Math Review
Summer 2016

Topic 7

7. Probability

We go through the basics of probability building from the foundations and notations we need to work with. This topic is very definition and information heavy – I have tried to add as many examples as I could, but still the lecture outweighs the exercises. It might be the nature of the coverage of this topic – we'll try to keep things interesting and cruise through the material. The main goal is to nail down the basic notions of probability and build tools to use in statistical inferences and modelling (eventually).

7.1 Random variables and distributions

An experiment is any procedure that can be infinitely repeated and has a well-defined set of outcomes.

Formally,

Definition. An *experiment* is a one-off or repeatable process or procedure for which

(a) there is a **well-defined** set of possible outcomes
(b) the actual outcome is **not known with certainty**.

We are not always interested in an experiment itself, but rather in some consequence of its random outcome. Such consequences, when real valued, may be thought of as random variables. A random variable is a variable, say X, that takes on numerical values and has an outcome that is determined by an experiment. We generally denote the set of all possible outcomes for a particular random variable as X (or some other upper case letter) and a particular outcome is denoted as x.

7.1.1. Discrete random variable

A **discrete random variable** is a random variable that takes on only a finite number of values. The simplest type of discrete random variable is called a Bernoulli random variable (or "binary") where $X = \{0, 1\}$. Typically an outcome of "1" denotes a success and an outcome of "0" denotes a failure.

The **sample space** of the experiment is the set of all possible outcomes, X. An **event** is a set of outcomes in the sample space, or a subset of X. For example, we could have the sample space $X = \{x_1, x_2, x_3, x_4\}$ and an event be $E = \{x_1, x_2\}$.

Example. If I plant ten bean seeds and count the number that germinate, the sample space is $S = \{0,1,2,3,4,5,6,7,8,9,10\}$

$\mathcal{Q}$: If you toss a 'fair' coin three times and record the result, what is the sample space?

$\mathcal{A}$: $S = \{HHH, HHT, HTH, HTT, THH, THT, TTH, TTT\}$

HTH means 'heads on the first toss, then tails, then heads again'.

Definition. An event, E, is a designated collection of sample outcomes. An event is, therefore, a subset of the sample space.

Example. As we saw, by definition, an event is a subset of the sample space. We can specify an event by listing all the outcomes that make it up. In the coin toss example, let A be the event 'more heads than tails.'

We can write out $A = \{HHH, HHT, HTH, THH\}$

$\mathcal{Q}$: Imagine you toss a 'fair' coin three times and record the result. What is the event 'heads on last throw'? Denote the event as B and write it out.

$\mathcal{A}$: B = {HHH,HTH,THH,TTH}

The events $E_1, E_2, \ldots, E_m$ are **mutually exclusive** if for each pair of events E_i and E_j in the collection we have $E_i \cap E_j = \emptyset$;, i.e. the two events have no outcomes in common. The complement of E, E^c, denotes all possible outcomes not in the event E.

7.1.2. What is probability?

No, we will not get into a philosophical discussion! But broadly speaking, it is hard to concretely define what a probability is. Some view it as a 'limiting frequency.' For example, if I toss a coin 1000 times, I am likely to get tails about ½ of the time. Other may view it as a more subjective matter. For our purposes, we will work with a mathematical construct centered around an experiment which satisfies some key axioms.

The probability of E, $P(E)$, is the probability that event E will occur and satisfies:

(1) For any event E, $0 \leq P(E) \leq 1$

(2) $P(X) = 1$,

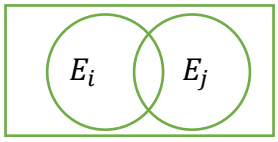
(3) If $E_1, E_2, \ldots, E_m$ are mutually exclusive events (or *pairwise disjoint*), then:

$$P(E_1 \cup E_2 \cup \ldots E_m) = P(E_1) + P(E_2) + \cdots + P(E_m) = \sum_{i=1}^{m} P(E_i)$$

Note that this is a union of events, not the union of probabilities ~~$(P(E_1) \cup P(E_2) \ldots)$~~

Now, if E_i and E_j are not mutually exclusive, then:

$P(E_i \cup E_j) = P(E_i) + P(E_j) - P(E_i \cap E_j)$. This must be a familiar result to many. This is really getting at the inclusion-exclusion principle:



(4) $P(E^c) = 1 - P(E)$

$\mathcal{Q}$: What is $E \cap E^c$?

$\mathcal{A}$: $E \cap E^c = \emptyset$

$\mathcal{Q}$: What is $P(E) + P(E^c) =$?

$\mathcal{A}$: $P(E) + P(E^c) = \quad P(E) + P(E^c) = P(X) = 1$

A discrete random variable has any number of $k < \infty$ possible outcomes. The possible outcomes are given by $X = \{x_1, x_2, \ldots, x_k\}$ and p_i is used to denote the probability that outcome x_i occurs:

$$p_i = P(X = x_i) \text{ for } i = 1,2, \ldots, k$$

where p_i must satisfy the following two conditions:

(1) $p_i \in [0,1]$, for all $i = 1,2,3 \ldots, k$ and
(2) $p_1 + p_2 + \cdots + p_k = \sum_{i=1}^{k} p_i = 1$

Notation use: Often we use capital letters at the end of the alphabet, X, Y, Z, to define the random variable. The corresponding lower case letters, x, y, z, to represent the actual

values. Thus, the expression $P(X = x)$ symbolizes the probability that the random variable X takes on the particular value x. Often, this is written simply as P(x). Likewise, $P(X \leq x)$ is the probability that the random variable X is less than or equal to the specific value x; $P(a \leq X \leq b)$ is the probability that X lies between values a and b.

In context of a binary variable, these conditions imply that we only need to state one of the two probabilities, i.e. $P(X = 1) = \alpha$ implies that $P(X = 0) = 1 - \alpha$.

We can use all possible probabilities to create the **probability density function (pdf).** The pdf is just a function where the input is any possible outcome, x_i, of the random variable X and the outcome is the probability of that outcome, p_i. The function is denoted as:

$$pdf = f(x_i) = p_i \text{ for } i = 1,2,\dots,k$$

$f(x) = 0$ if $x \neq x_i$ for some $i = 1,2,\dots,k$

Example. The probability density function for a standard die is given by:

$$f(x) = \begin{cases} \frac{1}{6} \text{ if } x = 1,2,3,4,5,6 \\ 0 \; otherwise \end{cases}$$

Let's try a more involved example as a group.

Group Example. Consider the simple experiment of tossing a coin three times. Let $\boldsymbol{X}$ = **number of times the coin comes up heads**. Write out the pdf for this experiment.

The 8 possible elementary events, and the corresponding values for X, are:

Event	Value of X
TTT	0
TTH	1
THT	1
HTT	1
THH	2
HTH	2
HHT	2
HHH	3

Therefore, the probability distribution for the number of heads occurring in three coin tosses is:

x	$f(x)$
0	1/8
1	3/8
2	3/8
3	1/8

Finally, we can write this out as:

$$f(x) = \begin{cases} \frac{1}{8} \text{ if } x = 0 \\ \frac{3}{8} \; if \; x = 1,2 \\ \frac{1}{8} if \; x = 3 \\ 0 \text{ otherwise} \end{cases}$$

The **cumulative distribution function** is then just the probability that the outcome is below some number:

$$cdf = F(x) = P(X \leq x)$$

For a discrete random variable this function reduces to:

$$F(x_i) = P(X \leq x_i) = f(x_1) + f(x_2) + \cdots + f(x_i)$$

Recall that outcomes of an experiment are individual outcomes that are **mutually exclusive.**

Example. Consider a baseball player who has a 68% chance of striking out, a 15% chance of getting to first base, a 10% chance of getting to second base, a 5% chance of getting to third base, and a 2% chance of hitting a home run. Then, the baseball players pdf and cdf are given by:

$$f(x) = \begin{cases} 0.68 & if\ x = 0 \\ 0.15 & if\ x = 1 \\ 0.10 & if\ x = 2 \\ 0.05 & if\ x = 3 \\ 0.02 & ifx = 4 \\ 0 & otherwise \end{cases} \qquad F(x) = \begin{cases} 0.68 & if\ x = 0 \\ 0.83 & if\ x = 1 \\ 0.93 & if\ x = 2 \\ 0.98 & if\ x = 3 \\ 1.00 & ifx = 4 \end{cases}$$

7.1.2. Continuous random variable

A *continuous random variable* is a random variable that can take on every point in some interval of the real numbers. A continuous random variable differs from a discrete random variable in that it takes on an uncountably infinite number of possible outcomes.

For example, if we let X denote the height (in meters) of a randomly selected maple tree, then X is a continuous random variable. In other words, the probability of any given number occurring is zero. This zero probability is because the continuous random variable can take on so many possible outcomes that we cannot count them (infinitely many). There is, however, a nonzero probability that the outcome falls in a certain range or interval. To compute this probability we generally use the pdf.

Note that each individual point and image on a pdf does not give us the probability of the event occurring as it does in the discrete case, i.e.:

$$f(x_i) \neq p_i$$

Instead, in the continuous case, we have:

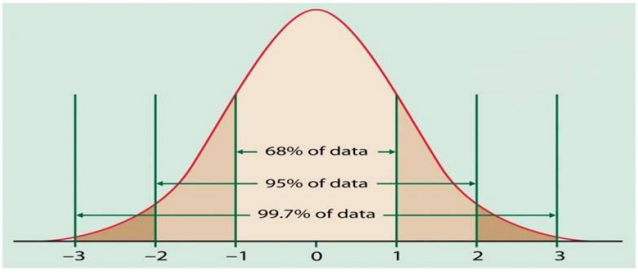
$$P(a \leq X \leq b) = \int_a^b f(x)dx$$

The cdf is then given by:

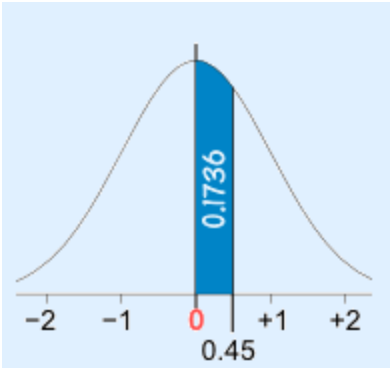
$$F(x) = P(X \leq x) = \int_0^x f(x)dx$$

Notice the differences between these two expressions.

One of the important continuous distribution is the Standard Normal Distribution. The Random Variable if often denoted by Z. The graph for Z is a symmetrical bell-shaped curve:



Example. Find $P(0 < z < 0.45)$. You can use the standard normal distribution table (for now you can believe me), that this is equal to:



$$P(0 < z < 0.45) = 0.6736 - 0.5 = 0.1736$$

Q: What is the $P(Z < -0.9)$? How about $P(-1.40 < z < -1.20)$?

$\mathcal{A}$:

$$P(Z < -0.9) = 1 - 0.8159$$
$$P(-1.40 < z < -1.20) = 0.91924 - 0.8849 = 0.03434$$

z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.5000	0.5040	0.5080	0.5120	0.5160	0.5199	0.5239	0.5279	0.5319	0.5359
0.1	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596	0.5636	0.5675	0.5714	0.5753
0.2	0.5793	0.5832	0.5871	0.5910	0.5948	0.5987	0.6026	0.6064	0.6103	0.6141
0.3	0.6179	0.6217	0.6255	0.6293	0.6331	0.6368	0.6406	0.6443	0.6480	0.6517
0.4	0.6554	0.6591	0.6628	0.6664	0.6700	0.6736	0.6772	0.6808	0.6844	0.6879
0.5	0.6915	0.6950	0.6985	0.7019	0.7054	0.7088	0.7123	0.7157	0.7190	0.7224
0.6	0.7257	0.7291	0.7324	0.7357	0.7389	0.7422	0.7454	0.7486	0.7517	0.7549
0.7	0.7580	0.7611	0.7642	0.7673	0.7703	0.7734	0.7764	0.7794	0.7823	0.7852
0.8	0.7881	0.7910	0.7939	0.7967	0.7995	0.8023	0.8051	0.8078	0.8106	0.8133
0.9	0.8159	0.8186	0.8212	0.8238	0.8264	0.8289	0.8315	0.8340	0.8365	0.8389
1.0	0.8413	0.8438	0.8461	0.8485	0.8508	0.8531	0.8554	0.8577	0.8599	0.8621
1.1	0.8643	0.8665	0.8686	0.8708	0.8729	0.8749	0.8770	0.8790	0.8810	0.8830
1.2	0.8849	0.8869	0.8888	0.8907	0.8925	0.8944	0.8962	0.8980	0.8997	0.90147
1.3	0.90320	0.90490	0.90658	0.90824	0.90988	0.91149	0.91309	0.91466	0.91621	0.91774
1.4	0.91924	0.92073	0.92220	0.92364	0.92507	0.92647	0.92785	0.92922	0.93056	0.93189
1.5	0.93319	0.93448	0.93574	0.93699	0.93822	0.93943	0.94062	0.94179	0.94295	0.94408
1.6	0.94520	0.94630	0.94738	0.94845	0.94950	0.95053	0.95154	0.95254	0.95352	0.95449
1.7	0.95543	0.95637	0.95728	0.95818	0.95907	0.95994	0.96080	0.96164	0.96246	0.96327
1.8	0.96407	0.96485	0.96562	0.96638	0.96712	0.96784	0.96856	0.96926	0.96995	0.97062
1.9	0.97128	0.97193	0.97257	0.97320	0.97381	0.97441	0.97500	0.97558	0.97615	0.97670
2.0	0.97725	0.97778	0.97831	0.97882	0.97932	0.97982	0.98030	0.98077	0.98124	0.98169
2.1	0.98214	0.98257	0.98300	0.98341	0.98382	0.98422	0.98461	0.98500	0.98537	0.98574
2.2	0.98610	0.98645	0.98679	0.98713	0.98745	0.98778	0.98809	0.98840	0.98870	0.98899

Properties of the CDF for a continuous random variable

(i) $F(x) \in [0,1] for\ all\ x \in \mathbb{R}$.

(ii) $If\ x_1 < x_2\ then\ P(X \leq x_1) = F(x_1) \leq F(x_2) = P(X \leq x2)$.

(iii) For any number $c, P(X > c) = 1 - F(c)$.

(iv) $P(X \geq c) = P(X > c)$, and

(v) $P(a < X < b) = P(a \leq X \leq b) = P(a \leq X < b) = P(a < X \leq b)$.

7.1.3. Multiple random variables - Joint distributions

We have thus far considered discrete and continuous distributions but we only focused on single random variables. Probability distributions can, however, be applied to grouped random variables which gives rise to joint probability distributions. We'll mostly focus on 2-dimensional distributions (i.e. only two random variables) but higher dimensions (more than two variables) are also possible.

Since all random variables are divided into discrete and continuous random variables, we have end up having both discrete and continuous joint probability distributions. These distributions are not so different from the one variable distributions we just looked at, but understanding some concepts might require one to have knowledge of multivariable calculus at the back of their mind.

Essentially, joint probability distributions describe situations where by both outcomes represented by random variables occur. While we only X to represent the random variable, we now have X and Y as the pair of random variables.

The discrete random variables X and Y have a *joint distribution* which is described by the joint **probability density function**:

$$f_{X,Y}(x,y) = P(X = x, Y = y) = P(X = x \text{ and } Y = y)$$

Note that, this joint distribution can be reduced to a very workable expression in the case of independence (for discrete random variables). What we mean here is that random variables X and Y are independent if, and only if:

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

The **Cumulative Distribution Function** (CDF) for a joint probability distribution is given by:

$$F(x,y) = P(X \leq x, Y \leq y)$$

Definition. The discrete random variables $X_1, X_2, \ldots, X_n$ are independent random variables if, and only if:

$$f(x_1, x_2, \ldots, x_n) = f_1(x_2) \cdot f_2(x_2) \cdot \cdots \cdot f_n(x_n).$$

Example. The table below represents the joint probability distribution obtained for the outcomes when a die is flipped and a coin is tossed.

$f(x,y)$	1	2	3	4	5	6	Row Totals
Heads	a	b	c	d	e	f	α
Tails	g	h	i	j	k	l	β
Column Totals	γ	δ	ε	ζ	θ	ψ	ω

In the table above, $x = 1, 2, 3, 4, 5, 6$ as outcomes when the die is tossed while $y = Heads, Tails$ are outcomes when the coin is flipped. The letters a through l represent the joint probabilities of the different events formed from the combinations of x and y while the Greek letters represent the totals and ω should equal to 1. The row sums and column sums are referred to as the marginal probability distribution functions (PDF).

$\mathcal{Q}$: In the above table, assume that flipping a coin and tossing a die are independent. What are the letters a through l equal to? How about α and δ for instance?

$\mathcal{A}$:

$f(x,y)$	1	2	3	4	5	6	Row Totals
Heads	1/12	1/12	1/12	1/12	1/12	1/12	1/2
Tails	1/12	1/12	1/12	1/12	1/12	1/12	1/12
Column Totals	1/6	1/6	1/6	1/6	1/6	1/6	1

We will not spend too much time on continuous joint distributions, but as one can expect, the results can be summarized as:

For a given interval A, we have:

$$P[(X,Y) \in A] = \iint_A f(x,y)dx\,dy$$

This tells us that in order to determine the probability of an event *A*, we must integrate the function $f(x,\ y)$ over the space defined by the event *A*.

Example and exercise:

A certain farm produces two kinds of eggs on any given day; organic and non-organic. Let these two kinds of eggs be represented by the random variables X and Y respectively. Given that the joint probability density function of these variables is given by

$$f(x,y) = \begin{cases} \frac{2}{3}(x+2y), 0 \le x \le 1, 0 \le y \le 1, \\ 0\ elsewhere \end{cases}$$

Find the $P(X\ \le \frac{1}{2}, Y\ \le \frac{1}{2})$

$$P\left(X\ \le \frac{1}{2}, Y\ \le \frac{1}{2}\right) = \int_0^{0.5}\int_0^{0.5} \frac{2}{3}(x+2y)dxdy$$

$$= \frac{2}{3}\int_0^{0.5} \left[\frac{x^2}{2} + 2xy\right]_0^{0.5} dy$$
$$= \frac{2}{3}\int_0^{0.5} \frac{1}{8} + y\, dy$$
$$= \frac{2}{3}\left[\frac{y}{8} + \frac{y^2}{2}\right]_0^{0.5}$$
$$= \frac{1}{8}$$

7.1.4. Conditional distributions

The *conditional distribution* of the random variable Y given the random variable X. We can use this information to find out how X affects Y. The conditional probability density function is defined as:

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

For continuous variables we compute $f_{Y|X}$ by computing the area underneath the conditional probability density function. If X and Y are independent random variables then we have:

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

$\mathcal{Q}$: Say you have two binary random variables given by X and Y with the following conditional probabilities. Are X and Y independent?

$$f_{Y|X}(1|1) = 0.85$$
$$f_{Y|X}(1|0) = 0.70$$
$$f_{Y|X}(0|1) = 0.15$$
$$f_{Y|X}(0|0) = 0.30$$

$\mathcal{A}$:
No, the variables are not independent because the value of X affects the probability of Y. If X and Y were independent we would have:

$f_{Y|X}(1|1) = f_{Y|X}(1|0)$ and $f_{Y|X}(0|1) = f_{Y|X}(0|0)$.

7.2 Moments and properties

The **expected value** of the random variable X is a weighted average of all possible values of X where the weights are determined by the probability density function. In the case of a discrete variable we have:

$$E[X] = \mu_x = x_1 f(x_1) + x_2 f(x_2) + \cdots + x_n f(x_n) = \sum_{j=1}^{n} x_j f(x_j)$$

If we have a continuous variable, the expected value is:

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

$\mathcal{Q}$: A fair six-sided die is tossed. You win \$2 if the result is a "1," you win \$1 if the result is a "6," but otherwise you lose \$1.
Find $E(X)$.

Hint: Let X = Amount Won or Lost.

$\mathcal{A}$:

X	+\$2	+\$1	-\$1
Probability	1/6	1/6	4/6

$$E(X) = \$2\left(\frac{1}{6}\right) + \$1\left(\frac{1}{6}\right) + (-\$1)\left(\frac{4}{6}\right) = -\$0.17$$

Here are some *key properties of the expected value*:

(i) If X is a random variable then the function g(X) is also a random variable. We have:

$$E[g(X)] = \sum_{j=1}^{n} g(x_j) f_X(x_j) \text{ if X is discrete}$$

$$E[g(X)] = \int_{-\infty}^{\infty} g(x) f_X(x) \text{ if X is continuous}$$

(ii) For any constant $c, E[c] = c$

(iii) For any constants a and $b, E[aX + b] = aE[X] + b.$

(iv) If $\{a_1, a_2, \dots, a_n\}$ are constants and $\{X_1, X_2, \dots, X_n)$ are random variables, then:

$$E[a_1X_1 + a_2X_2 + \cdots + a_nX_n] = a_1E[X_1] + a_2E[X_2] + \cdots + a_nE[X_n]$$

Or equivalently:

$$E\left[\sum_{j=1}^{n} a_iX_i\right] = \sum_{j=1}^{n} a_iE[X_i]$$

Definition. Let X be a random variable and let $k \geq 1$. If $E[X^k] < \infty$, then $E[X^k]$ is called the $k - th$ moment of X.

Example. For the random variable X the first moment is given by:

$$E[X^1] = E[X] = \mu_x$$

The second moment is:

$E[X^2] = \sum_{j=1}^{n} x_j^2 f(x_j)$ if X is discrete

&

$E[X^2] = \int_{j=1}^{n} x^2 f(x)dx$ if X is continuous

Some may be familiar with the formula for the variance.

The variance of the random variable X is a measure of the expected distance of X from its mean, given by: $var[X] = \sigma^2 \equiv E[(X - \mu)^2]$. By definition, the variance becomes a random variable where $g(X) = (X - \mu)$. Thus, the variance can be described as the second moment of $(X - E[X])$.

Properties of variance
(i) $var[X] = 0$ if, and only if, there exists a constant c, such that $P[X = c] = 1$, in which case $E[X] = c$,
(ii) For any constants a and b, $var[aX + b] = a^2var[X]$,
(iii) $var[aX + bY] = a^2var[X] + b^2 var[Y] + 2abcov[X, Y]$

The standard deviation of a random variable, denoted $sd[X]$, is the positive square root of the variance:

$$sd[X] = \sigma = +\sqrt{var[X]}$$

Properties of standard deviation

(i) For any constant $c, sd[c] = 0$
(ii) For any constants a and b, $sd[aX + b] = |a| \cdot sd[X]$

Here are couple of definitions that one can keep handy which we will not further expand on. In many statistical analyses the task is to really characterize the location and variability of a data set. A further characterization of the data includes skewness and kurtosis.

Skewness is a measure of symmetry, or more precisely, the lack of symmetry. A distribution, or data set, is symmetric if it looks the same to the left and right of the center point.

Kurtosis is a measure of whether the data are heavy-tailed or light-tailed relative to a normal distribution. That is, data sets with high kurtosis tend to have heavy tails, or outliers. Data sets with low kurtosis tend to have light tails, or lack of outliers.

Accordingly, the third and fourth moments of the the random variable $(X - \mu)$ can also be used to describe the distribution. We have:

Skewness: $E[(X-\mu)^3]$, with skewness coefficient = $\frac{\mu^3}{\sigma^3}$
Kurtosis: $E[(X-\mu)^4]$, with the degree of excess = $\frac{\mu^4}{\sigma^4} - 3$

These coefficients are the respective descriptive statistics for these moments.

Example. If you have the following data:

x	3	4	5
$f(x)$	0.3	0.4	0.3

We want to find the variance and standard deviation of X.

$$\sigma_x^2 = E[(X-\mu)^2] = (3-4)^2 0.3 + (4-4)^2 0.4 + (5-4)^2 0.3 = 0.6$$

$$\sigma_x = \sqrt{0.6} = 0.77$$

x	1	2	6	8
$f(x)$	0.4	0.1	0.3	0.2

$\mathcal{Q}$: What is the variance and standard deviation of X?
$\mathcal{A}$:

$$\sigma_y^2 = E[(Y-\mu)^2] = (1-4)^2(0.4) + (2-4)^2(0.1) + (6-4)^2(0.3) + (8-4)^2(0.2)$$
$$= 8.4$$

$$\sigma_y = \sqrt{8.4} = 2.9$$

7.3 Covariance and correlation

The covariance between two random variables X and Y is defined as:

$$cov[X,Y] = \sigma_{XY} = E[(X-\mu_x)(Y-\mu_y)]$$

The covariance measures the amount of linear dependence between X and Y. If $cov[X,Y\,] > 0$ then X and Y move in the same direction and if $cov[X,Y\,] < 0$ then X and Y move in opposite directions.

Properties of covariance

(i) If X and Y are independent, then $cov[X,Y\,] = 0$

(ii) For any constant a_1, b_1, a_2 and b_2:

$$cov[a_1X + b_1, a_2Y + b_2] = a_1a_2cov[X,Y]$$

(iii) $|cov[X,Y]| \leq sd[X].sd[Y]$

Another statistic you may be interested in is the correlation coefficient. Both covariance and correlation indicate whether variables are positively or inversely related. Correlation also tells you the degree to which the variables tend to move together.

$$corr[X,Y] = \frac{cov[X,Y]}{sd[X].sd[Y]} = \frac{\sigma_{XY}}{\sigma_X\sigma_Y}$$

Properties of correlation

(i)

$$-1 \leq corr[X,Y] \leq 1$$

$corr[X,Y\,] = 0$ can happen if X and Y are independent. We can say that X and Y are uncorrelated random variables.

(ii) For any constant a_1, b_1, a_2 and b_2, with $a_1.a_2 > 0$

$$corr[a_1X + b_1, a_2Y + b_2] = corr[X,Y]$$

with $a_1.a_2 < 0$

$$corr[a_1X + b_1, a_2Y + b_2] = -corr[X,Y]$$

Example. Suppose X and Y have the following joint probability function, let' s find the covariance and correlation between X and Y.

		y			
	$f(x,y)$	1	2	3	$f_X(x)$
x	1	0.25	0.25	0	0.5
	2	0	0.25	0.25	0.5
$f_Y(y)$		0.25	0.5	0.25	1

We can can easily calculate the mean.

$$\mu_x = x_1 f(x_1) + x_2 f(x_2) = (1)\frac{1}{2} + (2)\frac{1}{2} = \frac{1}{2} + 1 = \frac{3}{2}$$

$$\mu_y = y_1 f(y_1) + y_2 f(y_2) + y_3 f(y_3) = (1)\frac{1}{4} + (2)\frac{1}{2} + (3)\frac{1}{4} = \frac{1}{4} + 1 + \frac{3}{4} = 2$$

$$cov[X,Y] = \sigma_{XY} = E[(X - \mu_x)(Y - \mu_y)]$$

Covariance = $\left(1 - \frac{3}{2}\right)(1-2)\left(\frac{1}{4}\right) + \left(1 - \frac{3}{2}\right)(2-2)\left(\frac{1}{4}\right) + \left(1 - \frac{3}{2}\right)(3-2)(0) + \left(2 - \frac{3}{2}\right)(1-2)(0) + \left(2 - \frac{3}{2}\right)(2-2)\left(\frac{1}{4}\right) + \left(2 - \frac{3}{2}\right)(3-2)\left(\frac{1}{4}\right) = \frac{1}{8} + 0 + 0 + 0 + 0 + \frac{1}{8} = \frac{1}{4}$

$\mathcal{Q}$: What is the correlation? Can you use the formula to find out?
$\mathcal{A}$:

$$corr[X,Y] = \frac{cov[X,Y]}{sd[X].sd[Y]} = \frac{\sigma_{XY}}{\sigma_X \sigma_Y}$$

First,

$$\sigma_x^2 = \left(1 - \frac{3}{2}\right)^2 .0.5 + \left(2 - \frac{3}{2}\right)^2 .0.5 = \frac{1}{4}$$
$$\sigma_y^2 = (1-2)^2 .0.25 + (2-2).0.5 + (3-2).0.25 = \frac{1}{2}$$

Thus $\sigma_x = \frac{1}{2}$ and $\sigma_y = \sqrt{\frac{1}{2}}$

Thus, we have $corr[X,Y] = \frac{\frac{1}{4}}{\frac{1}{2} * \sqrt{\frac{1}{2}}} = 0.71$

7.4 Conditional expectation

The conditional expectation of Y given X computes the *expected value* of the random variable Y, given that we know $X = x$:

$$E[Y|X = x] = \begin{cases} \sum_{j=1}^{m} y_j f_{Y|X}(y_j|x) \\ \int_{-\infty}^{\infty} y f_{Y|X}(y|x) dy \end{cases}$$

The top is for discrete and bottom for continuous Y. This conditional expectation tells us how the expected value of Y varies with X.

Properties of conditional expectations

(i) $E[g(X)|X] = g(X)$ for any function $g(X)$
(ii) For any functions $g(X)$ and (X),
$$E[g(X)Y + h(X)|X] = g(X)E[Y|X] + h(X),$$
(iii) If X and Y are independent, then $E[Y|X] = E[Y]$.
(iv) $E[E[Y|X]] = E[Y]$ (*Law of Iterated Expectations - LIE*)

This is very important for Prof. Glewwe's section especially. I will provide an intuition and you can look at a proof (easily available) to convince yourself. You will mostly just apply the results afterwards.

Intuition: Think of X as a discrete vector taking on possible values $c_1, c_2, \ldots, c_m$, with probabilities $p_1, p_2, \ldots, p_m$. Then LIE says:

$$E[Y] = p_1 E(y|x = c_1) + p_2 E(y|x = c2) + \cdots + p_m E(y|x = c_m)$$

That is, $E[Y]$ is simply a weighted average of the $E(y|x = c_j)$, where the weight p_j is the probability that x takes on the value of c_j. In other words, a weighted average of averages. E.g., suppose we are interested in average IQ generally, but we have measures of average IQ by gender. We could figure out the quantity of interest by weighting average IQ by the relative proportions of men and women.

(v) $E[Y|X] = E[E[Y|X,Z]|X]$
(vi) If $E[Y|X] = E[Y]$, then $cov[X,Y] = 0$

The **conditional variance** is given by:

$$var[Y|X = x] = E[Y^2|x] - \left[E[Y|x]\right]^2$$

If X and Y are independent, however, $var[Y|X] = var[Y]$

7.5. Brief notes on common probability distributions:

Normal distribution: The normal distribution for the random variable X with mean μ and standard deviation σ is given by:

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

It is often denoted as $x \sim N[\mu,\sigma^2]$. The normal distribution has a useful property (*amongst others I imagine*).

$$X \sim N[\mu,\sigma^2], (aX+b) \sim N[a+b\mu, b^2\sigma^2]$$

Although I won't do a proof here, when standardized to variable $z = \frac{X-\mu}{\sigma}$, we have $E[Z] = 0$ and $var(Z) = 1$. As we saw earlier, the normal distribution is always a symmetric distribution. Then $Z \sim N[0,1]$

We will often work with normal distributions and we further define the Chi-squared, t and F-distribution.

Let Z_i for $i = 1, 2, \ldots, n$ be independent random variables each distributed as standard normal. A chi-square distribution with n degrees of freedom is given as:

$$X = \sum_{i=1}^{n} Z_i^2 \sim \mathcal{X}_n^2$$

Let $Z \sim N[0,1] and\ X \sim \mathcal{X}_n^2$ be independent random variables, then the ratio:

$T = \frac{Z}{\sqrt{\frac{X}{n}}} \sim t_n$. has the t-distribution with n degrees of freedom.

Let $X_1 \sim \mathcal{X}_{n1}^2$ and $X_2 \sim \mathcal{X}_{n2}^2$ be independent random variables, then the ratio:

$$F = \frac{X_1/n_1}{X_2/n_2} \sim F_{n_1,n_2}$$

has the F-distribution with n_1 and n_2 degrees of freedom. We say n_1 is the numerator degrees of freedom and n_2 is the denominator degrees of freedom.

You may also come across the following distributions but going through those is outside the scope of this class.

Lognormal distribution
The lognormal distribution is convenient for modeling size distributions, such as the distribution of firm sizes or distribution of income. It is given as:

$f(x) = \frac{1}{\sqrt{2\pi}\sigma x} e^{-\frac{1}{2}(\frac{(lnx-\mu)^2}{\sigma^2})}$, for $x > 0$

Gamma distribution
The Gamma distribution can also be used for the study of income distribution, we have:

$f(x) = \frac{\lambda^P}{\Gamma(P)} e^{-\lambda x} x^{P-1}$, for $x \geq 0, \lambda > 0, P > 0$
$\Gamma(P)$ is the gamma function which is equal to $\int_0^\infty t^{P-1} e^{-t} dt$

Logistic distribution
The logistic distribution is similar to the normal distribution but allows for more "thickness" in the tails of the distribution:

$$f(x) = \frac{1}{1+e^{-x}}(1 - \frac{1}{1+e^{-x}})$$

7.5. Hypothesis testing – very basic

Why do we care about hypothesis testing?

So far, we assumed to know the parameter values and investigated the properties of the distribution (i.e. we knew that μ and σ pretty straightforwardly). However, knowing the true values of the parameters is not a straightforward task. We can use the data to estimate the parameters, the second is to guess a value for the parameters and ask the data whether this value is true – this is close to a proof by contradiction in essence. The former approach is **estimation** and the latter is **hypothesis testing.**

Hypothesis testing will let us make decisions about specific values of parameters or relationships between parameters.

In hypothesis testing, we use sample data to choose between two competing hypotheses. There are 2 choices, the null hypothesis and the alternative hypothesis. We assume the null hypothesis is true until the alternative is shown beyond "chance".

Definition. A *type I error* is the rejection of the null hypothesis when the null hypothesis is true. The probability of a type I error is the size of the test. We denote the size of a test as α (it can also be called the significance level).

Definition. A *type II error* is the failure to reject the null hypothesis when the null hypothesis is false. The power of a test is the probability that it will correctly lead to rejection of a false null hypothesis.

More simply stated, a type I error is detecting an effect that is not present, while a type II error is failing to detect an effect that is present.

Power of a test: power = $1 - \beta = 1 - P[$ type II error $]$, β is simply the probability of making a type II error.

In any hypothesis test, you need to balance α and β. You can never get them both to be 0 with a sample. In practice, we choose the largest α that is tolerable - usually .01, .05 or .10. The precise relationship between α and β cannot be easily determined. That is to say, when α increases, β decreases (and power would increase), and vice versa. It hard to say by how much. Power really gets at the probability of correctly rejecting the null

hypothesis in favor of the alternative when the alternative is true. Ideally, you want power to be as large (close to 1) as possible. There are 2 quantities that you have control over that can help you increase power.

To increase power, you can:
1. Increase sample size, n
2. Increase α, the significance level. (Note you still don't want to set this too high).

To begin, a hypothesis is a claim about a population characteristic (parameter). A hypothesis test is set up as follows:

(1) State the null hypothesis: $H_0 = \mu = \mu_0$ where μ_0 is some value. Remember, we assume that the null hypothesis is true until we statistically prove that it is not true. This is often referred to as the status quo - initially assumed true.

(2) State the alternative hypothesis: $H_1 : \mu \neq \mu_0$ for a two-tailed test or $H_1 : \mu \geq \mu_0$, $H_1 : \mu \leq \mu_1$ for a one-tailed test. This is usually your proposal – what you want to show.

(3) Decide on the significance level α of the test. Standard values are 0.10, 0.05, and 0.01.

Obtain the critical value for the test.

(4) Calculate the test statistic.

(5) If the test statistic falls in the rejection region then reject the null hypothesis. If the test statistic does not fall in the rejection region then we fail to reject the null hypothesis.

The cases for hypothesis testing can be summarized as follows:

H_0	H_1	Test	Rejection area
$\mu = \mu_0$	$\mu \neq \mu_0$	Two-tailed	$\lvert t \rvert = \left\lvert \frac{\bar{\theta} - \mu_0}{\frac{s_\theta}{\sqrt{n}}} \right\rvert > C_{1-\frac{\alpha}{2}}$
$\mu = \mu_0$	$\mu > \mu_0$	One-tailed	$t = \frac{\bar{\theta} - \mu_0}{s_\theta/\sqrt{n}} > C_{1-\frac{\alpha}{2}}$
$\mu = \mu_0$	$\mu < \mu_0$	One-tailed	$t = \frac{\bar{\theta} - \mu_0}{s_\theta/\sqrt{n}} > -C_{1-\frac{\alpha}{2}}$

θ is a parameter and s_θ is the standard deviation of that parameter.

Note that $C_{1-\frac{\alpha}{2}}$ denotes the critical value for the corresponding distribution and confidence level.

We will do one quick example and get to the next section.

A principal at a certain school claims that the students in his school are above average intelligence. A random sample of thirty students IQ scores have a mean score of 112. Is there sufficient evidence to support the principal's claim? The mean population IQ is 100 with a standard deviation of 15.

State the null hypothesis:

$$H_0: \mu = 100$$

State the alternate hypothesis:

$$H_a: \mu > 100$$

Is this a one tailed or two tailed test?

One tailed

What would be an appropriate rejection area if you are given an $\alpha = 0.05$.

For those who remember, this is z=1.645

What is the t test statistic?

$$t = \frac{\bar{\theta} - \mu_0}{s_\theta/\sqrt{n}} = \frac{112.5 - 100}{15/\sqrt{30}} = 4.56$$

If the test statistic is greater than the z-score, we reject the null.

7.5 Brief review of large sample distribution theory

After some thought, I decided to do only a brief review of this section. I've been convinced that this part is nicely and thoroughly covered by Prof. McCullough, so I will let him take you on this ride. You will learn much more about this in class – so do not worry too much about this being only a brief introduction.

7.5.1. A note of unbiasedness and consistency

One philosophical foundation of our approach in thinking about estimation and hypothesis testing is that sample data, say, $X_1, X_2, \ldots, X_n$, of a sample of size n, are thought of as a (subset of an infinite) collection of independent, identically distributed (i.i.d.) random variables, following the probability distribution in question (for example, say a normal distribution). This does not always have to be true, sometimes things are not that nice. A good example is a succession of throws of a fair coin: The coin has no memory, so all the throws are "independent". And every throw is 50:50 (heads, tails), so the coin is and stays fair - the distribution from which every throw is drawn, so to speak, is and stays the same: "identically distributed". This a good example, but it should be not confused to mean equi-probable, i.e., all outcomes are equally likely. This is more related to this example.

Let θ be some unknown parameter. An estimator is then a rule or strategy that we use to estimate the parameter θ given our random sample of data. Let's denote our estimate as $\hat{\theta}$.

Definition. An estimator $\hat{\theta}$ of a parameter θ of a distribution is called unbiased estimator if

$$E[\hat{\theta}] = \theta$$

It should be noted that the estimator $\hat{\theta}$ will be a function of the measurements $(X_1, X_2, \ldots, X_n)$ on the sample, i.e. $\hat{\theta} = \widehat{\theta}\,(X_1, X_2, \ldots, X_n)$.

Definition. An unbiased estimator $\hat{\theta}_1$ is more efficient than another unbiased estimator $\hat{\theta}_2$ if the sampling variance of $\hat{\theta}_1$is less than that of $\hat{\theta}_2$:

$$var(\hat{\theta}_1) < var(\hat{\theta}_2)$$

Definition: The random variable x_n converges in probability to a consistent c if:

$$\lim_{n\to\infty} P([x_n - c] > \varepsilon) = 0$$

for any $\varepsilon > 0$. For x_n that converges in probability to c we can also write:

$$plim\, x_n = c$$

I personally, I'm not fan of this quite popular definition. I prefer this one:

$$\lim_{n\to\infty} P([x_n - b] < \varepsilon) = 1$$
$$plim\, x_n = b$$

An estimator $\hat{\theta}_n$ of a parameter θ is a consistent estimator of θ if and only if plim $\theta_n = \theta$.

We write $plim\ b_n$ = b, where $plim$ is short-hand for probability limit, or $b_n \xrightarrow{p} b$. The limit b may be a constant or a random variable.

7.5.2. Large numbers

One law of large numbers states that the mean of a random sample is always a consistent estimate of the true population average, that is:

$$plim\ \hat{x}_n = \mu$$

In simple terms, we can get arbitrarily close to estimating the true mean by using a large sample.

Properties of probability limits

Let $x_n$and y_n be random variables with $plim\ x_n = c$ and $plim\ y_n = d$, then:

(i) $plim(x_n + y_n) = c + d$,
(ii) $plim\ x_n y_n = cd$,
(iii) $plim \frac{x_n}{y_n} = \frac{c}{d}$
(iv) for a continuous function $g(x_n)$ that is not a function of n, $plim\ g(x_n) = g(plim\ x_n)$
(□) $plim\ b = b$, for b is a constant.

You will use these rules to show that Ordinary Least Squares (OLS) estimate is consistent for example.

7.6 Optimizing expected values

Initially, I had planned to cover probability before optimization, hence you have this in your syllabus. However, we will learn about optimization starting next week. Decision makers often have to make decisions in the presence of uncertainty. In this case, decision problems are often formulated as optimization problems to be solved in a setting in which the optimization depends on parameters which are unknown. I think it will make more sense if we cover this after doing some regular optimization problems, so we will keep this as an add on to our optimization section.