Experiment Design for Computer Sciences (01CH740)

Topic 05 - Statistical Inference III – Equality and Non-Normal testing

Claus Aranha
caranha@cs.tsukuba.ac.jp

University of Tsukuba, Department of Computer Sciences

May 29th, 2020

Version 2020.1 (Updated May 28, 2020)

Introduction

In this lecture, we deal with two special cases of Statistical Inference:

- **Equality Testing**: We saw last classes that failing to reject a null hypothesis does not imply that the null hypothesis is true. So how should we perform the analysis when we want to show that two quantities are equal?
- **Non-normal Data**: The tests discussed in lecture 3 and 4 assume that the sampling distribution is normal. What can we do when this assumption is breached?

Part I – Equality Testing

Testing equivalence

Introduction

The tests introduced in the preceding chapters deal with situations in which one is interested in detecting *differences* between a population parameter θ – e.g., a population mean μ or a difference between population means $(\mu_1 - \mu_2)$ – and its nominal value θ_0 under a null hypothesis;

Another useful class of experiments in engineering and science is one in which the experimenter is interesting in investigating *equivalence* (within a given margin of error), for instance:

- Conformity/compliance testing (industrial certification);
- Equivalence of effects (pharmaceutical industry);

Testing equivalence

Introduction

In principle, one could express this as a shift in focus from trying to establish whether a population parameter is different from a given reference to trying to determine whether it is equal to that reference.

In usual (two-sided) comparative studies, the alternative hypothesis (i.e., the one that presents novelty in relation to the current state of knowledge) is the one of difference between the parameters of interest - that is, unless there is strong evidence of differences, one cannot rule out the null hypothesis of equality;

Testing equivalence

Introduction

In equivalence testing, the situation is reversed: the (approximate) equality of two parameters is the novelty one hopes to establish. Consequently, the burden of proof shifts to providing evidence that there is no difference.

The term *equivalent* is not used strictly, but to mean the absence of practical differences - that is, any differences that might exist fall within an *equivalence margin* or *limit of practical significance* δ^*.

Using this approach, the equivalence of two parameters can be established if a sample provides enough evidence that the true difference is smaller than δ^* units.

Testing Non-inferiority

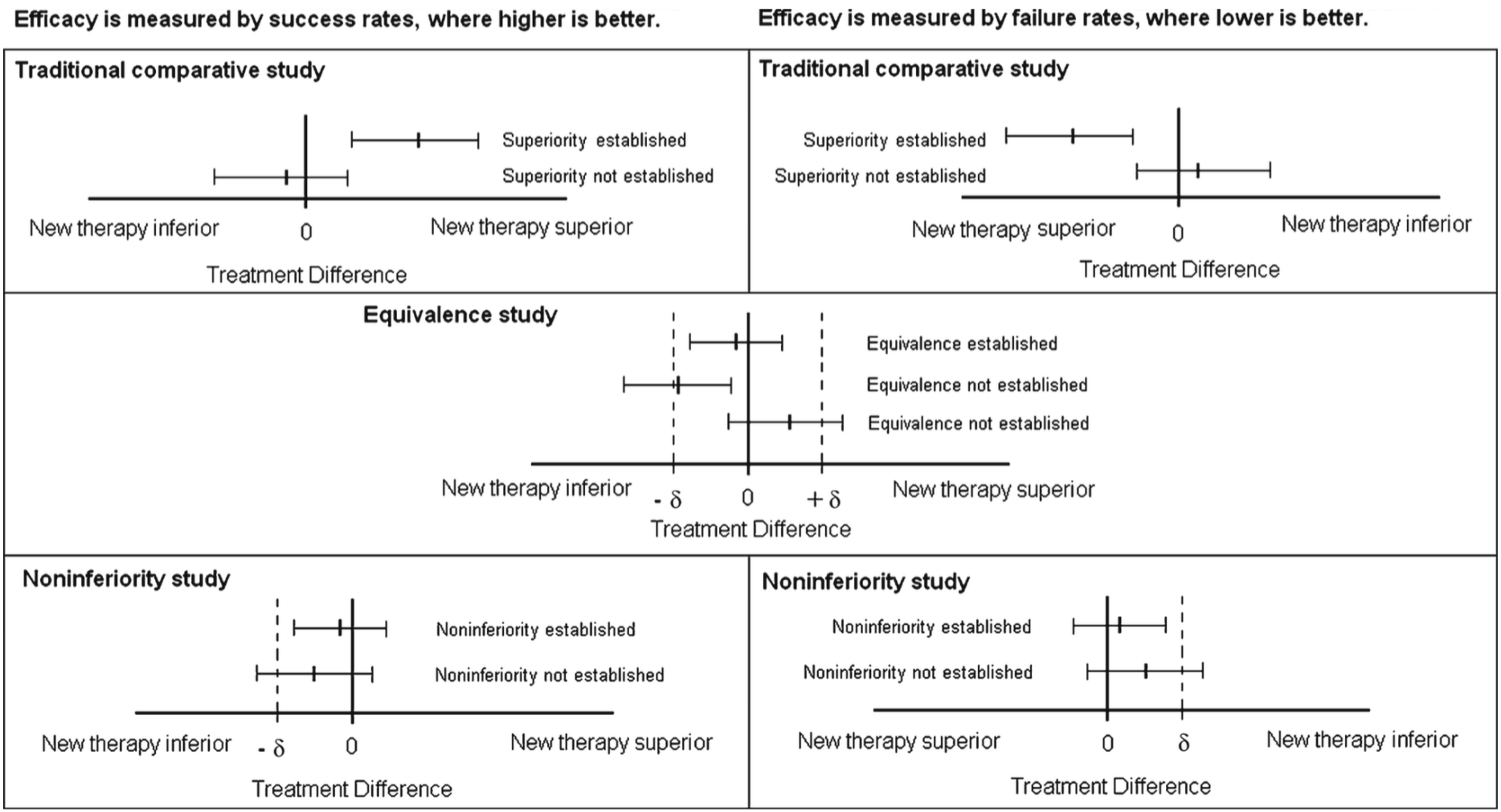
Definition

A similar concept to equivalence testing is the definition of non-inferiority of a given treatment/ process/ method in relation to another (e.g., a standard solution).

In non-inferiority tests, one can declare that a given process is not worse than a standard one only if enough evidence is provided to conclude that the performance of the proposed process is no more than δ^* units worse than that of the standard.

In the case of non-inferiority tests, one can in principle use a regular test of differences with a one-sided alternative (which would be equivalent to setting $\delta^* = 0$), or define the null hypothesis in a way that includes δ^* in its formulation.

Comparison of studies



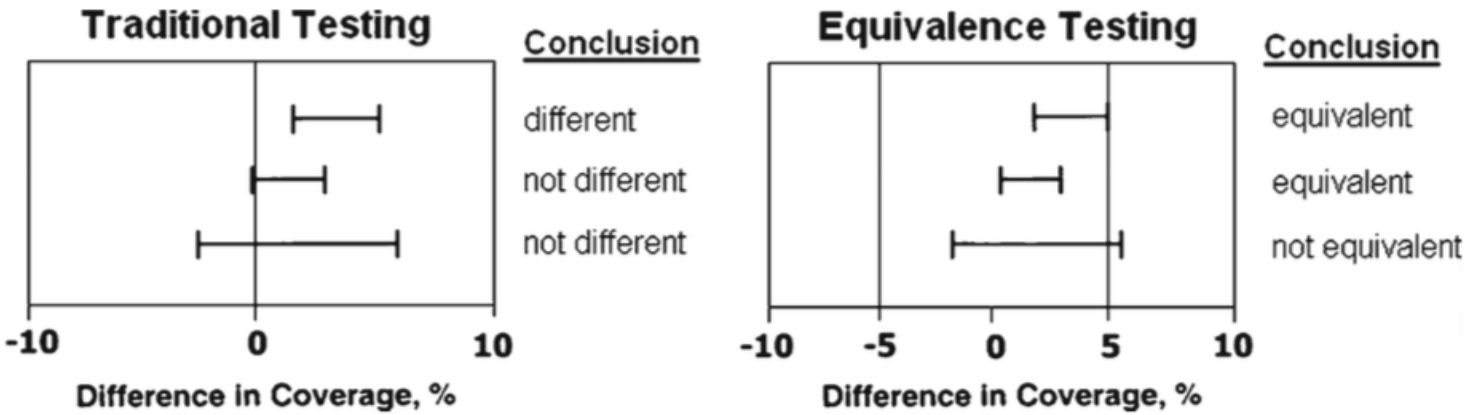
Testing Equivalence

Quick-and-dirty approach

A simple way of thinking about testing equivalence of two means is to observe confidence intervals instead of p-values:

> "*Equivalence can be established at the α significance level if a $(1-2\alpha)$-confidence interval for the difference between the two means is contained within a interval $\pm\delta^*$.*"

The difference between testing for differences and for equivalence can be easily illustrated using this approach:



Equivalence test for a single mean

Hypotheses

An equivalence test for a single population mean can be expressed by the hypotheses:

$$\begin{cases} H_0 : |\mu - \mu_0| = & \Delta\mu \geq \delta^* \\ H_1 : & \Delta\mu < \delta^* \end{cases}$$

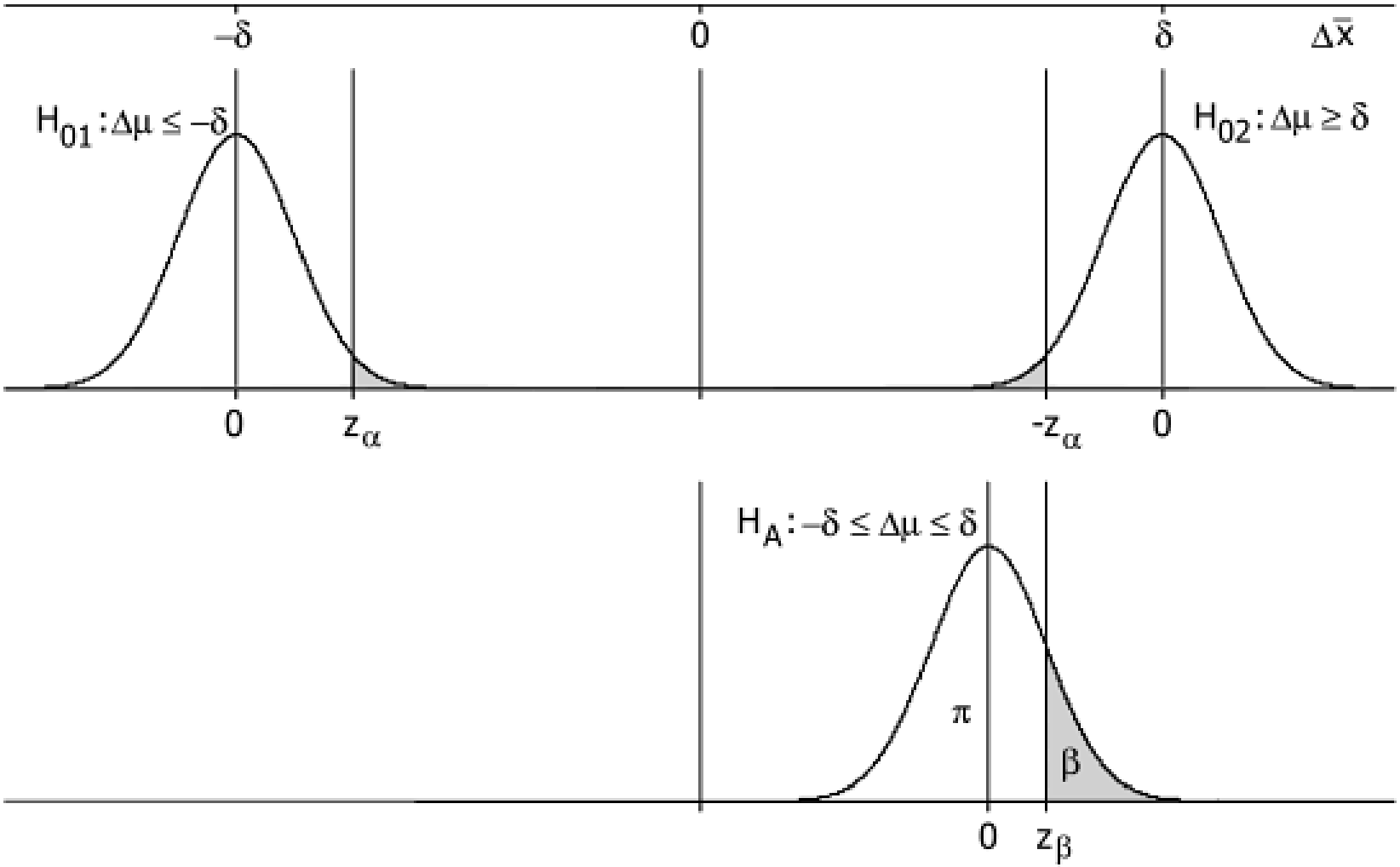
The most usual way of testing these hypotheses is the TOST (*two one-sided tests*) method. As the name suggests, two one-sided significance tests are constructed so that the desired statistical properties can be achieved. Using our standard notation:

$$\begin{cases} H_0^1 : & \Delta\mu = -\delta^* \\ H_1^1 : & \Delta\mu > -\delta^* \end{cases} \qquad \begin{cases} H_0^2 : & \Delta\mu = \delta^* \\ H_1^2 : & \Delta\mu < \delta^* \end{cases}$$

If both tests reject their respective H_0, then equivalence (within the equivalence margin δ^*) can be declared with significance level α.

Equivalence test for a single mean

Graphical interpretation



Equivalence of two means

Hypotheses

Analogously to the single sample test of equivalence, the hypotheses for testing the equivalence of two population means can be described as:

$$\begin{cases} H_0: & \mu_1 - \mu_2 \geq \delta^* \\ H_1: & \mu_1 - \mu_2 < \delta^* \end{cases}$$

$$\begin{cases} H_0^1: & \mu_1 - \mu_2 = -\delta^* \\ H_1^1: & \mu_1 - \mu_2 > -\delta^* \end{cases} \qquad \begin{cases} H_0^2: & \mu_1 - \mu_2 = \delta^* \\ H_1^2: & \mu_1 - \mu_2 < \delta^* \end{cases}$$

Just as in the previous case, both hypotheses are tested at the desired α value, and the rejection of both H_0 indicates evidence of equivalence.

Example – Laboratory certification

A ballistics laboratory is in the process of being certified for the evaluation of shielding technology, and needs to provide evidence of equivalence of a given callibration procedure with the reference equipment;

The certification authority demands that the mean hole area generated by this procedure in the lab be the same as the one from the reference equipment, and tolerates deviations no greater than $4mm^2$;

From previous measurements, the standard deviations can be roughly estimated as $\hat{\sigma}_{Lab} = 5mm^2$ and $\hat{\sigma}_{ref} = 10mm^2$.

The desired error levels for the comparison are $\alpha = 0.01$ and $\beta = 0.1$.

Example – Laboratory certification

To calculate the required sample size, assume that $\Delta\mu^* = 0.5$. Then:

```
> # load functions to calculate sample size for TOST
> source("calcN_tost.R")
>
> # Calculate sample size
> calcN_tost2(alpha = 0.01,
+             beta = 0.1,
+             diff_mu = 0.5,
+             tolmargin = 4,
+             s1 = 5,
+             s2 = 10)
[1] 144.1999
```

We'll need 145 observations from each group to test for equivalence with the desired experimental properties.

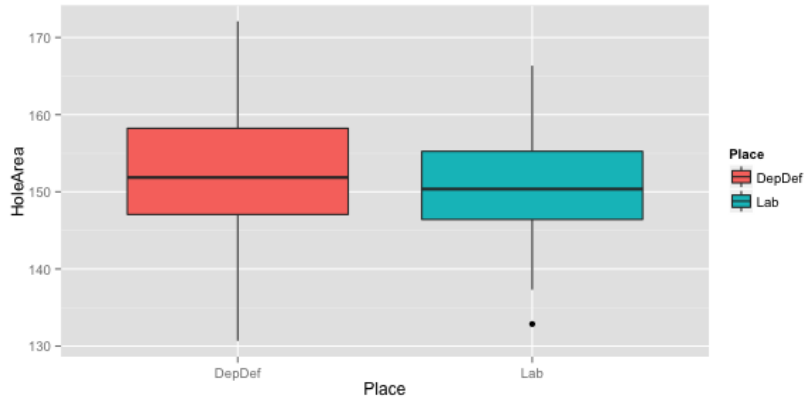
Certification – Data Analysis

After collecting the observations, we proceed to the analysis:

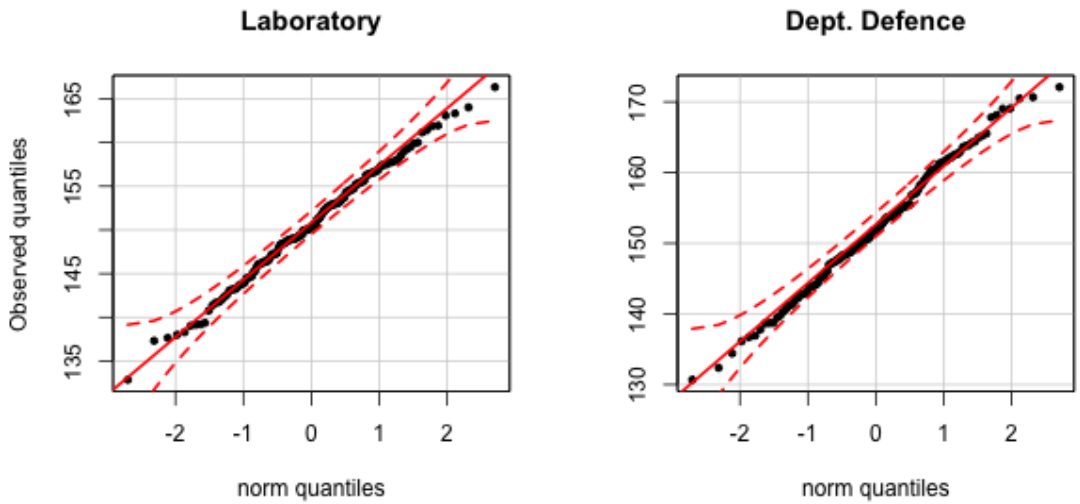
```
> data<-read.table("../data files/labdata-example.csv",
+                  header = T, sep = ",")

> # Two one-sided t-tests
> t.test(HoleArea~Place,  data = data,  alternative = "less", mu = 4,
+        conf.level = 0.99)$p.value
[1] 0.00304124
> t.test(HoleArea~Place, data = data, alternative = "greater", mu = -4,
+        conf.level = 0.99)$p.value
[1] 6.586193e-10

> # Get (1-2*alpha) CI
> t.test(HoleArea~Place, data = data,  conf.level = 0.98)$conf.int
[1] -0.5117627  3.6244386
```



Verification of test assumptions – Normalty

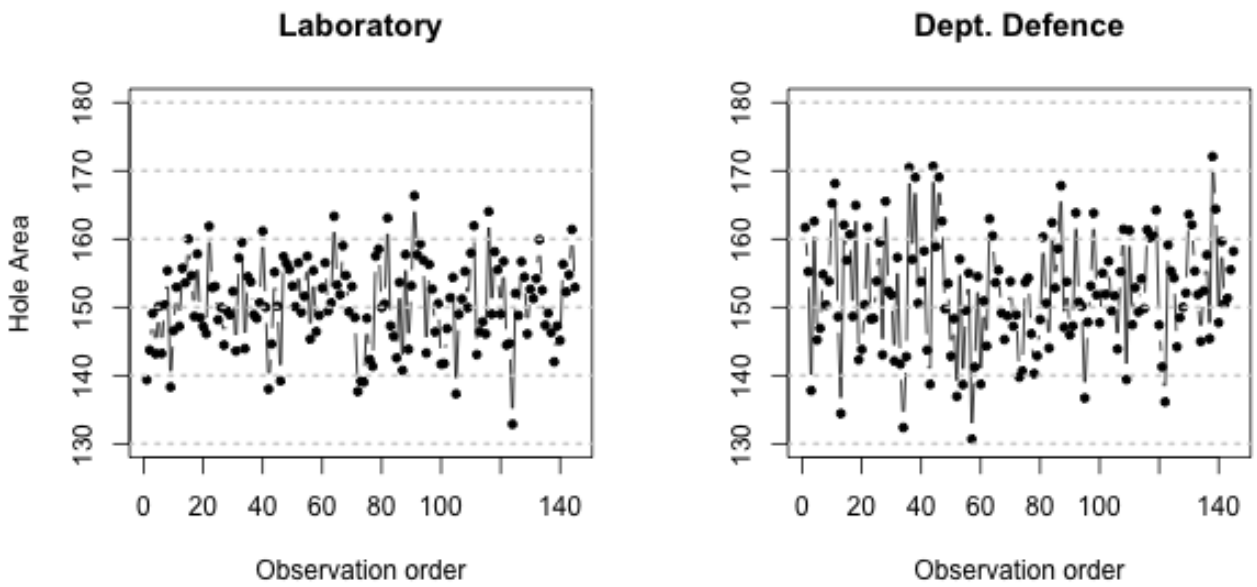



```
> par(mfrow=c(1,2))
> qqPlot(subset(data, Place=="Lab")[,2],
+        pch=20,
+        main = "Laboratory",
+        ylab = "Observed quantiles")
> qqPlot(subset(data, Place=="DepDef")[,2],
+        pch=20,
+        main = "Dept. Defence",
+        ylab = " ")
```

Verification of test assumptions – Independence

```
> dwtest(HoleArea~Place, data=data)
DW = 1.8116, p-value = 0.04757

> par(mfrow=c(1,2))
> plot(seq_along(subset(data, Place=="Lab")[,2]),
+      subset(data, Place=="Lab")[,2], ...)
> plot(seq_along(subset(data, Place=="DepDef")[,2]),
+      subset(data, Place=="DepDef")[,2], ...)
```



Non-normal Data

Non Normality

What is non-normality?

- Until now we studied methods which **assume** that the experimental data follows a normal distribution (or close enough).
- In many cases, this assumption **does not hold**. In this condition, how can we perform the statistical analysis of the results?

Non Normal data makes everything go wrong

Weight Loss Example

A researcher is examining two different diets, **Diet A** and **Diet B**, and wants to compare the weight loss by people following one diet or the other. They obtained the following data:

```
diet.a <- c(4,3,0,-3,-4,-5,-11,-14,-15,-300)
diet.b <- c(-8,-10,-12,-16,-18,-20,-21,-24,-26,-30)
```

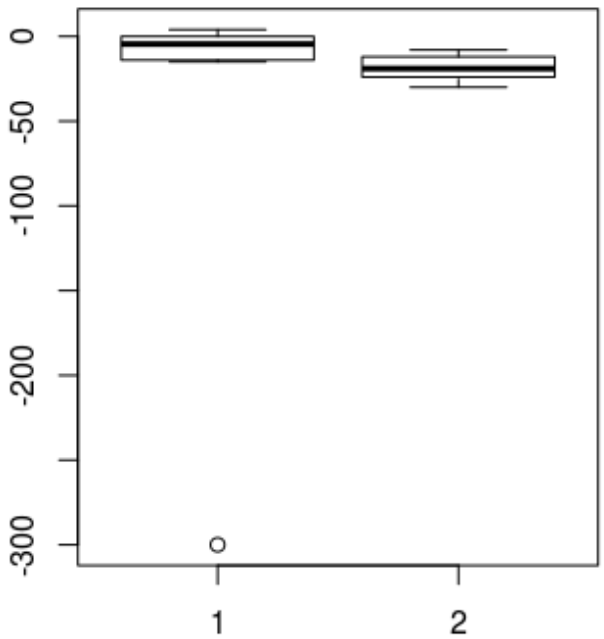
As you can see, Diet A has one big outlier[1] that makes the data not normal. How much does this affect the statistical test?

[1] Why does this outlier exist? Data input error? Very rare case?

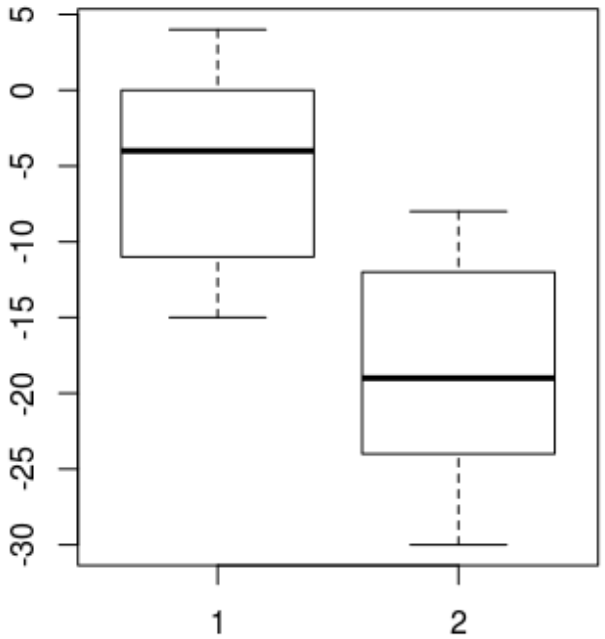
Non Normal data makes everything go wrong

Weight Loss Example

Data with Outlier



Data without Outlier



Checking a visualization, it seems like diet A has smaller losses than diet B overall. Except for that outlier. What happens with the T-test?

Non Normal data makes everything go wrong

Weight Loss Example

The standard T-test does not indicate a difference between these samples, and even suggests that the mean of the first sample is lower!

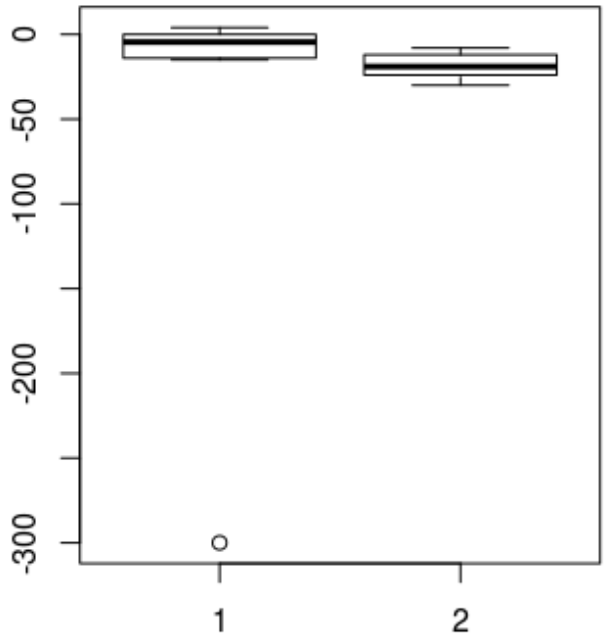
```
diet.a <- c(4,3,0,-3,-4,-5,-11,-14,-15,-300)
diet.b <- c(-8,-10,-12,-16,-18,-20,-21,-24,-26,-30)
t.test(diet.a,diet.b)

##  Welch Two Sample t-test
##
## data:  diet.a and diet.b
## t = -0.53945, df = 9.1048, p-value = 0.6025
## alternative hypothesis: true difference in
##                         means is not equal to 0
## 95 percent confidence interval:
##  -82.9774  50.9774
## sample estimates:
## mean of x mean of y
##     -34.5     -18.5
```

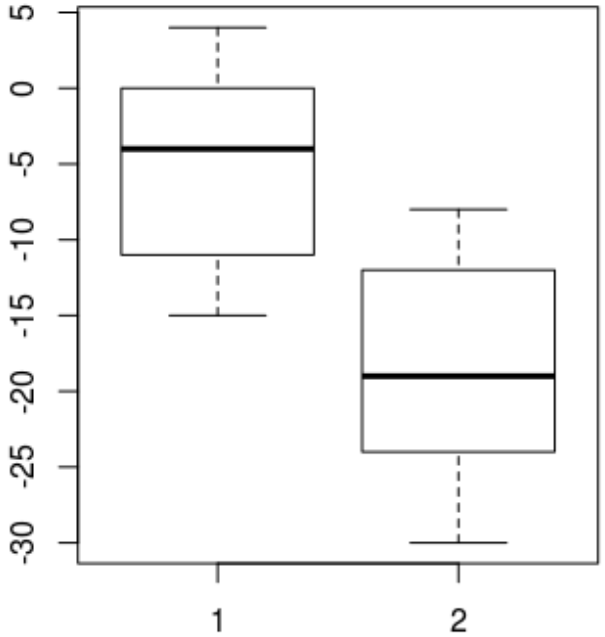
Non Normal data makes everything go wrong

Weight Loss Example

Data with Outlier



Data without Outlier



Remember that outliers are not always this obvious!

Non Normal data makes everything go wrong

Weight Loss Example

Using a **non-parametric test** solves the problem.

```
diet.a <- c(4,3,0,-3,-4,-5,-11,-14,-15,-300)
diet.b <- c(-8,-10,-12,-16,-18,-20,-21,-24,-26,-30)
wilcox.test(diet.a,diet.b)

##  Wilcoxon rank sum test
##
## data:  diet.a and diet.b
## W = 82, p-value = 0.01469
## alternative hypothesis: true location shift
## is not equal to 0
```

Non Normality

Examples of Non Normal data?

There are many different ways that data can violate the assumption of normality:

- *Special Data Cases: Outliers and Limits*: We saw one example of outlier that challenges the assumption of normalty. Limits in the data (minimum time) can have a similar effect.
- *Extreme Non-Normal Distributions.* Power Distribution, Cauchy Distribution, etc.
- *Ordinal Data.* Ordinal data is numeric, in the sense that it can be compared/ordered, but you can't rely on direct algebra. (eg: Human opinion scales)
- *Non-numerical data*: categorical data, class data, etc;

Non Normality

Example: Random Processes

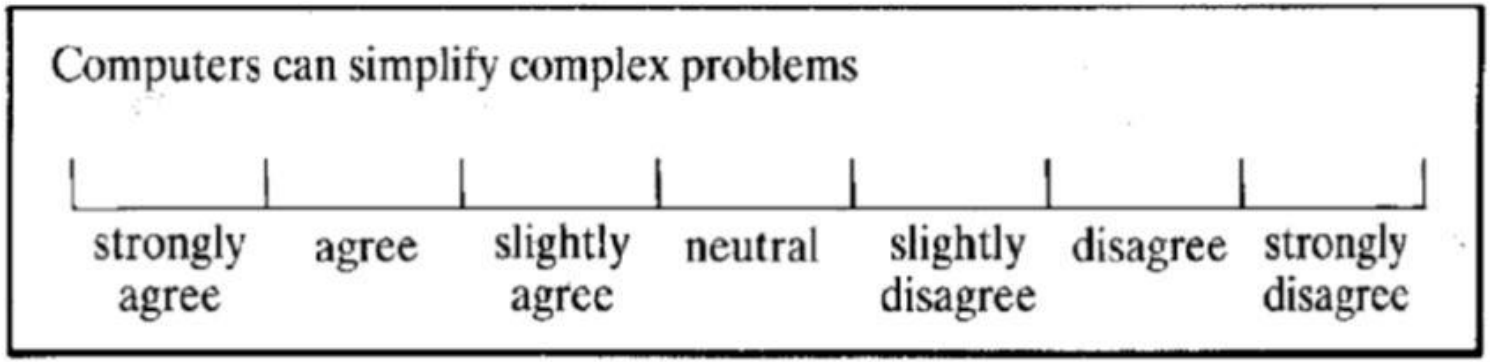
Random processes in nature (such as plant growth or shell formation) very often follow a normal distribution (or bell curve). However, random artificial processes do not always follow a normal distribution:

- Random numbers in computer very often use **Uniform distributions**. (Uniform distributions can be easily approximated to normal distributions by the use of bootstrapping).
- Random numbers from social processes very often show **Power distributions**. Power distributions are characterized by "90% of observations have 10% of values". Examples: Salary data, Social network data, etc.

Non Normality

Example: Likert Data

Likert data is often collected from surveys and interview questions. It is usually composed of multiple questions with 5 or 7 options, ranged from "Strongly Agree" to "Strongly Disagree", or "Always" to "Never".



Why can't we treat likert data directly as numerical?

- Values outside of the 0-5 range have no meaning;
- Algebra on likert data has no meaning (Neutral+Disagree=?)
- The difference between levels is not clear. Is "Agree" equally distant from "Slightly Agree" and "Neutral"?

Non Normality

Strategies for non normal data

Let's discuss three things that we can do about non-normal data, regarding statistical testing:

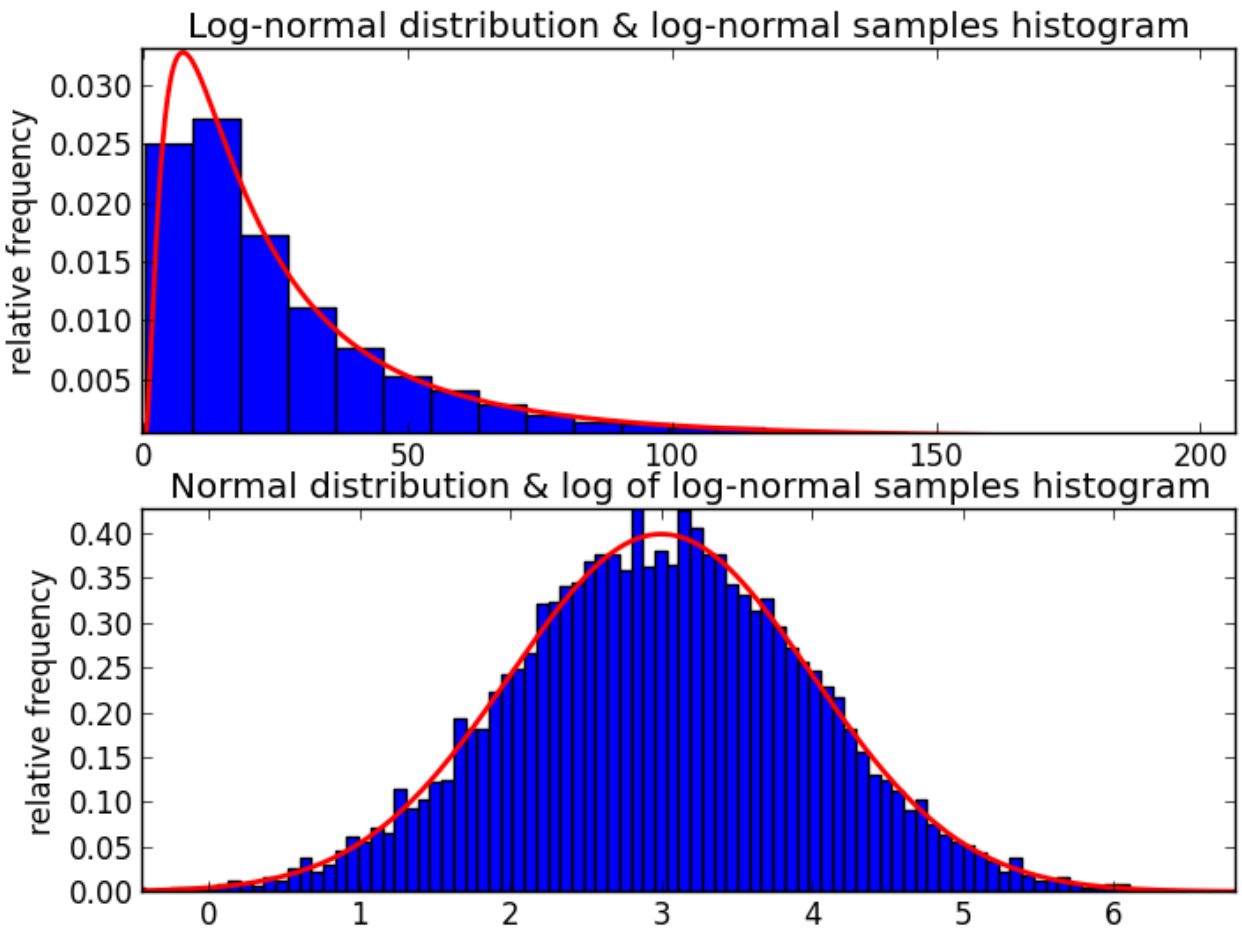
- Do nothing
- Transform the Data
- Non parametric Testing

Please note that this is a quick overview of the treatment of non-normal data. **Study your particular case carefully!**

Data Transformation

Log Transformation

We can apply certain transformations to "normalize" some data.



Log Transformation

```
# Generate lognormal data
#
set.seed(17)
z <- exp(rnorm(200, -2, 0.4))

# Log transformation
#
y <- log(z)
mu.hat <- mean(y)
sigma.hat <- sd(y)
```

Data Transformation

Other transformations

- For left skewed data:
  - square root, cube root, log
- For right skewed data:
  - square root (constant $-x$), cube root (constant $-x$)

Attention: Logarithm of 0 and negative data is not defined, so you may need to add a constant before the transformation.

Data Transformation

Be careful when transforming data

- Careful when reporting back the data on the paper.
  - Make sure to mention what data transformation was used in the analysis.
  - Apply **back-transformation** before discussing results.
- Beware of when Null hypotheses are not equivalent!
  - Example: Lognormal mean includes the variance. Transformed lognormal mean does not. Null hypothesis is only equivalent when variance is equal!

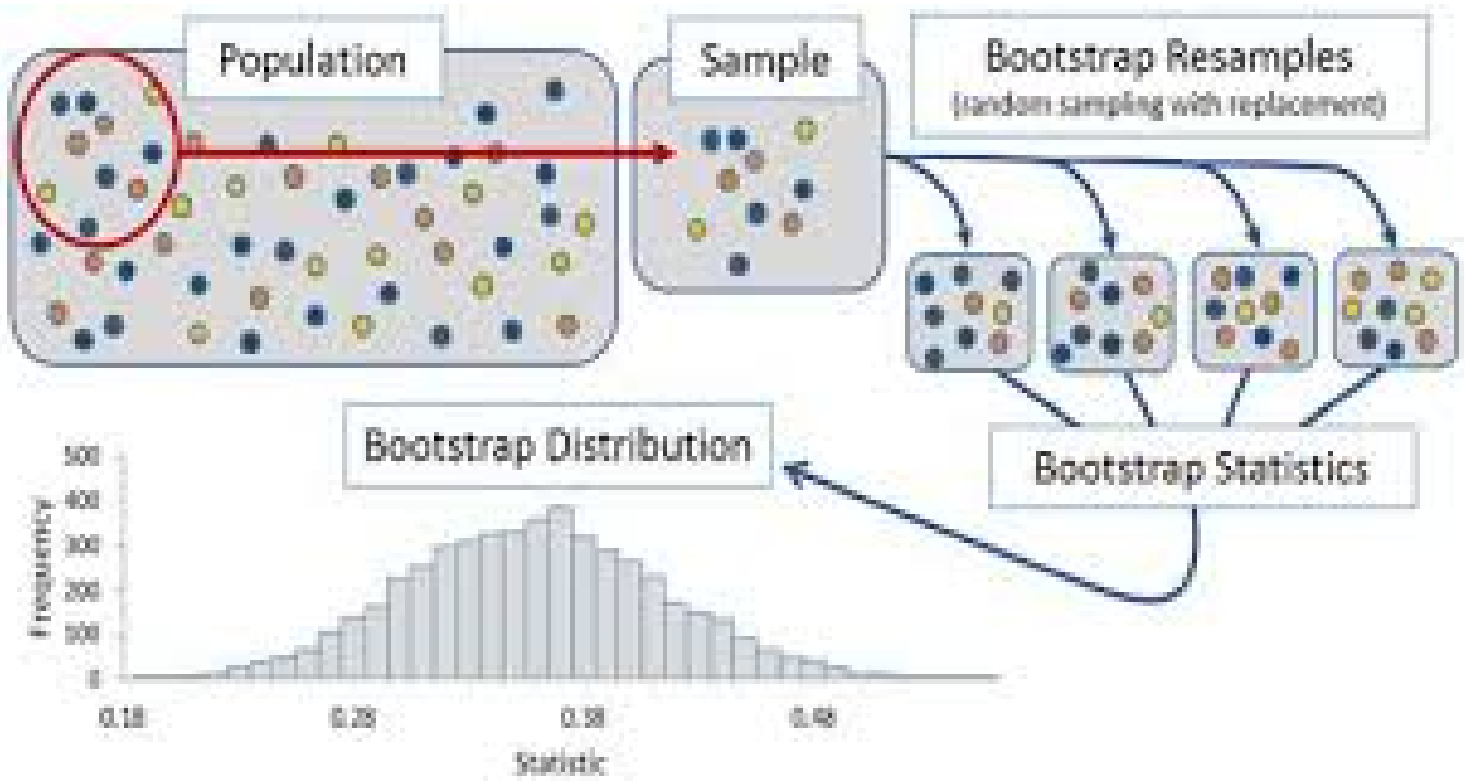
Bootstrapping

Another way to make data follow a normal distribution is to use the **Bootstrapping** procedure. Bootstrapping uses the idea of the CLT (central limit theorem) by creating a "sample mean distribution" that will usually follow the normal distribution.

Bootstrapping

The Bootstrapping Procedure

- Take n samples of size m (with repetition)
- Calculate the mean of each sample.
- Your bootstrapped data is the set of sample means.



Bootstrapping

Using Bootstrapped Data

The R package "boot" can help you appropriate create tests and confidence intervals from bootstrapped data.

Non Parametric Tests

Non-parametric tests remove the assumption of normality from the population distribution. On the other hand, they can be a bit weaker, and cannot estimate the real-world distance between two samples.

- Wilcoxon Signed Rank Test (1 sample)
- Wilcoxon Ranked Sum Test / Mann-whitney Test (2 samples)
- Kruskall-Wallis Test (multiple samples)

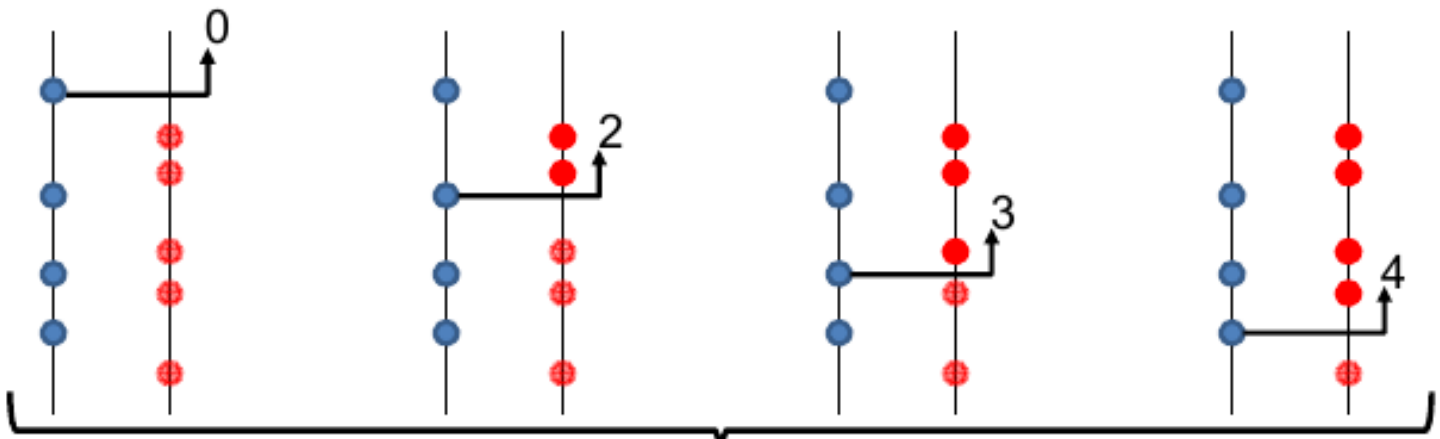
Non Parametric Tests

One or two Samples

- unpaired test: Mann-Whitney U-test;
- paired test: Wilcoxon signed-ranks test;

Mann-Whitney U-test

1. Calculate a *U* value.

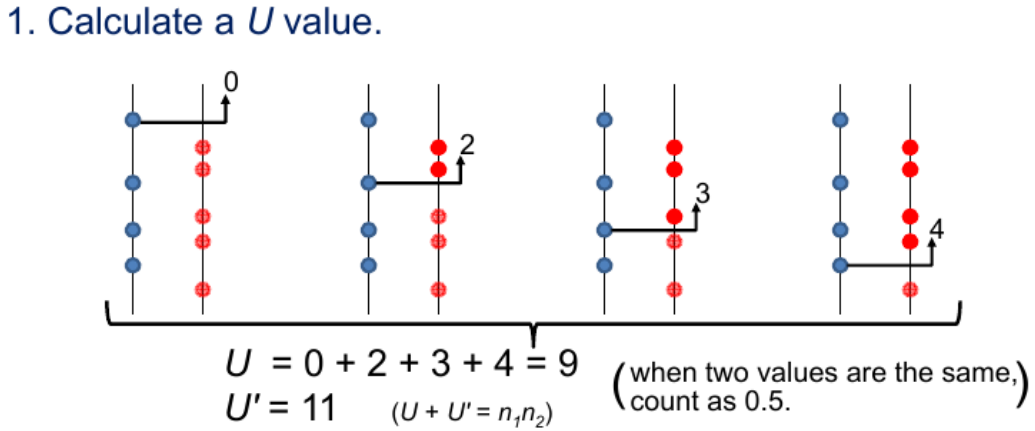


$U = 0 + 2 + 3 + 4 = 9$

$U' = 11$ $(U + U' = n_1 n_2)$

(when two values are the same, count as 0.5.)

Mann-Whitney U-test



- Choose the smaller value of U or U'
- Null Hypothesis: **Both samples come from the same distribution**
- Under the null hypothesis, for big enough n_1 and n_2, U follows roughly a normal distribution with mean $\frac{n_1 n_2}{2}$ and variance $\frac{n_1 n_2 (n_1+n_2+1)}{12}$
- Calculate the test statistic z, and find the p-value from the α-percentile in the z distribution.

Wilcoxon Signed Rank Test

data of 2 groups		# of winnings and losses	
173	174	-	+
143	137	+	-
158	151	+	-
156	143	+	-
176	180	-	+
165	162	+	-

- The Wilcoxon test takes the relative difference between pairs (positive or negative)
- Null hypothesis: **Positive and Negative signs are equally likely**
- The overall number of signs is compared against a binomial distribution under the Null hypothesis.

Wilcoxon Signed Rank Test

R example

```
## Hollander & Wolfe (1973), 29f.
## Hamilton depression scale factor measurements in 9
## patients with mixed anxiety and depression, taken at
## the first (x) and second (y) visit after initiation
## of a therapy (administration of a tranquilizer).

x <- c(1.83,  0.50,  1.62,  2.48, 1.68, 1.88,
       1.55, 3.06, 1.30)
y <- c(0.878, 0.647, 0.598, 2.05, 1.06, 1.29,
       1.06, 3.14, 1.29)

wilcox.test(x, y, paired = TRUE, alternative = "greater")

Wilcoxon signed rank test
data:  x and y
V = 40, p-value = 0.01953
alternative hypothesis: true location shift
is greater than 0
```

Lecture Summary

Equality Testing:

- You can test for equality of a quantity to a value by using a non-superiority test and a non-inferiority test together.

Non-Normality:

- Normality can be broken in several cases: Extreme outliers, data limits, ordinal data, and non-numerical data;
- For light cases of non-normality, we can just remove outliers or transform the data;
- For transforming the data, bootstrapping is a good technique;
- If transformation of the data is not feasible, non-parametric tests can easily substitute the parametric tests that we studied in the last classes;

Next Lecture – Q&A Session and Tutorial

The lecture next week will be an open Q&A session, followed by a tutorial.

- Ask questions about the lecture topics;
- Ask questions about your report 1 and 2;
- Ask questions about anything else;

The lecture will happen as a Zoom meeting. Details for the meeting will be published on Manaba. Stay tuned!

Recommended Reading

- E. Walker, A.S. Nowacki, "Understanding Equivalence and Noninferiority Testing", Journal of General Internal Medicine 26(2):192-196, 2011.
- Kristin L Sainani, "Dealing with Non-Normal Data." https://onlinelibrary.wiley.com/doi/full/10.1016/j.pmrj.2012.10.013
- Feng et al., "Log transformation and its implication for data analysis." https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4120293/
- Bommae Kim, "Should I always Transform My Variables to Make them Normal?" https://data.library.virginia.edu/normality-assumption/
- Hideyuki Takagi, "Tutorial on Statistical Tests" http://www.design.kyushu-u.ac.jp/~takagi/TAKAGI/downloadablefile.html#StatisticalTests

About these Slides

These slides were made by Claus Aranha, 2020. You are welcome to copy, re-use and modify this material.

These slides are a modification of "Design and Analysis of Experiments (2018)" by Felipe Campelo, used with permission.

Individual images in some slides might have been made by other authors. Please see the references in each slide for those cases.

Image Credits I

[Page 8] Summary of Equivalence Testing: Walker and Nowacki (2011), J. General Internal Medicine 26(2):192-196.
[Page 9] Difference vs Equivalence Image: Walker and Nowacki (2011), J. General Internal Medicine 26(2):192-196.
[Page 11] TOST errors image: Matthews (2010), Sample Size Calculations, MMB. pg. 46
[Page 13] Gun Shield Image from:
`http://www.everydaynodaysoff.com/2013/08/05/ballistic-shield-for-operators-only/`