Statistische Hypothesentests

Eine anschauliche Einführung

Markus Kalisch, Lukas Meier

Statistische Hypothesentests: Typische Fragen

Können wir **basierend auf Daten** entscheiden / nachweisen …

- … ob ein **Grenzwert** überschritten wird, z.B. bei Pestiziden?
- … ob ein Hersteller bei einem Produkt die **Spezifikationen** verletzt?
- … ob sich die gemessene Chlorid-Konzentration an der Oberfläche an einem Betontragwerk von unseren **Modellannahmen** unterscheidet?

Grundsätzliche Idee: **Falsifikationismus** (Popper):

Wissenschaftliches Prinzip: Kann wissenschaftliche Theorie nie wirklich als wahr beweisen, aber kann wissenschaftliche Theorie falsifizieren!

Lernziele

- Was ist ein **statistischer Hypothesentest** und wofür wird er verwendet?
  → «Rezept» um eine Behauptung (typischerweise über ein Modell) und eine Beobachtung (Daten) zu vergleichen, inkl. Berücksichtigung des Zufalls
- Was ist die **Struktur** eines Hypothesentests?
  → Immer 6 Schritte
- Ferner: Worum geht es bei **Fehler 1**. bzw. **2. Art** und **Macht**?
  → «Gütekriterien» eines Tests

Zudem:

- Beispiele, die ev. auch in der Schule verwendet werden können
- «Einfache» Anwendungen von R

Hypothesentests: 3 Beispiele

- **Test mit Simulation**: 
  Werden Panini-Bilder zufällig eingepackt?

- **Binomialtest**: 
  Ist der Würfel gezinkt?

- **Runs-Test**: 
  Ist eine Sequenz von 0/1 zufällig erzeugt worden?

- Allgemeine Empfehlung: R verwenden, sonst werden die Rechnungen schnell sehr aufwändig

Panini Album (Fussball WM 2010)

Südafrika 2010

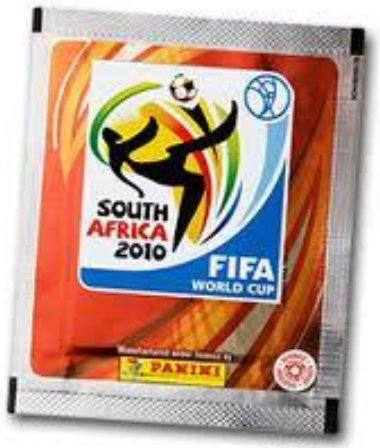
Die Schweiz schlägt den Weltmeister

Spanien holte sich 2 Jahre nach dem EM-Titel auch die WM-Krone. Der Schweiz gelang ein Coup, gefolgt von Enttäuschungen.

Sammelalbum:
661 Bilder

Panini Bilder: Kaufmöglichkeiten

Packung



5 Bilder

Sammel-Box

100 Packungen
= 500 Bilder


kkiosk
STIMOROL
STIMOROL


SOUTH AFRICA 2010
FIFA
PANINI

?

Beobachtung von Vorjahren

- Sammelbox: Erstaunlich wenige doppelte Bilder
- **Nullhypothese**:
  Bilder werden zufällig verpackt («Null», weil kein System hinter dem Verpacken steckt)
- **Alternativhypothese**:
  Die Bilder werden systematisch verpackt, sodass man weniger Doppelte hat.
- Wie könnte man zwischen diesen beiden Hypothesen unterscheiden?
  Typischerweise wollen wir die Nullhypothese verwerfen.

Hypothesentest

- Ich habe eine Box mit 500 Bildern gekauft. In ein leeres Album (661 mögliche Bilder) konnte ich 477 Bilder einkleben.

- Passen die **Nullhypothese «zufällig verpackt»** und die **Beobachtung «477 Bilder eingeklebt»** zusammen?

- Angenommen, die Nullhypothese stimmt:
  Ist es plausibel, dass ich dann 477 Bilder einkleben kann?

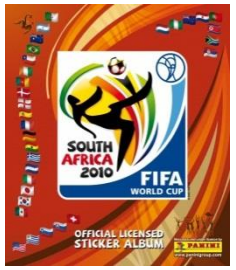
Problem: Wie viele Bilder sind «normal»?

- Theoretisch: Alles zwischen 1 und 500 möglich, aber unterschiedlich wahrscheinlich (Variante des sogenannten coupon collector's problem).
- Angenommen, die Nullhypothese stimmt (Bilder zufällig verpackt):
  Wie viele Bilder kann man «normalerweise» einkleben?
- Wenn wir (sehr) viel mehr Bilder als «normal» einkleben konnten, wurden die Bilder wohl **nicht** zufällig verpackt.
- **Signifikanzniveau** α:

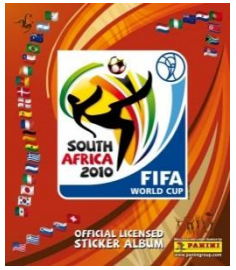
  Wie «extrem» muss die Beobachtung liegen, damit wir der Nullhypothese nicht mehr glauben?
- Z.B.: $\alpha = 10^{-6}$: Wir lehnen die Nullhypothese ab, wenn wir etwas beobachten, das zu den 10^{-6} extremsten Ausgängen gehört.

Lösung: Computersimulation

1: 


350 Bilder eingeklebt

2: 

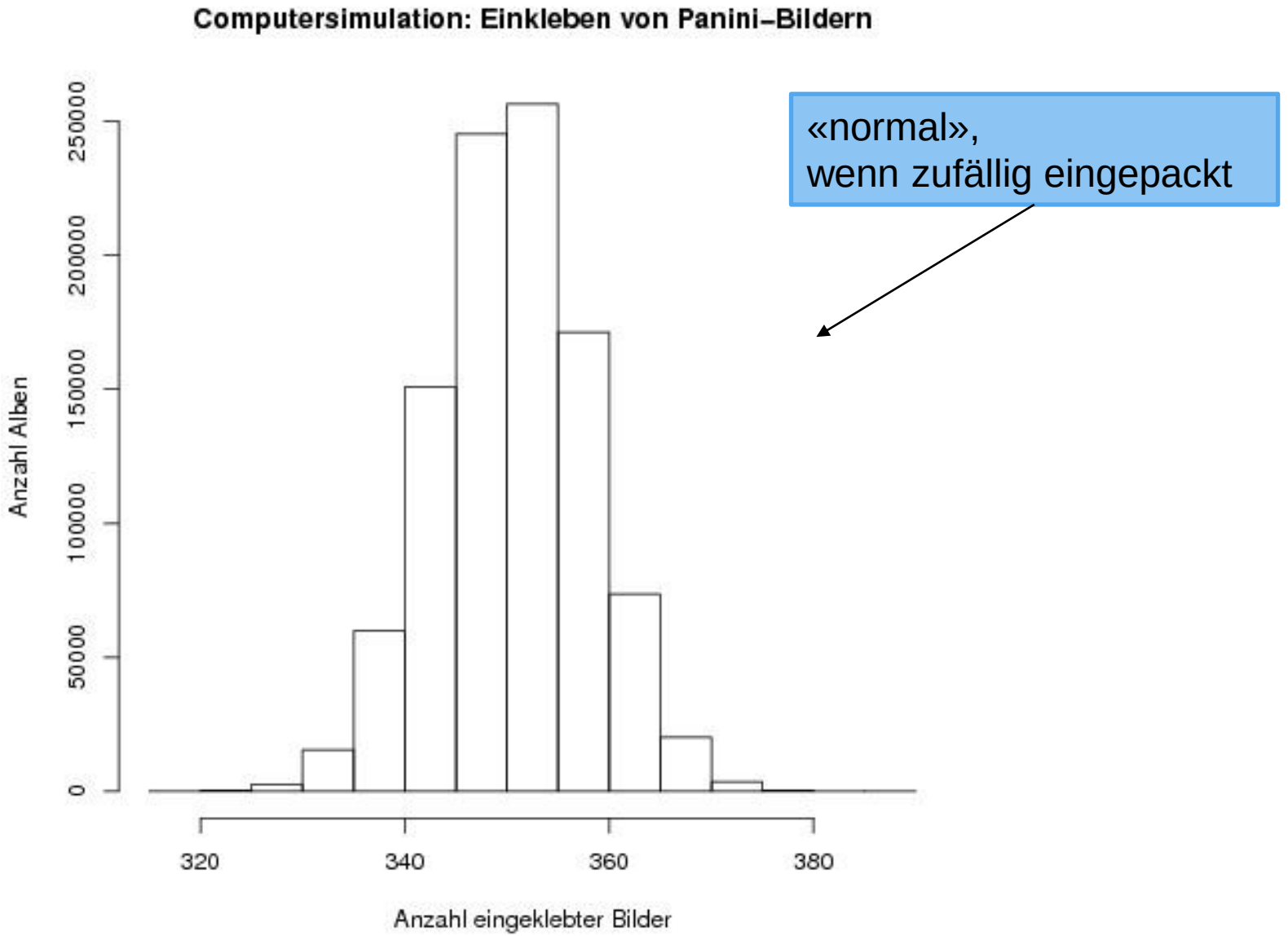

361 Bilder eingeklebt

⋮

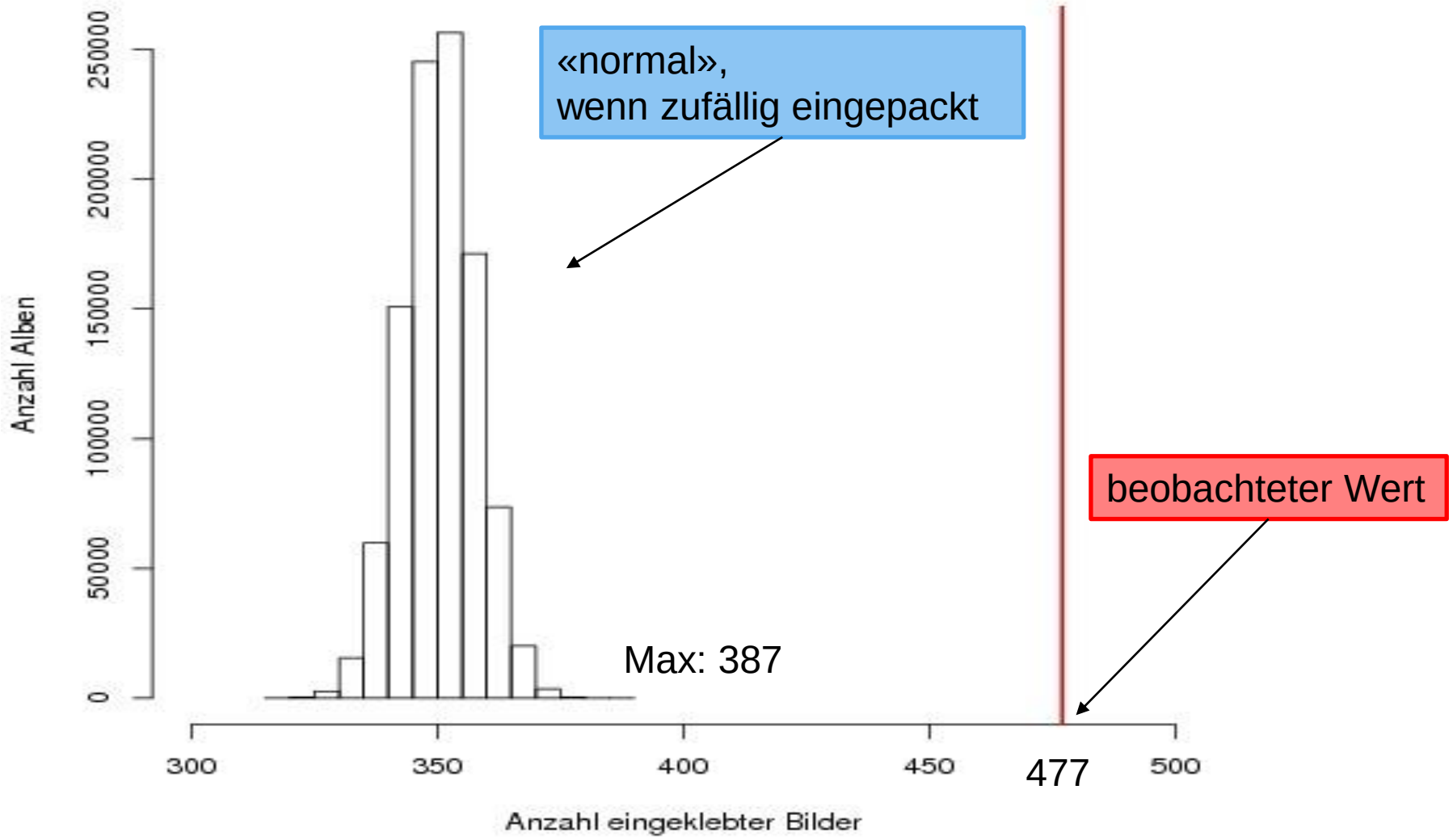
1 Mio: 


358 Bilder eingeklebt

Ergebnis der Computersimulation



Passt unsere Beobachtung zur Computersimulation?

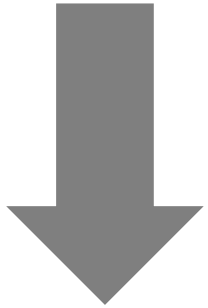


R-Tipps für Nullverteilung

```r
nA <- 661 ## Alle Bilder in Album
nB <- 500 ## Bilder in Box
nreps <- 10000
## "Platzhalter" für Ergebnis:
eingeklebt <- vector("numeric", nreps)
for (i in 1:nreps) {
  ## erzeuge fiktive Box: Ziehen mit Zurücklegen
  box <- sample(x = 1:nA, size = nB, replace = TRUE)
  ## zähle jedes Bild nur einmal
  eingeklebt[i] <- length(unique(box))
}
hist(eingeklebt)
```

Schlussfolgerung

- Angenommen, die Bilder werden zufällig verpackt. Die Wahrscheinlichkeit, 477 oder mehr Bilder einkleben zu können, ist kleiner als ein Millionstel!



- Beobachtung und Simulation passen **nicht** zusammen:

  Die Bilder werden wohl **nicht** zufällig eingepackt.

Zusammenfassung: Hypothesentest

1. **Modell:** Ziehen 500 Bilder mit Zurücklegen aus 661 Bildern

2. **Nullhypothese H_0:** “Panini-Bilder in Sammel-Box werden zufällig eingepackt”
   Alternativhypothese H_A: “Bilder werden systematisch eingepackt, sd. weniger Doppelte”

3. **Teststatistik:** Anzahl Bilder, die man in ein leeres Album einkleben kann (wenn man eine Box mit 500 Bildern hat).
   Verteilung der Teststatistik, wenn Nullhypothese stimmt: Computersimulation (hier)

4. **Signifikanzniveau:** $\alpha = 10^{-6}$

5. **Verwerfungsbereich** der Teststatistik (approximativ):
   Computer beobachtet bei 1 Mio Simulationen nie mehr als 387 eingeklebte Bilder
   Verwerfungsbereich: $K = \{388, 389, \ldots, 500\}$

6. **Testentscheid:** Der beobachtete Wert (477) liegt im Verwerfungsbereich der Teststatistik. Daher wird die Nullhypothese auf dem Signifikanzniveau $\alpha = 10^{-6}$ verworfen.

Hypothesentests: 3 Beispiele

- **Test mit Simulation**:
  Werden Panini-Bilder zufällig eingepackt?

- **Binomialtest**:
  Ist der Würfel gezinkt?

- **Runs-Test**:
  Ist eine Sequenz von 0/1 zufällig erzeugt worden?

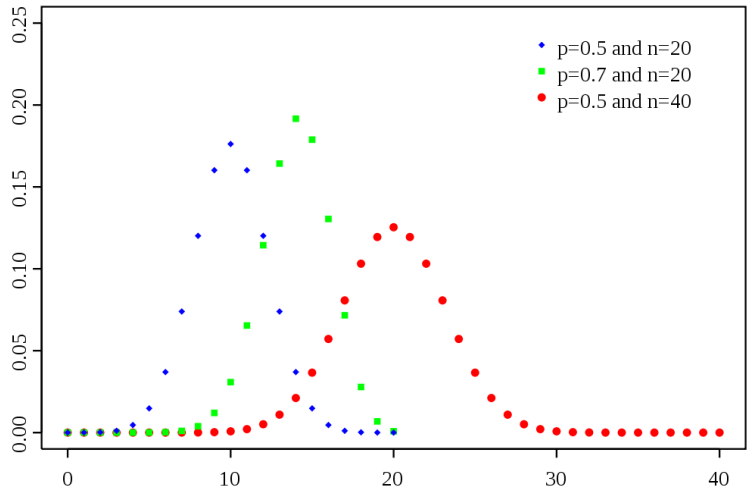
Binomialtest: Gezinkter Würfel

Tipp: «Poor man’s» gezinkter Würfel - Die «1» mit einer «6» überkleben

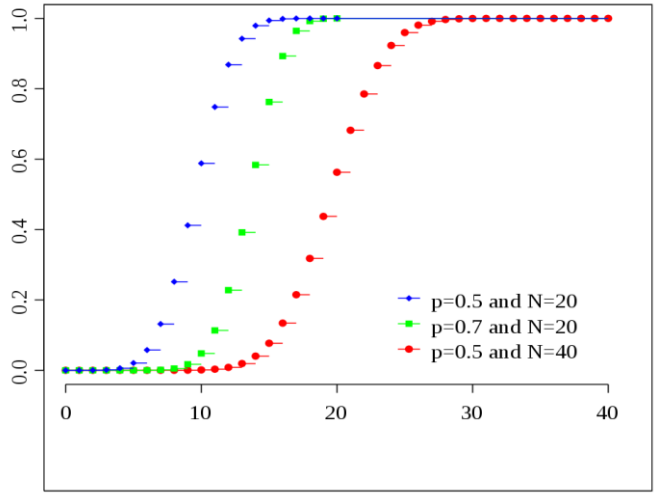
Binomialverteilung: Refresher

- **Situation**: Ziehe n Lose an **Losbude**; gleiche Gewinnw'keit für alle Lose; Lose unabhängig voneinander.
- X: Anzahl Gewinne unter n Losen
- $X \sim Bin(n, p)$
  "X ist binomial-verteilt mit Parametern n und p"
- $P(X = x) = \binom{n}{x} p^x (1-p)^{n-x},\ x \in \{0,1,\dots,n\}$
- $E(X) = n \cdot p,\ Var(X) = n \cdot p \cdot (1-p)$

"probability mass function" (pmf)



"cumulative distribution function" (cdf)



Binomialtest: Gezinkter Würfel

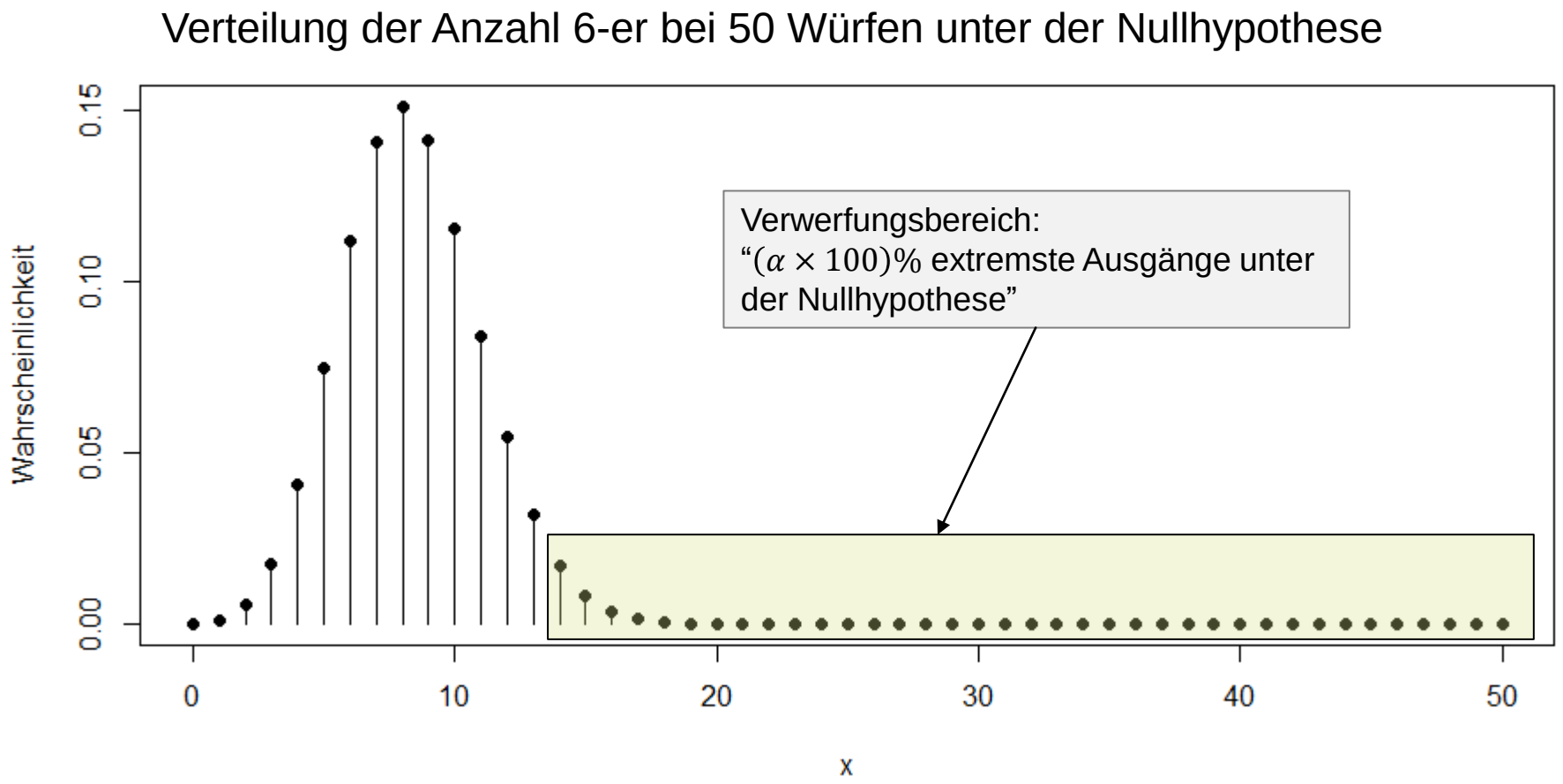
1. **Modell:** X: Anzahl 6er bei 50 Würfen; $X \sim Bin(n = 50, p)$
2. **Nullhypothese:** $H_0: p = 1/6$
   Alternativhypothese: $H_A: p > 1/6$ (einseitig)
3. **Teststatistik** T: Anzahl 6er bei 50 Würfen
   Verteilung der Teststatistik, wenn Nullhypothese stimmt: $T \sim Bin(50, 1/6)$
4. **Signifikanzniveau:** $\alpha = 0.05$ (Konvention)
5. **Verwerfungsbereich** der Teststatistik: "$(\alpha \times 100)\%$ extremste Ausgänge"
   $P(T = t) = \binom{n}{t} p^t (1-p)^{n-t}$; berechne $P(T \geq t)$

t	…	13	14	15	…
$P(T \geq t)$	…	0.06	0.03	0.01	…

Verwerfungsbereich:
Kleinste Zahl t, sd.
$P(T \geq t) \leq \alpha$

6. **Testentscheid:** Liegt die beobachtete Anzahl 6er bei 50 Würfen im Verwerfungsbereich der Nullhypothese?
   Falls ja: H_0 wird auf dem 5% Niveau verworfen
   Falls nein: H_0 kann auf dem 5% Niveau nicht verworfen werden

Illustration Verwerfungsbereich



Backup: Rechnen mit der Binomialverteilung

- $X \sim Bin(n = 10, p = \frac{1}{6})$; Gesucht: $P(X = 2) = ?$
- Von Hand:

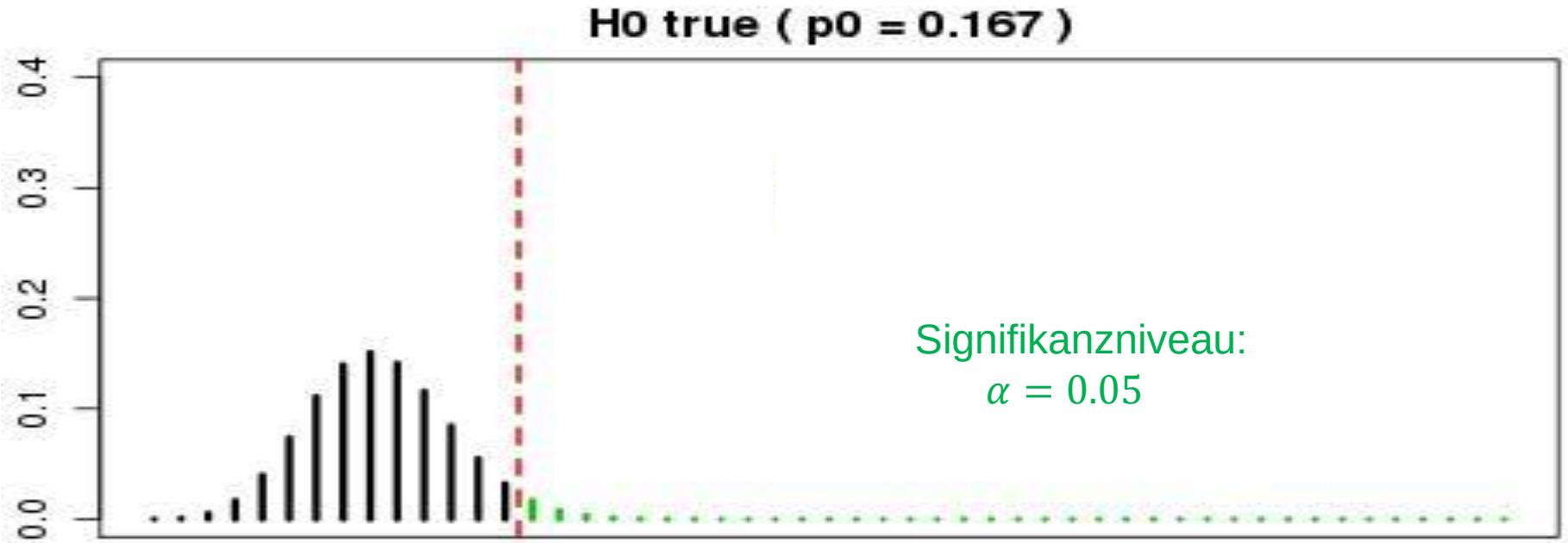
$$P(X = 2) = \binom{10}{2}\left(\frac{1}{6}\right)^2\left(\frac{5}{6}\right)^8 \approx 0.291$$

- Mit Taschenrechner
  Je nach Modell verschieden: binompdf, binompmf, etc.
- Mit R:

```
choose(10, 2) * (1 / 6)^2 * (5 / 6)^8
## [1] 0.29071
## Mit pmf:
dbinom(x = 2, size = 10, prob = 1 / 6)
## [1] 0.29071
```

Statistische Tests: Verschiedene Fehlerarten

Entscheidung / Wahrheit	H_0	H_A
H_0	✓	Fehler 1. Art
H_A	Fehler 2. Art	✓



Entscheidung / Wahrheit	H_0	H_A
H_0	✓	Fehler 1. Art
H_A	Fehler 2. Art	✓

Verwerfungsbereich

Fehler 1. Art:
Fairer Würfel, aber wir landen «aus Versehen» im Verwerfungsbereich
→ glauben an gezinkten Würfel (Fehlentscheidung!)
W'keit für Fehler 1. Art ist Summe der grünen Stabhöhen ($\leq \alpha$)

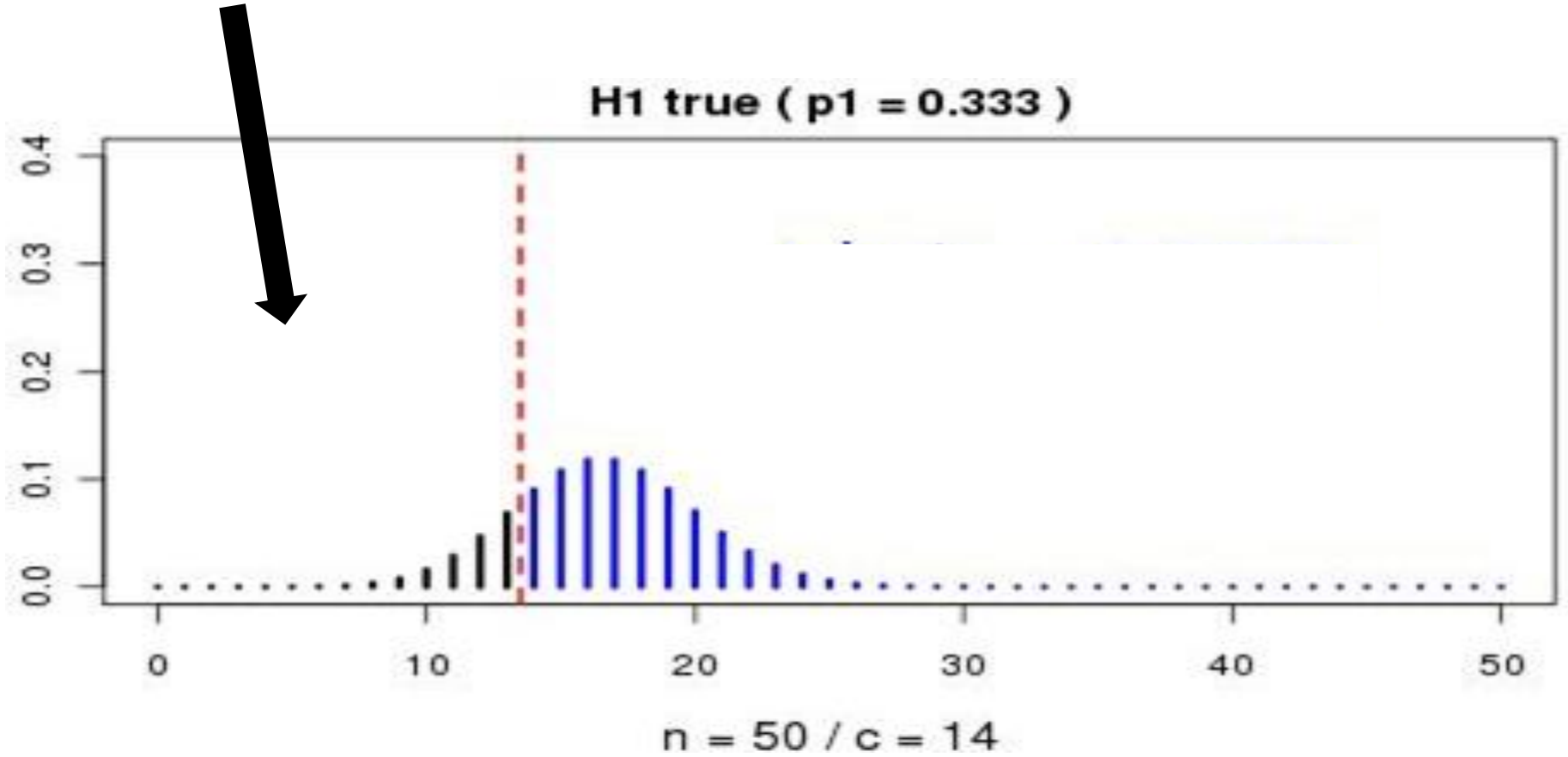
(Gedankenexperiment, «was wäre wenn …»)

- Angenommen Würfel gezinkt
- «Wie» gezinkt? Z.B.: zeigt 6 **mit W'keit** $\frac{1}{3}$ - was wäre dann zu erwarten ?

Entscheidung / Wahrheit	H_0	H_A
H_0	✓	Fehler 1. Art
H_A	Fehler 2. Art	✓

Fehler 2. Art:

- «aus Versehen» landen wir **nicht** im Verwerfungsbereich → glauben an fairen Würfel (**Fehler!**)
- W'keit für Fehler 2. Art ist **Summe der schwarzen Stabhöhen**
- muss berechnet werden und hängt von konkreter Wahl der Alternativhypothese ab.

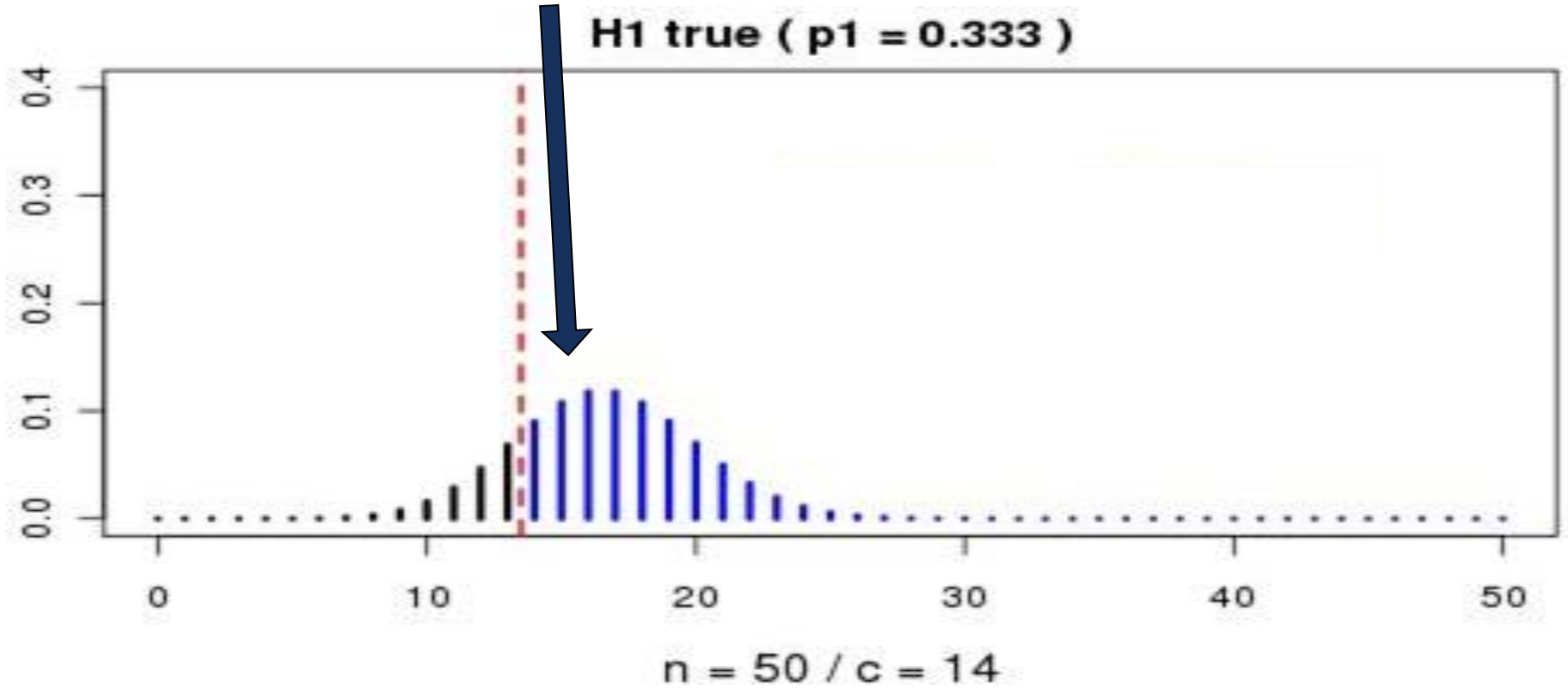


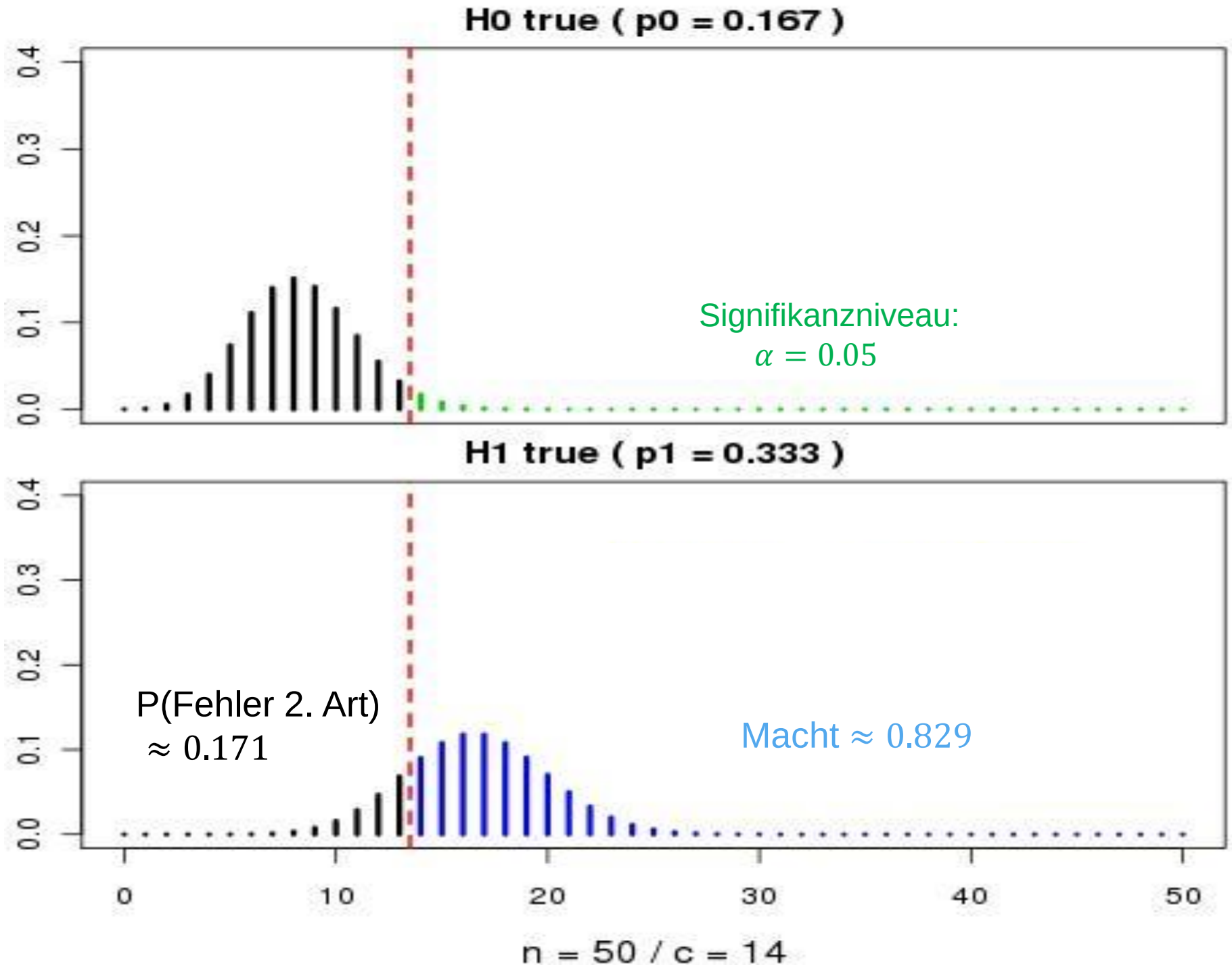
(Gedankenexperiment, «was wäre wenn …»)

- Angenommen Würfel gezinkt
- «Wie» gezinkt? Z.B.: zeigt 6 **mit W'keit** $\frac{1}{3}$ - was wäre dann zu erwarten ?

Macht:

- landen korrekterweise im Verwerfungsbereich → glauben an gezinkten Würfel (korrekt)
- W'keit dafür: «Macht», Summe der blauen Stabhöhen
- Macht = 1 – P(Fehler 2. Art)
- Macht: «Wie gut kann eine Abweichung von der Nullhypothese entdeckt werden?»




H0 true (p0 = 0.167)
0.0
0.1
0.2
0.3
0.4
Signifikanzniveau:
$\alpha = 0.05$
H1 true (p1 = 0.333)
P(Fehler 2. Art)
≈ 0.171
Macht ≈ 0.829
0
10
20
30
40
50
n = 50 / c = 14

R-Tipps für Macht

```r
## 1) Verwerfungsbereich bestimmen
c <- 14
## P(X >= c)
sum(dbinom(x = c:50, size = 50, prob = 1/6))
## [1] 0.03072356
## Wähle kleinste Zahl (hier 14), sodass <= 0.05

## 2) Macht:
## P(X_2 >= c)
sum(dbinom(x = c:50, size = 50, prob = 1/3))
## [1] 0.8285349
```

!!!

Kommentare zu Fehlerarten

- Die Wahrscheinlichkeit für einen Fehler 1. Art wird bei einem statistischen Test **immer kontrolliert** mit dem **Signifikanzniveau** α.

  → Wenn die Nullhypothese stimmt, wird diese nicht sehr oft verworfen.

- Die Berechnung der Macht ist ein «Gedankenexperiment». Sie hängt ab von der konkreten Wahl des Parameters unter der Alternativhypothese. Z.B. glauben wir, dass ein Würfel $p = 1/3$ hat?

- Die Berechnung der Macht ist nützlich für die **Planung von Experimenten**. Man will typischerweise die Nullhypothese verwerfen und die Macht sagt einem dann, was für Chancen man dafür hat (→ Stichprobe gross genug?).

Weiterführende Fragen

- Was passiert mit der Macht, wenn wir das Signifikanzniveau kleiner machen, d.h., Fehler 1. Art wird seltener?

- Was passiert mit der Macht, wenn man 100 mal statt 50 mal würfeln würde?

- Was passiert mit der Macht, wenn der Würfel die 6 mit W'keit $p = 1/2$ (statt wie bisher $p = 1/3$) zeigen würde? Was ist mit $p = 1$, $p = 0$ und $p = 1/6$?

- Eigentlich nützlicher: Kann man auch einen Bereich angeben von Werten für p, die alle kompatibel sind mit den beobachteten Daten? → **Vertrauensintervalle**

Hypothesentests: 3 Beispiele

- **Test mit Simulation**:
  Werden Panini-Bilder zufällig eingepackt?

- **Binomialtest**:
  Ist der Würfel gezinkt?

- **Runs-Test**:
  Ist eine Sequenz von 0/1 zufällig erzeugt worden?

0/1-Zufallssequenz

- Hausaufgabe: 200 mal Münze werfen, damit zufällige 0/1-Sequenz erzeugen
- LANGWEILIG !!! ☹
- Mogeln: Willkürlich auf Tasten 0/1 tippen
- Wie vergleichen sich die Mogelsequenzen mit echten Zufallssequenzen?
- Möglicher Exkurs: Pseudozufallszahlen
- Test-Batterien für Zufallszahlen: Diehard Tests
  Ein Test dabei ist der sog. “Runs-Test”
  https://en.wikipedia.org/wiki/Diehard_tests

Runs

00111010110

Sequenz hat 5 “runs”

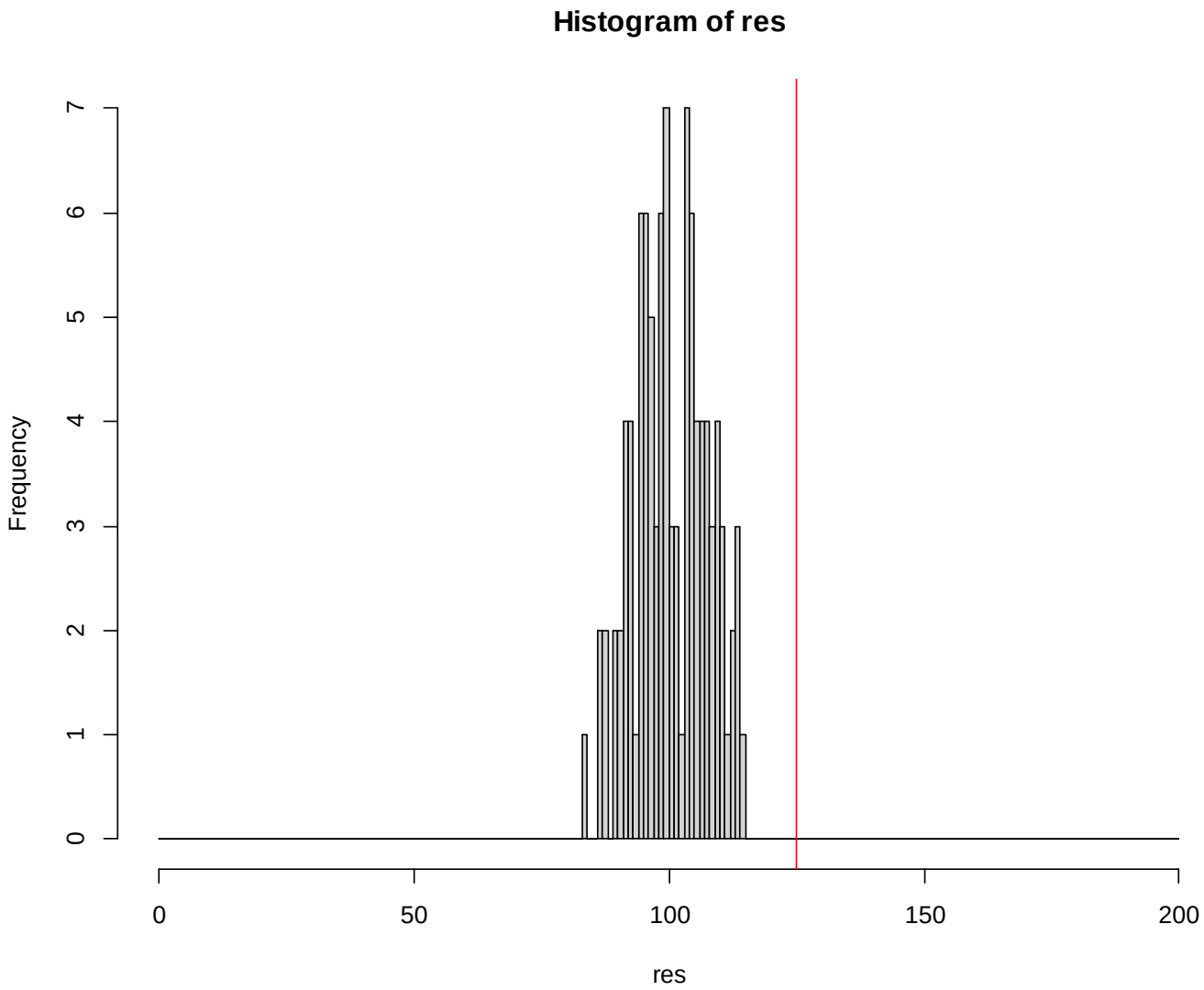
Runs-Test

- Schüler ermogeln Zufallssequenz der Länge 200.
- Wir zählen die Anzahl Runs.
- Wir simulieren 1000 echte Zufallssequenzen der Länge 200
- Wir zählen die Anzahl Runs pro Sequenz
- Wir vergleichen

Runs-Test: Zusammenfassung Hypothesentest

1. **Modell:** Sequenz aus 0/1 der Länge 200
2. **Nullhypothese H_0:** Mit Werfen von fairer Münze
   Alternative H_A: von Hand
3. **Teststatistik:** Anzahl Runs
   Verteilung der Teststatistik, wenn Nullhypothese stimmt: Computersimulation
4. **Signifikanzniveau** $\alpha = 0.01$
5. **Verwerfungsbereich** der Teststatistik:
   Computer beobachtet bei 100 Simulationen nur Sequenzen mit weniger Runs
   Verwerfungsbereich (approximativ):
   “weniger als 84” und “mehr als 115”
6. **Testentscheid**: Der beobachtete Wert (125) liegt im Verwerfungsbereich der Teststatistik.
   Daher wird die Nullhypothese auf dem Signifikanzniveau $\alpha = 0.01$ verworfen.

Runs-Test: Ergebnis



Schlussfolgerung

- Nullhypothese wurde (knapp) verworfen.
- Bei kurzen Sequenzen (<100) ist es relativ leicht, eine Sequenz zu erzeugen, die im Runs-Test nicht auffällt.
- Je länger die Sequenz, desto mehr Macht hat der Runs-Test, d.h., desto eher kann er Abweichungen vom echten Zufall entdecken. Empfehle ≥ 200 für die Schule.
- Kann jemand eine Sequenz der Länge 1000 eintippen, die nicht auffällig ist?
- Echte Test-Prozeduren für Pseudo-Zufallszahlengeneratoren enthalten mehrere solche Tests und sind daher noch viel schwieriger zu schlagen.

Lernziele

- Was ist ein **statistischer Hypothesentest** und wofür wird er verwendet?
  → «Rezept» um eine Behauptung (typischerweise über ein Modell) und eine Beobachtung (Daten) zu vergleichen, inkl. Berücksichtigung des Zufalls
- Was ist die **Struktur** eines Hypothesentests?
  → Immer 6 Schritte
- Ferner: Worum geht es bei **Fehler 1**. bzw. **2. Art** und **Macht**?
  → «Gütekriterien» eines Tests

Zudem:

- Beispiele, die ev. auch in der Schule verwendet werden können
- «Einfache» Anwendungen von R