# Chapter 2: Probability for Statistical Computing

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.



https://xkcd.com/892/
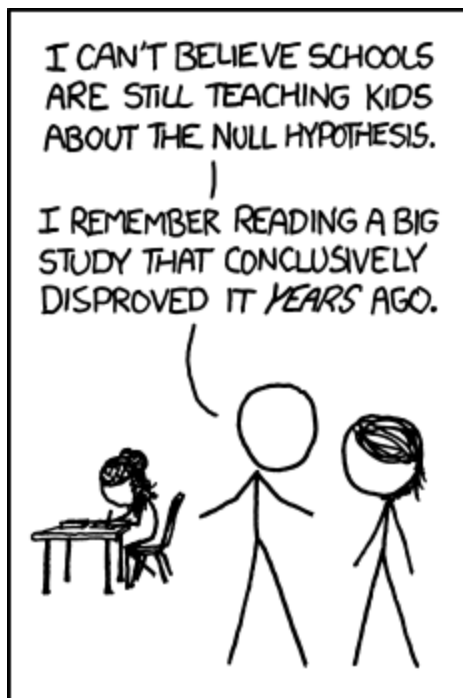
Alternative text: "Hell, my eighth grade science class managed to conclusively reject it just based on a classroom experiment. It's pretty sad to hear about million-dollar research teams who can't even manage that."

# 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 $\Omega$) to $\mathbb{R}$.

**Example 1.1**

**Example 1.2**

**Example 1.3**

Types of random variables –

**Discrete** take values in a countable set.

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

# 1.1 Distribution and Density Functions

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

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

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

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

1.

2.

3.

**Example 1.4** Let $\Omega =$ 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 $\in \{1, \ldots, 6\}$.

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

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

$X$ 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.

2.

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

**Example 1.5** Let

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

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

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

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

The cdf has the following properties

1.

2.

3.

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.

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

## 1.2 Two Continuous Random Variables

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

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

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

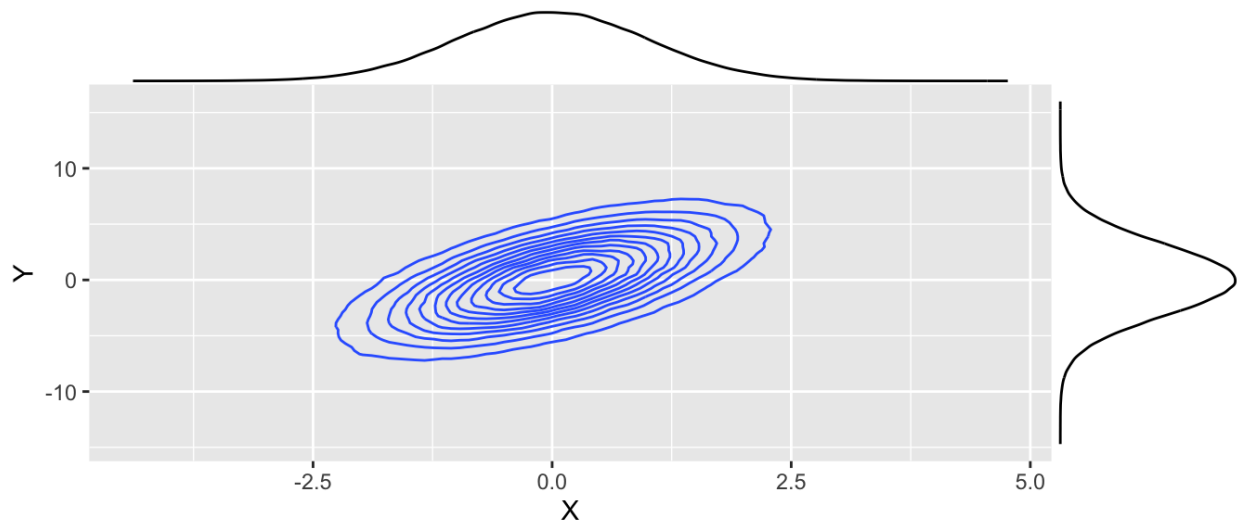Joint pdfs have the following properties

1.

2.

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

**Example 1.7**


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

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

**Example 1.8** (From Devore (2008) Example 5.3, pg. 187) A bank operates both a drive-up facility and a walk-up window. On a ramdonly 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

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

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

Find the marginal densities for $X$ and $Y$.

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

# 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

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

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 $E[X]$.

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

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

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

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

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

**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

$E[X]$

$Var[X]$

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

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

$$Cov[X, Y] = E\left[(X - E[X])(Y - E[Y])\right]$$
$$= E[XY] - E[X]E[Y]$$

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 $\$n = \{i = 1\}^\wedge n \ X\_i$.

# 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!

```
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)
```

# 7 Food for thought

Recall an indicator function is defined as

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

**Example 7.1**

**Example 7.2** 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 $c$?