Vorlesung 12b

Kann das Zufall sein?

Beispiele von statistischen Tests

1. Fishers exakter Test

“Passen die Verhältnisse in den Rahmen?”

(vgl. Buch S. 130/131

Zur Erinnerung: **12. S** Denken wir uns einen FBR, der aus 9 Vertretern der Informatik und 8 der Mathematik besteht.
In einem 5-köpfigen Komitee des FBR
findet sich nur ein Vertreter der Informatik.
Wie wahrscheinlich ist eine so extreme Zusammensetzung bei einer rein zufälligen Auswahl?

Genauer: Was ist die Wahrscheinlichkeit, dass *die Anzahl* X *der Informatiker in einem rein zufällig gebildeten 5-köpfigen Komitee des FBR* mindestes so weit entfernt von $\mu := \mathbf{E}[X]$ ist wie die beobachtete Anzahl 1?

Die Bilderbuchversion davon:

Aus einer Urne mit 9 roten und 8 blauen Kugeln
wurden 5 Kugeln entnommen .
1 davon war rot, und 4 waren blau.

Passt das zur Hypothese
einer rein zufälligen Entnahme der 5 Kugeln?

Was ist die Wahrscheinlichkeit eines
mindestens so extremen Ergebnisses wie das Beobachtete?

$$\mathbf{E}[X] = 5 \cdot \frac{9}{9+8} = 2.65$$

Mögliche Ergebnisse: $0, 1, 2, 3, 4, 5$.

Davon sind von 2.65 mindestens so weit entfernt wie 1:

0,1 und 5.

$$\mathbf{P}(X \in \{0, 1, 5\}) = 0.13$$

Unter der Hypothese des rein zufälligen Ziehens
kommt ein so extremer Ausgang nur mit W'keit 0.13 vor.

Im Jargon der Statistik:

Aufgrund des beobachteten Ergebnisses
kann man die Hypothese des rein zufälligen Ziehens
zum p-Wert 0.13 ablehnen.

Noch ein Beispiel:

Gleiches Modell, andere Zahlen, frische Anwendung:

Aus einer Urne mit 80 roten und 87 blauen Kugeln

wurden 113 Kugeln entnommen.

40 davon waren rot, und 73 waren blau.

Passt das zur Hypothese, dass die Kugeln

rein zufällig gezogen wurden?

Stimmen die Verhältnisse einigermaßen,

oder fallen sie aus dem Rahmen?

	gezogen	nicht gezogen	Summe
rot	40	40	80
blau	73	14	87
Summe	113	54	167

	gezogen	nicht gezogen	Summe
rot	40	40	80
blau	73	14	87
Summe	113	54	167

Unter den 113 gezogenen Kugeln erwartet man

ähnliche Verhältnisse wie in der gesamten Urne:

80: 167 für rot, 87 : 167 für blau.

Tatsächlich ergab sich in der Stichprobe

für rot ein **sehr** unterdurchschnittliches Ergebnis !

Wie lässt sich das quantifizieren?

	gezogen	nicht gezogen	Summe
rot	40	40	80
blau	73	14	87
Summe	113	54	167

Unter des Hypothese des rein zufälligen Ziehens

ist die Anzahl X der gezogenen roten Kugeln

hypergeometrisch verteilt

mit Parametern $n = 113$, $g = 167$, $r = 80$.

Dafür ergibt sich:

$$\mathbf{E}[X] = n \cdot \frac{r}{g} = 54.1 \ .$$

	gezogen	nicht gezogen	Summe
rot	40	40	80
blau	73	14	87
Summe	113	54	167

Unter des Hypothese des rein zufälligen Ziehens

ist die Anzahl X der gezogenen roten Kugeln

hypergeometrisch verteilt

mit Parametern $n = 113$, $g = 167$, $r = 80$.

Dafür ergibt sich:

$$\mathbf{E}[X] = n \cdot \frac{r}{g} = 54.1 \ .$$

$$g = 167,\ r = 80,\ n = 113$$

X ist Hyp(n, g, r)-verteilt

$$\mathbf{E}[X] = n \cdot \frac{r}{g} = 54.1\ .$$

Zur Erinnerung:

$$\mathbf{P}(X = k) = \frac{\binom{r}{k}\binom{g-r}{n-k}}{\binom{g}{n}}\ , \quad k = 0, 1, \ldots, n\ .$$

	gezogen	nicht gezogen	Summe
rot	**40**	40	80
blau	73	14	87
Summe	113	54	167

Die Wahrscheinlichkeit, ein Ergebnis zu erhalten,

das mindestens so weit von 54 weg ist

wie der beobachtete Wert 40, ist

$$\mathbf{P}\big(|X - 54| \geq |40 - 54|\big)$$

$$= \mathbf{P}(X \leq 40) + \mathbf{P}(X \geq 68)$$

$$= 5.57 \cdot 10^{-6} \, .$$

$$\mathbf{P}\big(|X - 54| \geq |40 - 54|\big) = 5.6 \cdot 10^{-6}$$

Was bedeutet das?

Fazit: Angenommen die Hypothese trifft zu.

Dann tritt ein Ergebnis, das so extrem ist wie das beobachtete, gerade 6 mal in einer Million auf.

Damit wird die Hypothese mehr als fragwürdig.

Man nennt die berechnete Wahrscheinlichkeit

den zu den Daten gehörigen p-Wert

oder auch das *beobachtete Signifikanzniveau*,

zu dem die Hypothese abgelehnt wird.

Wie passt unser Urnen-Beispiel in die Welt?

Es geht um die Fragestellung

“Passen die Proportionen

– oder sollte man an der Hypothese der reinen Zufälligkeit zweifeln?”

Zwei Verpackungen einer Botschaft – und eine Frage dazu:

A. Die sanfte Therapiemethode T1 brachte sogar
in 30% der Fälle keinen Heilungserfolg,
wohingegen die harte Therapiemethode T2
in immerhin 80 % der Fälle erfolgreich war.

B. Die harte Therapiemethode T2 brachte
in immer noch 20% der Fälle keinen Heilungserfolg,
wohingegen die sanfte Therapiemethode T1
in immerhin 70 % der Fälle erfolgreich war.

Welche Therapiemethode würden Sie (als Arzt) bevorzugen?

Von insgesamt 167 Ärzten

wurden **rein zufällig** 80 ausgewählt,

denen die Botschaft in der Form A vermittelt wurde,

die restlichen 87 bekamen die Botschaft in der Form B.

Jeder der Ärzte hatte sich daraufhin für die Bevorzugung

einer der beiden Therapiemethoden zu entscheiden.

Das Ergebnis war:

	für Methode T1	für Methode T2	Summe
A	40	40	80
B	73	14	87
Summe	113	54	167

Angenommen

die 113 Befürworter der (sanften) Behandlungsmethode T1

und die 54 Befürworter der (harten) Behandlungsmethode T2

sind zu iher Einstellung aufgrund der Fakten gekommen

(T1 in 70% erfogreich, T2 in 80% erfogreich), und nicht

aufgrund der “Verpackungen der Botschaft”. Das heißt dann,

dass die Zuteilung der 80 Formulare mit der Botschaft in der

Form A und der 87 Formulare mit der Botschaft in der Form B

auf die 113 Befürworter von T1

rein zufällig (durch Ziehen ohne Zurücklegen) erfolgt ist.

Für das Testen der Hypothese

“Die Verpackung der Botschaft

hat keinen Einfluss auf die Entscheidung”

eignet sich das vorher besprochene Urnenmodell.

Unter dieser Hypothese

kommt die Aufteilung der $80 + 87$ Formulare

auf die 113 Befürworter von T1

und die 54 Befürworter von T2

rein zufällig zustande.

So gesehen kann das Ergebnis “wohl kaum Zufall sein”:

unter unserer Hypothese tritt ein Ausgang,

der so extrem ist wie der beobachtete,

gerade mal 6 mal in einer Million auf.

Wenn (wie in diesem Beispiel)

der Stichprobenumfang einigermaßen groß ist,

bietet die *Normalapproximation*

eine weitere Möglichkeit des Testens der Hypothese

“Zwei Verhältnisse sind gleich”:

	gezogen	nicht gezogen	Summe
rot	40	40	80
blau	73	14	87
Summe	113	54	167

Anteilsschätzung über die Normalapproximation:

$X :=$ Anzahl roter Kugeln bei $n = 113$ Zügen

$$\mathbf{E}[X] = np$$

mit

$$p = \frac{80}{167}$$

	gezogen	nicht gezogen	Summe
rot	40	40	80
blau	73	14	87
Summe	113	54	167

$H = \frac{X}{n}$ ist approximativ normalverteilt

(trotz der schwachen Abhängigkeiten beim Ziehen ohne Zurücklegen),

$$\mu_H = p = \frac{80}{167}, \quad g = 167, \quad n = 113,$$

$$\sigma_H^2 = \frac{p(1-p)}{n} \frac{g-n}{g-1},$$

$$\sigma_H = 0.022.$$

	gezogen	nicht gezogen	Summe
rot	40	40	80
blau	73	14	87
Summe	113	54	167

$Z := \frac{H - \mu_H}{\sigma_H}$ ist approximativ N$(0, 1)$-verteilt

Der beobachtete Wert von Z war $z = -4.67$.

$$\mathbf{P}(|Z| > 4.67) = 3 \cdot 10^{-6}$$

ist hier der p-Wert,

zu dem die Hypothese abgeleht werden kann.

2. z-Test und t-Test

“Kann *diese* Verschiebung des Mittelwertes Zufall sein?”

n reelle Messwerte $x_1, \ldots, x_n$ haben den Mittelwert m

(alles gemessen auf einer bestimmten Skala.)

Unterscheidet sich der beobachtete Mittelwert m signifikant

von einem hypothetischen “Populationsmittelwert” μ_0?

Eine erste Auskunft gibt ein Vergleich

des Unterschiedes $|m - \mu_0|$

mit dem *Standardfehler*

$$f := s/\sqrt{n}.$$

Wir können fragen:

Um weches Vielfache des Standardfehlers

unterscheidet sich m von μ_0?

Dies erhält seinen theoretischen Unterbau

durch die goldene Idee der Statistik

(*man fasse die* x_i *auf als Realisierungen von unabhängigen, identisch verteilten Zufallsvariablen* X_i *mit Erwartungswert* μ *und Varianz* σ^2*)*

und den Zentralen Grenzwertsatz:

Für große n *ist der Stichprobenmittelwert* $\bar{X}$ *approximativ* $\mathrm{N}\left(\mu, \frac{\sigma^2}{n}\right)$*-verteilt*.

Für große n ist

$\bar{X} - \mu$ approximativ $\mathrm{N}(0, \frac{\sigma^2}{n})$-verteilt.

Bei bekanntem σ sei $z := \dfrac{m - \mu_0}{\sigma/\sqrt{n}}$.

Unter der Hypothese $\mu = \mu_0$ ist

$$\mathbf{P}(|\bar{X} - \mu_0| \geq |m - \mu_0|) \approx \mathbf{P}(|Z| \geq |z|),$$

mit $\mathrm{N}(0, 1)$-verteiltem Z.

Man spricht vom p-Wert für die Ablehnung

der Hypothese $\mu = \mu_0$ zugunsten der Alternative $\mu \neq \mu_0$.

Oft gibt man sich ein *Signifikanzniveau* α vor.

Wenn der p-Wert kleiner als α ist, sagt man: Die Hypothese $\mu = \mu_0$ kann zugunsten der Alternative $\mu \neq \mu_0$ zum Niveau α abgelehnt werden.

Populär ist die Wahl $\alpha = 0.05$.

Was tun bei unbekannten σ?

In der Praxis ist σ ja meist unbekannt.

Aber:

Für große n ist s mit großer W'keit nahe bei σ.

Also ist für große n

$$T := \frac{\bar{X} - \mu}{s/\sqrt{n}}$$ approximativ $\mathsf{N}(0,1)$-verteilt.

Sei $t := \dfrac{m - \mu_0}{s/\sqrt{n}}$.

Unter der Hypothese $\mu = \mu_0$ ist für große n

$$\mathbf{P}(|T| \geq |t|) \approx \mathbf{P}(|Z| \geq |t|),$$

mit $\mathsf{N}(0,1)$-verteiltem Z.

Und was kann man bei kleinem n sagen?

Unter der zusätzlichen Modellannahme

$X_1, \ldots, X_n$ sind (unabhängig und) $N(\mu, \sigma^2)$-verteilt

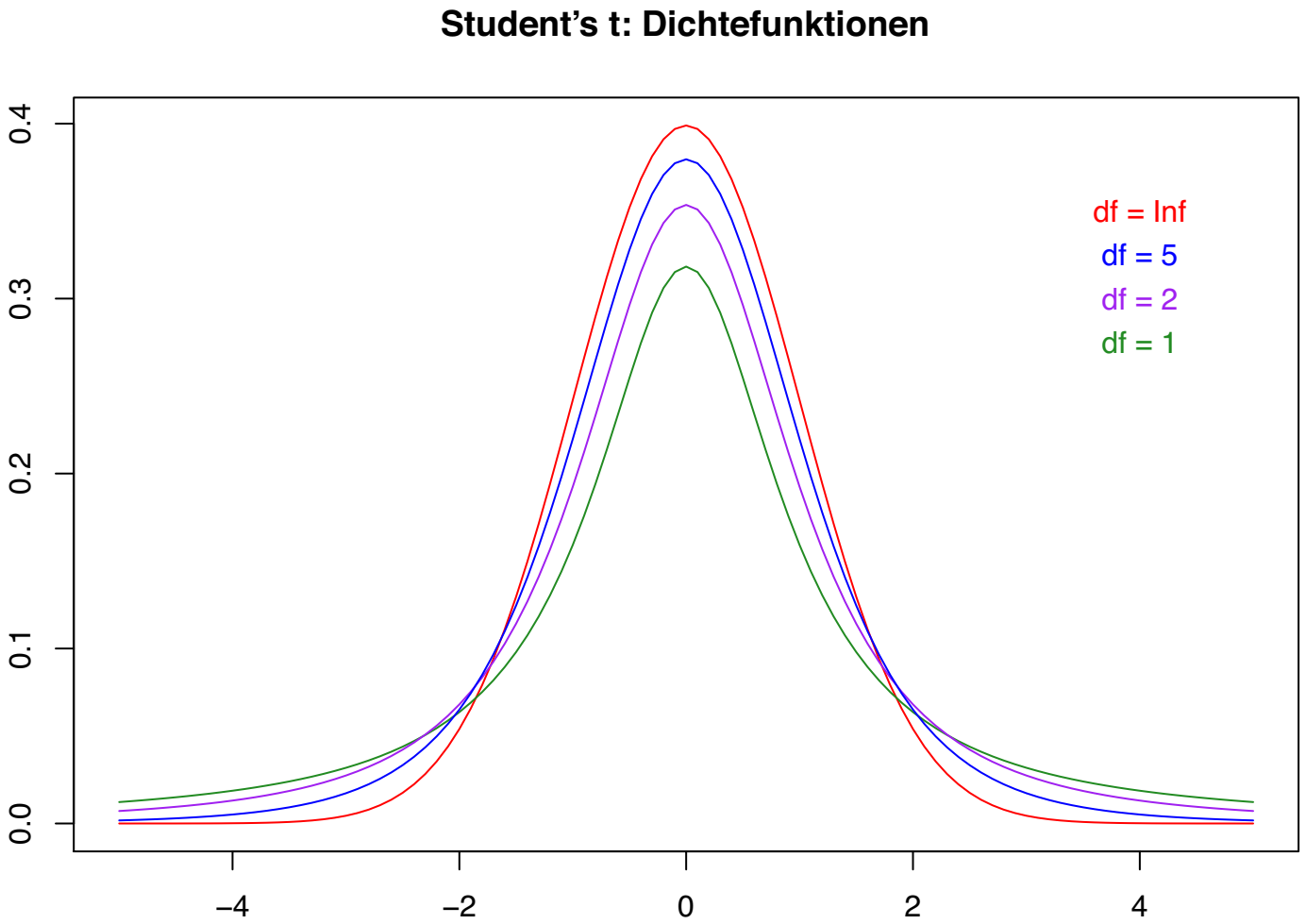
hängt die Verteilung von

$$T = \frac{\sqrt{n}\bar{X} - \mu}{s}$$

zwar von n, jedoch nicht von μ und σ ab.

Siehe dazu auch VL 12a Folie 19 (und Buch Seite 132).
Diese "Student-Verteilung mit $n - 1$ Freiheitsgraden" kennt R gut.
Z. B. ist der Befehl für Wert der Verteilungsfunktion an der Stelle b:

```
pt(b,n-1).
```



Dichten von T_{df}:= eine student-verteilte Zufallsvariable mit df Freiheitsgraden

Beispiel:

Ist für $n = 16$ der "Wert der t-Statistik"

$$t = \frac{m - \mu_0}{s/\sqrt{n}}$$

glech 2.5, dann ergibt sich

$$\mathbf{P}(|T_{15}| \geq 2.5) = 2\mathbf{P}(T_{15} \geq 2.5)$$

$$= 2(1 - \mathrm{pt}(2.5, 15)) = 0.025.$$

Beispiel: Für $n = 10$ und den Wert $t = 2$ ist

$$\mathbf{P}(|T_9 \geq 2|) = 2(1 - \mathrm{pt}(2, 9)) \approx 0.077.$$

In diesem Fall kann die Hypothese $\mu = \mu_0$ aufgrund des t-Tests nicht zum 5%-Signifikanzniveau abgelehnt werden; man kann sie nur zum (nicht sehr aussagekräftigen) p-Wert 0.077 ablehnen.

Merke: Weil die Verteilung von T_9 “breitschultriger” ist als $\mathrm{N}(0, 1)$, ist es für T_9 wahrscheinlicher, außerhalb der Grenzen ± 2 zu fallen als für eine standard-normalverteilte ZV’e.

3. z-Test und t-Test für ungepaarte Stichproben.

“Unterscheiden sich zwei Mittelwerte signifikant?”

Man stelle sich vor: 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?

Das kommt drauf an ...

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

mit "seinem Standardfehler" f.

Weil wir hier an unahängige Stichproben denken,

addieren sich die Varianzen:

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

Eine Maßzahl für den “relativen Unterschied” ist also

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

Anders gefragt: Wie groß ist der beobachtete Wert der Differenz der Stichprobenmittelwerte, gemessen in Einheiten der geschätzen Standardabweichung der Differenz der Stichprobenmittelwerte?

Mit Hilfe der asymtptotischen Normalität

wird die Antwort leicht: Unter der Hypothese $\mu_X = \mu_Y$

ist $\dfrac{m_y - m_x}{f}$ zu lesen als

Realisierung einer annähernd $N(0,1)$- verteilten

Zufallsvariablen.

Ist dieser Wert (etwa) 1.96,

dann bekommt man 0.05

als p-Wert für die Ablehnung der Hypothese $\mu_X = \mu_Y$.

Was tun für kleinere Stichprobenumfänge?

Interpretiert man wiederum 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 "Equal or 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 beobachtet man die *Rangsumme* $w := \sum_{i=1}^{n_x} R(x_i)$

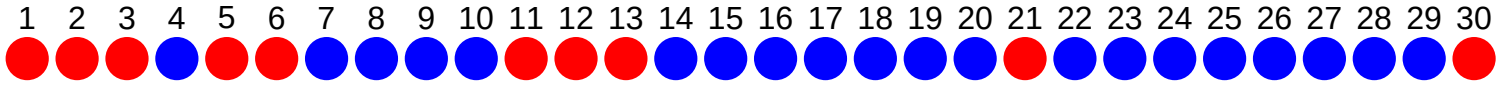
und fragt: Wie wahrscheinlich ist eine

mindestens so "randständige" Rangsumme

bei rein zufälliger Auswahl von n_x Elementen

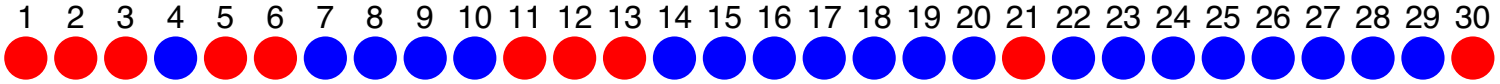
aus der Menge $\{1, \ldots, n_x + n_y\}$?

Die Raenge der x_i und der y_j



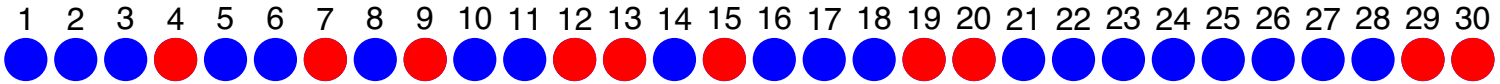
Rangsumme der x_i = 104

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

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

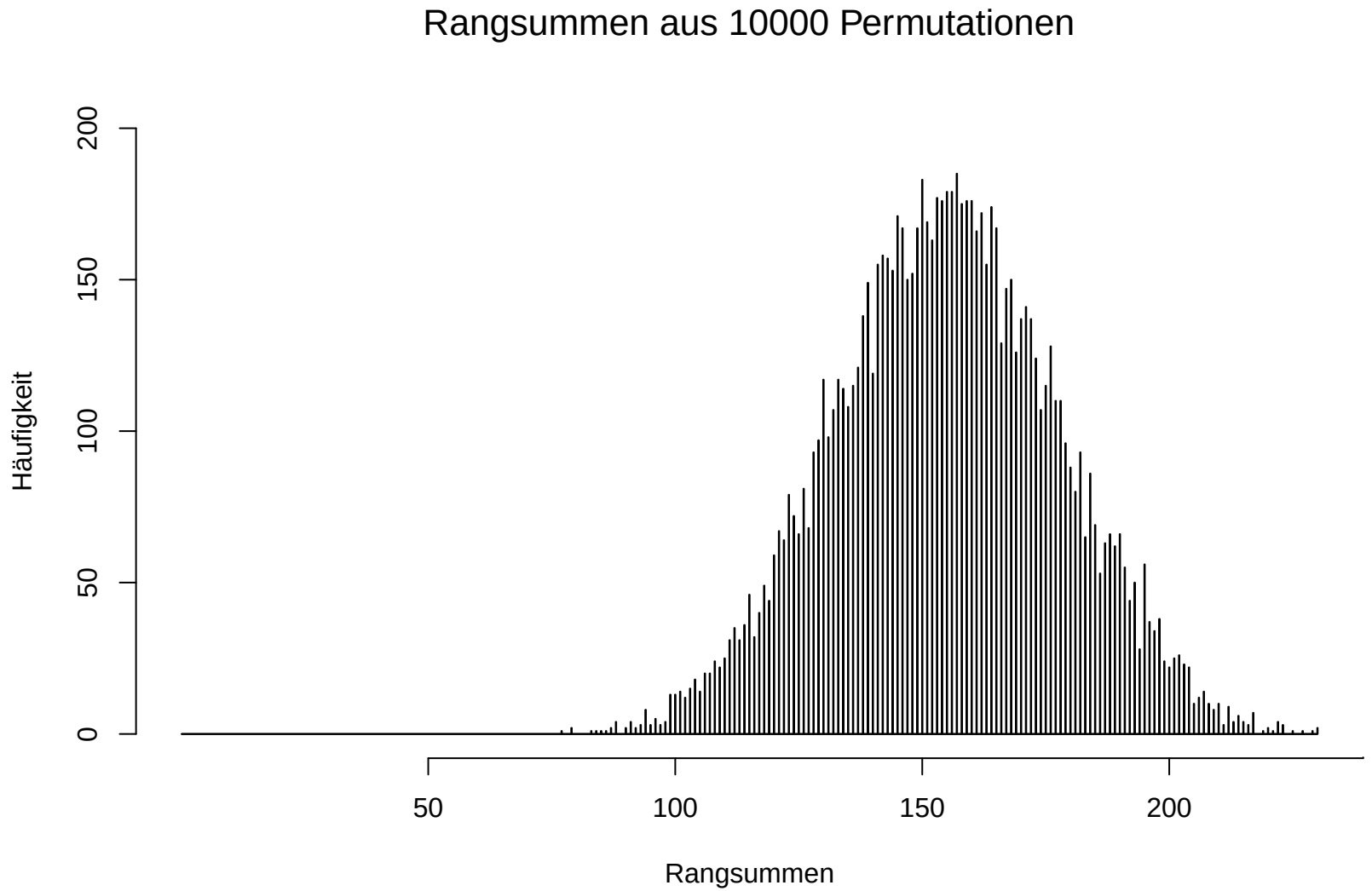
Wir ziehen 10000 mal eine Stichprobe der Größe 10 (aus 30)

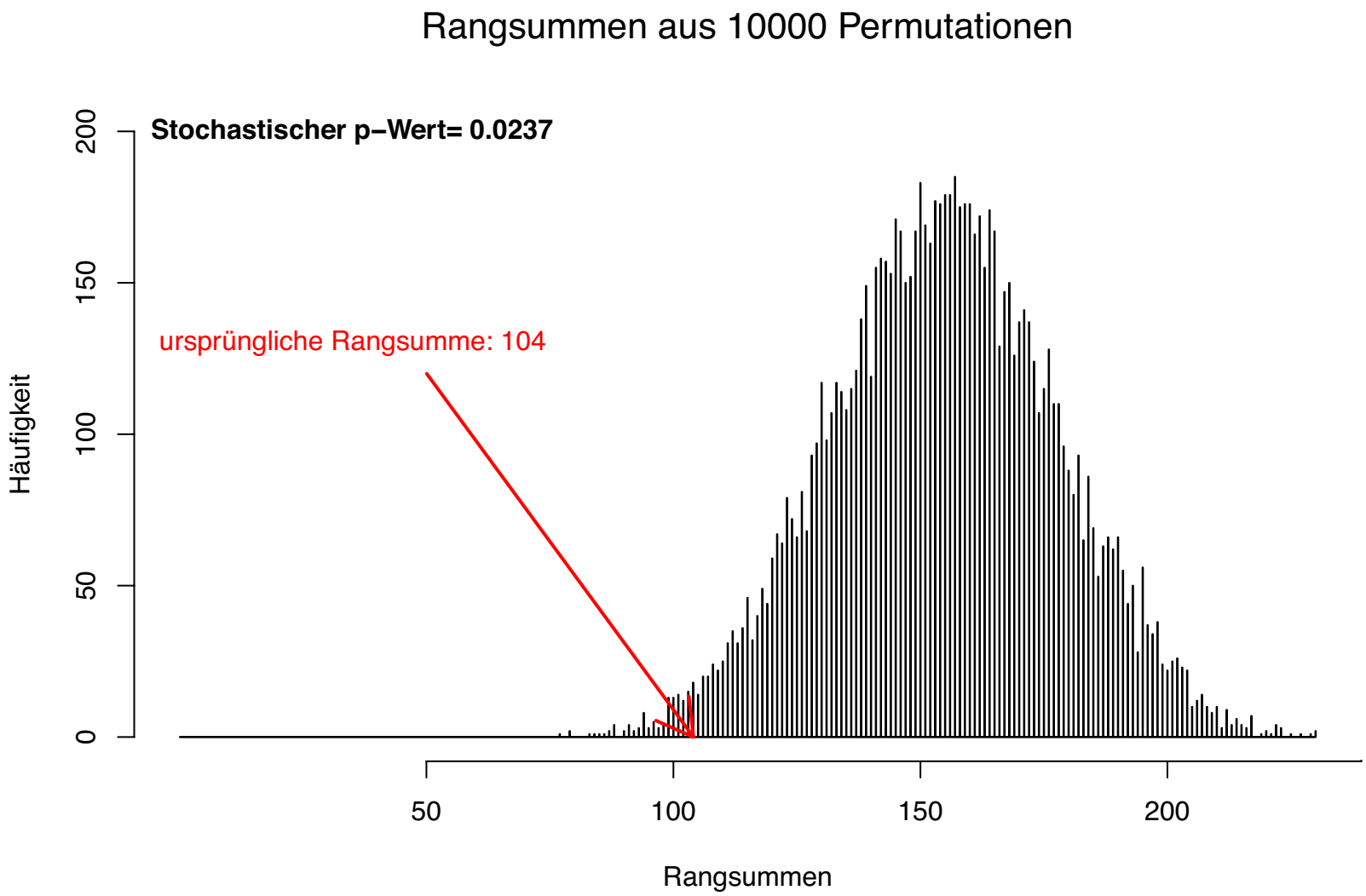
und notieren deren Rangsumme.

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