Vorlesung 13b

Statistische Tests

Zusammenfassung und weitere Beispiele

1. Die Grundidee

Die Daten werden aufgefasst als Realisierung

einer zufälligen Stichprobe

Man fragt:

Passt diese Realisierung zu einer bestimmten

Hypothese über ihre Verteilung

oder fällt sie diesbezüglich aus dem Rahmen?

Wie wahrscheinlich ist unter der Hypothese ein Ausgang,

der “mindestens so exotisch ist” wie der beobachtete?

Diese W’keit ist der p-Wert,

zu dem die Hypothese abgelehnt werden kann.

Die Weite des Rahmens, in dem die Daten

$$x = (x_1, \ldots, x_n) \in S$$

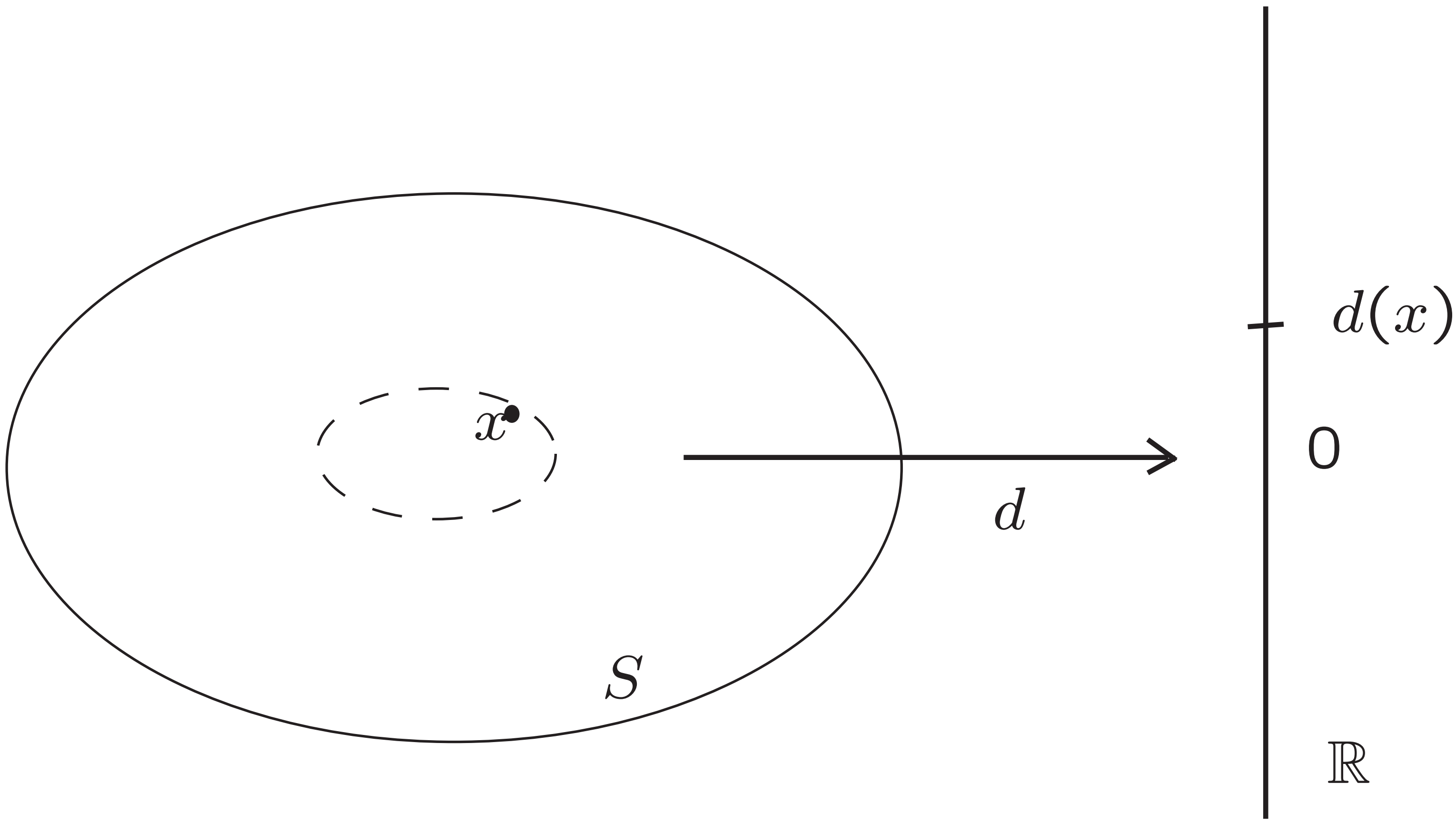
in Bezug auf die Hypothese

(gerade noch bzw. gerade nicht mehr) liegen,

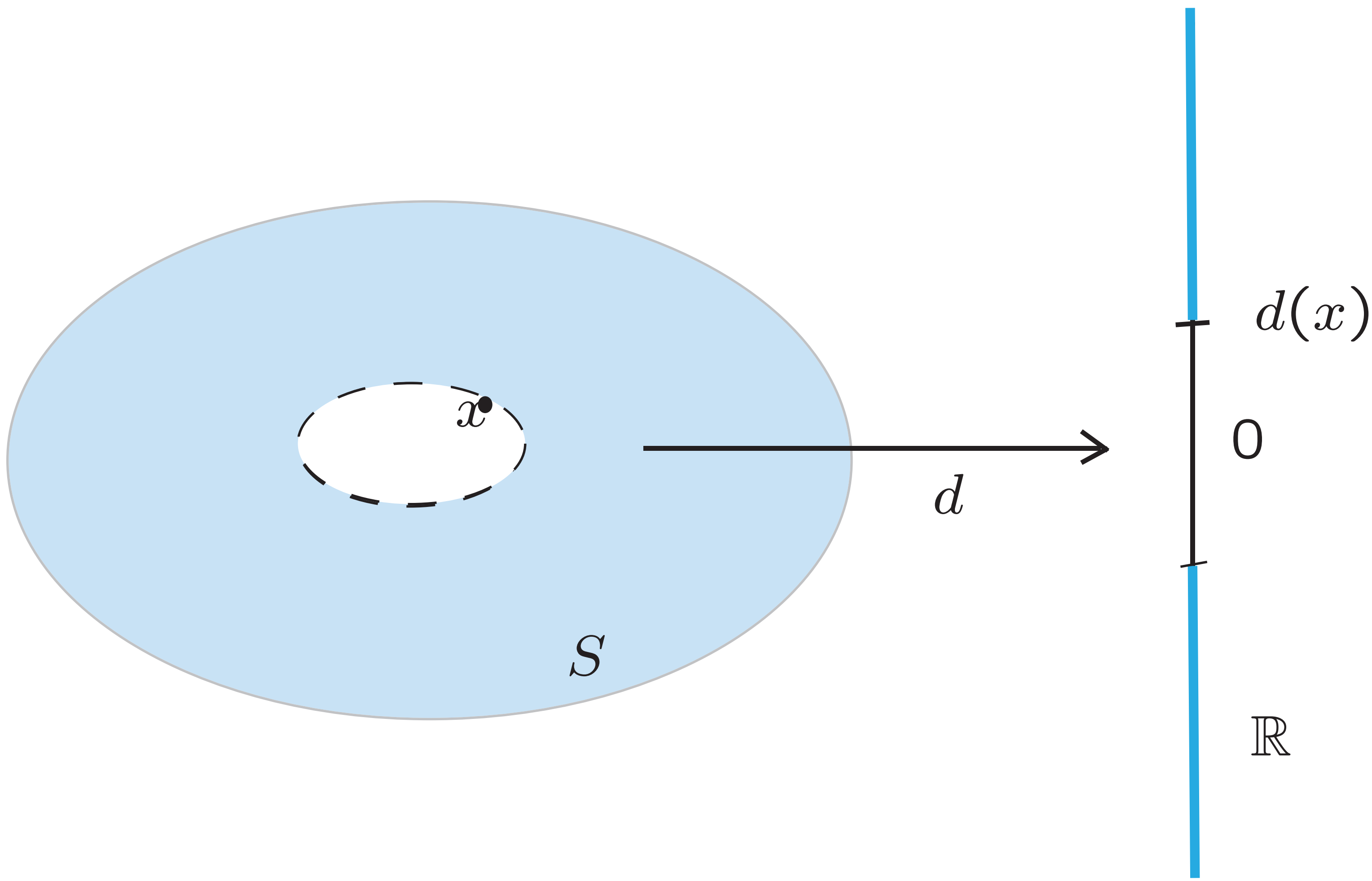
wird gemessen durch eine Teststatistik

$$d : S \to \mathbb{R}$$

Dabei ist S der *Datenraum* ("Stichprobenraum")



x
S
d
$d(x)$
0
$\mathbb{R}$


x
S
d
$d(x)$
0
$\mathbb{R}$

2. Beispiele

Beispiel 1

$p :=$ Anteil eines Typs in der Gesamtpopulation

$h :=$ Anteil dieses Typs in der Stichprobe

Zwei Szenarien (i) und (ii) haben wir schon kennengelernt:

(i) Die reine Zufälligkeit des Ziehens ist unstrittig.

Hypothese: $p = p_0$

(ii) $p = p_0$ ist bekannt.

Hypothese: "rein zufälliges Ziehen"

Beidemale eignet sich die Teststatistik

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

Beispiel 1

$p :=$ Anteil eines Typs in der Gesamtpopulation

$h :=$ Anteil dieses Typs in der Stichprobe

Zum Szenario (ii):

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

Man erinnere sich an das Beispiel mit den 167 Formularen,

80 davon rot und 87 davon blau “verpackt”

(d.h. $p_0 = 80/167$ *ist bekannt).*

Aber wurde wirklich rein zufällig gezogen? ...

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:

$$T := \frac{\bar{X} - \mu_0}{S/\sqrt{n}}$$

Beobachter Wert der Teststatistik:

$$t := \frac{\bar{x}-\mu_0}{f} \quad \text{mit } f := s/\sqrt{n} \quad \text{“Standardfehler”}$$

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

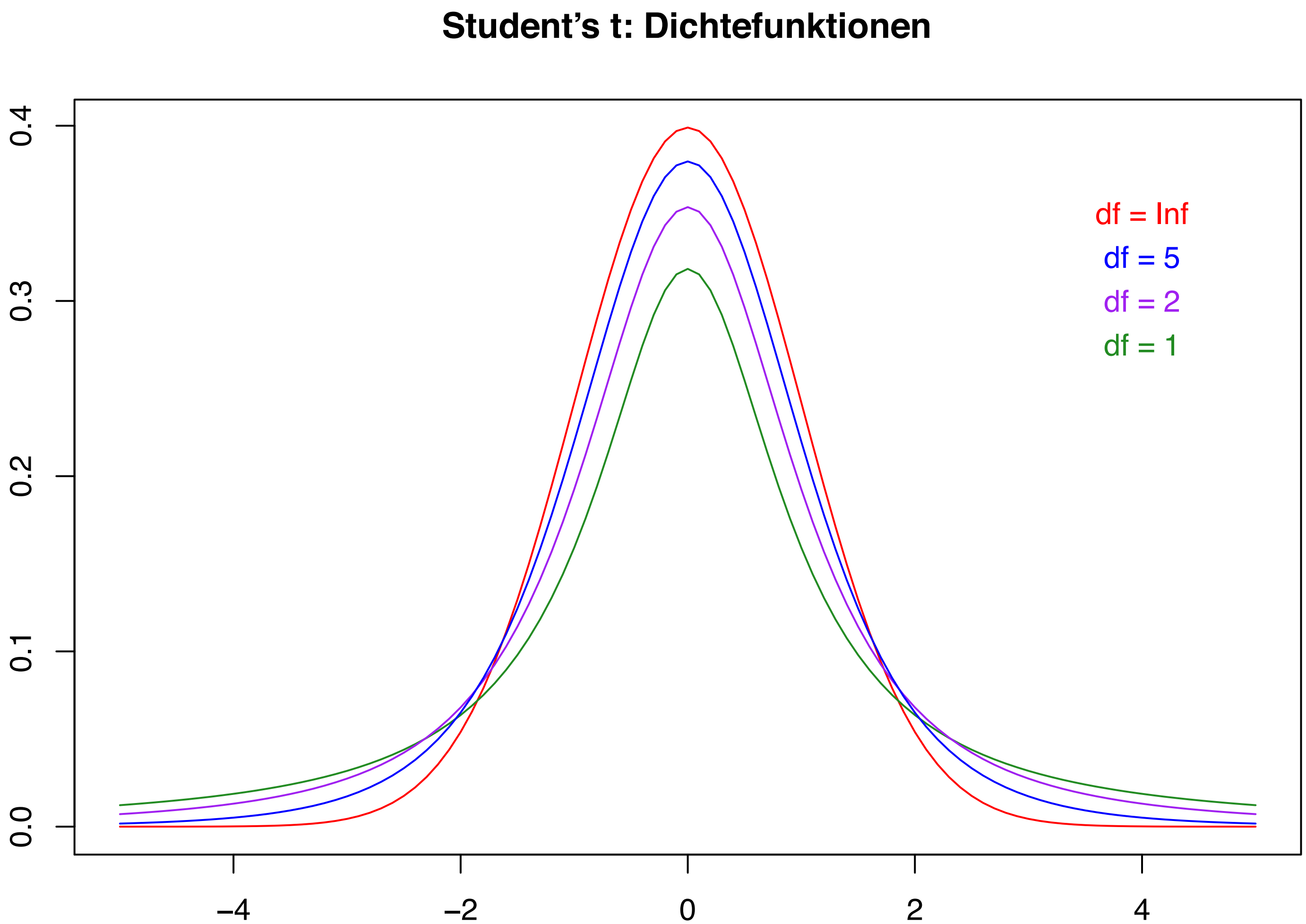
Satz (W. Gosset (alias "Student", 1908), R. Fisher (1924))
Sind $X_1, \ldots, X_n$ unabhängig und N(μ, σ^2)-verteilt,

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?”

Beispiel: 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 N$(0, 1)$-verteilt (wegen des Zentralen Grenzwertsatzes (im Zähler) und des Gesetzes der großen Zahlen (im Nenner)).

Was aber tun 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.

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

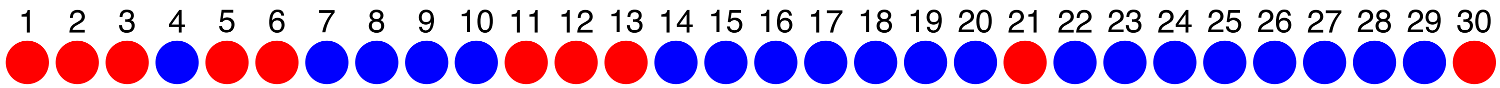
und fragt: Wie wahrscheinlich ist eine

mindestens so "randständige" Rangsumme

bei rein zufälliger Auswahl von n_x Elementen

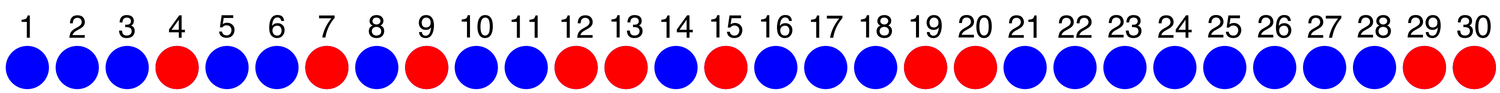
aus der Menge $\{1, \ldots, n_x + n_y\}$?

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

$21 + \ldots + 30 = 255$.

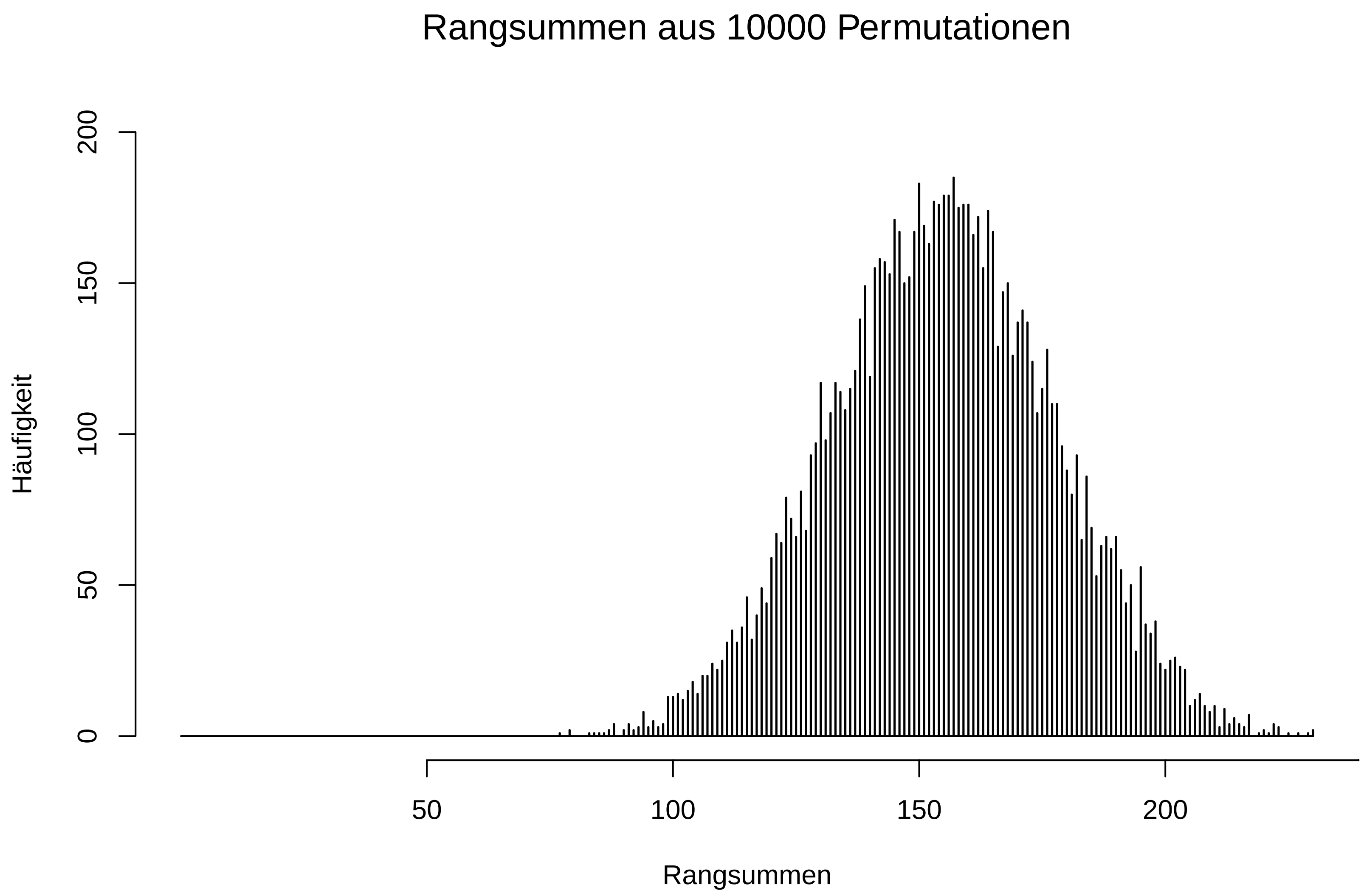
Wir ziehen 10000 mal eine Stichprobe der Größe 10 (aus 30)

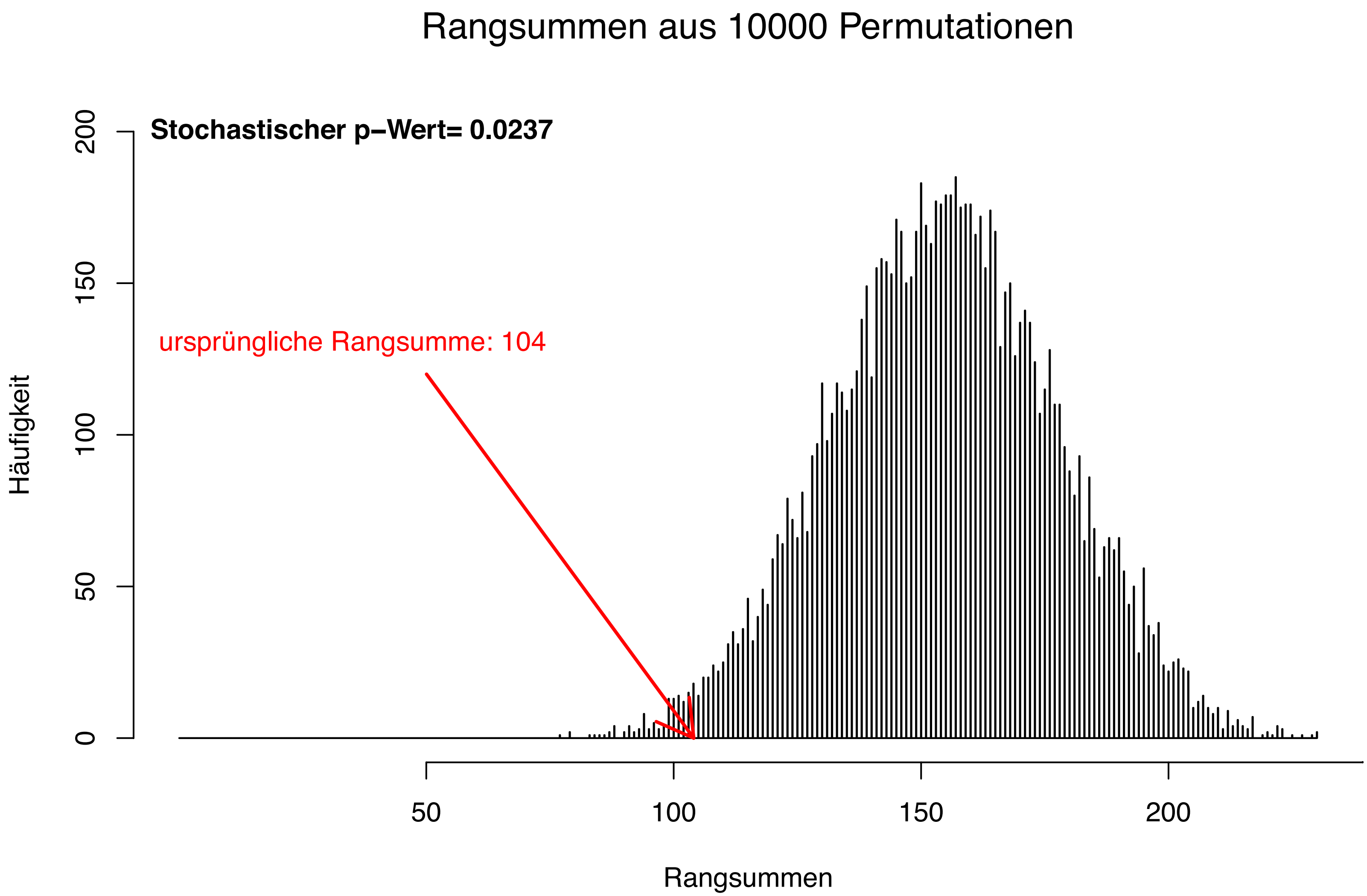
und notieren deren Rangsumme.

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
Häufigkeit
0
50
100
150
200
50
100
150
200
Rangsummen


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