

# STATS 260 Class 1

*Gavin Jaeger-Freeborn*

*Tue 05 May 2020 07:58:29 AM*

## 1. Course Outline

### Grading

R Assignments	6%
Weekly Quizzes	80%
Final (Time TBA)	14%

### Weekly Quizzes

There will be weekly quizzes. Approximately 10 worth 8% each. You are expected to be able to submit pdfs.

### Homework

There are 3 R assignments

### Important dates

Classes begin	Monday, May 4, 2020
Drop (100% Fee Reduction)	Saturday, May 16, 2020
Last Day to Add Courses	Saturday, May 16, 2020
Drop (50% Fee Reduction)	Saturday, June 6, 2020
Academic Drop Date	Wednesday, July 1, 2020
Reading Break (no classes)	July 1-2, 2020
Last day of classes	Friday, July 31, 2020
Examination period	August 4 to August 17, 2020

## STATS 260 Class 2

*Gavin Jaeger-Freeborn*

### 2. population ( $\mu$ )

#### Example

- All I-beams being made by a particular manufacturer.
- All Canadians who will be eligible to vote in an upcoming election.
- All people who will at some point take a particular blood pressure medication.

### 3. Parameter

Measurement of a population

### 4. Sample ( $x$ )

A subset of the population

### 5. Statistic

Measurement of a sample

#### 5.1. Descriptive Statistics

organize, summarize, display, and describe features of the data.

#### Example

Some sorts of questions descriptive statistics answers:

- What is the greatest tensile strength recorded? What is the range of recorded tensile strengths?
- What proportion of the sample of voters is older than 65?
- What is the average weight of the sample of people taking blood pressure medication? How spread out are the measurements for resting heart rate?

## 5.2. Inferential Statistics

draw conclusions about the population based on the measurements from the sample.

### Example

Some sorts of questions inferential statistics answers:

- What is a likely range of values of tensile strengths for all I-beams made by the manufacturer?

**NOTE:** all I-beams → population

- Based on our survey, which party is likely to win the election?
- Can we conclude that there a relationship between weight and blood pressure?

## 6. Examples

Determine whether the underlined words refer to a:

- We wish to study poplar trees, so we make a selection of 15 poplar trees in a forest.  
→ Sample
- From our selection of 15 poplar trees, we find the largest tree to have a height of 1.9 m.  
→ Statistic
- A newspaper wants to determine the feelings of Victoria residents regarding a bridge to the mainland.  
→ Population
- The newspaper phones 500 Victoria residents.  
→ Sample
- It is found that 95% of these people are in favor of a bridge.  
→ Statistic

## 7. Mean, Median, and Mode

### 7.1. Mean ( $\bar{x}$ )

Sample Mean	$\bar{x}$	average of a sample (an estimation of $\bar{\mu}$ )
Population Mean	$\bar{\mu}$	mean of a population

$$\bar{x} = \frac{x_1 + \dots + x_n}{n}$$

#### Example

Suppose the following is data taken from some sample. Calculate the sample mean.

10, 6, 12, 7, 3, 6

$$44/6 = 7.333, \therefore \bar{x} = 7.333$$

### 7.2. Median ( $\tilde{x}$ )

Middle of a sorted list

#### Example

Suppose we have the sample data: 6, 9, 3, 18, 11. Find the sample median of these data.

3, 6, 9, 18, 11

Median is then 9

**NOTE:** median is unaffected by outliers

### 7.3. Mode

The value that appears the most often

#### Example

Median of 3, 5, 9, 9, 9, 5 is 9

### Example

The data set 1, 2, 3, 3, 3, 4, 4, 4, 5, 5 has two modes (3 and 4).

The data set 1, 2, 3, 4, 5 has **no modes** (since there is no observation that occurs more frequently than any other observation).

### 8. Standard Deviation

sample variance	$s^2$	Sample Standard Deviation
population variance	$\sigma^2$	Population Standard Deviation

ith Deviation = difference between  $x_i$  and  $\bar{x}$

### Example

Find the variance and standard deviation of the following sample:

7, 7, 9, 15, 16, 17, 19, 21, 22, 40

$$\bar{x} = 17.3$$

$$\tilde{x} = 16.5$$

$$s = 9.5730$$

$$s^2 = 93.5667$$

$$\sum x_i^2 = 3835 \quad \bar{x} = 17.3$$

$$s^2 = \frac{\sum x_i^2 - n(\bar{x})^2}{n - 1}$$

### 9. coefficient of variation (cv)

used to compare 2 sets is a dimensionless quantity (i.e. no units of measurement) which can be used to assess the variability of a set of observations. The cv is calculated by

$$\frac{s}{\bar{x}}$$

#### Example

One set of observations has a mean of 35 with a standard deviation of 7. A second set of observations has a mean of 55 with a standard deviation of 9. Which data set has more variability?

More spread out

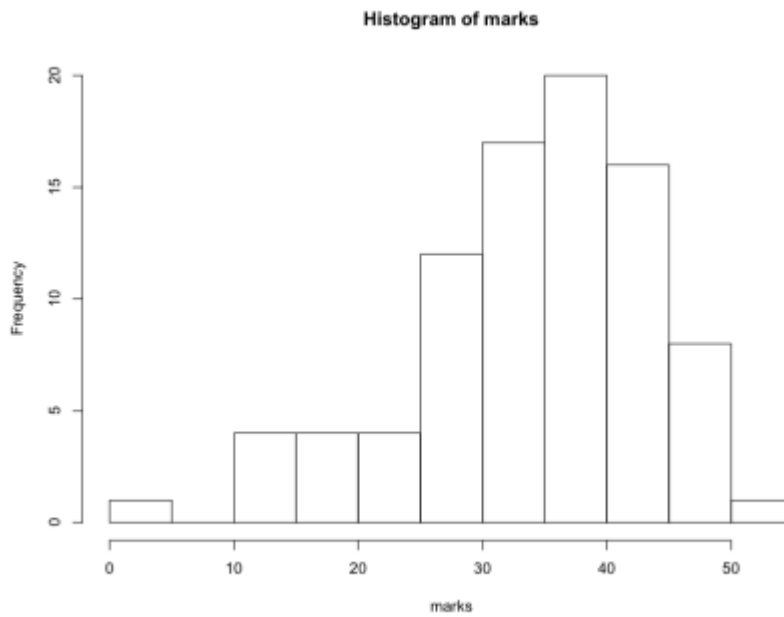
$$cv_1 = 7/35 = 0.2$$

$$cv_2 = 9/55 = 0.1636$$

### STATS 260 Class 3

*Gavin Jaeger-Freeborn*

### 10. Histograms



### Modal

Unimodal	only one mode
Bimodal	2 modes
Multimodal	more then 2

symmetric	even tail on both sides
asymmetric	uneven tail

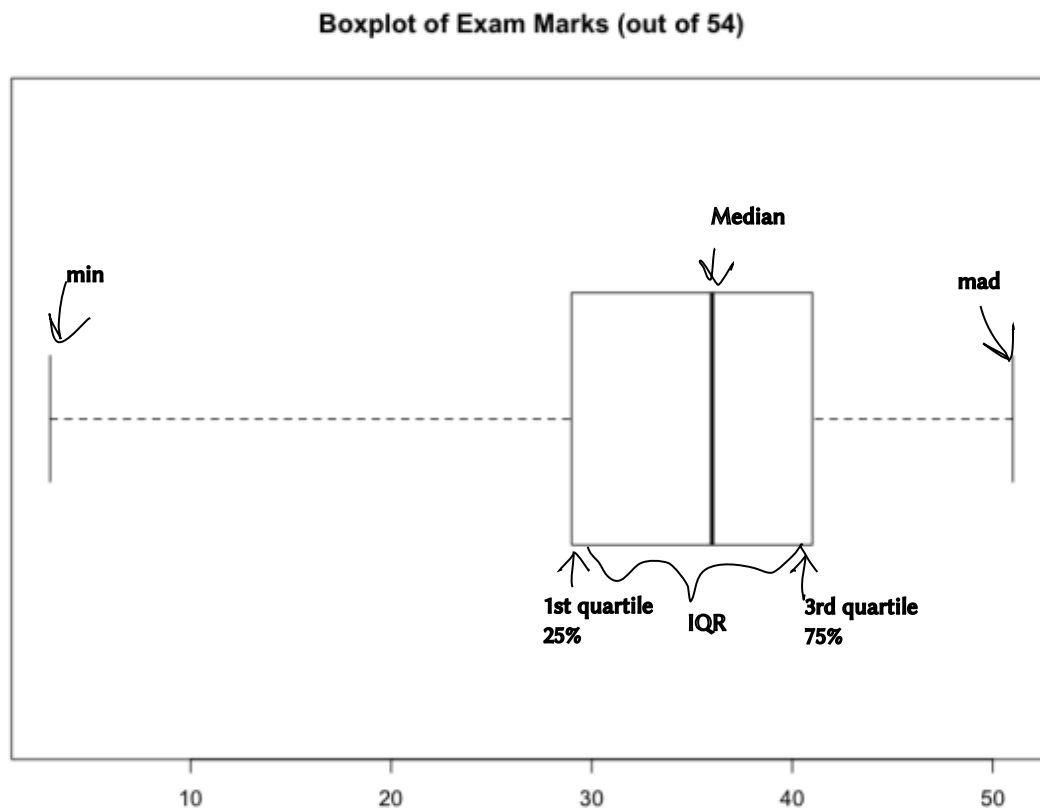
In this example it is unimodal

### Skewed

Where the data is mostly tailing in terms of the Mode(peak)

In this case it is negatively skewed

## 11. Boxplot



### 11.1. Interquartile

Range is the area between the 1st quartile and 3rd quartile

### 11.2. Outliers

Outside of the interval

$$[lowerquartile - 1.5 IQR, upperquartile + 1.5 IQR]$$



### 11.3. Example

Suppose I have the following sample data:

1. 7, 0. 9, 3. 8, 2. 1, 1. 9, 0. 6, 0. 5, 5. 0, 2. 4, 0. 1, 5. 0, 0. 3, 8. 8, 0. 3, 0. 3, 3. 3, 4. 8, 0. 2, 2. 2, 3. 5

I've used R to find that the lower quartile is **0.45**, the upper quartile is 3.575, and the IQR is  $3.575 - 0.45 = 3.125$

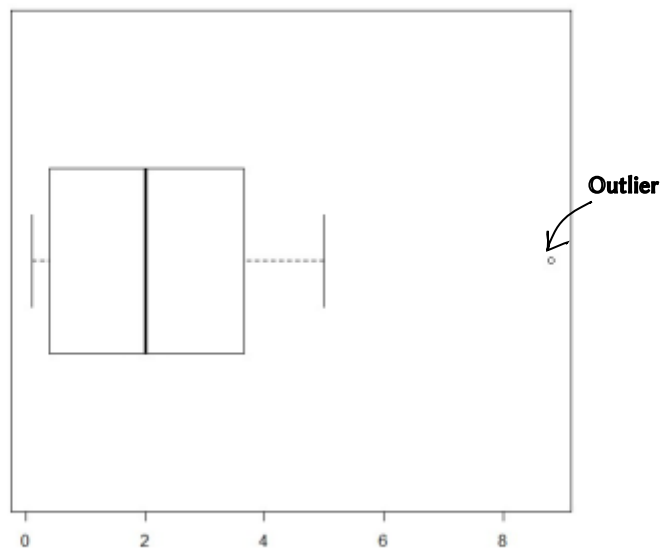
3.575, and the IQR is  $3.575 - 0.45 = 3.125$

**NOTE:** Sometimes outliers are actually an error

Data outside  $[0.45 - (1.5)(3.125), 3.575 - (1.5)(3.125)] = [-4.2375, 8.2625]$  Would be an **Outlier**.

8.8 which is outside that range. The right whisker ends at 5.0 (our largest non-outlier),

The Outlier 8.8 is indicated with a circle.



## 12. Bivariate data

Two variables. Set of pairs:  $(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$ .

Common question - Whether or not there is a relationship between the two variables.

## 13. Scatterplot

A **Scatterplot** is used to visually depict **bivariate** data. The observations are plotted as a set of points on the plane.

### Important

For a scatterplot to be appropriate, each pair of measure-

### Example

I select 20 people, and for each person, I record **x**, their age, and **y**, their maximum heart rate.

Here, the data is **clearly bivariate** (one sample of size  $n = 20$ , with pairs of measurements being made);

A scatterplot would be appropriate.

### Example

I select 20 people and put them on Diet A, and measure **x**, their blood pressure after two weeks. I select another 20 people and put them on Diet B, and measure **y**, their blood pressure after two weeks.

Here, we have two samples, of sizes  $n_1 = 20$  and  $n_2 = 20$ . The data is **not bivariate**;

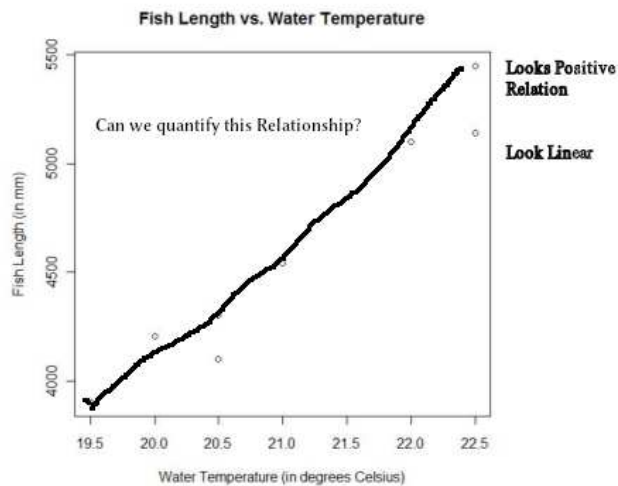
a scatterplot would be inappropriate.

### 13.1. Scatterplot Example

Several of a particular species of fish are grown from eggs in tanks set at particular temperatures. After a fixed number of days, all fish are measured.

We wish to investigate the relationship between  $y$ , the length of the fish (in mm), and  $x$ , the temperature of the tank (in degrees Celsius).

$y$	3900	4205	4100	4300	4540	5100	5450	5140		
$x$	19.5	20	20.5	20.5	21	22	22.5	22.5	$n = 8$	



### 13.2. Sample Correlation Coefficient ( $r$ )

Used to assess the **linearity** of **bivariate data**.

$$r = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2} \sqrt{\sum_{i=1}^n (y_i - \bar{y})^2}}$$

#### Computation Form

$$r = \frac{\sum_{i=1}^n x_i y_i - n\bar{x}\bar{y}}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2} \sqrt{\sum_{i=1}^n (y_i - \bar{y})^2}}$$

Denominator could be written in terms of  $s_x$  and  $s_y$  (the standard deviation of  $x$  and  $y$ , respectively).

$$r = \frac{\sum_{i=1}^n x_i y_i - n\bar{x}\bar{y}}{(n-1)s_x s_y}$$

#### Example: For our fish data:

We find  $\sum_{i=1}^n x_i y_i = (19.5)(3900) + \dots + (22.5)(5140) = 778165$ .

Then, we find  $\bar{x}$ ,  $\bar{y}$ ,  $s_x$ ,  $s_y$  using our calculator. We have  $r \approx 0.973$ .

**NOTE:** This is fairly linear

#### Interpretation

$r$  takes on values **between -1 and 1**. *no units*

- An  $r$  value of -1 indicates a perfect **decreasing linear** relationship.
- An  $r$  value of 1 indicates a perfect **increasing linear** relationship.
- An  $r$  value of 0 indicates a **non linear** relationship.

**Warning** An  $r$  value of 0 does not mean there is no relationship, only that the relationship is not linear.

## 14. Correlation Vs Causation

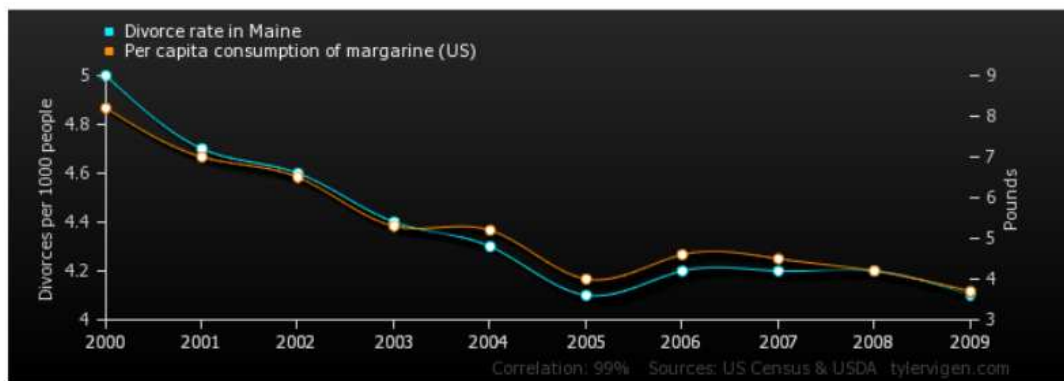
When we examine variables  $x$  and  $y$  and find there appears to be some correlation between them, there are many possible explanations:

- $x$  causes  $y$
- $y$  causes  $x$
- There is some other unexplored variable which relates to both  $x$  and  $y$
- The correlation is spurious (there's no actual relationship; the correlation is just a coincidence)

NOTE. Spurious = no real correlation

### Example

The image below shows that there appears to be a strong correlation between the divorce rate in Maine and the consumption of margarine. This is one of many examples of spurious correlation.



## 15. Introduction to Probability

### Experiment

An activity we measure, or observe the results **Example** - Flipping a coin three times and noticing the sequence of heads and tails is an experiment.

### Outcomes

The observations from our experiment.

### Sample Space $S$

The set of all possible outcomes. The sample space may contain a finite or an infinite number of outcomes.

### Sample Point

A single outcome in the sample space.

### Event

Any subset of  $S$  (i.e. any collection of outcomes).

### Simple event

An event consisting of one outcome.

### Compound event

An event consisting of more than one outcomes.

### 15.1. Example

Consider the experiment where we flip a coin three times and note the sequence of heads and tails.

For this experiment the sample space is as follows:

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

Each of these eight elements of  $S$  are sample points. Some examples of events are:

$$A = \{HHH, HHT, HTH, THH\} \leftarrow \text{at least 2 heads}$$

$$B = \{HHT, HTT, THT, TTT\}$$

$$C = \{HHH, TTT\}$$

Events are usually described in words. For example,  $B$  is the event that the third flip is tails

We say that an event **occurs** if one of its sample points is an observed when we carry out the experiment when we carry out the experiment.

### 16. Set Theory

A and B	The <b>intersection</b> of $A$ and $B$ is $A \cap B$
A or B	The <b>union</b> of $A$ and $B$ is $A \cup B$
not A	The <b>complement</b> of $A$ is $\bar{A}$ or $A'$

#### Example

Suppose we select an integer from 1 to 10 at random. Let  $A$  be the event that an even number is selected. Let  $B$  be the event that a number 7 or larger is selected.

Find  $A \cap B$ ,  $A \cup B$ , and  $\bar{B}$

$$S = \{1, 2, \dots, 10\}, A = \{2, 4, 6, 8, 10\}, B = \{7, 8, 9, 10\}$$

$$1 \rightarrow \text{outcome sample point}$$

$$A \cap B = \{8, 10\}$$

$$A \cup B = \{2, 4, 6, 8, 10, 7, 9\}$$

$$B' = \{1, 2, 3, 4, 5, 6\}$$

## STATS 260 Class 4

Gavin Jaeger-Freeborn

Thu 14 May 2020 11:43:27 PM

Guaranteed event	$S$	will always happen
Impossible/null event	$\emptyset$	will never happen

$S$  is called a **guaranteed** or **certain event**, because it will always occur.

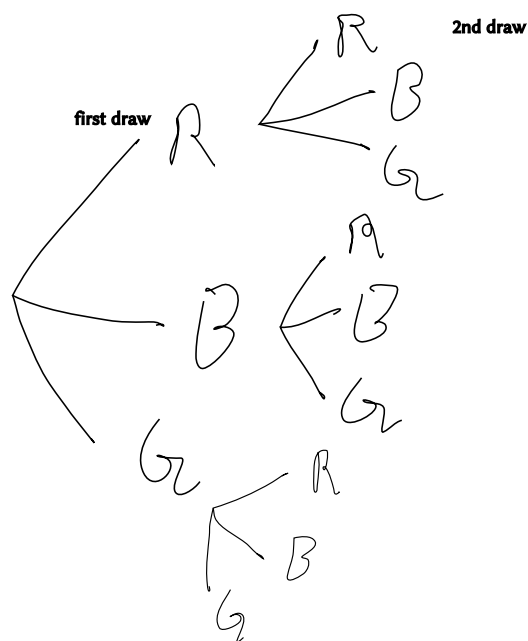
The event  $\emptyset$ , which consists of no outcomes, is called the **impossible event** or **null event**, because it never occurs.

If for events  $A$  and  $B$ , we have  $A \cap B = \emptyset$ , then we say that  $A$  and  $B$  are disjoint or mutually exclusive events.

We can often use tree diagrams to help us find all possible outcomes.

### Example

Suppose that a box contains red, blue, and green marbles (several of each color). Two marbles are selected one at a time from the box, and the sequence of colors is noted. What is the sample space?



$$S = \{ RR, RB, BR, BB, BG, GR, GB, GG \}$$



## 17. Probability ( $Pr(A)$ or $P(A)$ )

Likelihood that some event will or will not occur.

We measure probability on a scale from 0 to 1

0  $\rightarrow$  impossible for the event to occur

1  $\rightarrow$  event is guaranteed to occur.

### 17.1. Approaches

#### Experimentally

- repeat an experiment  $n$  times
- count  $f$ , the number of times the event in question occurs.
- then  $P(A) = f/n$

#### Classical (the one we will use)

Theoretically

### 17.2. Probability Axioms

1.  $P(S) = 1 \leftarrow$  Guaranteed
2.  $P(A) \geq 0$  for any event  $A$
3.  $P(A_1 \cup A_2 \cup \dots) = \sum P(A_i)$  for all **infinite** collection of **mutually exclusive** events.  
 $\therefore A_i \cap A_j = \emptyset$

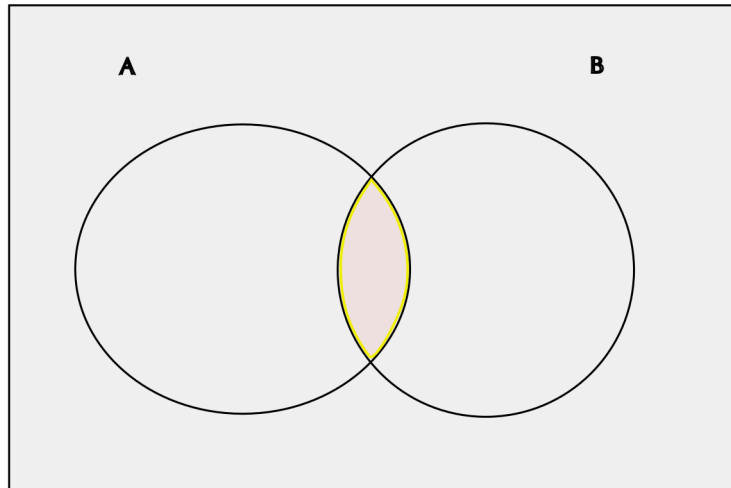
From these axioms, we can derive other properties of probability, including:

- $P(\emptyset) = 0$
- $P(A_1 \cup A_2 \cup \dots \cup A_k) = \sum_{i=1}^k P(A_i)$ . ( where the events are all mutually exclusive )
- $P(A) = 1 - P(\bar{A})$  for any event  $A$ .  $\leftarrow$  or  $P(\bar{A}) = 1 - P(A)$
- $P(A) \leq 1$  for any event  $A$
- $P(A \cup B) = P(A) + P(B) - P(A \cap B)$  for any events  $A$  and  $B$ .
- $P(A \cup B \cup C) = P(A) + P(B) + P(C) - P(A \cap B) - P(A \cap C) - P(B \cap C) + P(A \cap B \cap C)$  for any events  $A$ ,  $B$ , and  $C$ .

$$P(A \cup B) = P(A) + P(B) - P(A \cap B)$$

If we just did  $P(A) + P(B)$  we would over count so we -  $P(A \cap B)$

### Example



**NOTE:** End of first quiz

## 18. Uniform Sample Space

Each sample is equivalently likely to be picked

### Example

Since every element of S appears the same amount of times they are all equivalently likely to be picked.

$$S = \{1, 2, 3, 4, 5, 6, \}, P(\{1\}) = \frac{1}{6}$$

$$n(S) = 6$$

$n(S)$  = size of the sample space

$n(A)$  = size of event A

$$A = \{2, 4, 6\}$$

$$n(A) = 3$$

$n(S)$  sample events must have the same probability, and those probabilities must add to 1.

The probability of each event must be  $1/n(S)$

The **probability of any event A** in a **uniform, finite sample space S** is

$$\therefore P(A) = \frac{n(A)}{n(S)}$$

$$\frac{3}{6} = \frac{1}{2}$$

### Example

There are 80 students in a classroom. I will select one of the 80 students at random to answer a question. Of the 80 students, 7 are sitting in the front row. What is the probability that I select a student who is sitting in the front row?

$$n(S) = 80, n(A) = 7$$

$$P(A) = \frac{7}{80}$$

### Example

The 2001 Census found that in Tofino, there were 790 residents who traveled to work. Here are the results of this census question

Mode of Transportation	Total Numbers
Car/truck/van	435
Walk/bicycle	250
Other method	105

Suppose a Tofino resident who travels to work is selected at random. What is the probability that this resident walks or bikes to work?

$$435 + 250 + 105 = 790$$

### Example

Consider the results of the following survey of 250 single-crop farms:

	Wheat	Corn	Soy
Alberta	69	15	16
Saskatchewan	61	65	24

If we select one farm at random, what is the probability that the **farm grows wheat, or is in Saskatchewan**?

	Wheat	Corn	Soy
Alberta	69	15	16
Saskatchewan	61	65	24

$$\text{Prob} = \frac{69 + 61 + 65 + 24}{250}$$

### 19. $P(B|A)$

$P(B|A)$  = probability that B will occur if A occurs.

$$P(B|A) = \frac{n(B \cap A)}{n(A)} = \frac{P(B \cap A)}{P(A)}$$

### Example

Consider the results of the following survey of 250 single-crop farms:

	Wheat	Corn	Soy
Alberta	69	15	16
Saskatchewan	61	65	24

Suppose that a single-crop farm is selected at random. If the farm is in Alberta, what is the probability the farm grows soy?

$$P(\text{Soy}|\text{Alberta}) = \frac{16}{69 + 15 + 16}$$

### Example 2

If a farm which **grows soy** is selected, what is the probability that the farm is **in Alberta**?

$$P(\text{Alberta}|\text{Soy}) = \frac{16}{16 + 24}$$

**NOTE:**  $P(A|B) \neq P(B|A)$  - in general

### Example

Suppose 80% (**A**) of all Canadians exercise one or more days a week, and also, that 20% (**B**) of all Canadians exercise at five or more days a week. If we randomly select a Canadian who exercises at least one day a week, what is the probability that this Canadian exercises five or more days a week?

$$B \subseteq A$$

$$B \cap A = B$$

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

$$= \frac{P(B)}{P(A)} = \frac{0.2}{0.8}$$

$$\boxed{= 0.25}$$

### Example

Suppose we would like to know the probability that someone orders **chocolate ice cream in a waffle cone**.

- We want  $P(\text{Chocolate} \cap \text{Waffle})$

### Example

Suppose we would like to know the probability that someone **who wants a waffle cone will order chocolate ice cream**. Which of the following are we trying to find:

- We want  $P(\text{Chocolate}|\text{Waffle})$

## 20. Multiplication Rule

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

This is from

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

## STATS 260 Class 5

*Gavin Jaeger-Freeborn*

*Thu 21 May 2020 11:20:51 PM*

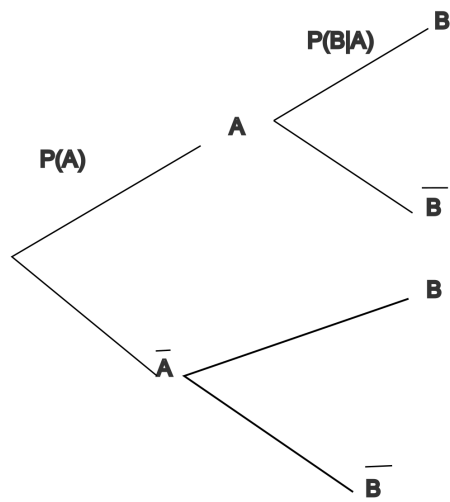
## 21. Multiplication Rule

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

This is from

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

**NOTE:** This is useful for tree diagrams



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

$$P(\bar{A} \cap B) = P(\bar{A})P(B|\bar{A})$$

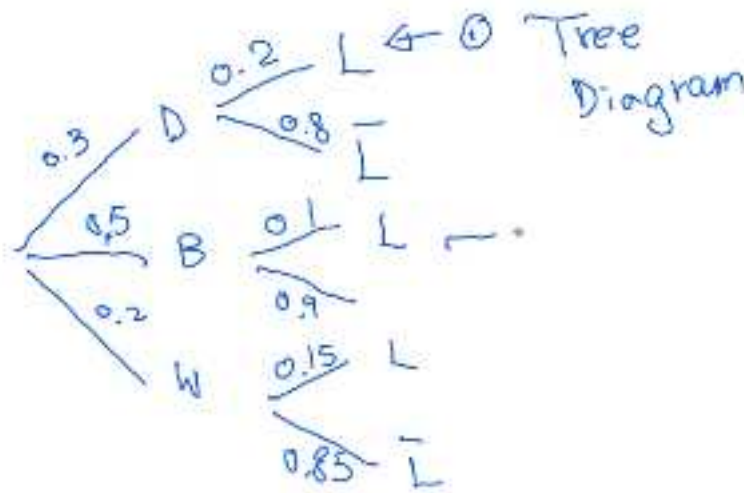


$$P(B) = P(A \cap B) + P(\bar{A} \cap B)$$

$$A \cup \bar{A} = S \quad A \cap \bar{A} = \phi$$

### Example

Suppose that 30% of all students drive to school, 50% take the bus, and 20% walk. Of those who drive, 20% are usually late for their first class of the day. Of those who take the bus, 10% are usually late for their first class of the day. Of those who walk, 15% are usually late for their first class of the day. What is the probability that a randomly selected student is regularly late for their first class?



$$P(L \cap D) = 0.3 * 0.2 = 0.06$$

$$P(B \cap L) = 0.5 * .1 = .05$$

$$P(W \cap L) = 0.2 * 0.15 = 0.03$$

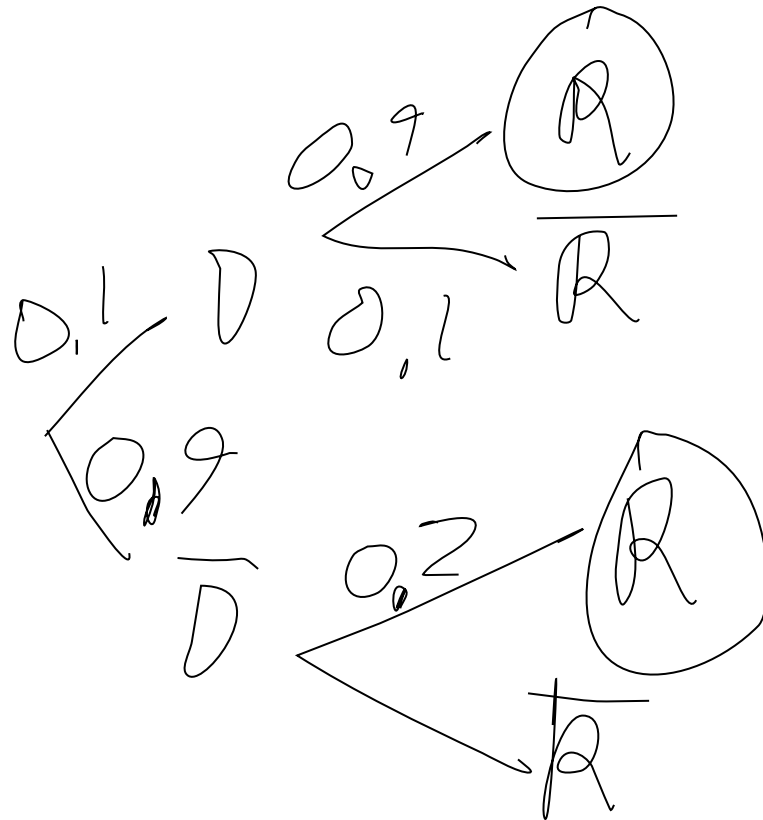
$$P(L) = P(L \cap D) + P(B \cap L) + P(W \cap L)$$

$$= 0.06 + 0.05 + 0.03 = 0.14$$

### Example

The probability of an item on a certain production line being defective is 0.1. If an item is defective, the probability that the inspector will remove it from the line is 0.9. If an item is not defective, the probability that the inspector will remove it from the line is 0.2.

What is the probability that a randomly selected item will be removed from the production line?



$$P(R) = (0.1)(0.9) + (0.9)(0.2) = 0.27$$

## 22. Law of Total Probability

if  $A_1, A_2, \dots, A_k$  are a collection of mutually exclusive and exhaustive events, then for any event  $B$  we have:

$$\begin{aligned} P(B) &= P(B \cap A_1) + P(B \cap A_2) + \dots + P(B \cap A_k) \\ &= P(B|A_1)P(A_1) + P(B|A_2)P(A_2) + \dots + P(B|A_k)P(A_k) \end{aligned}$$

### 23. Bayes Theorem

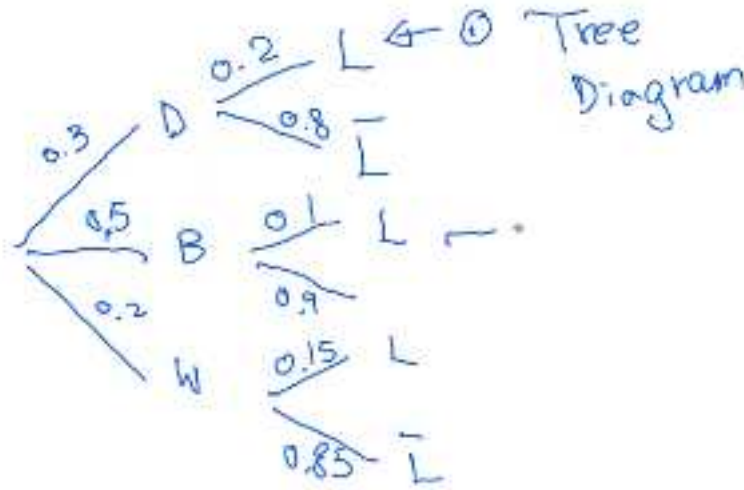
If  $A_1, A_2, \dots, A_k$  are a collection of mutually exclusive and exhaustive events, then for any event  $B$  (where  $P(B) \neq 0$ ) we have the following, for  $1 \leq i \leq k$ :

$$P(A_i|B) = \frac{P(A_i \cap B)}{P(B)}$$

$$= \frac{P(B|A_i)P(A_i)}{P(B|A_1)P(A_1) + P(B|A_2)P(A_2) + \dots + P(B|A_k)P(A_k)}$$

#### Example

using the previous tree calculate  $P(\text{Late})$



$$P(L \cap D) = 0.3 \cdot 0.2 = 0.06$$

$$P(B \cap L) = 0.5 \cdot 0.1 = 0.05$$

$$P(W \cap L) = 0.2 \cdot 0.15 = 0.03$$

$$P(L) = P(L \cap D) + P(B \cap L) + P(W \cap L)$$

$$P(L) = 0.06 + 0.05 + 0.03 = 0.14$$

### Example

Suppose that 30% of all students drive to school, 50% take the bus, and 20% walk. Of those who drive, 20% are usually late for their first class of the day. Of those who take the bus, 10% are usually late for their first class of the day. Of those who walk, 15% are usually late for their first class of the day. **Suppose that a student is late for class. What is the probability that this student walks to school?**

$$P(W|L) = \frac{P(W \cap L)}{P(L)}$$

$$P(W|L) = \frac{0.03}{0.14} = \frac{3}{14}$$

### 24. Set 7

### 25. Independent events

If A occurred but does not change the likelihood of B occurring, then A and B are Independent events.

If Independent then

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

### 26. Mutually Exclusive

The probability of A and B are mutually exclusive if and only if

$P(A \cap B) = 0$
-------------------

### Example

to check if a probability is independent or mutually exclusive just check

If  $P(A \cap B) = 0$  then it is **Mutually Exclusive**.

If  $P(B \cap A) = P(A)P(B)$  then it is **Independent**

## 27. Pairwise

if  $P(A_i \cap A_j) = P(A_i)P(A_j)$  for all  $i, j$ .)

These events A, B, C

Pairwise

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

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

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

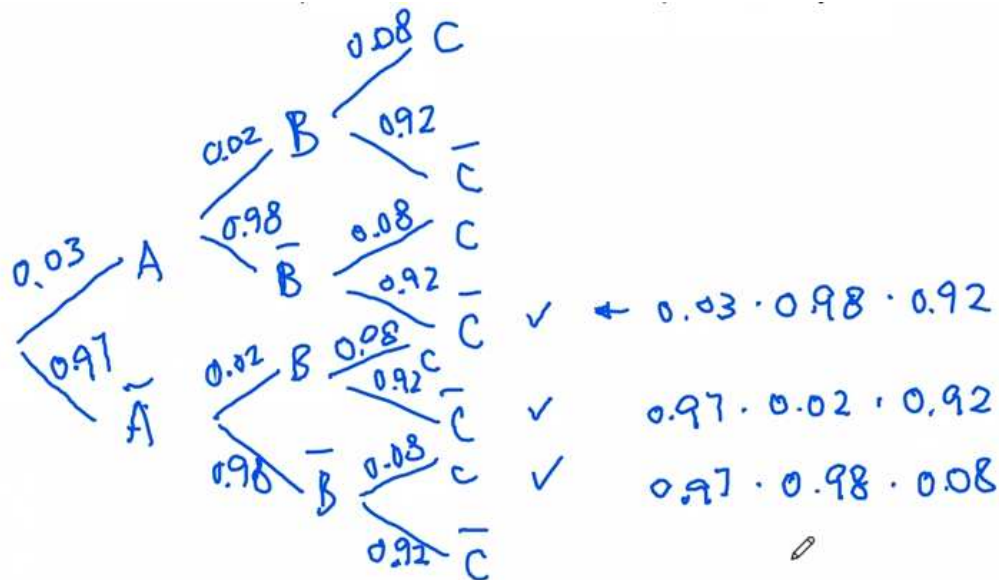
if Pairwise and

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

then it is just independent

### Example

A machine is made of three components (A,B,C) which function independently. The probability that components A,B,C will need to be repaired today is 0.03, 0.02, 0.08 (respectively). What is the probability **exactly one** of the three components will need to be repaired today?



$$P(A \cap \bar{B} \cap \bar{C}) + P(\bar{A} \cap B \cap \bar{C}) + P(\bar{A} \cap \bar{B} \cap C)$$

## STATS 260 Class 6

Gavin Jaeger-Freeborn

Thu 21 May 2020 11:20:51 PM

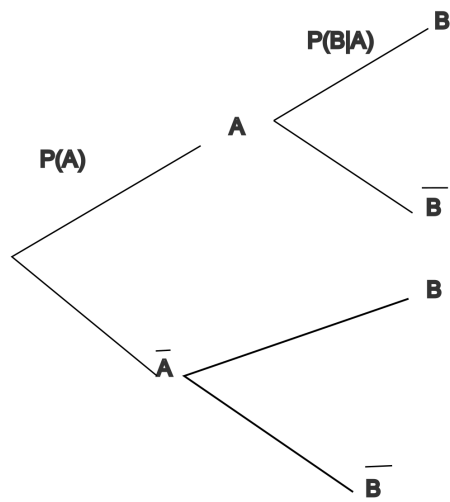
### 28. Multiplication Rule

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

This is from

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

**NOTE:** This is useful for tree diagrams



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

$$P(\bar{A} \cap B) = P(\bar{A})P(B|\bar{A})$$

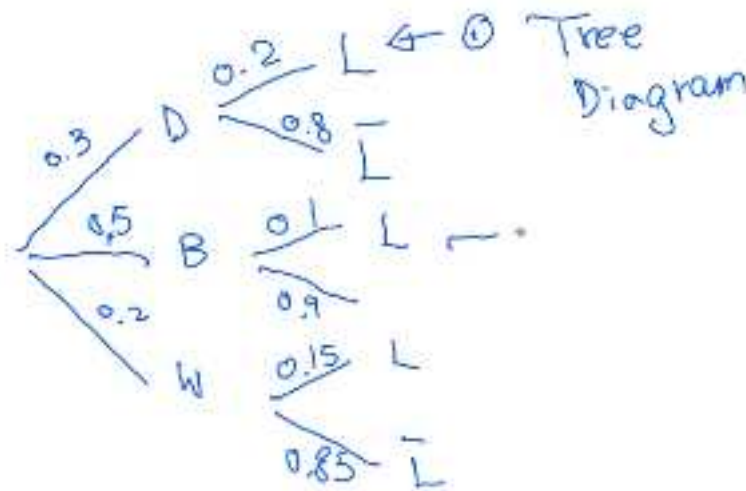


$$P(B) = P(A \cap B) + P(\bar{A} \cap B)$$

$$A \cup \bar{A} = S \quad A \cap \bar{A} = \phi$$

### Example

Suppose that 30% of all students drive to school, 50% take the bus, and 20% walk. Of those who drive, 20% are usually late for their first class of the day. Of those who take the bus, 10% are usually late for their first class of the day. Of those who walk, 15% are usually late for their first class of the day. What is the probability that a randomly selected student is regularly late for their first class?



$$P(L \cap D) = 0.3 * 0.2 = 0.06$$

$$P(B \cap L) = 0.5 * .1 = .05$$

$$P(W \cap L) = 0.2 * 0.15 = 0.03$$

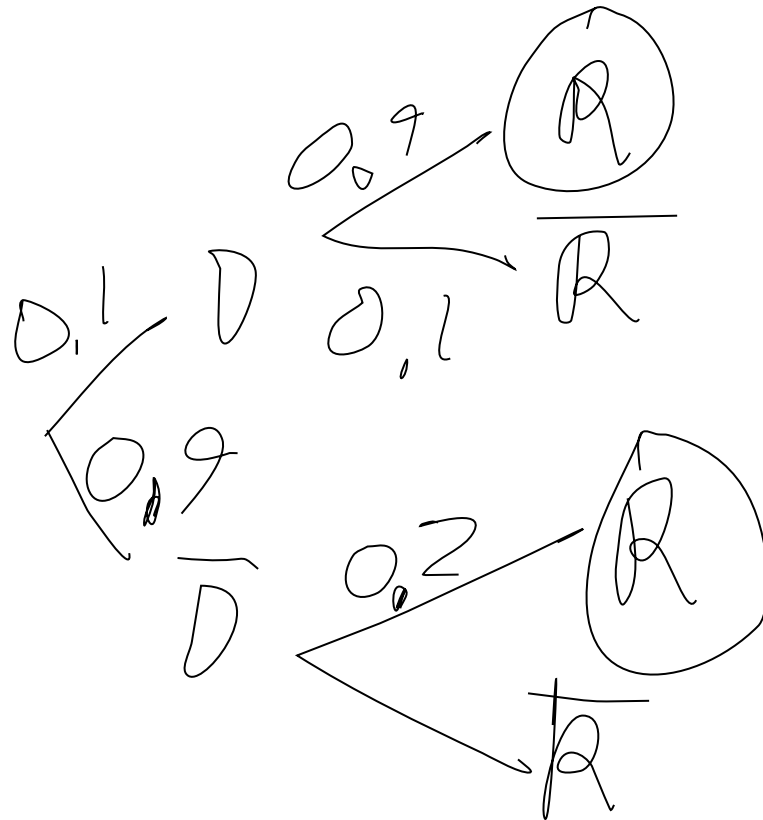
$$P(L) = P(L \cap D) + P(B \cap L) + P(W \cap L)$$

$$= 0.06 + 0.05 + 0.03 = 0.14$$

### Example

The probability of an item on a certain production line being defective is 0.1. If an item is defective, the probability that the inspector will remove it from the line is 0.9. If an item is not defective, the probability that the inspector will remove it from the line is 0.2.

What is the probability that a randomly selected item will be removed from the production line?



$$P(R) = (0.1)(0.9) + (0.9)(0.2) = 0.27$$

## 29. Law of Total Probability

if  $A_1, A_2, \dots, A_k$  are a collection of mutually exclusive and exhaustive events, then for any event  $B$  we have:

$$\begin{aligned} P(B) &= P(B \cap A_1) + P(B \cap A_2) + \dots + P(B \cap A_k) \\ &= P(B|A_1)P(A_1) + P(B|A_2)P(A_2) + \dots + P(B|A_k)P(A_k) \end{aligned}$$

### 30. Bayes Theorem

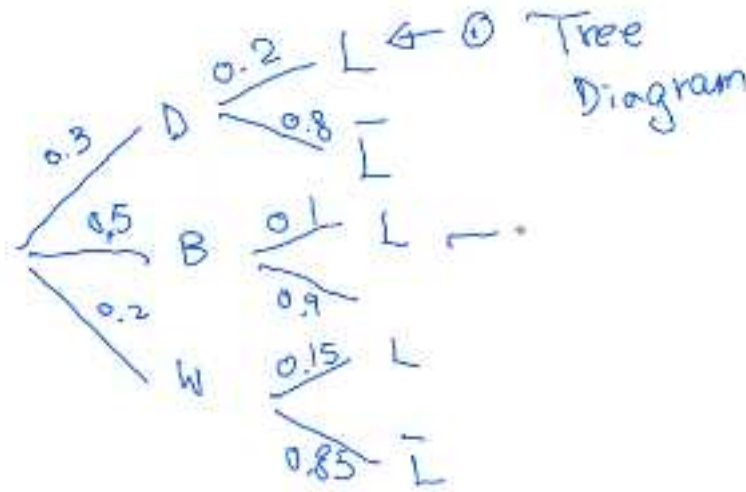
If  $A_1, A_2, \dots, A_k$  are a collection of mutually exclusive and exhaustive events, then for any event  $B$  (where  $P(B) \neq 0$ ) we have the following, for  $1 \leq i \leq k$ :

$$P(A_i|B) = \frac{P(A_i \cap B)}{P(B)}$$

$$= \frac{P(B|A_i)P(A_i)}{P(B|A_1)P(A_1) + P(B|A_2)P(A_2) + \dots + P(B|A_k)P(A_k)}$$

#### Example

using the previous tree calculate  $P(\text{Late})$



$$P(L \cap D) = 0.3 \cdot 0.2 = 0.06$$

$$P(B \cap L) = 0.5 \cdot 0.1 = 0.05$$

$$P(W \cap L) = 0.2 \cdot 0.15 = 0.03$$

$$P(L) = P(L \cap D) + P(B \cap L) + P(W \cap L)$$

$$P(L) = 0.06 + 0.05 + 0.03 = 0.14$$

### Example

Suppose that 30% of all students drive to school, 50% take the bus, and 20% walk. Of those who drive, 20% are usually late for their first class of the day. Of those who take the bus, 10% are usually late for their first class of the day. Of those who walk, 15% are usually late for their first class of the day. **Suppose that a student is late for class. What is the probability that this student walks to school?**

$$P(W|L) = \frac{P(W \cap L)}{P(L)}$$

$$P(W|L) = \frac{0.03}{0.14} = \frac{3}{14}$$

### 31. Set 7

### 32. Independent events

If A occurred but does not change the likelihood of B occurring, then A and B are Independent events.

If Independent then

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

### 33. Mutually Exclusive

The probability of A and B are mutually exclusive if and only if

$P(A \cap B) = 0$
-------------------

### Example

to check if a probability is independent or mutually exclusive just check

If  $P(A \cap B) = 0$  then it is **Mutually Exclusive**.

If  $P(B \cap A) = P(A)P(B)$  then it is **Independent**

### 34. Pairwise

if  $P(A_i \cap A_j) = P(A_i)P(A_j)$  for all  $i, j$ .)

These events A, B, C

Pairwise

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

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

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

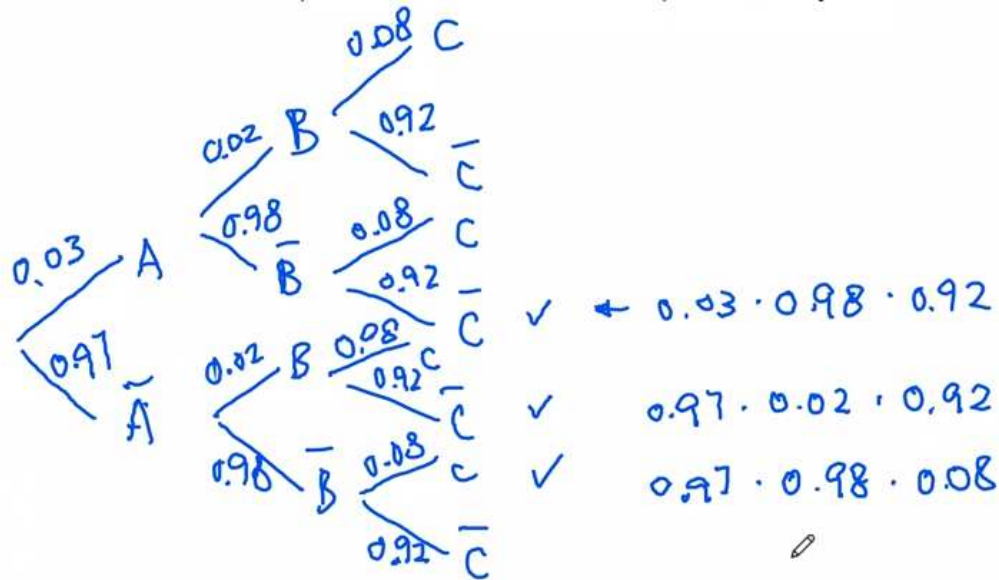
if Pairwise and

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

then it is just independent

### Example

A machine is made of three components (A,B,C) which function independently. The probability that components A,B,C will need to be repaired today is 0.03, 0.02, 0.08 (respectively). What is the probability **exactly one** of the three components will need to be repaired today?



$$P(A \cap \bar{B} \cap \bar{C}) + P(\bar{A} \cap B \cap \bar{C}) + P(\bar{A} \cap \bar{B} \cap C)$$

## STATS 260 Class 7

Gavin Jaeger-Freeborn

### 35. Probability Modeling

#### 35.1. Random Variable

a function which maps each outcome of an experiment to a number

$$\text{events} \rightarrow \#s$$

### Example

The number of defective items could be 0, 1,..., 10. Thus, X can take on the values 0, 1,..., 10.

$$X = \{0, 1, \dots, 10\}$$

Probability one item is defective is  $P(X=1)$

Probability at least 2 items are defective is  $P(X \geq 1)$

### Example

I randomly select a student and ask if they have taken Math 122. For this experiment, I have the random variable Y , which takes on two values: 0 and 1. The random variable Y will take a value of 1, if the answer is “Yes”, and will take on a value of 0 if the answer is “No”.

$$P(X = 0) \rightarrow NO, P(X = 1) \rightarrow YES, X \{0, 1, \}$$

### 35.2. Support

possible values it can take. In the last example question.

$$X = \{0, 1\}$$

#### 35.2.1. Continuous

Support is real numbers

#### 35.2.2. Discrete

Support is non real numbers

## 36. Probability Mass Function or Probability Distribution $f(X)$

$$f(x) = P(X = x)$$

### 36.1. Probability Distribution Table

x	0	1	...	10
f(x)	0.1	0.03	...	0.005

### Example

At a small taco shop, it has been noted that 80% of customers order beef tacos, and the other 20% of customers order veggie tacos. **Three customers** enter the store, and each customer independently orders one taco. Construct the probability distribution table for the random variable  $X$ , where  **$X$  is number of veggie tacos ordered** by the three customers.

Outcomes {BBB,VBB,BVB,BBV,VVB,VBV,BVV,VVV}

$X = 0 \rightarrow BBB$

$X = 1 \rightarrow VBB, BVB, BBV$

$X = 2 \rightarrow VVB, VBV, BVV$

$x = 3 \rightarrow VVV$

R.V.  $X$  Support of  $X = \{0, 1, 2, 3\}$

$$f(0) = P(X = 0) = P(BBB) = 0.8 \times 0.8 \times 0.8 = 0.512$$

$$f(1) = P(\{VBB, BVB, BBV\})$$

$$= (0.2)(0.8)(0.8) + (0.8)(0.2)(0.8) + (0.8)(0.8)(0.2) = 0.384$$

$$f(2) = P(\{VVB, VBV, BVV\})$$

$$= 3 \times (0.2)(0.2)(0.8) = 0.096$$

$$f(3) = P(\{VVV\})$$

$$= 0.2^3 = 0.008$$

x	0	1	2	3
f(x)	0.512	0.384	0.096	0.008

**NOTE:**

$$\sum_x f(x) = 1$$



What is the probability that exactly one veggie taco will be ordered?

$$P(x = 1) = f(1) = 0.384$$

What is the probability that at least two veggie tacos will be ordered?

$$\begin{aligned} P(X \geq 2) &= P(X = 2) + P(X = 3) \\ &= f(2) + f(3) \\ &= 0.96 + 0.008 = 0.104 \end{aligned}$$

Suppose we know that at **least one veggie taco** is ordered. What is the probability that **exactly two veggie tacos** will be ordered?

### Conditional Probability

$$P(X = 2 | X \geq 1)$$

$$\begin{aligned} P(A|B) &= \frac{P(A \cap B)}{P(B)} \\ &= \frac{P(X = 2 \cap X \geq 1)}{P(X \geq 1)} = \frac{P(X = 2)}{P(X \geq 1)} \end{aligned}$$

x	0	1	2	3
f(x)	0.512	0.384	0.096	0.008

$$\frac{0.096}{0.384 + 0.096 + 0.008} = \frac{0.096}{0.488} = \frac{12}{61}$$

### 37. Cumulative Distribution Function F(X) cdf

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

#### Example

Suppose the random variable X has the following probability distribution:

x	1	2	3	4	5
f(x)	0.3	0.15	0.05	0.2	0.3

Find the cdf for this random variable

$$F(1) = P(X \leq 1) = P(X = 1) = 0.3$$

$$F(2) = P(X \leq 2) = P(X = 1) + P(X = 2) = f(1) + f(2) = 0.3 + 0.15 = 0.45$$

$$F(3) = P(X \leq 3) = P(X = 1) + P(X = 2) + P(X = 3) = 0.3 + 0.15 + 0.05 = 0.5$$

$$F(4) = 0.7$$

$$F(5) = 1$$

x	1	2	3	4	5
F(x)	0.3	0.45	0.5	0.7	1

The easier way is to just add them

x	1	2	3	4	5
f(x)	0.3	0.15	0.05	0.2	0.3
F(x)	0.3	0.45	0.5	0.7	1

$$f(x) \rightarrow F(X)$$

### 37.1. Properties of a cdf

- $F(x)$  is monotone increasing.
- $\lim_{x \rightarrow -\infty} F(x) = 0$  and  $\lim_{x \rightarrow \infty} F(x) = 1$ .

#### Explanation

$$x \rightarrow \infty$$

$$P(X \leq x)$$

$$X \leq x \rightarrow \text{Sample Space}$$

Remember

$$P(S) = 1$$

When  $S$  is sample space

$$x \rightarrow -\infty$$

$\phi$  is the empty set

$$P(\phi) = 0$$

- $F(x)$  is right-continuous (continuous at each point  $x = k$  where  $x$  approaches  $k$  from the right)

**NOTE:** In the previous example, the support for the pmf was  $x = 1, 2, 3, 4, 5$ . As we've discussed previously, for any  $x$  which is not part of the support (i.e. impossible outcomes), the probability of that value of being observed is zero.

### Example

In the previous example, the event  $X = 3.5$  is an impossible event. Therefore,

$$f(3.5) = P(X = 3.5) = 0.$$

However, **this does not mean the cdf also has a value of zero :**

### Example

$$F(3.5) = P(X \leq 3.5)$$

x	1	2	3	4	5
F(X)	0.3	0.45	0.5	0.7	1

frame invis left solid bot solid label left "F(X)" "" aligned label bottom "X" line from 1,0.3 to 2,0.3 line from 2,.45 to 3,0.45 line from 3,.50 to 4,0.50 line from 4,.7 to 5,0.70 arrow from 5,1 to 6,1

$$\lim_{x \rightarrow k^+} F(X) = F(k)$$

### Example

Let the discrete random variable  $X$  count the number of classes a randomly selected UVic student is currently taking. The cdf for  $X$  is the following.

$x$	1	2	3	4	5	6	7
$F(x)$	0.15	0.25	0.4	0.6	0.75	0.90	1

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

- What is the probability that the student is taking no more than 4 classes?

$$P(X \leq 4) = F(4) = 0.6$$

- Calculate  $F(4.5)$ .

$$F(4.5) = F(4) = 0.6$$

- What is the probability that the student is taking at least 3 classes?

$$P(X \geq 3)$$

we can then use the complement of  $F(3)$  since  $F(3) = P(x \leq 3)$

$$P(X \geq 3) = 1 - P(x < 3)$$

$$= 1 - P(x \leq 2) = 1 - F(2)$$

$$= 1 - 0.25$$

$$\boxed{= 0.75}$$

- What is the probability that the student is taking exactly 3 classes?

$$P(x \leq 3) - P(x \leq 2) = F(3) - F(2) = 0.4 - 0.25$$

$$= 0.15$$

- What is the probability that the student is taking at **least 2 but no more than 5 classes**?

$$P(x \geq 2) \cap P(x \leq 5) = P(2 \leq x \leq 5)$$

$$F(5) = \{1, 2, 3, 4, 5\}, \text{ and } F(1) = \{1\}$$

$$F(5) - F(1) = 0.75 - 0.15 = 0.6$$

## STATS 260 Class 8

*Gavin Jaeger-Freeborn*

### 38. Population mean $E(X)$ (expected value) $\mu$

Let  $X$  be a discrete random variable. The **expected value**, or **mean**, of  $X$ , denoted by  $\mu$ , or by  $E(X)$  is

$$E(X) = \sum_{\text{all } x} x \cdot f(x)$$

$f(x)$  is the pmf of  $X$  probability mass function (pmf) or probability distribution

#### Example

Suppose that  $X$  has the following distribution.

$x$	5	15	100
$f(x)$	$\frac{1}{3}$	$\frac{1}{4}$	$\frac{5}{12}$

Find  $E(X)$  (center of distribution)

$$= 5(1/3) + 15(1/4) + 100(5/12)$$

$$= \frac{20}{12} + \frac{45}{12} + \frac{500}{12} = \frac{565}{12}$$

### Example

Approximately 40% of all laptops of a particular brand will need a battery replacement within 3 years of purchase. **Three laptops** of this brand are selected at random. What is the expected number of laptops (in each group of three laptops) which will need a battery replacement within 3 years of purchase?

R = replacement

N = no replacement

let  $X$  = # of laptops needing replacement

$E(X) = ?$

Guess : 3 laptops, 40% replacement

$$3 \times 0.4 = 1.2$$

$$P(X = 0) = P(NNN) = (0.6)^3 = 0.216 = f(0)$$

$$P(X = 1) = P(RNN, NRN, NNR) = 3 \times (0.6)^2(0.4) = 0.432 = f(1)$$

$$P(X = 2) = P(RRN, RNR, NRR) = 3 \times (0.6)(0.4)^2 = 0.288 = f(2)$$

$$P(X = 3) = P(RRR) = 0.4^3 = 0.064$$

$$E(X) = 0(0.216) + 1(0.432) + 2(0.288) + 3(0.064)$$

= 1.2 **NOTE:** this is exactly  $3 \times 0.4 = 1.2$

If  $X$  is a random variable., and  $Y = g(X)$  then:

$$E(Y) - E(g(X)) = \sum g(x)P(X = x) = \sum_x g(x)f(x) \equiv \mu_y \equiv \mu_{g(x)}$$

### Example

Using the pmf from the previous example, find  $E(X + 2)$ , and  $E(X^2)$ .

x	0	1	2	3
$g(x) = x + 2$	2	3	4	5
f(x)	0.216	0.432	0.288	0.064
$X^2$	0	1	4	9

$$E(X + 2) = 2 \cdot (0.216) + 3 \cdot (0.432) + 4 \cdot (0.288) + 5 \cdot (0.064)$$

$$= 3.2$$

$$[1.2 + 2]$$

$$E(X^2) = 0(0.216) + 1 \cdot (0.432) + 4 \cdot (0.288) + 9 \cdot (0.064)$$

$$= 2.16$$

$$[E(X)]^2 = 1.2^2 = 1.44$$

$$2.16 \neq 1.44$$

**NOTE:** In this example, and in most cases,  $E(X^2)$  is not the same thing as  $[E(X)]^2$ .

$$E(x^2) \neq [E(X)]^2$$

In general

$$E[g(x)] \neq g(E(X))$$

When is  $E[g(x)] = g[E(X)]$  ?

When  $g(x)$  is linear  $y = ax + b$

### 39. Laws of Expected Value: (a, b are constants)

$$1) E(b) = b$$

$$2) E(X + b) = E(X) + b$$

$$3) E(aX) = aE(X)$$



### 39.1. Notation

We may also express  $E(aX + b)$  as  $\mu_{aX+b}$ .

### 39.2. Proof of 2

$$\begin{aligned} E(X + b) &= \sum (x + b)f(x) \\ &= \sum xf(x) + \sum bf(x) \\ &= E(x) + b \sum f(x) \\ &= E(X) + B \end{aligned}$$

#### Example:

If the random variable  $X$  is known to have expected value 3.8, find  $E(7X + 3)$ .

$$\begin{aligned} E(X) &= 3.8 \\ &= 7E(X) + 3 \\ &7(3.8) + 3 \\ &= 29.6 \end{aligned}$$

#### Example:

For the laptop experiment, the cost for a replacement battery is \$30 per laptop. What is the expected cost for each group of three laptops? (Assume that each laptop will need at most one replacement battery.)

Let  $y$  = cost of each group of 3 laptops.

$$\begin{aligned} y &= 30X \\ E(Y) &= E(30x) = 30E(X) \\ &= 30 \times 1.3 = \$36 \end{aligned}$$

**Example**

Suppose a random variable  $X$  has the following cdf:

$x$	1	2	3
$F(x)$	0.3	0.8	1

- Find  $E(X)$ .

Select the closest to your unrounded answer:

must convert to  $f(x)$

$x$	1	2	3
$f(x)$	0.3	0.5	0.2

$$\sum x \cdot f(x) = 1.9$$

☒ (A) 2

(B) 3

(C) 4

(D) 5

- Find  $E(X^2)$ .

Select the closest to your unrounded answer:

$$E(X^2) = \sum x^2 f(x) = 4.1$$

(A) 3.5

☒ (B) 4

(C) 4.5

(D) 5

#### 40. Set 10

#### 41. Variance $V(X)$

The variance of  $X$  is written as  $\sigma^2$

REMEMBER this is related to the population not a sample

$$\sigma^2 = V(X) = E[(X - \mu)]$$

The **standard deviation** of  $X$  written  $\sigma_1$  is  $\sigma = \sqrt{\sigma^2}$

We can interpret  $V(X)$  in a similar way to  $E(X)$ : If we were to carry out the experiment many times, and each time keep track of the observed value of  $X$ , then the variance of these observed values would approach  $V(X)$ , as the number of repetitions of the experiment approaches infinity.

##### 41.1. Computational Formula for Variance

$$\sigma^2 = V(X) = E(X^2) - \mu^2$$

##### Laptop Example

$$E(X) = 1.2$$

$$E(X^2) = 2.16$$

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

$$= 2.16 - 1.2^2$$

$$V(X) = 0.72$$

#### STATS 260 Class 9

*Gavin Jaeger-Freeborn*

## 42. Set 10

### 43. Variance $V(X)$

The variance of  $X$  is written as  $\sigma^2$

REMEMBER this is related to the population not a sample

$$\sigma^2 = V(X) = E[(X - \mu)]$$

The **standard deviation** of  $X_1$  written  $\sigma_1$  is  $\sigma = \sqrt{\sigma^2}$

We can interpret  $V(X)$  in a similar way to  $E(X)$ : If we were to carry out the experiment many times, and each time keep track of the observed value of  $X$ , then the variance of these observed values would approach  $V(X)$ , as the number of repetitions of the experiment approaches infinity.

#### 43.1. Variance $V(X)$

$$\sigma^2 = V(X) = E(X^2) - \mu^2$$

#### Laptop Example

$$E(X) = 1.2$$

$$E(X^2) = 2.16$$

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

$$= 2.16 - 1.2^2$$

$$V(X) = 0.72$$

**NOTE:** standard deviation is just the  $\sqrt{V(X)} = \sigma$

**NOTE:**  $\sigma^2, \sigma \geq 0$

**43.2. Laws of Variance: (a, b are constants)**

1.  $V(b) = 0$
2.  $V(X + b) = V(X)$
3.  $V(aX) = (a)^2 V(X)$

**Example**

If the random variable X has  $V(X) = 2$ , then  $V(3X + 1) = (3)^2 V(X) = 9(2) = 18$ .

**Notation:**

We may write the variance of  $aX + b$  as either  $V(aX + b)$  or 2 as  $\sigma_{aX+b}^2$ .

We would write the standard deviation of  $aX + b$  as  $\sigma_{aX+b}$ .

**Important:**

These laws apply to variance, and **not** to standard deviation.

**Example:**

If the random variable X has  $\sigma = 5$ , find  $\sigma_{-2X+1}$ .

**Example:**

Suppose the random variable  $X$  has  $E(X) = 1.9$  and  $V(X) = 0.5$ .

Find  $E(3X + 2)$ .

Select the closest to your unrounded answer:

$$E(3X + 2) = 3E(X) + 2 = 3(1.9) + 2 = 7.7$$

(A) 2

(B) 4

(C) 6

☒ (D) 8

Find  $V(-4X + 8)$ .

Select the closest to your unrounded answer:

$$(-4)^2 V(X) = 16V(X) = 16 \times 0.5 = 8$$

(A) - 8

(B) 0

☒ (C) 8

(D) 16

## STATS 260 Class 10

*Gavin Jaeger-Freeborn*

### 44. Recap up to set 11

Capital letters	$X, Y, \dots$	Random Variables (r.v)
small letters	$x, y, \dots$	numerical values
discrete r.v	$P(X = x) = f(x)$	pmf
	$P(X \leq x) = F(x)$	cdf

## Parameters quantities about population

$E(X) = \mu$	mean, or expected value
$V(X) = \sigma^2$	variance
$SD(X) = \sigma$	standard deviation

## 45. Counting

### 45.1. Permutations

#### 45.1.1. n factorial

Permutations n distinct items is  $n!$ ,

$$n! = n(n-1)(n-2)\dots(2)(1)$$

**NOTE:** We define  $0!$  to be equal to 1.

#### Example

The number of different ways to arrange 4 people for a photograph is  $4! = 24$ .

$$4 \times 3 \times 2 \times 1$$

The number of arrangements of r items taken from a collection of n distinct items is:

$$P(n, r) = {}_n P_r = n^{(r)} = n \frac{n!}{(n-r)!}$$

#### Example

Suppose I have a class of 20 students. The number of ways I can select 4 of these students and arrange them for a photograph is:

$$\begin{aligned} {}_{20}P_4 &= \frac{20!}{16!} = 116280 \\ &= \frac{20 \times 19 \times 18 \times 17}{1} = \frac{20!}{16!} \end{aligned}$$

## 45.2. Combinations

The number of combinations (selections) of  $r$  items taken from a collection of  $n$  distinct items is:

$$C(n, r) = {}_n C_r = \binom{n}{r} = \frac{n!}{r!(n-r)!}$$

### Example

Suppose I have a class of 20 students. The number of ways I can select (but not arrange) 4 of these students is:

$$\binom{20}{4} = \frac{20!}{4!16!} = 4845$$

$${}_{20}C_4 = 4845$$

### Example

Suppose I have a box containing slips of paper, numbered 1, 2, . . . 30. If I select three of the thirty slips at random, what is the probability that all three slips show a number which is 9 or less?

$$n = 30$$

Select 3 means

$$= r = 3$$

$$\text{Prob} = \frac{n(A)}{n(S)}$$

$$= \frac{{}_9C_3}{{}_{30}C_3}$$



## 46. Set 11

### 47. Bernoulli Process

An experiment consisting of one or more trials, each having the following properties.

1. Each trial has exactly two outcomes, which we call success and failure.
2. The trials are independent of each other.
3. For all trials the probability of success,  $p$ , is a constant.

A **binomial experiment** is a Bernoulli process where  $n$ , the number of trials, is fixed in advance.

Let  $X$  count the number of successes in a binomial experiment. Then  $X$  is a binomial random variable, and we write  $X \sim \text{Bin}(n, p)$ , where  $n$  is the number of trials, and  $p$  is the probability of successes. For a binomial random variable,  $n$  and  $p$  are its parameters.

$\sim$  means  $X$  follows  $n, p$  parameters

#### Example

In a manufacturing process, each item has a probability of 0.05 of being defective, independent of all other items. Suppose 12 items are selected at random, and we let  $W$  denote the number of defective items.

**NOTE:** Defective is success

$$P = P(\text{success}) = 0.05$$

$$n = 12$$

$$W \sim \text{Bin}(12, 0.05)$$

$$w = 0, 1, \dots, 12$$

0 to all defective

#### 48. Binomial Probability Distribution

$$f(x) =$$

$$pmf =$$

$$P(X = x) = \binom{n}{x} p^x (1 - p)^{n-x} \text{ for } x = 0, 1, 2, \dots, n$$

where  $x$  = success

##### Example

On a multiple choice test, there are 10 questions, each with 8 possible responses. I will complete the test by randomly selecting answers. What is the probability that I will get 1 question correct?

*Assume Independence*

*Exactly*

Let  $X$  = # of correct answers

$$X = \text{Bin}(10, \frac{1}{8})$$

$$P(X = 1) = \binom{10}{1} \left(\frac{1}{8}\right)^1 \left(1 - \frac{1}{8}\right)^{10-1}$$

$$= \binom{10}{1} \left(\frac{1}{8}\right)^1 \left(\frac{7}{8}\right)^9$$

$$= 0.375$$

**Example**

In the manufacture of lithium batteries, it is found that 7% of all batteries are defective. Suppose that we test 6 randomly selected batteries. What is the probability that at least two batteries are defective?

$X$  = # of defective batteries

$$p = 0.07, n = 6$$

$$X \sim \text{Bin}(6, 0.07)$$

$$P(X \geq 2) = P(X = 2) + P(X = 3) + P(X = 4) + P(X = 5) + P(X = 6)$$

$$= 1 - P(X < 2)$$

$$= 1 - \binom{6}{0} 0.07^0 (0.93)^6 - \binom{6}{1} 0.07^1 (0.93)^5$$

$$= 0.0608$$

**48.1. Expected Value and Variance**

If  $X \sim \text{Bin}(n, p)$ , then:

$$E(X) = \sum_{x=0}^n x f(x) = \sum_{x=0}^n x \binom{n}{x} p^x (1-p)^{n-x}$$

$$E(X) = np \text{ and } V(X) = np(1-p)$$

**Example**

What is the expected number of defective lithium batteries per batch of 6? What is the variance?

$$n = 6, p = 0.07$$

$$E(X) = 6 \times 0.07 = 0.42$$

$$V(X) = 6 \times 0.07 \times 0.93 = 0.3906$$

## 48.2. Cumulative Distribution Tables F(X)

These tables give  $P(X \leq x)$  for “nice” values of  $n$  and  $p$

### Example

It is known that 20% of all tablet computers will need the touch-screen repaired within the first two years of use. Suppose we select 15 tablet computers at random.

$$n = 15, p = 0.2$$

$$X \sim \text{Bin}(15, 0.2)$$

What is the probability that no more than 6 tablets will need repairs to the touch-screen within the first two years of use?

$$P(X \leq 6)$$

n	r	p									
		0.10	0.20	0.25	0.30	0.40	0.50	0.60	0.70	0.80	0.90
14	0	0.2288	0.0440	0.0178	0.0068	0.0008	0.0001	0.0000			
	1	0.5846	0.1979	0.1010	0.0475	0.0081	0.0009	0.0001			
	2	0.8416	0.4481	0.2811	0.1608	0.0398	0.0065	0.0006	0.0000		
	3	0.9559	0.6982	0.5213	0.3552	0.1243	0.0287	0.0039	0.0002		
	4	0.9908	0.8702	0.7415	0.5842	0.2793	0.0898	0.0175	0.0017	0.0000	
	5	0.9985	0.9561	0.8883	0.7805	0.4859	0.2120	0.0583	0.0083	0.0004	
	6	0.9998	0.9884	0.9617	0.9067	0.6925	0.3953	0.1501	0.0315	0.0024	0.0000
	7	1.0000	0.9976	0.9897	0.9685	0.8499	0.6047	0.3075	0.0933	0.0116	0.0002
	8		0.9996	0.9978	0.9917	0.9417	0.7880	0.5141	0.2195	0.0439	0.0015
	9		1.0000	0.9997	0.9983	0.9825	0.9102	0.7207	0.4158	0.1298	0.0092
	10			1.0000	0.9998	0.9961	0.9713	0.8757	0.6448	0.3018	0.0441
	11				1.0000	0.9994	0.9935	0.9602	0.8392	0.5519	0.1584
	12					0.9999	0.9991	0.9919	0.9525	0.8021	0.4154
	13					1.0000	0.9999	0.9992	0.9932	0.9560	0.7712
	14						1.0000	1.0000	1.0000	1.0000	1.0000
15	0	0.2059	0.0352	0.0134	0.0047	0.0005	0.0000				
	1	0.5490	0.1671	0.0802	0.0353	0.0052	0.0005	0.0000			
	2	0.8159	0.3980	0.2361	0.1268	0.0271	0.0037	0.0003	0.0000		
	3	0.9444	0.6482	0.4613	0.2969	0.0905	0.0176	0.0019	0.0001		
	4	0.9873	0.8358	0.6865	0.5155	0.2173	0.0592	0.0093	0.0007	0.0000	
	5	0.9978	0.9389	0.8516	0.7216	0.4032	0.1509	0.0338	0.0037	0.0001	
	6	0.9997	0.9819	0.9434	0.8689	0.6098	0.3036	0.0950	0.0152	0.0008	
	7	1.0000	0.9958	0.9827	0.9500	0.7869	0.5000	0.2131	0.0500	0.0042	0.0000
	8		0.9992	0.9958	0.9848	0.9050	0.6964	0.3902	0.1311	0.0181	0.0003
	9		0.9999	0.9992	0.9963	0.9662	0.8491	0.5968	0.2784	0.0611	0.0022
	10		1.0000	0.9999	0.9993	0.9907	0.9408	0.7827	0.4845	0.1642	0.0127
	11			1.0000	0.9999	0.9981	0.9824	0.9095	0.7031	0.3518	0.0556
	12				1.0000	0.9997	0.9963	0.9729	0.8732	0.6020	0.1841
	13					1.0000	0.9995	0.9948	0.9647	0.8329	0.4510
	14						1.0000	0.9995	0.9953	0.9648	0.7941
	15							1.0000	1.0000	1.0000	1.0000

$$P(X \leq 6) = 0.9819$$

**Example**

What is the probability that exactly 5 tablets will need touch-

$$\begin{aligned}P(X = 5) &= \binom{15}{5} 0.02^5 (0.8)^{10} \\&= P(X = 5) - P(X \leq 4) \\&= 0.9389 - 0.8358 = 0.1031\end{aligned}$$

**Example**

What is the probability that at least 2 tablets will need touch- screen repairs?

$$\begin{aligned}P(X \geq 2) &= 1 - P(X < 2) = 1 - P(X \leq 1) \\&= 1 - 0.1671 \\&= 0.8329\end{aligned}$$

### Example

It is known that 30% of all laptops of a certain brand experience hard-drive failure within 3 years of purchase. Suppose that 20 laptops are selected at random. Let the random variable  $X$  denote the number of laptops which have experienced hard-drive failure within 3 years of purchase. If it is known that at least 3 laptops experience hard-drive failure, what is the probability that no more than 6 laptops will experience hard-drive failure?

$$X \sim \text{Bin}(20, 0.3)$$

Conditional Probability

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

$$P(X \leq 6 | X \geq 3) = \frac{P(X \leq 6 \cap X \geq 3)}{P(X \geq 3)}$$

$$= \frac{P(3 \leq X \leq 6)}{P(X \geq 3)}$$

don't forget to convert to  $F(X)$  aka cdf

$$= \frac{P(X \leq 6) - P(X \leq 2)}{1 - P(X \leq 2)}$$

look up in table

$$= \frac{0.6080 - 0.0355}{1 - 0.0355}$$

$$= 0.5936$$

## STATS 260 Class 11

*Gavin Jaeger-Freeborn*

### 49. Set 12

### 50. Poisson Experiment

An experiment having the following properties.

1. The number of successes that occur in any interval is independent of the number of successes occurring in any other interval. *non-overlapping interval*

2. The probability of success in an interval is proportional to the size of the interval. *Larger the interval larger the probability*
3. If two intervals have the same size, then the probability of a success is the same for both intervals.

### 51. Poisson Random Variable

If in a Poisson experiment,  $X$  counts the number of successes that occur in one interval of time/space, then  $X$  is a Poisson random variable. We write  $X \sim \text{Poisson}(\lambda)$ .

Where  $\lambda$  is the average number of successes per region/interval.

**NOTE:** Some books will use  $\mu$  rather than  $\lambda$  for the parameter of the Poisson random variable.

#### Example

At a bank, customers use the bank machine at an average rate of 40 customers per hour. Let  $X$  count the number of customers that use the machine in a 30-minute interval.

40 customers per hour

$$\lambda = 40 \text{ per hour}$$

we use 20 for a 30 minute interval

$$X \sim \text{Poisson}(\lambda = 20)$$

NO n

#### Example

At a busy intersection, it is noted that on average 5 cars pass through the intersection per minute. Let  $X$  count the number of cars which pass through the intersection in an hour.

$$X \sim \text{Poisson}(\lambda = 300)$$

### Example

Suppose that a typist makes on average 10 errors while typing 300 pages of text. Let  $X$  count the number of errors on one page of text.

Errors per page

$$X \sim \text{Poisson}(\lambda = \frac{10}{300})$$

### Example

We examine ten pages of text. Let  $Y$  count the number of pages with at least one error. The random variable  $Y$  is **not** Poisson. Why?

Assume pages are independent

$$n = 10, p = P(\text{at least one error per page})$$

Binary

$$y \sim \text{Bin}(10, p)$$

Poisson

$$X \sim \text{Poisson}(\lambda = \frac{1}{30})$$

The difference is that  $y$  counts the # of pages out of the 10 pages

## 52. Poisson Probability Distribution

$$f(x) = P(X = x) = \frac{e^{-\lambda} \lambda^x}{x!}$$

### Remember

Binomial has a set endpoint eg 1, 2 ,..., n

Poisson has no fixed end eg 1, 2 , ...



**Example**

Suppose a machine makes defective items at an average rate of 5 defective items per hour. What is the probability that the machine will make exactly 4 defective items in an hour?

$X$  = # of defective items per hour

$$X \sim \text{Poisson}(\lambda = 5)$$

$$\begin{aligned} P(X = 4) &= \frac{e^{-5} \cdot 5^4}{4!} \\ &= 0.1755 \end{aligned}$$

**52.1. Expected Value and Variance**

if  $X \sim \text{Poisson}(\lambda)$

$$E(X) = \lambda \text{ and } V(X) = \lambda$$

**Example**

What is the expected number of defective items made by the machine in an hour? What is the variance?

$$\lambda = \mu = E(X) = 5 \text{ defective items (per hour)}$$

$$\sigma^2 = V(X) = 5 \text{ item}^2$$

$$\sigma = \sqrt{5} \text{ defective items}$$

**52.2. Cumulative Distribution Tables**

These tables give  $P(X \leq x)$  for “nice” values of  $\lambda$

**Example**

Suppose the machine is watched for three hours. What is the probability that it will make no more than 12 defective items?

$$\lambda = 5 \text{ per hour}$$

$$X \sim \text{Poisson}(\lambda = 15)$$

(Recall that the machine makes on average 5 defective items per hour)

From table

$$P(X \leq 12) = 0.2676$$

### Example

What is the probability that at least 6 defective items will be made?

$$P(X \geq 6) = 1 - P(X \leq 5)$$

$\mu = 7$

$r$	10.0	11.0	12.0	13.0	14.0	15.0	16.0	17.0	18.0
0	0.0000	0.0000	0.0000						
1	0.0005	0.0002	0.0001	0.0000	0.0000				
2	0.0028	0.0012	0.0005	0.0002	0.0001	0.0000	0.0000		
3	0.0103	0.0049	0.0023	0.0011	0.0005	0.0002	0.0001	0.0000	0.0000
4	0.0293	0.0151	0.0076	0.0037	0.0018	0.0009	0.0004	0.0002	0.0001
5	0.0671	0.0375	0.0203	0.0107	0.0055	0.0028	0.0014	0.0007	0.0003

$$= 1 - 0.0028$$

$$= 0.9972$$

### Example

What is the probability that exactly 13 defective items will be made?

$$P(X = 13) = \frac{e^{-15} \cdot 15^{13}}{13!}$$

$$= P(X \leq 13) - P(X \leq 12)$$

$$= 0.3632 - 0.2676$$

$$= 0.0956$$

### Example

Suppose that a typist makes on average of 2 errors per page. [Poisson] Suppose the typist is creating a ten-page document. What is the probability that exactly three of the pages do not contain any errors?

let  $X$  be the number of errors per page

$$X \sim \text{Poisson}(\lambda = 2 \text{ per page})$$

let  $y$  be a number of pages that contain  
no errors (success)

Assuming they are independent

$$y \sim \text{Bin}(n = 10, p = P(X = 0)) = 0.1353$$

$$\begin{aligned} P(y = 3) &= \binom{10}{3} 0.1353^3 (1 - 0.1353)^7 \\ &= 0.1074 \end{aligned}$$

### 53. Poisson approximation to Binomial

If  $X$  is a binomial random variable where  $n$  is very large and  $p$  is very small then  $X$  can be approximated with a Poisson distribution with  $\lambda = np$ .

**NOTE:** Provided  $n \geq 100$  and  $np \leq 10$ , the approximation will be quite good. It will still be reasonably good when  $n \geq 20$ , as long as  $p \leq 0.05$ .

### Example

Brugada syndrome is a rare disease which afflicts 0.02% of the population. Suppose 10,000 people are selected at random and tested for Brugada syndrome. What is the probability that no more than 3 of the tested people will have Brugada syndrome?

$$X \sim \text{Bin}(n = 10000, p = 0.0002)$$

No table to look up

$$P(X \leq 3)$$

$$= P(X \leq 3)$$

$$X \sim \text{Poisson}(\lambda = 10000 \times 0.0002 = 2)$$

$$= 0.8571$$

#### 54. Sets 13 and 14

#### 55. Continuous Random Variable

A random variable which can assume an uncountable number of values (i.e. some interval of real numbers).

For a random variable, the **probability distribution** or **probability density function** (pdf) is a function  $f(x)$  satisfying

**NOTE:** Discrete random variable support is countable  
a.k.a finite number of outcomes or countably infinite [Poisson]

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

For any two numbers  $a$  and  $b$  with  $a \leq b$

Some immediate consequences

1.  $f(x) \geq 0$  for all  $x$
2.  $\int_{-\infty}^{\infty} f(x) dx = 1$

### STATS 260 Class 12

*Gavin Jaeger-Freeborn*

#### 56. Sets 13 and 14

### 57. Continuous Random Variable

A random variable which can assume an uncountable number of values (i.e. some interval of real numbers).

For a random variable, the **probability distribution** or **probability density function** (pdf) is a function  $f(x)$  satisfying

**NOTE:** Discrete random variable support is countable  
a.k.a finite number of outcomes or countably infinite [Poisson]

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

For any two numbers  $a$  and  $b$  with  $a \leq b$

Some immediate consequences

1.  $f(x) \geq 0$  for all  $x$
2.  $\int_{-\infty}^{\infty} f(x)dx = 1$

**Note:** Since a valid pdf must never be below the x axis, we can interpret  $P(a \leq X \leq b)$  as the area under  $f(x)$  on the interval  $[a, b]$ .

Some further consequences for any valid pdf:

1.  $P(X = a) = 0$  for any  $a$ .

$$P(X = a) = P(a \leq X \leq a) = \int_a^a f(x)dx = 0$$

Discrete =  $P(X = a) > 0$  a in support of  $X$ .

2.  $P(X \geq a) = P(X > a)$  and  $P(X \leq a) = P(X < a)$

$$= P(X < a) + P(X = a)$$

where  $P(X = a) = 0$

3.  $P(X \geq a) = 1 - P(X \leq a)$

if all Random Variables

$$= 1 - P(X < a)$$

Continuous

$$= 1 - P(X \leq a)$$

4.  $P(a \leq X \leq b) = P(X \leq b) - P(X < a)$  ( provided  $a \leq b$  )

$$= P(X \leq b) - P(X < a)$$

Example of a Continuous Random variable

### 58. Uniform Probability Distribution

For a uniform probability distribution, the pdf is:

$$f(x; a, b) = \frac{1}{b-a} \text{ where } a \leq x \leq b$$

**NOTE:**  $f(x) \neq P(X=x)$  in Continuous Random Variable

The graph of  $f(x)$  is a horizontal line segment from  $a$  to  $b$  with height  $1/(b-a)$ .

$$P(x_1 \leq X \leq x_2) = (\text{height}) \times (\text{width}) = \left( \frac{1}{b-a} \right) (x_2 - x_1)$$

eg

$$X \sim \text{Uniform}(1, 3)$$

$$f(x) = \begin{cases} 0 & x < 1 \\ 1/2 & 1 \leq x \leq 3 \\ 0 & x > 3 \end{cases}$$

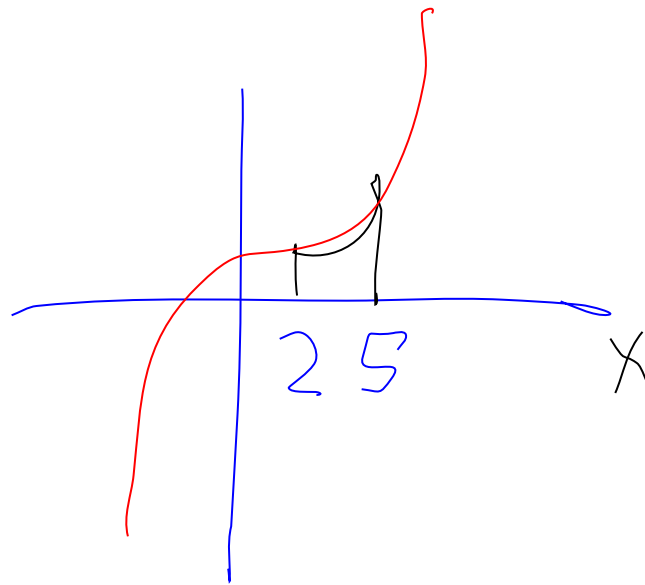
**Example**

Suppose that the continuous rv  $X$  has the following pdf:

$$X \sim \text{Uniform}(1, 3)$$

$$f(x) = \begin{cases} \frac{4}{609} x^3 & 1 \leq x \leq 3 \\ 0 & \text{otherwise} \end{cases}$$

Find  $P(3 \leq X \leq 4)$ .



$$\begin{aligned} &= \int_3^4 \frac{4}{609} x^3 dx \\ &= \frac{x^4}{609} \Big|_3^4 = \frac{4^4}{609} - \frac{3^4}{609} \\ &\quad \frac{25}{87} \end{aligned}$$

Check that

1.  $f(x) \geq 0$

1.  $\int_{-\infty}^{\infty} f(x) dx = \int_2^5 \frac{4x^3}{609} dx = 1$



**Example**

Find an expression for  $P(X \leq b)$ , where  $b$  is some number in  $[2, 5]$ .

$$F(b) = P(X \leq b)$$

$$= \int_2^b \frac{4}{609} x^3 dx$$

$$= \frac{x^4}{609} \Big|_2^b$$

$$= \frac{b^4}{609} - \frac{16}{609}$$

When  $b < 2$   $F(b) = 0$

When  $b > 5$

Put it together to get

$$f(x) = \begin{cases} 0 & x < 2 \\ \frac{x^4}{609} - \frac{16}{609} & 2 \leq x \leq 5 \\ 0 & x > 5 \end{cases}$$

**NOTE:** The fundamental theorem of calculus tells us that for every  $x$  at which  $F'(x)$  exists, that  $F'(x) = f(x)$ .

**Example**

Suppose the random variable X has the following cdf:

$$F(x) = \begin{cases} 0 & x < 0 \\ \frac{x}{x+1} & x \geq 0 \end{cases}$$

Find the pdf for the random variable X

$$\begin{aligned} f(x) - F'(x) &= \left( \frac{x}{x+1} \right)' \\ &= \frac{1(x+1) - x \cdot 1}{x+1^2} \\ &= \frac{1}{x+1^2} \geq 0 \\ f(x) &= \begin{cases} 0 & x < 0 \\ \frac{1}{x+1^2} & x \geq 0 \end{cases} \end{aligned}$$

**STATS 260 Class 13**

*Gavin Jaeger-Freeborn*

Let p be a number between 0 and 1. The 100p th percentile of a continuous random variable is the value  $\alpha$  such that  $F(\alpha) = p$ .

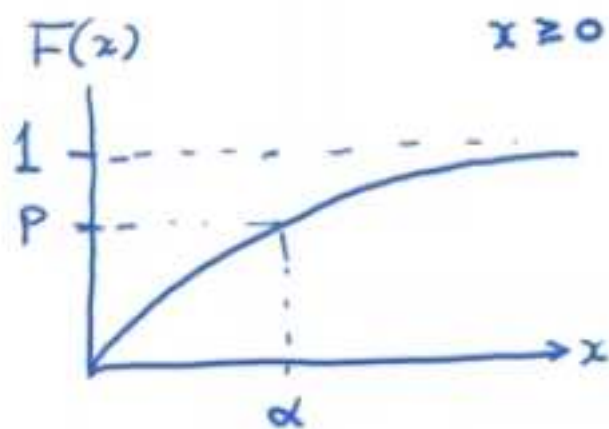
**Example**

For the random variable from the previous example, find the 90 th percentile.

$$F(\alpha) = \frac{\alpha}{1+\alpha} = 0.9$$

$$\alpha = 0.9 + 0.9\alpha$$

$$0.1\alpha = 0.9\alpha = 1$$



### 59. Mean and Variance of an Interval

The **expected value** or **mean** of a continuous random variable  $X$  with pdf  $f(x)$  is:

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

similar to

$$E(X) = \mu = \sum_x x f(x)$$

(provided this integral converges)

The **variance** of a continuous random variable  $X$  with pdf  $f(x)$  is:

$$V(X) = \sigma^2 = \int_{-\infty}^{\infty} (x - \mu)^2 f(x) dx$$

similar to

$$V(X) = \sigma^2 = \sum_x (x - \mu)^2 f(x)$$

(provided this integral converges) and the standard deviation,  $\sigma = \sqrt{\sigma^2}$ .

As with discrete random variables, we have the following:

- $V(X) = E(X^2) - \mu^2$
- $E(aX + b) = aE(X) + b$
- $V(aX + b) = a^2 V(X)$

**Example**

Suppose the random variable  $X$  has pdf

$$f(x) = \begin{cases} 2e^{-2x} & 0 \leq x < \infty \\ 0 & \text{otherwise} \end{cases}$$

Find the median of the distribution.

**NOTE:** The median,  $\tilde{\mu}$  of a continuous random variable is the 50<sup>th</sup> percentile.

$$\begin{aligned} \mu &= E(x) = \int_0^{\infty} x 2e^{-2x} dx \\ &= \lim_{b \rightarrow \infty} \int_0^b x \cdot 2e^{-2x} dx \\ &= \lim_{b \rightarrow \infty} \left[ (-be^{-2b} - \frac{e^{-2b}}{2}) - (0 - \frac{1}{2}) \right] \\ &= \lim_{b \rightarrow \infty} \left[ (-be^{-2b} - \frac{e^{-2b}}{2}) - \lim_{b \rightarrow \infty} (0 - \frac{1}{2}) \right] \\ &= \frac{1}{2} \end{aligned}$$

$$E(X^2) = \int_0^{\infty} x^2 2e^{-2x} dx$$

$$V(X) = \frac{1}{2} - \left(\frac{1}{2}\right)^2 = \frac{1}{4}$$

**STATS 260 Class 14**

*Gavin Jaeger-Freeborn*

### 61. Normal Density Function

if  $X$  is normally distributed with the mean  $\mu$  and standard deviation  $\sigma$ . then we write  $X \sim N(\mu, \sigma)$ . The pdf of  $X$  is:

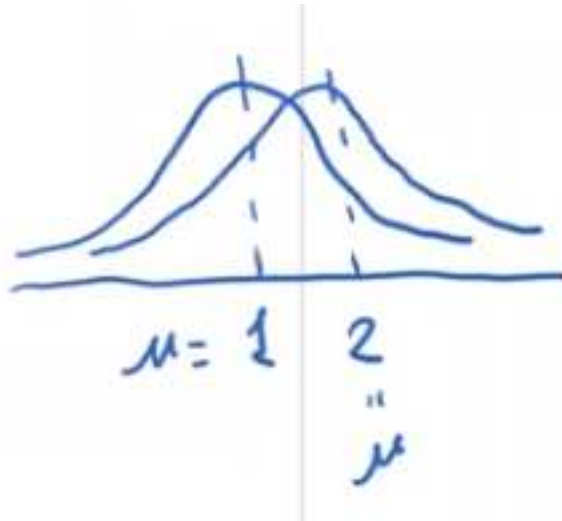
$$f(x; \mu, \sigma) = \frac{1}{\sigma \sqrt{2\pi}} e^{-\frac{1}{2} \left( \frac{x-\mu}{\sigma} \right)^2}$$

Properties of Normal Curves:

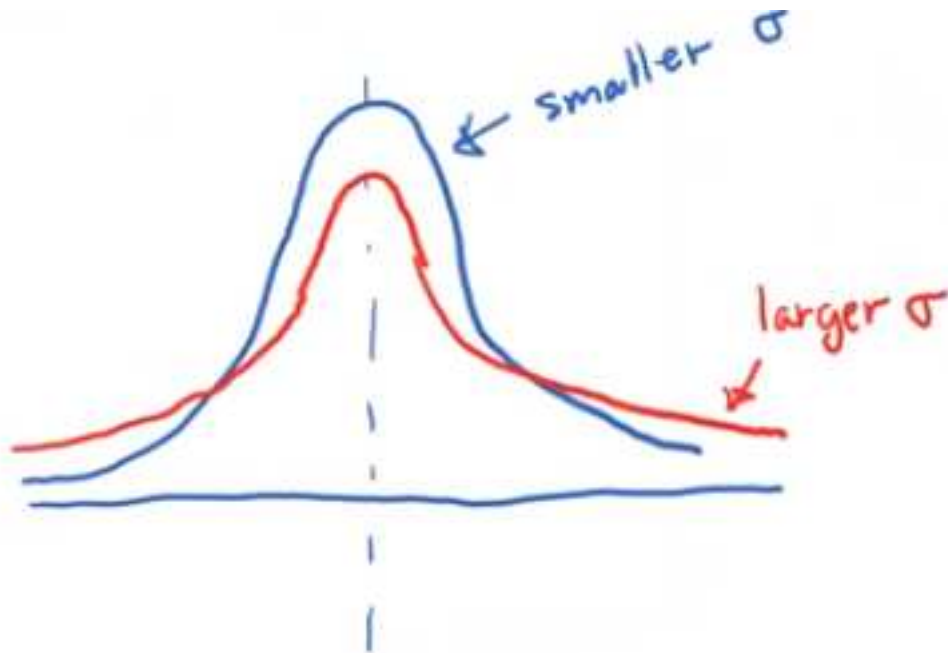
- All normal curves are defined on  $(-\infty, \infty)$  and is bell-shaped.
- There is a single peak at  $x = \mu$  and the curve is symmetric about this peak.

$\mu$  is the max of  $f(x)$

- The mean, median, and mode are all  $\mu$ ; the variance is  $\sigma^2$ .
- There are points of inflection at  $\mu - \sigma$  and  $\mu + \sigma$
- As  $\mu$  increases, the peak moves further to the right. As  $\mu$  decreases, the peak moves further to the left. ( $\mu$  is a **location parameter**)



- As  $\sigma$  increases, the peak becomes lower, and the curve becomes flatter. As  $\sigma$  decreases, the curve becomes more abruptly peaked, and the peak becomes taller. ( $\sigma$  is a **scale** parameter).



$$E(X) = \mu$$

$$V(X) = \sigma^2$$

## 62. Standard Normal Distribution

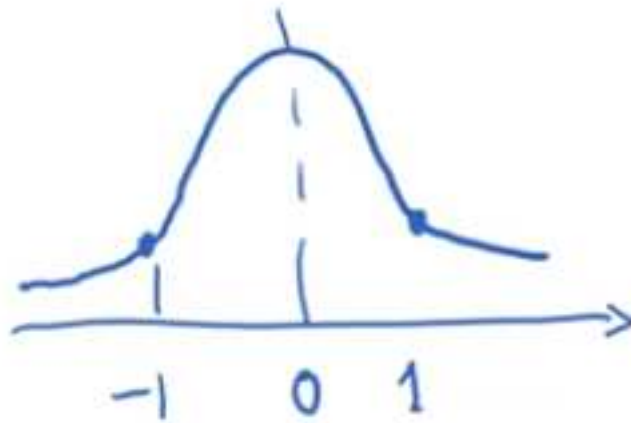
the standard normal random variable has mean 0 and standard deviation 1. We use the letter  $Z$  to denote the standard normal distribution.

$$N(0, 1), \mu = 0, \sigma = 1$$

$$f(z; 0, 1) = \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}}$$

The standard normal curve is:

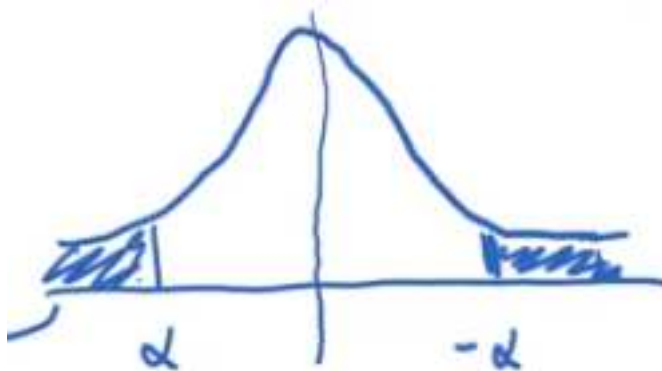
- has its peak at 0, and is symmetric about the y-axis
- has points of inflection at 1 and -1. If our random variable is  $Z$ , then we denote the cdf  $P(Z \leq z)$  by  $\Phi(z)$ .



### 63. Symmetry Property

Since the random variable  $Z$  is symmetric about  $Z = 0$ , then for any  $\alpha$

$$P(Z \leq \alpha) = P(Z \geq -\alpha)$$





### Example

Find  $P(Z \leq 2.56)$

### Solution

$$\int_{-\infty}^{2.56} f(x)dx$$

USE THE TABLE

→ 2.5	0.9938	0.9940	0.9941	0.9943	0.9945	0.9946	0.9948	0.9949
2.6	0.9953	0.9955	0.9956	0.9957	0.9959	0.9960	0.9961	0.9962
2.7	0.9965	0.9966	0.9967	0.9968	0.9969	0.9970	0.9971	0.9972
2.8	0.9974	0.9975	0.9976	0.9977	0.9977	0.9978	0.9979	0.9979
2.9	0.9981	0.9982	0.9982	0.9983	0.9984	0.9984	0.9985	0.9985
3.0	0.9987	0.9987	0.9987	0.9988	0.9988	0.9989	0.9989	0.9989
3.1	0.9990	0.9991	0.9991	0.9991	0.9992	0.9992	0.9992	0.9992
3.2	0.9993	0.9993	0.9994	0.9994	0.9994	0.9994	0.9994	0.9995
3.3	0.9995	0.9995	0.9995	0.9996	0.9996	0.9996	0.9996	0.9996
3.4	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997

Table D.3 (continued) Areas under the Normal Curve

$$P(Z \leq 2.56) = 0.9948$$

or in R

```
> pnorm(2.56)
```

### Example

Calculate  $P(Z \geq 0.16)$

### Solution

$$1 - P(Z \leq 0.16) \text{ or } P(Z \leq -0.16)$$

```
> 1 - pnorm(0.16)
[1] 0.4364405
```

$$\therefore P(Z \geq 0.16) = 0.4364405$$

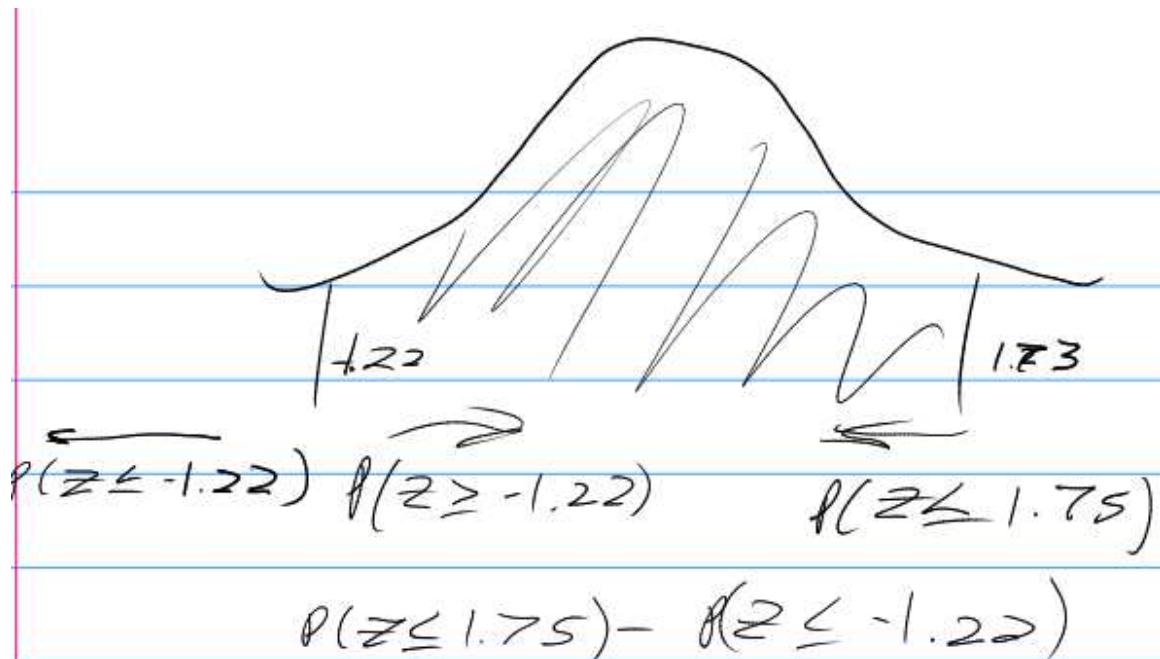
### Example

Calculate  $P(-1.22 < Z \leq 1.73)$

### Solution

$$P(Z \leq 1.73) - P(Z \leq 1.22)$$

$$pnorm(1.73) - pnorm(-1.22)$$



### Example

Suppose that the heights of Andean flamingos are normally distributed with a mean of 105 cm and a standard deviation of 2 cm. Let the random variable  $X$  denote the height of a randomly selected Andean flamingo

What is the **median** Andean flamingo height?

$$\mu = 105, \sigma = 2$$

$$X \sim N(\mu = 105, \sigma = 2)$$

$$\boxed{105 \text{ cm}}$$

is  $P(X \geq 100) = P(X \leq -100)$  ?

Only true for  $Z \sim N(0, 1)$

$\therefore$  FALSE

What is  $P(X = 105)$

As a property of all **continuous random variables**

$$P(X = 105) = 0$$

is  $P(X \leq 100) = P(x \geq 110)$

$$\boxed{\text{True}}$$

Due to symmetry about its mean

Remember

$$P(X \leq \mu - x) = P(X \geq \mu + x)$$

NOTATION:  $z_\alpha$  is the number such that  $P(Z > z_\alpha) = \alpha$  . Alternatively,  $z_\alpha$  is the  $100(1 - \alpha)$  percentile of the standard normal distribution.

**Example**

Find the 97.5<sup>th</sup> percentile of the standard normal distribution

$$100 - 97.5 = 2.5/100 = 0.025$$

$$Z_{0.025}$$

$$P(Z \leq Z_{0.025}) = 0.975$$

USE THE TABLE and find where the probability = 0.9750 and solve for z

$$Z_{0.025} = 1.96$$

in R

```
qnorm( 0.975 )
```

#### 64. Standardizing a Normal Random Variable

If  $X$  is normally distributed with mean  $\mu$  and standard deviation  $\sigma$ , then:

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

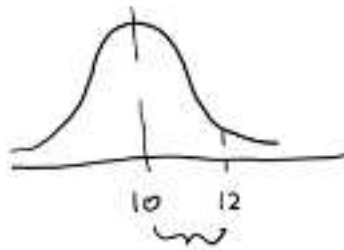
NOTE : this can be used for  $P(Z \leq z)$  eg  $X \sim N(105, 2)$

This basically means how many standard deviations away from the mean.

$$X = 12, \mu = 10 \sigma = 2$$

$$Z = \frac{12 - 10}{2}$$

= 1 standard deviation from the mean



**Example**

The masses of a certain type of bolt is approximately normally distributed with  $\mu = 15$  g, and  $\sigma = 2$  g. What is the probability that a **randomly selected bolt** has a mass between 14.3 g and 17.1 g?

$$P(14.3 \leq X \leq 17.1) = P(X \leq 17.1) - P(X \leq 14.3)$$

$$P\left(\frac{X - \mu}{\sigma} \leq \frac{14.3 - 15}{2}\right)$$

$$P\left(Z \leq \frac{14.3 - 15}{2}\right)$$

$$P(Z \leq -0.35)$$

$$P\left(Z \leq \frac{17.1 - 15}{2}\right)$$

$$P(Z \leq 1.05) - P(Z \leq -0.35)$$

USING THE TABLE

$$= 0.8531 - 0.3632 = 0.4899$$

What is the probability that a randomly selected bolt will have a mass of at least 20 g?

$$P(X \geq 20) = P\left(Z \geq \frac{20 - 15}{2}\right)$$

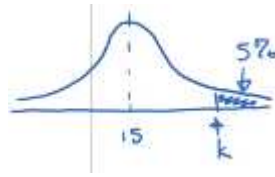
$$P(X \geq 20) = 1 - P\left(Z \leq \frac{20 - 15}{2}\right)$$

$$P(X \geq 20) = 1 - P(Z \leq 2.5)$$

FROM TABLE

$$1 - 0.9938 = 0.0062$$

What is the minimum mass of the heavy 5% of bolts ?



$$P(X \leq k) = 9.5$$

$$P(Z \leq \frac{k - \mu}{\sigma}) = 9.5$$

REVERSE ON TABLE

$$P(Z \leq 1.645) = 0.95$$

$$\frac{k - 15}{2} = 1.645$$

solve for k

$$k = 2 ( 1.645 ) + 15 = 18.29$$

### 65. empirical rule

The empirical rule states that if the distribution of a variable is approximately normal, then:

1. About 68% of values lie within  $\sigma$  of the mean.
2. About 95% of values lie within  $2\sigma$  of the mean.
1. About 99.7% of values lie within  $3\sigma$  of the mean.



From this, we can conclude that almost all bolts will have a mass within 6g of the mean 15 ( ie about 99.7% will have a mass between 9 g and 21 g ).

$$\sigma = 2$$





## 66. Approximating the Binomial Distribution with the Normal Distribution

Suppose  $X \sim \text{Bin}(n, p)$  where  $np$  and  $n(1 - p)$  are both at least 5.

Then  $X = N(\mu = np, \sigma^2 = np(1 - p))$

This means that:

if

$$np \geq 5, \text{ and } n(1 - p) \geq 5$$

$$P(X \leq x) = P\left(Z \leq \frac{x - np}{\sqrt{np(1 - p)}}\right)$$

Since we are using continuous distribution to approximate a discrete one, this approximation will be slightly off. If we wish to get a better approximation use the following, with a **continuity correction**

$$P(X \leq x) = P(X \leq x + 0.5)$$

The +0.5 is for correction

$$= P\left(Z \leq \frac{x - np + 0.5}{\sqrt{np(1 - p)}}\right)$$

eg

$$P(X \leq 3) = P(x \leq 3 + 0.5)$$

**Example**

Suppose it is known that 20% of batteries have a lifespan shorter than the advertised lifespan. Suppose that 100 batteries are selected at random.

What is the approximate probability (using the continuity correction) that at least 10 batteries will have a short lifespan?

$$X \sim \text{Bin}(100, 0.2)$$

$$np = 100(0.2) = 20$$

$$n(1 - p) = 100(0.8) = 80$$

$$\text{Both are } \geq 5$$

$$P(X \geq 10)$$

$$= 1 - P(X \leq 9)$$

$$= 1 - P(X \leq 9 + 0.5)$$

$$X \sim N(20, \sqrt{100(0.2)(0.8)})$$

$$= N(20, 4)$$

$$= 1 - P(X \leq 9.5)$$

$$= 1 - P(Z \leq -2.625)$$

USING THE TABLE

$$1 - 0.0043 = 0.9957$$

**Example**

Suppose it is known that the reaction time of type of voiceactivated robot is normally distributed with  $\mu = 6.3$  microseconds, and  $\sigma = 2$  microseconds.

Suppose I select one voice-activated robot at random. What is the probability that its reaction time is between 5 and 7 microseconds? Report your answer to three decimal places.

$$X \sim N(6.3, 2)$$

$$P(5 < X < 7)$$

$$= P(X < 7) - P(X < 5)$$

$$= P(Z < 0.35) - P(Z < -0.65)$$

FROM TABLE

$$= 0.6368 - 0.2578 = 0.379$$

**Example**

Suppose that I select five robots and test each of them. Assume the reaction time of each robot is independent of exactly three of the robots will have a reaction time between 5 and 7 microseconds? Report your answer to three decimal places.

$Y = \#$  of robots having reaction time between 5 and 7 microseconds

$$y \sim \text{Bin}(n = 5, p = 0.379)$$

$$P(y = 3) = \binom{5}{3} 0.379^3 (1 - 0.379)^2$$

$$= 0.210$$

**STATS 260 Class 15 and 16**

*Gavin Jaeger-Freeborn*

## 67. Sets 17 Gamma functions and exponential Distribution

### 68. Gama Function

the **gamma function**  $\Gamma(\alpha)$  is defined for  $\alpha > 0$  by:

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

Where  $\alpha$  is some positive real number

It can be shown through integration by parts that the gamma function satisfies the relation  $\Gamma(\alpha + 1) = \alpha\Gamma(\alpha)$  for all  $\alpha > 0$ . It can also be shown that  $\Gamma(1) = 1$ .

**NOTE:**  $\Gamma(\alpha + 1) = \alpha\Gamma(\alpha)$  is a recursive relation

Putting these two facts together yields the property that  $\Gamma(n) = (n - 1)!$  for any positive integer  $n$ .

A continuous random variable  $X$  has gamma distribution with parameters  $\alpha > 0$  and  $\beta > 0$  if the pdf is

**NOTE:**  $\alpha, \beta$  are fixed

$$f(x) = f(x; \alpha, \beta) = \begin{cases} \frac{1}{\beta^\alpha \Gamma(\alpha)} x^{\alpha-1} e^{-x/\beta} & x \geq 0 \\ 0 & \text{otherwise} \end{cases}$$

This means

$$X \sim \Gamma(\alpha, \beta) \quad \alpha, \beta > 0$$

Check that  $\int f(x) dx = 1$

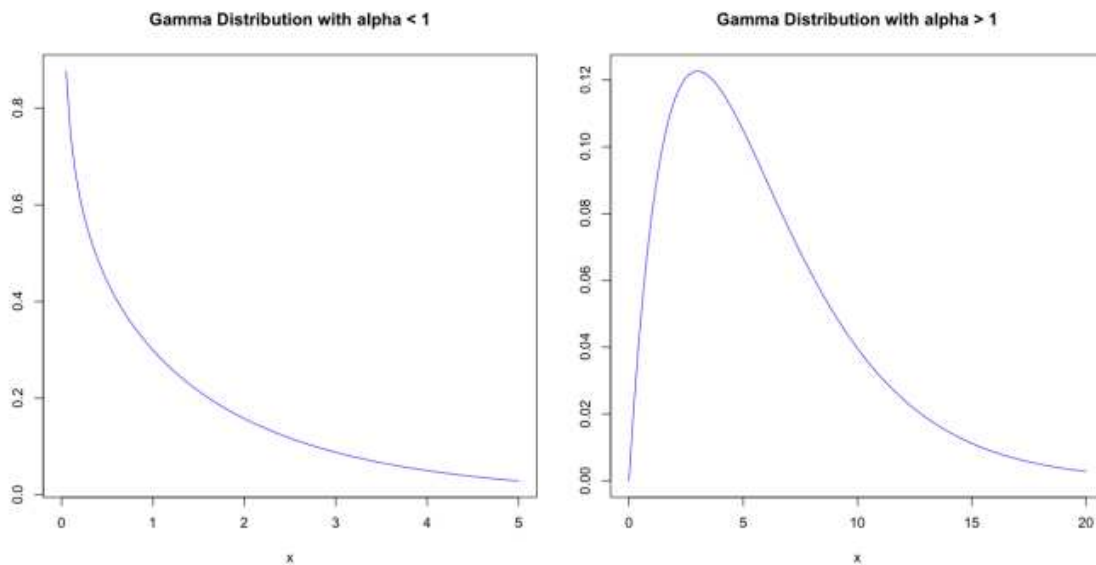
HINT: let  $u = \frac{x}{\beta}$  and integrate

**NOTE:**  $x \geq 0$  compared to normal distribution which is  $-\infty < x < \infty$

The gamma distribution is often used as a probability model for waiting times (e.g. time until death, time until failure).

We call  $\beta$  the **scale parameter** (since it stretches/compresses the pdf) and  $\alpha$  the **shape parameter** (since it determines the shape of the pdf).

- $E(X) = \alpha \beta$
- $V(X) = \alpha \beta^2$
- There are two basic shapes for the gamma distribution. The left image is the shape for  $\alpha \leq 1$ , and the right image is for  $\alpha > 1$



- For most values of  $\alpha$ ,  $\beta$  a closed-form expression for the cdf does not exist; tables or software packages are used. In cases where  $\alpha$  is an integer, however, we can calculate probabilities by integrating.

**Example**

Suppose  $X \sim \text{Gamma}(\alpha = 2, \beta = 3)$ . Calculate  $P(X \leq 5)$ .

$$f(x) = \frac{1}{\beta^\alpha \Gamma(\alpha)} x^{\alpha-1} e^{-\frac{x}{\beta}} dx$$

$$P(X \leq 5) = F(5)$$

$$= \int_0^5 \frac{1}{3^2 \Gamma(2)} x^{2-1} e^{-\frac{x}{3}} dx$$

Using Integration By Parts

$$= \frac{1}{9} \int_0^5 x^1 e^{-\frac{x}{3}} dx$$

in R

shape =  $\alpha$

scale =  $\beta$

```
pgamma ( 5 , shape = 2 , scale = 3 )  
= 0.4963
```

## 69. Exponential Distribution

The **exponential distribution** is a member of the gamma family when  $\alpha = 1$ . The random variable  $X$  has exponential distribution with parameter  $\lambda$  ( $\lambda > 0$ ) if the pdf is:

$$f(x) = f(x; \lambda) = \begin{cases} \lambda e^{-\lambda x} & x \geq 0 \\ 0 & \text{otherwise} \end{cases}$$

**NOTE:** now  $\beta = \frac{1}{\lambda}$

**NOTE:** Be aware that a second definition exists, with a parameter  $\theta$  where  $\theta = 1/\lambda$ . We will not be using this alternate definition.

We find  $E(X)$  and  $V(X)$  either by integrating, or by recognizing that if  $X \sim \text{Exp}(\lambda)$  then  $X \sim \text{Gamma}(\alpha = 1, \beta = 1/\lambda)$ . Either way gives us:

- $E(X) = \frac{1}{\lambda}$
- $V(X) = \frac{1}{\lambda^2}$

**NOTE:** same as  $E(X) = \alpha \beta$  and  $V(X) = \alpha \beta^2$  unlike other gamma distributions, the pdf of the exponential distribution can be easily integrated, giving us

$$P(X \leq x) = F(x; \lambda) = \begin{cases} 1 - e^{-\lambda x} & x \geq 0 \\ 0 & \text{otherwise} \end{cases}$$

$$\therefore F(x) = \int_0^x \lambda e^{-\lambda x}$$

$$P(X > x) = e^{-\lambda x}$$

### Example

During the lunch hour, the average waiting time to use an automatic bank machine is 6 minutes. Let the random variable  $X$  measure the time (in minutes) that a customer waits before service. It is known that  $X$  has exponential distribution. What is the probability that a customer will need to wait at least 9 minutes?

$$X \sim \text{Exp}(\lambda \frac{1}{6})$$

$$\mu = \frac{1}{\lambda} \quad \lambda = \frac{1}{\mu} = \frac{1}{6}$$

$$E(X) = 6 \text{ minutes} = \mu$$

$$P(X \geq 9) = P(X > 9) = e^{-\frac{9}{6}} = 0.2231$$

## 70. Relationship between Poisson and Exponential Distributions

Suppose we have a Poisson process, where events occur at a rate of  $\lambda$  occurrences per unit of time/space.

If random variable  $X$  denotes the number of occurrences of an event in a unit of time/space then  $X \sim \text{Poisson}(\lambda)$ .

If we now let the random variable  $Y$  measure the units of time/space until the next occurrence then  $Y \sim \text{Exp}(\lambda)$ .

Example: In our last example, the time (in minutes) between customers had exponential distribution with  $\lambda = 1/6$ .

If we now count the number of customers per minute, then this would have Poisson distribution with  $\lambda = 1/6$ . There is an average rate of is  $1/6$  customers per minute (or 1 customer per 6 minutes) for the machine.

**NOTE:** More generally, there is also a relationship between Poisson and Gamma distributions. Suppose again that we have a Poisson process, where events occur at a rate of  $\lambda$  occurrences per unit of time/space. If we let the random variable  $Y$  measure the units of time/space until the  $k$ th occurrence, then  $Y \sim \text{Gamma}(\alpha = k, \beta = 1/\lambda)$ .

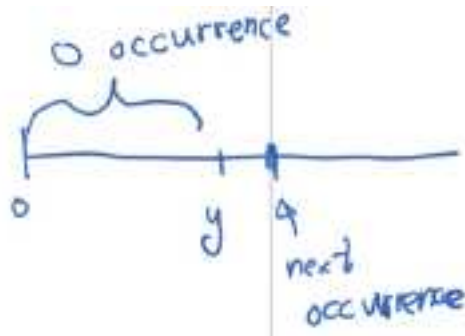


### Proof

let  $w$  = # of occurrences of an event over  $y$  units of time/space

$w \sim \text{Poisson}(\lambda y)$

$$\begin{aligned} p(y \leq y) &= 1 - P(Y > y) \\ &= 1 - P(W = 0) = \frac{e^{-\lambda} \cdot \lambda^x}{x!} = P(X = x) \\ &= 1 - e^{-\lambda y} \\ &\quad \uparrow \\ &\text{cdf for } \text{Exp}(\lambda) \end{aligned}$$



**NOTE:** This is useful since Exponential is easier to calculate than Poisson

### Example

It is known that accidents in a factory follow a Poisson process, with an average rate of 1 accident per week. What is the probability that the next accident at the factory will occur within the next two weeks?

$$\lambda = 1 \text{ per week}$$

$$Y \sim \text{Exp}(\lambda = 1)$$

$$P(y \leq 2) = 1 - e^{-2 \cdot 1}$$

$$= 1 - 0.1353 = 0.8647$$

## 71. Memoryless Property

Suppose  $X \sim \text{Exp}(\lambda)$ . Then for any  $a, b \geq 0$

$$P(X \geq a + b | X \geq b) = P(X \geq a)$$

This means that the probability of a person needing to wait at least  $a$  minutes more if they've already waited  $b$  minutes, is the same as the probability of a newly-arrived person needing to wait  $a$  minutes

### Example

Suppose I've already been waiting to use the bank machine for six minutes. What is the probability my total waiting time will be at least 10 minutes?

$$\begin{aligned} &P(X \geq a + b | X \geq b) \\ &= \frac{P(X \geq a + b \cap X \geq b)}{P(X \geq b)} \\ &= \frac{P(X \geq a + b)}{P(X \geq b)} \\ &= \frac{e^{-\lambda(a+b)}}{e^{-\lambda b}} = e^{-\lambda a} \\ &= P(X \geq a) \end{aligned}$$

For this example

$X$  = waiting time

$$X \sim \text{Exp}(\lambda = \frac{1}{6})$$

$$\begin{aligned} &P(X \geq 10 | X \geq 6) \\ &= P(X \geq 4 + 6 | X \geq 6) \\ &= P(X \geq 4) \\ &= P(\geq 4) \\ &= e^{-4/6} = 0.5134 \end{aligned}$$

## 72. Sets 18 and 19 Joint Distribution

Let  $X$  and  $Y$  be discrete random variables defined on some sample space  $S$ . The **joint probability function**  $f(x, y)$  is defined as:

$$(x, y) \leftarrow \text{ordered pairs}$$

**NOTE:** This is used for more than one random variable

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

let  $A$  be any set of  $(x, y)$  pairs, then:  $A$  is an event

$$P((X, Y) \in A) = \sum_{(x, y) \in A} f(x, y)$$

### Example

Suppose that we consider the manufacture of wind turbines. Before the turbines are shipped, they are checked for flaws and repaired (if necessary).

Let  $X$  denote the number of manufacturing flaws in a randomly selected turbine. Let  $Y$  denote the maximum number of days it takes to repair the flaws.

The following the **joint probability table** for the probability function  $f(x, y)$ :

$f(x, y)$		$y$		
		0	1	2
$x$	0	0.512	0.000	<b>0.000</b>
	1	0.000	0.102	<b>0.008</b>
	2	<b>0.000</b>	<b>0.175</b>	<b>0.089</b>
	3	<b>0.000</b>	<b>0.015</b>	<b>0.099</b>

**NOTE:**  $\sum_{all (x,y)} \sum F(x, y) = 1$  because if you add up all the probabilities in the table you should always get 1

$$x = 0, 1, 2, 3$$

$$y = 0, 1, 2$$

### Example

Based on the previous example calculate  $P(X \geq 2 \cap Y = 2)$

In english this is  $P(\text{there are at least 2 flaws and it will take exactly 2 days to repair})$

**NOTE:** in the table we bold all of the important data then add the intersection

$$P(X \geq 2 \cap Y = 2) = 0.089 + 0.099$$

$$= 0.188$$

### 73. Marginal Probability Function

The marginal probability function of  $X$  and  $Y$ , denoted  $f_X(x)$  and  $f_Y(y)$  are:

$$f_X(x) = \sum_y f(x, y), \quad f_Y(y) = \sum_x f(x, y)$$

#### Example

Find  $f_X(x)$  and  $f_Y(y)$  for the previous example.

$f(x, y)$		$y$			
		0	1	2	
$x$	0	0.512	0.000	<b>0.000</b>	$(y = 0) + (y = 1) + (y = 2) + (y = 3)$
	1	0.000	0.102	<b>0.008</b>	$(y = 0) + (y = 1) + (y = 2) + (y = 3)$
	2	<b>0.000</b>	<b>0.175</b>	<b>0.089</b>	$(y = 0) + (y = 1) + (y = 2) + (y = 3)$
	3	<b>0.000</b>	<b>0.015</b>	<b>0.099</b>	$(y = 0) + (y = 1) + (y = 2) + (y = 3)$
		$(x = 0)$ $+(x = 1)$ $+(x = 2)$	$(x = 0)$ $+(x = 1)$ $+(x = 2)$	$(x = 0)$ $+(x = 1)$ $+(x = 2)$	

$f(x, y)$		$y$			
		0	1	2	
$x$	0	0.512	0.000	<b>0.000</b>	0.512
	1	0.000	0.102	<b>0.008</b>	0.110
	2	<b>0.000</b>	<b>0.175</b>	<b>0.089</b>	0.264
	3	<b>0.000</b>	<b>0.015</b>	<b>0.099</b>	0.114
		0.512	0.292	0.196	

$x$	0	1	2	3
$f_X(x)$	0.512	0.110	0.264	0.114

$$E(X) = \mu_x = 0(0.512) + 1(0.110) + 2(0.264) + 3(0.114) = 0.98$$

$y$	0	1	2
$f_Y(y)$	0.512	0.292	0.196

$$E(Y) = \mu_y = 0(0.512) + 1(0.292) + 2(0.196) = 0.684$$

If  $X$  and  $Y$  are independent random variables, then  $f(x, y) = f_X(x)f_Y(y)$  for every  $(x, y)$  pair.

We can show without too much difficulty that  $X$  and  $Y$  are not independent in our turbine example.

We can extend our definitions quite naturally to any sequence  $X_1, X_2, \dots, X_n$  of random variables.

$$f(0, 0) = f_X(0) \cdot f_Y(0) ?$$

$$0.512 \neq 0.512 \cdot 0.512$$

$\therefore X, Y$  is not independent

### Example

Suppose that in a copy shop, three photocopiers work. Let  $X_i$  be the number of paper jams that copier  $i$  experiences in a day, where  $i = 1, 2, 3$ . Suppose that  $X_1, X_2, X_3$  are independent,  $X_1 \sim \text{Poisson}(\lambda = 4)$ ,  $X_2 \sim \text{Poisson}(\lambda = 3)$ ,  $X_3 \sim \text{Poisson}(\lambda = 10)$ .

Find the joint pmf  $f(x_1, x_2, x_3)$ .

$$f_{X_1}(x_1) = \frac{e^{-4} 4^{x_1}}{x_1!}$$

$$f_{X_2}(x_2) = \frac{e^{-3} 3^{x_2}}{x_2!}$$

$$f_{X_3}(x_3) = \frac{e^{-10} 10^{x_3}}{x_3!}$$

$$f(x_1, x_2, x_3) = \frac{e^{-10} 10^{x_3}}{x_3!} \cdot \frac{e^{-3} 3^{x_2}}{x_2!} \cdot \frac{e^{-4} 4^{x_1}}{x_1!}$$

$$= \frac{e^{-17} \cdot 4^{x_1} \cdot 3^{x_2} \cdot 10^{x_3}}{x_1! x_2! x_3!}$$

**NOTE:** remember that  $x_1, x_2$ , and  $x_3$  can be any thing from 0 to  $\infty$ .

## General Review 1

*Gavin Jaeger-Freeborn*

### 74. Probability distribution function ( not the same as pdf) for random variables

pmf	probability mass function	$f(x)$	$P(X = x)$ discrete
pdf	probability density function	$f(x)$	$P(X = x)$ continuous
cdf	cumulative distribution function	$F(x)$	$P(X \leq x)$ or $\int_{-\infty}^x f(x)dx = F(x)$

### 75. Variance and standard deviation

$$V(X) = \sigma^2_X = E((X - \mu_X)^2) = E(X^2) - \mu^2_X$$

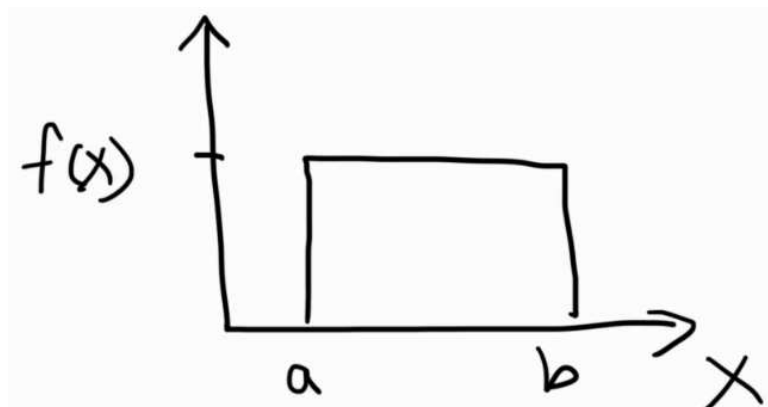
### 76. Uniform Distribution

Uniform distribution is a rectangle where  $a$  is the minimum and  $b$  is the maximum

$$X \sim U(a, b)$$

pmf (probability mass function)	$f(x) = \frac{1}{b-a}$
mean	$\mu = \frac{a+b}{2}$
standard deviation	$\sigma = \frac{b-a}{\sqrt{12}}$

**diagram of uniform distribution**



## 77. Poisson Distribution

If arrivals occur at random in time (or space) at the average rate of  $\delta$  per unit time (or space), and  $X$  = total number of arrivals that occur in a time (or space) window of size  $t$ , then the distribution of  $X$  is:  $Poisson(\lambda = \delta t)$

If  $X \sim Poisson(\lambda)$ , then

pmf (probability mass function)	$f(x) = \frac{\lambda^x}{x!} e^{-\lambda}$
mean	$\mu = \lambda$
standard deviation	$\sigma = \sqrt{\lambda}$

**NOTE:** for Poisson  $\mu = \lambda$  and  $\sigma^2 = \lambda$

**NOTE:** to use the table the probability must be of this form  $P(X \leq x)$

## 78. Binomial Distribution

if  $X$  = total number of successes out of  $n$  independent trials where  $P(\text{success}) = p$  on every trial, then the distribution of  $X$  is  $Binomial(n, p)$

If  $X \sim Binomial(n, p)$ , then

pmf (probability mass function)	$f(x) = \binom{n}{x} p^x q^{n-x}$
mean	$\mu = np$
standard deviation	$\sigma = \sqrt{npq}$

**NOTE:**  $\binom{n}{r} = \frac{n!}{r!(n-r)!} = nCr$

**NOTE:**  $q$  is just the probability of failure aka  $q = 1 - p$

**NOTE:** to use the table the probability must be of this form  $P(X \leq x)$



## 79. Normal Distribution

Every linear combination of independent normally distributed rev's is normally distributed

If  $X \sim N(\mu, \sigma)$ , then the distribution of

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

is: Standard Normal

**NOTE:**  $P(Z > z) = P(Z < -z)$  for using the table we must have  $P(Z < z)$  from

## 80. Statistics and Distribution

let  $X = X_1 + X_2 + \dots + X_n$  be random variables. Some common statistics include:

sample mean	$\bar{X} = \frac{X_1 + X_2 + \dots + X_n}{n}$
sample sum	$T = X_1 + X_2 + \dots + X_n$
sample variance	$S^2 = \frac{\sum (X_i - \bar{X})^2}{n - 1}$
sample median	$\tilde{X} = \text{median}(X_1, X_2, \dots, X_n)$

**NOTE:** the value is  $\bar{x}, t, x^2, \tilde{x}$  while  $\bar{X}, T, S^2, \tilde{X}$  is a random variable.

**NOTE:** we want  $X_i \sim \text{Some Distribution}$  and have the same  $\mu, \sigma^2, \sigma$

### 81. Set 21 The Importance of Normal Distribution

if  $X_1, X_2, \dots, X_n$  are **independent identical** ( meaning they have the same distribution )  
**random variables**

Then:

1. The sample mean ,  $\bar{X}$  , has mean  $\mu$  and standard deviation  $\sigma/\sqrt{n}$
2. The sample sum ,  $T$  , has mean  $n\mu$  and standard deviation  $\sigma\sqrt{n}$

All linear combinations of independent normal random variables are normally distributed.

If  $X_1, X_2, \dots, X_n$  are all iid (independent identical), and normally distributed then:

1.  $\bar{X}$ , has normal distribution with mean  $\mu$  and standard deviation  $\frac{\sigma}{\sqrt{n}}$

$$\bar{X} \sim N\left(\mu, \frac{\sigma}{\sqrt{n}}\right)$$

2.  $T$ , has normal distribution with mean  $n\mu$  and standard deviation  $\sigma\sqrt{n}$

$$T \sim N(n\mu, \sigma\sqrt{n})$$

**NOTE:** This is for any sample size

**Example**

Suppose it is known that the levels of fluid in soda bottles is normally distributed, with a mean of 355 mL, and standard deviation of 2 mL. Let  $X_1, X_2, X_3, X_4$  denote liquid content of four randomly selected bottles.

Find the probability that the average liquid content will be less than 356 mL.

$$\mu = 355ml$$

$$\sigma^2 = 2ml$$

$$n = 4$$

find

$$\bar{X} \sim N(355, 2/\sqrt{2})$$

$$P(\bar{X} < 356ml)$$

$$= P(Z < \frac{X - \mu}{\sigma})$$

$$= P(Z < \frac{356 - 355}{1})$$

$$= P(Z < 1)$$

FROM TABLE

$$0.8413$$

## 82. Central Limit Theorem

Let  $X_1, X_2, \dots, X_n$  be iid random variables, each with mean  $\mu$  and standard deviation  $\sigma$ . Provided that  $n \geq 30$  (rule of thumb).

1.  $\bar{X}$ , has *approximately* normal distribution with mean  $\mu$  and standard deviation  $\frac{\sigma}{\sqrt{n}}$ .

$$\bar{X} \approx N\left(\mu, \frac{\sigma}{\sqrt{n}}\right)$$

2.  $T$ , has *approximately* normal distribution with mean  $n\mu$  and standard deviation  $\sigma\sqrt{n}$

$$T \approx N(n\mu, \sigma\sqrt{n})$$

WE MUST KNOW WHAT  $\mu, \sigma$ , AND  $n$  ARE

**NOTE:** the larger the sample size the closer  $\bar{X}$  and  $T$  will be to a normal distribution.

### Example

The number of bacteria per mL sample of water has a Poisson distribution, with an average of 50 bacteria per sample. Suppose that 100 samples are tested. What is the probability that the average number of bacteria per sample is at least 52?

$$\lambda = 50 \text{ per sample}$$
$$n = 100$$

Using Central Limit Theorem since  $n > 30$

**NOTE:** when using  $\bar{X}$  you don't want to use Poisson

For Poisson Distribution recall that  $\mu = \lambda = 50$  and  $\sigma^2 = \lambda = 50 \therefore \sigma = \sqrt{\lambda}$

$$\bar{X} \approx N\left(\mu, \frac{\sigma}{\sqrt{n}}\right)$$
$$\bar{X} \approx N\left(50, \frac{\sqrt{50}}{\sqrt{100}}\right)$$
$$\bar{X} \approx N(50, 0.7071)$$
$$P(\bar{X} \geq 52) \approx P\left(Z > \frac{52 - 50}{0.7071}\right)$$
$$\approx P\left(Z > \frac{2}{0.7071}\right)$$
$$\approx P(Z > 2.828)$$
$$\approx P(Z > 2.83)$$

Since standard normal is symmetric about 0

$$\approx P(Z < -2.83) = 0.0023$$

FROM TABLE

$$\approx 0.0023$$

### Example

In a particular lake, the amount of pollutant in a 1 L sample is has a mean of 6 mg with a standard deviation of 1 mg. Suppose we take 50 randomly selected samples, each of 1 L of lake water. What is the probability that the total amount of pollutant will be between 295 mg and 305 mg?

$$n = 50$$

$$\mu = 6mg$$

$$\sigma = 1mg$$

$$P(295 < T < 305)$$

$$P(T < 305) - P(T < 295)$$

Applying CLT we get

$$T \approx N(n\mu, \sigma\sqrt{n})$$

$$T \approx N(50 \cdot 6, 1\sqrt{50})$$

$$T \approx N(300, \sqrt{50})$$

$$\approx P(Z < \frac{305 - 300}{\sqrt{50}}) - P(Z < \frac{295 - 300}{\sqrt{50}})$$

$$\approx P(Z < 0.7071) - P(Z < -0.7071)$$

$$\approx 0.7611 - 0.2389$$

$$0.5222$$

in R

```
> pnorm ( 305, 300, sqrt ( 50 ) ) - pnorm( 295 , 300, sqrt (50 ) )  
[1] 0.5204999
```

### Example

Pheasants in a particular region were found to have an appreciable mercury contamination. The mercury level in parts per million for these birds is normally distributed with mean 0.25 and standard deviation 0.08.

If I select 4 pheasants at random, what is the probability that the mean mercury level will be greater than 0.3 ppm?

$$n = 4$$

Since  $n = 4 < 30$  we cannot use CLT

$$\mu = 0.25$$

$$\sigma = 0.08$$

$$X_1, \dots, X_4 \sim N(0.25, 0.08)$$

$$P(\bar{X} > 0.3)$$

This is no longer an approximation since we are using a normal distribution

$$\bar{X} \sim N(0.25, \frac{0.08}{\sqrt{4}} = 0.04)$$

$$P(\bar{X} > 0.3)$$

$$P(Z > \frac{0.3 - 0.25}{0.04}) = P(Z > 1.25)$$

$$= P(Z < -1.25)$$

$$= 0.1056$$

in R

```
> pnorm( -1.25 )  
[1] 0.1056498
```

### Example

Suppose again that we select 4 pheasants at random. What is the probability that all of the pheasants will have a mercury level which is less than 0.2?

$y = \#$  of pheasants having mercury levels  $< 0.2$  ppm

$\therefore$  success = (having mercury levels  $< 0.2$  ppm)

$$\begin{aligned} P &= P(X < 0.2) \\ &= P\left(X < \frac{0.2 - 0.25}{0.08}\right) \\ Z &= \frac{x - \mu}{\sigma} \\ &= P(Z < -0.635) \\ &= 0.2643 \end{aligned}$$

Now apply this to a binomial distribution

$$y \sim \text{Bin}(4, 0.2643)$$

$$\begin{aligned} P(y = 4) &= \binom{4}{4} (0.2643)^4 (1 - 0.2643)^0 \\ &= 0.2643^4 \\ &= 0.0049 \end{aligned}$$

## STATS 260 Class 19

*Gavin Jaeger-Freeborn*

### 83. 22 and 23 Estimation

#### 84. Point estimate

Single number that serves as an estimate for the true value of the parameter.

The estimate comes from a statistic called the **estimator**



It is calculated from the sample data

### Example

We might not know the population mean  $\mu$  (also called the **true mean**) of some population. We take a sample of  $n$  observations, and then find the value of the sample mean  $\bar{X}$ , which is our estimator. The value we find,  $\bar{x}$ , is our point estimate for the true mean  $\mu$ .

Here we use  $\bar{X}$  to estimate  $\mu$

### 85. Properties of a good estimator

- The estimator should be **unbiased**: this means that the estimator does not tend to over-estimate, and does not tend to underestimate.

If  $\hat{\theta}$  is as unbiased estimator for some parameter  $\theta$  then:

$$E(\hat{\theta}) = \theta$$

In other words: the long- run average value of the estimate will be parameter which we are estimating.

- The estimator should be **consistent**; this means the value of the estimate will approach the true value of the parameter as  $n \rightarrow \infty$  where  $n$  is the sample size.

$$V(\hat{\theta}) \rightarrow 0$$

$$n \rightarrow \infty$$

### Example

Suppose we have a population with an unknown true mean  $\mu$ . We select  $n$  members of the population as our sample, and wish to use  $\bar{X}$ , the sample mean, as the estimator which will give us the point estimate of our true mean,

is  $\bar{X}$  unbiased for  $\mu$ ? Is  $\bar{X}$  as consistent estimator?

$$E(\bar{X}) = \mu \text{ unbiased}$$

$$V(\bar{X}) = \frac{\sigma^2}{n}, \quad n \rightarrow \infty \quad V(\bar{X}) \rightarrow 0$$

### Example

Suppose a population has unknown mean  $\mu$  and variance  $\sigma^2$ . We take  $n$  observations,  $X_1, \dots, X_n$  from this population.

We can show that  $S^2 = \frac{\sum (X_i - \bar{X})^2}{n-1}$  the sample variance, is an unbiased estimator for  $\sigma^2$  (i.e. we can show  $E(S^2) = \sigma^2$ ).

$S^2$  is unbiased for  $\sigma^2$

$E(S) \neq \sigma$  Therefore  $S$  is biased estimation for  $\sigma$

If we have a choice between estimators, we would prefer the estimator with the grater **efficiency**. The less variability an estimator has in estimating the parameter, the more efficient it is.

If two estimators are both unbiased, then the estimator with the smaller variance is the more efficient estimator.

### features of a good estimator

Unbiased	Does not over estimate aka $E(\hat{\theta}) = \theta$
Consistent	$\lim_{n \rightarrow \infty} V(\hat{\theta}) = 0$ as $n \rightarrow \infty$
Efficient	Smaller variance or smaller standard deviation

now use

$\bar{X}$  to estimate  $\mu$

$S$  to estimate  $\sigma$

$S^2$  to estimate  $\sigma^2$

### 85.1. interval estimate

An estimate between and interval e.g. Between 10 and 15 or (10, 15)

A **pivotal quantity** is a function of observations with a distribution that does not depend on the value of any unknown parameters.

#### Example

Suppose we have a random sample of  $n$ , either from population with mean  $\mu$  and standard deviation  $\sigma$ . Also, suppose the population is normal, or that  $n$  is large (or both).

Then  $\frac{\bar{X} - \mu}{\sigma/\sqrt{n}}$  is a pivotal quantity

does not depend on value of  $\mu$

$$\frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \sim N(0, 1)$$

**NOTE:**  $N(0,1)$  does not depend on  $\mu$  even though  $\frac{\bar{X} - \mu}{\sigma/\sqrt{n}}$  does depend on  $\mu$

By the Central Limit Theorem, we know that  $\bar{X}$  is normal with mean  $\mu$  and standard deviation  $\sigma/\sqrt{n}$ . so the expression above with standard normal distribution, regardless of what the value of  $\mu, \sigma$  are.

(i)  $\frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \sim N(0, 1)$  if  $X_i$  is normal ( population)  $n$  doesn't matter

(i)  $\frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \approx N(0, 1)$  if  $n \geq 30$  (rule of thumb for CLT)

Computation is the same if it is approximate.

What happens to  $\frac{\bar{X} - \mu}{\sigma/\sqrt{n}}$  if  $X_i$  is not normal and  $n < 30$

↑ Trick question we ignore this in this class

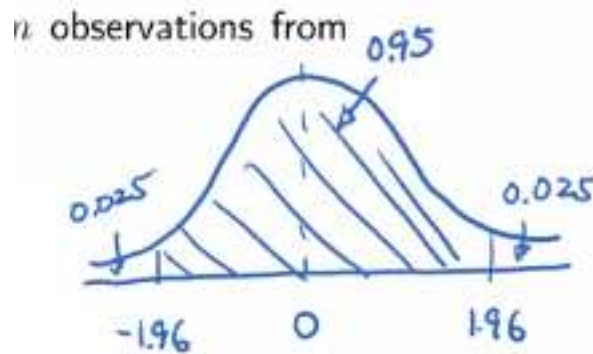
## 86. Random intervals

A random interval is an interval for real numbers whose endpoints are random variables.

We can use **pivotal quantities** to construct a **random interval** for one of the parameters.

### Example

Suppose we have a random sample of  $n$  observations from some normal population. We can find that



$$P\left(-1.96 \leq \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \leq 1.96\right) = 0.95$$

This can be re written as:

$$P\left(\bar{X} - 1.96 \frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{X} + 1.96 \frac{\sigma}{\sqrt{n}}\right) = 0.95$$

Written as a random interval this is  $\left(\bar{X} - 1.96 \frac{\sigma}{\sqrt{n}}, \bar{X} + 1.96 \frac{\sigma}{\sqrt{n}}\right) = 0.95$

### 86.1. Example from video

the average test scores in a physics class is normally distributed with a standard deviation of 5.4. 50 scores with a sample mean of 79 were selected at random.

(A) Find a 95% confidence interval for the population mean test score.

Given

$$\sigma = 5.4$$

$$n = 50$$

$$\bar{X} = 79$$

$$CL = 0.95\%$$

*CL is for confidence level*

The equation to determine the confidence interval is

$$\bar{X} - z_{\alpha} \frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{X} + z_{\alpha} \frac{\sigma}{\sqrt{n}}$$

and the actual interval is

$$CI \rightarrow (\bar{X} - \text{error bound for the mean}, \bar{X} + \text{error bound for the mean})$$

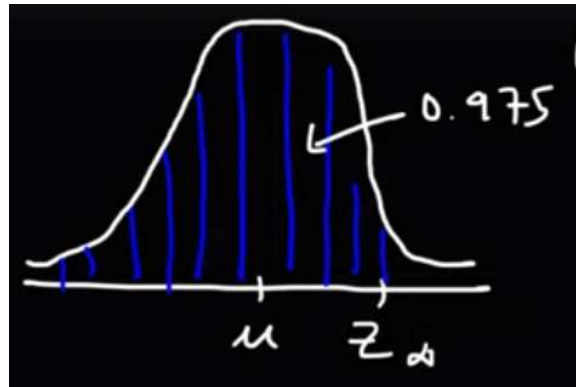
aka

$$CI \rightarrow \bar{X} \pm z_{\alpha} \frac{\sigma}{\sqrt{n}}$$

How do I find  $z_\alpha$

$z_\alpha$  is some point right of  $\mu$ .

as seen here



Here is how you calculate the area to the left

$$A_L = \frac{CL + 1}{2}$$

$$A_L = \frac{0.95 + 1}{2} = 0.975$$

Now simply find where in the positive z score table the probability = 0.975

In this case it is at 1.96

$\therefore$  the z score is 1.96 and  $z_\alpha = 1.96$

Next we need to calculate  $\sigma_{\bar{X}}$

$$\text{This is just } \sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}} = \frac{5.4}{\sqrt{50}}$$

$$0.76368$$

$$\begin{aligned} 95\%CI &\rightarrow 79 \pm 1.96(0.76368) \\ &\rightarrow 79 \pm 1.4968 \end{aligned}$$

$\therefore$  the confidence interval is (77.5032, 80.4968)

This means that we are 95% confident that the mean for the population falls somewhere between 77.5032 and 80.4968

(B) What is the value of the margin of error?

We need to find **error bound for the mean** aka EBM

$$\begin{aligned}\text{This is just } EBM &= Z_{\alpha} \frac{\sigma}{\sqrt{n}} \\ &= 1.96 \frac{(5.4)}{\sqrt{50}} = 1.4968\end{aligned}$$

### 86.2. Another way to approach this

**For a large-sample size scenario ( $n \geq 40$ ), the following is an approximate  $100(1 - \alpha)\%$  confidence interval:**

$$\left( \bar{x} - z_{\alpha/2}, \bar{x} + z_{\alpha/2} \frac{s}{\sqrt{n}} \right)$$

$Z_{\frac{\alpha}{2}}$  is known as the critical value for the confidence interval.

Here is how to find  $Z_{\frac{\alpha}{2}}$

If the confidence level is 95 then we know that  $95 = 100(1 - \alpha)\%$

$$\alpha = 0.05$$

$$\frac{\alpha}{2} = 0.025$$

$Z_{\frac{\alpha}{2}}$  is the 97.5th percentile

We get this using the formula

$$1 - \frac{\alpha}{2}$$

We then just need to find what Z value on the table has the probability of 0.975

$$\boxed{Z_{0.025} = 1.96}$$

### In summary

$$\bar{X} \pm \frac{estimate \pm (criticalvalue)(standarderror)}{Z_{\frac{\alpha}{2}}} \frac{\sigma}{\sqrt{n}}$$

### Common Critical Values:

% Confidence	Critical Value
90	$z_{\alpha/2} = z_{0.05} = 1.645$
95	$z_{\alpha/2} = z_{0.025} = 1.96$
99	$z_{\alpha/2} = z_{0.005} = 2.575$

**Note:** While these confidence levels are common, you should also be able to find the critical value for *any* confidence level.

### Interpretation

We should not interpret this as meaning that there is a 95% chance that the true mean height is between 1.66864 and 1.73136 meters.

### 86.3. Margin of error (d)

$$d = (\text{critical value}) (\text{standard error})$$

$$d = z_{\alpha/2} \frac{\sigma}{\sqrt{n}}$$

rearranging this we get

$$n = \left( \frac{Z_{\frac{\alpha}{2}} \cdot \sigma}{d} \right)^2$$

**NOTE:** Width of confidence interval, is the distance between the upper and lower confidence limits. This is  $2D$  is the width of a CI.



#### 86.4. Scaling of confidence level

CL $\uparrow$	$Z_{\frac{\alpha}{2}}$ $\uparrow$	$d$ $\uparrow$
	$\sigma$ $\uparrow$	$d$ $\uparrow$
	$n$ $\uparrow$	$d$ $\uparrow$

**NOTE:** we want a smaller  $d$  since this means we have more confidence

#### Example

Example: Suppose the lifetime of certain type of lightbulb is normally distributed and has a standard deviation of  $\sigma = 200$  hours. How many samples do we need to be create a 95% confidence interval for  $\mu$ , the mean lifespan, with a margin of error of 10 hours?

$$d = 10$$

$$Z_{\frac{\alpha}{2}} = 1.96$$

$$\sigma = 200$$

$$n = \left( \frac{1.96 \cdot 200}{10} \right)^2$$

$$= 1536.64 \Rightarrow 1.537$$

**NOTE:** Since  $n$  aka sample size must be a whole number we always round up.

### Example

How many observations do we need to create a 95% confidence interval for  $\mu$  with a width of 40 hours?

$$d = 20$$

$$n = \left( \frac{1.96 \cdot 200}{20} \right)^2 = 384.16$$

remember to round up

$$= 385$$

wider confidence interval requires smaller n

### 87. Estimated Standard Error

For a large-sample size scenario ( $n \geq 40$ ), the following is an approximate  $100(1 - \alpha)\%$  confidence interval:

We can replace  $\sigma$  with  $s$  to estimate

$$\left( \bar{x} - Z_{\frac{\alpha}{2}} \frac{s}{\sqrt{n}}, \bar{x} + Z_{\frac{\alpha}{2}} \frac{s}{\sqrt{n}} \right)$$

We call  $s/\sqrt{n}$  the **estimated standard error**

**Example**

At a particular location, fifty daily measurements of wind speed (in m/s) are made. It is found that  $\bar{x} = 15.9$  m/s and  $s = 7.7$  m/s. Find a 98% confidence interval for  $\mu$ , the average daily wind speed. Assume that the measurements constitute a random sample from the population of all wind speed measurements.

$$\alpha = 1 - 98 = 0.02$$

$$\frac{\alpha}{2} = 0.01$$

$$\text{reverse lookup for probability} = 1 - \frac{\alpha}{2} = 0.99$$

$$Z_{0.01} = 2.33$$

$$\bar{x} \pm Z_{\frac{\alpha}{2}} \frac{s}{\sqrt{n}}$$

$$15.9 \pm 2.33 \frac{7.7}{\sqrt{50}}$$

$$15.9 \pm 2.5$$

$$(13.363, 18.437)$$

**88. t-Distribution**

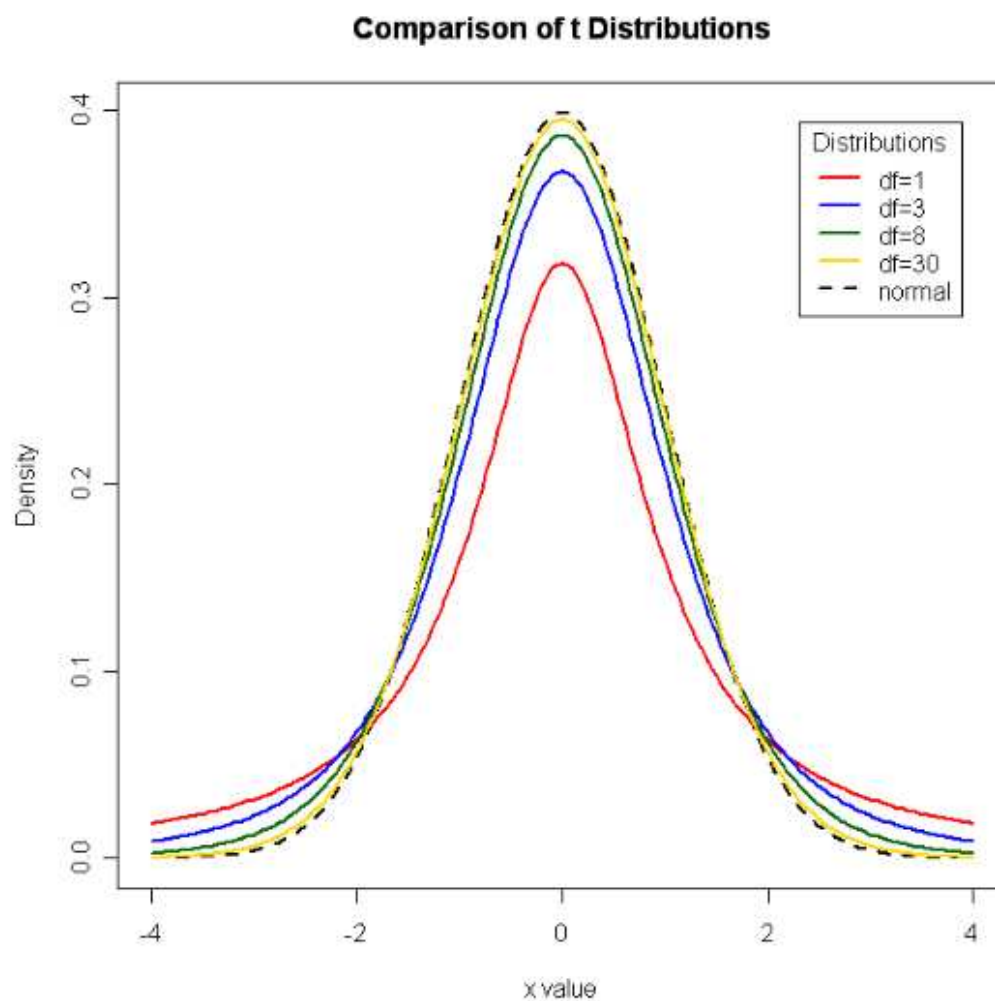
If the sample is  $n < 40$  then we use a t-distribution, with  $n - 1$  degrees of freedom.

$$\frac{\bar{X} - \mu}{S/\sqrt{n}} \sim t_{n-1}$$

### 88.1. Properties of the t distribution:

1. The t distribution is continuous, and defined on  $(-\infty, \infty)$ .
2. The t distribution is **symmetric**, **bell-shaped**, and **centered** at zero.
3. The number of degrees of freedom affect the shape of the distribution; as the number of degrees of freedom increases, the distribution becomes more peaked, and the tails become thinner.
4. When the number of degrees of freedom is large (30 or more), the t-distribution is approximately a standard normal distribution.

**NOTE:** To denote a t-distribution with k degrees of freedom, we write



## 88.2. Confidence Interval for Population Mean

for sample sizes  $< 40$  we use

$$\bar{x} \pm t_{n-1, \alpha/2} \cdot \frac{s}{\sqrt{n}}$$

$t_{n-1, \alpha/2}$  acts as the critical value for a t-distribution with  $n-1$  degrees

**NOTE:** sample sizes  $< 40$  we and sigma must not be known

### Example

The following data is collected on the mass (in grams) of adult white mice.

14.6, 13.2, 19.5, 10.1, 8.8, 15.5, 16.1

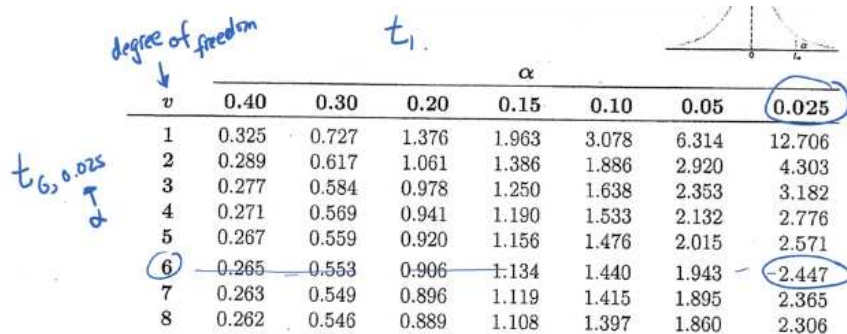
Assuming that the weights of mice are normally distributed, find a 95% confidence interval for  $\mu$ , the mean weight of adult white mice.

$$n = 7 < 40, \alpha = 1 - 0.95 = 0.05, \alpha/2 = 0.025, s = 3.655003$$

$$\mu = 13.97143$$

$$t_{7-1, 0.025} = t_{6, 0.025}$$

Using the table in Appendix D we use  $v$  = degrees of freedom and  $\alpha$  as  $\alpha$



$v$	$\alpha$						
	0.40	0.30	0.20	0.15	0.10	0.05	0.025
1	0.325	0.727	1.376	1.963	3.078	6.314	12.706
2	0.289	0.617	1.061	1.386	1.886	2.920	4.303
3	0.277	0.584	0.978	1.250	1.638	2.353	3.182
4	0.271	0.569	0.941	1.190	1.533	2.132	2.776
5	0.267	0.559	0.920	1.156	1.476	2.015	2.571
6	0.265	0.553	0.906	1.134	1.440	1.943	2.447
7	0.263	0.549	0.896	1.119	1.415	1.895	2.365
8	0.262	0.546	0.889	1.108	1.397	1.860	2.306

$$t_{6, 0.025} = 2.447$$

$$13.971 \pm 2.447 \cdot \frac{3.655}{\sqrt{7}}$$

$$(10.591, 17.351)$$

**in R**

```
> tmp=c(14.6, 13.2, 19.5, 10.1, 8.8, 15.5, 16.1)
> t.test(tmp, conf.level = .95)
```

One Sample t-test

```
data: tmp
t = 10.114, df = 6, p-value = 5.431e-05
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
 10.59111 17.35174
sample estimates:
mean of x
 13.97143
```

## STATS 260 Class 20

*Gavin Jaeger-Freeborn*

### 89. Review

small sample cs  $n < 40$

$$\bar{x} \pm t_{n-1, \frac{\alpha}{2}} \cdot \frac{s}{\sqrt{n}}$$

**NOTE:** d is still everything to the right of  $\pm$  therefor

$$d = t_{n-1, \frac{\alpha}{2}} \cdot \frac{s}{\sqrt{n}}$$

and

$$n = \left( \frac{z_{\frac{\alpha}{2}} s}{d} \right)^2$$

**NOTE:** the resulting n must be  $n < 40$  also round up

### 90. set 24

## 91. Population Proportion Estimation

if we have a **binomial distribution** with  $n < 40$ . we can estimate the value of the true proportion of success

if  $n$  is the number of observations and  $x$  is the number of successes, then  $\hat{p} = x/n$

$p$  is the **sample proportion**

$$\frac{\hat{p} - p}{\sqrt{\frac{p(1-p)}{n}}} \approx N(0, 1)$$

this can be rewritten as

$$\frac{\hat{p} - p}{\frac{\sqrt{p(1-p)}}{\sqrt{n}}} \approx N(0, 1)$$

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

**NOTE:**  $X$  is the total number of successes where  $X_1, X_2, \dots, X_n \sim \text{Bin}(n = 1, p)$  each one is either a success or a failure.

$$E(x_i) = P = \mu$$

$$V(X_i) = E(X_i^2) - \mu^2$$

$$= p(1 - p)$$

$$\sigma_{x_i} = \sqrt{p(1 - p)}$$

$$\frac{\hat{p} - p}{\frac{\sqrt{p(1-p)}}{\sqrt{n}}} \approx N(0, 1) = \frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}}}$$

Because of central limit Theorem

**Standard error** of  $\hat{p}$  is just  $\sqrt{p(1-p)/n}$  since the value of  $p$  is unknown

We cannot use the standard error in our confidence interval.

Instead, we use the **estimated standard error**

**estimated standard error** of  $\hat{p}$  is

$$\sqrt{\hat{p}(1-\hat{p})/n}$$

Using the same formula as before

$$(\text{estimate}) \pm (\text{critical value}) \cdot (\text{estimated standard error})$$

We get

$$\hat{p} \pm z_{\alpha/2} \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

**NOTE:** there must be at least 5 success and 5 failures



**Example**

A sample of 1380 randomly selected books produced by a publishing company finds that 25 have bookbinding errors. Find a 95% confidence interval for p, the proportion of books with bookbinding errors.

$$n = 1380$$

$$x = 25$$

$$cl = 95\%$$

$$\alpha = 1 - 0.95 = 0.05$$

$$\frac{\alpha}{2} = 0.025$$

Looking up z value for  $p = 1 - 0.025$

$$z_{0.025} = 1.96$$

$$\hat{p} = \frac{x}{n} = \frac{25}{1380}$$

$$\frac{25}{1380} \pm 1.96 \sqrt{\frac{\frac{25}{1380} \left(1 - \frac{25}{1380}\right)}{1389}}$$

$$0.018116 \pm 0.007036831$$

$$(0.011079, 0.02515)$$

We are 95% confident that the true proportion of books with errors is between 1.11% and 2.52%

If we are given  $d$  ( margin of error) we can estimate the sample size using

$$n = \frac{(z_{\alpha/2})^2 \hat{p}(1 - \hat{p})}{d^2}$$

### Option 1

Sometimes we use a previous study to estimate for  $\hat{p}$

### Option 2

$$\text{use } \hat{p} = \frac{1}{2} \text{ ( based on calculus )}$$

This gives you

$$n = \frac{(z_{\alpha/2})^2}{4d^2} \Leftarrow \hat{p}(1 - \hat{p}) = \frac{1}{4}$$

### Example

In an earlier study, it was found that 1.4% of all microchips made by a particular manufacturer were defective. Using this as a pilot study, estimate the sample size needed to create a 99% confidence interval for  $p$ , the true proportion of defective microchips, with a margin of error of 0.005.

$$n = \frac{(z_{\alpha/2})^2 \hat{p}(1 - \hat{p})}{d^2}$$

$$\frac{2.575^2(0.014)(1 - 0.014)}{0.005^2}$$

$$n = 3661.166$$

Remember to round up

$$n = 3662$$

**Example**

We wish to carry out a telephone survey to estimate  $p$ , the proportion of island residents who want a bridge to the mainland. How many people must we call in order to estimate  $p$  with 98% confidence, to within 0.01?

Margin of error

$$d = 0.01$$

$$cl = 98\%$$

$$\alpha = 1 - .98, \frac{\alpha}{2} = 0.01$$

$$z_{0.01} = 2.326348$$

$$\hat{p} = 1/2$$

$$n = \frac{(z_{\alpha/2})^2}{4d^2}$$

$$n = \frac{(2.326348)^2}{4(0.01)^2}$$

$$n = \frac{(5.4289)}{(4e-04)}$$

$$n = 13572.25$$

$$n = 13573$$

**STATS 260 Class 21**

*Gavin Jaeger-Freeborn*

**92. Sets 25 to 27 Hypothesis Testing****93. Single Sample Hypothesis**

### Example

A factory has a machine that dispenses 80ml of fluid in a bottle, an employee believes the average amount of fluid is not 80ml. Using 40 samples, he measures the average amount dispensed by the machine to be 78ml with a standard deviation of 2.5.

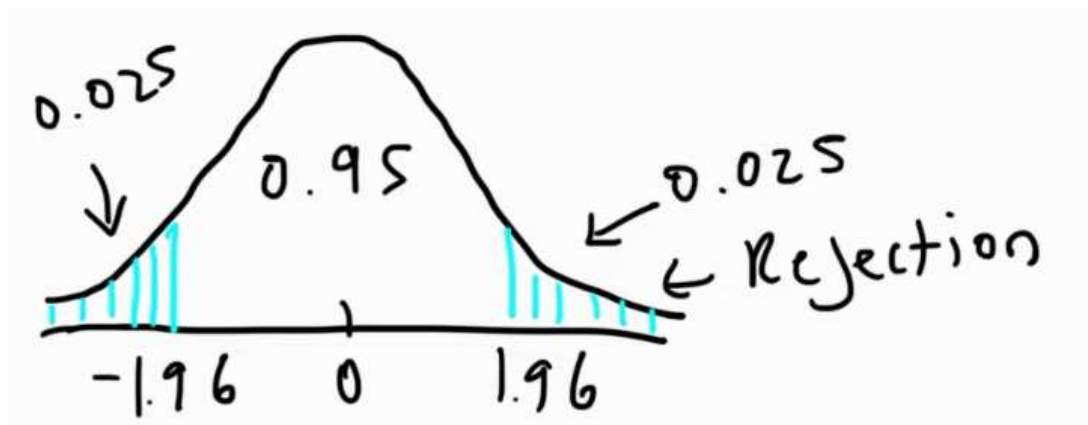
- (a) state the null and alternative Hypothesis.
- (b) at a 95% confidence level, is there enough evidence to support the idea that the machine is not working properly.

$$H_0: \mu \neq 80$$

Alternatively Hypothesis

$$H_a: \mu \neq 80$$

$$\bar{x} = 78, s = 2.5, n = 40 > 30$$

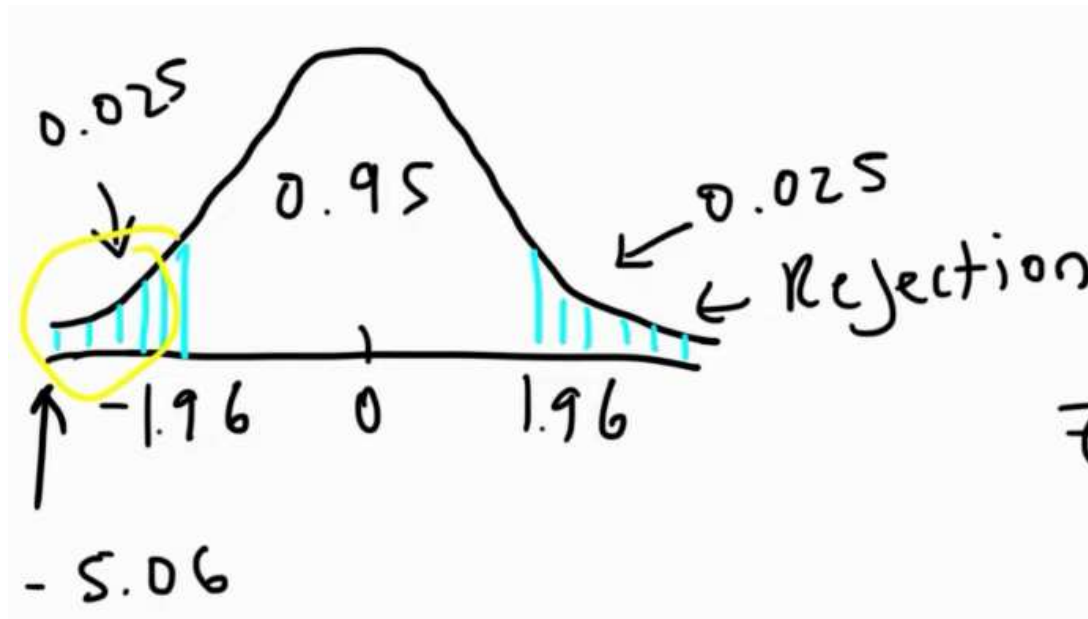


Since  $n > 30$  we can use

$$\begin{aligned} Z &= \frac{\bar{x} - \mu_0}{s/\sqrt{n}} \\ &= \frac{78 - 80}{2.5/\sqrt{40}} \\ &= \frac{-2}{0.39528} \end{aligned}$$

Calculated z value

$$z_c \approx -5.06$$



With a 95% confidence we can say that the machine is not working properly. Since our z value falls outside of our confidence interval.

#### 94. p-value

The **smaller the** the p-value the **stronger** the evidence against  $H_0$

The **larger** the p-value is, the **weaker** the evidence we have **against**  $H_0$

#### 95. The p-value approach

1. Define the parameters to be tested. eg  $\mu$  or  $P$
2. Define  $H_0$  and  $H_1$ .
3. Specify the test statistic and the distribution under  $H_0$ . (assume  $H_0$  is true)
4. Find the observed value of the test statistic.
5. Find the p-value.
6. Report the strength of evidence against  $H_0$  :
  - Very strong if  $p \leq 0.01$
  - Strong if  $0.01 < p \leq 0.05$
  - Moderate if  $0.05 < p \leq 0.1$
  - Little or none if  $0.1 < p$
7. Answer any other questions given (i.e. report the value of the estimate, report the value of the estimated standard error, etc.)

single tail	$\mu > or \mu < or \mu \geq or \mu \leq$
2 tail test	$\mu = or \mu \neq$

### Example

A certain medication is supposed to contain 350 mg of the active ingredient per pill. It is known from previous work that this content is normally distributed with a standard deviation of 3.5 mg. Suppose a random sample of 5 pills are taken, and the average content is 346.4 mg.

Is the mean pill content not 350 mg?

1. let  $\mu$  = true mean active ingredients per pill (mg)

2.  $H_0: \mu = 350$

$H_1: \mu \neq 350$  (therefore its 2 tail)

3 test statistic ad distribution

$\sigma = 3.5, n = 5$

$$Z_{obs} = \frac{\bar{x} - \mu}{\sigma / \sqrt{n}} \sim N(0, 1)$$

4

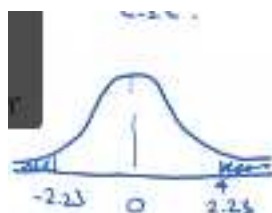
$$Z_{obs} = \frac{346.4 - 350}{3.5 / \sqrt{5}} = \frac{-3.6}{1.5652} = -2.23$$

**NOTE:** the 1.5652 is the estimated standard error (e.s.e)

5.  $p - value = 2 \times P(Z > 2.23) = 2 \times p(z < -2.23)$

$= 2 \times 0.0129$

$= 0.0258$



6. there is **strong** evidence against  $H_0$

7. Estimate = 346.4 , e.s.e = 1.5652

## 96. Relationship between Hypothesis and pvalue

Alternate Hypothesis	p-value
$h_1: \mu > \mu_0$	$p(z > z_{observed})$
$h_1: \mu < \mu_0$	$p(z < z_{observed})$
$h_1: \mu \neq \mu_0$	$2P(z < - z_{observed} ) \text{ or } 2P(z >  z_{observed} )$

## 97. Errors and Hypothesis Tests

Two types of errors are possible in a hypothesis test.

### Type I error

(Rejection Error) is made when we reject the null hypothesis when it is true.

### Type II error

(Acceptance Error) is made when we do not reject the null hypothesis when it is false.

### 97.1. Possibility of each type of error

$\alpha$  is the probability of making a Type 1 error

$\beta$  is the probability of making a Type 2 error

	$H_0$ true	$H_0$ false
Reject $H_0$	Type I	✓
do not reject $H_0$	✓	Type II

↓ $\alpha$	↑ $\beta$
↑ $\alpha$	↓ $\beta$



### 97.2. When do we reject $H_0$ ?

We are asked to test  $H_0$  at some significance level  $\alpha$ . We carry out the hypothesis test in much the same way: defining parameters, calculating the value of the test statistic, finding the p-value. Rather than giving the level of strength against  $H_0$ , as in the p-value

Approach, we instead either reject or don't reject  $H_0$  by the following rule:

- If  $p \leq \alpha$ , then reject the null hypothesis.
- If  $p > \alpha$ , then do not reject the null hypothesis. (Some will phrase this as "maintain the null hypothesis" or "fail to reject the null hypothesis")

#### Example

For the pill example, if we were asked to test our hypotheses at the level  $\alpha = 0.01$ , what would our conclusion be?

$$p - \text{value} = 0.0258 > \alpha = 0.01$$

Conclusion: maintain  $H_0$  at  $\alpha$

Example: What if we were testing at the level  $\alpha = 0.05$ ?

$$p - \text{value} < \alpha = 0.05$$

conclusion = reject  $H_0$

IMPORTANT: It is dishonest to set your value of  $\alpha$  after the data has been collected and examined; the value of  $\alpha$  should be made by taking into account the consequences of Type I and II errors before the study is carried out.

↑ set  $\alpha$  before the study is completed

## 98. Relationship between hypothesis testing and confidence intervals

Suppose we construct a  $(1 - \alpha)100\%$  confidence interval for  $\mu$ .

It is true that for any number  $k$  in this interval, that if we were to test  $H_0: \mu = k, H_1: \mu \neq k$ , we'd have a p-value greater than  $\alpha$ .

This means that if we were testing  $H_0: \mu = k, H_1: \mu \neq k$  at the level of  $\alpha$ , we would reject the null hypothesis if and only if  $k$  were not inside the  $(1 - \alpha)100\%$  confidence interval for  $\mu$ .

### Example

Using our pill data, we can find that a 95% confidence interval for  $\mu$  is (343.77, 349.03).

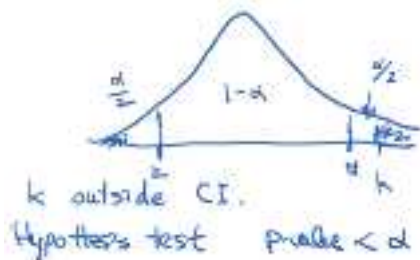
What would our conclusion be if we test  $H_0: \mu = 344, H_1: \mu \neq 344$  at the level  $\alpha = 0.05$ ?

$$\alpha = 1 - \text{confidence interval} = 1 - 95 = 0.05$$

344 is inside the CI

$$p\text{-value} > 0.05$$

$\therefore$  Retain  $H_0$



### Example

What would our conclusion be if we test  $H_0: \mu = 342, H_1: \mu \neq 342$  at the level  $\alpha = 0.05$ ?

(343.77, 349.03)

342 is outside the CI

$\therefore$  reject  $H_0$  because  $p\text{-value} < 0.05$

### Example

The lengths of mourning doves (from beak to tail) are known to be normally distributed. Suppose that 5 mourning doves are selected at random, and it is found that the average length of the mourning doves is 32.4 cm, with a standard deviation of 2.9 cm.

Let  $\mu$  denote the true mean length of mourning doves. Test the hypotheses  $H_0 : \mu = 30$ ,  $H_a : \mu > 30$  at the level  $\alpha = 0.1$ .

3. Test statistic and distribution:

Population is normal

$$n = 5, s = 2.9$$

Therefore we use  $t_{n-1}$

$$t_{obs} = \frac{\bar{x} - \mu}{s/\sqrt{n}} \sim t_{n-1} = t_4$$

$$s = 2.9 \text{ cm}$$

$$\bar{x} = 32.4 \text{ cm}$$

$$n = 5 < 30$$

$$H_0: \mu = 30$$

$$H_a: \mu > 30$$

$$\alpha = 0.1$$

4. 
$$t_{obs} = \frac{\bar{X} - \mu}{s/\sqrt{n}}$$

$$t_{obs} = \frac{32.4 - 30}{2.9/\sqrt{5}}$$

$$\frac{2.4}{1.296919}$$

$$t_{observed} = 1.85054$$

Now find p-value using t table

$$p - value = P(t_4 > 1.8505)$$

**NOTE:** the reason for using  $>$  is because the alternative is  $>$

In R

```
> 1 - pt(1.8505, 4)
[1] 0.06895478
```

Therefore the p-value is 0.06895478

$$0.05 < p - value < 0.1$$

6. Moderate evidence against  $H_0$

7. estimate = 32.4

e.s.e = 1.2969

p-value  $> 0.05$

$\therefore$  retain  $H_0$

### Example

In a sample of 46 people, we find the average blood glucose level upon waking up is 5.3 mmol/L with a standard deviation of 1.2 mmol/L. Is there reason to believe that the true mean blood glucose level upon waking for people is not 5 mmol/L?

let  $\mu$  denote true mean blood glucose level of people upon waking up

2.  $h_0: \mu = 5$   $h_1: \mu \neq 5$

2 tail

3. test statistic , distribution

$$z_o = \frac{\bar{x} - \mu}{s/\sqrt{n}} \sim N(0, 1)$$

4.  $z_{sub\ o} = \{5.3 - 5\} / \{1.2 / \sqrt{46}\} = 0.3 / 0.1764 = 1.6956$

$$\begin{aligned}
 5. \quad p\text{-value} &= 2P(z > 1.6956) \\
 &= 2P(z < -1.6956) \\
 &= 0.0892
 \end{aligned}$$

6. There is moderate evidence against  $H_0$
- $\bar{x} = 5.3 \text{ mol/L}$ ,  $\text{ese} = 0.1769 \text{ mmol/L}$

## STATS 260 Class 22

*Gavin Jaeger-Freeborn*

### 99. Set 30 Comparing Two Population Proportions

In this section we will consider scenarios where we take samples of 2 independent scenarios. Here we compare the two population proportions  $p_1$  and  $p_2$ .

To compare them we use  $p_1 - p_2$

$p_1 - p_2 \neq 0$	different
$p_1 - p_2 > 0$	larger
$p_1 - p_2 < 0$	smaller
$p_1 - p_2 = 0.1$	requires a reason to test this

To estimate  $p_1$  and  $p_2$  we use  $\hat{p}_1$  and  $\hat{p}_2$

Where  $\hat{p}_1$  and  $\hat{p}_2$  are 2 **sample proportions**.

$$\hat{p}_n = \frac{x_n}{\hat{n}_n}$$

### 100. Confidence Interval

CI: estimated  $\pm$  (c. v)(ese)

**100.1. Option 1** for if  $p_1 - p_2 \neq 0$  AKA unpooled

$$\hat{p}_1 - \hat{p}_2 \pm z_{\alpha/2} \sqrt{\frac{\hat{p}_1(1 - \hat{p}_1)}{n_1} + \frac{\hat{p}_2(1 - \hat{p}_2)}{n_2}}$$

**100.2. Option 2** for if  $p_1 - p_2 = 0$

AKA pooled

$$\hat{p} \pm z_{\alpha/2} \sqrt{\hat{p}(1 - \hat{p}) \left( \frac{1}{n_1} + \frac{1}{n_2} \right)}$$

**101. Test Statistic**

$$\text{test statistic} = \frac{\text{estimate} - \text{parameter value}}{\text{ese}}$$

$$Z = \frac{(\hat{p}_1 - \hat{p}_2) - (p_1 - p_2)}{\sqrt{\frac{p_1(1 - p_1)}{n_1} + \frac{p_2(1 - p_2)}{n_2}}} \sim N(0, 1)$$

Here the **estimated standard error** is since we don't have  $\hat{p}_1$  or  $\hat{p}_2$

$$\sqrt{\frac{\hat{p}_1(1 - \hat{p}_1)}{n_1} + \frac{\hat{p}_2(1 - \hat{p}_2)}{n_2}}$$

**101.1. Option 1** for if  $p_1 - p_2 \neq 0$

AKA unpooled

$$\therefore Z = \frac{(\hat{p}_1 - \hat{p}_2) - (p_1 - p_2)}{\sqrt{\frac{\hat{p}_1(1 - \hat{p}_1)}{n_1} + \frac{\hat{p}_2(1 - \hat{p}_2)}{n_2}}} \sim N(0, 1)$$

**101.2. Option 2** for if  $p_1 - p_2 = 0$

AKA pooled

We assume that  $p_1 = p_2 = p$  by combining them to form a single random sample.

$$\hat{p} = \frac{x_1 + x_2}{n_1 + n_2}$$

$$\therefore \text{ese} = \sqrt{\hat{p}(1 - \hat{p})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}$$

$\hat{p} \equiv$  pooled sample

**NOTE:**  $V(\hat{p}_1 - \hat{p}_2) = V(\hat{p}_1) + V(\hat{p}_2)$

**NOTE:**  $\hat{p}_{1or2} \approx N\left(p_n, \sqrt{\frac{p_{1or2}(1 - p_{1or2})}{n_1}}\right)$

### Example

Motherboards are made by one of two manufacturing processes. 300 motherboards made by the first process and 500 mother boards made by the second process are sampled at random. From the first process, 15 have flaws. From those made by the second process, 30 have flaws. Let  $p_1$ ,  $p_2$  denote the proportion of motherboards made by process one, two (respectively) which are defective.

- (a) What is the estimate for  $p_1 - p_2$ ?
- (b) What is the unpooled estimated standard error of  $\hat{p}_1 - \hat{p}_2$ ?
- (c) Test the research hypothesis that the first process makes a smaller proportion of defective items than the second process, using the e.s.e. from part (b).
- (d) Test the same hypotheses in (c), this time using the pooled estimated standard error.
- (e) Create a 93% confidence interval for  $p_1 - p_2$ .
- (f) What does the confidence interval tell you about  $p_1 - p_2$ ?
- (g) Suppose we wish to use these data as a pilot study to estimate the sample size we would need in the future to create a 95% confidence interval with a margin of error of 0.01. What sample size is needed (assuming that  $n_1 = n_2$ ).

## STATS 260 Class 22

*Gavin Jaeger-Freeborn*

### 102. Sets 28 and 29

We may wish to compare  $\mu_1$  and  $\mu_2$ , the population means for populations 1 and 2. We do so by examining the difference,  $\mu_1 - \mu_2$ .

**NOTE:** We want to find the difference between 2 means

### Example

Suppose we wish to compare the  $\mu_1$ , the mean lead content (in ppm) per mL in Victoria tap water with  $\mu_2$ , the mean lead content (in ppm) per mL in Vancouver tap water.

- If the means are equal, then  $\mu_1 - \mu_2 = 0$ .
- If the means are different, then  $\mu_1 - \mu_2 \neq 0$ .



- If the lead content is higher in Victoria, then  $\mu_1 - \mu_2 > 0$ .
- If the lead content is higher in Vancouver, then  $\mu_1 - \mu_2 < 0$ .
- If the lead content is higher in Victoria by at least 4 ppm than in Vancouver, then  $\mu_1 - \mu_2 > 4$
- If the lead content is higher in Vancouver by at least sub 2 ppm than in Victoria, then  $\mu_1 - \mu_2 < -2$

The points on estimation for  $\mu_1 - \mu_2$  we will use is  $\bar{x}_1 - \bar{x}_2$

As before the confidence interval we construct will have the form.

$$\text{estimate} \pm (c. v)(e. s. e. )$$

All pivotal quantities have the form

$$\frac{\text{estimate} - \text{parameter}}{e. s. e.}$$

### 103. Three Case Scenarios

#### 103.1. Large Sample Size Procedures $n > 40$

$$\bar{x}_1 \approx N\left(\mu_1, \frac{\sigma_1}{\sqrt{n_1}}\right)$$

$$\bar{x}_2 \approx N\left(\mu_2, \frac{\sigma_2}{\sqrt{n_2}}\right)$$

$$\bar{x}_1 - \bar{x}_2 \approx N\left(\mu_1 - \mu_2, \frac{\sigma_1}{\sqrt{n_1}} - \frac{\sigma_2}{\sqrt{n_2}}\right)$$

Test Statistic

$$Z = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\left(\frac{s^2_1}{n_1}\right) + \left(\frac{s^2_2}{n_2}\right)}}$$

#### Assumptions

- Independent random samples from two populations.
- **Both** sample sizes are large ( $n_1 \geq 40, n_2 \geq 40$ ) and the population standard deviations are unknown.

- Populations may have any distribution

**NOTE:**  $\sigma_1, \sigma_2$  are unknown and finite

**estimate**

$$\bar{x}_1 - \bar{x}_2$$

**e.s.e**

$$\sqrt{\left(\frac{s_1^2}{n_1}\right) + \left(\frac{s_2^2}{n_2}\right)}$$

#### 104. Cases 2 & 3 Small Sample Size

We Calculate

$$\frac{\max(s_1, s_2)}{\min(s_1, s_2)}$$

CASE 2	CASE 1
If $\leq 1.4$ , we assume $\sigma_1 = \sigma_2$ .	If $\geq 1.4$ , we assume $\sigma_1 \neq \sigma_2$ .
↓	↓
Then use pooled procedure.	Then use pooled procedure.

#### 105. Pooled Procedure

Test Statistic

$$t_{n_1+n_2-2} = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2} \left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

est - value under  $H_0$

$$(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)$$

ese (estimated standard error)

$$\sqrt{\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2} \left(\frac{1}{n_1} + \frac{1}{n_2}\right)}$$

### Assumptions

- Independent random samples from two populations. **At least one of the sample sizes is small**, and the population standard deviations are unknown.
- We know that (or assume that)  $\sigma_1 = \sigma_2$
- Both populations have normal (or approximately normal) distribution.

**NOTE:** The value is sometimes denoted  $\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}$  by  $s_p^2$ , and is called the pooled variance estimate.

Recall that for pooled procedures, we assumed that  $\sigma_1 = \sigma_2$ . The value of  $s_p^2$  is the estimate for both  $\sigma_1^2$  and  $\sigma_2^2$ .

### 106. Unpooled Procedure

Test Statistic

$$t_\gamma = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)}}$$

where  $\gamma$  is the number of degrees of freedom is the number of

$$v = \frac{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}{\frac{\left(\frac{s_1^2}{n_1}\right)^2}{n_1 - 1} + \frac{\left(\frac{s_2^2}{n_2}\right)^2}{n_2 - 1}}$$

**NOTE:**  $v$  = degrees of freedom

### Assumptions

- Independent random samples from two populations. At least one of the sample sizes is small, and the population standard deviations are unknown.
- We know that (or assume that)  $\mu_1 \neq \mu_2$
- Both populations have normal (or approximately normal) distribution.

**NOTE:** We always round down (i.e. take the integer part) as the number of degrees of freedom when we are using our tables.

If you are carrying out an unpooled test on  $\mu_1 - \mu_2$  using R (or other software), the p-value will be calculated using the unrounded value of  $v$  as the degrees of freedom.

### Example

We wish to compare the cube compressive strength (in  $\text{N/mm}^2$ ) of two types of concrete. The summary statistics are as follows:

	sample size	sample mean	sample sd
Type A	$n_1 = 70$	$\bar{x}_1 = 31.9$	$s_1 = 1.4$
Type B	$n_2 = 50$	$\bar{x}_2 = 35.6$	$s_2 = 2.1$

Test the research hypothesis that the two types of concrete have **different** mean cube compressive strengths.

1. let  $\mu_1$  = mean cube compressive strength ( $\text{inN/mm}^2$ ) of type A concrete let  $\mu_2$  = mean cube compressive strength ( $\text{inN/mm}^2$ ) of type B concrete
2.  $H_0: \mu_1 - \mu_2 = 0$   $H_1: \mu_1 - \mu_2 \neq 0$  ( 2-tailed test )
3. test statistic + distribution 1

$$n_1, n_2 > 40 \rightarrow \text{use CASE 1}$$

$$Z_{\text{obs}} = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\left(\frac{s_1^2}{n_1}\right) + \left(\frac{s_2^2}{n_2}\right)}} \sim N(0, 1)$$

$$4. \quad \text{compute } Z_{\text{obs}} = \frac{(31.9 - 35.6) - (0)}{\sqrt{\frac{1.4^2}{70} + \frac{2.1^2}{50}}} = -10.854$$

$$5. \quad \begin{aligned} p - \text{value} &= 2 \cdot P(Z < -|Z_{\text{obs}}|) \\ &= 2 \cdot P(Z < -(0.854)) \\ &\approx 0 (< 0.01) \end{aligned}$$

**NOTE:**  $\mu_1 - \mu_2 = 0$  are the same but still unknown

6. There is very strong evidence against  $H_0$  in R ,  $p - \text{value} = 2.01 \cdot 10^{-8}$

### Example

We wish to compare the lifespans of smart-phones produces by two companies. Let  $\mu_1$ ,  $\mu_2$  be the mean lifespan (in weeks) of smart phones produced by Company A, B (respectively). The summary of our study's observations are as follows:

Company A	Company B
$\bar{x}_1 = 148$	$\bar{x}_2 = 153$
$s_1 = 8.3$	$s_2 = 5.1$
$n_1 = 15$	$n_2 = 6$

- What is the estimated standard error of  $\bar{x}_1 - \bar{x}_2$ ?
- What probability distribution is used to calculate the p-value in a hypothesis test on  $\mu_1 - \mu_2$ ? (Note: It is not enough to just say "t-distribution"; you must also specify the number of degrees of freedom)
- Test  $H_0: \mu_1 - \mu_2 = 0$ ,  $H_a: \mu_1 - \mu_2 < 0$ , at the significance level  $\alpha = 0.1$ .

$$n_1, n_2 < 40$$

$$\frac{s_1}{s_2} = \frac{8.3}{5.1} = 1.627 > 1.4$$

$\therefore$  use CASE 3 unpooled  
use formula (I)

(a)

$$ese = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} = \sqrt{\frac{8.3^2}{15} + \frac{5.1^2}{6}} = 2.9879$$

(b)

We need to use the t-distribution

$$dof(orv) = \frac{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}{\frac{\left(\frac{s_1^2}{n_1}\right)^2}{n_1 - 1} + \frac{\left(\frac{s_2^2}{n_2}\right)^2}{n_2 - 1}} = \frac{79.703}{1.5066 + 3.7584} = 15.138$$

after rounding down we get  $v = 15$

(c)

$$t_{\text{obs}} = \frac{(148 - 153) - 0}{2.9879}$$

$$= -1.673$$

$$p\text{-value} = P(t_{15} < -1.673) = P(t_{15} > 1.673)$$

$$0.05 < p\text{-value} < 0.10 = \alpha$$

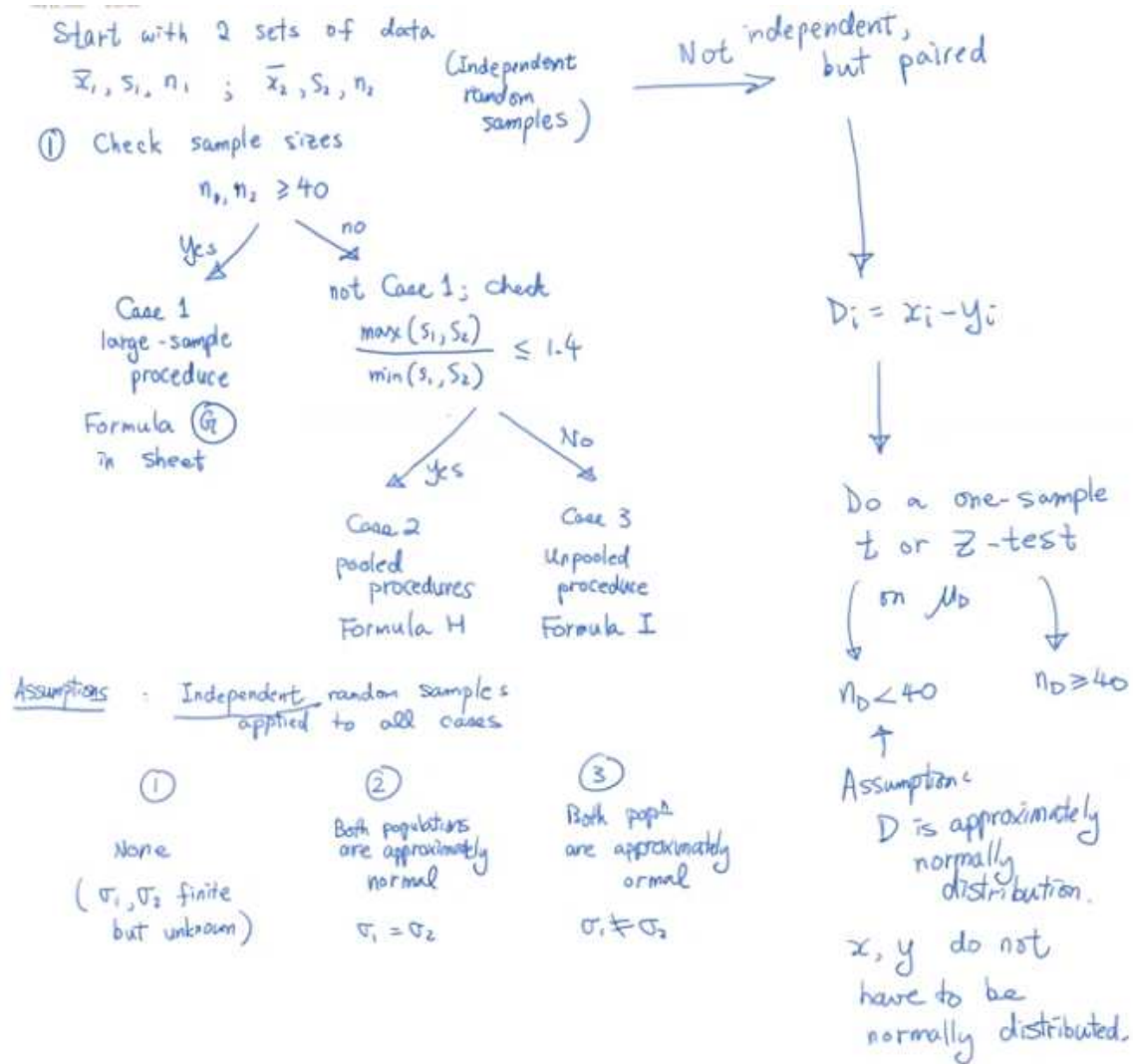
$$p\text{-value} < \alpha$$

Reject  $H_0$

## 107. Decision Tree For Comparing Two Population Means

Start with 2 sets of data

$\bar{x}_1, s_1, n_1; \bar{x}_2, s_2, n_2$



## 108. Set 31

## 109. Two - Sample Paired Test

Sometimes, the data we have collected forms a set of **matched pairs**. Rather than having two **independent samples (AKA not independent)**  $x_1, \dots, x_{n_1}$  and  $y_1, \dots, y_{n_2}$  we have pairs of observations  $(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$ . Some examples:

### Examples

- have a sample of cancer patients who will receive a new drug, we first measure the size of their tumor before receiving the medication, and again after receiving the medication.
- We have two appraisers working for an insurance company. The first appraiser examines ten works of art, and then the second appraiser examines the same ten works of art.

**NOTE:** We need to make sure that  $n_1 = n_2$  and that the samples are the same.

### Example

if the first appraiser examined ten works of art, and the second appraiser examined ten different works of art

*Here there are independent since they are completely independent pieces of art.*

Here we would use stuff from set 28, and 23

### Notation

If  $x_i$  is the  $i$ th observation from the first sample, and  $y_i$  is the  $i$ th observation from the second sample, then  $D_i$  is the difference between These two observations:

$$D_i = x_i - y_i$$

**NOTE:**  $D_1$  means single sample analysis if ( $n < 40$ )

The parameter of interest is  $\mu_D$  the mean difference between samples.

The estimate will be  $\bar{x}_D$ , the average of the observed differences.

$$t = \frac{\bar{x}_D - \mu_D}{s_D / \sqrt{n_D}}$$

(where  $s_D$  is the sample standard deviation of the  $n_D$  differences.)

This test statistic has  $t$  distribution with  $n_D - 1$  degrees of freedom.

**NOTE:** if the sample size is larger then ( $n_{sub D} \geq 40$ ) we can just use standard normal distribution.



### Example

An insurance company is worried about differences in the values of art objects as estimated by two appraisers. The company selects 5 works of art, and asks both appraisers to determine a value. The following are the appraised values (in millions of dollars). Is there a significant difference in appraised values?

		Object 1	Object 2	Object 3	Object 4	Object 5
$x_i$	Appraiser 1	22.10	92.70	2.76	75.60	4.13
$y_i$	Appraiser 2	21.30	92.10	1.54	78.90	4.78
	D	0.8	0.6	1.22	-3.3	-0.65

1. let  $\mu_0$  be mean of the difference in the values of art objects estimated by appraiser 1 and appraiser 2 ( $a_1 - a_2$ )

2.  $H_0: \mu_0 = 0$

$$H_0: \mu_0 \neq 0$$

**NOTE:** Two Tailed Test

3. test statistic and distribution

$$t_{\text{obs}} = \frac{\bar{x}_D - \mu_D}{\frac{S_D}{\sqrt{n_D}}} \sim t_{n_D-1} = t_4$$

$$\bar{x}_D = -0.266, s_D = 1.8335$$

- 4.

$$t_{\text{obs}} = -0.266 - \frac{0}{\frac{1.8335}{\sqrt{5}}} = -0.3244$$

- 5.

$$p\text{-value} = 2 \cdot P(t_4 > 0.3244)$$

$$p\text{-value} > 0.1$$

6. there is little evidence against the  $H_0$ .  $\therefore$  no evidence that two appraiser's appraisals are different

$$\bar{x}_D = 0.266, \text{ese} = 0.82$$

**Example**

Find a 99% confidence interval for  $\mu_D$ , the mean difference in the appraisal values.

$$\bar{x}_D \pm t_{4, 0.005} \cdot \frac{s_D}{\sqrt{n_D}}$$

$$0.266 \pm 4.604 \cdot 0.$$

$$(-4.041, 3.509)$$

**NOTE:** 0 is inside the confidence interval