Grundpraktikum Bioinfo - Week 1

Biological Data Analysis

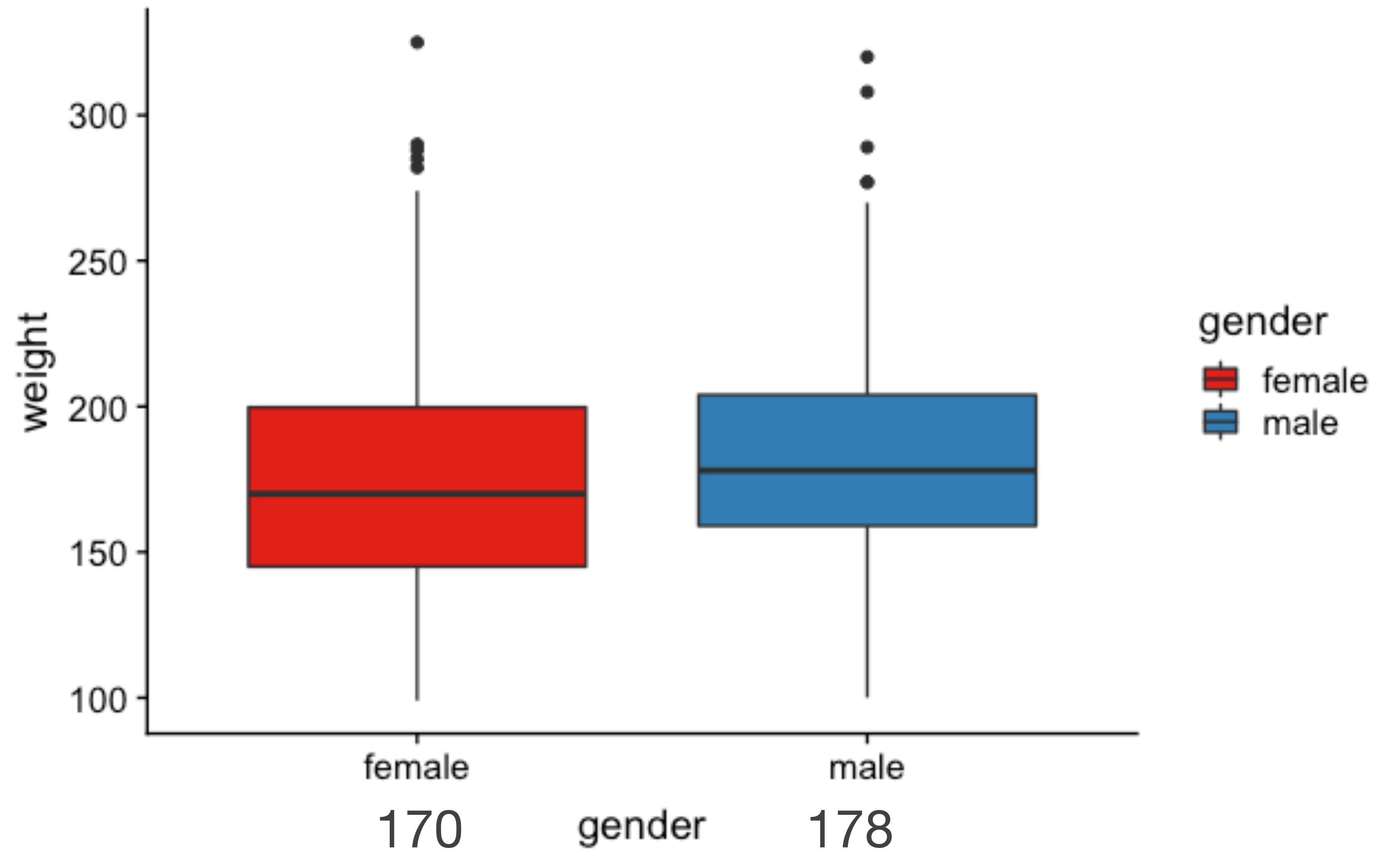
Carl Herrmann
IPMB - Universität Heidelberg

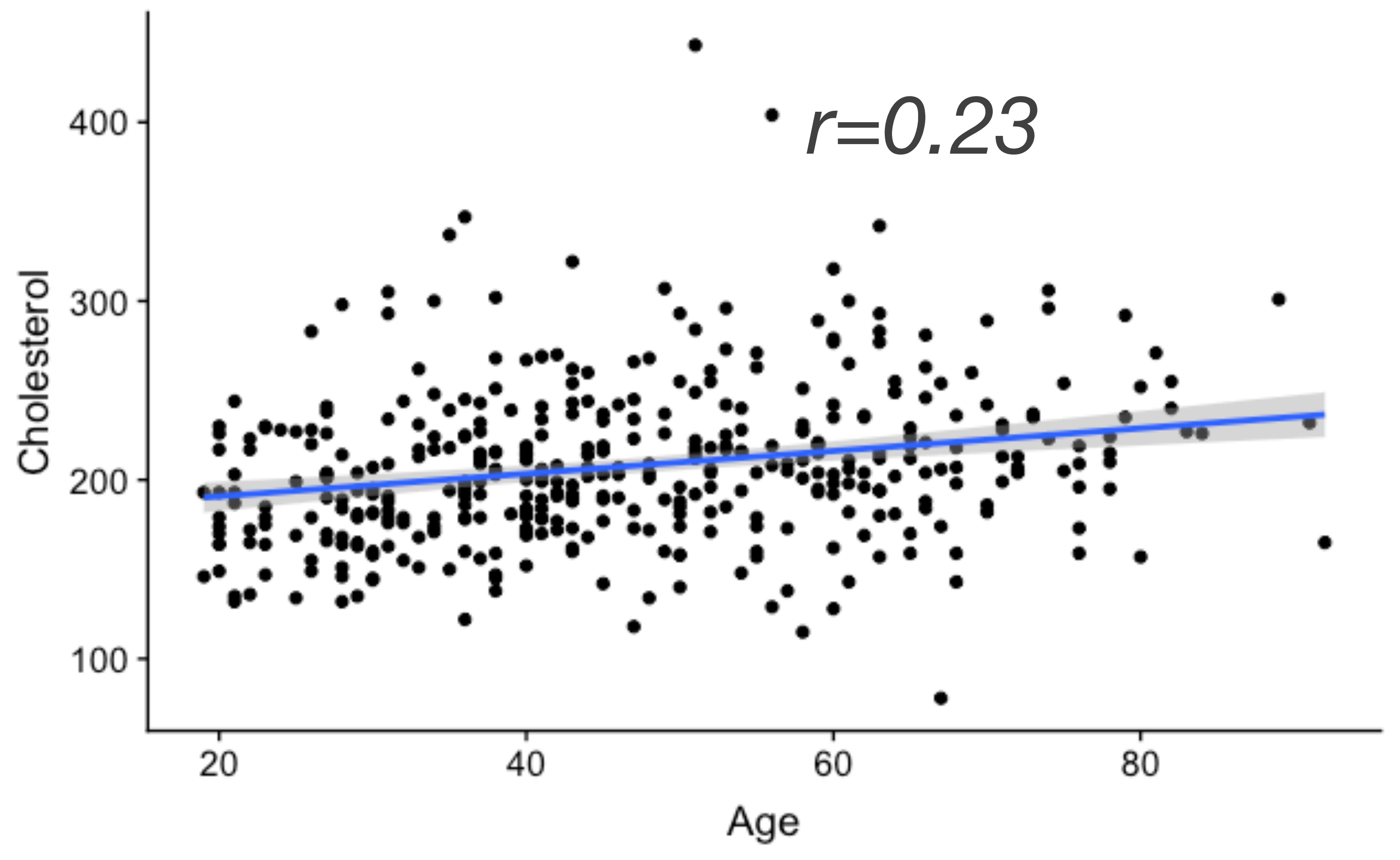
IPMB
Institut für Pharmazie und Molekulare Biotechnologie

UNIVERSITÄT HEIDELBERG
ZUKUNFT SEIT 1386

7. Hypothesis testing

Are observations significant?





- For the cohort, we observe:
  - a difference in man/women weights
  - a non-zero correlation between age and cholesterol
- But:
  - would we observe this in another cohort??
  - Does this hold for the entire (unknown) population?

    → ***is this difference/correlation significant?***

Hypothesis testing: what do we need?

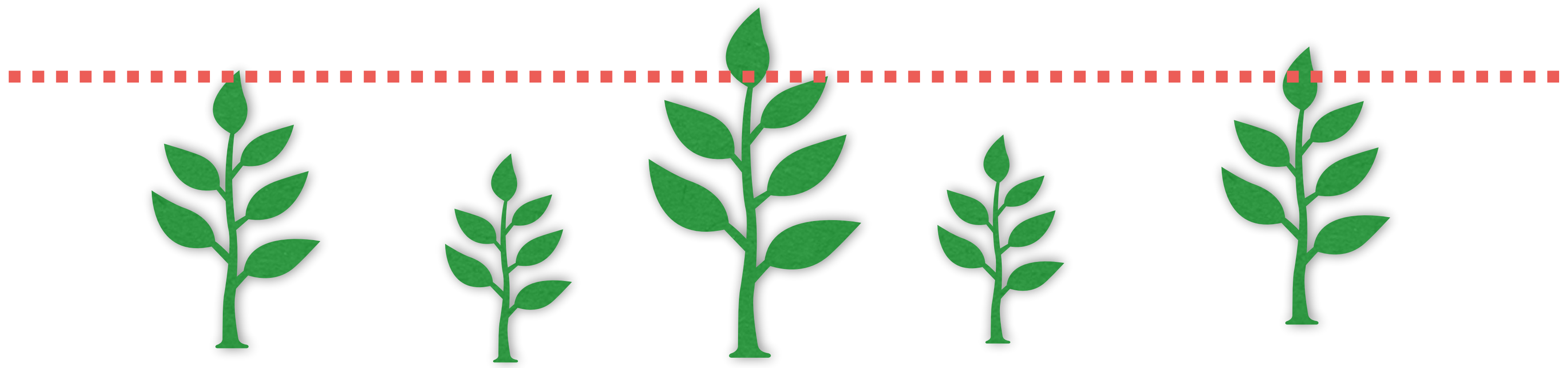
Question	is there a GENERAL weight difference between men/women?	is there a GENERAL non-zero correlation between age/cholesterol?
Random variables	X_m, X_w = weights men/women	X_{age}, X_{chol} : age/ cholesterol level
Null hypothesis (H_0)	no difference between the expectations of the random variables $E(X_m) = E(X_w)$	no correlation between age and cholesterol $cor(X_{age},X_{chol}) = 0$
Alternative hypothesis (H_1)	expectations of the random variables are different $E(X_m) \neq E(X_w)$	correlation of the random variable is not zero $cor(X_{age},X_{chol}) \neq 0$

We are considering the random variables, not the realizations!

Example 1

- **Study**: effect of fertilizer F1 on plant growth
  - no-fertilizer: ***h* = 1.5 m**
  - fertilizer on n = 10 samples:

    $x = \{1.47, 1.62, 1.51, 1.61, 1.27, 1.51, 1.55, 1.49, 1.44, 1.5\}$
- **Random variable**: plant height X after treatment with F1



- **Question: does the treatment with fertilizer enhance plant growth?**

$$\bar{x} = 1.497 \text{ m} \longleftrightarrow h = 1.5 \text{ m}$$

Example 1

- **Question: does the treatment with fertilizer enhance plant growth?**
- **Hypothesis**:
  - H0 : no → $E(X) \leq h = 1.5m$
  - H1: yes → $E(X) > h = 1.5m$
- **Effect size:** $\bar{x} - h = -0.003$
- **size of random effect:** $s/\sqrt{n} = 0.031$

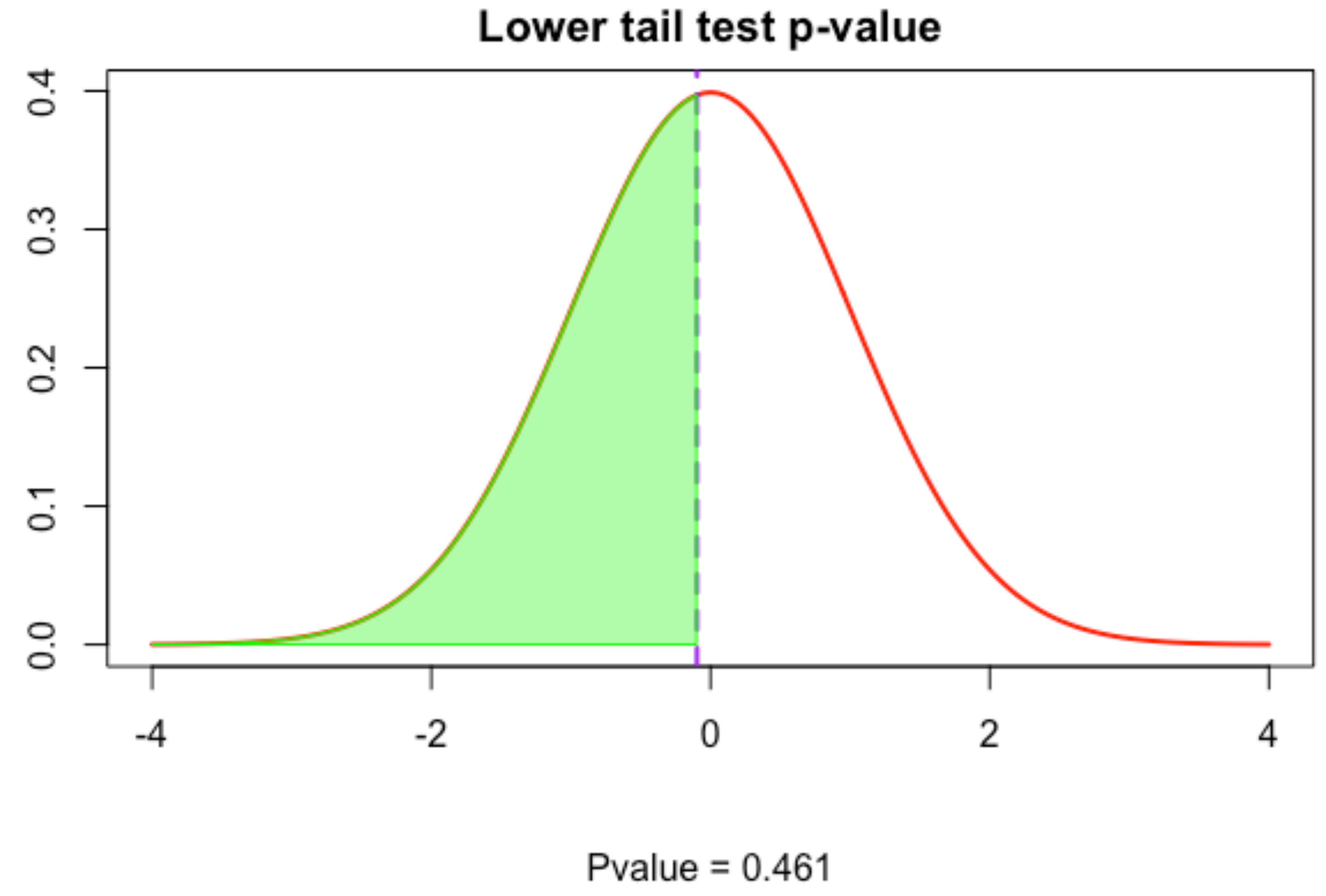
$$t = \frac{\bar{x} - h}{s/\sqrt{n}} = -0.09$$

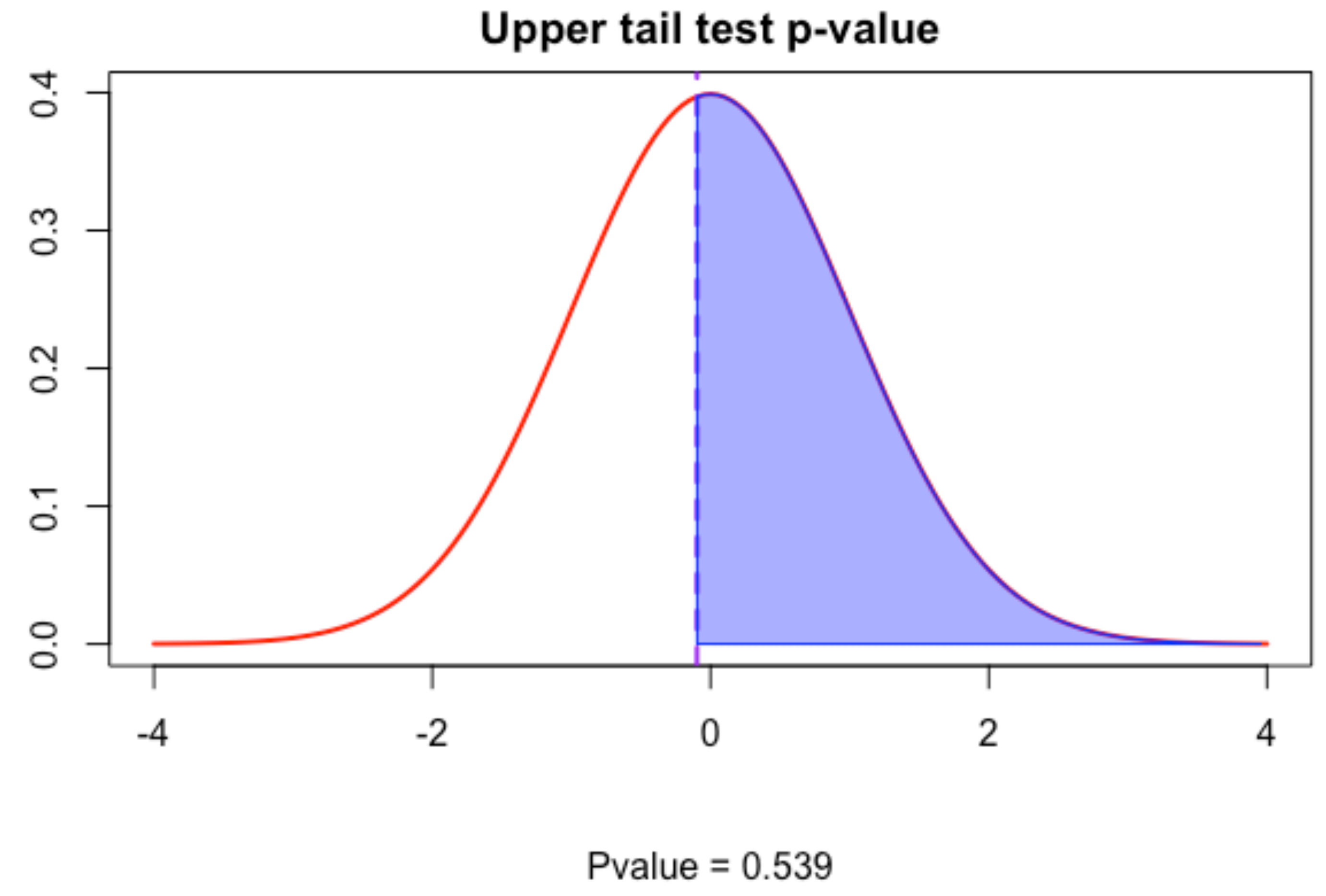
s = standard deviation of sample

- What are typical values of *t* **under the *H_0* hypothesis**?

Example 1

- Distribution of t under the H_0 hypothesis
- Vertical line = observed value of test statistics t
- **Green** = probability to observe under H_0 a lower value of t
- **Blue** = probability to observe under H_0 a larger value of t
- Here: Blue = 53.9% of total area





Conclusion: if H0 (= no effect) is true, there is a **53.9% probability** to observe a value of t larger or equal to the one observed

→ ***not unlikely, hence no reason to distrust H0 (= no effect)***

Example 2

- **Study**: effect of fertilizer F2 on plant growth
  - no-fertilizer: h = 1.5m
  - fertilizer on n = 10 samples: x = `{1.47,1.62,1.61,1.61,1.47,1.51,1.55,1.59,1.64,1.5}`
- **Random variable**: plant height X after treatment with F2
- **Question: does the treatment with fertilizer enhance plant growth?**
- **Hypothesis**:
  - H0 : no → $E(X) \leq h = 1.5m$
  - H1: yes → $E(X) > h = 1.5m$
- Effect size: $\bar{x} - h = 0.057$
- size of random effect: $s/\sqrt{n} = 0.02$

$$t = \frac{\bar{x} - h}{s/\sqrt{n}} = 2.77$$

s = standard deviation of sample

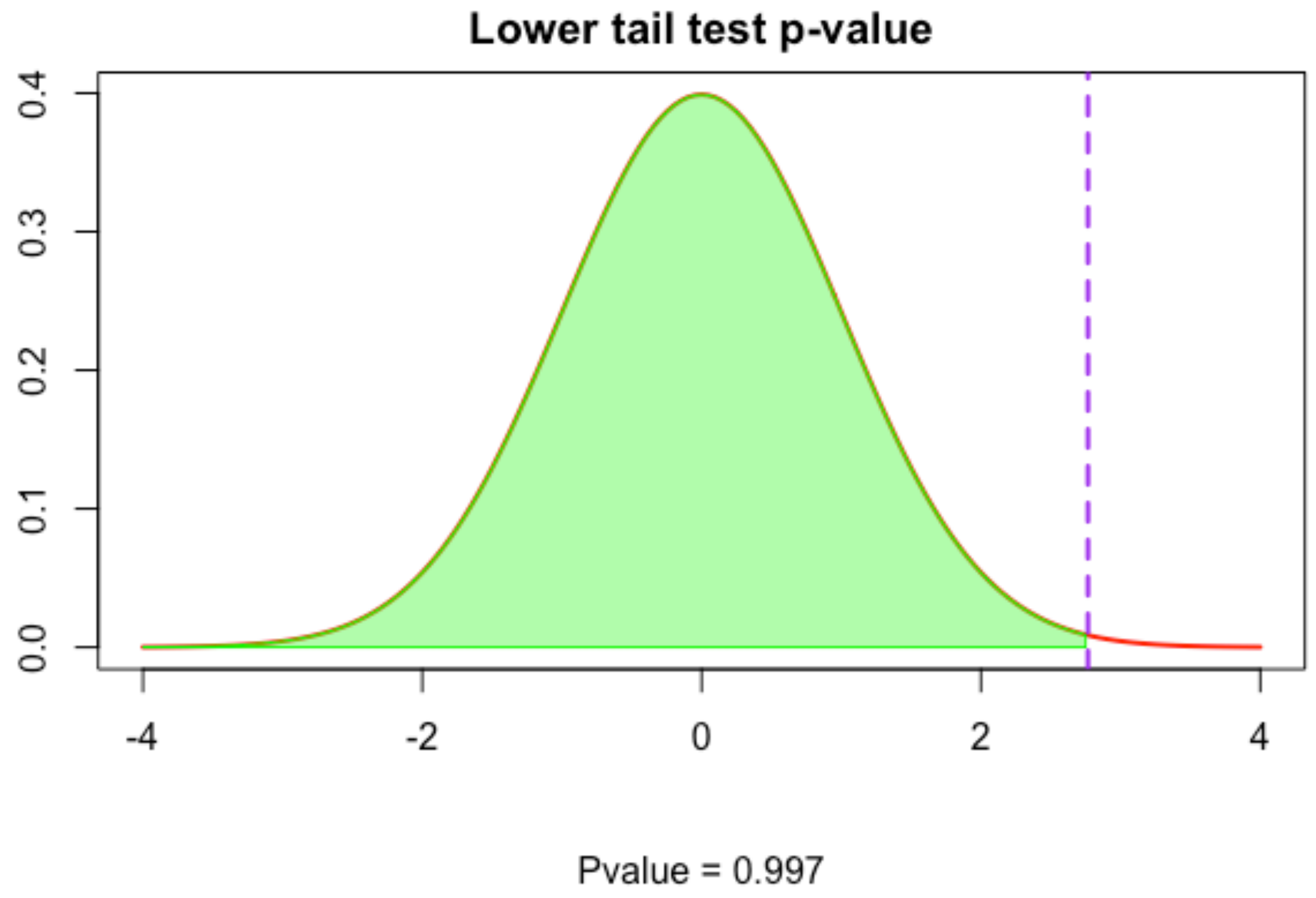
- What are typical values of t **under the H_0 hypothesis**?

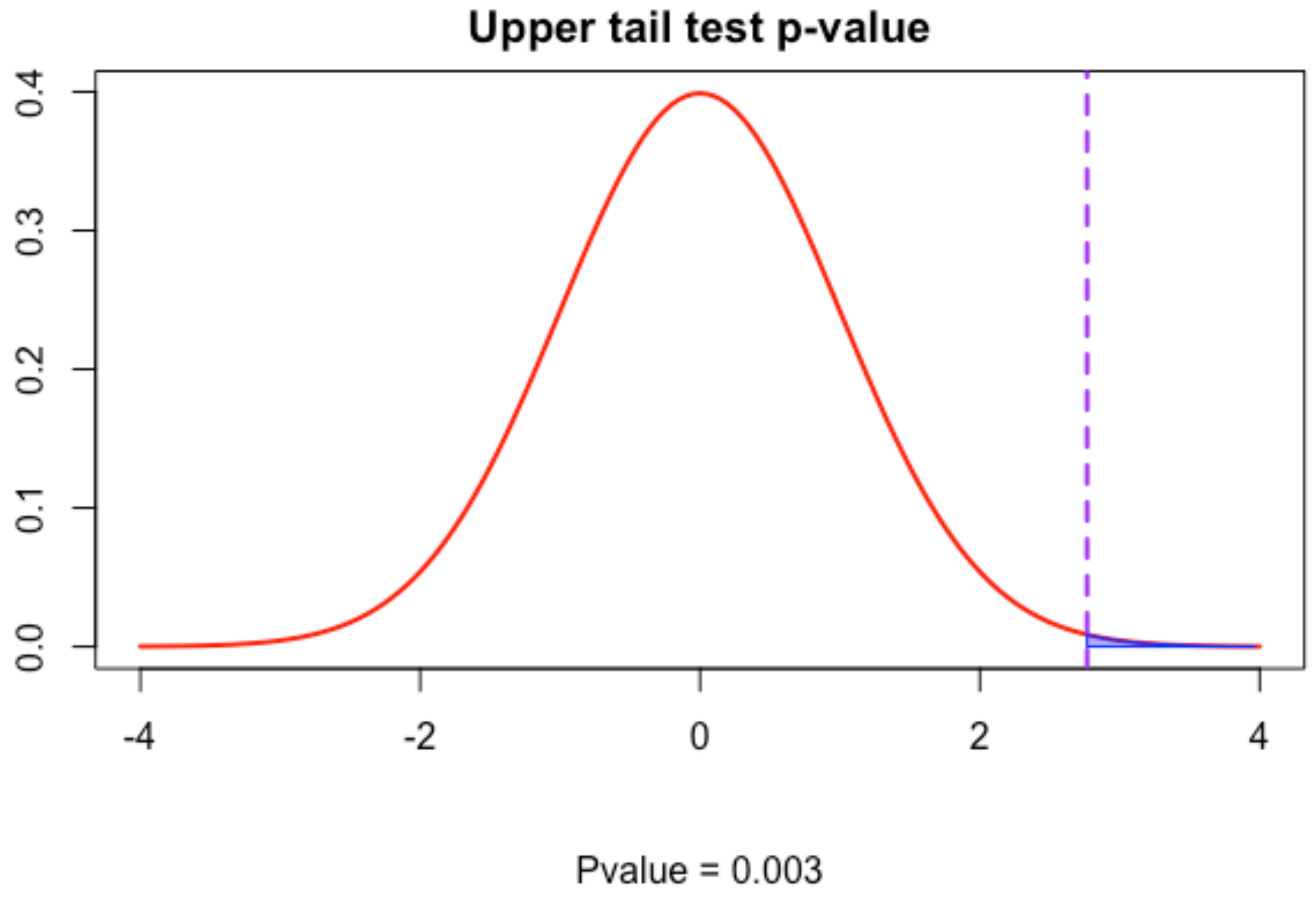
Example 2

- Distribution of t under the H_0 hypothesis
- Vertical line = observed value of test statistics t
- Green = probability to observe under H_0 a lower value of t
- Blue = probability to observe under H_0 a larger value of t
- Here: Blue = 0.3% of total area

Conclusion: if H0 (= no effect) is true, there is a **0.3% probability** to observe a value of t larger or equal to the one observed
→ ***very unlikely, H0 is probably not true and should be rejected***

$$t = 2.77$$





Example 3

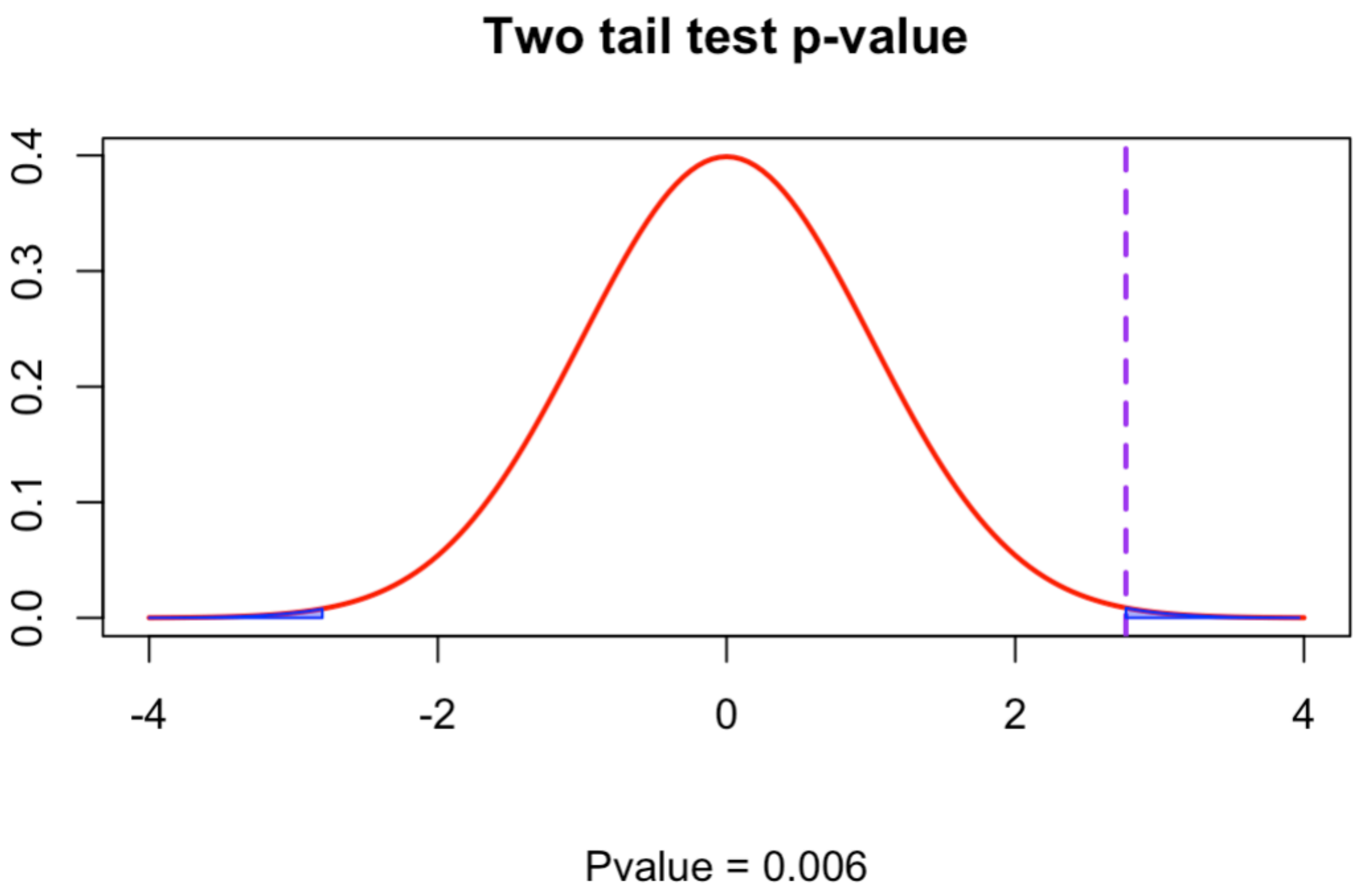
- **Study**: effect of fertilizer F2 on plant growth
  - no-fertilizer: h = 1.5m
  - fertilizer on n = 10 samples: x `= {1.47,1.62,1.61,1.61,1.47,1.51,1.55,1.59,1.64,1.5}`
- **Random variable**: plant height X after treatment with F2
- **Question: does the treatment with fertilizer <u>influence</u> plant growth?**
- **Hypothesis**:
  - H0 : no → $E(X) = h = 1.5m$
  - H1: yes → $E(X) \neq h = 1.5m$
- Effect size: $\bar{x} - h = 0.057$
- size of random effect: $s/\sqrt{n} = 0.02$

$$t = \frac{\bar{x} - h}{s/\sqrt{n}} = 2.77$$

s = standard deviation of sample

- What are typical values of t **under the H_0 hypothesis**?

What was the question again?



blue area = 0.6%: H0 very unlikely

P-value

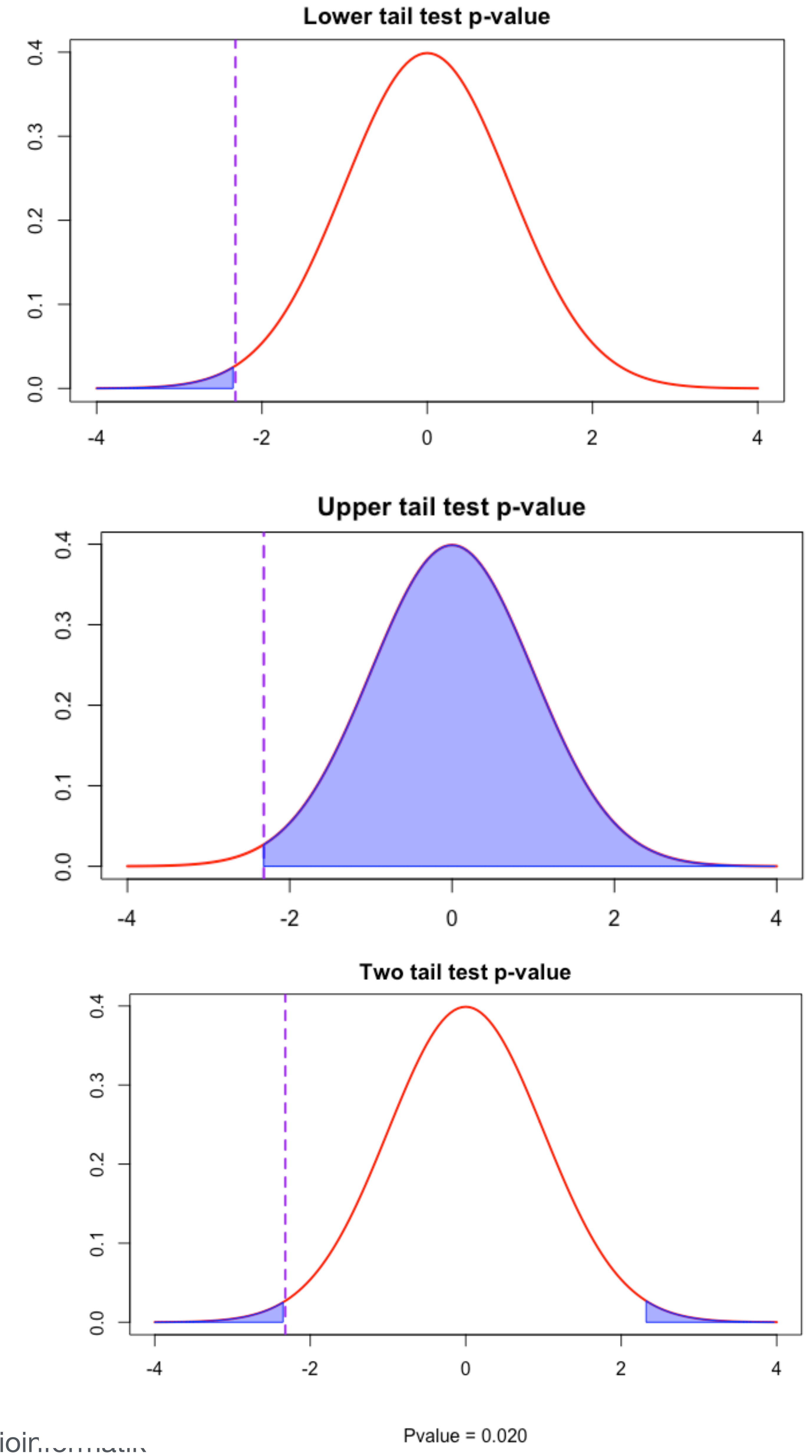
the p-value is the **probability** of obtaining a

- **larger** (one-sided upper tail)
- **smaller** (one-sided lower tail)
- **more extreme** (two-sided or two tailed)

value of the test statistics **if H_0 is valid!**

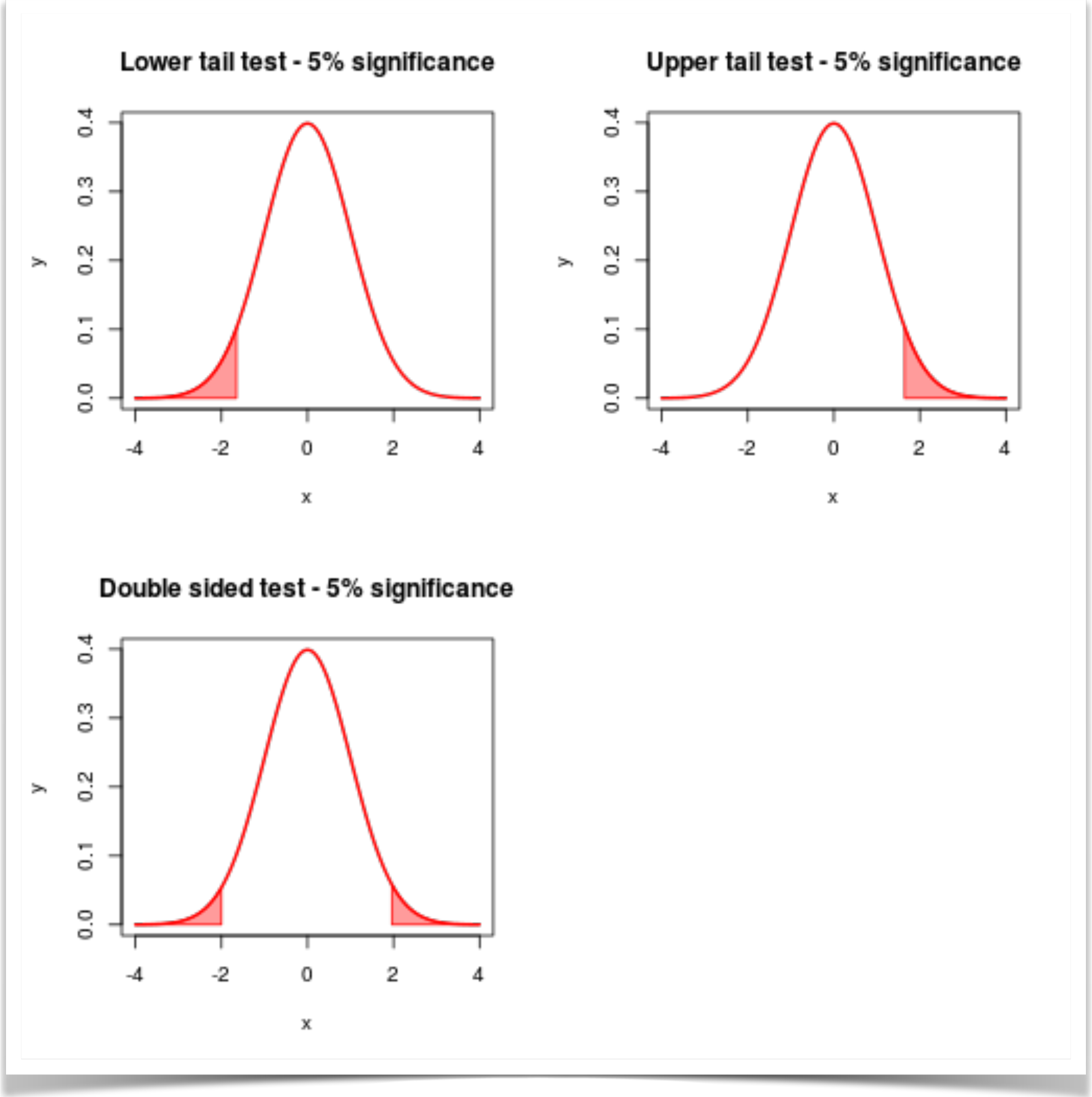
The probability of the two sided test is **twice** the smallest probability of the upper-tail or lower-tail test

$$p_{2sided} = 2\ min(p_{lower-tail}, p_{upper-tail})$$



Significance

- When is a probability low, very low, or high?
- Define a **significance level α**
- $p < \alpha$:
  - H_0 hypothesis can be rejected
  - the observed effect is significant
  - H_1 is statistically proven
- $p > \alpha$:
  - effect is not sufficient to reject H_0
  - observed effect is compatible with statistical fluctuations
  - H_0 is not proven, maybe with a larger sample, the effect could become significant
- $\alpha = 0.05$ has become a standard value (but no golden rule!)

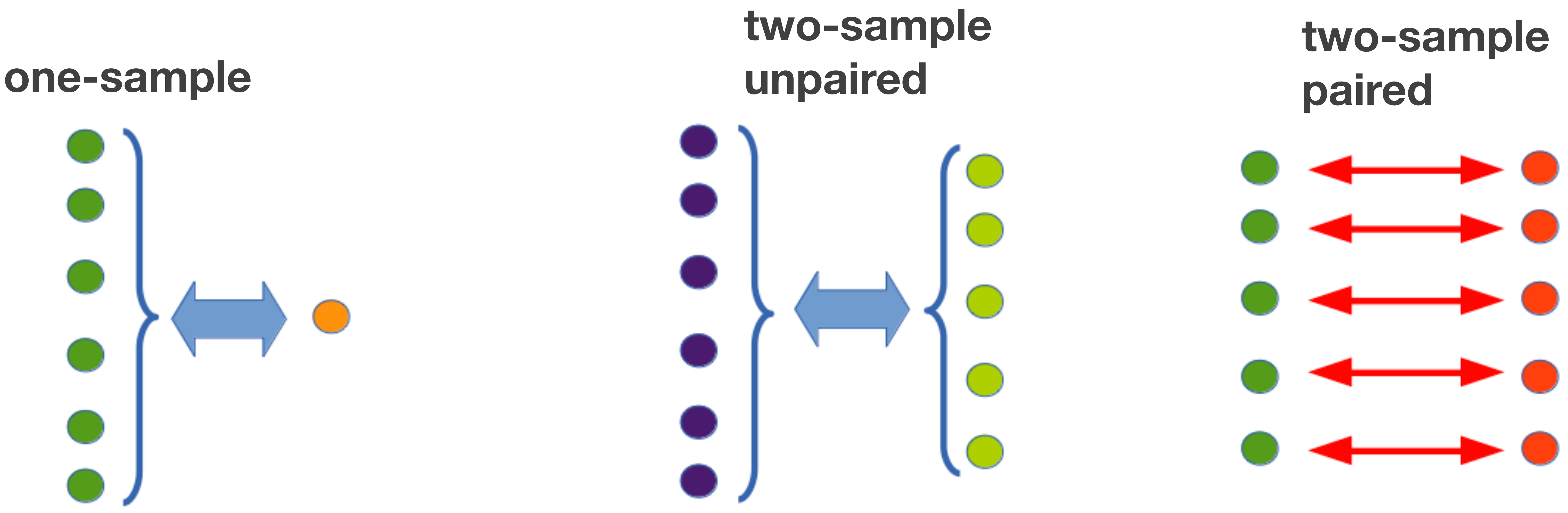


7. Hypothesis testing

Testing the mean - t-tests

Test on mean values

- Hypothesis on mean values can be investigated using a ***t-test***
- Family of tests with different version:
  - **one-sample test**: *is the mean body temperature 37.7 C?*
  - **two-sample test, unpaired**: *do men and women have different mean cholesterol levels?*
  - **two-sample test, paired**: *is there a change in cholesterol level after a one-month egg rich diet?*



(do both samples have equal variance?)

Running a t-test in R

two-sample unpaired, two-sided

t = test statistics
df = degrees of freedom

confidence interval differences of the means

```
> t.test(weight.m,weight.f,var.equal=TRUE)

        Two Sample t-test
data:  weight.m and weight.f

t = 1.8265, df = 400, p-value = 0.06852

alternative hypothesis: true difference in
means is not equal to 0

95 percent confidence interval:
 -0.5669448 15.4259192

sample estimates:
mean of x mean of y
 181.9167  174.4872
```

Running a t-test in R

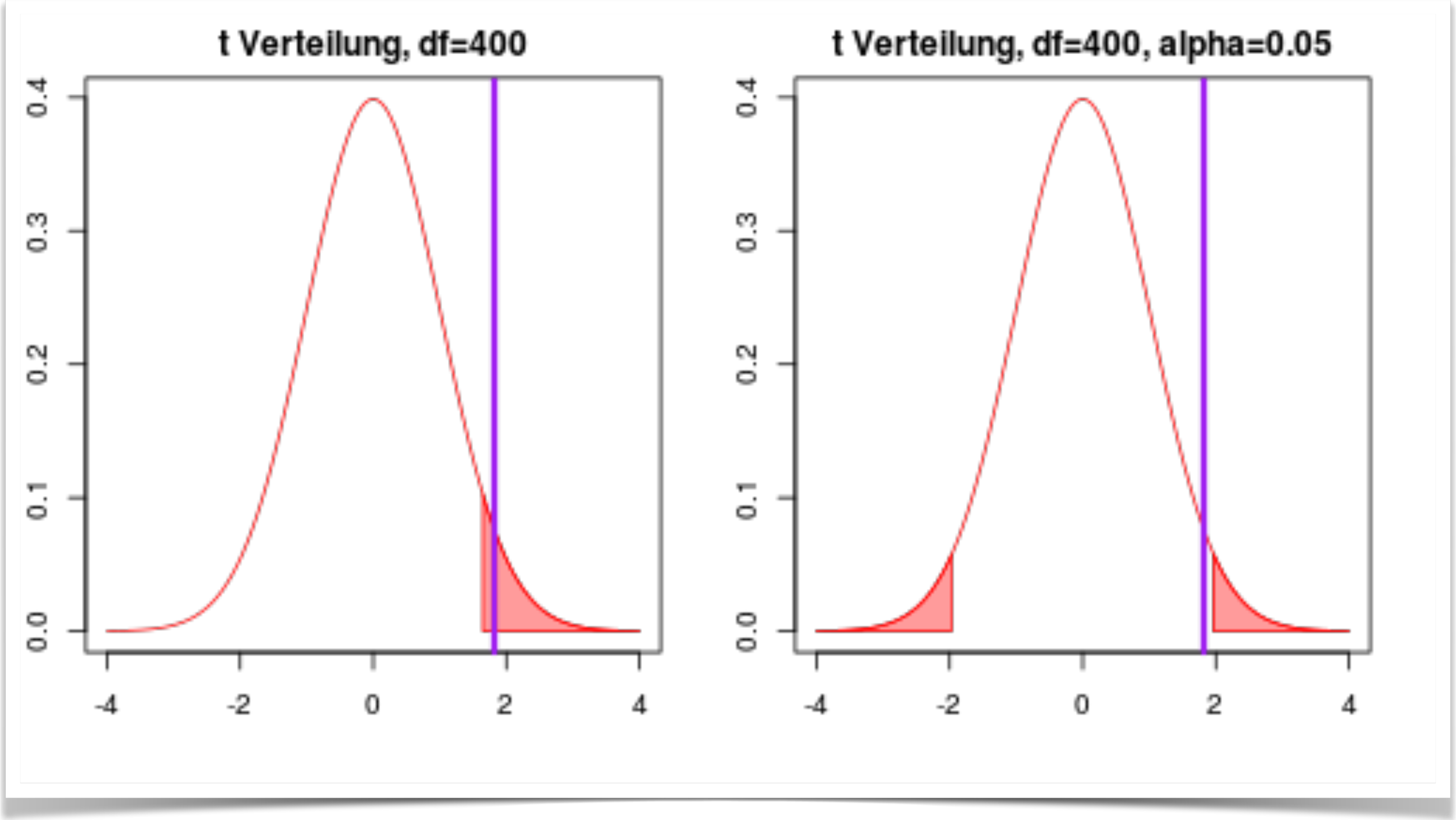
two-sample unpaired, one-sided

t = test statistics
df = degrees of freedom

confidence interval differences of the means

```
>t.test(weight.m,weight.f,alternative="greater",va
r.equal=TRUE)

        Two Sample t-test
data:  weight.m and weight.f

t = 1.8265, df = 400, p-value = 0.03426

alternative hypothesis: true difference in means
is greater than 0

95 percent confidence interval:
 0.723444       Inf
sample estimates:
mean of x mean of y
 181.9167  174.4872
```

Running a t-test in R



one-sided t-test
→ significant

two-sided t-test
→ non significant

Running a t-test in R

two-sample Welch unpaired, one-sided

t = test statistics
df = degrees of freedom

confidence interval differences of the means

```
>t.test(weight.m,weight.f,alternative=“greater")

        Welch Two Sample t-test
data:  weight.m and weight.f

t = 1.8453, df = 372.446, p-value = 0.0329

alternative hypothesis: true difference in means
is greater than 0

95 percent confidence interval:
 0.7903498       Inf

sample estimates:
mean of x mean of y
 181.9167  174.4872
```

Paired t-test

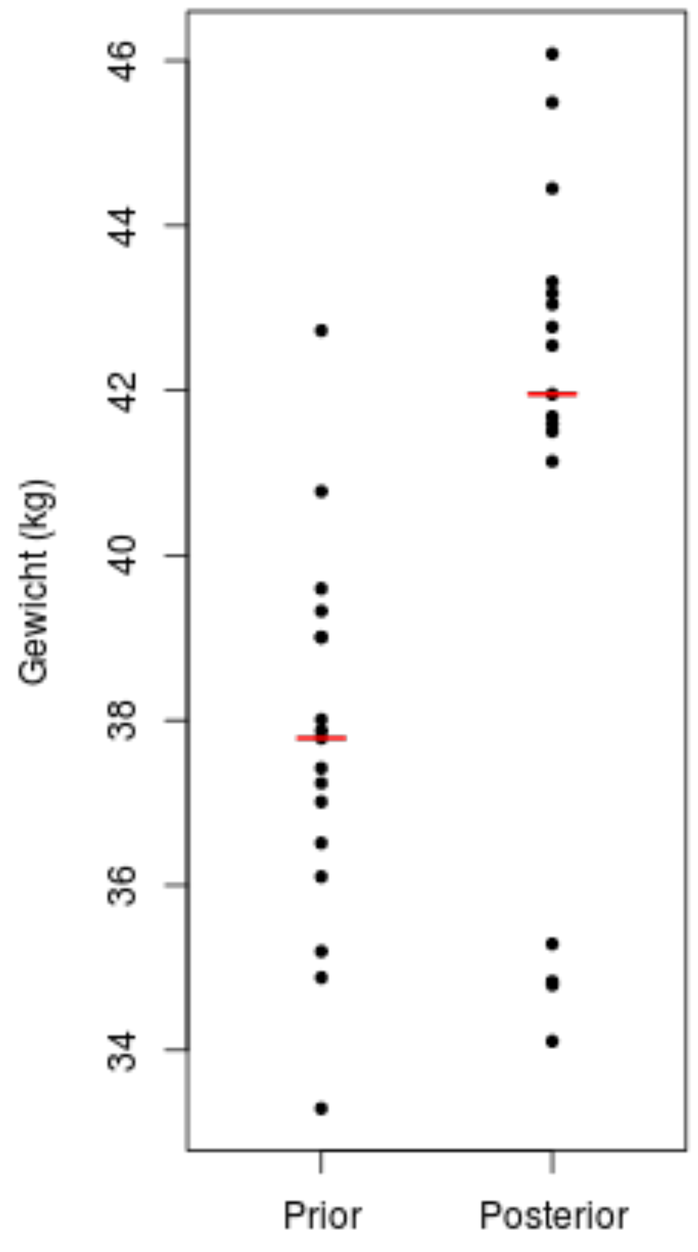
- 2 samples with equal number of elements
- each element of sample A can be associated to one element of sample B
  - patients before (A) and after (B) treatment
  - technical replicates

$$t = \frac{\bar{x_D} - \mu}{s_D/\sqrt{n}}$$

$\bar{x_D}$ = mean of differences

μ = expected difference

Treatment against anorexia
Weight before/after treatment



unpaired: $p = 5 \cdot 10^{-3}$

When can we apply t-test?

- There are several conditions that must be fulfilled to apply a t-test
- **Normality**: data must be (approximately) normaly distributed
  → check using
  - QQ-plot
  - statistical tests: Shapiro-Wilks / Kolmogorov-Smirnov
  - if not, apply non-parametrical test
- **Variance** of samples must be equal
  - if so: **Student** t-test
  - if not: **Welch** t-test
- **Independance**: independent samples: values in one sample should not be influenced by those in the second sample

7. Hypothesis testing

proportion tests

Proportion tests

- This class of tests can be used when searching for
  - **relation between different categorical variables**
    Is there a relation between social background and school grades?
  - comparison of **observed** vs. **expected** counts
    Is there a significant gender bias in the math department if 4 professors out of 10 are women?
- Two tests are generally used
  - **Fisher-Exact test** (FET): gives an exact p-value, used for small samples
  - **chi-square test**: for larger samples (n>5 in each category)
  - both tests are equivalent for large n

Fisher Exact Test

- Tests for a significant relationship between 2 variables
- Starting point: contingency table

	iPhone	other	Total
Men	**4**	**1**	5
Women	**2**	**3**	5
Total	6	4	10

Proportion iPhone/other:
- Men : 4/1 = 4
- Women: 2/3 = 0.66

Odds-Ratio:

OR = (4/1)/(2/3) = 6

If we would randomly distribute 6 iPhone and 4 other smartphones to 5 men and 5 women, how often would we get a larger/smaller*/more extreme** odds-ratio?

*smaller: < 1/6

**More extreme: > 6 or < 1/6

What is H0?

	iPhone	other	Total
Men	**3**	**2**	5
Women	**3**	**2**	5
Total	6	4	10

H_0: The proportion of men with iPhone is **equal** to the proportion of women with iPhones (2-sided) $OR = 1$

H_0:The proportion of men with iPhones is **not higher** that the proportion of women with iPhones (1-sided) $OR \leq 1$

H_0:The proportion of men with iPhones is **not lower** that the proportion of women with iPhones (1-sided) $OR \geq 1$

Random permutations

If I randomly distribute 6 iPhones and 4 other phones
to 5 women and 5 men, how likely it is to obtain this table?

	iPhone	other	Total
Men	**4**	**1**	5
Women	**2**	**3**	5
Total	6	4	10

MoBi students

	iPhone	other	Total
Men	**8**	**19**	27
Women	**16**	**16**	32
Total	24	35	59

```
Fisher's Exact Test for Count Data

data:  X
p-value = 0.1831
alternative hypothesis: true odds
ratio is not equal to 1
95 percent confidence interval:
 0.1230632 1.3943512
sample estimates:
odds ratio
 0.4273899
```

chi-square test

IPMB
Institut für Pharmazie und Molekulare Biotechnologie
UNIVERSITÄT HEIDELBERG ZUKUNFT SEIT 1386

- The chi-square test compares **observed** and **expected** counts
- Starting point is a **contingency table**
- Test statistics

$$\chi^2 = \sum_{i=1}^{n} \frac{(O_i - E_i)^2}{E_i}$$

  H_0: expected and observed proportions are equal

- H_0 distribution: chi2 distribution with *n-1* degrees of freedom for *n* observations
- Application possible when $O_i > 2$ and $O_i > 5$ in 80% of observations
- Note: the chi-square test is always a 1-sided upper tail test!

Observed

	iPhone	other	Total
Men	**14**	**30**	44
Women	**5**	**20**	25
Total	19	50	69

Observed proportions

	iPhone	other	Total
Men	**31.8%**	**68.2%**	100%
Women	**20%**	**80%**	100%
Total	27.5%	72.5%	100%

H0 proportions

	iPhone	other	Total
Men	**27.5%**	**72.5%**	100%
Women	**27.5%**	**72.5%**	100%
Total	27.5%	72.5%	100%

Expected counts under H0

	iPhone	other	Total
Men	**12.1**	**31.9**	44
Women	**6.9**	**18.1**	25
Total	19	50	69

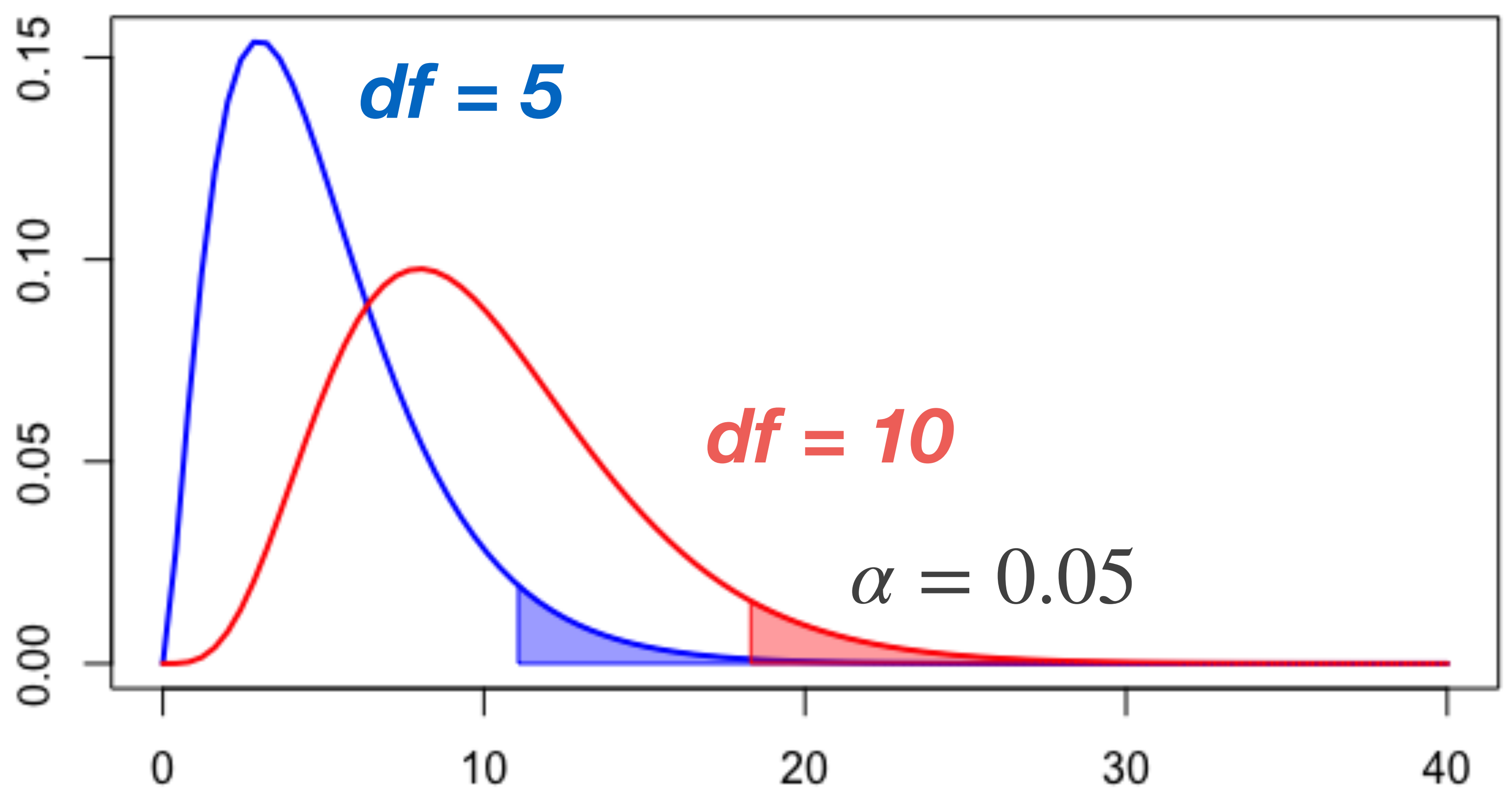
$$\chi^2 = \frac{(14-12.1)^2}{12.1} + \frac{(30-31.9)^2}{31.9} + \frac{(5-6.9)^2}{6.9} + \frac{(20-18.1)^2}{18.1}$$
$$= 0.6022$$

degrees of freedom = (rows-1) x (columns-1)

chi-square distribution

Critical values

	0.025	0.05	0.1
df = 1	5.02	3.84	2.71
df = 2	7.38	5.99	4.61
df = 3	9.35	7.81	6.25
df = 4	11.14	9.49	7.78
df = 5	12.83	11.07	9.24
df = 6	14.45	12.59	10.64
df = 7	16.01	14.07	12.02
df = 8	17.53	15.51	13.36
df = 9	19.02	16.92	14.68
df = 10	20.48	18.31	15.99



$\alpha = 0.05$

$\chi^2 = 0.6022$ **not significant...**

$df = 1$

More than 2 categories

Side effects

	weak	medium	strong	Total
Drug A	**25**	**11**	**13**	49
Drug B	**9**	**14**	**11**	34
Total	34	25	24	83

	weak	medium	strong	Total
Drug A	**51%**	**22.5%**	**26.5%**	100%
Drug B	**26.5%**	**41.2%**	**32.3%**	100%
Total	41%	30.1%	28.9%	100%

```
> table(sideeffect)
    SideEffect
Drug weak medium strong
   A     25     11     13
   B      9     14     11

> chisq.test(table(sideeffect))
        Pearson's Chi-squared test
data:  table(sideeffect)
X-squared = 5.5257, df = 2, p-value = 0.06311


> fisher.test(table(sideeffect))
        Fisher's Exact Test for Count Data
data:  table(sideeffect)
p-value = 0.06375
alternative hypothesis: two.sided
```

8. Power of a test

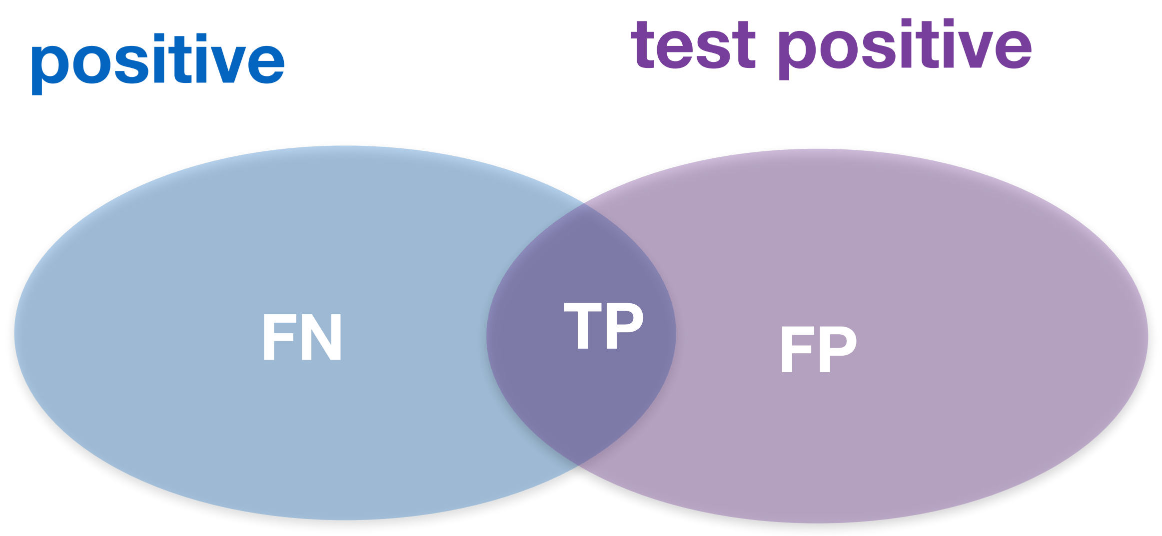
Reliability of statistical test

- A **reliable test** should have a small number of false-positives and false-negatives
- Increasing significance level leads to ???? false-positives and ??? false-negatives

	H_0 is valid	H_0 is NOT valid	
H_0 rejected ($p < \alpha$)	False-positive (type 1 error)	True positive	test positive
H_0 not rejected ($p > \alpha$)	True negative	False-negative (type 2 error)	test negative
	negative	positive	

Reliability of statistical test

	H_0 is valid	H_0 is NOT valid	
H_0 rejected ($p < \alpha$)	FP	TP	test positive
H_0 not rejected ($p > \alpha$)	TN	FN	test negative
	negative	positive	



$$\text{false-negative rate (FNR)} = \frac{FN}{\text{positives}} = \frac{FN}{FN + TP}$$

$$\text{false-positive rate (FPR)} = \frac{FP}{\text{negatives}} = \frac{FP}{FP + TN}$$

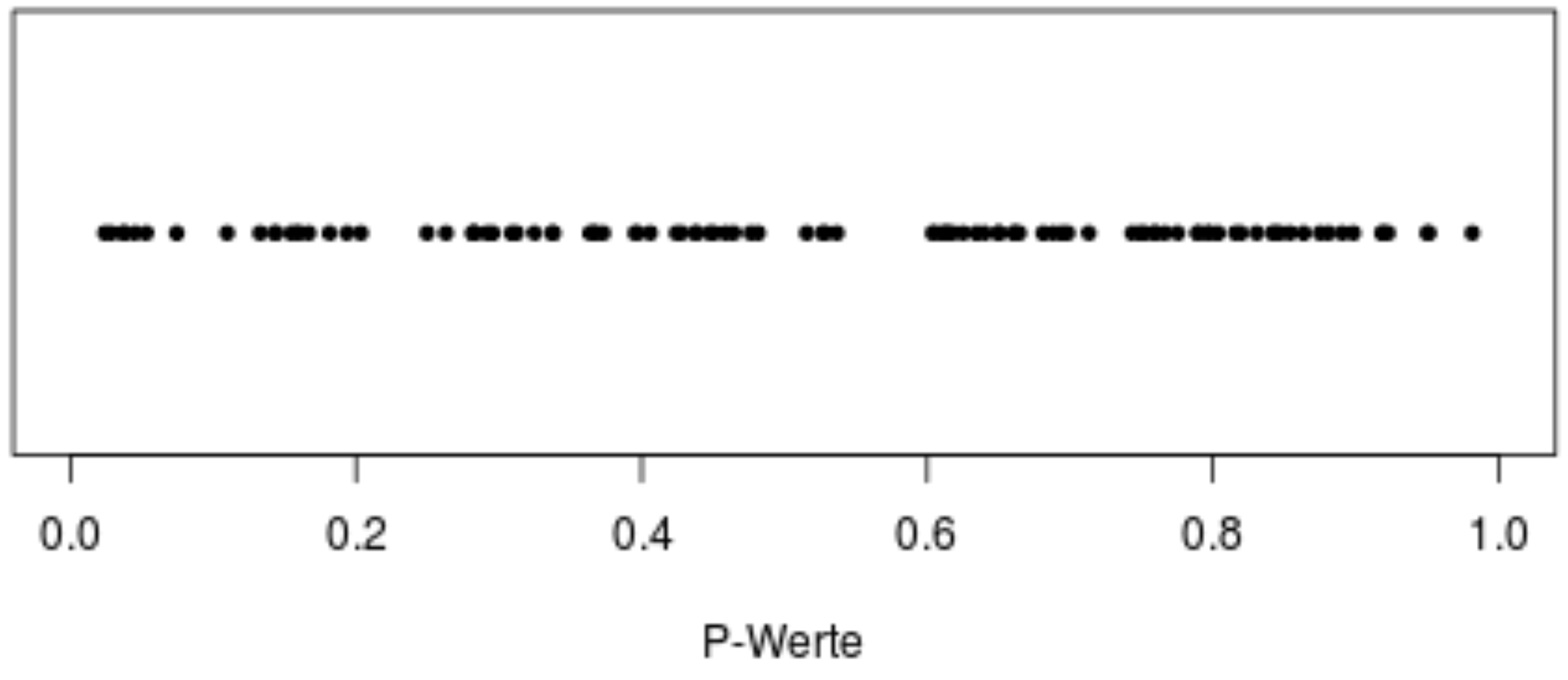
$$\text{false-discovery rate (FDR)} = \frac{FP}{\text{test positives}} = \frac{FP}{FP + TP}$$

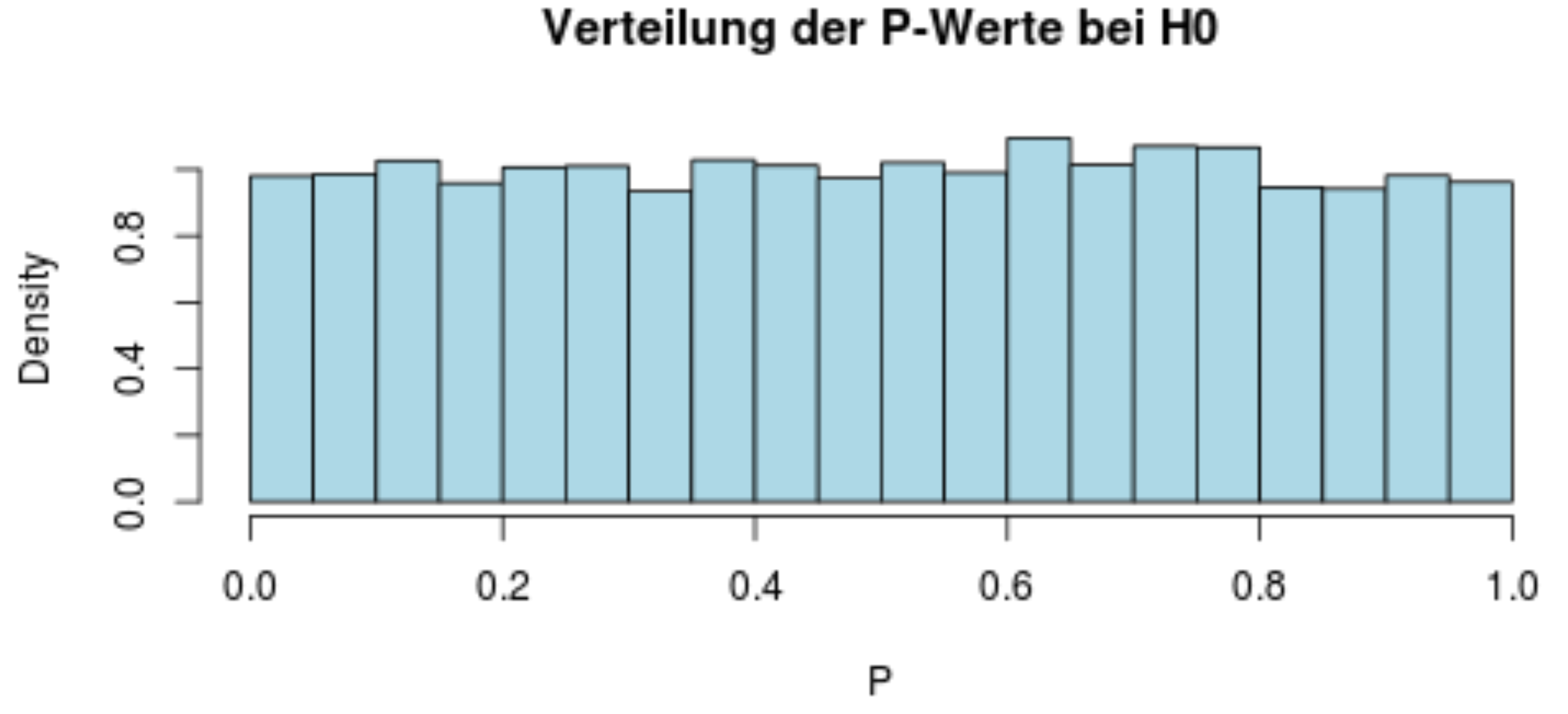
$$\text{precision} = \frac{TP}{\text{test positives}} = \frac{TP}{FP + TP}$$

$$\text{recall} = \frac{TP}{\text{positives}} = \frac{TP}{FN + TP}$$

P-value distribution under H_0

- ***What are typical p-values under H_0?***
- **Experiment**: draw 2 sets (S_1 & S_2) of 50 random numbers each **from the same distribution**
- ***H_0: the expectation of both distributions are equal (TRUE!)***
- Compute t-test between S_1 and S_2, and determine P-value
- Repeat this experiment 1000 times, and plot the distribution of the 1000 p-values





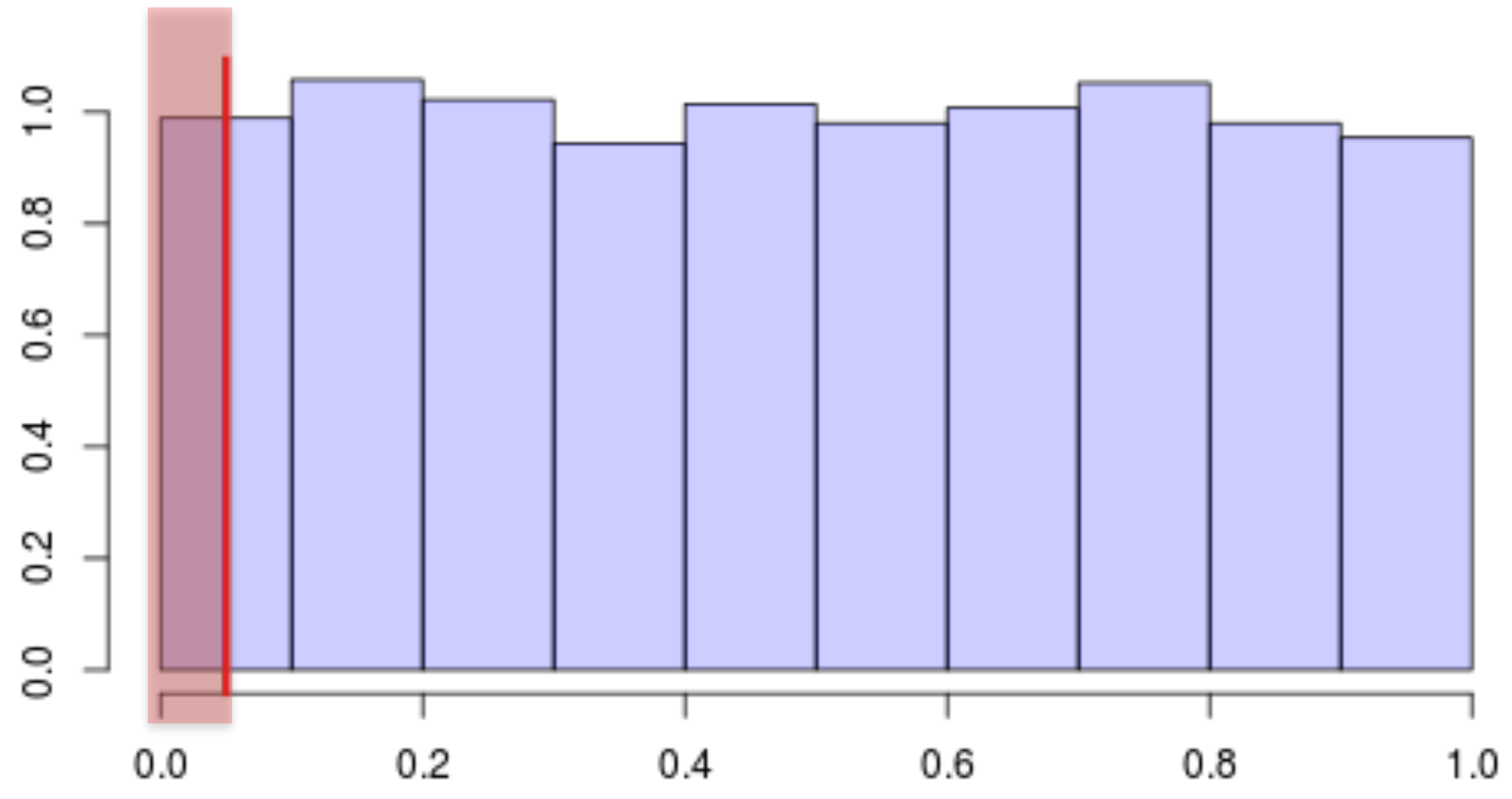
Distribution of p-values under H_0 = uniform distribution

Type 1 errors

- **Red area:**
  with α = 5%, we would have wrongly rejected H0
  → ***FALSE POSITIVE***

- ***How often would that occur?***
  → red area compared to the total area = 5% because uniform distribution

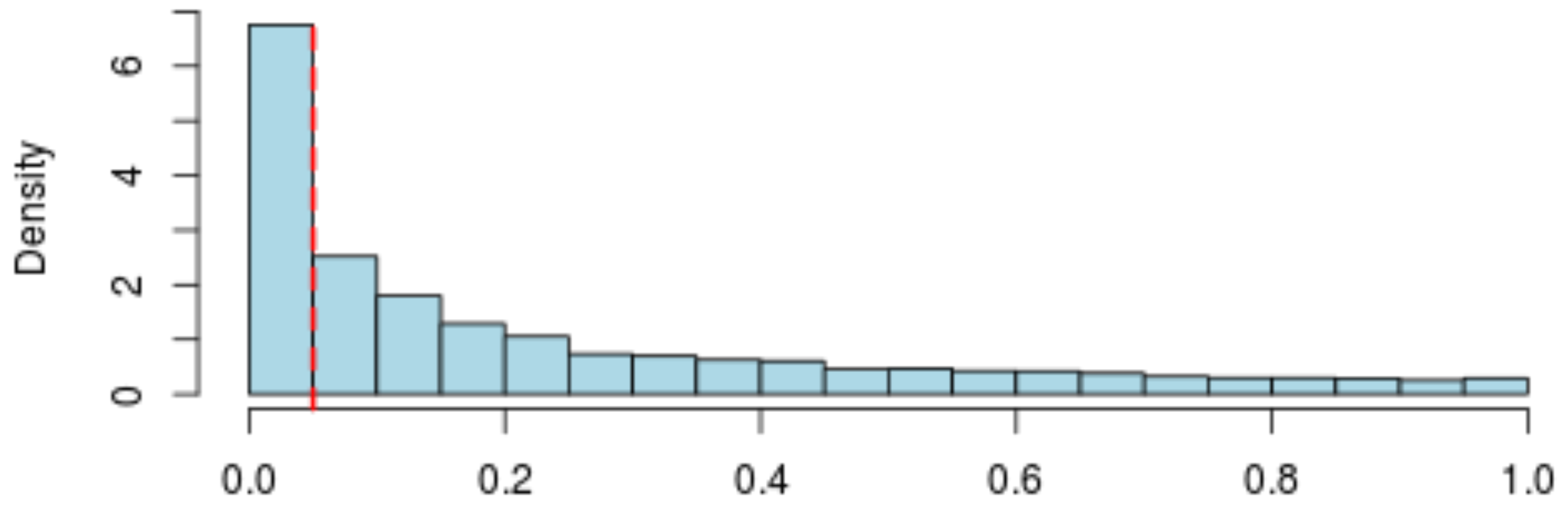
Distribution of p-values under H_0



α is the FALSE-POSITIVE RATE (FPR)

P-value distribution under H_1

- Experiment: draw 2 sets (S_1 & S_2) of 50 random numbers each **from two distributions with different expectation**
- ***H_0: the expectation of both distributions are equal (FALSE!)***
- compute p-value using a 2 sample t-test
- Repeat 1000 times and plot distribution of p-values



Many small p-values
→ H_0 would have been rejected

🙂

Some large p-values
→ H_0 would have NOT been rejected

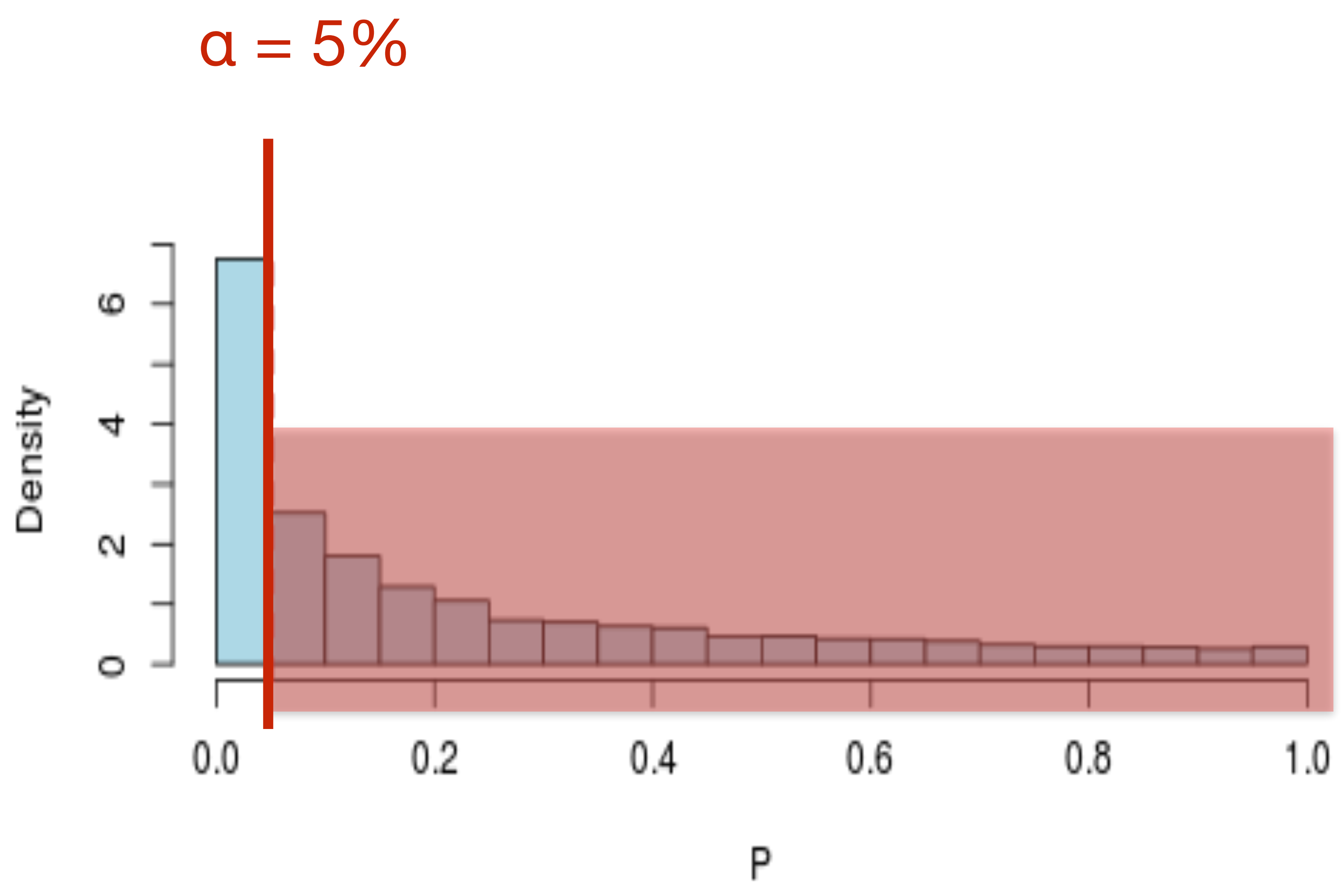
☹️

Type 2 errors

- Occur when a false H_0 hypothesis is **NOT rejected** by the test
  → False-negative (Type 2 errors)

- Probability of a type 2 error:
  β - value

- Probability for a type 2 error NOT to occur
  → **power of a test = 1- β**

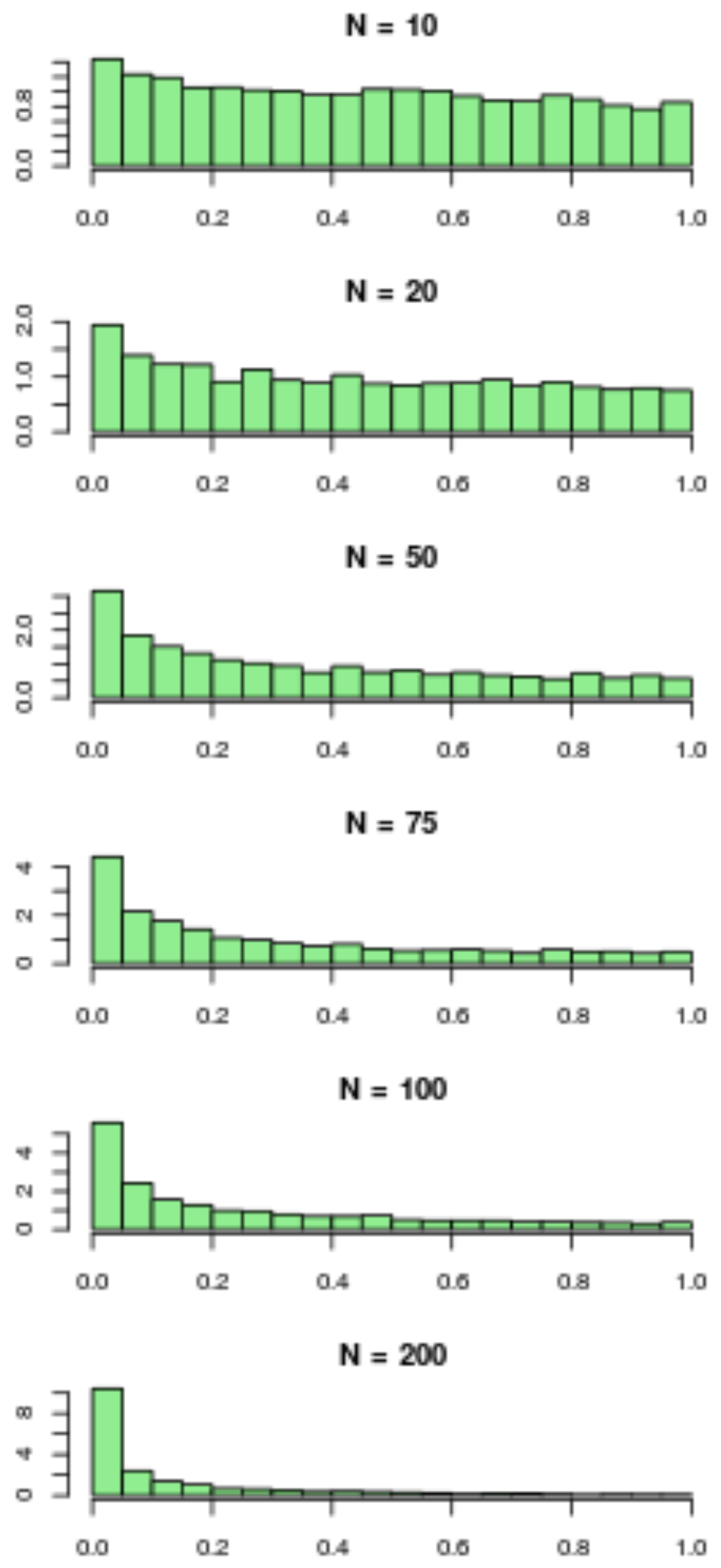


This area represents the cases for which H0 will not be rejected
→ ***false-negatives***

Power of a test

- Generate 2 datasets of length *n*
  - one from a normal distribution with mean 0
  - one from a normal distribution with mean 0.2
- **H_0: expectation of both underlying distributions is identical** (False!)
- perform t-test, compute p-values for various values of *n*

$$\beta \xrightarrow{n\to\infty} 0$$



Power of a test

- The power depends on:
  - **Significance level α**
  - **Sample size n**
  - **Effect-size**: how strong is the observed effect?