Biomedical Engineering Degree

1. Probability and Random Variables

Felipe Alonso Atienza

✉felipe.alonso@urjc.es

🐦@FelipeURJC

Escuela Técnica Superior de Ingeniería de Telecomunicación
Universidad Rey Juan Carlos

References

1. R. Bernard. *Fundamentals of Biostatistics*. Ed.: Thompson. Chapters 3-5
2. B. Caffo. *Statistical Inference for Data Science*. Leanpub. Chapters 2-6

Outline

1. Probability
   - Conditional probability

2. Random variables
   - Mass functions and densities
   - CDF and survival function
   - Mean and variance
   - Some common distributions

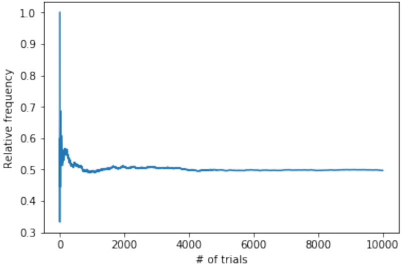
Intuition

- Let's run some random **experiments**:
  - Say you flip a (fair) coin, what's the probability of having a *HEAD*?
  - Say you roll a (fair) die, what's the probability of having 6?
- Our intuition says that these should be 0.5 and $1/6$, respectively.
- But, how can we prove it?
  - Flip the coin many times and divide the number of HEADs over the number of trials. So let's do it!

Flip a coin

Python code

```
N = 10000 #number of trials
trials = range(1,N+1)
p_heads = []; n_heads = 0
for i in trials:
 flip = random.randint(0, 1)
 n_heads += flip
 p_heads.append(n_heads/i)
plt.plot(trials,p_heads)
plt.xlabel('# of trials')
plt.ylabel('Relative frequency')
plt.show()
```



Homework

Write a python code to simulate the experiment of rolling a die

Definition

- **Sample space** (Ω): the set of all possible outcomes
  - Flipping a coin: $\Omega = \{\mathrm{HEAD}, \mathrm{TAIL}\}$
  - Rolling a die, $\Omega = \{1, 2, 3, 4, 5, 6\}$
- **Event** (A): any set of outcomes of interest
  - Flipping a coin: $A = \{\mathrm{HEAD}\}$
  - Rolling a die: $A = \text{odd number} = \{1, 3, 5\}$
- **Probability** of an event, denoted by $P(A)$, is the relative frequency of this set of outcomes over an indefinitely large (or infinite) number of trials (frequentist definition).
  - Flipping a coin: for $A = \{\mathrm{HEAD}\}$, $P(A) = 0.5$
  - Rolling a die: for $A = \{1, 3, 5\}$, $P(A) = 0.5$

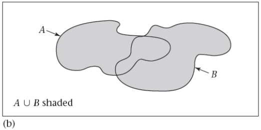
Axiomatic approximation

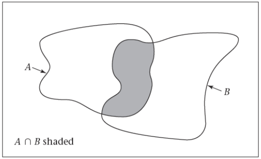
- Given a random experiment (say rolling a die) a probability measure is a **population quantity** that summarizes the randomness.
- Probability rules:
  1. It is a function that assigns a number to events so that $0 \leq P(A) \leq 1, \ \forall A$
  2. $P(\Omega) = 1$
  3. Be A and B two mutually exclusive events (they do not have nothing in common: $A \cap B = \emptyset$), then

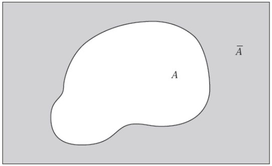
$$P(A \cup B) = P(A) + P(B)$$

Venn diagrams

- Probability can be understood using set operations, since $A \subset \Omega$.







- So, it can be stated that
  1. $P(A \cup B) = P(A) + P(B) - P(A \cap B)$
  2. $P(A^C) = 1 - P(A)$

Example

The National Sleep Foundation reports that around $3\,\%$ of the American population has sleep apnea. They also report that around $10\,\%$ of the North American and European population has restless leg syndrome. Does this imply that $13\,\%$ of people will have at least one sleep problems of these sorts?

Independence

- Statistical independence of events is the idea that the events are unrelated

Two events A and B are independent if

$$P(A \cap B) = P(A)P(B)$$

Example

What is the probability of getting two consecutive heads?

Conditional probability

- It is the probability of an event A, knowing that the event B has occurred (conditioned on B), and it is expressed as $P(A|B)$
  - What's the probability of getting a one if the die roll was an odd number?
- Let B be an event so that $P(B) > 0$, then

$$P(A|B) = \frac{P(A \cap B)}{P(B)}$$

- If A and B are independent, then

$$P(A|B) = \frac{P(A)P(B)}{P(B)} = P(A)$$

that is, the occurrence of B offers no information about the occurrence of A.

Total-probability rule

- For any events A and B

$$P(A) = P(A \cap B) + P(A \cap B^C)$$

and since $P(A \cap B) = P(A|B)P(B)$, then

$$\boxed{P(A) = P(A|B)P(B) + P(A|B^C)P(B^C)}$$

Bayes rule

- It allows us to reverse the conditioning set provided that we know some marginal probabilities.

$$P(B|A) = \frac{P(A|B)P(B)}{P(A|B)P(B) + P(A|B^C)P(B^C)}$$

- Note that the denominator corresponds to $P(A)$ so we might find bayes rules written as

$$\boxed{P(B|A) = \frac{P(A|B)P(B)}{P(A)}}$$

  where

  - $P(B)$ is the prior probability
  - $P(A|B)$ is the likelihood
  - $P(B|A)$ is the posterior probability

- It is very useful in terms of **diagnostic tests**.

Diagnostic tests (detection problem): definitions

- Let $+$ and $-$ be the events that the result of a diagnostic test is positive or negative, respectively
- Let D and D^C be the event that the subject of the test has or does not have the disease, respectively
- The **sensitivity** $P(+|D)$ is the probability that the test is positive given that the subject actually has the disease
- The **specificity** $P(-|D^C)$ is the probability that the test is negative given that the subject actually has the disease
- The **prevalence** of the disease $P(D)$, which is the marginal probability of disease

And the quantities that we'd like to know are the predictive values

- The **positive predictive value** $P(D|+)$ is the probability that the subject has the disease given that the test is positive, or
- The **negative predictive value** $P(D^C|-)$ is the probability that the subject does not have the disease given that the test is negative

Diagnostic tests: example

A study comparing the efficacy of HIV tests, reports on an experiment which concluded that HIV antibody tests have a sensitivity of $99.7\,\%$ and a specificity of $98.5\,\%$. Suppose that a subject, from a population with a $0.1\,\%$ prevalence of HIV, receives a positive test result. What is the positive predictive value?

Example solution

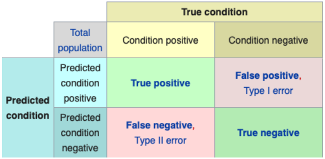
- Answer: Mathematically, we want $P(D \mid +)$ given the sensitivity, $P(+ \mid D) = .997$, the specificity, $P(- \mid D^c) = .985$ and the prevalence $P(D) = .001$.

$$\begin{aligned}
P(D \mid +) &= \frac{P(+ \mid D)P(D)}{P(+ \mid D)P(D) + P(+ \mid D^c)P(D^c)} \\
&= \frac{P(+ \mid D)P(D)}{P(+ \mid D)P(D) + \{1 - P(- \mid D^c)\}\{1 - P(D)\}} \\
&= \frac{.997 \times .001}{.997 \times .001 + .015 \times .999} \\
&= .062
\end{aligned}$$

- In this population a positive test result only suggests a $6\,\%$ probability that the subject has the disease, (the positive predictive value is $6\,\%$ for this test). If you were wondering how it could be so low for this test, the low positive predictive value is due to low prevalence of disease and the somewhat modest specificity

Confusion matrix

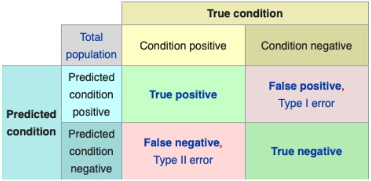
- There are four situations in a diagnostic test
  - True positive (TP): # cases for with the subject has the disease and the and the test is positive
  - True negative (TN): # cases for with the subject does not have the disease and the and the test is negative
  - False positive (FP): # cases for which the subject does not have the disease and the and the test is positive
  - False positive (FN): # cases for which the subject has the disease and the and the test is negative
- This information can be presented in matrix form



		True condition	
	Total population	Condition positive	Condition negative
Predicted condition	Predicted condition positive	True positive	False positive, Type I error
	Predicted condition negative	False negative, Type II error	True negative

Confusion matrix

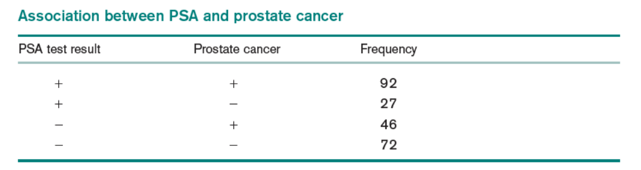
- Sensitivity (recall): $\frac{TP}{TP+FN}$
- Specificity: $\frac{TN}{TN+FN}$
- Positive predictive value: $\frac{TP}{TP+FP}$
- Negative predictive value: $\frac{TN}{TN+FN}$



		True condition	
	Total population	Condition positive	Condition negative
Predicted condition	Predicted condition positive	**True positive**	**False positive**, Type I error
	Predicted condition negative	**False negative**, Type II error	**True negative**

Example

Calculate the sensitivity, specificity and positive predictive value in the following study



Association between PSA and prostate cancer

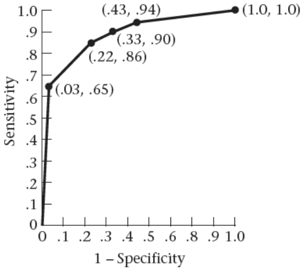
PSA test result	Prostate cancer	Frequency
+	+	92
+	−	27
−	+	46
−	−	72

ROC curves

- A **receiver operating characteristic** (ROC) curve is a plot of:
  - The sensitivity (on the y-axis)
  - ($1-$ specificity) (on the x-axis)

  of a screening test, where the different points on the curve correspond to different cutoff points used to designate test-positive.
- Extended use in (physiological) signal processing and machine learning in classification/detection scenarios.

ROC curve for the data in Table 3.4*



*Each point represents (1 – specificity, sensitivity) for different test-positive criteria.

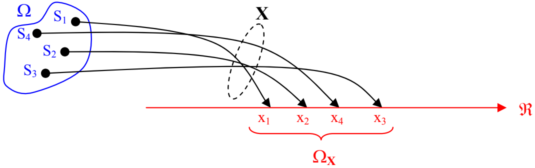
Outline

1. Probability
    - Conditional probability

2. Random variables
    - Mass functions and densities
    - CDF and survival function
    - Mean and variance
    - Some common distributions

Random variables (r.v.)

- A random variable X is a function that assigns numeric values to different events in a sample space $X : \Omega \to \mathbb{R}$



- There are two types of random variables: **discrete** or **continuous**
  - Discrete: take on only a countable number of possibilities
  - Continuous: take any value on the real line or some subset of the real line

Random variables: examples

1. Experiment: card from the Spanish deck drawn at random

   $$\Omega = \{\text{“as de oros”}, \text{“dos de oros”}, \ldots, \text{“rey de bastos”}\}, \text{ with } |\Omega| = 40$$

   - r.v. 1: card number, $\Omega_x = \{1, 2, 3, 4, 5, 6, 7, 10, 11, 12\}$
   - r.v. 2: gold vs no-gold card, $\Omega_x = \{0, 1\}$

2. Experiment: flipping a coin
   - r.v.: $\Omega_x = \{0, 1\}$

3. Experiment: blood pressure measurement
   - r.v.: $\Omega_x \in [80, 240]\,\mathrm{mmHg}$

Mass functions and densities

- We need convenient mathematical functions to model the probabilities of collections of realizations.
- These functions, called **mass functions** and **densities**, take possible values of the random variables, and assign the associated probabilities.
- These entities describe the **population of interest**
  - We might say that body mass indices follow a normal distribution, and this is a statement about the population of interest
  - Some of our goals will be to use our data to figure out things about that normal distribution, where it's centered, how spread out it is and even whether our assumption of normality is warranted!

Probability mass function

- A Probability mass function (pmf) evaluated at a value x_i corresponds to the probability that a random variable takes that value

$$f_X(x_i) = P(X = x_i)$$

- To be a valid pmf a function $f_X(x)$ must satisfy
  1. $f(x) \geq 0, \quad \forall x \in \mathbb{R}$
  2. $\sum_i f_X(x_i) = 1$

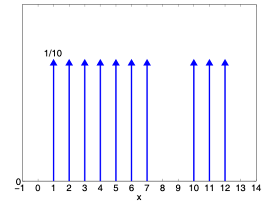
Probability mass function: example 1

- Let X be the card number from the Spanish deck so that

$$\Omega_x = \{1, 2, 3, 4, 5, 6, 7, 10, 11, 12\}$$

where

$$f_X(x) = P(X = x) = \sum_i P(X = x_i)\delta(x - x_i)$$



What's the probability of extracting a "sota"?

Probability mass function: example 2

- Let X be the result of a coin flip where $X = 0$ represents tails and $X = 1$ represents heads.

$$f_X(x) = (1/2)^x \, (1/2)^{1-x} \, , \quad x = 0, 1$$

- Suppose that we do not know whether or not the coin is fair; Let θ be the probability of a head expressed as a proportion (between 0 and 1), then

$$f_X(x) = \theta^x \, (1-\theta)^{1-x} \, , \quad x = 0, 1$$

Could we calculate the value of θ based on our data? Estimation theory

Probability density function (pdf)

- It is a function for characterizing continuous r.v. so that **areas under PDFs correspond to probabilities for that r.v.**

$$P(x_1 < X \leq x_2) = \int_{x_1}^{x_2} f_X(x)dx$$

- To be a valid pdf a function $f_X(x)$ must satisfy
  1. $f_X(x) \geq 0, \quad \forall x \in \mathbb{R}$
  2. $\int_{-\infty}^{\infty} f_X(x)dx = 1$

Probability density function: example

- The life span (in years) of today's electronic devices follows an exponential trend such that

$$f_X(x) = \frac{1}{5}e^{-x/5}, \quad x > 0$$

Your turn

1. Represent $f_X(x)$
2. Is it a valid pdf?
3. What is the probability that an electronic device following this distribution will survive more than 6 years?

Cumulative distribution function

- Certain areas of pdf's and pmf's are so useful, we give them names
- The **cumulative distribution function** (CDF) of a random variable X returns the probability that the random variable is less than or equal to the value x

$$F_X(x) = P(X \leq x)$$

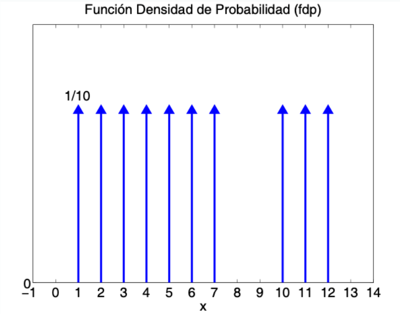
- This definition applies to both discrete and continuous r.v.'s.
- Properties:
  1. $F_X(-\infty) = 0$, $F_X(\infty) = 1$
  2. $0 \leq F_X(x) \leq 1$
  3. $P(x_1 < X < x_2) = F_X(x_2) - F_X(x_1)$
  4. $f_X(x) = \frac{dF_X(x)}{dx}$

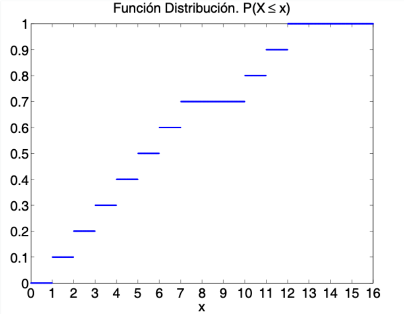
Cumulative distribution function: example

- Let X be the card number from the Spanish deck so that

$$\Omega_x = \{1, 2, 3, 4, 5, 6, 7, 10, 11, 12\}$$

then the cdf looks like this





Survival function

- The **survival function** of a random variable X is defined as the probability that the random variable is greater than the value x.

$$S_X(x) = P(X > x)$$

- Notice that $S_X(x) = 1 - F_X(x)$
- The survival function is often preferred in biostatistical applications while the distribution function is more generally used (though both convey the same information.)

Homework

1. What are the survival function and the cdf from the density considered before?

$$f_X(x) = \frac{1}{5}e^{-x/5}, \quad x > 0$$

2. Check cdf properties on your result

Quantiles and percentiles

- The α^{th} **quantile** of a cdf $F_X(x)$ is the point x_α so that

$$F_X(x_\alpha) = \alpha$$

  - As for example, the 0.95 quantile of a distribution is the point so that $95\,\%$ of the mass of the density lies below it

- A **percentile** is simply a quantile with α expressed as a percent rather than a proportion.
  - The (population) median is the 50^{th} percentile.
- Remember that
  - Percentiles are not probabilities!
  - Quantiles have units

Expected values

- Expected values characterize a distribution
- The **mean** characterizes the center of a pdf or pmf
- The **variance** characterizes how spread out a density is
- Yet another expected value calculation is the **skewness**, which considers how much a density is pulled toward high or low values
- Do not confuse the sample/empirical mean with the population mean

Population mean

- The **expected value** or (population) mean of a r.v. is the center of its distribution

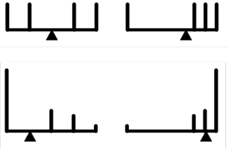
Discrete r.v.

$$\mu = E[X] = \sum_i x_i P(X = x_i) dx$$

Continuous r.v.

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

- It represents the center of mass (discrete case) of a collection of locations and weights $\{x_i, f_X(x_i)\}$



Expected value properties

1. Let X and Y be two r.v.'s and a and b two real-valued constants, then

$$E\left[aX + by\right] = aE[X] + bE[Y]$$

2. r.v.'s transformation

$$E\left[g(x)\right] = \int_{-\infty}^{\infty} g(x) f_X(x) dx$$

Examples

1. Suppose that a die is rolled and X is the number face up. What is the expected value of X?
2. Two dice are thrown, what is the expected value of averaging the results obtained?

Variance

- The variance, on the other hand, is a measure of spread of a distribution
- If X is a random variable with mean μ, the variance of X is defined as

$$\mathrm{Var}[X] = \sigma^2 = E[(X-\mu)^2]$$

that is, the expected (squared) distance from the mean

- **Standard deviation** $\sigma = \sqrt{\mathrm{Var}[X]} \geq 0$, which has the same units as X.

Properties

1. $\mathrm{Var}[X] = E[(X-\mu)^2] = E[X^2] - \cancel{E[X]^2} \rightarrow \mu^2$
2. $\mathrm{Var}[aX+b] = a^2\mathrm{Var}[X]$

Examples

1. What's the variance of a toss of a die?
2. What's the variance of the toss of a (potentially biased) coin with probability of heads (1) of p?

Bernoulli distribution

- The Bernoulli distribution arises as the result of a **binary outcome**, such as a coin flip.
- Thus, Bernoulli r.v.'s take (only) the values 1 and 0 with probabilities of (say) p and $1-p$, respectively
- Recall that the pmf for a Bernoulli r.v. variable X is

$$f_X(x) = p^x(1-p)^{1-x}$$

  with $E[X] = p$ and $\mathrm{Var}[X] = p(1-p)$
- If we let X be a Bernoulli random variable, it is typical to call $X = 1$ as a "success" and $X = 0$ as a "failure".
- If a random variable follows a Bernoulli distribution with success probability p we write that X ∼Bernoulli(p).

Binomial distribution

- The binomial r.v.'s are obtained as the **sum of independent Bernoulli trials**.
    - So if a Bernoulli trial is the result of a coin flip, a binomial random variable is the total number of heads.
- Let $X_1, \ldots, X_n$ be independent Bernoulli(p), then $X = \sum_{i=1}^{n} X_i$ is a binomial r.v. We write out that $X \sim \mathrm{Binomial}(n, p)$. The binomial mass function is

$$f_X(x) = \left(\begin{array}{c} n \\ x \end{array} \right) p^x (1-p)^{n-x}$$

with $E[X] = np$ and $\mathrm{Var}[X] = np(1-p)$

Poisson distribution

- The Poisson distribution is especially useful for modeling unbounded counts or counts per unit of time (rates)
  - The number of clicks on advertisements
  - The number of events in a period of time (like the number of people who show up at a bus stop).
- The Poisson mass function is:

$$f_X(x;\lambda) = \frac{\lambda^x e^{-\lambda}}{x!}, \, x = 0, 1, \ldots$$

with $E[X] = \lambda$ and $\mathrm{Var}[X] = \lambda$

Poisson distribution

- The Poisson distribution is useful for rates, counts that occur over units of time.
- Specifically, if $X \sim Poisson(\lambda t)$ then $\lambda = E[X/t]$ is the **expected count per unit of time** and t is the total monitoring time.

Example

The number of people that show up at a bus stop is Poisson with a mean of 2.5 per hour. If watching the bus stop for 4 hours, what is the probability that 3 or fewer people show up for the whole time?

Poisson approximation to the binomial

- The binomial distribution with large n and small p can be accurately approximated by a Poisson distribution.

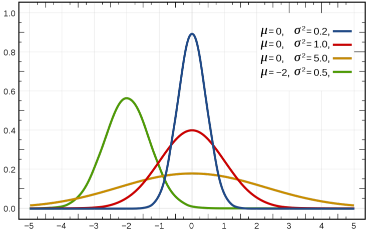
Formally, if $X \sim \text{Binomial}(n, p)$ then X is approximately Poisson where $\lambda = np$ provided that n is large p is small.

Normal distribution

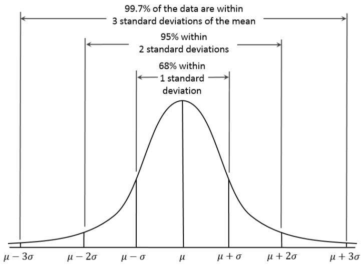
- The pdf of the **normal** or **gaussian distribution** distribution is given as

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

  with $E[X] = \mu$ and $\mathrm{Var}[X] = \sigma^2$
- We write $X \sim N(\mu, \sigma^2)$ to denote a normal random variable
- When $\mu = 0$ and $\sigma = 1$ the resulting distribution is called the **standard normal distribution**. Standard normal r.v.'s are often labeled Z



Reference quantiles for the standard normal



1. Approximately $68\,\%$, $95\,\%$ and $99\,\%$ of the normal density lies within 1, 2 and 3 standard deviations from the mean, respectively.
2. -1.28, -1.645, -1.96 and -2.33 are the 10^{th}, 5^{th}, 2.5^{th} and 1^{st} percentiles of the standard normal distribution, respectively.
3. By symmetry, 1.28, 1.645, 1.96 and 2.33 are the 90^{th}, 95^{th}, 97.5^{th} and 99^{th} percentiles of the standard normal distribution, respectively.

Shifting and scaling normals

- We can transform normal random variables to be standard normals and viceversa
- For example, if $X \sim N(\mu, \sigma^2)$ then:

$$Z = \frac{X - \mu}{\sigma} \sim N(0, 1)$$

- If Z is standard normal

$$X = \mu + \sigma Z \sim N(\mu, \sigma^2)$$

then $X \sim N(\mu, \sigma^2)$

Shifting and scaling normals

- We can use these facts to answer questions about non-standard normals by relating them back to the standard normal.

Example

1. Population mean body mass index (BMI) for men is reported as $29\mathrm{kg}/\mathrm{mg}^2$ with a standard deviation of 4.73. Assuming normality of BMI, what is the population 95^{th} percentile
2. What's the probability that a randomly drawn subject from this population has a BMI less than 24.27?