Hypothesis testing
S2

John Zavrakidis
j.zavrakidis@nki.nl

November 19, 2018

NKI-AVL
The Netherlands Cancer Institute
Antoni van Leeuwenhoek Hospital

Introduction

- Point estimation: use a sample statistic to estimate a population parameter γ of interest.
- Interval estimation: use a 95% CI to give a 95% probability that this interval contains γ.
- Hypothesis testing: assume a value for γ and make a probability statement about the value of the corresponding statistic.
  - A statistical hypothesis is an assumption about a population parameter.
  - Hypothesis testing refers to the formal procedures used to accept or reject statistical hypotheses.

NKI-AVL
The Netherlands Cancer Institute
Antoni van Leeuwenhoek Hospital

Introduction

Example: The average person sleeps 8 hours per day. But can we state that college students tend to sleep also 8 hours?

- Take random sample of 100 college students and ask them how long they sleep on an average day.
- Sample mean $\bar{x} = 6.5$ hours.
- Is this difference just due to sampling variation? Or do college students really sleep different hours than the general population?
- Would your decision change if $\bar{x}$ = 3? And if $\bar{x}$ = 8.5?
- What factors influence the decision?

NKI-AVL
The Netherlands Cancer Institute
Antoni van Leeuwenhoek Hospital

Null and alternative hypothesis

Null hypothesis (H_0): specifies (a) hypothesized true value/s for a parameter.

- Always a statement about independence, no effect or equality in populations (e.g., no difference between means; no relationship between variables).
- Very precise statement.
- Usually a statement that we wish to reject.

Keep it simple: if e.g. we compare the mean effect of a new treatment with a standard treatment, a simple H_0 is that the treatments have the same effect.

NKI-AVL
The Netherlands Cancer Institute
Antoni van Leeuwenhoek Hospital

Null and alternative hypothesis

Alternative hypothesis (H_1): specifies (a) true value/s for a parameter, that will be considered if H_0 rejected.

- Is always a statement about dependence, an effect or differences in populations (e.g., there is a difference between means; there is a relationship between variables).
- Not precise statement.

Null and alternative hypothesis

Null hypothesis:

H_0 : Parameter = Parameter value

Alternative hypothesis:

H_1: Parameter $\neq$ Parameter value

or

H_1: Parameter > Parameter value

or

H_1: Parameter < Parameter value

- **Parameter** is the population parameter; **Parameter value** is the hypothesized value of the parameter
- Example with number of hours college students sleep:

$$H_0 : \mu = 8$$
$$H_1 : \mu < 8$$

NKI-AVL
The Netherlands Cancer Institute
Antoni van Leeuwenhoek Hospital

Null and alternative hypothesis

We test the null hypothesis:
Assuming that the null hypothesis is correct, what is the probability of obtaining our pattern of results?

- We either "reject the null hypothesis" or "fail to reject the null hypothesis"
  - If the probability of obtaining our pattern of results is high, the result is considered to be likely so we do not reject the null hypothesis.

NKI-AVL
The Netherlands Cancer Institute
Antoni van Leeuwenhoek Hospital

Test statistic

- We construct a test statistic:

$$\text{Test statistic} = \frac{\text{Statistic} - \text{Parameter value}}{\text{Standard error of the statistic}} = \frac{\text{Effect}}{\text{Error}}$$

  - Statistic is the observed sample statistic, i.e. point estimate of the population parameter
  - Parameter value is the hypothesized value of the parameter

- Test statistic measures by how many SEs the statistic differs from its hypothesized value in H_0
- Test statistic has a sampling distribution
- Example with number of hours college students sleep where $H_0: \mu = 8$ & $\bar{x} = 6.5$, $SE = 1$:

$$\text{Test statistic} = \frac{6.5-8}{1} = -1.5$$

NKI-AVL
The Netherlands Cancer Institute
Antoni van Leeuwenhoek Hospital

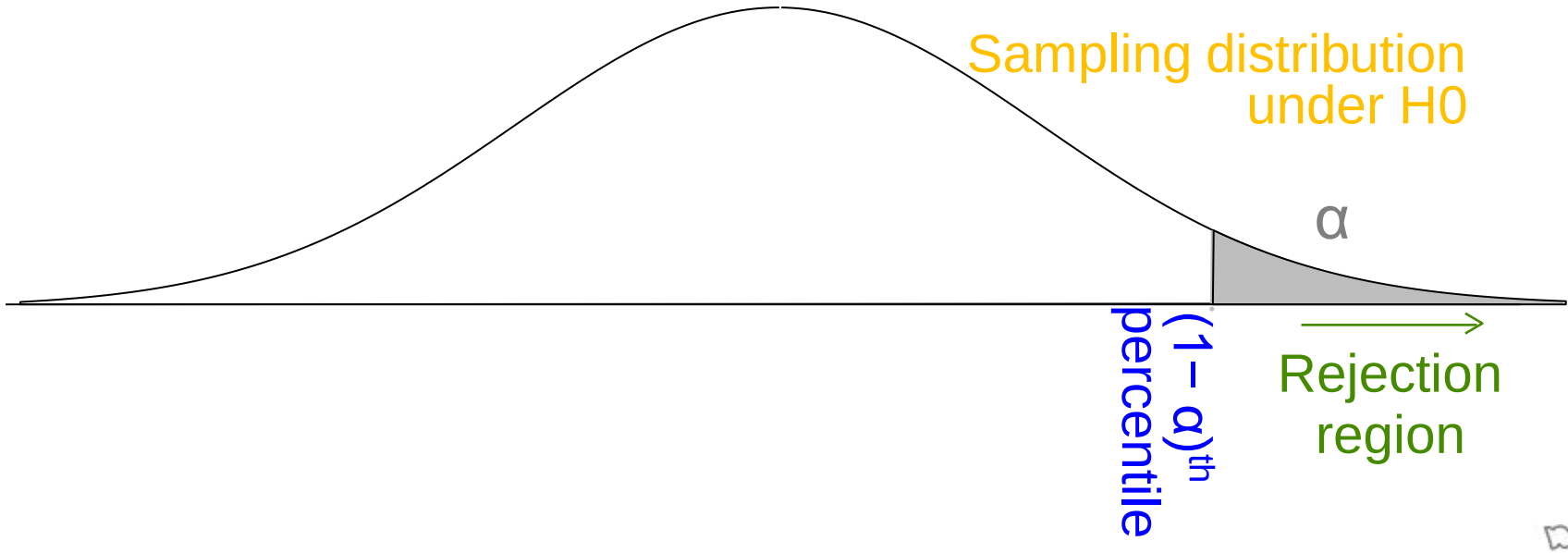
Significance level and critical value

- **Level of significance, α**, specifies how strongly the sample evidence must contradict the null hypothesis before you can reject it.
    - Usually $\alpha = .05$, sometimes $\alpha = .01$ or $\alpha = .1$
    - Level of significance should be chosen before conducting any data analysis.
- **Rejection region** is the set of values for the test statistic that leads to rejection of the null hypothesis
- **Non-rejection region** is the set of values not in the rejection region that leads to non-rejection of the null hypothesis
- **Critical value** is the value of the test statistic that separate the rejection and non-rejection regions

NKI-AVL
The Netherlands Cancer Institute
Antoni van Leeuwenhoek Hospital

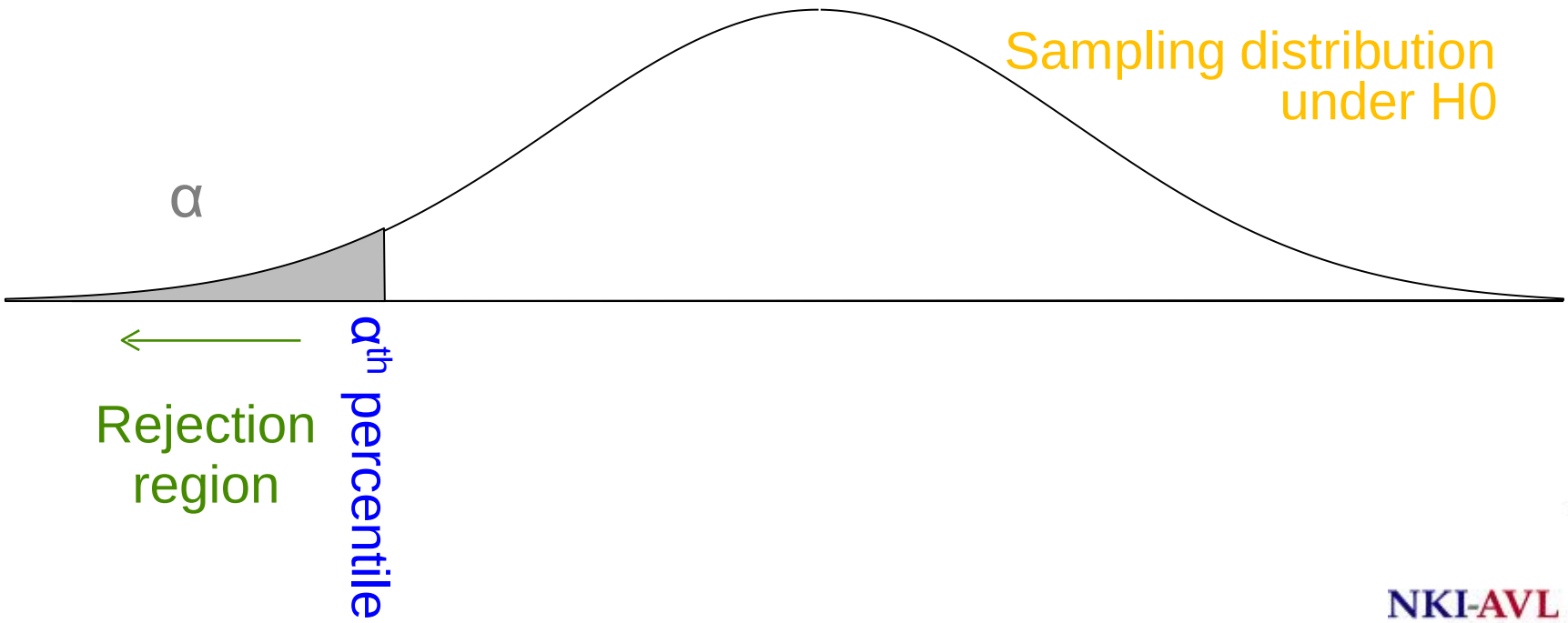
Significance level and critical value

- One-sided test, H_1: Parameter > Parameter value
- When Test statistic $\geq$ $(1-\alpha)^{th}$ quantile: we reject the null hypothesis



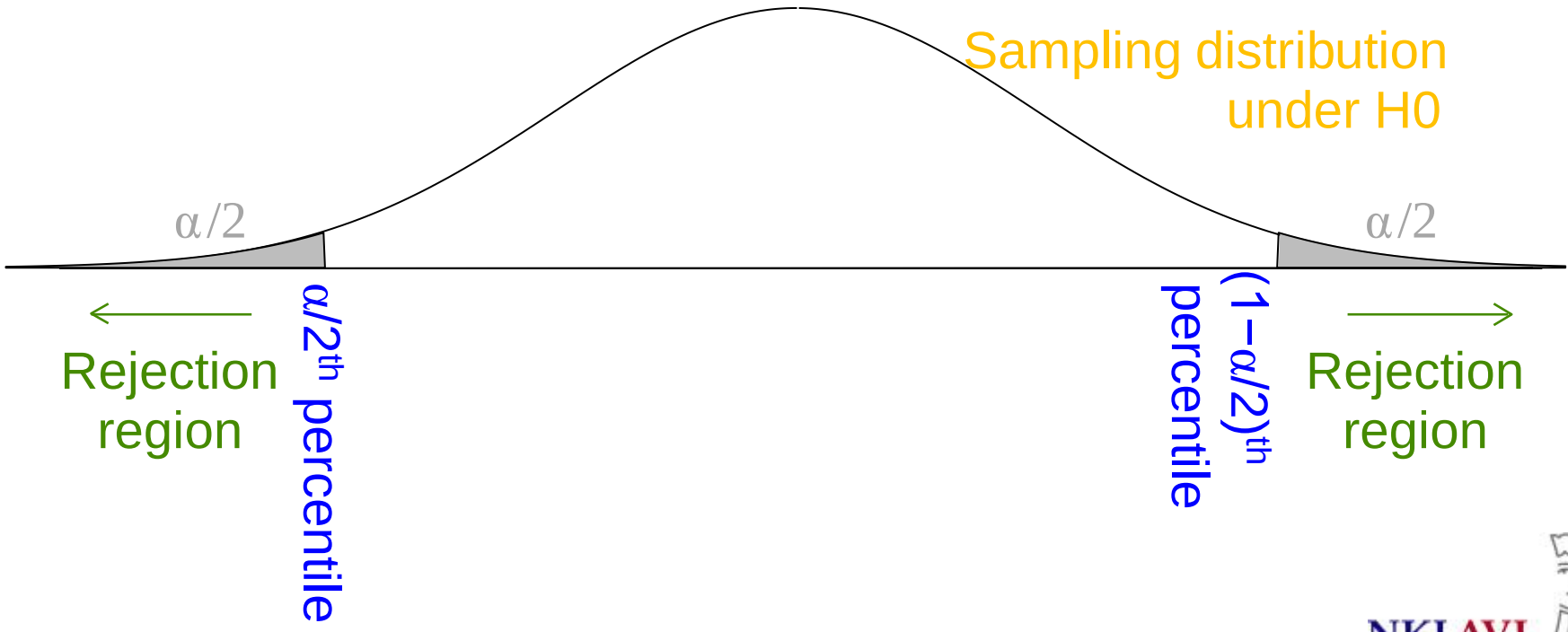
Significance level and critical value

- One-sided test, H_1: Parameter < Parameter value
- When Test statistic $\leq \alpha^{th}$ quantile: we reject the null hypothesis



Significance level and critical value

- Two-sided test, H_1: Parameter ≠ Parameter value
- When Test statistic ≤ $\alpha/2^{th}$ quantile or Test statistic ≥ $(1-\alpha/2)^{th}$ quantile: we reject the null hypothesis



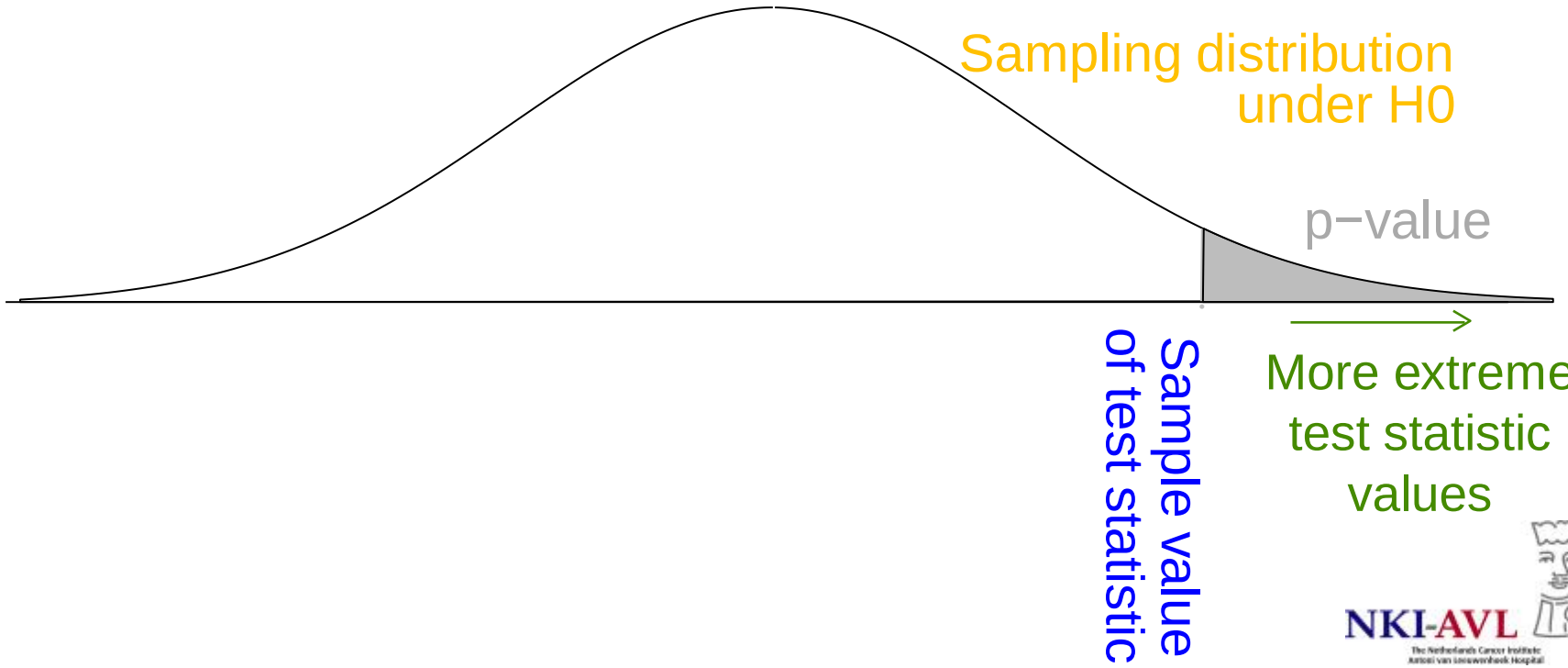
p– value

Using distribution of a test statistic we find probability of obtaining any value of the test statistic

- **p–value** is the probability of obtaining a test statistic result at least as extreme as the one that is actually observed, assuming that the null hypothesis is true
  - when p–value is small the test statistic is significant and the null hypothesis is rejected

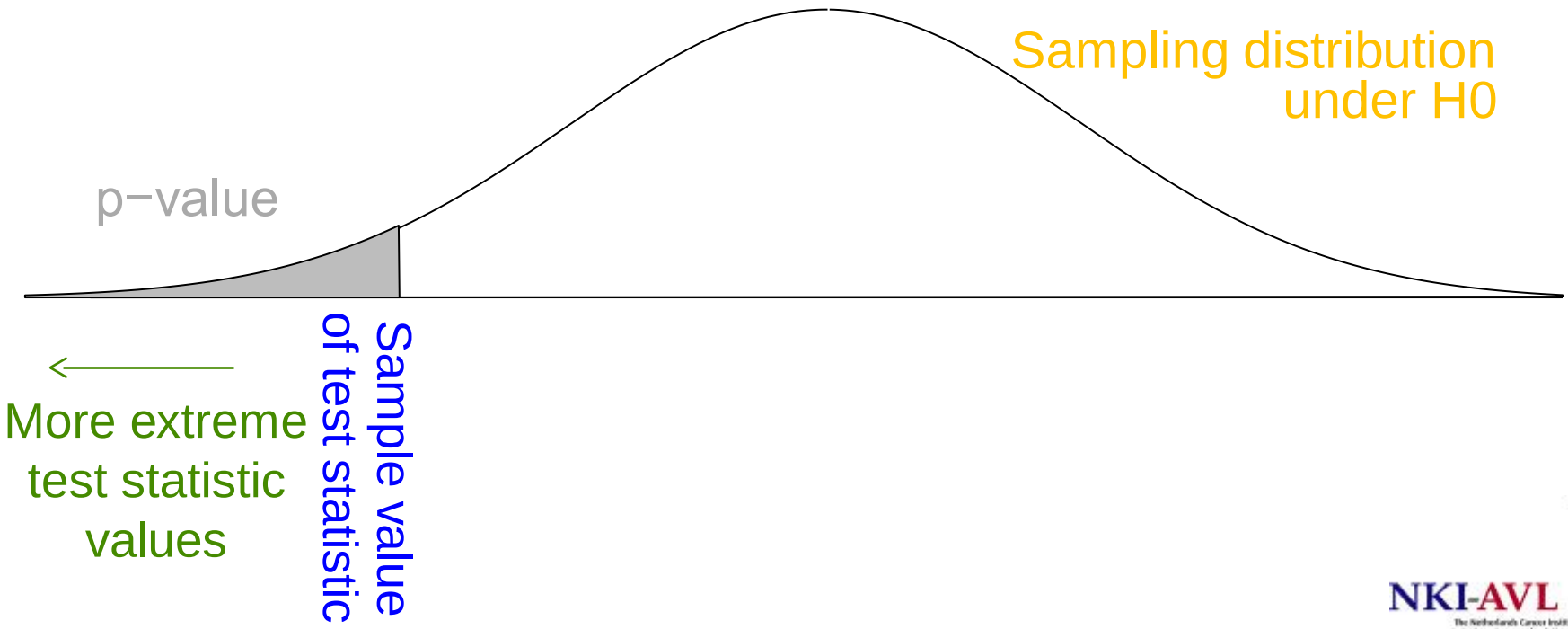
p– value

- One-sided test, H_1: Parameter > Parameter value
- When p–value < α: we reject the null hypothesis



Significance level and critical value

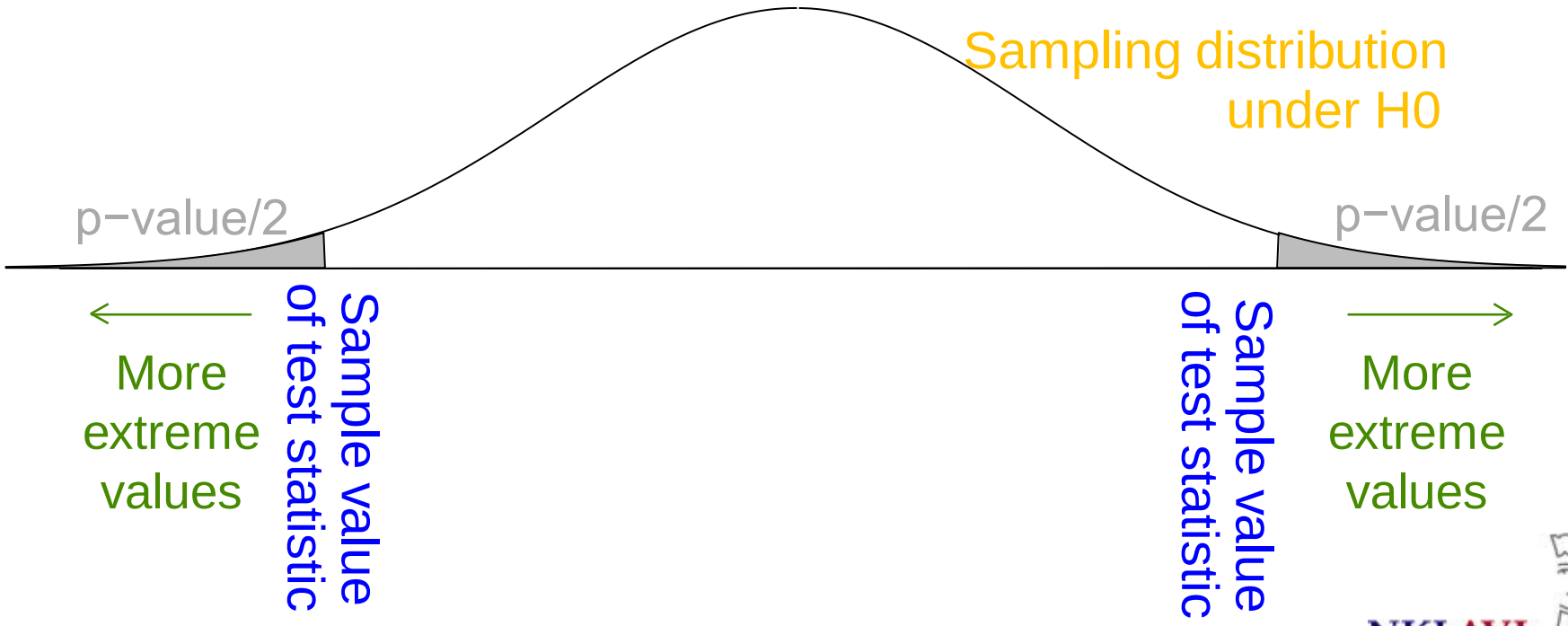
- One-sided test, H_1: Parameter < Parameter value
- When p−value < α: we reject the null hypothesis



NKI-AVL
The Netherlands Cancer Institute
Antoni van Leeuwenhoek Hospital

Significance level and critical value

- Two-sided test, H_1: Parameter ≠ Parameter value
- When p−value < α: we reject the null hypothesis



NKI-AVL
The Netherlands Cancer Institute
Antoni van Leeuwenhoek Hospital

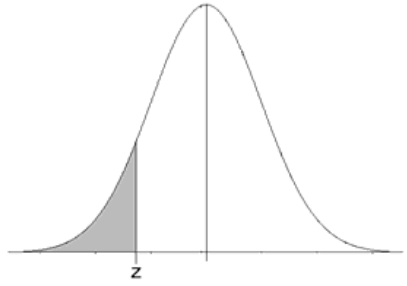
p– value

- p–value lower than 0.05 (0.1): results are "significant at the 5% (10%) level" (small enough to reject H_0)
  - Reject H_0 at a 5% significance level
  - There is evidence to reject H_0
- p–value not lower than 0.05 (0.1): results are “not significant at the 5% (10%) level”
  - Fail to reject H_0 at a 5% significance level
  - There is not enough evidence to reject H_0
  - But this does NOT mean that we accept that H_0 is true
- The cut-off level 0.05 (0.1) is the significance level of the test

NKI-AVL
The Netherlands Cancer Institute
Antoni van Leeuwenhoek Hospital

Statistical table with probabilities

Standard Normal Cumulative Probability Table



Cumulative probabilities for NEGATIVE z-values are shown in the following table:

z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
-3.4	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0002
-3.3	0.0005	0.0005	0.0005	0.0004	0.0004	0.0004	0.0004	0.0004	0.0004	0.0003
-3.2	0.0007	0.0007	0.0006	0.0006	0.0006	0.0006	0.0006	0.0005	0.0005	0.0005
-3.1	0.0010	0.0009	0.0009	0.0009	0.0008	0.0008	0.0008	0.0008	0.0007	0.0007
-3.0	0.0013	0.0013	0.0013	0.0012	0.0012	0.0011	0.0011	0.0011	0.0010	0.0010
-2.9	0.0019	0.0018	0.0018	0.0017	0.0016	0.0016	0.0015	0.0015	0.0014	0.0014
-2.8	0.0026	0.0025	0.0024	0.0023	0.0023	0.0022	0.0021	0.0021	0.0020	0.0019
-2.7	0.0035	0.0034	0.0033	0.0032	0.0031	0.0030	0.0029	0.0028	0.0027	0.0026
-2.6	0.0047	0.0045	0.0044	0.0043	0.0041	0.0040	0.0039	0.0038	0.0037	0.0036
-2.5	0.0062	0.0060	0.0059	0.0057	0.0055	0.0054	0.0052	0.0051	0.0049	0.0048

I.e. If z = -2.54

P-value = 0.0055

NKI-AVL

Statistical table with probabilities

Standard Normal Cumulative Probability Table

z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
-2.4	0.0082	0.0080	0.0078	0.0075	0.0073	0.0071	0.0069	0.0068	0.0066	0.0064
-2.3	0.0107	0.0104	0.0102	0.0099	0.0096	0.0094	0.0091	0.0089	0.0087	0.0084
-2.2	0.0139	0.0136	0.0132	0.0129	0.0125	0.0122	0.0119	0.0116	0.0113	0.0110
-2.1	0.0179	0.0174	0.0170	0.0166	0.0162	0.0158	0.0154	0.0150	0.0146	0.0143
-2.0	0.0228	0.0222	0.0217	0.0212	0.0207	0.0202	0.0197	0.0192	0.0188	0.0183
-1.9	0.0287	0.0281	0.0274	0.0268	0.0262	0.0256	0.0250	0.0244	0.0239	0.0233
-1.8	0.0359	0.0351	0.0344	0.0336	0.0329	0.0322	0.0314	0.0307	0.0301	0.0294
-1.7	0.0446	0.0436	0.0427	0.0418	0.0409	0.0401	0.0392	0.0384	0.0375	0.0367
-1.6	0.0548	0.0537	0.0526	0.0516	0.0505	0.0495	0.0485	0.0475	0.0465	0.0455
-1.5	0.0668	0.0655	0.0643	0.0630	0.0618	0.0606	0.0594	0.0582	0.0571	0.0559
-1.4	0.0808	0.0793	0.0778	0.0764	0.0749	0.0735	0.0721	0.0708	0.0694	0.0681
-1.3	0.0968	0.0951	0.0934	0.0918	0.0901	0.0885	0.0869	0.0853	0.0838	0.0823
-1.2	0.1151	0.1131	0.1112	0.1093	0.1075	0.1056	0.1038	0.1020	0.1003	0.0985
-1.1	0.1357	0.1335	0.1314	0.1292	0.1271	0.1251	0.1230	0.1210	0.1190	0.1170
-1.0	0.1587	0.1562	0.1539	0.1515	0.1492	0.1469	0.1446	0.1423	0.1401	0.1379
-0.9	0.1841	0.1814	0.1788	0.1762	0.1736	0.1711	0.1685	0.1660	0.1635	0.1611
-0.8	0.2119	0.2090	0.2061	0.2033	0.2005	0.1977	0.1949	0.1922	0.1894	0.1867
-0.7	0.2420	0.2389	0.2358	0.2327	0.2296	0.2266	0.2236	0.2206	0.2177	0.2148
-0.6	0.2743	0.2709	0.2676	0.2643	0.2611	0.2578	0.2546	0.2514	0.2483	0.2451
-0.5	0.3085	0.3050	0.3015	0.2981	0.2946	0.2912	0.2877	0.2843	0.2810	0.2776

Statistical table with probabilities

- Example with number of hours college students sleep:

$H_0 : \mu = 8$ vs $H_1 : \mu < 8$

$\bar{x} = 6.5$ & SE = 1:

Test statistic = $\frac{6.5-8}{1} = -1.5$

P-value = 0.0668

z	0.00
-2.4	0.0082
-2.3	0.0107
-2.2	0.0139
-2.1	0.0179
-2.0	0.0228
-1.9	0.0287
-1.8	0.0359
-1.7	0.0446
-1.6	0.0548
-1.5	0.0668

- If we assume that the test statistic has a normal distribution; when we change the alternative hypothesis to $H_1 : \mu \neq 8$, then p−value = 2*0.0668 = 0.1336

NKI-AVL
The Netherlands Cancer Institute
Antoni van Leeuwenhoek Hospital

Hypothesis testing step by step

1. Define the null and alternative hypotheses under study.
2. Collect relevant data from a sample of individuals.
3. Calculate the value of the test statistic.
4. Compare the value of the observed test statistic to a critical value of the distribution of the test statistic under H_0.
5. Interpret the results.

NKI-AVL
The Netherlands Cancer Institute
Antoni van Leeuwenhoek Hospital

Type I and Type II errors

In a hypothesis test we draw a conclusion about H_0 (there is strong/weak evidence to reject/not reject) based on a sample

	Decision	
Truth	Reject H_0	Do not reject H_0
H_0 true	Type I error (α)	Correct decision (1-α)
H_0 false	Correct decision (1 – β)	Type II error (β)

- $\boldsymbol{\alpha}$: significance level of the test.
  - Probability of rejecting H_0 when H_0 is true.
- $\boldsymbol{\beta}$: Probability of not rejecting H_0 when H_0 is false.
  - $\mathbf{1-\beta}$: power of the test. Probability of rejecting H_0 when H_0 is false. i.e. detecting a true effect

NKI-AVL
The Netherlands Cancer Institute
Antoni van Leeuwenhoek Hospital

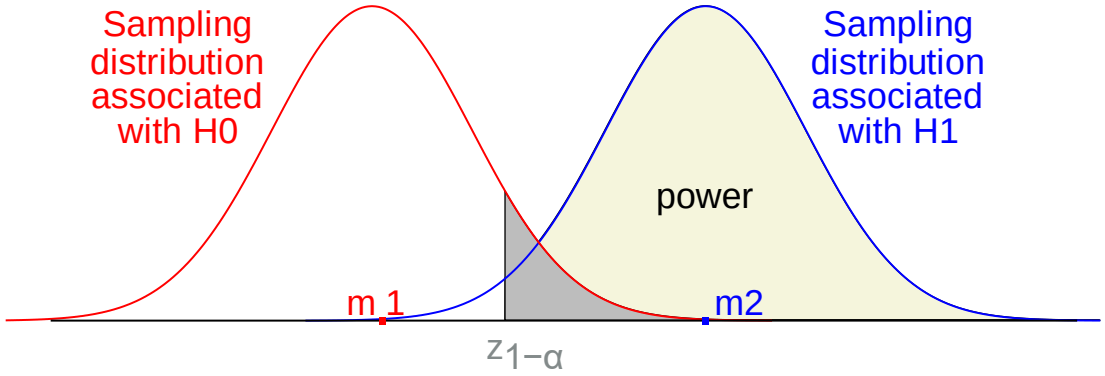
Power

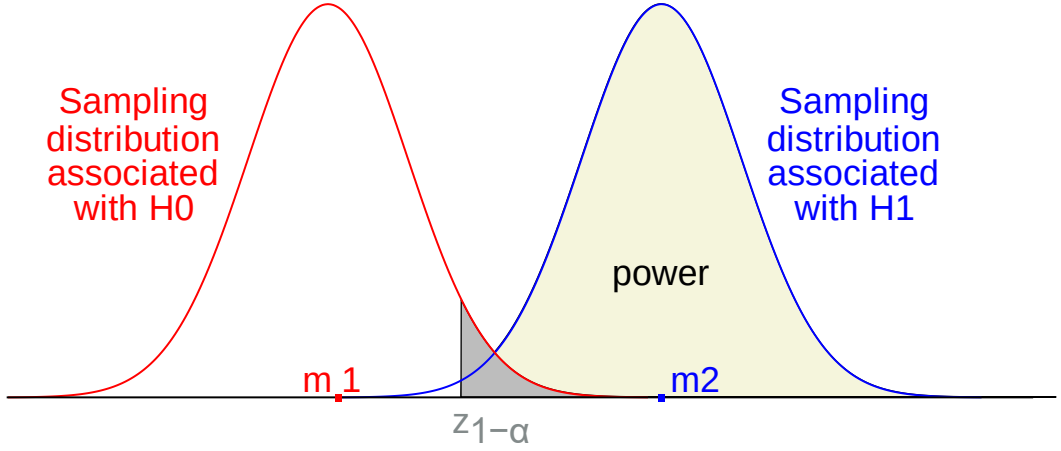
- Essential to know the power of a proposed test.
  - We want to be confident that our statistical method can correctly reject the null hypothesis.
- It is accepted that power should be 0.8 or greater.
  - Study with a smaller power might waste our time and resources.
  - If the power is smaller than 0.8 we might want to replicate our study.
- Factors related to power:
  - ↑ sample size ⇒ ↑ power.
  - ↓ variability of observations ⇒ ↑ power.
  - ↑ effects of interest ⇒ ↑ power
  - ↑ significance level α ⇒ ↑ power

NKI-AVL
The Netherlands Cancer Institute
Antoni van Leeuwenhoek Hospital

Power

Measurement variability influence

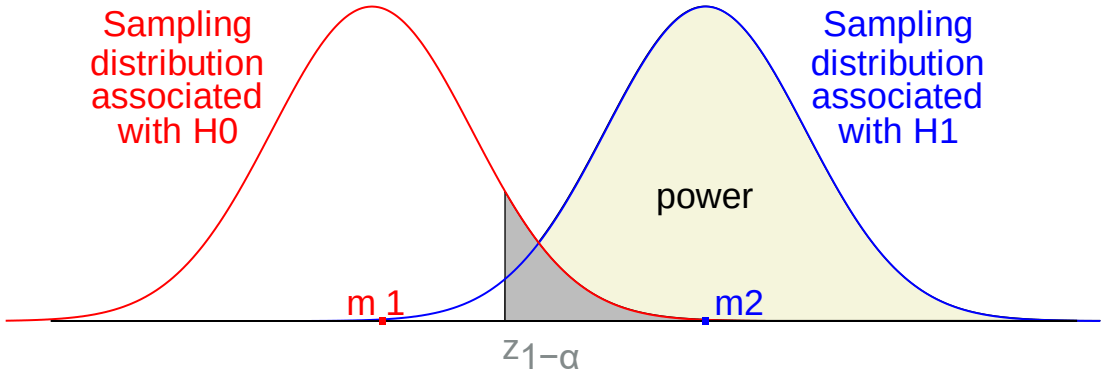


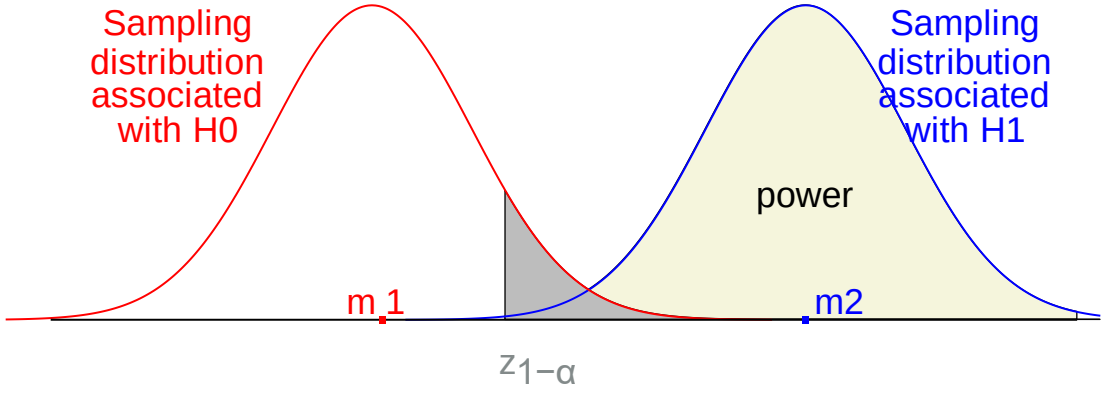


NKI-AVL
The Netherlands Cancer Institute
Antoni van Leeuwenhoek Hospital

Power

Effect size influence

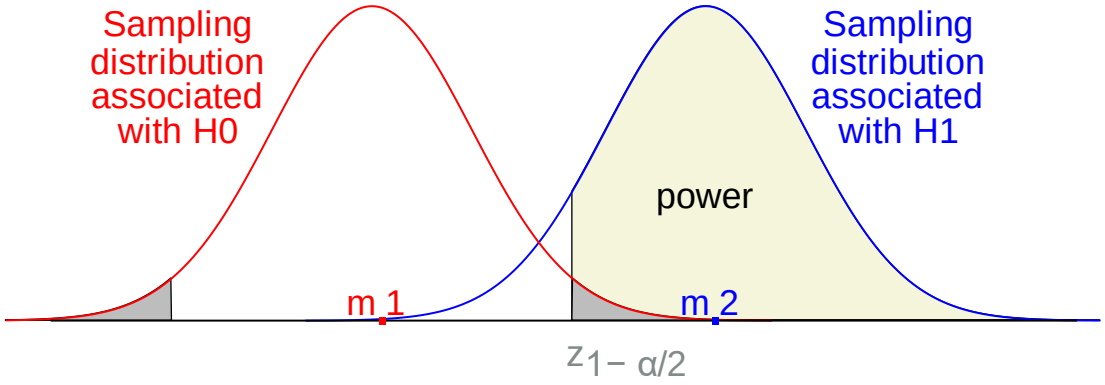


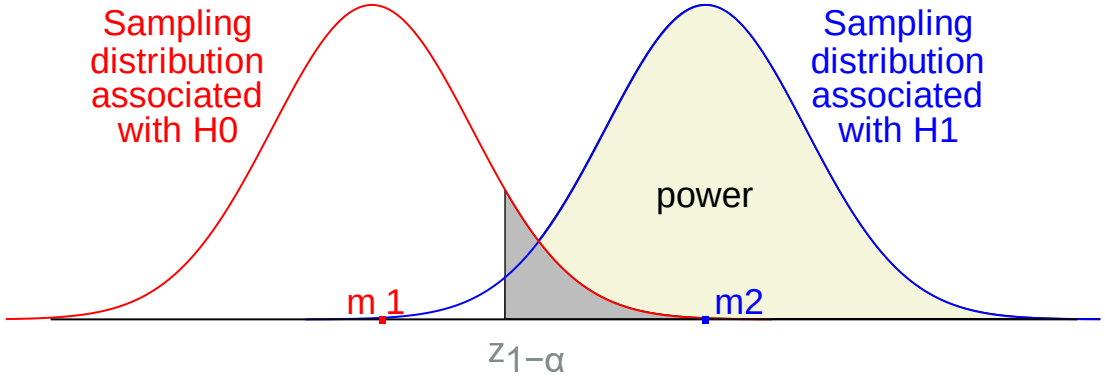


NKI-AVL
The Netherlands Cancer Institute
Antoni van Leeuwenhoek Hospital

Power

α and test type influence





NKI-AVL
The Netherlands Cancer Institute
Antoni van Leeuwenhoek Hospital

Power

Statistical significance vs. Clinical relevance

- Significant result ≠ effect meaningful or important.
- Underpowered studies fail to detect clinically relevant effects as statistically significant (False Negative)
- Overpowered studies might show small clinically irrelevant effects to be statistically significant.
  - Reduction of blood pressure by two points between treatment and placebo.
  - An expensive new psychiatric treatment technique reduces the average hospital stay from 60 days to 59 days.

NKI-AVL
The Netherlands Cancer Institute
Antoni van Leeuwenhoek Hospital

Power

- Power analysis calculation should be used at the planning stage of investigation

- For any power calculation we need to know:
    - Type of a test (e.g., independent t -test, paired t -test, ANOVA, regression, etc.)
    - The significance level, α
    - The expected effect size
    - The sample size

Power

How large should the sample be in order to detect a specific effect size?

- **One-sample case example:**

$H_0: \mu = m$
$H_1: \mu < m$
or
$\mu > m$

$$\rightarrow n \approx \sigma_x{}^2 \left(\frac{Z_{1-\alpha} + Z_{1-\beta}}{\bar{x} - m} \right)^2$$

$H_0: \mu = m$
$H_1: \mu \neq m$

$$\rightarrow n \approx \sigma_x{}^2 \left(\frac{Z_{1-\alpha/2} + Z_{1-\beta}}{\bar{x} - m} \right)^2$$

$\bar{x}$: The observed value of the sample mean
$\sigma_x{}^2$: Variance of the measured variable

Power

How large should the sample be in order to detect a specific effect size?

- **Two-sample case example:**

$H_0 : \mu 1 = \mu 2$
$H_1 : \mu 1 < \mu 2$ or $\mu 1 > \mu 2$ $\quad \rightarrow n_1 \approx \left(\sigma_{x_1}{}^2 + \sigma_{x_2}{}^2 / k\right)\left(\frac{z_{1-\alpha}+z_{1-\beta}}{\bar{x}_1-\bar{x}_2}\right)^2$

$H_0 : \mu_1 = \mu_2$
$H_1 : \mu_1 \neq \mu_2$ $\quad \rightarrow n_1 \approx \left(\sigma_{x_1}{}^2 + \sigma_{x_2}{}^2 / k\right)\left(\frac{z_{1-\alpha/2}+z_{1-\beta}}{\bar{x}_1-\bar{x}_2}\right)^2$

$\bar{x}$: The observed value of the sample mean
$\sigma_{x_1}{}^2$: Variance of the measured variable

k = n1/n2: ratio between the sample sizes of the two groups

NKI-AVL
The Netherlands Cancer Institute
Antoni van Leeuwenhoek Hospital

Power

- $z_{1-\alpha}$, $z_{1-\beta}$: critical values from the standard normal distribution

- $\alpha = 5\% \rightarrow z_{1-\alpha} = 1.65,\ z_{1-\alpha/2} = 1.96$
- $\alpha = 1\% \rightarrow z_{1-\alpha} = 2.33,\ z_{1-\alpha/2} = 2.58$

- power= 80% $\rightarrow z_{1-\beta} = 0.84$
- power= 85% $\rightarrow z_{1-\beta} = 1.04$
- power= 90% $\rightarrow z_{1-\beta} = 1.28$
- power= 95% $\rightarrow z_{1-\beta} = 1.65$

NKI-AVL
The Netherlands Cancer Institute
Antoni van Leeuwenhoek Hospital

Power

Example

A cardiologist is interested in difference in BMI levels between heart attack patients and healthy people. How many people should take part in the study to be able to assess that heart attack patients have on average different BMI levels that healthy ones?

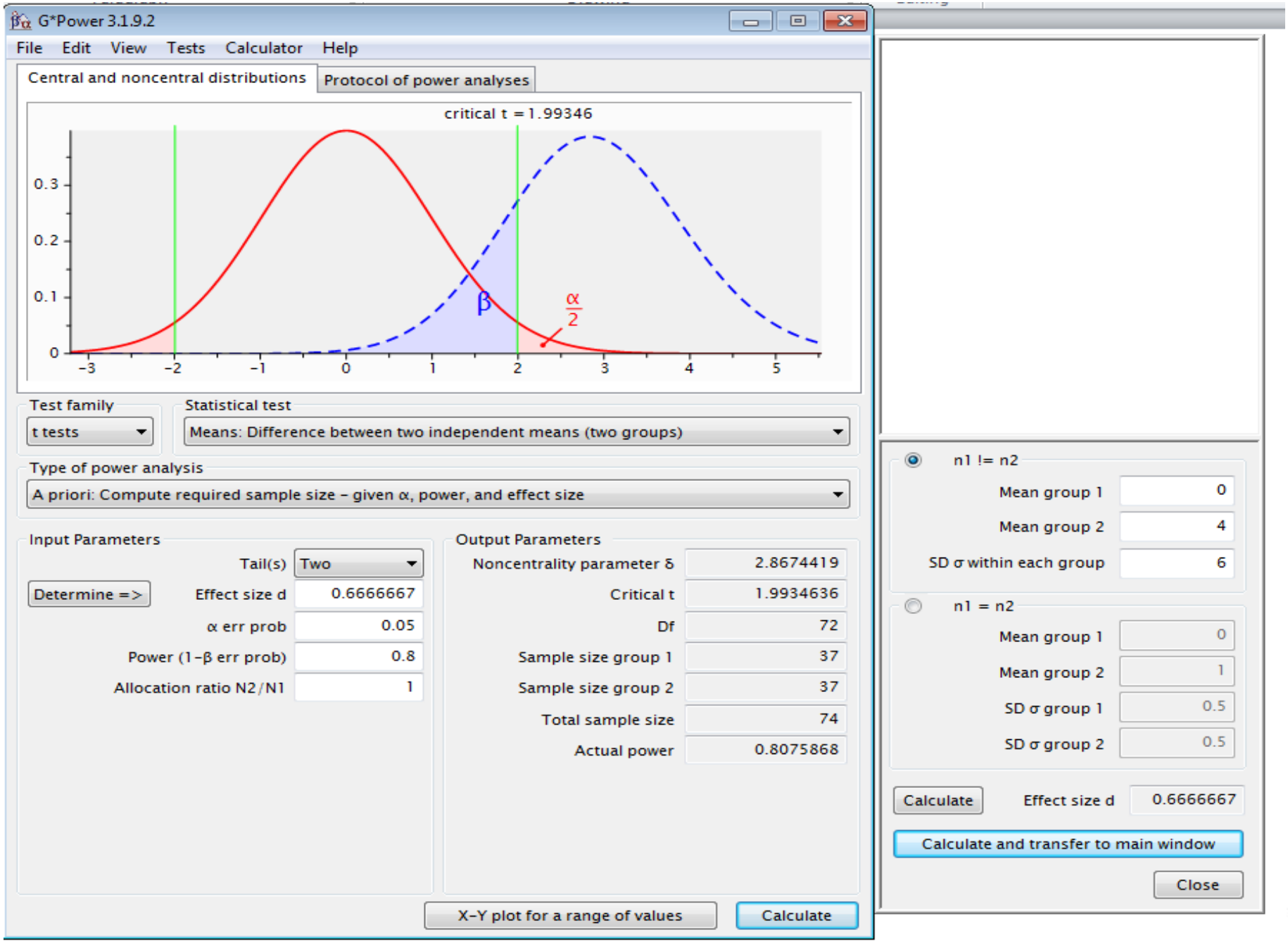
H0 : BMI levels are equal between heart attack patients and healthy people

H1 : BMI levels are not equal between heart attack patients and healthy people

- Expected effect size = 4
- Standard deviation = 6
- Equally sized groups, k = 1
- Desired power = 80%

$$n_1 \approx \left(\sigma_{x_1}{}^2 + \sigma_{x_2}{}^2 / k\right)\left(\frac{z_{1-\alpha/2}+z_{1-\beta}}{\bar{x}_1-\bar{x}_2}\right)^2 = (6^2 + 6^2/1)\left(\frac{1.96+0.84}{4}\right)^2$$

$$\approx 36$$



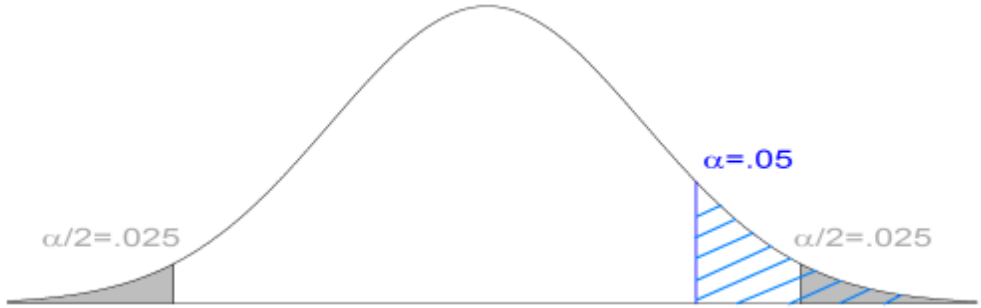
http://www.gpower.hhu.de/en.html

NKI-AVL
The Netherlands Cancer Institute
Antoni van Leeuwenhoek Hospital

One-tailed vs. two-tailed

If e.g. we wish to test parameter μ:

- $H_0 : \mu = 5$ vs. $H_1 : \mu \neq 5$ will require a two-tailed test (two-sided)
  - We test for the possibility of deviation from H_0 in both directions.
  - For any α, half of it is used to test statistical significance in one direction ($\mu > 5$) and the other half for the other direction ($\mu < 5$).
- $H_0 : \mu = 5$ vs. $H_1 : \mu > 5$ will require a one-tailed test (one-sided)
  - We completely disregard the possibility of $\mu < 5$.
  - This provides more power to detect an effect in direction $\mu > 5$ by not testing $\mu < 5$.
  - Hardly ever appropriate with biomedical data.



NKI-AVL
The Netherlands Cancer Institute
Antoni van Leeuwenhoek Hospital

Multiple hypothesis testing

Example: We have three groups of participants and we want to compare every pair of groups. We carry out three separate independent tests with $\alpha = 0.05$ for each:

- 1 test: probability of a correct decision (given H_0 is true) is 1-0.05=0.95, $\alpha = 1 - 0.95 = 0.05$

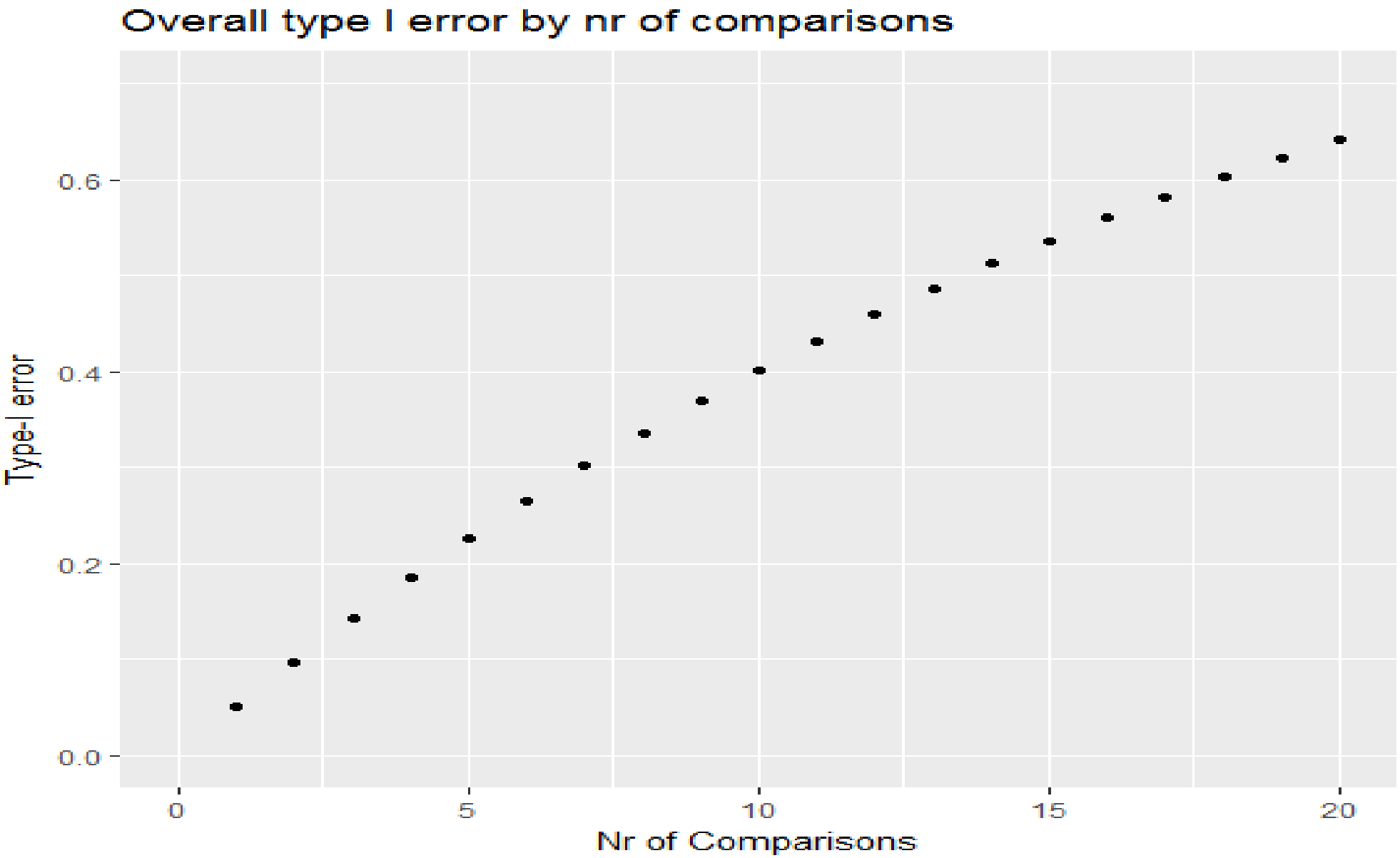
- 2 tests: probability of a correct decision (given H_0 is true) is (1-0.05)(1-0.05)=0.9025, overall $\alpha = 1 - 0.9025 = 0.0975$

- 3 tests: probability of a correct decision (given H_0 is true) is (1-0.05)(1-0.05)(1-0.05)=0.8573, overall $\alpha = 1 - 0.8573 = 0.1426$

❖ The probability of making a type 1 error (probability of falsely rejecting at least one null hypothesis) increases from 5% to 14.3%!

NKI-AVL
The Netherlands Cancer Institute
Antoni van Leeuwenhoek Hospital

Multiple hypothesis testing



Multiple hypothesis testing

❖ Familywise error rate or experimentwise error rate or overall α is the error rate across statistical tests conducted on the same data; i.e., this is the probability that at least one test erroneously rejects the null hypothesis:

$$\text{familywise error rate} = 1 - (0.95)^{\text{number of tests}}$$

- E.g. Carry out 20 tests on a dataset. The probability that at least one of them erroneously rejects H_0 is $1 - (1 - \alpha)^{20} = 64\%$! We cannot identify which one(s), if any, are false positives...
- We can use some post-hoc adjustment, to account for how many tests we perform and control the familywise error rate, (will be covered later this week)

- E.g. Bonferroni: divide α by the number of comparisons to ensure that the overall α is below 0.05. For 10 tests we use 0.005 as our criterion for significance

NKI-AVL
The Netherlands Cancer Institute
Antoni van Leeuwenhoek Hospital

Parametric and non-parametric tests

Parametric: data follows a parametric probability distribution.

Non-parametric: replace data with their ranks (numbers describing position in the ordered dataset) and make no assumption about probability distribution of the data/ranks.

Example:

scores:	5, 2, 6, 8, 7, 1, 10
ordered scores:	1, 2, 5, 6, 7, 8, 10
ranks:	1, 2, 3, 4, 5, 6, 7

- Useful when sample size is small and/or data are measured on a categorical scale.
- However: less power to detect a true effect than the equivalent parametric test (given assumptions of this are satisfied).

NKI-AVL
The Netherlands Cancer Institute
Antoni van Leeuwenhoek Hospital

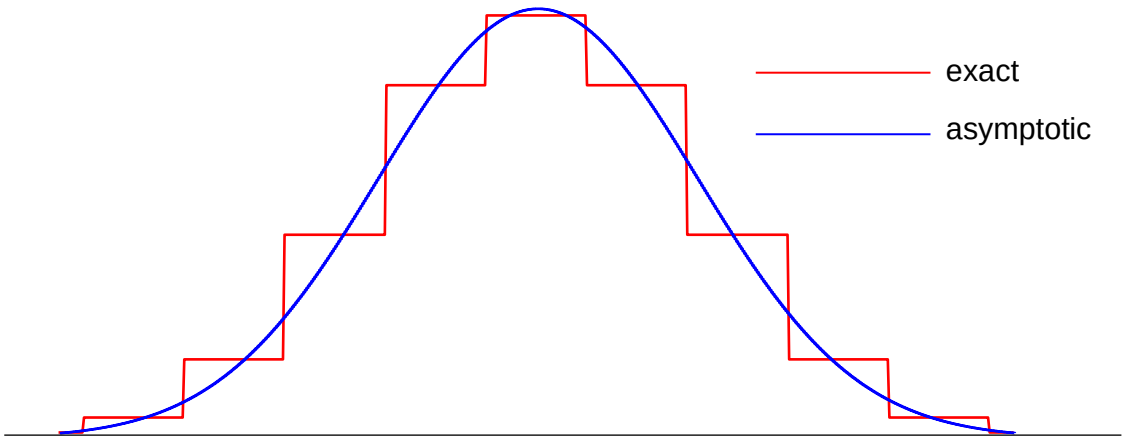
Asymptotic and exact tests

Asymptotic method:

- P-values estimated assuming the test statistic, given a sufficiently large sample size, conform to a particular parametric distribution; e.g. test statistic follows a normal distribution, binomial distribution, chi-square distribution, Wilcoxon distribution
- May not yield reliable results if the data is small, sparse, heavily tied, unbalanced or poorly distributed

Exact method:

- p-value estimated based on exact distribution of the test statistic



NKI-AVL
The Netherlands Cancer Institute
Antoni van Leeuwenhoek Hospital

Hypothesis tests vs. confidence intervals

- A CI can be used to make a decision in a similar way to a hypothesis test:
  - In the sleeping hours example: if the hypothesized mean value 8 lies outside the corresponding 95% CI for the mean then hypothesized value is implausible and enough evidence to reject H_0 (no p-value is used):
  - $\bar{x} = 6.5$, SEM=1:

    CI = [6.5-1.96*1 ; 6.5+1.96*1] = [4.54 ; 8.46]

    8 lies inside the 95% CI

- If a hypothesis test and a CI can do the same, why not use just one of them?

NKI-AVL
The Netherlands Cancer Institute
Antoni van Leeuwenhoek Hospital

Hypothesis tests vs. confidence intervals

Hypothesis testing

- Hypothesis testing provides no information about the probability of H_0, just gives the probability of finding a specific or more extreme results when the null hypothesis is true

- Statistical significance ≠ medical importance
    - Large sample sizes -> small & not interesting differences are statistically significant
    - Small samples sizes -> effects that are clinically relevant might not be statistically significant

- Hypothesis testing used to calculate required sample

NKI-AVL
The Netherlands Cancer Institute
Antoni van Leeuwenhoek Hospital

Hypothesis tests vs. confidence intervals

Confidence interval

- CI does not depend on a priori hypotheses
- CI displays the entire set of believable values of the parameter, so we can determine whether or not the difference between the true value and the H_0 value has practical importance

NKI-AVL
The Netherlands Cancer Institute
Antoni van Leeuwenhoek Hospital