Vorlesung 13a

Statistische Tests

Zusammenfassung und weitere Beispiele

1. Die Grundidee

Die Daten werden aufgefasst als Realisierung einer (mehrdimensionalen) Zufallsvariablen

Man unterstellt eine Hypothese über deren Verteilung und fragt:
Passen die Daten zur Hypothese
oder fallen sie aus dem Rahmen?

Wie wahrscheinlich ist unter der Hypothese ein Ergebnis,
das “mindestens so exotisch ist” wie das beobachtete?

Diese W’keit ist der p-Wert,
zu dem die Hypothese abgelehnt werden kann.

Die Weite des Rahmens, in dem die Daten x

in Bezug auf die Hypothese

(gerade noch bzw. gerade nicht mehr) liegen

wird gemessen durch eine Teststatistik

$$d : D \to \mathbb{R}_+$$

Dabei ist D der *Datenraum*.

2. Beispiele

Beispiel 1

$h :=$ Anteil eines Typs in der Stichprobe

$p :=$ Anteil dieses Typs in der Gesamtpopulation

Zwei Szenarien:

(i) $p = p_0$ ist bekannt. Hypothese: “rein zufälliges Ziehen”

(ii) Die reine Zufälligkeit des Ziehens ist unstrittig.

Hypothese: $p = p_0$

Beidemale eignet sich die Teststatistik

$$d(H) := |H - p_0|$$

Beispiel 2

$x = (x_1, \ldots, x_n)$ eine Stichprobe reellwertiger Daten

Modellannahme: Die x_i sind Realisierungen von unabhängigen, identisch verteilten Zufallsvariablen X_i mit Erwartungswert μ (und Standardabweichung σ)

Hypothese: $\mu = \mu_0$

Teststatistik:

$$T := \frac{|\bar{X} - \mu_0|}{S/\sqrt{n}} \text{ oder auch } \tilde{T} := \frac{|\bar{X} - \mu_0|}{\sigma/\sqrt{n}} \quad \text{(falls } \sigma \text{ bekannt)}$$

Beobachter Wert der Teststatistik:

$t := \frac{|\bar{x}-\mu_0|}{f}$ mit $f := s/\sqrt{n}$ "Standardfehler"

Zu welchem p-Wert kann man Hypothese $\mu = \mu_0$ ablehnen?

Ist n groß, dann nimmt man $\mathbf{P}(|Z| \geq t)$

für ein standardnormalverteiltes Z (z-Test)

Ist n nicht groß (und die Verteilung von X_i annähernd normal)

dann nimmt man $\mathbf{P}(|T_{n-1}| \geq t)$

für T_{n-1} student-verteilt mit $n-1$ Freiheitsgraden (t-Test)

Der theoretische Unterbau ist der

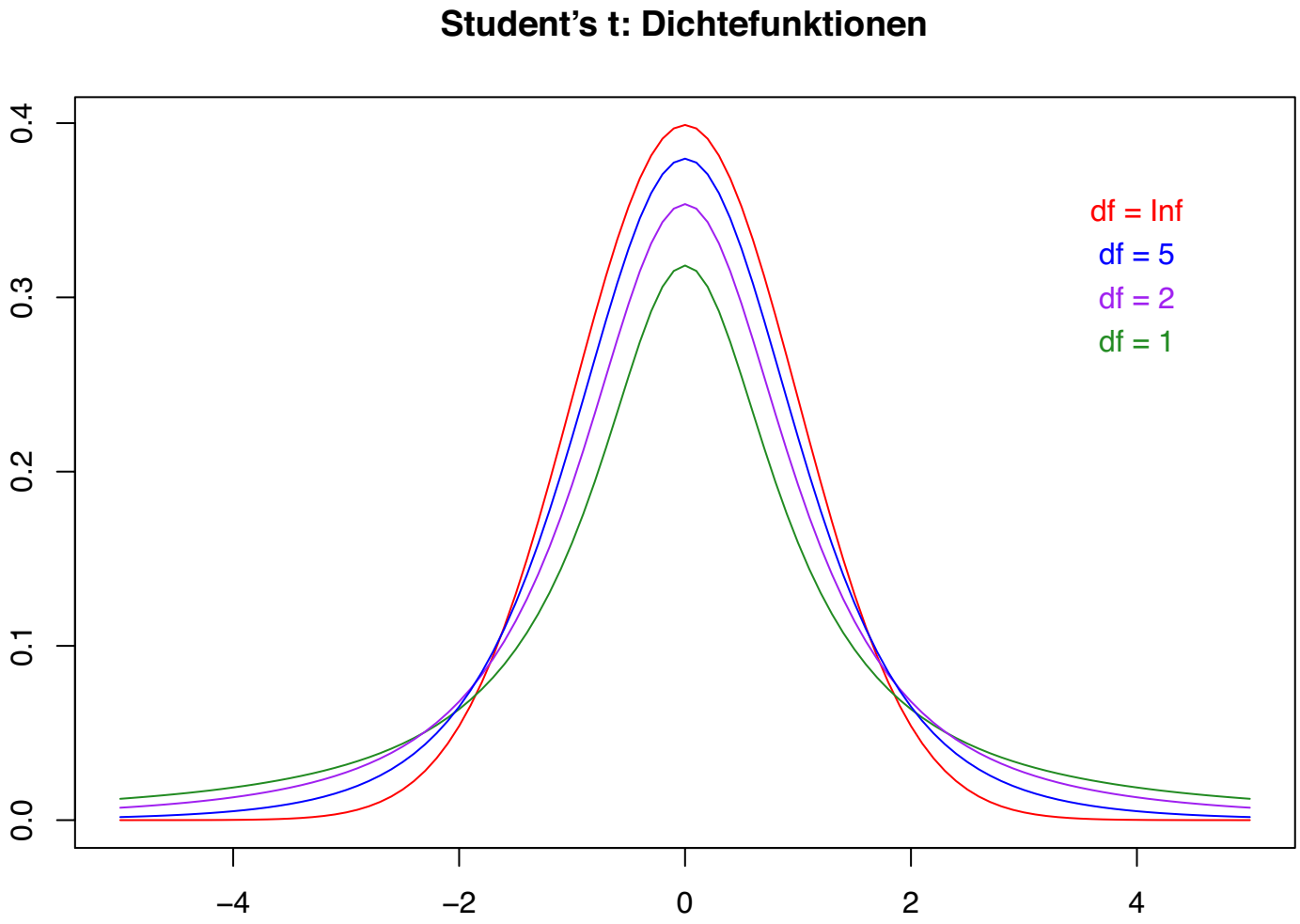
Satz (W. Gosset (alias “Student”, 1908), R. Fisher (1924))
Sind $X_1, \ldots, X_n$ unabhängig und N(μ, σ^2)-verteilt,

dann ist $\quad T := \dfrac{\bar{X} - \mu}{S/\sqrt{n}} \quad$ so verteilt wie

$$T_{n-1} := \frac{N_0}{\sqrt{\frac{1}{n-1}\left(N_1^2 + \cdots + N_{n-1}^2\right)}}$$

mit unabhängigen und N$(0, 1)$ verteilten $N_0, \ldots, N_{n-1}$.

Die Verteilung von T_{n-1} heißt
t-Verteilung (oder Student-Verteilung) mit $n - 1$ Freiheitsgraden.



Dichten von T_{df}

3. Der t-Test für ungepaarte Stichproben.

“Unterscheiden sich zwei Mittelwerte signifikant?”

Die Mittelwerte m_x und m_y

von zwei Stichproben des Umfangs n_x und n_y

unterscheiden sich um 0.5 Einheiten: $|m_y - m_x| = 0.5$.

Ist dieser Unterschied signifikant?

Nach bewährtem Rezept vergleichen wir $|m_x - m_y|$

mit "seinem Standardfehler" f.

Weil wir an unahängige Stichproben denken,

addieren sich die Varianzen:

$$f := \sqrt{f_x^2 + f_y^2}$$

Eine Maßzahl für den “relativen Unterschied” ist also

$$\frac{m_y - m_x}{f}.$$

Interpretiert man die x_i und die y_j

als Realisierungen von

unabhängigen Zufallsvariablen X_i, Y_j,

(mit (X_i) identisch verteilt, (Y_j) identisch verteilt)

dann stellt sich die Frage nach der Verteilung von

$$T := \frac{\bar{Y} - \bar{X}}{F} = \frac{\bar{Y} - \bar{X}}{\sqrt{\frac{S_X^2}{n_x} + \frac{S_Y^2}{n_y}}}$$

$$T := \frac{\bar{Y} - \bar{X}}{F} = \frac{\bar{Y} - \bar{X}}{\sqrt{\frac{S_X^2}{n_x} + \frac{S_Y^2}{n_y}}}$$

Für große n_x, n_y ist T annähernd $\mathrm{N}(0, 1)$-verteilt (wegen des Zentralen Grenzwertsatzes und des Gesetzes der großen Zahlen).

Was aber ist für kleine n_x, n_y?

Hier kommt man zumindest unter der zusätzlichen Annahme weiter, dass die X_i und Y_j normalverteilt sind.

Man kann zeigen, dass T dann annähernd t-verteilt ist

mit einer i.a. nicht ganzzahligen Anzahl von Freiheitsgraden.

Die Formel dafür (die man sich nicht merken muss) findet man auf
http://en.wikipedia.org/wiki/Student's_t-test
im Abschnitt "Unequal sample sizes, unequal variance"

Wichitger ist der praktische Umgang damit in R,

zu dem man dort auf die Frage ?t.test Auskunft bekommt.

Nimmt man überdies an (was nicht immer gerechtfertigt ist!) , dass die beiden Varianzen gleich sind ($\sigma_X^2 = \sigma_Y^2 =: \sigma^2$), dann gibt es eine effizientere Art für die Schätzung von σ:

$$s_{X,Y} := \sqrt{\tfrac{1}{n_x+n_y-2}\left((X_1-\bar{X})^2+\cdots+(Y_{n_y}-\bar{Y})^2\right)}$$

$$= \sqrt{\frac{(n_x-1)s_X^2+(n_y-1)^2 s_Y^2}{n_x+n_y-2}}$$

Der Charme davon ist, dass die entsprechende t-Statistik nicht nur approximativ, sondern exakt t-verteilt ist. Mit einem “Projektionsargument à la Fisher” zeigt man wie im Satz von Gosset und Fisher:

$$\frac{(\bar{X}-\mu_X)-(\bar{Y}-\mu_Y)}{\sqrt{\frac{1}{n_x}+\frac{1}{n_y}}\, s_{X,Y}} \quad \text{ist t}(n_x+n_y-2)\text{-verteilt.}$$

4. Der t-Test für gepaarte Stichproben.

“Ist der Mittelwert (von Differenzen)
signifikant von Null verschieden?”

Welches Schuhsohlenmaterial nutzt sich weniger ab? *

*nach einem Bsp. im Buch “Statistics for experimenters” von G.E.P. Box, W.G.Hunter, J.S. Hunter, Wiley 1978, 2005

Welches Schuhsohlenmaterial nutzt sich weniger ab? *

10 Jungen, 10 Paar Schuhsolen (je 5 vom Material A bzw. B)

2 Varianten der Versuchsplanung (eine schlechte, eine gute):

1) Wähle zufällig 5 der 10 Jungen und gebe ihnen Material A, die anderen bekommen Material B.

2) Jeder Junge erhält Material A für den einen und Material B für den anderen Fuß.

*nach einem Bsp. im Buch “Statistics for experimenters” von G.E.P. Box, W.G.Hunter, J.S. Hunter, Wiley 1978, 2005

Welches Schuhsohlenmaterial nutzt sich weniger ab?

Material \ Junge	1	2	3	4	5	6	7	8	9	10
A	13.2	8.2	10.9	14.3	10.7	6.6	9.5	10.8	8.8	13.3
B	14.2	8.8	11.2	14.2	11.8	6.4	9.8	11.3	9.3	13.6

Dieses Datenmaterial kommt aus einem

gemäß Variante 2 durchgeführten Versuch.

Welches Schuhsohlenmaterial nutzt sich weniger ab?

Material \ Junge	1	2	3	4	5	6	7	8	9	10
A	13.2	8.2	10.9	14.3	10.7	6.6	9.5	10.8	8.8	13.3
B	14.2	8.8	11.2	14.2	11.8	6.4	9.8	11.3	9.3	13.6

Betrachte die Differenzen $D_i := A_i - B_i$

als unabhängig und N(μ, σ^2)-verteilt.

Der t-Test mit der Hypothese $\mu = 0$

ergibt einen p-Wert von 0.008,

also einen hochsignifikanten Unterschied.

(Die Variabilität zwischen den Individuen
geht hier nicht in die t-Statistik ein - und das ist gut so!).

Welches Schuhsohle (A oder B) nutzt sich weniger ab?

Material/Junge	1	2	3	4	5	6	7	8	9	10
A	13.2	8.2	10.9	14.3	10.7	6.6	9.5	10.8	8.8	13.3
B	14.2	8.8	11.2	14.2	11.8	6.4	9.8	11.3	9.3	13.6

Eine unpassende Analyse wäre es, hier auf die Paarung zu vergessen.

Dann geht auch die Variabilität zwischen den Individuen
in die Schätzung der Varianz ein,
was den Wert von f ungebührlich groß
und den Wert der t-Statistik klein macht.
Der t-Test für ungepaarte Stichproben ergibt einen
p-Wert von 0.72 - damit kann man nichts anfangen.

5. Der Wilcoxon-Test.

Wie untypisch ist die Lage der Ränge?

Wie eben zuvor geht es um einen Test der Hypothese,

dass zwei Stichproben

aus derselben Verteilung (auf $\mathbb{R}$) kommen,

gegen die Alternative, dass sich die beiden Verteilungen

durch eine Verschiebung unterscheiden.

Die folgende Idee kommt ganz

ohne spezielle Verteilungsannahme aus:

Man ordnet die $n_x + n_y$ Werte der Größe nach

und ersetzt sie durch ihre Ränge $R(x_i)$, $R(y_j)$.

(der kleinste Wert bekommt den Rang 1, der zweitkleinste den Rang 2,...).

Dann beobachtet man die *Rangsumme* $w := \sum_{i=1}^{n_x} R(x_i)$

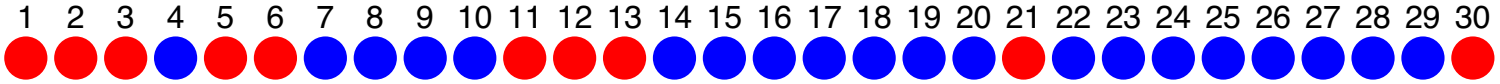
und fragt: Wie wahrscheinlich ist eine

mindestens so "randständige" Rangsumme

bei rein zufälliger Auswahl von n_x Elementen

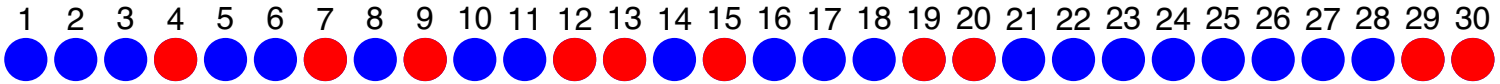
aus der Menge $\{1, \ldots, n_x + n_y\}$?

Die Raenge der x_i und der y_j



Rangsumme der x_i = 104

Eine zufaellige Permutation



Rangsumme der x_i in der Permutation = 158

Die “beobachtete” Rangsumme war 104.

Die minimale mögliche Rangsumme

einer “roten Teilstichprobe” ist $1 + \ldots + 10 = 55$.

Ihre maximale mögliche Rangsumme ist

$$21 + \ldots + 30 = 255.$$

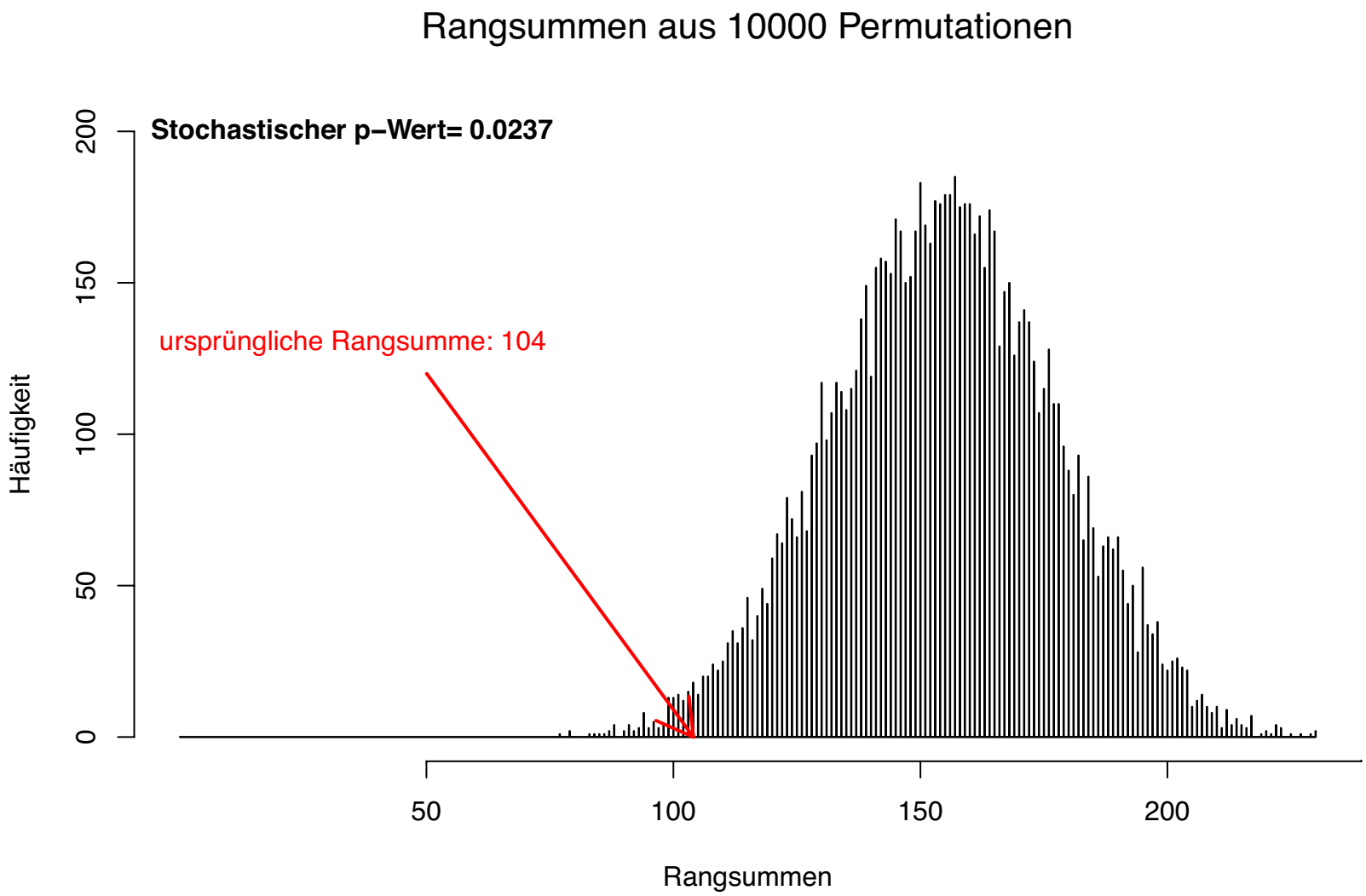
Wir ziehen 10000 mal eine Stichprobe der Größe 10 (aus 30)

und notieren deren Rangsumme.

Der *stochastische p-Wert* ist die relative Häufigkeit

der Ergebnisse, für die sich eine Rangsumme

≤ 104 oder $\geq 255 - (104 - 55)$ ergibt.


Rangsummen aus 10000 Permutationen
Stochastischer p−Wert= 0.0237
ursprüngliche Rangsumme: 104
Häufigkeit
0
50
100
150
200
50
100
150
200
Rangsummen