Experiment Design for Computer Sciences (01CH740)

Topic 03 - Statistical Inference

Claus Aranha
caranha@cs.tsukuba.ac.jp

University of Tsukuba, Department of Computer Sciences

Version 2021.1 (Updated April 21, 2021)

Part I (Intro) – What is Statistical Inference?

Summary and Outline

- In the last lecture, we talk about **descriptive statistics**:
  - Describe the system we want to study using experiment data and statistics;
  - Point Estimators: Calculate a specific value (parameter) of the population;
  - Interval Estimator: Shows the error/confidence of our estimation;
- In some cases, a description alone is not enough.
- We want to make an "informed decision" about the data.
  - "Given some experimental data, can I conclude **X**?"
- **Statistical Inference** is a powerful tool for this.

Example of when we need Statistical Inference

You are the owner of a factory that produces delicious chocolate.

The packages that you produce and sell should contain around 300g of chocolate. (Of course there is some variation)

Every 6 months, you want to check if everything is working correctly.

What should you do?

Example of when we need Statistical Inference

Using the knowledge of this class, you make an experiment.
You take a sample of **30 packages** from the facture and weight them.

First, you calculate the sample's average weight: **295g**

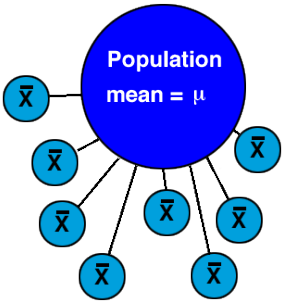
Then you remember last lecture, and calculate the
95% confidence interval: **283g to 307g**.

From this data, do you believe that the mean production of your factory is 300g? Or is it necessary to investigate further?

You need more information to make that decision!

Enter Statistical Inference:

Statistical inference is a process that uses data analysis to establish the (probable) truth of a statement. (compare with logical inference)



- We create a probabilistic model that describes our system of interest, and the possible outcomes of an experiment.
- The statistics calculated from sample data can be described as random variables, and analized.

Using the sample data, we can compare the characteristics of the sample (sample model) with the assumed characteristics of the population (population model).

Sample data $\rightarrow$ parameter estimation $\rightarrow$ compare with model $\rightarrow$ statistical inference

Statistical Hypothesis

(Statistical) Hypothesis is a key concept of statistical inference.

- **Hypothesis:** a statement that explains a phenomenom we observe;
- **Statistical Hypothesis:** an statement about a **statistical** model;

Back to our example. We want to know if our chocolate factory is working normally. Our packages should have, on average, at least 300g of chocolate (or we might get sued!).

Statistical Hypothesis

The population model that describes the weight of a package in our factory has a **mean** of at least 300g. ($\mu_{\text{weight}} \geq 300$)

Pay attention! Note how the *scientific question* leads to the *statistical hypothesis*.

Statistical Hypothesis

Common Hypotheses and Statistical Hypotheses

- **Common Hypothesis:** a general statement about what we believe about the world;
- **Statistical Hypothesis:** an statement about parameters in a **statistical** model;

Common Hypothesis: The factory is broken, and producing less cocoa than normal.
Statistical Hypothesis: The mean weight of packages produced is less than 300g

Common Hypothesis: The proposed algorithm is faster than standard algorithms.
Statistical Hypothesis: The difference in mean execution time between the proposed and the standard algorithm is above 2s

Common Hypothesis: Cats are more popular than dogs.

Statistical Hypothesis: The proportion of cat videos on Youtube that are watched until the end is 5% higher than the proportion of dog videos watched until the end.

One easy way to think about it, is that a statistical hypothesis can always be represented as an equation.

Statistical Hypothesis

What is a good hypothesis?

Here are some characteristics of a useful **scientific hypothesis** (statistical or not). Keep these characteristics in mind when you create your hypotheses:

- **Predictive power**: The hypothesis not only explains the existing data, it helps you predict future data.
  - There were mistakes in the production because the workers were tired yesterday.
  - There were mistakes in the production because the workers are tired on Mondays.
- **Principle of parsimony** (Ockam's razor): The hypothesis makes few assumptions about the system:
  - Energy usage pattern is described by this neural network with 1.000.000 parameters.
  - Energy usage pattern is described by this 3rd degree polynomial.
- **External consistency**: The hypothesis fits with existing, well accepted knowledge about the system.
  - Mean global temperature is correlated with the number of active pirates;
  - Mean global temperature is correlated with CO2 emmissions;

Hypothesis and Experiments

How do we use hypotheses?

General approach:

- Create multiple hypothesis for the phenomenom that we are studying.
- Run experiment and decide which hypothesis fits the data best.

1. Create Hypotheses:

- H1: The mean weight of packages produced by the factory is above 300g.
- H2: The factory is broken and producing packages with much less than 300g.
- H3: The packages produced by the factory follow a sine wave

2. Obtain Data: Collect 10 cocoa packages randomly, and weight them

- Weights: 293 325 271 313 309 298 284 304 248 296
- Sample average: 294g
- Minimum and maximum: 248g, 325g

Hypothesis, Experiments, and Statistical Inference

The statistical inference process helps us choose which hypothesis fits the experimental data best. Given the sample data x, we calculate the **probability that x is observed if the hypothesis H_i is true**: $P(x|H_1), P(x|H_2), P(x|H_3), \ldots$

We give more credibility for the hypothesis that maximizes the probability of the data.

Example: $x = \{293, 325, 271, 313, 309, 298, 284, 304, 248, 296\}$

Hypothesis 1: $\mu \geq 300$

What is the probability that we see the sample x when the mean production of the factory is 300g or more?

Hypothesis 2: $\mu < 300 - \delta$

What is the probability that we see the sample x when the mean production of the factory is δ less than 300g?

Null Hypothesis Significance Testing (NHST)

The NHST approach for statistical inference involves the contrast between a **null hypothesis** and an **alternate hypothesis**.

Null Hypothesis (H_0)

- Absence of effects;
- Conservative model;
- "nothing special is happening"

"As expected, the mean weight packages produced in the factory is at least 300g"

$H_0 : \mu \geq 300$

Alternative Hypothesis (H_1)

- Presence of some effect;
- Something "new" is happening;

"There is an anomaly in the factory, and the mean production is below 300g"

$H_1 : \mu < 300$

Null Hypothesis Significance Testing (NHST)

How to choose a null hypothesis?

- Use existing knowledge about the process being investigated;
- Values obtained from theory or models (model validation);
- System requirements (investigation of system compliance);

Chocolate factory example:

One client complained about our packages on twitter, so we suspect that there may be a problem in our chocolate production. We propose sampling 20 packages, and estimating the *mean* of the population from this sample:

- **Null Hypothesis:** $H_0 : \mu \geq 300$
- **Alternative Hypothesis:** $H_1 : \mu < 300$

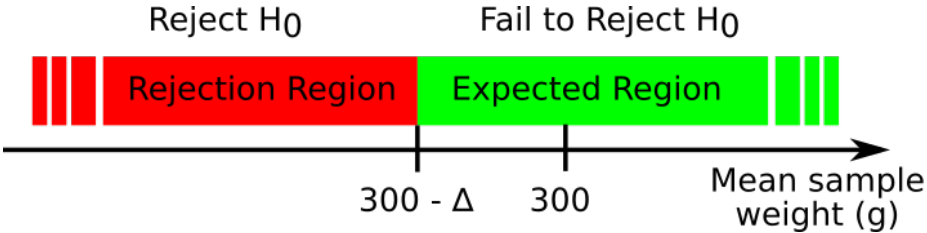
Part II – General Procedure of Statistical Testing

Outline of the Procedure for the NHST

Given a good H_0 and H_1, the objective of the statistical test is to answer the question: "Do we have enough evidence to prefer the Alternate Hypothesis to the Null Hypothesis"?

More usually, the phrase "reject the null hypothesis" is used.

The overall procedure is to calculate the parameter estimate for our parameter of interest, and see if this value falls into a Rejection Region. An estimate that falls in this region leads us to reject the null hypothesis as **unlikely.**

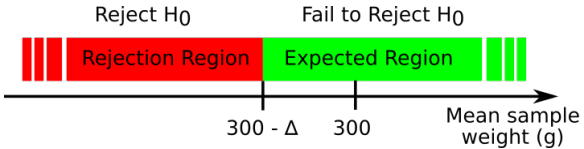


Outline of the Procedure for the NHST

Inference Errors

Remember that the statistic that estimate a parameter is a **random variable**. So there is an error associated with the statistic, and we can reach wrong conclusions.

- Sample error is too large: statistic is in the rejection region when the H_0 is true;
- Δ is too large: statistic is in expected region when H_0 is not true;



Designing the experiment correctly, we want to control the probability of these errors.

Type I Error (False Positive)

The data reject the H_0, when H_0 is true.

The probability of occurrence of a false positive in any hypothesis testing procedure is generally known as the significance level (α, or "alpha") of the test.

- Significance Level $\alpha = P(\text{type I error}) = P(\text{reject } H_0 | H_0 \text{ is true})$

It is also called the confidence level of a test, given by $(1 - \alpha)$ or $100(1 - \alpha)\%$

- Confidence Level $100 * (1 - \alpha)\%$

Type I Error (False Positive)

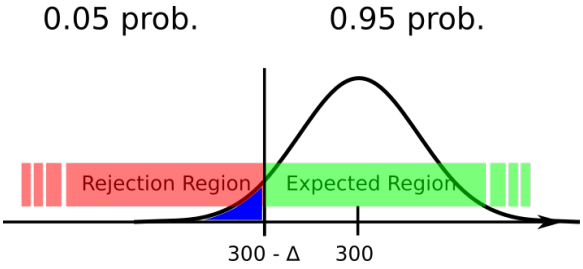
Controlling the probability of Type I Error by choosing α

For a given statistical model, H_0, and sample, the probability of a type I error (α) is the area of the sample distribution that falls under the threshold value.

So, by selecting the threshold value, we can control α.

In our example, assuming the package follows a normal distribution, we expect the sample mean to be μ with standard error ($\sigma/\sqrt{n}$).

We can calculate Δ so that the probability of the sample mean falling in the rejection region is α



Type II Error (False Negative)

The data does not reject H_0 when H_0 is false

The probability of occurrence of a false negative in any hypothesis testing procedure is generally represented by the Greek letter β:

$$\beta = P(\text{type II error}) = P(\text{not reject } H_0 | H_0 \text{ is false})$$

The quantity $(1 - \beta)$ is known as the **power of the test**. It quantifies the test's sensitivity to effects that violate the null hypothesis.

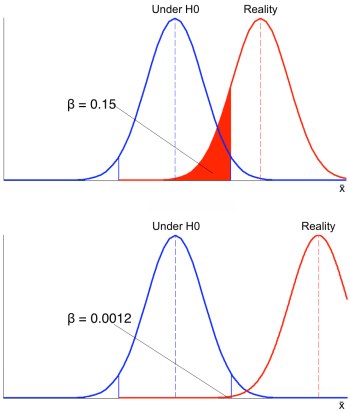
Type II Error (False Negative)

Statistical Interpretation

The probability of Type II error is harder to estimate precisely than the Type I error.

This is because it depends on the True Value of the parameter. But this value **is unknown when H_0 is false.**

The probability of Type II error is calculated as the area of the sampling region under the real parameter value that falls in the acceptance region.



Type II Error (False Negative)

Controlling the value of β

The power β of a test is governed by several factors:

- **Controllable factors**: significance level α, size of the sample;
- **Uncontrollable factors**: real value of the parameter, variance of the data;

In general, it is possible to estimate the power of a test for a target desired difference $|H_0 - H_1|$.

The interpretation of this calculation is "If the difference between the real value and H_0 is at least $|H_0 - H_1|$, the probability of a type II error is β"

Considerations on Inference Errors

Type I Error (α) **– significance**: depends only on the distribution of the null hypothesis – easier to control;

Type II Error (β) **– power**: depends on the real value of the parameter – more difficult to specify and control;

The two conclusions of a test of hypothesis are normally considered as follows:

- Rejection of H_0 (easy to control the error) - strong conclusion;
- Failure to Reject H_0 (hard to control the error) - weak conclusion;

It is important to remember that failing to reject H_0 does not mean that there is evidence in favor of H_0! – it only suggests that H_1 is not better than the normal assumptions (H_0).

Hypothesis testing

General Procedure

1. Identify the parameter of interest;
2. Define H_0 and H_1 (also define if the test is one-sided or two-sided);
3. Determined desired values for α and β;
4. Define minimal interesting effect size δ^*;
5. Calculate the sample size;
6. Determine the test statistic and the critical region (for rejecting H_0);
7. Obtain the data from the experiment and calculate the value of the test statistic;
8. Decide whether or not to reject H_0;

Hypothesis Testing Example 1

Estimate the mean of a normal distribution, the variance known

For a certain brand of peas, we want to determine if there is any significant deviation in the mean weight of sacks from an advertised amount. Assume (for now) that the true variance of the process is known. The test hypotheses are defined as:

- $H_0 : \mu = 50$kg
- $H_1 : \mu \neq 50$kg (Two sided test!)

Let the desired significance level be $\alpha = 0.05$.

Given these characteristics, we expect that the sampling distribution of $\bar{x}$ is normal, with $\text{Var}(\bar{x}) = \sigma^2/n$ and, **if H_0 is true**, a mean of $\mu_{\bar{x}} = \mu_0 = 50$;

Hypothesis Testing Example 1

Estimate the mean of a normal distribution, the variance known

Given these characteristics, we can define the test statistic Z_0:

$$Z_0 = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}}$$

Under the null hypothesis, the value of Z_0 follows a standard normal $N(0, 1)$. With probability $1 - \alpha$, the value of Z_0 will fall between the $\alpha/2$ and $1 - \alpha/2$ quantiles of N(0,1):

$$P(z_{\alpha/2} \leq Z_0 \leq z_{1-\alpha/2} | H_0 \text{ is true}) = 1 - \alpha$$

This results allows us to calculate the critical zone for H_0 and H_1:

- If $z_{\alpha/2} > Z_0$ or $z_{1-\alpha/2} < Z_0$: **reject** H_0, with confidence $(1 - \alpha)$
- If $z_{\alpha/2} \leq Z_0 \leq z_{1-\alpha/2}$: **fail to reject** H_0: There is not enough evidence to justify H_1.

Hypothesis Testing Example 1

Putting Numbers in the Equation

Assume that we took $n = 10$ observations, and obtained the mean estimate $\bar{x} = 49.65$kg. Assume too that we know that the variance is $\sigma = 1$kg. We calculate z_0 as:

$$z_0 = \frac{49.65 - 50}{1/\sqrt{10}} = -1.113$$

The critical values for the standard normal distribution at the significance level $\alpha = 0.05$ are $[z_{0.025}, z_{0.975}] = [-1.96, 1.96]$;

In this case, because the value of z_0 is inside the non-rejection interval, so we conclude that **the data does not support rejecting H_0 at the 95% confidence level.**

Hypothesis Testing Example 2

Mean of a normal distribution, variance unknown

Suppose now a more realistic situation, in which the real variance is unknown. Assume also that we want to be more conservative, so we pick a significance level of $\alpha = 0.01$.

The test hypotheses remain the same:

- $H_0 : \mu = 50\text{kg}$
- $H_1 : \mu \neq 50\text{kg}$

In this case, **if H_0 is true**, we have that

$$T_0 = \frac{\bar{x} - \mu_0}{s/\sqrt{n}} \sim t^{(n-1)}$$

where s is the sample error, and t^d is a *Student's t distribution* with d degrees of freedom.

Hypothesis Testing Example 2

Mean of a normal distribution, variance unknown

From the same data used in the example 1, $\bar{x} = 49.65$, $n = 10$, $s = 0.697$

$$t_0 = \frac{49.65 - 50}{0.697/\sqrt{10}} = -1.597$$

The critical value of this test statistic for the desired significance is $t_{\alpha/2}^{(n-1)} = t_{0.005}^{(9)} = -3.24$, which means that under H_0, there is a 99% chance that the test statistic will give a value that is greater than -3.24, and smaller than 3.24.

Given that $-3.24 < t_0 < 3.24$, we conclude that the evidence from the sample is insufficient to reject H_0 at the 99% confidence level.

Hypothesis Test: How to calculate?

You do not need to calculate test statistics by hand. Many programming tools can do this calculation for you. For example, let's repeat the calculation of last slide in R:

```
> my.sample <- read.table("rawdata/greenpeas.txt")
> t.test(my.sample, mu = 50, conf.level = 0.99)

One Sample t-test

data:  my.sample
t = -1.5969, df = 9, p-value = 0.1447
alternative hypothesis: true mean is not equal to 50
99 percent confidence interval:
 48.93166 50.36434
sample estimates:
mean of x
   49.648
```

Part III.a – Interpretation of the Statistical Test

What is the result of a statistical test?

The result of the statistical test procedure can be reported as:

> *Sufficient (or insufficient) evidence for rejecting H_0 at the significance level α.*

This report is correct, but not very informative:

- How strong is the evidence for rejection/non-rejection?
- The significance level is fixed and predetermined. Is it the best one?
- If H_0 is rejected, big is the difference observed? ("magnitude of the effect size")
- How sensitive is the test to effect sizes?

Hypothesis testing and the p-value

p-value: *The lowest significance level α that would lead to the rejection of H_0 for the available data.*

The p-value can be used to obtain more information about a statistical hypothesis test. It is the probability, under H_0, that the test statistic would assume a value at least as extreme as the one observed.

For the previous example, the p-value would be calculated as:

$$p = 2 * P(t_0 \leq -1.597 | H_0 = \text{TRUE}) = 2 * \int_{-\infty}^{-1.597} t^{(9)} dt = 0.1447$$

So, to reject H_0 in this experiment, we would have needed significance $\alpha = 0.1447$.

One interpretation of the p-value is "How surprised we are to see this result under H_0".

p-value problems

a priori definition of significance level

The p-value calculates the smallest α that would be necessary to reject H_0.

A wrong conclusion is that deciding α before the experiment is not necessary: We could just evaluate the strength of the conclusion from the p-value!

In reality, it is still important to define de desired α during the experiment design. This avoids the moving goal posts problem, where we decide what is a good result **after** we see the result.

p-value problems

number of repetitions and p-hacking

It is possible to inflate the p-value artificially by increasing the size of n.

Suppose an experiment where $H_0 : \mu = 500$, $H_1 : \mu \neq 500$.
Sample size is $n = 5000$, sample average $\bar{x} = 499$, error $s = 5$. p-value calculation:

- $t_0 = -14.142$
- $p = 1.02 \times 10^{-23}$

The p-value is minuscule, but the difference between sample mean and H_0 is **smaller than the error**! Is this result meaningful?

In CS, it is very easy to artificially inflate the p-value using multiple simulations.

p-hacking is bad. Don't do it.

Using the p-value responsibly

Significance and effect sizes

To "tell the whole story" of the experiment, it is necessary to use **effect size estimators** together with the tests of statistical significance.

While there are whole books on the subject[1], the main idea is quite simple: to quantify the magnitude of the observed deviation from the null hypothesis.

Examples of effect size estimators include the simple **point estimator for the difference** $\bar{x} - \mu_0$, or the dimensionless d **estimator**:

$$d = \frac{\bar{x} - \mu_0}{s}.$$

Alternatively, **report confidence intervals** together with p-values in your results!

[1] See, for instance, Paul D. Ellis' *"The Essential Guide to Effect Sizes"*, Cambridge University Press, 2010

Part III.b – Model Validation

Assumptions of the Null Hypothesis Statistical Testing

Notice that the *NHST* approach adpts a number of assumptions, both statistical and technical:

- **The mean is a good measure for the question of interest.**
  (i.e., the variance is small enough, the weigths of packages are indepentent, customers usually purchase many packages, so individual extreme values are not important, etc);
- **The sample is representative of our population of interest.**
  (i.e., the packages are from regular production (not specially produced for this test), they are not tampered with, etc)
- **The contents of the packages are actually chocolate**
  (the weight of the package is not a significant part of the measured weight, etc);
- ... and others ...

Assumptions of the Null Hypothesis Statistical Testing

Statistical Assumptions

More specifically, the testing and analysis procedure that we studied (calculation of test statistic, etc), assumes that our experiment can be described by a specific model:

- The sample distribution follows a normal curve (Assumption of Normality);
- The observations in the sample are independent (Assumption of Independence);
- The variance is constant (Assumption of variance);
- etc..

It is important to validate these assumptions, to make sure that our test, analysis and conclusions are valid!

Assumption of Normality

The assumption of Normality is required by the **z** and **t** tests described in this lecture:

"*The Assumption of Normality (note the upper case) that underlies parametric stats does not assert that the observations within a given sample are normally distributed, nor does it assert that the values within the population (from which the sample was taken) are normal. This core element of the Assumption of Normality asserts that* **the distribution of sample means (across independent samples) is normal**."

J. Toby Mordkoff, 2011[a]

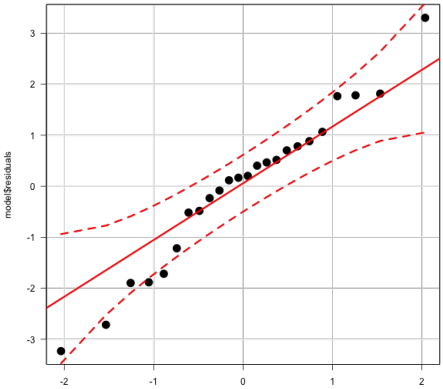
[a] J.T. Mordkoff, The assumption(s) of normality: `http://goo.gl/Z3w8ku`

The normality assumption

Visual Inspection

If we cannot assume the conditions for the CLT *a priori*, then we can perform normality tests on the data.

The **QQ plot** (quantile-quantile plot) plots the quantiles of two data sets against each other. If one of the data-sets is the theoretical normal quantiles, this plot can help visualize deviations from normality.



The normality assumption

Normality Tests

You can also perform statistical tests on the assumption of normality:

- **Shapiro-Wilk;** **<- recommended for this course;**
- Anderson-Darling;
- Lilliefors / Kolmogorov-Smirnov;

These procedures use different aspects of the sample distribution to test the following hypotheses:

- H_0: The population is normal;
- H_1: The population is not normal;

In this case, rejection of the null hypothesis suggests evidence that the **sample** came from a non-normal distribution. Although, for a large enough sample, the CLT might still guarantee a normally distributed **sample mean estimate**, a visual investigation of the distribution of sample's observations is very important in this case.

The independence assumption

The strongest assumption used for the t-test is the **independence assumption**. This assumption means that the value of observations are not dependent (biased) on the values of other observations.

Example of independence violation:

- You measure the speed of a robot in 10 trials. However, because the battery is low, the speed will progressively decay;
- You measure the accuracy of an algorithm in predicting 20 time series curves. However, 5 of those curves represent different instances of the same model, and are closely related to each other.

In general, we want to guarantee the independence assumption through careful experiment design. The **Durbin-Watson** test can be used to detect auto-correlation, but it is sensitive to the order of observations.

Part IV – Outro

Conclusion: A framework for statistical testing

In this lecture, we introduced the concept of "hypothesis testing" as a way to use data obtained from an experiment to make conclusions about a population. Let's think back to the steps of this procedure:

- Formulate the question of interest, and define the hypotheses;
- Define the minimally interesting effect;
- Define desired confidence and power for the test;
- Calculate required sample size; **<- Future Lecture**
- Collect the data;
- Perform Statistial Analysis, and validate the assumptions;
- Draw conclusions and recommendations;

In future lectures, we will study variations and special cases of this testing procedure;

Recommended Reading

- University of Guelph: "Statistical Significance vs Practical Significance: A tutorial."https://atrium.lib.uoguelph.ca/xmlui/bitstream/handle/10214/1869/A_Statistical_versus_Practical_Significance.pdf?sequence=7
- J.T. Mordkoff, "The Assumption(s) of Normality", 2016 http://www2.psychology.uiowa.edu/faculty/mordkoff/GradStats/part%201/I.07%20normal.pdf

Florence Nightingale

1820-1910 – "The Lady with the Lamp"

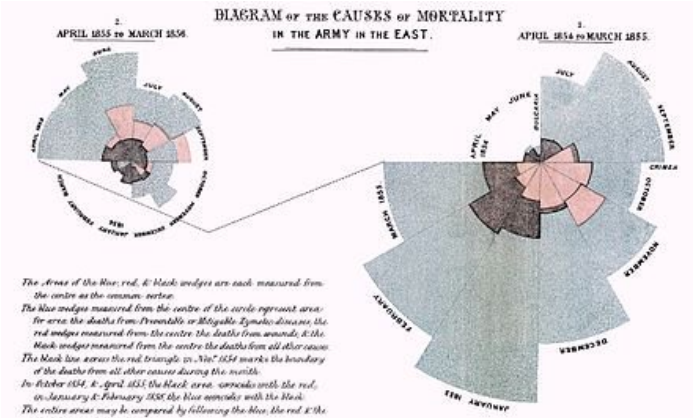
Let's talk about a scientist who made great contributions to evidence-based medicine and descriptive statistics: **Florence Nightingale**.

- British nurse and mathematician;
- Born in 05/12/1820, her parents were opposed to her careers;
- She was driven, a prolific writer, and knew several languages;
- Gave great contributions for the professionalization of nursing;

Florence Nightingale

Descriptive Statistics in Health

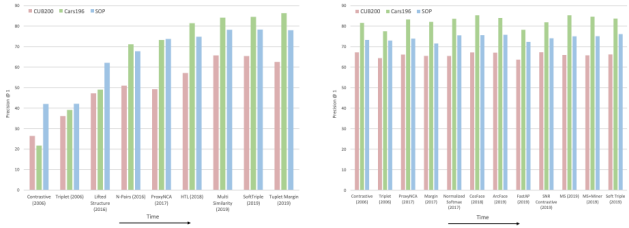
- Implemented the use of **hand washing** in hospitals for nurses:
- Pioneer of using of data visualization (infographics!) in medicine;



Experiment Design: Fair Comparisons

A sombering example

Musgrave et al (preprint): several ML methods for metric learning perform exaclty the same when the hyperparameters are properly tuned for all methods.



(a) The trend according to papers

(b) The trend according to reality

Figure from Musgrave et al. "A Metric Learning Reality Check"

Fair comparisons will help you avoid false conclusions!

Experiment Design: Fair Comparisons

What are fair comparisons?

The definition of a **fair** comparison, of course, depends on the field being studied and the experiment being conducted. In the comparison of algorithms in computer science, we can think of some points:

- Fine-tuning of algorithmic parameters;
- Discarding failed variations;
- Fine-tuning of the algorithm itself on the training data;
- Only comparing on data favorable to one of the algorithms;
- Coding with modern libraries vs old algorithms;
- Different computational environments;
- etc...

About these Slides

These slides were made by Claus Aranha, 2021. 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 following references for those cases.

Image Credits I

[Page 4] Cocoa image from https://www.irasutoya.com
[Page 5] Cocoa image from https://www.irasutoya.com
[Page 48] Figure from Musgrave et al. "A Metric Learning Reality Check"
https://arxiv.org/pdf/2003.08505.pdf