## Contents

1 Intro to Probability and Statistics 5
   1.1 Probability is Not Intuitive .......................... 6

2 Events as Sets 7
   2.1 Introduction ............................................. 7
   2.2 Venn Diagrams ........................................... 8
   2.3 Important Events ........................................ 8
   2.4 Operating on Events ..................................... 9
   2.5 Disjoint Sets ............................................. 9

3 Probability 9
   3.1 How to Assign Probabilities to Events ................... 9
   3.2 Properties of Probability Models ......................... 10

4 Total Probability 11
   4.1 Trees ................................................ 12

5 Counting 13
   5.1 Multiplication Rule .................................... 14
   5.2 Permutations ........................................... 14

6 Counting 15
   6.1 Birthday Paradox: a Permutation Example ............ 15
   6.2 Combinations ........................................... 16

7 Independence 17

8 Conditional Probability 18
   8.1 Independence and Conditional Probability ............ 21

9 Conditional Probability 22
   9.1 Examples ................................................. 22

10 Bayes Law 23
    10.1 Examples ............................................. 24
    10.2 Bayesian Philosophy ................................. 26

11 Poker 27
    11.1 First Deal ........................................... 28
    11.2 Selection of Discards ................................. 29

12 Random Variable 29

13 Probability Mass Functions 30
    13.1 Cumulative Distribution Function (CDF) ............. 32

14 Probability Density Functions 33
15 Expectation
   15.1 Expectation of a Function ..................................................... 38
   15.2 Examples ........................................................................ 39

16 Binomial pmf ........................................................................ 40

17 Uniform Distribution ............................................................... 41

18 Gaussian Distribution ............................................................. 42
   18.1 Cumulative Probability ........................................................... 43
   18.2 Inverse use of Normal Probability .......................................... 45

19 Joint Discrete Random Variables .............................................. 46
   19.1 Marginal Distributions ............................................................ 47
   19.2 Conditional Distributions ....................................................... 48
   19.3 Independence of Random Variables ...................................... 49

20 Joint Continuous Random Variables ....................................... 50
   20.1 Marginal and Conditional pdfs ............................................... 52
   20.2 Independence of Random Variables ...................................... 54

21 Correlation and Covariance ...................................................... 55

22 Linear Combinations of r.v.s .................................................... 58
   22.1 Expected Value for Product .................................................... 58
   22.2 Variance of a Linear Combination ......................................... 59
   22.3 Two r.v.s ........................................................................ 59
   22.4 Many r.v.s ..................................................................... 60

23 Examples ................................................................................ 61

24 Chebychev’s Theorem ............................................................... 64

25 Intro to Statistics .................................................................. 66
   25.1 What is the topic, “statistics”? ............................................... 69

26 Central Limit Theorem / Distribution of Sample Mean .............. 69

27 Using Sample Instead of Population Variance ........................... 72

28 Intro to Estimation ................................................................. 73
   28.1 Estimator Bias .................................................................. 74
   28.2 Estimator Variance ............................................................. 76

29 Confidence Interval on Mean .................................................... 76
   29.1 Two-sided ..................................................................... 77
   29.2 One-sided Bound ................................................................ 77
   29.3 Known Population Standard Deviation ................................. 77
   29.4 Unknown Population Standard Deviation .............................. 79
29.5 Procedure to Determine CI ......................................................... 80

30 Confidence Interval on Variance ........................................... 81

31 Hypothesis Testing ................................................................. 82

32 Hypothesis Tests for Change in Mean ................................... 84

33 Tests for Change in Variance ................................................... 86

34 Hypothesis Tests on Difference in Mean ............................... 89

35 What is a p-value? ................................................................. 92

36 Tests on Proportions ............................................................... 93
    36.1 Gaussian Approximation ................................................ 93
    36.2 Exact Solution .............................................................. 94
1 Intro to Probability and Statistics

Randomness is all around us; in many engineering and scientific applications.

- Communications: A transmitted signal is attenuated, corrupted, and is received with added noise. The attenuation, channel-induced corruption, and noise can be considered to be random.

- Controls: A controls system includes measuring the state of an engineered system and using actuators (e.g., motors) to react to that data. The measurement and the actuation are not exact; both are considered random.

- Manufacturing: There are process variations that are considered random. Manufacturing imprecision causes products to differ; some will even not work, or meet specs, in a way that can be considered random. If cost was no concern, you might measure precisely every single product; but if not, you don’t know for sure whether each product is “good”.

- Economics: Securities prices change in a random fashion; we don’t know in advance how it will change.

- Imaging: Medical images (e.g., CT scans) try to use as little energy as possible due to possible health effects. These images are corrupted by noise. For example, a system may be counting photons which arrive randomly over time, and the number which arrive in a given time period would be considered random.

- Biology: The spread of an infection; the growth of cells; genetic inheritance; can all be considered as random.

- Internet: The spread of an idea (via social network), the sequence of clicks while surfing the web, the arrival (or dropping) of packets at a destination or at a router, the delay between send and receive of a packet, downtime of a server, can all be considered as random. The randomness of passwords and keys are critical to security.

- Algorithms: Many algorithms incorporate randomness on purpose in order to compute a result or to store data.

In all of these applications, we have what we call random variables. These are things which vary across time in an unpredictable manner. Sometimes, these things we truly could never determine beforehand. For example, thermal noise in a receiver is truly unpredictable. In other cases, perhaps we could have determined if we had taken the effort. For example, whether or not a machine is going to fail today could have been determined by a maintenance checkup at the start of the day. But in this case, if we do not perform this checkup, we can consider whether or not the machine fails today as a random variable, simply because it appears random to us.

The study of probability is all about taking random variables and quantifying what can be known about them. Probability is a set of tools which take random variables and
output deterministic numbers which answer particular questions. So while the underlying variable or process may be random, we as engineers are able to ‘measure’ them.

For example:

- The expected value is a tool which tells us, if we observed lots of realizations of the random variable, what its average value would be.
- Probability of the random variable being in an interval or set of values quantifies how often we should expect the random value to fall in that interval or set.
- The variance is a tool which tells us how much we should expect it to vary.
- The correlation or covariance (between two random variables) tells us how closely two random variables follow each other.

The study of probability is simply the study of random variables, to provide tools which allow us to find deterministic measures of what can be known about the random variables.

The study of statistics is about learning about something from noisy data. Statistics is a key element of science, that is, how to test a hypothesis, or estimate an unknown value. As engineers, we build systems that then we need to understand. We do a lot of testing of systems to characterize them. We essentially are scientists of engineered systems, and as such, need to understand one of its key building blocks.

1.1 Probability is Not Intuitive

We are often mislead by our intuition when it comes to understanding probability. Thus a mathematical framework for determining probabilities is critical. To do well in this course you must follow the frameworks we set up to answer questions about probability.

Example: Rare Condition Tests

One example of how our intuition can betray us is in the understanding of tests for rare medical conditions. As an example, consider a common test for genetic abnormalities, the “nuchal translucency” (NT) test, on a 12-week old fetus. (I’m taking these numbers from the internet and I’m not a doctor, so please take this with a grain of salt.) About 1 in 1000 fetuses have such an abnormality. The NT test correctly detects 62% of fetuses that have the abnormality. But, it has a false positive rate of 5%, that is, 5% of normal fetuses are detected by the NT test as positive for the abnormality.

Given that a fetus has tested positive in the NT test, what is the probability that it is, in fact, abnormal?

Here’s my solution, using Bayes’ Law: 1.2%. That is, even after the positive on the NT test, it is highly unlikely that the fetus has a genetic abnormality. True, it is 12 times higher than before the test came back with a positive, but because the condition is so rare, and the false positive rate is reasonably high, the probability of the fetus having the abnormality is still very very low. We will learn Bayes’ Law and how to use it in the first segment of this course.

\[
P[A|NT+] = \frac{P[NT+A]P[A]}{P[NT+A]P[A] + P[NT+A']P[A']} = \frac{(0.001)(0.62)}{(0.001)(0.62) + (0.999)(0.05)} = 0.012
\]

Example: Conjunction Fallacy

Linda is 31 years old, single, outspoken and very bright. She majored in philosophy. As a student,
she was deeply concerned with issues of discrimination and social justice. Before entering the workforce, Linda spent time in the Peace Corps helping women gain access to health care in and around the Congo region of Africa. Which is most likely?

1. Linda is a bank teller.

2. Linda is a bank teller and writes for a feminism blog.

This question is adapted from a question posed to U. of British Columbia undergraduates in 1983 by Amos Tversky and Daniel Kahneman. They found 85% of students rated (2) as more likely than (1). However, if we define \( A \) = the event that Linda is a bank teller, and \( B \) = the event that Linda writes for a feminism blog, then (1) has probability \( P[A] \), and (2) has probability \( P[A \cap B] \). The laws of probability, as we will see, mandate that \( P[A \cap B] \leq P[A] \), with equality only if all bank tellers write for feminism blogs. In short, Tversky and Kahneman noticed that people are swayed by a narrative – the more detail, the more believable it seems.

Example: Probability of Failure in Disasters
Consider a hypothetical nuclear reactor located near an ocean that will leak radiation only if there is both a severe earthquake, and simultaneously a flood in the reactor facility. The probability of a severe earthquake is, on any given day, \( 1/10,000 \), and the probability of a flood in the facility is \( 1/5,000 \). What is the probability the nuclear reactor will leak radiation?

There is a tendency, without thinking about the events in question, to assume they are independent, and thus the probability \( P[A \cap B] = P[A]P[B] \). In this case would lead to a probability of \( \frac{1}{5 \times 10^7} \), or one in 50 million. That would be, about once every 130,000 years. Based on this analysis, an engineer might say, this facility will never fail. However, it would be more accurate to realize the events are closely related, that a tsunami can be both a severe earthquake, and cause major flooding at the same time. The rate of leaks would be close to the probability of a major tsunami, something likely to happen in an average person’s lifetime.

In summary, intuition is not a good way to approach problems in probability and statistics. It is important to follow the methods we will present in this course. It is also important to have lots of practice – repetition will help to train us on the proper procedure for analysis.

Lecture #2

2 Events as Sets

All probability is defined on sets. In probability, we call these sets events. A set is a collection of elements. In probability, we call these outcomes. The words we use for events is slightly different than what you did when you learned sets in a math class, see Table 2.

**Def’n: Event**
A collection of outcomes. Order doesn’t matter, and there are no duplicates.

2.1 Introduction

There are different ways to define an event (set):
Set Theory | Probability Theory | Probability Symbol
--- | --- | ---
universe | sample space (certain event) | \( S \)
element | outcome (sample point) | \( s \)
set | event | \( E \)
disjoint sets | disjoint events | \( E_1 \cap E_2 = \emptyset \)
null set | null event | \( \emptyset \)

Table 1: Set Terminology vs. Probability Terminology

Figure 1: Venn diagrams of sample space \( S \) and two sets \( A \) and \( B \) which (a) are not disjoint, (b) are disjoint, and (c) form a partition of \( S \).

- List them: \( A = \{0, 5, 10, 15, \ldots\}; B = \{\text{Tails, Heads}\} \).
- As an interval: \([0, 1], [0, 1), (0, 1), (a, b)\). Be careful: the notation overlaps with that for coordinates!
- An existing event set name: \( \mathbb{N}, \mathbb{R}^2, \mathbb{R}^n \).
- By rule: \( C = \{x \in \mathbb{R} | x \geq 0\}, D = \{(x, y) \in \mathbb{R}^2 | x^2 + y^2 < R^2\} \).

### 2.2 Venn Diagrams

Venn diagrams can be used to pictorially show whether or not there is overlap between two or more sets. They are a good tool for helping remember some of the laws of probability. We don’t use them in proofs, however, they’re particularly good to develop intuition.

### 2.3 Important Events

Here’s an important event: \( \emptyset = \{\} \), the null event or the empty set.

Here’s the opposite: \( S \) is used to represent the set of everything possible in a given context, the sample space.

- \( S = B \) above for the flip of a coin.
- \( S = \{1, 2, 3, 4, 5, 6\} \) for the roll of a (6-sided) die.
- \( S = \{\text{Adenine, Cytosine, Guanine, Thymine}\} \) for the nucleotide found at a particular place in a strand of DNA.
- \( S = C, \text{ i.e., } \) non-negative real numbers, for your driving speed (maybe when the cop pulls you over).
2.4 Operating on Events

We can operate on one or more events:

- Complement: \( A^c = \{ x \in S | x \notin A \} \). We must know the sample space \( S \)!
- Union: \( A \cup B = \{ x | x \in A \text{ or } x \in B \} \). Merges two events together.
- Intersection: \( A \cap B = \{ x | x \in A \text{ and } x \in B \} \). Limits to outcomes common to both events.

DO NOT use addition to represent the union, and DO NOT use multiplication to represent the intersection. An example of why this is confusing:

\[
\{1\} + \{1\} = \{1\}.
\]

This leads to one of the most common written mistakes – exchanging unions and plusses when calculating probabilities. Don’t write \( P[A] + P[B] \) when you really mean \( P[A \cup B] \). Don’t add sets and numbers: for example, if \( A \) and \( B \) are sets, don’t write \( P[A] + B \).

2.5 Disjoint Sets

**Def’n:** Disjoint

Two events \( A_1 \) and \( A_2 \) are disjoint if \( A_1 \cap A_2 = \emptyset \).

**Def’n:** Mutually exclusive

Multiple events \( A_1, A_2, A_3, \ldots \) are mutually exclusive if every pair of events is disjoint.

Some disjoint events: \( \{1, 2\} \) and \( \{6\} \); \( A \) and \( A^c \); \( A \) and \( \emptyset \).

3 Probability

You’re familiar with functions, like \( f(x) = x^2 \), which assign a number output to each number input. Probability assigns a number output to each event input.

3.1 How to Assign Probabilities to Events

As long as we follow three intuitive rules (axioms) our assignment can be called a *probability model*.

**Axiom 1:** For any event \( A \), \( P[A] \geq 0 \).

**Axiom 2:** \( P[S] = 1 \).

**Axiom 3:** For any two disjoint events \( A_1 \) and \( A_2 \),

\[
P[A_1 \cup A_2] = P[A_1] + P[A_2].
\]

The final axiom in the Walpole book is more complicated:

Axiom 3: If \( A_1, A_2, A_3, \ldots \) is a sequence of mutually exclusive events, then

\[
P[A_1 \cup A_2 \cup A_3 \cup \cdots] = P[A_1] + P[A_2] + P[A_3] + \cdots
\]
However, one may prove the more complicated Walpole Axiom 3 from the three axioms given above. For details, see [Billingsley 1986].

Example: DNA Measurement

Consider the DNA experiment above. We measure from a strand of DNA its first nucleotide. Let’s assume that each nucleotide is equally likely. Using axiom 3,

$$P[\{a, c, g, t\}] = P[\{a\}] + P[\{c\}] + P[\{g\}] + P[\{t\}]$$

But since $P[\{a, c, g, t\}] = P[S]$, by Axiom 2, the LHS is equal to 1. Also, we have assumed that each nucleotide is equally likely, so

$$1 = 4P[\{a\}]$$

So $P[\{a\}] = 1/4$.

**Def’n: Discrete Uniform Probability Law**

In general, for event $A$ in a discrete sample space $S$ composed of equally likely outcomes,

$$P[A] = \frac{|A|}{|S|}$$

### 3.2 Properties of Probability Models

1. $P[A^c] = 1 - P[A]$. Proof:
   
   First, note that $A \cup A^c = S$ from above. Thus
   
   $$P[A \cup A^c] = P[S]$$

   Since $A \cap A^c = \emptyset$ from above, these two events are disjoint.

   $$P[A] + P[A^c] = P[S]$$

   Finally from Axiom 2,

   $$P[A] + P[A^c] = 1$$

   And we have proven what was given.

   Note that this implies that $P[S^c] = 1 - P[S]$, and from axiom 2, $P[\emptyset] = 1 - 1 = 0$.

2. For any events $E$ and $F$ (not necessarily disjoint),

   $$P[E \cup F] = P[E] + P[F] - P[E \cap F]$$

   Essentially, by adding $P[E] + P[F]$ we double-count the area of overlap. Look at the Venn diagram in Figure 1(a). The $-P[E \cap F]$ term corrects for this. Proof: Do on your own using these four steps:

   (a) Show $P[A] = P[A \cap B] + P[A \cap B^c]$.
   (b) Same thing but exchange $A$ and $B$.
   (c) Show $P[A \cup B] = P[A \cap B] + P[A \cap B^c] + P[A^c \cap B]$. 


(d) Combine and cancel.

3. Subset rule: If \( A \subset B \), then \( P[A] \leq P[B] \). Proof:
Let \( B = (A \cap B) \cup (A^c \cap B) \). The two events, \((A \cap B)\) and \((A^c \cap B)\), are disjoint since \( A \cap B \cap A^c \cap B = \emptyset \). Thus:

\[
P[B] = P[A \cap B] + P[A^c \cap B] = P[A] + P[A^c \cap B] \geq P[A]
\]

Note \( P[A \cap B] = P[A] \) since \( A \subset B \), and the inequality in the final step is due to the Axiom 1.


5. Conjunction bound: \( P[A \cap B] \leq P[A] \). Proof: \( A \cap B \) is a subset of \( A \). This then follows from the subset rule above. See the “Conjunction Fallacy” from lecture 1.

---

Lecture #3

4 Total Probability

We mentioned in passing the very important rule that for any two events \( A \) and \( B \),

\[
P[A] = P[A \cap B] + P[A \cap B^c]. \tag{1}
\]

This is a simple example of the law of total probability applied to the partition, \( B \) and \( B^c \). Now, let’s present the more general law of total probability. First, the definition of partition:

**Def’n:** Partition
A countable collection of mutually exclusive events \( C_1, C_2, \ldots \) is a partition if \( \bigcup_{i=1}^{\infty} C_i = S \).

Examples:
1. For any set \( C \), the collection \( C, C^c \).
2. The collection of all simple events for countable sample spaces. Eg., \{1\}, \{2\}, \{3\}, \{4\}, \{5\}, \{6\}.

**Def’n:** Law of Total Probability
For a partition \( C_1, C_2, \ldots \), and event \( A \),

\[
P[A] = P[A \cap C_1] + P[A \cap C_2] + P[A \cap C_3] + \cdots \tag{2}
\]

Note: You must use a partition!!!, but \( A \) can be any event.

**Example: Die and Odd**
Let’s consider a fair 6-sided die again, but define the following sets: \( E_1 = \{1, 2, 3\} \), \( E_2 = \{4, 5\} \), and \( E_3 = \{6\} \). Also let \( F \) be the event that the number rolled is odd. First, find:

1. \( P[E_i \cap F] \) for \( i = 1, 2, 3 \).
2. Do $E_1, E_2,$ and $E_3$ form a partition?

3. Use (1.) and (2.) to calculate $P[F]$.

**Solution:**

1. Using the discrete uniform probability law, $P[E_1 \cap F] = 2/6$, $P[E_2 \cap F] = 1/6$, and $P[E_3 \cap F] = 0/6$.

2. Well, $E_1 \cup E_2 \cup E_3 = \{1, 2, 3, 4, 5, 6\} = S$ and the intersection of any two of the $E_i$ sets is the null set. So they do form a partition.

3. By the law of total probability, $P[F] = P[E_1 \cap F] + P[E_2 \cap F] + P[E_3 \cap F] = 2+1+0/6 = 0.5$.

**Example: Traits**

People may inherit genetic trait A or not; similarly, they may inherit genetic trait B or not. People with both traits A and B are at higher risk for heart disease. For the general population, $P[A] = 0.50$, and $P[B] = 0.10$. We know that $P[A \cap B^c]$, that is, the probability of having trait A but not trait B, is 0.48. What is the probability of having both trait A and B?

**Solution:** Use the law of total probability with partition $B$ and $B^c$:

$$P[A] = P[A \cap B] + P[A \cap B^c].$$

We can rearrange to: $P[A \cap B] = P[A] - P[A \cap B^c]$. Using the given information:

$$P[A \cap B] = 0.50 - 0.48 = 0.02.$$

A Venn diagram is particularly helpful here.

**4.1 Trees**

A graphical method for organizing information about how a multiple-stage experiment can occur. See Figure 2.

**Example: Customers Satisfied by Operator**

A tech support phone bank has exactly three employees. A caller is randomly served by one of the three. After a customer is served, he or she is surveyed. TData is kept as to the employee (by number, 1, 2, or 3) and whether the caller was satisfied (S). The data shows:

- $P[S \cap 1] = 0.40$
- $P[S \cap 2] = 0.20$
- $P[S \cap 3] = 0.18$

Questions:

1. What is the probability that a caller is satisfied?

2. Are the probabilities of a caller being served by each of the three employees equal?

**Solution:**
1. Since 1, 2, and 3 are a partition, \( P[S] = P[S \cap 1] + P[S \cap 2] + P[S \cap 3] = 0.4 + 0.2 + 0.18 = 0.78 \).

2. The three probabilities, \( P[1] \), \( P[2] \), and \( P[3] \) can’t be equal. Employee 1 has too high of a probability. By the conjunction bound, \( P[1] \geq P[S \cap 1] = 0.40 \). Note you can get the conjunction bound by considering that

\[
P[1] = P[1 \cap S] + P[1 \cap S^c] \geq P[1 \cap S].
\]

The latter line because the probability \( P[1 \cap S^c] \geq 0 \) by probability Axiom 1.

In any case, the probability a caller is served by 1 is at least 40%. If the three employees were equally likely, they’d each have probability of 0.333.

---

**Lecture #4**

## 5 Counting

This section covers how to determine the size, \( i.e., \) the number of outcomes, in the sample space or in an event.

**Example: Tablet Specs**

A tablet manufactured by Pear Computer Co can have 64 GB, 32 GB, or 16 GB of memory, and can either have 4G connectivity or not. How many different versions of tablet does Pear make?

**Solution:** The solution is 6: (64, 4G), (64, No), (32, 4G), (32, No), (16, 4G), (16, No).
5.1 Multiplication Rule

What is the general formula for this, so we don’t have to list outcomes?

**Def’n: Multiplication Rule**

If one operation can be performed in \( n_1 \) ways, and for each of those ways, a second operation can be performed in \( n_2 \) ways, then there are \( n_1 n_2 \) ways to perform the two operations together.

**Example: Form Entry**

A web form asks for 1) a person’s state of birth (or n/a if not born in the US) and 2) a pin code between 0 and 9999. How many different entries are possible?

**Solution:** There are 50 states and 1 n/a, for a total of 51 ways to answer the first question, and 10,000 ways to answer the second question. By the multiplication rule, there are \( 51 \times 10^4 \) different entries possible.

We don’t need to stop at two “operations” – many problems involve more.

**Def’n: Multiplication Rule**

If a first operation can be performed in \( n_1 \) ways, and for each of those ways, a second operation can be performed in \( n_2 \) ways, and for each of those ways to do both operations together, there are \( n_3 \) ways to do a third operation, and so forth through the \( k \)th operation, then there are \( n_1 n_2 \cdots n_k \) ways to perform the \( k \) operations together.

**Example: Phonics testing**

The state of Utah tests early elementary students on their ability in phonics by presenting them with a (potentially fake) three letter word and asking them to pronounce it. The first letter is a consonant, the second a vowel (excluding y), and the third a consonant. The consonant is allowed to repeat. How many different questions could there be?

**Solution:** There are 26 letters, five of them vowels (a e i o u). Thus there are 21 consonants. The first letter must be one of 21, the second letter must be one of five vowels, and the third letter is one of 21 consonants again. The answer is \( 21 \times 5 \times 21 = 2205 \).

**Example: Turn in order**

There are 115 students in a class taking an exam. Consider the first, second, and third person to turn in their exam. How many ways can this happen? Assume order is important, and each person can turn in their exam only once.

**Solution:** The first person can be any of the 115. The second person can’t be the same as the 1st, so is one of the remaining 114. The third is similarly one of the remaining 113. Thus there are \( 115 \times 114 \times 113 = 1,481,430 \) ways for this order to happen.

5.2 Permutations

In general, when we take elements of a set (such as the students in a class, above) and order a few (or all) of them, it is called a **permutation**.

The number of ways to put \( r \) elements in order, from a set of \( n \) unique possibilities, is denoted \( n P_r \), and is calculated as:

\[
n P_r = \frac{n!}{(n-r)!} = n \cdot (n-1) \cdot (n-2) \cdots (n-r+1)
\]
Note $0! = 1! = 1$.

If you’re going to order ALL of the elements of the set, then $r = n$, so there are $n \cdot (n-1) \cdot (n-2) \cdots 1 = n!$ ways to order them.

Example: Cards in the deck
Order is important in a card deck because it affects which player will get each card.

1. How many ways are there to order the 52 cards in a standard card deck?
2. How many ways are there to pick 4 cards, in order which you pick them, from a standard card deck?

Example: Three pills
You are sick and there are three pills in front of you, one which will do nothing (N), one which will kill you (K), and one which will cure you (C). You have no idea which one is which. You will put the three pills in order, and take one at a time, and stop when you either get cured or killed.

1. How many ways are there to put the pills in order from first to last? (You might stop before you get to the end, but just answer how many ways there are to put them in order.
2. How many ways are there to order the three pills that will cure you?

Lecture #5

6 Counting

6.1 Birthday Paradox: a Permutation Example

Example: Birthday Paradox
What is the probability that two people in this room will have the same birthday? Assume that “birthday” doesn’t include year, and that each day of the year is equally likely (and exclude the leap day), and that each person’s birthday is independent.

Solution:

1. How many ways are there for $n$ people to have their birthday? Answer: Each person can have their birthday happen in one of 365 ways, assuming 365 days per year. So: $365^n$.
2. How many ways are there to have all $n$ people have unique birthdays? The first one can happen in 365 ways, the second has 364 left, and so on: $365 \cdot 364 \cdot \ldots \cdot 365 - n)$. 
3. Discrete uniform probability law:

$$P[\exists \text{no duplicate birthdays}] = \frac{365!}{(365 - n)!} \cdot \frac{1}{365^n}$$

See Fig. 3, which ends at $n = 30$. At $n = 50$, there is only a 3% chance. At $n = 60$, there is a 0.6% chance. At 110 students (our entire class) the probability is on the order of $10^{-8}$. This is called the birthday paradox because most people would not guess that it would be so likely to have two people with a common birthday in a group as small as 20 or 30.
Figure 3: The probability of having no people with matching birth dates in a group of people, vs. the number in the group. Note the probability is less than half when the class is bigger than 23.

6.2 Combinations

Permutations require that the order matters. For example, in the three pills example where you take one pill at a time until you die or are cured, where K is the pill that will kill you, and C is the pill that will cure you, KC is different than CK (as different as life and death!). Certainly in words or license plates, as well, the order of the chosen letters matter. However, sometimes order doesn’t matter, ie, if you are counting how many different sets could exist (since order doesn’t matter in a set). As another example, in many card games you are dealt a hand which you can re-order as you wish, thus the order you were dealt the hand doesn’t matter.

**Def’n: Combinations**
The number of (unordered) combinations of size $r$ taken from $n$ distinct objects is

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

We can think of permutations as overcounting in the case when order doesn’t matter. If order doesn’t matter and we order $r$ distinct objects from a set of $n$, we’d count them as $n!/(n-r)!$. But for each ordering of length $r$, somewhere else in the permutation set is a different ordering of the same $r$ objects. Since there are $r!$ ways to order these $r$ objects, the formula $nP_r$ overcounts the number of combinations by a factor of $r!$, or:

$$\binom{n}{r} = nP_r \frac{1}{r!} = \frac{n!}{(n-r)!r!} \frac{1}{r!}.$$  

**Example: 5-card Poker**

How many ways are there to be dealt a 5-card hand from a standard 52-card deck, where order doesn’t matter? Note that in poker, all 52 cards are distinct.

**Solution:** We are taking $r = 5$ cards from a $n = 52$ distinct objects. Thus

$$\binom{52}{5} = \frac{52!}{(52-5)!5!} = \frac{52 \cdot 51 \cdot 50 \cdot 49 \cdot 48}{5 \cdot 4 \cdot 3 \cdot 2 \cdot 1} = 2,598,960$$
Example: Poker
Evaluate the probabilities of being dealt a “flush” in a five card poker hand. Note the The standard deck has 52 cards, 13 cards of each suit; and there are four suits (hearts, diamonds, clubs, and spades). The thirteen cards are, in order, A, 2, 3, 4, 5, 6, 7, 8, 9, 10, J, Q, K. The ace (A) can also be higher than the king (K). A flush is any 5 cards of the same suit. Note that this definition of flush includes the royal flush.

Solution:
1. How many different hands are there? Order doesn’t matter in a poker hand, so this is a combination problem. A: 52 choose 5, or 2,598,960.

2. \( P[\text{Flush}] \) (A flush is any 5 cards of the same suit, not including any straight flushes.) There are really two “operations” you do to build a flush. First, you select the suit. There are four suits, and you choose one. Then there are 13 of each suit, of which you choose five cards. so

\[
\binom{13}{5} \cdot \binom{4}{1} = 1287(4) = 5148
\]

ways to have a flush.

Example: Pass / Try Again / Fail
There are 9 PhD students who take a qualifying exam. Their grade will be, “pass”, “retake”, or “fail”. The program has determined that two students will pass, three will retake, and four will fail. How many ways are there for the eight students to receive grades?

Solution: There are two operations to be performed. First, we must select the two students from 9 who will pass, which can occur in \( \binom{9}{2} \) ways. Second, we must select the three students from the remaining 7 who will get a retake grade, which can occur in \( \binom{7}{3} \) ways. The remaining four will fail. Thus there are

\[
\binom{9}{2} \cdot \binom{7}{3} = \frac{9!}{(9-2)!2!} \cdot \frac{7!}{(7-3)!3!} = \frac{9!}{2!3!4!} = 1260
\]

Note there is a reason the 7! cancels in the two fractions. This is my preferred way of doing problems which the Walpole book describes as “partitioning \( n \) objects into \( r \) cells”, given in Theorem 2.5.

Lecture #6

7 Independence

Def’n: Independence of a Pair of Sets
Sets \( A \) and \( B \) are independent if and only if \( P[A \cap B] = P[A] P[B] \).

Example: Parking and Speeding.
The probability you get a ticket for an expired parking meter on any given day is 1/50. The probability that you get a speeding violation on any given day is 1/200. The probability that you
get a parking meter ticket and speeding violation on the same day is 1/10,000. Are the two events independent?

Solution: The question is the same as evaluating the truth of:

\[ P[A \cap B] = P[A]P[B] \]

where A is the event you get a parking meter ticket, and B is the event you get a speeding violation. Note \( A \cap B \) is the event both happen in the same day.

\[ \frac{1}{10000} = \frac{1}{50200} \]

This statement is true. Therefore the two events are independent.

Example: Set independence
Consider \( S = \{1, 2, 3, 4\} \) with equal probability, and events \( A = \{1, 2\} \), \( B = \{1, 3\} \), \( C = \{4\} \).

1. Are \( A \) and \( B \) independent?
2. Are \( B \) and \( C \) independent?

Solution:

1. \( P[A \cap B] = P[\{1\}] = 1/4 = (1/2)(1/2) \). Yes, \( A \) and \( B \) are independent.

2. \( P[B \cap C] = P[\emptyset] = 0 \neq (1/2)(1/4) \). \( B \) and \( C \) are not independent.

Are \( B \) and \( C \) disjoint? Yes; but they are not independent. Independence is NOT about non-overlapping; orthogonal might be a better word. Actually, as we will show when discussing conditional independence, two sets are independent if knowing whether one has happened does not affect the probability of the other.

8 Conditional Probability

Def'n: Conditional Probability, \( P[A|B] \)
The conditional probability of event \( A \) given that event \( B \) has occurred, denoted \( P[A|B] \), is defined as:

\[ P[A|B] = \frac{P[A \cap B]}{P[B]} \]

whenever \( P[B] > 0 \), and is undefined in the case when \( P[B] = 0 \).

You can read the vertical bar as the word “given”. There is an implicit “has occurred” at the end of that phrase (after the given event \( B \)).

Notes:

1. We’re defining a new probability model, knowing more about the world. Instead of \( P[\cdot] \), we call this model \( P[\cdot|B] \). See Figure 4. All of our Axioms STILL APPLY, but with \( B \) as the sample space.
Figure 4: The original sample space shown in (a) is reduced to only the outcomes in event $B$ when considering the probability of any event conditioned on $B$, as in (b). For example, part of the set $A$ is no longer possible; the remaining outcomes in $A$ can be called $A|B$.

2. NOT TO BE SAID OUT LOUD because its not mathematically true in any sense. But you can remember which probability to put on the bottom, by thinking of the $|$ as $/$ – you know what to put in the denominator when you do division.

We use this definition to get the very important result, by multiplying both sides by $P[B]$, that

$$P[A \cap B] = P[A|B] P[B]$$

This is true for ANY two sets $A$ and $B$. Recall $P[A \cap B] = P[A] P[B]$ if and only if $A$ and $B$ are independent. Thus, by the way, if $A$ and $B$ are independent, then $P[A|B] = P[A]$.

**Relationship to Multiplication Rule** The rule in Equation (5) is an extension of the multiplication rule, which was for counting events, to probabilities of events. Remember the definition of the multiplication rule says that “If there are $n_1$ ways of doing operation 1, and for each of those ways, there are $n_2$ ways of doing operation 2, then there are $n_1 n_2$ ways of doing the two together”. In Equation (5), it essentially says,

- If there is a probability $P[B]$ of a first thing ($B$) happening, and when $B$ occurs, there is a probability of $P[A|B]$ of a second thing ($A$) happening, then there is a probability $P[A \cap B] = P[A|B] P[B]$ of both events happening.

**Tree Diagram for Conditional Probabilities** The relationship in (5) makes explicit why we can multiply probabilities in a tree together to get the probability of events jointly occurring. See Figure 5.

**With Law of Total Probability** Given that $B$ occurs, now we know that either $A \cap B$ occurs, or $A^c \cap B$ occurs. For events $A$ and $B$, using the law of total probability and the partition $A, A^c$,

$$P[A|B] \triangleq \frac{P[A \cap B]}{P[B]} = \frac{P[A \cap B]}{P[A \cap B] + P[A^c \cap B]}$$

Note: Conditional probability almost always has the form, $\frac{x}{x+y}$. If $P[A|B] = \frac{x}{x+y}$ then $P[A^c|B] = \frac{y}{x+y}$. Note the two terms add to one.

**Example: Coin then Die**
Your assistant rolls a fair coin. If it is heads, he rolls a fair 4-sided die. If it is tails, he rolls a
Figure 5: The probability written onto each branch is the probability conditioned on all the events listed to its left. Thus the probability of the intersection of all of the events from root to a “leaf” (the ◦ at the right end of the branch) is the product of the probabilities written from root to leaf.

fair 6-sided die. In either case, he emails you the number on the die, but not what die he rolled or the coin outcome. You get the number 3 by email. What is the probability that his coin toss was heads?

Solution: Draw a tree diagram to help with this one. $A = \text{heads on the coin}$, $B = \text{getting the number 3}$. The probability is

$$P[A|B] = \frac{P[A \cap B]}{P[A \cap B] + P[A^c \cap B]}$$

Example: Three Card Monte

(Credited to Prof. Andrew Yagle, U. of Michigan.)

There are three two-sided cards: red/red, red/yellow, yellow/yellow. The cards are mixed up and shuffled, one is selected at random, and you look at one side of that card at random. You see red. **What is the probability that the other side is red?**

Three possible lines of reasoning on this:

1. Bottom card is red only if you chose the red/red card: $P = 1/3$.

2. You didn’t pick the yellow/yellow card, so either the red/red card or the red/yellow card: $P = 1/2$.

3. There are five sides which we can’t see, two red and three yellow: $P = 2/5$.

Which is correct?
Solution: BR = Bottom red; TR = Top red; BY = Bottom yellow; TY = Top yellow.

\[
P[BR|TR] = \frac{P[BR \text{ and } TR]}{P[TR]} = \frac{P[BR \text{ and } TR]}{P[BR \text{ and } TR] + P[BY \text{ and } TR]}
\]
\[
= \frac{2/6}{2/6 + 1/6} = \frac{1/3}{1/2} = 2/3.
\]

8.1 Independence and Conditional Probability

We know that for any two sets A and B, that \( P[A \cap B] = P[A|B] P[B] \). Recall that independent sets have the property \( P[A \cap B] = P[A] P[B] \). So, independent sets also have the property that

\[
\]
\[
P[A|B] = P[A]
\]
as long as \( P[B] > 0 \).

Thus the following are equivalent:

1. \( P[A \cap B] = P[A] P[B] \),
2. \( P[A|B] = P[A] \), and
3. \( P[B|A] = P[B] \),

If one is true, all of them are true. If one is false, all are false.

Example: Three different dice

There are three dice in a box, one a four-sided die, one a six-sided die, and one a 10-sided die. Each is fair, that is, each number on a die is equally likely to come up when rolled. One die is picked at random from the three, and then rolled.

1. What is the sample space? How many elements does it have?
2. What is the probability of getting the 4-sided die and rolling a 1?
3. What is the probability of rolling a 1?
4. What is the probability of rolling a 10?

Solution:

1. The sample space could be just \( \{1, \ldots, 10\} \). Or, it could be all of the dice-number pairs, eg, \( \{(4\text{-sided}, 1), \ldots (4\text{-sided}, 4), (6\text{-sided}, 1), \ldots (6\text{-sided}, 6), (10\text{-sided}, 1), \ldots (10\text{-sided}, 10) \} \) The former has 10, the latter 20.

2. The probability of getting the 4-sided die and rolling a 1 is

\[
P[\text{roll} = 1 \cap 4\text{-sided}] = P[4 - \text{sided}] P[\text{roll} = 1 | 4 - \text{sided}] = (1/3)(1/4) = 1/12.
\]
3. There are three ways of rolling a 1. Using the law of total probability:

\[ P[\text{roll} = 1] = P[\text{roll} = 1 \cap 4\text{sided}] + P[\text{roll} = 1 \cap 6\text{sided}] + P[\text{roll} = 1 \cap 10\text{sided}] \]

We have the \( P[\text{roll} = 1 \cap 4\text{sided}] \) above. Similarly, \( P[\text{roll} = 1 \cap 6\text{sided}] = (1/3)(1/6) \) and \( P[\text{roll} = 1 \cap 10\text{sided}] = (1/3)(1/10) \). So

\[ P[\text{roll} = 1] = \frac{1}{12} + \frac{1}{18} + \frac{1}{30} \approx 0.1722 \]

4. There’s only one way of rolling a 10.

\[ P[\text{roll} = 10] = P[\text{roll} = 10 \cap 10\text{sided}] = P[10\text{sided}] P[\text{roll} = 10 | 10\text{sided}] = (1/3)(1/10) = 1/30. \]

Note a tree diagram is really useful for this problem.

9 Conditional Probability

9.1 Examples

Many examples fit into the general shape of the tree diagram in Figure 6.

![Tree Diagram](image)

Figure 6: First, an event \( A \) happens with probability \( P[A] \) (or doesn’t happen with probability \( 1 - P[A] \)). Next, event \( B \) happens (or does not happen) with given conditional probabilities (conditioned on \( A \) or \( A^c \). The joint probabilities are given at the leaves of the tree; two joint probabilities must be summed to find either \( P[B] \) or \( P[B^c] \).

Recall that \( P[A \cap B] = P[A|B] P[B] \) for any two sets \( A \) and \( B \).

**Example: Target Marketing**

If I’ve moved into a new home in the last month, I have a 90% chance of buying furniture at BigBox furniture store. If I haven’t moved into a home that recently, I still have a 5% chance of buying furniture at this store. BigBox thinks they can sell the names and addresses of people
who buy furniture as a list of people who have recently moved. Assume people move, on average, with probability 1/40 in a given month. What is the probability that a person buys furniture at BigBox and they have recently moved? What is the total probability that a person buys furniture at BigBox?

**Solution:** Let $F$ be the event that I buy furniture at BigBox. Let $M$ be the event I moved into a new home in the last month. We can write down from the example:

- $P[F|M] = 0.90$
- $P[F|MC] = 0.05$
- $P[M] = 0.025$

a) Complete the tree diagram to see that

$$P[F \cap M] = P[F|M] P[M] = (0.9)(0.025) = 0.0225$$

b) From the law of total probability,

$$P[F] = P[F \cap M] + P[F \cap MC] = 0.0225 + P[F|MC] P[MC] = 0.0225 + (0.05)(0.975) = 0.071$$

---

**Lecture #8**

10 **Bayes Law**

Bayes Law is simply a re-writing of the conditional probability definition. Given any two events $A$ and $B$,

$$P[B|A] = \frac{P[B \cap A]}{P[A]}$$

(6)

The first line can be rewritten as $P[B \cap A] = P[B|A] P[A]$. But also note that we could have used the definition of conditional probability to write:

$$P[A|B] = \frac{P[B \cap A]}{P[B]}$$

So we can also write the joint probability as $P[B \cap A] = P[A|B] P[B]$. Thus $P[B|A] P[A] = P[A|B] P[B]$. The simplest definition of Bayes’ Law is to take this and divide both sides by $P[A]$ to get:

$$P[B|A] = \frac{P[A|B] P[B]}{P[A]}$$

(7)

In short, Bayes’ Law is a way to convert $P[A|B]$, to the other way around, $P[B|A]$, using the probabilities of the two events $A$ and $B$.

Often, we won’t be given $P[A]$. Instead, we’ll have to find it ourselves using the partition, $\{B, BC\}$, and the law of total probability:

In this case, Bayes’ Law becomes:


(8)

The Walpole book writes Bayes’ Law in its most drawn-out form, i.e., for the case where the partition has \( k \) sets \( B_1, B_2, \ldots, B_k \), instead of just the two sets \( B \) and \( B^C \). In this case, the conditional probability of the \( r \)th set, \( B_r \), is

\[ P[B_r|A] = \frac{P[A|B_r] P[B_r]}{P[A|B_1] P[B_1] + P[A|B_2] P[B_2] + \cdots + P[A|B_k] P[B_k]} \]  

(9)

There’s no magic, in Bayes’ Law. It is just a rewriting of the conditional probability definition, additionally using the law of total probability.

### 10.1 Examples

#### Example: Neural Impulse Actuation

A embedded sensor is used to monitor a neuron in a human brain. We monitor the sensor reading for 100 ms and see if there was a spike within the period. If the person thinks of flexing his knee, we will see a spike with probability 0.9. If the person is not thinking of flexing his knee (due to background noise), we will see a spike with probability 0.01. For the average person, the probability of thinking of flexing a knee is 0.001 within a given period.

1. What is the probability that we will measure a spike? Answer: Let \( S \) be the event that a spike is measured, and its complement as \( NS \). Let the event that the person is thinking about flexing be event \( T \), and its complement is event \( NT \).

   \[
P[S] = P[S|T] P[T] + P[S|NT] P[NT] = 0.9 \times 0.001 + 0.01 \times 0.999 = 0.0009 + 0.00999 = 0.01089\]

2. What is the probability that the person wants to flex his knee, given that a spike was measured? Answer:

   \[
P[T|S] = \frac{P[T \cap S]}{P[S]} = \frac{P[S|T] P[T]}{P[S]} = \frac{0.9 \times 0.001}{0.01089} = 0.0826\]

3. What is the probability that the person wants to flex his knee, given that no spike was measured? Answer:

   \[
P[T|NS] = \frac{P[T \cap NS]}{P[NS]} = \frac{P[NS|T] P[T]}{P[S]} = \frac{0.1 \times 0.001}{1 - 0.01089} \approx 1.01 \times 10^{-4}\]
It wouldn’t be a good idea to create a system that sends a signal to flex his knee, given that a spike was measured. Some other system design should be considered.

**Def’n: a priori**
prior to observation

For example, prior to observation, we know \( P[T] = 0.001 \), and \( P[NT] = 1 - 0.001 = 0.999 \).

**Def’n: a posteriori**
after observation

For example, after observation, we know \( P[T|S] = 0.0826 \), and \( P[NT|NS] = 1 - 0.0001 = 0.9999 \).

We did another example on the first day of class: refer to the *Rare Condition Tests* example.

Another one from the Mlodinow reading:

**Example: Mlodinow, p117**
What is the probability that an asymptomatic woman between 40 and 50 years of age who has a positive mammogram actually has cancer? You know that 7 percent of mammograms show cancer when there is none, and the actual incidence for this population is 0.8 percent, and the false-negative rate is 10 percent.

Note: False positive: Test is positive, but it is a “false alarm”, because the person does not have the disease. False negative: The test did not discover the disease, but the person actually has it.

**Solution:** Let \( M \) be the event that a mammogram is positive. Let \( C \) be the event that the person has cancer. Here is the provided information:

1. \( P[M|C^C] = 0.07 \). The “false positive rate”.
2. \( P[C] = 0.008 \). The prior information.
3. \( P[M|C] = 0.1 \).

Applying Bayes’ Law:

\[

From 3., \( P[M|C] = 1 - P[M^C|C] = 1 - 0.1 = 0.9 \). From 2., \( P[C^C] = 0.992 \). Thus

\[
P[C|M] = \frac{(0.9)(0.008)}{(0.9)(0.008) + (0.07)(0.992)} = 0.094.
\]

Note that it is ALWAYS possible to design a test to have a 0% false positive rate, or to have a 0% false negative rate.

- 0% false positive rate: Make the test always return negative.
- 0% false negative rate: Make the test always return positive.

Both are meaningless tests. But they point out, if you design a test, don’t simply ask for one rate to describe the accuracy of the test. You need BOTH the false negative and false positive (or their complements) to evaluate the effectiveness of a test.
Example: Boot Sequence
A computer at boot time must pass (P) three tests for a certain OS; failure (F) of any test stops the boot sequence. A certain type of computer passes the first test 90% of the time, passes the second test 80% of the time, and passes the third test 70% of the time. Given that the boot sequence stopped, what is the probability that it failed the first test?

Figure 7: A computer at boot time must pass (P) three tests for a certain OS; failure (F) of any test stops the boot sequence.

Solution: See Figure 7. Let’s call $P_1, P_2, P_3$ the events that the computer passes the 1st, 2nd, and 3rd tests, respectively; $F_1, F_2, F_3$ are that it failed those tests. Let $BS$ be the event that the boot sequence stopped. Then the question is asking for $P[F_1|BS]$. Using Bayes Law,

$$P[F_1|BS] = \frac{P[F_1 \cap BS]}{P[BS]} = \frac{P[F_1]}{P[F_1] + P[P_1 \cap F_2] + P[P_1 \cap P_2 \cap F_3]}$$

The numerator has $P[F_1 \cap BS] = P[F_1]$ because $F_1$ is a subset of $BS$. That is, $BS$ is the union of three disjoint events: $F_1, P_1 \cap F_2$, and $P_1 \cap P_2 \cap F_3$. Plugging in from Figure 7,

$$P[F_1|BS] = \frac{0.1}{0.1 + (0.9)(0.2) + (0.9)(0.8)(0.3)} = 0.202.$$ 

Note the answer is not the a priori probability of failing the first test, which is 10%.

10.2 Bayesian Philosophy
The Bayesian philosophy is that one can learn more about whether or not something is true from observation, even noisy observation. Thus it relates to the classic philosophical question, “How do we know what we know?”. However, for hundreds of years, it was sidelined by mathematicians and scientists. Why? Consider this application of Bayes’ Law:

- Let $B$ be the event that “voltage is equal to current times resistance”, and
- Let $A_1, A_2, \ldots$ be some imperfect binary measurements which are affected by the truth of this statement.

To understand the philosophical beef, From our existing understanding, assume we can easily come up with $P[A_1|B]$ and $P[A_1|B^C]$. We can use Bayes’ Law to come up with the $P[B|A_1]$, the probability that the statement is true, given $A_1$; but ONLY if we have the probability $P[B]$. That is, we need to know beforehand the probability that our statement is true. In real life, this statement is not random – it is either true or not true. However, before you know whether it is
true or not, Bayes would argue, you have to quantify how firmly you believe that $B$ is true or not. Keep track of $P[B]$, and then $P[B|A_1]$ when the first experiment is done, in order to quantify how sure you are about $B$. And if more measurement events $A_2, A_3, \ldots$ are recorded, you can keep track of $P[B|A_1 \cap A_2]$ and then $P[B|A_1 \cap A_2 \cap A_3]$, and so on. The probability of $B$ given all of the observations you’ve taken then gives you a quantitative way of tracking your belief in the truth of statement $B$. However, the only way to do this is, at first before any observations are recorded, to assume some $P[B]$. Bayes suggested setting it to $1/2$ when no information is available. Mathematicians and scientists of his time thought that was unjustified and thus the whole idea of using Bayes’ Law was nonsense. As a result, Bayesian statistics was sidelined, and almost unheard of until the 20th century.

However, there are many engineering applications in which Bayes’ Law is perfectly matched. For example, Bayes’ Law is fundamental math used by the British to decode German radio messages, as depicted in the movie on Alan Turing called *The Imitation Game* (2014). There was a large set of possible codes, that is, settings for the German coding machine. It was infeasible to try each possible code to decode a message. However, by observing characteristics of the coded message, one could say that certain sets of codes had higher likelihood. The machines designed to break the code thus applied Bayes Law, using the set of all codes as the partition $B_1, \ldots, B_k$, and the characteristics observed from the messages as the observations $A_1, A_2, \ldots$. By listening to many messages, eventually, a few posterior probabilities $P[B_r|A_1, A_2, \ldots]$ would become very high, and the rest relatively low; these would be tested by decoding the message with them. When a code could successfully decode a message, the code was broken. Bayes’ Law was perfect for this problem because, before observing any message, it was perfectly acceptable to assign equal prior probabilities to each possible code.

This historical and philosophical perspective is discussed both in the Mlodinow reading (posted on Canvas), and more extensively in the book, *The Theory That Would Not Die*, by S. B. McGrayne, Yale University Press, 2012.

Lecture #8

### 11 Poker

Poker provides many examples of counting and probability. In these notes, we study a variation of poker we play with “Phase 10” cards in order to have more practice with permutations, combinations, and probability. These are faceless cards, 48 cards in a deck: 12 each of yellow (Y), red (R), green (G), and blue (B). In each color, the 12 cards are numbered 1 through 12. The deck also comes with Wild cards, Skip cards, and an instruction card, which we will remove.

We will play the following game. The dealer will shuffle the cards and then deal you five cards, and the order they are dealt to you does not matter. You may discard any five of the cards — you can discard all five, or none, or any number in between. The discarded cards do not go back into the dealer’s deck. The dealer will deal you new cards to take the place of those discarded. After this, the person with the “best hand” out of the players in the group is declared the winner.

The “best” hand is the one least likely to be dealt to you (at the start, before discarding and dealing more cards). The types of hands are as follows, in order of decreasing probability:

1. One pair: the same number appears twice in your hand. Example: { B2, R2, G12, Y5, Y9}. 


2. Two pairs: one number appears twice, and a second number appears twice in your hand. Example: \{ B2, R2, G5, Y5, Y9 \}.

3. Three of a kind: the same number appears three times in your hand. Example: \{ B2, R2, G2, Y5, Y9 \}.

4. Straight: five cards of any color have sequential numbers. Example: \{ B2, R3, G4, G5, B6 \}.

5. Full house: one number appears three times, and a second number appears twice in your hand. Example: \{ B2, R2, G2, B9, Y9 \}.

6. Flush: the five cards are all of the same color. Example: \{ B2, B4, B8, B9, B12 \}.

7. Four of a kind: the same number appears four times in your hand. Example: \{ B2, R2, G2, Y2, Y9 \}.

8. Straight flush: five cards of a single color have sequential numbers. Example: \{ B2, B3, B4, B5, B6 \}.

9. Royal flush: Five cards of the same color numbered 12, 11, 10, 9, and 8 appear in your hand. Example: \{ G8, G9, G10, G11, G12 \}.

Note that some hands fit multiple types; you always take the best (least likely) type. Thus when we count the ways to get “one pair”, we exclude two pair, full house, and four of a kind.

If two people have the same hand (or none of the above), then the person with the highest numbered card wins.

### 11.1 First Deal

At first deal, how big is the sample space? That is, how many ways are there to get a hand of five cards using this deck? There are 48 distinct cards, and order doesn’t matter. Thus there are \( \binom{48}{5} \) or 1,712,304 ways.

How many ways are there to get a royal flush? There is only one for each color, so there are 4 ways.

How many ways are there to get a straight flush? First you pick one of four colors. Then you pick a starting number, from 1 to 7. Thus there are \( 4 \binom{7}{1} \) = 28.

How many ways are there to get four of a kind? First, pick a number that will be repeated 4 times in the hand from among 12. Then, pick one additional card from among the remaining 44. Thus \( 12 \binom{44}{1} \) = 528 ways.

How many ways are there to get a flush? First, pick one color from the four. Then, select 5 cards from the 12 available in that color. Thus \( 4 \binom{12}{5} \) = 3168. However, these include the straight and royal flush. Subtracting them, there are 3168 – 28 – 4 = 3,136 ways.

How many ways are there to get a full house? First, pick a number for the three-peat from among 12. Then, pick three out of four cards that have that number. Then, pick a second number from among the 11 remaining. Then, pick two out of the four cards that have that number. Thus there are \( (12) \binom{4}{3} (11) \binom{4}{2} \) = 3,168.

How many ways are there to get a straight? First pick a starting number, from 1 to 8. For each of the five cards, you may independently pick one of four colors. So there are \( 8 \binom{4}{5} \) = 8192 ways. However, these include the straight flush. Subtracting them out, there are 8,164 ways.
How many ways are there to get three of a kind? First, pick a number that will be repeated 3 times in the hand from among 12. Then, pick two additional cards from among the remaining 44 (of other numbers). Thus 12\(\binom{14}{2}\) = 11352 ways. These include the full house. Subtracting, there are then 8,184 ways.

How many ways are there to get two pairs? First, pick a number that will be repeated 2 times in the hand from among 12. Choose two out of the 4 with this number. Then, pick a different number that will be repeated 2 times in the hand from among the 11 left. Choose two out of the 4 with this number. Then, pick one of the remaining 40 cards. Thus 12\(\binom{4}{2}\)11\(\binom{4}{2}\)(40) = 190,080 ways.

How many ways are there to get one pair? First, pick a number that will be repeated 2 times in the hand from among 12. Choose 2 out of the 4 with this number. Then, pick three of the remaining 11 numbers, and each of these three cards can be one of the four colors. Thus 12\(\binom{4}{2}\)(11\(\binom{4}{3}\) = 760,320.

11.2 Selection of Discards

Given any hand, how many ways are there to discard? There are five cards, each one can either be discarded or not. Thus there are 2\(^5\) = 32 ways to discard. One may evaluate the probability of any particular first deal and selection of discards. (This is what is done in the reading for today, an excerpt from Prof. Stewart N. Ethier’s *The Doctrine of Chances: Probabilistic Aspects of Gambling* for two types of video poker.) This probability is essentially a conditional probability, because the probability is conditioned on the cards dealt in the first round (which you cannot receive in the 2nd round), and the cards selected to be discarded. The sample space is limited by these conditions, as well as the number of possibilities in each type of hand.

Lecture #11

12 Random Variable

This is Section 3.1 in the Walpole book.

**Def’n: Random Variable**

A random variable is a mapping from sample space \(S\) to the real numbers. In symbols, \(X\) is a random variable if \(X : S \to \mathbb{R}\).

We write \(S_X\) to indicate the set of values which \(X\) may take: \(S_X = \{x : X(s) = x, \forall s \in S\}\). This is the ‘range of \(X\’.

In other words, if the sample space is not already numeric, then a real number is assigned to each outcome. For example,

- For a coin-flipping experiment with \(S = \{\text{Heads, Tails}\}\) we might assign \(X(\text{Heads}) = 1\) and \(X(\text{Tails}) = 0\); thus \(S_X = \{0, 1\}\).

- For an experiment involving a student’s grade, \(S = \{F, D, C, B, A\}\), we might assign \(X(F) = 0, X(D) = 1, X(C) = 2, X(B) = 3, X(A) = 4\). Thus \(S_X = \{0, 1, 2, 3, 4\}\).

- For a 8-bit memory value, we might count the number of “1”s. Here, \(S_X = \{0, 1, \ldots, 8\}\).

- For products tested from an assembly line, we might count the number of items until the first defective product. Here, \(S_X = \{1, 2, 3, \ldots\}\).
For outcomes that already have numbers associated with them, i.e., temperature, voltage, or the number of a die roll, we can just use those numbers as the random variable.

**Def’n:** *Discrete Random Variable*

$X$ is a discrete-valued random variable if one can assign a unique integer to each possible value of $X$.

Examples: All of the random variables above are discrete random variables. Another example: The average of two randomly chosen positive integers has sample space $S_X = \{1, 1.5, 2, 2.5, \ldots\}$. It is still a discrete random variable because I could assign it the number $2X$, which is an integer.

**Example: Number of 1s in 8 bit memory**

Assume 0s and 1s are equally likely in the 8-bit memory above. What is the probability that $X$, the number of 1s, is 8? Solution: $1/2^8$.

**Def’n:** *Continuous r.v.*

$X$ is a continuous-valued random variable if its range $S_X$ contains as many values as the real line.

Examples: The voltage measured in a circuit. Power consumption. Height of people. Time duration to complete a task.

One may make the argument that any *measurement* of these will probably be done on a digital device with a finite precision, and thus it is a discrete random variable. This may be true; however, one may also argue that the actual value, if it could be known, would be an infinite number of decimal places, and thus would be a continuous random variable. Frankly, it is most important what is USEFUL, thus we represent values as continuous r.v.s when it provides a useful and accurate method to answer a question.

**Example: Probability of a Real value**

What is the probability that a vehicle’s speed is 85 mph? Solution: Zero. Why? Because there are an infinite number of possible values between 84.5 mph and 85.5 mph (or even 84.999 and 85.001 mph). (Don’t use this as a defense in court when you get a ticket for going 85 in a 65-mph zone).

### 13 Probability Mass Functions

This is Section 3.2 in the Walpole book.

**Def’n:** *probability mass function (pmf)*

The probability mass function (pmf) of the discrete random variable $X$ is $f_X(x) = P[X = x]$.

Note the use of capital $X$ to represent the random variable name, and lowercase $x$ to represent a particular value that it may take (a dummy variable). Eg., we may have two different random variables $R$ and $X$, and we might use $f_R(u)$ and $f_X(u)$ for both of them.

**Example: Die Roll**

Let $Y$ be the sum of the roll of two dice. What is $f_Y(y)$?
Die 1 \ Die 2 | 1 | 2 | 3 | 4 | 5 | 6
---|---|---|---|---|---|---
1 | 1/36 | 1/36 | 1/36 | 1/36 | 1/36 | 1/36
2 | 1/36 | 1/36 | 1/36 | 1/36 | 1/36 | 1/36
3 | 1/36 | 1/36 | 1/36 | 1/36 | 1/36 | 1/36
4 | 1/36 | 1/36 | 1/36 | 1/36 | 1/36 | 1/36
5 | 1/36 | 1/36 | 1/36 | 1/36 | 1/36 | 1/36
6 | 1/36 | 1/36 | 1/36 | 1/36 | 1/36 | 1/36

Noting the numbers of rolls which sum to each number, 2 through 12, we have:

\[
f_Y(y) = \begin{cases} 
  1/36 & \text{if } y = 2, 12 \\
  2/36 & \text{if } y = 3, 11 \\
  3/36 & \text{if } y = 4, 10 \\
  4/36 & \text{if } y = 5, 9 \\
  5/36 & \text{if } y = 6, 8 \\
  6/36 & \text{if } y = 7 \\
  0 & \text{o.w.} 
\end{cases}
\]

See Figure 8. Check: What is \( \sum y f_Y(y) \)?

\[
\frac{2(1+2+3+4+5)+6}{36} = \frac{2(15)+6}{36} = 1.
\]

Figure 8: Probability mass function for the sum of two dice.

- Always put, ‘zero otherwise’.
- Always check your answer to make sure the pmf sums to 1.

**Def’n: Bernoulli Random Variable**

A r.v. \( X \) is Bernoulli (with parameter \( p \)) if its pmf has the form,

\[
f_X(x) = \begin{cases} 
  1 - p & \text{if } x = 0 \\
  p & \text{if } x = 1 \\
  0 & \text{o.w.} 
\end{cases}
\]

This is the most basic, and most common pmf! Experiments are often binary. Eg., success/failure, in range / out of range, disease / no disease, packet or bit error / no error, etc.
Example: Fraud Detection using Benford’s Law
Here is a pmf you’ve probably never heard of, and a reason why an engineer might be interested in it. A set of numbers is said to satisfy Benford’s Law if the most significant digit $d$ ($d \in \{1, \ldots, 9\}$) has pmf:

**Def’n:** Benford Random Variable
A r.v. $X$ is Benford with sample space $S_X = \{1, \ldots, 9\}$ if its pmf has the form,

$$P_D(d) = \begin{cases} \log_{10}(d+1) - \log_{10} d & \text{if } x \in \{1,2,\ldots,9\} \\ 0 & \text{o.w.} \end{cases}$$

Electrical Engineer Frank Benford is given credit for the law. For example, the physical constants (speed of light, force of gravity, Boltzman’s constant, etc.) taken from a physics book follow this law. More relevant, numbers on tax returns (and more generally, accounting statements) follow this law. However, when people just make up numbers, they tend to have a most significant digit closer to uniform on $\{1, \ldots, 9\}$. Thus looking at the histogram of most significant digits on an accounting statement allows auditors one way to screen for fraud.

As a recent example, Greek accounting statements provided to the European Union (EU) from 1999-2009, as they desired to adopt the euro, have most-significant digits that deviate significantly from Benford’s Law. The numbers were “revised” several times over the years, and the E.U. more recently confirmed that data was manipulated.

### 13.1 Cumulative Distribution Function (CDF)

This is Section 3.3 in the Walpole book.

**Def’n:** Cumulative Distribution Function (CDF)
The CDF, $F_X(x)$ is defined as $F_X(x) = P[X \leq x]$.

---

For a discrete r.v. $X$, the CDF is a cumulative sum of the pmf, up to the value $x$. In math, $F_X(x) = \sum_{u \in S_X; u \leq x} f_X(u)$.

Properties of the CDF:

1. It always starts at 0, and always ends at 1. In math, $\lim_{x \to +\infty} F_X(x) = 1$, and $\lim_{x \to -\infty} F_X(x) = 0$.

2. The CDF is non-decreasing. In math, for all $b \geq a$, $F_X(b) \geq F_X(a)$.

3. The difference between the CDF at two values gives the probability that the r.v. is in the range between those two values. In math, $F_X(b) - F_X(a) = P\{a < X \leq b\}$.

**Example: Sum of Two Dice**

Plot the CDF $F_Y(y)$ of the sum of two dice. A: Figure 10.

![Figure 10: CDF for the sum of two dice.](image)

**Lecture #12**

### 14 Probability Density Functions

Previously, we discussed the probability mass function, that is, the probability of each possible value of a discrete random variable (r.v.). As described earlier, the probability of any single value of a continuous r.v. is zero, its pmf is meaningless. However, the cumulative distribution function (CDF) of a continuous random variable is meaningful. In these notes, we define and describe it, and then define the probability density function (pdf) in terms of the CDF.

**Example: Uniform Continuous r.v.**
The CDF for a continuous r.v. that is uniformly random in the interval 0 to 1 is given by

$$F_X(x) = P\{[0, x]\} = \begin{cases} 
0, & x < 0 \\
x, & 0 \leq x < 1 \\
1, & x \geq 1
\end{cases}.$$

See Figure 11.
The three properties of the CDF listed above are the same for a continuous r.v. and for a discrete r.v. However, the CDF is not a cumulative sum; instead, it is a cumulative integral of a probability density function (pdf). Though, this is because we define the pdf that way:

**Def’n:** Probability density function (pdf)

The pdf of a continuous r.v. \( X \), \( f_X(x) \), is the derivative of its CDF:

\[
f_X(x) = \frac{\partial F_X(x)}{\partial x}
\]

For the uniform continuous random variable with CDF above, the derivative is zero except for between 0 and 1, where the slope is 1. Thus the pdf has value 1 between 0 and 1 only, as seen in Figure 11.

**Properties:**

1. \( f_X(a) \) is the *density*. Not a probability!.

2. \( \epsilon \cdot f_X(a) \) is approximately the probability that \( X \) falls in an \( \epsilon \)-wide window around \( a \). Its a good approximation if \( \epsilon \approx 0 \).

3. \( F_X(x) = \int_{-\infty}^{x} f_X(u)du \). (Fundamental theorem of calculus)
4. \( \forall x, f_X(x) \geq 0 \). (from non-decreasing property of \( F_X \))

5. \( \int_{-\infty}^{\infty} f_X(u)du = 1 \). (from limit property of \( F_X \))

6. \( P[a < X \leq b] = \int_a^b f_X(x)dx \).

**Example: Exponential**

**Def’n: Exponential random variable**

A r.v. \( X \) is Exponential if it has CDF,

\[
F_X(x) = \begin{cases} 
1 - e^{-\lambda x}, & x > 0 \\
0, & o.w.
\end{cases}
\]

An exponential r.v. is often used to model a device’s lifetime; or how long you will have to wait for something to happen (e.g., bus arrival or arrival of the next phone call to a phone bank).

![Cumulative Distribution Function](a)

![Probability density function](b)

**Figure 12:** For a Exponential random variable \( X \) with parameter \( \lambda = 0.5 \), the (a) CDF and (b) pdf.

What is the probability that the lifetime of a product an Exponential r.v. \( X \) is less than 1 year,
when $\lambda = 0.5 \text{ 1/year}$? **Solution:** Using the CDF,

$$P [X < 1 \text{ year}] = F_x (1 \text{ year}) = 1 - e^{-0.5 \frac{1}{\text{year}} (1 \text{ year})} = 1 - e^{-0.5} = 0.39$$

What is the pdf for an Exponential r.v.?

**Solution:**

$$f_x(x) = \frac{\partial F_x(x)}{\partial x} = \frac{\partial}{\partial x} \left( 1 - e^{-\lambda x} \right) = \lambda e^{-\lambda x}$$

However, don’t stop there – this was only true for $x > 0$. Write your solution with the “0 otherwise”:

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

What is the probability that the product fails within 1 day after the end of the 1 year warranty?

**Solution:** Here’s an approximate solution using pdf property #2:

$$P \left[ 1 \leq X < 1 + \frac{1}{365} \right] \approx f_x(x) \epsilon = 0.5 e^{-0.5(1)} \frac{1}{365} = 8.3086 \times 10^{-4}$$

Here’s an exact solution using CDF property #3:

$$P \left[ 1 \leq X < 1 + \frac{1}{365} \right] = F_x \left( 1 + \frac{1}{365} \right) - F_x (1)$$

$$= \left( 1 - e^{-0.5(366/365)} \right) - \left( 1 - e^{-0.5(1)} \right) = 8.3029 \times 10^{-4}$$

---

**Lecture #13**

**15 Expectation**

We’ve been talking about expectation as an extension of the concept of averaging. For example, let’s say you have $n$ samples of a random variable $X$, called $x_1, x_2, \ldots, x_n$. You can average them and compute

$$\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i$$

As shown in Figure 13(a).

Intuitively, we want to find a way to come up with the average we’d have if we had an infinite number of samples of the random variable. We will call this expectation rather than the mean, because we will never compute exactly this number with a finite number of samples. In a sense, it is a mathematical limit of the average as $n \to \infty$ (although we are not providing enough theory to back this statement up).

See a problem? We don’t ever as engineers have an infinite amount of time or resources to be able to make an arbitrary number of measurements in order to find the expected value this way.
We write the expected value of $X$ as $E[X]$. The point of lecture 12 was that one may compute this from the distribution functions. For a continuous r.v.:

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

and for a discrete random variable:

$$E[X] = \sum_{x \in S_X} x f_X(x)$$

We also denote the expected value of $X$ as $\mu_X$.

Six notes:

1. The expected value is non-random! More specifically, once you take the expected value w.r.t. $X$, you will no longer have a function of $X$.

2. By analogy to statics, the expected value is where on the x-axis you’d need to put your finger to “balance” the mass or density on the pmf or pdf plot.

3. Note we sometimes write $E[X]$ as $E_X[X]$, where the subscript $X$ indicates which r.v. we’re taking the expectation of. When there’s only one thing in the expectation operator, there’s no ambiguity, so no need to write the $X$ in the subscript. When there’s any ambiguity, I try to write the subscript. The Walpole book never writes the subscript (as far as I’ve seen).

4. The Walpole book typically writes $E(\cdot)$, that is, using parentheses, but sometimes they write $E[\cdot]$ or $E\{\cdot\}$. I try to be consistent and use square brackets every time.
5. **The expected value is a linear operator.** Consider \( g(X) = aX + b \) for \( a, b \in \mathbb{R} \). Then

\[
E_X [g(X)] = E_X [aX + b] = \sum_{x \in S_X} (ax + b) f_X(x) = \sum_{x \in S_X} [ax f_X(x) + bf_X(x)] = \sum_{x \in S_X} ax f_X(x) + \sum_{x \in S_X} bf_X(x) = a \sum_{x \in S_X} x f_X(x) + b \sum_{x \in S_X} f_X(x) = aE_X [X] + b
\]

6. The expected value of a constant is a constant: \( E_X [b] = b \).

### 15.1 Expectation of a Function

In general, we can average any function of a random variable. For example, if \( X \) is a voltage, then \( X^2/R \) is the power dissipated across resistance \( R \). Perhaps we are interested in the power, but are making a measurement of voltage. No problem, we can average power:

\[
g^*(x) = \frac{1}{n} \sum_{i=1}^{n} g(x_i)
\]

where

\[
g(x) = \frac{x^2}{R}
\]

In this case, if we have the distribution function for \( X \), we could calculate the expected value of power. We’d denote this \( E [g(x)] \), which is defined as follows. For a continuous r.v.:

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

If this had been an example where \( X \) was a discrete random variable:

\[
E [g(X)] = \sum_{x \in S_X} g(x) f_X(x)
\]

Now, we used \( g(x) = x^2/R \) here, but any function may be used. Let’s consider some common functions \( g(X) \):

1. Let \( g(X) = X \). We’ve already done this! The mean \( \mu_X = E_X [X] \).
2. Let \( g(X) = X^2 \). The value \( E_X [X^2] \) is called the *second moment*.
3. Let \( g(X) = X^n \). The value \( E_X [X^n] \) is called the \( n \)th moment.
4. Let \( g(X) = (X - \mu_X)^2 \). This is the *second central moment*. This is also called the variance, or \( \sigma_X^2 \). What are the units of the variance?
5. Let $g(X) = (X - \mu_X)^n$. This is the $n$th central moment.

Some notes:

1. “Moment” is used by analogy to the moment of inertia of a mass. Moment of inertia describes how difficult it is to get a mass rotating about its center of mass, and is given by:

$$I \triangleq \int \int \int_V r^2 \rho \, dx \, dy \, dz$$

where $\rho$ is the mass density, and $r$ is the distance from the center.

2. Standard deviation $\sigma = \sqrt{\sigma^2_X}$.

3. Variance in terms of 1st and 2nd moments.

$$E [(X - \mu_X)^2] = E [X^2 - 2X \mu_X + \mu^2_X] = E [X^2] - 2E [X] \mu_X + \mu^2_X = E [X^2] - (E [X])^2.$$  

4. Note $E_X [g(X)] \neq g (E_X [X])$!

A summary:

<table>
<thead>
<tr>
<th>Expression</th>
<th>$X$ is a discrete r.v.</th>
<th>$X$ is a continuous r.v.</th>
</tr>
</thead>
<tbody>
<tr>
<td>$E [X]$</td>
<td>$\sum_{x \in S_X} x f(x)$</td>
<td>$\int_{x \in S_X} x f(x) , dx$</td>
</tr>
<tr>
<td>$E [g(x)]$</td>
<td>$\sum_{x \in S_X} g(x) f(x)$</td>
<td>$\int_{x \in S_X} g(x) f(x) , dx$</td>
</tr>
<tr>
<td>$E [aX + b]$</td>
<td>$a E [X] + b$</td>
<td>$a E [X] + b$</td>
</tr>
<tr>
<td>$E [X^2]$ 2nd moment</td>
<td>$\sum_{x \in S_X} x^2 f(x)$</td>
<td>$\int_{x \in S_X} x^2 f(x) , dx$</td>
</tr>
<tr>
<td>$\sigma^2_X$ $E [(X - \mu_X)^2]$, $\mu_X = E [X]$</td>
<td>$\sum_{x \in S_X} (x - \mu_X)^2 f_X(x)$</td>
<td>$\int_{x \in S_X} (x - \mu_X)^2 f_X(x) , dx$</td>
</tr>
</tbody>
</table>

15.2 Examples

Example: Variance of Uniform r.v.

Let $X$ be a continuous uniform r.v. on $(a, b)$, with $a, b > 0$.

1. What is $E [X]$? It is

$$\int_a^b \frac{x}{b - a} \, dx = \frac{1}{2(b - a)} \bigg|_{x=a}^{x=b} (b^2 - a^2) = \frac{b + a}{2}.$$  

2. What is $E \left[ \frac{1}{X} \right]$?

$$\int_a^b \frac{1}{b - a} \, dx = \frac{1}{b - a} (\ln b - \ln a) = \frac{1}{b - a} \ln \frac{b}{a}.$$  

3. What is $E [X^2]$? It is

$$\int_a^b \frac{x^2}{b - a} \, dx = \frac{1}{3(b - a)} \bigg|_{x=a}^{x=b} (b^3 - a^3) = \frac{b^2 + ab + a^2}{3}.$$  

4. What is $\sigma^2_X$? It is

$$\sigma^2_X = E_X^2 \left[ (E [X])^2 \right] = \frac{b^2 + ab + a^2}{3} - \frac{b^2}{4} = \frac{b^2}{12} - \frac{2ab + a^2}{12} = \frac{(b - a)^2}{12}.$$
Example: Bernoulli Moments

What are the mean, 2nd moment and variance of $X$, a Bernoulli r.v.?

Solution: Recall a Bernoulli r.v. $X$ has two possible outcomes, 0 and 1. The probability that $X = 1$ is denoted $p$ and thus $P[X = 0] = 1 - p$. The mean is:

$$\mu_X = E[X] = \sum_{x \in S_X} xf(x) = 0(1 - p) + 1(p) = p.$$ 

The 2nd moment is:

$$E[X^2] = \sum_{x \in S_X} x^2 f(x) = 0^2(1 - p) + 1^2(p) = p.$$ 

The variance could be calculated from the definition as:

$$\sigma^2_x = E[(X - \mu_X)^2] = \sum_{x \in S_X} (x - p)^2 f(x) = p^2(1 - p) + (1 - p)^2(p) = p(1 - p)(p + 1 - p) = p(1 - p).$$ 

Alternatively, we could use the relationship shown earlier that $\sigma^2_X = E[X^2] - (E[X])^2$:

$$\sigma^2_X = p - p^2 = p(1 - p).$$

Lecture #14

16 Binomial pmf

Consider the following experiment. We flip a coin $n$ times, and count the number of tails that we see. Let’s call the number of tails r.v. $K$.

What is the sample space? We can only get a number from 0 to $n$, so $S_K = \{0, 1, \ldots, n\}$.

What is the expected value of $K$? If it is a fair coin, I’d predict the mean number of tails is $n/2$.

What is the pmf? That is, what is $P[K = k]$ for some value $k$ in the sample space? Well, we are counting the number of ways to have $k$ tails and $n - k$ heads in our $n$ flips. There are $2^n$ total ways to have the coin flip experiment turn out (counting with order mattering). Then there are $\binom{n}{k}$ ways to “place” the $k$ tails among the $n$ total flip results. Once you pick where the tails are, that determines where the heads are. So the probability is

$$f_K(k) = P[K = k] = \binom{n}{k} / 2^n = \binom{n}{k} (0.5)^n$$

More generally, we may not have equally likely of “tails”, the event we are counting. For example, we may be counting the number of defective products out of $n$ products coming off of an assembly line. Hopefully we are more likely to have a working product than a defective product! For more generality, denote the probability of “tails” as $p$, for some value in $[0, 1]$. The solution for the pmf in this more general case is called the binomial pmf.
Example: We know 2% of ICs are defective. We buy three ICs. What is the probability of getting 0, 1, 2, and 3 defective ICs?

Solution: This can be solved by a tree diagram or as a permutation problem. Both will be presented in class.

Def’n: Binomial r.v.
A r.v. $K$ is binomial with success probability $p$ and number of trials $n$ if it has the pmf,

$$f_K(k) = \binom{n}{k} p^k (1-p)^{n-k}, \quad k = 0, \ldots, n$$  \hspace{1cm} (10)

A binomial r.v. stems from $n$ independent Bernoulli r.v.s. Specifically, it is the number of successes in $n$ trials (where each trial has success with probability $p$). Mathematically,

$$K = \sum_{i=1}^{n} X_i$$

where $X_i$ for $i = 1 \ldots n$ are independent Bernoulli r.v.s, each having probability $p$ of being 1, and probability $1-p$ of being 0.

- Since success (1) has probability $p$, and failure has probability $(1-p)$, what is the probability of this particular event: First, we have $k$ successes and then $n-k$ failures? A: $p^k (1-p)^{n-k}$.

- Order of the $n$ Bernoulli trials may vary; how many ways are there to arrange $k$ successes into $n$ slots? A: $\binom{n}{k}$.

Lecture #15

Nothing new for this lecture, we’re catching up with past lecture notes.

Lecture #16

17 Uniform Distribution

Def’n: Uniform r.v.
A r.v. $X$ is uniform between $a$ and $b$ (for some constants $b$ and $a$ where $b > a$) if it has pdf,

$$f(x; a, b) = \begin{cases} \frac{1}{b-a}, & a \leq x \leq b \\ 0, & \text{o.w.} \end{cases}$$

- A uniform r.v. is “flat” or “rectangular” over an interval $[a, b]$. Probabilities are particularly easy to calculate as a result of the pdf being flat.

- $a$ is the minimum possible value for $X$, and $b$ is the maximum.
• It is also a distribution used when nothing is known about $X$ except for its min and max. For example, if you are starting to tracking someone and know their $x$-coordinate must be between $a$ and $b$ (for example because they are in a building and you can’t see inside), but don’t know anything else, you would start by assuming a uniform distribution.

• Matlab will generate uniform r.v.s on $[0,1)$ using the “rand” command. You can generate a uniform $[a,b)$ r.v. using:

$$x = a + (b-a) \cdot \text{rand}$$

The mean of a uniform on $[a,b]$ r.v. is

$$E[X] = \frac{a + b}{2}$$

and its variance is

$$\text{Var}[X] = \frac{(b-a)^2}{12}$$

We did the derivation of these in a prior class.

18 Gaussian Distribution

The Gaussian distribution is perhaps the distribution used most often in engineering and science. It is often used to describe a voltage measured in thermal noise; measurement error when measuring a physical quantity (length, height, volume, etc.); velocity of molecules in a gas; changes in the price of a stock or security. It is also used as an approximation of the binomial pmf or the Poisson pmf under certain conditions.

Def’n: Gaussian r.v.

A r.v. $X$ is Gaussian with mean $\mu$ and variance $\sigma^2$ if it has pdf,

$$n(x; \mu, \sigma) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(x-\mu)^2/(2\sigma^2)}$$

Things to know:

• The expected value (mean), mode (where the highest value occurs), and median (where 50% of the data lie below), are all equal to $\mu$.

• The pdf is symmetric about the mean.

• The curve approaches zero as $x$ gets further and further from its mean, but is never exactly zero.

• The standard deviation is $\sigma$ and is always $>0$. The variance is $\sigma^2$.

• This pdf is also known as the Normal distribution (it is Gaussian only in engineering).

Changing the mean parameter $\mu$ up (or down) moves the same shape curve right (or left) on the x-axis. Changing the standard deviation up (or down) makes the curve wider (or sharper).
• Zero-mean unit-variance Gaussian is also known as the ‘standard normal’. 

$$f_X(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}$$

• Matlab generates standard normal r.v.s with the ‘randn’ command. To generate Gaussian r.v.s with mean m and standard deviation s:

$$x = m + s \cdot \text{randn}$$

Example: Prove that the expected value of a Gaussian r.v. is 

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

Starting by “bringing out” a μ, which will help later when I do the u substitution,

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

This last line is because the integral from −∞ to +∞ of any pdf is 1, and

$$\int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}\sigma} e^{-(x-\mu)^2/(2\sigma^2)} dx = \mu \int_{-\infty}^{\infty} f_X(x) dx = \mu.$$ Now, let $$u = \frac{(x-\mu)^2}{2\sigma^2}$$. Then $du = \frac{x-\mu}{\sigma^2} dx$. That is, $\sigma^2 du = (x-\mu) dx$. Note that I leave the limits in terms of x, it makes it easier in the end.

$$E[X] = \mu + \int_{-\infty}^{\infty} \frac{\sigma}{\sqrt{2\pi}} e^{-u} du$$

$$E[X] = \mu + \frac{\sigma}{\sqrt{2\pi}} \left[ -e^{-u} \right]_{x=-\infty}^{x=\infty}$$

$$E[X] = \mu + \frac{\sigma}{\sqrt{2\pi}} \left[ -e^{-(x-\mu)^2/(2\sigma^2)} \right]_{x=-\infty}^{x=\infty}$$

$$E[X] = \mu + \frac{\sigma}{\sqrt{2\pi}} [0 - 0] = \mu.$$

18.1 Cumulative Probability

What is the CDF of the Gaussian r.v.?

$$F_X(x) = \int_{-\infty}^{x} n(u; \mu, \sigma) du = \frac{1}{\sqrt{2\pi}\sigma^2} \int_{-\infty}^{x} e^{-(u-\mu)^2/(2\sigma^2)} du$$

Unfortunately, there is no regular analytical function that results from the integral. In general, we use a table to give us the values of this CDF. However, we don’t want to make a new table for every possible value of μ and σ. So instead, we just do it for the standard normal random variable Z.
Theorem: If $X$ is a Gaussian r.v. with mean $\mu$ and standard deviation $\sigma$, then

$$Z = \frac{X - \mu}{\sigma}$$

is a standard normal random variable.

Proof: We will do part of this proof later in the course.

So we can always “convert” $X$ into a standard normal r.v. by subtracting its mean and dividing by its standard deviation.

The CDF of $Z$, a standard normal r.v. is expressed as (even though the integral can’t be solved):

$$F_Z(z) = P[Z < z] = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{z} e^{-x^2/2} dx$$

The values of $F_Z(z)$ are given in Table A.3 “Normal Probability Table” in the Appendix in the Walpole book.

Example: Resistance Meets Minimum Requirement

A 3 Ohm resistor, from a particular manufacturer, is Gaussian with mean $\mu = 3$ Ohm and standard deviation 0.3 Ohm. Find the probability $X < 2.5$.

My method for getting a solution is what I call the DEFG method:

1. Draw: the pdf and the probability (area) which this question asks for.
2. Exchange: the interval on $X$ with the equivalent interval on $Z$, a standard normal r.v., using $Z = \frac{X - \mu}{\sigma}$.
3. $F_Z(z)$: Write the probability of that interval in terms of $F_Z()$, the CDF of the standard normal r.v.
4. Go: to Table A.3 and plug in the $F_Z(z)$ values from the table to get the solution.

For this problem my solution is:

1. Draw the pdf (see in class). We’re looking for the area to the left of 2.5 on the x axis.
2. Exchange. $X < 2.5$ we subtract the mean and then divide by the standard deviation on both sides to get:

$$\begin{align*}
X &< 2.5 \\
X - \mu &< 2.5 - 3 \\
\frac{X - \mu}{\sigma} &< \frac{2.5 - 3}{0.3} = \frac{-0.5}{0.3} = -1.67 \\
Z &< -1.67 \\
\end{align*}$$

3. $F_Z(z)$: We are looking for the $P[Z < -1.67]$ which is the same as $F_Z(-1.67)$.
4. Go to Table A.3: Look in row -1.6, column 0.07. It shows a value of 0.0475. Thus $P[X < 2.5] = 0.0475$. 
Example: What is the probability that a normal random variable $X$ will be within one standard deviation of its mean?

Solution:

1. Draw the pdf (see in class). We’re looking for the area between $\mu - \sigma$ and $\mu + \sigma$.

2. No matter what the mean and standard deviation are, we want the probability of $\mu - \sigma < X < \mu + \sigma$. We subtract $\mu$ from each limit to get $-\sigma < X - \mu < \sigma$, and then divide by $\sigma$ to get $-1 < Z < 1$.

3. $F_Z(z)$: We convert $P[-1 < Z < 1] = F_Z(1) - F_Z(-1)$.

4. Go to Table A.3: In row 1.0 column 0.00, $F_Z(1) = 0.8413$. In row -1.0 column 0.00, $F_Z(-1) = 0.1587$. Thus $P[-1 < Z < 1] = 0.8413 - 0.1587 = 0.6826$.

68%, or just over two-thirds, of the data is within one standard deviation of the mean for Gaussian random variables.

In general, for any random variable $Y$ (not just Gaussian r.v.s), you can find the $P[a \leq Y \leq b]$, for any two constants $b > a$, using the CDF $F_Y(y)$:

$$P[a \leq Y \leq b] = F_Y(b) - F_Y(a).$$

This is because $F_Y(b)$ is the probability $Y \leq b$, which includes the event that $\{Y \leq a\}$, which you don’t want. To compensate, you subtract out $F_Y(a)$, the probability of that event.

18.2 Inverse use of Normal Probability

Example: Find the Required Standard Deviation

We want to design our resistor manufacturing line so that 95% of our 3 Ohm resistors are between 2.9 and 3.1 Ohms. We assume that the resistances are normal with mean 3.0. What is the necessary standard deviation?

Solution: Draw a plot of the pdf. Because of symmetry, if w% are below 2.9 Ohms, then w% will be above 3.1 Ohms. Thus we need 2.5% of the probability to be below 2.9. That is,

$$P[X < 2.9] = 0.025.$$ 

Exchanging $X$ for a standard normal r.v., we have:

$$P \left[ Z < \frac{2.9 - 3.0}{\sigma} \right] = 0.025.$$ 

Using $F_Z$ in the terminology, we need to find the $z$ such that

$$F_Z(z) = 0.025$$

So we just look through the table to find 0.0250. We see it at $z = -1.96$. So

$$\frac{2.9 - 3.0}{\sigma} = -1.96$$
Solving for sigma,

\[ \sigma = \frac{-0.1}{-1.96} = 0.0510 \]

Thus the standard deviation required is 0.051 Ohms.

Lecture #18

19 Joint Discrete Random Variables

There are many applications in which we need to consider two random variables at once. Some examples:

- Communications systems: The transmitted bit and received bit, considered simultaneously, allow us to characterize the performance of the communication system. Further, modern modulations send multiple bits simultaneously and using multiple dimensions. A receiver measures and makes multiple bit decisions from each multidimensional measurement. Allowing higher and higher dimension of these joint random variables has allowed communication systems to get closer and closer to the theoretical limits on communication efficiency.

- Science: We are often interested in whether one variable affects another variable. Comparing data from both at once allow us to decide upon their relationship.

- Circuits: Many things on an IC are random. Evaluating system output may require considering the randomness from multiple circuit elements. For example, when two the resistors in a voltage divider are random, what is the distribution of the output voltage?

**Def’n:** Joint probability mass function

We define the joint probability mass function of two random variables X and Y as

\[ f_{X,Y}(x, y) = P\{X = x \cap Y = y\} \]

In short, we use \( f_{X,Y}(x, y) \) because it is a function of \( x \) and \( y \), and the entire expression is too long. There is a middle ground where we also write the joint probability using a comma rather than the two events and a intersection sign.

\[ f_{X,Y}(x, y) = P\{X = x, Y = y\} = P\{X = x\} \cap \{Y = y\}\]

These are all the same. There are just shorthand ways of writing the event.

The function \( f_{X,Y}(x, y) \) retains the properties of probability:

1. Sum to 1:

\[ \sum_{x \in S_X} \sum_{y \in S_Y} f_{X,Y}(x, y) = 1 \]

2. Non-negative and less than 1: \( 0 \leq f_{X,Y}(x, y) \leq 1 \), for all \( (x, y) \).
Example: Binary Communications System

In a communication system, a random bit \( X \) is 0 or 1. The bit is sent from the transmitter to a receiver, which demodulates the bit as random variable \( Y \), a 0 or 1. Ideally, \( X = Y \), but in the real world, there are errors. Consider the following pdf for the joint random variables \( f_{X,Y}(x, y) \) as given in this table:

<table>
<thead>
<tr>
<th>( f_{X,Y} )</th>
<th>( X = 0 )</th>
<th>( X = 1 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>( Y = 0 )</td>
<td>0.48</td>
<td>0.04</td>
</tr>
<tr>
<td>( Y = 1 )</td>
<td>0.02</td>
<td>0.46</td>
</tr>
</tbody>
</table>

We can plot a 2-variable pmf as a 3-D stem plot (Figure drawn in lecture).

The probability of any event is a sum of the pmf in for combinations \((x, y)\) which are in that event:

\[
P[A] = \sum_{(x,y) \in A} f_{X,Y}(x, y)
\]

For example, the probability of error is the event \( E = \{ X \neq Y \} \). This includes \((0, 1)\) and \((1, 0)\). So \( P[E] = 0.02 + 0.04 = 0.06 \).

19.1 Marginal Distributions

The "marginal" probability mass function is the pmf of just one of the random variables, either \( X \) or \( Y \), but not both. For example, \( f_X(x) = P[\{X = x\}] \) is a marginal pmf. The other example is \( f_Y(y) = P[\{Y = y\}] \). Its what we just to call just “the pmf”, but now we have a joint pmf and a marginal pmf, the different terms used to distinguish exactly what we mean. Note Walpole uses \( g(x) = f_X(x) \) for the marginal pmf of \( X \), and \( h(y) = f_Y(y) \) as the marginal pmf for \( Y \), but this notation is not standard - I’ve never seen written that way anywhere else. To find the marginal pmf from the joint pmf, just sum “out” the variable you want get rid of:

\[
f_X(x) = \sum_{y \in S_Y} f_{X,Y}(x, y)
\]

\[
f_Y(y) = \sum_{x \in S_X} f_{X,Y}(x, y)
\]

This is a direct result of the law of total probability, because the events \( \{ Y = y \} \) for all \( y \in S_Y \) form a partition of the sample space.

For example, in the communication system example:

<table>
<thead>
<tr>
<th>( f_{X,Y} )</th>
<th>( X = 0 )</th>
<th>( X = 1 )</th>
<th>SUM: ( f_Y(y) )</th>
</tr>
</thead>
<tbody>
<tr>
<td>( Y = 0 )</td>
<td>0.48</td>
<td>0.04</td>
<td>0.52</td>
</tr>
<tr>
<td>( Y = 1 )</td>
<td>0.02</td>
<td>0.46</td>
<td>0.48</td>
</tr>
<tr>
<td>SUM: ( f_X(x) )</td>
<td>0.50</td>
<td>0.50</td>
<td></td>
</tr>
</tbody>
</table>

The sum at bottom is \( f_X(x) \), the marginal pmf for \( X \). The sum at the right is \( f_Y(y) \), the marginal pmf for \( Y \). Note that bits 0 and 1 are equally likely to be sent, but for whatever reason, the receiver is a little biased and guesses \( Y = 0 \) more often.
19.2 Conditional Distributions

Recall the definition of conditional probability:

\[ P[A|B] = \frac{P[A \cap B]}{P[B]} \]

If we define \( A = \{X = x\} \) and \( B = \{Y = y\} \), this is the definition of a conditional probability mass function of \( X|Y \):

\[ P[X = x|Y = y] = \frac{P[\{X = x\} \cap \{Y = y\}]}{P[Y = y]} \]

We can write this using the joint and marginal pmf, and use some shorthand:

\[ f_{X|Y}(x|y) = \frac{f_{X,Y}(x, y)}{f_Y(y)} \]

Also,

\[ f_{Y|X}(y|x) = \frac{f_{X,Y}(x, y)}{f_X(x)} \]

**Example:** For the binary communication system, find \( f(y|x) \) and \( f(x|y) \).

**Solution:** There are 4 different possible combinations of \( (x, y) \) for each. To show how to do this, let’s do one example: \( f_{Y|X}(0|0) \).

\[ f_{Y|X}(y = 0|x = 0) = \frac{f(0,0)}{g(0)} = \frac{0.48}{0.50} = 0.96 \]

For the same given information, \( f_{Y|X}(1|0) \) is

\[ f_{Y|X}(y = 1|x = 0) = \frac{f(0,1)}{g(0)} = \frac{0.02}{0.50} = 0.04 \]

Note it’s a good idea to check your answer by seeing if the conditional pmf sums to one:

\[ f_{Y|X}(y = 0) = 0.96 + 0.04 = 1.0 \]

This conditional pmf says, if the transmitter sends bit 0, there is a 96% chance the receiver gets a 0, and a 4% chance of error. Similarly,

\[ f_{Y|X}(y = 0|x = 1) = \frac{f(1,0)}{g(1)} = \frac{0.04}{0.50} = 0.08 \]

\[ f_{Y|X}(y = 1|x = 1) = \frac{f(1,1)}{g(1)} = \frac{0.46}{0.50} = 0.92 \]

These also add to 1. Note that \( f_{Y|X}(1|0) + f_{Y|X}(1|1) = 0.04 + 0.92 = 0.96 \neq 1 \). These are two conditional probabilities from two different conditional pmfs, thus there is no reason they should add to 1.0.
For the opposite direction pmf $f_{X|Y}(x|y)$:

\[
\begin{align*}
f_{X|Y}(x = 0|y = 0) &= \frac{f(0,0)}{h(0)} = \frac{0.48}{0.52} = 0.923 \\
f_{X|Y}(x = 1|y = 0) &= \frac{f(1,0)}{h(0)} = \frac{0.04}{0.52} = 0.077 \\
f_{X|Y}(x = 0|y = 1) &= \frac{f(0,1)}{h(1)} = \frac{0.02}{0.48} = 0.042 \\
f_{X|Y}(x = 1|y = 1) &= \frac{f(1,1)}{h(1)} = \frac{0.46}{0.48} = 0.958
\end{align*}
\]

These two pmfs relate to the uncertainty in the receiver’s bit decision. Given that the receiver decides bit 0, there is a 92.3% chance the receiver was correct (that the transmitter sent a 0), and a 7.7% chance it was wrong (the transmitter really sent a 1). Given that the receiver decides bit 1, there is a 95.8% chance the receiver was correct (that the transmitter sent a 1), and a 4.2% chance it was wrong (the transmitter really sent a 0).

Make sure that you understand the MEANING of a conditional pmf as well as being very careful while calculating the probabilities within a conditional pmf.

### 19.3 Independence of Random Variables

Two random variables are independent if, for all $x, y$ in the sample space,

\[f_{X,Y}(x, y) = f_X(x)f_Y(y)\]

or, equivalently,

\[f_{X|Y}(x|y) = f_X(x)\]

The first says, the two events \{X = x\} and \{Y = y\} are independent if the joint probability $P\{X = x \cap Y = y\}$ is equal to $P[X = x]P[Y = y]$. This is the SAME definition as independence of two events, except that this statement must be true for ALL possible combinations of $X = x$ and $Y = y$, not just one of them. If you can find ANY exception, where $f_{X,Y}(x, y) \neq f_X(x)f_Y(y)$, then the two are NOT independent.

**Example: Binary Communications System**

Are $X$ and $Y$ independent in the binary communication system?

**Solution:** Let’s try a particular pair, $x = 0, y = 0$. Then the question is:

\[
\begin{align*}
f_{X,Y}(0,0) & \overset{?}{=} g(0)h(0) \\
0.48 & \overset{?}{=} (0.50)(0.52) \\
0.48 & \overset{?}{=} 0.26
\end{align*}
\]

so NO! The two r.v.s are NOT independent. We don’t need to check any further.

You can see it is actually harder to prove than disprove independence.

---

Lecture #19
20 Joint Continuous Random Variables

Def’n: Joint probability density function
\( f_{X,Y}(x,y) \) is the joint probability density function of two random variables \( X \) and \( Y \) if for any event \( A \in S_{X,Y} \),
\[
\int \int_A f_{X,Y}(x,y) \, dx \, dy = P[(X, Y) \in A]
\]

Note that when we do the double integral over \( A \), we are finding the volume under the surface of \( f_{X,Y}(x,y) \). See an example in Figure 14 (Image credit: Dwight L. Raulston, http://www2.sjs.org/raulston/mvc.10/Topic.4.Lab.1.htm)

![Figure 14: An example joint pdf](Image credit: Dwight L. Raulston, http://www2.sjs.org/raulston/mvc.10/Topic.4.Lab.1.htm)

Figure 14: An example joint pdf \( f(x,y) \). Consider the event shown on the \( XY \) plane as the square in the center of a \( 3 \times 3 \) grid of squares. The volume under the surface of \( f(x,y) \) and above the \( XY \) plane is the probability of that event.

The joint pdf has the properties:

1. Integrate to 1:
\[
\int_{x \in S_X} \int_{y \in S_Y} f_{X,Y}(x,y) \, dx \, dy = 1.
\]

2. Non-negative: \( f_{X,Y}(x,y) \geq 0 \), for all \((x,y)\).

Example: From Prof. Tolga Tasdizen: Cone pdf
Let \( X, Y \) be distributed with pdf:
\[
f_{X,Y}(x,y) = \begin{cases} 
  \frac{h^2 - \sqrt{x^2 + y^2}}{2}, & \sqrt{x^2 + y^2} < 2 \\
  0, & \text{o.w.}
\end{cases}
\]

Find the value of \( h \) which results in a valid pdf.

Solution: In short, \( h \) will be set so that the volume under the curve of the pdf is equal to 1. The volume of a cone of is \( \frac{1}{3} \pi r^2 h \), where \( r \) is the radius and \( h \) is the height. In this problem, the
height is $h$ because when you plug in $x = 0, y = 0$, the pdf is $h$. The radius $r = 2$, because when $\sqrt{x^2 + y^2} = 2$, the pdf becomes 0. So the volume is $h \frac{4\pi}{3}$. Setting this equal to 1, we have $h = \frac{3}{4\pi}$. Thus

$$f_{X,Y}(x, y) = \begin{cases} \frac{3}{4\pi} \frac{2-\sqrt{x^2+y^2}}{2}, & \sqrt{x^2+y^2} < 2 \\ 0, & \text{o.w.} \end{cases}$$ (11)

Note also that we could have integrated to find the volume of a cone. Let’s do the integration by finding the area of all of the circular disks of height $dz$ that make up the cone volume. A circular disk of height $dz$ has area $\pi r^2 dz$, for radius $r$. Here, the radius is a function of the height $z$. At height $z = 0$, the radius is 2, and at height $z = h$, the radius is 0. Between the two the radius is linear. The function of that line must be $r = mz + 2$ for some slope $m$ and $z$-intercept of 2. The slope $m$ satisfies $0 = mh + 2$, thus $m = -2/h$. Thus $r = 2 - \frac{2}{h}z$ and we can write the volume integral as:

$$1 = \int_{z=0}^{h} \pi r^2 dz = \int_{0}^{h} \pi \left(2 - \frac{2}{h} z \right)^2 dz$$ (12)
Multiplying out the quadratic and integrating,

\[
1 = \pi \int_0^h \left( 4 - 2 \frac{4}{h} z + \frac{4}{h^2} z^2 \right) dz
\]

\[
= \pi \left[ 4z - \frac{4}{h} z^2 + \frac{4}{3h^2} z^3 \right]_0^h
\]

\[
= \pi \left[ 4h - \frac{4}{h} h^2 + \frac{4}{3h^2} h^3 \right]
\]

\[
= \pi \left[ 4h - 4h + \frac{4}{3} h \right]
\]

\[
h = \frac{3}{4\pi}
\]

which is the same result as above.

**Example: Probability of radius 1 event**

Next consider the event \(A\) that the radius \(\sqrt{X^2 + Y^2}\) is less than 1. What is \(P[A]\)?

**Solution:** We are then finding the area under the surface \(f_{X,Y}(x, y)\) for \(\sqrt{x^2 + y^2} < 1\). Let’s do the problem in polar coordinates to remind ourselves how to do that. It involves setting \(r = \sqrt{x^2 + y^2}\), and putting in \(r dr d\theta\) instead of \(dxdy\).

\[
P[A] = \int_A f_{X,Y}(x, y)dxdy
\]

\[
= \int_{\theta=0}^{\pi/2} \int_{r=0}^1 \frac{3}{4\pi} \left( 2 - r \right) r dr d\theta
\]

\[
= \frac{3}{8\pi} \int_{\theta=0}^{\pi/2} \int_{r=0}^1 (2r - r^2) dr d\theta
\]

\[
= \frac{3}{8\pi} \left[ \frac{2r^2}{2} - \frac{1}{3} r^3 \right]_{r=0}^{r=1}
\]

\[
= \frac{3}{4} \left[ 1 - \frac{1}{3} \right] = 0.5.
\]

### 20.1 Marginal and Conditional pdfs

The marginal pdfs are defined as:

\[
f_X(x) = \int_{y \in S_Y} f_{X,Y}(x, y) dy
\]

\[
f_Y(y) = \int_{x \in S_X} f_{X,Y}(x, y) dx
\]
These are the single-variable pdfs derived from the joint pdf.

The conditional pdfs are defined as:

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

Note the conditional pdf is defined only where the marginal pdf is non-zero. As in the discrete case, this is not a limiting factor because anywhere the marginal is zero, the joint pdf will also be zero.

Example: Marginal and Conditional pdfs for the Cone example

See the attached plots (from Prof. Tolga Tasdizen) which show graphically the marginal and conditional pdfs that can be derived from the Cone joint pdf in Equation (11).

Example: A Joint Exponential with X and Y

Let a joint distribution of X and Y be defined as:

$$f_{X,Y}(x, y) = \begin{cases} 
    ye^{-y(x+1)}, & x > 0, y > 0 \\
    0, & o.w.
\end{cases}$$

1. What is the marginal distribution $f_Y(y)$?
2. What is the marginal distribution $f_X(x)$?
3. What is the conditional distribution $f_X(x)$?

Solution: (a) For $y > 0$,

$$f_Y(y) = \int_{x \in S_X} f_{X,Y}(x, y) dx$$
$$= \int_{x=0}^{\infty} ye^{-y(x+1)} dx$$
$$= y \int_{x=0}^{\infty} e^{-yx} e^{-y} dx$$
$$= ye^{-y} \int_{x=0}^{\infty} e^{-yx} dx$$
$$= ye^{-y} \left[ (-1/y)e^{-yx} \right]_{x=0}^{\infty}$$
$$= -e^{-y} [e^{-yx}]_{x=0}^{\infty}$$
$$= -e^{-y} [0 - 1]$$
$$= e^{-y}$$

This was only for $y > 0$. A complete pdf would be written as

$$f_Y(y) = \begin{cases} 
    e^{-y}, & y > 0 \\
    0, & o.w.
\end{cases}$$
Solution: (b) For $x > 0$, and using a table,

\[ f_X(x) = \int_{y \in S_Y} f_{X,Y}(x,y) \, dy \]
\[ = \int_{y=0}^\infty ye^{-y(x+1)} \, dy \]
\[ = \frac{1}{(x+1)^2} \left[ -e^{-y(x+1)}((x+1)y + 1) \right]_{y=0}^{y=\infty} \]
\[ = \frac{1}{(x+1)^2} \left[ 0 - 0 \cdot [0 + 1] \right] \]
\[ = \frac{1}{(x+1)^2} \]

For all $x$, the solution is:

\[ f_X(x) = \begin{cases} \frac{1}{(x+1)^2}, & x > 0 \\ 0, & \text{o.w.} \end{cases} \]

Solution: (c) For $x > 0$ and $y > 0$,

\[ f_{X\mid Y}(x\mid y) = \frac{f_{X,Y}(x,y)}{f_Y(y)} \]
\[ = \frac{ye^{-y(x+1)}}{e^{-y}} \]
\[ = \frac{ye^{-yx}}{e^{-y}} \]
\[ = ye^{-yx} \]

Again, this solution was for $x > 0$ and $y > 0$, so a complete conditional pdf is written as

\[ f_{X\mid Y}(x\mid y) = \begin{cases} ye^{-yx}, & y > 0, x > 0 \\ 0, & \text{o.w.} \end{cases} \]

20.2 Independence of Random Variables

Def’n: Independence of Continuous r.v.s
Two continuous random variables $X$ and $Y$ are independent if

\[ f_{X,Y}(x,y) = f_X(x)f_Y(y) \]

In short, if the joint pdf can be written as the product of the marginal pdfs, then the two r.v.s are independent.

Example: A Joint Exponential with $X$ and $Y$
Are $X$ and $Y$ independent?
Solution: The product of the two marginal pdfs is:
\[ f_X(x)f_Y(y) = \frac{e^{-y}}{(x + 1)^2} \]
which is NOT the same as the joint pdf. Thus the two r.v.s are not independent.

In the other direction, if two r.v.s are independent, you can write their joint pdf as a product of their pdfs.

Example: Independent Gaussian r.v.s
Assume that \( X \) and \( Y \) are both Gaussian random variables, with means \( \mu_X = 1, \mu_Y = 2, \sigma_X = 4, \) and \( \sigma_Y = 16, \) and that they are independent. What is their joint pdf?
Solution: We know from the Gaussian lecture that their marginal pdfs are as follows:
\[
\begin{align*}
    f_X(x) &= \frac{1}{\sqrt{2\pi(4)}} e^{-\frac{(x-1)^2}{2(4)}} \\
    f_Y(y) &= \frac{1}{\sqrt{2\pi(16)}} e^{-\frac{(y-2)^2}{2(16)}}
\end{align*}
\]
To write the joint pdf we take the product of the two:
\[
\begin{align*}
    f_{X,Y}(x,y) &= f_X(x)f_Y(y) \\
    &= \frac{1}{\sqrt{2\pi(4)}} e^{-\frac{(x-1)^2}{2(4)}} \cdot \frac{1}{\sqrt{2\pi(16)}} e^{-\frac{(y-2)^2}{2(16)}} \\
    &= \frac{1}{2\pi(2)(4)} e^{-\frac{(x-1)^2}{8}} e^{-\frac{(y-2)^2}{32}} \\
    &= \frac{1}{16\pi} e^{-\frac{(x-1)^2}{8}} e^{-\frac{(y-2)^2}{32}}
\end{align*}
\]

21 Correlation and Covariance

As before, we can find expected values using a joint pdf or joint pmf. The only difference is, we plug in the joint pmf or pdf where the marginal pmf or pdf used to be. For discrete r.v.s,
\[
E[g(X,Y)] = \sum_{x \in S_X} \sum_{y \in S_Y} g(x,y)f_{X,Y}(x,y)
\]
For continuous r.v.s,
\[
E[g(X,Y)] = \int_{x \in S_X} \int_{y \in S_Y} g(x,y)f_{X,Y}(x,y)dxdy
\]

Note that if you plug in \( g(X,Y) = X \) you get the expected value of \( X \), just as before. If the function \( g \) doesn’t have a \( y \) in it, that is, it is really \( g(X) \), then you can simplify the expected value:
\[
\begin{align*}
    E[g(X)] &= \int_{x \in S_X} \int_{y \in S_Y} g(x)f_{X,Y}(x,y)dxdy \\
    E[g(X)] &= \int_{x \in S_X} \int_{y \in S_Y} g(x)f_{X,Y}(x,y)dxdy \\
    E[g(X)] &= \int_{x \in S_X} g(x)f_X(x)dx
\end{align*}
\]
where the last line is because the integral w.r.t. \( y \) of the joint pdf simplifies to the marginal pdf. This result is the same as the \( E[g(X)] \) that we had prior in the notes.

Important to us is one particular function:

- \( g(X, Y) = (X - \mu_X)(Y - \mu_Y) \). The expected value of this function,

\[
\text{Cov}(X, Y) = E[(X - \mu_X)(Y - \mu_Y)]
\]

is called the covariance of \( X \) and \( Y \). It is abbreviated to \( \text{Cov}(X, Y) \). Just like the variance has a simpler form, the covariance has a simpler form:

\[
\text{Cov}(X, Y) = E[XY] - \mu_X \mu_Y
\]

The Walpole book calls the covariance \( \sigma_{XY} \), but I’ve never seen this notation in another book, and it is confusing to me. So I’m just going to use \( \text{Cov}(X, Y) \).

**Example: Covariance of Independent r.v.s**

What is \( \text{Cov}(X, Y) \) when \( X \) and \( Y \) are independent?

**Solution:** We know

\[
\text{Cov}(X, Y) = E[XY] - \mu_X \mu_Y
\]

Let’s find \( E[XY] \). Using the fact that \( f_{X,Y}(x, y) = f_X(x)f_Y(y) \) for independent r.v.s,

\[
E[XY] = \sum_{x \in S_X} \sum_{y \in S_Y} xyf_{X,Y}(x, y)
= \sum_{x \in S_X} \sum_{y \in S_Y} xyf_X(x)f_Y(y)
= \sum_{x \in S_X} xf_X(x) \sum_{y \in S_Y} yf_Y(y)
= E[X]E[Y] = \mu_X \mu_Y
\]

Thus for independent r.v.s,

\[
\text{Cov}(X, Y) = \mu_X \mu_Y - \mu_X \mu_Y = 0.
\]

Note that independence implies a covariance of 0. But a covariance of 0 does NOT imply independence. Independence is a stronger condition, and requires more than a covariance of 0.

Specifically, we get the most intuition about how \( X \) and \( Y \) vary together using a normalized version of the covariance called the correlation coefficient, \( \rho_{XY} \):

\[
\rho_{XY} = \frac{\text{Cov}(X, Y)}{\sqrt{\text{Var}[X] \text{Var}[Y]}} = \frac{\text{Cov}(X, Y)}{\sigma_X \sigma_Y}
\]

The denominator is the geometric mean of the variances of \( X \) and \( Y \).

- The covariance of \( X \) and \( Y \) will never exceed this denominator. In fact, the correlation coefficient is bounded between -1 and 1.
• When $\rho$ is +1, it indicates that all of the variation of $X$ can be explained by the variation in $Y$ (and vice versa). When it is $-1$ any increase in $X$ is deterministically explained by a decrease in $Y$ or vice versa.

• When the $\rho$ is positive, it indicates that when one variable is above its mean, the other is above its mean (on average); and that when one decreases below its mean, the other also decreases (on average).

**Example: Binary Communication System**

What is the covariance and correlation coefficient of the binary communication system? Recall:

<table>
<thead>
<tr>
<th>$f(X, Y)$</th>
<th>$X = 0$</th>
<th>$X = 1$</th>
<th>$f_Y(y)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$Y = 0$</td>
<td>0.48</td>
<td>0.04</td>
<td>0.52</td>
</tr>
<tr>
<td>$Y = 1$</td>
<td>0.02</td>
<td>0.46</td>
<td>0.48</td>
</tr>
<tr>
<td>$f_X(x)$</td>
<td>0.50</td>
<td>0.50</td>
<td></td>
</tr>
</tbody>
</table>

**Solution:** There are three things you need to find the covariance:

1. $\mu_X$. For a discrete r.v. $X$ this is:

$$E[X] = \sum_{x=0}^{1} x f_X(x) = (0)(0.5) + (1)(0.5) = 0.5$$

   General trick: When I have a Bernoulli r.v. (that takes values 0 and 1), the expected value of that r.v. is just the probability that it is equal to 1.

2. $\mu_Y$. Using the trick, $E[Y] = \sum_{y=0}^{1} y f_Y(y) = f_Y(1) = 0.48$.

3. $E[XY]$. This is

$$E[XY] = \sum_{x=0}^{1} \sum_{y=0}^{1} xy f_{X,Y}(x, y) = (1)(1) f_{X,Y}(1, 1) = 0.46$$

Now we can calculate the covariance as:

$$\text{Cov}(X, Y) = E[XY] - E[X] E[Y] = 0.46 - (0.5)(0.48) = 0.22.$$ 

Additionally, for the correlation coefficient, you need:

1. $\sigma_X^2$: We have $E[X^2] = \sum_{x=0}^{1} x^2 f_X(x) = (0^2)(0.5) + (1^2)(0.5) = 0.5$. Note the same general trick applies as for the calculation of $E[X]$. Next,

$$\text{Var}[X] = E[X^2] - (E[X])^2 = 0.5 - (0.5)^2 = 0.25$$

2. $\sigma_Y^2$: We have $E[Y^2] = 0.48$ using the same trick (that the $Y = 1$ probability is multiplied by 1 and the only term not multiplied by 0). Next

$$\text{Var}[Y] = E[Y^2] - (E[Y])^2 = 0.48 - (0.48)^2 = 0.2496$$
The correlation coefficient is
\[ \rho_{XY} = \frac{\text{Cov}(X, Y)}{\sqrt{\text{Var}[X] \text{Var}[Y]}} = \frac{0.22}{\sqrt{0.25(0.2496)}} = 0.881 \]

The fact that it is very close to 1 indicates that the transmitted bit and received bit are highly positively correlated. This makes sense, and we would hope they’d be that correlated, because we want the communication system to receive nearly the same thing that is transmitted.

Lecture #21

22 Linear Combinations of r.v.s

When we introduced the expected value, we mentioned particular properties. Assume \( X \) is a random variable, and \( a \) and \( b \) are non-random real-valued constants. Then

1. \( E[aX] = aE[X] \),
2. \( E[b] = b \),
3. \( E[aX + b] = aE[X] + b \).

The last one is the most general of the three. In general, whenever there are terms separated by plus signs, we can add together the expected values of each term. The general way to say this is, for two functions of \( X \), \( g(X) \) and \( h(X) \),

\[ E[g(X) + h(X)] = E[g(X)] + E[h(X)] \]

Moreover, let there be two random variables \( X \) and \( Y \), and let there be two functions \( g(X, Y) \) and \( h(X, Y) \). Then

\[ E[g(X, Y) + h(X, Y)] = E[g(X, Y)] + E[h(X, Y)] \]

Some simpler rules that can be derived from this:

1. \( E[g(X) + h(Y)] = E[g(X)] + E[h(Y)] \),
2. \( E[aX + bY] = aE[X] + bE[Y] \).

22.1 Expected Value for Product

Assume there are two random variables \( X \) and \( Y \). In general, we use \( E[XY] \) to find the covariance of \( X \) and \( Y \). This is

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

If \( E[XY] = E[X]E[Y] \) then \( \text{Cov}(X, Y) = 0 \), and we say that \( X \) and \( Y \) are uncorrelated. Remember that there some uncorrelated random variables are, further, independent, meaning that \( f_{X,Y}(x, y) = f_X(x)f_Y(y) \). Because independent r.v.s are always uncorrelated, they also have \( E[XY] = E[X]E[Y] \).
22.2 Variance of a Linear Combination

One of the most useful results of this section is the following. If $a$ and $b$ are non-random constants and $X$ is a random variable, then

$$\text{Var} [aX + b] = a^2 \text{Var} [X]$$

Recall that $E [aX + b]$ simplifies to $aE [X] + b$. So why does the $b$ disappear and the $a$ become $a^2$? First let’s see the proof.

**Proof:** Let $Y = aX + b$. The variance of $Y$ is defined as:

$$\text{Var} [Y] = E [(Y - E [Y])^2]$$

$$= E [(aX + b - E [aX + b])^2]$$

$$= E [(aX + b - aE [X] - b)^2]$$

$$= E [(aX - aE [X])^2]$$

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

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

$$= a^2 \text{Var} [X]$$

In short, the $b$ cancels out because in the calculation of variance, we always subtract the mean. Thus even though the distribution has shifted by $b$, we cancel out that shift by subtracting the mean. Intuitively, variance is a measure of the width of the distribution, and not the mean. Note that $a$ may be negative, however, the variance is always multiplied by the positive value $a^2$.

Note that this means that the standard deviation of $aX + b$ is $|a|$ times the standard deviation of $X$, where $|a|$ is the absolute value of $a$.

$$\sigma_{aX+b} = |a|\sigma_X$$

22.3 Two r.v.s

Now, let’s consider the more complicated case of $aX + bY$. In general, the variance of $aX + bY$ is:

$$\text{Var} [aX + bY] = a^2 \text{Var} [X] + b^2 \text{Var} [Y] + 2ab \text{Cov} (X, Y)$$

However, if $X$ and $Y$ are uncorrelated (or independent, which is a subset of uncorrelated), the $\text{Cov} (X, Y) = 0$ and thus

$$\text{Var} [aX + bY] = a^2 \text{Var} [X] + b^2 \text{Var} [Y]$$

**Example: Two resistors in series**

In a circuit, voltage drops $V_1$ and $V_2$, across two resistors in series, and are added together in $V = V_1 + V_2$. Both have mean 1.5 V and standard deviation 0.5 V.

1. What is the mean of $V$?

2. What is the variance of $V$ in the case where $V_1$ and $V_2$ are uncorrelated?

3. What is the variance of $V$ in the case where $V_1$ and $V_2$ have a correlation coefficient of 0.7?
Solution: The mean $E[V_1 + V_2] = E[V_1] + E[V_2] = 2(1.5) = 3$ Volts. The variance, when they are uncorrelated, is $\text{Var}[V_1 + V_2] = \text{Var}[V_1] + \text{Var}[V_2] = 2(0.5^2) = 0.5 \text{ V}^2$. Note this is a standard deviation of 0.707 Volt. When they have correlation coefficient of 0.7, then

$$0.7 = \frac{\text{Cov}(X, Y)}{\sqrt{\text{Var}[V_1] \text{Var}[V_2]}}$$

which implies that $\text{Cov}(X, Y) = 0.7(0.5^2) = 0.175$. Then

$$\text{Var}[V_1 + V_2] = \text{Var}[V_1] + \text{Var}[V_2] + 2\text{Cov}(X, Y) = 0.5^2 + 0.5^2 + 2(0.175) = 0.85$$

Note that the standard deviation is now $\sqrt{0.85} = 0.922$ V, thirty percent higher than in the uncorrelated case.

22.4 Many r.v.s

Consider a list of $n$ mutually uncorrelated r.v.s $X_1, X_2, \ldots, X_n$. The mean of the sum is:

$$E[X_1 + X_2 + \cdots + X_n] = E[X_1] + E[X_2] + \cdots + E[X_n]$$

The variance of the sum is:

$$\text{Var}[X_1 + X_2 + \cdots + X_n] = \text{Var}[X_1] + \text{Var}[X_2] + \cdots + \text{Var}[X_n]$$

That is, the means add together, and the variances add together.

Example: Commute Time

Your drive to work takes you through 10 stoplights. The additional time you wait at the $i$th stoplight is denoted $X_i$ and has a mean of 1 minute and a variance $\sigma^2_X = 0.5 \text{ minute}^2$. If there were no stoplights, it would be a 5 minute drive to work. What is the mean and variance of your commute? Your boss has a longer distance to drive, but only 5 stoplights. His commute would take 10 minutes with no stoplights. Assume that his stoplights have the same mean and variance of delay. You find that you both have the same average duration of your commute, 15 minutes. But he says that you should be more at work on time more reliably, because the extra stoplights in your commute average themselves out. Is he right or wrong?

Solution: We can write your commute time $C_{\text{you}}$ in minutes as:

$$C_{\text{you}} = X_1 + X_2 + \cdots + X_{10} + 5$$

The mean is

$$E[C_{\text{you}}] = E[X_1] + E[X_2] + \cdots + E[X_{10}] + 5 = 10(1) + 5 = 15$$

minutes and the variance is

$$\text{Var}[C_{\text{you}}] = \text{Var}[X_1] + \text{Var}[X_2] + \cdots + \text{Var}[X_{10}] = 10(0.5) = 5.0$$

His commute time is:

$$C_{\text{boss}} = X_1 + X_2 + \cdots + X_{5} + 10$$

The mean is

$$E[C_{\text{boss}}] = E[X_1] + E[X_2] + \cdots + E[X_{5}] + 10 = 5(1) + 10 = 15$$
minutes and the variance is

$$\text{Var}[C_{\text{boss}}] = \text{Var}[X_1] + \text{Var}[X_2] + \cdots + \text{Var}[X_5] = 5(0.5) = 2.5$$

Your commute time has a variance of 5, while his has a variance of 2.5. So your boss is wrong! Good luck telling him so.

**Example: Average**
The average of 3 random variables is $Y = \frac{X_1 + X_2 + X_3}{3}$. Assume that $X_1, X_2,$ and $X_3$ are Gaussian with identical mean $\mu_X$ and variance $\sigma^2_X$, but that they are independent. What is the mean and variance of $Y$?

**Solution:** The mean of $Y$ is:

$$E[Y] = \frac{1}{3} [E[X_1] + E[X_2] + E[X_3]] = \frac{1}{3} [3\mu_X] = \mu_X.$$

The variance of $Y$ is:

$$\text{Var}[Y] = \text{Var}\left[\frac{X_1 + X_2 + X_3}{3}\right] = \frac{1}{3^2} \text{Var}[X_1 + X_2 + X_3],$$

because when we take out the constant $1/3$ we have to square it. Then,

$$\text{Var}[Y] = \frac{1}{9} \left[\text{Var}[X_1] + \text{Var}[X_2] + \text{Var}[X_3]\right] = \frac{1}{9} [3\sigma^2_X],$$

and thus $\text{Var}[Y] = \frac{1}{3} \sigma^2_X$.

In general, if $Y$ is the average of $n$ independent, identically distributed random variables, then $E[Y] = \mu_X$ and $\text{Var}[Y] = \frac{1}{n} \sigma^2_X$, or equivalently, the standard deviation of $Y$ is $\frac{1}{\sqrt{n}}$ times the standard deviation of $X$.

---

**Lecture #22**

**23 Examples**

**Example: Joint continuous r.v.s**

Let $X$ and $Y$ have joint pdf,

$$f_{X,Y} = \begin{cases} a(2 - x)(1 - 0.5y), & 0 < x < y < 1 \\ 0, & \text{o.w.} \end{cases}$$  \quad (14)

Find:

1. What must $a$ be to make this a valid pdf?
2. Find the marginals $f_X(x)$ and $f_Y(y)$.
3. Are $X$ and $Y$ independent?

5. Find the covariance of $X$ and $Y$.

6. Find the correlation coefficient $\rho_{XY}$.

Solution:

1. To find $a$, use the fact that the pdf must integrate to 1:

$$
1 = \int_{y=0}^{1} \int_{x=0}^{y} a(2-x)(1-0.5y)dx
dy
$$

$$
1 = a \int_{y=0}^{1} (1-0.5y) \int_{x=0}^{y} (2-x)dx
dy
$$

$$
1 = a \int_{y=0}^{1} (1-0.5y) \left[ 2x - \frac{x^2}{2} \right]_{x=0}^{y} dy
$$

$$
1 = a \int_{y=0}^{1} (1-0.5y) \left[ 2y - \frac{y^2}{2} \right] dy
$$

$$
1 = a \int_{y=0}^{1} \left[ 2y - \frac{3y^2}{2} + \frac{y^3}{4} \right] dy
$$

$$
1 = a \left[ y^2 - \frac{y^3}{2} + \frac{y^4}{16} \right]_{y=0}^{1}
$$

$$
1 = a \left[ 1 - \frac{1}{2} + \frac{1}{16} \right]
$$

$$
1 = a \frac{9}{16}
$$

Thus $a = \frac{16}{9}$.

2. Next, find the marginals:

$$
f_X(x) = \int_{y=x}^{1} \frac{16}{9}(2-x)(1-0.5y)dy
$$

$$
f_X(x) = \frac{16}{9}(2-x) \int_{y=x}^{1} (1-0.5y)dy
$$

$$
f_X(x) = \frac{16}{9}(2-x) \left[ y - \frac{y^2}{4} \right]_{y=x}^{1}
$$

$$
f_X(x) = \frac{16}{9}(2-x) \left[ \frac{3}{4} - \left( x - \frac{x^2}{4} \right) \right]
$$

$$
f_X(x) = \frac{16}{9}(2-x) \left( \frac{3}{4} - x + \frac{x^2}{4} \right)
$$

$$
f_X(x) = \frac{16}{9} \left( \frac{3}{2} - 2x + \frac{x^2}{2} - \frac{3}{4}x + x^2 - \frac{x^3}{4} \right)
$$

$$
f_X(x) = \frac{16}{9} \left( \frac{3}{2} - \frac{11}{4}x + \frac{3}{2}x^2 - \frac{1}{4}x^3 \right)
$$

(16)
Note this is for $0 \leq x \leq 1$, and it is zero otherwise.

\[
f_Y(y) = \int_{x=0}^{y} \frac{16}{9} (2 - x)(1 - 0.5y) dx
\]

\[
f_Y(y) = \frac{16}{9} (1 - 0.5y) \int_{x=0}^{y} (2 - x) dx
\]

\[
f_Y(y) = \frac{16}{9} (1 - 0.5y) \left[ (2x - 0.5x^2) \right]_{x=0}^{y}
\]

\[
f_Y(y) = \frac{16}{9} (1 - 0.5y) \left[ 2y - 0.5y^2 \right]
\]

\[
f_Y(y) = \frac{16}{9} \left[ 2y - \frac{3}{2}y^2 + \frac{1}{4}y^3 \right]
\]

(17)

Note this is for $0 \leq y \leq 1$, and it is zero otherwise.

3. Since the product $f_X(x)f_Y(y)$ is not equal to $f_{XY}(x,y)$, the two r.v.s are not independent.

4. The expected value of $X$ is

\[
E[X] = \int_{x=0}^{1} x \frac{16}{9} \left( \frac{3}{2} - \frac{11}{4} x + \frac{3}{2} x^2 - \frac{1}{4} x^3 \right) dx
\]

\[
E[X] = \frac{16}{9} \int_{x=0}^{1} \left( \frac{3}{2} - \frac{11}{4} x^2 + \frac{3}{2} x^3 - \frac{1}{4} x^4 \right) dx
\]

\[
E[X] = \frac{16}{9} \left( \frac{3}{4} - \frac{11}{12} + \frac{3}{8} - \frac{1}{20} \right)
\]

\[
E[X] = \frac{16}{9} \left( \frac{90 - 110 + 45 - 6}{120} \right)
\]

\[
E[X] = 0.28
\]

(18)

The expected value of $Y$ is

\[
E[Y] = \int_{y=0}^{1} y \frac{16}{9} \left[ 2y - \frac{3}{2}y^2 + \frac{1}{4}y^3 \right] dy
\]

\[
E[Y] = \int_{y=0}^{1} \frac{16}{9} \left[ 2y^2 - \frac{3}{2}y^3 + \frac{1}{4}y^4 \right] dy
\]

\[
E[Y] = \frac{16}{9} \left[ \frac{2}{3} - \frac{3}{8} + \frac{1}{16} \right]
\]

\[
E[Y] = 0.63
\]

(19)

5. To find the covariance, the easiest way is probably to use the form, $\text{Cov}(X,Y) = E[XY] - E[X]E[Y]$. The expected value of the product is evaluated as the integral of $xy$ times the joint pdf:

\[
\int_{y=0}^{1} \int_{x=0}^{y} xy \frac{16}{9} (2 - x)(1 - 0.5y) dx dy.
\]
6. The correlation coefficient requires both of the variances, $\sigma_X^2$ and $\sigma_Y^2$. To find those, the easiest way is to find $E[X^2]$ and $E[Y^2]$:

$$E[X^2] = \int_{x=0}^{1} x^2 \frac{16}{9} \left( \frac{3}{2} - \frac{11}{4} x + \frac{3}{2} x^2 - \frac{1}{4} x^3 \right) dx$$

$$E[Y^2] = \int_{y=0}^{1} y^2 \frac{16}{9} \left[ 2y - \frac{3}{2} y^2 + \frac{1}{4} y^3 \right] dy \quad (20)$$

Then use $\sigma_X^2 = E[X^2] - (E[X])^2$ and $\sigma_Y^2 = E[Y^2] - (E[Y])^2$.

**Example: Joint discrete r.v.s**

Let $X$ and $Y$ have joint pmf,

<table>
<thead>
<tr>
<th>$f_{X,Y}(x,y)$</th>
<th>$X = 0$</th>
<th>$X = 1$</th>
<th>$X = 2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$Y = 0$</td>
<td>0.15</td>
<td>0.1</td>
<td>0.15</td>
</tr>
<tr>
<td>$Y = 1$</td>
<td>0.15</td>
<td>0.15</td>
<td>0.05</td>
</tr>
<tr>
<td>$Y = 2$</td>
<td>0.15</td>
<td>0.05</td>
<td>0.05</td>
</tr>
</tbody>
</table>

1. What is $f_X(1)$?

2. What is $f_{X|Y}(x = 1|y = 2)$?

3. True or False: $X$ and $Y$ are independent.

**Example: Stoplights**

The probability of having to stop at any given stoplight is 0.6. If your route includes 5 stoplights, what is the probability that you will have to stop at one or zero stoplights?

**Example: Load and Outage**

The current draw on a circuit is Gaussian with mean 12 Amps and standard deviation 1.55 Amps. The circuit breaker trips when the current exceeds 15 Amps. What is the probability that the circuit breaker will trip?

---

**Lecture #23**

### 24 Chebychev’s Theorem

As engineers, when we design a system, we’re often asked to provide some guarantees. We can often make these probabilistic, that is, with a probability of $AA$, characteristic (random variable) $X$ will fall within a particular range. Coming up with this range (and the probability) will be a major focus of the last 1/3 of this course.

If you know the CDF of $X$, you can find many such ranges. I don’t need to present to you any more on how to do that from a known CDF.

If you don’t know the distribution of $X$ but you know the standard deviation, you can still get some useful information about the probability of $X$ falling in a certain interval. Chebychev’s theorem states:
**Theorem:** The probability that any random variable $X$ will assume a value within $k$ standard deviations of its mean is at least $1 - \frac{1}{k^2}$. That is,

$$P[\mu_X - k\sigma_X < X < \mu_X + k\sigma_X] \geq 1 - \frac{1}{k^2}$$

**Proof:** Beyond the scope of this course.

The bound is usually very loose. That is, if you knew the distribution, you’d use the CDF to get a much more accurate value for the probability $X$ will assume a value within $k$ standard deviations of its mean.

**Example: Chebychev with mean 0 and std. deviation of 1**

What is the Chebychev bound on the probability that a zero-mean random variable with standard deviation 1 is:

- $-1 < X < 1$
- $-2 < X < 2$
- $-3 < X < 3$
- $-4 < X < 4$
- $-5 < X < 5$

Compare the first 3 with the result for $Z$, a standard normal r.v.

**Solution:**

- $P[-1 < X < 1] \geq 1 - \frac{1}{1^2} = 0$. Great. We already knew probabilities were non-negative.
- $P[-2 < X < 2] \geq 1 - \frac{1}{2^2} = 0.75$.
- $P[-3 < X < 3] \geq 1 - \frac{1}{3^2} = 0.889$.
- $P[-4 < X < 4] \geq 1 - \frac{1}{4^2} = 0.9375$.
- $P[-5 < X < 5] \geq 1 - \frac{1}{5^2} = 0.96$.

For a standard normal r.v.,

- $P[-1 < X < 1] = F_Z(1) - F_Z(-1) = 0.8413 - 0.1587 = 0.6826$.
- $P[-2 < X < 2] = F_Z(2) - F_Z(-2) = 0.9772 - 0.0228 = 0.9544$.
- $P[-3 < X < 3] = F_Z(3) - F_Z(-3) = 0.9987 - 0.0013 = 0.9974$.

**Example: 3 Ohm resistors**

A manufacturer assures you that the standard deviation of his 3 Ohm resistors is 0.1 Ohm, but that he doesn’t know their distribution. You’re sure that their mean is 3 Ohms, but your measurements indicate that 10% of the resistors are more than 0.5 Ohms below or above 3 Ohms. Does this give you enough information to assure yourself that the standard deviation is not 0.1 Ohm?
Solution: Given the 0.1 Ohm standard deviation, 0.5 Ohms is 5 standard deviations. So from Chebychev’s theorem, only 4% or less of values can be outside of the 2.5 to 3.5 range. The fact that 10% of the measurements do contradicts this.

Lecture #24

25 Intro to Statistics

Def’n: Population
The entire set of observations with which we are concerned.

For example, if we manufacture a product, this would be every one of the product we have manufactured or will manufacture. If we are interested in the blood pressure of Americans, it would be the blood pressure of every person in the country.

Def’n: Sample
A subset of the population.

Clearly, one may not observe everything or everyone. So we are interested in the probability of the observations on some small (random) sample.

Def’n: Joint Distribution on Random Sample
Assume a sample of \( n \) of random variables \( X_1, \ldots, X_n \). Assume the r.v.s are independent and identically distributed (i.i.d.), each with distribution function \( f_X(x) \). Then the joint probability distribution of \( X_1, \ldots, X_n \) is written

\[
    f_{X_1, \ldots, X_n}(x_1, \ldots, x_n) = \prod_{i=1}^{n} f_X(x_i)
\]

Def’n: Statistic
A function of \( X_1, \ldots, X_n \).

The most common statistic is the sample mean or equivalently, the average. These two words are interchangeable. The words “sample mean” are a little bit less ambiguous, because we tend in English to use average and mean interchangeably. By saying “sample mean”, we explicitly say that we are referring to a statistic of a sample.

Def’n: Sample Mean

\[
    \bar{X} = \frac{1}{n} \sum_{i=1}^{n} X_i
\]

Def’n: Sample Median
Let \( x_{(k)} \) be the \( k \)th biggest value in the sample \( X_1, \ldots, X_n \). In other words, \( x_{(1)}, x_{(2)}, \ldots, x_{(n)} \) is the same numbers as \( X_1, \ldots, X_n \) but in sorted order. The sample median \( \hat{x} \) is \( x_{(n/2)} \) if \( n \) is even, and the average of \( x_{((n-1)/2)} \) and \( x_{((n+1)/2)} \) if \( n \) is odd.
**Def’n: Sample Mode**  
The sample mode is the number that appears most often in $X_1, \ldots, X_n$.  

We also have a way of estimating the variance and standard deviation of distribution by using the sample. This is called the “sample variance” and “sample standard deviation”.  

**Def’n: Sample Variance**  
The sample variance, $S^2$, is  

$$S^2 = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \bar{X})^2.$$  

The sample standard deviation is $S = \sqrt{S^2}$.  

Notes:  
1. We use $n-1$ in the denominator rather than $n$. We won’t explain why in detail, at this point. In short, dividing by $n$ results in biased estimator, while dividing by $n-1$ is not biased.  
2. Another way to calculate the sample variance is:  

$$S^2 = \frac{n}{n-1} \left[ \left( \frac{1}{n} \sum_{i=1}^{n} X_i^2 \right) - (\bar{X})^2 \right]$$  

A proof of the second note:  

\begin{align*}  
S^2 &= \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \bar{X})^2 \\
&= \frac{1}{n-1} \sum_{i=1}^{n} (X_i^2 - 2X_i\bar{X} + \bar{X}^2) \\
&= \frac{1}{n-1} \left[ \sum_{i=1}^{n} X_i^2 - 2\bar{X} \sum_{i=1}^{n} X_i + n\bar{X}^2 \right] \\
&= \frac{n}{n-1} \left[ \frac{1}{n} \sum_{i=1}^{n} X_i^2 - 2\bar{X} \frac{1}{n} \sum_{i=1}^{n} X_i + \bar{X}^2 \right] \\
&= \frac{n}{n-1} \left[ \frac{1}{n} \sum_{i=1}^{n} X_i^2 - 2\bar{X} \bar{X} + \bar{X}^2 \right] \\
&= \frac{n}{n-1} \left[ \frac{1}{n} \sum_{i=1}^{n} X_i^2 - \bar{X}^2 \right]  
\end{align*}  

(21)  

An outlier is a value that is far away from the other values in the sample. There isn’t only one definition of an outlier. In general, this is a qualitative concept until someone provides an explicit definition.
Example: Grades
Ten grades on an exam are as follows:

\[ [85.6, 99.3, 76.3, 71.1, 74.7, 12.2, 79.6, 90.2, 90.0, 76.7] \]

1. What is the sample mean?
2. What is the sample median?
3. What is the sample standard deviation?

Note: I actually generated these numbers from a Gaussian distribution with \( \mu = 82.0 \) and standard deviation 10, but then added in the “12.2” value as an outlier.

Solution: The sample mean is 75.57. The sorted values in the array are:

\[ [12.2, 71.1, 74.7, 76.3, 76.7, 79.6, 85.6, 90.0, 90.2, 99.3] \]

Thus the median (the average of the 5th and 6th element, since \( n = 10 \) is even) is \( (76.7 + 79.6)/2 = 78.15 \). The sample variance is

\[
S^2 = \frac{n}{n-1} \left[ \frac{1}{n} \sum_{i=1}^{n} X_i - \left( \bar{X} \right)^2 \right] = \frac{10}{9} (6224.877 - 5526.282) = 776.2
\]

The sample standard deviation is thus \( S = \sqrt{776.2} = 27.8 \).

What if the outlier had not been in the data? That is, if there were nine values and the 12.2 excluded? We would have had a sample mean of 82.6, a sample median of 79.6, and a sample standard deviation of 8.7. You can see that one outlier has a big impact on the sample mean and standard deviation, but less effect on the median.

Def’n: Quantile
A quantile of a sample, \( q(f) \), is a value for which a fraction \( f \) of the data is less than or equal to the value \( q(f) \).

1. The median is \( q(0.5) \).
2. The 75th percentile is \( q(0.75) \).
3. The 25th percentile is \( q(0.25) \).

How to find the median, 25th, and 75th quantiles?

1. Sort the data from lowest to highest. For example, use Matlab “sort” or Python “sorted” functions.
2. The 25th percentile is the \( \lceil \frac{25n}{100} \rceil \)th element where \( \lceil x \rceil \) means round up to the next highest integer.

What is are the 25th and 75th percentiles of the grade data? Solution: There are \( n = 10 \) data points. Thus we are looking for the \( \lceil \frac{250}{100} \rceil = 3 \)rd element and \( \lceil \frac{750}{100} \rceil = 8 \)th element, that is, 74.7 and 90.0.
25.1 What is the topic, “statistics”?

Perhaps the key point about statistics is that they are random. For example, if you were to run
an experiment twice, even with the same \( n \), you’d get a different value for the sample mean. The
field of “statistics”, in large part, provides tools to quantify the randomness of any given statistic,
for example:

1. What is the mean and variance of the statistic?

2. What is the distribution of the statistic?

3. Given the first two answers, what is a confidence interval for the statistic, that is, a range in
which you know the statistic will fall with some given probability?

If we answer these questions, and the statistic falls far outside of our confidence interval, we have
evidence that, perhaps our answers to the first two questions are not correct. There are particular
hypothesis tests which we will perform which are simple threshold tests – if the statistic falls outside
of a range, we will say that it provides evidence that one hypothesis is not correct.

Lecture #25

26 Central Limit Theorem / Distribution of Sample Mean

Suppose we have a random sample of size \( n \)

\[ X_1, X_2, \ldots, X_n \]

each with the same mean \( \mu \) and standard deviation \( \sigma \). We talked about the sample mean, which is
called a statistic because it is a function of the sample data. The particular function is:

\[ \bar{X} = \frac{1}{n} \sum_{i=1}^{n} X_i = \frac{1}{n} \left[ X_1 + X_2 + \cdots + X_n \right]. \]

It is critical to realize that \( \bar{X} \) is a random variable itself. It is not the same thing as \( \mu \! \! 1 \)

Example: What is the mean and variance of \( \bar{X} \)?

Solution: The expected value is:

\[ E \left[ \bar{X} \right] = \frac{1}{n} E \left[ X_1 + X_2 + \cdots + X_n \right] = \frac{1}{n} \left\{ E \left[ X_1 \right] + E \left[ X_2 \right] + \cdots + E \left[ X_n \right] \right\} = \frac{1}{n} n \mu = \mu \]

Recall that the r.v.s in the sample are assumed to be independent, so the variance of a sum is the
sum of the variances. But recall that any constant multiplier is squared when taken out of the
variance operator.

\[ \text{Var} \left[ \bar{X} \right] = \text{Var} \left[ \frac{1}{n} \sum_{i=1}^{n} X_i \right] = \frac{1}{n^2} \sum_{i=1}^{n} \text{Var} \left[ X_i \right] = \frac{1}{n^2} \sum_{i=1}^{n} \sigma^2 = \frac{1}{n^2} n \sigma^2 = \frac{\sigma^2}{n}. \]
Note that the variance is reduced by a factor of \( n \) compared to the original samples \( X_i \). The standard deviation is reduced by a factor of \( \sqrt{n} \). We could call the standard deviation of \( \bar{X} \) as \( \sigma_{\bar{X}} \), and then
\[
\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}}.
\]

**Activity:** Within a group, add together the seven digits (excluding area code) of a randomly selected phone number, four times, so that you are adding 28 digits. Divide by the number of digits summed to find the average. The mean of a discrete uniform r.v. on \( \{0, \ldots, 9\} \) should be \( \mu = (b+a)/2 = 4.5 \). The variance of this discrete uniform r.v. should be \( \sigma^2 = [(b-a+1)^2 - 1]/12 = 99/12 = 8.25 \). This means the variance of the average should be \( 8.25/27 = 0.30 \) or that the standard deviation of the mean should be 0.55. If Gaussian, 68% of the averages should be within one standard deviation of the mean, that is, between 3.95 and 5.05.

What could be going wrong in the above analysis?

If, IN ADDITION, the distribution of \( X_i \) are Gaussian, then the distribution of \( \bar{X} \) is also Gaussian. This is an exact result, discussed in Walpole Section 7.3. The proof is not covered in this course, because we don’t describe derived distributions. (Take ECE 5510 to learn more.) Just know that when you add independent Gaussian r.v.s, you get a Gaussian r.v.

However, NO MATTER WHAT the distribution the sample data is, in the limit as the sample size increases, the distribution of \( \bar{X} \) becomes Gaussian. For any finite but large sample size \( n \) (typically \( n \geq 30 \) is good enough), we will approximate \( \bar{X} \) as Gaussian. Here is the precise theorem:

**Theorem: Central Limit Theorem:** If \( \bar{X} \) is the sample mean of a sample size \( n \) taken from a population with mean \( \mu \) and finite variance \( \sigma^2 \), then the limiting form of the distribution of
\[
Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}}
\]
as \( n \to \infty \), is the standard normal distribution.

**Proof:** Not covered.

Walpole says that \( n > 30 \) is good enough “if the data is not terribly skewed”. But it really depends on how accurate you need it to be, and how far down the tail you are going to look.

Note that one also must know the population mean and standard deviation, which are hard to know. At this point, the important point is that \( \bar{X} \) is Gaussian with some mean and variance, and we can further normalize it to a standard normal r.v. if we know the population mean and variance. Soon, we will make things more realistic (at the cost of making it more complicated) by replacing the population standard deviation \( \sigma \) with the sample standard deviation \( S \).

**Example: High Sample Mean of Resistances**
You buy a block of 1000 resistors which are supposed to have mean 100 \( \Omega \) and standard deviation of 2 \( \Omega \). You measure 10 resistors and find that \( \bar{X} = 101.2 \Omega \). What is the probability, given the manufacturers specs, that this test would measure a sample average of 101.2 or above?

**Solution:** The sample mean has an expected value of 100 and a standard deviation of \( 2/\sqrt{10} = 0.632 \).
\[
P[X > 101.2] = P \left[ Z > \frac{101.2 - 100}{0.632} \right] = 1 - F_Z(1.61) = 1 - 0.9463 = 0.0537.
\]
Activity: Based on this data, you decide your resistors don’t meet the advertised spec. You call up the manufacturer and demand a refund. List some potential problems with your analysis and decision?

Lecture #26

In the previous handout, we learned about the distribution of the random variable $\bar{X}$, the sample mean. In this section, we continue by studying the distribution of the sample variance $S^2$. In the same way as $\bar{X}$ is a random variable, $S^2$ is also a random variable. In Walpole Section 8.5, the following theorem is presented. Only a qualitative proof is given, and it relies on Chapter 7, which we did not cover.

**Theorem:** If $S^2$ is the sample variance of a random sample of size $n$ taken from a Gaussian population having variance $\sigma^2$, then

$$\chi^2 = \frac{(n-1)S^2}{\sigma^2} = \sum_{i=1}^{n} \frac{(X_i - \bar{X})^2}{\sigma^2}$$

has a chi-squared distribution with parameter $\nu = n - 1$.

**Proof:** Not covered.

The chi-squared distribution is

$$f_X(x; \nu) = \begin{cases} \frac{1}{2^{\nu/2} \Gamma(\nu/2)} x^{\nu/2-1} e^{-x/2}, & x > 0 \\ 0, & \text{o.w.} \end{cases}$$

where $\Gamma(x)$ is the Gamma function, which is like a continuous version of the factorial function. The Gamma function is defined on page 194 of Walpole. Note the mean of a chi-squared r.v. is $\nu$, and the variance is $2\nu$.

We won’t directly use the chi-squared pdf. We’ll be most interested in the CDF, which, similar to the Gaussian pdf, doesn’t lend itself to easy calculation. Instead, we need a table. Walpole has a chi-squared “critical values” table in Table A.5. Note this table is very different from the standard normal CDF table in Table A.2. First, its probabilities are the $P[\chi^2 > \chi^2_\alpha]$, that is, the area to the right of a threshold value $\chi^2_\alpha$. Second, the table is an inverse table. Each column is labeled with the probability $\alpha = P[\chi^2 > \chi^2_\alpha]$, and each row is labeled with $\nu$, the distribution’s “degrees of freedom” parameter. The value you read in the table is the threshold $\chi^2_\alpha$. Typically, in the problems we will need to solve, one knows the tail probability $\alpha$, and one needs to find the value of the threshold so that the area to the right of that threshold is $\alpha$.

**Example: Interval for $S^2$**

Consider calculating $S^2$ using $n = 10$ samples from a standard normal distribution. What is a range across which we can expect 95% of the calculated sample variances to lie within?

**Solution:** Since $n = 10$, the number of degrees of freedom is $\nu = 9$. This is an ambiguous question, because it doesn’t specify where the range is — it could be from 0 to $\chi^2_{0.05}$, or it could be from $\chi^2_{0.025}$. Or it could be from $\chi^2_{0.95}$ to $\chi^2_{0.05}$. From Table A.5, for $\nu = 9$ and the range, $\chi^2_{0.025}$ to $\chi^2_{0.975}$, we see $\chi^2_{0.975} = 2.700$ to $\chi^2_{0.025} = 19.023$. Since $\chi^2 = \frac{(n-1)S^2}{\sigma^2}$, we can convert the limits
on $\chi^2$ to limits on $S^2$ as follows:

$$2.700 < \chi^2 < 19.023$$
$$2.700 < \frac{(n-1)S^2}{\sigma^2} < 19.023$$
$$2.700 \frac{\sigma^2}{n-1} < S^2 < 19.023 \frac{\sigma^2}{n-1}$$

Since the data was Gaussian with variance of 1, and $n = 10$, $2.700/9 < S^2 < 19.023/9$ or $0.30 < S^2 < 2.11$. Taking the square root of all sides,

$$0.548 < S < 1.454$$

Note the wide range on the sample standard deviation. Note that this range was chosen to have equal 2.5% chance of being below the bottom threshold (0.548) as being above the top threshold (1.454).

### 27 Using Sample Instead of Population Variance

We discussed how

$$Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}}$$

has a standard normal distribution for large $n$, regardless of the underlying distribution of the sample data, due to the Central Limit Theorem. If one knows the true population standard deviation $\sigma$, one may use this statistic to determine the probabilities of seeing a $\bar{X}$ a given distance from the true population mean. However, the true population standard deviation $\sigma$ is not likely to be known, exactly. In many cases, we have to estimate $\sigma$ using the sample standard deviation $S$; that is, use the data itself to determine how much variation there is in the population.

If we replace $\sigma$ with $S$ in the above expression, we have:

$$\frac{\bar{X} - \mu}{S/\sqrt{n}}$$

But we no longer have a standard normal r.v. In fact, we no longer have a Gaussian r.v. This is because we’re dividing a Gaussian random variable by another random variable $S$. We noted in the previous section that $S^2$ has a chi-squared distribution. When $n \to \infty$, we would expect the fraction to approach a standard normal distribution. However, when we only have a few samples (i.e., $n$ is low), this makes a big difference. We need to account for the fact that $S$ is not particularly accurate for low $n$. What we end up with is that the above fraction has the “Student $t$” distribution, or the $t$-distribution, for short.

$$T = \frac{\bar{X} - \mu}{S/\sqrt{n}}$$

Here, $T$ has the $t$-distribution with $\nu = n - 1$ degrees of freedom, given as:

$$f_T(t) = \frac{\Gamma[(\nu + 1)/2]}{\Gamma(\nu/2)\sqrt{\pi\nu}} \left(1 + \frac{t^2}{\nu}\right)^{-(\nu+1)/2}$$

this formula is true for all $-\infty < t < \infty$. 

The t-distribution was derived by William S. Gosset, an employee of the Guinness brewery in Dublin, in 1908. The brewery was very interested in finding the best processes for brewing beer, including what the best varieties of barley were. Gosset was trying to determine accurate statistical tests to decide when one process was significantly better than another process. However, his employer would not permit its employees to publish. Gosset successfully argued that his distribution was of little use to any other brewery, and was allowed to publish anonymously. He submitted the publication under the name, Student.

The t-distribution looks like the Gaussian bell-shaped curve, but with heavier tails. It is also symmetric about \( t = 0 \). See Figure 8.8 and 8.9 in Walpole. As \( n \to \infty \), the t-distribution becomes the same as the standard normal.

Again, we don’t use the t-distribution pdf directly. Instead, we’ll use Table A.4, which has the “critical” values of the t-distribution complementary CDF. That is, we will know what probability \( \alpha = P [ T > t_\alpha ] \) we want at the tail of the distribution, and we will look up in the table that tail probability (the column) and the degrees of freedom \( \nu \) (the row) and find the threshold \( t_\alpha \) we need.

The t-distribution is symmetric about 0. Thus the median is always 0. Also, if we want the threshold on the negative side which has \( \alpha \) probability that \( P [ T < t_\alpha ] \), it is the same as finding the negative of the threshold on the positive side for \( P [ T > t_\alpha ] \).

Example: Mean of Group Data
In your group you measured a sample of 28 digits from phone numbers. Let’s go back to the time before I told you the mean and variance of the population distribution. You, as the engineer, need to test the assumption that the mean is 4.5. If the \( t \)-value is less than \(-t_{0.025}\) or greater than \( t_{0.025}\), then you will decide that the mean is not zero; if it falls between those two values, you will be satisfied that the mean is zero. Find the \( t \)-value and decide whether or not the mean is zero.

**Solution:** The solution depends on your data. Let’s say we calculated \( \bar{X} = 5.8 \) and \( S = 3.24 \) from our sample of \( n = 28 \) data values. Then, applying the t-statistic formula,

\[
t = \frac{\bar{X} - \mu}{S/\sqrt{n}} = \frac{5.8 - 4.5}{3.24/\sqrt{28}} = 2.12
\]

Looking in Table A.4 for a 0.025 percent tail probability and \( \nu = n - 1 = 9 \), \( t_{0.025} = 2.262 \). Since 2.12 falls in the range between -2.262 and +2.262, we are satisfied with the assumption that the mean is zero.

---

**Lecture #27**

28 Intro to Estimation

The statistics we’ve talked about so far (sample mean, sample variance, sample median, etc.) are also called *estimators* because they are used to estimate some parameter of the distribution. There might be many ways to estimate the same thing – for example, the sample mean and sample median both estimate the mean of the distribution. Why would we consider one to be “better” or “worse” than another? In this section, we discuss two metrics to evaluate the “performance” of an estimator: 1) *bias*, and 2) *variance*.

Bias and variance are two aspects of error, similar to the difference between accuracy and precision. In this case, bias is a measure of accuracy, and variance is a measure of precision. But
we will define both bias and variance in terms of probability.

In this section we generically refer to the true parameter as $\theta$ and the estimate of the parameter as $\hat{\Theta}$. Some examples of estimators and parameters are given in Table 28.

<table>
<thead>
<tr>
<th>Estimator $\Theta$</th>
<th>Formula for $\hat{\Theta}$</th>
<th>Parameter $\theta$ Being Estimated</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sample Mean $\bar{X}$</td>
<td>$\frac{1}{n} \sum_{i=1}^{n} X_i$</td>
<td>$\mu$</td>
</tr>
<tr>
<td>Sample Median $\tilde{X}$</td>
<td>$X_{\left(\frac{n+1}{2}\right)}$ for odd $n$</td>
<td>$\mu$</td>
</tr>
<tr>
<td>Sample Std. Dev. $S$</td>
<td>$\sqrt{\frac{1}{n-1} \sum_{i=1}^{n} (X_i - \bar{X})^2}$</td>
<td>$\sigma$</td>
</tr>
<tr>
<td>Sample Binomial Success Prob. $\hat{p}$</td>
<td>$K/n$ for $K$ successes</td>
<td>Binomial $p$</td>
</tr>
</tbody>
</table>

Table 2: Examples of estimator $\hat{\Theta}$ and the distribution parameter $\theta$ they are intended to estimate.

28.1 Estimator Bias

**Def’n: Bias**

The bias of an estimator $\hat{\Theta}$ of a parameter $\theta$ is $E[\hat{\Theta}] - \theta$, that is, the difference between the expected value of the estimate and the actual parameter value.

![Figure 16: An estimator $\hat{\Theta}$ is a random variable with a probability distribution function $f_{\hat{\Theta}}(\cdot)$. If it is unbiased the expected value of $\hat{\Theta}$ is equal to the parameter $\theta$ it is supposed to estimate. If not, its bias is the difference $E[\hat{\Theta}] - \theta$. The standard deviation of the estimator is related to its precision – the smaller the standard deviation, the more precise the estimator.](image-url)

**Def’n: Unbiased**

An estimator is unbiased if the bias is zero, i.e., $E[\hat{\Theta}] = \theta$. 
A good estimator is *accurate*, that is, even though it is a function of random variables, its average over many, many trials would return the correct parameter value. We call it “bias” because it makes us artificially believe that the parameter value is off, no matter how much data we have available. Some estimators are biased, but these are not estimators we generally want to use. If possible, we try to remove the bias.

Often we will limit our discussion of estimators to those that are unbiased. There may be many estimators of a parameter that are unbiased. For example, consider estimators of the mean $\mu$:

1. $\bar{X}$: The sample mean $\bar{X}$ we’ve already shown to have an expected value of $\mu$. Thus it is an unbiased estimator.

2. $\tilde{X}$: The median $\tilde{X}$ also has an expected value of $\mu$ when the data is Gaussian – or for that matter, any distribution symmetric about its mean. In this case, the median is unbiased. However, if the distribution is not symmetric, the median will be biased.

**Example: Is the Sample Variance Unbiased?**

Calculate the bias of the sample variance, and determine whether or not it is unbiased.

**Solution:** The sample variance $S^2$ is:

$$ S^2 = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \bar{X})^2. $$

$S^2$ is an estimator of the variance, $\sigma^2$. Finding the bias is finding the expected value of $S^2$ minus the parameter being estimated, $\sigma^2$.

$$ E[S^2] = \frac{1}{n-1} E \left[ \sum_{i=1}^{n} (X_i - \bar{X})^2 \right] $$

(22)

This is where we use a trick: we will expand the expression in the sum by using another relationship derived on p. 244 in Walpole. Here is the equality:

$$ \sum_{i=1}^{n} (X_i - \bar{X})^2 = -n(\bar{X} - \mu)^2 + \sum_{i=1}^{n} (X_i - \mu)^2 $$

Thus our expression in (22) becomes:

$$ E[S^2] = -\frac{n}{n-1} E[ (\bar{X} - \mu)^2 ] + \frac{1}{n-1} \sum_{i=1}^{n} E[(X_i - \mu)^2]. $$

Since the mean of $\bar{X}$ is $\mu$, the expression, $E[(\bar{X} - \mu)^2]$, is simply the variance of the sample mean, *i.e.*, $\text{Var} [\bar{X}]$. Similarly, the expression $E[(X_i - \mu)^2]$ is the variance of $X$. So,

$$ E[S^2] = -\frac{n}{n-1} \text{Var} [\bar{X}] + \frac{1}{n-1} \sum_{i=1}^{n} \text{Var} [X]. $$

We use $\sigma_X^2$ for the variance of $X$. Also, the variance of the sample mean is $\sigma_X^2/n$. So

$$ E[S^2] = -\frac{n}{n-1} \frac{\sigma_X^2}{n} + \frac{1}{n-1} n \sigma_X^2 = \sigma_X^2 \frac{-1 + n}{n-1} = \sigma_X^2. $$
Since $E[S^2] = \sigma_X^2$, the bias $E[S^2] - \sigma_X^2 = 0$, or in other words, the estimator is unbiased.

What if $S^2$ was defined as $\frac{1}{n} \sum_{i=1}^{n} (X_i - \mu)^2$ (instead of $\frac{1}{n-1}$ times the sum)? What would the bias be?

### 28.2 Estimator Variance

For an estimator, the lower the variance (or standard deviation), the better. The standard deviation describes how wide a distribution is. If we come up with two unbiased estimators for the same parameter, the one with the lower variance is better. This lower-variance unbiased estimator is called the more “efficient” estimator.

What’s the best we can do? Sometimes, there is an estimator that achieves the lowest variance possible for an unbiased estimator.

**Def’n: Most Efficient Estimator**

An unbiased estimator with the lowest possible variance among all unbiased estimators is called the most efficient estimator.

**Example: Sample Mean vs. Sample Median**

Assume a sample of $n$ Gaussian r.v.s with standard deviation $\sigma_X$ and mean $\mu$. Both sample mean and sample median are estimators of $\mu$. Which one is more efficient? Note that the variance of the sample median is $\frac{\pi \sigma_X^2}{2n}$, where $f_{X}(x)$ is the pdf of the r.v. $X$.

**Solution:** For Gaussian r.v., the pdf is $f_X(x) = \frac{1}{\sqrt{2\pi \sigma^2}} e^{-(x-\mu)^2/(2\sigma^2)}$. Plugging in $x = \mu$,

$f_X(\mu) = \frac{1}{\sqrt{2\pi \sigma^2}} e^{-(\mu-\mu)^2/(2\sigma^2)} = \frac{1}{\sqrt{2\pi \sigma^2}}$

Thus the variance of the median is

$$\frac{1}{4n[f_X(\mu)]^2} = \frac{2\pi \sigma^2}{4n} = \frac{\pi \sigma^2}{2n} = 1.57\frac{\sigma^2}{n}$$

The variance of the sample mean is $\frac{\sigma^2}{n}$. Thus the variance of the sample median is 57% higher than the variance of the sample mean. Thus the sample mean is the more efficient estimator.

---

### Lecture #28

#### 29 Confidence Interval on Mean

When estimating a distribution parameter, it is often a good idea to report an interval in which you believe the parameter will fall in, with some high probability.

You collect a sample $X_1, \ldots, X_n$, and use it to estimate a parameter of the distribution of $X$, say the population mean $\mu_X$. You calculate the sample mean $\bar{X}$ and it comes out to be some value. There’s a problem with just reporting the value of $\bar{X}$ you got and saying that it is the mean. Sure, it is your best estimate of the mean from your sample. But just reporting the sample mean doesn’t give another person an idea of how precise the estimate is. You could also report the sample standard deviation, but it is a somewhat hard-to-interpret measure of how wide the distribution of your estimate is, particularly for a non-statistician.
A confidence interval is a very useful way to report uncertainty in an estimate. It is easy to interpret and it is quantitative. It conveys a range in which the true parameter is likely to be, and the probability of it being in that range. We call the range the confidence interval (CI), and we denote the probability of it falling in that range as \( (1 - \alpha) \). Equivalently, the probability the population mean doesn’t fall in the CI is \( \alpha \). You want to make \( \alpha \) small so you’re not wrong very often, but you don’t want to make \( \alpha \) so close to zero that your CIs are so wide that they are essentially useless. Typically, we will calculate CIs for \( \alpha \) in the range of 0.001 to 0.1, that is, a 99.9% to 90% confidence interval.

29.1 Two-sided

A two-sided \( 100(1 - \alpha) \) confidence interval for the population mean \( \mu \) looks like this:

\[
P [\text{Lower limit} < \mu < \text{Upper limit}] = 1 - \alpha,
\]

where we set the lower limit so that there is a \( \alpha/2 \) probability that \( \mu < \) Lower limit, and we set the upper limit so that there is a \( \alpha/2 \) probability that \( \mu > \) Upper limit. See Figure 17(a).

29.2 One-sided Bound

Alternatively, we are sometimes interested in having only one limit on the interval, like in Figure 17(b) or (c). For example, we might be measuring power dissipation of an electronic device, or mercury concentration in a river (an example from Walpole). A high number is bad, so essentially, no one will care what the lower limit of the CI is — they will only look at the upper limit and take actions based on that number. In this case a one-sided \( 100(1 - \alpha) \) confidence interval for the population mean \( \mu \) looks like one of the following:

\[
P [\mu < \text{Upper limit}] = 1 - \alpha, \quad \text{or} \quad (23)
\]

\[
P [\text{Lower limit} < \mu] = 1 - \alpha. \quad \text{(24)}
\]

where we set the lower limit so that there is a \( \alpha \) probability that \( \mu < \) Lower limit, or we set the upper limit so that there is a \( \alpha \) probability that \( \mu > \) Upper limit. See Figure 17(b) and (c). The value in (23) is called the "upper bound" and the inequality \( \mu < \) Upper limit is called the upper one-sided confidence interval. The value in (24) is called the "lower bound" and the inequality \( \mu < \) Upper limit is called the lower one-sided confidence interval.

29.3 Known Population Standard Deviation

All of our analysis for the CI assumes that the Central Limit Theorem applies (that is, if the data is not itself Gaussian, then there is a large sample size). For the first (less realistic) case, we assume that we know the population standard deviation \( \sigma \). In this case, the statistic:

\[
Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}}
\]

is standard normal.

Let’s consider the two-sided case. We set \( z_{\alpha/2} \) to be the threshold at which \( P [Z > z_{\alpha/2}] = \alpha/2 \). We can do this from Table A.3, by looking up where the probability is closest to \( 1 - \alpha/2 \). However, this is harder than the following. Use the last row in Table A.4, labeled \( \nu = \infty \). Recall the limit
Figure 17: (a) Two-sided CI and (b,c) one-sided CIs for the population mean calculated from the sample mean $\bar{X}$. We want the probability that the CI contains the population mean to be $1 - \alpha$. The two-sided CI allows $\alpha/2$ probability of the mean falling on either side of the CI. The one-sided CI allows $\alpha$ probability of being either (b) above or (c) below the limit of the CI.
of the t-distribution as the degrees-of-freedom goes to infinity, is the standard normal. Thus the Table A.4 $\infty$ row gives us the critical values of the standard normal distribution. Write:

$$P\left[-z_{\alpha/2} < Z < z_{\alpha/2}\right] = 1 - \alpha$$

Exchanging $Z$ for $\bar{X} - \mu \sigma/\sqrt{n}$, we have:

$$P\left[-z_{\alpha/2} < \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} < z_{\alpha/2}\right] = 1 - \alpha$$

Manipulating the interval (multiplying and adding to all “sides” of the inequality):

$$P\left[-z_{\alpha/2} < \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} < z_{\alpha/2}\right] = 1 - \alpha$$

Note when you multiply all sides by $-1$ (as I did between lines 2 and 3 of the above derivation) you also need to exchange $<$ and $>$ signs. However, the next step (line 4 of the above derivation) is to flip the sides of the inequality so that it still reads in the standard “lowest thing $<$ middle thing $<$ highest thing” format we’re used to seeing.

Thus the lower and upper limits are $\bar{X} \pm z_{\alpha/2} \sigma/\sqrt{n}$.

The one-sided CIs use the $z_{\alpha}$ value instead of $z_{\alpha/2}$ value, and use only one of the limits, but are otherwise the same. See Table 3.

### 29.4 Unknown Population Standard Deviation

In the case where we don’t know $\sigma$, we can’t use it in the formula for the CI. However, we can still find a CI using $S^2$ and the $T$ statistic. In this case, the statistic:

$$T = \frac{\bar{X} - \mu}{S/\sqrt{n}}$$

has the student-t distribution with $\nu = n - 1$.

Let’s consider the two-sided case. We set $t_{\alpha/2}$ to be the threshold at which $P \left[ T > t_{\alpha/2} \right] = \alpha/2$. This requires Table A.4, looking in the row for the given $\nu = n - 1$. Write:

$$P \left[ -t_{\alpha/2} < T < t_{\alpha/2} \right] = 1 - \alpha$$

Exchanging $T$ for $\frac{\bar{X} - \mu}{S/\sqrt{n}}$ we have:

$$P \left[ -t_{\alpha/2} < \frac{\bar{X} - \mu}{S/\sqrt{n}} < t_{\alpha/2} \right] = 1 - \alpha$$
Table 3: 100(1 − α)% confidence intervals for the population mean µ, as a function of the sample mean \( \bar{X} \), sample size \( n \), and critical values from the standard normal distribution (\( z_\alpha \) or \( z_{\alpha/2} \)) or the t-distribution (\( t_\alpha \) or \( t_{\alpha/2} \)).

<table>
<thead>
<tr>
<th>Interval Type</th>
<th>( \sigma ) Known</th>
<th>( \sigma ) Unknown</th>
</tr>
</thead>
<tbody>
<tr>
<td>Two-sided</td>
<td>( \bar{X} - z_{\alpha/2} \frac{\sigma}{\sqrt{n}} &lt; \mu &lt; \bar{X} + z_{\alpha/2} \frac{\sigma}{\sqrt{n}} )</td>
<td>( \bar{X} - t_{\alpha/2} \frac{S}{\sqrt{n}} &lt; \mu &lt; \bar{X} + t_{\alpha/2} \frac{S}{\sqrt{n}} )</td>
</tr>
<tr>
<td>Upper One-sided</td>
<td>( \mu &lt; \bar{X} + z_{\alpha} \frac{\sigma}{\sqrt{n}} )</td>
<td>( \mu &lt; \bar{X} + t_\alpha \frac{S}{\sqrt{n}} )</td>
</tr>
<tr>
<td>Lower One-sided</td>
<td>( \mu &gt; \bar{X} - z_{\alpha} \frac{\sigma}{\sqrt{n}} )</td>
<td>( \mu &gt; \bar{X} - t_\alpha \frac{S}{\sqrt{n}} )</td>
</tr>
</tbody>
</table>

Manipulating the interval (multiplying and adding to all “sides” of the inequality):

\[
P \left[ -t_{\alpha/2} < \frac{\bar{X} - \mu}{s/\sqrt{n}} < t_{\alpha/2} \right] = 1 - \alpha
\]

\[
P \left[ -t_{\alpha/2} \frac{s}{\sqrt{n}} < \bar{X} - \mu < t_{\alpha/2} \frac{s}{\sqrt{n}} \right] = 1 - \alpha
\]

\[
P \left[ t_{\alpha/2} \frac{s}{\sqrt{n}} > \mu - \bar{X} > -t_{\alpha/2} \frac{s}{\sqrt{n}} \right] = 1 - \alpha
\]

\[
P \left[ -t_{\alpha/2} \frac{s}{\sqrt{n}} < \mu - \bar{X} < t_{\alpha/2} \frac{s}{\sqrt{n}} \right] = 1 - \alpha
\]

\[
P \left[ \bar{X} - t_{\alpha/2} \frac{s}{\sqrt{n}} < \mu < \bar{X} + t_{\alpha/2} \frac{s}{\sqrt{n}} \right] = 1 - \alpha
\]

Thus the lower and upper limits are \( \bar{X} \pm t_{\alpha/2} \frac{s}{\sqrt{n}} \).

The one-sided CIs use the \( t_\alpha \) value instead of \( t_{\alpha/2} \) value, and use only one of the limits, but are otherwise the same. See Table 3.

### 29.5 Procedure to Determine CI

1. Determine if the population mean is assumed to be known, or to be unknown.

2. Determine if a two-sided or one-sided CI is needed. If a one-sided, determine if an upper bound or a lower bound is required.

3. Find \( \bar{X} \), \( n \), and \( S^2 \) (if needed) from the problem statement or data.

4. Determine the level of confidence parameter \( \alpha \). Recall that the “confidence” is \((1 - \alpha) \times 100\%\), so a 99% CI means \( \alpha = 0.01 \).

5. Look up the appropriate critical value for \( z \) or \( t \). For two-sided and known \( \sigma \), find \( z_{\alpha/2} \); for two-sided and unknown \( \sigma \), find \( t_{\alpha/2} \); for one-sided and known \( \sigma \), find \( z_\alpha \); for one-sided and unknown \( \sigma \), find \( t_\alpha \).
6. Look up the appropriate formula in Table 3 and plug in.

**Example: Clean Room Standards**

An ISO-3 clean room can have at most 35 particles (of size $\geq 0.5$ micron) per cubic meter. You make 30 measurements in the clean room of the particle density, and find $\bar{X} = 30.4$.

1. Assume that $\sigma = 12 /m^3$. Find the two-sided 95% confidence interval.

2. Assume that $\sigma = 12 /m^3$. Find the upper one-sided 95% confidence interval.

3. Now assume that $\sigma$ is not known but the sample standard deviation is 12. Find the upper one-sided 95% confidence interval.

**Solution:** Looking in Table A.4, $t_{0.05} = 1.699$ and $t_{0.025} = 2.045$ for $\nu = 29$. Looking in the same table at the $\nu = \infty$ row, $z_{0.05} = 1.645$ and $z_{0.025} = 1.960$.

1. Recall the limits are $\bar{X} \pm z_{\alpha/2} \frac{\sigma}{\sqrt{n}}$. We can calculate $z_{\alpha/2} \frac{\sigma}{\sqrt{n}} = 1.960(12/\sqrt{30}) = 4.29$. So with 95% confidence, the population mean is between $30.4 - 4.29 = 26.11$ and $30.4 + 4.29 = 34.69$.

2. The upper limit is $\bar{X} + z_{\alpha} \frac{\sigma}{\sqrt{n}}$. Calculating, $z_{\alpha} \frac{\sigma}{\sqrt{n}} = 1.645(12/\sqrt{30}) = 3.60$. The CI is thus $\mu < 30.4 + 3.60 = 34.0$.

3. The upper limit is $\bar{X} + t_{\alpha} \frac{S}{\sqrt{n}}$. Calculating, $t_{\alpha} \frac{S}{\sqrt{n}} = 1.699(12/\sqrt{30}) = 3.72$. The CI is thus $\mu < 30.4 + 3.72 = 34.12$.

---

**Lecture #29**

### 30 Confidence Interval on Variance

This material is in the Walpole book Section 9.12. Similar to the confidence interval (CI) on the population mean $\mu$, in this section, we develop a CI for the population variance or standard deviation. In contrast, though, this CI requires that the data $X_1, \ldots, X_n$ come from a Gaussian distribution, regardless of the sample size. The confidence interval on $\mu$, in comparison, would allow non-Gaussian data as long as $n \geq 30$.

The same procedure is followed for finding a $100(1-\alpha)$% confidence interval on the variance as was followed for the mean:

1. Determine if a two-sided or one-sided CI is needed. If a one-sided, determine if an upper bound or a lower bound is required.

2. Find $n$ and $S^2$ from the problem statement or data.

3. Determine the level of confidence parameter $\alpha$. Recall that the “confidence” is $(1-\alpha) \times 100\%$, so a 99% CI means $\alpha = 0.01$.

4. Look up the appropriate critical values from the chi-squared distribution. For a two-sided CI find $\chi^2_{\alpha/2}$ and $\chi^2_{1-\alpha/2}$; for a one-sided CI, find $\chi^2_{\alpha}$ or $\chi^2_{1-\alpha}$ (for a lower bound or an upper bound, respectively).
5. Look up the appropriate formula in Table 4 and plug in.

Now that I’ve given the procedure, let’s derive the formula for the two-sided test show why it takes the form it does.

\[
P \left[ \frac{\chi^2_{1-\alpha/2} < X^2 < \chi^2_{\alpha/2}}{\chi^2_{1-\alpha/2}} \right] = 1 - \alpha
\]

\[
P \left[ \frac{\chi^2_{1-\alpha/2} < \frac{(n-1)s^2}{\sigma^2} < \chi^2_{\alpha/2}}{(n-1)S^2} \right] = 1 - \alpha
\]

\[
P \left[ \frac{\chi^2_{1-\alpha/2} > \frac{\sigma^2}{(n-1)S^2} > \frac{1}{\chi^2_{\alpha/2}}}{\chi^2_{1-\alpha/2}} \right] = 1 - \alpha
\]

\[
P \left[ \frac{(n-1)S^2}{\chi^2_{1-\alpha/2}} > \sigma^2 > \frac{(n-1)S^2}{\chi^2_{\alpha/2}} \right] = 1 - \alpha
\]

\[
P \left[ \frac{(n-1)S^2}{\chi^2_{\alpha/2}} < \sigma^2 < \frac{(n-1)S^2}{\chi^2_{1-\alpha/2}} \right] = 1 - \alpha
\]

Note that when you take the inverse of each side (between lines 2 and 3 of the above derivation), you must change the order of the inequality signs. However, the last step (line 5 of the above derivation) is to flip the sides of the inequality so that it still reads in the standard “lowest thing < middle thing < highest thing” format we’re used to seeing. Note that this leads to the counterintuitive fact that \( \chi^2_{\alpha/2} \) is in the fraction on the LHS of the inequality and \( \chi^2_{1-\alpha/2} \) is on the RHS.

In summary, the lower limit on the CI for the variance is

\[
\frac{(n-1)S^2}{\chi^2_{1-\alpha/2}}
\]

and the upper limit is

\[
\frac{(n-1)S^2}{\chi^2_{\alpha/2}}.
\]

Finally, if we want a CI on the standard deviation, we simply take the square root of all sides of the inequality:

\[
P \left[ \sqrt{\frac{(n-1)S^2}{\chi^2_{1-\alpha/2}}} < \sigma < \sqrt{\frac{(n-1)S^2}{\chi^2_{\alpha/2}}} \right] = 1 - \alpha.
\]

The one-sided CIs are shown in Table 4.

---

### Lecture #30

#### 31 Hypothesis Testing

In estimation, we use sample data to come up with a parameter of a distribution of the data, for example, the mean or standard deviation. We estimate a real-valued parameter, and we know we won’t be exactly right (even though we hope our estimate is close to the true value).

In contrast, in hypothesis testing, we are answering a yes or no question. We have two ideas about what the distribution might be, but we don’t know which one is correct. In hypothesis testing, we choose one or the other, based on our sample data.
<table>
<thead>
<tr>
<th>Probability of $\sigma^2$ in Interval $= 1 - \alpha$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Two-sided</td>
</tr>
<tr>
<td>$\frac{(n-1)S^2}{\chi^2_{\alpha/2}} &lt; \sigma^2 &lt; \frac{(n-1)S^2}{\chi^2_{1-\alpha/2}}$</td>
</tr>
<tr>
<td>Upper One-sided</td>
</tr>
<tr>
<td>$\sigma^2 &lt; \frac{(n-1)S^2}{\chi^2_{1-\alpha}}$</td>
</tr>
<tr>
<td>Lower One-sided</td>
</tr>
<tr>
<td>$\frac{(n-1)S^2}{\chi^2_{\alpha}} &lt; \sigma^2$</td>
</tr>
</tbody>
</table>

Table 4: 100(1 − $\alpha$)% confidence intervals for the population variance $\sigma^2$, as a function of the sample variance $S^2$, sample size $n$, and critical values from the chi-squared distribution.

Typically, one hypothesis represents the “status quo”, that is, what people believe right now. Or, it is what we would believe if we had no data. This is called the null hypothesis. We will always denote the null hypothesis $H_0$.

The alternative to the null hypothesis is called the alternate hypothesis, which is denoted $H_1$. It represents what we would be willing to decide if the sample data is convincing enough. If the sample is something that we’d reasonably expect to occur given the null hypothesis, we will decide $H_0$. Scientists call this “fail to reject $H_0$” because its not that we truly believe $H_0$ is true, we just don’t have enough evidence to reject it.

An analogy is used to a criminal trial, where a defendant is found not guilty or guilty. One should presume the defendant innocent until proven guilty beyond a reasonable doubt. In terms of hypothesis testing, we’d say that the null hypothesis $H_0$ is that the person is innocent. Only if we have sufficient data showing his guilt should we decide $H_1$, the alternate hypothesis that he is guilty. This analogy is also good because it is clear that acquittal (decide $H_0$, a.k.a. failure to reject $H_0$) doesn’t prove that the person is innocent ($H_0$ is true), it simply says that there was not enough evidence presented for a guilty verdict.

The criminal trial analogy is bad because lawyers, judges, and juries are notoriously bad at statistics, and often justify their decisions with invalid/incorrect application of probability

Certainly, hypothesis testing is the basis for science and medicine. Scientific knowledge is gained by making hypotheses, and testing them with experimental data. Engineers use hypothesis testing as well – we make many systems where data is collected and decisions are made (and perhaps actions are taken as a result) automatically from the data.

- **Anomaly or fault detection.** Automatic detection is required to detect changes in computers, networks, or engineered systems that would indicate that something is wrong, for example, that a virus or an denial-of-service attack is spreading. Or, for example, detecting fraud in credit card transactions, or detecting abnormally high temperatures in a system component. If one is testing continually to look for a change, this is also called “change detection”. In general, engineered detection systems (such as airplane radar systems) make measurements and automatically detect $H_1$ (for example, the presence of an airplane in a region of airspace) and then take an action (start tracking the airplane).

---

Table 5: Correct decisions vs. name of two types of incorrect decisions.

<table>
<thead>
<tr>
<th>Decision</th>
<th>True $H_0$</th>
<th>True $H_1$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Reject $H_0$</td>
<td>False alarm / Type I error</td>
<td>Correct Decision</td>
</tr>
<tr>
<td>Do not reject $H_0$</td>
<td>Correct Decision</td>
<td>Type II error / Missed detection</td>
</tr>
</tbody>
</table>

- **Communications receivers.** Receivers must decide which symbol was sent by a transmitter. Symbol detection is another word for hypothesis testing. Optimal detection is important in reducing error rates. In addition, receivers must also decide that a signal is being sent (vs. no signal currently being transmitted), and can do so by measuring received power.

Many of our engineered sensing systems are designed to “raise an alarm”, that is, inform a person when the data is significantly different from what would be expected under normal conditions. In this case, the null hypothesis is that the system is operating normally (with normal parameters). When data indicates that the null hypothesis is not true, the system raises the alarm. There are two types of errors we can make with a hypothesis test.

1. The null hypothesis $H_0$ is actually true, but we decide the alternate hypothesis $H_1$ is true. This is called a “false alarm”, because we have effectively raised an alarm when it did not need to be raised. This is also called a “type I error” in science.

2. The alternate hypothesis $H_0$ is actually true, but we don’t reject $H_0$. That is, we should have raised the alarm and didn’t. Engineers call this a “missed detection”, as we should have detected the alarm condition, but didn’t. Scientists call this a “type II error”.

There are often different costs associated with the two different types of errors. One may be a “worse” mistake to make. For example, in my research lab, we are developing sensors which use RF to detect breathing under a pile of rubble. Our $H_0$ is that there is no breathing present, and $H_1$ is that there is breathing present. An emergency responder would deploy the system at the site of a collapsed building to determine if there was anyone alive underneath. A false alarm or type I error would be bad because the responders would (presumably) work to clear the rubble in hopes of finding a breathing person when no living person was actually present. A missed detection or type II error would tell the responders not to clear the rubble, even though a breathing person actually was underneath.

We’d love to have no false alarms and no missed detections. But in real world noisy systems, this is often not possible. There is a tradeoff; if we accept more false alarms, we’d probably have fewer missed detections, or vice versa. Generally, engineers use $\alpha$ to denote the acceptable probability of a false alarm. We tend to set the maximum false alarm probability $\alpha$ we’re willing to accept.

### 32 Hypothesis Tests for Change in Mean

We often want to test whether, given our data, the current assumption about the population mean is true or false. To do this with a hypothesis test, we consider the following hypotheses:

\[
H_0 : \mu = \mu_0, \\
H_1 : \mu \neq \mu_0,
\]
where $\mu_0$ is called the “mean given $H_0$ is true”, and $\mu$ is called the “true population mean”. We can read the $H_0$ line as “The true population mean is equal to the value $\mu_0$”, and you’d be given a value $\mu_0$.

**Example: Packet Traffic**

Under normal conditions, traffic on an internet backbone link has mean $10^6$ packets per second, with a standard deviation of $0.4 \times 10^6$ packets per second. When traffic is extremely low or high (with false alarm probability $0.001$), it should be brought to the attention of a network administrator. Traffic is measured during 30 randomly-selected second-long periods. The sample mean is $\bar{X} = 0.71 \times 10^6$ packets per second. Should the administrator be alerted?

**Solution:** This is a two-sided test with $\mu_0 = 10^6$ packets per second with a known standard deviation. If $H_0$ is true, the following $Z$ statistic has the standard normal distribution:

$$Z = \frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}} = \frac{0.71 - 1.0}{0.4/\sqrt{30}} = -3.97,$$

where units of the measurements are in Mpackets/sec. To create a hypothesis test with a 0.001 probability of a false alarm, we would set the limits on $Z$ so that the tails have 0.0005 probability, that is,

$$P[-z_{0.0005} < Z < z_{0.0005}] = 1 - 0.001$$

Looking up in Table A.4 (on the $\nu = \infty$ line), we have $z_{0.0005} = 3.290$. Since our $Z = -3.97$ is outside of the range $-3.290$ to $3.290$, we have enough evidence (at an $\alpha = 0.001$ probability of false alarm) to reject $H_0$. We conclude that the population mean is not $10^6$.

**Example: Router Delay**

Under normal conditions, a router introduces (on average) a $1.5 \mu s$ delay, with a standard deviation of $1.0 \mu s$. In congested conditions, the delay is higher. An alarm should be raised if the router is congested so that an operator knows there is a problem. A random sample of 50 packet delays are measured, and $\bar{X}$ is found to be $1.8 \mu s$. For an acceptable false alarm probability $\alpha = 0.01$, should the alarm be raised?

**Solution:** Here, the hypotheses are:

$$H_0 : \quad \mu = 1.5 \mu s,$$

$$H_1 : \quad \mu > 1.5 \mu s,$$

We would only decide $H_0$ if the sample mean is very high compared to $1.5 \mu s$. We are given $\sigma$. To have a false alarm probability $\alpha = 0.01$, we would now set the upper threshold as:

$$P[Z > z_{0.01}] = 1 - 0.01.$$  

Looking at Table A.4, $z_{0.01} = 2.326$. What is $Z$? Leaving out the units of $\mu s$,

$$Z = \frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}} = \frac{1.8 - 1.5}{1.0/\sqrt{50}} = 2.12.$$  

Since $Z = 2.12$ is less than 2.326, we decide not to reject $H_0$.

Let’s enumerate the possible hypothesis tests on a population mean. For all such tests, $H_0$ is that $\mu = \mu_0$. The alternate hypothesis is one of:
1. Two-sided: \( H_1 : \mu \neq \mu_0 \). That is, the mean is different from \( \mu_0 \), but being far from the mean in either direction should “trigger the alarm”.

2. One-sided greater than: \( H_1 : \mu > \mu_0 \). That is, the mean is larger than \( \mu_0 \). Only a significant increase in the mean should trigger the alarm.

3. One-sided less than: \( H_1 : \mu < \mu_0 \). That is, the mean is less than \( \mu_0 \). Only a significant decrease in the mean should trigger the alarm.

In case 1, we will determine a low and a high threshold, each tail having probability \( \alpha/2 \). In cases 2 and 3, the single threshold will be set so the tail has probability \( \alpha \).

Next, we will either assume 1) a known standard deviation \( \sigma \); or 2) \( \sigma \) is unknown. When \( \sigma \) is known, then we will consider the \( Z \) value,

\[
Z = \frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}},
\]

and our threshold(s) will be determined from the standard normal distribution. When the \( \sigma \) is not known, we will compute the sample standard deviation \( s \) and then consider the \( T \) value,

\[
T = \frac{\bar{X} - \mu_0}{s/\sqrt{n}},
\]

and our threshold(s) will be determined from the student \( t \) distribution with \( \nu = n - 1 \) degrees-of-freedom.

For example, assume \( \sigma \) is unknown and we need a two-sided test with false alarm probability \( \alpha \). We’ll look up \( t_{\alpha/2} \) in Table A.4. We’ll compute \( T \) and see it is in the interval \(-t_{\alpha/2} < T < t_{\alpha/2}\). In other words,

\[
-t_{\alpha/2} < \frac{\bar{X} - \mu_0}{s/\sqrt{n}} < t_{\alpha/2}
\]

Note that this could also be solved for \( \bar{X} \):

\[
\mu_0 - t_{\alpha/2} \frac{s}{\sqrt{n}} < \bar{X} < \mu_0 + t_{\alpha/2} \frac{s}{\sqrt{n}}
\]

This version simply says, compute the sample mean, and if it is in between the two values, \( \mu_0 \pm t_{\alpha/2} \frac{s}{\sqrt{n}} \), then decide \( H_0 \).

Note the similarity with the confidence interval. However, the difference here is that we are finding a range for \( \bar{X} \) (in which we will decide \( H_0 \)), whereas a confidence interval is a range on \( \mu \).

Similar to CIs, though, we can present a table with all of the thresholds for possible hypothesis tests on the mean. We do this in Table 32.

Lecture #31

33 Tests for Change in Variance

Another useful test is to determine, given our data, whether or not the current assumption about the population variance is true or false.
Table 6: Ranges for the sample mean $\bar{X}$ which lead to rejection of $H_0$, as a function of sample size $n$, and critical values from the standard normal distribution ($z_\alpha$ or $z_{\alpha/2}$) or the t-distribution ($t_\alpha$ or $t_{\alpha/2}$).

For example, the average resistance of a box of resistors may be the correct value, but the standard deviation may have sharply increased. Since high variance would reduce the reliability of circuits which use them, we would want to detect that we have a “bad” batch of resistors and stop using them. In such a case, high variability is a bad thing.

As another example, regulators may wish to watch the variance (over a period of time) of a security (a stock, for example). Or, a credit card company may want to watch the variance of the prices of transactions charged to a card. A high or low variance, compared to past history, may indicate some event has happened, or that there is some fraudulent activity taking place. A credit card being used for many transactions, all near the same price, would lower the variance, and might indicate someone breaking a large transaction (larger than allowed) into smaller parts so not to be detected.

As another example, communications receivers continuously monitor the short-term variance of a received signal. When the variance dramatically increases, the receiver can assume that an interfering signal has suddenly appeared on the channel. It can attempt an interference-cancellation technique if interference is suspected, with the expense of reduced performance if it is wrong.

In such cases, we would use a null hypothesis of:

$$H_0 : \sigma^2 = \sigma_0^2,$$

where $\sigma_0^2$ is called the “variance given $H_0$ is true”, and $\sigma^2$ is called the “true population variance”. We can read the $H_0$ line as “The true population mean is equal to the value $\sigma_0^2$”, and you’d be given a value $\sigma_0^2$. We would use one of three possible alternate hypotheses:

1. Two-sided: $H_1 : \sigma^2 \neq \sigma_0^2$. That is, the variance is different from $\sigma_0^2$, but being far from the variance in either direction should “trigger the alarm”.

2. One-sided greater than: $H_1 : \sigma^2 > \sigma_0^2$. That is, the variance is larger than $\sigma_0^2$. Only a significant increase in the variance should trigger the alarm.

3. One-sided less than: $H_1 : \sigma^2 < \sigma_0^2$. That is, the variance is less than $\sigma_0^2$. Only a significant decrease in the variance should trigger the alarm.
When to Reject \( H_0 \)

<table>
<thead>
<tr>
<th>When to Reject ( H_0 )</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Two-sided</strong></td>
<td>( S^2 &lt; \chi^2_{1-\alpha/2} \left( \frac{\sigma^2_0}{n-1} \right) ), or ( S^2 &gt; \chi^2_{\alpha/2} \left( \frac{\sigma^2_0}{n-1} \right) )</td>
</tr>
<tr>
<td><strong>Upper One-sided</strong></td>
<td>( S^2 &gt; \chi^2_{\alpha} \left( \frac{\sigma^2_0}{n-1} \right) )</td>
</tr>
<tr>
<td><strong>Lower One-sided</strong></td>
<td>( S^2 &lt; \chi^2_{1-\alpha} \left( \frac{\sigma^2_0}{n-1} \right) )</td>
</tr>
</tbody>
</table>

Table 7: Ranges for the sample variance \( S^2 \) which lead to rejection of \( H_0 \), as a function of sample size \( n \), and critical values from the chi-squared distribution with \( \nu = n - 1 \) degrees of freedom.

We only study hypotheses tests on the variance when the data, \( X_1, \ldots, X_n \), are approximately Gaussian. In that case, the statistic,

\[
\chi^2 = \frac{(n-1)S^2}{\sigma^2_0} = \sum_{i=1}^{n} \frac{(X_i - \bar{X})^2}{\sigma^2_0}
\]

has a chi-squared distribution with parameter \( \nu = n - 1 \). As a result, we can use Table A.5 to find the critical values of the chi-squared statistic. If the true variance \( \sigma^2 \) is much higher than \( \sigma^2_0 \), then \( \chi^2 \) will be a very high number. Similarly, if the true variance \( \sigma^2 \) is much lower than \( \sigma^2_0 \), then \( \chi^2 \) will be a very low and close to zero.

A two-sided test on the variance would be designed as

\[
P \left[ \chi^2 < \chi^2_{1-\alpha/2} < \chi^2_{\alpha/2} \right] = 1 - \alpha,
\]

where the threshold value \( \chi^2_{1-\alpha/2} \) is the threshold with probability \( 1 - \alpha/2 \) to the right of the threshold; and \( \chi^2_{\alpha/2} \) is the threshold with probability \( \alpha/2 \) to the right of the threshold. Just as for tests on the mean, \( \alpha \) is the chosen probability of false alarm.

**Note that since the chi-squared distribution is not symmetric, the lower threshold is NOT \(-\chi^2_{\alpha/2}\).** We actually have to look up two thresholds. The lower threshold has \( \alpha/2 \) probability to the left of it, the same as saying \( 1 - \frac{\alpha}{2} \) to the right of it. Since Table A.5 is written in terms of probabilities to the right of the threshold, we write the lower threshold as \( \chi^2_{1-\alpha/2} \).

If the calculated value of \( \chi^2 = \frac{(n-1)S^2}{\sigma^2_0} \) is inside the range in (25), we’d decide \( H_0 \) (that is, fail to reject \( H_0 \)). If \( \chi^2 \) falls outside of this range, we’d decide \( H_1 \) (reject \( H_0 \)). We can rewrite the inequality in (25) using the formula for \( \chi^2 \):

\[
P \left[ \chi^2_{1-\alpha} \left( \frac{\sigma^2_0}{n-1} \right) < S^2 < \chi^2_{\alpha/2} \left( \frac{\sigma^2_0}{n-1} \right) \right] = 1 - \alpha
\]

\[
P \left[ \chi^2_{1-\alpha} \left( \frac{\sigma^2_0}{n-1} \right) < S^2 \right] = 1 - \alpha
\]

The one-sided tests have only one threshold, and Table 33 lists the three possible tests on variance and their thresholds on \( S^2 \).
We should end with another reminder that the $\chi^2$ tests described above will not have an accurate $\alpha$ probability of false alarm if the data itself is not Gaussian, because the $\chi^2_\alpha$ thresholds are only accurate when the data $X_1, \ldots, X_n$ are in fact Gaussian. Walpole says “The analyst should approach the use of this particular chi-squared test with caution.”

**Example: Detecting Interference**

Received signal values $X_1, \ldots, X_n$ are collected for $n = 16$ during reception of a packet. We can reliably assume that the sample values are zero mean and Gaussian, and based on past received signals from the same transmitter, we’d assume that the standard deviation is 1.1 V.

1. If an interferer starts operating on this band, we expect the standard deviation to increase. For $\alpha = 0.01$, design a hypothesis test. Now, $S^2 = 3 \text{ V}^2$ is recorded – should $H_0$ or $H_1$ be decided?

2. Abnormally low variance may indicate the signal is in a fade. For a two-sided test with the same $\alpha$, if $S^2 = 0.49 \text{ V}^2$ is recorded – should $H_0$ or $H_1$ be decided?

**Solution:**

1. We want an upper one-sided test, so $H_1: \sigma > 1 \text{ V}$. We should reject $H_0$ if $S^2 = 3 \text{ V}^2$ is greater than:

$$\chi^2_{0.01} \left( \frac{\sigma^2}{n-1} \right) = 30.578 \left( \frac{1.1^2}{15} \right) = 2.47.$$ 

Thus we should reject $H_0$ (and decide $H_1$).

2. The two thresholds are:

$$\chi^2_{1-\alpha/2} \left( \frac{\sigma^2}{n-1} \right) = 4.601 \left( \frac{1.1^2}{15} \right) = 0.371$$

and

$$\chi^2_{\alpha/2} \left( \frac{\sigma^2}{n-1} \right) = 32.801 \left( \frac{1.1^2}{15} \right) = 2.65$$

Since 0.49 is between 0.371 and 2.65, we decide $H_0$.

---

**Lecture #32**

**34 Hypothesis Tests on Difference in Mean**

Consider the case where we have two populations (call them population 1 and population 2), and we want to know if the mean of population 1, $\mu_1$ is different from the mean of population 2, $\mu_2$. For example, we may believe that our new drug will reduce a person’s cholesterol – in this case, we want to know the mean change in cholesterol level of people who take the drug, vs. people who take a placebo. As another example, we may have two different ways to manufacture our product, an original way and a less expensive way, and we want to make sure that there is no difference in the specs for our product if we use the less expensive way.
In any of these cases, we will take a sample from population 1, and a sample from population 2. We want to know, based on our sample data, is the mean of population 1 different from the mean of population 2? In this case, we will compute from our two samples (or be given in a problem statement) the following:

- $n_1$, the size of the sample of population 1;
- $n_2$, the size of the sample of population 1;
- $\bar{X}_1$, the sample average from the first sample;
- $\bar{X}_2$, the sample average from the second sample;
- $S_1$, the sample standard deviation from the first sample, or $\sigma_1$ if happens to be known beforehand; and
- $S_2$, the sample standard deviation from the second sample, or $\sigma_2$ if happens to be known beforehand.

For the following, we assume that either: (1) both populations have data that is Gaussian; or (2) that both $n_1$ and $n_2$ are large enough for the central limit theorem to apply.

Our null hypothesis is that the difference between the two means is equal to some value:

$$H_0 : \mu_1 - \mu_2 = d_0,$$

where $d_0$ is the “status quo” assumption about the difference between the means. For example, if the standard assumption is that the means are the same, then $d_0 = 0$.

There are two statistics we might use. If we know the population standard deviations, then

$$Z = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}}$$

has the standard normal distribution. Again, $\mu_1 - \mu_2$ is the difference in the sample means under $H_0$, so we can plug in $\mu_1 - \mu_2 = d_0$. If we don’t know the population standard deviations (and we don’t know that the standard deviations are equal), then we use the following $T$ statistic:

$$T' = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}}.$$

This $T'$ statistic is unique because its number of “degrees of freedom” is a real number given by the following formula:

$$\nu = \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\frac{1}{n_1-1} \left(\frac{s_1^2}{n_1}\right)^2 + \frac{1}{n_2-1} \left(\frac{s_2^2}{n_2}\right)^2}$$

Although the formula is complicated, one simply calculates $\nu$, rounds it to the nearest integer, and uses it to look it up the required critical value in Table A.4. In a sense, equation (27) quantifies the idea that, for example, you can’t have a really really large sample of population 1, and a really small sample for population 2, and expect to see the degrees-of-freedom simply be the sum of the two degrees-of-freedom. That is, if you define $\nu_1 = n_1 - 1$ and $\nu_2 = n_2 - 1$ as the degrees-of-freedom from each sample, we can’t simply expect the overall $\nu = \nu_1 + \nu_2$. Instead, a better judgment of how
When to Reject $H_0$, For $\sigma$ Known

<table>
<thead>
<tr>
<th></th>
<th>When to Reject $H_0$, For $\sigma$ Known</th>
</tr>
</thead>
<tbody>
<tr>
<td>Two-sided</td>
<td>$\bar{X}_1 - \bar{X}<em>2 &lt; d_0 - z</em>{\alpha/2}\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$, or $\bar{X}_1 - \bar{X}<em>2 &gt; d_0 + z</em>{\alpha/2}\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$</td>
</tr>
<tr>
<td>Upper One-sided</td>
<td>$\bar{X}_1 - \bar{X}<em>2 &gt; d_0 + z</em>{\alpha}\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$</td>
</tr>
<tr>
<td>Lower One-sided</td>
<td>$\bar{X}_1 - \bar{X}<em>2 &lt; d_0 - z</em>{\alpha}\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>When to Reject $H_0$, For $\sigma$ Unknown</th>
</tr>
</thead>
<tbody>
<tr>
<td>Two-sided</td>
<td>$\bar{X}_1 - \bar{X}<em>2 &lt; d_0 - t</em>{\alpha/2}\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}$, or $\bar{X}_1 - \bar{X}<em>2 &gt; d_0 + t</em>{\alpha/2}\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}$</td>
</tr>
<tr>
<td>Upper One-sided</td>
<td>$\bar{X}_1 - \bar{X}<em>2 &gt; d_0 + t</em>{\alpha}\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}$</td>
</tr>
<tr>
<td>Lower One-sided</td>
<td>$\bar{X}_1 - \bar{X}<em>2 &lt; d_0 - t</em>{\alpha}\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}$</td>
</tr>
</tbody>
</table>

Table 8: Ranges for the sample mean $\bar{X}_1 - \bar{X}_2$ which lead to rejection of $H_0$, as a function of sample sizes $n_1$ and $n_2$, and critical values from the standard normal distribution ($z_{\alpha}$ or $z_{\alpha/2}$) or the t-distribution ($t_{\alpha}$ or $t_{\alpha/2}$).

large our data sets are is a parallel combination of $\nu_1$ and $\nu_2$. Equation (27) is more complicated than a simple parallel combination, but this argument should provide a little intuition into why the degrees-of-freedom equation is as it is.

To start, let’s consider the two-sided alternate hypothesis,

$$H_1: \mu_1 - \mu_2 \neq d_0,$$

in the case where the standard deviation is known. The following interval would provide a probability of false alarm of $\alpha$ if $H_0$ is true:

$$P \left[ -z_{\alpha/2} < Z < z_{\alpha/2} \right] = 1 - \alpha$$

where $Z$ is given in (26). Solving for the difference in the sample means, the interval in which we’d decide $H_0$ is

$$d_0 - z_{\alpha/2}\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} < \bar{X}_1 - \bar{X}_2 < d_0 + z_{\alpha/2}\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

If $\bar{X}_1 - \bar{X}_2$ was lower than the lower limit, or higher than the higher limit, we’d decide $H_1$.

One-sided tests, and tests using the $T$ statistic, are derived in a similar manner.
Example: Drug Testing
A drug is being tested to see if it reduces cholesterol. We decide beforehand that we want a two-sided test with $\alpha = 0.05$, and that we will enroll 100 patients into a study. Randomly, and without them knowing (a double-blind study), half of them will be given the drug, and the other half will be given a placebo. From other studies, we know that $\sigma_1 = \sigma_2 = 60$ mg/dL. Our null hypothesis is that the means will be equal, $\mu_1 = \mu_2$. The alternate hypothesis is that $\mu_1 \neq \mu_2$. How far apart do the sample means need to be in order for the study to reject $H_0$?

**Solution:** For a two-sided test, the thresholds are $d_0 \pm z_{\alpha/2} \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$. Here $d_0 = 0$ (the means are assumed the same under $H_0$). Also, for $\alpha = 0.05$, $z_{\alpha/2} = 1.96$. Next, $n_1 = n + 2 = 50$. Thus the thresholds are plus or minus

$$1.96 \sqrt{\frac{60^2}{50} + \frac{60^2}{50}} = 23.52$$

Thus if the sample means are more than 23.52 mg/dL different, then we will decide to reject $H_0$. In other words, the drug will only be judged to be useful if the sample who received the drug had a mean 23.52 less than the other sample mean.

---

Lecture #33

35 What is a p-value?

One of the most difficult but pervasive concepts in statistics is a number reported with most hypothesis tests called the “p-value”. Note this $p$ is NOT the same as the $p$ we use in binomial random variables for the probability of success. This short discussion is meant to provide you with an understanding of the statistical $p$-value in hypothesis tests.

In our hypothesis tests, we decide ahead of time:

1. What we want to test (null vs. alternate hypotheses);
2. What statistic we will use to conduct the test (e.g., sample mean or sample variance);
3. What false alarm (or “type I error”) probability $\alpha$ we are willing to accept — how often we’re willing to deal with a false alarm, in other words, how often we are willing to decide $H_1$ when $H_0$ is really true; and
4. Whether we want a two-sided, lower one-sided, or upper one-sided test.

Based on this, and how much data we collect, we determine what the threshold(s) will be. Then, when we calculate the statistic from our data, we see if it is in the range, and decide $H_0$ or $H_1$. It is a binary decision, $H_0$ or $H_1$.

In science, this is also the procedure. However, when the data is reported, it is often accompanied by a report of the $p$-value. The $p$-value is the value of $\alpha$ which would have resulted in the statistic being equal to the threshold. That is, the $p$-value is the lowest $\alpha$ which would have resulted in a decision of $H_1$. This is useful information in many cases, because it tells us just how much we should believe in the null hypothesis.
For example, we may have set $\alpha = 0.05$, and have our decision rule that if $\frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}} > z_{0.05}$, then we’ll reject $H_0$. We collect the data, plug in $\bar{X}$ and see that the data is far, far above the threshold $z_{0.05}$. So we reject $H_0$. In fact, though the fraction is equal to the threshold $z_{0.0001}$. If all we report is that under $\alpha = 0.05$ that we reject $H_0$, then the person we’re reporting it to knows that the data is unlikely given $H_0$, but only knows it is less likely than 5%. In this case, the $p$-value is 0.0001, that is, we could have set $\alpha$ almost as low as 0.0001 and still decided $H_1$. Providing the information that the $p = 0.0001$ emphasizes that the data is very, very unlikely given $H_0$. In other words, you have even more confidence in your decision to reject $H_0$.

Another way to define $p$ is one used in statistics books:

$$p = P[\text{sample statistic or more extreme} | H_0 \text{ is true}].$$

If you’re interested in more perspective on the $p$-value there is a book called *What is a p-value anyway?*, by Andrew Vickers, 2009.

### 36 Tests on Proportions

Related to our tests on mean, we are sometimes interested in whether our assumption about the probability of success in a binomial random variable is correct. Recall this is what we defined as parameter $p$ when we discussed the binomial random variable. As an example, we ran coin-toss experiments in this class in which you tossed 10 coins. We discussed the numbers each team measured, but all we could do at that point was to look at the data and say that it looked right. The material in this section provides us with a method that would detect that the $p$ we assumed was incorrect.

**Example: Part Failures**

You sell an IC for which, after manufacturing, are 10% faulty. In quality control, you test them and throw out those that are faulty. (a) Today four out of the first ten are faulty. Does this provide enough evidence that the probability of a faulty IC is more than 10% faulty, with a 1% probability of false alarm? (b) Next, you test the next 100 and find that 19 are faulty. Does 19/100 faulty result provide enough evidence that the probability of a faulty IC is more than 10% faulty?

How can we address this problem? In either case, we’re interested in testing these hypotheses for $p_0 = 0.1$:

$$H_0 : \quad p = p_0,$$

$$H_1 : \quad p > p_0,$$

### 36.1 Gaussian Approximation

First, let’s address (b), the test of 100. The result that there are 19 faulty ICs is actually a sum of 100 random variables, $K = \sum_{i=1}^{100} X_i$ where $X_i$ is either 1 or 0. So even though $X_i$ are Bernoulli and the sum $K$ is a Binomial random variable, the sum obeys the central limit theorem and $K$ is approximately Gaussian. One can look up (in the Walpole book, for example) that since $K$ is Binomial, it has mean $\mu = np$ and variance $\sigma^2 = np(1 - p)$. Under $H_0$ the value of $p = 0.10$ and $n = 100$, so $\mu = 10$ and $\sigma = \sqrt{np(1-p)} = 3$. Thus the following $Z$ statistic is approximately standard normal:

$$Z = \frac{K - np}{\sqrt{np(1-p)}} = \frac{19 - 10}{3} = 3$$
Looking at Table A.3, \( P[Z > 3.00] = 1 - 0.9987 = 0.0013 \). At an \( \alpha = 0.01 \), we’d reject \( H_0 \).

Formalizing this, we have a binomial random variable with distribution \( b(k; n, p) \) where \( n \) is the number of trials in our sample, \( p \) is the probability of “success” of each trial, and \( k \) is the number of successes in our sample. If (a) \( n > 30 \) and (b) neither \( np \) nor \( n(1 - p) \) is close to 0.0, then under \( H_0 \), \( K \) is approximately Gaussian with mean \( \mu = np_0 \) and variance \( \sigma^2 = np_0(1 - p_0) \). We form a standard normal r.v. \( Z \) as:

\[
Z = \frac{k - np_0}{\sqrt{np_0(1 - p_0)}}
\]  

(28)

Then we form a two-sided, lower one-sided, or two-sided hypothesis (depending on when we want an alarm) as follows:

- **Two-sided**: Reject \( H_0 \) if \( Z < -z_{\alpha/2} \) or \( Z > -z_{\alpha/2} \).
- **Lower one-sided**: Reject \( H_0 \) if \( Z < -z_{\alpha} \).
- **Upper one-sided**: Reject \( H_0 \) if \( Z > z_{\alpha} \).

Compared to the test on the mean, there are a few minor differences:

1. The \( Z \) statistic looks slightly different. Note that we could write it identically if we divided top and bottom of (28) by \( n \), because \( k/n \) is the sample average of the \( X_1, \ldots, X_n \), and \( \sqrt{p_0(1 - p_0)} \) is the standard deviation of one of the observations \( X_i \).

2. There is no question about “unknown \( \sigma \)”. The standard deviation of a Bernoulli or binomial r.v. is a known function of \( p \). So under \( H_0 \), since \( p = p_0 \) is assumed, we know the standard deviation.

### 36.2 Exact Solution

When \( n \) is too low (below 30), we can’t accurately use the central limit theorem. However, we can still compute from the binomial pmf the probability of the observed data or worse, given \( H_0 \) is true. (This is the p-value.) If this conditional probability is lower than our pre-determined probability of false alarm \( \alpha \), we would reject \( H_0 \). Let’s first consider the upper one-sided hypothesis test, where \( H_1 \) is that \( p > p_0 \). We would calculate:

\[
P[K \geq k] = 1 - P[K < k] = 1 - \sum_{x=0}^{k-1} b(x; n, p),
\]

(29)

and if the probability is less than \( \alpha \), we’d reject \( H_0 \). Table A.1 can be used here, it has the sum \( \sum_{x=0}^{r} b(x; n, p) \) for \( 1 \leq n \leq 20 \), \( p \) from 0.1 to 0.9 in multiples of 0.1, and all \( 0 < r < n \). Here, because \( P[K < k] \) is looking for the probability that \( K \) is strictly less than \( k \), we have \( r = k - 1 \).

Solve part (a) of the example:

**Solution**: In part (a), we have \( K = 4, n = 10, \) and \( p_0 = 0.10 \),

\[
P[K \geq k] = 1 - \sum_{x=0}^{3} b(x; n = 10, p = 0.10) = 1 - 0.9872 = 0.0128.
\]

This value is above \( \alpha = 0.01 \), so we would not reject \( H_0 \). That is, we don’t have enough evidence to reject \( H_0 \) at the desired significance level. It would take getting 5 or more faulty ICs out of 10 to reject the null hypothesis at \( \alpha = 0.01 \) for \( n = 10 \).
Going back to the more general procedure: If it is a lower one-sided test, we’d compute

\[ P[K \leq k] = \sum_{x=0}^{k} b(x; n, p), \]  

(30)

and if the probability is less than \( \alpha \), we’d reject \( H_0 \). Note that for the \( P[K \leq k] \) case, \( r = k \).

If it is a two-sided test, to be complete, we could compute both. If either is less than \( \alpha/2 \), then we reject \( H_0 \). In reality, you don’t need to do both because one of them is going to be \( \geq 0.50 \) (since the two add to 1.0), and you don’t need to compute that one.