Lecture 4

Confidence intervals for non-normal distributions

Null Hypothesis testing

Introduction to significance tests

Theophanis Tsandilas

Complex designs

Experimental designs can be more complex in different aspects:

- Control for multiple **factors** (**independent variables**)
- Study more than two levels per factor
- Combine between-participants and repeated-measures designs

Example

A research team is interested in assessing the effect of a Geometry course on students' IQ performance. They randomly create two groups of students (*Control vs. Geometry*). Each student takes three IQ tests over three weeks.

Control Group

	P1	P2	P3	P4	P5	P6	P7	P8	P9	P10
Week 1	102	94	90	104	95	100	97	96	96	96
Week 2	105	97	93	96	99	105	100	93	103	98
Week 3	101	100	93	106	98	105	102	98	104	93

Geometry Class Group

	P11	P12	P13	P14	P15	P16	P17	P18	P19	P20
Week 1	105	98	100	97	103	104	103	103	100	99
Week 2	106	100	93	100	100	105	100	95	96	98
Week 3	103	101	98	106	98	100	101	95	104	97

Storing data on CSV files

```
group, participant, week, score
"control", 1, 1, 102
"control", 1, 2, 105
"control", 1, 3, 101
"control", 2, 1, 94
"control", 2, 2, 97
"control", 2, 3, 100
"control", 3, 1, 90
"control", 3, 2, 93
...
"geometry", 11, 1, 105
"geometry", 11, 2, 106
"geometry", 11, 3, 103
"geometry", 12, 1, 98
"geometry", 12, 2, 100
"geometry", 12, 3, 101
"geometry", 13, 1, 100
"geometry", 13, 2, 93
"geometry", 13, 3, 98
"geometry", 14, 1, 97
"geometry", 14, 2, 100
"geometry", 14, 3, 106
...
```

header

each raw represents a unique observation

Working with data frames on R

```
#read the data from a csv file into a data frame
data <- read.csv(file="IQ-tests.csv", header=TRUE, sep=“,")

# aggregate the scores by group and participant
data.aggr <- aggregate(score~group+participant, data,
   FUN = mean)

# split the aggregated data into two groups
data.control <- data.aggr[data.aggr$group==“control”,]
data.geometry <- data.aggr[data.aggr$group==“geometry”,]

# calculate the means
mean.control <- mean(data.control$score)
mean.geometry <- mean(data.geometry$score)
```

Some bad practices

A **bad practice** is to construct and graph confidence intervals over the full set of data

This approach is not appropriate as it treats all observations as independent, i.e., as observations from different participants.

Recommendation: Be always clear about what reported CIs represent and how they were calculated.

Example

Question 2: How does student performance evolve over time, from Week 1 to Week 3?

Control Group

	P1	P2	P3	P4	P5	P6	P7	P8	P9	P10
Week 1	102	94	90	104	95	100	97	96	96	96
Week 2	105	97	93	96	99	105	100	93	103	98
Week 3	101	100	93	106	98	105	102	98	104	93

Geometry Class Group

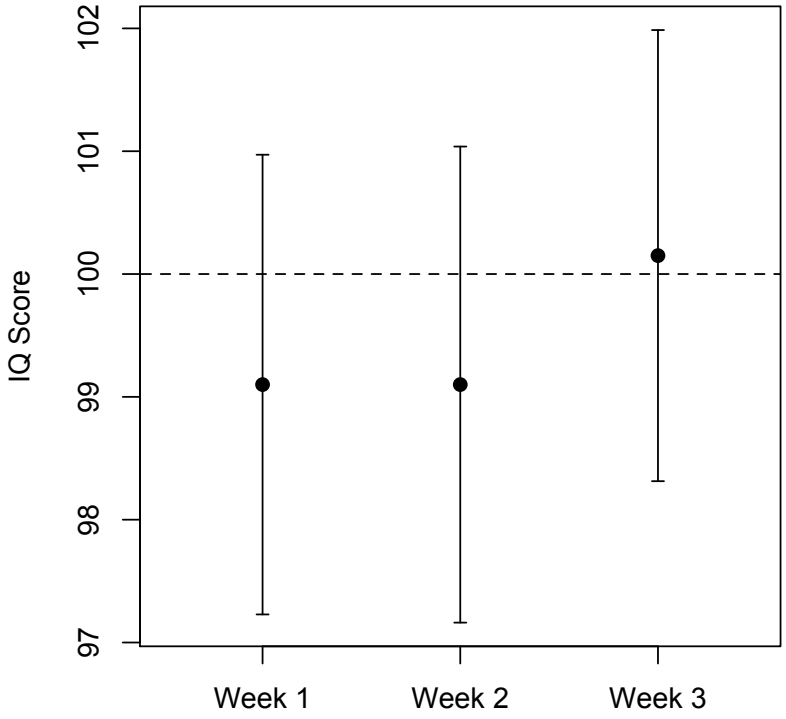
	P11	P12	P13	P14	P15	P16	P17	P18	P19	P20
Week 1	105	98	100	97	103	104	103	103	100	99
Week 2	106	100	93	100	100	105	100	95	96	98
Week 3	103	101	98	106	98	100	101	95	104	97

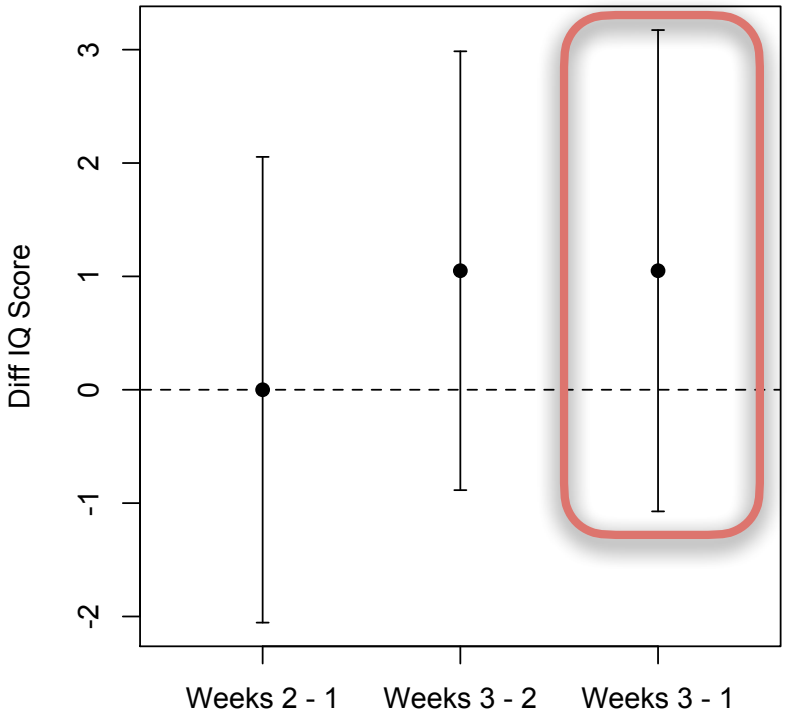
Example: Approach

We examine how IQ scores evolve for all 20 students.

We forget the student groups for now and focus on the repeated-measures variable, the **Week**.

Example: Results





Control Group

	P1	P2	P3	P4	P5	P6	P7	P8	P9	P10
Week 1	102	94	90	104	95	100	97	96	96	96
Week 2	105	97	93	96	99	105	100	93	103	98
Week 3	101	100	93	106	98	105	102	98	104	93
Week 3 -1	-1	6	3	2	3	5	5	2	8	-3

Geometry Class Group

	P11	P12	P13	P14	P15	P16	P17	P18	P19	P20
Week 1	105	98	100	97	103	104	103	103	100	99
Week 2	106	100	93	100	100	105	100	95	96	98
Week 3	103	101	98	106	98	100	101	95	104	97
Week 3 -1	-2	3	-2	9	-5	-4	-2	-8	4	-2

Non-normal distributions

How do we construct confidence intervals for non-normal distributions?

CIs for binomial proportions

Example: A basketball player has attempted a total of $n = 25$ three point shots and has succeeded in $x = 9$. Can you estimate the player's average success rate?

There are several alternatives for constructing CIs for binomial proportions, where some of them work well only for large samples (Control Limit Theorem) or proportions in the region of $.5$.

Baguley recommends the exact Blaker CI under most circumstances.

R Code

```
install.packages("exactci")
library(exactci)

ci <- binom.exact(9, 25, tsmethod=‘blaker',
                  conf.level=.95)$conf.int

cat(ci[1]*100, ci[2]*100, “\n”)
```

The success rate of the player is: 36%, 95% CI [19%, 56%]

CI for a difference in proportions

Example: A second basketball player has attempted a total of 20 three-point shots and has succeeded in 10. Estimate the difference between the success rate of the two players

Baguley recommends the **continuity corrected version of the Wilson CI**.

R Code

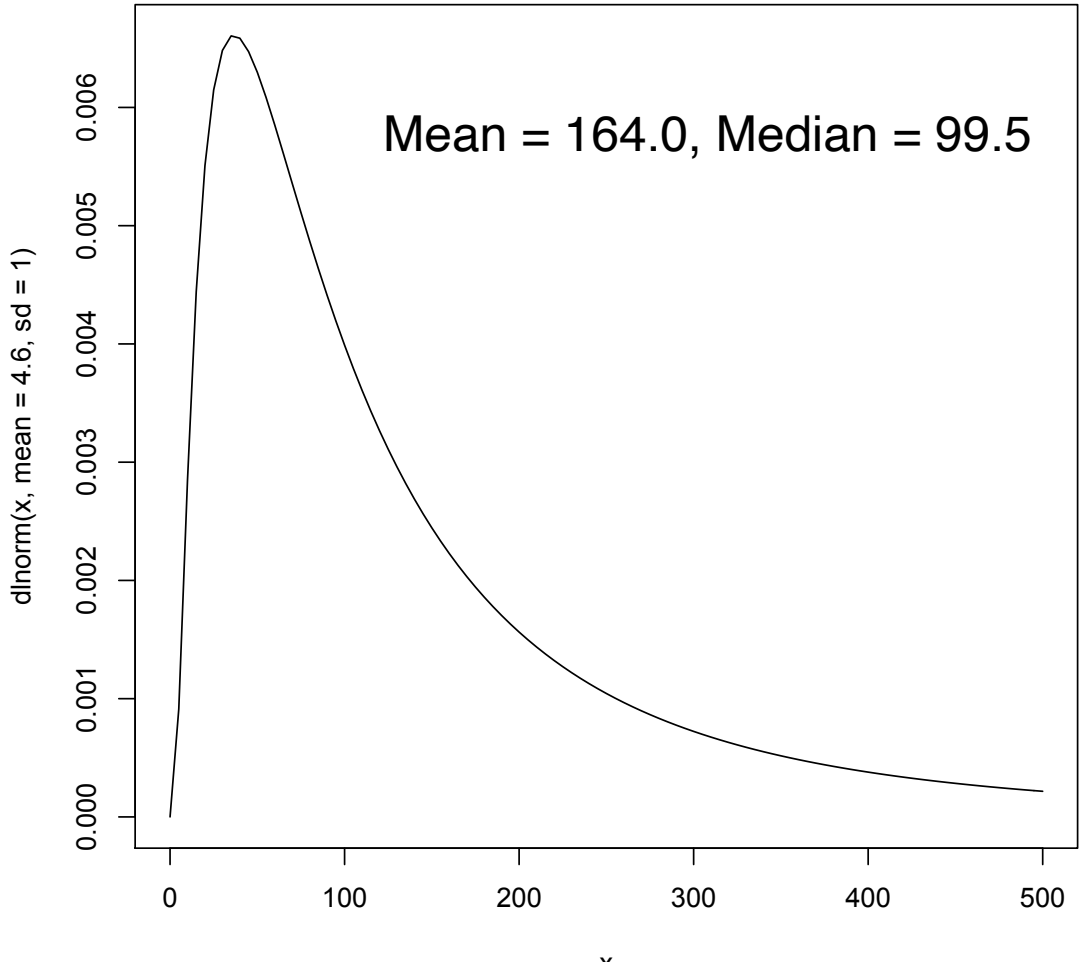
```
prop.test(c(25, 20), c(9, 8))$conf.int
```

The success rate of the 2nd player is higher by a

14%, 95% CI [-19%, 47%]

Lognormal distributions

Remember this example



R program

```r
# Parameters of the log-normal distribution
m <- 4.6   # Mean of log-transformed values (normal distribution)
sd <- 1    # SD of log-transformed values (normal distribution)

# True population mean (mean of the original skewed distribution)
M <- exp(m + sd^2/2)

n <- 12  # Size of the samples

N <- 10000 # Number of sampling repetitions
count <- 0 # Counter of how many times the CI fails

for(i in 1:N){
    sample <- rlnorm(n, meanlog = m, sdlog = sd)
    ci <- t.test(sample)$conf.int

    if(ci[2] < M || ci[1] > M) count <- count + 1
}

cat("Average number of failures = ", count/N, "\n")
cat("Coverage probability = ", (1 - count/N)*100, "%\n")
```

Lognormal distributions

1. Data values are first transformed to a logarithmic scale
2. CIs are computed over log-transformed values
3. CIs are then transformed back to the original scale

   Careful: The interpretation of a back-transformed 95% CIs of the mean is not the same. Those are not 95% CIs of the mean any more!!!

Simple math with logarithms

$$log_b(x) = a \iff b^a = x$$

$$log_b(1) = 0 \iff b^0 = 1$$

$$log_b(b) = 1 \iff b^1 = b$$

$$log_b(x) + log_b(y) = log_b(xy)$$

$$log_b(x) - log_b(y) = log_b \frac{x}{y}$$

Means under logarithmic transforms

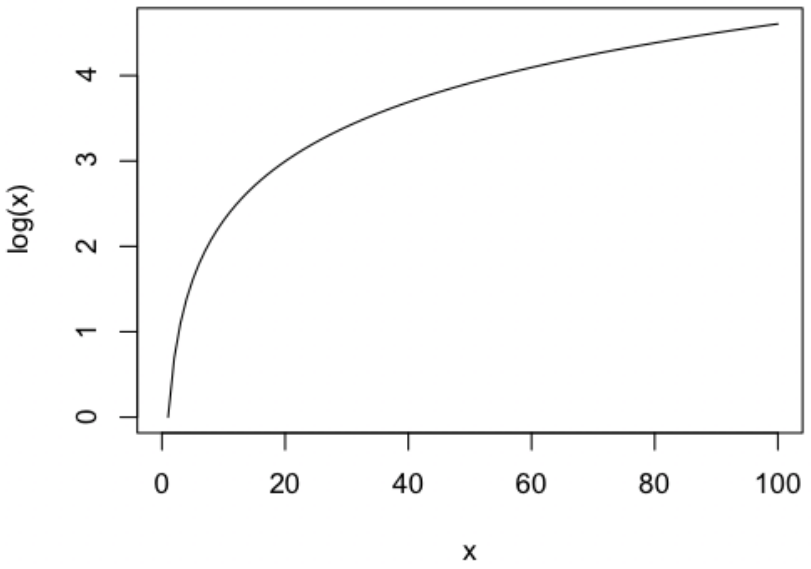
```
> sample <- c(4, 8, 3, 4, 6, 5, 15, 9, 5)

> mean(log(sample))

[1] 1.76295

> log(mean(sample))

[1] 1.880313
```

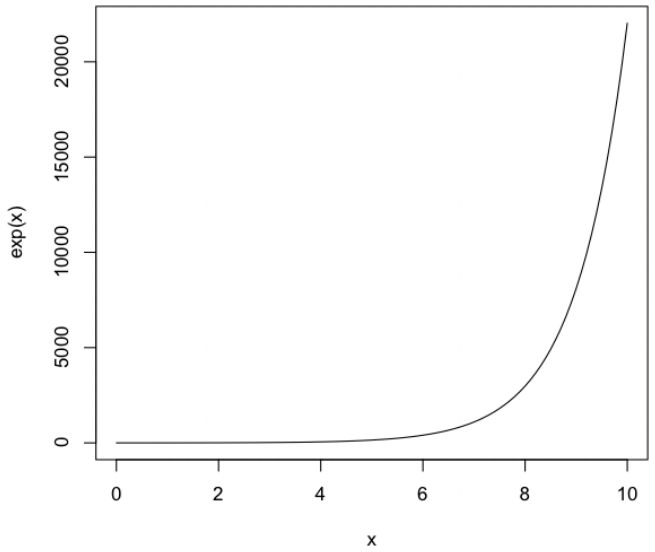
Means are not preserved under such non-linear transformations

What about medians?

The median is the middle value of a sample.

As long as the transformation function is *“monotonic”* (either increasing or decreasing), the order of values in a sample is preserved. Thus, the order of the middle value (the median) is also preserved.





What about medians?

```
> sample <- c(4,8,3,4,6,5,15,9,5)

> median(log(sample))

[1] 1.609438

> log(median(sample))

[1] 1.609438
```

Results would be slightly different if the sample had an even number of items as the median would be the mean of the two middle values.

Example

The following data show mean task completion times (in ms) of 10 participants for two selection techniques A and B. This is a repeated-measures (within-participants) design.

	P1	P2	P3	P4	P5	P6	P7	P8	P9	P10
Tech A	530	600	556	480	578	532	740	590	612	679
Tech B	511	552	430	455	610	520	731	483	610	539

The research team wants to compare their performance but suspects that task-completion times follow a skewed **log-normal** distribution.

Example

1. We log-transform the data, using natural logarithms.

	P1	P2	P3	P4	P5	P6	P7	P8	P9	P10
Tech A	6.27	6.39	6.32	6.17	6.36	6.28	6.61	6.38	6.42	6.52
Tech B	6.23	6.31	6.06	6.12	6.41	6.25	6.59	6.18	6.41	6.29

2. We compute 95% CIs by assuming that the sampling distribution of the means follow *t* distributions:

Tech A: 95% CI = [6.29, 6.46]

Tech B: 95% CI = [6.18, 6.40]

Since t-distributions are symmetric, means and medians coincide. Thus, we can also treat these intervals as **95% CIs of the median.**

Example

3. We then transform the CIs back to their original scale *(ms)* by using the inverse transformation $f(x) = e^x$:

Tech A: 95% CI = [535 ms, 640 ms]

Tech B: 95% CI = [481 ms, 602 ms]

Those intervals should now be interpreted as 95% CIs of the median (NOT of the mean).

Example

What about the time difference between the two techniques?

In logarithmic scale: 95% CI = [0.008, 0.161]

But if we now know apply the inverse transform $f(x) = e^x$, we get the following:

95% CI = [1.008, 1.175]

Clearly, these values do not represent milliseconds!

Remember: Differences in logarithmic scales correspond to ratios in the original scale.

$$log_b(x) - log_b(y) = log_b \frac{x}{y}$$

$$\Longrightarrow b^{(log_b(x) - log_b(y))} = \frac{x}{y}$$

if *b = e:* $e^{(ln(x) - ln(y))} = \frac{x}{y}$

Example

We interpret the results as follows:

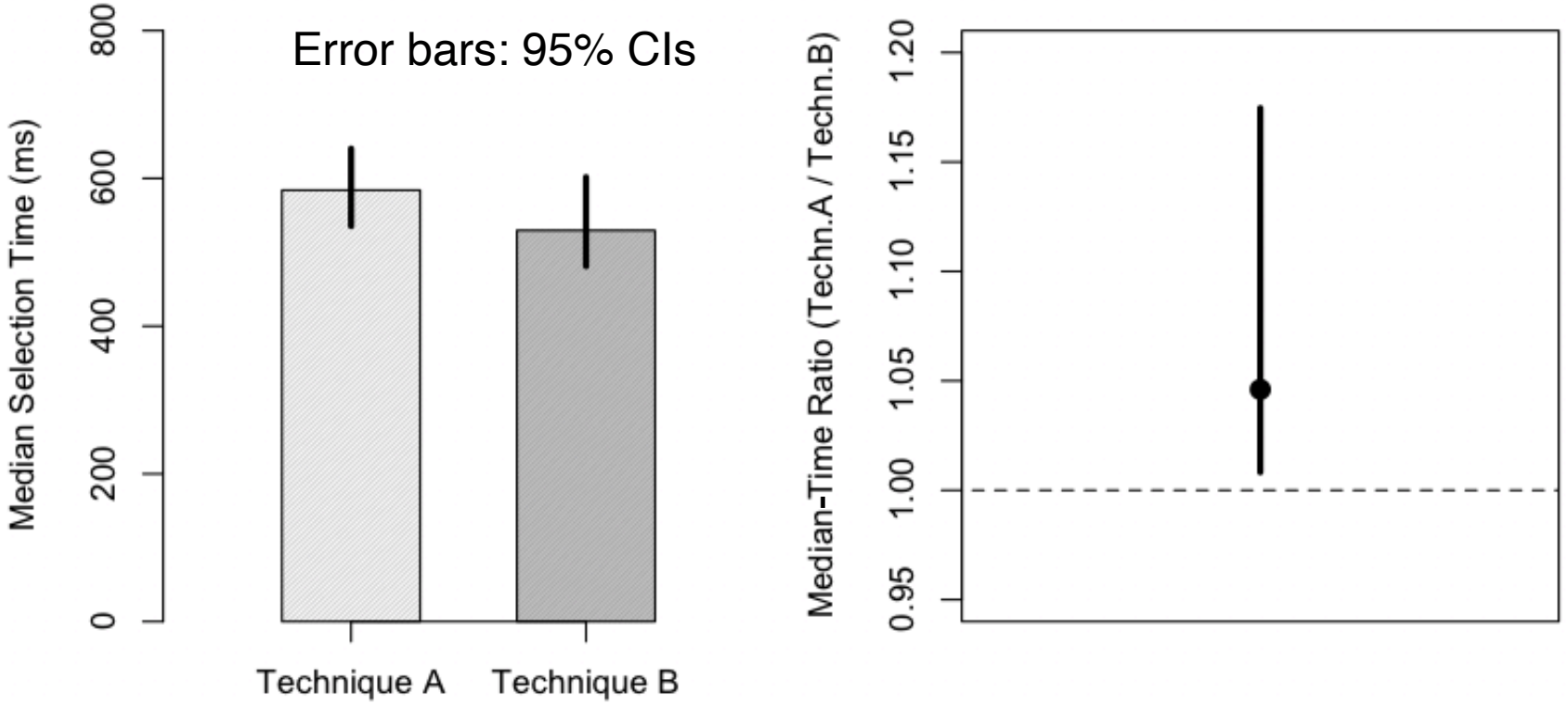
> The median selection time of Technique A is **109.4%, 95% CI [100.8%, 117.5%]** the median selection time of Technique B.

Or:

> The median selection time of Technique A is **9.4%, 95% CI [0.8%, 17.5%] higher** than the median selection time of Technique B.

Example

Graphing the results: choose your axes carefully!



Attention: If the parameters of interest are the means rather than the median, this approach is not appropriate.

Working with ratios

Interpreting results in terms of ratios (rather than differences) has several advantages: there are no units, and comparisons are based on relative values.

However, interpretation depends on the specific application context. Unfortunately, many people are not familiar with this approach. You may need to further justify it. Provide a clear interpretation of your results.

Log-normal distributions and means

There are several methods (Cox, modified Cox, etc.) that allow for constructing CIs for log-normal distributions when the parameter of interest is the mean:

http://ww2.amstat.org/publications/jse/v13n1/olsson.html

But they are out of the scope of this course.

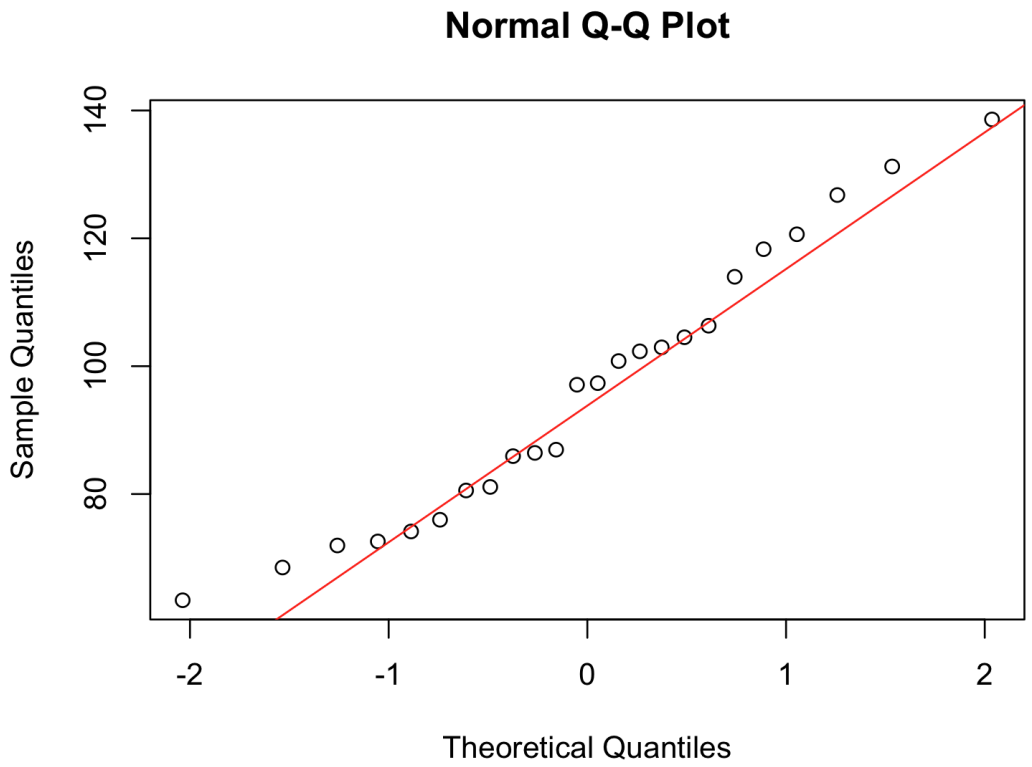
Note

If distributions are not too skewed and the size of the sample is not to small, assuming normality may still provide satisfying results. This is particularly the case for CIs of differences, since distributions of differences tend to be symmetric.

Exercise: Assess the **coverage probability** (percentage of times that the CI includes the true mean) of 95% CIs of the mean difference of samples drawn from skewed log-normal distributions, when we assume normality (i.e., if we do not apply any data transformation).

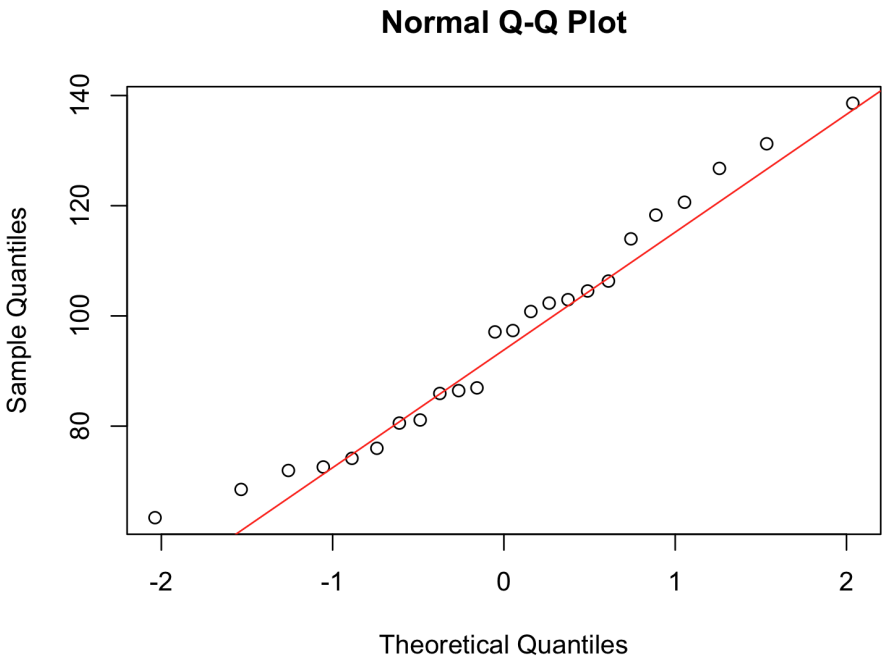
Quantile-Quantile (Q-Q) plots

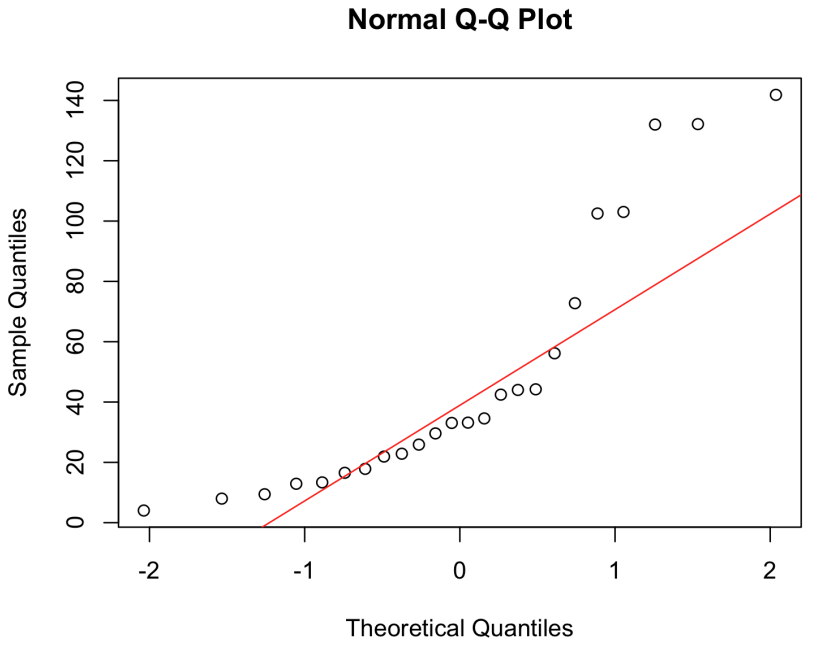
A useful type of plot for graphically inspecting how close the distribution of a sample is to normal.



See: https://www.lri.fr/~fanis/courses/Stats2019/lectures/distributions.html

Quantile-Quantile (Q-Q) plots






```
> x <- rnorm(24, mean = 100, sd = 20)
> qqnorm(x)
> qqline(x, col = "red")
```

```
> x <- rlnorm(24, meanlog = 3.6, sdlog = 0.8)
> qqnorm(x)
> qqline(x, col = "red")
```

Try the above code several times and check the results. Also try with smaller and with larger samples. How stable is the trend you observe in each case?

Rank transformations

We replace observations by their ranks. For example, the values (12.5, 6.8, 8.0, 11.2) become (4, 1, 2, 3)

They are often used when assumptions are violated (common choice for the analysis of questionnaire data).

Their interpretation can be tricky. They are not appropriate if the goal is to obtain confidence intervals or build a predictive model.

But they can be useful if the only goal is to construct a **significance test**.

From confidence intervals to significance tests

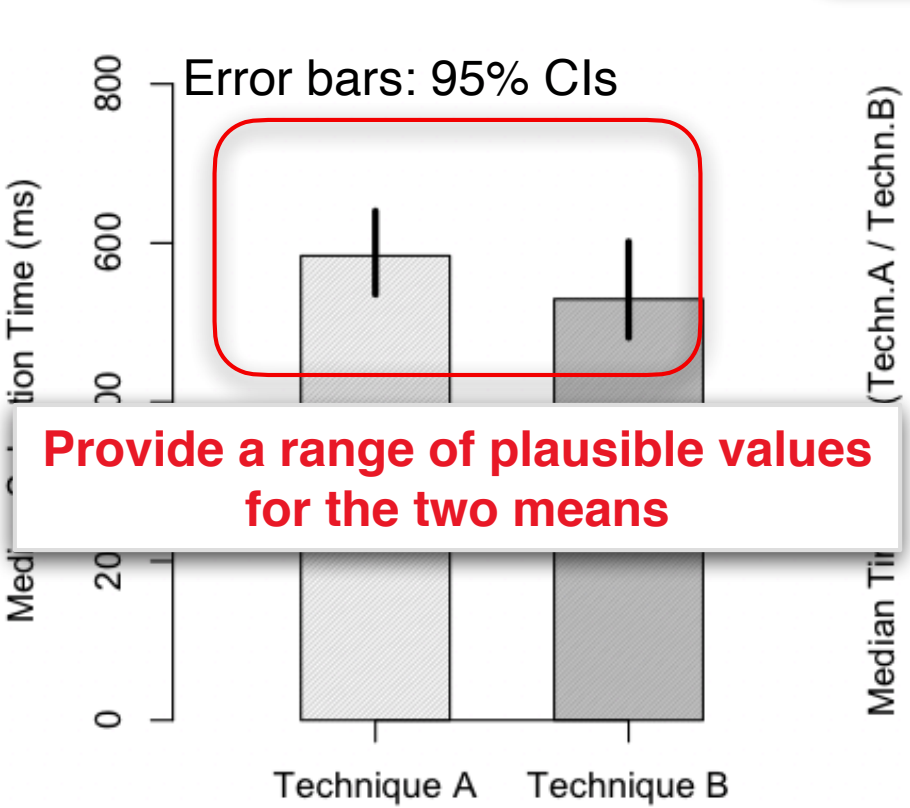
Uses of confidence intervals

1. To provide **an estimate** of plausible values that a population parameter may take.
2. To support **formal inference** about a parameter

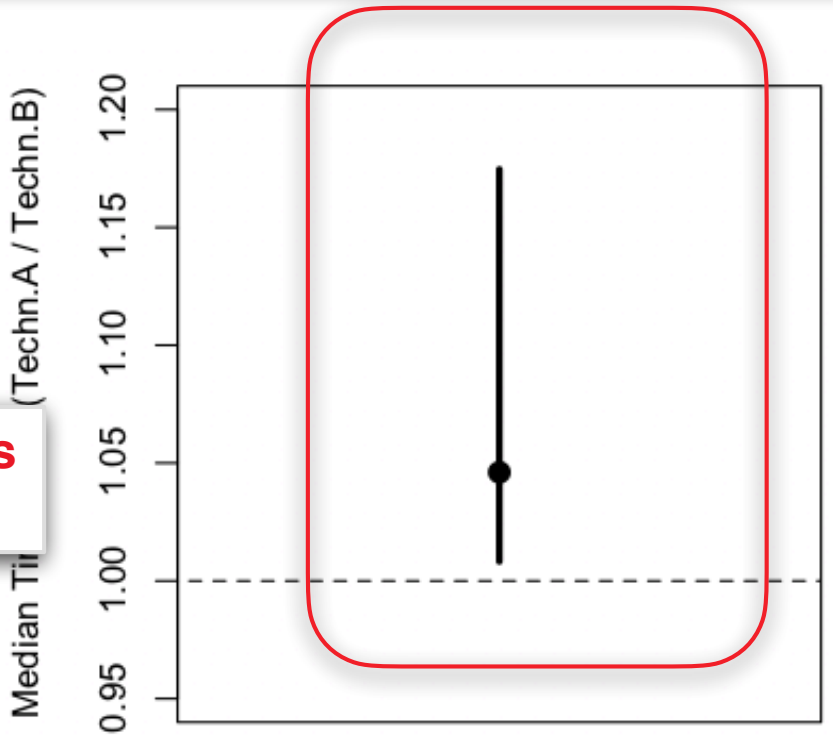
Example

Let's see these results:

It allows to conclude that "Technique A is slower than Technique B"



Provide a range of plausible values for the two means



Significance testing

Formal inference with a confidence interval is a form of **significance testing**.

A **significance test** involves explicitly or implicitly setting up a **hypothesis** about the value of a population parameter:

Hypothesis example:

"Median selection time of Technique A is equal to the median selection time of Technique B."

or

"The ratio of the median selection times of Technique A and Technique B is 1"

Null hypothesis

Such hypotheses that make a statement about a hypothetical value of a population parameter (a mean or median, a mean difference, etc.) are known as **null hypotheses.**

The goal of an experiment is commonly **to provide evidence against a null hypothesis.**

Examples of null hypotheses

"*The mean height of men is equal to the mean height of women.*"

“The mean IQ score of adults lacking enough sleep is 100.”

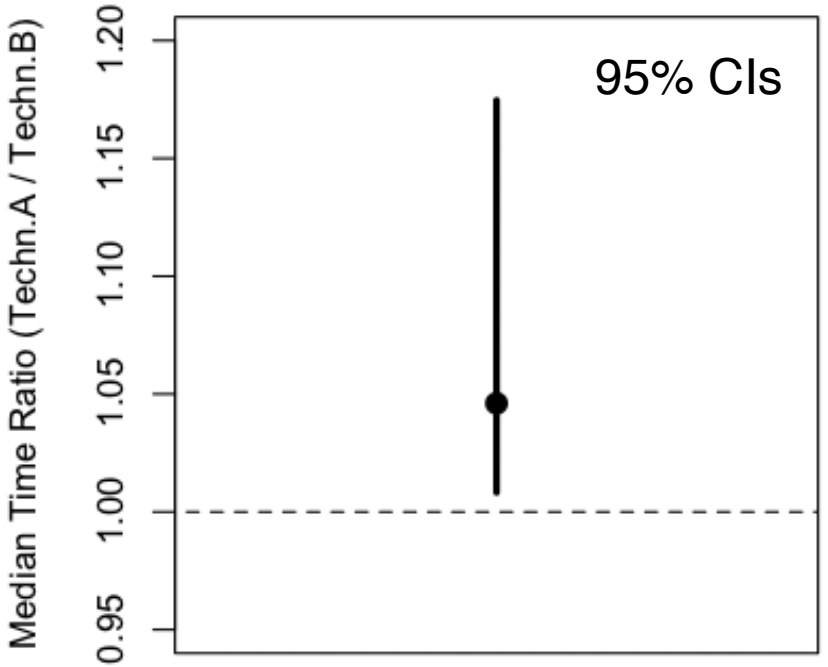
“Mean selection time with a mouse is equally fast to mean selection time with a trackpad.”

Rejecting a null hypothesis

If the *C%* confidence interval excludes the hypothesized population value, then the hypothesis is rejected (*C%* confidence level).

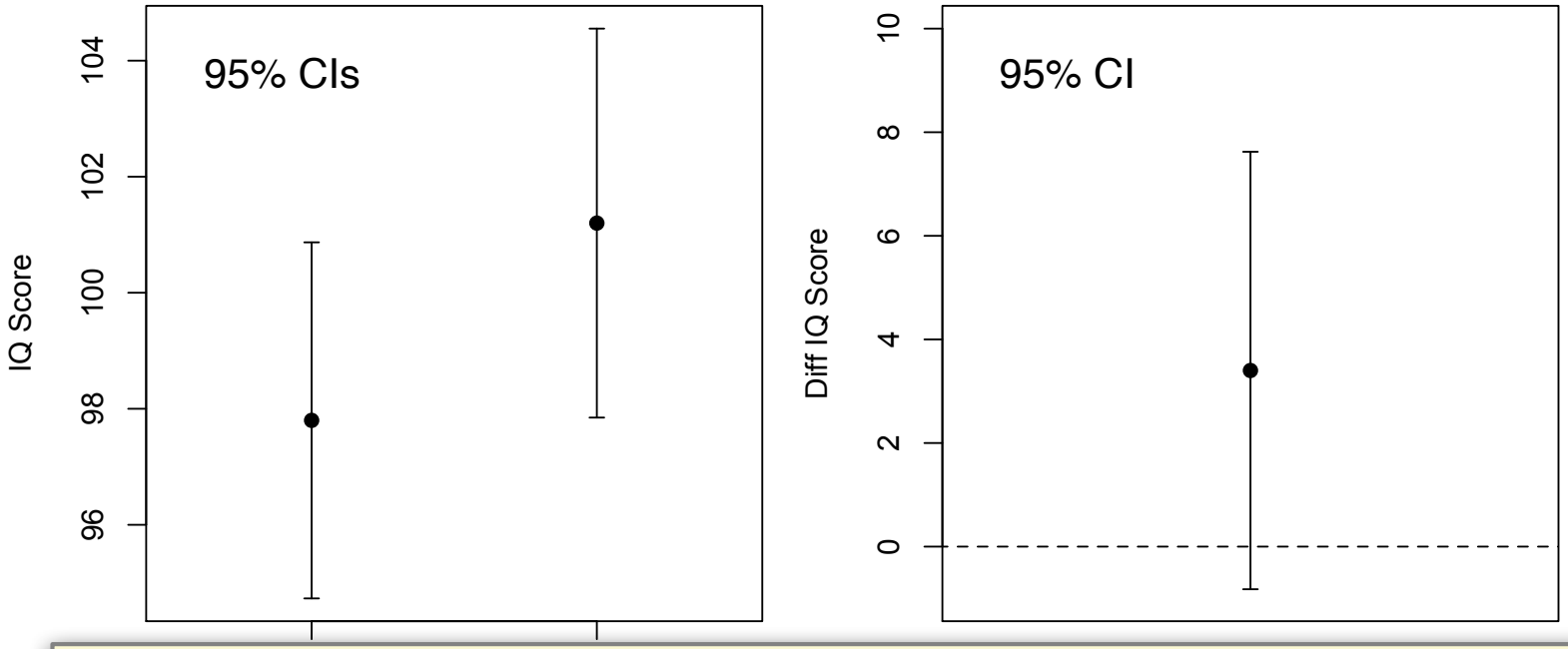
Example

Here, we reject the null hypothesis (95% confidence level) because the 95% CI does not include the value *1*.



Example

What about here? Can we reject the null hypothesis concerning the mean difference (in IQ scores) between the two groups?



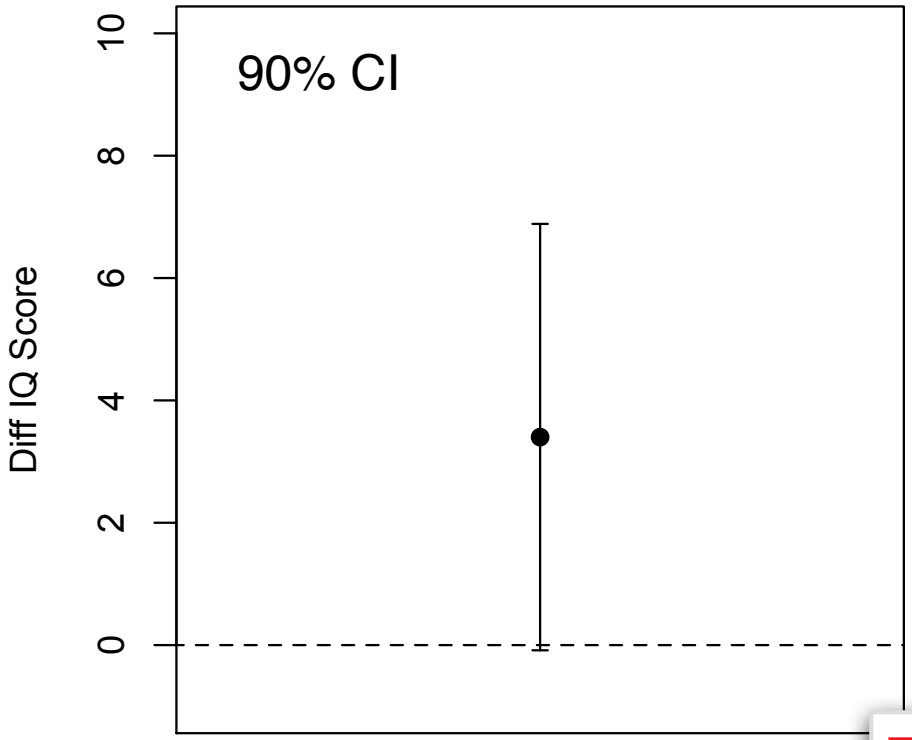
We cannot reject the null hypothesis.

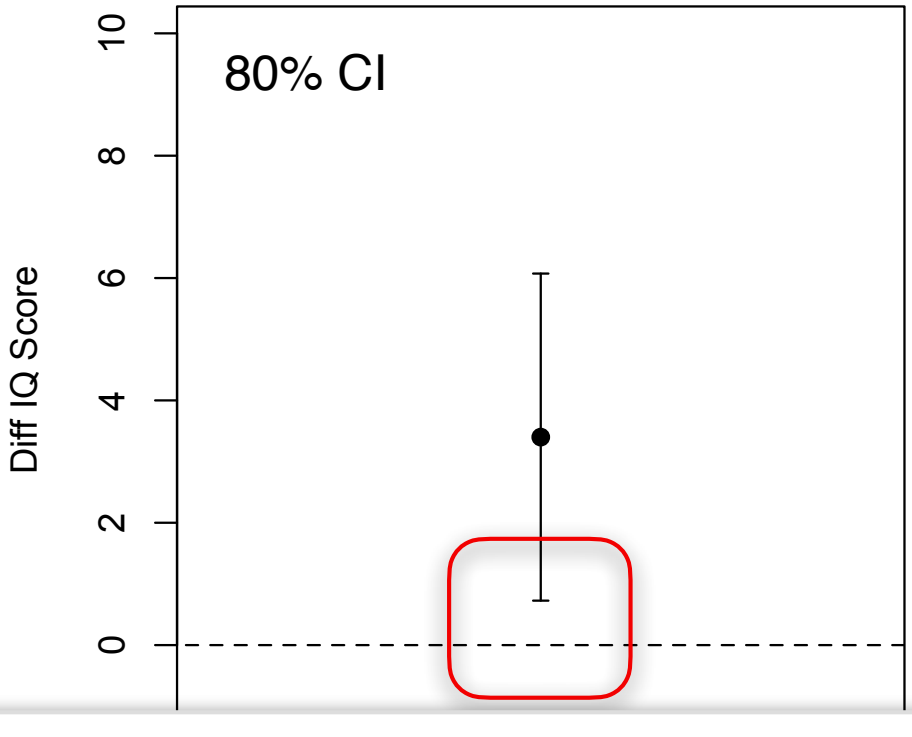
But this does not necessarily mean that the difference is zero!

We just do not have enough evidence to say anything more.

Example

And if we reduce the confidence level?





For this level, we can reject the null hypothesis!
But is a 80% confidence level high enough?

95% confidence level

It is commonly used as the threshold for rejecting a null hypothesis.

It is not a magic number and there is no reason why a different level (e.g., 92% or 97%) is not used.

But its use reflects a long tradition in science.

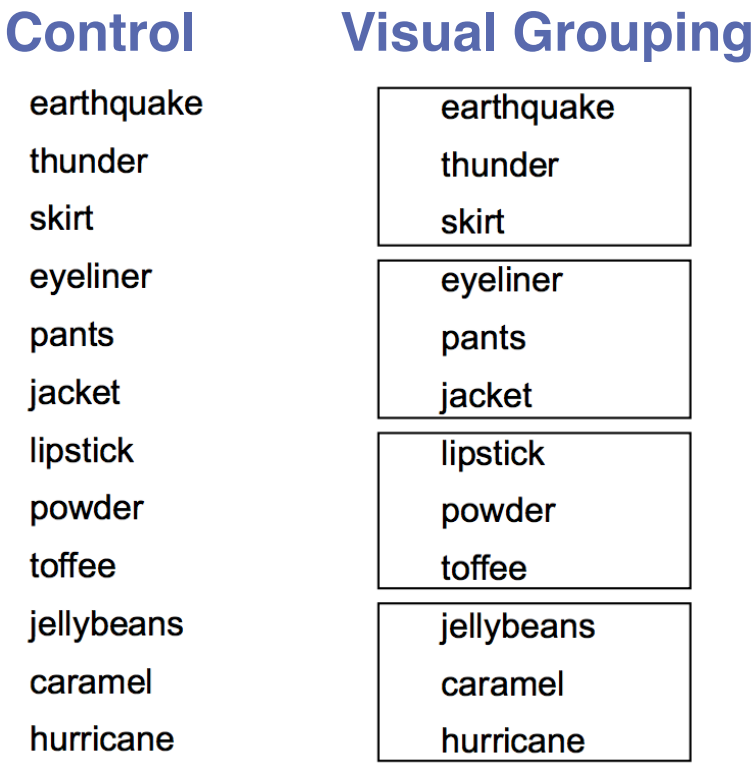
Null vs. Alternative hypothesis

Our goal is usually to find enough statistical evidence to **reject the null hypothesis (H_0)** in order to establish an **alternative hypothesis (H_1).**

The alternative hypothesis is the hypothesis of interest, i.e., what the researcher actually seeks to show by rejecting the null hypothesis.

Example

An HCI researcher studies whether visual grouping in menus help users locate menu items faster.



(Brumby & Zhuang, 2015)

Example

Null Hypothesis:

H_0: *“Mean selection time is the same for Control and Visual Grouping.”*

Alternative Hypothesis:

H_1: *“Mean selection is faster with Visual Grouping than with Control.”*

Significance tests

Null Hypothesis Significance Testing **(NHST)** can be supported by confidence intervals and **significance tests**.

Significance tests is a very common research tool. However, they have many limitations. They are also very frequently overused or misused.

Note that significance tests rely on similar assumptions as the ones that we have already discussed for constructing confidence intervals.

p value

The result of a significance test is a probability value *p,* which is commonly known as the ***p* value.**

Given an observed value of a statistic (e.g., the mean) of a sample, the p value gives **the probability that**:

> **if the null hypothesis H_0 was true, then a random sample of the same size would result in a value for the statistic that is equal or more extreme than the observed value.**

Example

An experiment studies the IQ scores of people lacking enough sleep.

H_1: The mean IQ score of people lacking sleep is lower than 100.

$H_{0:}$ The mean IQ score of people lacking sleep is equal to 100.

Example

Results from a sample of 15 participants are as follows:

94, 91, 96, 100, 103, 88, 98, 103, 87, 93, 97, 105, 99, 91, 92

The mean IQ score of the above sample is **M = 95.8**

The researchers conduct a significance test and find that ***p = .006***

Interpretation: *If the null hypothesis was true (M = 100)*, then the probability to draw a random sample of 15 people and find that the mean IQ score is *equal to or lower than 95.8* is 0.6%.

Interpreting the p value

Interpretation: *If the null hypothesis was true (M = 100)*, then the probability to draw a random sample of 15 people and find that the mean IQ score is *equal to or lower than 95.8* is 0.6%.

Notice that this is NOT the probability that the researcher's hypothesis H_1 is false.

An interpretation *"the probability that M is equal to 100 is 0.6%" is incorrect!*

Threshold for rejecting H_0

By tradition, the null hypothesis (H_0) is rejected when the p value is lower than **$\alpha = .05$**.

This alpha (α) is the same as the one we discussed for C% confidence intervals, where the confidence level is:

$$C = 100(1 - \alpha)$$

(Clearly, there is a correspondence between significance tests and confidence intervals.)

Back to our example

Results from a sample of 15 participants are as follows:

94, 91, 96, 100, 103, 88, 98, 103, 87, 93, 97, 105, 99, 91, 92

The mean IQ score of the above sample is **M = 95.8**

The researchers conduct a significance test and find that ***p = .006***

Since $p < \alpha$ ($\alpha = .05$), the **null hypothesis is rejected.**

The researchers conclude that the lack of sleep results in **statistically significantly** lower IQ scores.

Statistical significance

Since $p < \alpha$ ($\alpha = .05$), the **null hypothesis is rejected.**

The researchers conclude that the lack of sleep results in **statistically significantly** lower IQ scores.

It is common to say that such an experiment has resulted in a **“statistically significant”** result.

Interpreting statistical significance

The term **“significant”** can be very misleading. Statistical significance does not refer to the actual significance of the result!

A significance test may often characterize a very tiny (or practically insignificant) difference as statistically significant!

Make sure that you refer to “statistical significance” when you characterize the results of a study. Don’t use the term “significant” alone.

In the following lecture, I will explain why even stopping using the term “statistical significant” altogether may be a good practice.

Back to our example

Imagine that another researcher conducts the same experiment with 12 participants and finds the following IQ scores:

103, 97, 110, 99, 105, 96, 111, 108, 98, 106, 104, 102

The mean IQ score of the above sample is **M = 103.25**.

This mean is higher than 100, so clearly, the data do not support H_1.

Can we still test the null hypothesis (H_0) to check whether the lack of sleep leads to statistically significantly **higher IQ scores?**

Two-sided vs. one-sided tests

Answer: It depends on what type of significance tests the researcher uses.

One-sided (or one-directional) tests do not allow for that. If the hypothesized direction is not supported, the hypothesis cannot be rejected.

However, the common practice is to use **two-sided (or two-directional)** tests. In this case, the hypothesis can be rejected in both directions.

Back to our example

Imagine that another researcher conducts the same experiment with 12 participants and finds the following IQ scores:

103, 97, 110, 99, 105, 96, 111, 108, 98, 106, 104, 102

The mean IQ score of the above sample is **M = 103.25**.

A two-sided significance test results in ***p* = .047**

Interpretation: *If the null hypothesis was true (M = 100)*, then the probability to draw a random sample of 12 people and find that the absolute difference between the observed mean and 100 is equal to or higher than **ΔM = 3.25** is 4.7%.

p values are now interpreted differently

Back to our example

Imagine that another researcher conducts the same experiment with 12 participants and finds the following IQ scores:

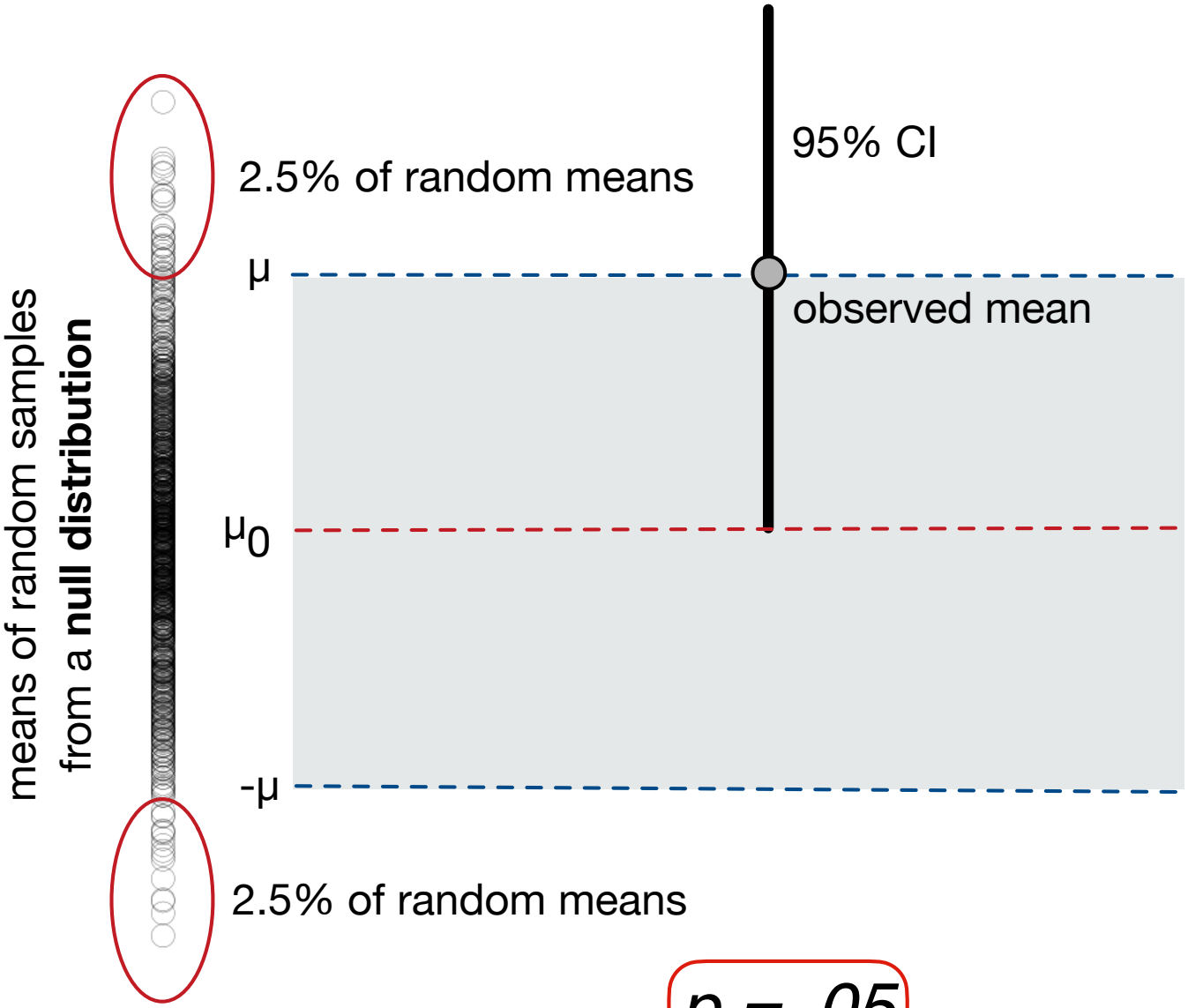
103, 97, 110, 99, 105, 96, 111, 108, 98, 106, 104, 102

The mean IQ score of the above sample is **M = 103.25**.

A two-sided significance test results in ***p = .047***

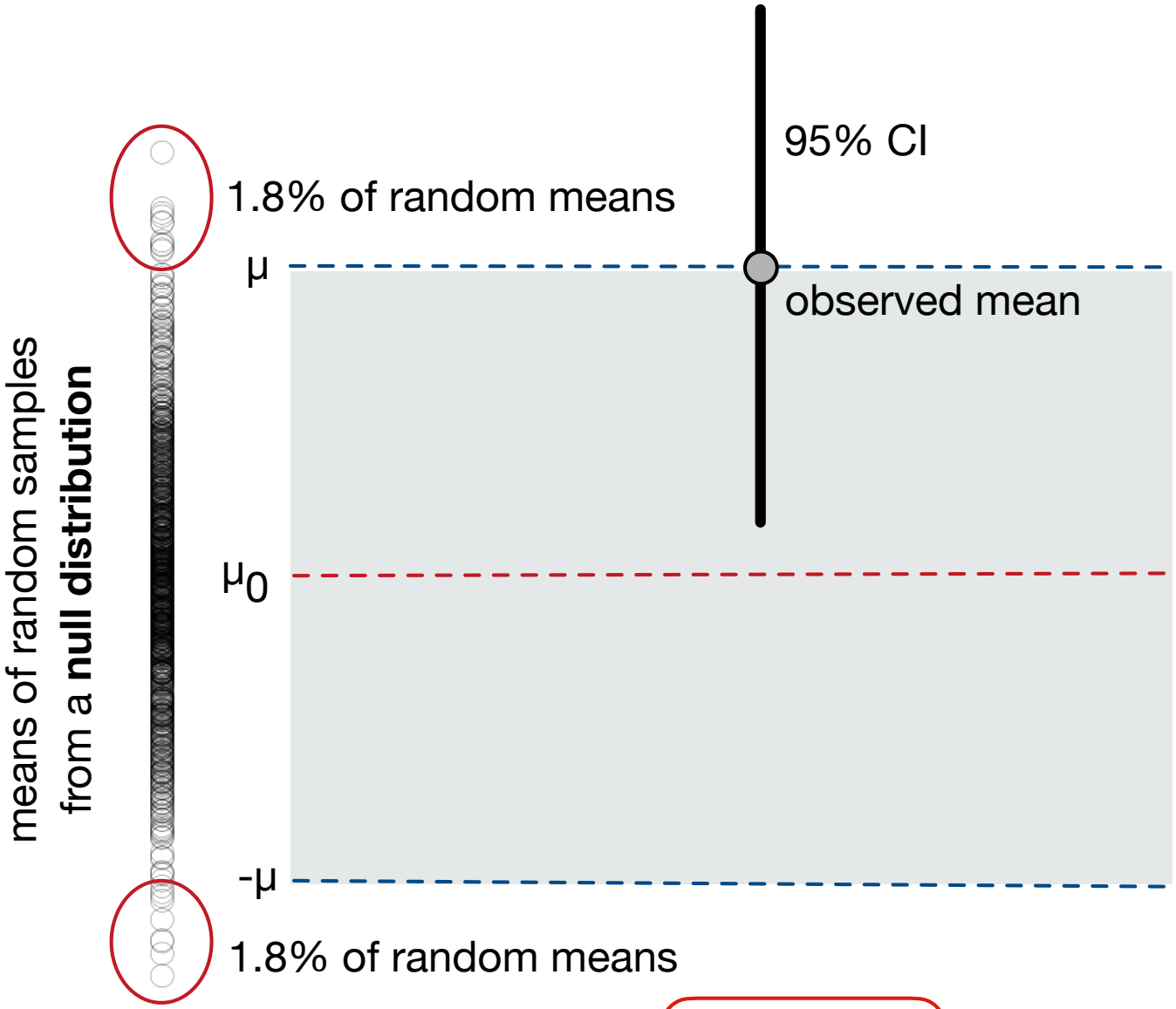
Conclusion: The null hypothesis is rejected (α = .05). The researchers conclude that the lack of sleep results in **statistically significantly** higher IQ scores.

p values vs. confidence intervals



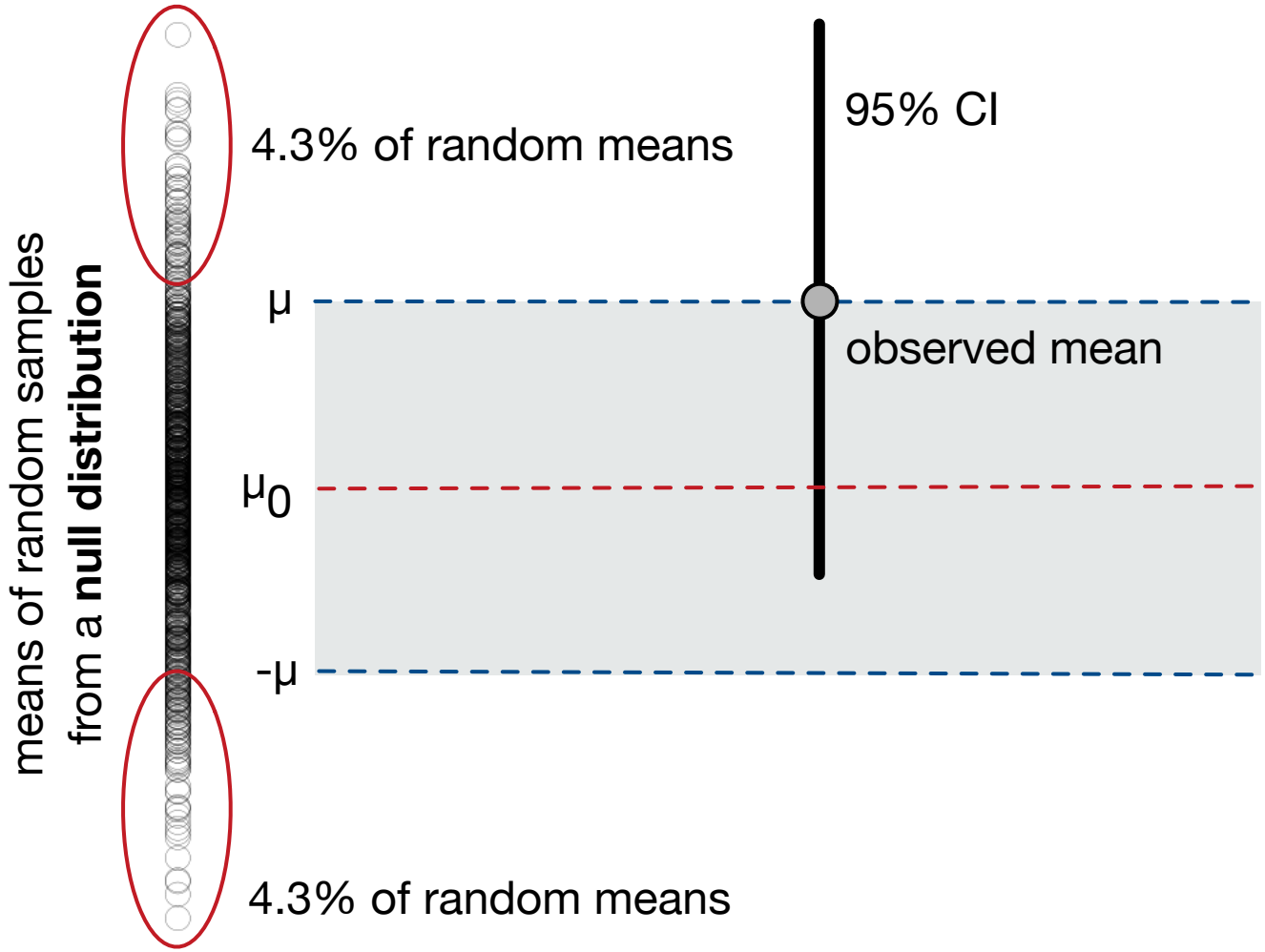
p = .05

p values vs. confidence intervals



p = .036

p values vs. confidence intervals



p = .086

Calculating *p*

The *p* value is the probability of obtaining a statistic *as extreme or more extreme than the one observed* if the null hypothesis was true.

When data are sampled from a known distribution, an exact *p* can be calculated.

If the distribution is unknown, it may be possible to estimate *p*.

Normal distributions

If the sampling distribution of the statistic is normal, we will use the standard normal distribution z to derive the p value

Example

An experiment studies the IQ scores of people lacking enough sleep.

$H_0: \mu = 100$ and $H_1: \mu < 100$ **(one-sided)**

or

$H_0: \mu = 100$ and $H_1: \mu \neq 100$ **(two-sided)**

Example

Results from a sample of 15 participants are as follows:

90, 91, 93, 100, 101, 88, 98, 100, 87, 83, 97, 105, 99, 91, 81

The mean IQ score of the above sample is **M = 93.6.**

Is this value **statistically significantly different** than 100?

Creating the test statistic

We assume that the population standard deviation is known and equal to SD = 15. Then, the standard error of the mean is:

$$\sigma_{\hat{\mu}} = \frac{\sigma}{\sqrt{n}} = \frac{15}{\sqrt{15}} = 3.88$$

Creating the test statistic

We assume that the population standard deviation is known and equal to SD = 15. Then, the standard error of the mean is:

$$\sigma_{\hat{\mu}} = \frac{\sigma}{\sqrt{n}} = \frac{15}{\sqrt{15}} = 3.88$$

The **test statistic** tests the **standardized** difference between the observed mean $\hat{\mu} = 93.6$ and $\mu_0 = 100$

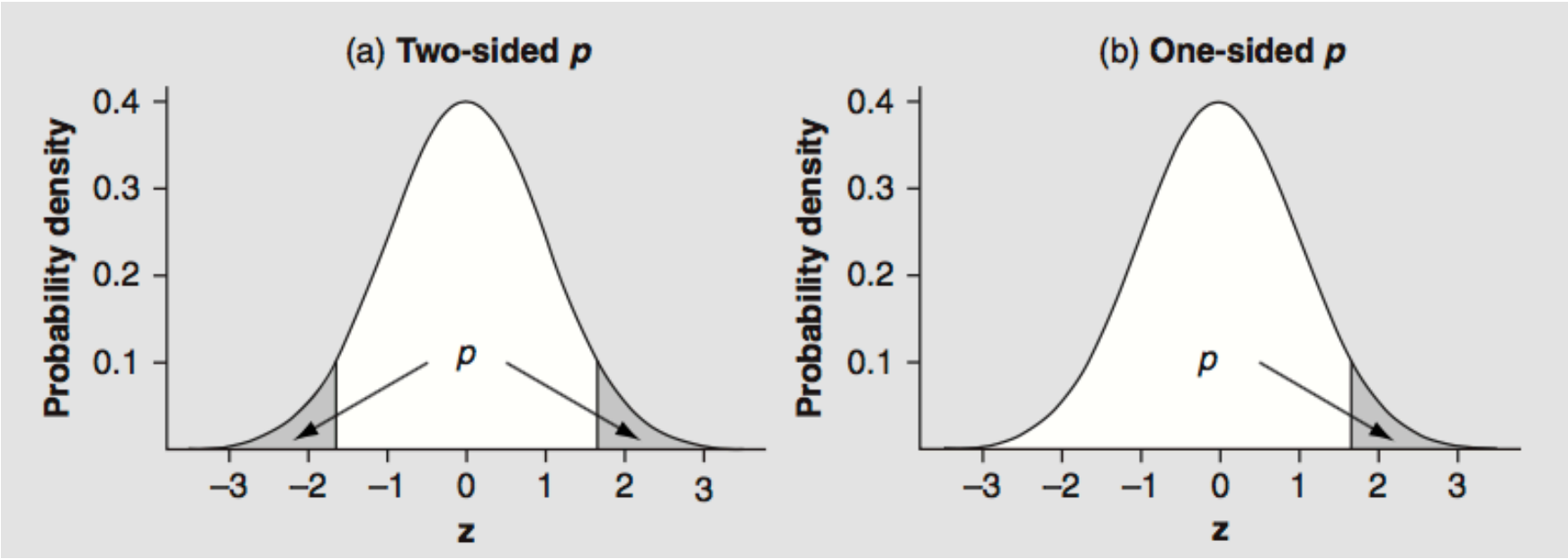
$$z = \frac{\hat{\mu} - \mu_0}{\sigma_{\hat{\mu}}} = \frac{93.6 - 100}{3.88} = -1.65$$

The *p* value is the probability of getting a z statistic as or more extreme than this value (given that H_0 is true)

Calculating the p value

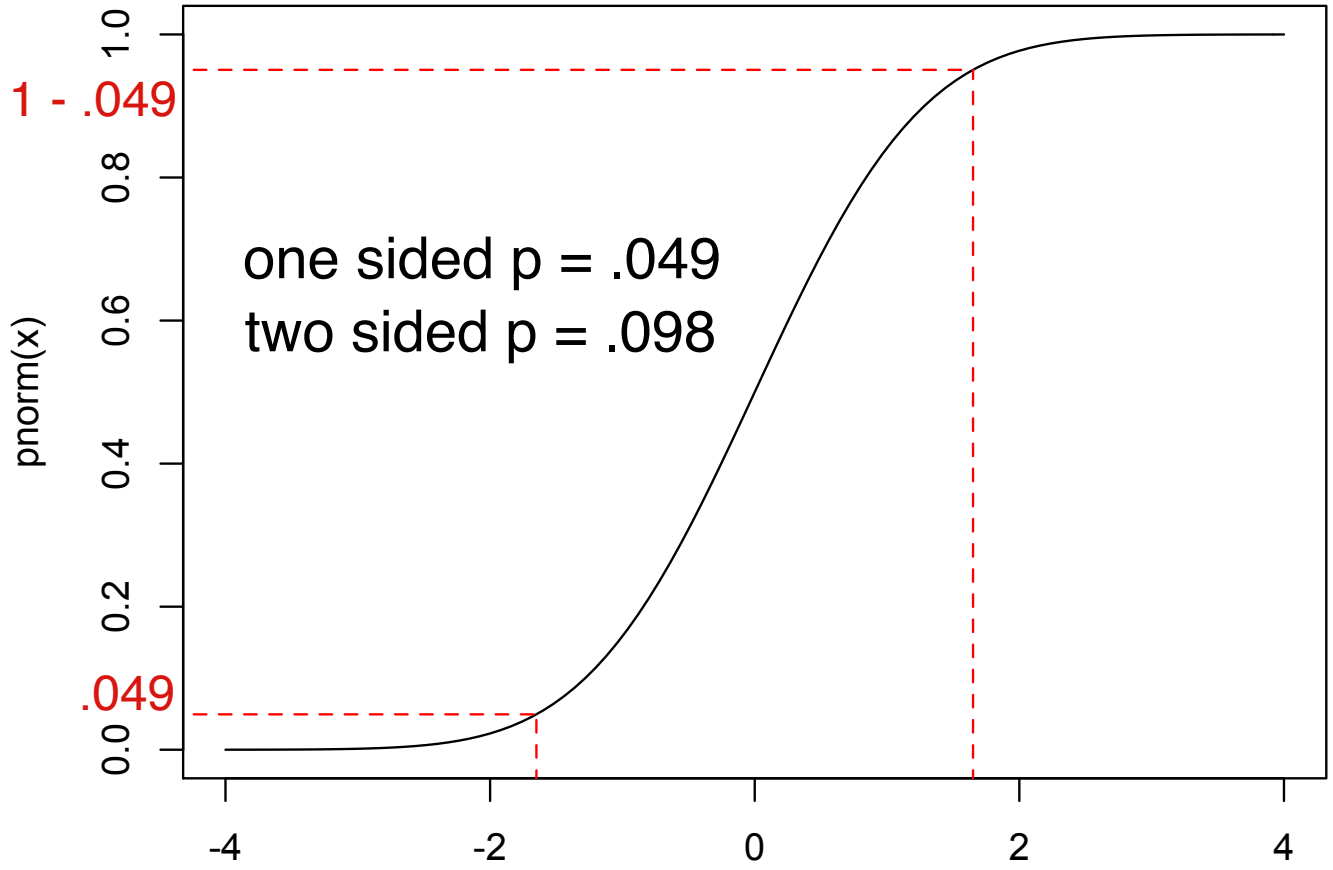
$$z = \frac{\hat{\mu} - \mu_0}{\sigma_{\hat{\mu}}} = \frac{93.6 - 100}{3.88} = -1.65$$

The *p* value is the probability of getting a z statistic as or more extreme than this value (given that H_0 is true)



Calculating the p value

To calculate the area in the distribution, we will work with the cumulative density probability function (cdf).



R code

```r
m0 <- 100 # mean for null hypothesis
sd <- 15 # We assume that the population sd is known

# These are the observed IQ scores
scores <- c(90, 91, 93, 100, 101, 88, 98, 100, 87, 83, 97, 105,
99, 91, 81)
m <- mean(scores)

# I calculate the standard error of the mean
se <- sd / sqrt(length(scores))

# I calculate the z statistic, i.e., the standardized mean
difference
z = (m - m0) / se

# This is the one side p value
pvalue <- pnorm(z)

cat("one sided p-value =", pvalue, "\n")
cat("two sided p-value =", 2*pvalue, "\n")
```

Example: conclusions

If we assume that the standard deviation of the population is known (SD = 15):

We reject the null hypothesis ($\alpha = .05$) if we apply a one-sided significance test ($p = .049$).

But we cannot reject it if we apply a two-sided significance test ($p = .092$)

One-sided vs. two sided

“A common misunderstanding of directional testing is to believe that it should be employed whenever a researcher predicts an effect in a particular direction.

This is almost always bad practice. *A one-sided test should be employed only if the direction of an effect is already known or if any outcome in the non-predicted direction would be ignored.*

The crucial question is not whether you think an effect lies in a particular direction, but whether you are willing to declare an effect in the wrong direction non-significant (no matter how interesting or how important it is). The answer to such a question is usually ‘no’ and for this reason ***one-sided tests should typically be avoided.****“*

from Baguley

t tests

If the population variance is unknown, it is better to work with the *t* distribution.

…but we will continue after the Christmas break!