Chapter 2: Probability for Statistical Computing

Just like we did w/ 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 $\mathbb{R}$.

Example 1.1

Toss 2 dice

X = sum of the dice

$\uparrow$

r.v.

Example 1.2

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

Sample space $\{+, -$ on CWD test$\}$.

$X = \{0$ or $1\}$ observe $X_1, \ldots, X_{25}$

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

Example 1.3

Today's high temperature $= X$

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 $\mathbb{R}$)

Ex 1.3 $\leftarrow$ $X_i \in \mathbb{R}$

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

1.1 Distribution and Density Functions

Definition 1.2 The *probability mass function (pmf)* of a random variable X is f_X defined by
for discrete R.V.'s.

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

sometimes when the r.v. is obvious we will omit the subscript.

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

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

requirements
1. $f(x) \geq 0 \quad \forall x.$
2. $\sum_{x} f(x) = 1$

not really a requirement
3. We call $\mathcal{X} = \{x : f(x) > 0\}$ the "support" of X.

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

$$\left.\begin{array}{l} f(1) = \frac{1}{6} \\ \vdots \\ f(6) = \frac{1}{6} \end{array}\right\} \geq 0 \checkmark_1 \implies \sum_{x \in \mathcal{X}} f(x) = \sum_{i=1}^{6} \frac{1}{6} = 1 \checkmark_2$$

valid pmf ✓

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

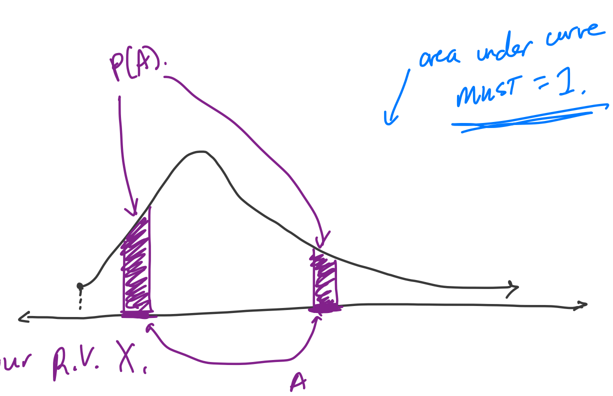
Definition 1.3 The *probability density function (pdf)* of a random variable X is f_X defined by

$$P(X \in A) = \int_{x \in A} f_X(x)dx. \qquad \text{for } A \subset \mathbb{R}.$$

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. $f_X(x) \geq 0 \quad \forall x \in \mathbb{R}.$
2. $\int_{\mathbb{R}} f_X(x)\,dx = 1$



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

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

Normal, Gamma, exponential, Beta, hypergeometric, Binomial, Poisson, ...

Example 1.5 Let

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

so that this is a valid pdf.

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

$$\int_{\mathbb{R}} f(x)\,dx = 1 \implies \int_0^2 c(4x-2x^2)\,dx + \int_{-\infty}^{0} \cancel{0}\,dx^{=0} + \int_2^{\infty} \cancel{0}\,dx^{=0}$$

$$= c\left[2x^2 - \frac{2x^3}{3}\right]_0^2 = c\left[\frac{8}{3}\right] \overset{\text{must}}{=} 1 \implies c = \frac{3}{8}$$

" normalizing constant"

$$P(X>1) = \int_1^{\infty} f(x)\,dx = \int_1^2 \frac{3}{8}(4x-2x^2)\,dx + \int_2^{\infty} \cancel{0}\,dx^{0} = \frac{3}{8}\left[2x^2 - \frac{2x^3}{3}\right]_1^2 = \frac{1}{2}.$$

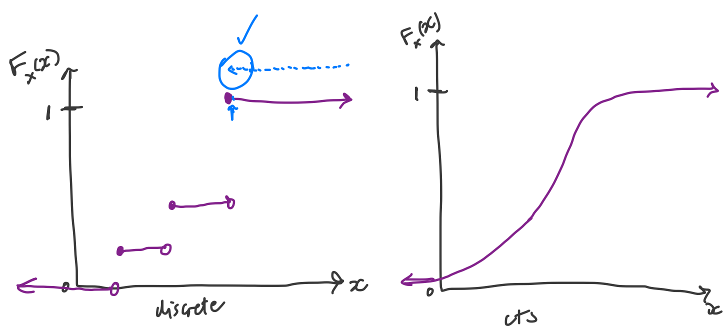
Definition 1.4 The *cumulative distribution function (cdf)* for a random variable X is F_X defined by

↳ for both cts and discrete r.v.

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

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.

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

if $x < 0$, $P(X \le x) = 0$

if $x \ge 2$, $P(X \le x) = 1$

if $x \in [0,2)$, $P(X \le x) = \int_0^x \frac{3}{8}(4y - 2y^2)\,dy = \frac{3}{8}\left[2y^2 - \frac{2y^3}{3}\right]_0^x = \frac{3}{4}x^2\left(1 - \frac{x}{3}\right)$

So $F_X(x) = \begin{cases} 0 & x < 0 \\ 3/4\,x^2(1-x/3) & x \in [0,2) \\ 1 & x \ge 2 \end{cases}$

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

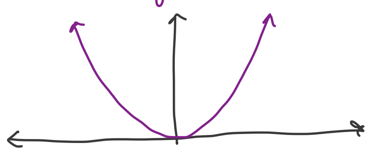
pdf deriv. of cdf.

Recall an indicator function is defined as

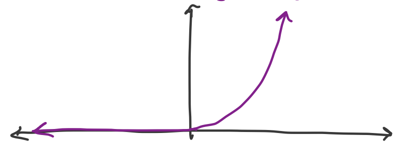
$$\mathbb{I}(A) = 1_{\{A\}} = \begin{cases} 1 & \text{if } A \text{ is true} \\ 0 & \text{otherwise} \end{cases}.$$

Example 1.7

$y = x^2$

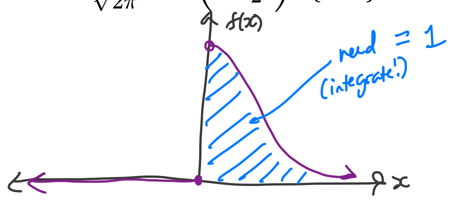


$$y = x^2 \mathbb{1}_{\{x \geq 0\}} = \begin{cases} x^2 \cdot 1 & x > 0 \\ x^2 \cdot 0 & x \leq 0 \end{cases}$$



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 $f(x) = \frac{c}{\sqrt{2\pi}} \exp\left(-\frac{x^2}{2}\right) 1_{\{x>0\}}$, what is c? so that f is a valid pdf



We know $N(0,1)$ symmetric around 0!

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

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

Need: $1 = \int_{-\infty}^{\infty} \frac{c}{\sqrt{2\pi}} \exp\left(-\frac{x^2}{2}\right) \mathbb{I}\{x \geq 0\}\, dx$

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

Need $1 = \frac{c}{2} \Rightarrow c = 2$

1.2 Two Continuous Random Variables

Definition 1.5 The *joint pdf* $f_{X,Y}(x,y)$ 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

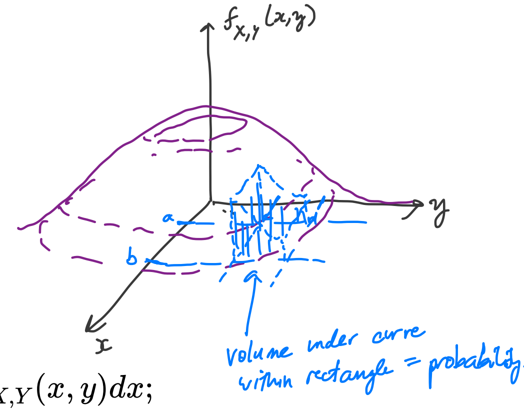
1. $f_{X,Y}(x,y) \geq 0 \quad \forall (x,y) \in \mathbb{R}^2$
2. $\iint_{\mathbb{R}^2} f_{X,Y}(x,y)\, dx dy = 1.$

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

Note we can also have joint discrete r.v.'s where $\sum_{\mathcal{X}} \sum f_{X,Y}(x,y) = 1.$

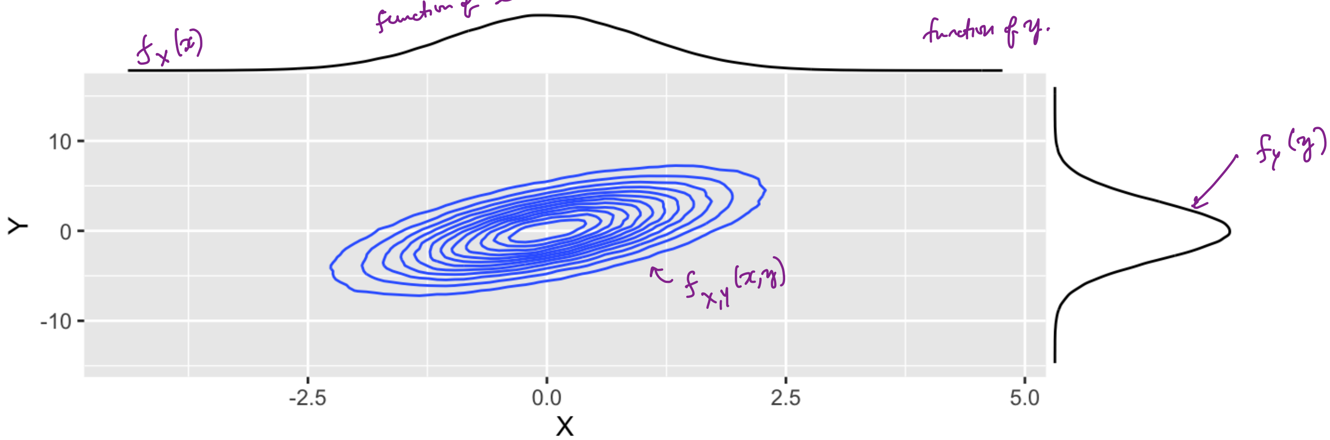
Example 1.9 X, Y w/ joint pdf $f_{X,Y}(x,y)$

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



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

$$f_X(x) = \underbrace{\int_{\infty}^{\infty} f_{X,Y}(x,y)dy}_{\text{function of } x} \qquad \text{and} \qquad f_Y(y) = \underbrace{\int_{\infty}^{\infty} f_{X,Y}(x,y)dx}_{\text{function of } y.};$$



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

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

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

$$P(0 \le X \le \tfrac{1}{4},\ 0 \le Y \le \tfrac{1}{4}) = \int_0^{\frac14}\int_0^{\frac14} \frac{6}{5}(x+y^2)\,dx\,dy$$

$$= \int_0^{\frac14} \frac{6}{5}\left[\frac{x^2}{2} + y^2 x\right]_{x=0}^{x=\frac14} dy$$

$$= \int_0^{\frac14} \frac{6}{5}\left[\frac{1}{32} + \frac{y^2}{4}\right] dy$$

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

Find the marginal densities for X and Y.

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

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

X

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

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

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

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

$$E[X] = \begin{cases} \sum\limits_{x \in \mathcal{X}} x f_X(x_i) & X \text{ is discrete} \\ \int\limits_{x \in \mathcal{X}} x f_X(x) dx & X \text{ is continuous.} \end{cases}$$

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 $X \sim \text{Bernoulli}(p)$. Find $E[X]$.

$$\Rightarrow X = \begin{cases} 1 & \text{w.p. } p \\ 0 & \text{w.p. } 1-p \end{cases} \Rightarrow f(x) = \begin{cases} p & \text{when } x = 1 \\ 1-p & \text{when } x = 0 \end{cases} \quad \text{or} \quad f(x) = p^x (1-p)^{(1-x)} \text{ for } x \in \{0,1\}.$$

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

Example 2.2 Let $X \sim \text{Exp}(\lambda)$. Find $E[X]$.

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

need integration by parts (HW 3)

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

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

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 do by hand (impossible?) $\Rightarrow$ need to compute our way out of this jam! (ch. 5).

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

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

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

$\leftarrow \mathcal{X}$

Find

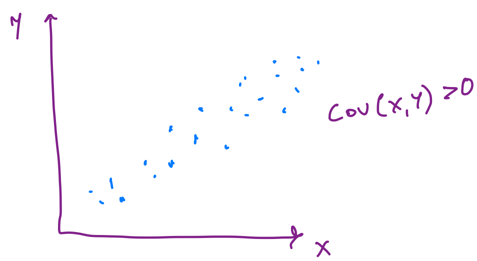
$E[X] = \sum_{\mathcal{X}} x f(x) = 4(0.5) + 6(0.3) + 8(0.2) = 5.4$

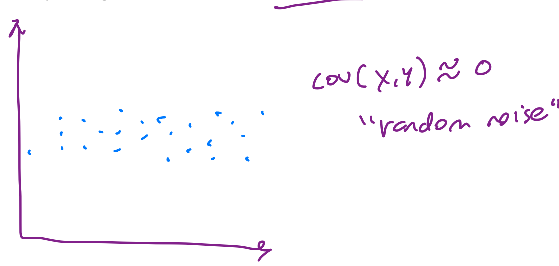
$Var[X] = E[X^2] - (E[X])^2$

$EX^2 = \sum_{\mathcal{X}} x^2 f(x) = 4^2(0.5) + 6^2(0.3) + 8^2(0.2) = 31.6$

$\Rightarrow Var(X) = 31.6 - 5.4^2 = 2.44$ 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

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

computationally convenient form.

Note: for 2 r.v. $E[g(X,Y)] = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} g(x,y) f_{X,Y}(x,y)\,dx\,dy$

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

$$\rho(X, Y) = \frac{Cov[X, Y]}{\sqrt{Var[X]Var[Y]}}. \in [-1, 1]$$

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

no linear relationship

3 Independence and Conditional Probability

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

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

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

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

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

$$P(A|B) \overset{\text{def'n}}{=} \frac{P(A \cap B)}{P(B)} = \frac{P(B|A)P(A)}{P(B)}.$$

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

$$\star \quad f_{X|Y=y}(x) \overset{\text{def'n}}{=} \frac{f_{X,Y}(x,y)}{f_Y(y)} \overset{\text{independence}}{=} \frac{f_X(x)\cancel{f_Y(y)}}{\cancel{f_Y(y)}} = f_X(x).$$

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. $E[aX + b] =$ $a\,E[X] + b$

2. $E[X + Y] =$ $E[X] + E[Y]$

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

4. $Var[b] =$ 0

5. $Var[aX + b] =$ $a^2\,Var[X]$

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

5 Random Samples

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

Example 5.1

$X_1, \ldots, X_n \overset{iid}{\sim} N(0, \sigma^2)$ "random sample" vs. $X_1 \sim N(\mu_1, \sigma^2)$, $X_2 \sim N(\mu_2, \sigma^2)$ may be independent but NOT identically distributed.

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

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

joint pdf

product of marginal pdf's. ← easier to deal with.

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

$$E[\bar{X}_n] \underset{\text{def'n}}{=} E\left[\frac{1}{n}\sum_{i=1}^{n} X_i\right] \overset{\text{property 1}}{=} \frac{1}{n} E\left[\sum_{i=1}^{n} X_i\right] \overset{\text{property 2}}{=} \frac{1}{n}\sum_{i=1}^{n} EX_i \overset{\text{identically distributed} \Rightarrow EX_1 = \ldots = EX_n}{=} \frac{1}{n}\sum_{i=1}^{n} EX_1 = \frac{n}{n} EX_1 = EX_1$$

$$Var[\bar{X}_n] \underset{\text{def'n}}{=} Var\left[\frac{1}{n}\sum_{i=1}^{n} X_i\right] \underset{\text{property 5}}{=} \frac{1}{n^2} Var\left[\sum_{i=1}^{n} X_i\right] \underset{X_i \text{ independent \& property 6}}{=} \frac{1}{n^2}\sum_{i=1}^{n} Var X_i \overset{\text{identically distributed} \Rightarrow Var X_1 = \ldots = Var X_n}{=} \frac{1}{n^2}\sum_{i=1}^{n} Var X_1 = \frac{n}{n^2} Var X_1 = \frac{Var X_1}{n}$$

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!

```
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.

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

densities
distributions (cdf)
quantiles.

may be useful for homework.