From Statistics to Data Mining

Master 1
COlour in Science and Industry (COSI)
Cyber-Physical Social System (CPS2)
Saint-Étienne, France

Fabrice MUHLENBACH
https://perso.univ-st-etienne.fr/muhlfabr/
e-mail: fabrice.muhlenbach@univ-st-etienne.fr

Statistics

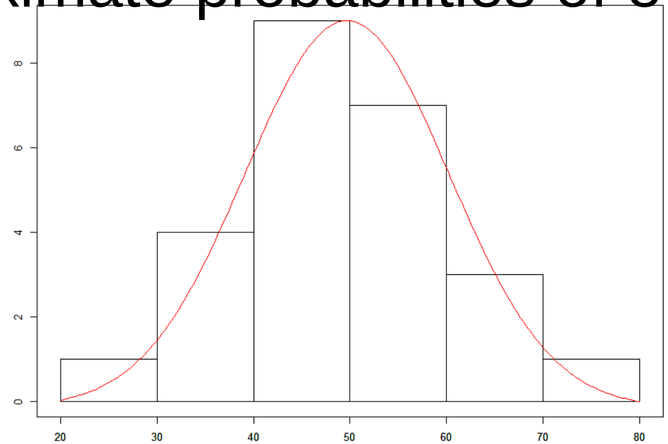
- Probability and Statistics

- ➢ Reminder: **Probability** is the branch of mathematics concerning numerical descriptions of how likely an event is to occur, or how likely it is that a proposition is true
  → random variables → probability distributions
- ➢ **Statistical methods** used in data analysis and data mining:
- ❑ **descriptive statistics**: summarize data from a sample using indexes such as the mean or standard deviation
- ❑ **inferential statistics**: draw conclusions from data that are subject to random variation → inferences are made under the framework of probability theory, which deals with the analysis of random phenomena

Statistics

- Normal Distribution and Discrete Random Variables

- often, a probability histogram can be well approximated by a normal curve (application which will be seen during tutorials)
- in such cases, it is customary to say that X has approximately a normal distribution
- the normal distribution can then be used to calculate in a simple way approximate probabilities of events involving X



Statistics

- Central Limit Theorem

➢ Let $X_1, \ldots, X_n$ be a random sample of size n –that is, a sequence of n independent and identically distributed (*i.i.d.*) random variables drawn from any distribution (not itself normal) of expected value μ and standard deviation σ

➢ If n is large, the sample average $\bar{X} = \frac{X_1,\ldots,X_n}{n}$ is approximately normally distributed of mean $\mu_{\bar{X}} = \mu$ and standard deviation $\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}}$

➢ therefore: $\frac{1}{\sqrt{n}}\left(\frac{X_1+X_2+\cdots+X_n-n\mu}{\sigma}\right) \rightarrow \mathcal{LG}(m=0\,;\sigma=1)$

➢ the Central Limit Theorem can safely be applied if $n \geq 30$

Statistics

- Estimation –Definitions

➢ In statistics, an **estimator** is a rule for calculating an estimate of a given quantity based on observed data:
→ thus the rule (the *estimator*), the quantity of interest (the *estimand*) and its result (the *estimate*) are distinguished

➢ **point** estimators → single-valued results, single vector-valued results, or results that can be expressed as a single function

➢ **interval** estimator → the result would be a range of plausible values (or vectors or functions)

➢ Quantified properties of the estimation:
→ error, mean squared error (MSE), sampling deviation, variance, bias...

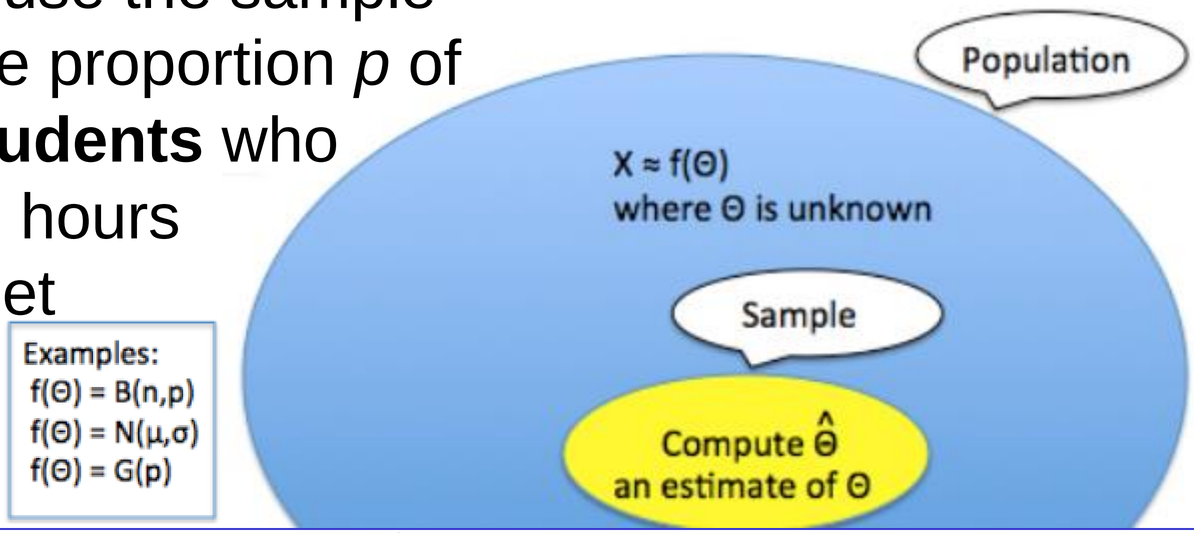
Statistics

- Estimation –Case Study

➢ Most American college students make use of the Internet for both academic and social purposes

➢ A study is done from a sample of 7421students at 40 colleges and universities

➢ The objective is to use the sample data to estimate the proportion *p* of **all U.S. college students** who spend more than 3 hours a day on the Internet



• Estimation –Case Study

- ➢ Choosing a Statistic for Computing an Estimate:
- ➢ X = 2998 of the n = 7421 students spend more than 3 hours per day on the Internet
- ➢ We can use this information to estimate the unknown population proportion p
- ➢ The statistic

$$\hat{p} = \frac{\text{Number of successes in the sample}}{\text{Size of the sample}} = \frac{2998}{7421} = 0.402$$

is an obvious choice for obtaining a point estimate of p

- ➢ **How to assess the quality of the estimate $\widehat{p}$?**

Statistics

- Estimation using a single sample

- ➢ The objective of inferential statistics is to use sample data to decrease our uncertainty about some characteristic of the corresponding population, such as a population mean μ or a population proportion p
- ➢ A statistic whose mean value is equal to the value of the population characteristic being estimated is said to be an **unbiased statistic**
- ➢ A statistic that is not unbiased is said to be **biased**
- ➢ More formally, a statistic $\hat{\theta}$ is an unbiased estimate of the population characteristic θ iff: $E\left(\hat{\theta}\right) = \theta$

Statistics

- Estimation –Example of an unbiased statistic

- $\hat{p}$ seems to be an obvious choice for obtaining a good estimate of p
- **Proof:**
- We know that $\hat{p} = \frac{X}{n}$, where *X* is the number of successes among *n*
- The distribution of *X* is a binomial distribution of success probability *p* and *n* is the number of independent trials
- Therefore, $E(X) = n.p$
- We can then deduce that $E(\hat{p}) = E\left(\frac{X}{n}\right) = \frac{1}{n}E(X) = \frac{n.p}{n} = p$
- Therefore $\hat{p}$ is an unbiased estimate of p

Statistics

- Convergence in probability of an estimate

➢ Given several unbiased statistics that could be used for estimating a population characteristic, the **best choice** to use is the statistic with the **smallest standard deviation**

➢ An unbiased estimate $\hat{\theta}$ of θ converges in probability iff:

$$\lim_{n\to\infty} V\left(\hat{\theta}\right) = 0$$

➢ $\hat{p}$ converges in probability towards p:

$$\lim_{n\to\infty} V(\hat{p}) = \lim_{n\to\infty} V\left(\frac{X}{n}\right) = \lim_{n\to\infty} \frac{1}{n^2} V(X) = \lim_{n\to\infty} \frac{p(1-p)}{n} = 0$$

Statistics

- Unbiased estimate of a Mean

➢ Let X_1, ..., X_n a set of n i.i.d. random variables of (unknown) mean μ and standard deviation σ

➢ We define $\bar{X}$, the sample mean, as follows: $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$

$$E(\bar{X}) = E\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{1}{n}\sum_{i=1}^{n} E(X_i) = \frac{1}{n} \times n \times \mu = \mu$$

➢ Therefore $\bar{X}$ is an unbiased estimate of μ

➢ Moreover, $\lim_{n\to\infty} V(\bar{X}) = \lim_{n\to\infty} \frac{1}{n^2} V(X_i) = \lim_{n\to\infty} \frac{\sigma^2}{n} = 0$

➢ Therefore $\bar{X}$ converges in probability towards μ

Statistics

- Estimation –Summary

➢ **Sample proportion:**

❑ Let $\hat{p}$ be the proportion of successes in a random sample of size *n* from a population whose proportion of successes is *p*

❑ Then, the following holds:

- The mean value of $\hat{p}$ is $\mu_{\hat{p}} = p$
- The standard deviation of $\hat{p}$ is $\sigma_{\hat{p}} = \sqrt{\frac{p(1-p)}{n}}$
- When *n* is large, the distribution of $\hat{p}$ is approximately normal

Statistics

- Estimation –Summary

➢ **Sample mean:**

❑ Let $\bar{X}$ be the mean of *n* observations drawn from a population having mean μ and standard deviation σ

❑ Then, the following holds:

- The mean value of $\bar{X}$ is $\mu_{\bar{X}} = \mu$
- The standard deviation of $\bar{X}$ is $\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}}$
- When the population distribution is normal, the distribution of $\bar{X}$ is also normal for any sample size *n*

Statistics

- Maximum Likelihood Estimation

➢ Let x_1, ..., x_n be a random set of n i.i.d. observations, coming from an unknown density function $f(x|\theta)$ where θ is a parameter (e.g., p for the binomial distribution, μ and σ for the normal distribution, etc.)

➢ The likelihood of the set corresponds to the joint density function $f(x_1, \dots, x_n)$ for all observations

➢ For an i.i.d. sample, we get: $f(x_1, \dots, x_n|\theta)$

$$= f(x_1|\theta) \times f(x_2|\theta) \times \cdots \times f(x_n|\theta) = \prod_i f(x_i|\theta)$$

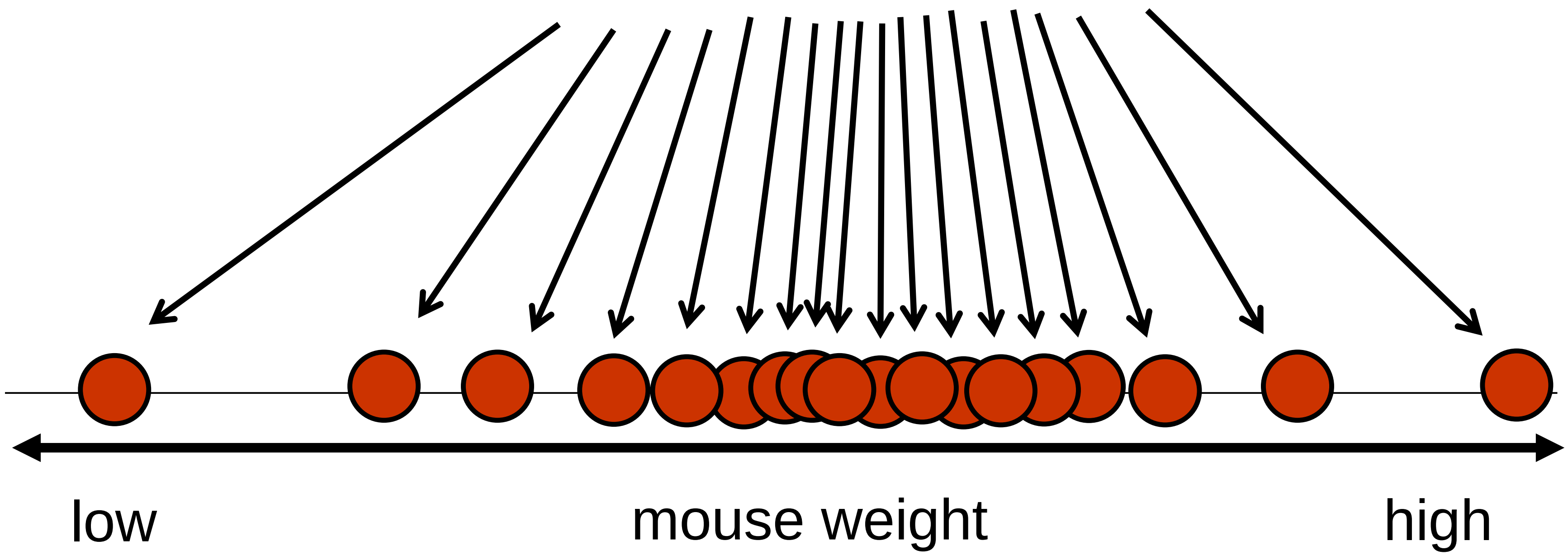
➢ The maximum likelihood estimation is to find an estimate $\hat{\theta}$ which would be as close to θ as possible

Statistics

- Maximum Likelihood Estimation –Example

➢ Experiment in biology → we weighted a bunch of mice

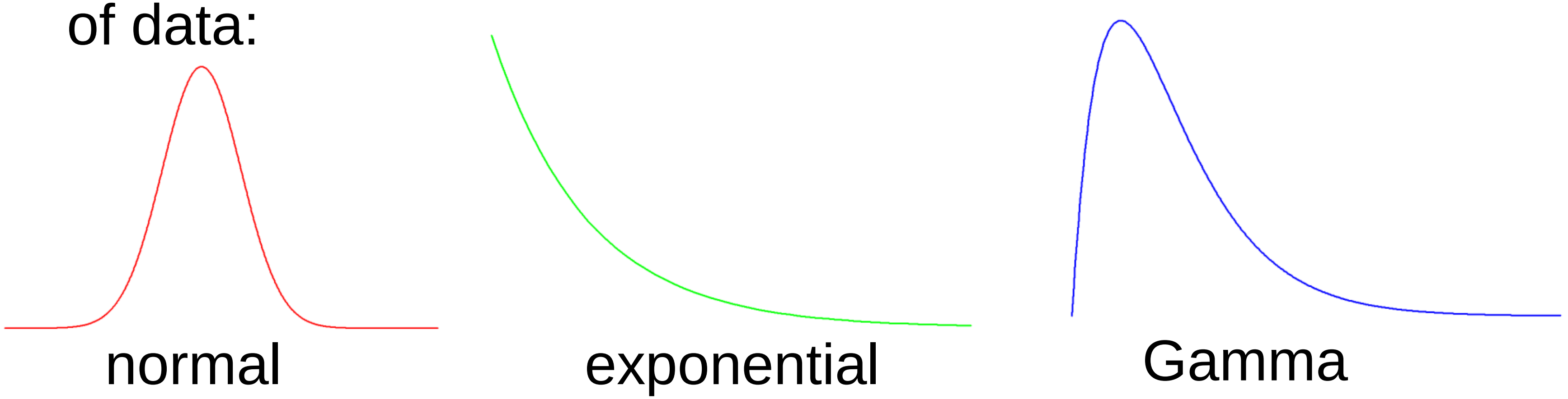
➢ the goal of maximum likelihood is to find the optimal way to fit a distribution to the data



Statistics

- Maximum Likelihood Estimation –Example

➢ Examples of different types of distributions for different types of data:

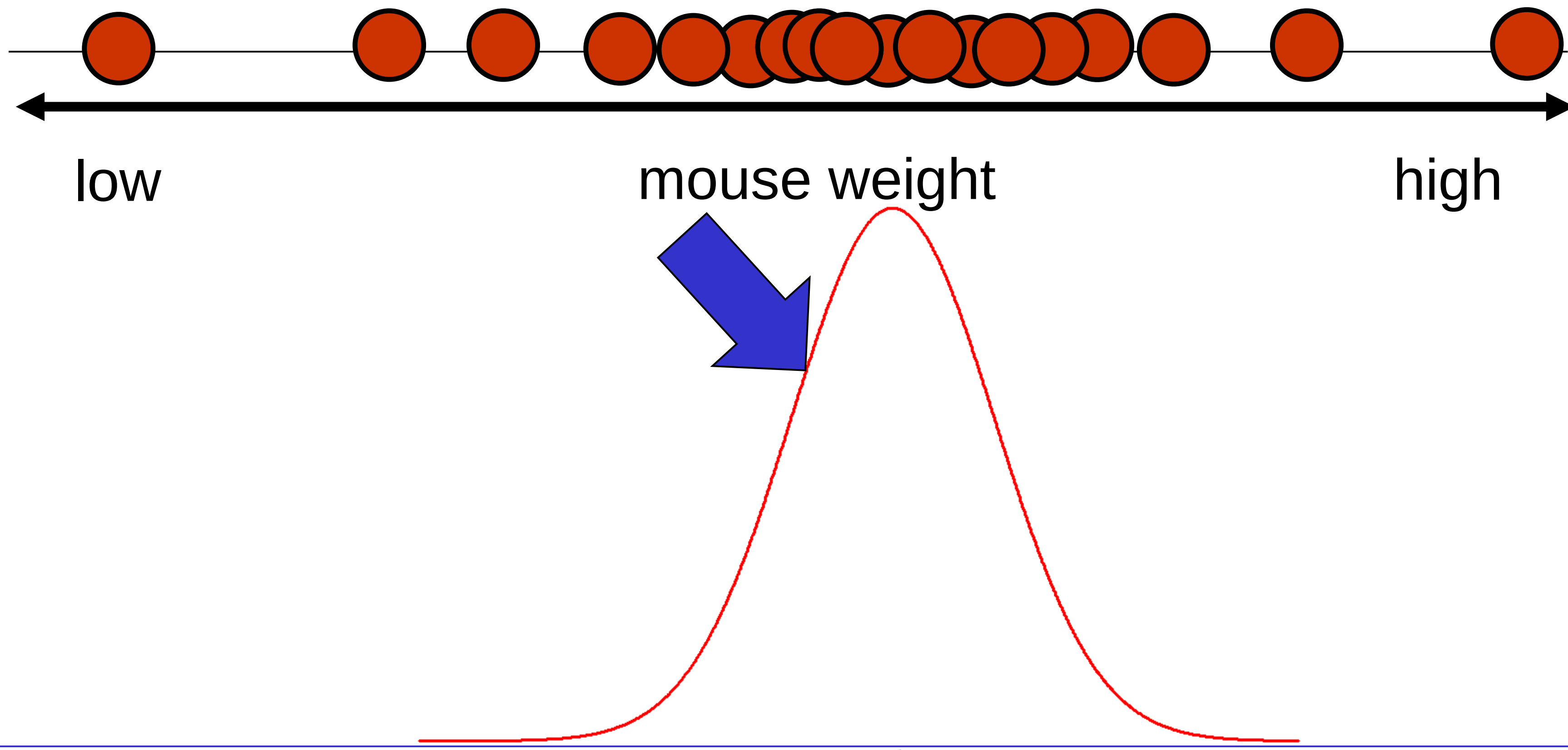


➢ The reason you want to fit a distribution to your data is it can be easier to work with

➢ A distribution is more general: it applies to every experiments of the same type

- Maximum Likelihood Estimation –Example

➢ Here, we think that the weights might be normally distributed



• Maximum Likelihood Estimation –Example

- “Normally distributed” means a number of things:

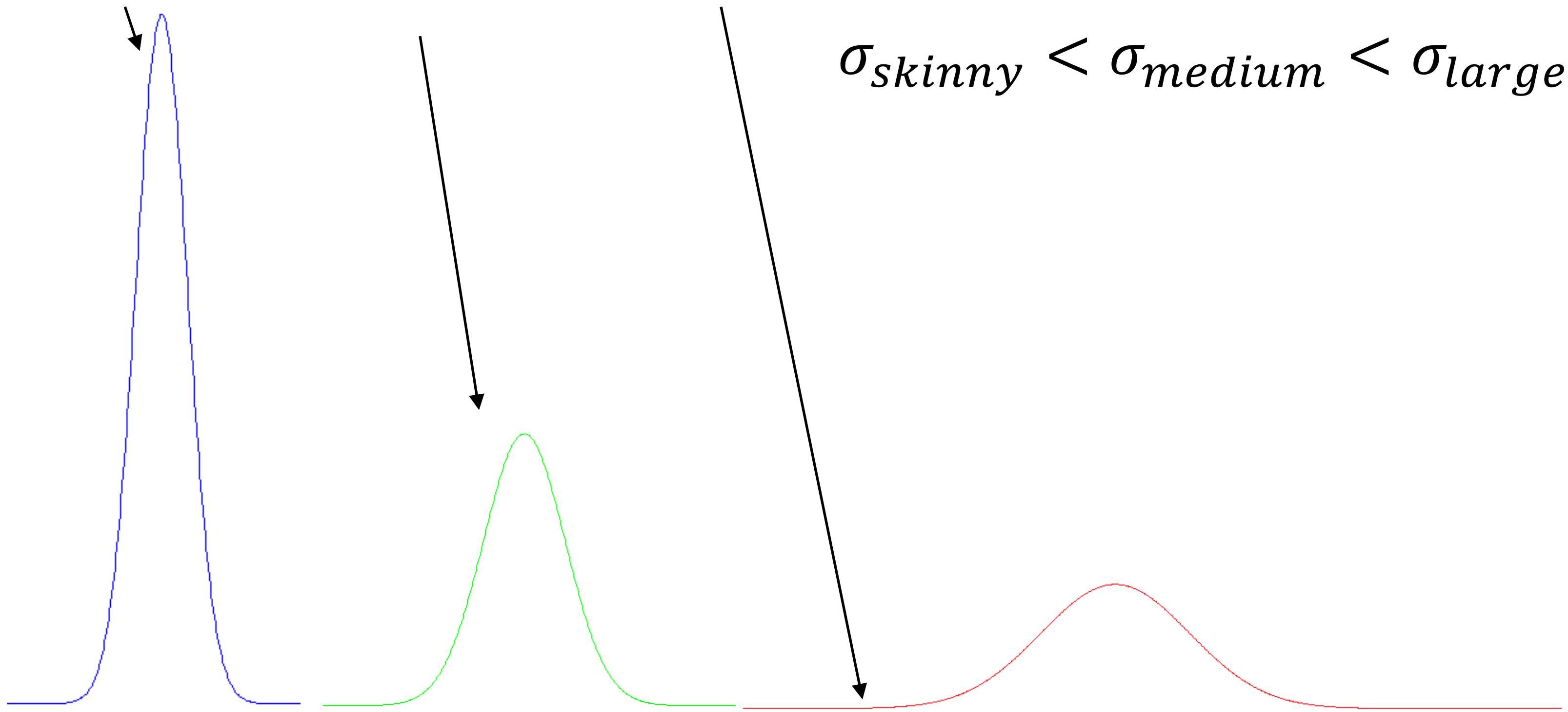
1. we expect most of the measurements (mouse weights) to be closed to the mean (average)
2. we expect the measurements to be relatively symmetrical around the mean

Statistics

- Maximum Likelihood Estimation –Example

➢ Normal distributions come in all kinds of shapes and sizes, e.g., "skinny", "medium" or "large" depending on the value of σ

$$\sigma_{skinny} < \sigma_{medium} < \sigma_{large}$$



Statistics

- Maximum Likelihood Estimation –Example

➢ For a normal distribution, we have to estimate 2 parameters:

1. the mean μ → location of the center of the distribution
2. the standard deviation σ → shape or size of the distribution

➢ It is possible to plot the likelihood of observing the data as a function of the location of the center of the distribution
→ we want the location that "maximizes the likelihood" of observing the weight we measured
→ maximum likelihood for the mean of the distribution

➢ In the same way, we can find the normal distribution that has been fit to the data by using the maximum likelihood estimations for the mean and the standard deviation

Statistics

- Statistical hypothesis testing

➢ A statistical hypothesis testing = a rule that indicates whether a statement about a population should be **accepted** or **rejected** based on the evidence provided by the data sample

➢ **principle**: establish two opposing hypotheses concerning a population → the null hypothesis & the alternative hypothesis

➢ **null hypothesis** (H_0) = statement tested: no difference/effect

➢ **alternative hypothesis** (H_1) = statement that we want to be able to conclude to be true from the evidence provided by the sample data on the basis of the sample data → the test determines whether or not to reject the null hypothesis based on the *p-value* and significance level threshold α

Statistics

- Statistical hypothesis testing

- ➢ if the p-value is less than the significance level (called α or *alpha*), we can reject the null hypothesis
- ➢ **warning**: statistical hypothesis tests do not aim to select the most probable hypothesis among two but to define a null hypothesis as being the one that we wish to reject
- ➢ by setting a low significance level before analysis (e.g., a value of 0.05), when we reject the null hypothesis, we have statistical evidence that the alternative is true
- ➢ on the other hand, if we fail to reject the null hypothesis, we do not have statistical evidence indicating that the null hypothesis is true (→ *p* at 5% too small)

Statistics

- Statistical hypothesis testing –Testing Process (1/2)

➢ There is an initial research hypothesis of which the truth is unknown

1. state the relevant **null** and **alternative** hypotheses
2. consider the statistical assumptions being made about the sample in doing the test; for example, assumptions about the statistical independence or about the form of the distributions of the observations
3. decide which test is appropriate, and state the relevant **test statistic *T***
4. derive the distribution of the test statistic under the null hypothesis from the assumptions

Statistics

- Statistical hypothesis testing –Testing Process (2/2)

5. select a significance level (α), a probability threshold below which the null hypothesis will be rejected (e.g., 5% or 1%)
6. the distribution of the test statistic under the null hypothesis partitions the possible values of T into those for which the null hypothesis is rejected and those for which it is not
7. compute from the observations the observed value t_{obs} of the test statistic T
8. decide to either reject the null hypothesis in favor of the alternative or not reject it. The decision rule is to reject the null hypothesis H_0 if the observed value t_{obs} is in the critical region, and to accept or "fail to reject" the hypothesis otherwise

Statistics

- Statistical hypothesis testing –Testing and Risks

- ➢ a fundamental notion regarding testing is the probability of being wrong
- ➢ there are two ways to go wrong in a statistical test:

1. reject the null hypothesis when it is true → error of the first kind, or “Type I” error α
2. retain the null hypothesis when it is false → Type II error β

- ➢ we try to minimize these errors but, in practice, a compromise must be found between these two types of error
- ➢ the probability 1 - β of choosing the alternative hypothesis H_1 rightly is called “the power of the test”

Statistics

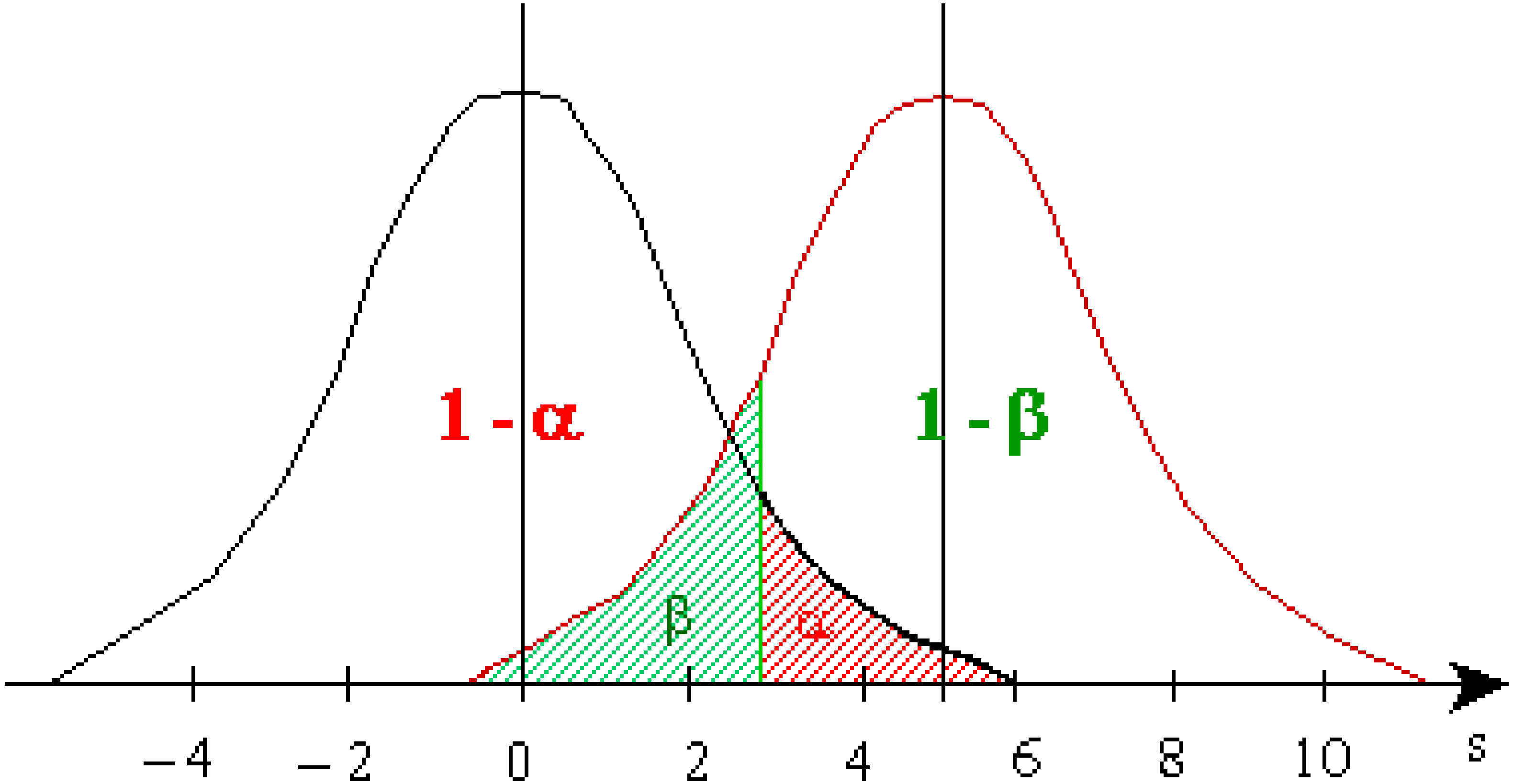
• Statistical hypothesis testing –Summary

➢ a statistical test procedure is similar to a criminal trial:
H_0: “the defendant is not guilty“
H_1: “the defendant is guilty”
→ the hypothesis of innocence is rejected only when an error is very unlikely, because one doesn’t want to convict an innocent defendant

	H_0 is true Truly not guilty	H_1 is true Truly guilty
Accept null hypothesis Acquittal	Right decision	Wrong decision Type II Error (β)
Reject null hypothesis Conviction	Wrong decision Type I Error (α)	Right decision power of the test (1 – β)

Statistics

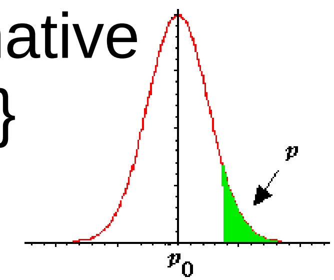
- Statistical hypothesis testing –Summary



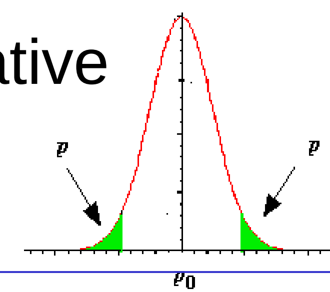
Statistics

- Statistical hypothesis testing –Test on one parameter

- ➢ the value of parameter θ (mean, variance, proportion) found on a sample can be related to an *a priori* value
- ➢ null hypothesis: $H_0 = \{\ \theta = \theta_0\ \}$
- ➢ alternative hypothesis: H_1 hypothesis different from H_0
- ➢ **one-tailed directional** (or *unilateral* test) → alternative hypothesis such as $H_1 = \{\ \theta < \theta_0\ \}$ or $H_1 = \{\ \theta > \theta_0\ \}$



- ➢ **two-tailed directional** (or *bilateral* test) → alternative hypothesis such as $H_1 = \{\ \theta \neq \theta_0\ \}$



Statistics

- Statistical hypothesis testing –Possible situations

➢ parameter to test: mean μ
distribution law of the population: normal
σ^2 known
test statistic: $\sqrt{n}\left(\frac{\bar{X}-\mu}{\sigma}\right)$
→ probability distribution: $\mathcal{N}(0\ ;\ 1)$ (normal distribution)

➢ parameter to test: mean μ
distribution law of the population: normal
σ^2 unknown
test statistic : $\sqrt{n}\left(\frac{\bar{X}-\mu}{s}\right)$
→ probability distribution: $\mathcal{S}(0\ ;\ 1)$ (Student's *t*-distribution)

Statistics

- Statistical hypothesis testing –Possible situations

➢ parameter to test: mean μ
distribution law of the population: ordinary with $n > 30$
σ^2 known
test statistic: $\sqrt{n}\left(\frac{\bar{X}-\mu}{\sigma}\right)$
→ probability distribution: $\approx\mathcal{N}(0\ ;\ 1)$ (normal approximation)

➢ parameter to test: mean μ
distribution law of the population: ordinary with $n > 30$
σ^2 unknown
test statistic: $\sqrt{n}\left(\frac{\bar{X}-\mu}{S}\right)$
→ probability distribution: $\approx\mathcal{N}(0\ ;\ 1)$ (normal approximation)

Statistics

- Statistical hypothesis testing –Possible situations

➢ parameter to test: variance σ^2
distribution law of the population: normal
μ known
test statistic : $\sum \frac{(X_i-\mu)^2}{\sigma^2}$
→ probability distribution: chi-square distribution χ^2 with n degrees of freedom

➢ parameter to test: variance σ^2
distribution law of the population: normal
μ unknown
test statistic: $\frac{(n-1)S^2}{\sigma^2}$
→ probability distribution : χ^2 with $(n – 1)$ degrees of freedom

Statistics

- Statistical hypothesis testing –Possible situations

➢ parameter to test: proportion p
distribution law of the population: $n > 50$
test statistic : $\sqrt{n}\dfrac{F-p}{\sqrt{p(1-p)}}$
→ probability distribution: $\approx \mathcal{N}(0\ ;\ 1)$ (normal approximation)

Statistics

- Statistical hypothesis testing –Possible situations

➢ it is a question of determining if two distinct populations have identical parameters

➢ parameter to test: means μ_1 and μ_2
$\sigma_1{}^2$ and $\sigma_2{}^2$ are known
distribution law of the population:
normal or ordinary with n_1 and n_2 > 5
test statistic: $\frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{\sigma_1{}^2}{n_1} + \frac{\sigma_2{}^2}{n_2}}}$
→ probability law under the assumption of equality of parameters: $\mathcal{N}(0\ ;\ 1)$

Statistics

- Statistical hypothesis testing –Possible situations

➢ parameter to test: means μ_1 and μ_2

variances $\sigma_1{}^2$ and $\sigma_2{}^2$ are unknown
distribution law of the population: normal with n_1 and n_2 > 20
or ordinary with n_1 and n_2 > 50

test statistic: $\dfrac{\bar{X}_1 - \bar{X}_2}{\sqrt{\dfrac{s_1{}^2}{n_1} + \dfrac{s_2{}^2}{n_2}}}$

→ probability law under the assumption of equality of parameters: $\mathcal{N}(0\,;\,1)$

Statistics

- Statistical hypothesis testing –Possible situations

➢ parameter to test: proportions p_1 and p_2
distribution law of the population: n_1 and n_2 > 50

test statistic: $$\frac{F_1 - F_2}{\sqrt{\frac{f_1(1-f_1)}{n_1} + \frac{f_2(1-f_2)}{n_2}}}$$

→ probability law under the assumption of equality of parameters: $\approx \mathcal{N}(0\,;\,1)$

Bibliography

- Peck R., C. Olsen, and J. L. Devore (2016). *Introduction to Statistics and Data Analysis*, 5th edition, Boston: Cengage Learning
- On YouTube: *StatQuest* with Josh Starmer:
- Maximum Likelihood, clearly explained!!! https://www.youtube.com/watch?v=XepXtl9YKwc
- Maximum Likelihood For the Normal Distribution, step-by-step! https://www.youtube.com/watch?v=Dn6b9fCIUpM
- P Values, clearly explained https://www.youtube.com/watch?v=5Z9OIYA8He8
- One or Two Tailed P-Values https://www.youtube.com/watch?v=bsZGt-caXO4