Experiment Design for Computer Sciences (01CH740)

Topic 03 - Statistical Inference

Claus Aranha
caranha@cs.tsukuba.ac.jp

University of Tsukuba, Department of Computer Sciences

May 15th, 2020

Version 2020.2 (Updated May 14, 2020)

Lecture 3: Statistical Inference

- In the last lecture, we described **descriptive statistics**, such as point and interval estimators.
- Point and interval estimators are very useful to define the value of parameters in the population, and to estimate their margins of errors.
- However, in some cases, we need **decision making tools**, in order to deal with information from random samples.
- **Statistical Inference** is one of these tools that can help us make a decision based on a random sample.

Statistical Inference

Motivating Example

Imagine you are the owner of a factory that produces delicious chocolate. Its packages should contain 300g of chocolate.

You take a sample of **30 packages** produced by your factory. Using the tools from lecture 2, you calculate the mean weight with a 95% confidence interval of **283g to 307g**.

Is everything normal? Or should you investigate the production line of your factory?

Remember!

The confidence interval does not say whether **any of the values inside it** are more or less likely than any other!

Florence Nightingale

1820-1910 – "The Lady with the Lamp"

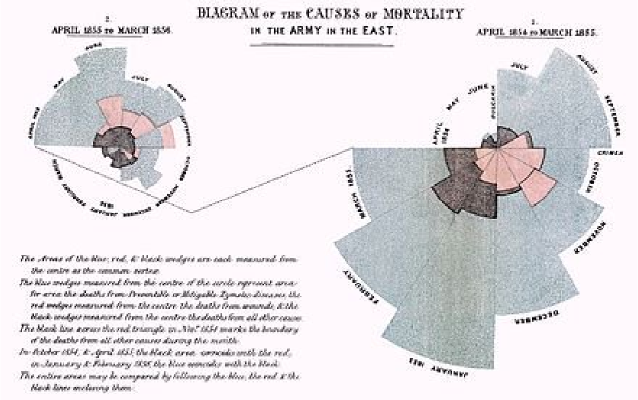
Let’s talk about a scientist who made great contributions to evidence-based medicine and descriptive statistics: **Florence Nightingale.**

- 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

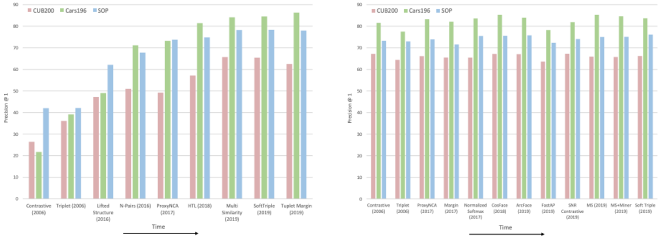
- Implemented the use of **hand washing** in hospitals for nurses:
- Pioneer of using of data visualization (infographics!) in medicine;



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;

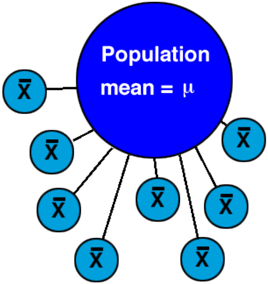
- Coding with modern libraries vs old algorithms;
- Different computational environments;
- etc...

Lecture 3: Outline

- What is Statistical Inference?
- Statistical Hypothesis and Errors;
- Z testing on a single sample;
- Statistical Testing Assumptions;

What is Statistical Inference?

Statistical inference is the process of using data analysis to deduce properties of an underlying statistical distribution.



Remember that a sample from a distribution is a *random variable*, defined by a *sampling distribution*.

The sampling distribution is characterized by parameters from the distribution. So when we deduce properties of the sampling distribution, we can use these properties to characterize the underlying population.

Sample data → Sampling distribution → Population parameters

What is a hypothesis?

A key tool of statistical inference is the **Statistical Hypothesis**:

Many meanings of hypothesis

Hypothesis: a proposed explanation to an observable phenomenon.
Scientific Hypothesis: must be *Testable* and *Falsifiabe*;
Statistical Hypothesis: explanation focus on statistical statements;

Example: Your cocoa factory is working normally, so the mean of its production is *no less than 300g*.

- **Testable**: analyze a produced sample;
- **Falsifiable**: estimated mean is much below 300g

Key question: how much is "*much below 300g*"?

Comparing Multiple Hypotheses

In general we can propose several hypothesis for the same observation. We compare these hyposeses against each other, and try to decide which one is **more likely to be true** based on the data.

Cocoa Sample (sample mean: 294)

293 325 271 313 309 298 284 304 248 296

Hypotheses:

- Sample mean under 300 is bad luck. Factory mean is $\approx$ 300g.
- Factory mean is than 300g. Factory has a problem.
- An evil employee sabotaged the sample, choosing bad packages.
- The balance I used to measure the sample is broken.
- Factory production depends on the hour of the day.

Comparing Multiple Hypotheses

Comparison criteria

When comparing multiple hypothesis, we want to keep several criteria in mind:

- **Predictive power**: The hypothesis allows you to predict future behavior of the system.
- **Principle of parsimony** (Ockham's razor): The hypothesis makes few assumptions about the system.
- **Fitting the data**: The data supports the hypothesis and has a high probability of being produced under it.
- **External consistency**: The hypothesis fits with existing, well accepted knowledge about the system.

Multiple hypotheses and statistical testing

One way that we compare multiple hypotheses is by calculating the probability of the observed sample under each competing hypothesis. In this sense, we give more credibility for the hypothesis that maximizes the probability of the observed sample.

- $P(\bar{x}|H_1), P(\bar{x}|H_2), P(\bar{x}|H_3), \ldots$

Hypothesis 1: $\mu \geq 300$

What is the probability that we see the sample: $X = \{293, 325, 271, 313, \ldots\}$ when the mean production of the factory is at least 300g?

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

What is the probability that we see the sample: $X = \{293, 325, 271, 313, \ldots\}$ when the mean production of the factory is much lower than 300g?

Null and alternate hypothesis

The *Null Hypothesis Significance Testing (NHST)* approach involves the contrast between a **null hypothesis** and an **alternative hypothesis**.

Null Hypothesis (H_0)

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

"The mean production of the chocolate factory is at least 300g"

$H_0 : \mu \geq 300$

Alternative Hypothesis (H_1)

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

"For some reason, the mean production is below 300g"

$H_1 : \mu < 300$

Null Hypothesis Testing

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:

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:** $H_1 : \mu < 300$

Null Hypothesis Testing

Testing Assumptions

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 of production is small enough, the weigths of items are indepentent, customers usually purchase many packages, so individual extreme values are not important, etc);
- The packages sampled are 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);
- etc ...

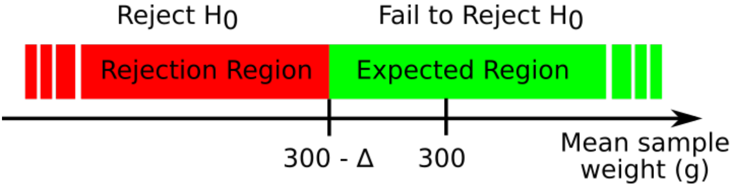
Null Hypothesis Testing

Testing Procedure

1. Obtain a sample (i.e. running the experiment);
2. Calculate test statistics from the sample data;
3. Make a decision based on the computed value;

As the sample mean ($\bar{X}$) is a good estimator for the population mean (μ), the decision to reject or not the null hypothesis could be made based on the difference Δ between the sample mean and μ_{H_0}.

But how do we define this **critical region** for rejecting H_0?

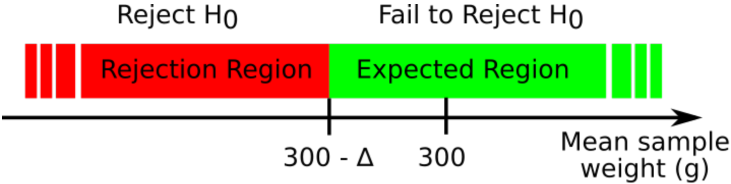


Inference Errors

Remember that a parameter estimator from a sample is a random variable, with an **estimator error** associated with it. Because of this, there is a chance that the hypothesis test reaches a wrong conclusion.

- If the estimatior error is too large, the sample mean could be in the rejection region, even if the null hypothesis is true;
- If Δ is too large, the sample mean could be in the expected region, even if the null hypothesis is not true;

We want to be able to estimate (and control!) the probability of these errors.



Type I Error

Type I Error (False Positive):
Reject the Null Hypothesis when it is true

The probability of occurrence of a false positive in any hypothesis testing procedure is generally known as the significance level of the test, represented by the Greek letter α:

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

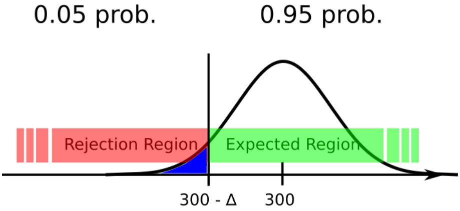
Another frequently used term is the confidence level of the test, which is given by $(1 - \alpha)$ or $100(1 - \alpha)\%$

Type I Error

Choosing α and the rejection region

For a given sample, the value selected for α defines the threshold for the rejection of H_0. If H_0 is true (i.e., $\mu = 300$g), we can expect that the distribution of mean estimates is approximately normal[1], with average 300g and standard error $(\sigma/\sqrt{n})$.

To control the probability of Type I error (e.g. $\alpha = 0.05$), we select the critical region so that the cumulative probability under the distribution of $\hat{x}$ inside that region is $1 - \alpha = 0.95$.



[1] Assuming the CLT conditions are met

Type II Error

Type II Error (False Negative):

Fail to reject the Null Hypothesis when it is false

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

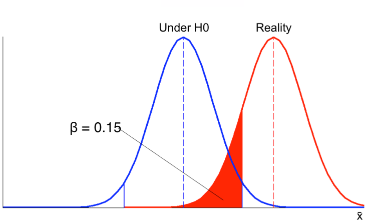
$$\beta = P(\text{type II error}) = P(\text{not reject } H_0 | H_0 \text{ is false})$$

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.

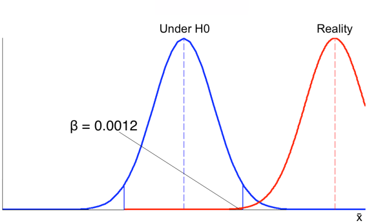
Type II Error

Interpretation

Unlike the Type-I error, the definition of the error rate for the Type-II error requires the specification of the value of the parameter of interest under the alternative hypothesis.



The probability of failing to reject a false H_0 strongly depends on the magnitude of the difference between the value of the parameter under H_0 and the real value of the parameter.



Type II Error
Controlling β

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

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

If H_0 is false, a smaller magnitude of the difference between the real value of the parameter and the one under the null hypothesis leads to a greater probability of a type II error. **On the other hand, the practical importance of the error gets smaller.**

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

Considerations on Inference Errors

Type I Error (α): depends only on the distribution of the null hypothesis – easier to control;
Type II Error (β): depends on the real value of the parameter – more difficult to specify and control;

These characteristics lead to the following classification of the conclusions obtained from a test of hypotheses:

- Rejection of H_0 - strong conclusion;
- Failure to Reject H_0 - weak conclusion (but we can strengthen it);

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 it is a better model than the alternative proposed.

Hypothesis testing

General Procedure

1. Identify the parameter of interest;
2. Define H_0 and H_1 (one- or two-sided);
3. Determined desired α and β;
4. Define minimal interesting effect size δ^*;
5. Calculate sample size;
6. Determine the test statistic and critical region;
7. Compute the statistic;
8. Decide whether or not to reject H_0;

Hypothesis Testing Example 1

Mean of a normal distribution, 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:

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

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

Mean of a normal distribution, variance known

Given these characteristics, we define a standardized random variable:

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

The values of Z_0 will be distributed, **under the null hypothesis**, according to a standard normal, $N(0, 1)$.

This implies that the probability of the value Z_0 to be between the $\alpha/2$ and $1 - \alpha/2$ quantiles of N(0,1) is $1 - \alpha$:

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

This results allows us to calculate a 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 $(1 - \alpha)$
- Else, there is not enough evidence to reject H_0;

Hypothesis Testing Example 1

Mean of a normal distribution, variance known

Assume that we took $n = 10$ observations, and obtained the mean estimate $\bar{x} = 49.65$kg. Assume too that we know that $\sigma = 1$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, 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 *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 Testing Example 2

Mean of a normal distribution, variance unknown

You can explore these values and statistic parameters in R:

```
> my.sample <- read.table("greenpeas.txt")

> t.test(my.sample,
+         alternative = "less",
+         mu = 50,
+         conf.level = 0.99)

One Sample t-test
data: my.sample
t = -1.5969, df = 9, p-value =0.07237
alternative hypothesis: true mean is less than 50
99 percent confidence interval:
-Inf 50.2699
sample estimates:
mean of x
49.648
```

Describing the results of a Hypothesis Test

(In)Sufficient evidence for rejecting H_0 at the significance level α.

Even though this description is correct, it is relatively poor:

- It does not provide information on the intensity of the evidence for rejection/non-rejection;
- It imposes a predetermined significance level to the reader;
- It does not provide information on the maginitude of the effect observed, or the sensitivity of the test.

The p-value

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

We can interpret the p-value as the probability, under H_0, that the test statistic assume a value at least as extreme as the one obtained.

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

$$p = P(t_0 \leq -1.597 | H_0 = \text{TRUE}) = \int_{-\infty}^{-1.597} t^{(9)} dt = 0.07237$$

One interpretation of the p-value is "How surprised we are to see this result under H_0". It quantifies the strength of rejection of the null hypothesis. However, a-priori definition of the significance level during the experiment design stage is still important!

The p-value

Significance and effect sizes

Remember that we can adjust the rejection area for the α by changing the number of observations (n) of the sample. This means that a p-value can be made arbitrarily small if n is big enough;

Suppose that a test for $H_0 : \mu = 500$ against $H_1 : \mu \neq 500$, with $n = 5000$, $\bar{x} = 499$, $s = 5$. In this case, we calculate the p-value:

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

The p-value is significant, but is the difference **meaningful?**

In experiments where the data comes from computer experiments, it can be very easy to inflate the p-value.

The p-value

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[2], 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**:

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

which quantifies the difference in terms of sample standard deviations.

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

The p-value

Effect sizes and confidence intervals

> *Point estimators + confidence intervals quantify the magnitude and accuracy of effects, and must be reported alongside the results of significance testing whenever possible.*

Suppose we are testing $H_0 : \mu = 50$ against $H_1 : \mu \neq 50$, with $n = 10$ and $\alpha = 0.01$. Assume also that the population is known to be normal, with unknown variance. Using the same data as before:

```
> t.test(my.sample, mu = 50, conf.level = 0.99)
(...)
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
```

Model Validation

As we mentioned before, the procedure of statistical testing includes several assumptions (technical and statistical) regarding the model being used:

- Assumption of Nomarlity;
- Assumption of Independence;
- Assumption on the value of variance;
- Assumptuons about the process;
- etc..

It is necessary to validate the assumptions to avoid bad surprises. The statistical assumptions can usually be validated through analysis of the sample data.

The normality assumption

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]

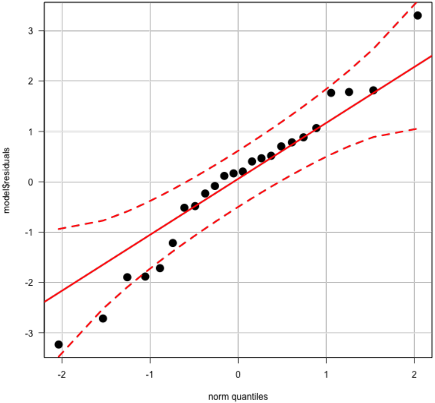
[a] J.T. Mordkoff, The assumption(s) of normality: `http://goo.gl/Z3w8ku`

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.

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 Statistical 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

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 3] Cocoa image from https://www.irasutoya.com
[Page 6] Figure from Musgrave et al. "A Metric Learning Reality Check" https://arxiv.org/pdf/2003.08505.pdf