Probability Review

36-600

Week 2 Tuesday – Fall 2021

Probability

What is probability?

- It is, for our purposes, the long-term frequency of the occurrence of an event.

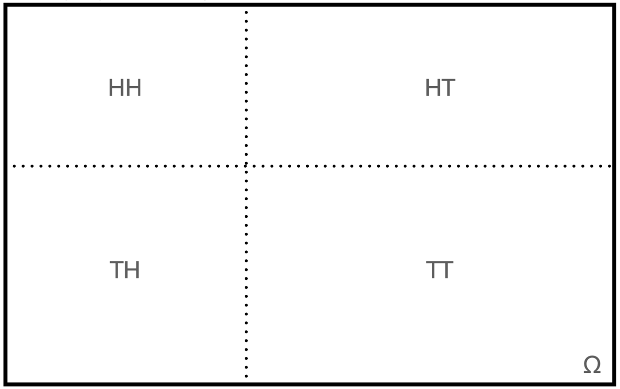
OK…what's an event?

- An experimental outcome, or a set of experimental outcomes.
- Example: observing heads then tails when flipping a coin twice…or observing at least one heads when flipping a coin twice.

How do we organize the information about experimental outcomes?

- For a given experiment, all possible separate outcomes (or *simple events*) comprise a *sample space*, conventionally denoted Ω.

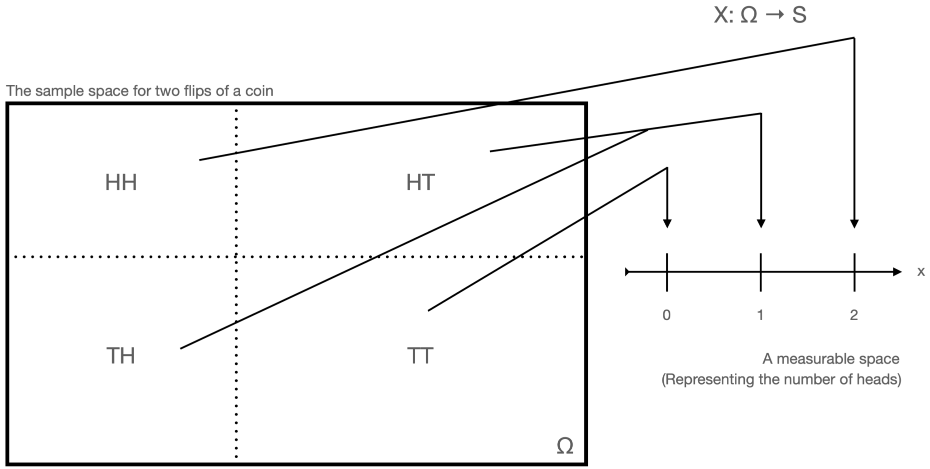
The sample space for two flips of a coin



Random Variables

Working directly with sample spaces is tedious. We can determine the probability associated with each simple event, sure, and we can write down a table of outcomes and probabilities, but is this really the best way to do things?

A *random variable* is a measurable function $X : \Omega \to S$ mapping the sample space to a measurable space. (Intuition: there is no "measurable distance" between `HH` and `TT` in Ω, but the distance is 2 in space illustrated below.)


X: Ω → S
The sample space for two flips of a coin
HH
HT
TH
TT
Ω
x
0
1
2
A measurable space
(Representing the number of heads)

A random variable may be

- continuously valued (perhaps representing time, or distance, etc.); or
- discretely valued (perhaps representing counts, or choices from a finite menu, etc.)

Probability Distributions

Neither a sample space, nor a random variable defined given that sample space, have any attached notion of probability. For instance, in the pictures above, we do not know the probability of, e.g., the event `HH`.

So we have to add an extra layer of information.

A *probability distribution* is a mapping $P : \Omega \to \mathbb{R}$ that describes how probabilities are distributed across the values of a random variable.

For instance, we can specify that the probability of seeing x heads in n coin flips is given by

$$p(x) = \binom{n}{x} p^x (1-p)^{n-x} \quad x \in [0, n] \,,$$

where

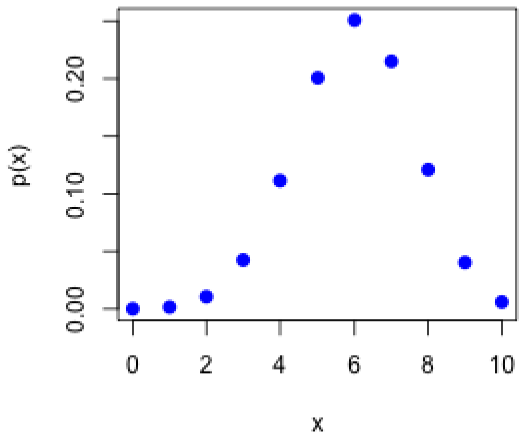
$$\binom{n}{x} = \frac{n!}{x!(n-x)!}$$

and, e.g., $x! = x \cdot (x-1) \cdot (x-2) \cdots 2 \cdot 1$. (This is called the factorial function.)

This is a particular probability distribution, the *binomial distribution*, and $p(x)$ represents its *probability mass function* or *pmf*. The binomial distribution is a discrete distribution (because the results are discretely valued), and like all discrete distributions, its masses sum to one. This is contrasted against continuous distributions (like the normal distribution), whose *probability density functions* (or *pdfs*) $f(x)$ integrate to one over the real-number line.

Probability Distributions

Here is an example of a binomial distribution:



This distribution encapsulates the probabilities associated with an experiment where you try something 10 times, with the probability of succeeding on any given try being 0.6. The value x represents the observed number of successes from any given experiment, and the value $p(x)$ represents the probablity that you would observe that value. For instance, the probability of seeing 6 successes is 0.251, i.e., if you kept rerunning the experiment, you'd see 6 successes roughly a quarter of the time.

Probability Distributions

In terms of the math, is there anything really special about probability distributions? Are they a class onto themselves?

No. *They are just functions*, albeit with some constraints.

1. They are non-negative. (For instance, $p(x)$ cannot be less than zero…as negative probabilities are non-sensical.)

2. As stated earlier, if discrete, the probability mass function (or pmf) sums to one,

   $$\sum_x p(x) = 1\,,$$

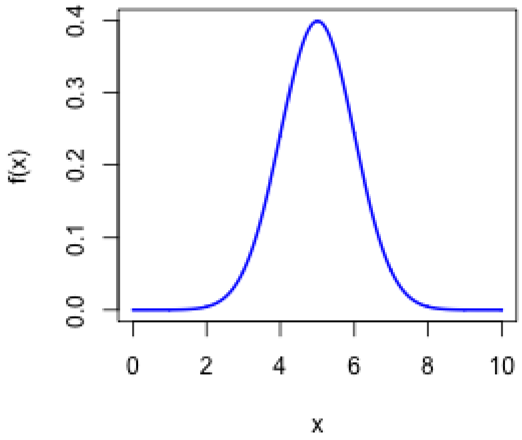
   whereas if continuous, the probability density function (or pdf) integrates to one:

   $$\int_{-\infty}^{\infty} f(x)dx = 1\,.$$

   In the end, the probability is 1 that the result of your experiment is one of the simple events contained in your sample space. That's all that these equations are saying.

Probability Distributions: Properties

A probability distribution governs what the next outcome of your experiment will be. If you have a (continuous) distribution govering your experiment that looks like this



then it stands to reason that the next datum you sample will probably be closer to 5 than to either 3 or 7. The height of the curve gives you that intuition. So: what is the average value of a sampled datum?

That average value is given by the *mean*, the *expected value*, or the *expectation*, $E[X]$ (sometimes rendered as μ):

$$E[X] = \mu = \sum_x x p(x) \;\; \text{or} \;\; E[X] = \int_{-\infty}^{\infty} x f(x) dx \, .$$

Probability Distributions: Properties

Another question that one may ask is "what is the 'width' of the distribution?"

That width is given by the *standard deviation*, σ_X:

$$\sigma = \sqrt{E[X^2] - (E[X])^2}\,.$$

The square of the standard deviation is dubbed the *variance*, $V[X]$.

Lastly, two common quantities that we are interested in, if we have two random variables X and Y, are the *covariance* and *correlation* between them:

$$\mathrm{Cov}(X,Y) = E[XY] - E[X]E[Y]$$

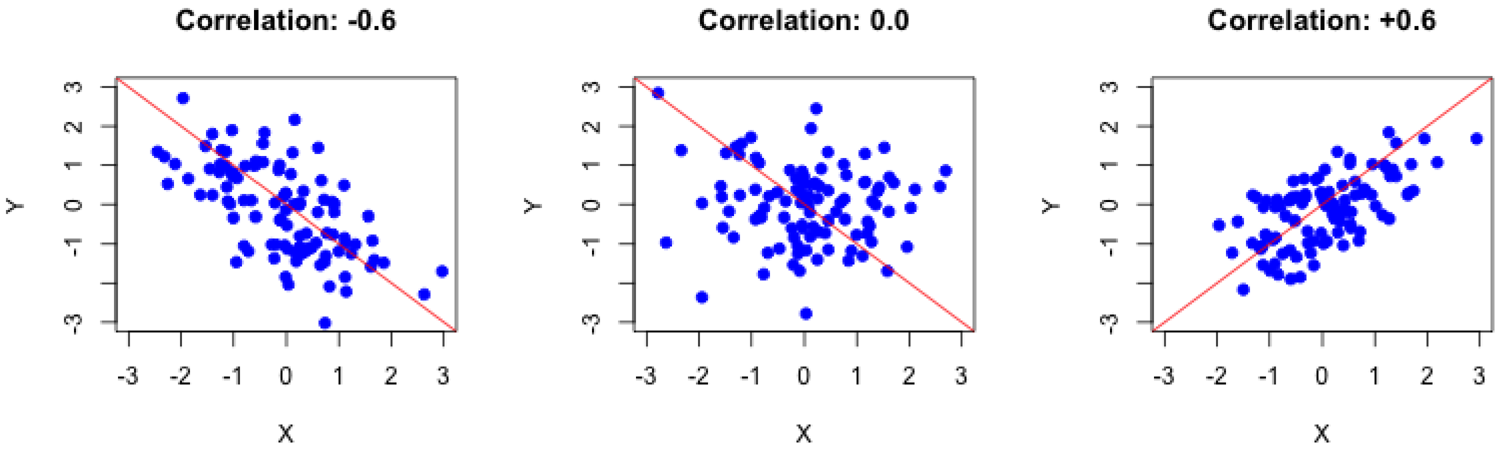
and

$$\rho_{XY} = \mathrm{Corr}(X,Y) = \frac{\mathrm{Cov}(X,Y)}{\sigma_X \sigma_Y}\,.$$

The correlation is bounded such that $-1 \leq \rho_{XY} \leq 1$ and it is a metric of the *linear* association between two random variables. If the correlation is negative, it means that as X increases in value, Y generally *decreases* in value, whereas if the correlation is positive, it means that as X increases, Y generally increases as well.

If there is no functional associated *at all* between two random variables, they are both *independent* and *uncorrelated*, but if all we know is that they are uncorrelated, they may still be dependent.

Probability Distributions: Correlation Examples



We will see correlation plots again later: they are an integral part of exploratory data analysis. (Are the data in column 1 correlated with the response? With the data in column 6? Etc.)

Probability Distributions: Motivation

You may be asking right now: "why exactly are we learning about probability distributions?"

Because they underlie

- the sampling of random variables, i.e., the data you wish to analyze;
- hypothesis tests; and
- (some) statistical learning models, like multiple linear regression.

For instance, one of the assumptions of linear regression is that, given a set of predictor variable values $\mathbf{x}$, the response variable value y is a random variable that is sampled from a normal distribution with mean $E[Y|\mathbf{x}]$ and standard deviation σ. It helps to have at least a passing idea of what these terms mean!

Also, methods of exploratory data analysis, while not *directly* indicating what probability distributions data are sampled from, at least give us ideas of what those distributions are or could be. (A histogram, for instance, is really just a simplistic nonparametric estimator of $p(x)$ or $f(x)$. It looks like a bell curve? The data are plausibly normally distributed…but could be distributed in another manner. Etc.)

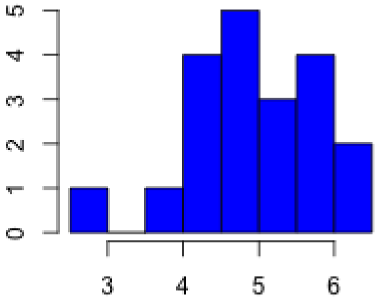
This is a natural segue to…

But...What Happens in the Wild?

When analyzing data, we do not "see" probability distributions, but the data sampled from them. For instance, we might run an experiment and observe the values

```
##  [1] 4.673964 5.552462 4.325056 5.214359 5.310769 6.173966 5.618790 4.887266 5.917028 4.776741 5.526448 4.205156 6
## [14] 3.533180 4.763317 4.806662 4.150245 5.058465 4.182330 2.949692
```

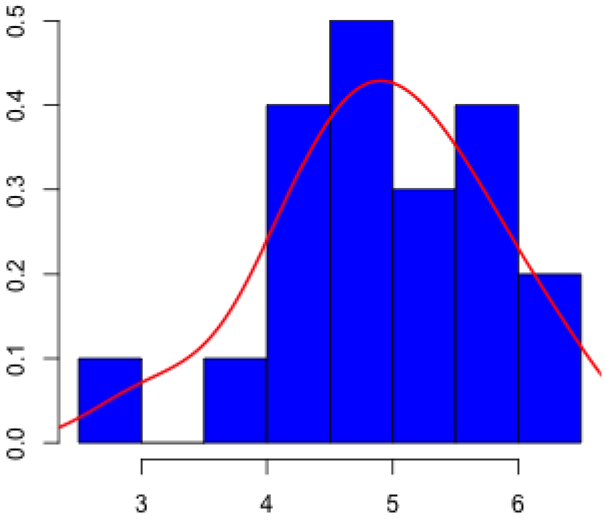
which we can visualize in, e.g., a histogram:



How might we summarize these data? With the sample analogues of the properties mentioned above, e.g., the *sample mean* $\bar{X}$ and the *sample standard deviation* s. (With pairs of data, we can also compute the *sample correlation* r_{XY}.) We use these sample quantities to make *inferences* about the distributions from which the data are sampled, or their properties.

What if I Have No Idea What the Underlying Distribution Might Be for My Data?

Well, there's always *density estimation*. Let's punt on this…but we might cover this later in the course.



(Example of an "off-the-shelf" density estimate done in R.)

Commonly Assumed Probability Distributions

Name	Type	Parameters	Domain (i.e., allowed x values)
Binomial	discrete	trials n , probability of success p	integers 0 to n
Poisson	discrete	expected number of counts λ	all integers ≥ 0
Normal	continuous	mean μ , standard deviation σ	$(-\infty, \infty)$
Standard Normal	continuous	none ($\mu = 0$ and $\sigma = 1$)	$(-\infty, \infty)$
t	continuous	number of degrees of freedom ν	$(-\infty, \infty)$
Exponential	continuous	event rate β	$(0, \infty)$
Chi-Square	continuous	number of degrees of freedom ν	$(0, \infty)$

Parameters: fixed constants that affect the location of the distribution on the real-number line, and its "shape."

Other distributions you might hear of but are less commonly used in basic data analysis situations are the geometric, negative binomial, and hypergeometric (discrete), and the uniform, gamma, and beta (continuous). Note that both the exponential and chi-square distributions are actually "sub-distributions" of the gamma, i.e., gamma distributions with constrained parameter choices.

In today's lab, you will play with some of these distributions, using `R`.