Rizzo

Chapter 2: Probability for Statistical Computing

just like we did for R

We will **briefly** review some definitions and concepts in probability and statistics that will be helpful for the remainder of the class.

Just like we reviewed computational tools (`R` and packages), we will now do the same for probability and statistics.

Note: This is not meant to be comprehensive. I am assuming you already know this and maybe have forgotten a few things.

i.e. you may need to do some refreshing outside of class as well.

https://xkcd.com/892/

Alternative text: "Hell, my eighth grade science class managed to conclusively reject it just based on a classroom experiment. It's pretty sad to hear about million-dollar research teams who can't even manage that."

1 Random Variables and Probability

Definition 1.1 A *random variable* is a function that maps sets of all possible outcomes of an experiment (sample space Ω) to $\mathbb{R}$.

real number line.

Example 1.1

Toss 2 dice

X = sum of the dice

r.v.

Example 1.2

Randomly select 25 deer & test for CWD (chronic wasting disease)

Sample {+, − CWD test}

X -- $\{0, 1\}$ observe $X_1, \ldots, X_{25}$ each a r.v.

Note $P = \sum_{i=1}^{25} X_i / 25$ is also a r.v.!

Example 1.3

Today's high temperature $= X_i$

Types of random variables –

Discrete take values in a countable set.

Ex. 1.1 and X_i from Ex 1.2

Continuous take values in an uncountable set (like $\mathbb{R}$)

real numbers $(-\infty, \infty)$

Ex. 1.3 $x_i \in \mathbb{R}$

P from ex. 1.2, $p \in [0, 1]$.

1.1 Distribution and Density Functions

Definition 1.2 The *probability mass function (pmf)* of a random variable X is f_X defined by

only defined for discrete r.v.'s

Sometimes when the r.v. is obvious we'll omit the subscript, just write $f(x)$.

$$f_X(x) = P(X = x)$$

where $P(\cdot)$ denotes the probability of its argument.

There are a few requirements of a **valid** pmf

1. $f(x) \geq 0$ for all x
2. $\sum_x f(x) = 1$

Not a requirement

3. We call $\mathcal{X} = \{x : f(x) > 0\}$ the "support" of X.

Example 1.4 Let $\Omega =$ all possible values of a roll of a single fair die $= \{1, \ldots, 6\}$ and X be the outcome of a single roll of one die $\in \{1, \ldots, 6\}$.

$f(1) = P(X=1) = \frac{1}{6}$

$\vdots$

$f(6) = \frac{1}{6}$

1. $f(x) \geq 0 \ \forall x$. ✓
2. $\sum_{x \in \mathcal{X}} f(x) = \sum_{i=1}^{6} \frac{1}{6} = 1$ ✓

valid pmf.

support $\mathcal{X} = \{1, 2, \ldots, 6\}$.

A pmf is defined for **discrete variables**, but what about **continuous**? Continuous variables do not have positive probability mass at any single point.

Definition 1.3 The *probability density function (pdf)* of a random variable X is f_X defined by

$$P(X \in A) = \int_{x \in A} f_X(x)dx.$$

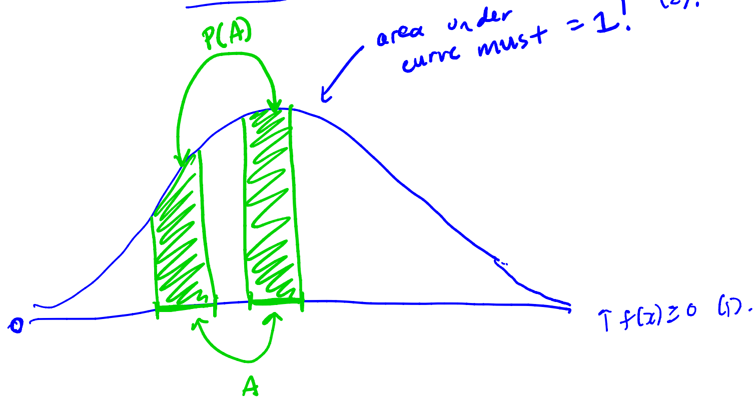
$A \subseteq \mathbb{R}$

X is a continuous random variable if there exists this function $f_X \geq 0$ such that for all $x \in \mathbb{R}$, this probability exists.

For f_X to be a valid pdf,

1. $f(x) \geq 0 \ \forall x$
2. $\int_{\mathbb{R}} f(x)dx = 1$

Again $\mathcal{X} = \{x : f(x) > 0\}$ is the "support" of X.



There are many named pdfs and cdfs that you have seen in other class, e.g.

Gamma, Poisson, Normal, Uniform, student t, snedecor's F, χ^2, binomial, exponential, Beta, hypergeometric, ...

Example 1.5 Let

shape

$$f(x) = \begin{cases} c(4x - 2x^2) & 0 < x < 2 \\ 0 & \text{otherwise} \end{cases}$$

← the support "(0,2)."

to make f a valid pdf

Find c and then find $P(X > 1)$

$$1 = \int_0^2 c(4x - 2x^2)\,dx = c\left[2x^2 - \frac{2x^3}{3}\right]_0^2 = c\left[\frac{8}{3}\right] \Rightarrow c = \frac{3}{8}$$

"normalizing constant"

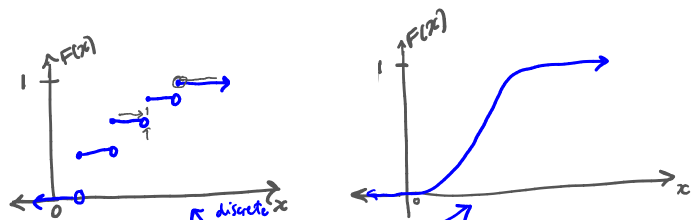
$$P(X>1) = \int_1^\infty f(x)\,dx = \int_1^2 \frac{3}{8}(4x - 2x^2)\,dx = \frac{3}{8}\left[2x^2 - \frac{2x^3}{3}\right]_1^2 = \frac{1}{2}$$

Definition 1.4 The *cumulative distribution function (cdf)* for a random variable X is F_X defined by

↓ for both cts and discrete.

$$F_X(x) = P(X \le x), \quad x \in \mathbb{R}.$$

The cdf has the following properties



1. F_X is non decreasing
2. F_X is right-continuous
3. $\lim_{x \to -\infty} F_X(x) = 0$ and $\lim_{x \to \infty} F_X(x) = 1$

A random variable X is *continuous* if F_X is a continuous function and *discrete* if F_X is a step function.

Example 1.6 Find the cdf for the previous example.

$$F(x) = P(X \le x) = \begin{cases} 0 & x \le 0 \\ \frac{3}{4}x^2\left(1 - \frac{x}{3}\right) & x \in (0,2) \\ 1 & x \ge 2 \end{cases}$$

For $x \in (0,2)$, $P(X \le x) = \int_0^x \frac{3}{8}(4y - 2y^2)\,dy$

$$= \frac{3}{8}\left[2y^2 - \frac{2y^3}{3}\right]_0^x$$

$$= \frac{3}{4}x^2\left(1 - \frac{x}{3}\right)$$

Note $f(x) = F'(x) = \frac{dF(x)}{dx}$ in the continuous case.

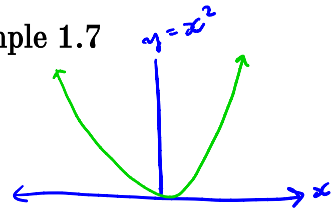
derivative of cdf wrt x = pdf.

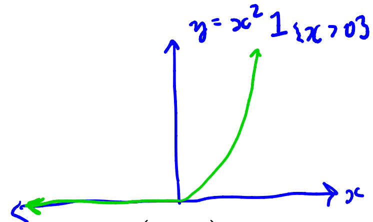
Recall an indicator function is defined as

$$1_{\{A\}} = \begin{cases} 1 & \text{if } A \text{ is true} \\ 0 & \text{otherwise} \end{cases}.$$

$1_A(x) = \begin{cases} 1 & \text{if } x \in A \\ 0 & \text{otherwise} \end{cases}$

Example 1.7

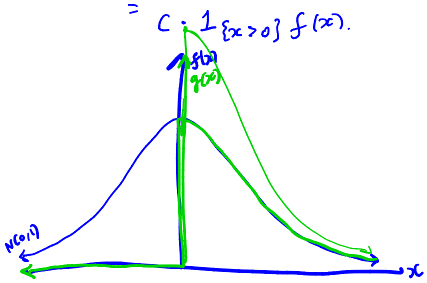




Example 1.8 If $X \sim N(0,1)$, the pdf is $f(x) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{x^2}{2}\right)$ for $-\infty < x < \infty$.

If $g(x) = \frac{c}{\sqrt{2\pi}} \exp\left(-\frac{x^2}{2}\right) 1_{\{x>0\}}$, what is c? in order for g to be a valid pdf?

$= c \cdot 1_{\{x>0\}} f(x).$



We know by symmetry of $N(0,1)$ around 0,

$$c \int_0^\infty \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{x^2}{2}\right) dx = c\,\frac{1}{2}$$

$\Rightarrow$ need $c = 2$.

Need: $1 = \int_{-\infty}^{\infty} g(x)\,dx = \int_0^\infty \frac{c}{\sqrt{2\pi}} \exp\left(-\frac{x^2}{2}\right) dx$

1.2 Two Continuous Random Variables

Definition 1.5 The *joint pdf* of the continuous vector (X, Y) is defined as

$$P((X,Y) \in A) = \iint_A f_{X,Y}(x,y)\,dx\,dy$$

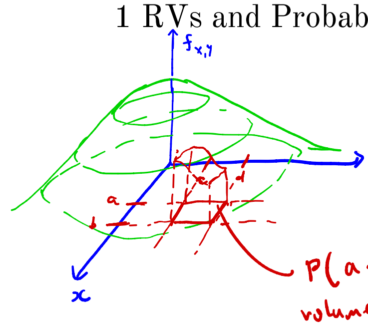
for any set $A \subset \mathbb{R}^2$.

Joint pdfs have the following properties

1. $f_{X,Y}(x,y) \geq 0 \quad \forall x, y \in \mathbb{R}^2$
2. $\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f_{X,Y}(x,y)\,dx = 1$

Note: can also have joint discrete r.v.'s where $\sum_x \sum f(x,y) = 1.$

and a support defined to be $\mathcal{X} = \{(x,y) \in \mathbb{R}^2 : f_{X,Y}(x,y) > 0\}$.



$P(a \le X \le b,\ c \le Y \le d)$
volume of the rectangle $[a,b] \times [c,d]$ under the curve.

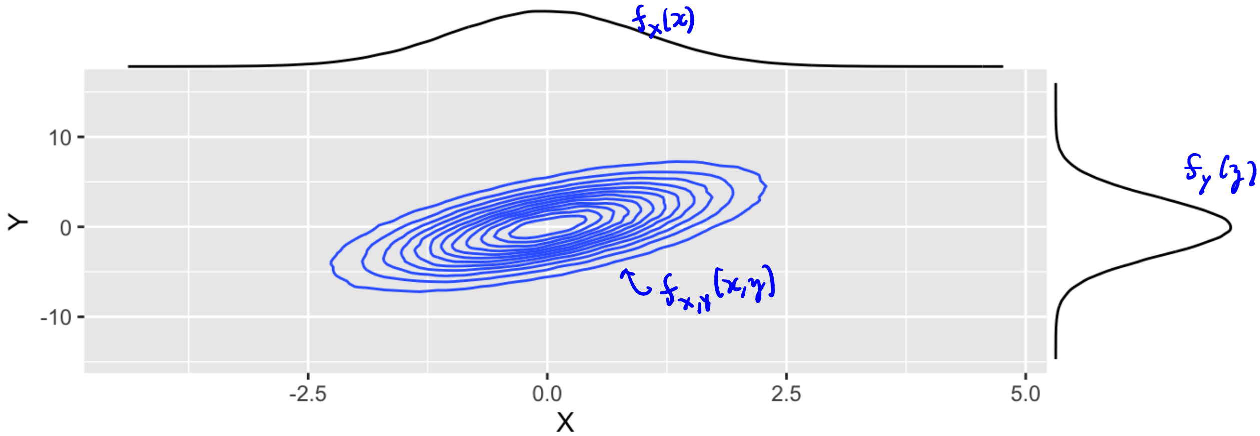
Example 1.9

$$P(a \le X \le b,\ c \le Y \le d) = \int_a^b \int_c^d f_{X,Y}(x,y)\,dy\,dx$$

The *marginal densities* of X and Y are given by

$$f_X(x) = \int_{\infty}^{\infty} f_{X,Y}(x,y)dy \quad \text{and} \quad f_Y(y) = \int_{\infty}^{\infty} f_{X,Y}(x,y)dx;$$

$x \in \mathbb{R}$ $\qquad$ $y \in \mathbb{R}$



Example 1.10 (From Devore (2008) Example 5.3, pg. 187) A bank operates both a drive-up facility and a walk-up window. On a randomly selected day, let X be the proportion of time that the drive-up facility is in use and Y is the proportion of time that the walk-up window is in use.

The the set of possible values for (X, Y) is the square $D = \{(x,y) : 0 \le x \le 1, 0 \le y \le 1\}$. Suppose the joint pdf is given by

$$f_{X,Y}(x,y) = \begin{cases} \frac{6}{5}(x+y^2) & x \in [0,1], y \in [0,1] \\ 0 & \text{otherwise} \end{cases}$$

NOTE: $\int_0^1 \int_0^1 \frac{6}{5}(x+y^2)\,dx\,dy = 1.$ valid joint pdf.

Evaluate the probability that both the drive-up and the walk-up windows are used a quarter of the time or less.

$$P\left(0 \le X \le \tfrac{1}{4},\ 0 \le Y \le \tfrac{1}{4}\right) = \int_0^{1/4} \int_0^{1/4} \frac{6}{5}(x+y^2)\,dx\,dy$$

$$= \int_0^{1/4} \frac{6}{5}\left[\frac{x^2}{2} + xy^2\right]_{x=0}^{x=1/4} dy$$

$$= \int_0^{1/4} \frac{6}{5}\left(\frac{1}{32} + \frac{y^2}{4}\right) dy$$

$$= \frac{6}{5}\left(\frac{y}{32} + \frac{y^3}{12}\right)\Bigg]_0^{1/4} = \frac{6}{5}\left(\frac{1}{32}\cdot\frac{1}{4} + \frac{1}{12}\left(\frac{1}{4}\right)^3\right] = \frac{7}{640} = 0.0109$$

Find the marginal densities for X and Y.

$$f_X(x) = \int_0^1 \frac{6}{5}(x+y^2)\,dy = \frac{6}{5}\left[xy + \frac{y^3}{3}\right]_{y=0}^1 = \begin{cases} \frac{6}{5}\left(x+\frac{1}{3}\right) & x \in [0,1] \\ 0 & \text{o.w.} \end{cases}$$

$$f_Y(y) = \int_0^1 \frac{6}{5}(x+y^2)\,dx = \frac{6}{5}\left[\frac{x^2}{2} + xy^2\right]_{x=0}^1 = \begin{cases} \frac{6}{5}\left(\frac{1}{2}+y^2\right) & y \in [0,1] \\ 0 & \text{o.w.} \end{cases}$$

X

Compute the probability that the drive-up facility is used a quarter of the time or less.

$$P\left(X \le \frac{1}{4}\right) = \int_0^{1/4} f_X(x)\,dx = \int_0^{1/4} \frac{6}{5}\left(x+\frac{1}{3}\right)dx = \frac{6}{5}\left[\frac{x^2}{2} + \frac{1}{3}x\right]_0^{1/4} = \frac{6}{5}\left[\frac{1}{4^2}\cdot\frac{1}{2} + \frac{1}{3}\cdot\frac{1}{4}\right]$$

$$= \frac{11}{80} = 0.1375.$$

2 Expected Value and Variance

Definition 2.1 The *expected value* (average or mean) of a random variable X with pdf or pmf f_X is defined as

$$E[X] = \begin{cases} \sum_{x \in \mathcal{X}} x f_X(x_i) & X \text{ is discrete} \quad (f_X \text{ is pmf}) \\ \int_{x \in \mathcal{X}} x f_X(x) dx & X \text{ is continuous.} \quad (f_X \text{ is pdf}) \end{cases}$$

Where $\mathcal{X} = \{x : f_X(x) > 0\}$ is the support of X.

This is a weighted average of all possible values $\mathcal{X}$ by the probability distribution.

Example 2.1 Let $X \sim$ Bernoulli(p). Find $E[X]$.

$$\Rightarrow \underset{\text{r.v.}}{X} = \begin{cases} 1 & \text{w.p. } p \\ 0 & \text{o.w.} \end{cases} \Rightarrow f(x) = \begin{cases} p & x=1 \\ 1-p & x=0 \\ 0 & \text{o.w.} \end{cases} \text{ or } \underbrace{f(x)}_{P(X=x) \,//\, \text{pmf}} = p^x(1-\underbrace{p}_{p \text{ parameter}})^{1-x} \text{ for } \underbrace{\boxed{x \in \{0,1\}}}_{\text{support}}.$$

$$E[X] = \sum_{x \in \{0,1\}} x f(x) = \sum_{x \in \{0,1\}} x \cdot p^x (1-p)^{1-x} = 0(1-p) + 1 \cdot p = \underline{p}$$

Example 2.2 Let $X \sim$ Exp(λ). Find $E[X]$.

parameter

$$f(x) = \begin{cases} \lambda e^{-\lambda x} & x \geq 0 \\ 0 & \text{o.w.} \end{cases}$$

integration by parts (HW 3)

$$\int u\, dv = uv - \int v\, du$$

pick u and dv

$$E[X] = \int_0^\infty \underbrace{\boxed{x}}_{u} \underbrace{\boxed{\lambda e^{-\lambda x} dx}}_{dv} = \frac{1}{\lambda}$$

CTS, analogous definition for pmf's.

Definition 2.2 Let $g(X)$ be a function of a continuous random variable X with pdf f_X. Then,

$$* \qquad E[g(X)] = \int_{x \in \mathcal{X}} g(x) f_X(x) dx.$$

Sometimes this is hard to compute by hand!
$\Rightarrow$ will need stat computing to estimate this value. (ch. 5).

Definition 2.3 The *variance* (a measure of spread) is defined as

$$\begin{aligned} Var[X] &= E\left[(X - E[X])^2\right] \quad \leftarrow \text{def'n} \\ &= E[X^2] - (E[X])^2 \quad \leftarrow \text{computational def'n} \end{aligned}$$

Example 2.3 Let X be the number of cylinders in a car engine. The following is the pmf function for the size of car engines.

x	4.0	6.0	8.0
f	0.5	0.3	0.2

Find

$E[X] = \sum_x x f(x) = 4(0.5) + 6(0.3) + 8(0.2) = 5.4$

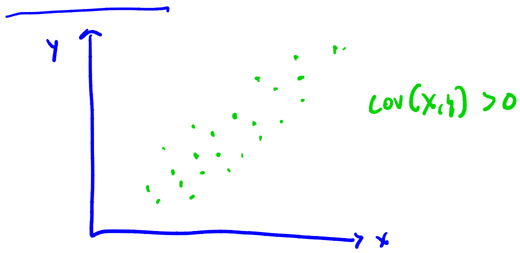
$Var[X] = \underline{E(X^2)} - (EX)^2$

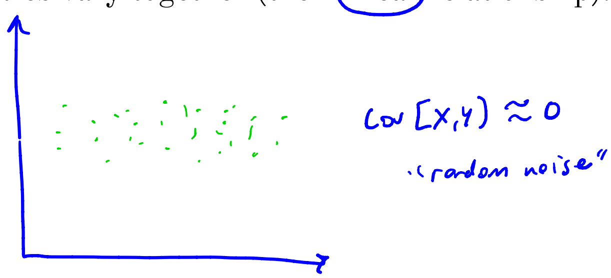
$E(X^2) = \sum_x x^2 f(x) = 4^2(0.5) + 6^2(0.3) + 8^2(0.2) = 31.6$

$g(x) = x^2$

$\Rightarrow Var[X] = 31.6 - (5.4)^2 = 2.44$ easier to interpret: $sd = \sqrt{Var X} = 1.56$

Covariance measures how two random variables vary together (their linear relationship).





Definition 2.4 The *covariance* of X and Y is defined by

$$\begin{aligned} Cov[X,Y] &= E\left[(X - E[X])(Y - E[Y])\right] \\ &= E[XY] - E[X]E[Y] \end{aligned}$$

and the *correlation* of X and Y is defined as

$$\rho(X,Y) = \frac{Cov[X,Y]}{\sqrt{Var[X]Var[Y]}}.$$

Two variables X and Y are *uncorrelated* if $\rho(X,Y) = 0$.

3 Independence and Conditional Probability

In classical probability, the *conditional probability* of an event A given that event B has occured is

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

Definition 3.1 Two events A and B are *independent* if $P(A|B) = P(A)$. The converse is also true, so

$$A \text{ and } B \text{ are independent} \Leftrightarrow P(A|B) = P(A) \Leftrightarrow P(A \cap B) =$$

Theorem 3.1 (Bayes' Theorem) Let A and B be events. Then,

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

3.1 Random variables

The same ideas hold for random variables. If X and Y have joint pdf $f_{X,Y}(x, y)$, then the conditional density of X given $Y = y$ is

$$f_{X|Y=y}(x) = \frac{f_{X,Y}(x, y)}{f_Y(y)}.$$

Thus, two random variables X and Y are independent if and only if

$$f_{X,Y}(x, y) = f_X(x)f_Y(y).$$

Also, if X and Y are independent, then

$$f_{X|Y=y}(x) =$$

4 Properties of Expected Value and Variance

Suppose that X and Y are random variables, and a and b are constants. Then the following hold:

1. $E[aX + b] =$

2. $E[X + Y] =$

3. If X and Y are independent, then $E[XY] =$

4. $Var[b] =$

5. $Var[aX + b] =$

6. If X and Y are independent, $Var[X + Y] =$

5 Random Samples

Definition 5.1 Random variables $\{X_1, \ldots, X_n\}$ are defined as a *random sample* from f_X if $X_1, \ldots, X_n \overset{iid}{\sim} f_X$.

Example 5.1

Theorem 5.1 If $X_1, \ldots, X_n \overset{iid}{\sim} f_X$, then

$$f(x_1, \ldots, x_n) = \prod_{i=1}^{n} f_X(x_i).$$

Example 5.2 Let $X_1, \ldots, X_n$ be iid. Derive the expected value and variance of the sample mean $\overline{X}_n = \frac{1}{n} \sum_{i=1}^{n} X_i$.

6 `R` Tips

From here on in the course we will be dealing with a lot of **randomness**. In other words, running our code will return a **random** result.

> But what about reproducibility??

When we generate "random" numbers in `R`, we are actually generating numbers that *look* random, but are *pseudo-random* (not really random). The vast majority of computer languages operate this way.

This means all is not lost for reproducibility!

```r
set.seed(400)
```

Before running our code, we can fix the starting point (`seed`) of the pseudorandom number generator so that we can reproduce results.

Speaking of generating numbers, we can generate numbers (also evaluate densities, distribution functions, and quantile functions) from named distributions in `R`.

```r
rnorm(100)
dnorm(x)
pnorm(x)
qnorm(y)
```