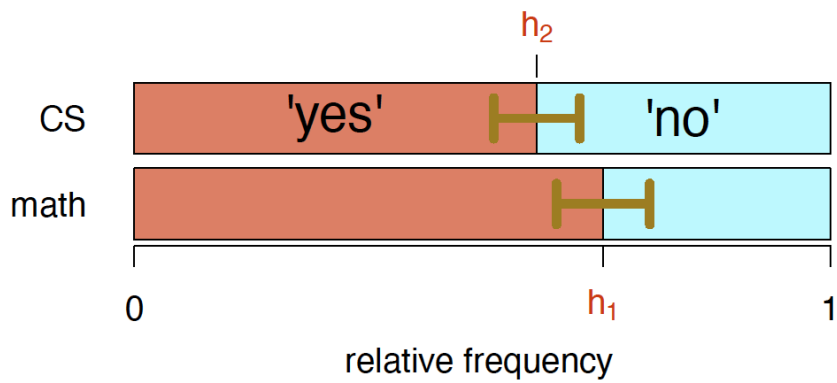
TECHNISCHE UNIVERSITÄT WIEN

Proportions



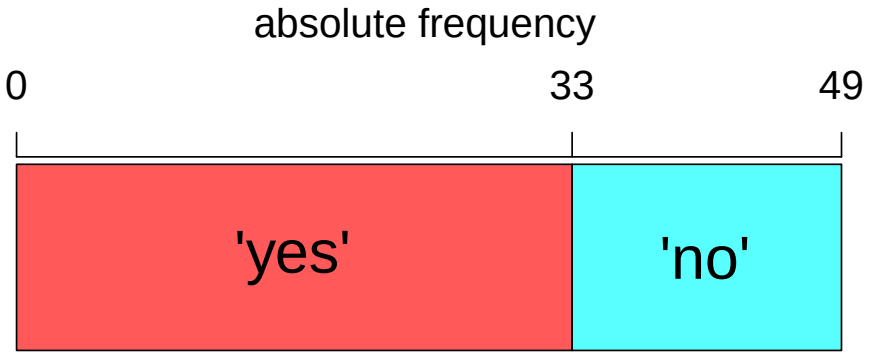
Slides were developed together with Dr. Michael Messer, former ASTAT member
Please send your comments to: T. Levajković, tijana.levajkovic@tuwien.ac.at

All examples are fictitious. All data are simulated and the graphics were created with the statistical program package R.

Sämtliche Beispiele sind frei erfunden. Alle Daten sind simuliert und die Grafiken wurden mit statistischen Programmpaket R erstellt.

Motivation

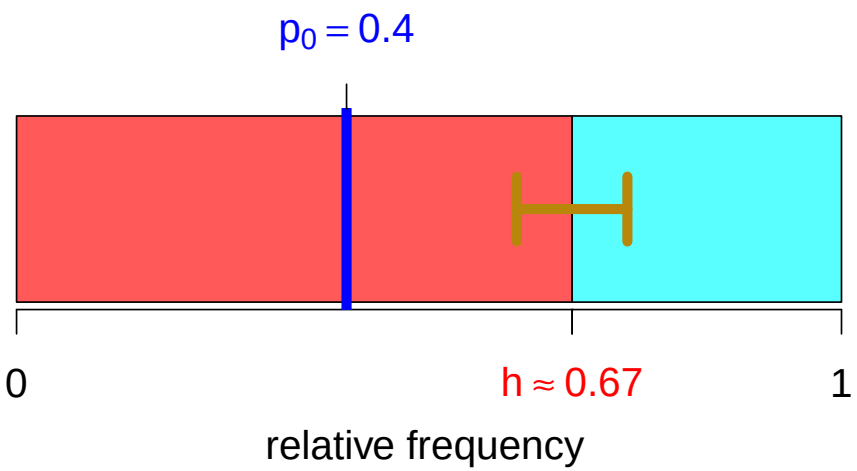
- Somebody claims that among the students with math major the proportion that primarily uses a certain operating system is $p_0 = 40\%$
- How can we, the statisticians, deals with this assertion?
- Collect data!
- Survey among $n = 49$ students: 'Do you use this operating system?'
- Possible answers: *yes* or *no*
  $\rightarrow$ categorical data, two categories
- the survey yields:
  yes, yes, yes, no, yes, no, no, yes, yes, yes, yes, yes, yes, yes, no, yes, no, no, yes, no, no, yes, yes, yes, yes, yes, yes, yes, no, yes, yes, yes, yes, yes, no, yes, no, yes, no, yes, no, yes, no, yes, yes, no, yes, yes, no
- We do understand: nothing?
- Thus, graphical representation, e.g., in a *barplot* (in R: `barplot()`)



Relative frequencies

- Somebody claims that among the students with math major the proportion that primarily uses a certain operating system is $p_0 = 40\%$
- $n = 49$ students interviewed
- In the survey the *absolute frequency* of the users was 33
- This gives a *relative frequency (proportion)* of $h = 33/49 \approx 0.67$

  We use h referring to the German *Häufigkeit* (as f or p are already used elsewhere)
- Question:
  Is the observed proportion h 'far' away from the assertion p_0?
- 'Answer':
  The discrepancy can be judged in the context of a statistical model...
  ...in which we can speak of the variability of the proportion



Relative frequency

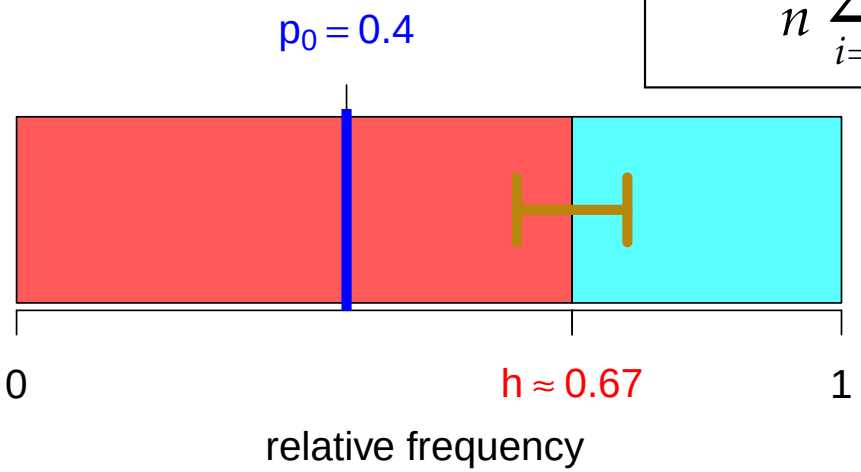
Notation

- n data in two categories (here: 'yes' and 'no')
- Construct binary variables $y_1, \cdots, y_n$ via

$$y_i := \begin{cases} 1, & \text{if the } i\text{-th observation lies in the first category (here: 'yes')} \\ 0, & \text{else} \end{cases}$$

- Then the relative frequency (of the data in the first category) is

$$h := \frac{1}{n} \sum_{i=1}^{n} y_i$$



Note that we deal here with categorical (or nominal) data which is a completely different data type as the metric data considered in the previous lectures. However, we mention here that the proportion h is nothing but a *mean*. Thus, despite the different data type, we will methodologically proceed analogously to the t-tests, comparing means. In that sense, nothing new is going to happen in the following...

Statistical model

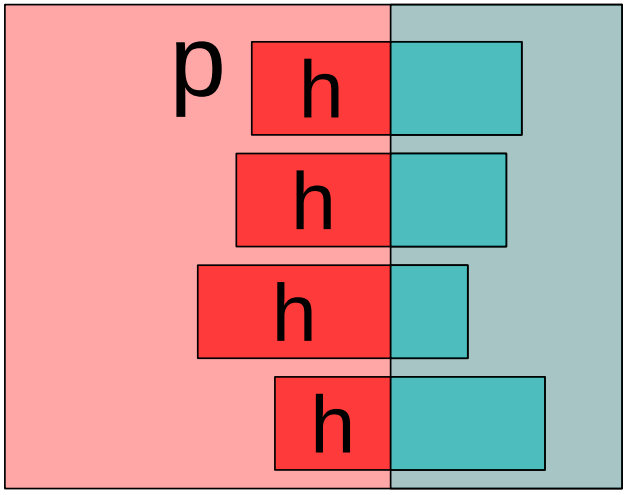
- Model: Let $Y_1, \ldots, Y_n$ be i.i.d. RVs with $Y_1 \sim ber(p)$, and $p \in (0,1)$
- This means

$$Y_i = \begin{cases} 1, & \text{with probability } p \\ 0, & \text{with probability } 1-p \end{cases}$$

- The relative frequency (of 'successes') is given as

$$\boxed{H := \frac{1}{n} \sum_{i=1}^{n} Y_i}$$

- H is a random variable... and has a standard deviation
- Interpretation: h based on the data is a realization of H



Asymptotic normality of H

- Let $Y_1, \ldots, Y_n$ be i.i.d. RVs with $Y_1 \sim ber(p)$. Let

  note that $\sum^n Y_i \sim b(n,p)$

$$H = \frac{1}{n} \sum_{i=1}^{n} Y_i$$

- It holds for the expectation and the variance

  'linearity of the expectation and independence'

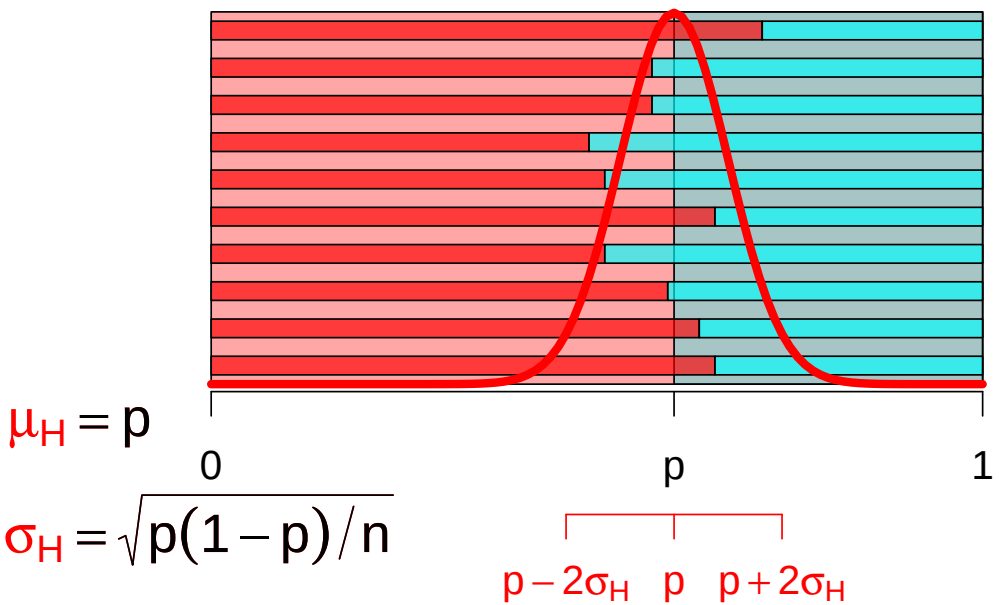
$$\mathbb{E}[H] = p \qquad \text{and} \qquad \mathbb{V}ar(H) = \frac{p(1-p)}{n}$$

- and for $n \to \infty$

$$\frac{H - p}{\sqrt{p(1-p)/n}} \xrightarrow{d} N(0,1)$$

'central limit theorem'

Note: in the t-test we found the rescaled mean to be exactly(!) t-distributed, $(\bar{X} - \mu)/\sqrt{S^2/n} \sim t(n-1)$, when assuming normally distributed RVs. Here, we do not make the normal assumption on the RVs and thus will make use of the normal approximation of H



The standard error of the relative frequency

- Let $Y_1, \ldots, Y_n$ be i.i.d. RVs with $Y_1 \sim ber(p)$ and $p \in (0,1)$

$$H = \frac{1}{n} \sum_{i=1}^{n} Y_i$$

- Variance

$$\mathbb{V}ar(H) = \frac{p(1-p)}{n}$$

- Problem: p unknown in practice
- Solution: Estimate p via H
- Definition: The standard error of H is

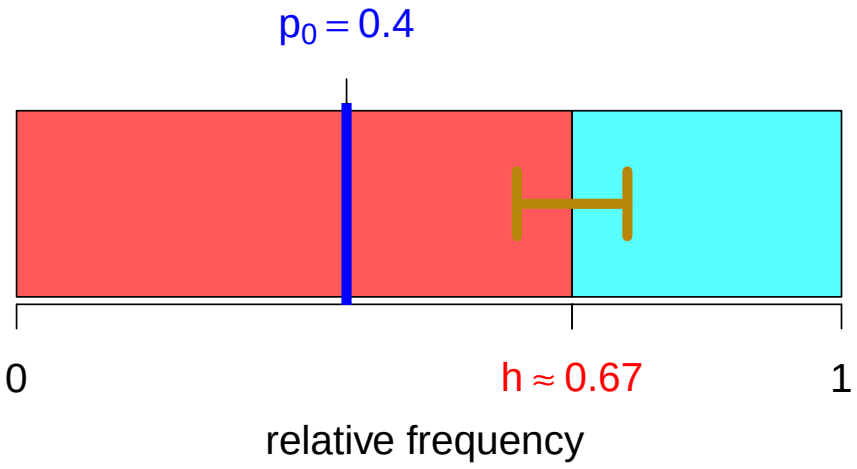
$$\boxed{SE_H := \sqrt{\frac{H(1-H)}{n}}}$$

- Summary:
  - H is approximately normally distributed, for large n
  - in expectation H hits the true unknown parameter p
  - the estimated standard deviation of H is SE_H
  - deviations from the expectation of order $1 \cdot SE_H$ are not unlikely
  - but deviations of 'many' SE_H are unlikely

The standard error of the relative frequency

- the estimation based on the data yields

$$se_h = \sqrt{\frac{h(1-h)}{n}} \approx 0.07$$



- the discrepancy of the observed frequency $h = 0.67$ and the claimed proportion $p_0 = 0.4$ is

$$|h - p_0| \approx 4.1 \cdot se_h$$

- this is extremely far, given the typical deviation of H to be about $1 \cdot SE_H$

Asymptotic one-sample test for frequencies

- Let $Y_1, \ldots, Y_n$ be i.i.d. RVs with $Y_1 \sim ber(p)$ and $p \in (0,1)$
- and let $q_{1-\alpha/2}$ be the $(1-\alpha/2)$-quantile of the $N(0,1)$-distribution

Under $H_0 : p = p_0$ it holds approximately for large n that

$$Z := \frac{H - p_0}{SE_H} \overset{d}{\approx} N(0,1)$$

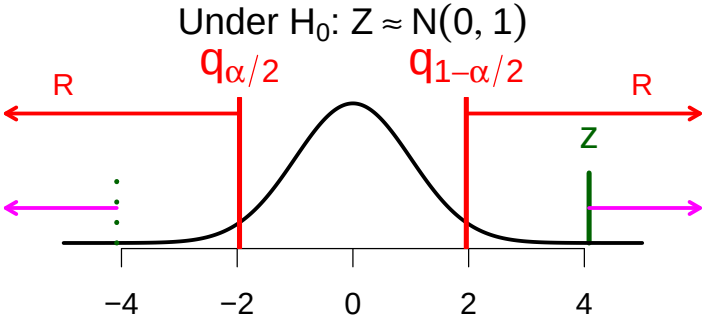
and equivalently: the confidence interval

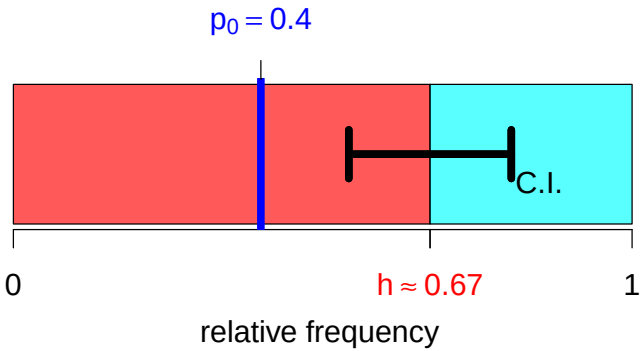
$$I := \left(H - q_{1-\alpha/2} \cdot SE_H,\ H + q_{1-\alpha/2} \cdot SE_H\right)$$

overlaps the parameter p_0 with probability about $1-\alpha$

- 'Structure' as in the t-test: $Z = (\spadesuit - \spadesuit)/\spadesuit$ and $I = (\spadesuit - q \cdot \spadesuit,\ \spadesuit + q \cdot \spadesuit)$
- Thus again: Equivalence of test and confidence interval
  $\alpha = \mathbb{P}_{H_0}(Z \in R) = \cdots = \mathbb{P}_{H_0}(I \not\ni p_0)$ (while R denotes the rejection area of the two-sided test)

Evaluation of the data



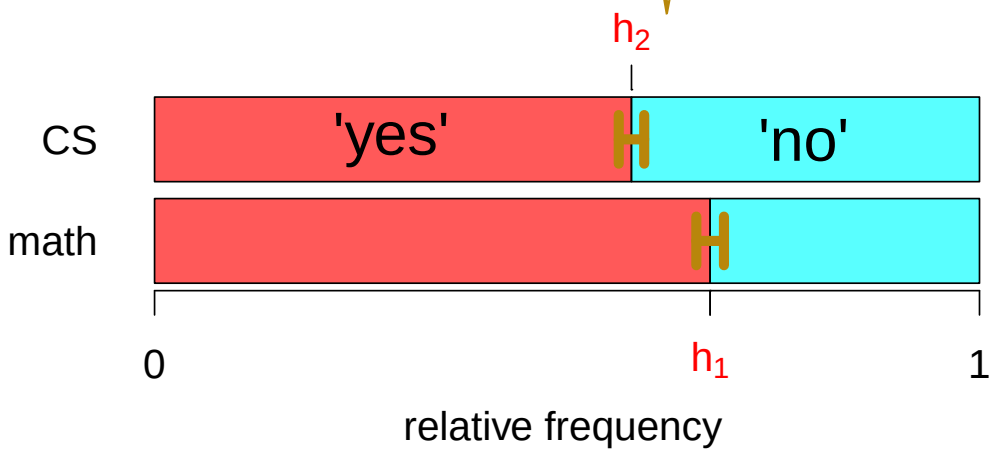


- For the significance level $\alpha = 5\%$ it is $q_{1-\alpha/2} \approx 1.96$
- $\rightarrow$ rejection area $R \approx (-\infty, -1.96] \cup [1.96, \infty)$ (two-sided)
- Evaluation of the data
$$z = \frac{h - p_0}{se_h} \approx 4.1$$
- as $z \in R$, we reject H_0 on the 5%-level
- the p-value is $p = \mathbb{P}_{H_0}(|Z| \geqslant |z|) \approx 4.5 \cdot 10^{-5}$ (tiny)
- equivalently: the 95%-confidence interval
$$i \approx (0.54, 0.80)$$
does not overlap the claimed parameter p_0
- Interpretation: the proportion observed in the data is barely compatible with the claimed population proportion. If H_0 holds true, then we observe in less than one of 2000 cases a discrepancy which is at least as extreme as in the data ($p < 1/2000$) $\rightarrow$ we are very inclined to doubt H_0

Transition to the two-sample situation

- Somebody claims that the proportion p_1 of students with a math major that primarily use a certain operating system equals the proportion p_2 of students with a computer science (CS) major that use it
- $n_1 = 49$ students with a math- and $n_2 = 64$ with a CS major interviewed
- observed absolute frequencies of users are 33 and 37
- this gives the relative frequencies $h_1 \approx 0.67$ and $h_2 \approx 0.58$
- Question: Are the proportions h_1 and h_2 far away?
- Answer: No! The standard errors se_{h_1} and se_{h_2} (typical variability) are large in relation to the distance

It is $se_{h_1} = \sqrt{\frac{h_1(1-h_1)}{n_1}}$ and $se_{h_1} = \sqrt{\frac{h_2(1-h_2)}{n_2}}$



Frequencies without standard errors are way less meaningful
Example:16times more students asked→quatering of the se_h → distance 'large'

Statistical model

- Model: Let $Y_{1,1}, \ldots, Y_{1,n_1}, \; Y_{2,1}, \ldots, Y_{2,n_2}$ be independent RVs with $Y_{1,i} \sim ber(p_1)$ for $i = 1, \ldots, n_1$ and $Y_{2,j} \sim ber(p_2)$ for $j = 1, \ldots, n_2$, and $(p_1, p_2) \in (0,1)^2$
  - particularly both groups have their own success probability
- Null hypothesis $H_0 : p_1 = p_2$
  - no difference in the success probabilities
- For the construction of the test statistic we need
  - first: the relative frequencies (of 'successes') in both groups

$$H_1 := \frac{1}{n_1} \sum_{i=1}^{n_1} Y_{1,i} \qquad \text{and} \qquad H_2 := \frac{1}{n_2} \sum_{j=1}^{n_2} Y_{2,j}$$

  - second: a standard error of the difference of H_2 and H_1, via

$$\sqrt{SE_{H_2}^2 + SE_{H_1}^2}$$

while

$$SE_{H_1} = \sqrt{\frac{H_1(1-H_1)}{n_1}} \qquad \text{and} \qquad SE_{H_2} = \sqrt{\frac{H_2(1-H_2)}{n_2}}$$

Two-sample test for frequencies

- Let $Y_{1,1}, \ldots, Y_{1,n_1}$, $Y_{2,1}, \ldots, Y_{2,n_2}$ be independent RVs with $Y_{1,i} \sim ber(p_1)$ for $i = 1, \ldots, n_1$ and $Y_{2,j} \sim ber(p_2)$ for $j = 1, \ldots, n_2$,
  and $(p_1, p_2) \in (0,1)^2$
- let $q_{1-\alpha/2}$ be the $(1-\alpha/2)$-quantile of the $N(0,1)$-distribution

Under $H_0 : p_2 - p_1 = 0$ it holds (approximately)

$$Z := \frac{(H_2 - H_1) - 0}{\sqrt{SE_{H_2}^2 + SE_{H_1}^2}} \overset{d}{\approx} N(0,1)$$

and equivalently: the confidence interval

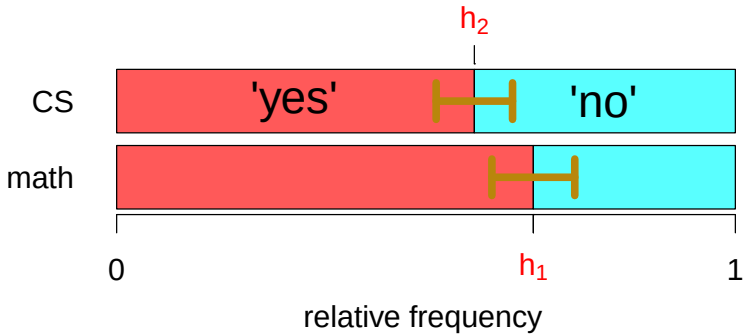
$$I := \left((H_2 - H_1) - q_{1-\alpha/2} \cdot \sqrt{SE_{H_2}^2 + SE_{H_1}^2},\ (H_2 - H_1) + q_{1-\alpha/2} \cdot \sqrt{SE_{H_2}^2 + SE_{H_1}^2} \right)$$

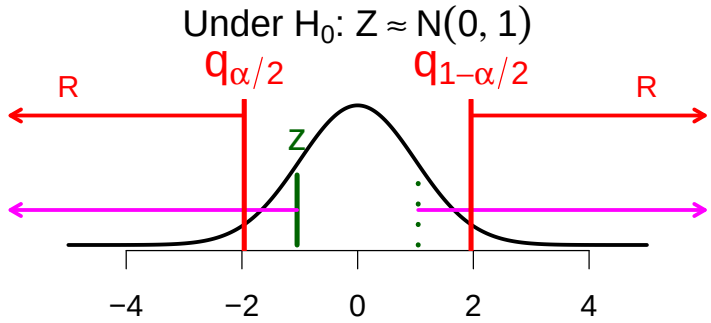
overlaps 0 with probability about $1 - \alpha$

Again:

- known structure: $Z = (\spadesuit - \spadesuit)/\spadesuit$ and $I = (\spadesuit - q \cdot \spadesuit,\ \spadesuit + q \cdot \spadesuit)$
- more precisely, Z has the structure of the Welch-statistic
- direct generalization for the difference $p_2 - p_1 = d_0 \neq 0$

Evaluation of the data





- For the significance level $\alpha = 5\%$ it is $q_{1-\alpha/2} \approx 1.96$
- $\rightarrow$ rejection area $R \approx (-\infty, -1.96] \cup [1.96, \infty)$ (two-sided)
- Evaluation of the data

$$z = \frac{h_2 - h_1}{\sqrt{se_{h_2}^2 + se_{h_1}^2}} \approx -1.05$$

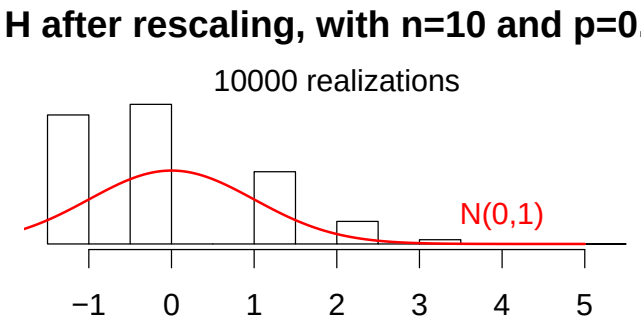
- Because $z \notin R$ we cannot reject H_0 on the 5%-level
- the p-value is $p = \mathbb{P}_{H_0}(|Z| \geqslant |z|) \approx 0.3$
- equivalently: the 95%-confidence interval

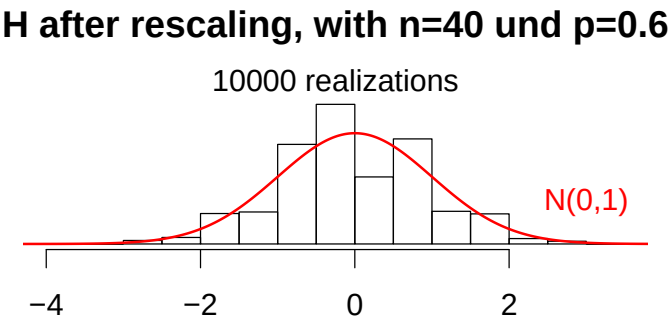
$$i \approx (-0.27, 0.08)$$

  does overlap zero
- Interpretation: the discrepancy of the observed proportions barely gives us reason to doubt the null hypothesis. If H_0 holds true, then in about every third case we will observe a discrepancy, that is at least as large as in our data ($p \approx 1/3$)

Remark

- In R the associated test and confidence intervals are implement in e.g., `prop.test()`
- However, some statistics are slightly adjusted
- the main reason: the approximation through the normal distribution can be quite 'rough'
- In general, one should be cautious if either the sample size n is small, or if the frequencies h (resp. p) are close to either 0 or 1





- The rescaled frequency H is

$$\frac{H-p}{\sqrt{p(1-p)/n}} \overset{d}{\approx} N(0,1)$$

- left: approximation unreasonable :-(right: quite plausible :-)

Multiple-choice questions

(1) In the one sample situation for $n > 1$ binary data let h denote the relative frequencies of 'successes'. Which statement is in general **correct**?

a. $h \geqslant h(1-h)$

b. $1/h \in (0,1]$, if $h > 0$

c. $h^2 \geqslant h(1-h)$

d. $1/h^2 \in (0,1]$, if $h > 0$

Multiple-choice questions

(2) A study is to be performed to estimate the proportion of voters who believe the economy is "heading in the right direction." Which of the following pairs of sample size n and population proportion p will result in the smallest variance for the sampling distribution of $\hat{p}$?

a. $n = 100$ and $p = 0.1$
b. $n = 100$ and $p = 0.99$
c. $n = 1000$ and $p = 0.1$
d. $n = 1000$ and $p = 0.5$

Multiple-choice questions

(3) An association of realtors estimates that 23% of all homes purchased in 2022 were considered investment properties. If a sample of 800 homes sold in 2022 is obtained and it was noted that 248 homes were to be used as investment property. Are the sample data compatible with the claim of the association of realtors?

a. Yes, because the p-value is larger than 10%.
b. No, because the sample size is not large enough.
c. No, because the p-value is larger than 10%.
d. Cannot be determined.

Multiple-choice questions

(4) In general, how does halving the sample size change the confidence interval size?

a. Doubles the interval size.
b. Halves the interval size.
c. Divides the interval size by $\sqrt{2}$.
d. Multiplies the interval size by $\sqrt{2}$.

Multiple-choice questions

(5) A coin is tossed 1000 times and 540 heads appeared. Anna wants to test the claim that this is not a biased coin. What is the p-value of this hypothesis testing?

a. 0.001
b. 0.011
c. 0.110
d. none of the above

Multiple-choice questions

(6) Suppose $H_0 : p = 0.4$ and the power of the test for the alternative hypothesis $p = 0.35$ is 0.75. Which one of the following statements is a valid conclusion?

a. The probability of committing a Type I error is 0.05.
b. The probability of committing a Type II error is 0.65.
c. If the null hypothesis is false, the probability of failing to reject it is 0.65.
d. If the alternative $p = 0.35$ is true, the probability of failing to reject H_0 is 0.25.

Multiple-choice questions

(7) Choosing a smaller level of significance α results in

a. a lower Type II error and lower power.
b. a lower Type II error and higher power.
c. a higher Type II error and lower power.
d. a higher Type II error and higher power.

Multiple-choice questions

(8) A pharmaceutical company claims that 8% or fewer of the patients taking their new statin drug will have a heart attack in a five-year period. In a government-sponsored study of 2300 participants taking the new drug, 198 have heart attacks in a five-year period. Is this strong evidence against the company claim?

a. yes, because the p-value is 0.005657
b. yes, because the p-value is 0.086087
c. no, because the p-value is only 0.005657
d. no, because the p-value is over 0.10.

Multiple-choice questions

(9) A private opinion poll is conducted for a politician to determine what proportion of the population favors adding more national parks. How large a sample is needed in order to be 99% confident that the sample proportion will not differ from the true proportion by more than 3%?

a. 22
b. 1844
c. 1509
d. 3684

Multiple-choice questions

(10) Two features of a novel operating system are compared using a two-sample t-test. The statistics for the first feature are $\bar{x} = 15$, $s_x^2 = 55$ and $n_x = 5$ and those for the second feature are $\bar{y} = 21$, $s_y = 10$ and $n_y = 4$. The rejection region is given through $R = (-\infty, -q] \cup [q, \infty)$. Then it holds degrees of freedom for the t-distribution increase, the distribution approaches

a. we reject for $q = 2.5$ but not for $q = 1.5$.
b. we reject for both $q = 2.5$ and $q = 1.5$.
c. we do not reject for $q = 2.5$ but for $q = 1.5$.
d. we do neither reject for $q = 2.5$ nor for $q = 1.5$.

Thank you for your attention!