

# Biostatistics

## Week 5

- **Binomial test for proportion**
- **Terms in hypothesis tests**
  - **Significance level, error type I and II, rejection zone, power**
- **sample size calculation / power analysis**
- **multiple testing**
  - **Bonferroni correction (for  $\ll 100$  tests)**
  - **False discovery rate, p-value histogram (for  $>100$  tests)**

## Example: Does new treatment lower the relaps risk?

- **Question:** Does a new treatment lower the proportion of 35% relapses that is typical for the state-of-the-art treatment.  
**Data:** we observe 26 relapses among 100 treated patients
- $H_0$ : „nothing changed“, proportion of relapses is  $\pi_0 = 0.35$
- How is the number of relapses among 100 treated patients distributed und  $H_0$ ?

$X$  : Number of relapses in 100 treated patients

$X \sim \text{Bin}(\pi=0.35, n=100)$

# How to assess if the observation is consistent with $H_0$ ?

- Probability to observe  $x = 26$  relapses under 100 treated patients under  $H_0$ :
- $n=100, x=26, \pi=0.35$

$$\binom{n}{x} \cdot \pi^x \cdot (1 - \pi)^{n-x}$$

In R:

```
> dbinom(26, 100, 0.35)
[1] 0.01400205
```

- What about, if we would have observed  $x=25$  relapses under 100 patients? That would indicate even stronger that  $H_0$  might be wrong.
- **Question (in 2 phrasings)**
  - What is the rejection zone for  $H_0$ ?
  - How large is the probability to observe 26 or less relapses among 100 patients, if  $H_0$  would be correct?

# Rejection zone

Look for the maximal range from 0 to k relapses among 100 patients, until the probability of 5% (significance level  $\alpha$ ) is taken.

```
> pbinom(25:27,100,p=0.35)
[1] 0.021 0.035 0.055
```

Probabilities:

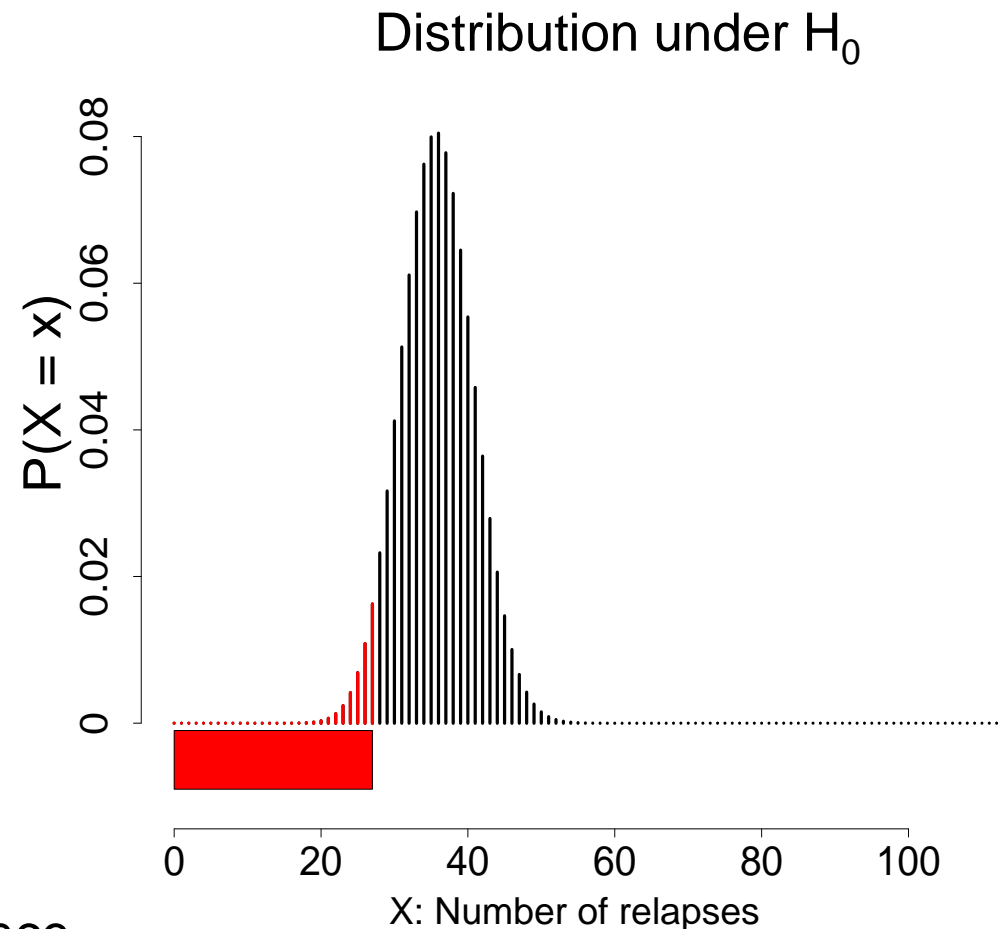
$0 < X \leq 25$  0.021 # below 5%

$0 < X \leq 26$  0.035 # below 5%

$0 < X \leq 27$  0.055 # too large

If the observed number  $x$  is between 0 and 26, then we have enough evidence to reject  $H_0$ .

→ Accept  $H_A$



Note: in case of a Binomial distributed variable  $X$ , we know the distribution of  $X$  under  $H_0$ , proposing a certain  $\pi_0$ .

Opposed to that, if  $X$  is Normal distributed and  $H_0$  proposed a certain  $\mu_0$ , we don't know how  $X$  is distributed under  $H_0$  because of the unknown  $\sigma$  which needs to be estimated; we just know that the test statistic  $T = \frac{\bar{x} - \mu_0}{s/\sqrt{n}}$  is  $t_{df=n-1}$  distributed which we can use to construct from the CI for  $T$  a CI for  $\mu$ .

# Binomial test in R

# Perform a Binomial test R:

```
> binom.test(x=26,n=100,p=0.35,  
             alternative="less")
```

one-sided test  
(not recommended in clinical trials)

Exact binomial test

```
data: 26 and 100  
number of successes = 26, number of  
trials = 100, p-value = 0.03514  
alternative hypothesis: true probability  
of success is less than 0.35  
95 percent confidence interval:  
 0.0000000 0.3419918  
sample estimates:  
probability of success  
          0.26
```

# compute the p-value yourself:

```
> pbinom(26, 100, 0.35)
```

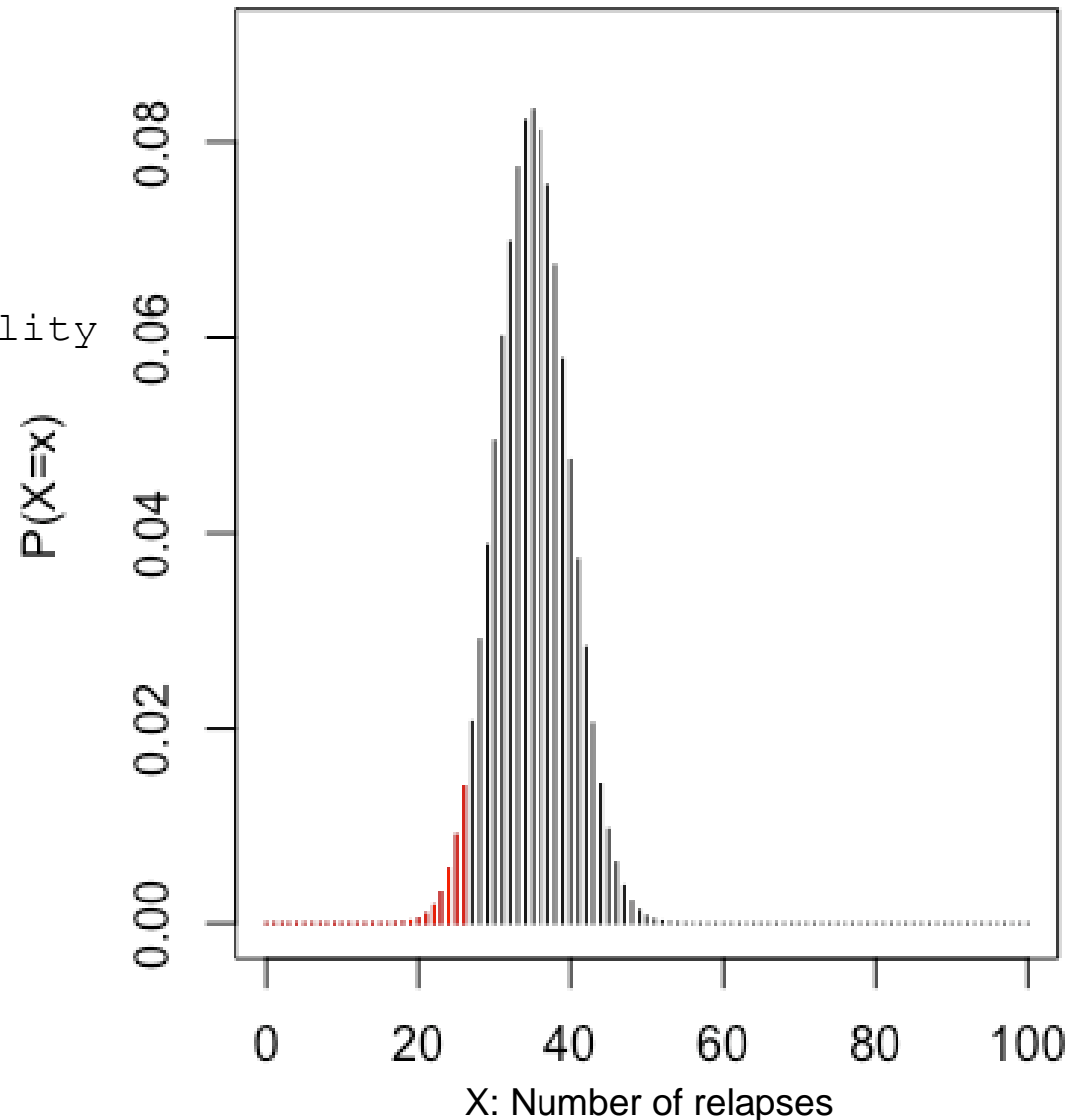
```
> 0.0351 # „p-Wert“
```

# or calculate p-value via

```
> sum(dbinom(0:26, 100, 0.35))
```

```
> 0.0351
```

Distribution of X under  $H_0$



# Recall: Construct CI for $\mu$ from CI for T

## In case of a 2-sided t-test

$X_i \text{ i.i.d. } \sim N(\mu, \sigma^2), E(X) = \mu_x, \text{Var}(X) = \sigma_x^2$

$$\Rightarrow T = \frac{\bar{X} - \mu_x}{s_x / \sqrt{n}} \sim t_{df=n-1}$$

$$P(q'_{\frac{\alpha}{2}} \leq \frac{\bar{X} - \mu_x}{\sigma_x / \sqrt{n}} \leq q'_{1-\frac{\alpha}{2}}) = 1 - \alpha$$

$$P(-\frac{\sigma_x}{\sqrt{n}} \cdot q'_{1-\frac{\alpha}{2}} \leq \bar{X} - \mu_x \leq \frac{\sigma_x}{\sqrt{n}} \cdot q'_{1-\frac{\alpha}{2}}) = 1 - \alpha$$

$$P(\bar{X} - \frac{\sigma_x}{\sqrt{n}} \cdot q'_{1-\frac{\alpha}{2}} \leq \mu_x \leq \bar{X} + \frac{\sigma_x}{\sqrt{n}} \cdot q'_{1-\frac{\alpha}{2}}) = 1 - \alpha$$

```
> sample = rnorm(100, mean=-0.3, sd=1.2)
> t.test(sample, alternative = "two.sided",
  mu=0, conf.level = 0.95)
```

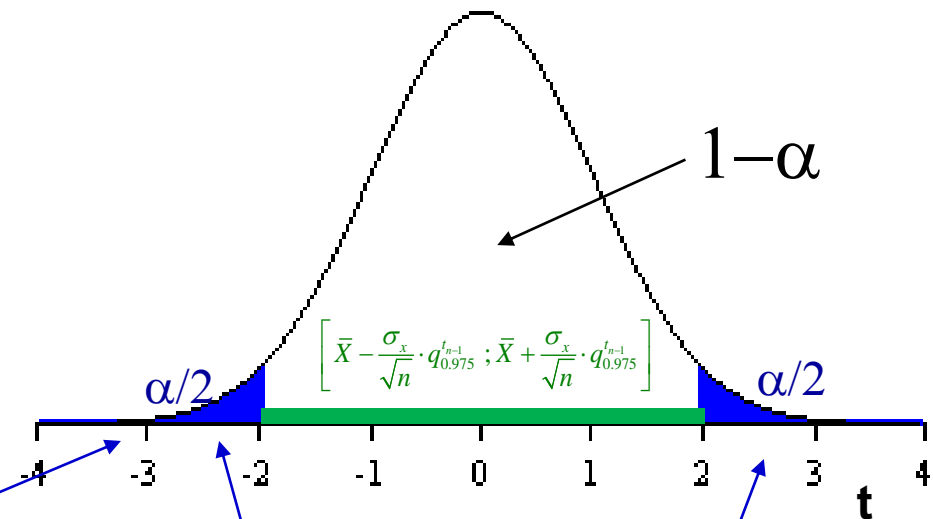
One Sample t-test

```
data: sample
t = -3.1524, df = 99, p-value = 0.002144
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
 -0.6206213 -0.1411401
sample estimates:
 mean of x
-0.3808807
```

$0 \notin \text{CI} \Rightarrow \text{reject } H_0$

CI for  $\mu$

Distribution of T (not X!) under  $H_0$



CI for T: if t falls into this region, don't reject  $H_0$

Note: in the test output we get the CI for  $\mu$ , not for T

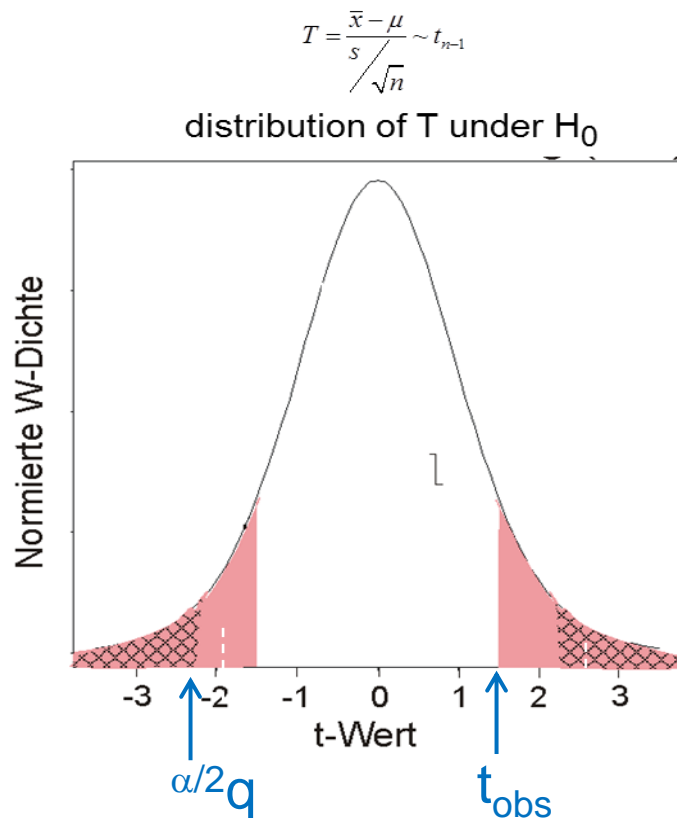
Rejection zone: if t falls into this region, then reject  $H_0$

Note: this corresponds to the case where the postulated parameter  $\mu_0$  is not covered by the CI for  $\mu$ ,

# Recall: Interpretation of the p-value

The p-value corresponds to the probability to get an at least such extreme result as the seen one if we assume that the Null-Hypothesis is valid.  
(Therefore we reject  $H_0$  if this probability is small)

Graphically: the p-value corresponds to the area in the extreme tails (from the observed t-value outwards) under the density of the test-statistic distribution which is taken for a true  $H_0$ .



$$p = P(|t| \geq |t_c| \mid H_0 \text{ is true})$$
$$= P(p_{new} \leq p \mid H_0 \text{ is true})$$

p-value > 0.1 : no evidence for  $H_A$

p-value < 0.1 : weak evidence for  $H_A$

p-value < 0.05 : evidence for  $H_A$

p-value < 0.01 : clear evidence for  $H_A$

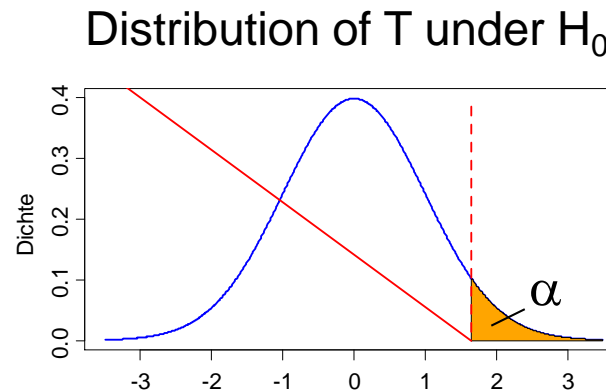
p-value < 0.001 : strong evidence for  $H_A$

# Right-, left- and two-sided Test

z.B.:  $T = \frac{\bar{x} - \mu_0}{s/\sqrt{n}} \stackrel{a}{\sim} N(0,1)$

Right-sided Test:

$H_A: \mu > \mu_0$

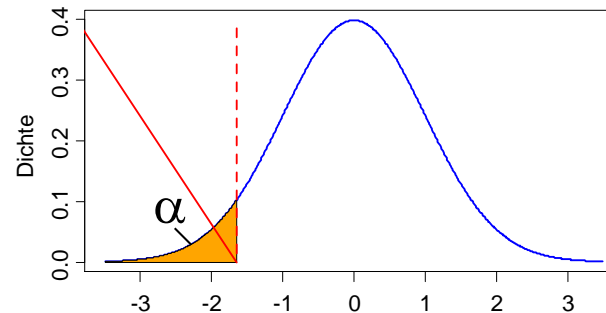


$(1-\alpha)$ -CI for  $\mu$

$\left[ -\infty , q_{1-\alpha} \right]$

Left-sided Test:

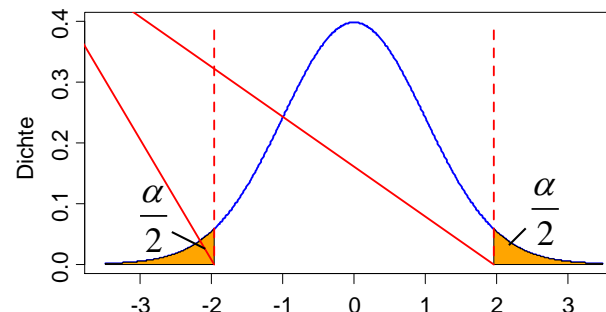
$H_A: \mu < \mu_0$



$\left[ q_{\alpha} , \infty \right]$

Two-sided Test:

$H_A: \mu \neq \mu_0$



$\left[ q_{\alpha/2} , q_{(1-\alpha/2)} \right]$



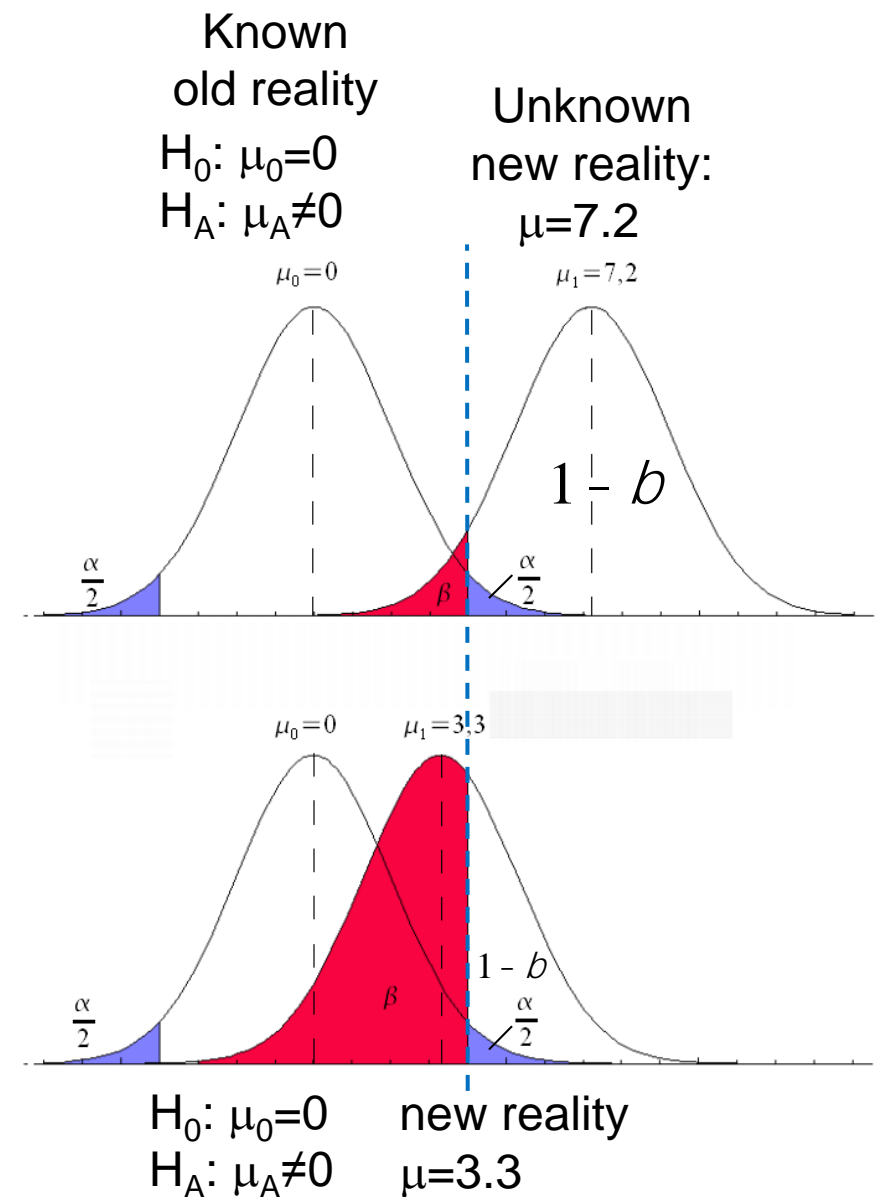
# Decision errors revisited

	negative test accepting $H_0$	positive test rejecting $H_0$
$H_0$ is true	<b>True Negative</b> (the probability for this correct test decision is $(1-\alpha)$ )	<b>False Positive</b> (the probability for a type-I error is $\alpha$ )
$H_0$ is false	<b>False Negative</b> (the probability for a type-II error is $\beta$ )	<b>True Positive</b> (the probability for this correct test decision is $(1-\beta)$ )

$P(\text{reject } H_0 \mid H_0 \text{ true}) = \alpha$  probability for type I error

$P(\text{accept } H_0 \mid H_0 \text{ false}) = \beta$  probability for type II error

power =  $1-\beta$



Effect size

=

difference between  $H_0$  and unknown new reality

# Power Analysis & Sample Size Calculation

# What is the power of a test

The power ( $1-\beta$ ), of a test is the probability to reject  $H_0$ , if  $H_A$  is true.

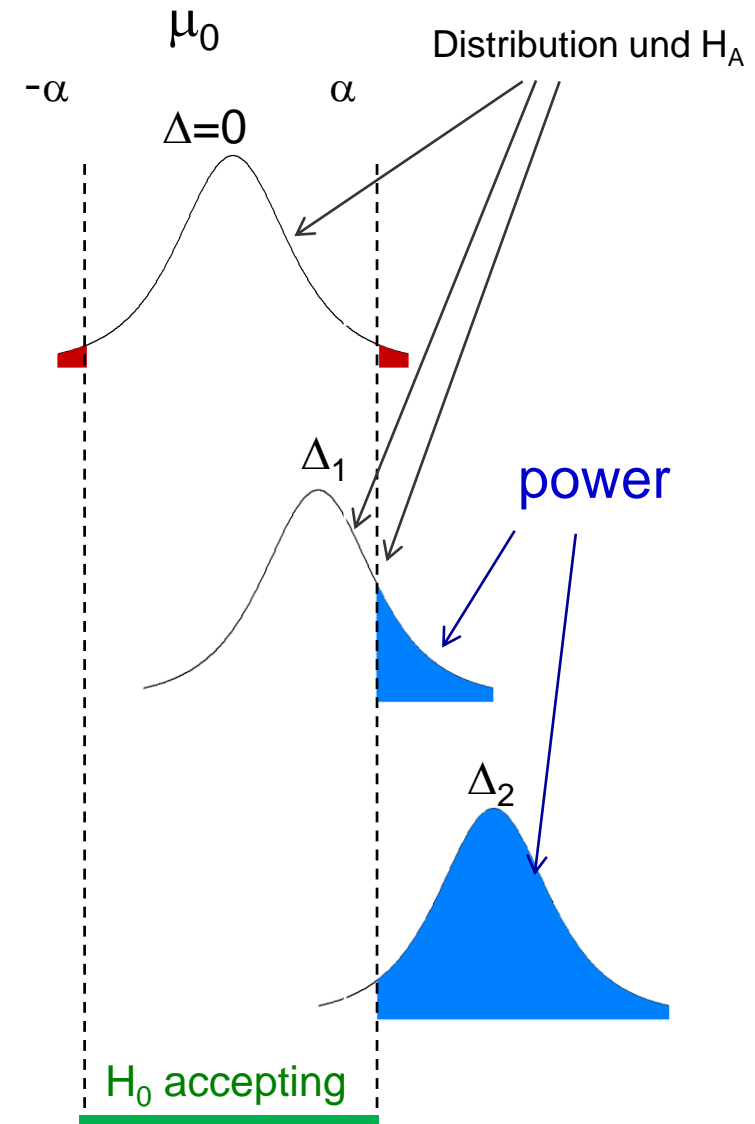
The power is given by the blue area.

The larger the difference  $D$  between  $H_0$  and reality is, the larger gets the power.

However, “reality” is not known -> it is hard to estimate the power.

For a given difference  $D$  between  $H_0$  and reality the power gets bigger if the width of the **distribution of the Test-Statistic, e.g. mean, gets smaller** which can be achieved by increasing the sample size.

Since the reality can not be changed, **in praxis the only way to increase the power is to increase the sample size.**



# Sample size calculation

## Situation:

Your group has developed a new drug for sleeping time elongation.



The new drug is only interesting if its sleeping time elongation surpasses the one from the «golden standard» by at least 1h (relevant effect).

From a pilot study we know the standard deviation of the sleeping time in individual patients is:  $sd=1$  (1h).

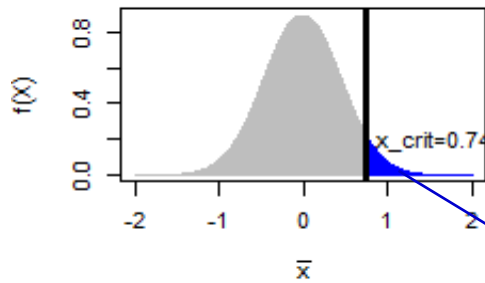
Given your drug surpasses the old drug by a mean sleeping time extension of 1h - how big should the sample size be chosen, so that you have a **power of 80%** and simultaneously an  $\alpha=5\%$  that your test rejects the  $H_0$  and proves the superiority of the new drug?

# Simulation with $n=5, 10, 20$

X: Sleep elongation compared the golden standard drug

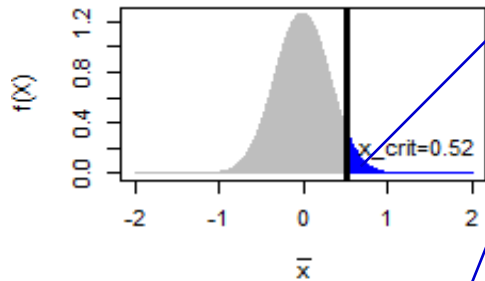
$$H_0: E(X)_0 = \mu_{\Delta 0} = 0$$

Distribution of  $\bar{X}$  under  $H_0$   
 $n = 5$



With a sample of size 5 we would reject  $H_0$ , if  
 $\bar{X} > 0.74$

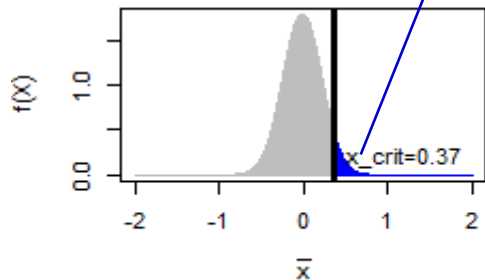
$n = 10$



$\alpha = 5\%$  is fixed

With a sample of size 10 we would reject  $H_0$ ,  
if  
 $\bar{X} > 0.52$

$n = 20$



With a sample of size 20 we would reject  $H_0$ ,  
if  
 $\bar{X} > 0.37$

# Simulation with $n=5, 10, 20$

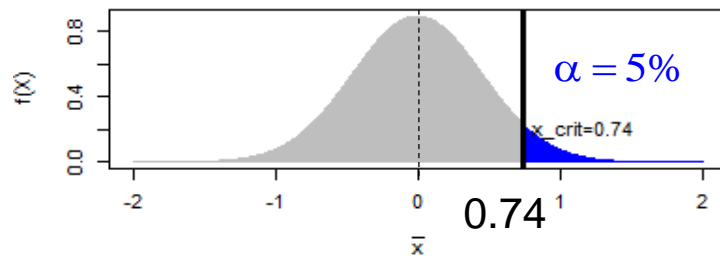
$X$ : Sleep elongation compared the golden standard drug

$$H_0: E(X)_0 = \mu_{\Delta 0} = 0$$

$$H_A: \mu_{\Delta} = 1$$

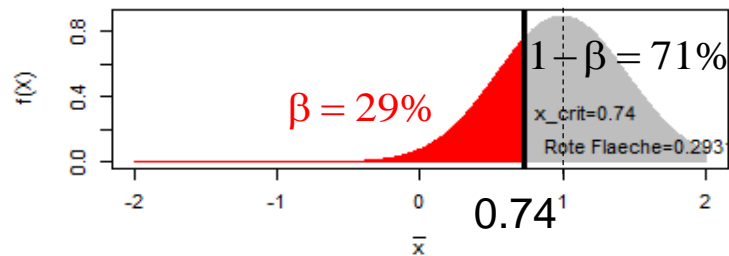
Distribution of  $\bar{X}$  under  $H_0$

$n = 5$



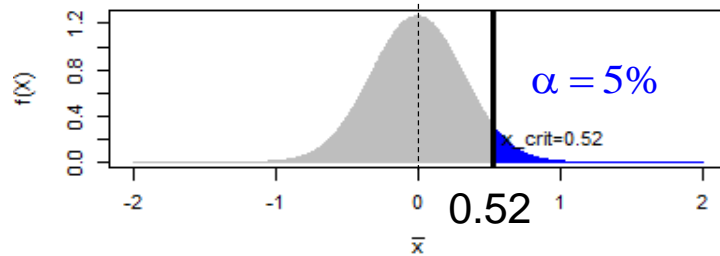
Distribution of  $\bar{X}$  given  $\mu_{\Delta}=1$

$n = 5$

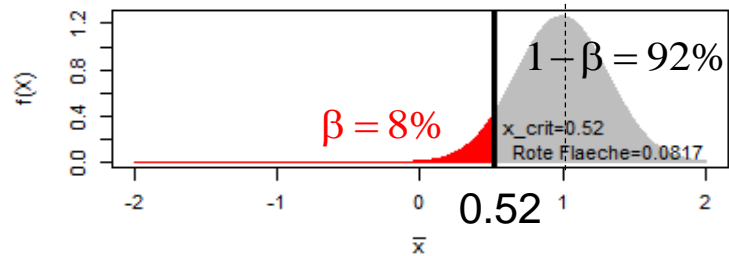


$n=5$ : Reject  $H_0$  in 71%  
of all simulation runs  
**Power=1- $\beta$ =71%**

$n = 10$

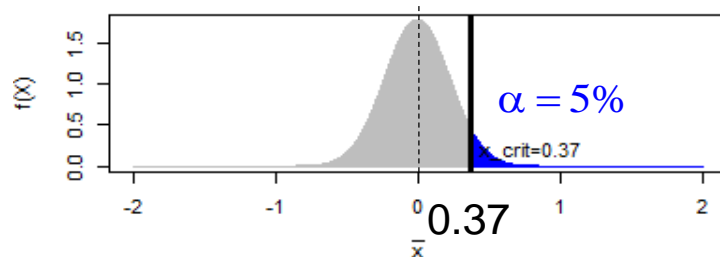


$n = 10$

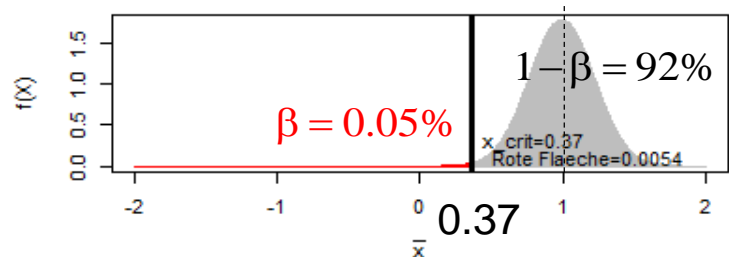


$n=10$ : Reject  $H_0$  in 92%  
of all simulation runs  
**Power=1- $\beta$ =92%**

$n = 20$

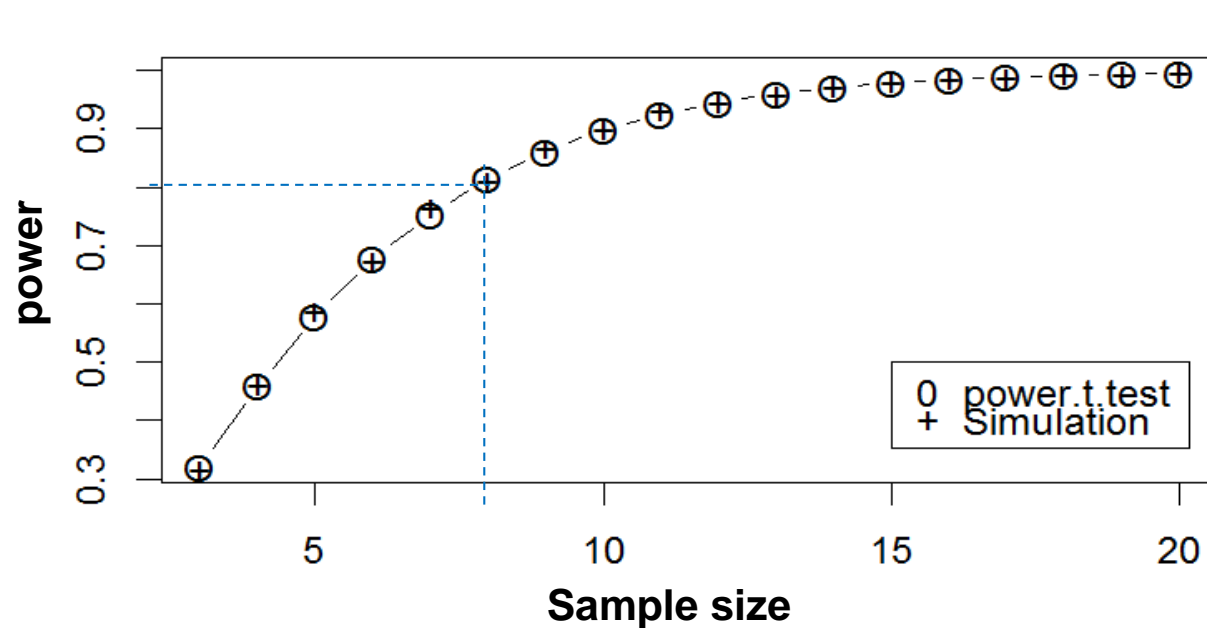


$n = 20$



$n=20$ : Reject  $H_0$  in 99.5%  
of all simulation runs  
**Power=1- $\beta$ =99.5%**

# Results



n	power.simu
3	0.3171
4	0.4631
5	0.5880
6	0.6737
7	0.7652
8	0.8118
9	0.8685
10	0.9001
11	0.9308
12	0.9452
13	0.9579
14	0.9712
15	0.9803
16	0.9862
17	0.9893
18	0.9926
19	0.9939
20	0.9960

```
> power.t.test(power=0.8, delta=1, sd=1, sig.level=0.05,  
               alternative="one.sided", type="one.sample")
```

One-sample t test power calculation

```
      n = 7.727622  
delta = 1  
  sd = 1  
sig.level = 0.05  
  power = 0.8  
alternative = one.sided
```

NOTE: n is number in *each* group

# How to plan the size of a study?

- Choose the test you want to use in your analysis
- Determine/Estimate the variation of the observations
- fix significance level  $\alpha$  (the accepted risk for a type-1-error, typically 5%)
- Fix relevant effect size (the minimal effect which is still relevant)
- Fix the power which gives the probability to detect an relevant effect (typically 80%)
  - Choose  $1-\beta$

Perform a sample-size calculation to derive the needed sample size at which the required power is given.

Good webpage for sample size calculations – menu based, but shows corresponding R-code:

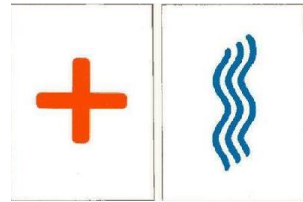
<http://powerandsamplesize.com/Calculators/>



# Multiple testing: Rhine Paradox

- The parapsychologist **Joseph Rhine** hypothesized in the 1950's that some people had *Extra-Sensory Perception (ESP)*.
- He tested for ESP by an experiment where people were asked to guess the color of 10 hidden cards:

red or blue.



- He discovered that almost 1 in 1000 had ESP – they were able to get all 10 right. **Surprised???**

$$P(10 \text{ correct answers} \mid \text{just guessing}) = \frac{1}{2} \cdot \frac{1}{2} \cdot \frac{1}{2} \cdot \frac{1}{2} \cdot \frac{1}{2} \cdot \frac{1}{2} \cdot \frac{1}{2} \cdot \frac{1}{2} \cdot \frac{1}{2} \cdot \frac{1}{2} = \left(\frac{1}{2}\right)^{10} = \frac{1}{2^{10}} = \frac{1}{1024}$$

No, 1 in 1024 is what we would expect to get by chance if everybody is just guessing

# Multiple testing: Rhine Paradox

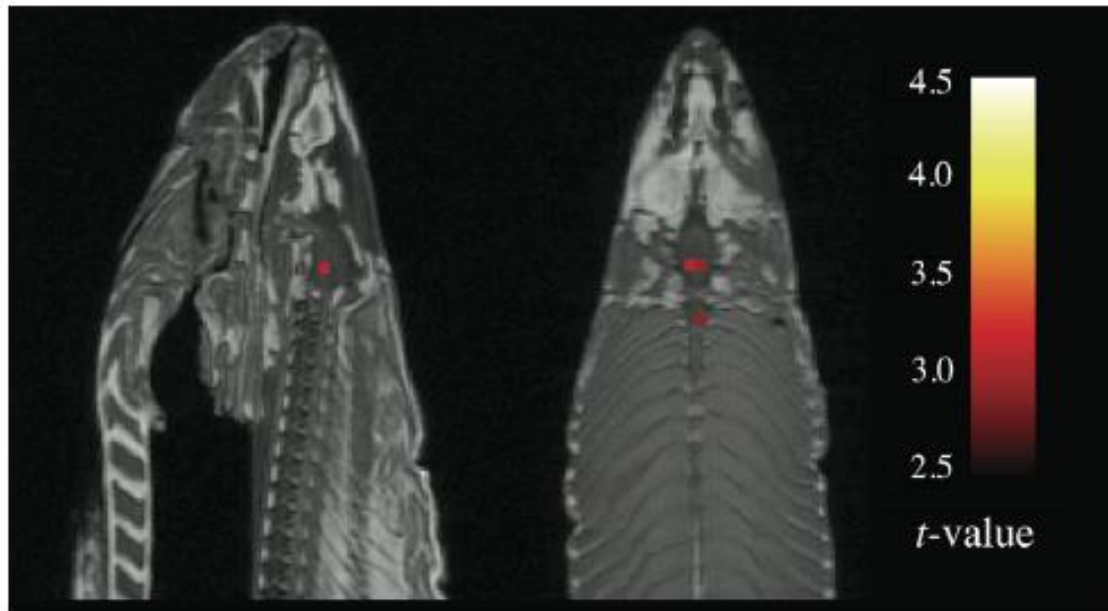
- He told these people they had ESP and called them in for another test of the same type.
- He discovered that all of them had lost their ESP.
- What did he conclude???

## Multiple testing: Rhine Paradox

He concluded that you shouldn't tell people that they have ESP, because it causes them to loose it.



# fMRI revealed brain response to trans-species emotional stimuli in a dead salmon



A **dead salmon** was repeatedly confronted with 2 different human emotional stimuli.

Out of 8064 brain voxels in 16 voxels a significant different activity ( $p \leq 0.001!$ ) was observed

This study received 2012 the IG nobel price (for *ignoble*, improbable research).

# The probability to get by chance a significant test result

The risk to get in **one test** an false positive result (that is  $p < \alpha$  under  $H_0$ ) is only

$$P(\text{reject } H_0 \mid H_0 \text{ true}) = \alpha$$

$$P(\text{accept } H_0 \mid H_0 \text{ true}) = 1 - \alpha$$

Assume **n independent test's** (with n independent samples) where the null-hypothesis  $H_0$  is always valid (no effect nowhere)

– the probability draw always the right test decision is:

$$P(\text{accept } n\text{-times } H_0 \mid H_0 \text{ true}) = (1 - \alpha)^n$$

– the probability of coming up with at least 1 false positive effect is:

$$P(\geq 1 \text{ rejecting } H_0 \mid H_0 \text{ true}) = 1 - (1 - \alpha)^n$$

the probability of making at least 1 type one error at **n trials**, when  $\alpha = 0.01\%$

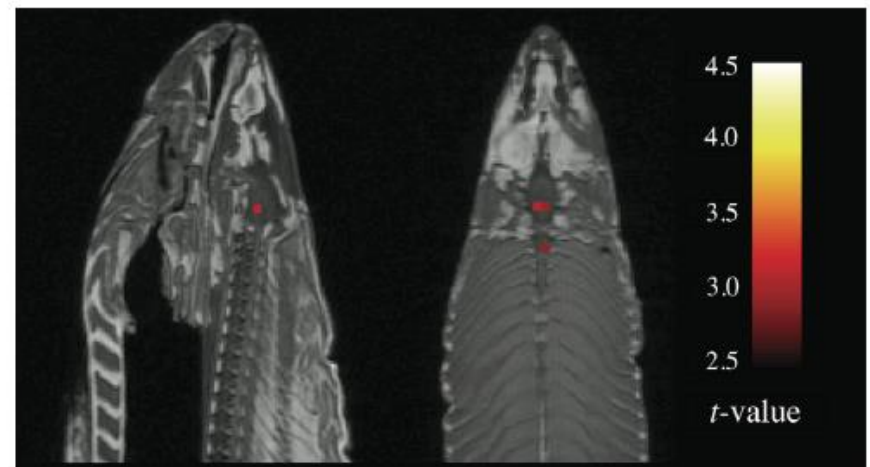
n	50	100	200	400	800
P(>1FP)	39%	63%	87%	98%	100%

# Bonferroni correction for multiple testing and its effect on the dead salmons reaction

Bonferroni: when performing  $n$  independent tests, conduct each test at significance level  $\frac{\alpha}{n}$  !

When applying Bonferroni's rule, we only have a risk of  $\alpha$ , to come up with  $\geq 1$  false positive effects (that is a significant test result although  $H_0$  is true).

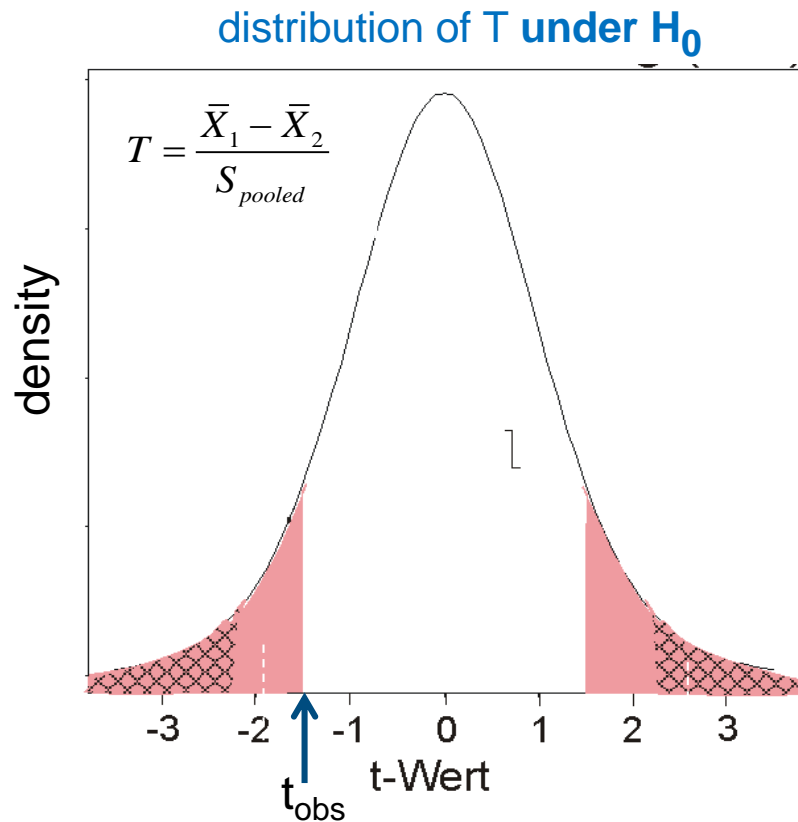
No brain region of the dead salmon showed a significant reaction to smiling people after Bonferroni correction.



# The p-value is uniformly distributed under $H_0$

The p-value corresponds to the probability to get an at least such extreme result as the observed result assuming that the Null-Hypothesis is valid

-> the p-value corresponds to the area in the extreme tails



$$p = Prob(|t| > |t_{obs}| \mid H_0 \text{ true})$$
$$= Prob(|p| \leq |p_{obs}| \mid H_0 \text{ true})$$

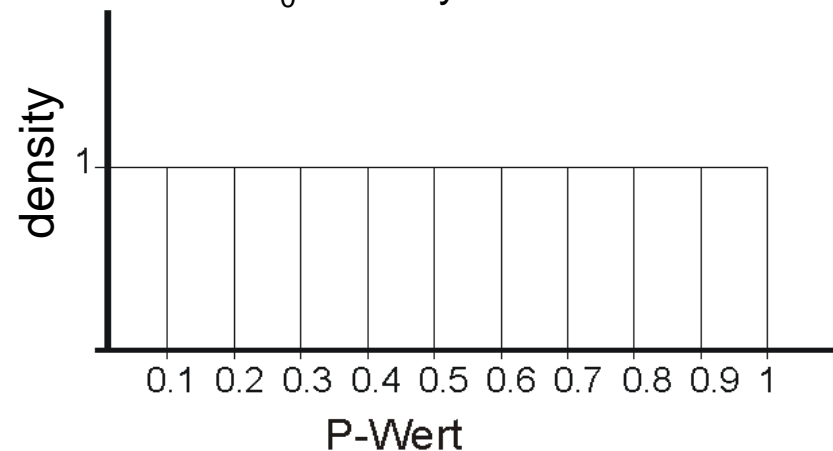
Given  $H_0$  is true in all tests:

$p=0.1$ : 10% of all tests get a  $p\text{-value} \leq 0.1$  if  $H_0$  is always true

$p=0.2$ : 20% of all tests get a  $p\text{-value} \leq 0.2$  if  $H_0$  is always true

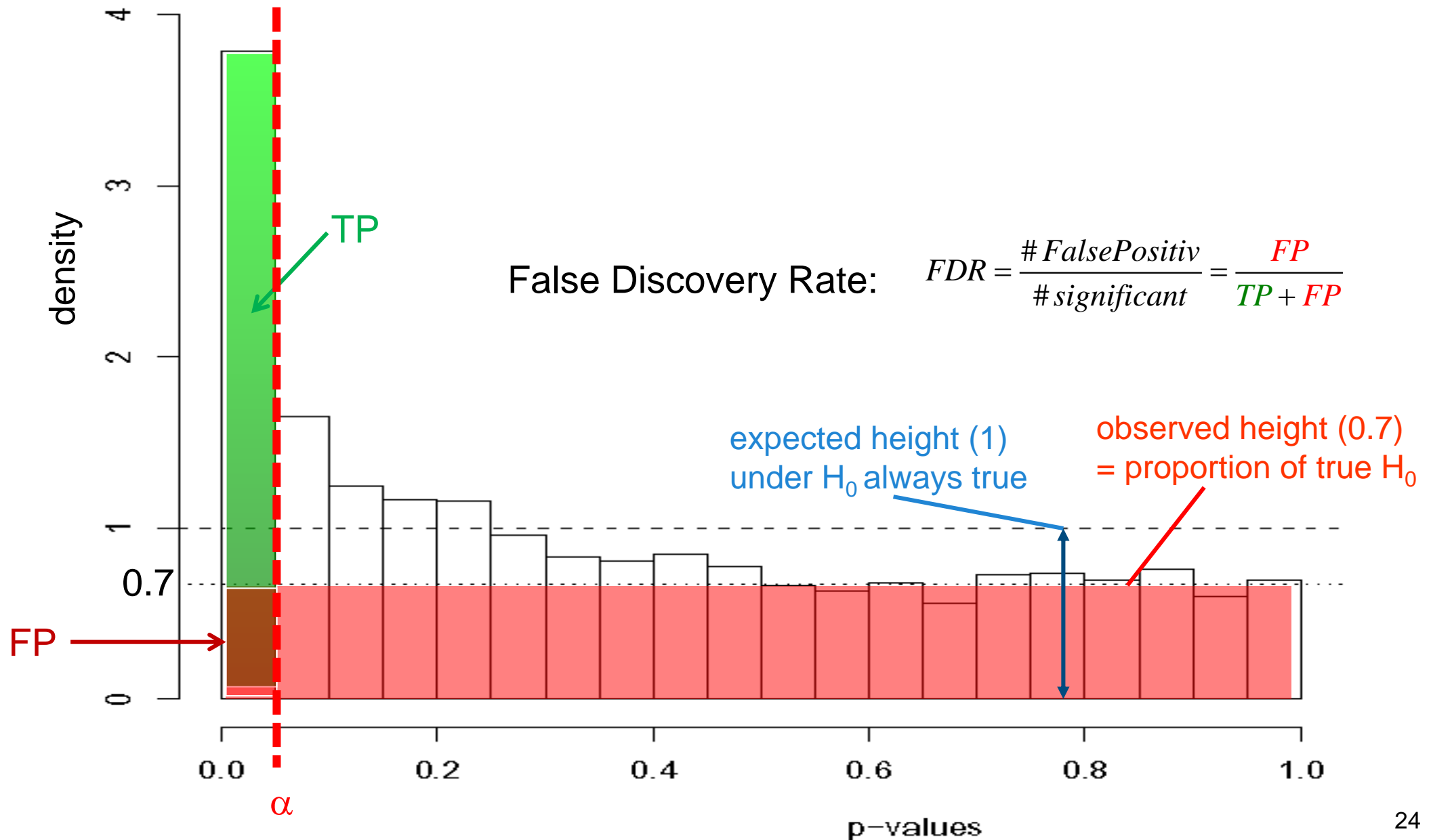
...

p-Wert Histogramm  
if  $H_0$  is always true



# How to estimate the ratio $p_0$ of truly null voxels?

## How to estimate the false discovery rate?



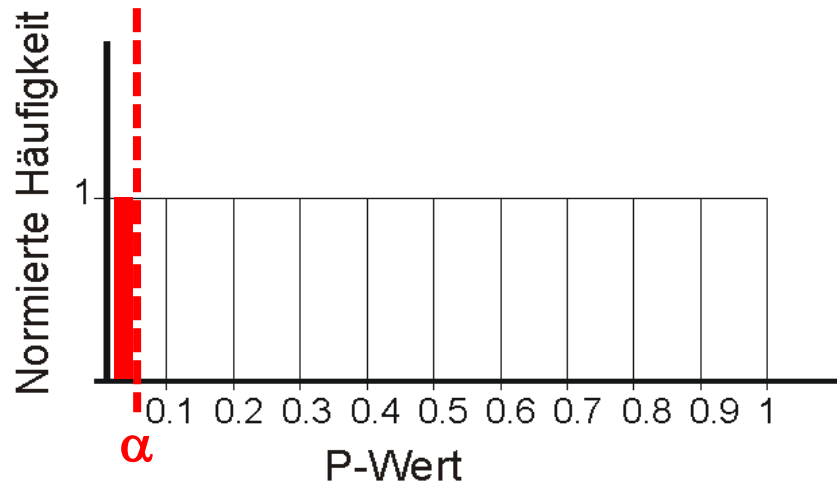


# Judging a p-value histogram

The p-value histogram helps to judge the results from many *independent* tests

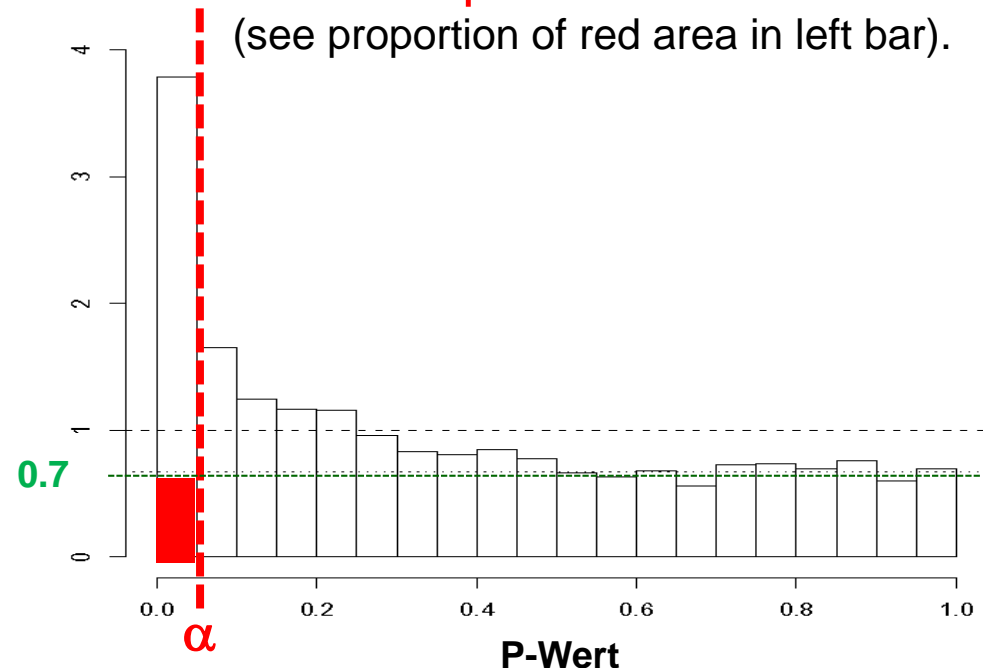
## Flat is Bad!

For all tests  $n$  independent  $H_0$  is true  
~ 100% of all significant findings are false positive.  
(see proportion of red area in left bar).



## The peak we seek!

For 70% of all tests  $H_0$  is true.  
Only ~20% of all significant findings are false positive  
(see proportion of red area in left bar).



## Take home message

### Multiple Testing

It is tempting, but not o.k. to forget about all non-significant tests and just publish the significant effects.

You **need to take account for the multiple testing**

- by p-value correction such as Bonferroni correction or
- using other measures like FDR or
- confirm the “found effect” in a new control experiment.