Looking back

Introduction to Statistical Learning
Bachelor in *Global Governance*
University of Rome - Tor Vergata

Marco Stefanucci
Department of Economics and Finance
University of Rome - Tor Vergata
marco.stefanucci@uniroma2.it

Prerequisites

- Descriptive Statistics
- Probability theory
- Statistical Inference

PROBABILITY

- Let Ω be a set $\mathcal{F}$ a collection of subsets of ω. A *probability measure*, or simply a *probability* on $(\Omega, \mathcal{F})$ is a function

$$P \, : \mathcal{F} \to [0, 1]$$

- To be a probability P must satisfy
  1. $\forall A \in \mathcal{F}, \quad 0 \leq P(A) \leq 1$
  2. $P(\Omega) = 1$
  3. if A_1 e A_2 are disjoint then $P(A_1 \cup A_2) = P(A_1) + P(A_2)$

- We can show that property 3 holds for any finite collection of disjoint sets

$$P(\bigcup_{i=1}^{n} A_i) = \sum_{i=1}^{n} P(A_i)$$

- It is common practice to assume that 3 hold for countable collections of disjoint sets

$$P(\bigcup_{i=1}^{\infty} A_i) = \sum_{i=1}^{\infty} P(A_i)$$

Properties of the probability

- $P(A) = 1 - P(\bar{A})$
- If $A \subseteq B$ $P(A) \leq P(B)$
- For general set A and B

$$P(A \cup B) = P(A) + P(B) - P(A \cap B)$$

Random variables

- A random variable (r.v.) is a numerical expression of the outcome of a statistical experiment or more generally the numerical manifestation of a phenomenon that can have different outcomes

- Formally, given Ω and a probability P on Ω,
    - a r.v. X is a function $X(\omega)$ defined on Ω and taking values in $\mathbb{R}$

$$X : \Omega \to \mathbb{R}$$

    - and $\forall B \subseteq \mathbb{R}$

$$P(X \in B) = P(\omega \in \Omega : X(\omega) \in B)$$

- Random variables will be generally indicated with the letters X, Y, Z...

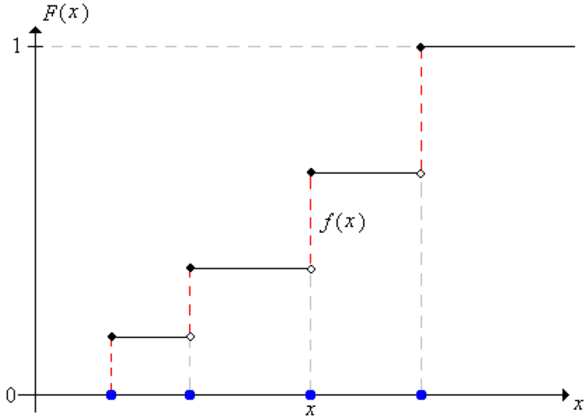
- A r.v. that may assume only a finite number or an infinite sequence of values is said to be discrete;
- A r.v. that may assume any value in some interval on the real number line is said to be continuous.
- For instance, a random variable representing the number of new cases of COVID-19 on one day would be discrete, while a random variable representing the weight of a person in kilograms would be continuous.

- Very often we work directly with random variables without knowing (or caring toknow) the underlying probability P on the space Ω
- In fact we will specify (model) directly the probabilities of the outcomes of the r.v.

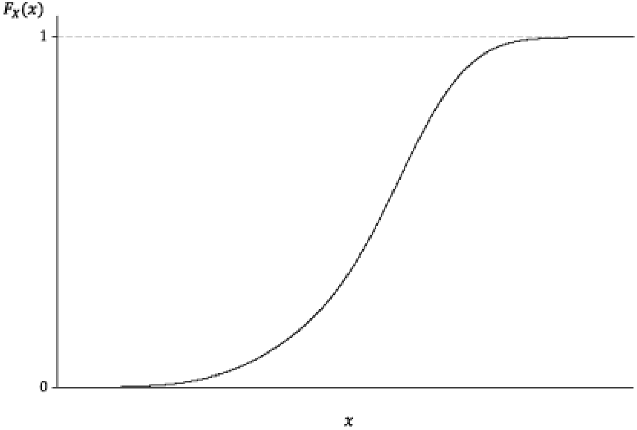
- The (cumulative) distribution function $F_X(x)$ of X is

$$F_X(x) = P(X \leq x) \quad x \in \mathbb{R}$$

- Note that
  1. $F_X(-\infty) = 0$
  2. $F_X(\infty) = 1$
  3. $x < x' \Rightarrow F(x) \leq F(x')$



Cumulative distribution function of a discrete random variable



Cumulative distribution function of a continuous random variable

- By the distribution function and the rules of probability we can obtain other probabilities on the r.v. X, e.g.

  $$P(X > a) = 1 - P(X \leq a) = 1 - F_X(a)$$

  and since for $a < b$, $(-\infty, b] = (-\infty, a] \cup (a, b]$

  $$P(X \leq b) = P(X \leq a) + P(a < X \leq b)$$

  and

  $$P(a < X \leq b) = F_X(b) - F_X(a)$$

Discrete random variables

- A discrete random variable takes value only in some countable subset D of $\mathbb{R}$
- Commonly this subset D is a subset of the integers
- The probability that X takes some given value x in D is

$$p(x) = P(X = x) = P(\omega \in \Omega : X(\omega) = x)$$

- The function $p(x)$ is be called *probability distribution of X* or probability function
- Example: Suppose we toss an unbiased coin 2 times in succession. What is the probability of obtaining x heads $(x = 0, 1, 2)$? Let X be the r.v. describing the result of such experiments. The probability function is

	x	$Pr(X = x)$
(T,T)	0	$\frac{1}{4}$
(T,H), (H,T)	1	$\frac{1}{2}$
(H,H)	2	$\frac{1}{4}$

- Note that

  $$p(x) \in [0,1]$$

  and

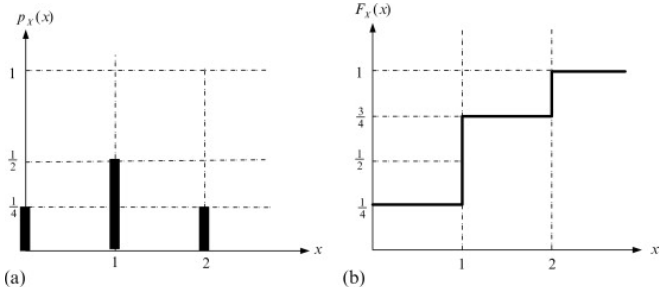
  $$\sum_{x \in D} p(x) = 1$$

  In fact,
  $\Omega = \bigcup_x \{\omega : X(\omega) = x\}$ and $\{\omega : X(\omega) = x\} \cap \{\omega : X(\omega) = x'\} = \emptyset$ otherwise we would have that $\exists \omega : X(\omega) = x$ and $X(\omega) = x'$ which is impossible, then

  $$\begin{aligned} 1 &= P(\Omega) = P\left(\bigcup_x \{\omega : X(\omega) = x\}\right) \\ &= \sum_{x \in D} P(\{\omega : X(\omega) = x\}) = \sum_{x \in D} p(x) \end{aligned}$$

- Moreover

  $$F(x) = Pr(X \leq u) = \sum_{u \leq x} p(u)$$



Density function and cumulative distribution function for the previous example

Famous probability distributions

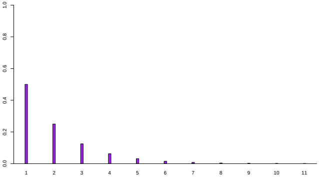
- $X \sim$ Bernoulli(p) for $0 \leq p \leq 1$

$$p(1) = p \text{ and } p(0) = 1 - p$$

- $X \sim$ Geometric(p) for $0 \leq p \leq 1$

$$p(x) = p(1-p)^{x-1} \text{ for } x = 1, 2, ..$$

This r.v. represents, for example, the number of coin flips until the first head sshows up (assuming independent coin flips)

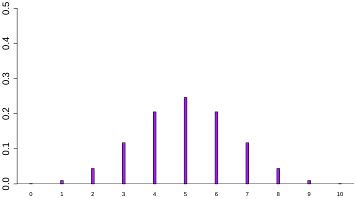


Probability distribution function for a Geometric r.v. with $p = 0.5$

- $X \sim$ Binomial(n, p) for $n > 0$ and $0 \leq p \leq 1$

$$p(x) = \binom{n}{x} p^x (1-p)^{n-x} \quad \text{for } x = 0, 1, 2, \ldots n$$

The binomial r.v.
represents, for example, the number of heads in n independent coin flips

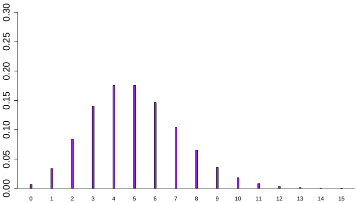


Probability distribution function for a Binomial r.v. with $n = 10$ and $p = 0.5$

- $X \sim$ Poisson(λ) for $\lambda > 0$

$$p(x) = \frac{\lambda^x}{x!} e^{-\lambda} \quad \text{for } x = 0, 1, 2, \dots$$

The Poisson r.v. often represents the number of random events, e.g. number of customers, email, COVID-19 cases etc., in some time interval



Probability distribution function for a Binomial r.v. with $n = 10$ and $p = 0.5$

Continuous random variables

- A continuous r.v. can take values on a interval, either of finite or infinite length
  1. Medical trials: the time until a patient experience a relapse or the time until healing
  2. Economics: the income of a family
  3. Health economics: the cost of a treatment

- Since the elements x of a real interval are uncountable, for a continuous r.v. we must have $P(X = x) = 0$

- Formally, a r.v. X is continuous if $\forall B \subset \mathbb{R}$

$$P(X \in B) = \int_B f_X(x)dx$$

for some function $f_X(x)$ that will be called probability **density function** or simply **density**

- Every density function satisfy the following two properties
    - $f_X(x) \geq 0$
    - $\int_{-\infty}^{\infty} f_X(x) = 1$

- In fact if $f_X(x) < 0$ on the interval (a, b) then $P(X \in (a, b)) = \int_a^b f_X(x)dx < 0$ and we can't have probabilities less than 0

- Moreover $1 = P(X \in (-\infty, \infty)) = \int_{-\infty}^{\infty} f_X(x)dx$

- Note that we effectively have $P(X = a) = 0 \; \forall a \; \in \; \mathbb{R}$
    - In fact $P(X = a) = \lim_{\epsilon \to 0} P(X \in [a, a + \epsilon])$
      $= \lim_{\epsilon \to 0} \int_a^{a+\epsilon} f_X(x)dx = 0$

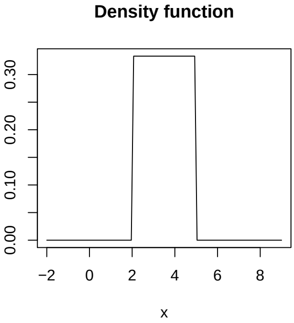
Popular continuous random variables

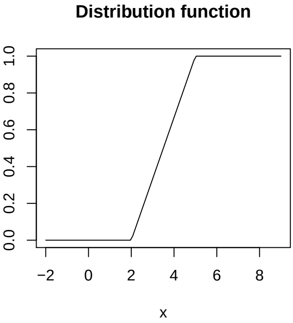
- Uniform: $X \sim \text{Unif}(a, b)$ for $b > a$ has density function

$$f_X(x) = \begin{cases} \frac{1}{b-a} & \text{for } a \leq x \leq b \\ 0 & \text{otherwise} \end{cases}$$

The cumulative distribution function is

$$F_X(x) = \begin{cases} 0 & \text{for } x < a \\ \frac{x-a}{b-a} & \text{for } a \leq x \leq b \\ 1 & \text{for } x > b \end{cases}$$


Density function
0.00
0.10
0.20
0.30
−2
0
2
4
6
8
x

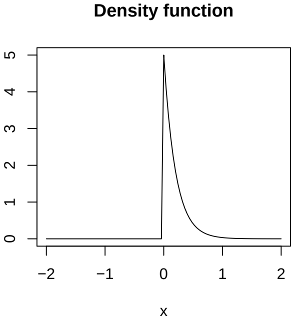

Distribution function
0.0
0.2
0.4
0.6
0.8
1.0
−2
0
2
4
6
8
x

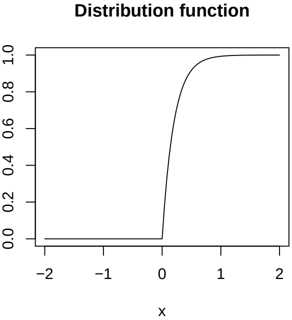
- Exponential: $X \sim \text{Exp}(\lambda)$ for $\lambda > 0$ has density function

$$f_X(x) = \begin{cases} \lambda e^{-\lambda x} & \text{for } x > 0 \\ 0 & \text{otherwise} \end{cases}$$

The cumulative distribution function is

$$F_X(x) = \begin{cases} 0 & \text{for } x < 0 \\ 1 - e^{-\lambda x} & \text{for } x \geq 0 \end{cases}$$

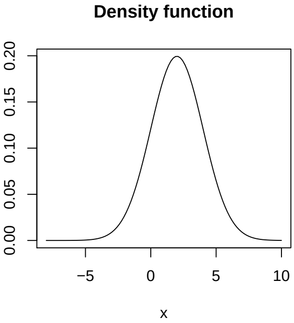

Density function
5
4
3
2
1
0
−2
−1
0
1
2
x

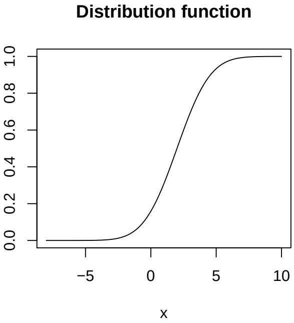

Distribution function
1.0
0.8
0.6
0.4
0.2
0.0
−2
−1
0
1
2
x

- Normal: $X \sim \mathsf{N}(\mu, \sigma^2)$ for $-\infty < \mu < \infty$ and $\sigma > 0$ has density function

$$f_X(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma^2}\right)}$$

  The cumulative distribution function cannot be obtained analytically





Mean of a random variable

- Let X be a discrete r.v. with probability distribution function $p_X(x)$ for $x \in S$. The expected value of X (or mean of X) is

$$E(X) = \sum_{x \in S} x \, p_X(x)$$

- Let X be a Bernoulli r.v. . Then $p(1) = p$, $p(0) = 1 - p$ and

$$E(X) = 1 \times p + 0 \times (1 - p) = p$$

- Let X be a discrete r.v. with probability $p(x) = 1/3$ for $x = -1, 0, 1$

$$E(X) = \sum_{i=1}^{3} x_i p_i = -1\frac{1}{3} + 0\frac{1}{3} + 1\frac{1}{3} = 0$$

- Let X be a continuous r.v. with density $f_X(x)$. The expected value of X (or mean of X) is

$$E(X) = \int_{-\infty}^{\infty} x\, f_X(x) dx$$

- Let X be a Unif$(0,1)$ r.v.

$$E(X) = \int_{-\infty}^{\infty} x\, f_X(x) dx = \int_0^1 x dx = 1/2$$

- $X \sim Exp(\lambda)$, $E(X) = 1/\lambda$. In fact

$$\begin{aligned} E(X) &= \int_0^{\infty} x\lambda e^{-\lambda x} dx \\ &= -xe^{-\lambda x}\Big|_0^{\infty} + \int_0^{\infty} e^{-\lambda x} dx \\ &= [0-0] + \frac{1}{\lambda}\int_0^{\infty} \lambda e^{-\frac{1}{\lambda}x} dx = \frac{1}{\lambda} \end{aligned}$$

- $X \sim Binomial(N, p)$, $E(X) = Np$
- $X \sim Poisson(\lambda)$, $E(X) = \lambda$
- $X \sim Exp(\lambda)$, $E(X) = 1/\lambda$
- $X \sim N(0, 1)$, $E(X) = 0$

Mean of a trasformation of a r.v.

- To calculate the mean of a trasformation $Y = g(X)$ we can obtain the p.d.f. or the density of Y and apply the definition of expected value
- Alternatively, we can use directly the following

  Theorem Let the random variables X and Y satisfy $Y = g(X)$ where $g(\cdot)$ is a real-valued function on $\mathbb{R}$.

  1. If X is discrete with probability distribution $p_X(x)$

  $$E(Y) = \sum_x g(x)p_X(x)$$

  2. If X is continuous with density $f_X(x)$

  $$E(Y) = \int_{-\infty}^{\infty} g(x)f_X(x)$$

- Linearity

$$E(a + bX) = a + bE(X)$$

In fact

$$\begin{aligned} E(a + bX) &= \int_{-\infty}^{\infty} (a + b\,x) f_X(x) dx \\ &= a \int_{-\infty}^{\infty} f_X(x) dx + b \int_{-\infty}^{\infty} x f_X(x) dx \\ &= a + bE(X) \end{aligned}$$

Variance

- Let X be a discrete r.v. with probability distribution function $p_X(x)$ for $x \in S$. The variance of X is

$$Var(X) = \sum_{x \in S} (x - E(X))^2 \, p_X(x)$$

- Let X be a continuous r.v. with density $f_X(x)$. The variance of X is

$$Var(X) = \int_{-\infty}^{\infty} (x - E(X))^2 \, f_X(x)$$

- $Var(a+bX) = b^2 Var(X)$. In fact

$$\begin{aligned}
Var(a+bX) &= E(((a+bX)-E(a+bX))^2) \\
&= E((a+bX-a-bE(X))^2) \\
&= E((b(X-E(X)))^2) \\
&= E(b^2(X-E(X))^2) = b^2 E((X-E(X))^2) \\
&= b^2 Var(X)
\end{aligned}$$

- $Var(X) = E(X^2) - E(X)^2$. In fact

$$\begin{aligned}
Var(X) &= \int (x-E(X))^2 f_X(x)dx = \\
&= \int (x^2 + E(X)^2 - 2xE(X)) f_X(x)dx \\
&= \int x^2 f_X(x)dx + E(X)^2 - 2E(X)\int x f_X(x)dx \\
&= E(X^2) + E(X)^2 - 2E(X)^2 = E(X^2) - E(X)^2
\end{aligned}$$

Bivariate random variables

- When the outcome of the random experiment is a pair of numbers (X, Y) we call (X, Y) a bivariate (or two-dimensional) random variable
- Example: result of a football match, cases of COVID19 today and tomorrow, income of husband and income of wife, cost and outcome of a medical treatment...
- When X and Y are both discrete we call $Z = (X, Y)$ a bivariate discrete random variable
- When X and Y are both continuous r.v. we call $Z = (X, Y)$ a bivariate continuous random variable

- The **joint** probability function of a discrete bivariate r.v. $Z = (X, Y)$ is the function

$$p_{X,Y}(x, y) = P(X = x, Y = y)$$

where x and y run over the possible values of X and Y

- Note that $P(X = x, Y = y) = P(X = x \cap Y = y)$
- If C is a subset of the possible values of $Z = (X, Y)$

$$P(Z \in C) = \sum_{(x,y) \in C} p_{X,Y}(x, y)$$

and if D is the set of all possible values of $Z = (X, Y)$ we have

$$P(Z \in D) = \sum_{(x,y)} p_{X,Y}(x, y) = 1$$

- By the joint probability function of (X, Y) we obtain $p_X(X = x) = p_X(x)$, i.e. the probability function of X that we call **marginal distribution of X**

$$p_X(x) = \sum_y p_{X,Y}(x, y)$$

$$\begin{aligned} p_X(x) &= P\left([X = x] \bigcap [\text{whatever result for } Y]\right) \\ &= P\left([X = x] \bigcap \left[\bigcup_y [Y = y]\right]\right) \\ &= P\left(\bigcup_y [X = x, Y = y]\right) \\ &= \sum_y p_{X,Y}(x, y) \end{aligned}$$

Similarly the **marginal distribution of Y** is given by

$$p_Y(y) = \sum_x p_{X,Y}(x, y)$$

- By applying the multiplication rule we have $\forall(x,y) : p_Y(y) > 0$

  $$p_{X,Y}(x,y) = P(Y=y)P(X=x|Y=y) = p_Y(y)P(X=x|Y=y)$$

  where

  $$P(X=x|Y=y) = \frac{P(X=x \cap Y=y)}{P(Y=y)}$$

  We will call

  $$p_{X|Y}(x|y) = P(X=x|Y=y)$$

  the **conditional distribution** of X given $Y=y$

- The joint distribution of (X,Y) can be obtained by specifying the marginal distribution of Y and the conditional distributions of $X|Y=y$ given all the values y

Continuous random variables

- Suppose now that X and Y are continuous random variables.
- More formally, let $B = \{x : x \in B_x, y : y \in B_y\}$ be the product of the intervals B_x and B_y, i.e. $B = B_x \times B_y$. We say that $Z = (X, Y)$ is a bivariate continous r.v. if it exists a real function $f_{X,Y}(x, y)$ such taht

$$f(x, y) \geq 0 \quad \forall (x, y) \in R^2$$

$$\begin{aligned} P[(X,Y) \in B] &= P(X \in B_x, Y \in B_y) = \int_B f_{X,Y}(x,y)dxdy \\ &= \int_{B_y} \left[\int_{B_x} f_{X,Y}(x,y)dx \right] dy \end{aligned}$$

The function $f_{X,Y}(x, y)$ is called the density of the the r.v.

- $f(x) = \int_{B_y} f_{X,Y}(x, y)dy$
- $f(y) = \int_{B_x} f_{X,Y}(x, y)dx$

Statistical Inference

- Statistics aims to extract information about the system that generated the data
  - Are the data at hand the best representation of the system?
  - What about the data variability?
- Statistical models: mathematical cartoons (with unknown quantities...parameters) describing how data might have been generated
  - If the unknowns were known, a *good model* should *generate* data that resemble the observed ones...reproducing their variability
  - Statistical inference goes in the reverse direction: given a statistical model we take the unknown values of the model that are consistent with the observed data

- The class of statistical models is huge:
  time series models, non linear models, generalized regression models, Markov models, mixture models, hidden Markov models, latent variable models, spatial models, spatiotemporal models, hierarchical models, change point models, extreme value models.. non parametric models ....
- Statisticians often mix up different models to improve the adequacy of the resulting model to the data $\boldsymbol{y}$
- More complicated models may lead to $f(\mathbf{y};\boldsymbol{\theta})$ that can be only numerically evaluated... or that cannot be evaluated at all.

Inferential questions

- Given some data $\boldsymbol{y}$ and a statistical model with parameters $\boldsymbol{\theta}$ we may ask
  - What values for $\boldsymbol{\theta}$ are most consistent with $\mathbf{y}$? → *point estimation*
  - Is some prespecifed restriction on $\boldsymbol{\theta}$ consistent with $\mathbf{y}$? → *hypothesis testing*
  - Which of several alternative models/hypothesis is most consistent with $\boldsymbol{y}$? → *model selection*
  - What ranges of values of $\boldsymbol{\theta}$ are consistent with $\boldsymbol{y}$? → *interval estimation*
  - Is the model consistent with $\boldsymbol{y}$ for any values of $\boldsymbol{\theta}$? → *model cheking*
  - The data gathering process can be optimized? → *experimental or sampling design*
- Statistics zigzag up and down across these questions

Point estimation tecniques that we will use

- Maximum likelihood estimation:
    - assume a *statistical model* $f(x;\theta)$ basing on the observed data;
    - define and maximize the *likelihood function* $L(\theta)$ w.r.t. θ.

- Ordinary least squares (linear regression)
    - minimize the sum of the squares of the differences between the observed dependent variable and those predicted by the linear function of the independent variable.

Likelihood and log-likelihood function

- Let x be a realization of a random variable or vector X with probability function (in the book is called mass function) or density function $f(x;\theta)$
- The function $f(x;\theta)$ depends on x and on (unknown) parameter θ. Note that f is assumed known, for instance it can be the density of a random sample or of a more complex statistical model
- The Greek letter θ will be used for general notation. In specific examples we will adopt different Greek letters
  - Let X be a Poisson random sample and let $x = (x_1, \dots, x_n)$ be the realization of X. Then

$$f(x;\lambda) = \prod_{i=1}^{n} \frac{e^{-\lambda}\lambda^{x_i}}{x_i!} = \frac{e^{-\lambda n}\lambda^{\sum_{i=1}^{n} x_i}}{\prod_{i=1}^{n} x_i!}$$

- The parameter θ can be a scalar or a vector (for example in the normal case we have (μ, σ^2)). Vector parameters will be denoted with $\boldsymbol{\theta}$

- The space of all possible realizations of X will be denoted with $\mathcal{T}$ and called *sample space*
- The parameter θ takes values in the *parameter space* Θ
- **Definition** The likelihood function

  $$L(\theta) = L(\theta; x) = f(x; \theta) \quad \theta \in \Theta$$

  is the probability function or density function of the observed data x, viewed as a function of the unknown parameter θ
- Parameter values that make the observed data appear relatively probable are more likely to be correct than parameter values that make the observed data appear relatively improbable

Example: Inference for a proportion

- Inference for a proportion. Consider X ∼Binomial(n, π): For example $X = x$ may represent the observed number of students with a mac computer among n randomly selected students.
- The number n of selected students n is hence known, while the true proportion π of mac users in the student population is unknown.
- When $n = 20$ and $x = 8$ the likelihood is

$$L(\pi) = f(x; \pi) = \binom{n}{x} \pi^x (1 - \pi)^{n-x} \quad for\ \pi \in (0, 1)$$

Maximum likelihood estimate

- Plausible values of θ should have a relatively high likelihood.
- **Definition. Maximum likelihood estimate**. The maximum likelihood estimate (MLE) $\hat{\theta}_{ML}$ of a parameter θ is the point where the likelihood assumes the maximum value

$$\hat{\theta}_{ML} = \text{arg}\max_{\theta \in \Theta} L(\theta)$$

- In order to compute the MLE we can ignore multiplicative constant.

- **Definition. Likelihood kernel**. The likelihood kernel is obtained by removing from a likelihood function all multiplicative constants. We will use the same symbol $L(\theta)$ both for likelihoods, likelihood kernels and each function $a \cdot L(\theta)$

- In the binomial example

$$L(\pi) = \binom{n}{x} \pi^x (1-\pi)^{n-x}$$

but also

$$L(\pi) = \pi^x (1-\pi)^{n-x}$$

where the last expression is the likelihood kernel. We will also use the notation

$$L(\pi) = \binom{n}{x} \pi^x (1-\pi)^{n-x} \propto \pi^x (1-\pi)^{n-x}$$

- It is very often convenient to use the **log-likelihood** function

$$\ell(\theta) = \log L(\theta)$$

In fact the logarithm is a strictly monotone function and

$$\hat{\theta}_{ML} = \operatorname{arg}\max_{\theta \in \Theta} \ell(\theta)$$

- Multiplicative constant in $L(\theta)$ turn to additive constants in $\ell(\theta)$. For example in the binomial example we have

$$\ell(\pi) = \log\left\{\binom{n}{x}\pi^x(1-\pi)^{n-x}\right\} =$$

$$= \log\binom{n}{x} \; + x\log\pi + (n-x)\log(1-\pi)$$

  and additive constant in the loglikelihood can be ignored so we have also

$$\ell(\pi) = x\log\pi + (n-x)\log(1-\pi)$$

- Note that in the binomial case we have

$$\ell'(\pi) = \frac{d\ell(\pi)}{d\pi} = \frac{x}{\pi} - \frac{n-x}{1-\pi}$$

  and $\ell'(\pi) = 0$ when $\pi = x/n$ and $\ell''(\pi) < 0$ so x/n is MLE

Confidence Intervals

A confidence interval (CI) is a range of estimates for an unknown parameter. A confidence interval is computed at a designated confidence level.

The confidence level represents the long-run proportion of corresponding CIs that contain the true value of the parameter. For example, out of all intervals computed at the 95% level, 95% of them should contain the parameter's true value.

Typically a rule for constructing confidence intervals is closely tied to a particular way of finding a point estimate of the quantity being considered.