Lecture 1

Recommended Readings: WMS Chapter 1

Learning objectives

1. In examples of real world experiments, and as a contrast to noiseless mathematical variables, students will correctly recognize the existence of uncertainty.
2. After studying the definitions, students will be able to recall the relationship between probability and statistics (in 1 sentence).
3. In practical examples, students will correctly identify the inferential objective, population of interest, and practical collection methods.
4. After a brief in class review, students will correctly identify the axes of a frequency distribution (histogram).
5. Given their mathematical definitions, students will accurately recall and interpret the definitions of sample mean and sample variance.
6. After reviewing theory of normally (Gaussian) distributed random variables, using mean and variance, students will correctly make inferences in simple experiments.

Part 1

Outside of probability, mathematical variables are noiseless. Meaning for example:

$$y = (3 + x)^2,$$

if we know $x = 6$, we arrive at the answer, which is always 81.
What about computer variables, like floating point numbers, are those any different?

Definition: Experiment. An experiment is the process by which an observation is made.

Examples of experiments include uncontrollable situations like the price of a particular stock, or observations under controlled laboratory conditions like a clinical trial.

Problem # 1

Recognize the uncertainty (if any) in the following experiments:
a) Distance of bike path from Pittsburgh to DC as reported by google maps.
b) Time to complete (1) for 100 cyclists traveling independently.

Part 2

We can think of a population as an entity which encompasses all possible data that can be observed in an experiment. This can be a tangible collection of objects or a more abstract entity.
Examples of populations are:

- the deers in Frick Park;
- the runners of all the marathons in the USA;
- the subscribers of HBO;
- the number of burritos sold daily at Chipotle.

Most often, we perform an experiment because we are interested in learning about a particular feature or parameter of a population. Examples of parameters for the above populations are respectively:

- the total number;
- the average time of the finishers;
- the proportion of subscribers that watched the final episode of GoT;
- the variability of the sales.

The objective of statistics is to make an inference (answer a scientific question) about a population based on information contained in a sample taken from that population.
Probability is the foundation of the theory of statistics.

Using probability we can predict from a known population to the outcome of a single experiment (sample). Statistics utilizes the probability of an observed sample to infer the characteristics of an unknown population.

We have mentioned probability several times, but what do we mean by that? There are two major interpretations:

1. objective probability: the long-run frequency of the occurrence of a given event (e.g. seeing heads in a coin flip)
2. subjective probability: a subjective *belief* in the rate of occurrence of a given event.

Methods based on objective probability are usually called *frequentist* or *classical* whereas those based on subjective probability are usually called *Bayesian* as they are associated to the Bayesian paradigm, although Larry argues that frequentist and Bayesian should be characterized in terms of their goals, and

not in terms of their methods.[1] In this class, we focus on frequentist methods (but we will discuss later in the course the basic idea at the basis of Bayesian Statistics).

Part 3

In general, the study of statistics is concerned with the design of experiments to obtain a specified quantity of information at a minimum cost and the optimum use of this information in making inferences about a population. The objective of statistics is to make an inference about a population based on information contained in a sample from that population and to provide a measure of goodness for the inference (more on this in future lectures).

Start with the question of interest (inferential objective), then identify the population of interest, and find a way to collect experimental data.

This class will provide you with the tools to approach these problems in a statistical manner. Statistics is a result of the following process:

TRUTH $\rightarrow$ MODELS $\rightarrow$ STATISTICS

There are situations in which the quantity that we want to study evolves with time (or with space, or with respect to some other dimension). For example, suppose that you are interested in the number of people walking into a store on a given day. We can model that quantity as a random quantity X_t varying with respect to time t. Most often there exists some kind of 'dependence' between the number of people in the store at time t and the number of people in the store at time t' (especially if $|t - t'|$ is small). There is a branch of Probability Theory that is devoted to study and model this type of problems. We usually refer to the collection $\{X_t : t \in \mathcal{T}\}$ as a 'random process' or as a 'stochastic process'. The set $\mathcal{T}$ can represent a time interval, a spatial region, or some more elaborate domain.
Another example of a random process is observing the location of robberies occurring in a given city. (Question: what is T)?
Another typical example of a random process is the evolution of a stock price in Wall Street as a (random) function of time. (Question: what is T)?
We will devote part of this course to study (at an introductory level) some of the theory related to random processes.

Part 4

Once you have collected some data, you can start by doing some exploratory data analysis. A useful graphical method is the relative frequency distribution

[1] https://normaldeviate.wordpress.com/2012/11/17/what-is-bayesianfrequentist-inference/

(histogram) of the data. The frequency is typically shown on the vertical axis, and the domain of the data are shown on the horizontal axis. This should be a review from previous introductory courses.

Part 5

Compared to histograms, numerical descriptive measures are often more useful when we wish to make an inference and measure the goodness of that inference. Here we focus on two types of descriptive numbers: measures of central tendency and measures of dispersion or variation.

The *mean* of a sample of n measured responses $y_1, y_2, ..., y_n$ is given by

$$\bar{y} = \sum_{i=1}^{n} y_i. \tag{1}$$

The corresponding population mean is denoted as μ.
Similarly the *variance* of n measured responses is the sum of the square of the differences between the measurements and their mean, divided by n-1.

$$s^2 = \frac{1}{n-1} \sum_{i=1}^{n} (y_i - \bar{y})^2 . \tag{2}$$

The corresponding population variance is denoted by the symbol σ^2.

Part 6

Many distributions of data in real life are bell-shaped. The truth need not be exactly bell-shaped, but the normal (Gaussian) approximation is often useful. For a distribution of measurements that is approximately Normal (Gaussian) it follows that the interval with end-points:

- $\mu \pm \sigma$ contains approximately 68% of the measurements.
- $\mu \pm 2\sigma$ contains approximately 95% of the measurements.
- $\mu \pm 3\sigma$ contains almost all of the measurements.

Lecture 2

Problem # 2

Weekly maintenance costs for a factory, recorded over a long period of time and adjusted for inflation, tend to have an approximately normal distribution with an average of $420 and a standard deviation of $30. If $450 is budgeted for next week, what is an approximate probability that this budgeted figure will be exceeded?

Recommended Readings: WMS Chapter 2.1-2.5
Learning objectives

1. Based on the concept of relative frequency, students will be able to identify the long series of observations used by a given frequentist probability model.
2. Given examples and discussion, students will correctly categorize (1) theoretical probability models and (2) empirical observations.
3. After reviewing basic set theory (including Venn diagrams), students will be able to correctly perform elementary set operations in the context of real world applications.
4. Given their definitions in the context of discrete probability, students will be able to summarize the 3 axioms of probability (in 1 sentence).
5. Given examples of real world applications, students will be able to correctly calculate the probability of simple discrete random events.

Part 1

In everyday use, probability describes the occurrence of a future event. In particular, we consider events that are noisy, meaning random (stochastic) events that cannot be predicted with certainty. For example, the exact number of hours that a battery will last cannot be predicted with certainty.
Frequentist (sometimes called classical) probability models are based on the relative frequency with which events occur in a long series of observations. This long-term relative frequency is often stable and provides an intuitive measure that helps us predict the occurrence of a random event in a future observation. A simple example is the outcome of a single toss of a coin, whether it turns out heads or tails is impossible to predict with certainty; however over many observations ($y_1, y_2, y_3, ..., y_n$, for a sufficiently large n) it is safe to assume that the relative frequency is 50% heads and 50% tails. Then we can use this information as a measure of the chance of winning in a single toss.

Part 2

As discussed on the previous lecture, probability models are theoretical descriptions of the population (true random events). As such we can theorize, for example, that the probability of each face from a fair die (cube with six faces) is 1/6. Thus we can draw a frequency distribution with exactly 1/6 for each of the 6 faces. If you wished to confirm that your own die is fair, you could throw it many n times $(y_1, y_2, ..., y_n)$ and on each time record the outcomes and then create an *empirical* histogram. The word empirical here means that it is the result of an observation from nature as opposed to theory. How many observations would you say are needed to confirm that your own die is fair? Once you have collected observations, the role of statistics is to make the trip in reverse: infer the probability model, and use probability to produce an inference (estimation or decision) with an associated degree of goodness.

Part 3

In probability theory, the notion of an 'event' is usually described in terms of a set (of outcomes). Therefore, now it is worthwhile reviewing some basic facts about set theory.
Throughout the course we will adopt the convention that Ω denotes the universal set (the superset of all sets) and $\emptyset$ denotes the empty set (i.e. a set that does not contain any elements).
Please pay attention to notation, there are many types of notations used by different authors; for instance, the WMS textbook uses S for the universal set, and $^-$ instead of ^{c} for complement.
Recall that a set A is a subset of another set B (or A is contained in B) if for any element x of A we have $x \in A \implies x \in B$. We denote the relation A is a subset of B by $A \subset B$ (similarly, $A \supset B$ means A is a superset of B or A contains B). Clearly, if $A \subset B$ and $B \subset A$, then $A = B$. Let $\wedge$ and $\vee$ stand for 'and' and 'or' respectively. Given three subsets A, B, C of Ω, recall the following set properties:

- commutativity: $A \cup B = B \cup A$ and $A \cap B = B \cap A$;
- associativity: $A \cup (B \cup C) = (A \cup B) \cup C$ and $A \cap (B \cap C) = (A \cap B) \cap C$;
- distributive laws: $A \cap (B \cup C) = (A \cap B) \cup (A \cap C)$ and $A \cup (B \cap C) = (A \cup B) \cap (A \cup C)$;
- De Morgan's laws: $(A \cup B)^c = A^c \cap B^c$ and $(\cap_{i=1}^n A_i)^c = \cup_{i=1}^n A_i^c$.

and the following basic set operations:

- union of sets: $A \cup B = \{x \in \Omega : x \in A \vee x \in B\}$
- intersection of sets: $A \cap B = \{x \in \Omega : x \in A \wedge x \in B\}$
- complement of a set: $A^c = \{x \in \Omega : x \notin A\}$

- set difference $A \setminus B = \{x \in \Omega : x \in A \wedge x \notin B\}$
- symmetric set difference $A \Delta B = (A \setminus B) \cup (B \setminus A) = (A \cup B) \setminus (A \cap B)$.

These can be extended to the union and the intersection of any $n > 0$ sets:

- $(\cup_{i=1}^{n} A_i)^c = \cap_{i=1}^{n} A_i^c$
- $(\cap_{i=1}^{n} A_i)^c = \cup_{i=1}^{n} A_i^c$.

The same strategy for the proof can be used for the second statement.
We typically use *Venn's diagrams* to represent logical relations between sets.

Notice that, for any set $A \subset \Omega$, $A \cup A^c = \Omega$ and $A \cap A^c = \emptyset$. Two sets $A, B \subset \Omega$ for which $A \cap B = \emptyset$ are said to be disjoint or mutually exclusive.
A **partition** (disjoint union) of Ω is a collection of subsets $A_1, \cdots, A_n$ that satisfy

1. $A_i \cap A_j = \emptyset$, $\forall i \neq j$
2. $\cup_{i=1}^{n} A_i = A_1 \cup A_2 \cup \cdots \cup A_n = \Omega$

We used before the word 'experiment' to describe the process of collecting data or observations. An experiment is, broadly speaking, any process by which an observation is made. This can be done actively, if you have control on the apparatus that collects the data, or passively, if you only get to see the data, but you have no control on the apparatus that collects them. An experiment generates observations or outcomes. Consider the following example: you toss a fair coin twice (your experiment). The possible outcomes (simple events) of the experiment are HH, HT, TH, TT (H: heads, T: tails). The collection of all possible outcomes of an experiment forms the 'sample space' of that experiment. In this case, the sample space is the set $\Omega = \{HH, HT, TH, TT\}$. Any subset of the sample space of an experiment is called an event. For instance, the event 'you observe at least one tails in your experiment' is the set $A = \{HT, TH, TT\} \subset \Omega$.
Note: These are *discrete* events, i.e. they are a collection of sample points from a *discrete sample space*. A discrete sample space Ω is one that contains either a finite or countable number of distinct sample points. An event in a discrete sample space is simply a collection of sample points – any subset of Ω.
In this part of the course, just assume we deal with discrete events from a discrete sample spaces. Later, we will deal with *continuous* sample spaces.

Part 4

Modern Probability Theory is built on a set of axioms[2] formulated by the Russian mathematician Andrey Kolmogorov in his masterpiece *Foundations of the Theory of Probability*.
A *probability measure* P on Ω is a set function[3] satisfying the following axioms:

[2] An axiom is a statement believed to be true.
[3] A set function is a function whose domain is a collection of sets.

1. for any $A \subset \Omega$, $P(A) \in [0,1]$; The relative frequency of occurrence of any event must be greater than or equal to zero. A negative relative frequency does not make sense.

2. (norming) $P(\Omega) = 1$; The relative frequency of the whole sample space Ω must be unity. Because every possible outcome of the experiment is a point in Ω, it follows that Ω must occur every time the experiment is performed.

3. (countable additivity) for any countable collection of disjoint events $\{A_i\}_{i=1}^{\infty} \subset \Omega$, $P(\cup_{i=1}^{\infty} A_i) = \sum_{i=1}^{\infty} P(A_i)$. If two events are mutually exclusive, the relative frequency of their union is the sum of their respective relative frequencies. (For example, if the experiment of tossing a balanced die yields a 1 on 1/6 of the tosses, it should yield a 1 or a 2 on $1/6 + 1/6 = 1/3$ of the tosses.)

From the second and third axiom, it follows that $P(\emptyset) = 0$.

Now, what do these axioms imply about $P(A \cup B)$? Can you show that $P(A \cup B) = P(A) + P(B) - P(A \cap B)$.?
In case of $A \cap B = \emptyset$ we clearly obtain

$$\begin{aligned} P(A \cup B) &= P(A) + P(B) - P(A \cap B) \\ &= P(A) + P(B) - P(\emptyset) \\ &= P(A) + P(B). \end{aligned}$$

This can be extended to more than 2 events. For instance, given $A, B, C \subset \Omega$, we have

$$\begin{aligned} P(A \cup B \cup C) = P(A) + P(B) + P(C) - P(A \cap B) \\ - P(A \cap C) - P(B \cap C) + P(A \cap B \cap C) \end{aligned}$$

and, in general, for $A_1, \ldots, A_n \subset \Omega$ we have the so called inclusion-exclusion formula (note the alternating signs for the summands):

$$\begin{aligned} P(\cup_{i=1}^{n} A_i) = \sum_{i=1}^{n} P(A_i) - \sum_{1 \le i < j \le n} P(A_i \cap A_j) \\ + \sum_{1 \le i < j < k \le n} P(A_i \cap A_j \cap A_k) - \cdots + (-1)^{n-1} P(A_1 \cap \cdots \cap A_n). \end{aligned}$$

Notice that in general we have the following *union bound* for any $A_1, \ldots, A_n \subset \Omega$:

$$P(\cup_{i=1}^{n} A_i) \le \sum_{i=1}^{n} P(A_i).$$

Part 5

For a given experiment with m possible outcomes, how can we compute the probability of an event of interest? We can always do the following:

1. define the experiment and describe its simple events E_i, $i \in \{1, \ldots, m\}$;
2. define reasonable probabilities on the simple events, $P(E_i)$, $i \in \{1, \ldots, m\}$;
3. define the event of interest A in terms of the simple events E_i;
4. compute $P(A) = \sum_{i:E_i \in A} P(E_i)$.

Lecture 3

Recommended Readings: WMS Chapter 2.6

Learning objectives

1. Given definitions and examples involving the uniform probability distribution on finite spaces, students will correctly count the number of sample points in A, the number of total possible sample points in Ω, and correctly compute P(A), the probability of A.

2. Given their definitions and examples, students will be able to correctly solve basic counting problems using combinatorial principles such as the multiplication rule, permutations, and combinations.

Example: There are 5 candidates for two identical job positions: 3 females and 2 males. What is the probability that a completely random selection process will appear discriminatory? (i.e. exactly 2 males or exactly 2 females are chosen for these job positions) We can approach this problem in the following way:

1. we introduce 5 labels for each of the 5 candidates: M_1, M_2, F_1, F_2, F_3. The sample space is then

$$\begin{aligned}\Omega =&\{M_1M_2, M_1F_1, M_1F_2, M_1F_3, M_2M_1, M_2F_1, M_2F_2, M_2F_3, F_1M_1, F_1M_2,\\ &F_1F_2, F_1F_3, F_2M_1, F_2M_2, F_2F_1, F_2F_3, F_3M_1, F_3M_2, F_3F_1, F_3F_2\}\end{aligned}$$

2. because the selection process is completely random, each of the simple events of the sample space is equally likely. Therefore the probability of any simple event is just $1/|\Omega|$

3. the event of interest is $A = \{M_1M_2, M_2M_1, F_1F_2, F_1F_3, F_2F_1, F_2F_3, F_3F_1, F_3F_2\}$

4. $P(A) = P(M_1M_2) + P(M_2M_1) + \cdots + P(F_3F_2) = |A|/|\Omega| = 8/20 = 2/5 = 0.4$.

If the simple events are equally likely to occur, then the probability of a composite event A is just $P(A) = |A|/|\Omega|$.

Questions to ask when you define the sample space and the probabilities on the simple events:

- is the sampling done with or without replacement?
- does the order of the labels matter?
- can we efficiently compute the size of the sample space, $|\Omega|$?

This leads to our next topic, which will equip us with *tools* to conveniently calculate probability.

Tools for counting sample points

Basic techniques from combinatorial analysis come handy for this type of questions when the simple events forming the sample space are equally likely to occur. Let's take a closer look to some relevant cases.

Suppose you have a group of m elements $a_1, \ldots, a_m$ and another group of n elements $b_1, \ldots b_n$. You can form mn pairs containing one element from each group. That is, if $A = \{a_1, \ldots, a_m\}$ and $B = \{b_1, \ldots, b_n\}$, then we are interested in the elements of the cartesian product $A \times B$. Of course you can easily extend this multiplication rule reasoning to more than just two groups.
As an example, let us use the multiplication rule in the following scenario where we are sampling with replacement and the order matters. Suppose You toss a six-sided die twice. You have $m = 6$ simple outcomes on the first roll and $n = 6$ possible outcomes in the second roll. The sample space is

$$\Omega = \{(1,1), (1,2), (1,3), \ldots, (6,5), (6,6)\}.$$

Therefore the size of Ω is $|\Omega| = mn = 6^2 = 36$.
Is the experiment performed with replacement? Yes, if the first roll is a 2, nothing precludes the fact that the second roll is a 2 again. Does the order matter? Yes, the pair (2,5) corresponding to a 2 on the first roll and a 5 on the second roll is not equal to the pair (5,2) corresponding to a 5 on the first roll and a 2 on the second roll.

Sampling without replacement when order matters
An ordered arrangement of r distinct elements is called a *permutation*. The number of ways of ordering n distinct elements taken r at a time is denoted P_r^n where

$$P_r^n = n(n-1)(n-2)\ldots(n-r+1) = \frac{n!}{(n-r)!}. \tag{3}$$

Consider the following example. There are 30 members in a student organization and they need to choose a president, a vice-president, and a treasurer. In how many ways can this be done? The size of the n distinct elements (persons) is $n = 30$, the number of persons to be appointed is $r = 3$. The sampling is done without replacement (a president is chosen out of 30 people, then among the 29 people left a vice-president is chosen, and finally among the 28 people left a treasurer is chosen). Does the order matter? Yes, the president is elected first, then the vice-president, and finally the treasurer (think about 'ranking' the 30 candidates). The number of ways in which the three positions can be assigned is therefore $P_3^{30} = 30!/(30-3)! = 30 * 29 * 28 = 24360$.

Sampling without replacement when order does not matter
The number of *combinations* of n elements taken r at a time corresponds to the number of subsets of size r that can be formed from the n objects. This is

denoted C_r^n where

$$C_r^n = \binom{n}{r} = \frac{P_r^n}{r!} = \frac{n!}{(n-r)!r!}. \tag{4}$$

How did we get such a formula? To get some intuition, first think about all the ordered sets that contain the same r objects: this number is given by P_r^n. Fixed this r elements, let the objects be $x_1, \ldots, x_r$. Then we will have $r!$ different possible permutations of these r objects. Why? It is clear that x_1 will appear in any of r positions in the set, hence we have r choices. For any choice, x_2 will appear in any of the $r-1$ positions left. And so on .... This is exactly equal to the permutation of r objects. Therefore we divide P_r^n by $r!$ to obtain the number of combinations.

Here is an example. How many different subsets of 3 people can become officers of the organization formed by 30 people, if chosen randomly? Order doesn't matter here (we are not interested in the exact appointments for a given candidate). The answer is therefore $C_3^{30} = 30!/(27! * 3!) = 30 * 29 * 28/6 = 4060$.

The *binomial coefficient* $\binom{n}{r}$ can be extended to the multinomial case in a straightforward way. Suppose that we want to partition n distinct elements in k distinct groups of size $n_1, \ldots, n_k$ in such a way that each of the n elements appears in exactly one of the groups. This can be done in

$$\binom{n}{n_1 \ldots n_k} = \frac{n!}{n_1! n_2! \ldots n_k!} \tag{5}$$

possible ways.

The connection to the combinations seen above is more clear through the following decomposition:

$$\binom{n}{n_1 \ldots n_k} = C_{n_1}^n C_{n_2}^{n-n_1} C_{n_3}^{n-n_1-n_2} \ldots C_{n_k}^{n-\sum_{i=1}^{k-1} n_i}.$$

Lecture 4

Recommended Readings: WMS Chapter 2.6
This lecture will continue the material from lecture 3.

Lecture 5

Recommended Readings: WMS Chapter 2.7-2.13

Learning objectives

1. Given their definitions, students will be able to write one sentences descriptions of the following concepts: conditional probability, independence, and law of total probability.

2. Students will be able to write the three axioms of probability for $P(\cdot|B)$ in mathematical and one sentence form, and thus recognize that the conditional probability $P(\cdot|B)$ with respect to an event B with $P(B) > 0$ is as a proper probability.

3. Given definitions and examples, students will be able to frame and solve conditional probability problems involving two or a small number of events.

Conditional Probability

The probability of an event A may vary as a function of the occurrence of other events. Then, it becomes interesting to compute *conditional* probabilities, e.g. the probability that the event A occurs *given* the knowledge that another event B occurred.
As an example, suppose you toss 2 dice, and that each of the 36 possible outcomes has probability 1/36. If the first die turns up 3, then given this information, what is the probability that the sum of the two dice equals 5? We can proceed by noting that given that the first die is a 3, the only possible outcomes now are: (3,1), (3,2), (3,3), (3,4), (3,5), and (3,6). Before observing the outcome of the first die, these outcomes had equal probabilities, so it seems intuitive that now they also continue to have equal probabilities. Hence, given that the first die turns up 3, these six outcomes have each (conditional) probability of 1/6. Conditionally, the other 30 outcomes in the sample space now have zero probability. The probability of the sum of the two dice being 5, given that the first dies is a 3, is 1/6.

The conditional probability of an event A given another event B is usually denoted as $P(A|B)$. The mathematical definition is:

$$P(A|B) = \frac{P(A \cap B)}{P(B)}. \tag{6}$$

Observe that the quantity $P(A|B)$ is well-defined only if $P(B) > 0$.

Consider the following table, where each entry denotes the probability of the events in the respective row and column:

	B	B^c
A	0.3	0.1
A^c	0.4	0.2

The probability of the intersection of A and B is $P(B \cap A) = 0.3$. Using this table, we can compute the unconditional probabilities of the events A and B: $P(A) = 0.3 + 0.1 = 0.4$ and $P(B) = 0.3 + 0.4 = 0.7$.

The conditional probability of event A given event B is $P(A|B) = P(A \cap B)/P(B) = 0.3/0.7 = 3/7$, the conditional probability of event A^c given event B is $P(A^c|B) = P(A^c \cap B)/P(B) = 0.4/0.7 = 4/7$.
In the same way, we can compute the probability of event B given event A is $P(B|A) = P(B \cap A)/P(A) = 0.3/0.4 = 3/4$. The conditional probability of event A given that B^c occurs is $P(A|B^c) = P(A \cap B^c)/P(B^c) = 0.1/0.3 = 1/3$. Does the probability of A vary as a function of the occurrence of the event B?

The conditional probability $P(\cdot|B)$ with respect to an event B with $P(B) > 0$ is a proper probability measure, therefore the three axioms of probability hold for $P(\cdot|B)$ as well:

- $0 \leq P(A|B) \leq 1$
- $P(\Omega|B) = 1$
- if A_i, $i = 1, ...,$ are *mutually exclusive events*, then $P(\cup_{i=1}^{\infty} A_i|B) = \sum_{i=1}^{\infty} P(A_i|B)$

For instance, for disjoint events $A_1, A_2 \subset \Omega$
we have $P(A_1 \cup A_2|B) = P(A_1|B) + P(A_2|B)$.

Exercise
There is 20% chance that you go to Craig Street to have lunch at Sushi Fuku, a 30% chance that you get a coffee at Starbucks, and a 10% chance that you both have lunch at Sushi Fuku and get a coffee at Starbucks. What's the probability that you get a coffee at Starbucks if you have been to Sushi Fuku? What about the probability that you get lunch at Sushi Fuku given that you have been to Starbucks?

Independence

The event $A \subset \Omega$ is said to be independent of the event $B \subset \Omega$ if $P(A \cap B) = P(A)P(B)$. This means that the occurrence of B does not alter the chance that A happens. In fact, assuming that $P(B) > 0$, we easily see that this is equivalent to $P(A|B) = P(A \cap B)/P(B) = P(A)P(B)/P(B) = P(A)$.
Furthermore, assuming that also $P(A) > 0$, we have that $P(A|B) = P(A)$ is equivalent to $P(B|A) = P(B)$ (independence is a symmetric relation!).

Exercise
Let $A, B \subset \Omega$ and $P(B) > 0$.

- What is $P(A|\Omega)$?
- What is $P(A|A)$?
- Let $B \subset A$. What is $P(A|B)$?
- Let $A \cap B = \emptyset$. What is $P(A|B)$?

Notice that from the definition of conditional probability, we have the following:

$$P(A \cap B) = P(B)P(A|B) = P(A)P(B|A). \tag{7}$$

This can be generalized to more than two events. For instance, for three events $A, B, C \subset \Omega$, we have

$$P(A \cap B \cap C) = P(A)P(B|A)P(C|A \cap B) \tag{8}$$

and for n events $A_1, \ldots, A_n$

$$P(A_1 \cap \cdots \cap A_n) = P(A_1)P(A_2|A_1)P(A_3|A_1 \cap A_2) \ldots P(A_n|A_1 \cap \cdots \cap A_{n-1}). \tag{9}$$

Independent sequence of events
Whenever A_i is an event whose occurrence is completely determined by the outcome of the ith subexperiment, then the subexperiments A_1, A_2,...,A_n are necessarily an independent sequence of events. If each subexperiment has the same set of possible outcomes, then the subexperiments are often called trials.

Warning: Independence and Disjoint are not the same.
Two events being disjoint simply means that they do not share any jointly occurring elements. For instance, in a single coin flip example, $A = \{\mathbf{H}\}$ and $B = \{\mathbf{T}\}$ are disjoint events, but A gives *perfect* knowledge of B, as we know that $P(A|B) = 0$ and $P(A|B^C) = 1$, so that $P(A|B) \neq P(A) = 1/2$.
Exercise
Consider the following events and the corresponding table of probabilities:

	B	B^c
A	0	0.2
A^c	0.4	0.4

Are the events A and B disjoint?
Are the events A and B independent?

Exercise
Consider the following events and the corresponding table of probabilities:

	B	B^c
A	1/4	1/12
A^c	1/2	1/6

Are the events A and B disjoint?
Are the events A and B independent?

Law of Total Probability and Bayes' Rule

Assume that $\{B_i\}_{i=1}^{\infty}$ is a partition of Ω, i.e. for any $i \neq j$ we have $B_i \cap B_j = \emptyset$ and $\cup_{i=1}^{\infty} B_i = \Omega$. Assume also that $P(B_i) > 0 \, \forall i$. Then, for any $A \subset \Omega$, we have the so-called *law of total probability*

$$P(A) = \sum_{i=1}^{\infty} P(A|B_i)P(B_i). \tag{10}$$

Indeed, $A = \cup_{i=1}^{\infty}(A \cap B_i)$ when $\cup_{i=1}^{n} B_i = \Omega$. But this time the sets $(A \cap B_i)$ are also disjoint since $\{B_i\}_{i=1}^{\infty}$ is a partition of Ω. Furthermore, by equation (7) we have $P(A \cap B_i) = P(A|B_i)P(B_i)$. Thus, $P(A) = \sum_{i=1}^{\infty} P(A \cap B_i) = \sum_{i=1}^{\infty} P(A|B_i)P(B_i)$.

We can now use the law of total probability to derive the so-called Bayes' rule. We have

$$P(B_i|A) = \frac{P(B_i \cap A)}{P(A)} = \frac{P(A|B_i)P(B_i)}{\sum_{i=1}^{\infty} P(A|B_i)P(B_i)}. \tag{11}$$

For two events A, B this reduces to

$$P(B|A) = \frac{P(A|B)P(B)}{P(A)} = \frac{P(A|B)P(B)}{P(A|B)P(B) + P(A|B^c)P(B^c)} \tag{12}$$

Exercise
You are diagnosed with a disease that has two types, A and B. In the population, the probability of having type A is 10% and the probability of having type B is 90%. You undergo a test that is 80% accurate, i.e. if you have type A disease, the test will diagnose type A with probability 80% and type B with probability 20% (and vice versa). The test indicates that you have type A. What is the probability that you really have the type A disease?
Let A ='you have type A', B='you have type B', T_A ='the test diagnoses type A', and T_B ='the test diagnoses type B'. We know that $P(A) = 0.1$ and $P(B) = 0.9$. The test is 80% accurate, meaning that

$$P(T_A|A) = P(T_B|B) = 0.8$$
$$P(T_B|A) = P(T_A|B) = 0.2$$

We want to compute $P(A|T_A)$. We have

$$\begin{aligned} P(A|T_A) &= \frac{P(A \cap T_A)}{P(T_A)} = \frac{P(T_A|A)P(A)}{P(T_A|A)P(A) + P(T_A|B)P(B)} \\ &= \frac{0.8 * 0.1}{0.8 * 0.1 + 0.2 * 0.9} = \frac{8}{26} = \frac{4}{13}. \end{aligned}$$

Lecture 6

Recommended Readings: WMS Chapter 2.7-2.13
This lecture will finish the material from lecture 5. We will talk about independent sequence of events, the law of total probability, and Bayes' rule.

Additional topic: conditional Independence

We are now equipped to talk about another 'parallel' concept to independence. Let C be an event with $P(C) > 0$. We say that the events A and B are conditional independent of C if

$$P(A \cap B|C) = P(A|C)P(B|C).$$

This implies that

$$P(A|B \cap C) = P(A|C)$$

To see this notice that conditional independence of A,B given C implies that

$$\begin{aligned} P(A|C)P(B|C) &= P(A \cap B|C) \\ &= \frac{P(A \cap B \cap C)}{P(C)} \\ &= \frac{P(C)P(B|C)P(A|B \cap C)}{P(C)} \\ &= P(B|C)P(A|B \cap C). \end{aligned}$$

Remarks

- Independence does not imply conditional independence
- Conditional independence does not imply independence
- Pairwise independence does not imply independence.
  Consider $A_1, \ldots, A_n \subset \Omega$. If for every $1 \leq i < j \leq n$ we have $P(A_i \cap A_j) = P(A_i)P(A_j)$, the events are said to be pairwise independent. Mutual independence, instead, is defined as $P(\cap_{i=1}^n A_i) = \prod_{i=1}^n P(A_i)$. Notice that mutual independence implies pairwise independence, but the converse is not true.

Lecture 7

Recommended Readings: WMS Chapter 3.1,3.1,4.1,4.2

Learning objectives

1. In one sentence, students will be able to correctly recall the definition and utility of random variables.
2. Given a simple experiment, students will be able to correctly construct a random variables and for each one of its values compute its associated probabilities.

Random variables

Let's start with an example. Suppose that you flip a fair coin three times. The sample space for this experiment is

$$\Omega = \{HHH, HHT, HTH, HTT, THH, THT, TTH, TTT\}.$$

Since the coin is fair, we have that

$$P(\{HHH\}) = P(\{HHT\}) = \cdots = P(\{TTT\}) = \frac{1}{|\Omega|} = \frac{1}{8}.$$

Suppose that we are interested in a particular quantity associated to the experiment, say the number of tails that we observe. This is of course a random quantity. In particular, we can conveniently define a function $X : \Omega \to \mathbb{R}$ that counts the number of tails. Precisely,

$$\begin{aligned} &X(HHH) = 0 \\ &X(HHT) = 1 \\ &X(HTH) = 1 \\ &X(HTT) = 2 \\ &\vdots \\ &X(TTT) = 3. \end{aligned}$$

We say that P induces through X a *probability distribution* on $\mathbb{R}$. In particular, we can easily see that the probability distribution of the random variable X is

$$\begin{aligned} &P(X = 0) = P(\{HHH\}) = 1/8 \\ &P(X = 1) = P(\{HHT, HTH, THH\}) = 3/8 \\ &P(X = 2) = P(\{HTT, THT, TTH\}) = 3/8 \\ &P(X = 3) = P(\{TTT\}) = 1/8 \\ &P(X = \text{any other number}) = 0. \end{aligned}$$

Remark: although the probability measures are denoted by P on both sides, they do refer to probability measure on different spaces. The measure on the LHS is on $\mathbb{R}$, while the one on the RHS is on Ω. This is more clear if we rewrite it, with some abuse of notation, as $P_{\mathbb{R}}(X = 0) = P_{\Omega}(X^{-1}(X = 0)) = P_{\Omega}(\{HHH\}) = 1/8$. Ω is typically referred to as the *underlying probability space*. Therefore $P_{\mathbb{R}}$ is the probability distribution induced by P_{Ω} through X.

Why is X said to be random if it is just a function? First, the outcome of the experiment is random, so the value that the function X takes is also random. Second, we could modify the underlying probability space Ω without modifying X; therefore "random" aims at highlighting the fact that we are not truly interested in Ω, but in the distribution of X.

Formal definition of a random variable

Suppose we have a sample space Ω. A **random variable** X is a function from Ω into the real line. In other words, $X : \Omega \to \mathbb{R}$ or

$$\omega \in \Omega \mapsto X(\omega) \in \mathbb{R}$$

Random variables will be denoted with capital letters, such as X. This is a consistent, standard notation for a random variable.
Why we care about random variables? After all, based on the example, it seems that we still have to clearly define and specify the sample space in order to determine $P(X = x)$ for $x \in \mathbb{R}$. Actually, it turns out when we model a random quantity of interest (such as the number of tails in the coin tossing example) most of the times one assumes (or knows) a distribution to use. It's just easier and more natural. Of course, this assumption must be reasonable and needs to be rigorously checked by the modeller. By modelling the randomness of a phenomenon as a random variable whose distribution is known, we can bypass the trouble of defining/specifying a sample space. In the example above, if we simply assume that X has a Binomial distribution (to be defined in upcoming lectures), then the sample space is automatically the discrete space of 3-length binary outcomes (HHH, HHT, etc.), and the probability of events of interest (e.g. X=1, which corresponds to one tail out of three throws, which is the subset of the sample space $\{THH, HTH, HHT\}$) is precisely calculable.

To summarise, one should specify Ω and a probability measure on this space, map Ω to $\mathbb{R}$ through X, and analyse the induced probability measure on $\mathbb{R}$. However, one can skip the first step, that is leave Ω undefined, and just specify the induced probability measure on $\mathbb{R}$, which is the object of interest. This is enough to guarantee the existence of some Ω for which this probability measure exists.

At this point it is worthwhile making an important distinction between two types of random variables, namely *discrete* random variables and *continuous* random variables.
We say that a random variable is discrete if the set of values that it can take is at most countable. On the other hand, a random variable taking values in an uncountably infinite set is called continuous.

Question
Consider the following:

- you draw a circle on a piece of paper and one diameter of the circle; at the center of the circle, you keep your pencil standing orthogonal to the plane where the circle lies. At some point you let go the pencil. X is the random variable corresponding to the angle that the pencil forms with the diameter of the circle that you drew after the pencil fell on the piece of paper.

- you roll a die. Y is the random variable corresponding to the number of rolls needed until you observe tail for the first time.

What are the possible values that X and Y can take? Are X and Y discrete or continuous random variables?

Depending on whether a given random variable X is discrete or continuous, we use two special types of functions in order to describe its distribution.

Lecture 8

Recommended Readings: WMS Chapter 3.1,3.1,4.1,4.2

Learning objectives

1. In one sentence students will be able to write the definition and properties of the cumulative distribution function (c.d.f).
2. For simple experiments, students will be able to define a suitable random variable, write down its associated probability function, cumulative distribution function, and support.
3. Students will correctly apply the binomial distribution to model the outcome of n coin flips.

Depending on whether a given random variable X is discrete or continuous, we use two special types of functions in order to describe its distribution.

If X is discrete, let's define its support as the set
$\text{supp}(X) = \{x \in \mathbb{R} : P(X = x) > 0\}$, for a discrete random variable X, $\text{supp}(X)$ is either a finite or countable set. We can describe the probability distribution of X in terms of its *probability mass function* (p.m.f.), i.e. the function

$$p(x) = P(X = x) \tag{13}$$

mapping $\mathbb{R}$ into $[0, 1]$. The probability mass function satisfies the following properties:

1. $p(x) \in [0, 1] \quad \forall x \in \mathbb{R}$
2. $\sum_{x \in \text{supp}(X)} p(x) = 1.$

Another way to describe the distribution of a random variable is by means of its *cumulative distribution function* (c.d.f). Again we will separate the discrete case and the continuous case. If X is a discrete random variable, its CDF is defined as the function $F_X : \mathbb{R} \to [0, 1]$:

$$F(x) = P(X \leq x) = \sum_{x_i \leq x} p(x_i).$$

Notice that for a discrete random variable, F is not a continuous function.
But for all types of random variables, discrete or continuous, the c.d.f satisfies the following properties:

1. Normalized: $\lim_{x \to -\infty} F(x) = 0$ and $\lim_{x \to +\infty} F(x) = 1$
2. Nondecreasing: $x \leq y \implies F(x) \leq F(y)$
3. Right-continuous: for any $x \in \mathbb{R}$ we have $\lim_{y \to x^+} F(y) = F(x)$.

What is the relationship between the c.d.f. of a random variable and its p.m.f? For a discrete random variable X, let $x_{(i)}$ denote the i-th largest element in $\text{supp}(X)$. Then,

$$p(x_{(i)}) = \begin{cases} F(x_{(i)}) - F(x_{(i-1)}) \text{ if } i \geq 2 \\ F(x_{(i)}) \text{ if } i = 1 \end{cases}$$

Probability calculations using sums and products

We can make good use of some basic probability concepts to invoke some tools for probability calculations. We know that if random events A and B are independent, then $P(A \cap B) = P(A)P(B)$. Also, if A and B are mutually exclusive, then $P(A \cup B) = P(A) + P(B)$. Take an example of 5 coin flips of an uneven coin that lands on heads with probability p. Then, the probability of 3 heads happening, $P(X = 3)$ for the random variable X which counts the number of heads in this scenario, can be calculated using these two tricks (instead of counting, which we have been doing so far). Let's do this in two steps:

1. What is the probability of a particular outcome $HHHTT$? This is simply obtained by multiplying p three times and $(1-p)$ two times. Why is this? because *within* a particular draw, those five coin flips are independent, so each single outcome's probability should be multiplied. They are not disjoint, as one event does not preclude the possibility of another event.

2. What about $HHTHT$? This outcome is *disjoint* from the first outcome $HHHTT$, or for that matter, any other outcome who is a combination of 3 H's and 2 T's. They can never happen together, so they are *disjoint* events (they are not independent, because they will never happen together – in fact, if you know that $HHTHT$ happened, then you definitely know that $HHHTT$ didn't happen). Also, we know how to count how many such events can happen – it is $\binom{5}{3}$. So, we can calculate the aggregate probability of 3 heads and 2 tails happening by adding the single event's probability $p^3(1-p)^2$ up exactly $\binom{5}{3}$ times, i.e.

$$P(\text{3 successes in 5 trials}) = \binom{5}{3} p^3(1-p)^2 \tag{14}$$

But in general, we are interested in the probability distribution of the r number of heads out of n trials.

Notation

The symbol $\sim$ is an identifier for a random variable, and specifies its probability function (p.m.f for discrete random variables). In statistical jargon, we will say that a random variable has a certain distribution to signify that it has a a certain probability function.

The Binomial distribution (outcome of n coin flips)

Consider again tossing a coin (not necessarily fair) n times in such a way that each coin flip is independent of the other coin flips. By this, we mean that if H_i denotes the event 'observing heads on the i-th toss', then $P(H_i \cap H_j) = P(H_i)P(H_j)$ for all $i \neq j$. Suppose that the probability of seeing heads on each flip is $p \in [0,1]$ (and let's call the event 'seeing heads' a *success*). Introduce the random variables

$$Y_i = \begin{cases} 1 \text{ if the i}-th \text{ flip is a success} \\ 0 \text{ otherwise} \end{cases}$$

fo $i \in \{0, 1, 2, \ldots, n\}$. The number of heads that we observe (or the number of successes in the experiment) is

$$X = \sum_{i=1}^{n} Y_i.$$

Under the conditions described above, the random variable X is distributed according to the Binomial distribution with parameters $n \in \mathcal{Z}_+$ (number of trials) and $p \in [0,1]$ (probability of success in each trial). We denote this by $X \sim \text{Binomial}(n,p)$. The p.m.f. of X is

$$p(x) = \begin{cases} \binom{n}{x} p^x (1-p)^{n-x} \text{ if } x \in \{0, 1, 2, \ldots, n\} \\ 0 \text{ if } x \notin \{0, 1, 2, \ldots, n\} \end{cases} \tag{15}$$

Lecture 9

Recommended Readings: WMS Chapter 3.1,3.1,4.1,4.2

Learning objectives

1. Students will correctly recall the definitions and properties of a probability density function (p.d.f) and the relationship to its cumulative distribution function (c.d.f).

2. Students will correctly compute integrals and derivatives in the context of probability distributions of continuous random variables.

If X is continuous, we can describe the probability distribution of X by means of the *probability density function* (p.d.f.) $f : \mathbb{R} \to \mathbb{R}_+$. Define in this case $\text{supp}(X) = \{x \in \mathbb{R} : f(x) > 0\}$. The function f satisfies the following properties:

1. $f(x) \geq 0 \quad \forall x \in \mathbb{R}$

2. $\int_{\mathbb{R}} f(x)\,dx = \int_{\text{supp}(X)} f(x)\,dx = 1$.

We use f to compute the probability of events of the type $\{X \in (a, b]\}$ for $a < b$. In particular, for $a < b$, we have

$$P(X \in (a, b]) = \int_a^b f(x)\,dx \tag{16}$$

Notice that this implies that, for any $x \in \mathbb{R}$, $P(X = x) = 0$ if X is a continuous random variable.

Also, if X is a continuous random variable, it is clear from above that

$$P(X \in (a, b]) = P(X \in [a, b]) = P(X \in [a, b)) = P(X \in (a, b)). \tag{17}$$

In general, for any set $A \subset \mathbb{R}$, we have

$$P(X \in A) = \int_A f(x)\,dx.$$

If X is a continuous random variable, its c.d.f is defined as the function

$$F(x) = f(X \leq x) = \int_{-\infty}^{x} f(y)\,dy.$$

Notice that for a continuous random variable, F is a continuous function.
What is the relationship between the c.d.f. of a continuous random variable and its p.d.f.? For a continuous random variable X,

$$f(x) = \frac{d}{dy}F(y)\bigg|_{y=x}$$

for any x at which F is differentiable.

Exercise
Consider the following p.d.f. for the random variable X:

$$f(x) = e^{-x}\mathbb{1}_{[0,\infty)}(x) = \begin{cases} e^{-x} \text{ if } x \geq 0 \\ 0 \text{ if } x < 0. \end{cases}$$

What is the support of X? Compute $P(2 < X < 3)$. Graph the p.d.f. of X.
The support of X is clearly the set $[0, \infty)$. We have

$$P(2 < X < 3) = P(X \in (2,3)) = \int_2^3 f(x)\,dx = -e^{-x}\big|_2^3 = e^{-2} - e^{-3}.$$

Furthermore, notice that for a continuous random variable X with c.d.f. F and for $a < b$ one has

$$\begin{aligned} P(a < X \leq b) &= P(a \leq X \leq b) = P(a \leq X < b) = P(a < X < b) \\ &= \int_a^b f(x)\,dx = \int_{-\infty}^b f(x)\,dx - \int_{-\infty}^a f(x)\,dx = F(b) - F(a). \end{aligned}$$

This is easy. However, for a discrete random variable one has to be careful with the bounds. Using the same notation as before, for $i < j$ one has

$$\begin{aligned} &P(x_{(i)} < X \leq x_{(j)}) = F(x_{(j)}) - F(x_{(i)}) \\ &P(x_{(i)} \leq X \leq x_{(j)}) = F(x_{(j)}) - F(x_{(i)}) + p(x_{(i)}) \\ &P(x_{(i)} \leq X < x_{(j)}) = F(x_{(j-1)}) - F(x_{(i)}) + p(x_{(i)}) \\ &P(x_{(i)} < X < x_{(j)}) = F(x_{(j-1)}) - F(x_{(i)}). \end{aligned}$$

Suppose that the c.d.f. F of a continuous random variable X is strictly increasing. Then F is invertible, meaning that there exists a function $F^{-1} : (0,1) \to \mathbb{R} \cup \{-\infty, +\infty\}$ such that for any $x \in \mathbb{R}$ we have $F^{-1}(F(x)) = x$. Then, for any $\alpha \in (0,1)$ we can define the *α-quantile* of X as the number

$$x_\alpha = F^{-1}(\alpha) \tag{18}$$

with the property that $P(X \leq x_\alpha) = \alpha$. This can be extended to c.d.f. that are not strictly increasing and to p.m.f.'s, but for this class we will not use that extension.

Exercise
Consider again a random variable X with p.d.f.

$$f(x) = e^{-x}\mathbb{1}_{[0,\infty)}(x) = \begin{cases} e^{-x} \text{ if } x \geq 0 \\ 0 \text{ if } x < 0. \end{cases}$$

Compute the α-quantile of X.

We know from above that

$$F(x) = \begin{cases} 0 \text{ if } x < 0 \\ 1 - e^{-x} \text{ if } x \geq 0. \end{cases}$$

For $\alpha \in (0,1)$, set $\alpha = F(x_\alpha) = 1 - e^{-x_\alpha}$. We then have that the α-quantile is

$$x_\alpha = -\log(1-\alpha).$$

Lecture 10

Recommended Readings: WMS Chapter 3.3, 4.3

Learning objectives

1. Students will correctly recall and apply the definitions and properties of expectation and variance.
2. Given a probability function, or a simple expression involving random variables, students will correctly distinguish between constants, parameters, observed values, and random variables.

The *expectation* (or expected value or mean) is an important operator associated to a probability distribution. Given a random variable X with p.m.f. p (if X is discrete) or p.d.f. f (if X is continuous), its expectation $E(X)$ is defined as

- $E(X) = \sum_{x \in \text{supp}(X)} xp(x)$, if X is discrete
- $E(X) = \int_{\mathbb{R}} xf(x)\,dx = \int_{x \in \text{supp}(X)} xf(x)\,dx$, if X is continuous.

Roughly speaking, $E(X)$ is the 'central tendency' of the distribution of X.

Exercise
Consider the random variable X and its p.m.f.

$$p(x) = \begin{cases} 0.2 \text{ if } x = 0 \\ 0.3 \text{ if } x = 1 \\ 0.1 \text{ if } x = 2 \\ 0.4 \text{ if } x = 3 \\ 0 \text{ if } x \notin \{0, 1, 2, 3\}. \end{cases}$$

What is $E(X)$?

By definition we have

$$E(X) = \sum_{x \in \text{supp}(X)} xp(x) = 0*0.2+1*0.3+2*0.1+3*0.4 = 0.3+0.2+1.2 = 1.7.$$

Exercise
Consider the random variable X and its p.d.f

$$f(x) = 3x^2 \mathbb{1}_{[0,1]}(x).$$

What is $E(X)$?

Again, by definition

$$E(X) = \int_{\mathbb{R}} xf(x)\,dx = \int_{\text{supp}(X)} xf(x)\,dx =$$
$$= \int_0^1 x * 3x^2\,dx = 3\int_0^1 x^3\,dx = \frac{3}{4}\,x^4\Big|_0^1 = \frac{3}{4}.$$

Consider a function $g : \mathbb{R} \to \mathbb{R}$ of the random variable X and the new random variable $g(X)$. The expectation of $g(X)$ is simply

- $E(g(X)) = \sum_{x\in\text{supp}(X)} g(x)p(x)$, if X is discrete
- $E(g(X)) = \int_{\mathbb{R}} g(x)f(x)\,dx = \int_{x\in\text{supp}(X)} g(x)f(x)\,dx$, if X is continuous.

Exercise
Consider once again the two random variables above and the function $g(x) = x^2$. What is $E(X^2)$?

- In the discrete example, we have

$$E(X^2) = \sum_{x\in\text{supp}(X)} x^2p(x) = 0^2{*}0.2{+}1^2{*}0.3{+}2^2{*}0.1{+}3^2{*}0.4 = 0.3{+}0.4{+}3.6 = 4.3.$$

- In the continous example, we have

$$E(X^2) = \int_{\mathbb{R}} x^2f(x)\,dx = \int_{\text{supp}(X)} x^2f(x)\,dx =$$
$$= \int_0^1 x^2 * 3x^2\,dx = 3\int_0^1 x^4\,dx = \frac{3}{5}\,x^5\Big|_0^1 = \frac{3}{5}.$$

A technical note: for a random variable X, the expected value $E(X)$ is a well-defined quantity whenever $E(|X|) < +\infty$. However, the opposite is not true.

The expected value operator E is a linear operator: given two random variables X, Y and scalars $a, b \in \mathbb{R}$ we have

$$\begin{aligned} E(aX) &= aE(x) \\ E(X+Y) &= E(X) + E(Y). \end{aligned} \tag{19}$$

For any scalar $a \in \mathbb{R}$, we have $E(a) = a$. To see this, consider the random variable Y with p.m.f.

$$p(y) = \mathbb{1}_{\{a\}}(x) = \begin{cases} 1 \text{ if } x = a \\ 0 \text{ if } x \neq a. \end{cases}$$

From the definition of $E(Y)$ it is clear that $E(Y) = a$.

Exercise
Consider again the continuous random variable X of the previous example. What is $E(X + X^2)$?

We have $E(X + X^2) = E(X) + E(X^2) = \frac{3}{4} + \frac{3}{5} = \frac{27}{20}$.

Beware that in general, for two random variables X, Y, it is not true that $E(XY) = E(X)E(Y)$. However, we shall see later in the class that this is true when X and Y are independent.
Remark: If X and Y are independent, then $f(x)$ and $g(Y)$ are independent. However, the opposite is not true.

The letter μ is often used to denote the expected value $E(X)$ of a random variable X, i.e. $\mu = E(X)$.
Another very important operator associated to a probability distribution is the *variance*. The variance measures how 'spread' the distribution of a random variable is. The variance of a random variable X is defined as

$$V(X) = E[(X - \mu)^2] = E(X^2) - \mu^2. \tag{20}$$

Equivalently, one can write

- if X is discrete:

$$V(X) = \sum_{x \in \text{supp}(X)} (x - \mu)^2 p(x) = \sum_{x \in \text{supp}(X)} x^2 p(x) - \mu^2$$

- if X is continuous:

$$V(X) = \int_{x \in \text{supp}(X)} x^2 f(x)\, dx - \mu^2.$$

Exercise in class
Consider the random variables of the previous examples. What is $V(X)$?

- In the discrete example, $V(X) = E(X^2) - [E(X)]^2 = 4.3 - (1.7)^2 = 4.3 - 2.89 = 1.41$.
- In the continuous example $V(X) = E(X^2) - [E(X)]^2 = 3/5 - (3/4)^2 = 3/5 - 9/16 = 94/80 = 47/40$.

σ^2 is often used to denote the variance $V(X)$ of a random variable X, i.e. $\sigma^2 = V(X)$. The variance of a random variable X is finite as soon as $E(X^2) < \infty$.
Here are two important properties of the variance operator. For any $a \in \mathbb{R}$ we have

- $V(aX) = a^2 V(X)$
- $V(a + X) = V(X)$.

Exercise
Consider the continuous random variable of the previous examples. What is $V\left(\sqrt{40/47}X\right)$?

We have

$$V\left(\sqrt{40/47}X\right) = \frac{40}{47}V(X) = \frac{40}{47}\frac{47}{40} = 1.$$

Beware that in general, for two random variables X, Y, it is not true that $V(X+Y) = V(X) + V(Y)$. However, we shall see later in the course that this is true when X and Y are independent.

Lecture 11

Recommended Readings: WMS Chapter 3.4, 3.11, 4.4, 4.5
Learning objectives

1. Students will learn the basic properties of Gaussian distributed random variables.

2. Students will use Chebyshev's Inequality theorem to draw conclusions about a random variable.

Notation

The symbol $\sim$ is an identifier for a random variable, and specifies its probability function. In statistical jargon, we will say that a random variable has a certain distribution to signify that it has a a certain pmf/pdf.

Binomial distribution in relationship to Bernoulli distribution

Suppose X is distributed according to the Binomial distribution with parameters $n \in \mathcal{Z}_+$ (number of trials) and $p \in [0,1]$ (probability of success in each trial). We denote this by $X \sim \text{Binomial}(n,p)$. The p.m.f. of X is

$$p(x) = \begin{cases} \binom{n}{x} p^x (1-p)^{n-x} \text{ if } x \in \{0,1,2,\ldots,n\} \\ 0 \text{ if } x \notin \{0,1,2,\ldots,n\} \end{cases}$$

Its expectation is $E(X) = np$ and its variance is $V(X) = np(1-p)$.

In particular, when $n = 1$, we usually say that X is distributed according to the Bernoulli distribution of parameter p, denoted $X \sim \text{Bernoulli}(p)$. In this case $E(X) = p$ and $V(X) = p(1-p)$. Every Y_i is a Bernoulli distributed random variable for the case of a binomial random variable $X = \sum_{i=1}^{n} Y_i$.

The Normal (Gaussian) distribution

The Normal family: this is a class of distributions that is commonly used to describe physical phenomena where the quantity of interest takes values (at least in principle) in the range $(-\infty, \infty)$. The Normal distribution is one of the most relevant distributions for applications in Statistics. We say that a random variable X has a Normal distribution with parameters $\mu \in \mathbb{R}$ and $\sigma^2 \in \mathbb{R}_+$ if its p.d.f. is

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}. \tag{21}$$

In this case, we write $X \sim \mathcal{N}(\mu, \sigma)$. The parameters of X correspond to its expectation $\mu = E(X)$ and its variance $\sigma^2 = V(X)$. The c.d.f. of $X \sim \mathcal{N}(\mu, \sigma^2)$

can be expressed in terms of the *error function* $\text{erf}(\cdot)$. However, for our purposes, we will not need to investigate this further.

The Standardized Normal Distribution
Given $X \sim \mathcal{N}(\mu, \sigma^2)$, we can always obtain a 'standardized' version Z of X such that Z still has a Normal distribution, $E(Z) = 0$, and $V(Z) = 1$ (i.e. $Z \sim \mathcal{N}(0, 1)$). This can be done by means of the transformation

$$Z = \frac{X - \mu}{\sigma}. \tag{22}$$

The random variable Z is said to be *standardized*. Of course, one can also standardize random variables that have other distributions (as long as they have finite variance), but unlike the Normal case the resulting standardized variable may not belong anymore to the same family to which the original random variable X belonged to.

The Normal family is *closed* with respect to translation and scaling:
if $X \sim \mathcal{N}(\mu, \sigma^2)$, then $aX + b \sim \mathcal{N}(a\mu + b, a^2\sigma^2)$ for $a \neq 0$ and $b \in \mathbb{R}$.
We finally mention that it is common notation to indicate the c.d.f. of $Z \sim \mathcal{N}(0, 1)$ by $\Phi(\cdot)$. Notice that $\Phi(-x) = 1 - \Phi(x)$ for any $x \in \mathbb{R}$.

Markov's Inequality and Chebyshev's Inequality

Sometimes we may want to compute or approximate the probability of certain events involving a random variable X even if we don't know its distribution (but given that we know its expectation or its variance). Let $a > 0$. The following inequalities are useful in these cases:

$$\begin{aligned} P(|X| \geq a) &\leq \frac{E(|X|)}{a} \text{ (Markov's inequality)} \\ P(|X - \mu| \geq a) &\leq \frac{V(X)}{a^2} \text{ (Chebyshev's inequality)} \end{aligned} \tag{23}$$

There are some conditions! The Markov's inequality can be used if X is a positive random variable. The Chebyshev's Inequality can be used for any random variable with finite (literally, not infinite) expected value. Let $\sigma^2 = V(X)$ as usual so that σ is the standard deviation of X. Note that we can conveniently take $a = k\sigma$ and the second inequality then reads

$$P(|X - \mu| \geq k\sigma) \leq \frac{\sigma^2}{k\sigma^2} = \frac{1}{k^2} \tag{24}$$

where $\mu = E(X)$. Thus, we can easily bound the probability that X deviates from its expectation by more than k times its standard deviation.

Let's first prove Markov's inequality.

$$aE[\mathbb{1}(|X| \geq a)] = E[a\mathbb{1}(|X| \geq a)] = \begin{cases} 0 \text{ if } |X| < a \\ a \text{ if } |X| \geq a \end{cases}$$

therefore we can upper bound it by $E[|X|]$ and the proof is completed.
Chebyshev's inequality can be proved via Markov's. Indeed,

$$P(|X - \mu| \geq a) = P\left(\frac{|X - \mu|}{a} \geq 1\right) = P\left(\frac{|X - \mu|^2}{a^2} \geq 1\right) \leq \frac{E[|X - \mu|^2]}{a^2}$$

where the last inequality is thanks to Markov's.

Exercise
A call center receives an average of 10,000 phone calls a day, with a standard deviation of $\sqrt{(}2000)$. What is the probability that that there will be more than 15,000 calls ?
Call X the number of phone calls (a random variable) with

1. Using Markov's inequality, we knot that

   $$P(X \geq 15,000) \leq \frac{E(X)}{15,000} = 2/3$$

   This is quick and easy, but we can do better

2. Using Chebyshev's, we get

   $$P(X \geq 15,000) = P(X - 10,000 \geq 5,000) \leq P(|X - 10,000| \geq 5,000) \leq \frac{2,000}{5,000^2} = 0.000008.$$

   This is *much* better than the previous result.

Lecture 12

Recommended Readings: WMS Chapter 3.5, 3.8
Learning objectives

1. Students will be able to correctly describe the random variables and parameters of the distributions covered so far.
2. Students will learn the basic properties of Geometric, and Poisson distributed random variables.

Geometric distribution (# times to 1st success)

Consider counting the number of coin flips needed before the first success is observed and let X denote the corresponding random variable. Then X has the Geometric distribution with parameter $p \in [0,1]$, denoted $X \sim \text{Geometric}(p)$. Its p.m.f. is given by

$$p(x) = \begin{cases} (1-p)^{x-1}p \text{ if } x \in \{1,2,3,\dots\} \\ 0 \text{ if } x \notin \{1,2,3,\dots\} \end{cases} \tag{25}$$

The expected value of X is $E(X) = 1/p$, while its variance is $V(X) = (1-p)/p^2$. The Geometric distribution is one of the distribution that have the so-called 'memoryless property'. This means that if $X \sim \text{Geometric}(p)$, then

$$P(X > x + y | X > x) = P(X > y)$$

for any $0 < x \leq y$. To see this, let's first compute the c.d.f. of X. We have

$$F(x) = P(X \leq x) = \begin{cases} 0 \text{ if } x < 1 \\ p\sum_{y=1}^{\lfloor x \rfloor}(1-p)^{y-1} \text{ if } x \geq 1 \end{cases} = \begin{cases} 0 \text{ if } x < 1 \\ p\sum_{y=0}^{\lfloor x \rfloor - 1}(1-p)^{y} \text{ if } x \geq 1 \end{cases}$$
$$= \begin{cases} 0 \text{ if } x < 1 \\ p\frac{1-(1-p)^{\lfloor x \rfloor}}{1-(1-p)} \text{ if } x \geq 1 \end{cases} = \begin{cases} 0 \text{ if } x < 1 \\ 1-(1-p)^{\lfloor x \rfloor} \text{ if } x \geq 1. \end{cases}$$

Let's look for example to the case $1 < x < y$. We have

$$P(X > x + y | X > x) = \frac{P(X > x+y)}{P(X > x)} = \frac{1 - F(x+y)}{1 - F(x)} = \frac{(1-p)^{\lfloor x+y \rfloor}}{(1-p)^{\lfloor x \rfloor}}$$
$$= (1-p)^{\lfloor y \rfloor} = P(X > y).$$

Exercise
Suppose you roll a fair die. What is the probability that the first 6 is rolled on the 4-th roll?
Let X denote the random variable counting the number of rolls needed to observe the first 6. We have $X \sim \text{Geometric}(p)$ with $p = 1/6$. Then,

$$P(X = 4) = p(4) = p(1-p)^{4-1} = \frac{1}{6}\frac{5^3}{6^3} = \frac{5^3}{6^4} \approx 0.096.$$

Poisson distribution (counts of rare events)

The Poisson distribution can be thought of as a limiting case of the Binomial distribution. Consider the following example: you own a store and you want to model the number of people who enter in your store on a given day. In any time interval of that day, the number of people walking in the store is certainly discrete, but in principle that number can be any non-negative integer. We could try and divide the day into n smaller subperiods in such a way that, as $n \to \infty$, only one person can walk into the store in any given subperiod. If we let $n \to \infty$, it is clear however that the probability p that a person will walk in the store in an infinitesimally small subperiod of time is such that $p \to 0$. The Poisson distribution arises as a limiting case of the Binomial distribution when $n \to \infty$, $p \to 0$, and $np \to \lambda \in (0, \infty)$.
The p.m.f. of a random variable X that has the Poisson distribution with parameter $\lambda > 0$, denoted $X \sim \text{Poisson}(\lambda)$ is

$$p(x) = \begin{cases} e^{-\lambda}\frac{\lambda^x}{x!} \text{ if } x \in \{0, 1, 2, 3, \dots\} \\ 0 \text{ if } x \notin \{0, 1, 2, 3, \dots\} \end{cases} \tag{26}$$

Both the expected value and the variance of X are equal to λ,
$E(X) = V(X) = \lambda$.

Exercise
At the CMU USPS office, the expected number of students waiting in line between 1PM and 2PM is 3. What is the probability that you will see more than 2 students already in line in front of you, if you go to the USPS office in that period of time?
Let $X \sim \text{Poisson}(\lambda)$ with $\lambda = 3$ be the number of students in line when you enter into the store. We want to compute $P(X > 2)$. We have

$$P(X > 2) = 1 - P(X \leq 2) = 1 - \sum_{x=0}^{2} p(x) = 1 - e^{-3}\sum_{x=0}^{2}\frac{3^x}{x!} =$$
$$= 1 - e^{-3}\left(1 + 3 + \frac{9}{2}\right) = 1 - e^{-3}\frac{17}{2} \approx 0.577.$$

Lecture 13

Exam review

Lecture 14

Recommended Readings: WMS Chapter 4.6, 4.7, 5.4
Learning objectives

1. Students will begin to familiarize themselves with the families of Gamma and Beta probability distributions and summarize their utility in one sentence.

2. Given an un-normalized distribution, students will be able to identify its normalization constant either by integration or by kernel pattern matching to a known distribution.

3. Students will learn to construct joint probability distributions for independent random variables.

Summary:

1. the Gamma family: this is a class of distributions which are frequently used to describe physical phenomena where the quantity of interest takes non-negative values, i.e. it takes values in the interval $[0, \infty)$

2. the Beta family: this is a class of distributions that is commonly used to describe physical phenomena where the quantity of interest takes values in some interval $[a, b]$ of the real line.

The Gamma Distribution

We say that the random variable X has a Gamma distribution with parameters $\alpha, \beta > 0$ if its p.d.f. is

$$f(x) = \frac{1}{\beta^\alpha \Gamma(\alpha)} x^{\alpha-1} e^{-\frac{x}{\beta}} \mathbb{1}_{[0,\infty)}(x). \tag{27}$$

Notice that the p.d.f. of the Gamma distribution includes the Gamma *function*

$$\Gamma(\alpha) = \int_0^\infty x^{\alpha-1} e^{-x}\, dx \tag{28}$$

for $\alpha > 0$. Useful properties of the Gamma function:

- For $\alpha \in \mathbb{Z}_+$, we have that $\Gamma(\alpha) = (\alpha - 1)!$.
- For any $\alpha > 0$, we have $\Gamma(\alpha + 1) = \alpha\Gamma(\alpha)$. [4]

Make sure not to confuse the Gamma distribution, described by the p.d.f. of equation (27), and the Gamma function of equation (28).
The expectation and the variance of X are respectively $E(X) = \alpha\beta$ and $V(X) = \alpha\beta^2$.

[4] (This recursive property often comes in handy for computations that involve Gamma-distributed random variables.)

The c.d.f. of a Gamma-distributed random variable can be expressed explicitly in terms of the *incomplete Gamma function*. Once again, for our purposes, we don't need to investigate this further.
We saw that the Normal family is closed with respect to translation and scaling. The Gamma family is closed with respect to positive scaling only.
If $X \sim \text{Gamma}(\alpha, \beta)$, then $cX \sim \text{Gamma}(\alpha, c\beta)$, provided that $c > 0$.

The Exponential Distribution

The Exponential distribution constitutes a subfamily of the Gamma distribution. In particular, X is said to have an Exponential distribution with parameter $\beta > 0$ if $X \sim \text{Gamma}(1, \beta)$. In that case, we write $X \sim \text{Exponential}(\beta)$.
The p.d.f. of X is therefore

$$f(x) = \frac{1}{\beta} e^{-\frac{x}{\beta}} \mathbb{1}_{[0,\infty)}(x).$$

Because the Exponential distribution is a subfamily of the Gamma distribution, we have $E(X) = \alpha\beta = \beta$ and $V(X) = \alpha\beta^2 = \beta^2$.

We will learn a bit more about this when we revisit Poisson processes and exponential wait times of events, but we note here that the exponential distribution also has the property of being 'memoryless'. For an exponential random variable X with parameter λ, the following hold

$$P(X > t + s | X > t) = P(X > s) \tag{29}$$

or, equivalently

$$P(X > t + s) = P(X > s)P(X > t) \tag{30}$$

Exercise
Prove this. Hint: consider using 1-CDF of exponential distributions.

The Beta distribution

Suppose that you are interested in studying a quantity Y that can only take values in the interval $[a, b] \subset \mathbb{R}$. We can easily transform Y in such a way that it can only take values in the standard unit interval $[0, 1]$: it is enough to consider the normalized version of Y

$$X = \frac{Y - a}{b - a}.$$

Then, a flexible family of distributions that can be used to model Y is the Beta family. We say that a random variable X has a Beta distribution with parameters $\alpha, \beta > 0$, denoted $X \sim \text{Beta}(\alpha, \beta)$, if its p.d.f. is of the form

$$f(x) = \frac{x^{\alpha-1}(1-x)^{\beta-1}}{B(\alpha, \beta)} \mathbb{1}_{[0,1]}(x). \tag{31}$$

where

$$B(\alpha,\beta)=\int_0^1 x^{\alpha-1}(1-x)^{\beta-1}\,dx=\frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}.$$

The c.d.f. of X can be expressed explicitly in terms of the *incomplete Beta function*

$$B(x;\alpha,\beta)=\int_0^x y^{\alpha-1}(1-y)^{\beta-1}\,dx,$$

but we don't need to investigate this further for our purposes.
The expected value of X is $E(X)=\frac{\alpha}{\alpha+\beta}$ and its variance is $V(X)=\frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}$.

A Note on the Normalizing Constant of a Probability Density Function

Most frequently, a given p.d.f. f takes the form

$$f(x)=cg(x)$$

where c is a positive constant and g is a function. The part of f depending on x, i.e. the function g, is usually called the *kernel* of the p.d.f. f. Very often one can guess whether f belongs to a certain family by simply inspecting g. Then, if f is indeed a density, $c>0$ is exactly the 'right' constant which makes f integrate to 1. Therefore, if for any reason c is unknown, but you can guess that $X\sim f$ belongs to a certain family of distributions for a particular value of its parameters, in order to figure out c one does not necessarily have to compute

$$c=\left(\int_{\mathrm{supp}(X)} g(x)\,dx\right)^{-1}.$$

Let us illustrate this by means of two examples:

- Let $f(x)=ce^{-\frac{x}{2}}\mathbb{1}_{[0,\infty)}(x)$ and we want to figure out what c is. Here $g(x)=e^{-\frac{x}{2}}\mathbb{1}_{[0,\infty)}(x)$ is the kernel of an Exponential p.d.f. with parameter $\beta=2$. We know therefore that c must be equal to $c=1/\beta$.
- Let $f(x)=cx^4(1-x)^5\mathbb{1}_{[0,1]}(x)$ and again suppose that we want to figure out the value of c. Here $g(x)=x^4(1-x)^5\mathbb{1}_{[0,1]}(x)$ which is the kernel of a Beta p.d.f. with parameters $\alpha=5$ and $\beta=6$. We know therefore that c must be the number

$$c=\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}=\frac{\Gamma(5+6)}{\Gamma(5)\Gamma(6)}=\frac{10!}{4!5!}=1260.$$

Independent Random Variables

Consider a collection of n random variables $X_1,\ldots,X_n$ and their *joint probability distribution*. Their joint probability distribution can be described in terms of the joint c.d.f.

$$F_{X_1,\ldots,X_n}(x_1,x_2,\ldots,x_n)=P(X_1\leq x_1\cap X_2\leq x_2\cap\cdots\cap X_n\leq x_n),\qquad(32)$$

in terms of the joint p.m.f. (if the random variables are all discrete)

$$p_{X_1,\dots,X_n}(x_1, x_2, \dots, x_n) = P(X_1 = x_1 \cap X_2 = x_2 \cap \dots \cap X_n = x_n), \qquad (33)$$

or in terms of the joint p.d.f. $f_{X_1,\dots,X_n}$ (if the random variables are all continuous).
The random variables $X_1, \dots, X_n$ are said to be independent if either of the following holds:

- $F_{X_1,\dots,X_n}(x_1, x_2, \dots, x_n) = \prod_{i=1}^{n} F_{X_i}(x_i)$
- (discrete case) $p_{X_1,\dots,X_n}(x_1, x_2, \dots, x_n) = \prod_{i=1}^{n} p_{X_i}(x_i)$
- (continuous case) $f_{X_1,\dots,X_n}(x_1, x_2, \dots, x_n) = \prod_{i=1}^{n} f_{X_i}(x_i)$.

Then, if we consider an arbitrary collection of events $\{X_1 \in A_1\}, \{X_2 \in A_2\}, \dots, \{X_n \in A_n\}$, we have that

$$P(X_1 \in A_1 \cap X_2 \in A_2 \cap \dots \cap X_n \in A_n) = \prod_{i=1}^{n} P(X_i \in A_i).$$

If the random variables also share the same *marginal distribution*, i.e. we have

- $F_{X_i} = F \quad \forall i \in \{1, \dots, n\}$
- $p_{X_i} = p \quad \forall i \in \{1, \dots, n\}$ (if the random variables are all discrete)
- $f_{X_i} = f \quad \forall i \in \{1, \dots, n\}$ (if the random variables are all continuous)

then the random variables $X_1, \dots, X_n$ are said to be *independent and identically distributed*, usually shortened in *i.i.d.*.

Lecture 15

Recommended Readings: WMS Chapter 5.1, 5.2, 5.4
Learning objectives

1. Students will correctly formulate pairs of random variables to represent a simple discrete sample space.
2. Given a pair of random variables, students will be able to compute probability of events using probability theory and summation/integration techniques.

Multivariate Probability Distributions

So far we have focused on *univariate* probability distributions, i.e. probability distributions for a single random variable. However, when we discussed independence of random variables in Lecture 14, we introduced the notion of joint c.d.f., joint p.m.f., and joint p.d.f. for a collection of n random variables $X_1, \ldots, X_n$. In this lecture we will elaborate more on these objects.
Let $X_1, \ldots, X_n$ be a collection of n random variables.

- Regardless of whether they are discrete or continuous, we denote by $F_{X_1,\ldots,X_n}$ their joint c.d.f., i.e. the function

$$F_{X_1,\ldots,X_n}(x_1, \ldots, x_n) = P(X_1 \leq x_1 \cap \cdots \cap X_n \leq x_n).$$

- If they are all discrete, we denote by $p_{X_1,\ldots,X_n}$ their joint p.m.f., i.e. the function

$$p_{X_1,\ldots,X_n}(x_1, \ldots, x_n) = P(X_1 = x_1 \cap \cdots \cap X_n = x_n).$$

- If they are all continuous, we denote by $f_{X_1,\ldots,X_n}$ their joint p.d.f..

The above functions satisfy properties that are similar to those satisfied by their univariate counterparts.

- The joint c.d.f. $F_{X_1,\ldots,X_n}$ satisfies:
  - $F_{X_1,\ldots,X_n}(x_1, \ldots, x_n) \in [0, 1]$ for any $x_1, \ldots, x_n \in \mathbb{R}$.
  - $F_{X_1,\ldots,X_n}$ is monotonically non-decreasing in each of its variables
  - $F_{X_1,\ldots,X_n}$ is càdlàg (right-continuous with left limits with respect to every variable)
  - $\lim_{x_i \to -\infty} F_{X_1,\ldots,X_n}(x_1, \ldots, x_i, \ldots, x_n) = 0$ for any $i \in \{1, \ldots, n\}$
  - $\lim_{x_1 \to +\infty, \ldots, x_n \to +\infty} F_{X_1,\ldots,X_n}(x_1, \ldots, x_n) = 1$
- The joint p.m.f. satisfies:
  - $p_{X_1,\ldots,X_n}(x_1, \ldots, x_n) \in [0, 1]$ for any $x_1, \ldots, x_n \in \mathbb{R}$.

- $\sum_{x_1 \in \text{supp}(X_1)} \cdots \sum_{x_n \in \text{supp}(X_n)} p_{X_1,\ldots,X_n}(x_1,\ldots,x_n) = 1.$

- The joint p.d.f. satisfies:
  - $f_{X_1,\ldots,X_n}(x_1,\ldots,x_n) \geq 0$ for any $x_1,\ldots,x_n \in \mathbb{R}$.
  - $\int_{\mathbb{R}} \cdots \int_{\mathbb{R}} f_{X_1,\ldots,X_n}(x_1,\ldots,x_n)\, dx_1 \ldots dx_n$
    $= \int_{\text{supp}(X_1)} \cdots \int_{\text{supp}(X_n)} f_{X_1,\ldots,X_n}(x_1,\ldots,x_n)\, dx_n \ldots dx_1 = 1.$

Furthermore we have

$$F_{X_1,\ldots,X_n}(x_1,\ldots,x_n) = \sum_{\substack{y_1 \leq x_1 \\ y_1 \in \text{supp}(X_1)}} \cdots \sum_{\substack{y_n \leq x_n \\ y_n \in \text{supp}(X_n)}} p(y_1,\ldots,y_n)$$

and

$$F_{X_1,\ldots,X_n}(x_1,\ldots,x_n) = \int_{-\infty}^{x_1} \cdots \int_{-\infty}^{x_n} f_{X_1,\ldots,X_n}(y_1,\ldots,y_n)\, dy_n \ldots dy_1$$

Exercise
You are given the following bivariate p.m.f.:

$$p_{X_1,X_2}(x_1,x_2) = \begin{cases} \frac{1}{8} \text{ if } (x_1,x_2) = (0,-1) \\ \frac{1}{4} \text{ if } (x_1,x_2) = (0,0) \\ \frac{1}{8} \text{ if } (x_1,x_2) = (0,1) \\ \frac{1}{4} \text{ if } (x_1,x_2) = (2,-1) \\ \frac{1}{4} \text{ if } (x_1,x_2) = (2,0) \\ 0 \text{ otherwise.} \end{cases}$$

What is $P(X_1 \leq 1 \cap X_2 \leq 0) = F_{X_1,X_2}(1,0)$? We have

$$F_{X_1,X_2}(1,0) = p_{X_1,X_2}(0,-1) + p_{X_1,X_2}(0,0) = \frac{1}{8} + \frac{1}{4} = \frac{3}{8}.$$

Practice drawing this p.m.f

Exercise
You are given the following bivariate p.d.f.:

$$f_{X_1,X_2}(x_1,x_2) = e^{-(x_1+x_2)} \mathbb{1}_{[0,\infty)\times[0,\infty)}(x_1,x_2)$$

Practice drawing this p.d.f.

- What is $P(X_1 \leq 1 \cap X_2 > 5)$?
- What is $P(X_1 + X_2 \leq 3)$?

We have

$$\begin{aligned} P(X_1 \leq 1 \cap X_2 > 5) &= \int_{\infty}^{1} \int_{5}^{\infty} f_{X_1,X_2}(x_1, x_2)\, dx_1 dx_2 = \int_{0}^{1} \int_{5}^{\infty} e^{-(x_1+x_2)}\, dx_1 dx_2 \\ &= \int_{0}^{1} e^{-x_1}\, dx_1 \int_{5}^{\infty} e^{-x_2}\, dx_2 = \left(-\, e^{-x_1}\big|_0^1\right)\left(-\, e^{-x_2}\big|_5^{\infty}\right) \\ &= \left(1 - e^{-1}\right) e^{-5}. \end{aligned}$$

and

$$\begin{aligned} P(X_1 + X_2 \leq 3) &= \int_{0}^{3} \int_{0}^{3-x_1} e^{-(x_1+x_2)}\, dx_2 dx_1 = \int_{0}^{3} e^{-x_1} \int_{0}^{3-x_1} e^{-x_2}\, dx_2 dx_1 \\ &= \int_{0}^{3} e^{-x_1} \left(-e^{-x_2}\big|_0^{3-x_1}\right) dx_1 = \int_{0}^{3} e^{-x_1} \left(1 - e^{x_1-3}\right) dx_1 \\ &= \int_{0}^{3} \left(e^{-x_1} - e^{-3}\right) dx_1 = \left(-\, e^{-x_1}\big|_0^3\right) - 3e^{-3} = 1 - 4e^{-3}. \end{aligned}$$

Lecture 16

Recommended Readings: WMS Chapter 5.3
Learning objectives

1. Students will correctly obtain marginal distributions from joint p.m.f's or p.d.f's.

Marginal Distributions

Given a collection of random variables $X_1, \ldots, X_n$ and their joint distribution, how can we derive the *marginal* distribution of only one of them, say X_i? The idea is summing or integrating the joint distribution over all the variables except for X_i.

Practice marginalizing a bivariate pmf from a table.

Thus, given $p_{X_1,\ldots,X_n}$ we have that

$$p_{X_i}(x_i) = \sum_{y_1 \in \text{supp}(X_1)} \cdots \sum_{y_{i-1} \in \text{supp}(X_{i-1})} \sum_{y_{i+1} \in \text{supp}(X_{i+1})} \cdots \sum_{y_n \in \text{supp}(X_n)} p_{X_1,\ldots,X_n}(x_1, \ldots, x_{i-1}, x_i, x_{i+1}, \ldots, x_n).$$

and given $f_{X_1,\ldots,X_n}$ we have

$$f_{X_i}(x_i) = \int_{\mathbb{R}} \cdots \int_{\mathbb{R}} \int_{\mathbb{R}} \cdots \int_{\mathbb{R}} f_{X_1,\ldots,X_n}(x_1, \ldots, x_{i-1}, x_i, x_{i+1}, \ldots, x_n)\, dx_n \ldots dx_{i+1} dx_{i-1} \ldots dx_1$$
$$= \int_{\text{supp}(X_1)} \cdots \int_{\text{supp}(X_{i-1})} \int_{\text{supp}(X_{i+1})} \cdots \int_{\text{supp}(X_n)} f_{X_1,\ldots,X_n}(x_1, \ldots, x_{i-1}, x_i, x_{i+1}, \ldots, x_n)$$
$$dx_n \ldots dx_{i+1} dx_{i-1} \ldots dx_1.$$

Exercise
Consider again the bivariate p.m.f.

$$p_{X_1,X_2}(x_1, x_2) = \begin{cases} \frac{1}{8} \text{ if } (x_1, x_2) = (0, -1) \\ \frac{1}{4} \text{ if } (x_1, x_2) = (0, 0) \\ \frac{1}{8} \text{ if } (x_1, x_2) = (0, 1) \\ \frac{1}{4} \text{ if } (x_1, x_2) = (2, -1) \\ \frac{1}{4} \text{ if } (x_1, x_2) = (2, 0) \\ 0 \text{ otherwise.} \end{cases}$$

Derive the marginal p.m.f. of X_2.

Notice first that $\text{supp}(X_2) = \{-1, 0, 1\}$. We have

$$p_{X_2}(-1) = p_{X_1,X_2}(0,-1) + p_{X_1,X_2}(2,-1) = \frac{1}{8} + \frac{1}{4} = \frac{3}{8}$$
$$p_{X_2}(0) = p_{X_1,X_2}(0,0) + p_{X_1,X_2}(2,0) = \frac{1}{4} + \frac{1}{4} = \frac{1}{2}$$
$$p_{X_2}(1) = p_{X_1,X_2}(0,1) = \frac{1}{8}.$$

Thus,

$$p_{X_2}(x_2) = \begin{cases} \frac{3}{8} \text{ if } x_2 = -1 \\ \frac{1}{2} \text{ if } x_2 = 0 \\ \frac{1}{8} \text{ if } x_2 = 1 \\ 0 \text{ if } x_2 \notin \{-1, 0, 1\} \end{cases}$$

Exercise
Consider again the bivariate p.d.f.

$$f_{X_1,X_2}(x_1, x_2) = e^{-(x_1+x_2)} \mathbb{1}_{[0,\infty)\times[0,\infty)}(x_1, x_2).$$

Derive the marginal p.d.f. of X_1.
Notice first that $\text{supp}(X_1) = [0, \infty)$. For $x_1 \in [0, \infty)$ we have

$$f_{X_1}(x_1) = \int_{\mathbb{R}} f_{X_1,X_2}(x_1, x_2)\, dx_2 = e^{-x_1} \int_0^\infty e^{-x_2}\, dx_2 = e^{-x_1}.$$

Thus,

$$f_{X_1}(x_1) = e^{-x_1} \mathbb{1}_{[0,\infty)}(x_1).$$

Lecture 17

Recommended Readings: WMS Chapter 5.3, 5.4
Learning objectives

1. Students will correctly obtain conditional distributions from joint p.m.f's or p.d.f's.

Conditional Distributions

We will limit our discussion here at the bivariate case, but all that follows is easily extended to more than two random variables.
Suppose that we are given a pair of random variables X_1, X_2 and we want to compute probabilities of the type

$$P(X_1 \in A_1 | X_2 = x_2)$$

for a particular fixed value x_2 of X_2. To do this, we need either the conditional p.m.f. (if X_1 is discrete) or the conditional p.d.f. (if X_1 is continuous) of X_1 given $X_2 = x_2$. By definition, we have

$$p_{X_1|X_2=x_2}(x_1) = \frac{p_{X_1,X_2}(x_1, x_2)}{p_{X_2}(x_2)}$$
$$f_{X_1|X_2=x_2}(x_1) = \frac{f_{X_1,X_2}(x_1, x_2)}{f_{X_2}(x_2)}.$$

Notice the two following impotant facts:

1. $p_{X_1|X_2=x_2}$ and $f_{X_1|X_2=x_2}$ are not well-defined if x_2 is such that $p_{X_2}(x_2) = 0$ and $f_{X_2}(x_2) = 0$ respectively (i.e. $x_2 \notin \text{supp}(X_2)$)
2. given $x_2 \in \text{supp}(X_2)$, $p_{X_1|X_2=x_2}$ and $f_{X_1|X_2=x_2}$ are null whenever $p_{X_1,X_2}(x_1, x_2) = 0$ and $f_{X_1,X_2}(x_1, x_2) = 0$ respectively.

So whenever you are computing a conditional distribution, it is good practice to 1) determine the support of the conditioning variable X_2 and clearly state that the conditional distribution that you are about to compute is only well-defined for $x_2 \in \text{supp}(X_2)$ and 2) given $x_2 \in \text{supp}(X_2)$, clarify for which values of $x_1 \in \mathbb{R}$ the conditional distribution $p_{X_1|X_2=x_2}$ and $f_{X_1|X_2=x_2}$ is null.

Exercise:
Consider again the bivariate p.m.f.

$$p_{X_1,X_2}(x_1, x_2) = \begin{cases} \frac{1}{8} \text{ if } (x_1, x_2) = (0, -1) \\ \frac{1}{4} \text{ if } (x_1, x_2) = (0, 0) \\ \frac{1}{8} \text{ if } (x_1, x_2) = (0, 1) \\ \frac{1}{4} \text{ if } (x_1, x_2) = (2, -1) \\ \frac{1}{4} \text{ if } (x_1, x_2) = (2, 0) \\ 0 \text{ otherwise.} \end{cases}$$

Derive the conditional distribution of X_1 given X_2.
First of all notice that the conditional p.m.f. of X_1 given $X_2 = x_2$ is only well-defined for $x_2 \in \text{supp}(X_2) = \{-1, 0, 1\}$. Then we have

$$p_{X_1|X_2=-1}(x_1) = \begin{cases} \frac{p_{X_1,X_2}(0,-1)}{p_{X_2}(-1)} \text{ if } x_1 = 0 \\ \frac{p_{X_1,X_2}(2,-1)}{p_{X_2}(-1)} \text{ if } x_1 = 2 \\ 0 \text{ if } x_1 \notin \{0,2\} \end{cases} = \begin{cases} \frac{1/8}{3/8} \text{ if } x_1 = 0 \\ \frac{1/4}{3/8} \text{ if } x_1 = 2 \\ 0 \text{ if } x_1 \notin \{0,2\} \end{cases}$$
$$= \begin{cases} \frac{1}{3} \text{ if } x_1 = 0 \\ \frac{2}{3} \text{ if } x_1 = 2 \\ 0 \text{ if } x_1 \notin \{0,2\} \end{cases}$$

$$p_{X_1|X_2=0}(x_1) = \begin{cases} \frac{p_{X_1,X_2}(0,0)}{p_{X_2}(0)} \text{ if } x_1 = 0 \\ \frac{p_{X_1,X_2}(2,0)}{p_{X_2}(0)} \text{ if } x_1 = 2 \\ 0 \text{ if } x_1 \notin \{0,2\} \end{cases} = \begin{cases} \frac{1/4}{1/2} \text{ if } x_1 = 0 \\ \frac{1/4}{1/2} \text{ if } x_1 = 2 \\ 0 \text{ if } x_1 \notin \{0,2\} \end{cases}$$
$$= \begin{cases} \frac{1}{2} \text{ if } x_1 = 0 \\ \frac{1}{2} \text{ if } x_1 = 2 \\ 0 \text{ if } x_1 \notin \{0,2\} \end{cases}$$

$$p_{X_1|X_2=1}(x_1) = \begin{cases} \frac{p_{X_1,X_2}(0,1)}{p_{X_2}(1)} \text{ if } x_1 = 0 \\ 0 \text{ if } x_1 \neq 0 \end{cases} = \begin{cases} \frac{1/8}{1/8} \text{ if } x_1 = 0 \\ 0 \text{ if } x_1 \neq 0 \end{cases}$$
$$= \begin{cases} 1 \text{ if } x_1 = 0 \\ 0 \text{ if } x_1 \neq 0. \end{cases}$$

Exercise:
Consider again the bivariate p.d.f.

$$f_{X_1,X_2}(x_1, x_2) = e^{-(x_1+x_2)} \mathbb{1}_{[0,\infty)\times[0,\infty)}(x_1, x_2).$$

Derive the conditional p.d.f. of X_2 given X_1.
First of all notice that the conditional p.m.f. of X_2 given $X_1 = x_1$ is only well-defined for $x_1 \in \text{supp}(X_1) = [0, \infty)$. For $x_1 \in [0, \infty)$, we have

$$f_{X_2|X_1=x_1}(x_2) = \frac{f_{X_1,X_2}(x_1, x_2)}{f_{X_1}(x_1)} = \frac{e^{-(x_1+x_2)} \mathbb{1}_{[0,\infty)\times[0,\infty)}(x_1, x_2)}{e^{-x_1} \mathbb{1}_{[0,\infty)}(x_1)}$$
$$= \frac{e^{-(x_1+x_2)} \mathbb{1}_{[0,\infty)}(x_2)}{e^{-x_1}} = e^{-x_2} \mathbb{1}_{[0,\infty)}(x_2).$$

Technical note: you may wonder why, if X_1, X_2 are continuous r.v's, $P(X_1 \in A_1 | X_2 = x_2)$ is even defined after all. Indeed, the event $P(X_2 = x_2)$ has probability zero. However, intuitively this probability should still make sense.

In order to overcome this kind of issue, we should dive into *regular* conditional probabilities. In its simplest form, think about the following decompoition

$$F_{X_1}(x_1) = \int_{-\infty}^{+\infty} F_{X_1|X_2=x_2}(x_1) f_{X_2}(x_2) dx_2.$$

We also know that

$$F_{X_1}(x_1) = \int_{-\infty}^{x_1} f_{X_1}(y_1) dy_1 = \int_{-\infty}^{+\infty} \int_{-\infty}^{x_1} f_{X_1,X_2}(y_1, y_2) dy_1 dy_2.$$

Therefore we would like the following equality to hold

$$F_{X_1|X_2=x_2}(x_1) f_{X_2}(x_2) = \int_{-\infty}^{x_1} f_{X_1,X_2}(y_1, y_2) dy_1.$$

We call exactly this expression the conditional density function of X_1 with respect to X_2.

A Test for the Independence of Two Random Variables

Suppose that you are given $(X_1, X_2) \sim f_{X_1,X_2}$. If

1. the support of f_{X_1,X_2} is a 'rectangular' region and
2. $f_{X_1,X_2}(x_1, x_2) = f_{X_1}(x_1) f_{X_2}(x_2)$ for all $x_1, x_2 \in \mathbb{R}$

then X_1 and X_2 are independent. The same is true for the case where (X_1, X_2) is discrete, after obvious changes to 1). Notice that 1) can be conveniently captured by appropriately using indicator functions.

Exercise:
Consider the following bivariate p.d.f.:

$$f_{X_1,X_2}(x_1, x_2) = 6(1 - x_2) \mathbb{1}_{\{0 \leq x_1 \leq x_2 \leq 1\}}(x_1, x_2).$$

Are X_1 and X_2 independent?

A Note on Independence and Normalizing Constants

Frequently, the p.d.f. (and the same holds true for a p.m.f.) of a pair of random variables takes the form

$$f_{X_1,X_2}(x_1, x_2) = c g(x_1) h(x_2) \tag{34}$$

where $c > 0$ is a constant, g is a function depending only on x_1 and h is a function depending only on x_2. If this is the case, the two random variables X_1 and X_2 are independent and g and h are the kernels of the p.d.f. of X_1 and X_2 respectively. Thus, by simply inspecting g and h we can guess to which family of distributions X_1 and X_2 belong to. We do not need to worry about

the constant c, as we know that if f_{X_1,X_2} is a bona fide p.d.f., then $c = c_1 c_2$ is exactly equal to the product of the normalizing constants c_1 and c_2 that satisfy

$$c_1 \int_{\mathbb{R}} g(x)\,dx = c_2 \int_{\mathbb{R}} h(x)\,dx = 1. \tag{35}$$

This easily generalizes to a collection of $n > 2$ random variables.

Lecture 18

Recommended Readings: WMS Chapter 5.5, 5.6,5.8
Learning objectives

1. Students will learn to compute expectations of functions of multiple random variables.

2. Students will be be able to recall the sum and multiplication rule for Expectations involving multiple random variables.

Previously we introduced the concept of expectation or expected value of a random variable. We will now introduce other operators that are based on the expected value operator and we will investigate how they act on bivariate distributions. While we focus on bivariate distributions, it is worthwile to keep in mind that all that we discuss in this lecture can be extended to collections of random variables $X_1, \ldots, X_n$ with $n > 2$.
For a function

$$\begin{aligned} g :& \mathbb{R}^2 \to \mathbb{R} \\ & (x_1, x_2) \mapsto g(x_1, x_2) \end{aligned}$$

and a pair of random variables (X_1, X_2) with joint p.m.f. p_{X_1,X_2} (if discrete) or joint p.d.f. f_{X_1,X_2} (if continuous), the expected value of $g(X_1, X_2)$ is defined as

$$E(g(X_1, X_2)) = \sum_{x_1 \in \text{supp}(X_1)} \sum_{x_2 \in \text{supp}(X_2)} g(x_1, x_2) p_{X_1,X_2}(x_1, x_2) \text{ (discrete case)}$$

$$E(g(X_1, X_2)) = \int_{\mathbb{R}} \int_{\mathbb{R}} g(x_1, x_2) f_{X_1,X_2}(x_1, x_2)\, dx_1 dx_2 \text{ if } X_1 \text{ and } X_2 \text{ (continuous case).}$$

Expectation and Independence

In general, given two random variables X_1 and X_2, it is not true that $E(X_1X_2) = E(X_1)E(X_2)$. We said, however, that it is true as soon as X_1 and X_2 are independent. Let's see why. Without loss of generality, assume that X_1 and X_2 are continous (for the discrete case, just change integration into summation as usual). If they are independent, $f_{X_1,X_2} = f_{X_1} f_{X_2}$. Take $g(x_1, x_2) = x_1x_2$. Then, we have

$$\begin{aligned} E(X_1X_2) &= \int_{\mathbb{R}} \int_{\mathbb{R}} x_1x_2 f_{X_1,X_2}(x_1, x_2)\, dx_1 dx_2 \\ &= \int_{\mathbb{R}} \int_{\mathbb{R}} x_1x_2 f_{X_1}(x_1) f_{X_2}(x_2)\, dx_1 dx_2 = \int_{\mathbb{R}} x_1 f_{X_1}(x_1)\, dx_1 \int_{\mathbb{R}} x_2 f_{X_2}(x_2)\, dx_2 \\ &= E(X_1)E(X_2). \end{aligned}$$

The above argument is easily extended to show that for any two functions $g_1, g_2 : \mathbb{R} \to \mathbb{R}$, if X_1 and X_2 are independent, then $E(g_1(X_1)g_2(X_2)) =$

$E(g_1(X_1)g_2(X_2))$.

Therefore remember that independence implies $E[XY] = E[X]E[Y]$, but the opposite does not hold in general.

Exercise:
Consider the following bivariate p.d.f.:

$$f_{X_1,X_2}(x_1, x_2) = \frac{3}{2}e^{-3x_1}\mathbb{1}_{[0,\infty)\times[0,2]}(x_1, x_2).$$

Are X_1 and X_2 independent? Why? What are their marginal distributions? Compute $E(3XY + 3)$.

Exercise:
Consider the pair of random variables X_1 and X_2 with joint p.d.f.

$$f_{X_1,X_2}(x_1, x_2) = \frac{1}{8}x_1 e^{-(x_1+x_2)/2}\mathbb{1}_{[0,\infty)\times[0,\infty)}(x_1, x_2)$$

and consider the function $g(x, y) = y/x$. What is $E(g(X_1, X_2))$?
First of all, notice that X_1 and X_2 are independent. Therefore, we know already that $E(g(X_1, X_2)) = E(X_2/X_1) = E(X_2)E(1/X_1)$. Moreover, the kernel of the p.d.f. of X_1 is

$$x_1 e^{-x_1/2}\mathbb{1}_{[0,\infty)}(x_1)$$

while that of the p.d.f. of X_2 is

$$e^{-x_2/2}\mathbb{1}_{[0,\infty)}(x_2).$$

Thus, X_1 and X_2 are distributed according to Gamma$(2, 2)$ and Exponential(2) respectively. It follows that $E(X_2) = \beta = 2$. We only need to compute $E(1/X_1)$. This is equal to

$$\begin{aligned} E\left(\frac{1}{X_1}\right) &= \int_0^\infty \frac{1}{x_1} f_{X_1}(x_1)\, dx = \int_0^\infty \frac{1}{x_1}\frac{1}{4}x_1 e^{-\frac{x_1}{2}}\, dx \\ &= \frac{1}{4}\int_0^\infty e^{-\frac{x_1}{2}}\, dx = \frac{1}{4}2 = \frac{1}{2}. \end{aligned}$$

It follows that $E(X_2/X_1) = E(X_2)E(1/X_1) = 2(1/2) = 1$.

Lecture 19

Recommended Readings: WMS Chapter 5.7
Learning objectives

1. Students will recall the definitions of covariance and correlation.
2. Students will be able to apply theory from expectations to prove properties of covariance.
3. Students will recall that covariance is only a measure of linear dependence, and that zero covariance does not imply independence in general.

Recall that $V(Z) = E[Z^2] - (E[Z])^2$. For two random variables X and Y, $V(X+Y) \neq V(X) + V(Y)$ usually, but that the result is true if X and Y are independent. Now, we can easily see that. Let $Z = X + Y$;

$$\begin{aligned} V(X+Y) &= E[(X+Y)^2] - (E[X+Y])^2 \text{ apply sum prop. of } E \\ &= E[(X+Y)^2] - (E[X] + E[Y])^2 \\ &= E[X^2 + 2XY + Y^2] - \left[(E[X])^2 + 2E[X]E[Y] + (E[Y])^2\right] \text{ linearity of } E \\ &= E[X^2] + 2E[XY] + E[Y^2] - (E[X])^2 - 2E[X]E[Y] - (E[Y])^2 \\ &= V(X) + V(Y) + 2\left(E[XY] - E[X]E[Y]\right) \\ &\text{apply product prop. of E for independent } X \text{ and } Y \\ &= V(X) + V(Y) + 2\left(E[X]E[Y] - E[X]E[Y]\right) \\ &= V(X) + V(Y) \text{ for independent } X \text{ and } Y \end{aligned} \tag{36}$$

Question: what about $V(X-Y)$?

Covariance

The *covariance* of a pair of random variables X and Y is a measure of their linear dependence. By definition, the covariance between X and Y is

$$Cov(X,Y) = E[(X - E(X))(Y - E(Y))] = E(XY) - E(X)E(Y). \tag{37}$$

It is easy to check that the covariance operator satisfies, for $a, b, c, d \in \mathbb{R}$

$$Cov(a + bX, c + dY) = bdCov(X,Y).$$

Also, notice that $Cov(X,Y) = Cov(Y,X)$ and $Cov(X,X) = V(X)$.

Question: suppose that X and Y are independent. What is $Cov(X,Y)$?

There exists a scaled version of the covariance which takes values in $[-1,1]$. This is the *correlation* between X and Y. The correlation between X and Y is defined as

$$Cor(X,Y) = \frac{Cov(X,Y)}{\sqrt{V(X)V(Y)}}. \tag{38}$$

It is easy to check that for $a, b, c, d \in \mathbb{R}$ with $b, d \neq 0$, we have $Cor(a + bX, c + dY) = Cor(X, Y)$. So, unlike the covariance operator, the correlation operator is not affected by affine transformations of X and Y.
Beware that while we saw that X and Y are independent $\implies Cov(X, Y) = 0$, the converse is not true (i.e. independence is stronger than *uncorrelation*). Consider the following example. Let X be a random variable with $E(X) = E(X^3) = 0$. Consider $Y = X^2$. Clearly, X and Y are not independent (in fact, Y is a deterministic function of X). However,

$$Cov(X, Y) = E(XY) - E(X)E(Y) = E(X * X^2) - 0 = E(X^3) = 0.$$

So X and Y are uncorrelated, because $Cor(X, Y) = 0$, but they are not independent!

Exercise:
Look back at equation (36). Without assuming that X and Y are independent, but rather only making the weaker assumption that X and Y are *uncorrelated* (i.e. $Cov(X, Y) = 0$), show that

$$V(X + Y) = V(X - Y) = V(X) + V(Y).$$

From the discussion above it is clear that in general

$$\begin{aligned} V(X + Y) &= V(X) + V(Y) + 2Cov(X, Y) \\ V(X - Y) &= V(X) + V(Y) - 2Cov(X, Y). \end{aligned}$$

This generalizes as follows. Given $a_1, \ldots, a_n \in \mathbb{R}$ and $X_1, \ldots, X_n$,

$$\begin{aligned} V\left(\sum_{i=1}^{n} a_i X_i\right) &= \sum_{i=1}^{n} a_i^2 V(X_i) + \sum_{i=1}^{n} \sum_{\substack{j=1 \\ j \neq i}}^{n} a_i a_j Cov(X_i, X_j) \\ &= \sum_{i=1}^{n} a_i^2 V(X_i) + \sum_{1 \leq i < j \leq n} 2a_i a_j Cov(X_i, X_j). \end{aligned}$$

If the random variables $X_1, \ldots, X_n$ are all pairwise uncorrelated, then obviously the second summand above is null.

Exercise:
You are given three random variables X, Y, Z with $V(X) = 1$, $V(Y) = 4$, $V(Z) = 3$, $Cov(X, Y) = 0$, $Cov(X, Z) = 1$, and $Cov(Y, Z) = 1$.
Compute $V(3X + Y - 2Z)$.
We have

$$\begin{aligned} V(3X + Y - 2Z) &= V(3X) + V(Y) + V(-2Z) + 2Cov(3X, Y) + 2Cov(3X, -2Z) + 2Cov(Y, -2Z) \\ &= 9V(X) + V(Y) + 4V(Z) + 6Cov(X, Y) - 12Cov(X, Z) - 4Cov(Y, Z) \\ &= 9 + 4 + 12 - 12 - 4 = 9. \end{aligned}$$

Remark: Why is $\rho(X,Y) := Cor(X,Y) \in [-1,1]$? In order to answer this question, we will need the *Cauchy-Schwarz* (CS) inequality. An application of this inequality gives us the following bound:

$$|E[ZQ]|^2 \leq (E[Z^2])(E[Q^2])$$

In particular, notice the absolute value. The more general version of this inequality is known as *Holder inequality*. CS inequality, although simple, is one of the most powerful tools used in Statistics. Now, if we take $Z := X - E[X]$ and $Q := Y - E[Y]$, we have

$$|cov(X,Y)| \leq \sqrt{Var(X)}\sqrt{Var(Y)}.$$

It follows that

$$-\sqrt{Var(X)}\sqrt{Var(Y)} \leq cov(X,Y) \leq \sqrt{Var(X)}\sqrt{Var(Y)}$$

hence

$$Cor(X,Y) \in [-1,1].$$

Lecture 20

Recommended Readings: WMS Chapter 5.7, 5.11
Learning objectives

1. Students will be able to recognize that functions of random variables are also random variables.
2. Students will be able to recall the definitions of conditional expectation and conditional variance.
3. Students will be able to apply the law of total expectation and law of total variance.

Using conditional distributions, we can define the notion of *conditional expectation*, *conditional variance*, and *conditional covariance*.

Introduction Conditional expectation can first be thought of as simply the expectation of a (discrete for now) random variable X given an event A:

$$E(X|A) = \sum_{x \in \text{supp}(X)} x p_{X|A}(x|A)$$

Conditional expectation $E(X|A)$ has all the properties of regular expectation. In particular:

1. $E(f(X)|A] = \sum_{x \in \text{supp}(X)} f(x) p_{X|A}(x|A)$
2. $E[aX + b|A] = aE[X|A] + b \ \forall a, b$

Set $A = \{Y = y\}$:

$$\Rightarrow E[X|Y = y] = \sum_{x \in \text{supp}(X)} x p_{X|Y}(x|y)$$

Where $E[X|Y = y]$ is the *conditional expectation* of X given $Y = y$. Now, Write $E[X|Y]$ where X and Y are r.v.'s *as the random variable*, which is a function of Y, whose value when $Y = y$ is $E[X|Y = y]$. So think of $E[X|Y] = g(Y)$ for some function g.
From here on, the distinction between how we refer to the two may be blurred i.e. we will call both $E(X|A)$ and $E(X|Y)$ conditional expectations, but it should be clear which it is referring to, from context.

Conditional Expectation

Given a conditional p.m.f. $p_{X_1|X_2=x_2}$ or a conditional p.d.f. $p_{X_1|X_2=x_2}$ the conditional expectation of X_1 given $X_2 = x_2$ is defined as

$$E(X_1|X_2 = x_2) = \sum_{x_1 \in \text{supp}(X_1)} x_1 p_{X_1|X_2=x_2}(x_1) \tag{39}$$

in the discrete case and as

$$E(X_1|X_2 = x_2) = \int_{x_1 \in \text{supp}(X_1)} x_1 f_{X_1|X_2=x_2}(x_1)\, dx_1 \tag{40}$$

in the continuous case.
It is clear from the definition that the conditional expectation is a function of the value of X_2. Since X_2 is a random variable, this means that the conditional expectation (despite its name) is itself a random variable! This is a fundamental difference with respect to the standard concept of expectation for a random variable (for instance $E(X_1)$) which is just a constant depending on the distribution of X_1. In terms of notation, we often denote the random variable corresponding to the conditional expectation of X_1 given X_2 simply as $E(X_1|X_2)$.
Note that if X_1 and X_2 are independent, then $E(X_1|X_2) = E(X_1)$.
Conditional expectation has the so-called 'tower property', or law of total expectation, which says that

$$E(E(X_1|X_2)) = E(X_1)$$

where the outer expectation is taken with respect to the probability distribution of X_2. This is easier to understand in the context of discrete random variables.
Exercise: Prove this, for discrete and continuous case. Hint: carefully write down the expression for $E(E(X_1|X_2))$, where $E(X_1|X_2)$ can be treated as a random variable and a function of X_2. You may even write it as $g(X_2)$ if it makes it clear. Then, work on the double sum (discrete) or double integral (continuous).

Conditional Variance

We can also define the conditional variance of X_1 given X_2. We have

$$V(X_1|X_2) = E[(X_1 - E(X_1|X_2))^2|X_2] = E(X_1^2|X_2) - [E(X_1|X_2)]^2 \tag{41}$$

Obtaining the unconditional variance from the conditional variance is a little 'harder' than obtaining the unconditional expectation from the conditional expectation:

$$V(X_1) = E[V(X_1|X_2)] + V[E(X_1|X_2)].$$

Note that if X_1 and X_2 are independent, then $V(X_1|X_2) = V(X_1)$.

Conditional Covariance

It is worthwhile mentioning that we can also define the conditional covariance between X_1 and X_2 given a third random variable X_3. We have

$$\begin{aligned} Cov(X_1, X_2|X_3) &= E[(X_1 - E(X_1|X_3))(X_2 - E(X_2|X_3))|X_3] \\ &= E(X_1X_2|X_3) - E(X_1|X_3)E(X_2|X_3). \end{aligned} \tag{42}$$

To get the unconditional covariance we have the following formula:

$$Cov(X_1, X_2) = E[Cov(X_1, X_2|X_3)] + Cov[E(X_1|X_3), E(X_2|X_3)]. \tag{43}$$

Note that if X_1 and X_3 are independent, and X_2 and X_3 are independent, then $Cov(X_1, X_2|X_3) = Cov(X_1, X_2)$.
In terms of computation, everything is essentially unchanged. The only difference is that we sum or integrate against the conditional p.m.f./conditional p.d.f. rather than the marginal p.m.f./marginal p.d.f..

Exercise:
Consider again the p.d.f. of exercise 7 in Homework 5:

$$f_{X_1,X_2}(x_1, x_2) = 6(1 - x_2)\mathbb{1}_{\{0\leq x_1\leq x_2\leq 1\}}(x_1, x_2).$$

What is $E(X_1|X_2)$? Compute $E(X_1)$.
We first need to compute $f_{X_1|X_2=x_2}$ for $x_2 \in \text{supp}(X_2)$. For $x_2 \in [0, 1]$ we have

$$f_{X_2}(x_2) = \int_0^{x_2} 6(1 - x_2)\mathbb{1}_{[0,1]}(x_1)\, dx_1 = 6x_2(1 - x_1)\mathbb{1}_{[0,1]}(x_1)$$

Notice that from the marginal p.d.f. of X_2 we can see that $X_2 \sim \text{Beta}(2, 2)$. For $x_2 \in (0, 1)$ we have

$$\begin{aligned} f_{X_1|X_2=x_2}(x_1) &= \frac{f_{X_1,X_2}(x_1, x_2)}{f_{X_2}(x_2)} = \frac{6(1 - x_2)\mathbb{1}_{[0,x_2]}(x_1)}{6x_2(1 - x_2)} \\ &= \frac{1}{x_2}\mathbb{1}_{[0,x_2]}(x_1) \end{aligned}$$

Notice that $f_{X_1|X_2=x_2}$ is the p.d.f. of a Uniform$(0, x_2)$ distribution. It follows that $E(X_1|X_2 = x_2) = x_2/2$. We can write this more concisely (and in a way that stresses more the fact that the conditional expectation is a random variable!) as $E(X_1|X_2) = X_2/2$.
We have $E(X_1) = E[E(X_1|X_2)] = E(X_2/2) = E(X_2)/2 = [(2/(2+2)]/2 = 1/4$.

Exercise:
Let $N \sim \text{Poisson}(\lambda)$ and let $Y_1, \ldots, Y_N \overset{\text{i.i.d.}}{\sim} \text{Gamma}(\alpha, \beta)$ with N independent of each of the Y_i's. Let $T = \sum_{i=1}^N Y_i$. Compute $E(T|N)$, $E(T)$, $V(T|N)$, and $V(T)$.
We have that

$$\begin{aligned} E(T|N = n) &= E\left(\sum_{i=1}^N Y_i|N = n\right) = \sum_{i=1}^n E(Y_i|N = n) \\ &= \sum_{i=1}^n E(Y_i) = n\alpha\beta. \end{aligned}$$

Thus, $E(T|N) = N\alpha\beta$. Now,

$$E(T) = E[E(T|N)] = E(N\alpha\beta) = \alpha\beta E(N) = \alpha\beta\lambda.$$

The conditional variance is

$$\begin{aligned} V(T|N=n) &= V\left(\sum_{i=1}^{N} Y_i | N=n\right) = V\left(\sum_{i=1}^{n} Y_i | N=n\right) \\ &= \sum_{i=1}^{n} V(Y_i|N=n) = \sum_{i=1}^{n} V(Y_i) = n\alpha\beta^2. \end{aligned}$$

Thus, $V(T|N) = N\alpha\beta^2$. The unconditional variance of T is then

$$\begin{aligned} V(T) &= V[E(T|N)] + E[V(T|N)] = V(N\alpha\beta) + E(N\alpha\beta^2) \\ &= \alpha^2\beta^2 V(N) + \alpha\beta^2 E(N) \\ &= \alpha^2\beta^2\lambda + \alpha\beta^2\lambda \\ &= \alpha\beta^2\lambda(1+\alpha). \end{aligned}$$

Exercise:
Let $Q \sim \text{Uniform}(0,1)$ and $Y|Q \sim \text{Binomial}(n, Q)$. Compute $E(Y)$ and $V(Y)$. We have

$$E(Y) = E[E(Y|Q)] = E(nQ) = nE(Q) = n/2$$

and

$$\begin{aligned} V(Y) &= V[E(Y|Q)] + E[V(Y|Q)] = V(nQ) + E(nQ(1-Q)) \\ &= n^2 V(Q) + nE(Q - Q^2) = \frac{n^2}{12} + nE(Q) - nE(Q^2) \\ &= \frac{n^2}{12} + \frac{n}{2} - n(V(Q) + [E(Q)]^2) \\ &= \frac{n^2}{12} + \frac{n}{2} - \frac{n}{12} - \frac{n}{4} \\ &= \frac{n^2}{12} + \frac{n}{6} = \frac{n}{6}(n/2+1). \end{aligned}$$

Lecture 21

Recommended Readings: WMS Chapter 5.9, 5.10
Learning objectives

1. Students will be able to interpret a correlation coefficient in the context of a pair of normally distributed random variables.
2. Students will be able to recall the definition and properties of iid random variables.
3. Students will be able to recall the definition of the multinomial distribution and its relationship to the binomial distribution.

Bivariate normal

A bivariate normal distribution is determined by a pair of means, a pair of standard deviations, and a correlation coefficient. We say that X and Y are continuous random variable with a bivariate normal (Gaussian) distribution when they have the following p.d.f.:

$$f(x,y) = \frac{1}{2\pi\sigma_X\sigma_Y\sqrt{1-\rho^2}} \exp\left\{ -\frac{1}{2}Q \right\}$$

Where $\rho = Corr(X,Y)$, $-1 < \rho < 1$, $\sigma_X = \sqrt{(V(X))} > 0$, $\sigma_Y = \sqrt{(V(Y))} > 0$, and

$$Q = \frac{1}{1-\rho^2}\left[\frac{(x-\mu_X)^2}{\sigma_X^2} + \frac{(y-\mu_Y)^2}{\sigma_Y^2} - \frac{2\rho(x-\mu_X)(y-\mu_Y)}{\sigma_X\sigma_Y} \right],$$

with $\mu_X = E[X]$ and $\mu_Y = E[Y]$.
If X, Y are bivariate normal, it follows that: (1) the marginal p.d.f's are normal with $X \sim \mathcal{N}(\mu_X, \sigma_X^2)$, $Y \sim \mathcal{N}(\mu_Y, \sigma_Y^2)$, and (2) the conditional distributions of $X|Y$ and $Y|X$ are also normal. For example,
$(X|Y=y) \sim \mathcal{N}(\mu_X + \rho\sigma_X \frac{(y-\mu_Y)}{\sigma_Y}, (1-\rho^2)\sigma_x^2)$.

Special case: two random variables X and Y that have a bivariate normal distribution are independent if and only if they are uncorrelated.

IID samples

Consider a collection of n random variables $X_1, \ldots, X_n$ and their *joint probability distribution*. Their joint probability distribution can be described in terms of the joint c.d.f. or in terms of the joint p.d.f. $f_{X_1,\ldots,X_n}$ (if the random variables are all continuous, discrete case is analogous).
The random variables $X_1, \ldots, X_n$ are said to be independent if either of the following holds:

- $F_{X_1,\ldots,X_n}(x_1, x_2, \ldots, x_n) = \prod_{i=1}^{n} F_{X_i}(x_i)$
- $f_{X_1,\ldots,X_n}(x_1, x_2, \ldots, x_n) = \prod_{i=1}^{n} f_{X_i}(x_i)$.

If the random variables are independent and also share the same *marginal distribution*, i.e. we have

- $F_{X_i} = F \quad \forall i \in \{1, \ldots, n\}$
- $f_{X_i} = f \quad \forall i \in \{1, \ldots, n\}$

then the random variables $X_1, \ldots, X_n$ are said to be *independent and identically distributed*, usually shortened in *i.i.d.*. We also call $X_1, \ldots, X_n$ a random sample of size n from F.

Multinomial Distribution

The multinomial distribution is a generalization of the binomial distribution. Suppose n i.i.d. experiments are performed. Each experiment can lead to r possible outcomes with probability $p_1, p_2, \cdots, p_r$ such that

$$p_i > 0, \sum_{i=0}^{r} p_i = 1$$

Now, let X_i be the number of experiments resulting in outcome $i \in [1, r]$. Then, the multinomial distribution characterizes the joint probability of $(X_1, X_2, \cdots, X_n)$; in other words, it fully describes

$$P(X_1 = x_1, X_2 = x_2, \cdots, X_r = x_r) \tag{44}$$

The probability mass function is

$$p_{X_1,\cdots,X_r}(x_1, \cdots, x_r) = \begin{cases} \binom{n}{x_1\ x_2 \cdots x_r} p_1^{x_1} \cdots p_r^{x_r} & \text{if } \sum_{i=1}^{r} x_i = n \\ 0 & \text{otherwise} \end{cases} \tag{45}$$

What is the mean? The mean should be defined on each of the X_i, and is $E(X_i) = np_i$. The variance is $V(X_i) = np_i(1 - p_i)$. The variance of each X_i is $V(X_i) = Cov(X_i, X_i) = np_i(1 - p_i)$. The covariance of X_i and X_j is $Cov(X_i, X_j) = -np_ip_j$.
How should we understand the expectation and variance? We can see that the *marginal* probability distribution of X_i as simply $\text{Binom}(n, p_i)$ – if we only focus on one variable at a time, each is simply the number of successes (each success having probability p_i) out of n trials.
The outcomes of X_i and X_j are 'pitted' against each other; if X_i is high, then $X_j (j \neq i)$ should be low, since there are only n draws in total. Indeed, that explains how they are negatively correlated (and have negative covariance).

Exercise
A fair die is rolled 9 times. What is the probability of 1 appearing 3 times,

2 appearing 2 times, 3 appearing 2 times, 4 appearing 1 times, 5 appearing 1 time, and 6 appearing 0 times?

Exercise
According to recent census figures, the proportion of adults (18 years or older of age) in the U.S. associated with 5 age categories are as given in the following table.

Age	Proportion
18-24	.18
25-34	.23
35-44	.16
45-64	.27
65+	.16

If the figures are accurate and five adults are randomly sampled, find the probability that the sample contains one person between the ages of 18 and 24, two between the ages of 25 and 34, and two between 45 and 64. (Hint: see WMS textbook)

Exercise:
Suppose that $(X_1, \ldots, X_k)$ has a Multinomial$(p_1 \ldots, p_k)$ distribution. We want to prove:

$$Cov[X_i, X_j] = -np_ip_j.$$

The trick is to treat a multinomial experiment as a sequence of n independent trials, Y_t. $1_{Y_t=i}$ is the random variable which takes the value of 1 if the t'th outcome Y_t took value i, and notice that $X_i = \sum_{t=1}^{n} 1_{Y_t=i}$ and $X_j = \sum_{t=1}^{n} 1_{Y_t=j}$. Then, proceed as follows: Now, if $s = t$,

$$Cov[1_{Y_t=i}, 1_{Y_s=j}] = -\mathbb{E}[1_{Y_t=i}]\mathbb{E}[1_{Y_t=j}] = -p_ip_j,$$

while if $s \neq t$

$$Cov[1_{Y_t=i}, 1_{Y_s=j}] = 0$$

by independence.

$$Cov[X_i, X_j] = -np_ip_j = \sum_{t=1}^{n}\sum_{s=1}^{n} Cov[1_{Y_t=i}, 1_{Y_s=j}].$$

Hence the result.

Lecture 22

Recommended readings: WMS, sections 3.9, 4.9, 6.1-6.5
Learning objectives

1. Students will recall the definition and utility of m.g.f.'s.
2. Students will learn techniques to obtain the distributions of functions of a random variables.

Moments

Remember how we learned that a CDF or a PDF/PMF *completely* characterizes the distribution of a random variable – in other words, it contains all you need to know about the law of randomness of that random variable. If two random variables (from different origins, or sampling procedure) give the same CDF, then they have the same distribution.
What is a *moment* of F? In mathematics/statistics/mechanics, a moment is a specific quantitative measure of the shape of a set of points. (For simplicity, think of such points as potential realizations random variables.) Would you agree that, if you have a set of points, and you knew (1) the center of balance (2) how spread out they are, then you already have a pretty good rough idea of the distribution of variables? (1) corresponds to the first moment (and the mean). The spread (variance) is the second moment minus the square of the first moment. The story goes on – what if you knew which side of the mean the points are *more* populated/concentrated in? This is given by the third moment, plus some multiples of the first and second moment.
Would you agree that, as you go further and learn more such information about the shape of the points, that you learn more about the exact *distribution* of the random variables? Indeed, knowing all the moments is equivalent to knowing the CDF (from which we know all we need to know about the distribution)!
In this lecture, we introduce a particular function associated to a probability distribution F which *uniquely* characterizes F. This function is called the *moment-generating function* (henceforth shortened to m.g.f.) of F because, if it exists, it allows to easily compute any *moment* of F, i.e. any expectation of the type $E(X^k)$ with $k \in \{0, 1, \dots\}$ for $X \sim F$.

Moment generating functions

There are other functions that are similar to the m.g.f.s which we will not discuss in this course: these include the *probability-generating function* for discrete probability distributions (which is a compact power series representation of the p.m.f.), the *characteristic function* (which is the inverse Fourier transform of a p.m.f./p.d.f. and always exists) and the *cumulant-generating function* (which is the logarithm of the m.g.f.).

The moments $\{E(X^k)\}_{k=1}^{\infty}$ associated to a distribution $X \sim F$ completely characterize F (if they exist). They can all be encapsulated in the m.g.f.

$$m_X(t) = E(e^{tX}) = \begin{cases} \int_{\text{supp}(X)} e^{tx} f_X(x)\, dx \text{ if } X \text{ is continuous} \\ \sum_{\text{supp}(X)} e^{tx} p_X(x) \text{ if } X \text{ is discrete.} \end{cases}$$

If two random variables X and Y are such that $m_X = m_Y$ then their c.d.f.'s F_X and F_Y are equal at almost all points (i.e. they can differ at at most countably many points). If you are interested in the proof, I suggest you to look it up in more advanced books.
We say that the moment generating function m_X of $X \sim F$ exists if there exists an open neighborhood around $t = 0$ in which $m_X(t)$ is finite. [5]Note that it is always true that, for $X \sim F$ with an arbitrary distribution F, $m_X(0) = 1$.
The name of this function comes from the following feature: suppose that m_X exists, then for any $k \in \{0, 1, \dots\}$

$$\frac{d^k}{dt^k} m_X(t)\bigg|_{t=0} = E(X^k). \tag{46}$$

This means that we can 'generate' the moments of $X \sim F$ from m_X by differentiating m_X and evaluating its derivatives at $t = 0$. This is more clear if we rewrite the m.g.f. in terms of its series expansion: $\forall t \in \mathbb{R}$

$$\begin{aligned} E[e^{tX}] &= E\left[1 + tX + \frac{t^2X^2}{2!} + \frac{t^3X^3}{3!} + \cdots + \frac{t^nX^n}{n!} + \dots\right] \\ &= 1 + tE[X] + \frac{t^2E[X^2]}{2!} + \frac{t^3E[X^3]}{3!} + \cdots + \frac{t^nE[X^n]}{n!} + \dots \end{aligned}$$

Exercise:
Show that the m.g.f. of $X \sim \text{Binomial}(n, p)$ is

$$m_X(t) = [pe^t + 1 - p]^n$$

and use it to compute $V(X)$. Let's see how.

$$\begin{aligned} m_X(t) &= \sum_{x=0}^{n} e^{tx} p_X(x) = \sum_{x=0}^{n} e^{tx} \binom{n}{x} p^x (1-p)^{n-x} \\ &= \sum_{x=0}^{n} \binom{n}{x} (pe^t)^x (1-p)^{n-x} = [pe^t + (1-p)]^n \end{aligned}$$

where we used the binomial theorem

$$(a+b)^n = \sum_{i=0}^{n} \binom{n}{i} a^i b^{n-i}.$$

[5]For instance, the mgf of a Cauchy distribution does not exists.

Notice that the function above is well-defined and finite for any $t \in \mathbb{R}$. Then,

$$E(X) = \frac{d}{dt} m_X(t)\Big|_{t=0} = npe^t[pe^t + (1-p)]^{n-1}\big|_{t=0} = np$$

and

$$\begin{aligned} E(X^2) = \frac{d^2}{dt^2} m_X(t)\Big|_{t=0} &= np^2e^{2t}(n-1)[pe^t + (1-p)]^{n-2} + npe^t[pe^t + (1-p)]^{n-1}\big|_{t=0} \\ &= np^2(n-1) + np. \end{aligned}$$

Thus,

$$V(X) = E(X^2) - [E(X)]^2 = np^2(n-1) + np - n^2p^2 = -np^2 + np = np(1-p).$$

It is easy to verify that a m.g.f. m_X satisfies

$$m_{aX+b}(t) = e^{bt} m_X(at). \tag{47}$$

M.g.f.'s also provide a tool which is often useful to identify the distribution of a linear combination of random variables. In the m.g.f. world sums becomes products! In particular, consider a collection of n independent random variables $X_1, \ldots, X_n$ with m.g.f's $m_{X_1}, \ldots, m_{X_n}$. It is easy to check from the definition of m.g.f. that, if we consider $Y = \sum_{i=1}^n (a_i X_i + b_i)$, then

$$m_Y(t) = e^{\sum_{i=1}^n b_i t} \prod_{i=1}^n m_{X_i}(a_i t). \tag{48}$$

Exercise:

1. What is the mgf of $X \sim \text{Poisson}(\lambda)$?

   $$m_X(t) = e^{\lambda(e^t - 1)}$$

   for $t \in \mathbb{R}$.

2. Consider $Y \sim \text{Poisson}(\mu)$ with X and Y independent. What is the distribution of $X + Y$?

   Since X and Y are independent, we have

   $$m_{X+Y}(t) = m_X(t) m_Y(t) = e^{\lambda(e^t-1)} e^{\mu(e^t-1)} = e^{(\lambda+\mu)(e^t-1)}$$

   and we recognize this as the m.g.f. of a Poisson$(\lambda + \mu)$ distribution.

Why are moment generating functions useful?
First, for X and Y whose moment generating functions are finite, $m_X(t) = m_Y(t)$ for all t if and only if $P(X \leq x) = P(Y \leq x)$ for all x.
We will briefly prove this, for discrete distributions X and Y. One direction is trivial; if the two have the same distribution, then of course

$$m_X(t) = E(e^{tX}) = E(e^{tY}) = m_Y(t)$$

is true. The other direction is harder. Call $A = \text{supp}(X) \cup \text{supp}(Y)$, and $a_1, \cdots, a_n$ the elements of A. Then, the mgf of X is

$$\begin{aligned} m_X(t) &= E(e^T X) \\ &= \sum_{x \in \text{supp}(X)} e^{tx} \cdot p_X(x) \\ &= \sum_{i=1 \cdots, n} e^{ta_i} \cdot p_X(a_i) \end{aligned}$$

and likewise, $m_Y(t) = \sum_{i=1,\cdots,n} e^{ta_i} \cdot p_X(a_i)$. Subtract the two, to get

$$m_X(t) - m_Y(t) = \sum_{i=\cdots,n} e^{ta_i} \cdot [p_X(a_i) - p_Y(a_i)] = 0$$

must be true when t is close to zero (because it is true for all t). If t is close to zero, then no matter what a_i is, e^{ta_i} must be close to 1. So, it must be the case that

$$p_X(a_i) = p_Y(a_i)$$

for all i; hence, the pmfs are the same, and the distributions are the same.

Second, if $m_{X_n}(t) \to m_X(t)$ for all t, and $P(X \leq x)$ is continuous in x, then $P(X_n \leq x) \to P(X \leq x)$. You can prove the central limit theorem with moment generating functions, assuming the moments are finite.

Functions of Random Variables and Their Distributions

Suppose that you are given a random variable $X \sim F_X$ and a function $g : \mathbb{R} \to \mathbb{R}$. Consider the new random variable $Y = g(X)$. How can we find the distribution of Y?
Besides m.g.f.'s, there are two other approaches that one can typically use to find the distribution of a function of a random variable $Y = g(X)$: the first approach is based on the c.d.f., the other approach is based on the change of variable technique. The former is general and works for any function g, the latter requires some assumptions on g, but it is generally faster than the general approach based on the c.d.f..
Remark: recall that thanks to the *law of the unconscious statistician* (LOTUS), you already know how to compute quantities such as $E[g(X)]$, etc... Here we only focus on computing the *distribution* of $g(X)$. The computation of quantities such as $E[g(X)]$ is typically easier.

The method of the cumulative distribution function

The idea is pretty simple. You know that $X \sim F_X$ and you want to find F_Y. The process is as follows:

1. $F_Y(y) = P(Y \leq y)$ by definition
2. $P(Y \leq y) = P(g(X) \leq y)$, since $Y = g(X)$
3. now you want to express the event $\{g(X) \leq y\}$ in terms of X, since you know the distribution of X only
4. let $A = \{x \in \mathbb{R} : g(x) \leq y\}$; then $\{g(X) \leq y\} = \{X \in A\}$
5. it follows that $P(g(X) \leq y) = P(X \in A)$ which can typically be expressed in terms of F_X
6. (EXTRA) once you have F_X, the p.m.f. or the p.d.f. of X can be easily derived from it.

Exercise:
Let $Z \sim \mathcal{N}(0,1)$ and $g(x) = x^2$. Consider $Y = g(Z) = Z^2$. What is the probability distribution of Y?
Following the steps above, we have

$$F_Y(y) = P(Y \leq y) = P(Z^2 \leq y) = \begin{cases} 0 \text{ if } y < 0 \\ P(|Z| \leq \sqrt{y}) \text{ if } y \geq 0. \end{cases}$$

$$= \begin{cases} 0 \text{ if } y < 0 \\ P(Z \in [-\sqrt{y}, \sqrt{y}]) \text{ if } y \geq 0. \end{cases}$$

Notice that in this case $A = [-\sqrt{y}, \sqrt{y}]$. It follows that

$$F_Y(y) = \begin{cases} 0 \text{ if } y < 0 \\ P(Z \in [-\sqrt{y}, \sqrt{y}]) \text{ if } y \geq 0. \end{cases} = \begin{cases} 0 \text{ if } y < 0 \\ \Phi(\sqrt{y}) - \Phi(-\sqrt{y}) \text{ if } y \geq 0. \end{cases}$$

$$= \begin{cases} 0 \text{ if } y < 0 \\ 2\Phi(\sqrt{y}) - 1 \text{ if } y \geq 0. \end{cases}$$

Let ϕ denote the standard normal p.d.f.

$$\phi(x) = \frac{1}{\sqrt{2\pi}} e^{\frac{x^2}{2}}.$$

The p.d.f. of $Y = Z^2$ is therefore

$$f_Y(y) = \begin{cases} 0 \text{ if } y \leq 0 \\ \frac{1}{\sqrt{y}} \phi(\sqrt{y}) \text{ if } y > 0 \end{cases} = \begin{cases} 0 \text{ if } y \leq 0 \\ \frac{1}{\sqrt{2\pi}} y^{-\frac{1}{2}} e^{-\frac{y}{2}} \text{ if } y > 0 \end{cases}$$

Notice that this is the p.d.f. of a Gamma$(\frac{1}{2}, 2) \equiv \chi^2(1)$ distribution.

Exercise: the probability integral transform
Consider a continuous random variable $X \sim F_X$ where F_X, the distribution of X is strictly increasing. Consider the random variable $Y = F_X(X)$. What is the probablity distribution of Y?
The transformation $Y = F_X(X)$ is usually called the *probability integral transform*. To see why, notice that

$$Y = F_X(X) = \int_{-\infty}^{X} f_X(y)\, dy.$$

We have

$$F_Y(y) = P(Y \leq y) = P(F_X(X) \leq y) = \begin{cases} 0 \text{ if } y < 0 \\ P(X \leq F_X^{-1}(y)) \text{ if } y \in [0,1) \\ 1 \text{ if } y \geq 1. \end{cases}$$

In this case, therefore, $A = (-\infty, F_X^{-1}(y)]$. Then,

$$F_Y(y) = \begin{cases} 0 \text{ if } y < 0 \\ P(X \leq F_X^{-1}(y)) \text{ if } y \in [0,1) \\ 1 \text{ if } y \geq 1 \end{cases} = \begin{cases} 0 \text{ if } y < 0 \\ F_X(F_X^{-1}(y)) \text{ if } y \in [0,1) \\ 1 \text{ if } y \geq 1 \end{cases}$$

$$= \begin{cases} 0 \text{ if } y < 0 \\ y \text{ if } y \in [0,1) \\ 1 \text{ if } y \geq 1. \end{cases}$$

The p.d.f. of Y is therefore

$$f_Y(y) = \mathbb{1}_{[0,1]}(y),$$

i.e. $Y \sim \text{Uniform}(0,1)$.

The change of variable method

In order to apply the method of the change of variable, the function g must be strictly increasing continuously differentiable and it must also admit an inverse g^{-1} (this is not required by the method based on the c.d.f.). A sufficient condition is that g is strictly monotone (increasing or decreasing).
Suppose that you are given $X \sim F_X$ and you want to compute the probability distribution of $Y = g(X)$. First of all, determine the support of Y. Then, use this formula to obtain f_Y on the support of Y from f_X and g:

$$f_Y(y) = f_X(g^{-1}(y)) \left| \frac{d}{dz} g^{-1}(z) \Big|_{z=y} \right|.$$

Exercise:
Consider $X \sim \text{Beta}(1,2)$ and $g(x) = 2x - 1$. What is the p.d.f. of $Y = g(X)$?
First of all, notice that (with a little abuse of notation) $\text{supp}(Y) = g(\text{supp}(X))$. Since $\text{supp}(X) = [0,1]$, it follows that $\text{supp}(Y) = [-1,1]$.
We have

$$f_X(x) = 2(1-x)\mathbb{1}_{[0,1]}(x),$$

$$g^{-1}(y) = \frac{y+1}{2},$$

and

$$\frac{d}{dx}g^{-1}(x) = \frac{1}{2}.$$

Thus, for $y \in [-1,1]$,

$$\begin{aligned} f_Y(y) &= f_X(g^{-1}(y)) \left| \frac{d}{dx}g^{-1}(x)\Big|_{x=y} \right| \\ &= 2\left(1 - \frac{y+1}{2}\right)\left|\frac{1}{2}\right| = 1 - \frac{y+1}{2} = \frac{1-y}{2}. \end{aligned}$$

The complete description of the p.d.f. of Y is therefore

$$f_Y(y) = \frac{1-y}{2}\mathbb{1}_{[-1,1]}(y).$$

Other exercises:
• $X \sim Exp(\lambda)$ and $Y = 3X + 1$. Find the pdf of Y and $E[Y]$.
• $X \sim \text{Gamma}(\alpha, \beta)$. Prove that $cX \sim \text{Gamma}(\alpha, c\beta)$ where β is the scale parameter.

Inverse transform method

Sometimes it is useful to *simulate* $X_1, \ldots, X_n \overset{iid}{\sim} F_X$. However, this would require the knowledge and developments of methods to simulate from every distribution F_X, which would be unbelievably expensive. Instead, there is a simple way to do it. It is called the *inverse transform method*. We will only see the case of continuous distributions; for the discrete case, the sampling strategy is analogous and one needs to take into account the partitioned space.

The method is fairly simple and it is based on the results about the transformations of random variables that we have already studied. Let $U \sim \text{Uniform}(0,1)$, and F_X any strictly increasing cumulative distribution function (cdf) admitting inverse. Let X be the transformation of U through F_X^{-1}, that is $X = F_X^{-1}(U)$. We would like to show that the equality $X \sim F_X$. Notice that

$$F_U(F_X(x))P(U \leq F_X(x)) = F_X(x) \qquad \forall x \text{ s.t. } F_X(x) \in [0,1]s$$

as already shown for the uniform distribution. Now, the event $\{U \leq F_X(x)\}$ occurs if and only if the event $\{F_X^{-1}(U) \leq F_X^{-1}(F_X(x))\}$ occurs. This is thanks

to the strict monotonicity condition on F_X. Moreover, the latter event is equivalent to $\{F_X^{-1}(U) \leq x\}$, again due to this condition. Therefore we can finally conclude that

$$P(X \leq x) = P(F_X^{-1}(U) \leq x) = F_U(F_X(x)) = F_X(x).$$

This means that $X \sim F_X$. In other words, now we know how to generate every random variable just knowing (1) its *inverse cdf* and (2) how to sample form the uniform distribution. The theory behind random number generation (RNG) is wide, and here we have only touched the surface.
The requirement for the cdf to be monotonic and have a closed-form inverse is not always satisfied. The monotonicity requirement can be easily relaxed. The proof is the same once we define

$$F^{-1}(u) = \inf\{x : F_X(x) \geq u\} \qquad \forall 0 < u < 1.$$

Regarding the sample problem, recall, for instance, that the cdf of the normal distribution does not have a closed-form form. Typically approximate methods are developed.

Exercises:
• Let $F_X(x) = 1 - e^{-\sqrt{x}}$ for $x \in [0, \infty)$. Find the transformation g such that $X = g(U)$ where $X \sim F_X$ and $U \sim \text{Uniform}(0, 1)$.
In other words, we want to find F_X^{-1} since we know that $P(F^{-1}(U) \leq x) = F_X(x)$ thanks to the previous result. Let's find it. Then

$$F_X(x) = 1 - e^{-\sqrt{x}} \iff x = [\log(1 - F_X(x))]^2$$

therefore, taking $F(x) = u$,

$$g(u) = F^{-1}(u) = [\log(1 - u)]^2 .$$

• Let $F_X(x) = 1 - e^{-\lambda x}$ for $x \in [0, \infty)$ and $\lambda > 0$. Find the transformation g such that $X = g(U)$ where $X \sim F_X$ and $U \sim \text{Uniform}(0, 1)$.
Here, similarly as in the example above, we need to have $F(F^{-1}(u)) = u$, therefore

$$1 - e^{-\lambda F^{-1}(u)} = u \iff F^{-1}(u) = -\frac{1}{\lambda} \log(1 - u).$$

therefore $g(U) = F^{-1}(U) = \frac{1}{\lambda} \log(1 - U)$.

Lecture 23

Recommended readings: WMS, sections 7.1 → 7.4
Learning objectives

1. Students will recall the definition of convergence for random variables.
2. Students will recall the CLT and LLN, identify their types of convergence, and use them in practical problems.

The Empirical Rule, Sampling Distributions, the Central Limit Theorem, and the Delta Method

In this lecture we will focus on some basic statistical concepts at the basis of statistical inference. Before starting, let's quickly discuss a useful rule of thumb associated to the Normal distribution which is often mentioned and used in practice.

The Empirical Rule Based on the Normal Distribution

Suppose that you collect data $X_1, \ldots, X_n \sim f$ where f is an unknown probability density function which, however, you know is 'bell-shaped' (or you expect to be 'bell-shaped' given the information you have on the particular phenomenon you are observing, or you can show to 'bell-shaped' using, for example, an histogram).

We know that we can estimate the mean μ_f and the standard deviation σ_f of f by means of the sample mean and the sample standard deviation $\bar{X}$ and S, respectively. On the basis of these two statistics, you may be interested in approximately quantifying the probability content of intervals of the form $[\mu_f - k\sigma_f, \mu_f + k\sigma_f]$ where k is a positive integer. Because for $X \sim \mathcal{N}(\mu, \sigma^2)$ one has

$$\begin{aligned} &P(\mu - \sigma \leq X \leq \mu + \sigma) \approx 68\% \\ &P(\mu - 2\sigma \leq X \leq \mu + 2\sigma) \approx 95\% \\ &P(\mu - 3\sigma \leq X \leq \mu + 3\sigma) \approx 99\%, \end{aligned}$$

one would expect that, based on $\bar{X}$ and S,

$$\begin{aligned} &P([\bar{X} - S, \bar{X} + S]) \approx P([\mu_f - \sigma_f, \mu_f + \sigma_f]) \approx 68\% \\ &P([\bar{X} - 2S, \bar{X} + 2S]) \approx P([\mu_f - 2\sigma_f, \mu_f + 2\sigma_f]) \approx 95\% \\ &P([\bar{X} - 3S, \bar{X} + 3S]) \approx P([\mu_f - 3\sigma_f, \mu_f + 3\sigma_f]) \approx 99\% \end{aligned} \tag{49}$$

if the probability density f is 'bell-shaped'. The approximations of equation (49) are frequently referred to as the *empirical rules* based on the Normal distribution.

Sampling Distributions

To illustrate the concept of sampling distribution, we will consider the Normal model, which assumes that the data $X_1, \ldots, X_n \overset{\text{i.i.d.}}{\sim} \mathcal{N}(\mu, \sigma^2)$ are i.i.d. with a Normal distribution with parameters μ and σ^2 that are usually assumed to be unknown. In order to estimate μ and σ^2, we saw that we can use the two statistics $\bar{X}$ and S^2 (the sample mean and the sample variance). These statistics, or *estimators*, are random variables too, since they depend on the data $X_1, \ldots, X_n$. If they are random variables, then they must have their own probability distributions! The probability distribution of a statistic (or an appropriate stabilizing transformation of it) is called the *sampling distribution* of that statistic. We have the following results:

1. $\frac{\sqrt{n}(\bar{X}-\mu)}{\sigma} \sim \mathcal{N}(0,1)$
2. $\frac{(n-1)S^2}{\sigma^2} \sim \chi^2(n-1)$
3. $\frac{\sqrt{n}(\bar{X}-\mu)}{S} \sim t(n-1)$.

Some comments: 1. is easy to understand and prove (it's just standardization of a Normal random variable!). 1. is useful when we are interested in making inferences about μ and we know σ^2. 2. is a little harder to prove. We use this result when we are interested in making inferences about σ^2 and μ is unknown. 3. is a classical result. It is useful when we want to make inferences on μ and σ^2 is unknown. We won't study in detail the t distribution. However, this distribution arises when we take the ratio of a standard Normal distribution and the square root of a χ^2 distribution divided by its degrees of freedom (under the assumption that the Normal and the χ^2 distributed random variables are also independent). The t distribution looks like a Normal distribution with 'fatter' tails. As the number of degrees of freedom of the t distribution goes to ∞, the t distribution 'converges in distribution' to a standard Normal distribution.
For the purposes of this course, we say that a sequence of random variables $X_1, X_2, \ldots X_n, \ldots$ *converges in distribution* to a certain distribution F if their c.d.f.'s $F_1, F_2, \ldots, F_n, \ldots$ are such that

$$\lim_{n\to\infty} F_n(x) = F(x)$$

for each $x \in \mathbb{R}$ at which F is continuous. We will use the notation $\overset{\text{d}}{\to}$ to denote convergence in distribution. The idea here is that the limiting c.d.f. F (often called the *limiting distribution* of the X's) can be used to approximate probability statements about the random variables in the sequence. This idea will be key to the next result, the Central Limit Theorem.

The Central Limit Theorem

The Central Limit Theorem is a remarkable result. Simply put, suppose that you have a sequence of random variables $X_1, X_2, \ldots, X_n, \ldots \overset{\text{iid}}{\sim} F$ distributed

according to some c.d.f. F, such that their expectation μ and their variance σ^2 exist and are finite. Then, the standardized version of their average converges in distribution to a standard Normal distribution.
In other words,

$$\frac{\sqrt{n}(\bar{X}_n - \mu)}{\sigma} \xrightarrow{\text{d}} Z \tag{50}$$

as $n \to \infty$, where $Z \sim \mathcal{N}(0,1)$. Another way to express the Central Limit Theorem is to say that, for any $x \in \mathbb{R}$,

$$\lim_{n\to\infty} P\left(\frac{\sqrt{n}(\bar{X}_n - \mu)}{\sigma} \leq x\right) \to \Phi(x). \tag{51}$$

This result is extremely frequently used and invoked to approximate probability statements regarding the average of a collection of i.i.d. random variables when n is 'large'. However, rarely the population variance σ^2 is known. The Central Limit Theorem can be extended to accomodate this case. We have

$$\frac{\sqrt{n}(\bar{X}_n - \mu)}{S} \xrightarrow{\text{d}} Z \tag{52}$$

where $Z \sim \mathcal{N}(0,1)$. Its typical proof involves characteristic functions, that we will study in a few classes.

The Delta Method

If we have a sequence of random variables $X_1, X_2, \ldots, X_n, \ldots$ which converges in distribution to a standard Normal and a differentiable function $g : \mathbb{R} \to \mathbb{R}$, then the Delta Method allows us to find the limiting distribution for the sequence of random variables $g(X_1), g(X_2), \ldots, g(X_n), \ldots$. Assume that

$$\frac{\sqrt{n}(\bar{X}_n - \mu)}{\sigma} \xrightarrow{\text{d}} \mathcal{N}(0,1)$$

and that $g : \mathbb{R} \to \mathbb{R}$ is differentiable with $g'(\mu) \neq 0$. Then,

$$\frac{\sqrt{n}(g(X_n) - g(\mu))}{|g'(\mu)|\sigma} \xrightarrow{\text{d}} \mathcal{N}(0,1).$$

Exercise:
Let $X_1, \ldots, X_n \overset{\text{iid}}{\sim} F$ where F is some c.d.f. such that both the expectation μ and the variance σ^2 of the X's exist and are finite. By the Central Limit Theorem,

$$\frac{\sqrt{n}(\bar{X}_n - \mu)}{\sigma} \xrightarrow{\text{d}} \mathcal{N}(0,1).$$

Consider the function $g(x) = e^x$. Find the limiting distribution of $g(\bar{X}_n)$.

We have $g'(x) = g(x) = e^x > 0$ for any $x \in \mathbb{R}$; thus, $g(\mu) \neq 0$ necessarily. By applying the Delta Method, we have that

$$\frac{\sqrt{n}[g(\bar{X}_n) - g(\mu)]}{|g'(\mu)|\sigma} = \frac{\sqrt{n}\left(e^{\bar{X}_n} - e^{\mu}\right)}{e^{\mu}\sigma} \xrightarrow{\text{d}} \mathcal{N}(0,1).$$

Thus, we can use the distribution

$$\mathcal{N}\left(e^{\mu}, \frac{e^{2\mu}\sigma^2}{n}\right).$$

to approximate probability statements about $g(\bar{X}_n) = e^{\bar{X}_n}$ when n is 'large'.

The Law of Large Numbers

We have so far studied the Central Limit Theorem and a particular mode of convergence, the convergence in distribution. Another important result on the convergence of the average of a sequence of random variables is the *law of large numbers*[6]. Simply put, the law of large numbers says that the sample mean of a sequence of i.i.d. random variables converges to the common expectation of the random variables in the sequence. More precisely, let $X_1, \ldots, X_n, \ldots$ be a sequence of iid random variables with common mean μ. Then, for any $\epsilon > 0$,

$$\lim_{n\to\infty} P(|\bar{X}_n - \mu| > \epsilon) = 0. \tag{53}$$

In this case, we use the notation $\bar{X}_n \xrightarrow{\text{p}} \mu$.

Example:
Suppose that you repeatedly flip a coin which has probability $p \in (0,1)$ of showing head. Introduce the random variables

$$X_i = \begin{cases} 1, & \text{if the i-th flip is head} \\ 0, & \text{if the i-th flip is tails.} \end{cases}$$

Consider the random variable $\bar{X}_n$ describing the observed proportion of heads. Then, as $n \to \infty$, $\bar{X}_n \xrightarrow{\text{p}} \mu$. In English, this means that for any arbitrarily small $\epsilon > 0$, the probability of the event 'the proportion of heads differs from p by more than ϵ' converges to 0 as $n \to \infty$. Let's prove it. By the central limit theorem, we know that

$$\bar{X} \sim \mathcal{N}(\mu, \mu(1-\mu)/n)$$

[6]This is also called the (weak) law of large numbers; the stronger version exists, but we do not discuss this in this course.

therefore, for any $\epsilon > 0$,

$$\begin{aligned} P(|\bar{X}_n - \mu| > \epsilon) &= 1 - P(-\epsilon \leq \bar{X}_n - \mu \leq \epsilon) \\ &= 1 - P\left(-\frac{\epsilon\sqrt{n}}{\sqrt{\mu(1-\mu)}} \leq Z \leq \frac{\epsilon\sqrt{n}}{\sqrt{\mu(1-\mu)}}\right) \\ &= 1 - \phi\left(\frac{\epsilon\sqrt{n}}{\sqrt{\mu(1-\mu)}}\right) + \phi\left(-\frac{\epsilon\sqrt{n}}{\sqrt{\mu(1-\mu)}}\right) \overset{n\to\infty}{\longrightarrow} 0 \end{aligned}$$

Summary: two ways in which Random Variables converge.

We introduced two notions of convergence for random variables: *convergence in distribution* (which we associated to the Central Limit Theorem) and *convergence in probability* (which we associated to the weak law of large numbers). There are other ways of convergence of random variables, all of which have the common goal: What does the distribution of $Y_n = f(X_1, \cdots, X_n)$ look like as we collect more and more samples? ($n \to \infty$)? These results allow us to understand the *precision* or *accuracy* in which our statistics (estimators, necessarily functions of the data) estimate some unknown quantity (usually a parameter). For further study, read a good probability textbook or consider taking a graduate class!

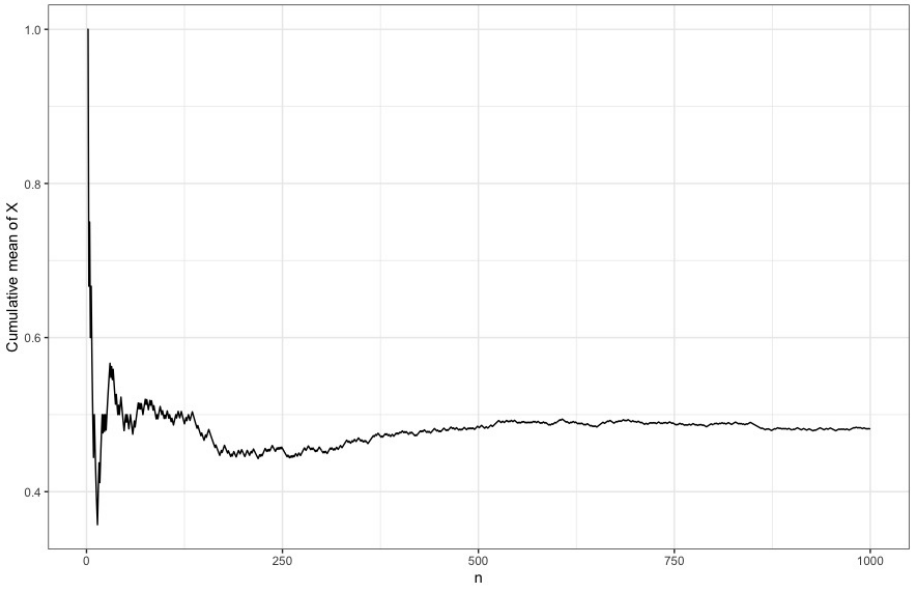

Figure 1: Convergence of $\bar{X}_n$.

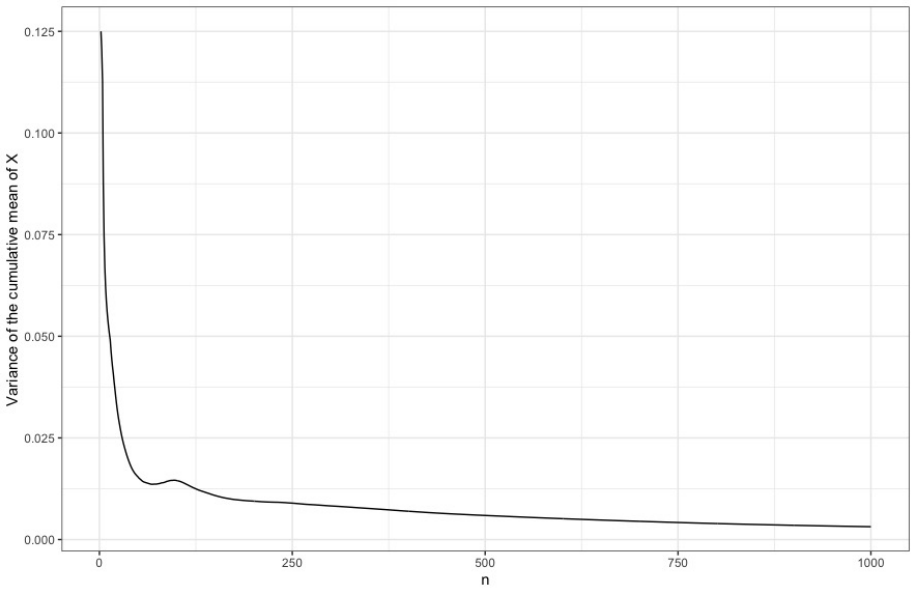

Figure 2: Convergence of $\frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X}_n)^2$.

Figure 3: Convergence of cumulative sample mean and sample variance of Bernoulli trials as a function of n.

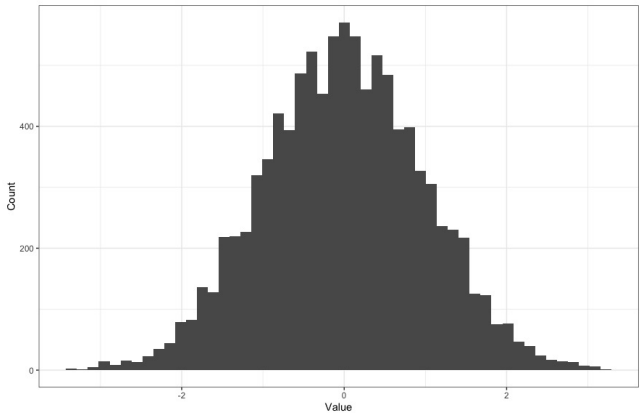

Figure 4: Histogram (counts) of the values taken by $\sqrt{n}(\bar{X}_n - \mu)/\sigma$ for Bernoulli trials with $\mu = 0.5$, and consequently variance $\sigma^2 = \mu(1-\mu)$. Each value is computed with 10^4 Bernoulli trials. As you see, $\sqrt{n}(\bar{X}_n - \mu)/\sigma \xrightarrow{d} \mathcal{N}(0,1)$.