

# 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:** In the last two class, we learned 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 statistics that we discussed until now make the assumption that the sampling distribution is normal. What can we do when this assumption is breached?

## Part I – Equality Testing

# Testing equivalence

## Introduction

In the previous lectures, we introduced tests that focus on detecting **differences** between a population parameter  $\theta$  and its nominal value  $\theta_0$  under a null hypothesis.

In engineering and science, we are sometimes interested in investigating the **equivalence**, within a given margin of error. Some example cases:

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

# Testing equivalence

## Question of interest in usual studies

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

## Question of interest in equivalence studies

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*  $\delta^*$ .

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

# Testing Non-inferiority

## Equivalence and non-inferiority

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).

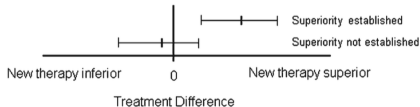
In non-inferiority tests, one can declare that a given process is not worse than a standard one only if enough evidence is provided to conclude that the performance of the proposed process is no more than  $\delta^*$  units worse than that of the standard.

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  $\delta^*$  in its formulation.

# Comparison of studies

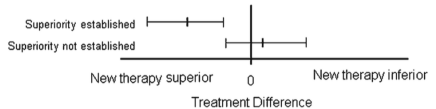
Efficacy is measured by success rates, where higher is better.

**Traditional comparative study**

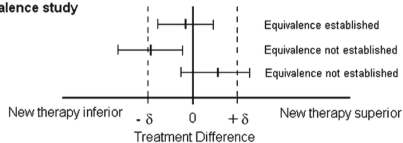


Efficacy is measured by failure rates, where lower is better.

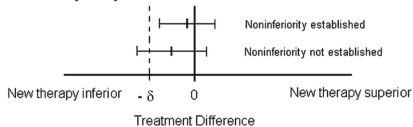
**Traditional comparative study**



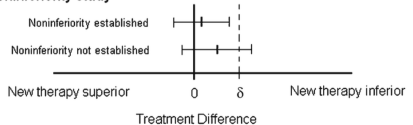
**Equivalence study**



**Noninferiority study**



**Noninferiority study**





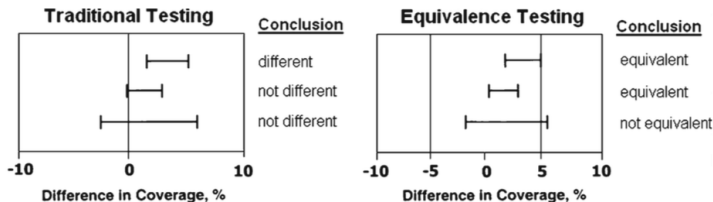
# Testing Equivalence

## 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  $\alpha$  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:

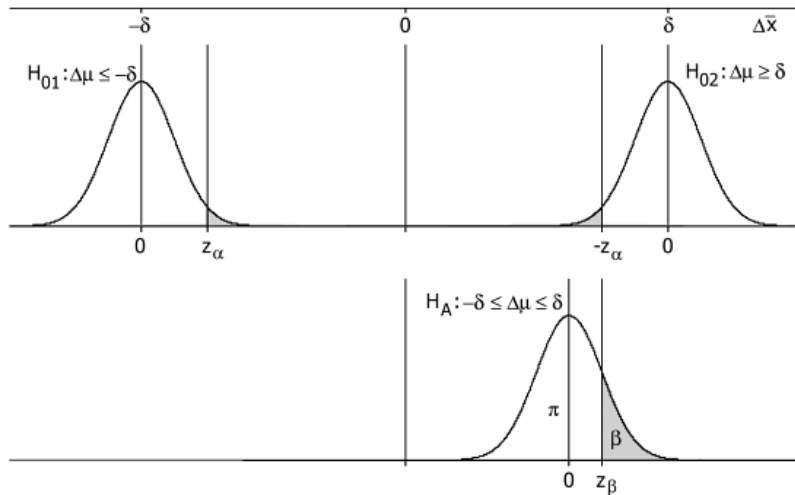
$$\begin{cases} H_0 : |\mu - \mu_0| = \Delta\mu \geq \delta^* \\ H_1 : \Delta\mu < \delta^* \end{cases}$$

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:

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

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

# Interpretation of the Two One Sided Hypothesis



# Equivalence Testing by TOST

## Hypotheses for two samples

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

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


---

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

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

Just as in the previous case, both hypotheses are tested at the desired  $\alpha$  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 calibration 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  $4\text{mm}^2$ ;

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

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:

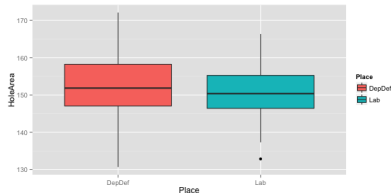
```
> # load functions to calculate sample size for TOST
> source("calcN_tost.R")
>
> # Calculate sample size
> calcN_tost2(alpha = 0.01,
+             beta = 0.1,
+             diff_mu = 0.5,
+             tolmargin = 4,
+             s1 = 5,
+             s2 = 10)
[1] 144.1999
```

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

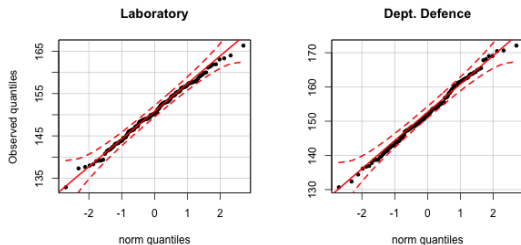
# Certification – Data Analysis

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 – Normality



```
> 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

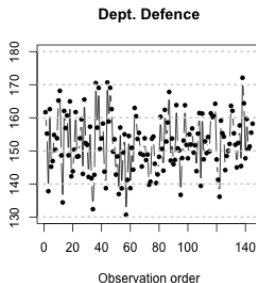
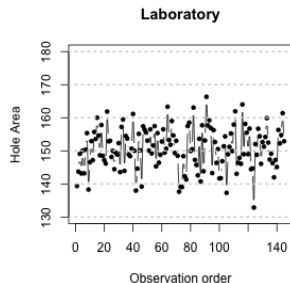
```
> dwtest(HoleArea~Place, data=data)
```

```
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], ...)
```



## **Part II – 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<sup>1</sup> that makes the data not normal. How much does this affect the statistical test?

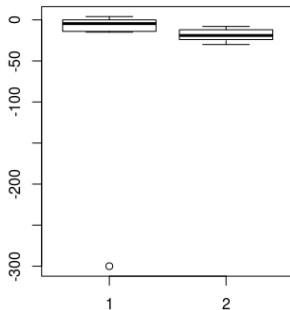
---

<sup>1</sup>Why does this outlier exist? Data input error? Very rare case?

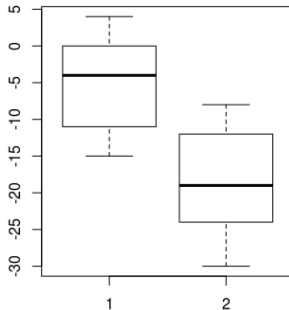
# Non Normal data makes everything go wrong

## 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

How can we solve this problem?

- Identify and remove the outlier?
  - If the outlier is an experimental error, it makes sense to remove it;
  - On the other hand, maybe the outlier is some important effect worth being investigated;
- Alternatively, we could use a test that is not sensitive to the outlier.

# Non Normal data makes everything go wrong

## Weight Loss Example

**A non-parametric test** will indicate a rejection of the null hypothesis of equality.

```
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
```



# Examples of Non-Normal Data

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

- *Special Observations in the Data:*
  - Outliers, data collection errors;
  - Absolute limits in the data (measuring time);
- *Extreme Non-Normal Distributions:*
  - Power Distribution, Cauchy Distribution, etc.
- *Ordinal Data:*
  - Ordinal data is data that can be ordered and compared by some criteria, but you cannot apply traditional algebra on it (ex: subjective scores);
- *Completely Non-numerical data:*
  - categorical data, class data, etc; (ex: colors)

# Non-Normal Data Example: Random Processes

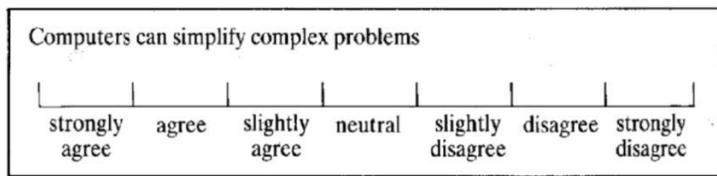
Random processes in nature, such as plant growth or shell formation, are often seen to follow a normal distribution or bell curve. On the other hand, random artificial processes not always do.

- In general, Pseudo Random Number Generators will use an **Uniform Distribution**. Because of the CLT, aggregations of these results will tend to normal distributions.
- On the other hand, random social processes will often show **Power Distributions** (salaries, social networks) or **Binomial Distributions** (queues);
- It is important to study and understand the process being researched to know its characteristics;

# 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

When our data does not follow the normality assumption, there are many different strategies that we can apply, depending on the type of data, and the type of normality violation:

- **Do Nothing**

- We can remove outliers that break the normality assumption, or trust that test will be robust for small deviations of normality;

- **Transform the Data**

- Transformation of the data can restore the normal property to data;

- **Non parametric Testing**

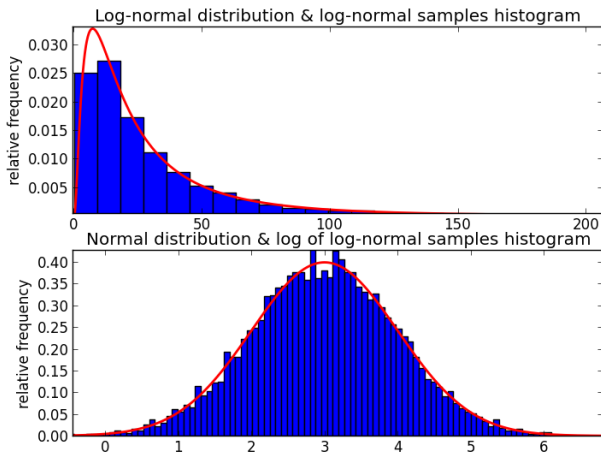
- Some statistical tests do not make the assumption of normality of the sampling distribution (in exchange for less power);

- **Look for a new statistics textbook**

# 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

## Skew Transformation

A strong skew in the sampling distribution can be a larger problem for the standard statistical tests. It is possible to remove these through data 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

- Pay attention when describing the analysis on a paper or report:
  - Any transformation used must be explained in the analysis;
  - The discussion of the results must be done on the transformed data as well as the original data;
- Beware that the hypotheses may not be equivalent!
  - Example: The lognormal mean includes the variance. But the transformed lognormal mean does not. In this case, the null hypothesis is only equivalent when the variance of the transformed distribution is equal!



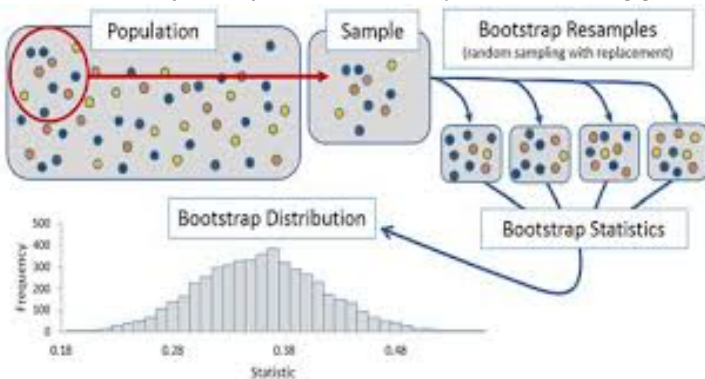
# Bootstrapping

The Bootstrapping procedure is used to obtain an approximation of the "sample mean distribution" from the sample data.

Following the properties of the Central Limit Theorem, the sample mean distribution will usually follow a normal distribution, even when the underlying distribution of observation values is not normal;

# The Bootstrapping Procedure

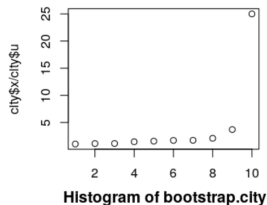
- Take an initial sample with  $m$  observations;
- Create  $n$  **bootstrap samples** by selecting  $m_b < m$  from the initial sample  $n$  times;
- Calculate the mean of each bootstrap sample.
- The set of bootstrap sample means are your **bootstrapped data**;



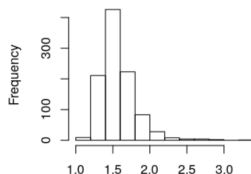
# Generating Bootstrapped Data

The R package "boot" creates bootstrapped data, c.i.s, and tests.

```
> city
      u      x
1  138  143
2   93  104
3   61   69
4  179  260
5   48   75
6   37   63
7   29   50
8   23   48
9   30  111
10   2   50
```



Histogram of bootstrap.city



```
> library(boot)
> ratio <- function(d, w) sum(d$x * w) / sum (d$u * w)
> bootstrap <- boot(city, ratio, R = 999, stype = "w")
> city.bootstrap <- bootstrap[[2]]
```

# Non Parametric Tests

Non-parametric Tests involve statistics that do not assume normality from the population distribution.

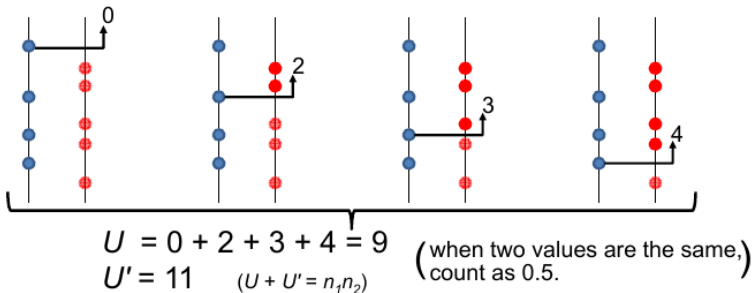
Weak assumptions about the population, however, causes the non-parametric tests to be less strong than parametric ones. Also, usually non-parametric statistics usually do not calculate the distance between the parameter estimate and the hypothesis values.

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

# Mann-Whitney U-test

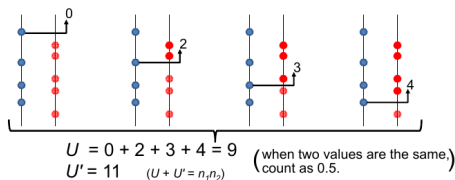
Unpaired test for two samples

1. Calculate a  $U$  value.



# Mann-Whitney U-test

1. Calculate a  $U$  value.



- 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  $\alpha$ -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/>