Sets and Notation

uOttawa - MAT2377

Fall 2020

Real-World to Math (1)

We start by turning informal descriptions of the real world into mathematical jargon. First is *experiment:*

Ex. 1: Dice

I roll a standard, fair 6-sided die.

Ex. 2: Pictures

I get up at 6 AM, pick up my phone, and take a picture out of my bedroom window.

The first “sounds like math," the second less so - but both are “experiments."

Real-World to Math (2)

Result: The thing that you write down at the end of the experiment. This "is math":

Ex. 1: Dice

Experiment: I roll a standard, fair 6-sided die.
Result: The number on top of the die.

Ex. 2: Pictures

Experiment: I get up at 6 AM, pick up my phone, and take a picture out of my bedroom window.
Result: An image file on my computer.

Real-World to Math (2)

Result: The thing that you write down at the end of the experiment. This "is math":

Ex. 1: Dice

Experiment: I roll a standard, fair 6-sided die.
Result: The number on top of the die.

Ex. 2: Pictures

Experiment: I get up at 6 AM, pick up my phone, and take a picture out of my bedroom window.
Result: An image file on my computer.
Result (Alternative): A binary "has bird" or "no bird."

You can choose what to call "the" result.

Real-World to Math (3)

Sample Space: The set of all possible results.

Ex. 1: Dice

Experiment: I roll a standard, fair 6-sided die.
Result: The number on top of the die.
Sample Space: $\Omega = \{1, 2, 3, 4, 5, 6\}$.

Ex. 2: Pictures

Experiment: I get up at 6 AM, pick up my phone, and take a picture out of my bedroom window.
Result: An image file on my computer. *Sample Space:* $\Omega = \{1, 2, \ldots, 1073741824\}^{12897485}$ (for 12897485 pixel image, 9 bits per pixel).

Real-World to Math (4)

An *event* is any subset of the sample space.

Ex. 1: Dice

Experiment: I roll a standard, fair 6-sided die.
Result: The number on top of the die.
Sample Space: $\Omega = \{1, 2, 3, 4, 5, 6\}$.
Events: Roll a 1; Roll an even; ...

Lots of choices.

Sample Spaces, Events, and Sets

In probability, we always have a *sample space* (big set) and many *events* (subsets). Doing calculations means getting comfortable with sets, and diagrams like:

Set Notation (1)

1. $\emptyset$ is the empty set, Ω is the entire sample space.
2. $A \subset B$ means 'every element of A is an element of B.' **Examples:** $\{1,4\} \subset \{1,4,5\}$ but $\{1,6\} \not\subset \{4,5,6\}$.
3. $A \cup B$ means "everything in A **or** B." **Example:** $\{1,2\} \cup \{1,6\} = \{1,2,6\}$.
4. $A \cap B$ means "everything in A **and** B." **Example:** $\{1,2,3\} \cap \{3,4,5,6\} = \{3\}$.
5. A' or A^c means "everything **not** in A." **Example:** In our dice example, $\{1,2,4\}^c = \{3,5,6\}$.

Warning: A^c only makes sense if you know Ω!

Set Notation (2)

We often write sets using notation that looks like:

$$A = \{x \in \Omega : \phi(x)\}.$$

The stuff after the colon gives a condition that the stuff before the colon has to meet. The set is everything that meets the condition.

examples

$$\{x \in \{1, 2, 3, 4, 5, 6\} : x \text{ is even.}\} = \{2, 4, 6\}.$$

You can define a set this way even if it is hard to figure out what is actually in the set:

$$\{x \in \mathbb{R} : x^4 - 17x^3 + 101x^2 - 247x + 210 = 0\} = \{2, 3, 5, 7\}.$$

Set Manipulation (1)

One of the main skills to pick up in this class is *set manipulation* - being able to go between different descriptions of the same set, using formulas like:

$$(A \cup B)^c = A^c \cap B^c$$

Set Manipulation (2)

Set manipulation is like algebra: you have some operations (e.g. $+, -, \times, \ldots$ for algebra, $\cap, \cup, \ldots$ for sets) and some rules for moving them around, like:

Some Algebra Rules

$$x + (y + z) = (x + y) + z;\ x(y + z) = (xy) + (xz).$$

Some Set Rules

$$A \cup (B \cup C) = (A \cup B) \cup C;\ A \cap (B \cup C) = (A \cap B) \cup (A \cap C).$$

Like with algebra, "simplifying" sets will feel natural eventually (but probably not immediately).

Set Manipulation (3)

Where do set manipulation rules come from?

1. A big list in your textbook.
2. Some complicated algebra proofs in your textbook.
3. Venn diagrams.
4. ...

I think (3) is easiest for most people, so concentrate on that for now.

Venn Diagrams and Set Formulas

Let's "prove" the formula $(A \cup B)^c = A^c \cap B^c$ by comparing pictures.

Many More Formulas for Practice

- $A \cap A^c = \emptyset$.
- $A \cup A^c = \Omega$.
- $A \cap (B \cap C) = (A \cap B) \cap C$.
- $A \cap (B \cup C) = (A \cap B) \cup (A \cap C)$.
- $A \cup (B \cap C) = (A \cup B) \cap (A \cup C)$.
- $(A \cap B)^c = A^c \cup B^c$.
- $(A \cup B)^c = A^c \cap B^c$.

Some Properties of Groups of Sets

Considering a collection $A_1, \ldots, A_n$ of sets:

1. They are *mutually exclusive* or *disjoint* if $A_i \cap A_j = \emptyset$ for $i \neq j$.
2. They are *exchaustive* if $\cup_{i=1}^{n} A_i = \Omega$.
3. They form a *partition* if they are both.

Axioms of Probability

uOttawa - MAT2377

Fall 2020

Viewpoints on Probability 1: Counting

Goal: Assign probability to some event A in some experiment.

1. **Example:** I roll a 6-sided die; what is the chance of getting an even number?
2. **Natural guess:** 3 even numbers, 6 numbers, so

$$P[even] = \frac{3}{6}.$$

3. **Generalization**: Probability of A is *number of ways for A to happen* divided by *number of things that can happen.*

Some big problems - outside of card and dice games, not all events are equally likely.

Viewpoints on Probability 2: Long-Run Frequencies

Goal: Assign probability to some event A in some experiment.

1. Perform *same experiment* many times; say experiment *succeeds* if event occurs.
2. Let S_n be number of successes in first n experiments.
3. Define $P[A] = \lim_{n\to\infty} \frac{S_n}{n}$.

Still some problems - what about one-off events, like sports competitions?

Viewpoints on Probability 3: Subjective Belief

Goal: Assign probability to some event A in some experiment.

1. Say $P[A]$ is *my* belief about its likelihood.
2. Different people may have different probability functions for the same experiment! I'm from here, and tend to assign

$$P[\text{Ottawa Senators win Stanley Cup}] > 0.9$$

most years. However, other people may disagree.

Fuzziness leads to possible problems - what if I have inconsistent beliefs, like $P[\text{Sens win}] = 0.9$ and also $P[\text{Leafs win}] = 0.8$?

Viewpoints on Probability 4: Axioms of Probability

Goal: Assign probability to some event A in some experiment.

1. **Mathematicians don't want to decide your viewpoint for you.**
2. **Main Idea:** Say $P : 2^{\Omega} \mapsto [0, 1]$ is a *probability function* as long as it describes *some* consistent worldview.

Notation aside: I write 2^{Ω} for the collection of all subsets of the state space Ω - that is, 2^{Ω} is the set of events.

Axioms of Probability

We say $\mathbb{P} : 2^{\Omega} \mapsto [0, 1]$ is a *probability measure* if:

1. $P[\Omega] = 1$.
2. For any countable sequence $\{A_i\}_{i \in \mathbb{N}}$ of pairwise mutually exclusive events, $P[\cup_{i \in \mathbb{N}} A_i] = \sum_{i \in \mathbb{N}} P[A_i]$.

Example 1: Fair Die

Let $\Omega = \{1, 2, 3, 4, 5, 6\}$. For $A \subset \Omega$, define $P[A] = \frac{|A|}{6}$. Then for any event A, the probability $P[A]$ is exactly the probability that the outcome of a fair die roll is in A.

In-Class Exercise: Show that P satisfies the axioms of probability.

Example 1: Unfair Die

Let $\Omega = \{1, 2, 3, 4, 5, 6\}$. For $A \subset \Omega$, define $P[\{i\}] = \frac{i}{21}$ and $P[A] = \sum_{i \in A} P[\{i\}]$.
This is also a probability, but it describes a very unfair die.

Example 1: Calculation

For both the “fair" and “unfair" dice, calculate $P[\{1,3\}]$.

1. In the fair case,

$$P[\{1,3\}] = P[\{1\}] + P[\{3\}] = \frac{1}{6} + \frac{1}{6} = \frac{1}{3}.$$

2. In the unfair case,

$$P[\{1,3\}] = P[\{1\}] + P[\{3\}] = \frac{1}{21} + \frac{3}{21} = \frac{4}{21}.$$

Note that the *answers* are different but *calculations* are essentially identical.

Example 2: Normalizing Constant

Let $\Omega = \{1, 2, 3\}$ and $P[\{k\}] = Ck^2$. What is C?

$$1 = P[\Omega] = P[\{1\}] + P[\{2\}] + P[\{3\}] = C(1 + 4 + 9) = 14C.$$

Therefore,

$$C = \frac{1}{14}.$$

Optional: Further Probability Frameworks

I'll give two less-obvious frameworks that can be useful in reframing some calculations:

1. Dartboard Model (simple, surprisingly useful in this course).
2. Gambling Model (complicated, not needed in this course, sort of fun).

Viewpoints on Probability 4: Dartboard Model

Following is completely general, and surprisingly helpful for later subjects such as conditional probabilities.

1. Draw Ω as unit square $[0,1]^2$ dartboard.
2. Draw any events in Ω.
3. Define

$$P[A] = \textit{Area of } A = \text{Probability of hitting A with dart.}$$

In-class: Doodle for $\Omega = \{1,2,3,4,5,6\}$, $A = \{1,2\}$.

Viewpoints on Probability 5: Gambling Model

1. A *bet on A with payoff X* means: you pay 1 dollar; if A happens you are then given X dollars.
2. **Example:** for a *fair* coin flip, a *fair* bet on Heads would have payoff $X = 2$.
3. Somebody comes up to you and proposes the following game:
   - 3.1 *You* propose a payoff X for event A.
   - 3.2 *They* force you to bet on either A (with payoff X) or A^c (with payoff $\frac{X}{X-1}$).
4. Define your "fair bet" probability:

$$P[A] = \frac{1}{X}.$$

Fun Fact: P is a probability function if you don't *deterministically* lose money for some sequence of events $A_1, A_2, \ldots$.

Probability Formulas and Independence

uOttawa - MAT2377

Fall 2020

Recall

Axioms of Probability

We say $\mathbb{P} : 2^{\Omega} \mapsto [0, 1]$ is a *probability measure* if:

1. $P[\Omega] = 1$.
2. For any countable sequence $\{A_i\}_{i \in \mathbb{N}}$ of pairwise mutually exclusive events, $P[\cup_{i \in \mathbb{N}} A_i] = \sum_{i \in \mathbb{N}} P[A_i]$.

Some Set Union/Intersection/Complement Formulas

$A \cap (B \cup C) = (A \cap B) \cup (A \cap C)$, $(A \cup B)^c = A^c \cap B^c, \ldots$

- **Today:** formulas for *probabilities* of set union/intersection/complement.
- **Key tool:** Venn diagrams + Dartboard Model.

First Formula

Simple Union Formula

For any sets A, B with $A \cap B = \emptyset$, we have $P[A \cup B] = P[A] + P[B]$.

Proof: Doodle Venn diagram and use Dartboard Model.

Consequence of First Formula

Complement Formula

For any set A, $P[A^c] = 1 - P[A]$.

Proof: Recall, for any sets A, B with $A \cap B = \emptyset$, we have $P[A \cup B] = P[A] + P[B]$.
Using $B = A^c$,

$$P[A^c] = P[A \cup A^c] - P[A] = P[\Omega] - P[A] = 1 - P[A].$$

Example

Roll a fair die. What is the probability of *not* getting a 1?

$$P[\{1\}^c] = 1 - P[\{1\}] = 1 - \frac{1}{6} = \frac{5}{6}.$$

Improving First Formula

Intersection/Union Formula

For any sets A, B with $A \cap B = \emptyset$, we have
$P[A \cup B] = P[A] + P[B] - P[A \cap B]$.

Proof: Doodle Venn diagram and use Dartboard Model.

Example

I roll two fair dice. What is the probability of getting *at least one* 6?

1. Let A_1, A_2 be the event that the first (respectively second) die is a 6.
2. We compute:

$$P[A_1 \cup A_2] = P[A_1] + P[A_2] - P[A_1 \cap A_2] = \frac{1}{6} + \frac{1}{6} - \frac{1}{36}.$$

Study Suggestion

The formula

$$P[A \cup B] = P[A] + P[B] - P[A \cap B]$$

is one of the most important formulas in the course.

Independence

1. **Informal Idea:** Sometimes two events don't seem to "influence each other."
2. **Example:** The two dice in the last example!
3. **Math Definition:** Events A, B are independent if

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

Example of Independence

1. We just did one!
2. **Number one error in this class:** using the formula $P[A \cap B] = P[A]P[B]$ when A, B are *not* independent.
3. **How can I tell?**
   - 3.1 For games, can *assume* that dice, coins, etc are independent unless I say otherwise.
   - 3.2 For other situations, I will *tell you* if I want you to *assume* independence.
   - 3.3 Otherwise, don't *assume* independence (though you may be able to *prove* it.

Weird Example of Independence (1)

1. The definition of independence is very good, but may not match your intuition.
2. If events are "obviously" independent (*e.g.* die rolls), they are independent in this definition.
3. Sometimes events are independent in a "nonobvious" way too!

Weird Example of Independence (2)

Let X_1, X_2 be the results of two independent die rolls. Let $A = \{X_1 = 1\}$, $B = \{X_1 + X_2 = 2\}$, and $C = \{X_1 + X_2 = 7\}$. Are A, B independent? What about A, C?

1. We have $A \cap B = B$ as sets, so

$$P[A \cap B] = P[B] \neq P[A]P[B].$$

   Thus A, B are *not* independent.

2. We have

$$P[A \cap C] = P[X_1 = 1, X_2 = 6] = \frac{1}{36} = P[A]P[C].$$

   Thus A, C *are* independent.

A, C don't look independent - but they are!

Weird Example of Independence (3)

Let X_1, X_2 be the results of two independent die rolls. Let $A = \{X_1 = 1\}$, $B = \{X_1 + X_2 = 2\}$, and $C = \{X_1 + X_2 = 7\}$. A, B are dependent, but similar-looking A, C are independent.

1. This "coincidental" independence is a *feature*, not a *bug.*
2. **Intution:** knowing that A occurs tells you that B is much more likely - but (by coincidence) knowing that A occurs tells you *exactly nothing* about the probability that C occurs.
3. These "coincidences" are rare in nature, but can be useful in designed systems.

Further Reading

1. Textbook had many examples of set formulas...
2. There are also many probability formulas.
3. **The important ones can be derived from the ones in this class.**

Elementary Combinatorics

uOttawa - MAT2377

Fall 2020

Goals

1. A common model for probability is:

$$P[A] = \frac{|A|}{|\Omega|}.$$

2. You need to be able to count $|A|$, $|\Omega|$; in math we call this "combinatorics."

Why is this a topic? (1)

We learned how to count in elementary school...

Why is this a topic? (2)

There is counting, and there is *counting*:

1. I roll one die; how many ways are there to get an even number? **In your head:** 3.
2. I roll two dice; how many ways for the *sum* to be multiple of 3? **With a piece of paper:** 12.
3. I roll 10 dice; how many ways for the *sum* to be a multiple of 5? **???**
4. I roll 10000 dice; how many ways for *sum* to be a multiple of 12 *and* sum-of-squares to be over 32000? **???!!!???**

We want to be able to count *in a reasonable amount of time.*

First General Principle

Multiplication Principle 1

There are $|A|\,|B|$ ways to choose one item from set A and then one item from set B.

Multiplication Principle 2

Let $\Omega = A_1 \times A_2 \times \ldots A_k$. Then

$$|\Omega| = |A_1| \times |A_2| \times \ldots \times |A_k|.$$

Example

1. A pizza store has 15 toppings and 4 types of crust. How many single-topping pizzas are possible?
2. **Answer:** By multiplication principle, there are

$$(15)(4) = 60$$

possible pizzas.

Second General Principle

1. The counting principle is good for *very structured* sets.
2. On the opposite extreme, we can use *tree diagrams* to count *very unstructured* sets.
3. Best to go by example.

Tree Diagram Example

1. I wish to count the number of ways to rearrange the letters AABBB, subject to the constraint that there are no repeated A's (so ABBAB is legal, but BAABB is not).
2. This set looks a bit weird - I can't think of any better strategy than just listing its elements.
3. You *could* do this "just by looking"; a tree diagram is a way to make sure you didn't miss anything.

Next: we'll doodle a solution.

Putting It Together (1)

I roll three fair dice, getting X_1, X_2, X_3. What is

$$P[\{X_1 + X_2 + X_3 = 5\}]?$$

Putting It Together (2)

Define $\Omega = \{1, 2, 3, 4, 5, 6\}^3$. By multiplication principle,

$$|\Omega| = 6^3 = 216.$$

Putting It Together (3)

Define $A = \{(x_1, x_2, x_3) \in \{1, 2, 3, 4, 5, 6\}^3 \, : \, x_1 + x_2 + x_3 = 5\}$. By tree diagram (or just looking carefully),

$$A = \{(2, 2, 1), (2, 1, 2), (1, 2, 2), (3, 1, 1), (1, 3, 1), (1, 1, 3)\}.$$

Thus, $|A| = 6$.

Putting It Together (4)

We have computed $|A| = 6$ and $|\Omega| = 216$, so

$$P[\{X_1 + X_2 + X_3 = 5\}] = \frac{|A|}{|\Omega|} = \frac{1}{36}.$$

Next Videos

We'll look at *intermediate-difficulty* problems, where the multiplication principle doesn't apply but we can do a lot better than tree diagrams.

Permutations and Combinations

uOttawa - MAT2377

Fall 2020

Goals

We'll jump right into some examples...

Permutation Example 1

How many ways are there to *rank* 5 people?

Permutation Example 1: Solution

In class - box doodle.

Permutation Example 2

How many ways are there to rank *the first 3* of 5 people?

Permutation Example 2: Solution

In class - box doodle.

Permutation Principle

There are

$$\frac{n!}{(n-k)!}$$

ways to rank k things in a group of n.

Combination Example

How many ways are there to *choose* 3 of 5 people?

Combination Example: Solution

1. There were (5)(4)(3) ways to *rank* 3 people.
2. Some rankings are the same choice: $\{1, 3, 5\} = \{3, 5, 1\}$.
3. **Key insight:** each choice can be permuted into 3! rankings!
4. Thus there are $\frac{(5)(4)(3)}{3!} = 10$ ways to *choose* 3 people.

Permutation Principle

There are

$$\frac{n!}{(n-k)!k!} \equiv \binom{n}{k}$$

ways to choose k things from a group of n.

Advanced Counting

uOttawa - MAT2377

Fall 2020

Goals

We have seen several related principles for counting, and for relating *counting* to *probability*. In this note, we put some of them together to answer more advanced questions.

Quick Review

Multiplication Principle

There are $|X|\,|Y|$ ways to choose one item from set X and then one item from set Y.

Counting Choices

There are $\binom{|X|}{k} = \frac{|X|!}{k!(|X|-k)!}$ ways to choose k items from a set X.

Uniform Probability

The *uniform probability on* X is $P[A] = \frac{|A|}{|X|}$.

Example: Netflix Day (1)

Netflix suggests 10 movies to me. If I watched them all, I'd like 3 (and dislike the other 7). I can't get off the couch on Sunday and watch 5 movies in a row, chosen at random from Netflix's suggestion; what is the probability that I dislike *all* of them? What is the probability that I dislike *all but one*?

Example: Netflix Day (2)

Let N be the number of ways to choose 5 movies out of 10, let n_k be the number of ways to choose k *bad* movies out of 7, and let m_k be the number of ways to choose k *good* movies out of 3. By the definition of *uniform probability*,

$$p_1 =$$

Using this and *multiplication principle*,

$$p_2 =$$

Example: Netflix Day (3)

Let's evaluate N and n. By the *choosing formula*,

$$n_k = \binom{7}{k}$$

and similarly

$$N =$$

Example: Netflix Day (4)

Putting this all together,

$$p_1 = \frac{\binom{7}{5}}{\binom{10}{5}}.$$

$$p_2 =$$

Example: Poker Hands (1)

A *full house* in Poker is a collection of 5 cards that includes both a 3-of-a-kind and a 2-of-a-kind (*e.g.* $\{10, 10, 10, 4, 4\}$).
If I draw 5 cards out of a 52-card deck at random, what is the probability of getting a full house?

Example: Poker Hands (2)

As before, we must count *number of ways to get a full house* and *number of ways to get 5 cards*. The second number is easy, by choosing formula:

$$N = \binom{52}{5}.$$

Example: Poker Hands (3)

To count the number of ways to full house, we do the following:

1. Choose the value of the 3-of-a-kind (there are
2. Choose the suits for the 3-of-a-kind (there are
3. Choose the value of the 2-of-a-kind (there are
4. Choose the suits for the 2-of-a-kind (there are

Note: the *number of choices remaining* don't depend on *which choices are already made* - whether my 3-of-a-kind is 10 or 7, I still have 12 choices for my 2-of-a-kind.

Example: Poker Hands (4)

Putting this together,

$$p = \frac{\binom{13}{1}\binom{4}{3}\binom{12}{1}\binom{4}{2}}{\binom{52}{5}}.$$

Example: Breakups and Movies (1)

You go to a movie with 11 friends and arrive a little late. There are 11 seats left, in groups of $(4, 4, 3)$. Unfortunately, Charlie and Sam just broke up, and don't want to be in the same group.
How many ways are there for your 11 friends to be seated, if Charlie and Sam must be separated?

Example: Breakups and Movies (2)

Let's start with an easier question. *Without* the condition on Charlie and Sam, the answer is:

$$N = 11!.$$

Example: Breakups and Movies (3)

For the main question, let's try the same procedure as with Poker hands:

1. Seat Charlie (11 choices).
2. Seat Sam (
3. Seat everyone else (9! choices).

Example: Breakups and Movies (4)

For the main question, let's try the same procedure as with Poker hands:

1. Seat Charlie (11 choices).
2. Seat Sam Problem! The number of choices depends on Charlie's seat!
3. Seat everyone else (9! choices).

Let's draw:

Example: Breakups and Movies (5)

Refining:

1. Seat Charlie (11 choices).
2. Seat Sam (8 choices if Charlie chose the 3-seat group, 7 choices otherwise).
3. Seat everyone else (9! choices).

Total:

$$N = (3)(8)(9!) + (8)(7)(9!).$$

Conditional Probability

uOttawa - MAT2377

Fall 2020

Goals

1. We know what we mean by "the chance that the Sens win the Stanley Cup *given that* they make the playoffs."
2. **Today:** turning "given that" into precise math.
3. **Starting out:** an example where we "know" the right definition.

Conditional Probability from Tables (1)

A University decided to study the impact of studying and joining a sports team on grades, obtaining the data:

Sports Team	Study Lots	Don't Study
Good Grades	85	500
Bad Grades	15	1000

No Team	Study Lots	Don't Study
Good Grades	3500	1500
Bad Grades	1000	3400

For the sports team members, calculate the probability of getting good grades *given that* you studied lots - in our new notation, $P[GG|SL]$

Conditional Probability from Tables (2)

Sports Team	Study Lots	Don't Study
Good Grades	85	500
Bad Grades	15	1000

1. Look at 100 people on sports teams who studied a lot.
2. 85 got good grades...
3. ... so the chance that a *random* person from that hundred got good grades is:

$$P[GG|SL] = \frac{85}{85+15} = 0.85$$

Conditional Probability from Tables (3)

Write “GG" for good grades, and “SL" and “DS" for “study lots" and “didn’t study." For both members and non-members, calculate $P[GG|SL]$, $P[GG|DS]$ and $P[GG]$. Members:

$$P[GG|SL] = \frac{85}{85+15} = 0.85.$$

$$P[GG|DS] = \frac{500}{500+1000} \approx 0.33.$$

$$P[GG] = \frac{85+500}{85+15+500+1000} \approx 0.37.$$

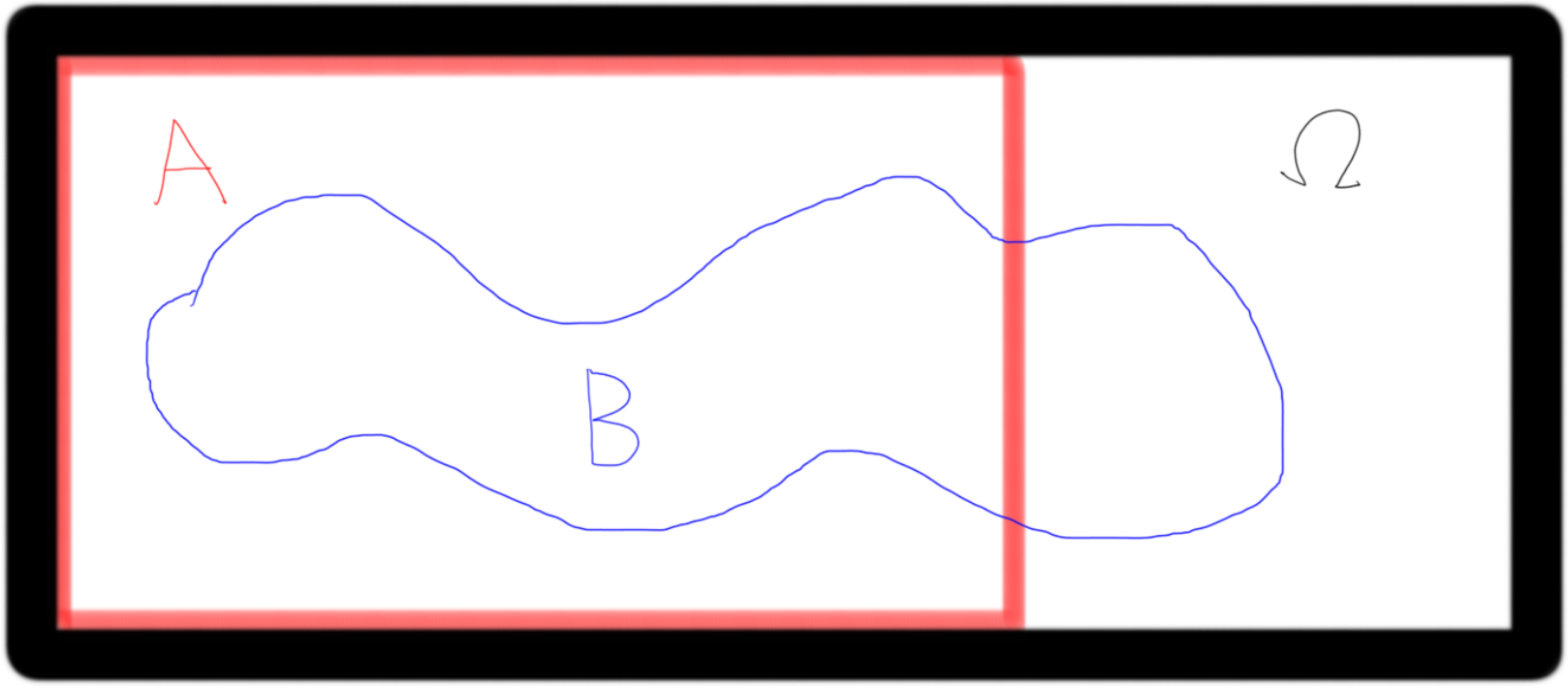
Non-members:

$$P[GG|SL] = \frac{3500}{3500+1000} \approx 0.78.$$

$$P[GG|DS] = \frac{1500}{1500+3400} \approx 0.31.$$

$$P[GG] = \frac{3500+1500}{3500+1000+1500+3400} \approx 0.51.$$

Dartboard Model of Probability



Conditioning = shrinking the dartboard.

Formal Definition

Say *the conditional probability of B given A is*

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

Simple Example

Let A, B satisfy $P[A] = 0.5$, $P[B] = 0.8$ and $P[A \cap B] = 0.3$.
Calculate $P[B|A]$.
Using the formula,

$$P[B|A] = \frac{P[A \cap B]}{P[A]} = \frac{0.3}{0.5} = 0.6.$$

Longer Example (1)

- **Question:** An urn contains 7 balls: 4 white and 3 black. Two balls are selected at random without replacement, and you are told that at least one is white. What is the probability that both are white?
- The hardest part here is setting up the question! **General tips:**
  1. Write down lots of elementary events; you can always combine them later.
  2. Even if you don't know all of the calculations that you will have to do, the first step is normally obvious. If you've defined your events will, this will let you break up a problem into little bits. Breaking the problem up into little bits is the main skill you will learn!

Longer Example (2)

- **Question:** An urn contains 7 balls: 4 white and 3 black. Two balls are selected at random without replacement, and you are told that at least one is white. What is the probability that both are white?
- Let WW, WB, BW and BB be the balls chosen, and A be the event that there is at least one white ball. Have:

$$P[WW|A] = \frac{P[A \cap WW]}{P[A]} = \frac{P[WW]}{P[WW] + P[WB] + P[BW]}.$$

Recall:

$$P[WW] = \frac{4}{7}\frac{3}{6} = \frac{2}{7}, \; P[WB] = P[BW] =$$

Plug-in to conclude:

$$P[WW|A] = \frac{1}{3}.$$

Tricky Question

- Let A, B satisfy $P[A] = 0.8$, $P[B] = 0.9$. Is it possible that $P[A|B] = 0.2$?
- Assuming $P[A|B] = 0.2$,

$$P[A \cap B] = P[A|B]P[B] = (0.2)(0.9) = 0.18.$$

Thus, we would have

$$1 \geq P[A \cup B] = P[A] + P[B] - P[A \cap B] = 0.8 + 0.9 - 0.18 = 1.52 > 1.$$

Contradiction!

Bayes' Rule

uOttawa - MAT2377

Fall 2020

Goals

How to go between pairs of sentences like:

1. “The probability that I am sick *given* the test says I am sick."
2. “The probability that the test says I am sick *given* I am sick."

Law of Total Probability (Example)

80 of 100 people get a flu shot on Oct. 10. Assume $P[Flu|Shot] = 0.03$, $P[Flu|Shot^c] = 0.06$. Compute $P[Flu]$.

Law of Total Probability (Example)

80 of 100 people get a flu shot on Oct. 10. Assume $P[Flu|Shot] = 0.03$, $P[Flu|Shot^c] = 0.06$. Compute $P[Flu]$.

$$\begin{aligned} P[F] &= P[F \cap S] + P[F \cap S^c] \\ &= P[F|S]P[S] + P[F|S^c]P[S^c] \\ &= (0.03)(0.8) + (0.06)(0.2) \\ &= 0.036. \end{aligned}$$

Law of Total Probability (Theorem)

Say $A_1, \ldots, A_n$ *partition* Ω if

1. They are pairwise mutually exclusive, *and*
2. $A_1 \cup \ldots \cup A_n = \Omega$.

For a partition,

$$P[B] = \sum_i P[B \cap A_i] = \sum_i P[B|A_i]P[A_i].$$

In-class: Doodle proof.

Bayes' Rule (Theorem)

Let $A_1, \ldots, A_n$ be a partition. Then

$$\begin{aligned} P[A_1|B] &= \frac{P[A_1 \cap B]}{P[B]} \\ &= \frac{P[A_1 \cap B]}{\sum_i P[B \cap A_i]} \\ &= \frac{P[B|A_1]P[A_1]}{\sum_i P[B|A_i]P[A_i]}. \end{aligned}$$

Note: Last is most common - we've "exchanged" A, B.

Bayes' Rule (Example)

80 of 100 people get a flu shot on Oct. 10. Assume $P[Flu|Shot] = 0.03$, $P[Flu|Shot^c] = 0.06$. Compute $P[Shot|Flu]$.

Bayes' Rule (Example)

80 of 100 people get a flu shot on Oct. 10. Assume $P[Flu|Shot] = 0.03$, $P[Flu|Shot^c] = 0.06$. Compute $P[Shot|Flu]$. We calculate:

$$\begin{aligned} P[S|F] &= \frac{P[F|S]P[S]}{P[F|S]P[S] + P[F|S^c]P[S^c]} \\ &= \frac{0.024}{0.036} \\ &= \frac{2}{3}. \end{aligned}$$

Note: even though flu shots make you less likely to get the flu, most people who have the flu also had a flu shot.

Combined Example (1)

- People attending a conference have a choice of three hotels: A, B, C. 60 percent of attendees stay at hotel A, 30 percent at B and 10 percent at C. 5 percent of showers are broken at A, 10 at B and 50 at C.
- Calculate the probability that a random attendee has a broken shower. Also calculate the probability that a random attendee was at hotel A, given that they had a broken shower.

Combined Example (2)

Observe: The choice of hotel is a *partition.* Want $P[S], P[A|S]$.

- Let S denote the event that the random attendee's shower is broken.

$$\begin{aligned} P[S] &= P[S|A]P[A] + P[S|B]P[B] + P[S|C]P[C] \\ &= (0.05)(0.6) + (0.1)(0.3) + (0.5)(0.1) = 0.11. \end{aligned}$$

- For the second part,

$$\begin{aligned} P[A|S] &= \frac{P[S|A]P[A]}{P[S]} \\ &= \frac{(0.05)(0.6)}{0.11} = \frac{3}{11} \approx 0.273. \end{aligned}$$

Standard Machine

uOttawa - MAT2377

Fall 2020

Goals

1. We've seen how to do probability problems in an ad-hoc way.
2. **Today:** An algorithm for "breaking down" elementary probability questions.

Intersection Possibilities

- A, B are *mutually exclusive* if $A \cap B = \emptyset$.
- A is *contained in* B is $A \cap B = A$.
- A, B are *independent* if $P[A \cap B] = P[A]P[B]$ (equivalent: $P[A|B] = P[A]$).

Other possibility: $P[A \cap B]$ given in question.

Standard Machine (picture)

$$P[A \cup (B \cap C)]$$

$$\downarrow \text{(just intersections)}$$

$$P[A] + P[B \cap C] - P[A \cap B \cap C]$$

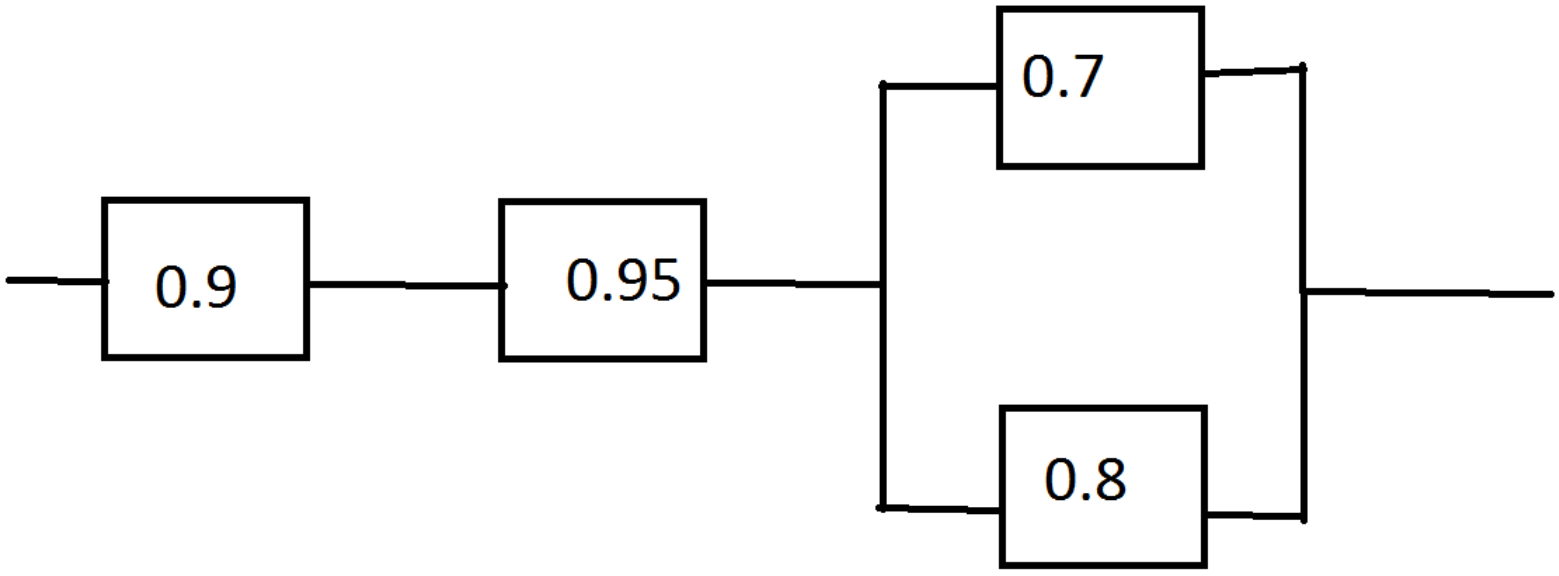
$$\downarrow \text{(assumptions)}$$

$$P[A] + P[B]P[C] - P[A]P[B]P[C]$$

Standard Machine (text)

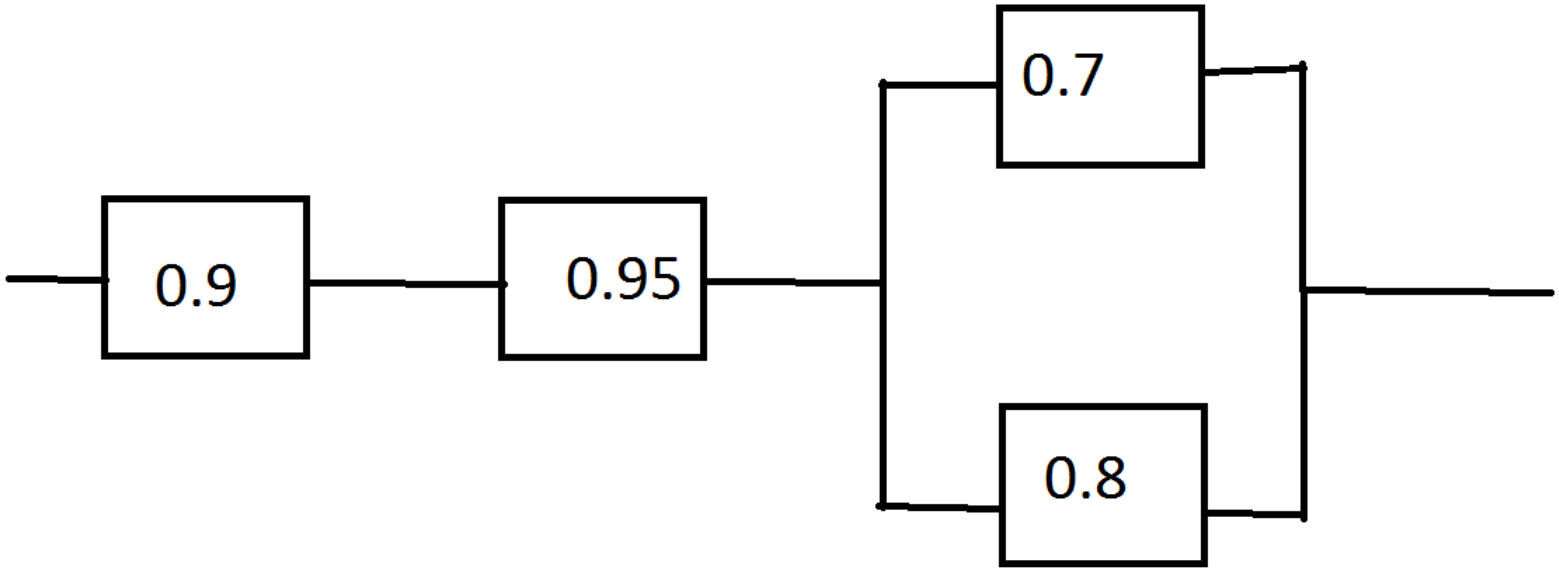
1. *Write down* the event of interest, A, in terms of a bunch of *independent* events $A_1, \ldots, A_k$.
2. *Expand* the expression for A and *Replace* all unions with intersections by the formula
   $P[A_i \cup A_j] = P[A_i] + P[A_j] - P[A_i \cap A_j]$.
3. *Replace* all intersections according to previous possibilities.

Standard Machine Example (1)



Circuit *works* if there is a left-right path that works; **assume independence.**

Standard Machine Example (2)



$$P[W] = P[C_1 \cap C_2 \cap (C_{top} \cup C_{bottom})]$$

Standard Machine Example (3)

$$
\begin{aligned}
P[W] &= P[C_1 \cap C_2 \cap (C_{top} \cup C_{bottom})] \\
&= P[(C_1 \cap C_2 \cap C_{top}) \cup (C_1 \cap C_2 \cap C_{bottom})] \\
&= P[C_1 \cap C_2 \cap C_{top}] + P[\cup(C_1 \cap C_2 \cap C_{bottom}] \\
&\quad - P[C_1 \cap C_2 \cap C_{top} \cap C_1 \cap C_2 \cap C_{bottom}] \\
&= P[C_1 \cap C_2 \cap C_{top}] + P[\cup(C_1 \cap C_2 \cap C_{bottom}] \\
&\quad - P[C_1 \cap C_2 \cap C_{top} \cap C_{bottom}] \\
&= P[C_1]P[C_2]P[C_{top}] + P[C_1]P[C_2]P[C_{bottom}] \\
&\quad - P[C_1]P[C_2]P[C_{top}]P[C_{bottom}] \\
&= (0.9)(0.95)(0.7) + (0.9)(0.95)(0.8) - (0.9)(0.95)(0.7)(0.8) \\
&= 0.8037.
\end{aligned}
$$

Standard Machine Slogan

These problems are *messy* but *easy* - just need to keep simplifying!

Elementary Probability Review

uOttawa - MAT2377

Fall 2020

Main Ideas

1. Axioms of probability.
   - We learned how to do set operations.
   - We learned important identities, like $P[A \cup B] = P[A] + P[B] - P[A \cap B]$.
   - Not many direct questions; have to know that $0 \leq P[A] \leq 1$, $P[\Omega] = 1$.
2. Rules for counting.
   - Relationship between counting and probability.
   - Multiplication rule.
   - Binomial coefficients.
3. Conditional probability.
   - A 'standard machine' for conditional probability questions.
   - Several tricky questions, and the value of just doing calculations when you're stuck.
4. Bayes rule.
   - A 'standard machine' for Bayes' rule questions.
   - More tricky questions and a sanity check.
   - Our first theorem with conditions: we need a partition in order to use Bayes' rule.

Random Variables

uOttawa - MAT2377

Fall 2020

Goals

1. Set up better notation for more complicated situations.
2. Talk about *probability* without spelling out *sample space.*

Informal Definition

Informal Random Variables

A random variable X is a number that depends on the outcome of an experiment. It may or may not give you all of the details of the outcome of the experiment.

An Example

Say we roll two dice. Then the sum of the numbers rolled is a random variable. Note that it isn't *everything* about the results of the two dice, since $(1, 6), (6, 1)$ and $(3, 4)$ all have the same sum.

Formal Definition

Random Variable

Fix a sample space Ω. A random variable X is a function from Ω to some other set.

In this course, we generally have:

Real Random Variable

A *real* random variable X is a function from Ω to the set $\mathbb{R}$ of real numbers.

Let's unpack this with an example:

Familiar Example

Let $\Omega = \{1, 2, 3, 4, 5, 6\}^2$. Then $X : \Omega \mapsto \{1, 2, 3, \ldots, 12\}$ given by

$$X[\omega_1, \omega_2] = \omega_1 + \omega_2$$

is a random variable.
If we view Ω as the possible outcomes when we roll two dice, then X is just the sum of the two dice.

Another Familiar Example

Let Ω be the collection of all possible 5-card hands of cards. For $\omega \in \Omega$, let

$$X(\omega) = 1, \qquad \text{all cards in } \omega \text{ have the same colour}$$
$$X(\omega) = 0, \qquad \text{otherwise.}$$

So X tells us if our hand is a 'flush' in poker, but doesn't tell us anything else about our hand.

Random Variables and Probabilities

We use the shorthand:

$$P[X = x] = P[\{\omega \in \Omega : X(\omega) = x\}].$$

Note:

- RHS is already defined, LHS didn't have any clear meaning.
- **You will never need to "use" this - this notation is supposed to "get out of the way."**

Using Random Variables

- Let X be the sum of 3 fair dice. Calculate $P[X = 4]$.
- Let $\Omega = \{1, 2, 3, 4, 5, 6\}^3$. We have

$$\begin{aligned} P[X(\omega) = 4] &= P[\omega \in \{(1, 1, 2), (1, 2, 1), (2, 1, 1)\}] \\ &= P[\omega = (1, 1, 2)] + P[\omega = (1, 2, 1)] + P[\omega = (2, 1, 1)] \\ &= 3(\frac{1}{6})^3 \approx 0.014. \end{aligned}$$

Note: The calculation is not new!

Why Use Random Variables? (1)

We “talk" in random variables, not sample spaces - our notation should match this.

Why Use Random Variables? (2)

Let's consider a silly experiment:

1. I choose a city at random in Canada, then
2. I choose a local restaurant at random on Yelp, then
3. I choose an item at random on the menu.
4. Let X be the price of this item.

The natural **sample space** is enormous and complicated; the **random variable** is easy to understand.

Notation Caveat

- In the rest of this class, we will define *random variables* without defining *sample spaces.*
- **Example:** I'll say "let X be the result of a fair die roll" - we all know this means $P[X = i] = \frac{1}{6}$ for $i \in \{1, 2, 3, 4, 5, 6\}$.
- **If that doesn't bother you, great! Don't read the rest of this slide!**

Notation Caveat

- In the rest of this class, we will define *random variables* without defining *sample spaces.*
- **Example:** I'll say "let X be the result of a fair die roll" - we all know this means $P[X = i] = \frac{1}{6}$ for $i \in \{1, 2, 3, 4, 5, 6\}$.
- **If that doesn't bother you, great! Don't read the rest of this slide!**
- If it does bother you: yes, we've defined a *function* without writing its *domain* - weird!
- I promise this doesn't get us into trouble.

Miscellany

With the basics out of the way, I introduce two definitions that we'll come back to repeatedly.

Cumulative Distribution Functions

The following will turn out to be a useful way to summarize real-valued random variables:

Cumulative Distribution Function (CDF)

For a random variable X, the CDF is the function

$$F_X(a) = P[X \leq a].$$

Note: the $\leq$ sign is not a $<$ sign.

Cumulative Distribution Functions

Let X be the result of a fair die roll. Then

$$
\begin{aligned}
F_X(a) &= 0, && -\infty < a < 1 \\
F_X(a) &= \frac{1}{6}, && 1 \le a < 2 \\
F_X(a) &= \frac{2}{6}, && 2 \le a < 3 \\
F_X(a) &= \frac{3}{6}, && 3 \le a < 4 \\
F_X(a) &= \frac{4}{6}, && 4 \le a < 5 \\
F_X(a) &= \frac{5}{6}, && 5 \le a < 6 \\
F_X(a) &= 1, && 6 \le a < \infty.
\end{aligned}
$$

Independence (1)

The main ideas are just like independence of sets:

- **Informally:** Random variables are independent if "knowing about one doesn't tell you about the other."
- **Main formula (and definition):** They satisfy

$$P[\{X \in A\} \cap \{Y \in B\}] = P[X \in A]P[Y \in B]$$

  for all sets A, B.
- **Grading:** On an exam, I'll say if random variables are independent.

Independence (2)

Let X, Y be the results of two fair dice rolls. What is the probability that *at least one* is 4 or higher?

$$\begin{aligned} P[\{X \geq 4\} \cup \{Y \geq 4\}] &= P[X \geq 4] + P[Y \geq 4] - P[X, Y \geq 4] \\ &= P[X \geq 4] + P[Y \geq 4] - P[X \geq 4]P[Y \geq 4] \\ &= \frac{1}{2} + \frac{1}{2} - \frac{1}{2}\frac{1}{2} = \frac{3}{4}. \end{aligned}$$

Independence (Full Definition)

Independence

Say that a collection of random variables $X_1, \ldots, X_n$ is independent if

$$P[\{X_1 \in A_1\} \cap \ldots \cap \{X_n \in A_n\}] = P[X_1 \in A_1] \ldots P[X_n \in A_n]$$

for all sequences of sets $A_1, \ldots, A_n$.

Independence (Fun Optional Example)

- Let $X_1, \ldots, Xn$ be independent random variables with $P[X_i = 1] = P[X_i = 0] = 0.5$.
- Let $X_{n+1} = 1$ if $\sum_{i=1}^{n} X_i$ is even, and 0 otherwise.
- **Short exercise:** Show that any size-n subset of $\{X_1, \ldots, X_{n+1}\}$ is independent.
- **Short exercise:** Show that the full set $\{X_1, \ldots, X_{n+1}\}$ is *not* independent.

Note: This is one of the calculations underlying the simplest *error-correcting codes.*

Probability Mass Functions

uOttawa - MAT2377

Fall 2020

Goals

1. Introduce the “probability mass function."
2. TL;DR of lecture: replace $P[X = a]$ by $p_X(a)$.

Formal Definition

Probability Mass Function

The *probability mass function* (or PMF) of a discrete random variable X is defined by $\mathrm{p}_X(a) = P[X = a]$.

Unsurprisingly, this is a lot like the definition of a probability.

PMFs vs Probabilities

Axioms of Probability

1. $P[\Omega] = 1$.
2. For any countable sequence $\{A_i\}_{i \in \mathbb{N}}$ of pairwise mutually exclusive events, $P[\cup_{i \in \mathbb{N}} A_i] = \sum_{i \in \mathbb{N}} P[A_i]$.

'Axioms' of Random Variables

Let X be a random variable with range S. Then

1. $\sum_{s \in S} p_X(s) = 1$.
2. For any subset $A \subset S$, $P[X \in A] = \sum_{s \in A} p_X(s)$.

Before getting to examples...

Again, note that there is nothing really *new* here - similar calculations with slightly-better notation.

Normalizing Constant Example

- Consider the PMF $f_X(x) = c(x+1)^2$, $x \in \{1, 2, 3, 4\}$. What is c?
- We know

$$\begin{aligned} 1 &= \sum_{x=1}^{4} c(x+1)^2 \\ &= c(4+9+16+25) \\ &= 54c, \end{aligned}$$

so $c = \frac{1}{54}$.

Dice Example (1)

- Roll 2 fair 4-sided dice, and let X be the larger number that is rolled. Calculate the PMF of X.
- Let X_1, X_2 be the two die rolls, so that $X = \max(X_1, X_2)$. We calculate

$$
\begin{aligned}
f_X(1) &= P[X_1 = X_2 = 1] = \frac{1}{16}, \\
f_X(2) &= P[X_1 = 1, X_2 = 2] + P[X_1 = 2, X_2 = 1] + P[X_1 = X_2 = 2] \\
&= 3 \times 4^{-2} = \frac{3}{16}, \\
f_X(4) &= P[X_1 = 4] + P[X_2 = 4] - P[X_1 = X_2 = 4] \\
&= \frac{1}{4} + \frac{1}{4} - \frac{1}{16} \\
&= \frac{7}{16}, \\
f_X(3) &= 1 - f_X(1) - f_X(2) - f_X(4) = \frac{5}{16}.
\end{aligned}
$$

Dice Example (2)

- Why don't these calculations all look the same? Pause
- **I chose shorter calculations when I could.**
- If you calculate $f_X(3)$ like $f_X(2)$, get 5 terms

$$f_X(3) = P[(X_1, X_2) \in \{(1,3),(2,3),(3,3),(3,2),(3,1)\}],$$

and $f_X(4)$ has 7. Annoying!

Interlude: PMF and CDF

They are linked by the formula:

$$p_X(a) = F_X(a) - F_X(a-1).$$

This is surprisingly helpful!

Harder Dice Example (1)

- Roll 3 fair 6-sided dice and let X be the smallest number rolled. Calculate the CDF and PMF for X.
- This looks harder than the last question! 3 dice instead of 2, 6-sided instead of 4...
- ... but the CDF will save us!

Harder Dice Example (2)

Let X_1, X_2, X_3 be the results of the three rolls. For $1 \leq i \leq 6$, we have

$$F_X(i) = 1 - P[X_1, X_2, X_3 > i] = 1 - (\frac{7-i}{6})^3.$$

This lets us immediately calculate the PMF. For $1 \leq i \leq 6$,

$$f_X(i) = F_X(i) - F_X(i-1) = (\frac{7-i}{6})^3 - (\frac{6-i}{6})^3.$$

Probability Density Functions

uOttawa - MAT2377

Fall 2020

Goals

1. Introduce Probability Density Functions (PDFs), the analogue of PMFs for continuous random variables.

Technical Issues (1)

- Let's say X is chosen "uniformly at random" on the interval $[0, 1]$.
- What is $P[X \geq 0.5]$?

Technical Issues (1)

- Let's say X is chosen "uniformly at random" on the interval $[0, 1]$.
- What is $P[X \geq 0.5]$?
- What is $P[X > 0.5]$?
- What is $P[X = 0.5]$?

Weird: We have probabilities of *intervals* but not *points*.

Technical Issues (2)

- **Problem:** We can't use formulas like $P[X > 0.5] = \sum_a P[X = a]$, because they don't make sense.
- **Solution:** We'll use integrals instead of sums!
- **Caveat:** There are some lurking paradoxes, but we'll skirt them.

Informal PDF

If X is "uniformly distributed" on $[0, 1]$, we want

$$P[X \in [a, b]] = b - a = \int_a^b 1 dx.$$

Formal PDF

PDF

Say that X has PDF f_X if

$$P[X \in [a, b]] = \int_a^b f_X(x)dx.$$

If X is “uniformly distributed" on $[0, 1]$, we have

$$f_X(x) = 1, \qquad 0 \leq x \leq 1$$
$$f_X(x) = 0, \qquad \text{otherwise.}$$

More Normalizing Constants

$$f(x) = ax^2, \qquad -2 \leq x \leq 4$$
$$g(x) = b\frac{1}{x}, \qquad 1 \leq x \leq 10$$
$$h(x) = c\frac{1}{x^2}, \qquad 1 \leq x < \infty.$$

We have:

$$1 = a\int_{-2}^{4} x^2 = \frac{a}{3}(4^3 - (-2)^3) = 24a$$
$$1 = b\int_{1}^{10} \frac{1}{x}dx = b(\log(10) - \log(1))$$
$$1 = c\int_{1}^{\infty} x^{-2}dx = -c(0 - (1)^{-1}) = c,$$

so

$$a = \frac{1}{24},\ b = \frac{1}{\log(10)},\ c = 1.$$

Uniform Distribution

- Let X be chosen uniformly at random from $[a, b]$. What is its PDF?

Uniform Distribution

- Let X be chosen uniformly at random from $[a, b]$. What is its PDF?
- We don't have a way to calculate PDFs from descriptions! But we guess:

$$f(x) = c\mathbf{1}_{x \in [a,b]}$$

  and can easily compute $c = \frac{1}{b-a}$.
- This works, and becomes our *definition* of the uniform distribution.

CDF and PDF

The CDF has the the same definition as before:

$$F_X(a) = P[X \leq a]$$

and relates to the PDF by:

$$f_X(a) = F'_X(a),\ F_X(a) = \int_{-\infty}^{a} f_X(x)dx.$$

This is surprisingly useful!

Simple CDF Calculation

The CDF for a random variable is

$$\begin{aligned} F(x) &= 0, && x \leq 0 \\ F(x) &= \frac{1}{25}x^2, && 0 \leq x \leq 5, \\ F(x) &= 1, && x \geq 5. \end{aligned}$$

This lets us calculate the PDF. For $0 \leq x \leq 5$,

$$f(x) = F'(x) = \frac{2}{25}x.$$

For $x \notin [0, 5]$, we have $f(x) = 0$.

Interlude: Notation

- We write $X \sim f$ as shorthand for "X has PDF/PMF f."
- We write $X_1, \ldots, X_n \overset{i.i.d.}{\sim} f$ as shorthand for "each X_i has PDF/PMF f, and they are independent."

Harder CDF Calculation

Let $X_1, X_2, X_3, X_4 \overset{i.i.d.}{\sim} \mathit{Unif}[0,1]$ and $Y = \max(X_1, X_2, X_3, X_4)$. We calculate f_Y using the "CDF trick." For $0 \leq x \leq 1$,

$$\begin{aligned} P[Y \leq x] &= P[X_1, X_2, X_3, X_4 \leq x] \\ &= P[X_1 \leq x]P[X_2 \leq x]P[X_3 \leq x]P[X_4 \leq x] \\ &= x^4. \end{aligned}$$

Thus, the PDF of Y is:

$$f_Y(y) = F_Y'(y) = 4x^3 \mathbf{1}_{x \in [0,1]}.$$

Exam Tip

1. The CDF trick is: calculate a complicated PDF by instead calculating a simple CDF, then differentiating.
2. This is very good for functions of random variables, especially min/max/etc.
3. This is one of only two “calculation tricks" in this course, so expect it on an exam!

Multiple Sums - Optional Review

uOttawa - MAT2377

Fall 2020

Goals

1. Remember how “multiple sums" work.

Notation (Square Regions)

1. Remember that $\sum_{a=1}^{5}\sum_{b=1}^{10} f(a,b)$ really means:

$$\sum_{a=1}^{5}\sum_{b=1}^{10} f(a,b) = \sum_{a=1}^{5}(\sum_{b=1}^{10} f(a,b)).$$

2. To emphasize that there is nothing "new" in a double-sum, we can unpack into single sums:

$$\sum_{a=1}^{5}\sum_{b=1}^{10} f(a,b) = \sum_{a=1}^{5} g(a),$$

$$g(a) \equiv \sum_{b=1}^{10} f(a,b).$$

Example (Square Regions)

Let $f(a, b) = ab$. Then

$$\begin{aligned}\sum_{a=1}^{3}\sum_{b=1}^{2} f(a,b) &= \sum_{a=1}^{3}(a+2a) \\ &= 3\sum_{a=1}^{3} a \\ &= 3(1+2+3) = 18.\end{aligned}$$

Notation (Non-Square Regions)

We could imagine writing

$$\sum_{a=1}^{3}\sum_{b=1}^{2} f(a,b) = \sum_{(a,b)\in\{1,2,3\}\times\{1,2\}} f(a,b).$$

- We're just saying which terms to sum over!
- In original double-sum notation, need to sum over a "square" region like $\{1,2,3\} \times \{1,2\}$ - but we can imagine summing over any set R.

Example (Non-Square Regions)

Let $f(a, b) = ab$, and let $R = \{(1, 1), (1, 3), (2, 5), (3, 12)\}$. Then

$$\sum_{(a,b)\in R} f(a, b) =$$

Example (Non-Square Regions)

Let $f(a, b) = ab$, and let $R = \{(1, 1), (1, 3), (2, 5), (3, 12)\}$. Then

$$\sum_{(a,b)\in R} f(a, b) = (1)(1) + (1)(3) + (2)(5) + (3)(12) = 50.$$

Joint Distributions

uOttawa - MAT2377

Fall 2020

Goals

1. Set up notation for PDFs and PMFs of dependant random variables.

Definition

The **joint probability mass function** (also called the joint distribution) of X and Y is

$$p_{XY}(x, y) = P(X = x, Y = y) = P(\{X = x\} \cap \{Y = y\}).$$

The **range** of the random vector (X, Y) is

$$R_{XY} = \{(x, y) \ : \ p_{XY}(x, y) \neq 0\}.$$

Some Properties

1. (non-negative probability)

$$p_{XY}(x, y) \geq 0$$

2. (total mass =1)

$$\sum_{(x,y) \in R_{XY}} p_{XY}(x, y) = 1$$

3. (addition property)

$$P((X, Y) \in A) = \sum_{(x,y) \in A \cap R_{XY}} p_{XY}(x, y)$$

Marginal PMF

The **marginal probability mass functions** of X and Y are respectively

$$p_X(x) = P(X = x) = \sum_y p_{XY}(x, y)$$

$$p_Y(y) = P(Y = y) = \sum_x p_{XY}(x, y).$$

Independence:

We say that X and Y are **independent** if

$$p_{XY}(x, y) = p_X(x)\, p_Y(y)$$

for all x, y.

Extended Example

Consider the following joint probability mass function:

x	y	$p_{XY}(x, y)$
1	1	1/30
1	2	5/30
1	3	6/30
2	1	2/30
2	2	10/30
3	1	6/30

$$P(X < 2, Y < 2) =$$

Extended Example

Consider the following joint probability mass function:

x	y	$p_{XY}(x,y)$
1	1	1/30
1	2	5/30
1	3	6/30
2	1	2/30
2	2	10/30
3	1	6/30

$$P(X=1) = p_{XY}(1,1)+p_{XY}(1,2)+p_{XY}(1,3) = \frac{1}{30}+\frac{5}{30}+\frac{6}{30} = \frac{12}{30}$$

Extended Example

x	y	$p_{XY}(x,y)$
1	1	1/30
1	2	5/30
1	3	6/30
2	1	2/30
2	2	10/30
3	1	6/30

$$\begin{aligned} P(X > 1, Y \leq 2) &= p_{XY}(2,1) + p_{XY}(2,2) + p_{XY}(3,1) \\ &= \frac{2}{30} + \frac{10}{30} + \frac{6}{30} = \frac{18}{30} \end{aligned}$$

Extended Example

x	y	$p_{XY}(x,y)$
1	1	1/30
1	2	5/30
1	3	6/30
2	1	2/30
2	2	10/30
3	1	6/30

$$P(Y > 2|X = 1) = \frac{P(\{Y > 2\} \cap \{X = 1\})}{P(X = 1)}$$
$$= \frac{p_{XY}(1,3)}{12/30} = \frac{6}{12}$$

Extended Example

x	y	$p_{XY}(x,y)$
1	1	1/30
1	2	5/30
1	3	6/30
2	1	2/30
2	2	10/30
3	1	6/30

The marginal distribution of X is:

$$p_X(x) = \begin{cases} 1/30 + 5/30 + 6/30; & x = 1 \\ 2/30 + 10/30; & x = 2 \\ 6/30; & x = 3 \end{cases}$$

Extended Example

x	y	$p_{XY}(x, y)$
1	1	1/30
1	2	5/30
1	3	6/30
2	1	2/30
2	2	10/30
3	1	6/30

Are X and Y independent?

$$p_X(3) \times p_Y(3) = \frac{6}{30} \times \frac{6}{30}$$

but

$$P_{XY}(3, 3) = 0 \neq p_X(3)p_Y(3).$$

Note: We just needed to find *one* failure!

Means of Random Variables

uOttawa - MAT2377

Fall 2020

Goals

1. Start discussing “typical" results of experiments.

Gambling Example (1)

Let's consider a game with payoff given by the distribution:

$$p_X(4) = 0.5$$
$$p_X(-2) = 0.5.$$

The *average* winnings are:

$$\text{Avg} = \frac{1}{2}4 + \frac{1}{2}(-2) = 1.$$

You might *expect* that the amount you win after n games is about

$$\text{Win}(n) \approx n \times \text{Avg} = n$$

for n large.

Gambling Example (2)

The average isn't everything. Consider another game with payoff:

$$p_Y(10^{12}) = 10^{-6}$$
$$p_Y(-10^6) = 1 - 10^{-6} \approx 0.999999.$$

The mean is still 1, and you would expect to win about n dollars over n games - but you would usually lose about a million dollars before getting *anything* back.

Formal Definition

Expectation

For a discrete random variable X with PMF p_X, the expectation is:

$$E[X] = \sum_x x p_X(x).$$

For a continuous random variable with PDF f_X, it is:

$$E[X] = \int x f_X(x) dx.$$

Simple Example (1)

- Let X have PMF $p_X(x) = \frac{x}{14}$, $x \in \{2, 3, 4, 5\}$. Find $E[X]$.
- We calculate

$$\begin{aligned} E[X] &= \sum_{x=2}^{5} x p_X(x) \\ &= \frac{1}{14} \sum_{x=2}^{5} x^2 \\ &= \frac{1}{14}(4 + 9 + 16 + 25) \\ &= \frac{54}{14} \approx 3.86. \end{aligned}$$

Simple Example (2)

- Let X have PDF $f_X(x) = \frac{x}{16}$, $x \in [2, 6]$. Find $E[X]$.
- We calculate

$$\begin{aligned} E[X] &= \int_2^6 x f_X(x) \\ &= \frac{1}{16} \int_2^6 x^2 dx \\ &= \frac{1}{16} \frac{1}{3} (216 - 8) \\ &= \frac{208}{48} \approx 4.33. \end{aligned}$$

Longer Example

- You pay 1 dollar to play a game, and roll two dice. If the sum of the dice is 8 or more, you get 2 dollars. Otherwise, you get nothing. Calculate the expected payout.
- Let X be the expected payout and let Y be the sum of two dice. We note $P[Y = 7] = \frac{1}{6}$, and $P[Y \geq 8] = P[Y \leq 6]$. Thus,

$$\begin{aligned} P[Y \geq 8] &= \frac{1}{2}(1 - P[Y = 7]) \\ &= \frac{1}{2}\frac{5}{6} = \frac{5}{12}. \end{aligned}$$

Thus,

$$\begin{aligned} E[X] &= 1P[Y \geq 8] - 1P[Y \leq 7] \\ &= \frac{5}{12} - \frac{7}{12} = \frac{-1}{6}. \end{aligned}$$

Symmetric Example

- You play the following poker-like game with a friend. You each draw a hand of 5 cards. If you have a better poker hand than your friend, you gain a dollar. If your friend has a better hand, you lose a dollar. What is your expected payout?
- By symmetry, the expected payout must be exactly 0.

Variances and Higher Moments

uOttawa - MAT2377

Fall 2020

Goals

1. Discuss how large “typical" deviations are.

Expectations and Functions

- If $X_1, \ldots, X_n$ are random variables, so is any function $f(X_1, \ldots, X_n)$ of them! So the definition "makes sense" as written.
- There is a nice trick for calculating expectations of functions: if $Y = h(X)$,

$$E[Y] = \int y f_Y(y) dy = \int h(x) f_X(x) dx;$$

analogous formula holds for discrete random variables.

Expectations and Functions (Example)

Let $X \sim \mathrm{Unif}[0, 1]$ and $Y = X^2$. Then

$$E[Y] =$$

Moments

For $k \in \mathbb{N}$,

- The expected value $E[X^k]$ is the k'th *moment* of X. $k = 1$ has the special name *mean*.
- The expected value $E[(X - E[X])^k]$ is the k'th *central moment* of X. $k = 2$ has the special name *variance*.

Variances are almost always calculated with the special formula:

$$\mathrm{Var}[X] = E[(X - E[X])^2] = E[X^2] - E[X]^2.$$

Heuristic: Variance Describes "Typical" Fluctuations

Define the *standard deviation*

$$\sigma = \sqrt{\mathrm{Var}[X]}.$$

For "nice" random variables X,

$$E[|X - E[X]|] \approx \sigma.$$

Note: This *can* fail miserably, but statisticians often have this in the back of their mind when discussing standard deviations.

Gambling Revisited

Last video, we considered the random variables:

$$f_X(M^2) = \frac{1}{M}$$
$$f_X(-M) = 1 - \frac{1}{M}.$$

for $M \in [2, \infty)$. The mean and variance are:

Gambling Revisited

Last video, we considered the random variables:

$$f_X(M^2) = \frac{1}{M}$$
$$f_X(-M) = 1 - \frac{1}{M}.$$

for $M \in [2, \infty)$. The mean and variance are:

$$E[X] = M^2 M^{-1} + (-M)(1 - M^{-1}) = M - M + 1 = 1.$$

However,

$$E[X^2] = M^4 M^{-1} + (M^2)(1 - M^{-1}) = M^3 + M^2 - M,$$

so

$$\mathrm{Var}[X] = E[X^2] - E[X]^2 = M^3 + M^2 - M - 1.$$

Mean is constant but *fluctuations* grow quickly!

Moment Formulas and Tricks

uOttawa - MAT2377

Fall 2020

Goals

1. State and use special “tricks" used in calculating expectations.
2. **Note:** You typically need to practice a little to get used to these!

Expectations and Functions (Review)

- If $X_1, \ldots, X_n$ are random variables, so is any function $f(X_1, \ldots, X_n)$ of them! So the definition "makes sense" as written.
- There is a nice trick for calculating expectations of functions: if $Y = h(X)$,

$$E[Y] = \int y f_Y(y) dy = \int h(x) f_X(x) dx;$$

analogous formula holds for discrete random variables.

Expectations and Functions (Example)

Let $X \sim \mathrm{Unif}[0, 1]$ and $Y = X^2$. Then

$$E[Y] =$$

Variance Formula (Review)

Recall

$$Var[X] = E[(X - E[X])^2] = E[X^2] - E[X]^2.$$

Variance Formula (Application)

Let $X \sim \mathrm{Unif}[0,1]$. Then

$$E[X] = \int_0^1 x dx = \frac{1}{2}$$
$$E[X^2] = \int_0^1 x^2 dx = \frac{1}{3},$$

so

$$Var[X] = E[X^2] - E[X]^2 = \frac{1}{3} - \frac{1}{4} = \frac{1}{12}.$$

Expectations and Integration-by-Parts

Integration by Parts Formula

Let $X \geq 0$ be a discrete random variable. Then

$$E[X] = \sum_{x=1}^{\infty} P[X \geq x].$$

Note: The proof of this identity is quite similar to the more familiar integration-by-parts formula

$$\int udv = uv - \int vdu.$$

Expectations and Integration-by-Parts (Example)

- Flip a fair coin, and let X be the number of coins flipped until your first head. What is $E[X]$?
- Let R_i be 1 if the i'th flip is tails, 0 otherwise. We have

$$\begin{aligned} P[X \geq i] &= P[R_1 = R_2 = \ldots = R_{i-1} = 1] \\ &= 2^{-(i-1)}. \end{aligned}$$

By the integration-by-parts formula,

$$\begin{aligned} E[X] &= \sum_{i=1}^{\infty} 2^{-(i-1)} \\ &= \sum_{i=0}^{\infty} 2^{-i} = 2. \end{aligned}$$

Expectations and Integration-by-Parts (Example)

This example is hard without our trick! We would get

$$\begin{aligned} E[X] &= \sum_{i=1}^{\infty} P[X = i] \\ &= \sum_{i=1}^{\infty} i 2^{-i}. \end{aligned}$$

Expectations and Linearity

Linearity of Expectation

Let X, Y be two random variables. Then

$$E[X + Y] = E[X] + E[Y].$$

Note: I don't assume independence!

Expectations and Linearity (Example)

- Let X be a random variable with $E[X] = 3$ and $E[X^2] = 22$. Calculate $E[(2X - 7)^2]$.
- Using linearity,

$$\begin{aligned} E[(2X-7)^2] &= E[4X^2 - 28X + 49] \\ &= 4E[X^2] - 28E[X] + 49 \\ &= (4)(22) - (28)(3) + 49 = 53. \end{aligned}$$

Expectations and Independence

If X, Y are independent,

$$E[XY] = E[X]E[Y].$$

Note: The converse is not true!

Summative Example (1)

Let $X_1, \ldots, X_n \overset{i.i.d.}{\sim} b$, where b is the PMF

$$b(1) = b(0) = \frac{1}{2}.$$

Let $X = \sum_{i=1}^{n} X_i$. Then

$$E[X] = \sum_{i=1}^{n} E[X_i] = \frac{n}{2}.$$

Summative Example (2)

Similarly,

$$
\begin{aligned}
E[X^2] &= \sum_{i=1}^{n} E[X_i^2] + \sum_{1 \le i \ne j \le n} E[X_i X_j] \\
&= \sum_{i=1}^{n} \frac{1}{2} + \sum_{1 \le i \ne j \le n} E[X_i]E[X_j] \\
&= \sum_{i=1}^{n} \frac{1}{2} + \sum_{1 \le i \ne j \le n} \frac{1}{4} \\
&= \frac{n}{2} + \frac{n(n-1)}{4}.
\end{aligned}
$$

Summative Example (3)

Finally,

$$\begin{aligned} Var[X] &= E[X^2] - E[X]^2 \\ &= \frac{n^2 - n}{4} + \frac{n}{2} - \frac{n^2}{4} \\ &= \frac{n}{4}. \end{aligned}$$

Reference: Sum Formulas

The following formulas are analogues to the familiar formulas for $\int_0^n x^k dx$, $k = 0, 1, 2, 3$.

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

$$\sum_{i=1}^{n} i = \frac{n(n+1)}{2} \approx \frac{n^2}{2}$$

$$\sum_{i=1}^{n} i^2 = \frac{n(n+1)(2n+1)}{6} \approx \frac{n^3}{3}$$

$$\sum_{i=1}^{n} i^3 = (\frac{n(n+1)}{2})^2 \approx \frac{n^4}{4}.$$

No need to memorize, but useful to remember these exist and are similar to the formulas from calculus.

Binomial

uOttawa - MAT2377

Fall 2020

Goals

1. Introduce the “binomial" random variable and “Bernoulli trials."
2. More generally: this is the first in a sequence of *named* random variables.

Review: Combination

In general, if you randomly choose k objects out of n objects without replacement ($k \leq n$), the total number of outcomes (regardless of order) is given by

$$\binom{n}{k} = \frac{n \cdot (n-1) \cdot (n-2) \cdot \ldots \cdot (n-k+1)}{k!},$$

where

- $k! = k \cdot (k-1) \cdot (k-2) \cdot \ldots \cdot 3 \cdot 2 \cdot 1$ is called **k factorial**.
- $\binom{n}{k}$ is called **n choose k**.

Bernoulli trials example (1)

There are 3 white balls and 7 black balls in a bag. Someone randomly picks a ball *with replacement* 4 times. What is the probability for him to see a white ball $0, 1, 2, 3,$ or 4 times?

Bernoulli trials example (1)

There are 3 white balls and 7 black balls in a bag. Someone randomly picks a ball *with replacement* 4 times. What is the probability for him to see a white ball $0, 1, 2, 3,$ or 4 times?

- Let X denote the number of times he see a white ball. X is a (discrete) random variable which takes value in $\{0, 1, 2, 3, 4\}$. It is also reasonable to assume the outcomes of his draws are **independent**.

Bernoulli trials example (1)

There are 3 white balls and 7 black balls in a bag. Someone randomly picks a ball *with replacement* 4 times. What is the probability for him to see a white ball $0, 1, 2, 3,$ or 4 times?

- Let X denote the number of times he see a white ball. X is a (discrete) random variable which takes value in $\{0, 1, 2, 3, 4\}$. It is also reasonable to assume the outcomes of his draws are **independent**.
- If $X = 0$, he sees no white ball in all 4 draws.

$$\begin{aligned} P(X = 0) &= P(1B \cap 2B \cap 3B \cap 4B) = P(1B)P(2B)P(3B)P(4B) \\ &= (0.7)(0.7)(0.7)(0.7) = \binom{4}{0} 0.7^4 = 0.2401 \end{aligned}$$

Bernoulli trials example (1)

There are 3 white balls and 7 black balls in a bag. Someone randomly picks a ball *with replacement* 4 times. What is the probability for him to see a white ball $0, 1, 2, 3,$ or 4 times?

- Let X denote the number of times he see a white ball. X is a (discrete) random variable which takes value in $\{0, 1, 2, 3, 4\}$. It is also reasonable to assume the outcomes of his draws are **independent**.
- If $X = 0$, he sees no white ball in all 4 draws.

$$\begin{aligned} P(X = 0) &= P(1B \cap 2B \cap 3B \cap 4B) = P(1B)P(2B)P(3B)P(4B) \\ &= (0.7)(0.7)(0.7)(0.7) = \binom{4}{0} 0.7^4 = 0.2401 \end{aligned}$$

- If $X = 1$, he sees a white ball in only one of the 4 draws.

$$\begin{aligned} P(X = 1) &= P(1W \cap 2B \cap 3B \cap 4B) + P(1B \cap 2W \cap 3B \cap 4B) \\ &\quad + P(1B \cap 2B \cap 3W \cap 4B) + P(1B \cap 2B \cap 3B \cap 4W) \\ &= 4(0.3)(0.7)(0.7)(0.7) = \binom{4}{1} (0.3)(0.7)^3 = 0.4116 \end{aligned}$$

Bernoulli trials example (2)

- If $X = 2$, he sees a white ball in only two of the 4 draws.

$$\begin{aligned} P(X=2) &= P(1W \cap 2W \cap 3B \cap 4B) + P(1W \cap 2B \cap 3W \cap 4B) \\ &\quad + P(1W \cap 2B \cap 3B \cap 4W) + P(1B \cap 2W \cap 3W \cap 4B) \\ &\quad + P(1B \cap 2W \cap 3B \cap 4W) + P(1B \cap 2B \cap 3W \cap 4W) \\ &= 6(0.3)(0.3)(0.7)(0.7) = \binom{4}{2}(0.3)^2(0.7)^2 = 0.2646 \end{aligned}$$

Bernoulli trials example (2)

- If $X = 2$, he sees a white ball in only two of the 4 draws.

$$\begin{aligned} P(X=2) &= P(1W \cap 2W \cap 3B \cap 4B) + P(1W \cap 2B \cap 3W \cap 4B) \\ &\quad + P(1W \cap 2B \cap 3B \cap 4W) + P(1B \cap 2W \cap 3W \cap 4B) \\ &\quad + P(1B \cap 2W \cap 3B \cap 4W) + P(1B \cap 2B \cap 3W \cap 4W) \\ &= 6(0.3)(0.3)(0.7)(0.7) = \binom{4}{2}(0.3)^2(0.7)^2 = 0.2646 \end{aligned}$$

- If $X = 3$, he sees a black ball in only one of the 4 draws.

$$\begin{aligned} P(X=3) &= P(1W \cap 2W \cap 3W \cap 4B) + P(1W \cap 2W \cap 3B \cap 4W) \\ &\quad + P(1W \cap 2B \cap 3W \cap 4W) + P(1B \cap 2W \cap 3W \cap 4W) \\ &= 4(0.7)(0.3)(0.3)(0.3) = \binom{4}{3}(0.3)^3(0.7) = 0.0756 \end{aligned}$$

Bernoulli trials example (3)

- If $X = 4$, he sees a white ball in all 4 draws.

$$\begin{aligned} P(X = 4) &= P(1W \cap 2W \cap 3W \cap 4W) \\ &= P(1W)P(2W)P(3W)P(4W) \\ &= (0.3)(0.3)(0.3)(0.3) = \binom{4}{4} 0.3^4 = 0.0081 \end{aligned}$$

Bernoulli and Binomials

- Fix $p \in [0,1]$. Say X has *Bernoulli distribution with parameter* p if
$$P[X = 1] = p = 1 - P[X = 0].$$
An i.i.d. sequence $X_1, X_2, \ldots$ is a *Bernoulli process with parameter* p.
- Fix $n \in \mathbb{N}$. Then $Y = \sum_{i=1}^{n} X_i$ has *binomial distribution with parameter* p *and* n *trials.* Write $Y \sim Bin(n, p)$.
- By previous example,
$$P[Y = a] = \binom{n}{a} p^a (1-p)^{n-a}, \qquad a \in \{0, 1, 2, \ldots, n\}.$$

Binomial distribution: expectation and variance

We calculated last class:

- Expectation:

$$E[X] = E[\sum_{i=1}^{n} X_i] = \sum_{i=1}^{n} E[X_i] = np.$$

- Variance:

$$Var[X] = np(1-p).$$

Example (1)

Suppose a certain surgery has a success rate at 80%. A group of 6 patients are scheduled to have this surgery next month. Assume the outcomes for different patients are independent. Let X be the number of patients who receive a successful surgery. (a) What is the probability that $X = 3$? (b) What is the probability that $X \geq 2$? (c) What is the expectation and variance of X?

- From the question, what distribution does X follow?
- (a) $P(X = 3) =$

Example (2)

- (b)

$$\begin{aligned} P(X \geq 2) &= 1 - P(X < 2) \\ &= 1 - (P(X = 0) + P(X = 1)) \\ &= \end{aligned}$$

Example (2)

- (b)

$$\begin{aligned} P(X \geq 2) &= 1 - P(X < 2) \\ &= 1 - (P(X = 0) + P(X = 1)) \\ &= \end{aligned}$$

- (c)

$$\begin{aligned} E(X) &= \\ \text{Var}(X) &= \end{aligned}$$

Common Jargon

- In binomial distribution, we say $X_i = 1$ corresponds to the "success" of trial i. A "success" can refer to any event of interest - it need not be "good" in any obvious sense!

Common Jargon

- In binomial distribution, we say $X_i = 1$ corresponds to the "success" of trial i. A "success" can refer to any event of interest - it need not be "good" in any obvious sense!
  - When you toss a coin, success may mean that you get a head (or tail).
  - When you randomly select an Ottawa resident, success may mean that resident speaks French.

Summary

Let X denotes the number of successes in a sequence of n independent yes/no experiments, each of which yields success with probability p. Then,

$$X \sim \text{Binomial}(n, p)$$

- For $0 \leq k \leq n$,

$$P(X = k) = \binom{n}{k} p^k (1-p)^{n-k}$$

-

$$E(X) = np, \qquad Var(X) = np(1-p)$$

Multinomial

uOttawa - MAT2377

Fall 2020

Goals

1. Binomials describe the following situation: we do n identical experiments, each of which have *two* possible outcomes.
2. Multinomials describe the following situation: we do n identical experiments, each of which have k possible outcomes for some $k > 2$.
3. **Note:** formulas get longer, but the situation is not any more complicated!

Review: Binomial

1. Let $X_1, X_2, \ldots$ be i.i.d., with
$$P[X_i = 1] = p = 1 - P[X_i = 0].$$
2. We say that $X = \sum_{i=1}^{n} X_i$ is *binomial* with parameter p and n trials.
3. **Silly observation:** $X = |\{i \in \{1, 2, \ldots, n\} : X_i = 1\}|$.

Rephrased: Binomial

1. Let $X_1, X_2, \ldots$ be i.i.d., with

$$P[X_i = 1] = p_1, P[X_i = 2] = p_2 = 1 - p_1.$$

2. Write $Y_m = |\{i \in \{1, 2, \ldots, n\} : X_i = m\}|$, $m \in \{1, 2\}$.
3. We say that (Y_1, Y_2) is *binomial* with parameter p_1 and n trials; only Y_1 matters.
4. The joint PMF is:

$$P[Y_1 = y_1, Y_2 = y_2] = P[Y_1 = y_1] = \frac{n!}{y_1! y_2!} p_1^{y_1} p_2^{y_2}$$

Extension: Multinomial

1. Let $X_1, X_2, \ldots$ be i.i.d., with

$$P[X_i = 1] = p_1, P[X_i = 2] = p_2, \ldots, P[X_i = k] = p_k.$$

2. Write $Y_m = |\{i \in \{1, 2, \ldots, n\} : X_i = m\}|$.
3. We say that $(Y_1, \ldots, Y_k)$ is *multinomial* with parameter $(p_1, \ldots, p_k)$ and n trials.
4. The joint PMF is:

$$P[Y_1 = y_1, \ldots, Y_k = y_k] = \frac{n!}{y_1! \ldots y_k!} p_1^{y_1} \ldots p_k^{y_k}.$$

Example

In an election year, parties A, B, C are supported by 45, 35 and 20 percent of the population respectively. What is the probability that a sample of 100 people has exactly $n_A = 45, n_B = 35, n_C = 20$ supporters of these parties?

$$P[n_A = 45, n_B = 35, n_C = 20] = \frac{100!}{45!35!20!} 0.45^{45} 0.35^{35} 0.2^{20} = 0.0089.$$

Hypergeometric

uOttawa - MAT2377

Fall 2020

Goals

1. Describe the *hypergeometric* distribution, another close cousin to binomials.

Review: Binomial

1. Let $X_1, X_2, \ldots$ be i.i.d., with

$$P[X_i = 1] = p = 1 - P[X_i = 0].$$

2. We say that $X = \sum_{i=1}^{n} X_i$ is *binomial* with parameter p and n trials.

Rephrased: Binomial

1. Consider a bin with k red balls and m blue balls.
2. Draw a sequence of balls at random *with replacement*, and let $X_i = 1$ if the i'th ball is red ($X_i = 0$ otherwise).
3. $X = \sum_{i=1}^{n} X_i$ is binomial with parameter $p = \frac{k}{k+m}$ and n trials.

Analogue: Hypergeometric

1. Consider a bin with k red balls and m blue balls.
2. Draw a sequence of balls at random without replacement, and let $X_i = 1$ if the i'th ball is red ($X_i = 0$ otherwise).
3. $X = \sum_{i=1}^{n} X_i$ is hypergeometric with parameter $p = \frac{k}{k+m}$ and n trials.
4. We've already seen this process! PMF is:

$$p_X(x) = \frac{\binom{k}{x}\binom{m}{n-x}}{\binom{k+m}{n}}$$

Statistics

1. For binomial,

$$E[X] = np, \qquad Var[X] = np(1-p)$$

2. For hypergeometric,

$$E[X] = np, \qquad Var[X] = np(1-p)\frac{k+m-n}{k+m-1}.$$

Example

I have a bin with 10 red balls and 25 blue balls. I sample 5 without replacement; what is the *variance* in the number of red balls that I've drawn?

$$\begin{aligned} Var[X] &= np(1-p)\frac{k+m-n}{k+m-1} \\ &= n\frac{k}{k+m}\frac{m}{k+m}\frac{k+m-n}{k+m-1} \\ &= \frac{10}{35}\frac{25}{35}\frac{35-5}{35-1}. \end{aligned}$$

Big Picture: Studying Named Random Variables

The real work in these problems is realizing which distribution is being described. Some hypergeometric examples:

1. I have a bin with red and blue balls...
2. I have a class with science students and engineering students...
3. I have a box with defective and non-defective widgets...

and sample *without replacement.*

Geometric

uOttawa - MAT2377

Fall 2020

Goals

1. Describe the *geometric* distribution, another close cousin to binomials.

Review: Binomial

1. Let $X_1, X_2, \ldots$ be i.i.d., with

$$P[X_i = 1] = p = 1 - P[X_i = 0].$$

2. We say that $X = \sum_{i=1}^{n} X_i$ is *binomial* with parameter p and n trials.

New: Geometric

1. Let $X_1, X_2, \ldots$ be i.i.d., with

$$P[X_i = 1] = p = 1 - P[X_i = 0].$$

2. We say that $\tau = \min\{i : X_i = 1\}$ is *geometric* with parameter p.

Geometric PMF and Statistics

1. PMF is simple exercise:

$$p_X(x) = (1-p)^{x-1}p.$$

2. We have computed expectation in previous slides (see integration-by-parts trick):

$$E[X] = p^{-1}.$$

3. Variance is more difficult, but also computable by integration-by-parts trick:

$$Var[X] = \frac{1-p}{p^2}.$$

Example

Assume that a bus arrives in any given minute with probability 0.06, independently. What is the expected time τ to wait for a bus? What is the probability that I will need to wait at least an hour?

$$E[\tau] =$$

$$P[\tau > 60] = P[X_1 = \ldots = X_{60} = 0] =$$

Big Picture

- Geometric is easy to remember - so far, it is the only *unbounded* random variable!
- **NB** weird naming conventions: hypergeometric is like binomial without replacement; negative binomial is like geometric without replacement.

Poisson

uOttawa - MAT2377

Fall 2020

Goals

1. Describe the *Poisson* distribution, another close cousin of the binomial.

Review: Binomial

1. Let $X_1, X_2, \ldots$ be i.i.d., with

$$P[X_i = 1] = p = 1 - P[X_i = 0].$$

2. We say that $X = \sum_{i=1}^{n} X_i$ is *binomial* with parameter p and n trials.

A Thought Experiment (1)

Let's look at incoming requests to IT helpdesk. We model this as binomial, with n being the number of customers and p being the probability of calls in any given day. What's the difference between the following situations?

1. $n = 1000$ clients, $p = 0.01$ of calling.
2. $n = 10000$ clients, $p = 0.001$ of calling.
3. $n = 100000$ clients, $p = 0.0001$ of calling.

All have 10 calls per day on average, with very similar probabilities of $0, 1, \ldots, 100$ calls.

A Thought Experiment (2)

- It is *almost impossible* to tell the difference between these situations by looking at the number of calls.
- We only care about the number of calls, and sometimes *don't even know* exactly what n is.

Poisson Definition

Let $X_n \sim Binom(n, \frac{\lambda}{n})$ and define

$$p_X(x) = \lim_{n \to \infty} p_{X_n}(x) = \frac{\lambda^x e^{-\lambda}}{x!}$$

to be the PMF of the *Poisson* distribution.

Poisson Statistics

Let $X \sim Pois(\lambda)$. Then

$$E[X] = \lambda, \qquad Var[X] = \lambda.$$

Poisson Parameters

1. Poisson is determined by *one parameter*, the expected number of occurrences.
2. *Whenever* we discuss events occurring "at some rate," we are implicitly talking about the Poisson random variable.

Poisson Example

- A call center receives calls at an average rate of 4 per hour. What is the probability that the call center receives *no* calls between 10 AM and 11:30 AM?

Poisson Example

- A call center receives calls at an average rate of 4 per hour. What is the probability that the call center receives *no* calls between 10 AM and 11:30 AM?
- Recognize this is discussing a Poisson random variable, with mean
$$\lambda = (4/hour) \times (1.5 hours) = 6.$$
- Use the formula:
$$P[X = 0] = \frac{\lambda^0 e^{-\lambda}}{0!} = e^{-6}.$$

Uniform

uOttawa - MAT2377

Fall 2020

Goals

1. Describe the *uniform* distribution.

Paired Definitions

1. For a *finite* set $S = \{a, a+1, \ldots, b\}$, the *uniform distribution on* S has PMF
$$p_X(x) = \frac{1}{b-a+1}, \qquad x \in S.$$
2. For an *interval* $S = [a, b]$, the *uniform distribution on* S has PDF
$$f_X(x) = \frac{1}{b-a}, \qquad x \in S.$$

Paired Statistics

1. For a *finite* set $S = \{a, a+1, \ldots, b\}$, the *uniform distribution on* S has statistics

$$E[X] = \frac{a+b}{2}, \qquad Var[X] = \frac{(b-a+1)^2 - 1}{12}.$$

2. For an *interval* $S = [a, b]$, the *uniform distribution on* S has statistics

$$E[X] = \frac{a+b}{2}, \qquad Var[X] = \frac{(b-a)^2}{12}.$$

Example

Just plug in! No new types of questions here - just a new definition.

Exponential

uOttawa - MAT2377

Fall 2020

Goals

Describe the *exponential* distribution, the continuous analogue to the geometric.

Exponential Definition

The exponential distribution with parameter λ has PDF

$$f_X(x) = \lambda e^{-\lambda x}$$

and statistics

$$E[X] = \lambda^{-1}, \qquad Var[X] = \lambda^{-2}.$$

Natural question: What is this and who cares?

Review: Bernoulli, Binomial and Geometric

Fix $p \in [0,1]$ and let $X_1, X_2, \ldots \overset{i.i.d.}{\sim} \textit{Bern}(p)$. Then:

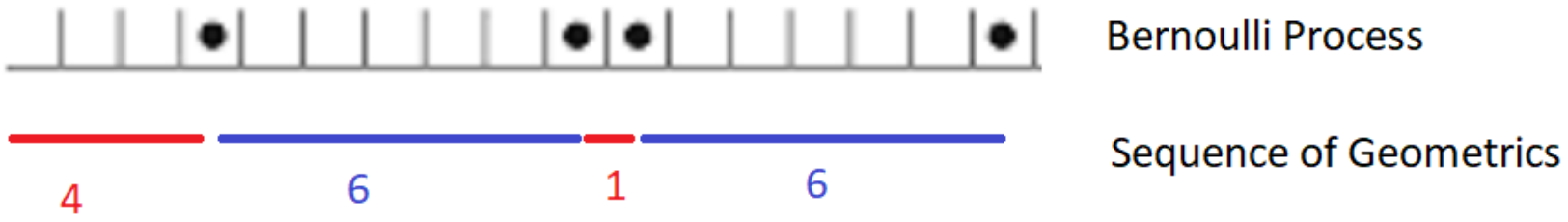
1. $X_1, X_2, \ldots$ is a *Bernoulli trial*.
2. $\sum_{i=1}^{n} X_i$ is *Binomial*(n, p).
3. $\tau = \min\{i : X_i = 1\}$ is *Geometric*(p^{-1}).

Interpretation: Something happens with an underlying rate; *geometric* measures time between successes, *binomial* measures number of successes by time n.

Binomial and Geometric to Poisson and Exponential

- *Discrete Interpretation:* Something happens with an underlying rate; *geometric* measures time between successes, *binomial* measures number of successes by time n.
- *Continuous Interpretation:* Something happens with an underlying rate; *exponential* measures time between successes, *Poisson* measures number of successes by time n.

Bernoulli, Binomial and Geometric in Pictures



Total: 4 Successes on 17 trials is *binomial(n,?)*.

Poisson and Exponential in Pictures

Poisson

~~Bernoulli~~ Process

Sequence of ~~Geometrics~~

Exponentials

Total: 4 Successes ~~on 17 trials~~ is ~~*binomial(n,?)*~~.

Poisson(?)

Note: Of course, the exponentials shouldn't *really* be integers.

Example

- A call center opens at 9 AM and receives calls at a rate of 4 per hour. What is the expected time of the first call?
- We note that the first call X is exponential with parameter $\lambda = 4$. Thus,

$$E[X] = \lambda^{-1} = \frac{1}{4}.$$

Caveat: Exponential random variables have units - in this case, hours!

Gaussians and Binomial Approximations

uOttawa - MAT2377

Fall 2020

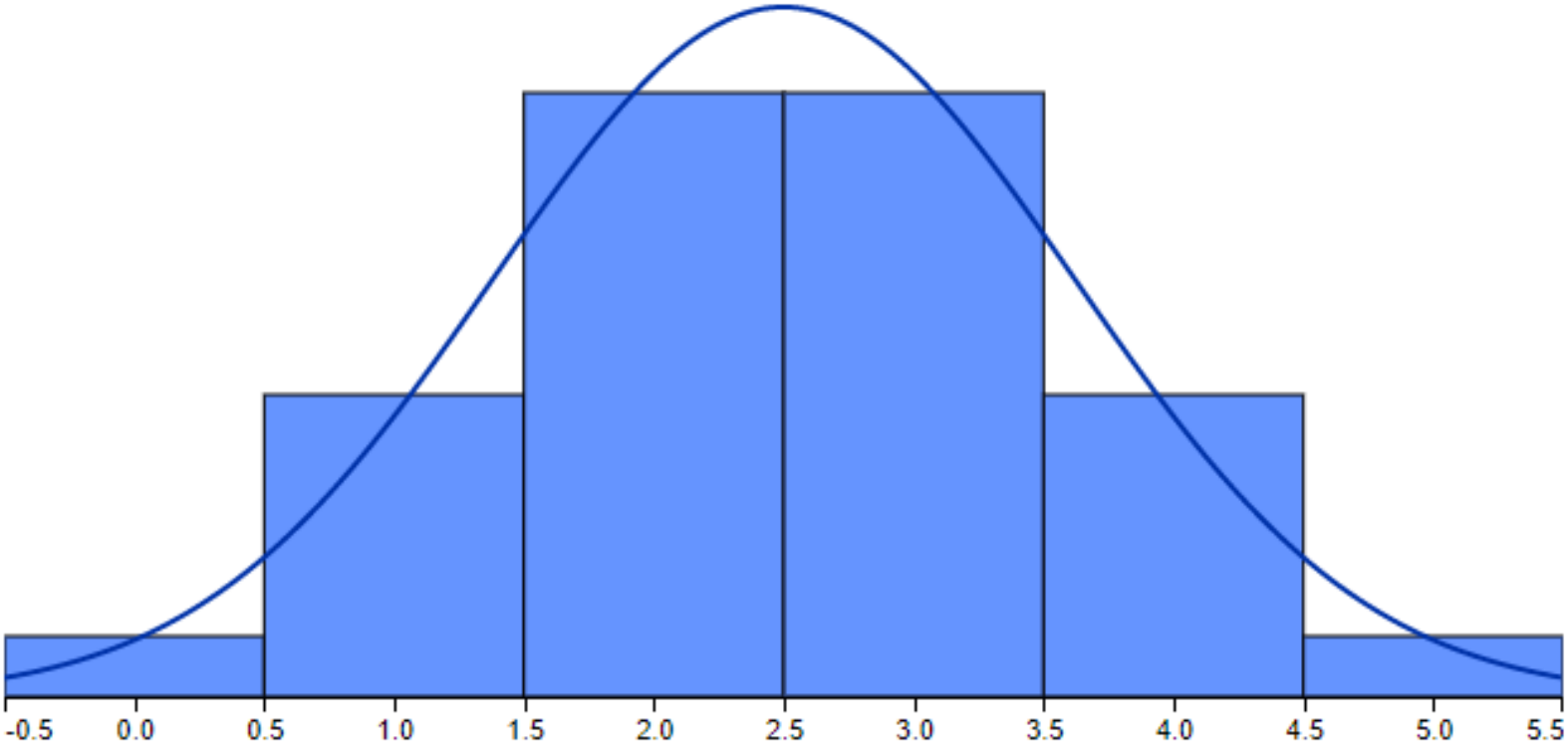
Goals

Introduce the *Gaussian* distribution, which describes all “nice" and “large" experiments.

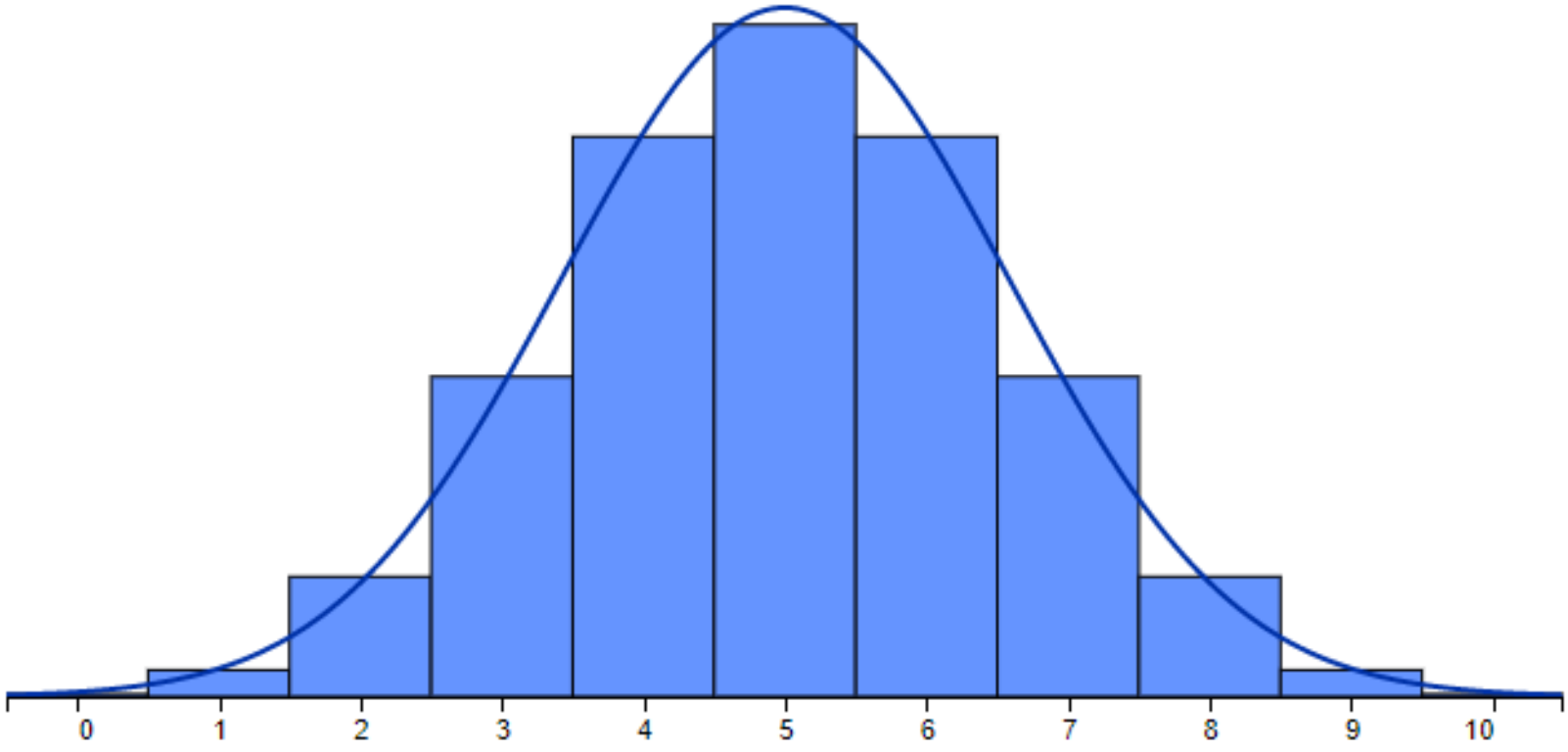
Let's do Gaussians by Pictures

Before *defining* the Gaussian, I'll plot its PDF vs. the Binomial PMF...

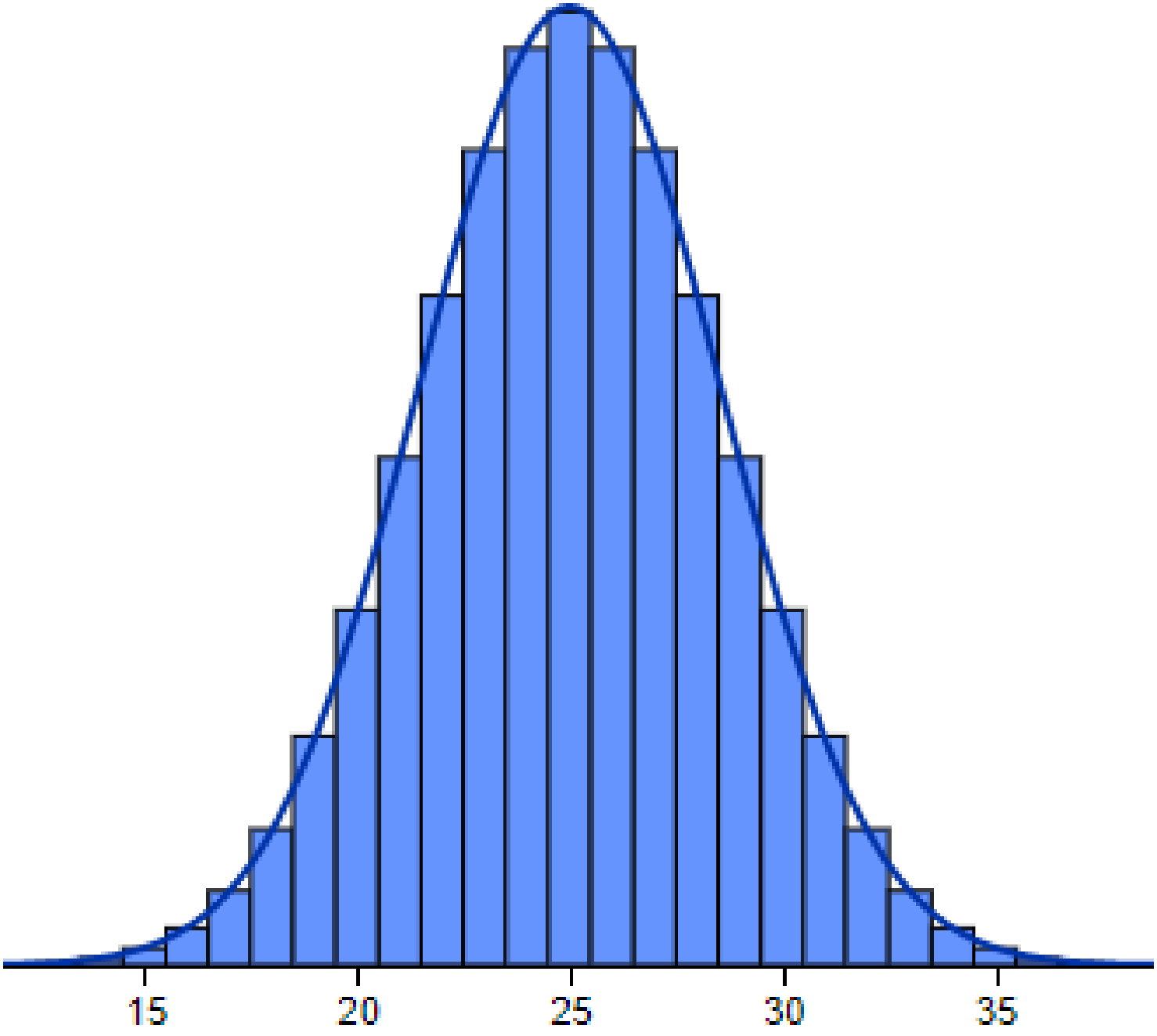
$p = 0.5, n = 5$



$p = 0.5, n = 10$



$p = 0.5, n = 50$



Main Idea (Rough)

1. Looking at the pictures, all of the binomials "have the same shape" for n big.
2. The Gaussian distribution is just that shared shape.
3. **Caveat:** Need to do some algebra (rescaling) to make PMF, PDF line up like in the pictures.

Before *giving a formula* for the Gaussian, I'll define it according to this observation.

Main Idea (Details)

- Fix $0 < p < 1$, let $X_n \sim Binom(n, p)$.
- The standard Gaussian has PDF f satisfying:

$$\int_a^b f(x)dx = \lim_{n\to\infty} \sum_{x=np+a\sqrt{np(1-p)}}^{np+b\sqrt{np(1-p)}} p_{X_n}(x)$$

$$= \lim_{n\to\infty} P[a \leq \frac{X_n - np}{\sqrt{np(1-p)}} \leq b].$$

- **Roughly:** The Gaussian PDF $f_X(x)$ is closely related to the Binomial PMF $p_{X_n}(y)$, where $y = np + x\sqrt{np(1-p)}$.

Why Bother?

- By some miracle, the Gaussian describes *almost all* repeated trials, not just those with $\{0, 1\}$-valued outcomes!
- Rather than doing big messy sums, we can just use the Gaussian formula for the limit!

What are other repeated trials?

- Let $Y_1, Y_2, \ldots$ be i.i.d. with $E[Y_1] = \mu$, $Var[Y_1] = \sigma^2$.
- Let $X_n = \sum_{i=1}^{n} Y_i$.
- If Y_i are Bernoulli(p),

$$\int_a^b f(x)dx = \lim_{n\to\infty} P[a \leq \frac{X_n - np}{\sqrt{np(1-p)}} \leq b].$$

- In the general case,

$$\int_a^b f(x)dx = \lim_{n\to\infty} P[a \leq \frac{X_n - n\mu}{\sqrt{n\sigma^2}} \leq b].$$

We didn't care about any of the details of Y_1!

Definition

We say a continuous random variable X has a **normal distribution** with parameters μ and σ, if its PDF is given by

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

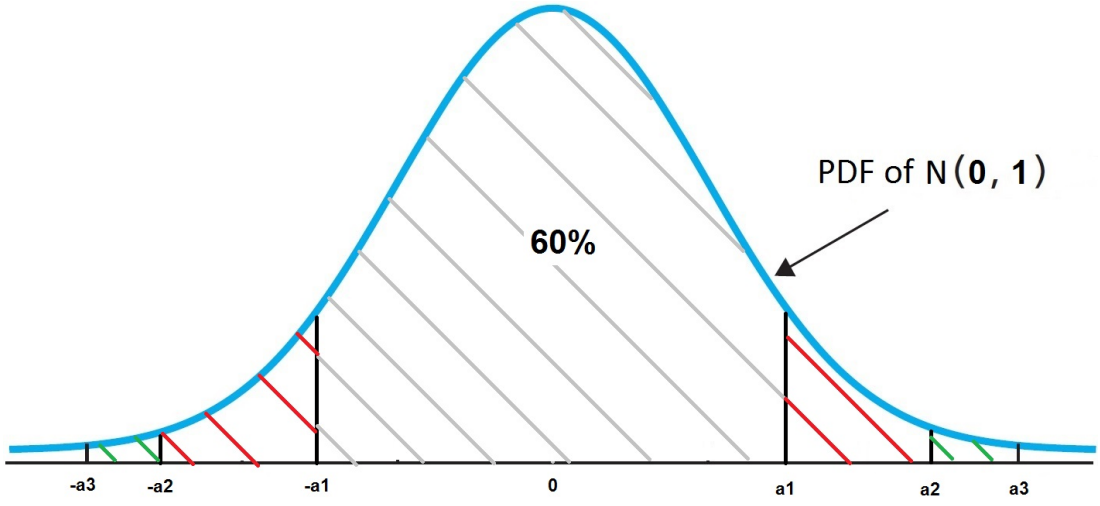
where $\mu \in (-\infty, +\infty)$, $\sigma > 0$, and $e = 2.71828...$ is Euler's number.

- Meaning of μ and σ:

$$E(X) = \mu \qquad \text{and} \qquad Var(X) = \sigma^2$$

- We use $N(\mu, \sigma^2)$ to denote a normal distribution.
- $\mu = 0$, $\sigma^2 = 1$ is the *standard* normal.

Normal PDF



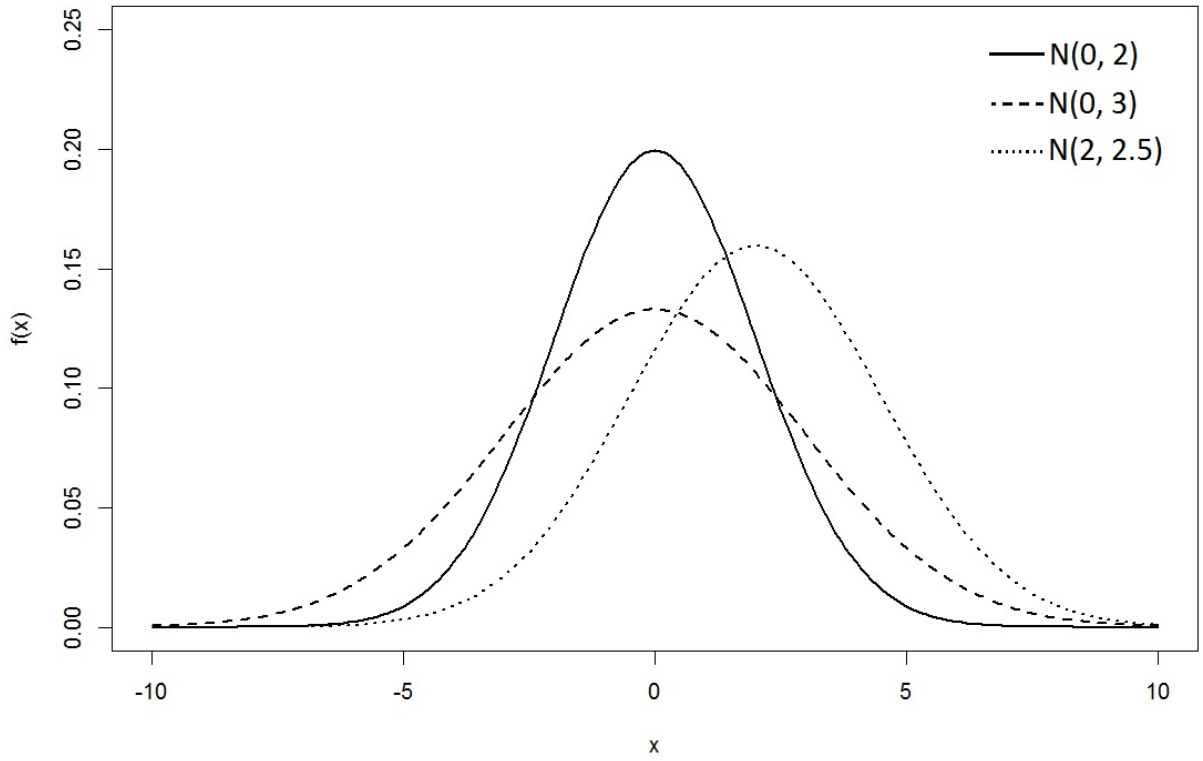
A normal PDF $f(x)$ is

- bell-shaped
- symmetric about the mean μ.

$$f(\mu + a) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(\mu+a-\mu)^2}{2\sigma^2}} = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{a^2}{2\sigma^2}} = f(\mu - a)$$

- strictly positive $f(x) > 0$; $f(x) \to 0$ as $|x| \to \infty$, but $f(x) \neq 0$

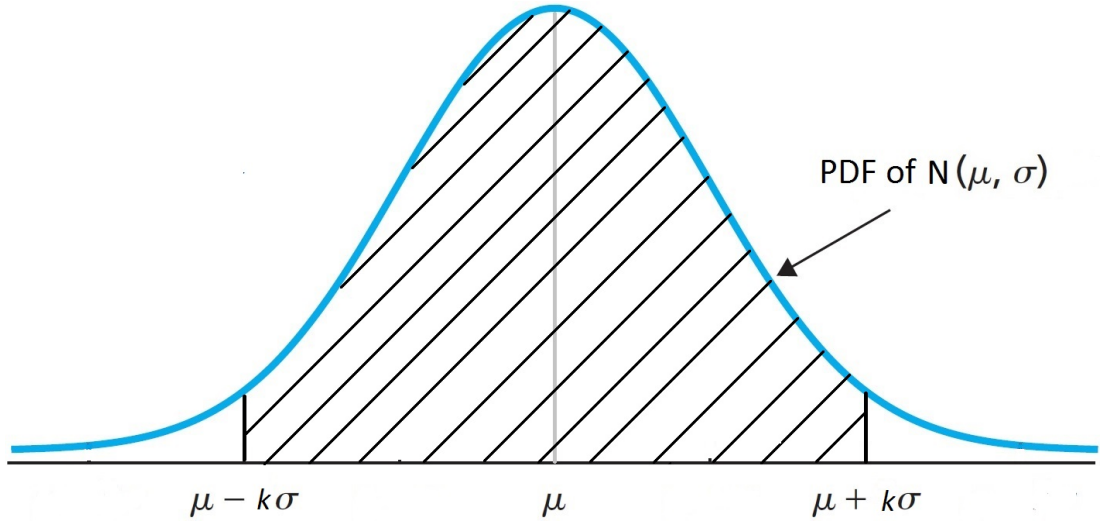
Normal PDF



A combination of (μ, σ) determines the shape of a normal curve $f(x)$.

- $\mu = E(X)$ is the center of $f(x)$.
- $\sigma = sd(X)$ is the "width" for $f(x)$ to spread around μ.

Three-sigma rule



If $X \sim N(\mu, \sigma^2)$, the probability of being within k standard deviation from the mean does **not** depend on μ and σ.

$$\begin{aligned} P(\mu - k\sigma < X < \mu + k\sigma) &= \int_{\mu-k\sigma}^{\mu+k\sigma} \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} dx \\ &= \int_{-k}^{k} \frac{1}{\sqrt{2\pi}} e^{\frac{-z^2}{2}} dz, \end{aligned}$$

where $z = (x - \mu)/\sigma$ (substitution method).

Example?

Deliberately omitted - see next set of notes!

Gaussian Calculations

uOttawa - MAT2377

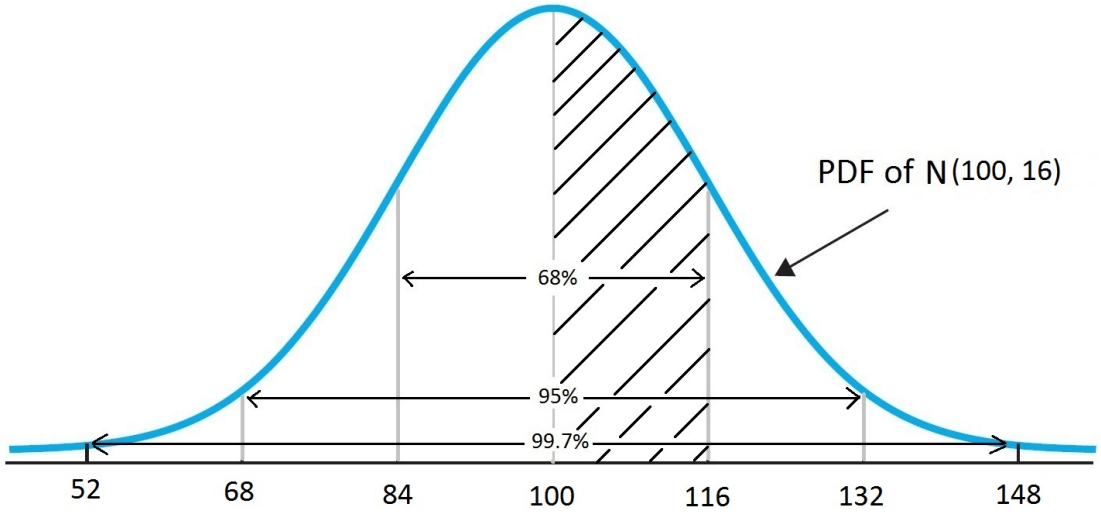
Fall 2020

Goals

Actually do calculations with the *Gaussian* distribution.

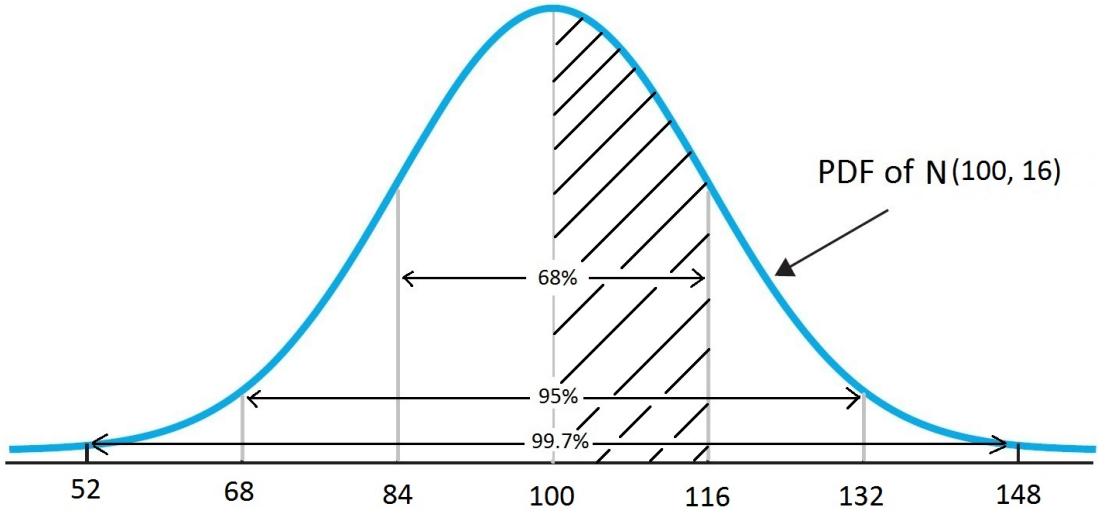
Example 1: IQ

Let X be the Intelligence Quotient (IQ) of a randomly selected Canadian. Assume $X \sim N(100, 16^2)$. $P[X < 90] = ?$



Example 1: IQ

Let X be the Intelligence Quotient (IQ) of a randomly selected Canadian. Assume $X \sim N(100, 16^2)$. $P[X < 90] = ?$

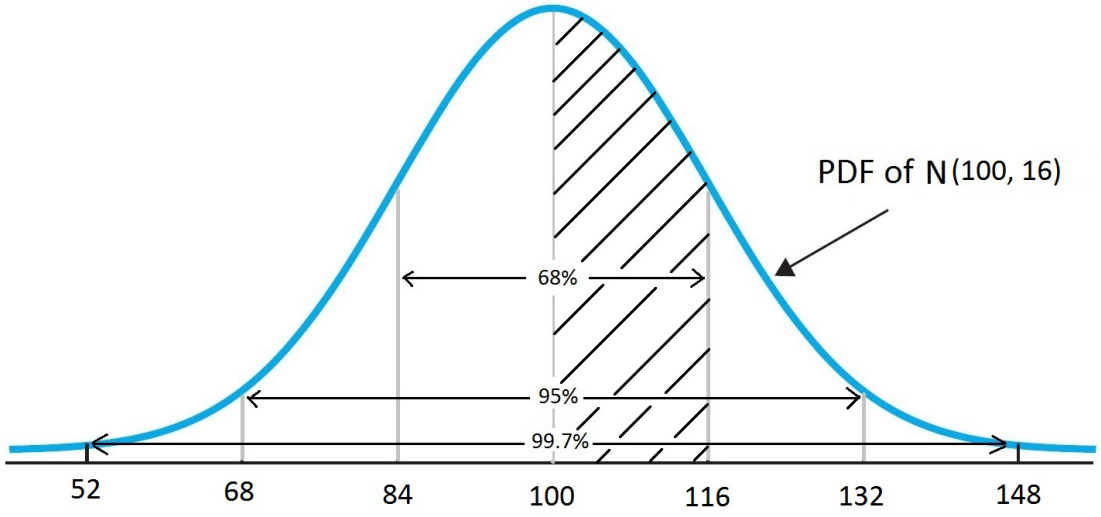


- By definition,

$$P(X < 90) = \int_{-\infty}^{90} f(x)dx = \int_{-\infty}^{90} \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} dx = ?$$

Example 1: IQ

Let X be the Intelligence Quotient (IQ) of a randomly selected Canadian. Assume $X \sim N(100, 16^2)$. $P[X < 90] = ?$



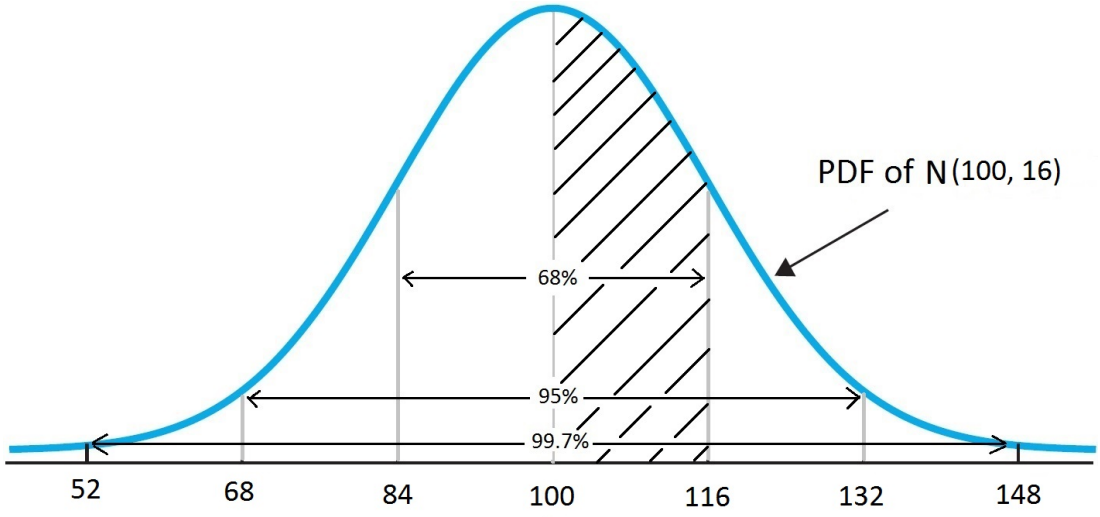
- By definition,

$$P(X < 90) = \int_{-\infty}^{90} f(x)dx = \int_{-\infty}^{90} \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} dx = ?$$

- Oops, my IQ is not high enough to do this integral...

Example 1: IQ

Let X be the Intelligence Quotient (IQ) of a randomly selected Canadian. Assume $X \sim N(100, 16^2)$. $P[X < 90] = ?$



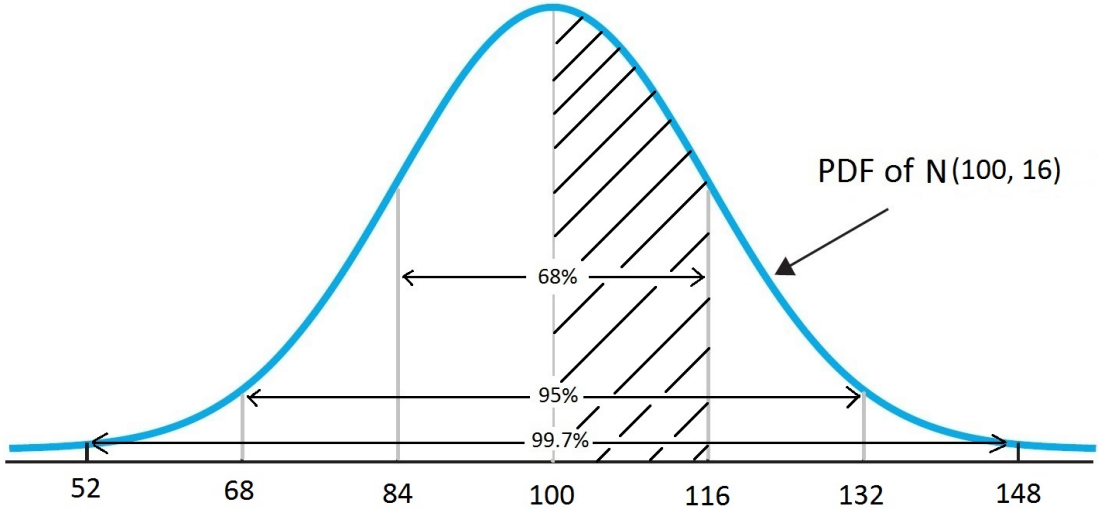
- By definition,

$$P(X < 90) = \int_{-\infty}^{90} f(x)dx = \int_{-\infty}^{90} \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} dx = ?$$

- Oops, my IQ is not high enough to do this integral...
  - Option 1: use R software

Example 1: IQ

Let X be the Intelligence Quotient (IQ) of a randomly selected Canadian. Assume $X \sim N(100, 16^2)$. $P[X < 90] = ?$



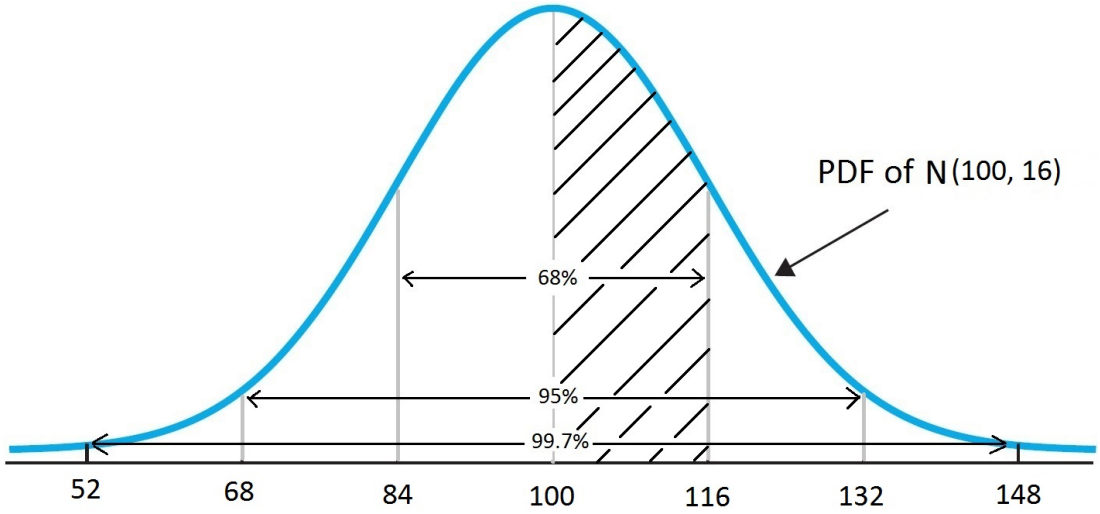
- By definition,

$$P(X < 90) = \int_{-\infty}^{90} f(x)dx = \int_{-\infty}^{90} \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} dx = ?$$

- Oops, my IQ is not high enough to do this integral...
  - Option 1: use R software
  - Option 2: use normal table

Example 1: IQ

Let X be the Intelligence Quotient (IQ) of a randomly selected Canadian. Assume $X \sim N(100, 16^2)$. $P[X < 90] = ?$



- By definition,

$$P(X < 90) = \int_{-\infty}^{90} f(x)dx = \int_{-\infty}^{90} \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} dx = ?$$

- Oops, my IQ is not high enough to do this integral...
  - Option 1: use R software
  - Option 2: use normal table
  - Option 3: forget about it...

Standard normal distribution

If $Z \sim N(0, 1)$, we say it follows a **standard normal distribution**. Its PDF is given by

$$f(z) = \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}}$$

Standard normal distribution

If $Z \sim N(0,1)$, we say it follows a **standard normal distribution**. Its PDF is given by

$$f(z) = \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}}$$

- Obviously, $E(Z) = 0$ and $Var(Z) = 1$.
- Its CDF is denoted by

$$\Phi(z) = P(Z \leq z).$$

Standard normal distribution

If $Z \sim N(0,1)$, we say it follows a **standard normal distribution**. Its PDF is given by

$$f(z) = \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}}$$

- Obviously, $E(Z) = 0$ and $Var(Z) = 1$.
- Its CDF is denoted by

$$\Phi(z) = P(Z \leq z).$$

- The values of $\Phi(z)$ are given in normal table (1st page of the textbook).

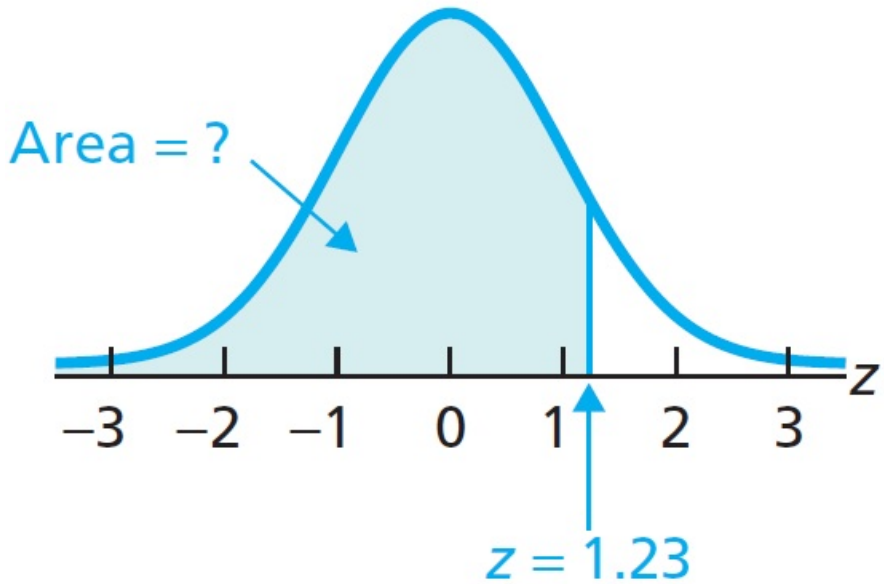
Example 2: Standard normal table

Suppose $Z \sim N(0, 1)$. Find the probability that $Z < 1.23$.

Example 2: Standard normal table

Suppose $Z \sim N(0, 1)$. Find the probability that $Z < 1.23$.

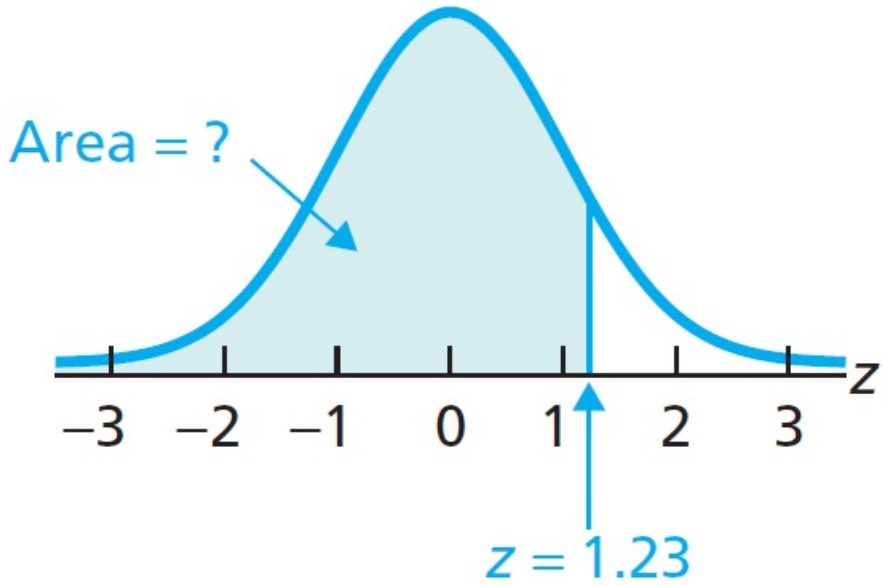
- $P(Z < 1.23) = P(Z \leq 1.23) = \Phi(1.23) =$?



Example 2: Standard normal table

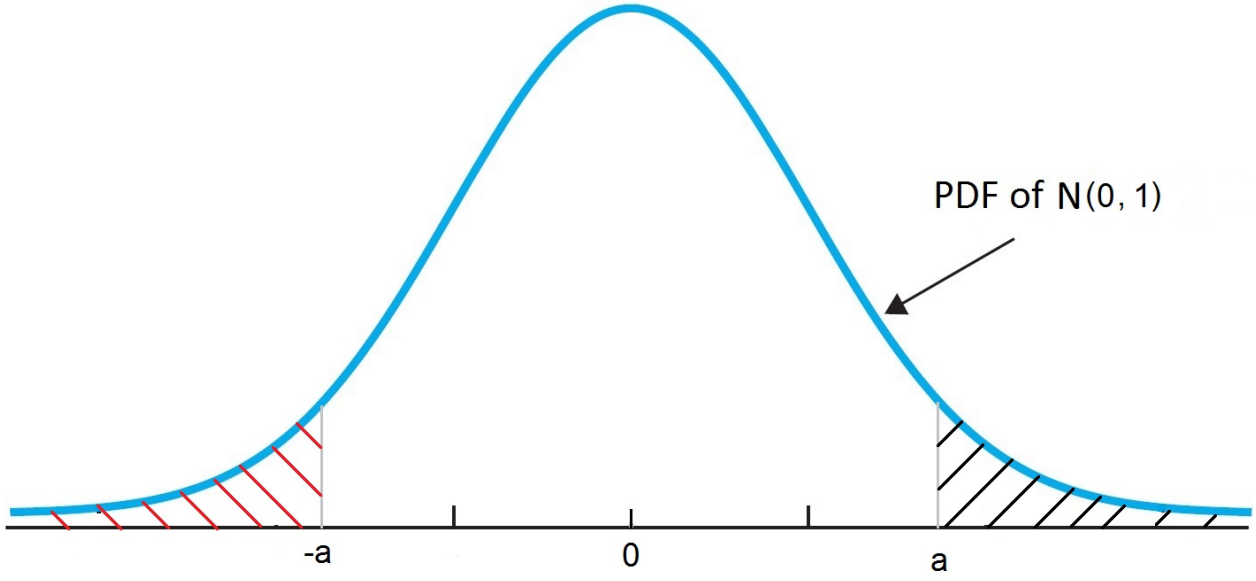
Suppose $Z \sim N(0, 1)$. Find the probability that $Z < 1.23$.

- $P(Z < 1.23) = P(Z \leq 1.23) = \Phi(1.23) =$?



- $\Phi(1.23) = 0.8907$.

Symmetry of $N(0,1)$

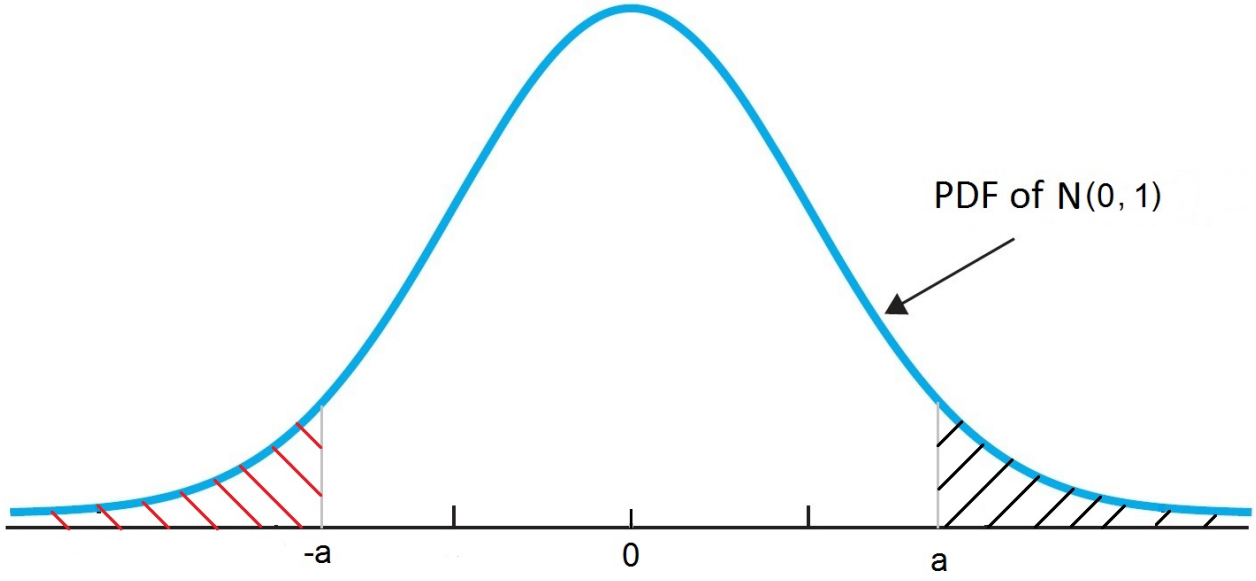


If $Z \sim N(0,1)$, $f(z)$ is symmetric about zero. This implies

$$P(Z < -a) = P(Z > a)$$

$$P(Z > -a) = P(Z < a)$$

Symmetry of $N(0,1)$



If $Z \sim N(0,1)$, $f(z)$ is symmetric about zero. This implies

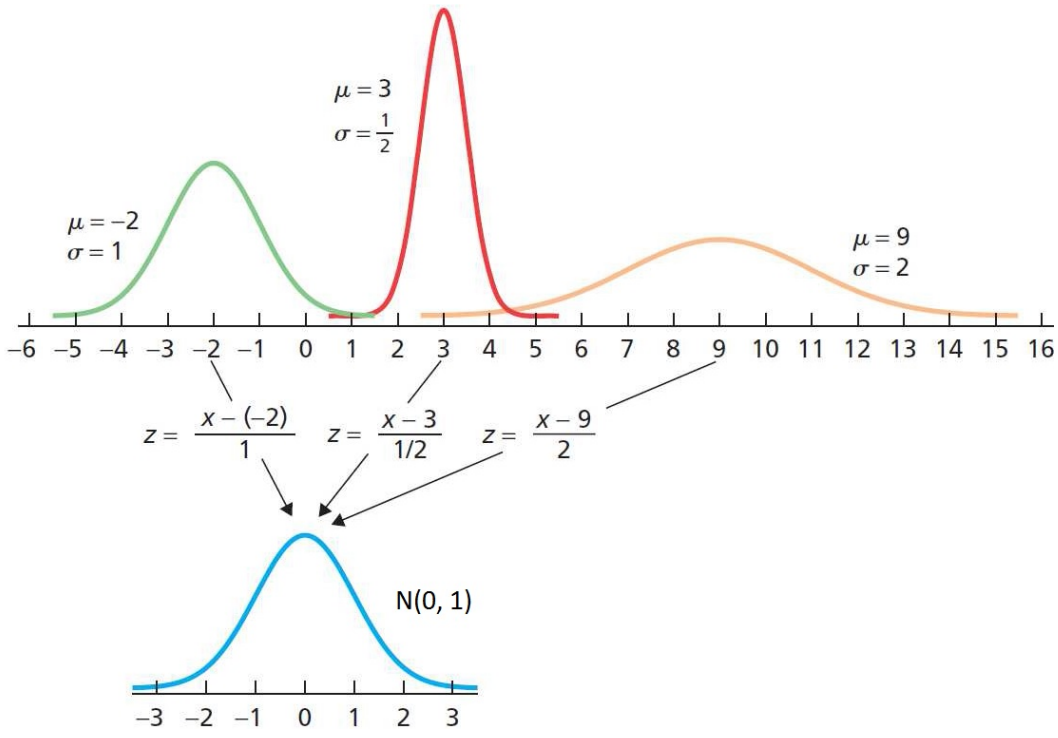
$$P(Z < -a) = P(Z > a)$$

$$P(Z > -a) = P(Z < a)$$

- Check

$$\begin{aligned} P(Z < 1.22) &= \Phi(1.22) = ? \\ P(Z > -1.22) &= 1 - \Phi(-1.22) = ? \end{aligned}$$

Standardization of normal distribution



Standardization Theorem

Suppose $X \sim N(\mu, \sigma^2)$. Then

$$Z = \frac{X - \mu}{\sigma} \sim N(0, 1).$$

Standardization of normal distribution

Suppose $X \sim N(\mu, \sigma^2)$. Using standardization, we have

$$\begin{aligned} P(X \leq a) &= P(X - \mu \leq a - \mu) \\ &= P(\frac{X - \mu}{\sigma} \leq \frac{a - \mu}{\sigma}) \\ &= P(Z \leq \frac{a - \mu}{\sigma}) \\ &= \Phi(\frac{a - \mu}{\sigma}) \end{aligned}$$

Standardization of normal distribution

Suppose $X \sim N(\mu, \sigma^2)$. Using standardization, we have

$$\begin{aligned} P(X \leq a) &= P(X - \mu \leq a - \mu) \\ &= P(\frac{X - \mu}{\sigma} \leq \frac{a - \mu}{\sigma}) \\ &= P(Z \leq \frac{a - \mu}{\sigma}) \\ &= \Phi(\frac{a - \mu}{\sigma}) \end{aligned}$$

- Since the values of $\Phi(.)$ are known, standardization enable us to calculate $P(X \leq a)$ for a general normal random variable X.

Example 3: IQ

Let X be the Intelligence Quotient (IQ) of a randomly selected Canadian. Assume $X \sim N(100, 16^2)$. (a) What is the probability that $X < 90$? (b) What is the probability that $115 < X < 140$?

Example 3: IQ

Let X be the Intelligence Quotient (IQ) of a randomly selected Canadian. Assume $X \sim N(100, 16^2)$. (a) What is the probability that $X < 90$? (b) What is the probability that $115 < X < 140$?

- Let $Z = (X - 100)/16$. We know $Z \sim N(0, 1)$, and thus

$$\begin{aligned} P(X < 90) &= P(\frac{X - 100}{16} \leq \frac{90 - 100}{16}) = P(Z \leq \frac{90 - 100}{16}) \\ &\approx \Phi(-0.63) \approx 0.2643 \end{aligned}$$

Example 3: IQ

Let X be the Intelligence Quotient (IQ) of a randomly selected Canadian. Assume $X \sim N(100, 16^2)$. (a) What is the probability that $X < 90$? (b) What is the probability that $115 < X < 140$?

- Let $Z = (X - 100)/16$. We know $Z \sim N(0, 1)$, and thus

$$\begin{aligned} P(X < 90) &= P(\frac{X - 100}{16} \leq \frac{90 - 100}{16}) = P(Z \leq \frac{90 - 100}{16}) \\ &\approx \Phi(-0.63) \approx 0.2643 \end{aligned}$$

-

$$\begin{aligned} P(115 < X < 140) &= P(\frac{115 - 100}{16} < \frac{X - 100}{16} < \frac{140 - 100}{16}) \\ &\approx P(0.94 < Z < 2.5) \\ &= \Phi(2.5) - \Phi(0.94) \\ &\approx 0.9938 - 0.8264 \\ &= 0.1674 \end{aligned}$$

Example 4: Runner

As reported in Runner's World magazine, the finish times of the runners in the New York City 10-km run are normally distributed with mean 61 minutes and standard deviation 9 minutes. (a) What is the probability that a runner finishes the run with time more than 75 minutes? (b) Find a time t such that 25% runners finish the run within t. code

- Let X be the finish time of a runner and $Z = (X - 61)/9$.

$$\begin{aligned} P(X > 75) &= P(\frac{X - 61}{9} > \frac{75 - 61}{9}) \approx P(Z > 1.55) \\ &= P(Z < -1.55) = \Phi(-1.55) = 0.0606 \end{aligned}$$

Example 4: Runner

As reported in Runner's World magazine, the finish times of the runners in the New York City 10-km run are normally distributed with mean 61 minutes and standard deviation 9 minutes. (a) What is the probability that a runner finishes the run with time more than 75 minutes? (b) Find a time t such that 25% runners finish the run within t. code

- Let X be the finish time of a runner and $Z = (X - 61)/9$.

$$\begin{aligned} P(X > 75) &= P(\frac{X-61}{9} > \frac{75-61}{9}) \approx P(Z > 1.55) \\ &= P(Z < -1.55) = \Phi(-1.55) = 0.0606 \end{aligned}$$

- From the question,

$$0.25 = P(X \leq t) = P(Z \leq \frac{t-61}{9}) = \Phi(\frac{t-61}{9}).$$

From the normal table, we know $\Phi(-0.67) \approx 0.25$, which implies that

$$\frac{t-61}{9} \approx -0.67.$$

Accordingly, $t \approx (-0.67 \times 9) + 61 = 54.97$.

Summary

- A continuous random variable X follows $N(\mu, \sigma)$, if its PDF takes the form

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

where $E(X) = \mu$ and $sd(X) = \sigma$.

- For any $X \sim N(\mu, \sigma)$,

$$Z = \frac{X - \mu}{\sigma} \sim N(0, 1),$$

Where $N(0, 1)$ is the standard normal distribution.

Samples and Statistics

uOttawa - MAT2377

Fall 2020

Goals

1. Start the second half of our class: *statistics.*
2. More specifically: learn about *sample* vs *population* statistics.

Typical Statistical Setup (Realistic Setup)

1. There is some particular group of interest, called the *population*. (Example: all Canadians).
2. We wish to learn the value of some function of this population. (Example: the true average height of all Canadians).
3. Rather than measuring all Canadians, we measure the heights of a random *sample* of all Canadians. (Example: by walking out the front door and asking the first 30 people we run into).
4. We then try to *infer* the *true* value based on this observed sample. (Example: next unit!).

Practical problem: Hard to do calculations with realistic sampling mechanisms!

Typical Statistical Setup (Toy Setup)

1. **Annoyance:** Since the true population is finite and sampling is complicated, our calculations get messy.
2. **Simple setup:** Pretend our observations are i.i.d., corresponding to perfect random samples from an "infinite" population.
3. **Toy setup:** Typically pretend our observations are exactly Gaussian.
4. **N.B.:** The toy setup is often OK.

Sampling Vocabulary (Toy Setup)

- **Sample:** A sequence $X_1, \ldots, X_n$ of i.i.d. random variables.
- **Population:** The shared distribution of the samples.
- **Sample Mean:** $\frac{1}{n}\sum_{i=1}^{n} X_i$.
- **Population Mean:** $\mu \equiv E[X_1]$.
- **Population Variance:** $\sigma^2 \equiv Var[X_1]$.

Sampling Vocabulary: Mean vs. Population

Call a function f of the data a *statistic*.

- **Sample Statistic:** $\frac{1}{n}\sum_{i=1}^{n} f(X_i)$. This is a *random variable* that *you can compute.*
- **Population Statistic:** $E[f(X_1)]$. This is a *deterministic number* that *you want to estimate.*

Note

1. These are often *close* but rarely *equal.*
2. We can already see this isn't completely general - what should sample variance be?

Introductory Statistics in a Nutshell

We get to see a *sample mean*, and infer a set of *plausible values* for the *population mean*.

Distribution of Sample Mean

uOttawa - MAT2377

Fall 2020

Goals

Find efficient formulas for computing with samples.

Reminder: Statistical Setup

1. **Sample:** An i.i.d. sequence $X_1, \ldots, X_n$ of random variables, drawn from a distribution called the **Population.**
2. **Sample mean:** $\frac{1}{n}\sum_{i=1}^{n} X_i$.

We'll do some calculations related to the sample mean.

Linear Combinations

Let $X_1, X_2, \ldots, X_n$ be random variables and $a_1, \ldots, a_n$ be constants. A *linear combination* is a random variable of the form

$$Y = a_1 X_1 + a_2 X_2 + \ldots + a_n X_n = \sum_{i=1}^{n} a_i X_i.$$

Computing with Linear Combinations

1. The mean of $Y = a_1 X_1 + a_2 X_2 + \ldots + a_n X_n$ is

$$E[Y] = a_1 E[X_1] + a_2 E[X_2] + \ldots + a_n E[X_n] = \sum_{i=1}^{n} a_i E[X_i].$$

2. If $X_1, X_2, \ldots, X_n$ are **independent**, then

$$V[Y] = a_1^2 V[X_1] + a_2^2 V[X_2] + \ldots + a_n^2 V[X_n] = \sum_{i=1}^{n} a_i^2 V[X_i].$$

3. If $X_1, X_2, \ldots, X_n$ are **independent** and **normal**, then

$$Y \sim N(E[Y]; V[Y]).$$

   In other words, Y is also normal.

Computing with Sample Means

The sample mean $\overline{X} = \sum_{i=1}^{n} X_i/n$ is a linear combination with $a_i = 1/n$. So, if the population is $N(\mu, \sigma^2)$, then

$$\overline{X} \sim N\left(\mu_{\overline{X}} = \mu, \sigma^2_{\overline{X}} = \frac{\sigma^2}{n}\right) \quad \text{and} \quad Z = \frac{\overline{X} - \mu_{\overline{X}}}{\sigma_{\overline{X}}} = \frac{\overline{X} - \mu}{\sigma/\sqrt{n}} \sim N(0, 1).$$

Example (1)

Suppose that the lifetime of a battery is normally distributed with a mean of 150 hours and a standard deviation of 25 hours. We take a random sampling of $n = 15$ batteries. What is the probability the sample mean lifetime is greater than 150 hours?

Solution: $\mu = 150$, $\sigma = 25$, $n = 15$

$$\overline{X} \sim N\left(\mu_{\overline{X}} = \mu, \sigma^2_{\overline{X}} = \frac{\sigma^2}{n}\right)$$

Now,

$$\mu_{\overline{X}} = \mu = 15$$

and

$$\sigma^2_{\overline{X}} = \frac{\sigma^2}{n} = \frac{25^2}{15} = 41.67$$

so

$$\overline{X} \sim N\left(\mu_{\overline{X}} = 150, \sigma^2_{\overline{X}} = 41.67\right).$$

Example (2)

Since

$$\overline{X} \sim N\left(\mu_{\overline{X}} = 150, \sigma^2_{\overline{X}} = 41.67\right),$$

We calculate:

$$P(\overline{X} > 150) = P\left(\frac{\overline{X} - \mu_{\overline{X}}}{\sigma_{\overline{X}}} > \frac{150 - \mu_{\overline{X}}}{\sigma_{\overline{X}}}\right)$$

$$= 1 - P\left(\frac{\overline{X} - \mu_{\overline{X}}}{\sigma_{\overline{X}}} \leq \frac{150 - \mu_{\overline{X}}}{\sigma_{\overline{X}}}\right)$$

$$= 1 - P\left(\frac{\overline{X} - \mu_{\overline{X}}}{\sigma_{\overline{X}}} \leq \frac{150 - 150}{41.67}\right)$$

$$= 1 - P\left(\frac{\overline{X} - \mu_{\overline{X}}}{\sigma_{\overline{X}}} \leq 0\right)$$

$$= 1 - \Phi(0) = 1 - \frac{1}{2} = \frac{1}{2}$$

Recap

1. If $X \sim \mathcal{N}(\mu, \sigma^2)$, then $Z = \frac{X-\mu}{\sigma} \sim \mathcal{N}(0,1)$ has "standard" normal distribution.
2. If $X_1, \ldots, X_n$ are independent Gaussians with means $\mu_1, \ldots, \mu_n$ and variances $\sigma_1^2, \ldots, \sigma_n^2$, then $\sum_{i=1}^n X_i \sim \mathcal{N}(\sum_{i=1}^n \mu_i, \sum_{i=1}^n \sigma_i^2)$.
3. To actually calcualte $P[X > x]$ for a Gaussian X, we write

$$P[X > x] = P[\frac{X-\mu}{\sigma} > \frac{x-\mu}{\sigma}] = P[Z > \frac{x-\mu}{\sigma}],$$

then look up the RHS in a table.

Central Limit Theorem

uOttawa - MAT2377

Fall 2020

Goals

State the most important theorem in probability and statistics.

Central Limit Theorem

Let $X_1, X_2, \ldots$ be an i.i.d. sequence with mean μ and variance σ^2. Let

$$Z_n = \frac{\overline{X}_n - \mu}{\sigma/\sqrt{n}}.$$

Then

$$\lim_{n\to\infty} F_{Z_n}(z) = \Phi(z).$$

Interpretation and Use

1. **Interpretation:** If the sample size n is large,

$$\overline{X} \sim N\left(\mu; \frac{\sigma^2}{n}\right) \quad \text{approximately.}$$

2. **NB:** There are no new calculations - the theorem just tells us we can pretend *any* sample is Gaussian and use the calculations from last video.
3. **How big is "large"?** In practice - should try to check. In this class - $n \geq 50$ is an OK rule of thumb.

Introduction to Inference

uOttawa - MAT2377

Fall 2020

Goals

Introduce *confidence intervals*.

Simple Setup

1. Observe $X_1, \ldots, X_n \overset{i.i.d.}{\sim} N(\mu, \sigma^2)$, σ known.
2. Want to find μ.
3. **Obvious problem:** *can't* compute μ exactly.
4. **Statistics idea:** *can* show that some values are *very implausible given the data.*

Precise Solution

1. Let $X_1, \ldots, X_n \overset{i.i.d.}{\sim} N(\mu, \sigma^2)$, σ known.
2. *Regardless of* μ,

$$P[\mu \in [\overline{X} - 1.96\frac{\sigma}{\sqrt{n}}, \overline{X} + 1.96\frac{\sigma}{\sqrt{n}}]] \approx 0.95.$$

3. **Interpretation:** we are *confident* that μ is in the interval $[\overline{X} - 1.96\frac{\sigma}{\sqrt{n}}, \overline{X} + 1.96\frac{\sigma}{\sqrt{n}}]$.

Formalization: Confidence Intervals

1. Let $X_1, \ldots, X_n \overset{i.i.d.}{\sim} f_\theta$.
2. Let $L = L(X_{1:n}), R = R(X_{1:n})$ be two functions.
3. We say that $[L, R]$ is *a confidence interval at level* α if

$$P[\theta \in [L, R]] \geq 1 - \alpha.$$

4. **Example:** we saw that $[\overline{X} - 1.96\frac{\sigma}{\sqrt{n}}, \overline{X} + 1.96\frac{\sigma}{\sqrt{n}}]$ is a confidence interval at level 0.05 for the Gaussian mean.
5. **Next few videos:** finding more confidence intervals.

Silly Confidence Intervals

1. Fix $0 < \alpha < 1$. Suppose we choose $(L, U) = \mathbb{R}$ with probability $1 - \alpha$ and $(L, U) = (0, 0)$ with probability α. Then (L, U) is a confidence interval according to our definition - but it is a useless one.
2. We don't do silly intervals in this class - but they exist in the rest of the world, sometimes by accident.

Studying Suggestion

Sections 6 and 7 of the course feature many calculations that look *very* similar. The main difficulty is recognizing what is going on. I suggest:

1. Glancing ahead at the review material for an overview, and perhaps making your own overview charts. More can easily be found online.
2. When practicing, consider just checking that you know *which procedure to carry out* without *actually doing all the computations.* This will let you do many more problems in the same amount of time, focusing on the part that most students find difficult.
3. **Always remember,** it doesn't count as practice if you don't check the results!

Mean Estimation (One Sample)

uOttawa - MAT2377

Fall 2020

Goals

1. Derive confidence intervals for the Gaussian mean problem.
2. Using this simple setting, explore some important definitions related to confidence intervals.

Useful Notation

Let Z be a standard normal random variable. Its upper quantile of order A is a value z_A such that

$$P(Z > z_A) = A,$$

i.e., the area under the probability density function of Z to the right of z_A is A.

Estimating μ (when σ is known)

Let $X_1, \ldots, X_n \overset{i.i.d.}{\sim} N(\mu, \sigma^2)$ and fix $0 < \alpha < 1$.

$$
\begin{aligned}
1 - \alpha &= P\left(-z_{\alpha/2} \leq \frac{\overline{X} - \mu}{\sigma/\sqrt{n}} \leq z_{\alpha/2}\right) \\
&= P\left(\overline{X} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \leq \mu \leq \overline{X} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right),
\end{aligned}
$$

so a confidence interval for μ at level α is

$$
\overline{x} \pm z_{\alpha/2}\frac{\sigma}{\sqrt{n}}.
$$

Example

A sample $(X_1, \ldots, X_{50})$ is taken from a population with distribution $\mathcal{N}(\mu, 1)$ for some unknown mean μ. The sample average is $\overline{X} = 12$. Calculate the standard 95 percent confidence interval.

We have $\overline{X} = 12$, $z_{0.025} = 1.96$, $\sigma = 1$ and $n = 50$. Thus, using the formula, our confidence interval is

$$(12 - (1.96)\frac{1}{\sqrt{50}}, 12 + (1.96)\frac{1}{\sqrt{50}}) = (11.72, 12.28).$$

Precision

The *precision* of a confidence interval is just its length - in our current case,

$$2\,\frac{z_{\alpha/2}\sigma}{\sqrt{n}}.$$

Remarks:

- The precision is a function of the confidence level, the true standard deviation, and the sample size.
- As we increase the confidence level and the true standard deviation, the estimation is less precise.
- As we increase the sample size, we increase precision.

Choosing sample size

If $\overline{x}$ is used as a point estimate of μ, we can be $100\,(1-\alpha)\%$ confident that the error $|\overline{x}-\mu|$ will not exceed a specified amount E when the sample size is

$$n \geq \left(\frac{z_{\alpha/2}\sigma}{E}\right)^2.$$

- If the value of n that you compute is not an integer, then it must be rounded-up to the nearest integer.
- If σ is unknown, then it is common practice to collect a preliminary sample and use the sample standard deviation s instead of σ in the formula to compute n.

Example

Consider a population with standard deviation $\sigma = 9.6$. Suppose that we want to be 95% confident that the sample mean $\overline{x}$ has an error less than 3 units when estimating the mean μ. What sample size should we use?

Solution: $\alpha = 0.05$. We want $|\overline{x} - \mu| < E$, where $E = 3$. So we need

$$n \geq \left(\frac{z_{\alpha/2}\sigma}{E}\right)^2.$$

but

$$z_{\alpha/2} = z_{0.025} = 1.96$$

so

$$n \geq \left(\frac{z_{\alpha/2}\sigma}{E}\right)^2 = \left(\frac{1.96 \times 9.6}{3}\right)^2 = 39.34,$$

so we need to use at least $n = 40$ samples.

6.3: Mean Estimation (One Sample, Unknown Variance)

uOttawa - MAT2377

Fall 2020

Goals

Find confidence intervals with *unknown variance.*

t distribution

1. Let $X_1, \ldots, X_n \overset{i.i.d.}{\sim} N(\mu, \sigma^2)$.
2. Let $\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$, $S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2$.
3. **Old Theorem:** $\frac{\overline{X}-\mu}{\sigma/\sqrt{n}} \sim N(0,1)$.
4. **Definition/New Theorem:** $\frac{\overline{X}-\mu}{S/\sqrt{n}}$ has t distribution with $(n-1)$ degrees of freedom.

Upshot: can take all of our old formulas, replacing (i) σ by S and (ii) z by t.

Confidence Intervals for Mean Estimation

σ known:

$$[\overline{X} - z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}, \overline{X} + z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}].$$

σ unknown:

$$[\overline{X} - t_{\frac{\alpha}{2},n-1} \frac{S}{\sqrt{n}}, \overline{X} + t_{\frac{\alpha}{2},n-1} \frac{S}{\sqrt{n}}].$$

Compute quantiles of t distribution by table lookup, just like Gaussian.

Example

We sample $n = 10$ springs at random. The sample has mean $\overline{X} = 8.231$ and standard deviation $S = 0.02424413$. Construct a 95% confidence interval for the mean spring diameter.

Solution: $\overline{x} = 8.231$, $s = 0.0242$, $n = 10$
$\alpha = 0.05$, so $\frac{\alpha}{2} = 0.025$. so

$$t_{\alpha/2,n-1} = t_{0.025,9} = 2.262$$

and a 95% confidence interval is

$$\overline{x} \pm t_{\alpha/2,n-1} \frac{s}{\sqrt{n}} = 8.231 \pm 2.262 \frac{0.0242}{\sqrt{10}} = 8.231 \pm 0.0152$$

Fridge Confusion:

We have two confidence intervals:

1. σ known:
$$[\overline{X} - z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}, \overline{X} + z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}].$$
2. σ unknown:
$$[\overline{X} - t_{\frac{\alpha}{2},n-1} \frac{S}{\sqrt{n}}, \overline{X} + t_{\frac{\alpha}{2},n-1} \frac{S}{\sqrt{n}}].$$

Why can't we always use the second?

6.4: Difference of Means (Paired and Unpaired)

uOttawa - MAT2377

Fall 2020

Goals

1. New problem: estimating the *difference of means* in two ways.
2. **NB:** this is probably the most important applied idea in the course.

Setup

1. We wish to *compare* two populations $N(\mu_x, \sigma_x^2)$, $N(\mu_y, \sigma_y^2)$ - typically to see which has a bigger mean.
2. Simple independent datasets: $X_1, \ldots, X_n \sim \mathcal{N}(\mu_x, \sigma_x^2)$, $Y_1, \ldots, Y_m \sim \mathcal{N}(\mu_y, \sigma_y^2)$.
3. Observe:

$$\overline{X} - \overline{Y} \sim \mathcal{N}(\mu_x - \mu(y), \frac{\sigma_x^2}{n} + \frac{\sigma_y^2}{m}),$$

so a $(1-\alpha)$ confidence interval for $\mu_x - \mu_y$ is

$$(\overline{X} - \overline{Y} - z_{\frac{\alpha}{2}}\sqrt{\frac{\sigma_x^2}{n} + \frac{\sigma_y^2}{m}}, \overline{X} - \overline{Y} + z_{\frac{\alpha}{2}}\sqrt{\frac{\sigma_x^2}{n} + \frac{\sigma_y^2}{m}}).$$

Example

We wish to compare the sheer strength of soil at two sites. Assume the testing machine is known to produce errors that are normally distributed with standard deviation of 35 pounds per square foot. Eleven samples were taken at both sites; the first had sample average 1800 and the second had sample average 1762. Compute 95 percent confidence intervals for both samples, and also for the difference between samples.

Solution: Using our formulas, the confidence intervals for μ_x, μ_y and $\mu_x - \mu_y$ are:

$$I_X = (1800 - (1.96)\frac{35}{\sqrt{11}}, 1800 + (1.96)\frac{35}{\sqrt{11}}) = (1779, 1821)$$

$$I_Y = (1762 - (1.96)\frac{35}{\sqrt{11}}, 1762 + (1.96)\frac{35}{\sqrt{11}}) = (1741, 1783)$$

$$I_{X-Y} = (38 - 1.96\sqrt{\frac{35^2}{11} + \frac{35^2}{11}}, 38 + 1.96\sqrt{\frac{35^2}{11} + \frac{35^2}{11}}) = (9, 67).$$

Note: $I_X \cap I_Y \neq \emptyset$, but $0 \notin I_{X-Y}$!

Not Appearing In This Course

We could extend this analysis to deal with unknown standard deviation, but won't do so in class. We will consider a closely related problem: **comparison with paired observations.**

Data with Paired Observations

1. We assumed that our X, Y samples were *independent.*
2. Sometimes, we get to observe *different procedures* on *same object* - e.g. we have n drivers, and let *each* driver drive in *the same two cars* we wish to compare.
3. Formalization: $X_1, \ldots, X_n \overset{i.i.d.}{\sim} N(\mu_x, \sigma_x^2)$, and $Y_1, \ldots, Y_n \overset{i.i.d.}{\sim} N(\mu_y, \sigma_y^2)$, and $X_i - Y_i \sim \mathcal{N}(\mu_{x-y}, \sigma^2)$ **with** $\sigma^2 \neq \sigma_x^2 + \sigma_y^2$.

Justification: Performances X_i, Y_i of driver i are clearly dependant!

Confidence Intervals for Paired Observations

1. Data: $X_1, \ldots, X_n \overset{i.i.d.}{\sim} N(\mu_x, \sigma_x^2)$, and $Y_1, \ldots, Y_n \overset{i.i.d.}{\sim} N(\mu_y, \sigma_y^2)$, and $X_i - Y_i \sim \mathcal{N}(\mu_{x-y}, \sigma^2)$ **with** $\sigma^2 \neq \sigma_x^2 + \sigma_y^2$.
2. Confidence Interval:

$$(\overline{X} - \overline{Y} - z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}, \overline{X} - \overline{Y} + z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}})$$

or

$$(\overline{X} - \overline{Y} - t_{\frac{\alpha}{2}, n-1} \frac{s}{\sqrt{n}}, \overline{X} - \overline{Y} + t_{\frac{\alpha}{2}, n-1} \frac{s}{\sqrt{n}})$$

where

$$s = \frac{1}{n-1} \sum_{i=1}^{n} ((X_i - Y_i) - (\overline{X} - \overline{Y}))^2.$$

Example

We wish to compare the sheer strength of soil at the *same* collection of 7 sites, both before and after rainfall. The following measurements were taken before and after rainfall:

$$\begin{aligned} \overline{X} &= 1692.9, \quad s_X = 79.9 \\ \overline{Y} &= 1793.9, \quad s_Y = 110.8 \\ \overline{X} - \overline{Y} &= 101, \quad s_{X-Y} = 33.8. \end{aligned}$$

Thus, the paired-differences 0.95-confidence interval is:

$$I_{X-Y} = (\overline{X} - \overline{Y} - t_{0.025,6}\frac{s_{X-Y}}{\sqrt{7}}, \overline{X} - \overline{Y} + t_{0.025,6}\frac{s_{X-Y}}{\sqrt{7}}) = (67.2, 134.8).$$

Very important note: old confidence interval

$$I_{X-Y} = (-25.5, 227.5).$$

This is *vastly* worse!

Estimating Proportions

uOttawa - MAT2377

Fall 2020

Goals

Compute confidence intervals for Bernoulli data.

Setup

1. We have data $X_1, \ldots, X_n \overset{i.i.d.}{\sim} Bern(p)$, and wish to estimate p.
2. Recall $[\overline{X}] = p$ and $V[\overline{X}] = n^{-1}p(1-p)$, so CLT suggests

$$\overline{X} \sim \mathcal{N}(p, \frac{p(1-p)}{n})$$

3. Usual confidence interval suggests:

$$(\overline{X} - z_{\frac{\alpha}{2}}\sqrt{\frac{\overline{X}(1-\overline{X})}{n}}, \overline{X} + z_{\frac{\alpha}{2}}\sqrt{\frac{\overline{X}(1-\overline{X})}{n}}).$$

4. Might worry: do we "know σ" or not? This is OK (textbook does a little better).

Example

The *Fantastic Party* polls its support in Metropoville. Among 2000 respondents, 1700 said they planned to vote for the FP. Calculate a 95 percent confidence interval for the true proportion of FP supporters in Metropoville.
We calculate:

$$I = (\overline{X} - z_{\frac{\alpha}{2}}\sqrt{\frac{\overline{X}(1-\overline{X})}{n}}, \overline{X} + z_{\frac{\alpha}{2}}\sqrt{\frac{\overline{X}(1-\overline{X})}{n}}) = (0.834, 0.866).$$

Not Covered In This Class

We won't cover *differences of proportions.* The suggested textbook contains a partial discussion.

Introduction to Hypothesis Testing (Single Mean Test)

uOttawa - MAT2377

Fall 2020

Goals

Introduction to a new topic: *hypothesis testing*.

What is hypothesis testing?

Hypothesis: A statement or conjecture about the population parameters

Hypothesis testing is a procedure that judges whether the data have evidence or not to support the hypothesis.

The result of a test is expressed in terms of a probability that measures how well the data and the hypothesis agree.

Example: GMAT training program

More than 200,000 people worldwide take the GMAT exam each year as they apply for MBA programs; their average score is about 525. A college designs a training program that helps students to raise their GMAT scores. To evaluate whether this program is successful, the college randomly select 20 students in their program and find that their average score is 533.

Question: Is the mean score of the students who enrolled in this training program higher than the average score 525?

Clearly, $533 > 525$ and there is a 8-point increase. Problem solved.

Example: GMAT training program

More than 200,000 people worldwide take the GMAT exam each year as they apply for MBA programs; their average score is about 525. A college designs a training program that helps students to raise their GMAT scores. To evaluate whether this program is successful, the college randomly select 20 students in their program and find that their average score is 533.

Question: Is the mean score of the students who enrolled in this training program higher than the average score 525?

Clearly, $533 > 525$ and there is a 8-point increase. Problem solved.

- Question: $\mu > 525$? What we know: $\bar{x} = 533$ with $n = 20$.

Example: GMAT training program

One way to answer this question is to compute $P(\bar{x} \geq 533)$, under the assumption that $\mu = 525$.

Suppose $P(\bar{x} \geq 533) = 1/5$. Because it is not very small, observing $\bar{x} = 533$ is not surprising when $\mu = 525$. The observed 8-point increase may be just due to chance rather than a real difference. We have no evidence to question the assumption $\mu = 525$, that is the students in this training program have the same mean GMAT score as others.

Be careful. The observed $\bar{x} = 533 > 525$ may not be sufficient for us to conclude that $\mu > 525$.

Example: GMAT training program

Suppose we observe $\bar{x} = 549$ from 20 students in this training program. Under the assumption that $\mu = 525$, we have $P(\bar{x} \geq 549) = 0.001$.

Because the probability of 0.001 is so low, it is unlikely to observe $\bar{x} \geq 549$, when $\mu = 525$. Thus, the assumption that $\mu = 525$ is likely to be false. We have sufficient evidence in the data to conclude that $\mu > 525$, that is the students in this training program have a higher mean GMAT score than others.

How we reached a testing conclusion?

Let us summarize what we just did in the previous example.

- Begin with a question on whether a particular group of students has a mean GMAT score higher than the others.
- Setup a **Hypothesis** that there is no difference in the mean GMAT scores between this group and general population.
- Assuming the hypothesis is true, compute the probability of observing a sample mean score as high or higher than the observed one.
- We reach two probabilities: 0.2 in the first case; 0.001 in the second case.

How we reached a testing conclusion?

We then make different conclusions:

Case 1 Since the probability of 0.2 is not small, we do not have enough evidence to question/reject the hypothesis.

Case 2 Since the probability of 0.001 is very small, we have two possible explanations.

1. We observed something that is very unusual, or
2. the hypothesis "*there is no difference in the mean GMAT scores between this group and general population*" is **not** true.

We prefer the second explanation: this group of students has a higher mean GMAT score than others.

Insight: rules in a criminal trial

What we did for hypothesis testing is analogous to the rules used in a criminal trial.

- Setup a hypothesis that the defendant is innocent.
- A jury assess the evidence against the hypothesis.
- If there is evidence to show that the hypothesis is false, the judge sentences that the defendant is guilty.
- If there is not enough evidence to show that the hypothesis is false, the judge acquits the defendant.

Formalization: setting up a hypothesis

Let $\{x_1, \ldots, x_n\}$ be a random sample of size n drawn from a population with mean μ and standard deviation σ. We are interested in testing whether μ is smaller/larger than some given value μ_0.

Setup a null hypothesis that we will try to find evidence **against**.

$$H_0 : \mu = \mu_0$$

Choose an alternative hypothesis, often the effect that we hope to find evidence **for**; common choices:

$$H_1 : \mu > \mu_0, \qquad H_1 : \mu < \mu_0, \qquad H_1 : \mu \neq \mu_0$$

Setup a hypothesis

If you do not have a specific direction firmly in mind on whether μ should be larger or smaller than μ_0, you should use a two-sided alternative.

$$H_1 : \mu \neq \mu_0$$

- In applications, we often begin with H_1 and then set up H_0 as the statement that the hoped-for effect is not present.
- H_1 should be set before the data are collected. **It is cheating to first look at the data and then frame H_1 to fit what the data show**.

Test statistic for the population mean

Suppose $X \sim N(\mu, \sigma^2)$ with an unknown μ and a **known** σ. Let $\{X_1, \ldots, X_n\}$ be a random sample. We use $\overline{X}$ to estimate μ.

- $E(\overline{X}) = \mu$, $sd(\overline{X}) = \sigma/\sqrt{n}$
- $$\frac{\overline{X} - E(\overline{X})}{sd(\overline{X})} = \frac{\overline{X} - \mu}{\sigma/\sqrt{n}} \sim N(0,1)$$

When $H_0 : \mu = \mu_0$ is true, we expect $\overline{X}$ is close to μ_0 and

$$Z = \frac{\overline{X} - \mu_0}{\sigma/\sqrt{n}} \sim N(0,1)$$

- It serves as a **test statistic** that measures the compatibility between H_0 and the data.

The p-value

When $\{x_1, \ldots, x_n\}$ is actually observed, we have an observed value for the test statistics Z:

$$z_0 = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}}$$

- We then assess H_0 by computing the probability that, under H_0, the test statistic would take a value as extreme or more extreme than what is actually observed. Such a probability is called ***p*-value**.
- The direction that counts as "extreme" is determined by H_1.
  - $H_1 : \mu > \mu_0$: $P(Z > z_0)$
  - $H_1 : \mu < \mu_0$: $P(Z < z_0)$
  - $H_1 : \mu \neq \mu_0$: $P(Z > |z_0| \text{ or } Z < -|z_0|)$
- The p-value judges how likely H_0 is against by the data. The smaller p-value, the stronger evidence against H_0.

Statistical significance

Once a *p*-value is obtained, we can compare it with a pre-specified threshold value $\alpha \in (0, 1)$ that we regard as decisive. Such a threshold value α is called **significance level**.

Suppose we set $\alpha = 0.05$, we are requiring that the data give evidence against H_0 so strong that it would happen no more than 5% of the time when H_0 is true.

- If the *p*-value is as small or smaller than α, we say that the test is statistically significant at level α. We have evidence to reject H_0.
- If the *p*-value is larger than α, the test is not statistically significant. We do not have evidence to reject H_0.
- α is chosen by users. It is common to set $\alpha = 0.05$ in many cases.

Hypothesis testing for the population mean

A hypothesis test is a process for assessing the significance of the evidence provided by data against a null hypothesis. It has four typical steps as follows.

1. State the null hypothesis H_0 and the alternative hypothesis H_1.

2. Calculate the observed value of the test statistic on which the test will be based.

3. Find the p-value for the observed value of the test statistic.

4. Make a conclusion. if a p-value is smaller than α, we reject H_0; if p-value is larger than α, we fail to reject H_0.

Example: one-sided z test

Suppose over 1000 athletes participated in a Marathon game; their average finish time is 256 min. A researcher wants to know whether African athletes ran faster than the general average. She draw a random sample of $n = 40$ African athletes and find their mean finish time $\bar{x} = 245$ min. Assume that the finish time of African athletes follows $N(\mu, 33^2)$. How should her make a conclusion under the significance level $\alpha = 0.05$?

1. Setup hypothesis:

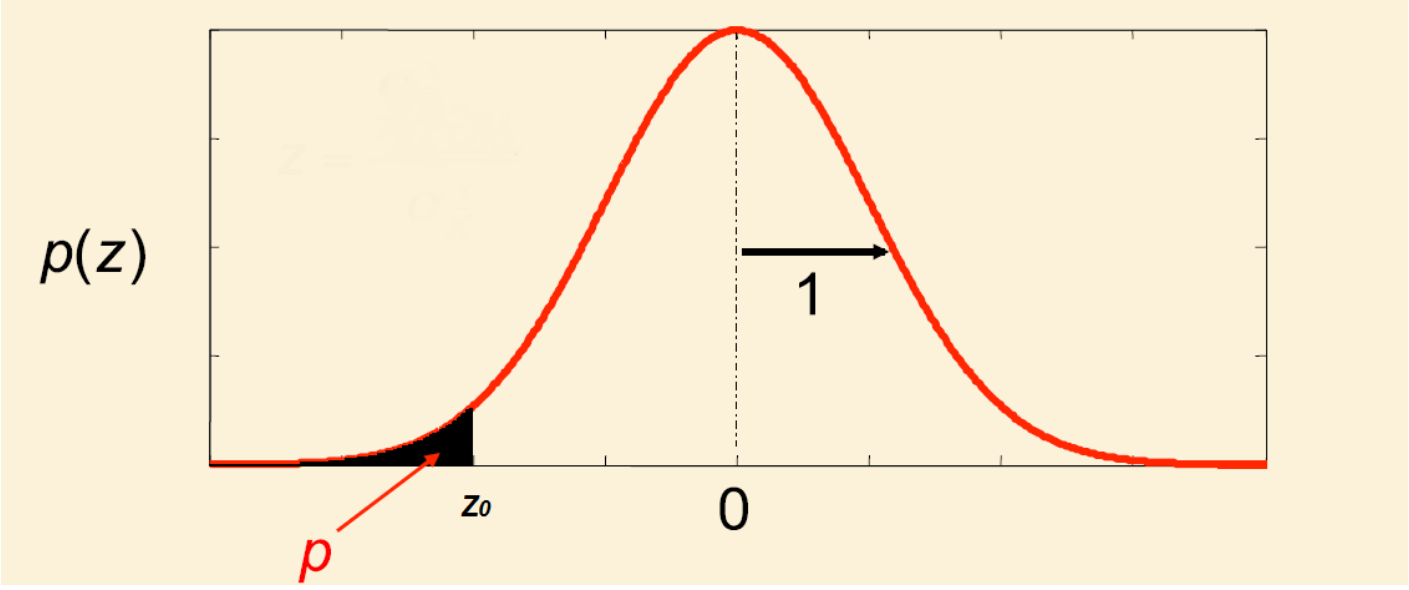
$$H_0 : \mu = 256 \quad \text{vs.} \quad H_1 : \mu < 256$$

   - H_0: the average finish time of African athletes is equal to 256 min.
   - H_1: the average finish time of African athletes is less than 256 min.

2. Compute the observed value of test statistic:

$$z_0 = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}} = \frac{245 - 256}{33/\sqrt{40}} \approx -2.11$$

Example: one-sided z test



3 Find the p-value (using Z-table):

$$P(Z < z_0) = P(Z < -2.11) \approx 0.02$$

4 Make a conclusion: Since the p-value is 0.02, which is smaller than $\alpha = 0.05$. We have evidence to reject H_0 and we conclude that the mean finish time of African athletes is less than the overall average (256 min).

Example: two-sided z test

In the Marathon example, the researcher turn to study on the performance of Asian athletes. However, she has no firm idea in mind on whether Asian athletes potentially run faster or not than the overall average. She randomly collected 30 Asian athletes and found that their mean finish time is 267 min. Assume that the finish time of Asian athletes follows $N(\mu, 33^2)$. Based on 0.05 significance level, can we conclude that the mean finish time of Asian athletes is different from the overall average 256 min?

1. Setup hypothesis:

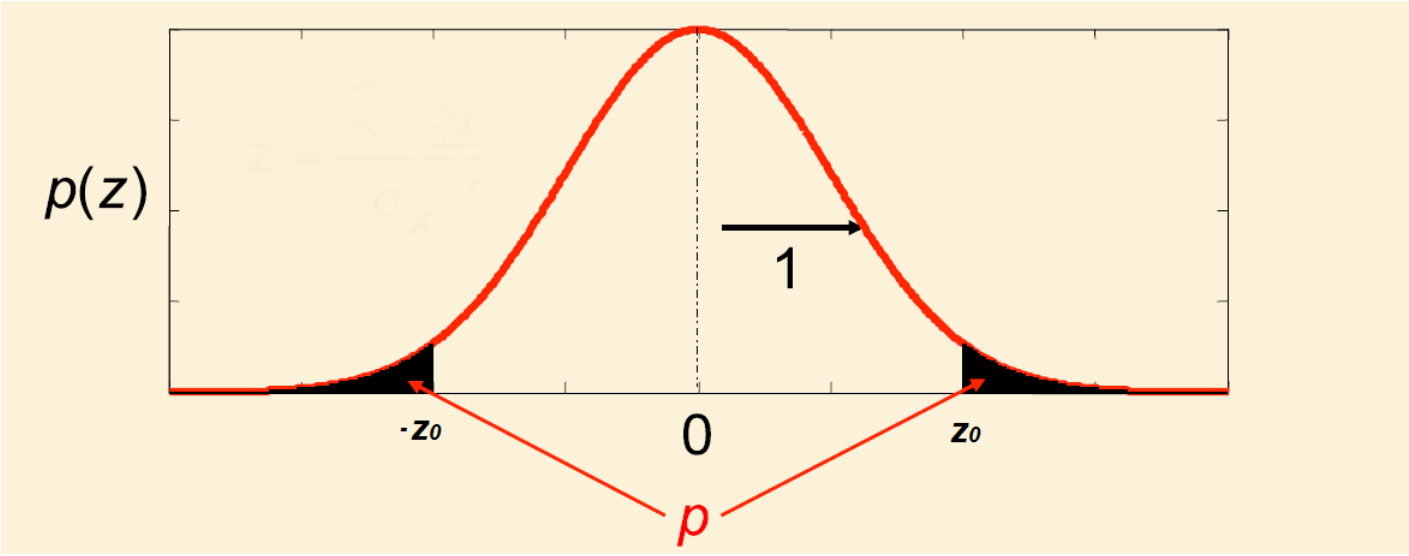
$$H_0 : \mu = 256 \quad \text{vs.} \quad H_1 : \mu \neq 256$$

   - H_0: the average finish time of Asian athletes is equal to 256 min.
   - H_1: the average finish time of Asian athletes is not equal to 256 min.

2. Compute the observed value of test statistic:

$$z_0 = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}} = \frac{267 - 256}{33/\sqrt{30}} \approx 1.83$$

Example: two-sided z test



3 Find the p-value (using Z-table):

$$P(Z < -|z_0| \text{ or } Z > |z_0|) = P(Z < -1.83) + P(Z > 1.83) \approx 0.07$$

4 Make a conclusion: Since the p-value is 0.07, which is larger than $\alpha = 0.05$. We fail to reject H_0; the data do not provide enough evidence to conclude that the mean finish time of Asian athletes is different from the overall average. (**Do not say H_0 is true!**)

z test for a population mean

To test $H_0 : \mu = \mu_0$ based on a random sample of size n from a normal population X with an unknown mean μ and a **known** standard deviation σ, we compute the observed value of the test statistic

$$z_0 = \frac{\bar{x} - u_0}{\sigma/\sqrt{n}}$$

Compute the p-value:

- H_1: $\mu > \mu_0$, $P(Z > z_0)$
- H_1: $\mu < \mu_0$, $P(Z < z_0)$
- H_1: $\mu \neq \mu_0$, $P(Z < -|z_0| \text{ or } Z > |z_0|)$

Let α be a given significance level. If $p < \alpha$, we reject H_0 and conclude H_1 is true. If $p \geq \alpha$, we fail to reject H_0 and do not have enough evidence to justify H_1.

Two types of error

Conclusions of hypothesis testing is based on probability. We reject H_0, if we have probability evidence from the data; we fail to reject H_0, if we do not have enough evidence from the data. We hope our decision will be correct, but sometimes it could be wrong.

	Truth	
Decision	H_0 true	H_0 false
Retain H_0	Correct retention	Type II error
Reject H_0	Type I error	Correct rejection

Two types of error

Type I error: we reject H_0 when in fact H_0 is true.

- Example: *The jury convicts an innocent defendant.*
- $P(\text{Type I error}) = P(\text{reject } H_0 | H_0 \text{ is true}) = \alpha =$ significance level

Type II error: we fail to reject H_0 when in fact H_0 is false.

- Example: *The jury acquits a guilty defendant.*
- $P(\text{Type II error}) = P(\text{accept } H_0 | H_0 \text{ is false})$
- $\text{Power} = 1 - P(\text{Type II error}) = P(\text{reject } H_0 | H_0 \text{ is false}) = \beta$

For a given testing method, there is no “ideal" decision rule to minimize both Type I and Type II errors. In practice, we often set a test with a small α, because usually Type I errors are more serious. Increasing sample size is an effective way to increase power (reduce Type II error) with a given α.

Statistical significance $\neq$ Practical significance

When H_0 can be rejected at usual level $\alpha = 0.05$, there is good evidence that an effect is present. However, such an effect can be very small in practice.

Consider the GMAT example we used before, we want to test

$$H_0 : \mu = 525 \quad \textit{vs.} \quad H_1 : \mu > 525$$

Suppose $X \sim N(\mu, 100^2)$ with $\mu = 528$, so H_1 is true. We observe $\bar{x} = 528$ with $n = 5000$. The p-value for the z test is $P(Z > 2.12) \approx 0.02$, which is statistical significant at $\alpha = 0.05$. However, only 3 points increase in GMAT score is not practically significant.

Lack of significance $\neq H_0$ is true

When you fail to reject H_0, this does not imply that H_0 is true. It may be due to Type II error of your testing method.

Again, in the GMAT example, we want to test

$$H_0 : \mu = 525 \quad \textit{vs.} \quad H_1 : \mu > 525$$

Suppose $X \sim N(\mu, 100^2)$ with $\mu = 550$, and we observe $\bar{x} = 541$ with $n = 15$. The p-value for the z test is $P(Z > 0.62) \approx 0.27$, which is not statistically significant at $\alpha = 0.05$. But in fact $\mu - \mu_0 = 25$, which is a significant increment. You just fail to detect it due to the lack of power.

Hypothesis Testing and Confidence Intervals

uOttawa - MAT2377

Fall 2020

Goals

Relate hypothesis tests to our previous subject: confidence intervals.

Simple NHST Framework

1. **Data:** $X_1, \ldots, X_n \sim f_\theta$.
2. **Hypotheses:** $H_0 : \theta = \theta_0$, some H_1, some α.
3. **Decision rule:** *reject* H_0 if $\overline{X}$ is not in some interval $I = I(\theta_0)$.

Simple NHST Framework (CI version)

1. **Data:** $X_1, \ldots, X_n \sim f_\theta$.
2. **Hypotheses:** $H_0 : \theta = \theta_0$, some H_1, some α.
3. **Decision rule:** *reject* H_0 if θ_0 is not in the confidence interval for θ.

Note:

1. All of our hypothesis test calculations are really just confidence interval calculations!
2. The formal decision does not depend on H_1 *at all*. H_1 only suggests the choice of confidence interval.

Rest of unit on NHST

1. We will go over NHST examples - but *no new calculations at all* until we get to regression.
2. You must still learn new things, such as (i) which CI is appropriate and (ii) the NHST framework itself.

Testing and Power

uOttawa - MAT2377

Fall 2020

Goals

Illustrate several hypothesis tests for the single-mean problem.

Recall: Testing population mean (σ^2 known)

Let $X \sim N(\mu, \sigma^2)$ with an unknown mean μ and a **known** standard deviation σ.

1. Setup hypothesis:

$$H_0 : \mu = \mu_0 \quad \text{vs.} \quad H_1 : \mu < \mu_0$$

If H_0 is true,

$$Z = \frac{\overline{X} - \mu}{\sigma/\sqrt{n}} = \frac{\overline{X} - \mu_0}{\sigma/\sqrt{n}} \sim N(0, 1)$$

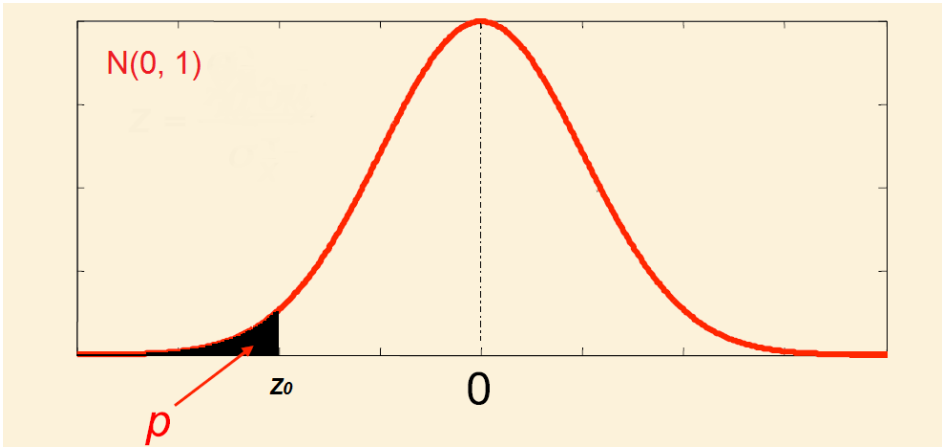
2. Compute the observed value for Z:

$$z_0 = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}}$$

3. Check whether z_0 is an **unusual** value for Z: $P(Z < z_0) = p$.

4. Let α be a given significance level. If $p < \alpha$, we reject H_0 and conclude H_1 is true. If $p > \alpha$, we fail to reject H_0 and do not have enough evidence to justify H_1. (Do **not** say H_0 is true.)

p-value: one-sided vs. two-sided



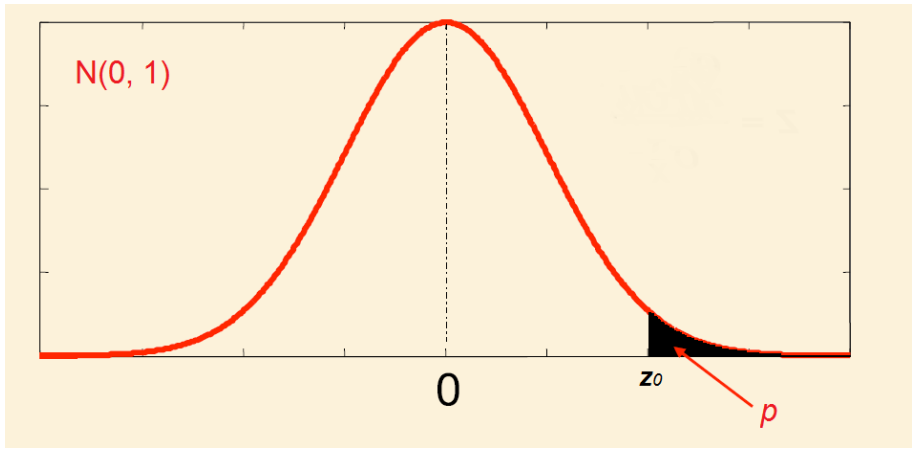
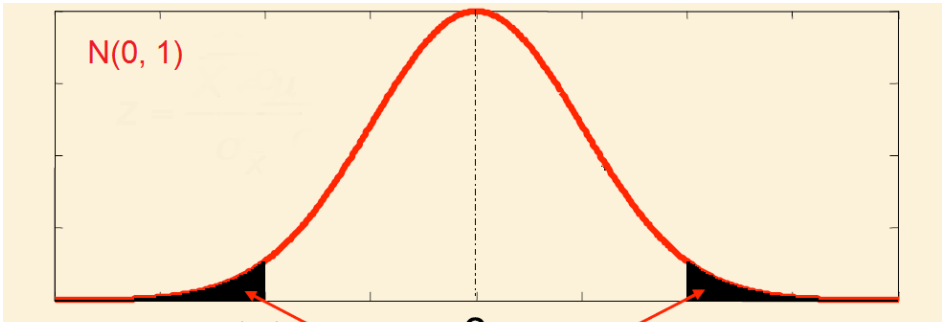


Figure 1: $H_1 : \mu < \mu_0$: $P(Z < z_0)$, $H_1 : \mu > \mu_0$: $P(Z > z_0)$



Example: student height (one-sided)

Let X denote the height (cm) of female students at uOttawa. Assume that $X \sim N(\mu, 5^2)$. A random sample of size 7 is drawn from X with a sample mean $\bar{x} = 166.86$. Can we claim that $\mu > 163.5$? (significance level $\alpha = 0.05$)

1. Setup hypothesis:

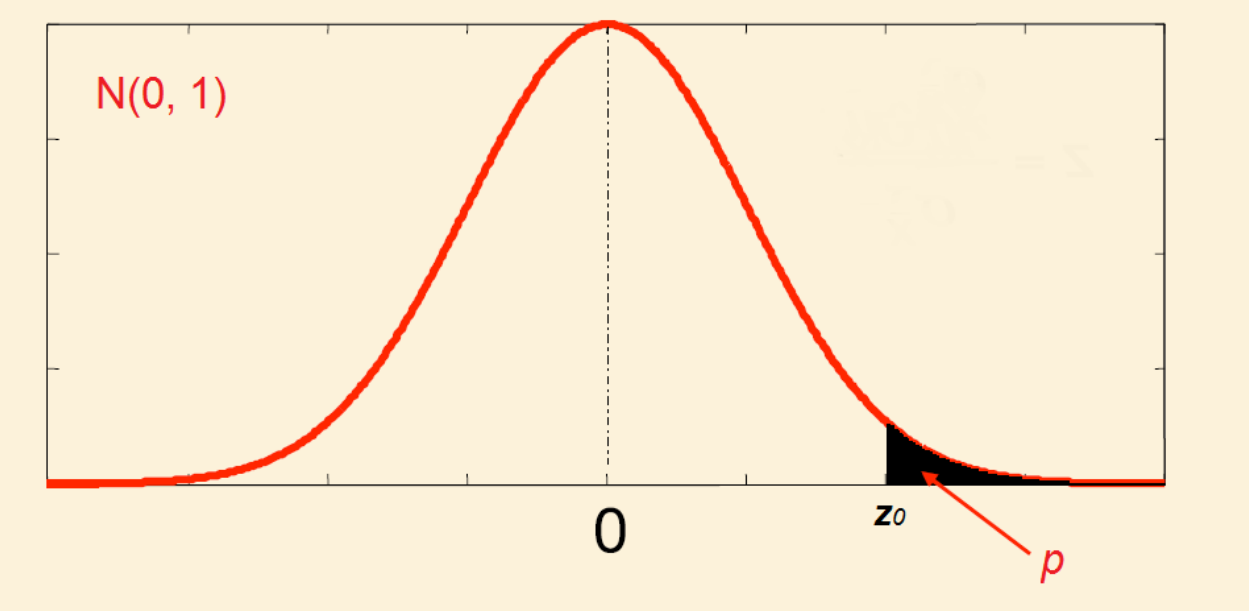
$$H_0 : \mu = 163.5 \quad \text{vs.} \quad H_1 : \mu > 163.5$$

   - H_0: the average height of female students is 163.5 cm.
   - H_1: the average height of female students is greater than 163.5 cm.

2. Compute the observed value of test statistic:

$$z_0 = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}} = \frac{166.86 - 163.5}{5/\sqrt{7}} \approx 1.78$$

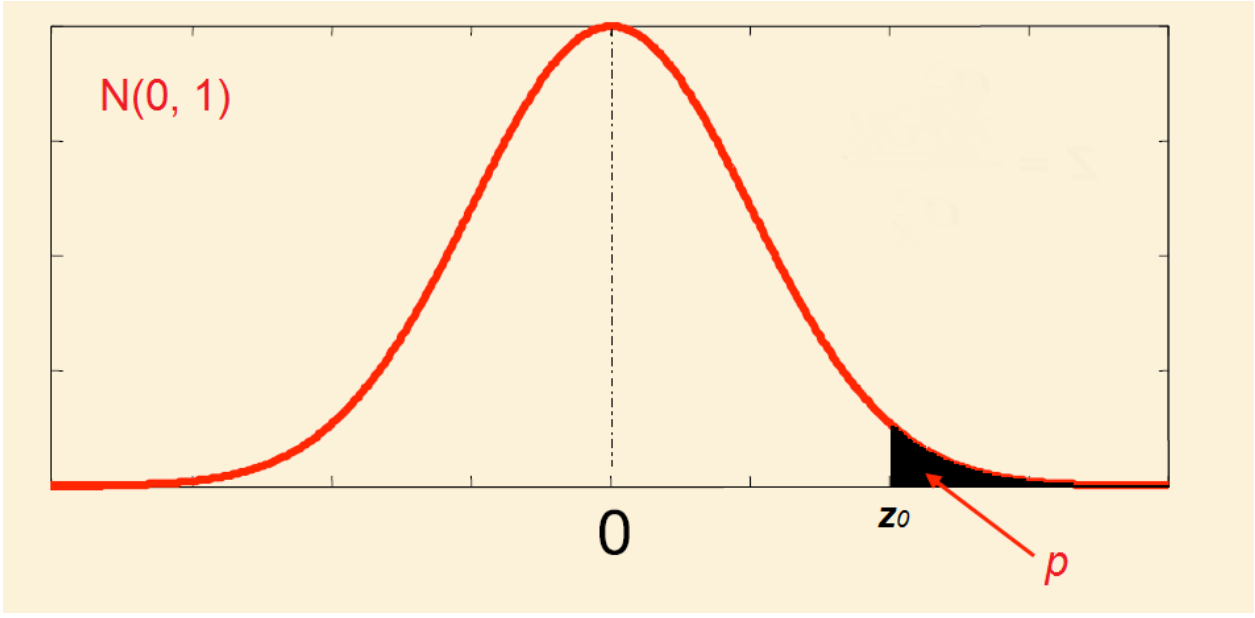
Example: student height (one-sided)



3 Find the p-value (using Z-table):

$$P(Z > z_0) = P(Z > 1.78) = P(Z < -1.78) = 0.0375$$

Example: student height (one-sided)



3 Find the p-value (using Z-table):

$$P(Z > z_0) = P(Z > 1.78) = P(Z < -1.78) = 0.0375$$

4 Make a conclusion: Since the p-value is smaller than $\alpha = 0.05$. We reject H_0 and conclude that the average height of female students at uOttawa is greater than 163.5 cm.

Example: student height (two-sided)

Let X denote the height (cm) of female students at uOttawa. Assume that $X \sim N(\mu, 5^2)$. A random sample of size 7 is drawn from X with a sample mean $\bar{x} = 166.86$. Can we claim that $\mu \neq 163.5$? (significance level $\alpha = 0.05$)

1. Setup hypothesis:

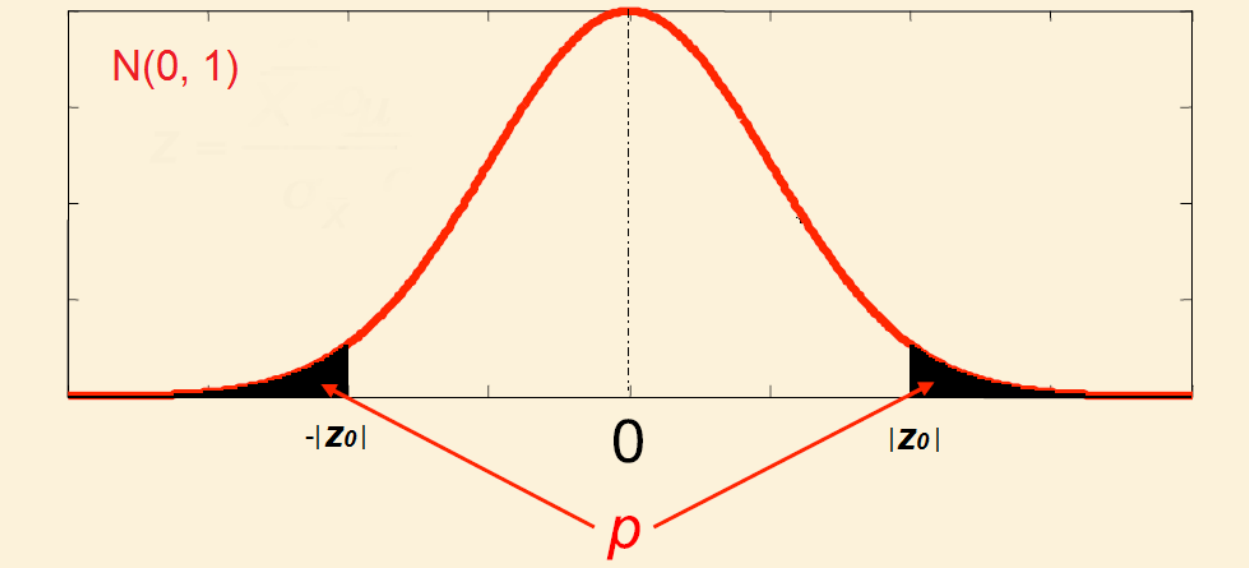
$$H_0 : \mu = 163.5 \quad \text{vs.} \quad H_1 : \mu \neq 163.5$$

   - H_0: the average height of female students is 163.5 cm.
   - H_1: the average height of female students is different from 163.5 cm.

2. Compute the observed value of test statistic:

$$z_0 = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}} = \frac{166.86 - 163.5}{5/\sqrt{7}} \approx 1.78$$

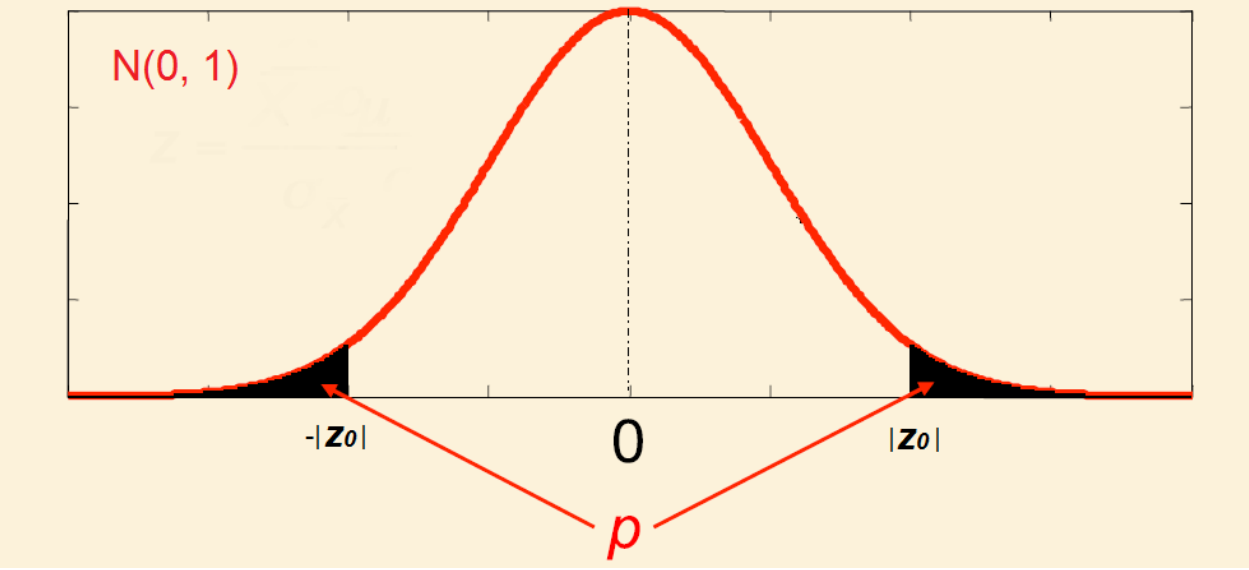
Example: student height (two-sided)



3 Find the p-value (using Z-table):

$$P(Z > |z_0| \text{ or } Z < -|z_0|) = 2P(Z < -1.78) = 0.075$$

Example: student height (two-sided)



3 Find the p-value (using Z-table):

$$P(Z > |z_0| \text{ or } Z < -|z_0|) = 2P(Z < -1.78) = 0.075$$

4 Make a conclusion: Since the p-value is larger than $\alpha = 0.05$. We fail to reject H_0 and do not have evidence to conclude that the average height of female students at uOttawa is greater than 163.5 cm.

One-sided vs. Two-sided

In one-sided test:

1. $H_0 : \mu = 163.5$ vs. $H_1 : \mu > 163.5$
2. $z_0 = \frac{\bar{x}-\mu_0}{\sigma/\sqrt{n}} = \frac{166.86-163.5}{5/\sqrt{7}} \approx 1.78$
3. $p = P(Z > z_0) = P(Z > 1.78) = 0.0375$
4. Since $0.0375 < \alpha = 0.05$, we reject H_0 and conclude $\mu > 163.5$.

In two-sided test:

1. $H_0 : \mu = 163.5$ vs. $H_1 : \mu \neq 163.5$
2. $z_0 = \frac{\bar{x}-\mu_0}{\sigma/\sqrt{n}} = \frac{166.86-163.5}{5/\sqrt{7}} \approx 1.78$
3. $p = P(Z > z_0 \text{ or } Z < -z_0) = 2P(Z > 1.78) = 0.075$
4. Since $0.075 > \alpha = 0.05$, we fail to reject H_0 and do not have enough
   evidence to conclude $\mu \neq 163.5$.

One-sided vs. Two-sided

- H_1 reflects the prior knowledge about the scientific question; we need to set up H_1 before analyzing the data.
- $H_1 : \mu > 163.5$ is more informative than $H_1 : \mu \neq 163.5$.
- With the same test statistic, an one-sided test is more powerful than the two-sided test in detecting mean change in **that particular direction**.

One-sided vs. Two-sided

- H_1 reflects the prior knowledge about the scientific question; we need to set up H_1 before analyzing the data.
- $H_1 : \mu > 163.5$ is more informative than $H_1 : \mu \neq 163.5$.
- With the same test statistic, an one-sided test is more powerful than the two-sided test in detecting mean change in **that particular direction**.
- A wrong direction would make things worse. Suppose $\mu = 165$.
  - $H_0 : \mu = 163.5 \quad \text{vs.} \quad H_1 : \mu < 163.5$.
  - $p = P(Z < z_0) = P(Z < 1.78) = 0.9625 \gg 0.05$
  - A larger p-value indicates less evidence against H_0.

Type I and Type II errors

	Truth	
Decision	H_0 true	H_0 false
Retain H_0	Correct retention	Type II error
Reject H_0	Type I error	Correct rejection

Type I error: we reject H_0, when in fact H_0 is true.

- $H_1 : \mu > 163.5$: We conclude that the average height of female students is greater than 163.5 cm, when in fact the average height is 163.5 cm.
- $H_1 : \mu \neq 163.5$: We conclude that the average height of female students is not 163.5 cm, when in fact the average height is 163.5 cm.

Type II error: we fail to reject H_0, when in fact H_0 is false.

- $H_1 : \mu > 163.5$: We are unable to conclude that the average height of female students is greater than 163.5 cm, when in fact it is greater than 163.5 cm.
- $H_1 : \mu \neq 163.5$: We are unable to conclude that the average height of female students is different from 163.5 cm, when in fact it is not

Type I and Type II errors

Two properties of a test:

- Significance level $= P(\text{Type I error}) =$ $P(\text{reject } H_0 | H_0 \text{ is true}) = \alpha$
- Power $= 1 - P(\text{Type II error}) = P(\text{reject } H_0 | H_0 \text{ is false}) = \beta$

Type I and Type II errors

Two properties of a test:

- Significance level $=$ P(Type I error) $=$ P(reject $H_0 | H_0$ is true) $= \alpha$
- Power $= 1 -$ P(Type II error) $=$ P(reject $H_0 | H_0$ is false) $= \beta$

- A good test is expected to have low type I and II error rates. However, they can not be minimized at the same time.
  - If we 100% reject all H_0, we have $\beta = 1$ (P(Type II error) $= 0$), but may have a large α.
  - If we 100% accept all H_0, we have $\alpha = 0$, but may have a small β (high P(Type II error)).

Type I and Type II errors

Two properties of a test:

- Significance level $=$ P(Type I error) $=$ P(reject $H_0 | H_0$ is true) $= \alpha$
- Power $= 1 -$ P(Type II error) $=$ P(reject $H_0 | H_0$ is false) $= \beta$

- A good test is expected to have low type I and II error rates. However, they can not be minimized at the same time.
  - If we 100% reject all H_0, we have $\beta = 1$ (P(Type II error) $= 0$), but may have a large α.
  - If we 100% accept all H_0, we have $\alpha = 0$, but may have a small β (high P(Type II error)).

- In practice, we control P(Type I error) at a small level (e.g. $\alpha = 0.05$) and prefer a test with high power (low P(Type II error)).

Example: student height (sample size)

Let X denote the height (cm) of female students at uOttawa. Assume that $X \sim N(\mu, 5^2)$. A random sample of size n is drawn from X with a sample mean $\bar{x} = 166.86$. Can we claim that $\mu \neq 163.5$ when $n = 7$? Can we claim that $\mu \neq 163.5$ when $n = 20$? (significance level $\alpha = 0.05$)

1. Setup hypothesis:

$$H_0 : \mu = 163.5 \quad \text{vs.} \quad H_1 : \mu \neq 163.5$$

2. Compute the observed value of test statistic:

   - When $n = 7$,

$$z_0 = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}} = \frac{166.86 - 163.5}{5/\sqrt{7}} \approx 1.78$$

   - When $n = 20$,

$$z_0 = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}} = \frac{166.86 - 163.5}{5/\sqrt{20}} \approx 3.01$$

Effect of sample size

3 p-value:

- When $n = 7$,
  $p = P(Z > 1.78 \text{ or } Z < -1.78) = 2P(Z < -1.78) = 0.075$
- When $n = 20$,
  $p = P(Z > 3.01 \text{ or } Z < -3.01) = 2P(Z < -3.01) = 0.0026$

4 Conclusion:

- When $n = 7$, $p = 0.075 > 0.05$, we fail to reject H_0 and do not have enough evidence to conclude $\mu \neq 163.5$.
- When $n = 20$, $p = 0.0026 < 0.05$, we reject H_0 and conclude $\mu \neq 163.5$.

Effect of sample size

3 p-value:

- When $n = 7$,
  $p = P(Z > 1.78 \text{ or } Z < -1.78) = 2P(Z < -1.78) = 0.075$
- When $n = 20$,
  $p = P(Z > 3.01 \text{ or } Z < -3.01) = 2P(Z < -3.01) = 0.0026$

4 Conclusion:

- When $n = 7$, $p = 0.075 > 0.05$, we fail to reject H_0 and do not have enough evidence to conclude $\mu \neq 163.5$.
- When $n = 20$, $p = 0.0026 < 0.05$, we reject H_0 and conclude $\mu \neq 163.5$.

- Under a given confidence level, a larger sample size increase the power of a test in detecting the mean change/difference.

Effect of sample size

Assume that $X \sim N(\mu, 5^2)$ with $\mu = 163$. Suppose $\bar{x} = 162.9$ with $n = 2000$.

1. $H_0 : \mu = 163.1 \quad \text{vs.} \quad H_1 : \mu < 163.1$
2. $z_0 = z_0 = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}} = \frac{162.9 - 163.1}{5/\sqrt{2000}} \approx -1.79$
3. $p = P(Z < -1.79) = 0.037 < 0.05$
4. We reject H_0 and conclude that $\mu < 163.1$.

Effect of sample size

Assume that $X \sim N(\mu, 5^2)$ with $\mu = 163$. Suppose $\bar{x} = 162.9$ with $n = 2000$.

1. $H_0 : \mu = 163.1 \quad$ vs. $\quad H_1 : \mu < 163.1$
2. $z_0 = z_0 = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}} = \frac{162.9 - 163.1}{5/\sqrt{2000}} \approx -1.79$
3. $p = P(Z < -1.79) = 0.037 < 0.05$
4. We reject H_0 and conclude that $\mu < 163.1$.

- When n is very large, even a little difference between μ_0 and μ is detectable. Here $\mu_0 - \mu = 0.1$.
- We have statistical evidence to reject $H_0 : \mu = 163.1$; this does **not** mean μ is practically far from 163.1.

Testing population mean (σ^2 unknown)

Let $X \sim N(\mu, \sigma^2)$ with an unknown μ and an **unknown** σ. Let $\overline{X}$ and S be the sample mean and sample standard deviation based on a random sample of size n.

1. Setup hypothesis:

$$H_0 : \mu = \mu_0 \quad \text{vs.} \quad H_1 : \mu < \mu_0$$

If H_0 is true,

$$T = \frac{\overline{X} - \mu_0}{S/\sqrt{n}} \sim T_{n-1}$$

2. Compute the observed value for T:

$$t_0 = \frac{\bar{x} - \mu_0}{s/\sqrt{n}}$$

3. Check whether t_0 is an **unusual** value for T_{n-1}: $P(T_{n-1} < t_0) = p$.

4. Let α be a given significance level. If $p < \alpha$, we reject H_0 and conclude H_1 is true. If $p > \alpha$, we fail to reject H_0 and do not have enough evidence to justify H_1.

Example: student height (one-sided)

Let X denote the height (cm) of female students at uOttawa. Assume that $X \sim N(\mu, \sigma^2)$. A random sample of size 7 is drawn from X with a sample mean $\bar{x} = 166.86$ and sample sd $s = 5.2$. Can we claim that $\mu > 163.5$? (significance level $\alpha = 0.05$)

1. Setup hypothesis:

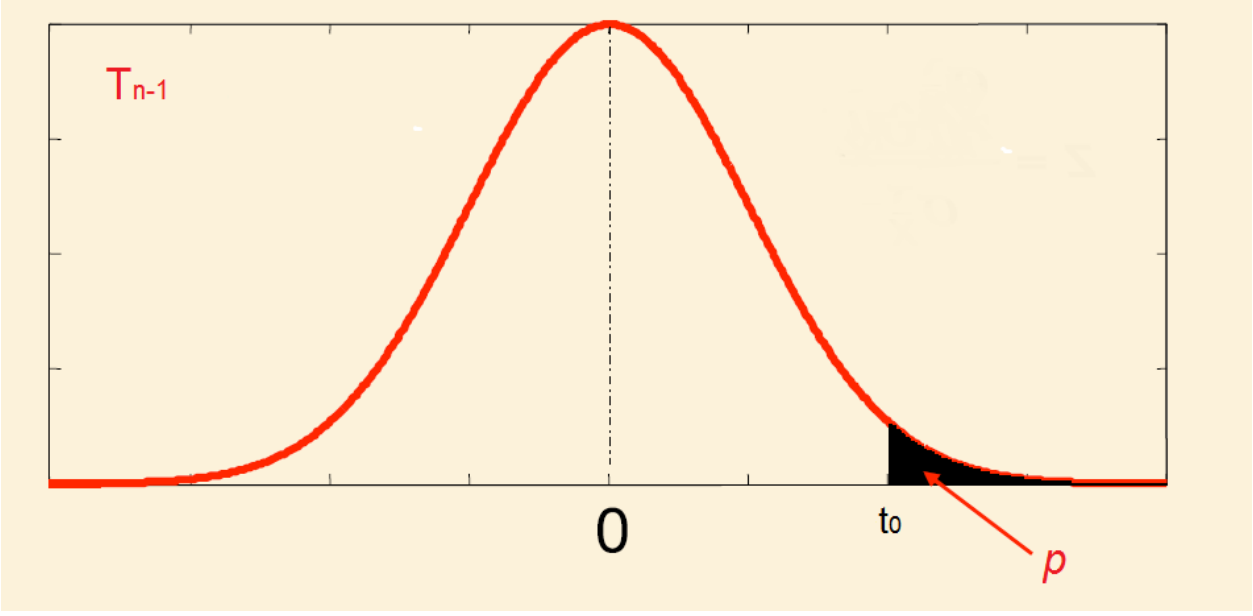
$$H_0 : \mu = 163.5 \quad \text{vs.} \quad H_1 : \mu > 163.5$$

   - H_0: the average height of female students is 163.5 cm.
   - H_1: the average height of female students is greater than 163.5 cm.

2. Compute the observed value of test statistic:

$$t_0 = \frac{\bar{x} - \mu_0}{s/\sqrt{n}} = \frac{166.86 - 163.5}{5.2/\sqrt{7}} \approx 1.71$$

Example: t-test (one-sided)



3 Find the p-value (using T-table):

$$0.1 > P(T_6 > t_0) = P(T_6 > 1.71) > 0.05$$

4 Make a conclusion: Since the p-value is larger than $\alpha = 0.05$. We fail to reject H_0 and do not have enough evidence conclude that $\mu < 163.3$ cm.

Example: student height (two-sided)

Let X denote the height (cm) of female students at uOttawa. Assume that $X \sim N(\mu, \sigma^2)$. A random sample of size 7 is drawn from X with a sample mean $\bar{x} = 166.86$ and sample sd $s = 5.2$. Can we claim that $\mu \neq 163.5$? (significance level $\alpha = 0.05$)

1. Setup hypothesis:

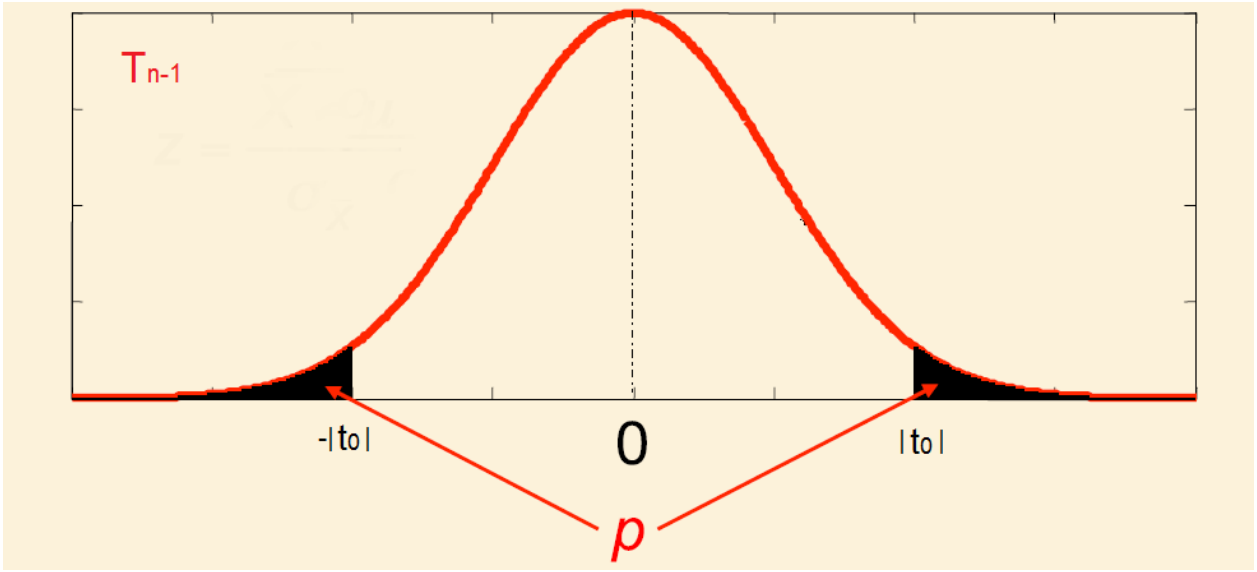
$$H_0 : \mu = 163.5 \quad \text{vs.} \quad H_1 : \mu \neq 163.5$$

   - H_0: the average height of female students is 163.5 cm.
   - H_1: the average height of female students is different from 163.5 cm.

2. Compute the observed value of test statistic:

$$t_0 = \frac{\bar{x} - \mu_0}{s/\sqrt{n}} = \frac{166.86 - 163.5}{5.2/\sqrt{7}} \approx 1.71$$

Example: t-test (two-sided)



3 Find the p-value (using T-table):

$$0.2 > P(T_6 > |t_0| \text{ or } T_6 < -|t_0|) = 2P(T_6 > 1.71) > 0.1$$

4 Make a conclusion: Since the p-value is larger than $\alpha = 0.05$. We fail to reject H_0 and do not have enough evidence conclude that $\mu < 163.3$ cm.

σ known vs. σ unknown ($n = 7$)

If $\sigma = 5$ is known,

1. $H_0 : \mu = 163.5 \quad \text{vs.} \quad H_1 : \mu > 163.5$
2. $z_0 = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}} = \frac{166.86 - 163.5}{5/\sqrt{7}} \approx 1.78$
3. $p = P(Z > z_0) = P(Z > 1.78) = 0.0375$
4. Since $0.0375 < \alpha = 0.05$, we reject H_0 and conclude $\mu > 163.5$.

If σ is unknown,

1. $H_0 : \mu = 163.5 \quad \text{vs.} \quad H_1 : \mu > 163.5$
2. $t_0 = \frac{\bar{x} - \mu_0}{s/\sqrt{n}} = \frac{166.86 - 163.5}{5.2/\sqrt{7}} \approx 1.71$
3. $p = P(T_6 > t_0) = P(T_6 > 1.71) = 0.07$
   $(P(Z > 1.71) = 0.044)$
4. Since $0.07 > \alpha = 0.05$, we fail to reject H_0 and do not have enough
   evidence to conclude $\mu > 163.5$.

σ known vs. σ unknown

- In general, a t-test is less powerful than a z-test in detecting a mean change/difference.
- When σ is unknown, we need to estimate it first. The cost for estimating σ is paid by the power of the test.
- If you conduct a z-test when σ is unknown, the Type I error rate of the test tends to be larger than the given α.
- When n is large (e.g. $n > 30$), a t-test would be very close to a z-test; a large sample size also helps to relax the normality assumption on X for using a z-test.

Large sample case: $n = 50$, $\bar{x} = 164$, $s = 5.1$

If $\sigma = 5$ is known,

1. $H_0 : \mu = 165.5 \quad \text{vs.} \quad H_1 : \mu < 165.5$
2. $z_0 = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}} = \frac{164 - 165.5}{5/\sqrt{50}} \approx -2.12$
3. $p = P(Z < z_0) = P(Z < -2.12) = 0.017$
4. Since $0.017 < \alpha = 0.05$, we reject H_0 and conclude $\mu < 165.5$.

If σ is unknown,

1. $H_0 : \mu = 165.5 \quad \text{vs.} \quad H_1 : \mu < 165.5$
2. $t_0 = \frac{\bar{x} - \mu_0}{s/\sqrt{n}} = \frac{164 - 163.5}{5.1/\sqrt{50}} \approx -2.08$
3. $p = P(T_{49} < -2.08) = 0.021 \quad (P(Z < -2.08) = 0.019)$
4. Since $0.021 < \alpha = 0.05$, we reject H_0 and conclude $\mu < 165.5$.

Exercise

Assume $X \sim N(\mu, \sigma^2)$. Suppose a sample of size 10 is drawn with $\bar{x} = 16$ and $s = 3$. Can we conclude that $\mu < 18$ at $\alpha = 0.05$?

Exercise

Assume $X \sim N(\mu, \sigma^2)$. Suppose a sample of size 10 is drawn with $\bar{x} = 16$ and $s = 3$. Can we conclude that $\mu < 18$ at $\alpha = 0.05$?

1. Setup hypothesis:

$$H_0 : \mu = 18 \quad \text{vs.} \quad H_1 : \mu < 18$$

Exercise

Assume $X \sim N(\mu, \sigma^2)$. Suppose a sample of size 10 is drawn with $\bar{x} = 16$ and $s = 3$. Can we conclude that $\mu < 18$ at $\alpha = 0.05$?

1. Setup hypothesis:

$$H_0 : \mu = 18 \quad \text{vs.} \quad H_1 : \mu < 18$$

2. Compute the observed value for T:

$$t_0 = \frac{\bar{x} - \mu_0}{s/\sqrt{n}} = \frac{16 - 18}{3/\sqrt{10}} \approx -2.11$$

Exercise

Assume $X \sim N(\mu, \sigma^2)$. Suppose a sample of size 10 is drawn with $\bar{x} = 16$ and $s = 3$. Can we conclude that $\mu < 18$ at $\alpha = 0.05$?

1. Setup hypothesis:

$$H_0 : \mu = 18 \quad \text{vs.} \quad H_1 : \mu < 18$$

2. Compute the observed value for T:

$$t_0 = \frac{\bar{x} - \mu_0}{s/\sqrt{n}} = \frac{16 - 18}{3/\sqrt{10}} \approx -2.11$$

3. p-value: $p = P(T_9 < t_0) = P(T_9 > -t_0) = P(T_9 > 2.11)$. Using T-table, we find $0.025 < p < 0.05$.

Exercise

Assume $X \sim N(\mu, \sigma^2)$. Suppose a sample of size 10 is drawn with $\bar{x} = 16$ and $s = 3$. Can we conclude that $\mu < 18$ at $\alpha = 0.05$?

1. Setup hypothesis:

$$H_0 : \mu = 18 \quad \text{vs.} \quad H_1 : \mu < 18$$

2. Compute the observed value for T:

$$t_0 = \frac{\bar{x} - \mu_0}{s/\sqrt{n}} = \frac{16 - 18}{3/\sqrt{10}} \approx -2.11$$

3. p-value: $p = P(T_9 < t_0) = P(T_9 > -t_0) = P(T_9 > 2.11)$. Using T-table, we find $0.025 < p < 0.05$.

4. Conclusion: Since $p < \alpha = 0.05$, we reject H_0 and conclude $\mu < 18$.

Summary

- An one-sided test is more powerful than a two-side test in one particular direction.
- A t-test is less powerful than a z-test.
- A larger sample size helps to increase the power of a test.

Testing Differences of Means

uOttawa - MAT2377

Fall 2020

Goals

Practice hypothesis tests for differences of means.

Comparison of Two Samples

Scientists often want to test the effect of a treatment by comparing groups under different experimental conditions.

- In botany, researchers want to know whether a new fertilizer produces taller plant.
- In zoology, researchers want to know whether low temperature affects the wing size of a certain fruit fly.
- In clinical studies, researchers want to know whether a new medicine is effective for treating hypertension.

When comparing the results from random experiments, scientists need statistical tools to make a conclusion.

Paired sample problem

A new drug is developed for reducing the level of blood cholesterol. A sample of 10 people take this drug for two weeks. Their blood cholesterol are measured (in mg/dl) before and after the treatment. The data are as follows. plant

Before	13.1	12.3	10.0	17.7	19.4	10.1	11.5	12.6	9.5	12.1
After	12.0	7.5	11.7	12.5	18.6	12.3	15.2	16.3	10.7	9.8

Let X, Y denote the level of blood cholesterol before and after the treatment respectively. Let $E(X) = \mu_x$ and $E(Y) = \mu_y$. Do we have enough evidence to claim that the new drug is effective (i.e. $\mu_x > \mu_y$) ?

- $H_0 : \mu_x = \mu_y, \quad$ vs. $\quad H_1 : \mu_x > \mu_y$

Paired sample problem

Let $d_i = x_i - y_i$.

x_i	13.1	12.3	10.0	17.7	19.4	10.1	11.5	12.6	9.5	12.1
y_i	12.0	7.3	11.7	12.5	18.6	12.3	15.2	16.3	10.7	9.8
$\mathbf{d_i}$	**1.1**	**5.0**	**-1.7**	**5.2**	**0.8**	**-2.2**	**-3.7**	**-3.7**	**-1.2**	**2.3**

- $\bar{X} - \bar{Y} = 0.19$, and $s^2_{x-y} = 3.26$.
- $\mu_{X-Y} = \mu_x - \mu_y$
- To check $\mu_x > \mu_y$, it is equivalent to test

$$H_0 : \mu_{x-y} = 0, \quad \text{vs.} \quad H_1 : \mu_{x-y} > 0$$

Paired t-test

Assume that D is normally distributed.

1. Setup hypothesis:

$$H_0 : \mu_{x-y} = 0 \quad \text{vs.} \quad H_1 : \mu_{x-y} > 0$$

   If H_0 is true,

$$T = \frac{\overline{X} - \overline{Y}}{S_{x-y}/\sqrt{n}} \sim T_{n-1}$$

2. Compute the observed value t_0 for T.
3. Find p-value: $P(T > t_0) = p$. (Check whether t_0 is an **unusual** value)
4. Conclusion: If $p < \alpha$, we reject H_0 and conclude H_1 is true. If $p > \alpha$, we fail to reject H_0 and do not have enough evidence to justify H_1.

Paired t-test

In a paired-sample problem,

- The observations come in pairs.
  - measurements made on the same individuals before and after a treatment.
  - measurements made on the same individuals using two different treatments.
- X and Y are dependent.
  - Let X and Y denote the body weight (lb) before and after a diet program. Suppose the program is successful in reducing 20 lbs in 3 months.

x_i	160	200	300
y_i	140	180	280

  - $Var(X - Y)$ is unknown. $Var(X - Y) \neq Var(X) + Var(Y)$.

Unpaired samples

A new fertilizer is developed to increase the height of plants at maturity. Researchers assign 16 seedlings to two plots (8 each), and use this fertilizer in plot 1. After 8 months, the plants are measured at maturity (in cm). The data are as follows.

plot 1	44.1	47.8	59.5	54.4	49.5	44.5	37.7	56.6
plot 2	47.4	35.0	35.7	50.2	43.4	51.1	40.6	36.7

Let X_1, X_2 denote the height of plants in plots 1 and 2 respectively. Let $E(X_1) = \mu_1$ and $E(X_2) = \mu_2$. Do we have enough evidence to claim that on average the plants in plot 1 are taller than the plants in plot 2 ?

- $H_0 : \mu_1 = \mu_2, \quad$ vs. $\quad H_1 : \mu_1 > \mu_2$

Unpaired samples

x_1	44.1	47.8	59.5	54.4	49.5	44.5	37.7	56.6
x_2	47.4	35.0	35.7	50.2	43.4	51.1	40.6	36.7

Let $D = X_1 - X_2$ and $\mu_d = E(D)$. To check $\mu_1 > \mu_2$, it is equivalent to test

$$H_0 : \mu_d = 0, \quad \text{vs.} \quad H_1 : \mu_d > 0$$

Unpaired samples

x_1	44.1	47.8	59.5	54.4	49.5	44.5	37.7	56.6
x_2	47.4	35.0	35.7	50.2	43.4	51.1	40.6	36.7

Let $D = X_1 - X_2$ and $\mu_d = E(D)$. To check $\mu_1 > \mu_2$, it is equivalent to test

$$H_0 : \mu_d = 0, \quad \text{vs.} \quad H_1 : \mu_d > 0$$

- Since x_1 and x_2 are **unpaired** measurements, it is meaningless to take $d_i = x_{1i} - x_{2i}$. blood
- We do not have appropriate data to conduct one-sample t-test.
- In this example, it is OK to assume X_1 and X_2 are independent.

Review: expectation and variance

Let X_1 and X_2 be two random variables. Let a, b be two constants.

- $E(aX_1 + bX_2) = aE(X_1) + bE(X_2)$
  - $E(X_1 + X_2) = E(X_1) + E(X_2)$.
  - $E(X_1 - X_2) = E(X_1) - E(X_2)$.
- $Var(aX_1 + b) = a^2 Var(X_1)$
  - $Var(2X_1 + 1) = 2^2 Var(X_1) = 4Var(X_1)$.
  - $Var(-2X_1 - 3) = (-2)^2 Var(X_1) = 4Var(X_1)$.

Review: expectation and variance

Let X_1 and X_2 be two random variables. Let a, b be two constants.

- $E(aX_1 + bX_2) = aE(X_1) + bE(X_2)$
  - $E(X_1 + X_2) = E(X_1) + E(X_2)$.
  - $E(X_1 - X_2) = E(X_1) - E(X_2)$.
- $Var(aX_1 + b) = a^2 Var(X_1)$
  - $Var(2X_1 + 1) = 2^2 Var(X_1) = 4Var(X_1)$.
  - $Var(-2X_1 - 3) = (-2)^2 Var(X_1) = 4Var(X_1)$.
- When X_1 and X_2 are independent (uncorrelated),

$$Var(aX_1 + bX_2) = a^2 Var(X_1) + b^2 Var(X_2)$$

  - $Var(X_1 + X_2) = 1^2 Var(X_1) + 1^2 Var(X_2) = Var(X_1) + Var(X_2)$
  - $Var(X_1 - X_2) = 1^2 Var(X_1) + (-1)^2 Var(X_1) = Var(X_1) + Var(X_2)$

Review: normal distribution

Suppose $X \sim N(\mu, \sigma^2)$. Let $\{x_1, \ldots, x_n\}$ be a random sample of X.

- Standardization:

$$\frac{X - \mu}{\sigma} \sim N(0, 1)$$

- Let $\overline{X} = \frac{1}{n} \sum X_i$ be the sample mean. We know

$$E(\overline{X}) = \mu, \quad Var(\overline{X}) = \sigma^2/n$$

$$\overline{X} \sim N(\mu,\ \sigma^2/n) \quad \text{and} \quad \frac{\overline{X} - E(\overline{X})}{sd(\overline{X})} = \frac{\overline{X} - \mu}{\sigma/\sqrt{n}} \sim N(0, 1)$$

Review: sample mean

Let $\{x_{1i}\}$ be a random sample of size n_1 from X_1 and $\{x_{2i}\}$ be a random sample of size n_2 from X_2. Let $E(X_1) = \mu_1$, $Var(X_1) = \sigma_1^2$ and $E(X_2) = \mu_2$, $Var(X_2) = \sigma_2^2$.

- When X_1 and X_2 are normally distributed, we have

$$\overline{X}_1 \sim N(\mu_1,\ \sigma_1^2/n_1), \quad \overline{X}_2 \sim N(\mu_2,\ \sigma_2^2/n_2)$$

- When X_1 and X_2 are independent,

$$\overline{D} = \overline{X}_1 - \overline{X}_2 \sim N(\delta, \tau^2),$$

where

$$\delta = E(\overline{D}) = E(\overline{X}_1 - \overline{X}_2) = \mu_1 - \mu_2$$

$$\tau^2 = Var(\overline{D}) = Var(\overline{X}_1 - \overline{X}_2) = Var(\overline{X}_1) + Var(\overline{X}_2) = \sigma_1^2/n_1 + \sigma_2^2/n_2$$

Review: CLT

Let $\{x_{1i}\}$ be a random sample of size n_1 from X_1 and $\{x_{2i}\}$ be a random sample of size n_2 from X_2. Let $E(X_1) = \mu_1$, $Var(X_1) = \sigma_1^2$ and $E(X_2) = \mu_2$, $Var(X_2) = \sigma_2^2$. Assume X_1 and X_2 are **independent**.

- Suppose X_1 and X_2 are normally distributed. We have

$$\frac{\overline{D} - \delta}{\tau} = \frac{\overline{X}_1 - \overline{X}_2 - (\mu_1 - \mu_2)}{\sqrt{\sigma_1^2/n_1 + \sigma_2^2/n_2}} \sim N(0,1)$$

- When n_1 and n_2 are **both** large (> 30), we **approximately** have

$$\frac{\overline{X}_1 - \overline{X}_2 - (\mu_1 - \mu_2)}{\sqrt{\sigma_1^2/n_1 + \sigma_2^2/n_2}} \sim N(0,1), \quad \frac{\overline{X}_1 - \overline{X}_2 - (\mu_1 - \mu_2)}{\sqrt{S_1^2/n_1 + S_2^2/n_2}} \sim N(0,1)$$

Two-sample z-test

Suppose that (1) X_1 and X_2 are normally distributed or (2) n_1 and n_2 are large. Assume X_1 and X_2 are independent.

1. Setup hypothesis:

$$H_0 : \mu_1 - \mu_2 = \delta_0 \quad \text{vs.} \quad H_1 : \mu_1 - \mu_2 > \delta_0$$

   If H_0 is true,

$$Z = \frac{\overline{X}_1 - \overline{X}_2 - \delta_0}{\sqrt{\sigma_1^2/n_1 + \sigma_2^2/n_2}} \sim N(0,1)$$

2. Compute the observed value for Z:

$$z_0 = \frac{\bar{x}_1 - \bar{x}_2 - \delta_0}{\sqrt{\sigma_1^2/n_1 + \sigma_2^2/n_2}}$$

3. Find p-value: $P(Z > z_0) = p$. (Check whether z_0 is an **unusual** value)

4. Conclusion: If $p < \alpha$, we reject H_0 and conclude H_1 is true. If $p > \alpha$, we fail to reject H_0 and do not have enough evidence to justify H_1.

Example: plant height

x_1	44.1	47.8	59.5	54.4	49.5	44.5	37.7	56.6
x_2	47.4	35.0	35.7	50.2	43.4	51.1	40.6	36.7

Suppose that X_1 and X_2 are independent and normally distributed with $\bar{x}_1 = 49.3$, $\bar{x}_2 = 42.5$ and $\sigma_1 = 7.5$ and $\sigma_2 = 6.5$. Can we claim $\mu_1 > \mu_2$? ($\alpha = 0.05$)

1. Setup hypothesis:

$$H_0 : \mu_1 - \mu_2 = 0 \quad \text{vs.} \quad H_1 : \mu_1 - \mu_2 > 0$$

2. Compute the observed value of test statistic:

$$z_0 = \frac{\bar{x}_1 - \bar{x}_2 - 0}{\sqrt{\sigma_1^2/n_1 + \sigma_2^2/n_2}} = 1.94$$

Example: plant height

3. Find the p-value: $p = P(Z > 1.94) = P(Z < -1.94) \approx 0.026$
4. Make a conclusion: Since $p < \alpha = 0.05$, we reject H_0 and conclude that $\mu_1 > \mu_2$.

A 95% confidence interval for $\delta = \mu_1 - \mu_2$ is given by

$$\begin{aligned} \bar{x}_1 - \bar{x}_2 \pm 1.96\sqrt{\frac{\sigma_1^2}{n} + \frac{\sigma_2^2}{n}} &= 49.3 - 42.5 \pm 1.96\sqrt{\frac{7.5^2}{8} + \frac{6.5^2}{8}} \\ &= [-0.08, 13.7] \end{aligned}$$

We are 95% confident that δ falls in [-0.08, 13.7]. Since $\delta_0 = 0 \in [-0.08, 13.7]$, we fail to reject H_0.

7.5: Testing Proportions

uOttawa - MAT2377

Fall 2020

Goals

Use the hypothesis testing framework to study proportions.

Review: Testing population mean (σ^2 known)

Let $X \sim N(\mu, \sigma^2)$ with an unknown mean μ and a **known** standard deviation σ. Let α be a given significance level.

1. Setup hypothesis:

$$H_0 : \mu = \mu_0 \quad \text{vs.} \quad H_1 : \mu < \mu_0$$

If H_0 is true,

$$Z = \frac{\overline{X} - \mu_0}{\sigma/\sqrt{n}} \sim N(0, 1)$$

2. Compute the observed value for Z:

$$z_0 = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}}$$

3. Find p-value: $P(Z < z_0) = p$. (Check whether z_0 is an **unusual** value)

4. Conclusion: If $p < \alpha$, we reject H_0 and conclude H_1 is true. If $p > \alpha$, we fail to reject H_0 and do not have enough evidence to justify H_1.

Testing a population proportion

Let p be the proportion of individuals who share a common characteristic in a given population. Recall that we estimate p by

$$\hat{p} = \frac{Y}{n},$$

where Y is the number of individuals with the desired characteristic in a sample of size n.

Testing a population proportion

Let p be the proportion of individuals who share a common characteristic in a given population. Recall that we estimate p by

$$\hat{p} = \frac{Y}{n},$$

where Y is the number of individuals with the desired characteristic in a sample of size n.

- Example: p is the proportion of MAT 2379 students who attend today's lecture. We take section B students as a sample with size $n = 150$ and estimate p by Y/n, where Y is the number of students in this classroom.

Testing a population proportion

Let p be the proportion of individuals who share a common characteristic in a given population. Recall that we estimate p by

$$\hat{p} = \frac{Y}{n},$$

where Y is the number of individuals with the desired characteristic in a sample of size n.

- Example: p is the proportion of MAT 2379 students who attend today's lecture. We take section B students as a sample with size $n = 150$ and estimate p by Y/n, where Y is the number of students in this classroom.
- You may treat p as the success rate and Y as the number of "successes" over n independent "yes/no" trails. What distribution does Y follow?

Testing a population proportion

Since $Y \sim Binomial(n, p)$, $E(Y) = np$ and $Var(Y) = np(1-p)$. This implies that

$$\begin{aligned} E(\hat{p}) &= E(\frac{Y}{n}) = \frac{1}{n}E(Y) = \frac{np}{n} = p \\ Var(\hat{p}) &= Var(\frac{Y}{n}) = \frac{1}{n^2}Var(Y) = \frac{np(1-p)}{n^2} = \frac{p(1-p)}{n} \end{aligned}$$

When n is large, we approximately have

$$\frac{\hat{p} - E(\hat{p})}{sd(\hat{p})} = \frac{\hat{p} - p}{\sqrt{p(1-p)/n}} \sim N(0, 1).$$

- For proportion, we usually need n to be very large for CLT. One commonly used rule is to require $n\hat{p} > 10$ and $n(1-\hat{p}) > 10$.

Testing a population proportion

1. Setup hypothesis:

$$H_0 : p = p_0 \quad \text{vs.} \quad H_1 : p < p_0$$

If H_0 is true and $np_0 > 10$, $n(1 - p_0) > 10$,

$$Z = \frac{\hat{p} - p_0}{\sqrt{p_0(1 - p_0)/n}} \sim N(0, 1).$$

2. Compute the observed value for Z:

$$z_0 = \frac{\hat{p} - p_0}{\sqrt{p_0(1 - p_0)/n}}$$

3. Check whether z_0 is an **unusual** value for Z: $P(Z < z_0) =$ p-value.

4. Let α be a given significance level. If $p < \alpha$, we reject H_0 and conclude H_1 is true. If $p > \alpha$, we fail to reject H_0 and do not have enough evidence to justify H_1.

Example: seatbelt

In 2001 Youth Risk Behavior survey, 747 out of 1168 female 12th graders said they always use a seatbelt when driving. Let p denote the proportion of 12th grade females who always use a seatbelt when driving. Can we conclude that $p < 0.67$? ($\alpha = 0.05$)

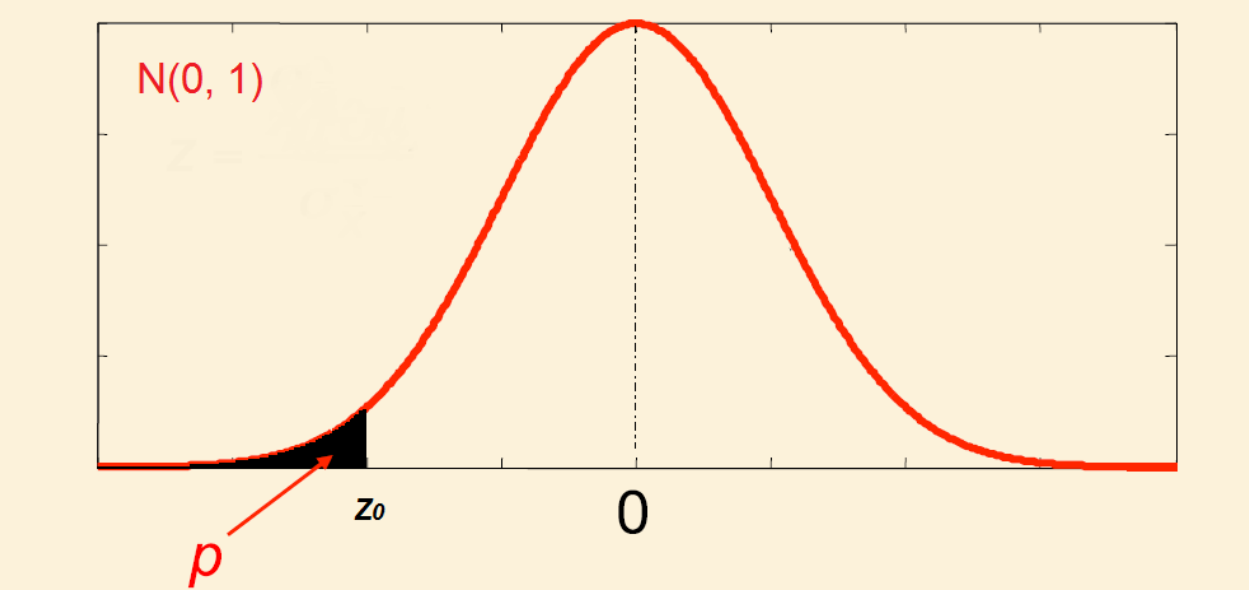
1. Setup hypothesis: $H_0 : p = 0.67$ vs. $H_1 : p < 0.67$

   Since $np_0 = 1168 \times 0.67 = 782.56 > 10$ and $n(1 - p_0) = 1168 \times 0.33 = 385.44 > 10$, it is OK to conduct a z-test for H_0.

2. Compute the observed value of test statistic:

$$z_0 = \frac{\hat{p} - p_0}{\sqrt{p_0(1 - p_0)/n}} = \frac{747/1168 - 0.67}{\sqrt{0.67(1 - 0.67)/1168}} = -2.18$$

Example: seatbelt



3 Find the p-value (using Z-table):

$$P(Z < z_0) = P(Z < -2.18) = 0.015$$

4 Make a conclusion: Since the p-value is smaller than $\alpha = 0.05$. We reject H_0 and conclude that the proportion of 12th grade females who always use a seatbelt when driving is less than 0.67.

7.5: Testing Review

uOttawa - MAT2377

Fall 2020

NHST Solutions

1. Set up framework (choose H_0, H_1, α, etc).
2. Choose *appropriate* confidence interval (or equivalent computation: p-value, etc).
3. *Compute* confidence interval.
4. Make decision.

Note Steps (1,4) are routine; Step (3) is *exactly* from last unit; Step (2) requires some thought!

Choice of CI

1. Paired difference? Unpaired difference? Not a difference?
2. Normal (or large-n)? Proportion? Normal (but small-n)?
3. Left-interval, right-interval, or symmetric interval?

8.1: Correlations and Scatterplots

uOttawa - MAT2377

Fall 2020

Goal

Introduce *correlations* as a warmup for *linear regression.*

Review: Independence

Let A and B be two random events with $P(A > 0)$ and $P(B > 0)$. We say A and B are independent, if

- $P(A \cap B) = P(A)P(B)$ or
- $P(A|B) = P(A)$ or
- $P(B|A) = P(B)$

Let X and Y be two random variables. We say X and Y are independent if

$$P(X \leq a \cap Y \leq b) = P(X \leq a)P(Y \leq b)$$

- Independence makes our life easier; it is also an important condition in statistics.

Association

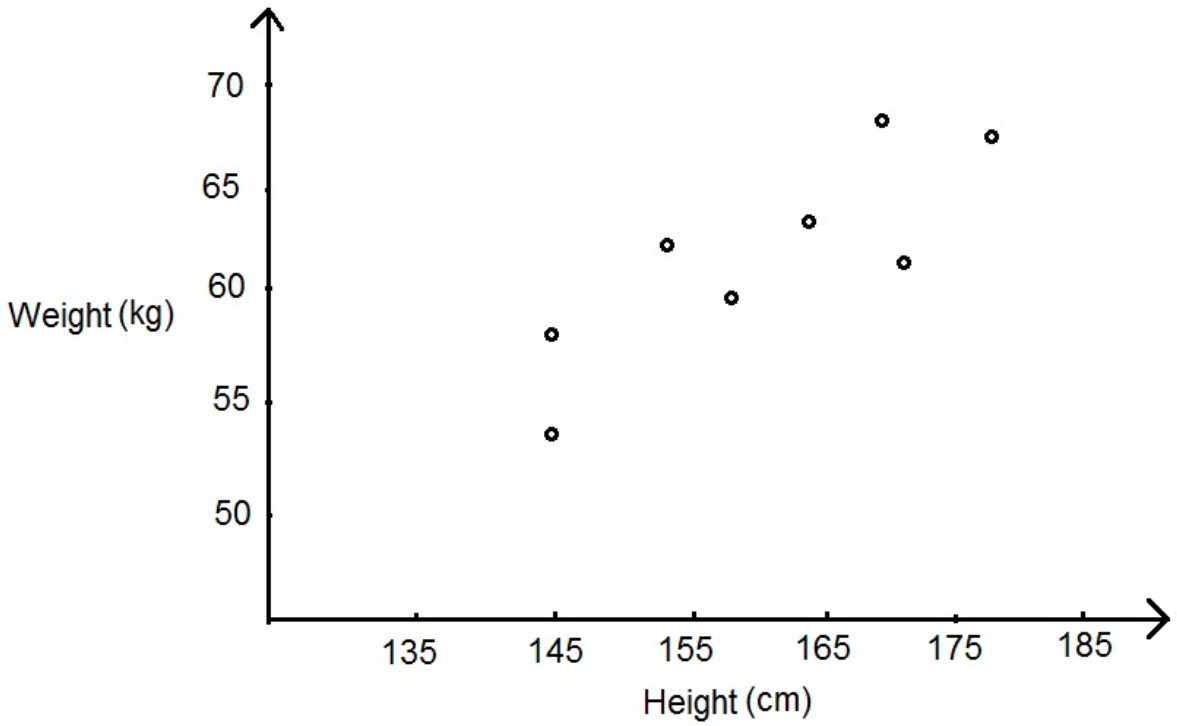
In fact, many things are not independent; they are associated (dependent).

- Height $\leftrightarrow$ Weight
- Years of education $\leftrightarrow$ Income
- Smoking $\leftrightarrow$ Lung cancer
- Vitamin C $\leftrightarrow$ Cold
- Weather $\leftrightarrow$ Electricity demand

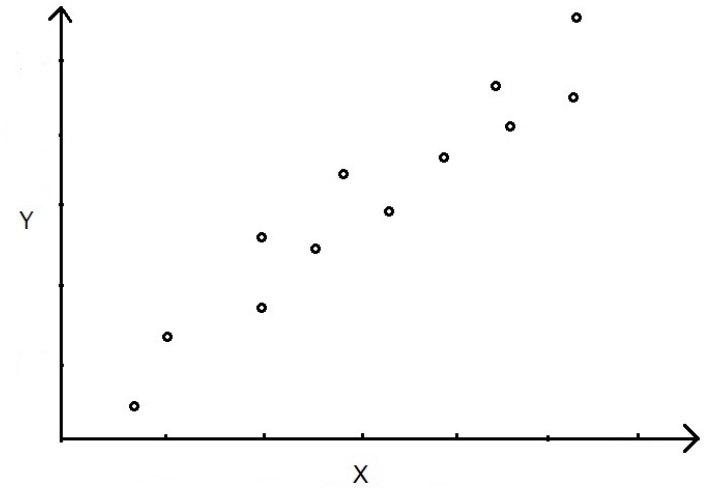
Sometimes, an association between two (random) variables can be detected visually from a **scatter plot**.

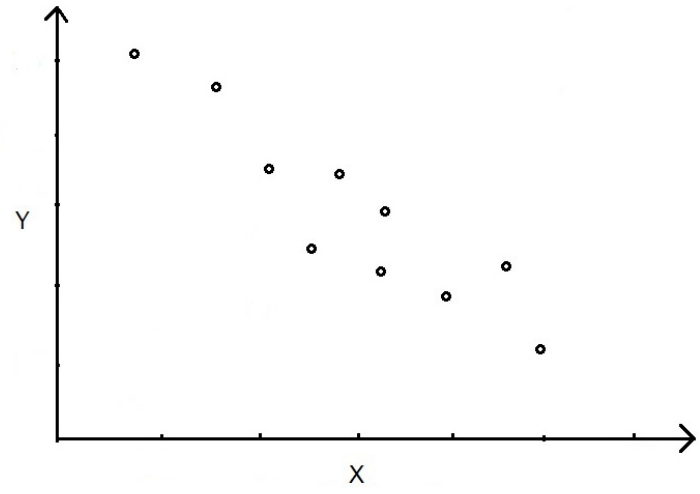
Scatter plot

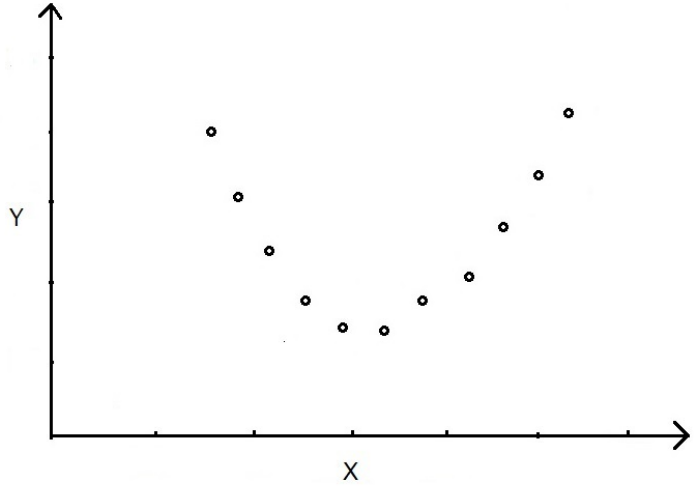
Let $\{(x_1, y_1), \ldots, (x_n, y_n)\}$ be n pairs of observations from random variables X and Y. A **scatter plot** shows each pair of the data as a point using their values as two coordinates.

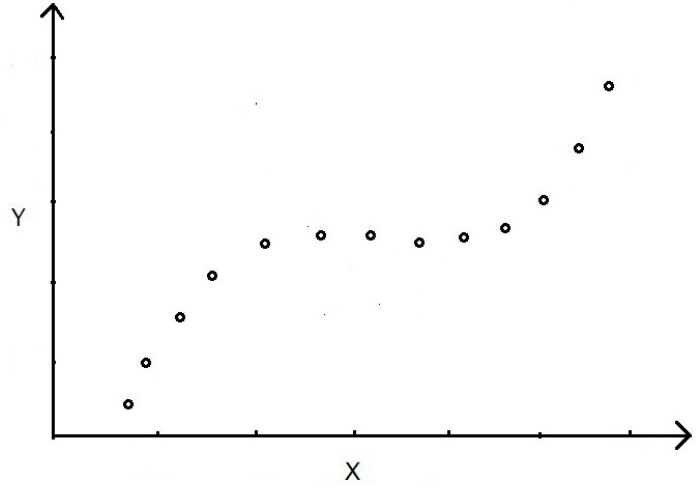


Shape of associations



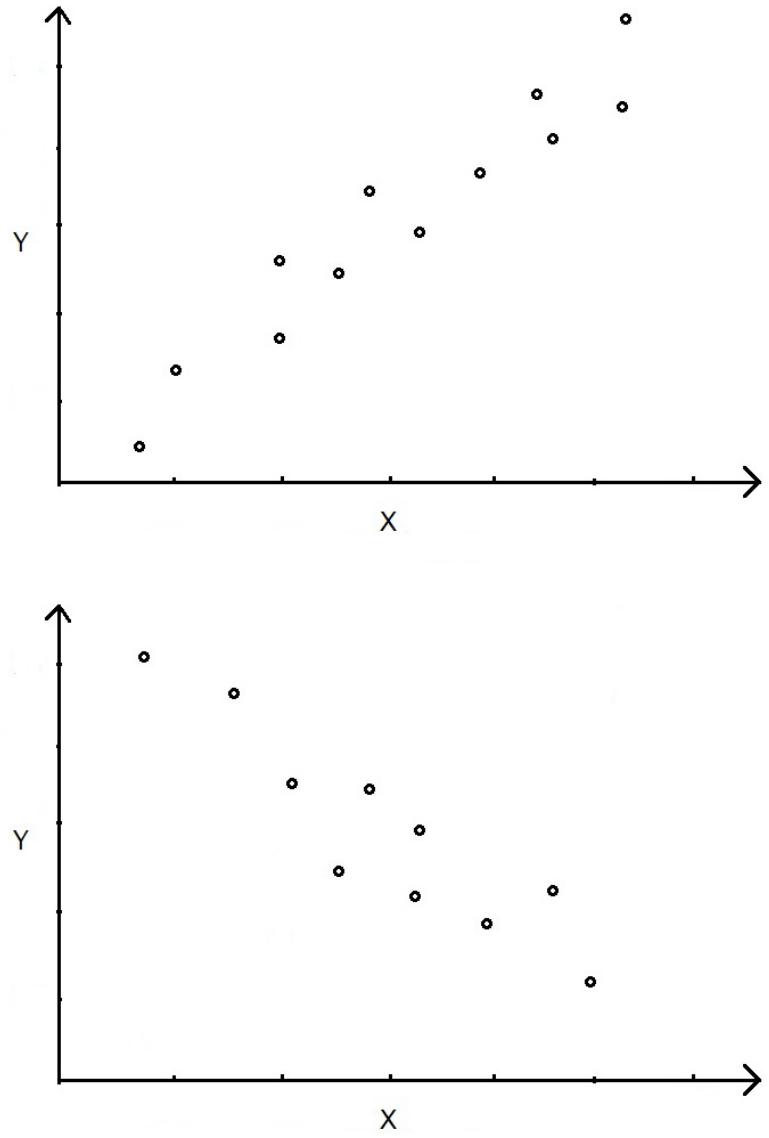






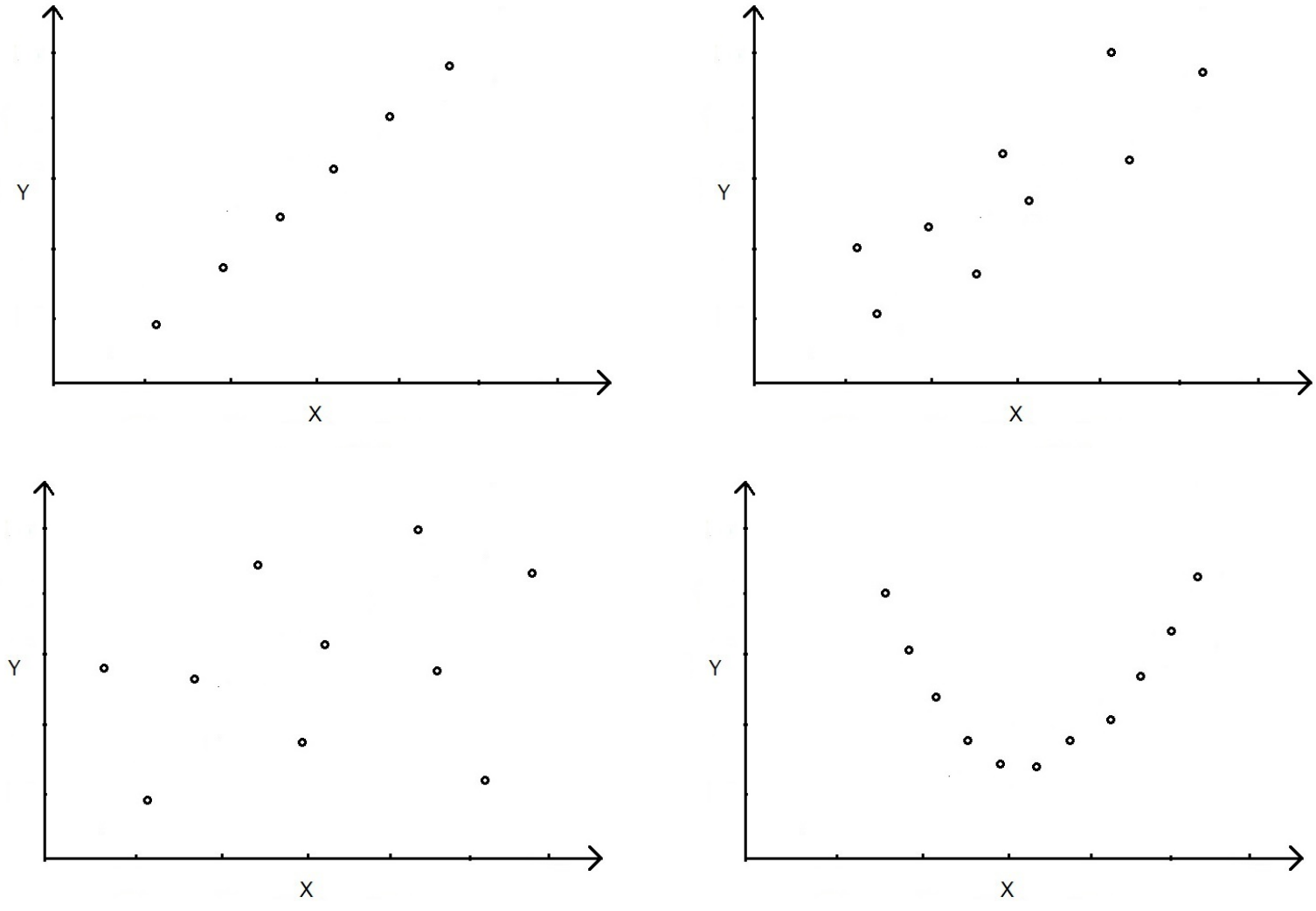
- Linear association vs. Non-linear association

Orientation of associations



- Positive association: larger X indicates larger Y
- Negative association: larger X indicates smaller Y

Intensity of associations



- Strong, Moderate, and Little association

Covariance

Let X and Y be two random variables with $E(X) = \mu_x$and $E(Y) = \mu_y$. To quantify the linear association between X and Y, we use **covariance**

$$Cov(X, Y) = E[(X - \mu_x)(Y - \mu_y)]$$

In practice, we can use **sample covariance** to estimate $Cov(X, Y)$

$$s_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{n - 1}$$

- $Cov(X, X) = E[(X - \mu_x)^2] = Var(X)$.
- In practice, $s_{xy} > 0$ indicates positive association; $s_{xy} < 0$ indicates negative association.

Correlation

Let $Var(X) = \sigma_x^2$ and $Var(Y) = \sigma_y^2$. A linear association can be better described by **correlation**

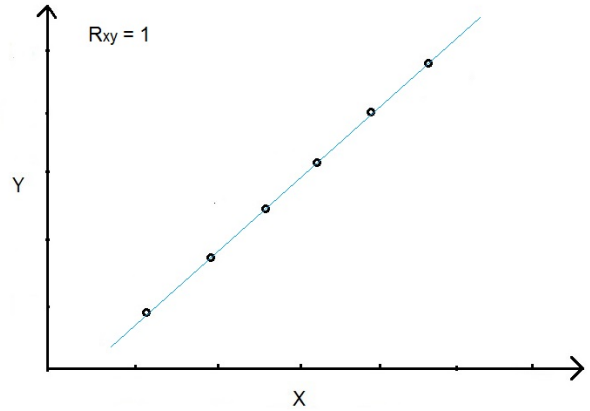
$$Cor(X, Y) = \frac{Cov(X, Y)}{\sigma_x \sigma_y}$$

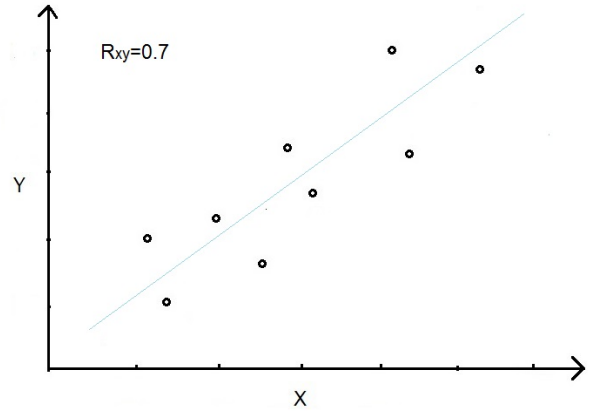
In practice, we can use **sample correlation** to estimate $Cor(X, Y)$

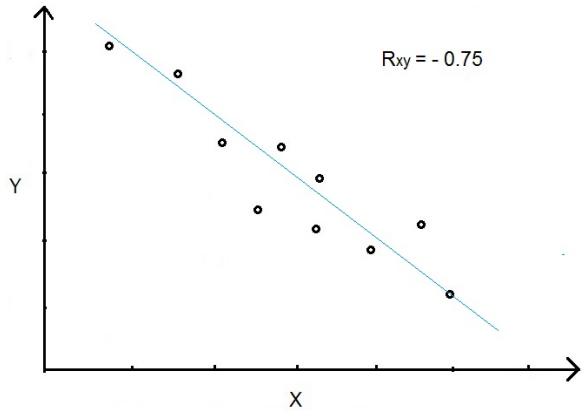
$$r_{xy} = \frac{s_{xy}}{s_x s_y}$$

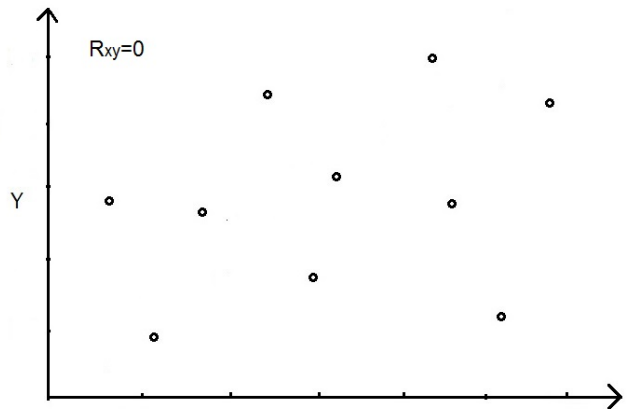
- r_{xy} is called Pearson's correlation, which has the same sign as s_{xy}.
- $-1 \leq r_{xy} \leq 1$. When $|r_{xy}| = 1$, X and Y has a perfect linear association; when $r_{xy} = 0$, X and Y has **no linear association**.
- r_{xy} is invariant to linear transformation.

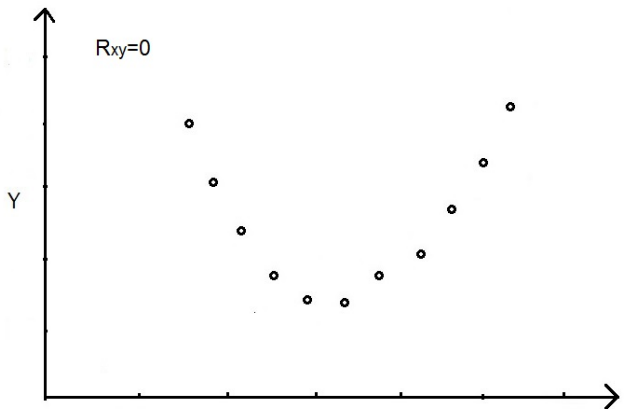
Association: intensity, orientation, shape











Independence

Let A and B be two random events with $P(A > 0)$ and $P(B > 0)$. We say A and B are independent, if

- $P(A \cap B) = P(A)P(B)$ or
- $P(A|B) = P(A)$ or
- $P(B|A) = P(B)$

Let X and Y be two random variables. We say X and Y are independent if

$$P(X \leq a \cap Y \leq b) = P(X \leq a)P(Y \leq b)$$

- Independence makes our life easier; it is also an important condition in statistics.

Association

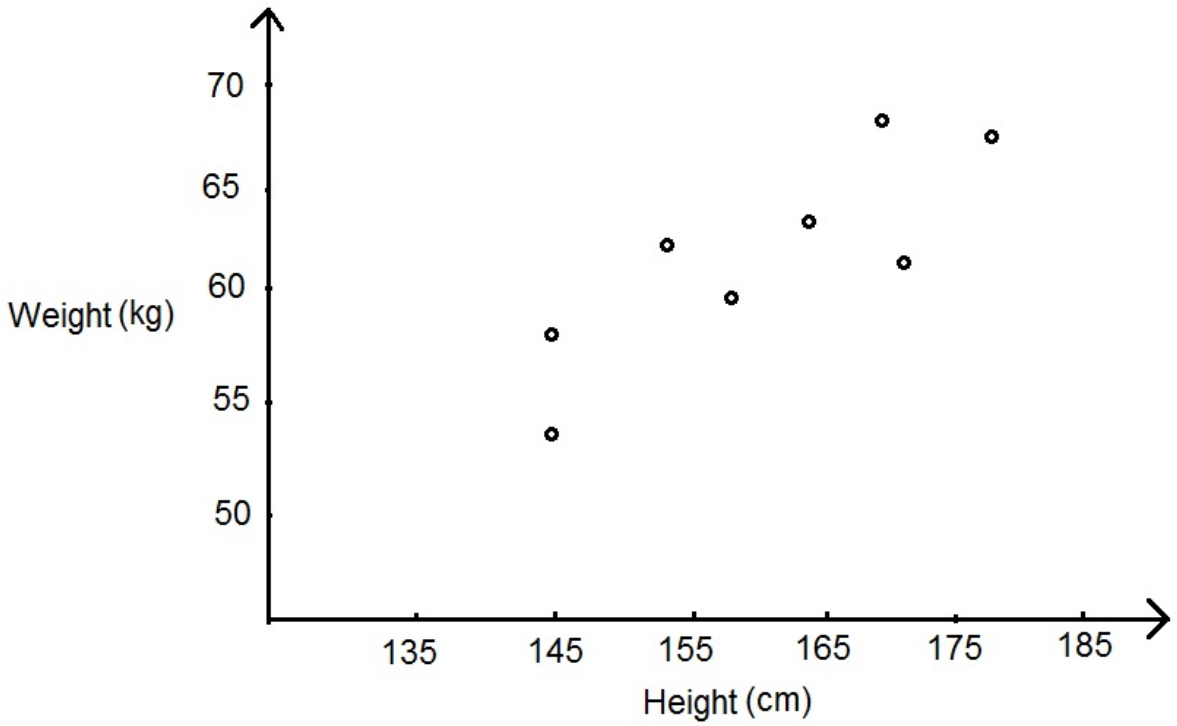
In fact, many things are not independent; they are associated (dependent).

- Height $\leftrightarrow$ Weight
- Years of education $\leftrightarrow$ Income
- Smoking $\leftrightarrow$ Lung cancer
- Vitamin C $\leftrightarrow$ Cold
- Weather $\leftrightarrow$ Electricity demand

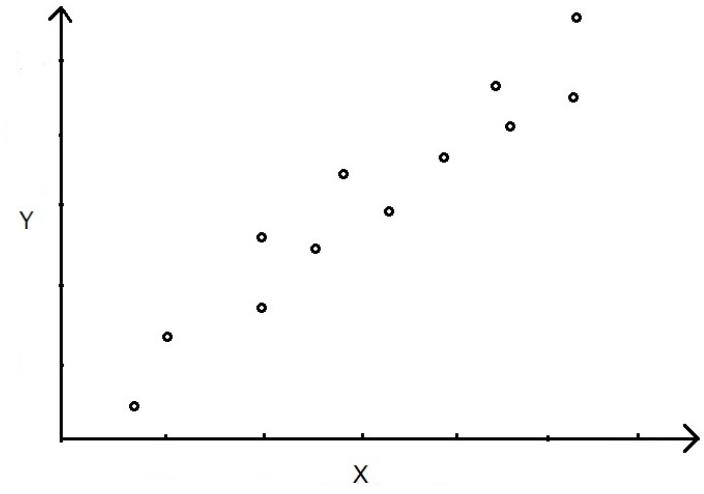
Sometimes, an association between two (random) variables can be detected visually from a **scatter plot**.

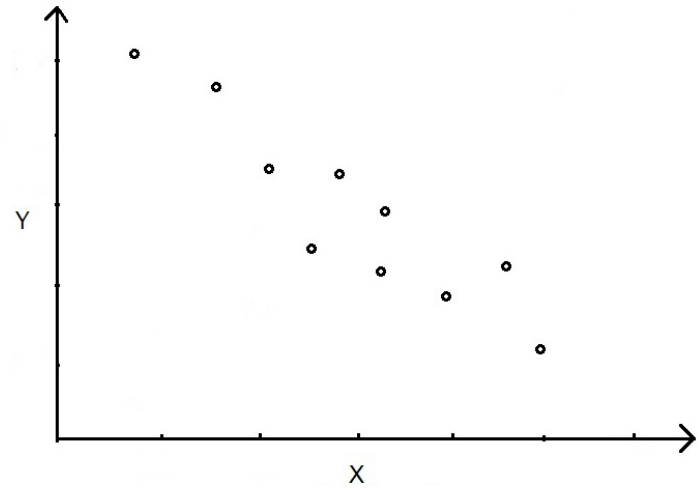
Scatter plot

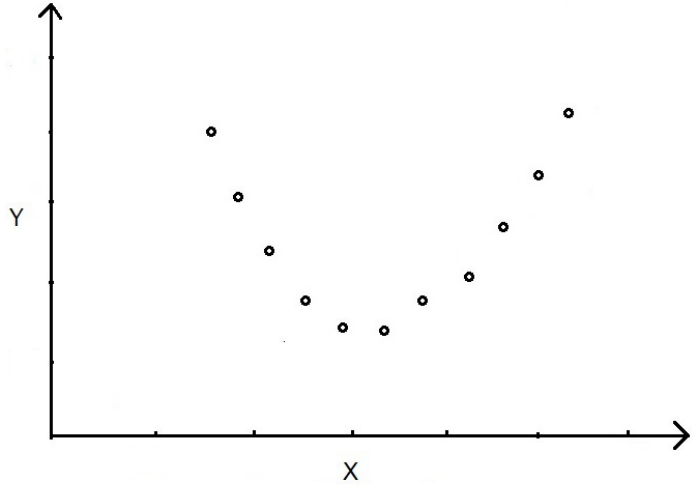
Let $\{(x_1, y_1), \ldots, (x_n, y_n)\}$ be n pairs of observations from random variables X and Y. A **scatter plot** shows each pair of the data as a point using their values as two coordinates.

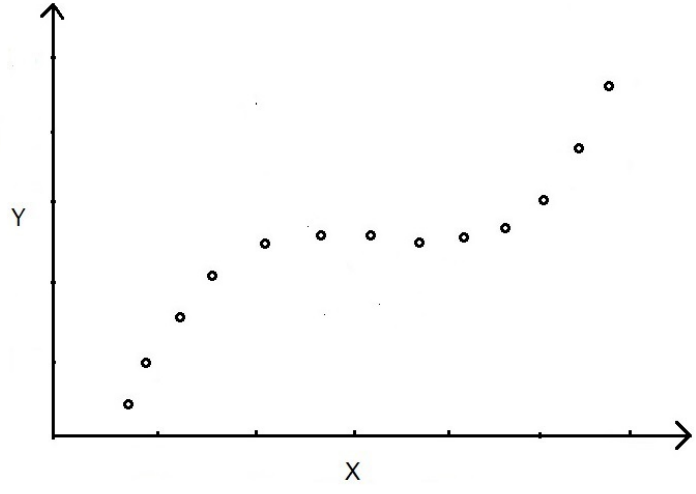


Shape of associations



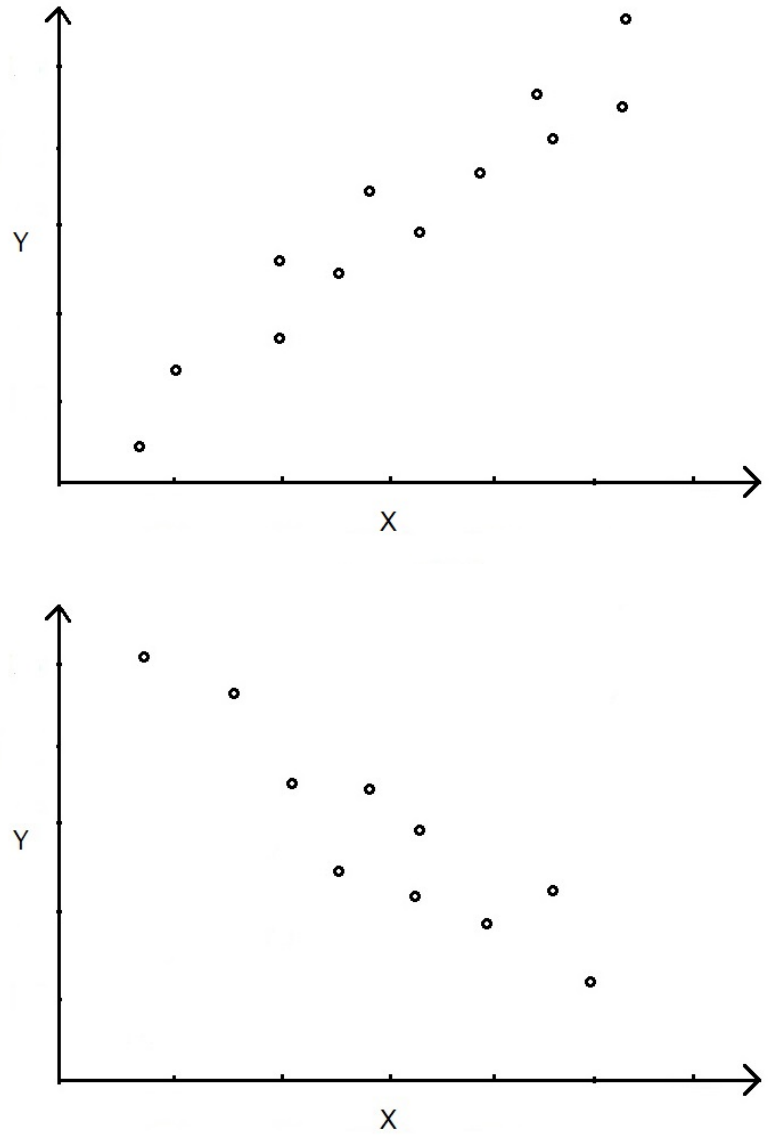






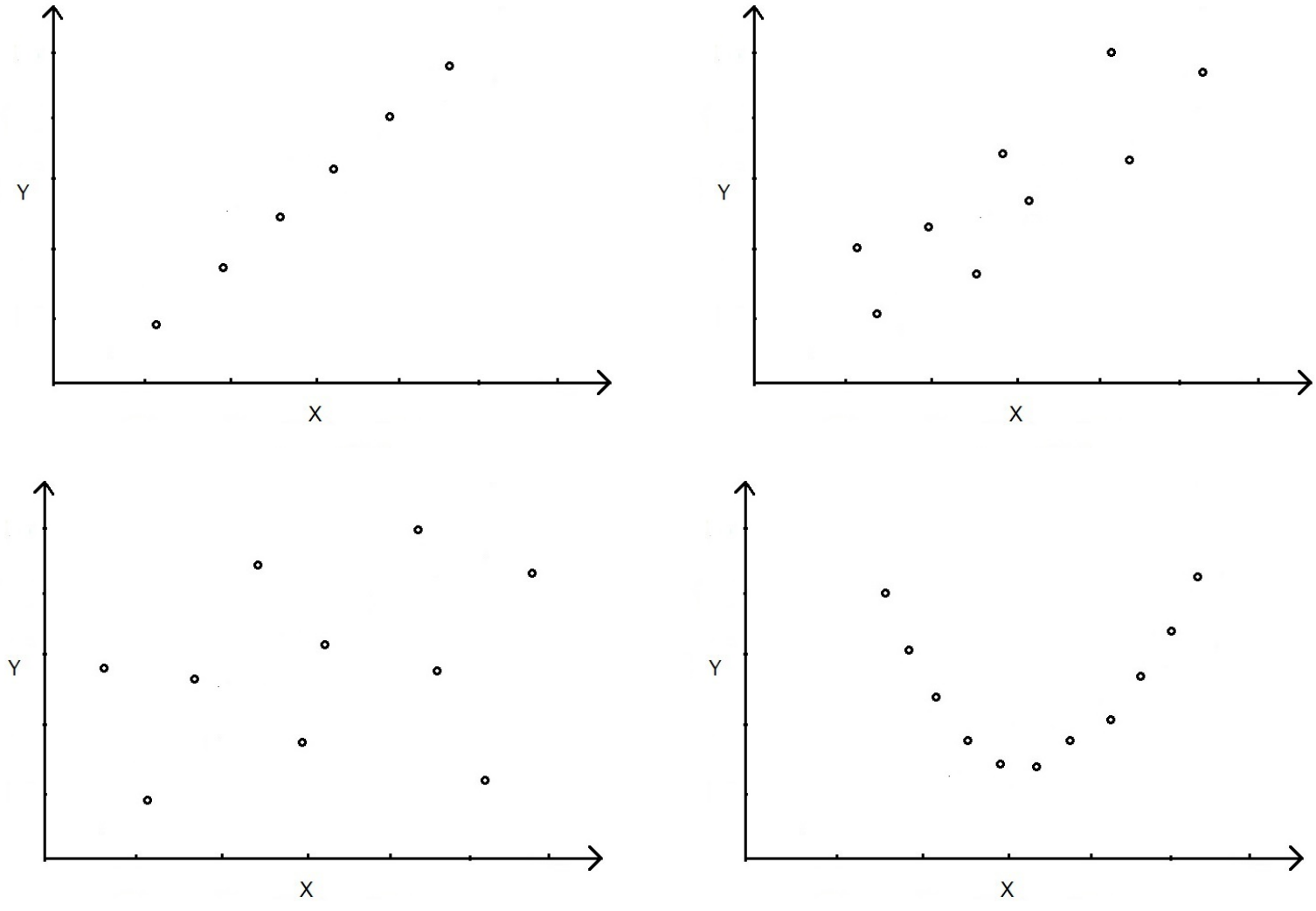
- Linear association vs. Non-linear association

Orientation of associations



- Positive association: larger X indicates larger Y
- Negative association: larger X indicates smaller Y

Intensity of associations



- Strong, Moderate, and Little association

Covariance

Let X and Y be two random variables with $E(X) = \mu_x$and $E(Y) = \mu_y$. To quantify the linear association between X and Y, we use **covariance**

$$Cov(X, Y) = E[(X - \mu_x)(Y - \mu_y)]$$

In practice, we can use **sample covariance** to estimate $Cov(X, Y)$

$$s_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{n - 1}$$

- $Cov(X, X) = E[(X - \mu_x)^2] = Var(X)$.
- In practice, $s_{xy} > 0$ indicates positive association; $s_{xy} < 0$ indicates negative association.

Correlation

Let $Var(X) = \sigma_x^2$ and $Var(Y) = \sigma_y^2$. A linear association can be better described by **correlation**

$$Cor(X, Y) = \frac{Cov(X, Y)}{\sigma_x \sigma_y}$$

In practice, we can use **sample correlation** to estimate $Cor(X, Y)$

$$r_{xy} = \frac{s_{xy}}{s_x s_y}$$

- r_{xy} is called Pearson's correlation, which has the same sign as s_{xy}.
- $-1 \leq r_{xy} \leq 1$. When $|r_{xy}| = 1$, X and Y has a perfect linear association; when $r_{xy} = 0$, X and Y has **no linear association**.
- r_{xy} is invariant to linear transformation.

From scatterplots to statistics

- No linear association does not mean no association (independence).
- However, linear association is common enough to be interesting!
- *Regression* is about estimating the strength of linear associations.
- **Regression is not a substitute for plotting!**

8.2: Line of Best Fit

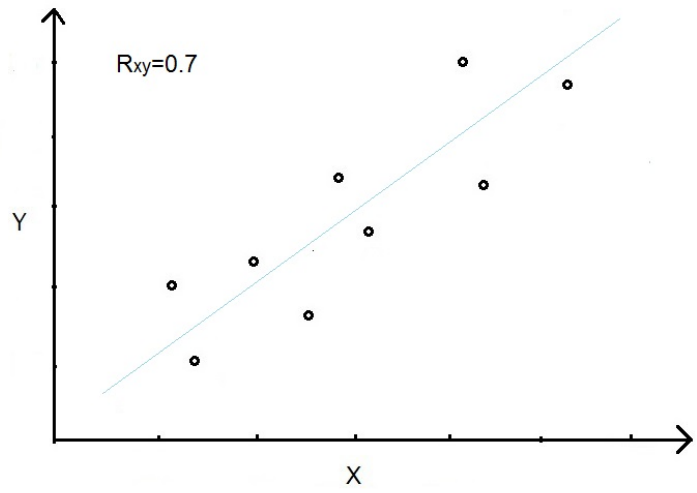
uOttawa - MAT2377

Fall 2020

Goal

Introduce *lines of best fit.*

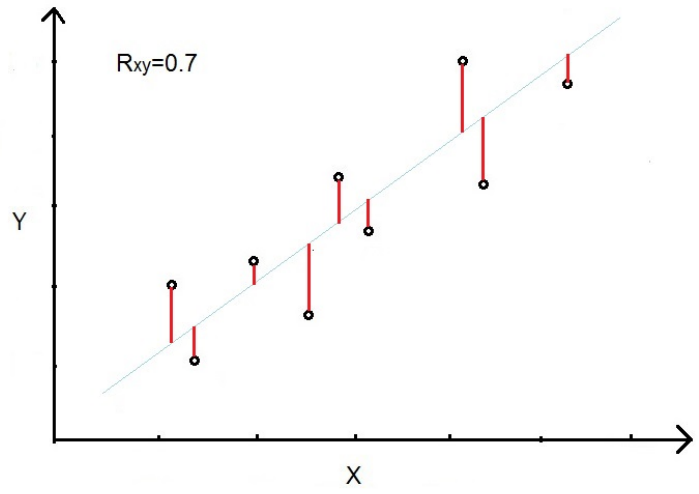
Least squares line



Let $\{(x_1, y_1), \ldots, (x_n, y_n)\}$ be n observations. We might hope to find a line (linear function) so that most data is close to the line

$$y = \hat{\alpha} + \hat{\beta}x.$$

Least squares line



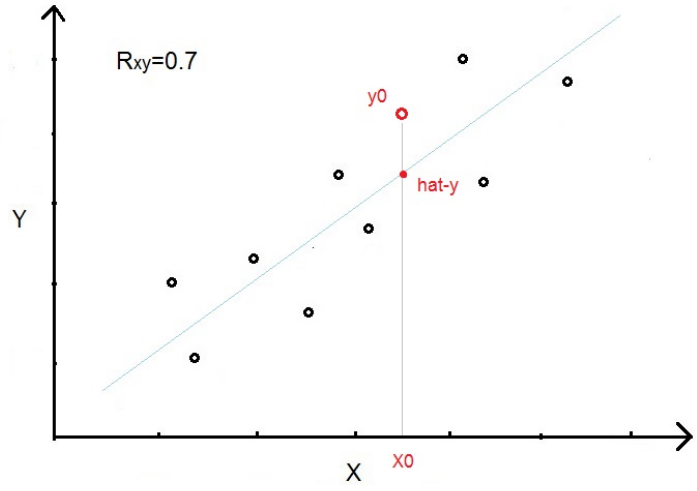
One way to get the "closest" line is to minimize

$$L(\alpha, \beta) = \sum_{i=1}^{n} [y_i - (\alpha + \beta x_i)]^2 ,$$

which leads to the formula

$$\hat{\alpha} = \bar{y} - \hat{\beta}\bar{x}, \quad \hat{\beta} = r_{xy} \frac{s_y}{s_x}$$

Prediction with Least squares line



In the least squares line,

$$\hat{y} = \hat{\alpha} + \hat{\beta}x$$

- Suppose x_0 is a new input, we predict the associated y value by $\hat{y}_0 = \hat{\alpha} + \hat{\beta}x_0$.
- $\hat{\alpha}$ is the intercept; it is the predicated y value when $x = 0$.
- $\hat{\beta}$ is the slope; One unit change in x would leads to $\hat{\beta}$ units change in $\hat{y}$.

Example: fish

Scientists find that the number of eggs laid per year (Y) of a particular fish has an linear association with its body length (X) in cm. A sample of size 30 is collected from (X, Y). The data are summarized is as follows

$$\sum_{i=1}^{30}(x_i - \bar{x})^2 = 119, \quad \sum_{i=1}^{30}(y_i - \bar{y})^2 = 587,$$
$$\sum_{i=1}^{30}(x_i - \bar{x})(y_i - \bar{y}) = 221$$
$$\sum_{i=1}^{30} x_i = 1179, \quad \sum_{i=1}^{30} y_i = 2972$$

Predict the y value at $x_0 = 36$.

Example: fish

Based on the summary, we can compute

$$s_x = 2.02, s_y = 4.5$$
$$s_{xy} = 7.6, \quad r_{xy} = \frac{s_{xy}}{s_x s_y} = 0.84$$
$$\bar{x} = 39.3, \quad \bar{y} = 99.1$$

Plugging in,

$$\hat{\alpha} = \bar{y} - \hat{\beta}\bar{x} = 99.1 - 1.87 \times 39.3 = 25.6, \quad \hat{\beta} = r_{xy}\frac{s_y}{s_x} = 0.84\frac{4.5}{2.02} = 1.87$$

Thus, the least squares line is $\hat{y} = 25.6 + 1.87x$.

Based on the least squares line, the predicted value of y at $x_0 = 36$ is given by

$$\hat{y}_0 = \hat{\alpha} + \hat{\beta}x_0 = 25.6 + 1.87 \times 36 = 92.92 \text{ (eggs/year)}$$

8.3: Regression and Statistics

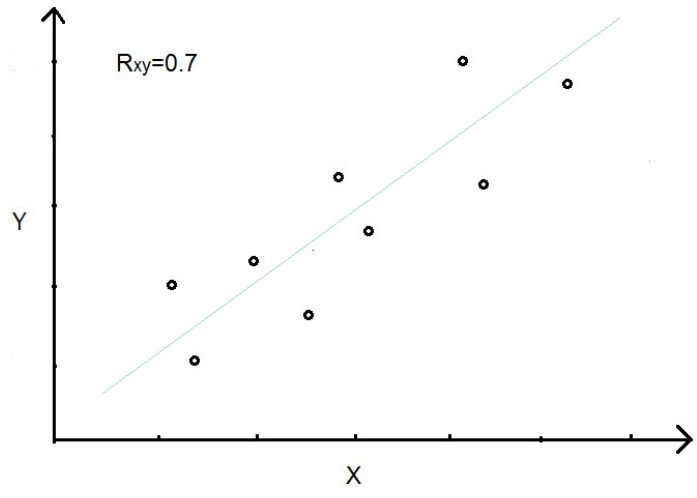
uOttawa - MAT2377

Fall 2020

Goal

Go from *line-fitting* to *statistics*.

Recall: Least squares line



We found the “line of best fit"

$$\hat{\alpha} = \bar{y} - \hat{\beta}\bar{x}, \quad \hat{\beta} = r_{xy}\frac{s_y}{s_x}.$$

Question: how to estimate error for $\hat{\alpha}, \hat{\beta}$?

Recall: Paired Differences Model

1. **Typically write:** $(Y_i - X_i) \overset{iid}{\sim} \mathcal{N}(\mu_{y-x}, \sigma^2_{y-x})$.
2. **Equivalent model:** X_i fixed, $\epsilon_i \overset{iid}{\sim} \mathcal{N}(0, \sigma^2_{y-x})$, $Y_i = X_i + \mu_{y-x} + \epsilon_i$;
3. **Regression model:** X_i fixed, $\epsilon_i \overset{iid}{\sim} \mathcal{N}(0, \sigma^2)$, $Y_i = \beta X_i + \alpha + \epsilon_i$.
4. **Note:** Paired differences is just regression with $\alpha = \mu_{y-x}$, $\beta = 1$ known; $\hat{\alpha} = \overline{Y} - \overline{X}$.

Models to Statistics

1. *Under this model*, we view $\hat{\beta}, \hat{\alpha}$ as random variables (just like $\overline{Y} - \overline{X}$ is for paired differences).
2. Can ask all the same questions:
   - 2.1 Distribution of $\hat{\beta}, \hat{\alpha}$?
   - 2.2 Confidence intervals for $\hat{\beta}, \hat{\alpha}$?
   - 2.3 Hypothesis testing for $\hat{\beta}, \hat{\alpha}$?
3. Really just need to know distribution.

Distribution of Regression Estimates: Mean

Unsurprisingly,

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

and

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

Distribution of Regression Estimates: Variance

More surprisingly,

$$Var[\hat{\beta}] = \frac{\sigma^2}{\sum_{i=1}^{n}(x_i - \overline{X})^2}$$

and

$$Var[\hat{\alpha}] = \frac{\sigma^2 \sum_{i=1}^{n} x_i^2}{n \sum_{i=1}^{n}(x_i - \overline{X})^2}.$$

Distribution of Regression Estimates: Rest of Distribution

1. Inevitably, both are Gaussian - so mean and variance tell you everything.
2. **Everything is Gaussian with known mean and variance**, so we can immediately write down confidence intervals!

Example (1)

a group of students are investigating Hook's law. They hang 25 different weights $X_1, \ldots, X_{25}$ from a spring and measure the amount $Y_1, \ldots, Y_{25}$ that the spring displaces. Their measurements have the following summary statistics:

$$\overline{X} = 39, \quad \overline{Y} = 115.73$$

$$\sum_{i=1}^{25} (x_i - \overline{X})^2 = 11700$$

$$\sum_{i=1}^{25} (y_i - \overline{Y})^2 = 103810.2$$

$$\sum_{i=1}^{25} (x_i - \overline{X})(Y_i - \overline{Y}) = 34733.64$$

$$\sum_{i=1}^{n} x_i^2 = 49725.$$

Calculate 95-percent confidence intervals for α, β assuming $\sigma = 5$.

Example (2)

As before, we start with the point estimates:

$$\begin{aligned}\hat{\beta} &= \frac{\sum_{i=1}^{n}(x_i - \overline{X})(y_i - \overline{Y})}{\sum_{i=1}^{n}(x_i - \overline{X})^2} \\ &= \frac{34733.64}{11700} = 2.969\end{aligned}$$

and

$$\begin{aligned}\hat{\alpha} &= \overline{Y} - \hat{\beta}\overline{X} \\ &= 115.73 - (2.969)(39) = -0.061.\end{aligned}$$

Example (3)

We now calculate the variances:

$$\sigma_\beta^2 = \frac{\sigma^2}{\sum_{i=1}^n (x_i - \overline{X})^2} = \frac{25}{11700} \approx 0.00214$$

$$\sigma_\alpha^2 = \frac{\sigma_\beta^2}{n} \sum_{i=1}^n x_i^2 = \frac{0.00214}{25}(49725) \approx 4.25.$$

Thus, our confidence intervals are:

$$\begin{aligned} I_\beta &= (\hat{\beta} - z_{0.025}\sigma_\beta, \hat{\beta} + z_{0.025}\sigma_\beta) \\ &= (2.969 - (1.96)\sqrt{(0.00214)}, 2.969 + (1.96)\sqrt{(0.00214)}) \\ &= (2.88, 3.06) \\ I_\alpha &= (-4.10, 3.98). \end{aligned}$$

8.4: Further Regression Topics

uOttawa - MAT2377

Fall 2020

Goal

Finish our discussion of regression (and the course!)

Review: Regression Models

Recall the basic regression model:

- We fix data $X_1, X_2, \ldots, X_n$ that is considered "nonrandom."
- We consider parameters α, β and σ^2.
- We assume that the observations $Y_1, \ldots, Y_n$ are random, with distribution given by:

$$Y_i = \alpha + \beta X_i + \epsilon_i,$$

  where $\epsilon_1, \ldots, \epsilon_n \sim \mathcal{N}(0, \sigma^2)$ are i.i.d.
- **Last video:** saw distribution of $\hat{\alpha}, \hat{\beta}$ given σ; used to do CI, NHST.
- **Next:** What if σ is not known?

Estimating σ

Set notation:

$$S_{xx} = \sum_{i=1}^{n}(x_i - \overline{X})^2, \quad S_{yy} = \sum_{i=1}^{n}(y_i - \overline{Y})^2, \quad S_{xy} = \sum_{i=1}^{n}(x_i - \overline{X})(y_i - \overline{Y})$$

and define the estimator

$$s^2 = \frac{S_{yy} - b_1 S_{xy}}{n-2}.$$

Theorem: $E[s^2] = \sigma^2$.

Example (1)

A group of students are investigating Hook's law. They hang 25 different weights $X_1, \ldots, X_{25}$ from a spring and measure the amount $Y_1, \ldots, Y_{25}$ that the spring displaces. Their measurements have the following summary statistics:

$$\overline{X} = 39, \quad \overline{Y} = 115.73$$

$$\sum_{i=1}^{25} (x_i - \overline{X})^2 = 11700$$

$$\sum_{i=1}^{25} (y_i - \overline{Y})^2 = 103810.2$$

$$\sum_{i=1}^{25} (x_i - \overline{X})(Y_i - \overline{Y}) = 34733.64$$

$$\sum_{i=1}^{n} x_i^2 = 49725.$$

Calculate an unbiased estimate of the parameter σ^2.

Example (2)

Last video we calculated:

$$\begin{aligned}\hat{\beta} &= \frac{\sum_{i=1}^{n}(x_i - \overline{X})(y_i - \overline{Y})}{\sum_{i=1}^{n}(x_i - \overline{X})^2} \\ &= \frac{34733.64}{11700} = 2.969.\end{aligned}$$

We then use our formula:

$$\begin{aligned}s^2 &= \frac{S_{yy} - \hat{\beta} S_{xy}}{n-2} \\ &= \frac{103810.2 - (2.969)(34733.64)}{25-2} \\ &= 29.8.\end{aligned}$$

Confidence Intervals (σ not known)

$$I_\beta = (\hat{\beta} - t_{\frac{\alpha}{2},n-2}\frac{s}{\sqrt{S_{xx}}}, B_1 + t_{\frac{\alpha}{2},n-2}\frac{s}{\sqrt{S_{xx}}})$$

$$I_\alpha = (\hat{\alpha} - t_{\frac{\alpha}{2},n-2}\frac{s}{\sqrt{n\,S_{xx}}}\sqrt{\sum_{i=1}^{n} x_i^2}, \hat{\alpha} + t_{\frac{\alpha}{2},n-2}\frac{s}{\sqrt{n\,S_{xx}}}\sqrt{\sum_{i=1}^{n} x_i^2}),$$

Note that we are using the *t*-distribution with *two* degrees of freedom, not one.

Example (3)

Continuing last example: calculate a 95-percent confidence interval for β. Recall:

$$\hat{\beta} = 2.969,\ \hat{\alpha} = -0.061,\ s^2 = 29.8.$$

From a table lookup, $t_{0.025,23} = 2.069$. Thus, we obtain the confidence intervals:

$$\begin{aligned} I_\beta &= (\hat{\beta} - t_{\frac{\alpha}{2},n-2}\frac{s}{\sqrt{S_{xx}}}, \hat{\beta} - + t_{\frac{\alpha}{2},n-2}\frac{s}{\sqrt{S_{xx}}}) \\ &= (2.969 - (2.069)\frac{\sqrt{29.8}}{\sqrt{11700}}, 2.969 + (2.069)\frac{\sqrt{29.8}}{\sqrt{11700}}) \\ &= (2.86, 3.07) \end{aligned}$$

Example (4)

Coontinuing the same example: test the hypothesis that $\beta = 0$. Formally, we have the null and alternative hypotheses:

$$H_0 : \beta_0 = 0,\ H_1 : \beta_0 \neq 0.$$

We note that $0 \notin (2.86, 3.07) = I_\beta$. Thus, we *reject* the null hypothesis.