Statistical Learning. Statistical Inference

Jose Ameijeiras Alonso
Departamento de Estatística, An. Mat. e Otimización (USC)

Máster Interuniversitario en Tecnologías de Análisis de Datos Masivos: Big Data

Introduction

- Objective of statistical methods: to use empirical evidence to improve our knowledge about the target population from representative members (sample).
- We study the population of interest by measuring a set of characteristics (variables)
- With the methods of statistical inference, we can infer properties about the population from the information in the sample

Introduction

- Variables are classified as:
  - Qualitative: take on values that are names or labels
    - Nominal
    - Ordinal
  - Quantitative: take numeric values
    - Discrete
    - Continuous

Introduction

- Statistical analysis begins with a scientific problem:
    - identifying possible relationships among different variables
    - explaining or predicting how a variable changes with respect to some other variables
    - examining a scientific statement that explains a phenomenon (hypothesis testing)
    - ...

Exploratory Data Analysis (EDA)

- After collecting the data, and before performing any statistical inference or decision making we need to perform data exploration:
  - Frequency tables and data visualization
  - Summary Statistics

Exploratory Data Analysis (EDA)

Spam	Characters	Format	Attached	Number
0	11.37	HTML	0	big
0	8.596	HTML	1	small
1	0.11	text	0	none
0	10.504	HTML	0	small
0	7.773	HTML	0	small
0	13.256	HTML	0	small
0	1.231	text	0	none
1	0.171	text	0	none
1	0.341	text	2	none
...	...	...	...	...

EDA: frequency tables

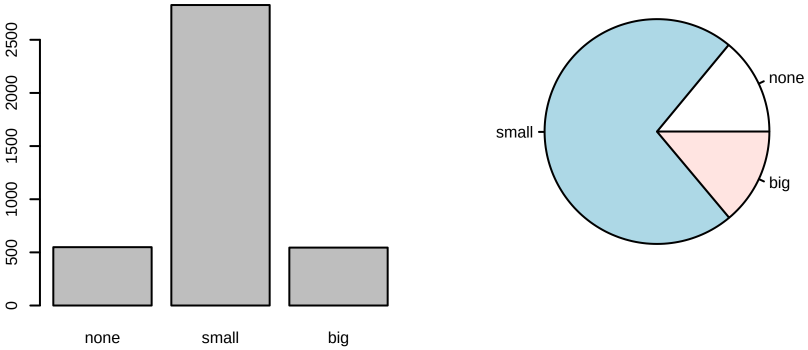
Format	n_i	f_i
HTML	2726	0.695
text	1195	0.305

Number	n_i	f_i	N_i	F_i
None	549	0.140	549	0.140
Small	2827	0.720	3376	0.861
Big	545	0.138	3921	1

c_i	n_i	f_i	N_i	F_i
c_1	n_1	f_1	N_1	F_1
c_2	n_2	f_2	N_2	F_2
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
c_m	n_m	f_m	N_m	F_m

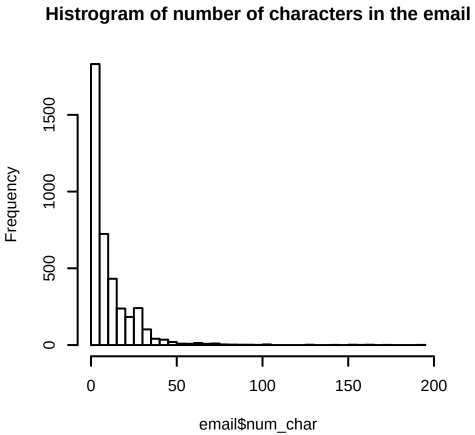
- $0 \leq n_i \leq n$, $\sum_{i=1}^m n_i = n$
- $0 \leq f_i \leq 1$, $\sum_{i=1}^m f_i = 1$
- $0 \leq N_i \leq n$, $N_m = n$
- $0 \leq F_i \leq 1$, $F_m = 1$

EDA: data visualization



Bar plots and pie charts are a common way to display a categorical or discrete variable.

EDA: data visualization



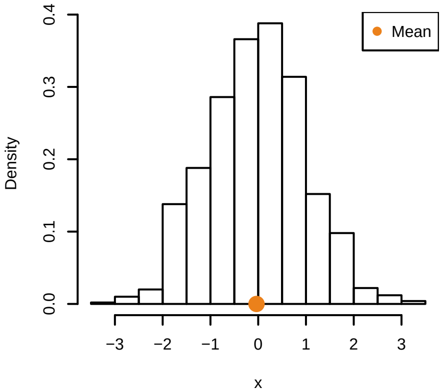
Histograms are used to display a continuous variable. They provide a view of the data density and are especially useful for describing the shape of the data distribution.

EDA: summary statistics

Summary statistics are numbers that summarize certain characteristics of the data.

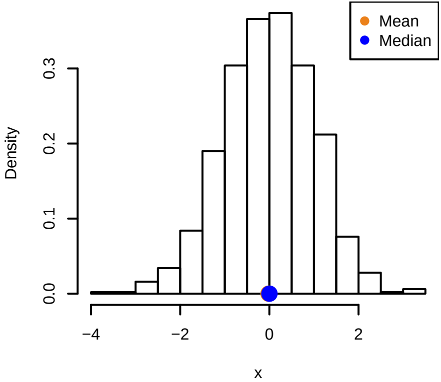
- Measures of location:
  - Mean
  - Median
  - Mode
  - Quantiles (quartiles, deciles, percentiles,. . .)
- Measures of variability
  - Variance
  - Standard deviation
  - Interquartile range (IR)
- Measures of skewness and kurtosis

EDA: summary statistics



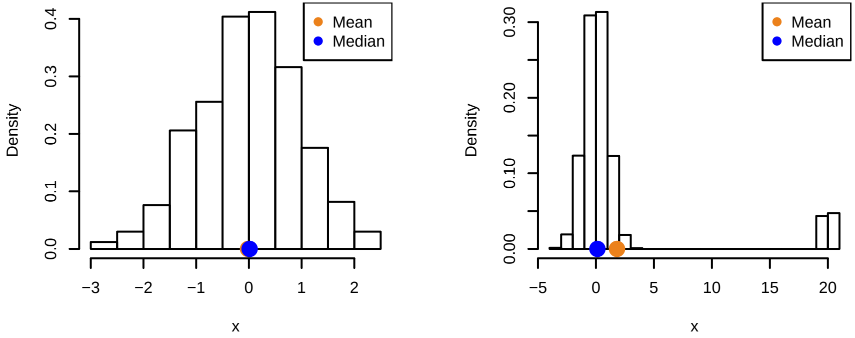
$$\boxed{\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}}$$

EDA: summary statistics



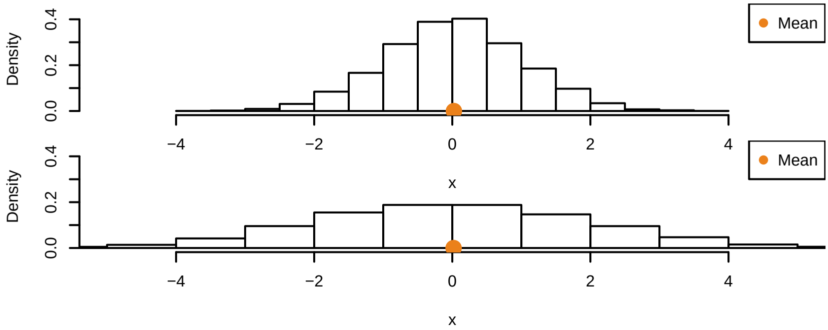
The median splits the data in half

EDA: summary statistics



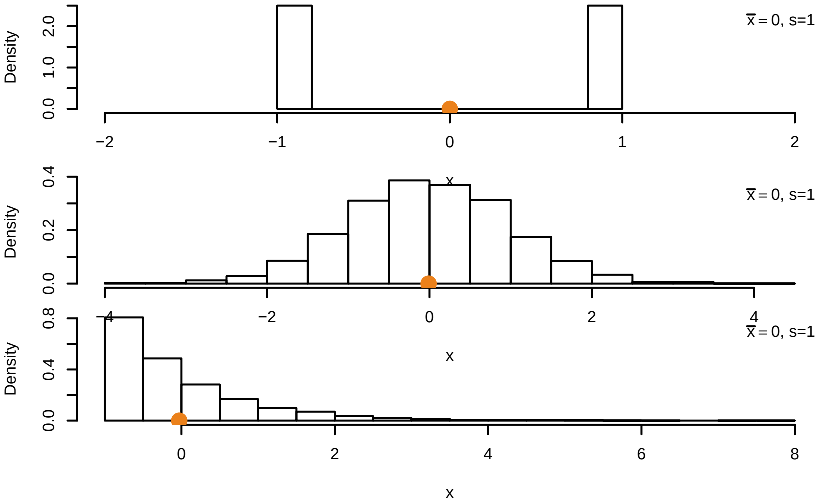
The median splits the data in half

EDA: summary statistics



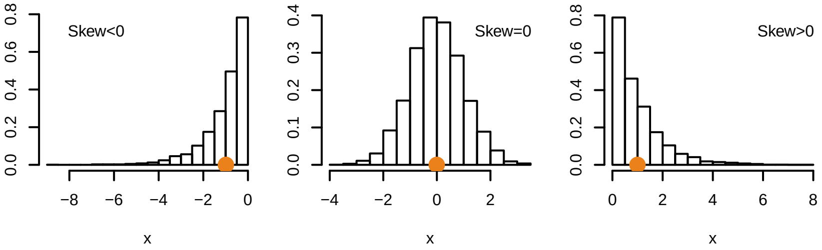
$$\boxed{s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

EDA: summary statistics



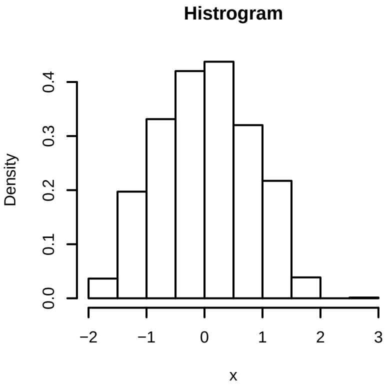
Very different population distributions can have the same mean and variance

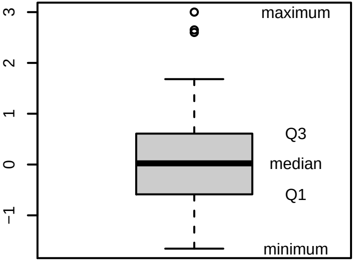
EDA: summary statistics



$$\boxed{Skew = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^3}{ns^3}}$$

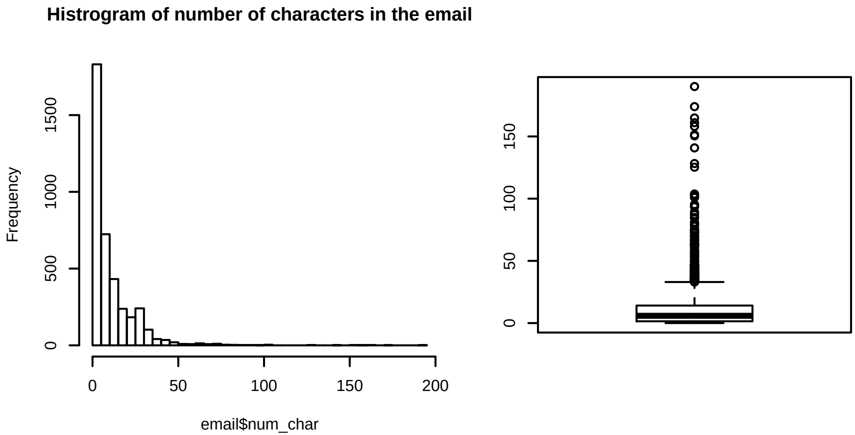
EDA: summary statistics





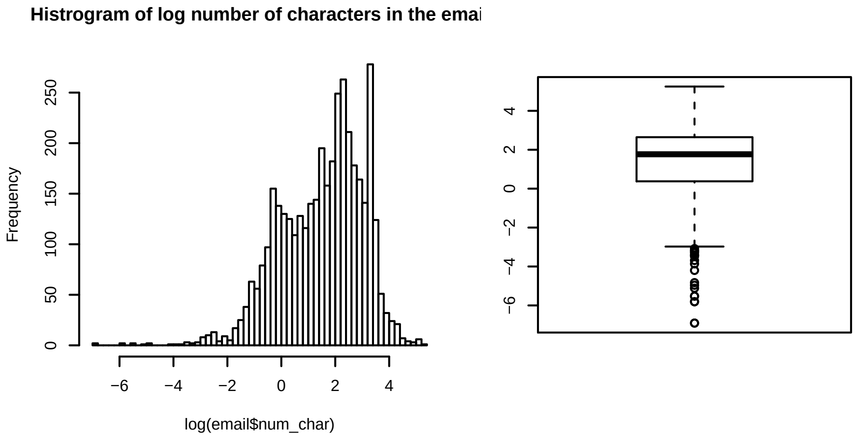
A box plot summarizes a data set using five statistics. It also represents unusual observations

EDA: summary statistics



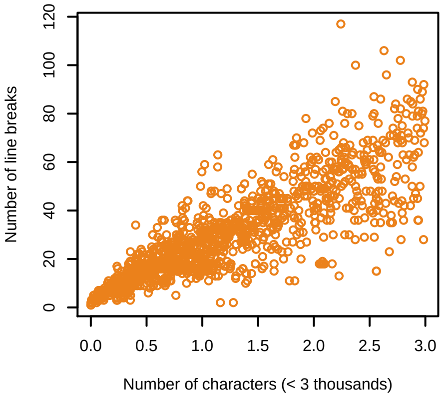
A box plot summarizes a data set using five statistics. It also represents unusual observations

EDA: summary statistics



A box plot summarizes a data set using five statistics. It also represents unusual observations

EDA: relationships between two variables



Scatterplots are one of the graphs used to analyze the relationship between two numerical variables.

Foundations for inference

- Statistical inference is concerned primarily with drawing conclusions on the population based on data (we use the information in the sample to infer facts about the population).
  - Parameters: population characteristics
  - Statistics: sample characteristics

Foundations for inference

- Point estimation: a single value that estimates the parameter
- Confidence Intervals: intervals within which the unknown parameter is expected to fall (with a given degree of confidence)
- Hypothesis testing: decision making in the presence of uncertainty

Foundations for inference

- Example: As part of a study on the behaviour of some of the functions in an R library, we generate 14 random matrices of size 500×450 and time the calculation of the pseudo-inverse with a given algorithm.

0.455	0.456	0.447	0.447	0.456	0.441	0.463
0.437	0.440	0.464	0.456	0.476	0.448	0.434

- For these data, $\bar{x} = 0.45$ and $s = 0.01$.
- Think of these observations as a random sample from a population
- The population could be described by its mean μ and its standard deviation σ
  - μ population mean time for the calculation of the pseudo-inverse
  - σ population standard deviation time for the calculation of the pseudo-inverse

Foundations for inference

- Example: As part of a study on the behaviour of some of the functions in an R library, we generate 14 random matrices of size 500×450 and time the calculation of the pseudo-inverse with a given algorithm.

0.455	0.456	0.447	0.447	0.456	0.441	0.463
0.437	0.440	0.464	0.456	0.476	0.448	0.434

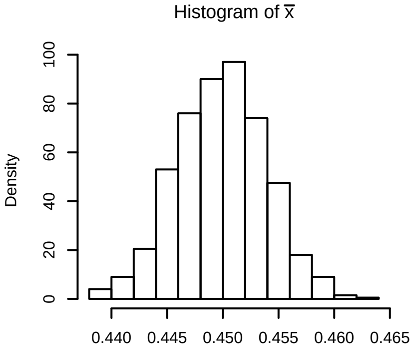
- For these data, $\bar{x} = 0.45$ and $s = 0.01$.
- It is natural to estimate μ by the $\bar{x}$ and σ by s
  - $\bar{x} = 0.45$ is an point estimation of μ
  - $s = 0.01$ is an point estimation of s
- These estimates depend on the sample and are subject to sampling error, no matter how accurately each time was computed.

Foundations for inference

- Suppose our goal is to estimate μ. How to determine the reliability of the estimate $\bar{x}$?

What if we repeat N times the experiment?
(N samples of size n)

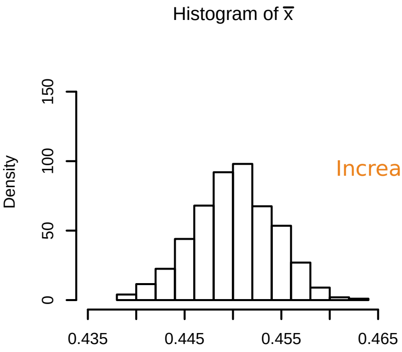
$$\begin{array}{cccccc}
X_1^1 & X_2^1 & \dots & X_n^1 & \rightarrow & \bar{x}^1 \\
X_1^2 & X_2^2 & \dots & X_n^2 & \rightarrow & \bar{x}^2 \\
\dots & \dots & \dots & \dots & \rightarrow & \dots \\
X_1^N & X_2^N & \dots & X_n^N & \rightarrow & \bar{x}^N
\end{array}$$



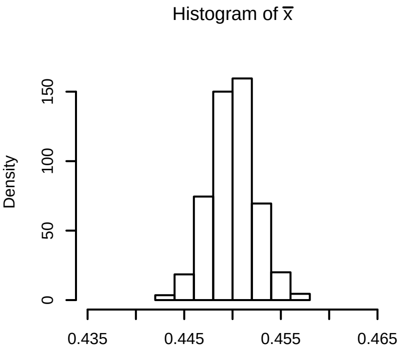
- If n is large, then the sampling distribution of $\bar{X}$ is approximately normal (Central Limit Theorem)

Foundations for inference

- The mean of the distribution of $\bar{X}$ is equal to μ [$\bar{X}$ is unbiased]
- The standard deviation of the distribution of $\bar{X}$ is equal to $\sigma/\sqrt{n}$ [standard error of the mean (SE)]
- The SE describes the variability (due to sampling error) in the mean of the sample as an estimate of the mean of the population.
- A natural estimate of $\sigma/\sqrt{n}$ would be $s/\sqrt{n}$



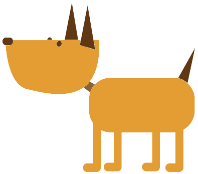
Increasing $n \rightarrow$



Foundations for inference

- In general, We would like to know the value of a population parameter θ but we cannot see it directly
- We choose an statistic, that is, a quantity $\hat{\theta}$ calculated from the sample to estimate the unknown parameter
- Two important characteristics of an statistic are the bias and the standard error
  - A statistic is biased if the expected value of the statistic is not the unkown parameter.
  - The standard error of an statistic is the standard deviation of the sampling distribution of the statistic.

Confidence interval: basic idea

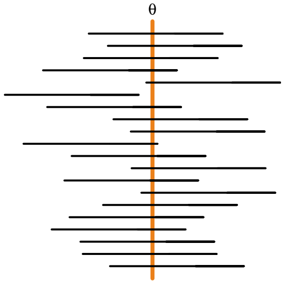


This figure is a drawing of an invisible man walking his dog. The dog, which is visible, is on an invisible spring-loaded leash. The tension on the spring is such that the dog is within 2 standard errors (SE) of the man 95% of the time. Only 5% of the time is the dog more than 2 SEs from the man-unless the leash breaks, in which case the dog could be anywhere.

We can see the dog, but we would like to know where the man is. Since the man and the dog are usually within 2 SEs of each other, we can take the interval "dog$\pm 2SE$" as an interval that typically would include the man. Indeed, we could say that we are 95% confident that the man is in this interval.

From Samuels et al. (2012) Statistics for the life sciences

Confidence interval



- In general, we have an unknow population parameter θ and $\alpha \in [0, 1]$.
- A confidence interval with confidence level $1-\alpha$ gives an estimated range of values $[L_1, L_2]$ such that

$$P(L_1 \leq \theta \leq L_2) \geq 1-\alpha$$

- Note tha L_1 y L_2 **depend on the sample!!!!**.

Confidence interval

- A confidence interval has this form:

$$\boxed{\text{IC} = \text{Point estimate} \pm \text{Margin of error}}$$

- The margin of error can also be subdivided into two parts and:

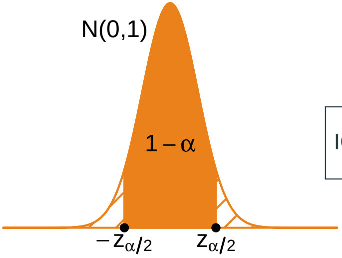
$$\boxed{\text{IC} = \text{Point estimate} \pm \text{Critical value} \times \text{Standard error}}$$

- The critical value depends on the the sampling distribution of the statistic
- The standard error of the point estimate

Confidence interval

$$\boxed{\text{IC} = \text{Point estimate} \pm \text{Critical value} \times \text{Standard error}}$$

- For example, the confidence interval for the mean μ of a normal pupulation with known σ^2 is:

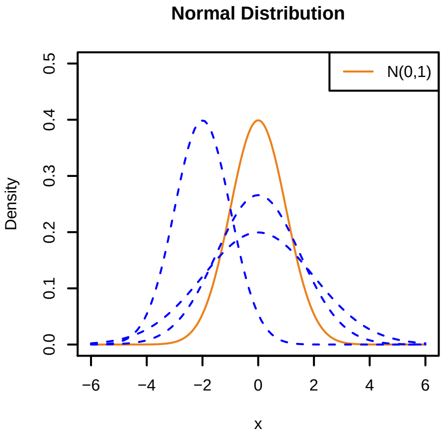


$$\boxed{\text{IC} = \left(\bar{X} - z_{\alpha/2}\,\frac{\sigma}{\sqrt{n}}\;,\;\bar{X} + z_{\alpha/2}\,\frac{\sigma}{\sqrt{n}}\right)}$$

The Normal distribution

- The probability density function of a normal distribution with mean μ and variance σ^2 is:

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}, \quad x \in \mathbb{R}$$



The Normal distribution

- The probability density function of a normal distribution with mean μ and variance σ^2 is:

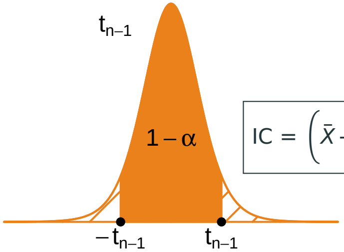
$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}, \quad x \in \mathbb{R}$$

- Its shape is symmetric
- The mean and the median are the same
- About 68% of its values lie within one standard deviation of the mean
- About 95% of its values lie within two standard deviations of the mean
- About 99.7% of its values lie within three standard deviations of the mean

Confidence interval

$$\boxed{\text{IC} = \text{Point estimate} \pm \text{Critical value} \times \text{Standard error}}$$

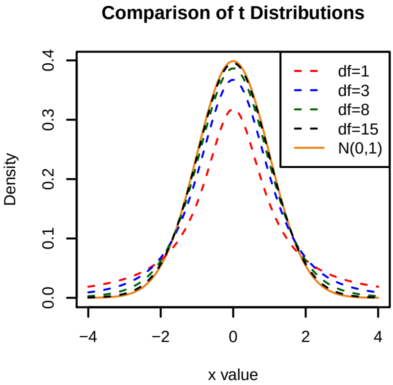
- For example, the confidence interval for the mean μ of a normal pupulation with unknown σ^2 is:



$$\boxed{\text{IC} = \left(\bar{X} - t_{n-1,\alpha/2}\,\frac{s}{\sqrt{n}}\ ,\ \bar{X} + t_{n-1,\alpha/2}\,\frac{s}{\sqrt{n}}\right)}$$

The Student's t distribution

- The Student's t distributions are theoretical continuous distributions that are used for different statistical analyses
- The shape of a Student's distribution depends on the parameter "degrees of freedom" (df)
- It is symmetric and bell-shaped, like the $N(0, 1)$, but with heavier tails. As the df increase, the curves approach the normal curve



Hypothesis testing

- When we seek to understand or explain something, we usually formulate our research question in the form of a hypothesis
- In statistics, a hypothesis is a statement about a distribution, an underlying parameter, a statement about the relationship between probability distributions, . . .

Hypothesis testing: the Lady tasting tea

A Lady declares that by tasting a cup of tea made with milk she can discriminate whether the milk or the tea infusion was first added to the cup.
How one should test the claim?

Ronald A. Fisher (1935) The Design of Experiments

Hypothesis testing

- Null hypothesis H_0: The hypothesis to be tested
- Alternative hypothesis H_1: The hypothesis that contradicts the null hypothesis

A hypothesis test is a decision-making process that examines the data, and on the basis of expectation under H_0, leads to a decision as to whether or not to reject H_0

"The null hypothesis... is never proved or established, but is possibly disproved, in the course of experimentation" Ronald A. Fisher (1935)

Hypothesis testing

- Example: The Lady testing tea

H_0 : The lady can not really tell the difference between teas, and she is just guessing

- Suppose we give her eight cups, four of each variety, in random order

- If she correctly identifies the mixing procedure, will we be convinced of her claim?
- Under the null hypothesis assumption (that she is guessing), what is the probability of this outcome?

Hypothesis testing

		Decision	
		Not to reject H_0	Reject H_0
The truth	H_0 is true	Correct decision	Type I error
	H_0 is not true	Type II error	Correct decision

- Type I error: incorrect rejection of a true null hypothesis
  - $\alpha = P(\text{Reject } H_0 / H_0 \text{ is true})$ → Significance level
- Type II error: failure to reject a false null hypothesis
  - $\beta = P(\text{Not to reject } H_0 / H_0 \text{ is false})$
  - $\text{Power} = P(\text{Reject } H_0 / H_0 \text{ is false}) = 1 - \beta$

Hypothesis testing

1. Specify the null hypothesis H_0 (and the alternative hypothesis H_A)
2. Specify the significance level α
3. Collect the sample
4. Calculate a test statistic
5. Determine the Acceptance/Rejection regions
6. Draw a conclusion about H_0

Hypothesis testing

1. Specify the null hypothesis H_0 (and the alternative hypothesis H_A)
2. Specify the significance level α
3. Collect the sample
4. Calculate a test statistic
5. Determine the Acceptance/Rejection regions
6. Draw a conclusion about H_0

1. Specify the null hypothesis H_0 (and the alternative hypothesis H_A)
2. Specify the significance level α
3. Collect the sample
4. Calculate a test statistic
5. Compute the p-value

Hypothesis testing

- The p-value is the probability of obtaining a more extreme statistic than we did if the null hypothesis were true
- It is the smallest level of significance at which the null hypothesis H_0 can be rejected
- A small p-value indicates that it is unlikely to observe such value of the statistic if the null hypothesis is true (the observed data are inconsistent with the assumption that the null hypothesis is true)