}}(\vartheta_1) \leq \beta_\varphi(\vartheta_1)$.

Neyman–Pearson–Test

Lemma (Neyman-Pearson)

Zwei einfache Hypothesen $H_0 : \vartheta = \vartheta_0$ gegen $H_1 : \vartheta = \vartheta_1$

Neyman–Pearson–Test:

$$\varphi(x) = \begin{cases} 1\,, & \text{falls } q(x) < c_\alpha \\ \gamma\,, & \text{falls } q(x) = c_\alpha, \\ 0\,, & \text{falls } q(x) > c_\alpha \end{cases} \quad \text{mit } \; q(x) = \frac{L(\vartheta_0|x)}{L(\vartheta_1|x)}$$

Likelihood–Quotient, ist gleichmäßig bester Test zum Niveau α.

$\gamma \in (0,1)$ so, dass $\sup_{\vartheta \in \Theta_0} \mathbb{P}_\vartheta \left(\varphi(X) = 1\right) = \alpha$.

Likelihood–Quotient monoton in T: kritische Werte für T(x)

p–Wert:

Der *p*–Wert zu einem Test φ gibt das kleinste Niveau α_{min} an, zu dem bei Vorliegen der Stichprobe x die Nullhypothese gerade noch abgelehnt werden kann. Manchmal bezeichnet man dies auch als beobachtetes Signifikanzniveau von φ.

In Statistik–Programmen gibt man das Niveau eines Tests nicht explizit vor, sondern erhält diesen *p*–Wert als Ausgabe.

Likelihood–Quotienten–Test:

Beim allgemeinen Testansatz $H_0 : \vartheta \in \Theta_0$ gegen $H_1 : \vartheta \in \Theta_1$ werden beim LQ–Test kritische Werte für

$$\Lambda(x) = \frac{\sup_{\vartheta \in \Theta_0} L(\vartheta|x)}{\sup_{\vartheta \in \Theta} L(\vartheta|x)} \text{ bestimmt.}$$

Übersicht

Beispiel zum Binomialtest

Test der Hypothese, dass eine Münze fair ist, also $H_0 : p = p_0 = 0.5$ versus $H_1 : p \neq p_0$.

```
> binom.test(15,20,p=0.5,alternative="two.sided")

        Exact binomial test

number of successes = 15, number of trials = 20, p-value = 0.04139
alternative hypothesis: true probability of success is not equal to 0.5
95 percent confidence interval: 0.5089541 0.9134285
sample estimates: probability of success = 0.75
```

Die kritischen Werte k und ℓ für den zweiseitigen Binomialtest erhält man mit dem Ansatz

$$\sum_{i=0}^{k} \binom{n}{i} p_0^i (1-p_0)^{n-i} + \sum_{j=n-\ell}^{n} \binom{n}{j} p_0^j (1-p_0)^{n-j} \leq \alpha.$$

Übersicht

Einstichproben–*t*–Test

$X_1, \ldots, X_n$ i. i. d. $\sim \mathbf{N}(\mu, \sigma^2)$, $\sigma^2 > 0$ unbekannt.
$H_0 : \mu = \mu_0$ (oder einseitige Hypothesen)

$$\text{Es gilt:} \quad \Lambda(x) = \left[\frac{\sum_{i=1}^n (x_i - \bar{x})^2}{\sum_{i=1}^n (x_i - \mu_0)^2} \right]^{\frac{n}{2}},$$

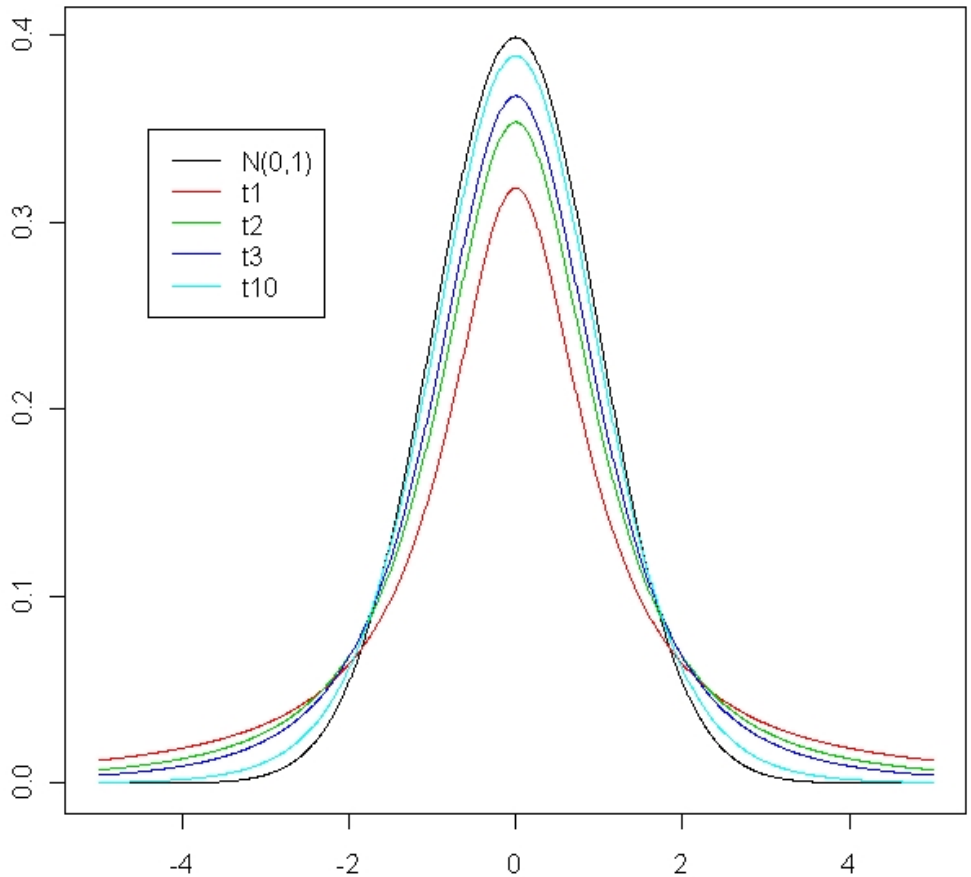
da $\hat{\mu}_{MLE} = \bar{x} = \sum_{i=1}^n x_i/n$ und $\hat{\sigma}^2_{MLE} = \sum_{i=1}^n (x_i - \bar{x})^2/n$.

Als Teststatistik verwendet man nun

$$T(X) = \frac{\sqrt{n(n-1)}(\bar{X} - \mu_0)}{\sqrt{\sum_i (X_i - \bar{X})^2}},$$

da $T(x)$ monotone Transformation von $\Lambda(x)$ und $T(X) \underset{H_0}{\sim} t_{n-1}$ unabhängig von μ und σ (Student, 1908).

Plot: *t*–Verteilungen



Kritische Bereiche

Ablehnbereiche Γ_α des Niveau α *t*–Tests
für einseitige bzw. zweiseitige Hypothesen:

H_0	H_1	Γ_α
$\mu \leq \mu_0$	$\mu > \mu_0$	$\{x \in \Omega : T(x) > t_{n-1;1-\alpha}\}$
$\mu \geq \mu_0$	$\mu < \mu_0$	$\{x \in \Omega : T(x) < -t_{n-1;1-\alpha}\}$
$\mu = \mu_0$	$\mu \neq \mu_0$	$\{x \in \Omega : \lvert T(x)\rvert > t_{n-1;1-\alpha/2}\}$

Einstichproben *t*–Tests in R

Temperaturdatenstichprobe gegeben
Nullhypothese: mittlere Temperatur nicht größer als 37°C

```
> temp<-c(36.8,37.2,37.5,36.9,37.0,37.4,37.9,38.0)
> t.test(temp,alternative="greater",mu=37)

        One Sample t-test

data:  temp
t = 2.1355, df = 7, p-value = 0.03505
alternative hypothesis: true mean is greater than 37
95 percent confidence interval:
 37.03807      Inf
sample estimates:
mean of x
  37.3375
```

Fallzahlabschätzung

Beispielhaft:

Einseitiges Hypothesenpaar $H_0 : \mu \leq \mu_0$ versus $H_1 : \mu > \mu_0$
Analysiere Trennschärfe (power) als Funktion in n

Eine Differenz $\delta = \mu - \mu_0$ soll mit Wahrscheinlichkeit $(1 - \beta) \in (0, 1)$ aufgedeckt werden können, d. h., unter $\mu_1 = \mu_0 + \delta$ soll $\beta_\varphi(\mu_1) = 1 - \beta$ gelten. Sei $\nu := n - 1$.

$$\begin{aligned} 1 - \beta &\stackrel{!}{=} \mathbb{P}_{\mu_1}\left(\sqrt{n}\frac{(\bar{X} - \mu_0)}{s} > t_{\nu;1-\alpha}\right) \\ &= \mathbb{P}_{\mu_1}\left(\sqrt{n}\frac{(\bar{X} - \mu_1)}{s} > t_{\nu;1-\alpha} - \sqrt{n}\frac{\delta}{s}\right), \text{ also} \end{aligned}$$

$$t_{\nu;1-\alpha} - \sqrt{n}\delta/s \stackrel{!}{=} t_{\nu;\beta} \Longleftrightarrow n = (s/\delta)^2 \cdot (t_{\nu;\alpha} + t_{\nu;\beta})^2.$$

Problem: Man benötigt eine Vorschätzung von s !

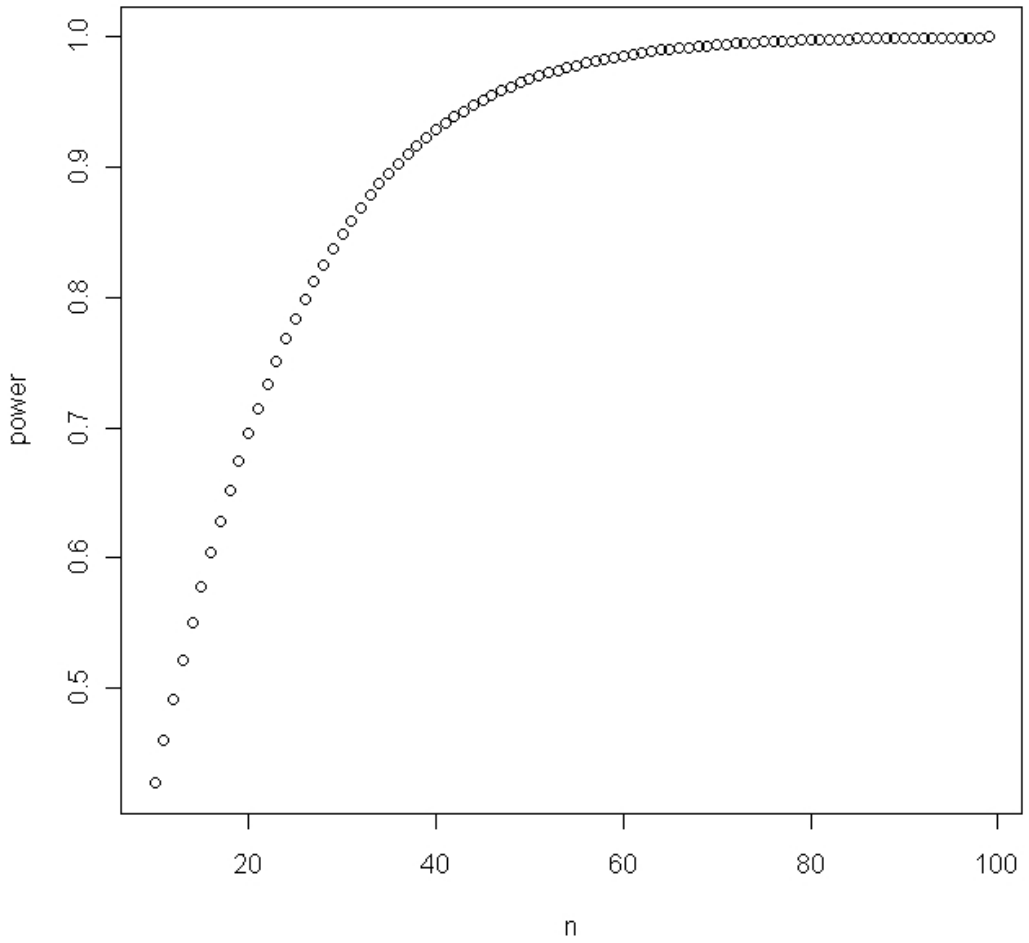
Powerberechnung in R

```
> d<-15;s<-30;r<-d/s;alpha<-0.05;beta<-0.2
> n1<-ceiling((qnorm(1-alpha)+qnorm(1-beta))^2/(r^2));n1
[1] 25
> n2<-ceiling((qt(1-alpha,n1-1)+qt(1-beta,n1-1))^2/(r^2));n2
[1] 27
> n3<-ceiling((qt(1-alpha,n2-1)+qt(1-beta,n2-1))^2/(r^2));n3
[1] 27
> power.t.test(delta=15,sd=30,sig.level=0.05,power=0.80,
+ type="one.sample",alternative="one.sided")

     One-sample t test power calculation

              n = 26.13751
          delta = 15
             sd = 30
      sig.level = 0.05
          power = 0.8
    alternative = one.sided
```

Grafik: Fallzahlabschätzung



Zweistichproben–*t*–Test

Seien $X_1, \ldots, X_n$ i. i. d. $\sim \mathbf{N}(\mu_1, \sigma^2)$ und $Y_1, \ldots, Y_m$ i. i. d. $\sim \mathbf{N}(\mu_2, \sigma^2)$.

Der *t*–Test wird hier für den Vergleich der Mittelwerte angewendet. Teststatistik :

$$T(\vec{X}, \vec{Y}) = \sqrt{\frac{nm}{n+m}} \frac{\overline{X} - \overline{Y}}{\sqrt{\frac{1}{m+n-2}\left[\sum_i (X_i - \overline{X})^2 + \sum_j (Y_j - \overline{Y})^2\right]}}.$$

$T(\vec{X}, \vec{Y}) \underset{H_0}{\sim} t_{m+n-2}$. Die zweiseitige Nullhypothese $H_0 : \mu_1 = \mu_2$ wird verworfen, falls

$$|T(\vec{X}, \vec{Y})| > t_{m+n-2;1-\alpha/2}.$$

Behrens–Fisher–Problem, Welch–Test

Zwei heteroskedastische Stichproben: Behrens-Fisher-Problem

Likelihood–Schätzer:

$$\hat{\sigma}_{MLE} = \sqrt{\frac{1}{n(n-1)}\sum_i (X_i - \overline{X})^2 + \frac{1}{m(m-1)}\sum_j (Y_j - \overline{Y})^2}\,.$$

$T(\vec{X}, \vec{Y}) = (\overline{X} - \overline{Y})/\hat{\sigma}_{MLE}$ ist Behrens–Fisher–verteilt.
Kein exakter Niveau–α–Test bestimmbar!

Welch–Test: Approximative Lösung, Verteilung von T wird approximiert durch eine t–Verteilung mit $\hat{k}$ Freiheitsgraden (Satterthwaite-Approximation):

$$\hat{k} = \frac{\hat{\sigma}^4_{MLE}}{\frac{1}{n^2(n-1)^3}\left[\sum_i (X_i - \overline{X})^2\right]^2 + \frac{1}{m^2(m-1)^3}\left[\sum_j (Y_j - \overline{Y})^2\right]^2}$$

Übersicht

F-Test zum Vergleich zweier Varianzen

$(X_i)_{i=1}^n$ iid. $\sim \mathbf{N}(\mu_1, \sigma_1^2)$ und $(Y_j)_{j=1}^m$ iid. $\sim \mathbf{N}(\mu_2, \sigma_2^2)$; $H_0 : \sigma_1^2 = \sigma_2^2$
Teststatistik: Quotient der empirischen Varianzen

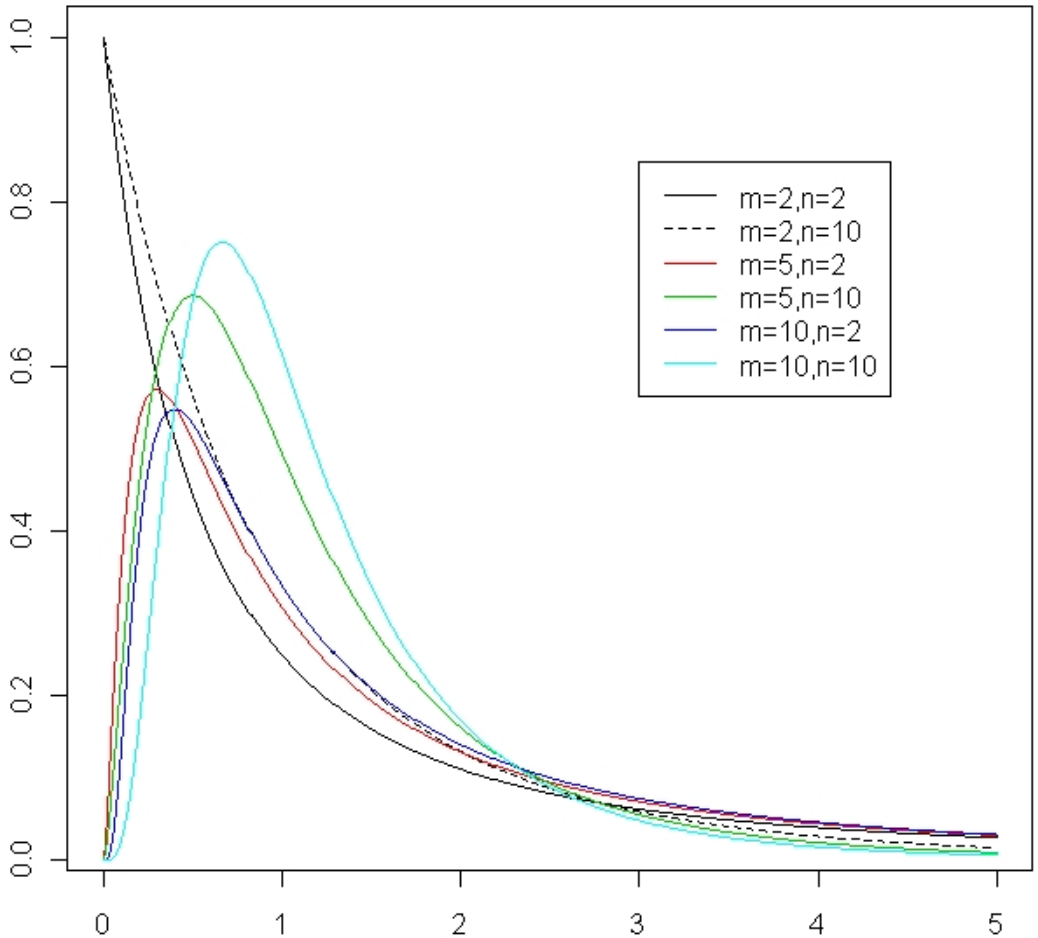
$$F(\vec{X}, \vec{Y}) = \frac{\sum_i (X_i - \overline{X})^2/(n-1)}{\sum_j (Y_j - \overline{Y})^2/(m-1)} \underset{H_0}{\sim} F_{n-1,m-1}.$$

Die Verteilung von F heißt Fisher– oder kurz F–Verteilung mit $n-1$ und $m-1$ Freiheitsgraden.

Annahmebereiche des Niveau α F–Tests:

H_0	Annahmebereiche $\mathbb{R}_{\geq 0} \setminus \Gamma_\alpha$ für F
$\sigma_1^2 = \sigma_2^2$	$\left[\left(F_{n-1,m-1;1-\alpha/2}\right)^{-1}, F_{n-1,m-1;1-\alpha/2}\right]$
$\sigma_1^2 \leq \sigma_2^2$	$\left[0, F_{n-1,m-1;1-\alpha}\right]$
$\sigma_1^2 \geq \sigma_2^2$	$\left[\left(F_{n-1,m-1;1-\alpha}\right)^{-1}, \infty\right)$

Veranschaulichung F–Verteilungen



Übersicht

Nichtparametrische Tests:

Keine konkrete Annahme einer Verteilungsfamilie, also $\nexists k \in \mathbb{N} : \Theta \subseteq \mathbb{R}^k$.

Nichtparametrische Tests sind in parametrischen Modellen nicht effizient.

Das heißt:

Nichtparametrische Tests zum gleichen Signifikanzniveau benötigen größeren Stichprobenumfang als der beste parametrische Test für gleiche Güte (Pitman-Effizienz),

falls das parametrische Modell richtig spezifiziert ist.

χ^2–Anpassungstest

Test der Hypothese $H_0 : F = F_0$ gegen $H_1 : F \neq F_0$

Beobachtungen aufgeteilt in m Klassen mit Häufigkeiten n_j, $j = 1, \ldots, m$.

Vergleiche diese mit den unter F_0 erwarteten Häufigkeiten $n_{j,0}$, $j = 1, \ldots, m$.

Teststatistik:

$$\chi^2 = \sum_{j=1}^{m} \frac{(n_j - n_{j,0})^2}{n_{j,0}} \xrightarrow[H_0]{\mathcal{L}} \chi^2_{m-1}.$$

Ablehnbereich des zugehörigen χ^2-Tests:

$$\Gamma_\alpha = (\chi^2_{m-1;1-\alpha}, \infty)$$

Anwendungsbeispiel: χ^2–Test

Problem: Teste nach 60 Würfen, ob ein Würfel fair ist.

```
> AZ<-c(1,2,3,4,5,6)
> H<-c(7,12,9,15,7,10)
> chisq.test(H)

        Chi-squared test for given probabilities

data:  H
X-squared = 4.8, df = 5, p-value = 0.4408
```

Wilcoxon signed rank Test

Nichtparametrische Alternative zum Einstichproben–t–Test.

$H_0 : m = m_0$ für den Median m.

Definition

Für eine Stichprobe $X_1, \ldots, X_n$ ist die zugehörige i-Ordnungsstatistik definiert durch

$$X_{(i)} := X_k : \ |\{j \in \{1, \ldots, n\} : X_j \leq X_k\}| = i, \quad i = 1, \ldots, n.$$

Die Rangstatistiken geben die Positionen der X_i's in der geordneten Stichprobe an:

$$R_i := |\{1 \leq j \leq n : X_j \leq X_i\}| \ .$$

Wilcoxon signed rank Test (II)

Teststrategie:

- Bilde Rangstatistiken $|X_i - m_0|$.
- Teststatistik:
  Differenz der Summe über alle Ränge mit positivem und der Rangsumme mit negativem Vorzeichen.
- Kritische Werte (einseitig und zweiseitig) sind vertafelt.
- In `R`: Funktion `wilcox.test`.

Mann–Whitney U–Test

$H_0 : \mathbb{P}_X = \mathbb{P}_Y$, also Gleichheit zweier Verteilungen.

Daten: Zwei unabhängige Stichproben $(X_i)_{i=1}^n$ und $(Y_j)_{j=1}^m$

Vorgehen:

- Berechne gruppenweise für (X_i) und (Y_j) Rangsummen R_X und R_Y, wobei die Ränge in der gepoolten Stichprobe $(X_1, \ldots, X_n, Y_1, \ldots, Y_m)$ genommen werden.
- Es gilt **immer:** $R_X + R_Y = (n+m)(n+m+1)/2$ (Gauß). Damit kann man R_X auch schreiben als $R_X = n(n+1)/2 + U$ mit der U–Statistik

$$U = \sum_{i=1}^{n} \sum_{j=1}^{m} \mathbb{1}_{\{X_i > Y_j\}}.$$

Mann–Whitney *U*-Test (II)

Durchführung des Tests:

- Ablehnbereich Γ_α für R_X entspricht Ablehnbereich für U, lediglich um $n(n+1)/2$ verschoben.
- Theorie von *U*-Statistiken geht zurück auf Hoeffding (1948), kritische Werte sind vertafelt.
- In R: `qwilcox(0.025,10,10)`
- Der Mann–Whitney *U*-Test wird auch „Wilcoxon's rank sum test“ genannt.

Weitere wichtige nichtparametrische Tests

- Exakter Fisher-Test (Assoziation kategorieller Merkmale)
- Chi-Quadrat Test auf Assoziation kategorieller Merkmale
- Pitman'scher Permutationstest für allgemeine Zweistichprobenprobleme
- Anpassungstests: Kolmogorov-Smirnov, Cramér-von Mises
- Resamplingverfahren (vgl. SoSe 2011!)