Imperial College
London

16th October 2018

Inference

`https://bitbucket.org/mfumagal/statistical_inference`

Matteo Fumagalli

Imperial College London

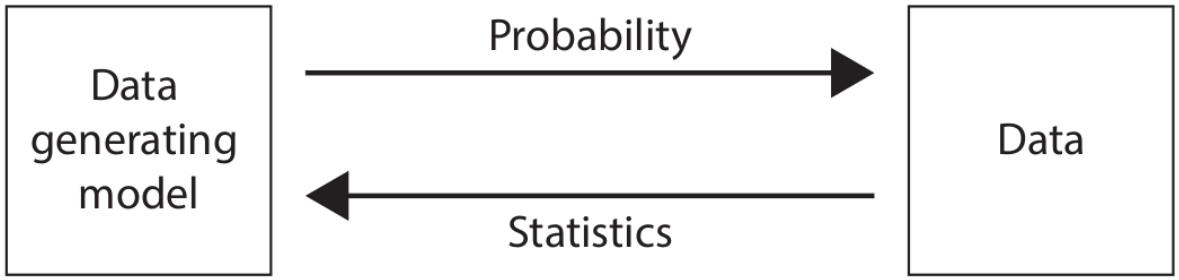
Intended Learning Outcomes

By the end of this session, you will be able to:

- Explain the difference between population and sample statistics
- Describe data using a range of descriptive and graphical summaries
- Illustrate the properties of estimators and principles of hypothesis testing

Imperial College London

From probability theory to statistics



Populations and random samples

A *population* is the set of all units or objects one intends to study:

- UK human population (e.g. height)
- Human T-lymphotropic Virus-1 (HTLV-1) infected T-cells
- Gut microbiota
- ...

Imperial College London

Populations and random samples

Although we would like to study the whole population, observing all the units in the population is often not possible:

- Measurement costs are too high (e.g. census)
- No access to all units (e.g. online survey respondents, T-cells in a blood sample, bacteria in faecal sample)

A *random sample* is a subset of the population that is representative of the population.

Imperial College London

Stages of a statistical analysis

There are four broad stages:

1. Select measurement variables
2. Perform random sampling
3. Construct one or more statistical models
4. Perform data analysis:
    - Descriptive statistics
    - Inferential statistics

1. Measurement variables

The initial stage of any experimental analysis involves the selection of variables to observe and their measurement scale.

There are two types of data:

- Quantitative

1. Measurement variables

The initial stage of any experimental analysis involves the selection of variables to observe and their measurement scale.

There are two types of data:

- Quantitative
  - Continuous (e.g. concentration) vs. Discrete (e.g. cell counts)
  - Univariate (e.g. light intensity) vs. Multivariate (e.g. chemotaxis: 2D movement vector, microarray data)
- Qualitative

1. Measurement variables

The initial stage of any experimental analysis involves the selection of variables to observe and their measurement scale.

There are two types of data:

- Quantitative
  - Continuous (e.g. concentration) vs. Discrete (e.g. cell counts)
  - Univariate (e.g. light intensity) vs. Multivariate (e.g. chemotaxis: 2D movement vector, microarray data)
- Qualitative
  - Nominal: No logical ordering (categories, classes, binary data. e.g. healthy/diseased, male/female, smoker/non-smoker, etc)
  - Ordinal: Codes with logical ordering (e.g. exam grades)

2. Random sampling

Most statistical analyses assume the following:

- Extract a random sample of n units
- All the measurements are collected in sample data $D = \{x_1, ..., x_n\}$
- The elements of D are random realisations of n random variables $\{X_1, ..., X_n\}$
- Assume variables are independent and identically distributed (i.i.d)
- The underlying probability function or density function is $f_X(x) \equiv f_X(x; \theta)$ where θ is a parameter or parameter vector.

3. Statistical models

A *statistical model* is a set $\{f_X(x;\theta)|\theta \in \Theta\}$ of probability measures, one of which corresponds to the true, unknown, probability measure $p(x;\theta^*)$ that produced the data.

θ is the parameter of the model, and Θ is the parameter space.

- θ^* is the true or population parameter such that $p(x;\theta^*)$ is the true probability measure for the data.
- We decide on the appropriate model either by using prior knowledge of the data generating process or by using exploratory statistical tools.
- The data space $\mathcal{X}$ and parameter space Θ are related by different spaces!

4. Data analysis

- *Descriptive statistics* is the discipline of summarising and describing data (e.g. quantitative summaries and visual representations of the data)
- *Inferential statistics* is the branch of statistics which attempts to draw conclusions about the the population from random samples

Descriptive statistics: summaries of the data

Given a random sample, one produces numerical and visual summaries of the data in order to:

- Detect trends or features in the observed data
- Detect outliers
- Suggest appropriate statistical models, often in the absence of prior knowledge (e.g. bi-modality in histograms suggests mixture models)

Typical quantitative summaries of data fall into several classes: location or central tendency measures, mode, scale (or spread of the data), skewness (or asymmetry).

Central tendencies

- arithmetic mean (or sample mean)
- weighted arithmetic mean
- geometric mean
- harmonic mean

Imperial College London

Central tendencies

There are several measures of central tendencies. For a dataset of n objects $D = \{x_1, \ldots, x_n\}$,

1. *Arithmetic mean*

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

   - Most common measure
   - Also called *sample mean*

2. *Weighted arithmetic mean*

$$\bar{x} = \frac{\sum_{i=1}^{n} w_i x_i}{\sum_{i=1}^{n} w_i}, \quad \text{, with} \quad \sum_{i=1}^{n} w_i = 1$$

   - Arises in *importance sampling* where one samples a probability distribution via samples from another distribution

Imperial College London

Central tendencies

3. *Geometric mean*

$$\bar{g} = \left(\prod_{i=1}^{n} x_i \right)^{\frac{1}{n}}$$

- Often used for positive-valued data with values varying significantly
- *Compound average growth rates*: the geometric mean of exponentials is the exponential of the arithmetic mean

$$\left(e^{x_1} e^{x_2} \cdots e^{x_n}\right)^{\frac{1}{n}} = e^{\frac{x_1 + \cdots + x_n}{n}} \equiv e^{\bar{x}}$$

4. *Harmonic mean*

$$\bar{h} = \left(\frac{1}{n} \sum_{i=1}^{n} x_i^{-1} \right)^{-1}$$

- Averages of ratios

NB: $\bar{h} \leq \bar{g} \leq \bar{x}$

Scale

For univariate sample data, the common summary measures for scale are

- sample minimum
- sample maximum
- sample range
- p^{th}-quantile of the empirical distribution function: median, lower and upper quartiles, inter-quartile range.
- (unbiased) sample variance

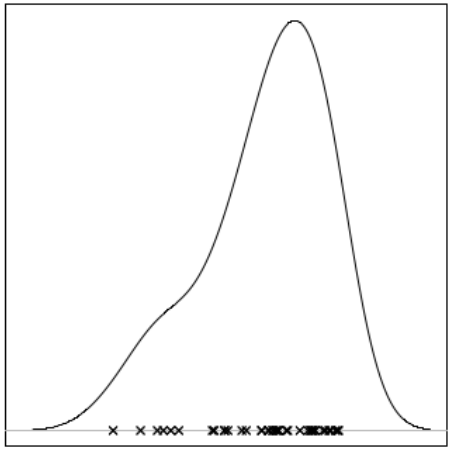
$$s^2_{n-1} = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \hat{x})^2$$

Imperial College London

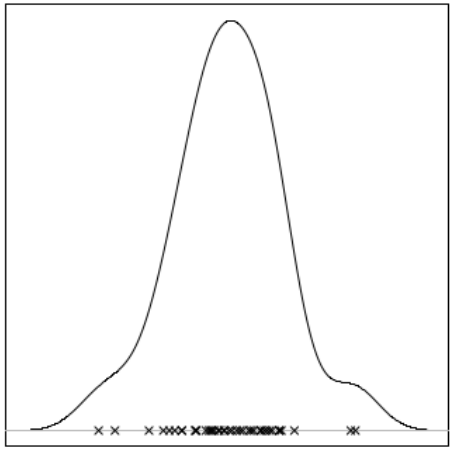
Skewness

Third and higher moments of a distribution can provide useful descriptions.

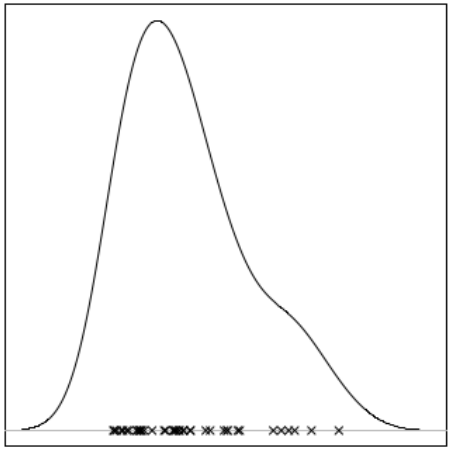
Skewness measures the departure from symmetry of the probability distribution of a real-valued random variable.



Negative skewness



Zero skewness



Positive skewness

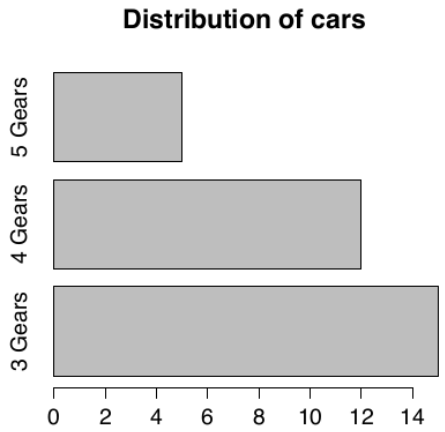
Graphical summaries

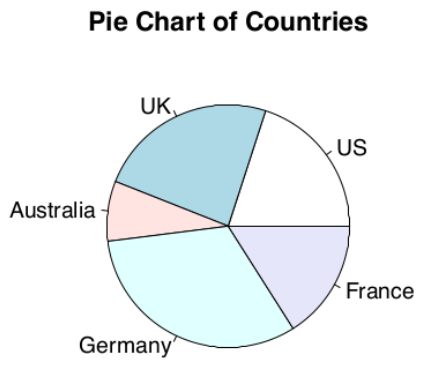
In addition to the preceding quantitative summaries, it is often useful to provide visual or graphical summaries.

- The type of graphs/plots depend on the data obtained, e.g. Discrete or continuous, Uni- or multivariate.
- Besides providing a summary of the data, graphical summaries can also illustrate any testing we may wish to perform.

Imperial College London

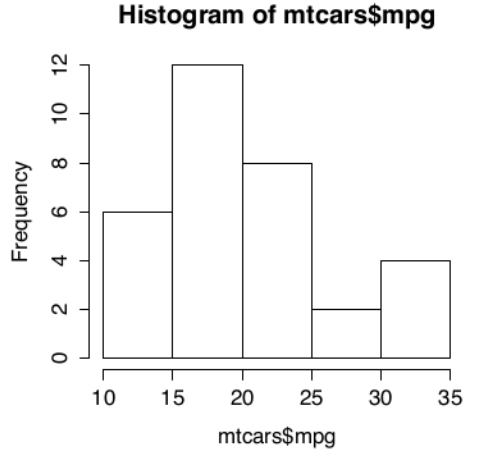
Discrete data visualisation - Barplot and piecharts

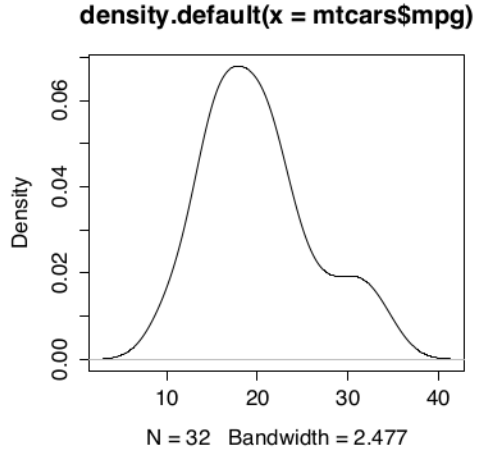




Imperial College London

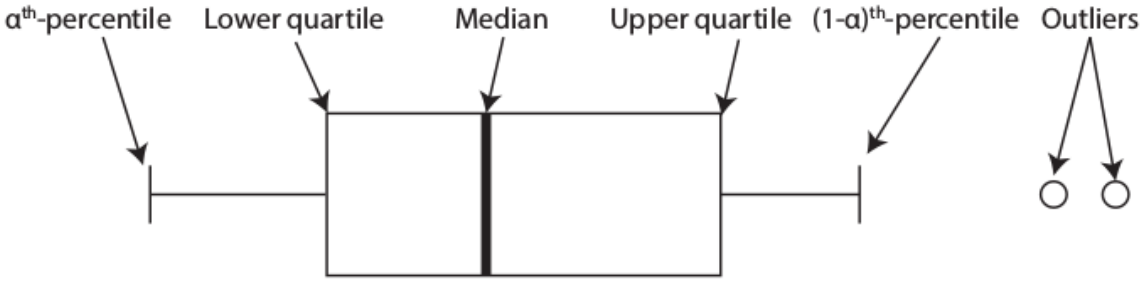
Continuous data visualisation – Histograms and density estimates

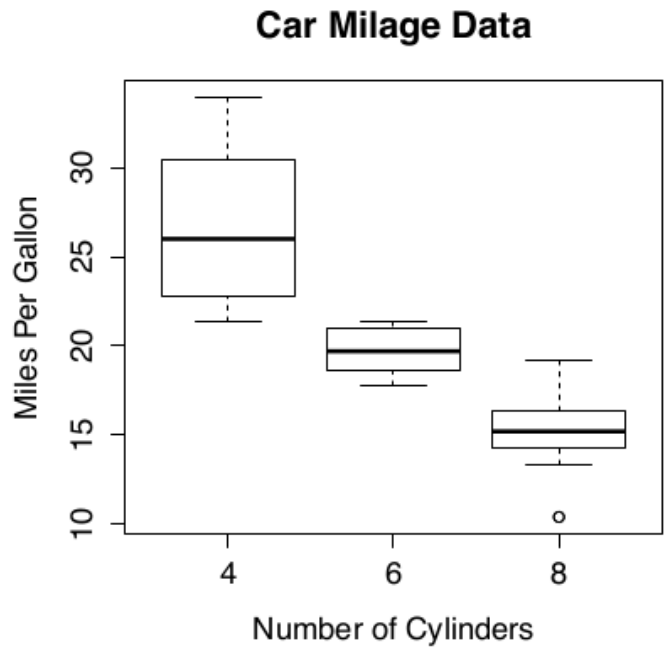




Imperial College London

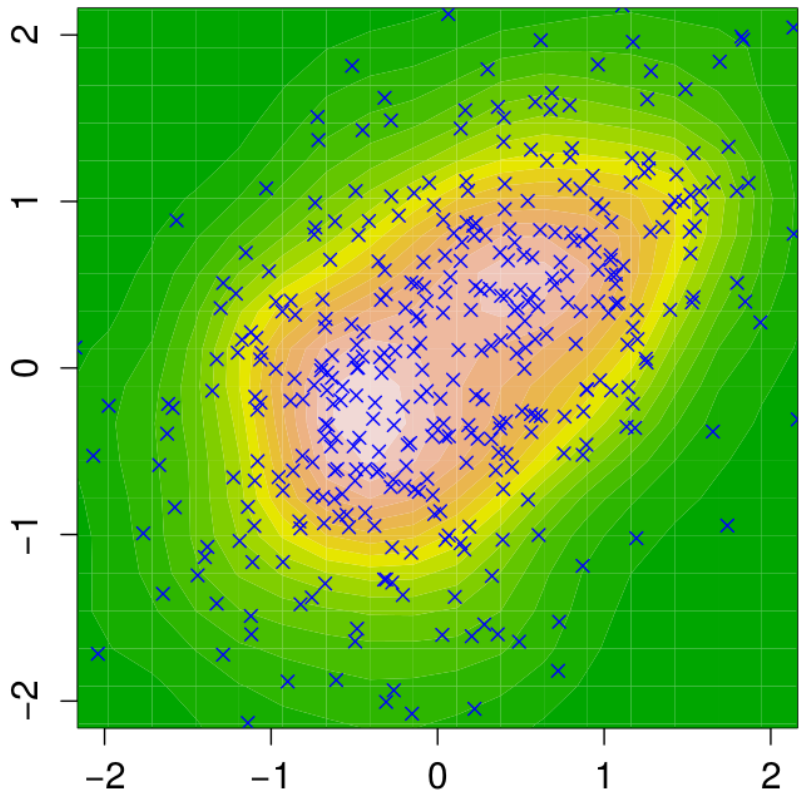
Continuous data visualisation – Boxplots





Imperial College
London

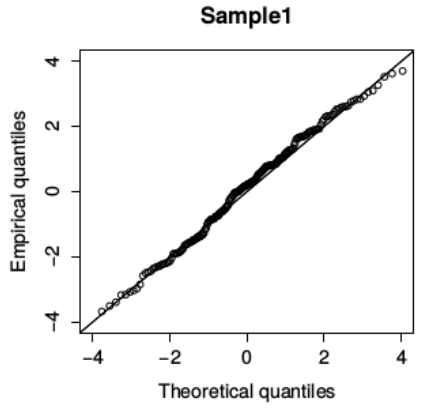
Bivariate data visualisation – Scatterplots



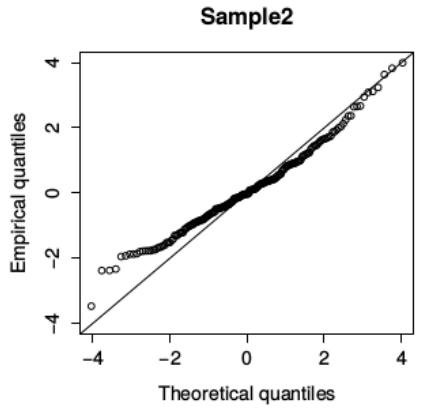
Imperial College London

Probability plotting techniques - QQ plot

Normal Data



Non-normal data



Inferential statistics

It provides the mathematical theory for inferring properties of an unknown distribution from data generated from that distribution.

- In the *parametric* approach, one selects a suitable distribution and attempt to infer the parameters from the data.
- In a *non-parametric* approach, one does not make any assumption about the underlying distribution of the data.

There are several approaches to statistical inference: frequentist, likelihoodist, Bayesian.

Imperial College London

Statistical inference

Assume that we know and accept a statistical model $f_X(x;\theta)$.

The main objective of parametric statistical inference is to learn something about the unknown population parameter θ from information contained in sample data.

We will use an estimator or *statistic*, a function of the sample data, to learn about θ.

Statistical inference

Common tasks are:

1. Point estimation: obtain a single estimate $\hat{\theta}$ of the population parameter θ
2. Interval estimation: obtain an interval having a certain probability to contain the unknown population parameter θ
3. Hypothesis testing: test a specific hypothesis about θ, i.e. do the observed data support the hypothesis?

Estimators

An estimator or statistic is a function of the random sample $D = \{x_1, ..., x_n\}$, say $t(D)$.

- $t(D)$ depends on the data sample alone
- We have already encountered several statistics, e.g. the sample mean and the sample variance
- Because an estimator is a function of random variables, it is itself a random variable with its own distribution. We usually refer to the latter as the *sampling distribution of the sample statistic*.

Imperial College London

Estimators

Estimators are used to estimate the unknown population parameters.

For instance, if we assume some parametric model with population mean parameter μ, we may want to use $\hat{x}$ as an estimate for μ.

Often, it is not trivial to construct estimators. One approach is the maximum likelihood estimator (more on this later).

Imperial College London

Constructing and characterising estimators

We encountered several measures of central tendency – the (arithmetic) sample mean, median, geometric mean, etc.

Which estimator is the "best" one for estimating the population mean parameter θ? How is one estimator "better" than another?

Imperial College London

Constructing and characterising estimators

We encountered several measures of central tendency – the (arithmetic) sample mean, median, geometric mean, etc.

Which estimator is the "best" one for estimating the population mean parameter θ? How is one estimator "better" than another?

- Since estimators are random variables, we compare them by assessing their respective sampling distribution (when known)
- We will look at the following properties: 1. Bias, 2. Mean Squared Error (MSE), 3. Efficiency, 4. Consistency

Bias

- Suppose we could repeat the same experiment a number of times, say B, and collect new data D each time.
- Each time we use our estimator $t(D)$ of choice to obtain an estimate $\hat{\theta}$ of an unknown population parameter θ.
- We have a set of different estimates which are of samples from the sampling distribution of the estimator/sample-statistic $t(D)$.

Bias

- Suppose we could repeat the same experiment a number of times, say B, and collect new data D each time.
- Each time we use our estimator $t(D)$ of choice to obtain an estimate $\hat{\theta}$ of an unknown population parameter θ.
- We have a set of different estimates which are of samples from the sampling distribution of the estimator/sample-statistic $t(D)$.

The bias of an estimator $t(D)$ is defined as

$$bias(t(D)) = E_\theta(t(D)) - \theta$$

An *unbiased* estimator is one with zero bias, i.e. $E_\theta(t(D)) = \theta$.

Example on height in R.

Imperial College London

Mean squared error (MSE) and standard error of an estimate

How much variability do we expect to see in $\hat{\theta}$ under repeated sampling from the assumed distribution?

Imperial College London

Mean squared error (MSE) and standard error of an estimate

How much variability do we expect to see in $\hat{\theta}$ under repeated sampling from the assumed distribution?

A common measure of the spread of the sampling distribution $t(D)$ around the true parameter θ is given by the mean squared error (MSE)

$$MSE(t(D)) = E(t(D) - \theta)^2$$

It measures the average *spread* difference between the estimator and θ.

Imperial College London

Mean squared error (MSE) and standard error of an estimate

The MSE has two components:

- the variability of the estimator (precision)
- its bias (accuracy)

For an unbiased estimator, the MSE equal its variance and the standard error is the square root of the MSE.

What's the interplay between bias and MSE?

Efficiency

All things equal, we choose the unbiased estimator with the smallest variance, i.e. with higher precision.

The efficiency of $t_1(D)$ relative to $t_2(D)$ is

$$efficiency = \frac{Var(t_1(D))}{Var(t_2(D))}$$

Consistency

Consistency is an asymptotic property of an estimator.

It describes the behaviour of the estimator $t(D)$ as the sample size n gets larger and larger. Hence it involves a sequence of estimators.

Two sufficient conditions for consistency are:

- $\lim_{n\to\infty} E(t_n) = \theta$
- $\lim_{n\to\infty} Var(t_n) = 0$

Imperial College London

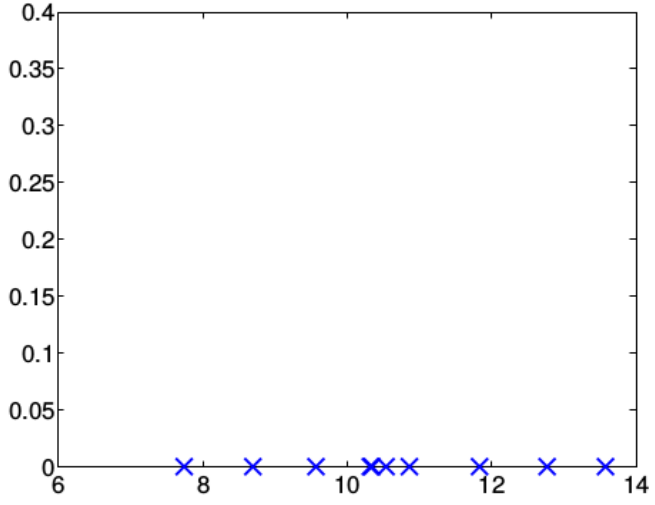
Wrap up

1. Stages of a statistical analysis:
    1. Define measurement variables
    2. Perform random sampling
    3. Choose a statistical model
    4. Perform data analysis
2. Descriptive statistics
    - data summaries
    - visualisation
3. Statistical inference
    - estimators and their properties
    - ...

Imperial College London

Model fitting

How do we fit models to data? A model will have one or more parameters, θ, that we need to estimate to get a good fit.



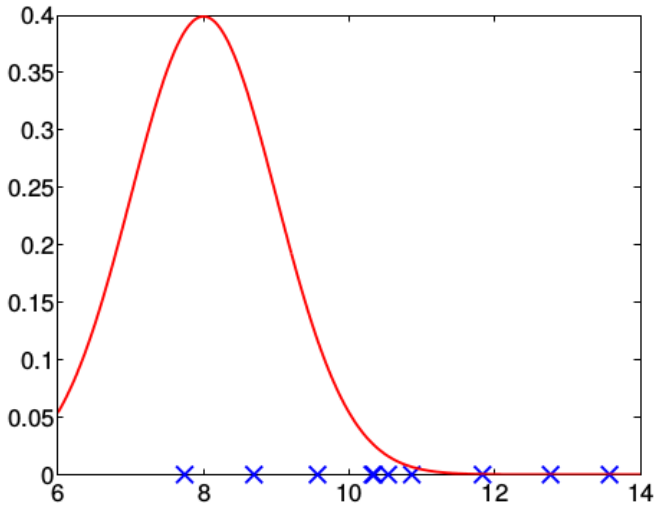
e.g. we may want to model some observations as being normally distributed with mean μ and standard deviation 1.

What is μ?

Imperial College London

Model fitting

What is μ?

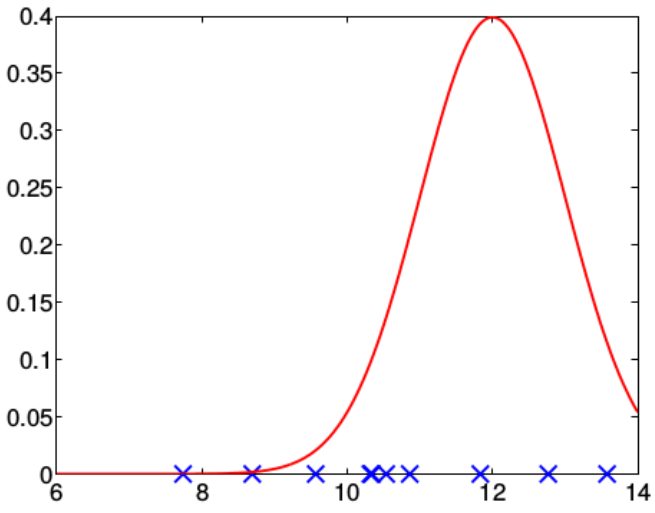


$\mu = 8$?

Imperial College London

Model fitting

What is μ?

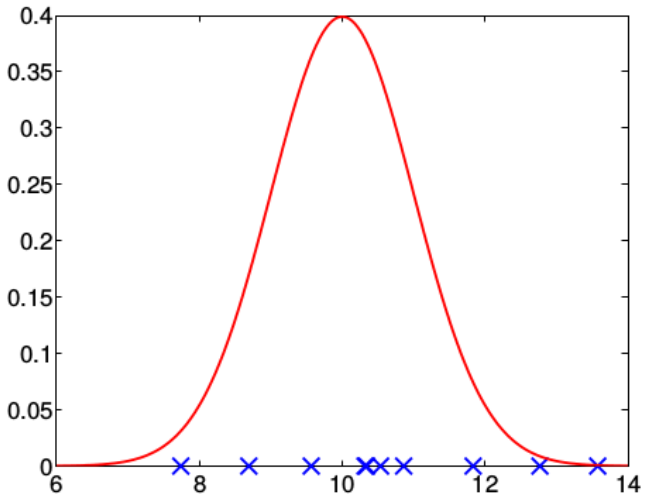


$\mu = 12$?

Imperial College London

Model fitting

What is μ?



$\mu = 10$?

Likelihood

- The concept of likelihood provides us with a formal framework for estimating parameters.
- In particular, maximum likelihood estimation is a general method for estimating the parameters of a (probability) model.
- The likelihood function is also one of the key ingredients for Bayesian inference.

Imperial College London

Likelihood

Suppose we have a *biased coin* such that:

$$p(H) = 0.7 \text{ and } p(T) = 0.3.$$

We toss the coin 3 times.
What is the probability of *HHH*?
(3 independent Bernoulli trials, so...) the probability is

$$p(H) \times p(H) \times p(H) = 0.7^3 = 0.343.$$

What is the probability of *HTH*?
What is the probability of *TTH*?

Imperial College London

Likelihood

Given our probability model, $p(H) = 0.7$, we can write down the probability of any outcome:

$$
\begin{aligned}
p(HHH) &= 0.343 & \qquad p(HTT) &= 0.063 \\
p(HHT) &= 0.147 & \qquad p(TTH) &= 0.063 \\
p(HTH) &= 0.147 & \qquad p(THT) &= 0.063 \\
p(THH) &= 0.147 & \qquad p(TTT) &= 0.027
\end{aligned}
$$

(Note that these probabilities sum to 1!)

Imperial College London

Likelihood

Suppose we consider another biased coin. This time, $p(H) = 0.6$.

Then we can again write down the probability of any outcome:

$$
\begin{aligned}
p(HHH) &= 0.216 & \qquad p(HTT) &= 0.096 \\
p(HHT) &= 0.144 & \qquad p(TTH) &= 0.096 \\
p(HTH) &= 0.144 & \qquad p(THT) &= 0.096 \\
p(THH) &= 0.144 & \qquad p(TTT) &= 0.064
\end{aligned}
$$

(Note that these probabilities sum to 1!)

Imperial College London

Likelihood

More generally, suppose $p(H) = \rho$.

Given ρ, we can write down the probability associated with any outcome, D (e.g. $D = HHH$).

We have:

$$p(D|\rho) = \rho^{n_H}(1-\rho)^{n_T},$$

where n_H is the number of heads and n_T is the number of tails that appear in D.

ρ is a *parameter* of our probability model.
If we know ρ, we can write down the probability of any D.

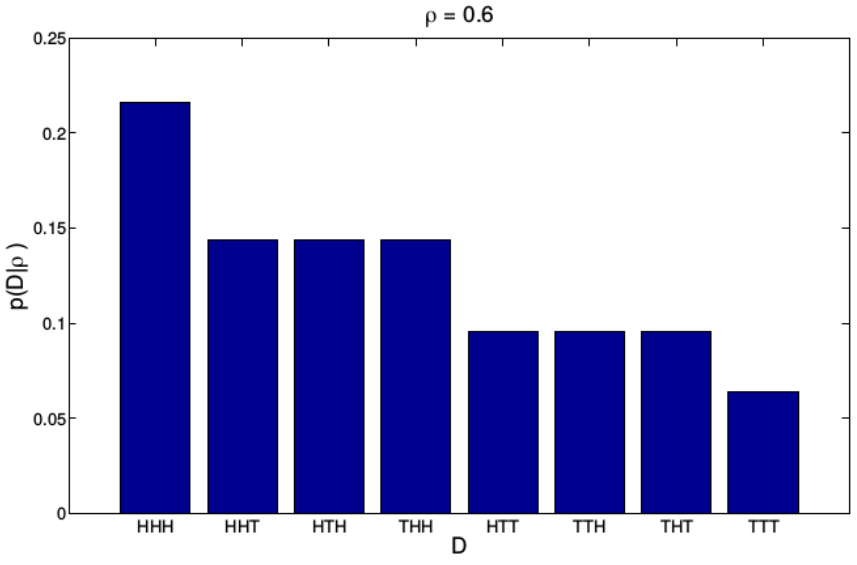
In practice, we often have the *inverse problem*. We observe the outcome D and want to estimate ρ.

Imperial College London

Likelihood

For a given ρ, we can consider $p(D|\rho)$ as a function of D (i.e. we fix ρ and can think about how $p(D|\rho)$ varies as we change D).

e.g. $\rho = 0.6, p(D|\rho = 0.6) = 0.6^{n_H}(1-0.6)^{n_T}$



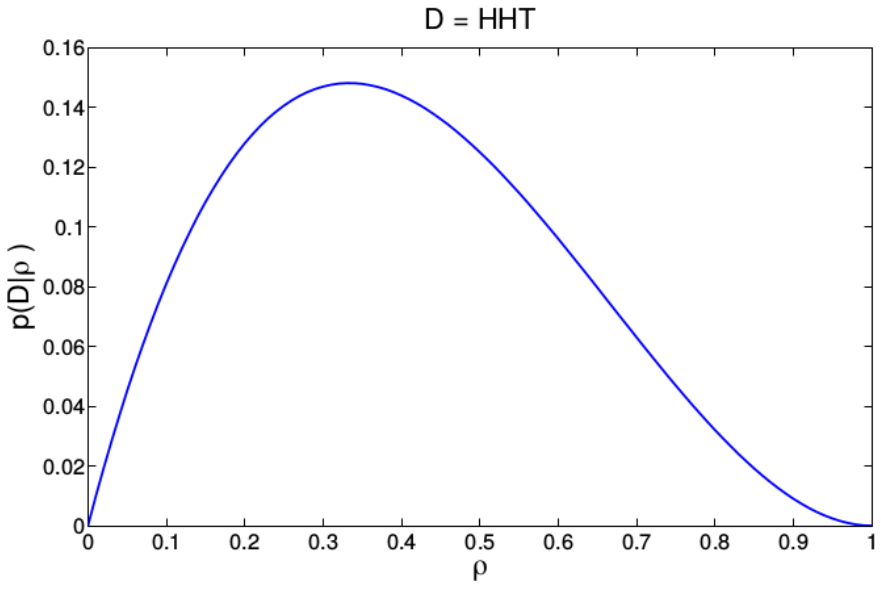
If we fix ρ and consider $p(D|\rho)$ as a function of D, $p(D|\rho)$ is a *probability mass function* (or *probability density function* in the continuous case – from now on, I will say "*probability function*").

Imperial College London

Likelihood

For an observed D, we can consider $p(D|\rho)$ as a function of ρ.

e.g. $D = HTT$, $p(D = HTT|\rho) = \rho^1(1-\rho)^2$



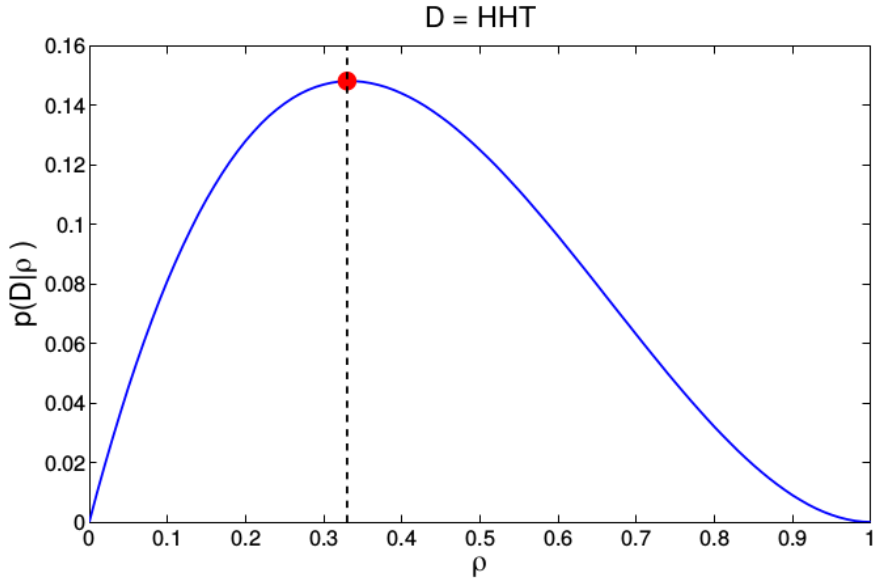
If we observe D and consider $p(D|\rho)$ as a function of ρ, $p(D|\rho)$ is a *likelihood function*.

Imperial College London

Likelihood

For an observed D, we can consider $p(D|\rho)$ as a function of ρ.

e.g. $D = HTT$, $p(D = HTT|\rho) = \rho^1(1-\rho)^2$



If we observe D and consider $p(D|\rho)$ as a function of ρ, $p(D|\rho)$ is a *likelihood function*. One way of estimating ρ is to find the value that maximises the likelihood function.

Imperial College London

Likelihood

Definition. Let $(X_1, \ldots, X_n)$ have joint probability function $f(x_1, \ldots, x_n; \theta)$, where $\theta \in \Theta$ is a parameter. The *likelihood function* L is a function of θ defined by

$$L(\theta; x_1, \ldots, x_n) = f(x_1, \ldots, x_n; \theta),$$

where $\{x_1, \ldots, x_n\} = D$ is fixed and θ varies in Θ.

If the data are i.i.d., we can factorize the likelihood as

$$L(\theta; x_1, \ldots, x_n) = \prod_{i=1}^{n} f(x_i; \theta) \qquad [\textit{i.i.d. case only.}]$$

Imperial College London

Maximum likelihood estimate

Definition. The value/vector θ_{ML} is called the *maximum likelihood estimate* of the parameter θ if, given the observed data D,

$$L(\theta_{ML}|D) = \max_{\theta} L(\theta|D),$$

i.e. θ_{ML} is the value/vector that maximises the value of the likelihood function.

To find the MLE, we need to solve an optimisation problem: we maximise the likelihood function $L(\theta|D)$ over θ in the parameter space Θ.

- In a few cases this optimization problem can be solved in closed-form
- In ~~other~~ *most* cases we need to resort to numerical methods

Imperial College London

Wrap up

1. The likelihood function *is* a function of θ
2. The likelihood function is *not* a probability density function for θ.
3. If the data are i.i.d. then the likelihood is

$$L(\theta; x_1, \ldots, x_n) = \prod_{i=1}^{n} f(x_i; \theta) \qquad [\textit{i.i.d. case only.}]$$

4. The likelihood is only defined up to a constant of proportionality

Statistical tests

- In our study of statistical inference, we have focused on point estimation of unknown parameters of a given statistical model.
- The residual uncertainty is expressed as uncertainty in these point estimates (e.g. sampling distribution of estimator).
- We will now look at *hypothesis testing*
  - Make a definitive hypothesis about θ
  - Uncertainty is reflected in the expected probability of being wrong (p-value)

Statistical test

We will encounter a wide range of statistical tests:

- Parametric vs. Non-parametric
- One-sided vs. Two-sided tests
- Z-test, t-test, goodness-of-fit tests, likelihood ratio test, etc...

Statistical test

We will encounter a wide range of statistical tests:

- Parametric vs. Non-parametric
- One-sided vs. Two-sided tests
- Z-test, t-test, goodness-of-fit tests, likelihood ratio test, etc...

However, there is a common approach to all statistical tests:

1. Generate a Null Hypothesis and an Alternative Hypothesis
2. Obtain the sampling distribution of the estimator under the null hypothesis: the Null distribution
3. Decide whether to reject the null hypothesis

Null and Alternative Hypotheses

- Suppose we want to know if the use of a drug is associated to a symptom.
- We take some mice and randomly divide them into two groups.
- We expose one group to the drug and leave the second group unexposed.
- We then compare the symptom rate in the two groups, θ_1, θ_2 respectively.

Null and Alternative Hypotheses

Consider the following two hypotheses:

- The Null Hypothesis: The symptom rate is the same in the two groups, $H_0 : \theta_1 = \theta_2$
- The Alternative Hypothesis: The symptom rate is not the same in the two groups, $H_1 : \theta_1 \neq \theta_2$

If the exposed group has a much higher rate of symptom than the unexposed group, we will **reject** the null hypothesis and conclude that the data favours the alternative hypothesis.

Null and Alternative Hypotheses

Let Θ be the parameter space of a statistical model.
We partition Θ into two disjoint sets Θ_0 and Θ_1 and test

$$H_0 = \theta \in \Theta_0 \text{ vs. } H_1 = \theta \in \Theta_1$$

H_0 and H_1 are the Null and Alternative hypotheses, respectively.

Null and Alternative Hypotheses

- A hypothesis of the form $\theta = \theta_0$ is called a *simple* hypothesis.
- A hypothesis of the form $\theta > \theta_0$ (or $\theta < \theta_0$) is called a *composite* hypothesis.

- A test of the form $H_0 : \theta = \theta_0$ vs. $H_1 : \theta \neq \theta_0$ is called a *two-sided test*.
- A test of the form $H_0 : \theta = \theta_0$ vs. $H_1 : \theta < \theta_0$ (or $>$) is called a *one-sided test*.

Imperial College London

The Null distribution

Let D be a random sample of data.
We test a hypothesis by finding a set of outcomes R called the *rejection* region.

- If $D \in R$ we reject the null hypothesis
- if $D \notin R$, we do not reject H_0

Can we ever accept H_0?

The Null distribution

R can be in the form $R = \{D : t(D) > c\}$ where D is the data, $t(D)$ is the test statistic, and c is the *critical value(s)*.

The critical value is determined from two things:

- The null distribution, i.e. the sampling distribution of $t(D)$, assuming the null hypothesis is true.
- The significance level α (typically $\alpha << 1$).

Significance and confidence intervals

- When there is enough evidence to reject H_0 , we say that the result from a statistical test is "statistically significant at the given significance level α".
- The significance level α controls how stringent the test is.
- When α is very small, the acceptance region is larger and the test is more stringent because the null hypothesis will be rejected less frequently.
- The probability of making an error and wrongly rejecting the null hypothesis when it is in fact true is exactly α.

Significance and confidence intervals

Statistical significance does not imply *scientific* significance!

It is often more informative to give *confidence intervals*.
A $(1-\alpha)$-confidence interval (C.I.) is an interval in which the true parameter lies with probability $1-\alpha$.

p-values

We have established a decision rule: we reject H_0 when the value of the test statistic falls outside the acceptance region.
Other than a reject/accept decision, the result of a statistical test is often reported in terms of a p-value.

A p-value is

- the probability, under the null hypothesis, of a result as or more extreme than that actually observed.
- the smallest significance level at which the null hypothesis would be rejected.

Imperial College London

p-values

A p-value is NOT the probability that the null hypothesis is true.
i.e. do not confuse the p-value with $P(H_0|D)$.

A large p-value can occur for two reasons:
(i) H_0 is true or
(ii) H_0 is false but the test has low power (e.g. too few samples).

Example on jupyter notebook

Imperial College London

Type I and II errors

When we test the null hypothesis versus the alternative hypothesis of a population parameter, there are four possible outcomes, two of which are erroneous:

	H_0 True	H_1 True
Do not reject H_0	–	Type II Error
Reject H_0	Type I Error	–

- A Type I error is made when the null hypothesis is wrongly rejected.
- A Type II Error is made when we conclude that we do not have enough evidence to reject the H_0 , but in fact the alternative hypothesis H_1 is true.

The power of a test

The power of a test represents its ability to correctly reject the null hypothesis. It is expressed as the probability

$$\begin{aligned} P(\text{Reject } H_0 | H_1 \, True) &= 1 - P(\text{Do not reject } H_0 | H_1 \, True) \\ &= 1 - P(\text{Type II error}) \end{aligned}$$

Statistical tests

There are several widely used tests:

- The Wald test: test the true value of the parameter based on the sample estimate.
- t-test: to determine if two sets of data are significantly different from each other, assuming normality.
- Wilcoxon signed-rank test: non-parametric, to compare two related samples to assess whether their population means are different.
- Chi-squared goodness-of-fit test: to test whether observed sample frequencies differ from expected frequencies.
- The likelihood ratio test.

Imperial College London

Chi-squared goodness-of-fit test

The problem of testing whether a data sample $x_1, \ldots, x_n$ is well modelled by a specified probability distribution can be approached from a *goodness-of-fit* perspective

Given a sample size n, suppose the data are categorised into K *bins*, and we count the number of observations that fall into each one of these K bins,

$$O_i, \qquad \text{for } i = 1, \ldots, K.$$

Suppose that under an *hypothesised probability model* with mass/density function $p_X(x)$, let the probability that a data point falls into bin i be given by

$$p_i$$

with $i = 1, \ldots, K$, which are calculated from p_X *after parameter estimation*.

Imperial College London

Chi-squared goodness-of-fit test

If the hypothesised probability model were true, and given n data points, the expected number E_i of observations in bin i is

$$E_i = np_i$$

We define the *Chi-squared statistic*

$$\chi^2 = \sum_{i=1}^{K} \frac{(O_i - E_i)^2}{E_i}$$

This is a test statistic measuring goodness-of-fit

The null distribution of this statistic is approximately a *chi-squared distribution* with

$$K - d - 1$$

degrees of freedom.

d is the number of parameters in p_X that were estimated in order to calculate the probabilities p_i.

Chi-squared goodness-of-fit test

Example: we observe two alleles, A and G, for a specific genomic locus in a population. Specifically, we observe 14 genotypes AA, 4 genotypes AG and 2 genotypes GG.
Can we reject the hypothesis of Hardy Weinberg Equilibrium (HWE, i.e. random mating and no natural selection) for this locus?
Note that, under HWE, we expect the following frequencies for the associated genotypes:

- homozygous f^2
- heterozygous $2f(1-f)$
- homozygous $(1-f)^2$

Imperial College London

Likelihood Ratio Test

Consider the test: $H_0 : \theta \in \Theta_0 \quad vs. \quad H_1 : \theta \notin \Theta_0$
Let $\theta = (\theta_1, \ldots, \theta_q, \theta_{q+1}, \ldots, \theta_r)$ and

$$\Theta_0 = \{\theta : (\theta_{q+1}, \ldots, \theta_r) = (\theta_{0,q+1}, \ldots, \theta_{0,r})\},$$

where $\theta_{0,*}$ are constants. Define the *Likelihood test statistic*

$$\lambda = 2 \log\left(\frac{\mathcal{L}(\theta_{MLE})}{\mathcal{L}(\theta_{MLE,0})} \right)$$

where θ_{MLE} is the MLE and $\theta_{MLE,0}$ is the MLE when θ is restricted to Θ_0.

Then as the sample size $n \to \infty$ the *null distribution* is

$$\lambda \sim \chi^2_{r-q}$$

where $r - q \equiv \dim(\Theta) - \dim(\Theta_0)$

Bootstrapping

It relies on random *sampling with replacement* to assign measure of accuracy to sample estimates.

The idea is to infer population parameter by resampling the data and performing inferences on the sample from the *resampled data*.

It assumes that samples are independent (otherwise use block bootstrap for correlated samples).

Wrap up

General procedure to test statistical hypotheses about a population parameter θ:

1. Set up the null and alternative hypotheses for θ, H_0 and H_1 (one-sided or two-sided).

Wrap up

General procedure to test statistical hypotheses about a population parameter θ:

1. Set up the null and alternative hypotheses for θ, H_0 and H_1 (one-sided or two-sided).
2. Compute the test statistic and obtain the null distribution, i.e. the sampling distribution under the null hypothesis.

Wrap up

General procedure to test statistical hypotheses about a population parameter θ:

1. Set up the null and alternative hypotheses for θ, H_0 and H_1 (one-sided or two-sided).
2. Compute the test statistic and obtain the null distribution, i.e. the sampling distribution under the null hypothesis.
3. Choose a significance level α or confidence level $1 - \alpha$.

Wrap up

General procedure to test statistical hypotheses about a population parameter θ:

1. Set up the null and alternative hypotheses for θ, H_0 and H_1 (one-sided or two-sided).
2. Compute the test statistic and obtain the null distribution, i.e. the sampling distribution under the null hypothesis.
3. Choose a significance level α or confidence level $1 - \alpha$.
4. Determine the rejection region for the test statistic, which depends on α.

Wrap up

General procedure to test statistical hypotheses about a population parameter θ:

1. Set up the null and alternative hypotheses for θ, H_0 and H_1 (one-sided or two-sided).
2. Compute the test statistic and obtain the null distribution, i.e. the sampling distribution under the null hypothesis.
3. Choose a significance level α or confidence level $1 - \alpha$.
4. Determine the rejection region for the test statistic, which depends on α.
5. Apply the decision rule: reject H_0 in favour of H_1 if the test statistic falls in the rejection region. Otherwise conclude that there is insufficient evidence to reject H_0.

Wrap up

General procedure to test statistical hypotheses about a population parameter θ:

1. Set up the null and alternative hypotheses for θ, H_0 and H_1 (one-sided or two-sided).
2. Compute the test statistic and obtain the null distribution, i.e. the sampling distribution under the null hypothesis.
3. Choose a significance level α or confidence level $1 - \alpha$.
4. Determine the rejection region for the test statistic, which depends on α.
5. Apply the decision rule: reject H_0 in favour of H_1 if the test statistic falls in the rejection region. Otherwise conclude that there is insufficient evidence to reject H_0.
6. Report the p-value and construct the confidence interval.