1. Probability

Probability is the measure of the likelihood of a random phenomenon or chance behaviour. When we calculate the probability of an event occurring, it is the long-term proportion with which a certain outcome will occur.

1.1 Properties of Probability

1. The probability of any event A must be between 0 and 1 inclusive, $0 \le P(A) \le 1$

2. The sum of the probabilities of all disjoint events must equal 1,
$$P(A_1)+P(A_2)+...+P(A_k)=P(S)=1$$

3. The **complement** of any event A is equivalent to the event that does not include A and is denoted by A'. The probability of the complementary event is 1 – *P*(A).

4. **Addition Rule for Combined Events**
$$P(A \text{ or } B)=P(A \cup B)=P(A)+P(B)-P(A \cap B)$$

5. **Addition Rule for Disjoint Events**
If the events A and B have no common outcomes, then they are said to be disjoint (or **mutually exclusive**) and
$$P(A \text{ or } B)=P(A \cup B)=P(A)+P(B).$$
Also the probability that at least one of them occurs.

6. The **conditional probability** is the probability of event A given the occurrence of event B and is determined by
$$P(A \mid B)=\frac{P(A \cap B)}{P(B)}$$

7. **Independence**: If event A and event B are independent,
$$P(A)=P(A \mid B)=P(A \mid B^c)$$
Note: Independence ≠ Disjoint

8. **Dependence**: If event A and event B are dependent,
$$P(A \mid B) \neq P(A \mid B^c)$$
Note: Dependence ≠ Causality

9. **Multiplication Rule**: Two events, A and B, are **independent** if the occurrence of event A does not affect the probability of event B. Then,
$$P(A \text{ and } B)=P(A \cap B)=P(A)P(B)$$

10. **Bayes' Rule** can be applied when an observed event A occurs with any one of several mutually exclusive and exhaustive events, B and B^c.
$$P(A \mid B)=\frac{P(B \mid A)P(A)}{P(B \mid A)P(A)+P(B \mid A^c)P(A^c)}$$

1.2 Practice Problems

Q1. A fair die is tossed, and the up face is noted. If the number is even, the die is tossed again; if the number is odd; a fair coin is tossed. Define the events:

A: {A head appears on the coin}

B: {The die is tossed only one time}.

a) List the sample space.

b) Give the probability for each of the outcome in the sample space.

c) Find P(A) and P(B).

d) Find $P(A^c)$, $P(B^c)$, $P(A \cap B)$, $P(A \cup B)$, $P(A|B)$, and $P(B|A)$.

e) Are A and B mutually exclusive events? Independent events? Why?

$$P(A)P(B) \neq P(A \cap B)$$
$$P(A)P(B) = 1/8$$
$$P(A \cap B) = 1/4$$

Q2. Suppose that it is known that, among students taking a certain entrance exam, 10% cheat on the exam and 90% do not cheat on the exam. Also, about 70% of the cheaters get a perfect score, while only 20% of the non-cheaters get a perfect score.

a) What is the probability that a randomly selected student taking the exam will get a perfect score?

b) Given that the student has a perfect score on the exam, what is the probability that it was obtained by cheating?

Q3. Common sources of caffeine are coffee and tea. Suppose that 55% of the students at University drink coffee, 30% drink tea and 10% drink both.

a) If we randomly choose one student, what is the probability that this student drinks neither tea nor coffee?

b) If we randomly choose two students, find the probability that at least one of these students is a student who drinks only tea.

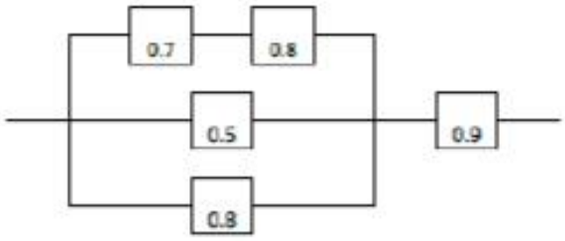
c) In addition to the information given above, suppose now that 70% of the students who drink coffee are females while only 35% of those who don't drink coffee are females. If we randomly select one student, what is the probability of selecting a female?

d) Suppose that the randomly selected student is a female, what is the probability that she is NOT a coffee drinker?

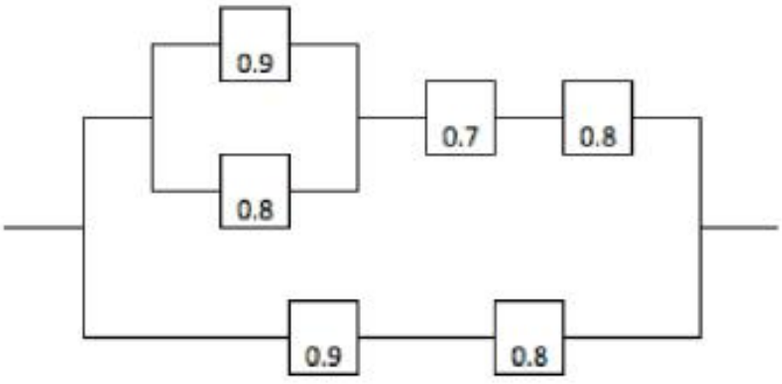
Q4. There are two types of basic systems introduced in the fundamentals of electric circuits, (a) **series/sequential** and (b) **parallel** systems. The following figure shows both systems including 3 electronic elements (A, B, and C). *A series system works only if all elements are simultaneously working, while a parallel system works if at least one of the elements works* (so the parallel system fails if all elements fail).

(a) Series system (b) Parallel system

a) In the following figure, the labeled values are the probabilities that each element works well.
If elements work **independently**, what is the probability that the whole system **fails**?



b) In the following figure, the labeled values are the probabilities that each element works well. If elements work **independently**, what is the probability that the whole system **fails**?



Q4. A diagnostic test for drug use among Olympic athletes based on a blood sample was thought to be excellent. Among drug users, 98% had a positive test. Among nondrug users, 95% had a negative test. Suppose that 1% of athletes are, indeed, drug users.

a) What proportion of athletes tested would have a positive test?

b) Of those athletes whose test is positive, what proportion are really drug users?

Q5. Professors know that 25% of students get perfect on the exam. Also, about 70% of the cheaters get a perfect score, while only 20% of the non-cheaters get a perfect score.
What is the probability that a student cheats on an exam?

Q6. An urn contains 5 marbles labelled from 1 to 5. You select one marble and from the remaining four marbles, you select another one.

a) What is the probability that you select a marble with an even number on your first selection?

b) What is the probability that the second marble has an even number on it?

c) What is the probability that you select one even and one odd marble.

Q7. An urn contains 15 blue marbles and 5 red marbles. You select 3 marbles from the urn in sequence without replacement. What is the probability that the third marble you selected is a Blue marble.

Q8. A particular element can be manufactured in three different workshops. It is manufactured in workshop A and workshop B with 80 and 15 percent, respectively. The elements made by workshop A have a 4% rate of defects, workshop B elements have a 6% rate of defects, and workshop C elements have a 9% rate of defects. If a randomly selected element is tested and is found to be defective, find the probability that it was made in workshop A.

Q9. There are 15 female and 20 male students in a class. They have this option to enroll in either math or chemistry course (not both). Reports show that 12 female and 15 male students have enrolled in the math course. If one student is randomly selected, what is the probability that this student is female given that she/he is enrolled in the chemistry course?

2. Probability Distributions

2.1 Random Variables

A function that maps a random event into a number.
I'll explain.

A **discrete probability distribution** lists each possible value the random variable can assume, together with its probability. A probability distribution must satisfy the following conditions:

1. The probability of each value of the discrete variable is between 0 and 1

$$0 \le P(x) \le 1$$

2. The sum of all the probabilities is 1

$$\sum P(x) = 1$$

3. Cumulative Distribution Function

$$F(x) = P(X \le x) = \sum_{t \le x} P(X = t)$$

The **expected value** or mean of a discrete random variable is found as follows:

$$\mu_X = E(X) = \sum_x xP(X = x)$$

Each value of x is multiplied by its corresponding probability and the products are added.

The variance of a discrete random variable is found as follows:

$$\sigma_x^2 = V(X) = \sum_x (x - \mu_x)^2 P(X = x) \quad \text{OR} \quad \sigma_x^2 = V(X) = \sum_x x^2 P(X = x) - \mu_x^2$$

Each value of x is subtracted by the mean, squared, then multiplied by its corresponding probability. The products are then added.

A simple example of a discrete random variable and its distribution:

Continuous Random Variables
A random variable is continuous if its probabilities are given by areas under a curve. This curve is called the probability density function (pdf) and usually denoted by f(x).

1. Probabilities are calculated as follows:

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

$$P(X < b) = P(X \leq b) = \int_{-\infty}^{b} f(x)dx$$
$$P(X > a) = P(X \geq a) = \int_a^{+\infty} f(x)dx$$

2. Total area under the pdf is 1

$$\int_{-\infty}^{+\infty} f(x)dx = 1$$

3. Cumulative Distribution function of X is

$$F(x) = P(X \leq x) = \int_{-\infty}^{x} f(t)dt$$

4. Mean of X or Expected Value of X

$$\mu_X = E(X) = \int_{-\infty}^{+\infty} xf(x)dx$$

5. Variance of X

$$\sigma_x^2 = \int_{-\infty}^{+\infty}(x-\mu_X)^2 f(x)dx \quad \text{OR} \quad \sigma_x^2 = \int_{-\infty}^{+\infty} x^2 f(x)dx - \mu_x^2$$

6. Median (x_m) and Percentiles (x_p *is the* p^{th}*percentile*)

$$F(x_m) = P(X \leq x_m) = \int_{-\infty}^{x_m} f(x)dx = 0.5$$

$$F(x_p) = P(X \leq x_p) = \int_{-\infty}^{x_p} f(x)dx = \frac{p}{100}$$

Rules for Means

Rule 1. If X is a random variable and *a* and *b* are constants,

$$\mu_{a+bX} = a + b\mu_X$$
$$E[a+bX] = a + bE[X]$$

Rule 2. If X and Y are both random variables and *a* and *b* are constants,

$$\mu_{aX+bY} = a\mu_X + b\mu_Y$$
$$E[aX+bY] = aE[X] + bE[Y]$$

Rules for Variances and Standard Deviations

Rule 1. If X is a random variable and *a* and *b* are constants,

$$\sigma^2_{a+bX} = b^2\sigma^2_X$$
$$Var[a+bX] = b^2 Var[X]$$

Rule 2. If X and Y are independent,

$$\sigma^2_{X+Y} = \sigma^2_X + \sigma^2_Y \quad \text{or} \quad Var[X+Y] = Var[X] + Var[Y]$$
$$\sigma^2_{X-Y} = \sigma^2_X + \sigma^2_Y \quad \text{or} \quad Var[X-Y] = Var[X] + Var[Y]$$

2.2 Practice Questions

Q1. Flip a fair coin. You win $1 if it lands Heads. You lose $1 if it lands Tails.

Q2. We roll a fair die and if up face is even, you win the dollar amount corresponding to the up face. If the up face is odd, you lose the dollar amount corresponding to the up face. Find the expected value and standard deviation of your winning. Would you ever see this kind of game at a casino?

Q3. An urn contains 10 marbles, each of which shows a number. Five marbles show 1, two show 2, and three show 3. A marble is drawn at random.

a) If X is the number that shows, determine the mean, variance, and standard deviation.

b) What would happen to the mean, variance and standard deviation if I were to replace the three marbles showing a 3 by ones showing a 30?

Q4. The Springville School System pays bonuses to its physical-education teachers according to their body types. Eighty percent of Springville's physical-education teachers are mesomorphs, fifteen percent are endomorphs, and the remaining five percent are ectomorphs.
Seventy percent of mesomorphic physical-education teachers have tenure, and get bonuses of $10,000. Sixty percent of endomorphic physical-education teachers have tenure, and get bonuses of $7500. Thirty percent of ectomorphic physical-education teachers have tenure, and get bonuses of $2000. The non-tenured physical-education teachers all get bonuses of $1000.

a) Find the mean and standard deviation in the bonuses of Eurelian physical-education teachers.

b) Of those who get bonuses less than $6000 (Eurelian), what proportion have tenure?

c) Are the events 'having tenure' and 'endomorph' are statistically independent?

2.3 Jointly Distributed Random Variables

Discrete Joint Distribution

When we have probabilities that depend on two random variables (X and Y), we call them joint probabilities.

For discrete random variables, the joint probability mass function is denoted as:

$$p(x, y) = P(X = x \; and \; Y = y)$$

Marginal probability mass functions of X and Y are given by:

$$p_X(x) = P(X = x) = \sum_y p(x, y) \text{ and } p_Y(y) = P(Y = y) = \sum_x p(x, y)$$

The join probability mass function sums up to 1

$$\sum_x \sum_y p(x, y) = 1$$

Examples
In a dental plan participants can opt either for one or two general checkups a year. A record has been kept of the number of additional claims made by participants, and the following joint-probability table has been calculated.

		X = number of annual checkups	
		1	2
Y = number of additional claims	0	0.09	0.17
	1	0.29	0.18
	2	0.12	0.12
	3	0.00	0.03

a) Find the marginal probability distribution of X and Y. Find the mean and standard deviation of X and Y. (Solve together)

b) For those who opt for one checkup per year, what is their mean number of additional claims?
(Solve together)

Continuous Joint Distribution
For continuous random variables, the joint probabilities are calculated as follows:

$$P(a \leq X \leq b \text{ and } c \leq Y \leq d) = \int_a^b \int_c^d f(x,y)dydx$$

$$\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f(x,y)dydx = 1$$

The marginal probability density functions of f(x,y) are given by:

$$f_X(x) = \int_{-\infty}^{\infty} f(x,y)dy \quad \text{AND} \quad f_Y(y) = \int_{-\infty}^{\infty} f(x,y)dx$$

Conditional probability mass function

$$p_{Y|X}(y|x) = \frac{p(x,y)}{p(x)}$$

Examples
Let's go solve some past midterm questions.

2.4 Binomial Distribution

There are many probability experiments for which the results of each trial can be reduced to two outcomes: success and failure. For instance, when a basketball player attempts a free throw, he or she either makes the basket or does not. Probability experiments such as these are called binomial experiments.

The binomial distribution is a discrete probability distribution function with 4 essential properties:

1. The sample consists of a fixed number of observations, n.
2. Each observation is classified into one of two mutually exclusive and collectively exhaustive categories, usually called *success* and *failure*.
3. The probability of an observation being classified as success, p, is constant from observation to observation. Thus, the probability of an observation being classified as failure, $1 - p$, is constant over all observations.
4. The outcome (i.e. success or failure) of any observation is independent of the outcome of any other observation.

The binomial probability distribution for obtaining a number of successes (X), given a knowledge of this distribution's parameters n and p is given by:

$$P(x) = f(x) = \binom{n}{x} p^x (1-p)^{n-x}$$

where P(x) = probability of k successes
n = sample size
p = probability of success
$1 - p$ = probability of failure
x = number of successes in the sample

The mean m of the binomial distribution is:

$$E(x) = \mu = np$$

The standard deviation s of the binomial distribution is:

$$\sigma = \sqrt{Var(x)} = \sqrt{np(1-p)}$$

Let's use a simple example to motivate the Binomial Distribution.

Example 1 (Solve in Class)
Suppose we flip a unfair coin (probability of tossing a Heads is 60%) three times. Let X denote the number of Heads we end up with.

What is the Sample Space of this Random Experiment?

What values can x take?

What is the Probability of P(x=0)?

What is the Probability of P(x=1)?

What is the Probability of P(x=2)?

What is the Probability of P(x=3)?

2.5 Geometric Distribution

The requirement of the Geometric Distribution is exactly the same as the Binominal Distribution. Our random experiment still consists of a series of independent Bernoulli trials with a constant probability of success/failure.

But this around, instead of being interested in the number of successes in n-trials, we are actually interested in the number of trials required until the first success.

Example 1 (Solve in Class)
I have an unfair coin (probability of tossing a Heads is 60%).
What is the probability of the first heads coming on the first trial? Second trial? Fifth trial? Tenth trial? N'th trial?

Let x denote the number of trials needed until the first Heads.

So in general, if the Bernoulli trial has a probability of success of p and a probability of failure of 1-p, then the probability distribution is:

$$P(x) = f(x) = (1-p)^{x-1} p$$

The mean and variance are

$$\mu = E(x) = \frac{1}{p} \quad \text{and} \quad \sigma^2 = Var(x) = \frac{1-p}{p^2}$$

2.6 Negative Binomial Distribution

The requirement of the Geometric Distribution is exactly the same as the Binominal Distribution. Our random experiment still consists of a series of independent Bernoulli trial with a constant probability of success/failure.

But this time around, instead of being interested in the number of successes in n-trials, we are actually interested in the number of trials required until the k^{th} success.

Example 1 (Solve in Class)
Same unfair coin (P(Heads)=0.60)
What is the probability of the second head coming on the third trial? Fourth trial? 10^{th} trial?

Let x denote the number of trials needed until the k^{th} Heads

The probability of needing x trials to get k successes is given by:

$$P(x) = f(x) = \binom{x-1}{k-1}(1-p)^{x-k}\,p^k$$

The mean and variance are

$$\mu = E(x) = \frac{k}{p} \quad \text{and} \quad \sigma^2 = Var(x) = \frac{k(1-p)}{p^2}$$

2.7 Hypergeometric Distribution

The binomial distribution and the hypergeometric distribution are concerned with the same thing: the number of success in a sample containing *n* observations. What's different between these two distributions is that in the binomial model, the sample data are drawn with replacement whereas in the hypergeometric model, the sample data are drawn without replacement. In other words, each outcome in a binomial distribution has a constant probability of occurring (*p*) while in a hypergeometric distribution, the outcome of one observation is affected by the outcomes of the previous observations.

The probability of having *x* successes in a sample of size *n* from a population with *r* successes of size *N* is:

$$P(x) = f(x) = \frac{\binom{k}{x}\binom{N-k}{n-x}}{\binom{N}{n}}$$

The mean of the hypergeometric distribution is

$$\mu = E(x) = np \text{ where } p = {}^{k}\!/_{N}$$

The standard deviation of the hypergeometric distribution is

$$\sigma^2 = Var(x) = np(1-p)\left(\frac{N-n}{N-1}\right)$$

Note: **The binomial distribution approximates the hypergeometric well when n is small compared to N. Rule of thumb is when** ${}^{n}\!/_{N} \le 0.05$.

2.8 The Poisson Distribution

The **Poisson distribution** is a discrete probability distribution. It expresses the probability of a number of events occurring in a fixed period of time/space if these events occur with a known average rate, and are independent of the time since the last event.

The mathematical expression for the Poisson distribution for obtaining *x* successes given that λ are expected is:

$$P(k) = \frac{e^{-\lambda}\lambda^{k}}{k!}$$

Both the mean and the variance are equal to the Poisson parameter:

$$\mu = \sigma^2 = \lambda$$

If X has a Poisson distribution with parameter λ , we denote it as:

$$X \sim Poisson(\lambda)$$

Note: If n is large and p is small, we can use the Poisson Distribution to approximate the Binomial Distribution.
This approximation works best when n > 50 and np < 5.

2.9 Practice Questions

Q1. The CRA audits 20% of all income tax returns submitted by working Canadians.

a) We randomly select 10 returns. What is the probability that at least 2 of them are audited?

b) If we take a look at the tax returns one at a time, what is the probability that the first return that is audited occurs on the fourth chosen tax return?

c) If we take a look at the tax returns one at a time, what is the probability that the 3rd tax return that is audited occurs on the 6th chosen tax return?

Q2. A major hurricane is a hurricane with wind speeds of 111 miles per hour or greater. During the 20th century, the mean number of major hurricanes to strike the U.S. mainland per year was about 0.7. Find the probability that in a given year

a) exactly one major hurricane will strike the U.S. mainland

b) at most one major hurricane will strike the U.S. mainland

c) more than one major hurricane will strike the U.S. mainland

d) What is the standard deviation of the number of hurricanes in 5 years?

e) What is the probability of the number of hurricanes in 5 years being within 1 standard deviations of the mean?

Q3. a) A manufacturer will accept a large shipment of a particular item from a supplier if a random sample of 12 items contains no more than two defective items. Find the probability that a large shipment will be rejected if the shipment actually contains exactly 5% defective items. After calculating the probability, round-off the answer to two decimals.

b) If eight iterations of the acceptance/rejection procedure in part (a) are carried out, what is the probability that at least two of the eight iterations would result in rejecting the shipment, again assuming that the actual percentage of defective items in the shipment is 5%?

Q4. Company XYZ has hundreds of thousands of employees world wide. The following join-probability table has been prepared based on employee data.

	Female	Male
University Degree	0.4	0.3
No Degree	0.2	0.2

Six employees are selected at random to get employee feedbacks.

a) What is the probability that at least three of the six selected will be women with degrees?

b) What will be the average number of university grads in the samples? What will be the standard deviation?

c) Once a year a special sample of 10 female employees is selected. What is the probability that at least 9 of them will have university degrees?

Q5. The number of bad particles in a cubic millimetre of air at a plant has a Poisson distribution with a mean of 4.8.

a) What is the probability that there will be more than 1 particle in a randomly selected cubic millimetre?

b) What is the standard deviation of the number of bad particles in 5 mm^3?

Q6. A random variable X has the following probability distribution

$P(x) = \binom{100}{x}(0.19)^x(0.81)^{100-x}$. Find the mean and standard deviation of X.

Q7. At Company XYZ, it is claimed that the probability of an accident during a work day is eight-tenths of one percent. There are 225 work days in a year.

a) What are the mean and standard deviation in the number of accidents per year.

b) What is the probability that more than the mean number of accidents occur in a year?

c) What is the probability of having to wait until after the first hundred work days in the year before the first accident occurs?

Q8. The quality control manager of Marilyn's Cookies is inspecting a batch of chocolate-chip cookies that has just been baked. If the production process is in control, the average number of chip parts per cookie is 6.

a) What is the probability that in any particular cookie inspected fewer than five chip parts will be found?

b) What is the probability that in any particular cookie being inspected exactly five chip parts will be found?

c) What is the probability that in any particular cookie being inspected four or five chip parts will be found?

Q9. The number of claims for missing baggage for a well known airline in a small city averages nine per day. What is the probability that, on a given day, there will be

a) fewer than three claims made?

b) exactly three claims made?

c) three or more claims made?

d) more than three claims made?

Q10. Orange Inc has tens of thousands of employees world wide. The employees are classified according to their gender (male or female), and highest education level attained (elementary, secondary, or college). The following table summarizes the results.

		Education		
		Elementary	*Secondary*	*College*
Gender	*Male*	0.19	0.14	0.11
	Female	0.225	0.25	0.085

Select 8 adults at random.

a) What is the probability that the sample includes at least 2 college grads?

b) We now select 8 females at random. What is the probability that at least 2 of the women in the sample have university degrees?

Q11. The Flames and Leafs meet in the finals – the first team to win 4 games will win the Stanley Cup. The Flames have a 60% chance of winning a game against the Leafs.

a) What is the probability that the Flames will win it in 7 games?

b) What is the probability that the Flames will win?
(Concept: Negative Binomial)

Q12. You have to inspect computer chips that were produced for the day – the factory can produce 1000 chips per day. It is known that the 1% of chips produced in the factory are defective.

Inspecting every computer chips will take too long so you decide to sample n chips such that the probability of having at least 1 defective chip is your sample is at least 95%. Find n.

Q13. There are 40 letters in the Panemian alphabet consisting of 10 vowels and 30 consonants.

a) We select 10 letters to construct a string that contains no repeated letters. What is the probability that the string contains at most 3 vowels?

b) We now select a 10 letter string where repetition is allowed. What is the probability that the string contains at most 3 vowels?

c) In a string of 10 letters, what is the probability of getting 3 vowels followed by 7 consonants if repetition is allowed?

d) In a string of 10 letters, what is the probability of getting 3 vowels followed by 7 consonants if repetition is not allowed?

Q14. Fifty employees at a branch office of Company XYZ have been categorized as follows:

	Female	Male
University Degree	20	15
No Degree	10	5

Six employees are then selected at random.

a) What is the probability that the sample includes at least two men.

b) We now select 5 men at random. What is the probability that at least two of these men have degrees?

Q15. In a shipment of 15 hard disks, 5 are defective. If 4 of the disks are inspected,

a) what is the probability that exactly 1 is defective?

b) what is the probability that at least 1 is defective?

c) What is the average number of defective hard disks that you would expect to find in the sample of 4 disk drives?

d) If we select the disks one by one, what is the probability that the first defect occurs on the 4th selection? What is the probability that the 2nd defect occurs on the 6th selection?

3. Continuous Probability Distributions

3.1 Uniform Distribution

If X takes values on some interval (a,b) with all values being equally likely, then X has a continuous Uniform Distribution.

We use the notation X~Uniform(a,b).

The probability density function is given by:

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

So the probability of X taking a value between x1 and x2 is simply:

$$P(x1 < X < x2) = (x2 - x1) \times \frac{1}{b-a}$$

$$\mu = \frac{1}{2}(a+b)$$

$$\sigma^2 = \frac{1}{12}(b-a)^2$$

3.2 Practice Questions

Q1. A couple agrees to meet for dinner at 8 pm. The woman arrives at 8:15. The time that her date arrives is continuous random variable that is uniformly distributed on the interval between 7:30 and 9:30.
a) What is the probability that the man arrives before the woman?

Solution
X ~Uniform(0,120)

$= P(0 < X < 45)$
$= 1/120 (45 - 0)$
$= 0.375$

b) What is the probability that she is kept waiting for less than 45 minutes?

Solution:
X ~Uniform(0,120)

If we don't make any assumptions about when the man arrives, the woman waits less than 45 minutes if the man arrives b/w 7:30pm to 9:00pm

P(man arriving b/w 7:30pm to 9:00pm)
$= P(0 < X < 90) = \frac{1}{120}(90 - 0) = 0.75$

c) What is the probability that she is kept waiting for less than 45 minutes (assume the man arrives after the woman)?

Solution:
X ~Uniform(0,120)
If we assume that the man arrives after the woman,

P(man arriving within 45minutes | man arrives after woman)

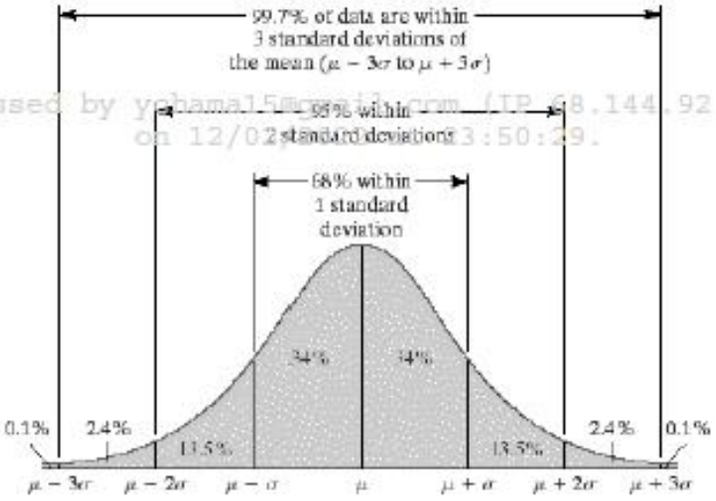
$= P(45 < X < 90 \mid X > 45)$
$= \frac{P[(45 < X < 90) \cap (X > 45)]}{P(X > 45)}$
$= \frac{P[(45 < X < 90)]}{P(X > 45)}$
$= \frac{1/120 (90 - 45)}{1/120 (120 - 45)} = \frac{45}{75} = 0.6$

3.3 Normal Distribution

The normal probability distribution is a continuous distribution that is symmetric and bell-shaped.

Properties of the Normal Density Curve:

1. It is symmetric about the mean μ
2. The area under the curve is 1.
3. **The Empirical Rule:** About 68% of the area under the graph is within one standard deviation of the mean; about 95% of the are under the graph is within two standard deviations of the mean; about 99.7% of the area under the graph is within three standard deviations of the mean.

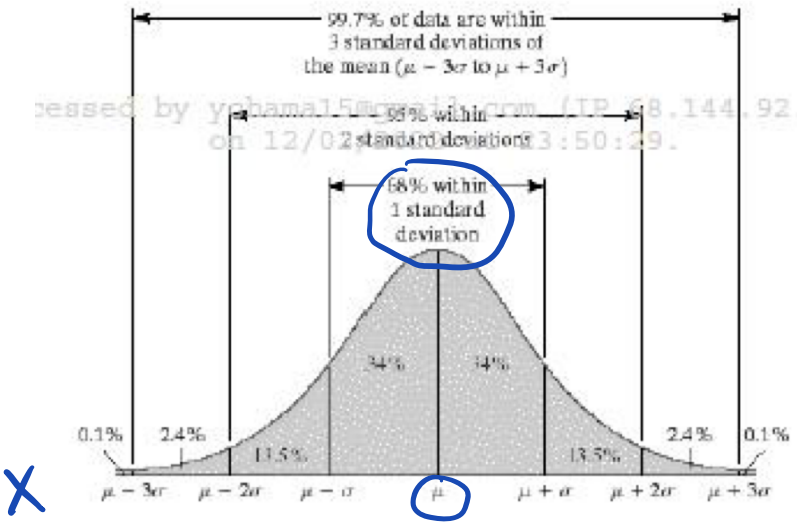


For a Random Variable $X \sim Normal(\mu, \sigma)$, the Probability Distribution Function is:

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

This eqn is rarely used in this course

- Normal distributions is continuous random variables denoted by $X \in \mathbb{R}$



- Symmetric bell shaped curve
- Normal distribution is centered @ the mean μ of your random variable X.

- The empirical rule states that 68% of observations or dat where the probability falls in one standard deviations below & above the mean

↳ what this means is : $P(\mu - \sigma < X < \mu + \sigma) \approx 68\%$.

Furthermore the empirical rule also tells us that X is 2 std above & below the mean :

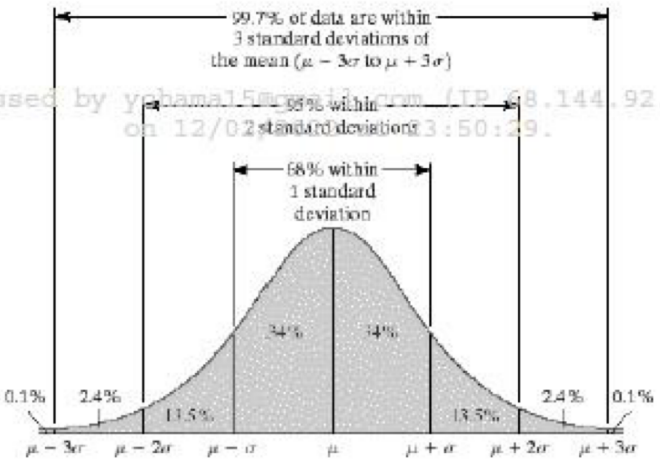
$$P(\mu - 2\sigma < X < \mu + 2\sigma) \approx 95\%.$$

If I were to give you a Random variable X, where

X: height of males @UofC

$X \sim N(176, 7)$

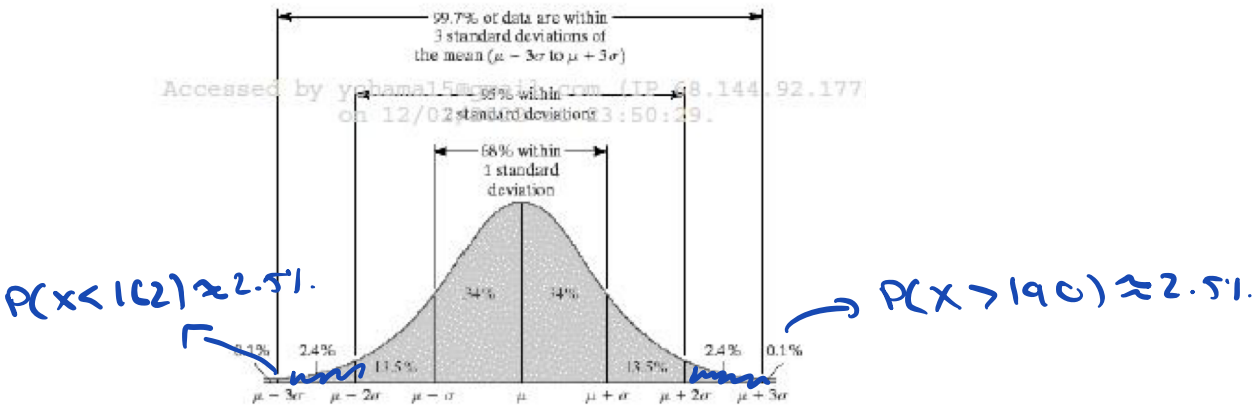
what the empirical rule says is 68% of the males will be between (169, 183) cm



162 169 183 190

using the same distribution, the empirical rule also says that 95%. fall between (162, 190)cm

Is this a sensible probability? if you walk around campus, do 2/3 of males fall between (169, 183)cm



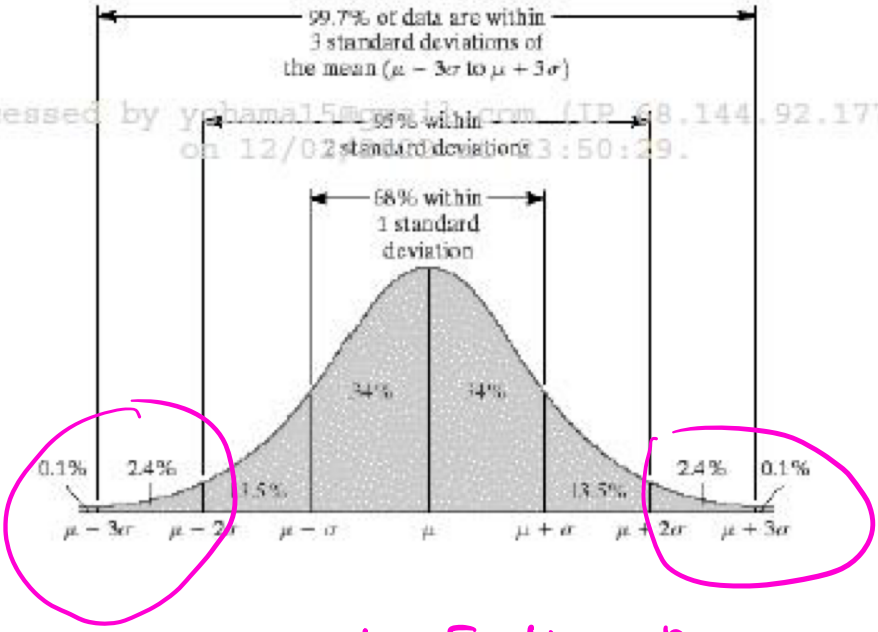
$P(x < 162) \approx 2.5\%$.

$P(X > 190) \approx 2.5\%$.

The last thing that the empirical rule states that the probability that males have a height > 190 is $\approx 2.5\%$.

- if a male has a height > 190cm is highly unlikely.
- _____ < 162cm is _____

↳ there is only a 2.5% chance of observing this.



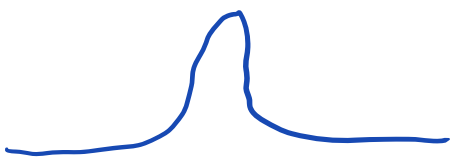
In statistics we spend a lot of time hanging around the tails of the region. The reason for that is we want to be able to make some kind of conclusion.

↳ To make a conclusion about an unknown population, we are going to use probability to assess the likelyhood of something equaling something, something being greater than something etc.

$X \sim (\mu, \sigma)$

The normal distribution is centered @ μ, and how skinny or how wide the bell shape curve is, is determined by σ.

if σ was 3 instead of 7, then the normal distribution is much skinnier



σ, or variance is a measure of dispersion in your data. dispersion being how all over the place your data is.

3.4 The Standard Normal Distribution

The standard normal distribution is the normal distribution N(0, 1) with mean $\mu = 0$ and standard deviation $\sigma = 1$.

Suppose the random variable X is normally distributed with mean μ and standard deviation σ, then we can use the standard normal distribution by finding the standardized value (called a z-score):

$$z = \frac{x - \mu}{\sigma}$$

So random variable Z is also normal and denoted as $Z \sim Normal(0,1)$ with probability distribution function given by:

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

It is important that you know the difference between *x* and *z*. The random variable *x* is sometimes called the raw score and represents value in a *non-standard* normal distribution, whereas *z* represents values in the *standard* normal distribution.

To transform a standard *z*-score to a data value *x* in a given population, use the formula (also called de-standardizing)

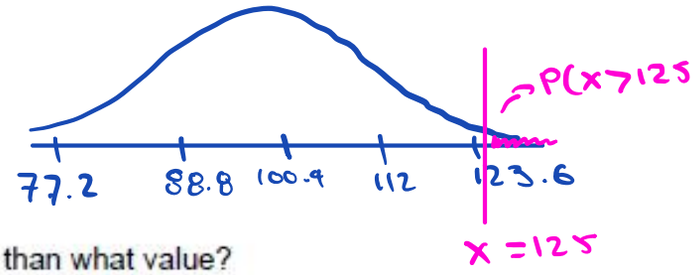
$$x = \mu + z\sigma$$

3.5 Practice Questions

Q1. The IQs of a large population of children are normally distributed with mean 100.4 (μ) and standard deviation 11.6 (σ)

a) What percentage of the children have IQs greater than 125?

$P(X > 125)$



b) About 90% of the children have IQs greater than what value?

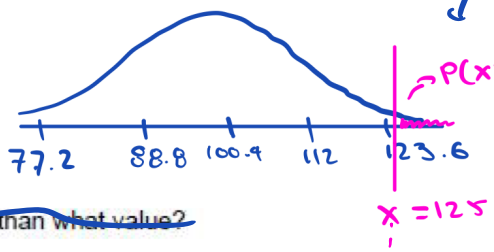
3.5 Practice Questions

Q1. The IQs of a large population of children are normally distributed with mean 100.4 (μ) and standard deviation 11.6 (σ).

a) What percentage of the children have IQs greater than 125?

$P(X > 125)$

the probability of interest



b) ~~About 90% of the children have IQs greater than what value?~~

To find this area in magenta, we need to go and transform the X into a standard normal distribution ie go look for X's sibling in the Z world 🙂 that is called the Z-score.

Z denotes a random variable

$Z \sim N(0, 1)$

125 > 2.? std above the mean.

∴ the corresponding Z must be greater than 2.

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

$$= \frac{125 - 100.4}{11.6}$$

$$= 2.12$$

An IQ of 125 is 2.12 standard deviations above the mean

Mathematically we are looking for this area here, this area is the same as the area in the X world.

$P(X > 125)$

$= P(Z > 2.12)$

$= P(Z < 2.12) = 0.9830$

$= P(Z > 2.12) = 1 - 0.9830 = 0.017$

A takeaway is an IQ of 125 corresponds to an IQ that is 2.12 std above the mean.

∴ A child that has an IQ that is >125 is a very smart child, an observation that lies either 2 std above or below the mean puts them in a very small minority, in otherwords it is a very unlikely observation. only 1.7% of children have an IQ >125

b) About 90% of the children have IQs greater than what value?

$X \sim N(1004, 11.6)$

90% of the children have an IQ > than what x

It has to be on the left tail because if x is on the right tail we don't have enough for 90% here. Only a small percentage

Here we are given a probability and are required to find an x.

we need to consult the z world and find the corresponding z value s.t this area is 90%.

Once we find the z-value we can always recover its corresponding x but doing the inv of z formula

Z-value that makes this area = 10%.

when $z = -1.28$ ensures that the shaded area is = 10%.

an IQ we are looking for is 1.28 std below the mean.

$x = \mu + z\sigma$ (inv of z-score formula)

↳ This eqn is trying to tell us how many standard deviations below the mean (it is 1.28σ below the mean)

$= 100.4 - 1.28(11.6)$

$= 85.6$

↳ 90% of children have IQ's greater than 85.6.

The key takeaway is how to interpret this eqn

$$\boxed{x = \mu + z\sigma}$$

It gives us how many standard deviations above and below the mean x is.

This x is actually 1.28 standard deviations below the mean of 100.4

Q2. The length of human pregnancies from conception to birth varies according to a distribution that is approximately normal with mean 266 days and standard deviation 16 days.

a) What percent of pregnancies last less than 240 days?

b) What percent of pregnancies last between 240 and 270 days?

c) How long do the longest 20% of pregnancies last?

$X \sim N(73, 8)$

Q3. A set of final examination grades in an introductory statistics course was found to be normally distributed, with a mean of 73 and a standard deviation of 8.

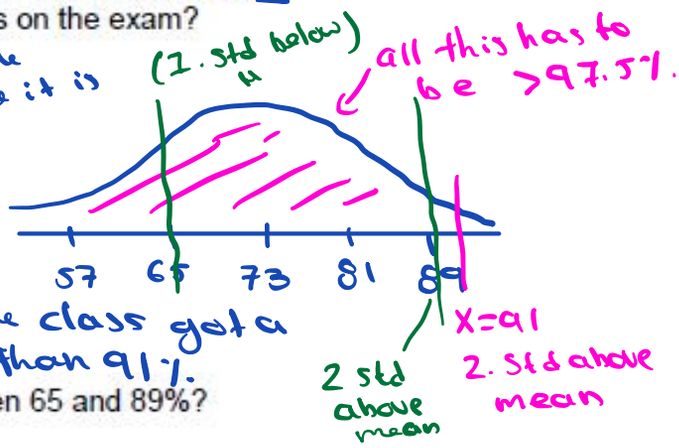
a) What is the probability of getting a 91 or less on the exam?

$P(X<91)$

$= P\left(Z < \frac{91-73}{8}\right)$

$= P(Z<2.25) = 0.9878$

as shown by the empirical rule it is $\rightarrow 97.5\%$.



$\therefore 99\%$ of the class got a mark less than 91%.

b) What percentage of students scored between 65 and 89%?

$P(65 < X < 89)$

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

65 1 std below mean

89 2 std above mean

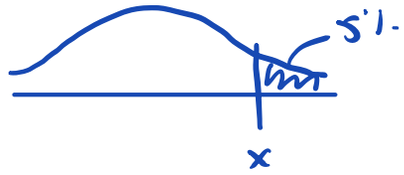
$= P(Z<2) - P(Z<-1) = 0.9772 - 0.1587$

$= 0.8185$

c) Only 5% of the students taking the test scored higher than what grade?

To get x we de-standardize it

$x = \mu + z\sigma$ (how many std above the mean is x)

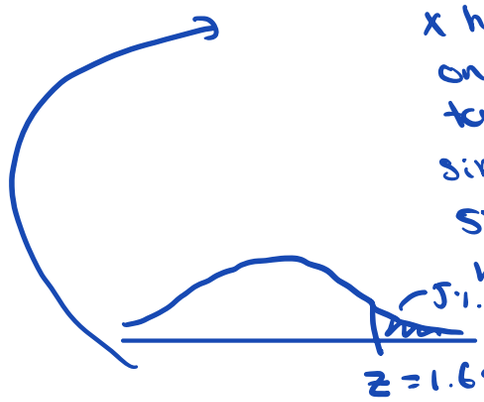


$x = \mu + 1.645\sigma$

$= 73 + 1.645(8)$

$= 86.16\%$

x has to be on the right tail since there only 5% that scored higher than x.

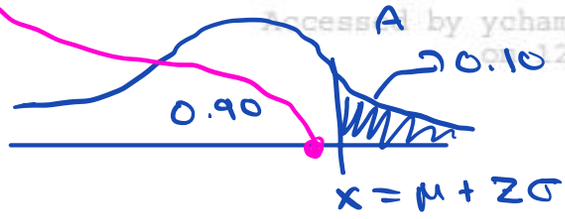


d) If the professor grades on a curve (i.e. gives A's to the top 10% of the class regardless of the score), are you better off with a grade of 81 on this exam or a grade of 68 on a different exam, where the mean is 62 and the standard deviation is 3?

let's say your mark
$x = 81 \Rightarrow$ good enough for an A?

$X \sim (73, 8)$

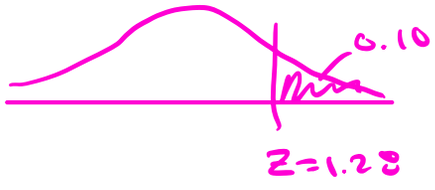
and you only get an A if you mark is the top 10% of the class



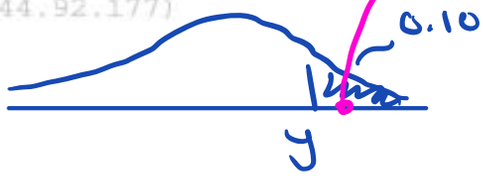
we want to find how many std above the mean x is

$= 73 + 1.28(8)$
$= 83.24$

81 falls just shy of the threshold mark to get an A.



whereas you have a friend that has a mark
$y = 68 \Rightarrow$ good enough for an A?
$Y \sim N(62, 3)$



$= \mu + z\sigma$
$= 62 + 1.28(3)$
$= 65.83$

So a mark of 68 gets an A since its above the threshold to get an A.

Although $68 < 81$ the professor marks on a curve and gives an A to the top 10%. a 68 here gives an A while an 81 doesn't.

3.6 Sampling Distribution for Mean

A **sampling distribution** is the probability distribution of a sample statistic that is formed when samples of size *n* are repeatedly taken from a population. If the sample statistic is the sample mean, then the distribution is the **sampling distribution of sample means**.

Properties of Sampling Distributions of Sample Means.

1. The mean of the sample means $\mu_{\bar{x}}$ is equal to the population mean μ.

$$\mu_{\bar{x}} = \mu$$

2. The standard deviation of the sample means $\sigma_{\bar{x}}$ is equal to the population standard deviation σ divided the square root of n.

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$$

The standard deviation of the sampling distribution of the sample means is called the **standard error of the mean.**

3.7 The Central Limit Theorem

The Central Limit Theorem forms the foundation for the inferential branch of statistics. This theorem describes the relationship between the sampling distribution of sample means and the population that the samples are taken from.

1. If the population itself is normally distributed, the sampling distribution of sample means is normally distributed for *any* sample size *n*.

Mathematically speaking,

$$\text{if } X \sim N(\mu, \sigma)$$

$$\text{then } \overline{X} \sim N(\mu, \frac{\sigma}{\sqrt{n}})$$

2. If samples of size *n*, where n>30, are drawn from any population with a mean μ and a standard deviation σ, then the sampling distribution of sample means approximates a normal distribution. The greater the sample size, the better the approximation.

In other words, regardless of whether X is normally distributed or not,

$\overline{X} \sim N(\mu, \frac{\sigma}{\sqrt{n}})$ if *n* is sufficiently large enough.

$X \sim N(176, 7.5)$

3.8 Practice Questions

Q1. The heights of adult males in Ontario follow a Normal Distribution with a mean of 176cm and a standard deviation of 7.5cm.

a) What is the probability that the next male walking into this room is greater than 183.5cm?

$P(X > 183.5)$ — 1.5 std from mean

$= P(z > 1)$

$= 1 - P(z < 1)$

$= 1 - 0.8413$

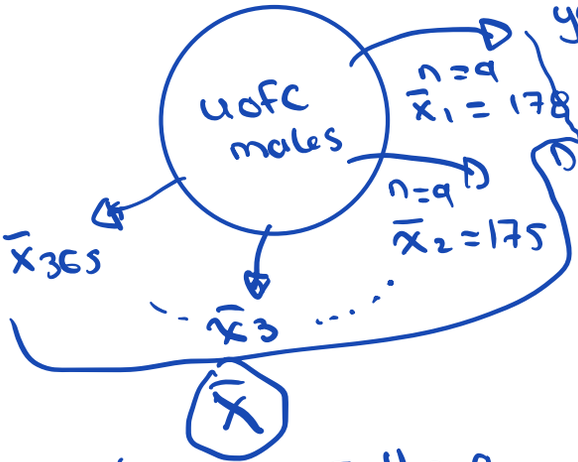
$= 0.1587$

straight forward normal distribution question

b) What is the probability that the average height of the next 9 males walking into this room will be greater than 183.5cm?

Here our random variable is no longer X, but our random variable is $\bar{X}$

Let's say this circle contains all uofc males.



suppose 9 of these guys walk in today and you measure all of them (heights) and take the average ie sample mean. ($\bar{x}$)

Do the same thing the next day. is there avg heights going to be the same as the previous day? No because these are 9 diff males. so $\bar{x}_2$ is a diff value.

if you do this over and over again The fact is your quantity $\bar{x}$ is a random variable because it varies over day, because it varies we represent all these small x's with an $\bar{X}$. Just like X represents the height of one dude that comes in your room.

All the avgs of the 9 dudes height can be represented by $\bar{X}$. Because of that $\bar{X}$ is a random variable

Now if we know X is Normal, the $\bar{X}$ is also normal

Not only will it also be normal but its mean will be the same as X's mean, but its std will be related to the standard deviation of X, it will be smaller by a factor of $\sqrt{n}$.

ie

$\bar{X} \sim \text{Normal}\left(\mu, \frac{\sigma}{\sqrt{n}}\right)$ ← The sampling distribution

↑ Std of $\bar{X}$ because $\bar{X}$ varies from sample to sample. It varies less than your X

$\bar{X} \sim \text{Normal}\left(176, \frac{7.5}{\sqrt{9}}\right)$

$= \bar{X} \sim \text{Normal}(176, 2.5)$

$P(\bar{X} > 183.5)$ ← 3 std above the mean.

$= P(Z > 3)$ Anything greater than 3 is a very unlikely proposition

$= 0.00135$

Let's think about this intuitively. 183.5 is about 6ft,

a) the probability that you are going to observe the next dude walking in being taller than 6ft is about 16%.

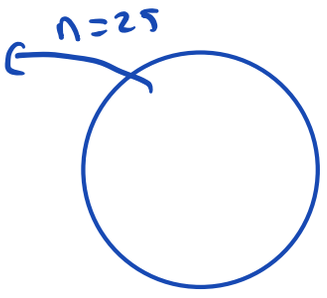
b) But if 9 guys were to walk-into this room, is the likehood that their avg height is 6ft. Is it higher or lower than 183.5? Your intuition tells you that if 9 dudes walkin to this room that their avg height > 6ft is a highly unlikely proposition.

↳ In fact the probability that 1 dudes height being > 6ft is 16%, the probability that the avg of all 9 dudes being taller than 6ft is < 1%..

c) What is the probability that the average height of the next 25 males walking into this room will be greater than 183.5cm?

25 dudes walkin wats the probability that their avg height is > 6ft.

Intution says that it is less than a and b.



$P(\bar{X} > 183.5)$ ← 5 std above the mean

$= P(Z > 5)$

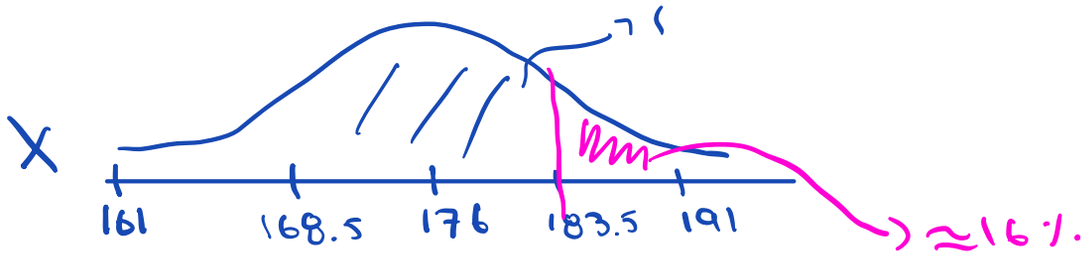
$= 0$ (5 doesn't exist on standard normal table)

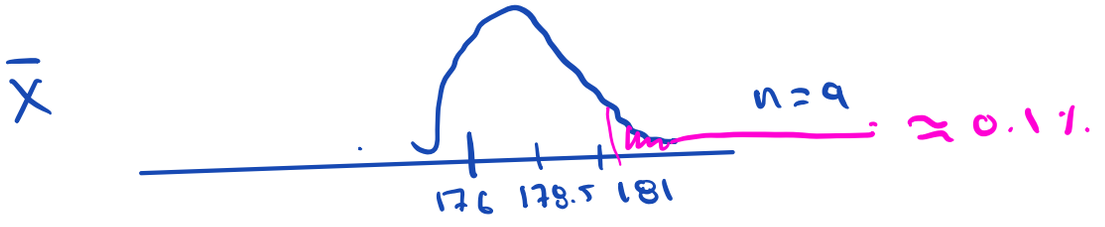
This is almost impossible.

$\bar{X}$ = avg height of the next 25 males

$\bar{X} \sim N(176, \frac{7.5}{\sqrt{25}})$

1.5





why as these get skinnier they get taller?

The area under the curve has to be 1 for that to happen they get taller as they get skinnier.

Let's appreciate that X has a normal distribution, if we sample x from this population (uofc males ×) all these males they follow a normal distribution, if you repeatedly sample dudes and compute their average, from sample to sample they will vary, if they vary we are wondering what kind of distribution do they have?

In fact is if X is normal then $\bar{X}$, they share the same mean of X but the σ will be smaller it will be related to σ of X but smaller by a factor of n.

$\bar{X} \sim N(\mu, \frac{\sigma}{\sqrt{n}})$ This is called the Sampling distribution of $\bar{X}$ (sampling mean).

To summarize if this is my population

uofc males $x \sim N(176, 7.5)$cm → what this is telling you is X follows

$n = 100$ if you were to sample 100 males and measure all there heights and draw a histogram

the histogram will also have a bell shaped curve.

less of these guys

where a lot of males will congregate around the mean.

you will see more males hanging around the mean than not.

μ can be thought of as a 2 directional magnet where the poles, pull all the males heights towards it, so it creates a mode. In your data distribution. That's why we call μ a measure of centrality.

The heights are centered around 176 cm.

A thought provoking question in reality do we really know (this value)

$X \sim N(176, 7.5)$

↳ this value (μ)

UofC has a lot of males does anyone have a true idea of what exact μ is? where μ is the avg height of all UofC males.

Do we know the avg height of all UofC males? Nobody knows μ=? because this is a huge population and we haven't measured the heights of all males @ UofC

if we expand the population to all of Calgary, do we know the true value of μ. The ans is no. Nobody knows.

But in statistics our goal is to get a better idea of the μ. And what will be a good sensible estimate for μ.

It is the sample mean $\bar{x}$.

In hypothesis testing and CI our goal is to infer about this unknown value of μ, by collecting observations $x_1, x_2, \ldots x_n$ using these observations, also known as your data, and compute $\bar{x}$ (sample mean) to get a better idea of the unknown population μ or an unknown population parameter.

- So this sample statistic $\bar{x}$ computed from your observation (from a sample collected from your population) helps us infer about an unknown value pertaining to a population. The word for this is a population parameter.

what is centeral limit theorm? (CLT)

CLT says what if X was not normal. Ex: what if heights were not normal.

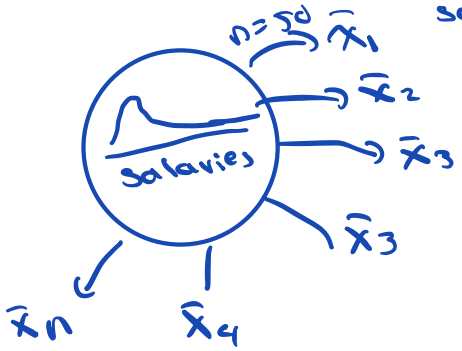
Let's talk about salaries

- are salaries normally distributed?

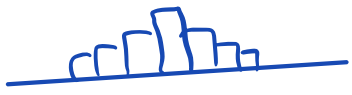
↳ salaries are right skewed.

You can't have someone whos height is 10 std above the mean but you can have someone whos salary is 10 std above the mean. (this is why they have a right skewed distribution)

what CLT says is if you were to repeadtly sample say 50 salaries and compute the $\bar{x}$ (sample mean) if you were to repeat this, what happens is the $\bar{x}$ themselves even if X is not normal or right skewed, what happens to $\bar{X} \Rightarrow$ Normal as your sample size gets bigger and bigger.



<u>summary</u>

CLT says that if you were to draw a histogram of your $\bar{x}$ they will look like  a bell shaped curve

as n gets bigger and bigger. $n > 30$ is a sufficantly large sample, you can assume $\bar{X}$ follows a normal distribution.

$\bar{X} \sim$ Normal C.L.T.

3.9 Exponential Distribution

The length of **time or the distance until the next occurrence** is often modeled using an Exponential Distribution.

Just like the Normal Distribution, the area underneath the exponential probability function will give you the probabilities associated with different values of the random variable.

The probability distribution function is given by:

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

And the following Cumulative distribution function:

$$F(x)=P(X\le x)=1-e^{-\frac{1}{\beta}x}$$

So its complement is simply:

$$P(X>x)=1-F(x)=e^{-\frac{1}{\beta}x}$$

where

x : length of time or space until the next occurrence (from any starting point)

$\beta=\frac{1}{\lambda}$: is the average length of time or space until the next occurrence

$\lambda=\frac{1}{\beta}$: is the average number of occurences per unit time or area/distance (rate)

The mean and standard deviation of the exponential distribution are:

$$\mu=\sigma=\beta$$

Also note that $\beta=\frac{1}{\lambda}$ where λ is the average number of occurrences per time/space from a Poisson distribution

Memoryless Property

$P(X<t_1+t_2\,|\,X>t_1)=P(X<t_2)$

Or

$P(X>t_1+t_2\,|\,X>t_1)=P(X>t_2)$

3.10 Practice Questions

Q1. Telephone calls arrive at the information desk of a large computer software company at the rate of 15 per hour.

a) What is the probability that the next call will arrive within 3 minutes?

b) What is the probability that the next call will arrive within 15 minutes?

e) What is the probability that the time until the next call is between 5 to 10minutes?

d) Determine the interval of time such that the probability of no phone calls occurring is 90%.

e) What is the mean time until the next call? What is the standard deviation until the next call?

f) Given that no calls arrive in the first three minutes of the day, what's the probability that it will arrive in the next 3minutes following the first 3minutes?

Q2. Autos arrive at a toll booth located at the entrance to a bridge at a rate of 50 per minute during the 5 to 6 p.m. hour.

a) If an auto has just arrived, what is the probability that the next auto will arrive within 3 seconds?

b) If an auto has just arrived, what is the probability that the next auto will arrive within 1 second?

Q3. Customers arrive at the drive-up window of a fast-food restaurant at the rate of 2 per minute during the lunch hour.

a) What is the probability that the next customer will arrive within 1 minute?

b) What is the probability that the next customer will arrive within 5 minutes?

c) During the dinner time period, the arrival rate is 1 per minute. What would be your answers to (a) and (b) for this period?

Q4. Golfers arrive at the starter's booth of a public golf course at the rate of 8 per hour during the Monday-to-Friday midweek period.

a) If a golfer has just arrived, what is the probability that the next golfer will arrive within 15 minutes?

b) If a golfer has just arrived, what is the probability that the next golfer will arrive within 3 minutes?

c) Suppose the actual arrival rate on Fridays is 15 per hour. What would be your answers to (a) and (b) on Fridays?

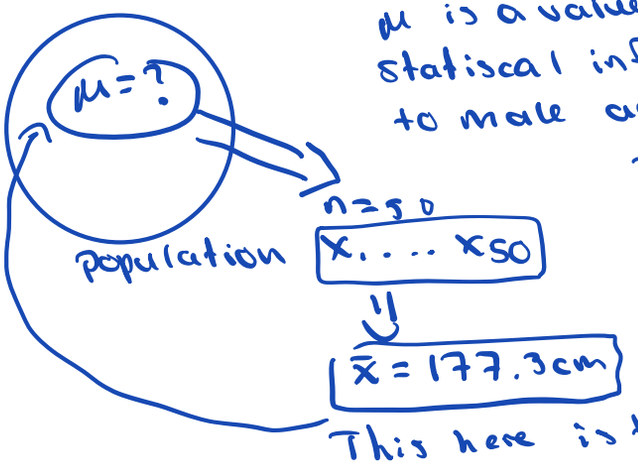
4. One-Sample Inference

This section concerns itself with the use of statistics to make inferences about population parameters, such as the population mean or population standard deviation. In this section, confidence intervals and hypothesis testing of one sample are reviewed.

4.1 Confidence Intervals for Means

When a sample is taken in order to estimate a population parameter, such as the mean or standard deviation, one can naturally expect that the sample statistic to not be the exact value of the population parameter. For this reason **confidence intervals** are used to provide a range of possible values for which we are certain (to a determined probability) that the true value of the population parameter lies within that interval. For example, a 95% confidence interval for the mean of a population would provide a range of values for which we are 95% certain that the interval contains the true population mean.

In general, a $(1-\alpha)$% confidence interval estimate is interpreted as follows: If all possible sample of the sample size n are taken and their sample means are computed $(1-\alpha)$% of the interval include the true population mean within their interval around their sample means and only α% of them do not.

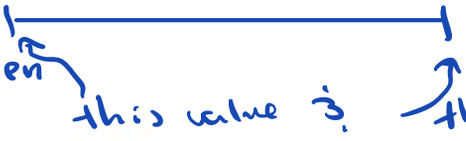


μ is a value nobody knows. Our goal in statiscal inference, In Chapt 5 & 6 we want to make an inference about μ.

To get a better understanding of μ we say sample 50 males in calgary. we measure all their heights, and w/ the data we compute $\bar{x}$.

This here is the simplest form of inference about μ. Using this we can say that the avg. height of all calgary males is 177.3cm.

But what we want to do in building CI is we want to make a statement like we are 95% confident that the value of μ



between this value & this value This is the CI

$\bar{x} = 177.3$cm

The CI is always centred @ $\bar{x}$ that you have computed from the 1 sample availble to you. the goal is to find the margin of error on both sides.

This CI is given by

$$\bar{x} \pm z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}$$

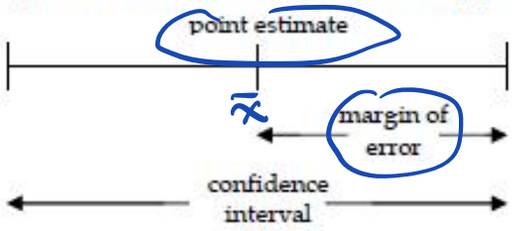
Once we compute this we can for example say that we are 95% confident that the avg height of all calgary males lies between some number & sum number.

↳ what are those 2 #'s, that is the objective of CI.

Depending on the sample from which data is collected, a different formula may be required.

General Form of a Confidence Interval

point estimate ± margin of error (E)



Some Notations

If $Z \sim N(0,1)$ then $P(Z > z_\alpha) = \alpha$

Common z_α values are:

$z_{0.05} = 1.645$

$z_{0.025} = 1.96$

$z_{0.01} = 2.326$

$z_{0.025} = 1.96$

$\bar{x}$: sample mean (point estimate of μ)

s: sample standard deviation (estimate of σ)

Confidence Interval for a Population Mean

Case 1: When σ known

$$\bar{x} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \le \mu \le \bar{x} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \qquad \bar{x} \pm z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}$$

Case 2: When σ unknown and n is large (n > 30)

$$\bar{x} - z_{\alpha/2}\frac{s}{\sqrt{n}} \le \mu \le \bar{x} + z_{\alpha/2}\frac{s}{\sqrt{n}} \qquad \bar{x} \pm z_{\alpha/2} \cdot \frac{s}{\sqrt{n}}$$

sample std computed from observations

Case 3: When σ unknown and n is not large (n < 30)

$$\bar{x} - t_{\alpha/2,(n-1)}\frac{s}{\sqrt{n}} \le \mu \le \bar{x} + t_{\alpha/2,(n-1)}\frac{s}{\sqrt{n}} \qquad \bar{x} \pm t_{\alpha/2,n-1} \cdot \frac{s}{\sqrt{n}}$$

Sample std.

variance of heights is 25 $\Leftrightarrow$ S.t.d is 5.

$\sigma^2 = 25$

variance and std capture how wide or narrow your normal dist is.

$X \sim N(\mu, \sigma)$ or $X \sim N(\mu, \sigma^2)$

The required sample size to ensure a certain margin of error (E) is:

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

One-Sided Confidence Bounds for Population Mean

$$\bar{x} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \le \mu \quad \text{or} \quad \mu \le \bar{x} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$$

4.2 Practice Questions

Q1. The manager of a paint supply store wants to estimate the actual amount of paint contained in 5-litre cans purchased from a nationally known manufacturer. It is known from the manufacturer's specifications that the standard deviation of the amount of paint is equal to 0.1 litres. A random sample of 50 cans is selected, and the sample mean amount of paint per 5-litre can is 4.975 litres.

a) Set up a 99% confidence interval estimate of the true population mean amount of paint included in 5-litre cans.

b) On the basis of your results, do you think that the manager has a right to complain to the manufacturer? Why?

4.2 Practice Questions

Q1. The manager of a paint supply store wants to estimate the actual amount of paint contained in 5-litre cans purchased from a nationally known manufacturer. It is known from the manufacturer's specifications that the standard deviation of the amount of paint is equal to 0.1 litres. A random sample of 50 cans is selected, and the sample mean amount of paint per 5-litre can is 4.975 litres.

a) Set up a 99% confidence interval estimate of the true population mean amount of paint included in 5-litre cans.

$\bar{x}$

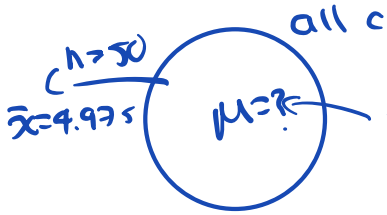
μ

$\sigma = 0.1$

$\bar{x} = 4.975$

our population of intrest

all cans of paint



the true population mean is a value that nobody knows, this including the manufacter themselves

They claim is that $\mu = 5$ (this is only a claim nobody knows this for sure).

- you don't know if this claim is true or not
- so what you did is sampled 50 cans of paint, and measured the amount of paint in those 50 cans and computed that the sample mean $\bar{x}$ is 4.975.

↳ This is our first form of inference we can say that it seems like the avg amount of paint per can is 4.975 L.

↳ But we want to be more certain and compute a 99% CI.

Soln:

σ is known $\therefore (\sigma = 0.1), n = 50, \bar{x} = 4.975$

$\bar{x} \pm Z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}$

$99 = 100(1-\alpha)$

$\therefore \alpha = 0.01$

$\frac{\alpha}{2} = 0.005.$

$Z_{\alpha/2} = 2.575$

$4.975 \pm 2.575 \left(\frac{0.1}{\sqrt{50}}\right)$

$= 4.975 \pm 0.0361$ → $(4.939, 5.011)$

so we are 99% confident that μ is between these two values

b) On the basis of your results, do you think that the manager has a right to complain to the manufacturer? Why?

if we are 99% confident that μ is between (4.939, 5.011) do we call the manfacturer?

No, because the manufacter claims that amount of paint in the can is on average 5L and 5L is in this interval.

Since, our 99% CI contains $\mu = 5$, we don't have much evidence suggesting that μ is not 5L.

if our 99% CI was (4.7, 4.75)

and we are 99% confident that μ falls in here then we claim.

we have strong evidence that $\mu \neq 5$L it is actually less than 5L.

CI is a form of statistical inferena.

Q2. You want to rent an unfurnished one-bedroom apartment for next semester. The mean monthly rent for a random sample of 10 apartments advertised in the local newspaper is $540. Assume that the standard deviation is $80.
a) Find a 95% confidence interval for the mean monthly rent for unfurnished one-bedroom apartments available for rent in this community.

b) Find a 99% confidence interval for the rent of the one-bedroom apartments. Is the 99% confidence interval wider or narrower than the 95% interval found in (a)? Explain why.

c) How large a sample of one-bedroom apartments would be needed to estimate the mean within ±$20 with 90% confidence?

Q2. You want to rent an unfurnished one-bedroom apartment for next semester. The mean monthly rent for a random sample of 10 apartments advertised in the local newspaper is $540. Assume that the standard deviation is $80. and rents are normally distributed
a) Find a 95% confidence interval for the mean monthly rent for unfurnished one-bedroom apartments available for rent in this community.

apartments.

* σ known

$\bar{x} + z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}$

$540 + (1.96) \cdot \frac{80}{\sqrt{10}}$

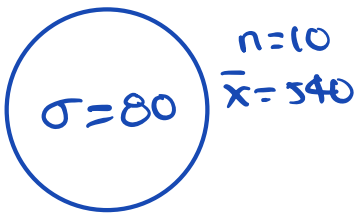
$= (490.42, 589.58)$

$95 = 100(1-\alpha)$

$\alpha = 0.05$

$\alpha/2 = 0.025$

$z_{\alpha/2} = 1.96$



is σ is known (even if $n > 30$)

b) Find a 99% confidence interval for the rent of the one-bedroom apartments. Is the 99% confidence interval wider or narrower than the 95% interval found in (a)? Explain why.

$540 \pm 2.575 \left(\frac{80}{\sqrt{10}}\right)$

$z_{0.005} = 2.575$

This will be wider than the CI in a) because $z_{\alpha/2}$ is bigger. with same amount of data available to us, the only way to up the confidence is to literally to make the CI wider.

$= (474.86, 605.14)$

c) How large a sample of one-bedroom apartments would be needed to estimate the mean within ±$20 with 90% confidence?

we want 90% CI s.t $n = ?$

what does n have to be to satisfy this eqn.

90% CI

$\alpha = 0.10$

$\frac{\alpha}{2} = 0.05$

$z_{\alpha/2} = z_{0.05} = 1.645$

$\bar{x} \pm z_{\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}}$

this is the margin of error.

and we want this to be $20.

∴ we want the 90% CI to be $20.

$1.645 \cdot \frac{80}{\sqrt{10}} < 20$

$\left[\frac{(1.645)(80)}{20}\right]^2 < n$

$\therefore n > 43.3$

$\therefore n = 44$

a sample size of 44 will ensure that our CI will be within ± 20$.

Q3. The personnel department of a large corporation wants to estimate the family dental expenses of its employees to determine the feasibility of providing a dental insurance plan. A random sample of ten employees reveals the family dental expenses for the preceding year with a mean of $261 and a sample standard deviation of $59. Assuming the data come from Normal Distribution. Set up a 90% confidence interval estimate of the mean family dental expenses for all employees of this corporation.

Q4. An advertising agency that serves a major radio station wants to estimate the mean amount of time that the station's audience spends listening to the radio on a daily basis. From past studies, the standard deviation is estimated as 45 minutes.
a) What sample size is needed if the agency wants to be 90% confident of being correct to within 5 minutes?

b) If 99% confidence is desired, what sample size is necessary?

Q3. The personnel department of a large corporation wants to estimate the family dental expenses of its employees to determine the feasibility of providing a dental insurance plan. A random sample of ten employees reveals the family dental expenses for the preceding year with a mean of $261 and a sample standard deviation of $59. Assuming the data come from Normal Distribution. Set up a 90% confidence interval estimate of the mean family dental expenses for all employees of this corporation.

$n = 10$ $(n < 30)$ ∴ σ is not given here.

$\bar{x} = 261\$$

$s = \$59$

$\bar{x} \pm t_{\alpha/2,\, n-1} \cdot \frac{s}{\sqrt{n}}$ (CI)

$90 = 100(1-\alpha)$

$\alpha = 0.1$

$\frac{\alpha}{2} = 0.05$

# of degrees of freedom is $n-1 \Rightarrow 9$.

$t_{0.05,9} = 1.833$

∴

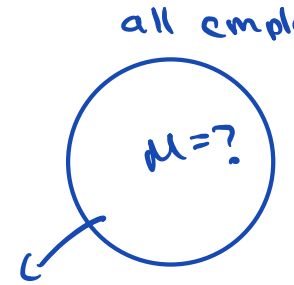
$\bar{x} \pm t_{0.05,9} \frac{59}{\sqrt{10}}$

$261 \pm (1.833) \cdot 18.65$

261 ± 34.185

$(226.815, 295.19)$

So we are 90% confident that μ (aug dental expenses) is in this range.



μ = aug family dental expenses of employees

To infer about μ what you have done is you sampled 10 employees and from those 10 employees they revealed $\bar{x} = 261$ & $s = 59\$$

$\bar{x}$ comes from 10 families

4.3 Hypothesis Testing of Means

A **hypothesis test** is a process that uses sample statistics to test a claim about the value of a population parameter.

Steps in Hypothesis Testing

1. State the null hypothesis, H_0. The null hypothesis is a statistical hypothesis that contains a statement of equality, such as $\leq$, =, or $\geq$.
2. State the alternative hypothesis, H_a. The alternative hypothesis is the complement of the null hypothesis. It is a statement that must be true if the null hypothesis is false and it contains a statement of inequality, such as >, ≠, or <.
3. Choose the level of significance, α. The level of significance is specified according to the relative importance of the risks of committing Type I and Type II errors In the problem.
4. Determine the appropriate statistical technique and corresponding test statistic.
5. Set up the critical values that divide the rejection and nonrejection regions.
6. Determine whether the test statistic has fallen into the rejection or the nonrejection region.
7. Make the statistical decision. If the test statistic falls into the nonrejection region, the null hypothesis cannot be rejected. If the test statistic falls into the rejection region, the null hypothesis is rejected.

Case 1: When σ is known, the test statistic is:

$$z = \frac{\bar{x} - \mu_o}{\sigma/\sqrt{n}}$$

Case 2: When σ is unknown but n is large (ie. n>30), the test statistic is

$$z = \frac{\bar{x} - \mu_o}{s/\sqrt{n}}$$

Case 3: When σ is unknown and n is small (ie. n<30), the test statistic is

$$t = \frac{\bar{x} - \mu_o}{s/\sqrt{n}}$$

with degrees of freedom $n - 1$ and the assumption that the sample is approximately normal.

P-Value – the probability of obtaining a sample statistic that is more extreme than the one observed from our sample.

4.4 Practice Questions

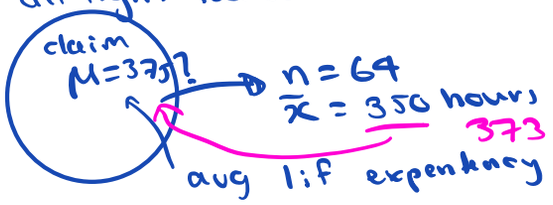
Q1. The quality-control manager at a light bulb factory needs to determine whether the average life of a large shipment of light bulbs is equal to the specified value of 375 hours. The population standard deviation is known to be 100 hours. A random sample of 64 light bulbs indicates a sample mean life of 350 hours.

a) At the 5% level of significance is there evidence that the mean life is less than 375 hours?

Let's motivate why we do HT using an example

Let's put ourselves in the shoes of this quality control manager for this light bulb factory. You are trying to determine whether the average of a large shipment of light bulbs is equal to the specified value of 375 hours.

Our population of intrest is all light bulbs



The company you work for claims that $\mu = 375$, as a quality control manager your job is to question that. Your job is to ensure that this happens and if it is not happenning your job is to catch it. Obviously nobody knows the true value of μ. In order to know the true value of μ, you basically have to burn through all of your light bulbs.

↳ what you can do is collect samples. In this case you have sampelled 64 light bulbs and you probably burnt through all 64 of them to see how long they lasted and you see that of these 64 they lasted on average 350 hours. Right off the bat this is a little bit of cause for concern. if your company claims $\mu = 375$ and you observe a sample mean of $\bar{x} = 350$ you are a little bit concerned.

you would like to see something closer to 375, like 373, 376, 380 (something in the neighborhood of 375, of the claim your company makes

Based on our pt estimate of 350 hours there seems to be a little bit of evidence that suggests that our $\mu < 375$ but the question is, is this due to random chance (maybe you got unlucky in sampling) or is there statisical evidence for us to conclude that there is something wrong in the factory that in truth μ is not equal to 375.

This is what a) is asking you to do.

is there evidence @ a 5% lvl of Significance that the mean life is less than 375 hrs.

To establish statistical Significance we conduct a Hypothesis test.

Soln'

Step(1) Form H_0 (null hypothesis)
—— H_a (alternative hypothesis)

The null hypothesis is <u>always</u> out of equality

$H_0: \mu = 375$ => nothing wrong with factor => not guilty

The alternative is a statement that we want to establish evidence for. Here its asking is there evidence that the mean life > 375.

$H_a: \mu < 375$ (is there evidence that this is happening)
↳ there's something wrong. => guilty

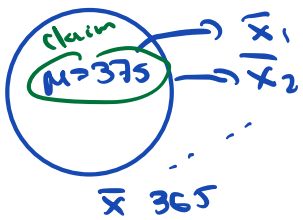
In statisical testing we want to know whether we have enough evidence to reject H_0 in favour of H_a. Or whether we have no evidence to reject H_0.

Goal: to see whether we can reject H_0. we want to see whether our evidence, observations or data supports H_a or not.

evidence: is the 64 observations, we summarized this evidence into one # $\bar{x} = 350$ (summary statistic)
↳ This is what we will use to reject H_0 and accept H_a or vice-versa. (Fail to reject)

Think of yourself as a prosecutor you are trying to establish if this evidence is strong enough to convict the factory of being guilty of producing bad light bulbs.

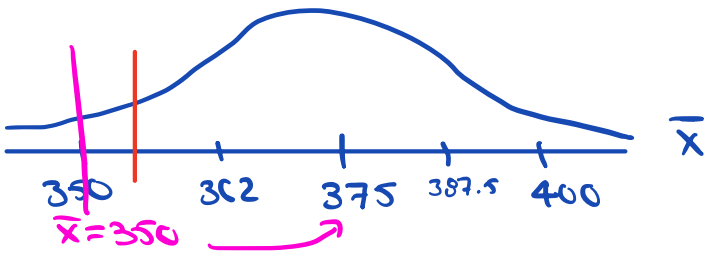
② Always assume innocence (innocent till proven guilty)
ie Assume H_0 is true $\Rightarrow$ $\mu = 375$ $\bar{x} \sim$ Normal



if it was true we would see all $\bar{x}$ congregate around 375.

ie $\bar{X} \sim N(375, \frac{100}{\sqrt{64}})$
$\bar{X} \sim N(375, 12.5)$

using 1 sample how do we establish this.



if everything was to be operating normally and μ was indeed equal to 375, and we repeatedly sample $\bar{x}$ would have this distribution.

From the 1 sample we see

what does this say about our assumption

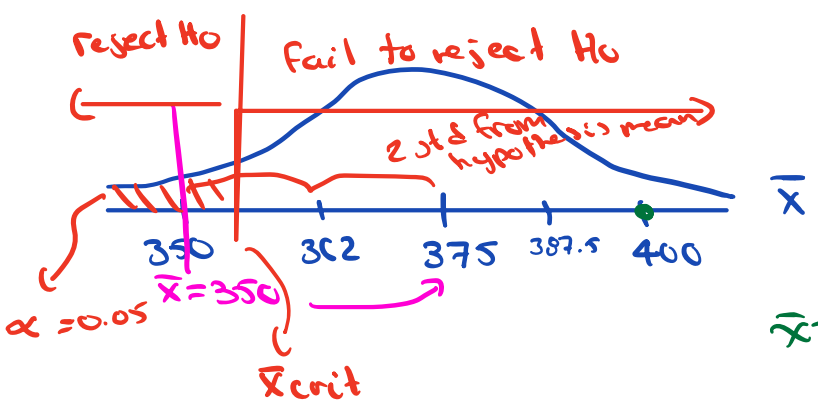
That line we drew is called the rejection region or the critical region that region is determined by our α of 5%.

stronger evidence that the factory is guilty (H_0 ↓) ⇒

What-if	guilty
$\bar{x}=200$	smaller $\bar{x}$, the stronger the evidence is in favour of H_0.
$\bar{x}=250$	
$\bar{x}=300$	
$\bar{x}=325$	
$\bar{x}=350$	
$\bar{x}=370$	

as you go down evidence is less strong

Where do we draw the line to statistically conclude the factory is guilty.



$\bar{x}=400$ evidence against H_0
fail to reject to H_0.

↳ its the value of $\bar{x}$ s.t if you observe evidence more extreme than it then we reject H_0 in favour of H_A and conclude that the factory is guilty. vs everything to the right of it, we fail to reject H_0

The question is where is $\bar{x}_{crit}$ is determined s.t the area to the left of our normal distribution is $\alpha = 0.05$ (5%.)

Well $\bar{X} \sim N$ so we can find the value of $\bar{x}$ s.t the left tail area is equal to 5%.

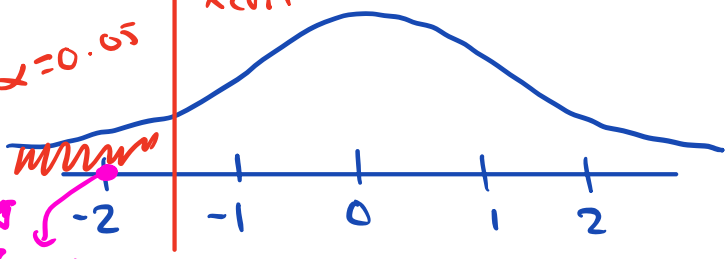
$\bar{x} = 350$

reject Ho

$\bar{x}_{crit} = 375 - 1.645(12.5) = 354.44$

← if you observe something less than this you raise an alarm

$\alpha = 0.05$



$z_{test} = z_{obs} = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}}$

$z_{crit} = -1.645$

∴ reject Ho in favour of Ha.

← how far away from what it was hypothized (the further this is the more extreme the z-score will be

$\frac{350 - 375}{12.5} = -2.$

- our observed $\bar{x} = 350$ falls in the rejection region it is more extreme than the critical value of 354.44. Using a 5% level of significance that determined this 5% critical value to make our decision had we observed 64 light bulbs that yielded a sample mean < 354.44 ($\bar{x}_{crit}$) we say we have statistical evidence to reject the null hypothesis and conclude that there is something wrong in the factory, and light bulbs are lasting less than 375 hours.

Test statistics.

Part 2 (48.14)

Let's introduce P values, how do we conduct P values, what is the meanin of the p values and how can we conduct a test using Pvalues.

Goal: to see whether we can reject Ho.

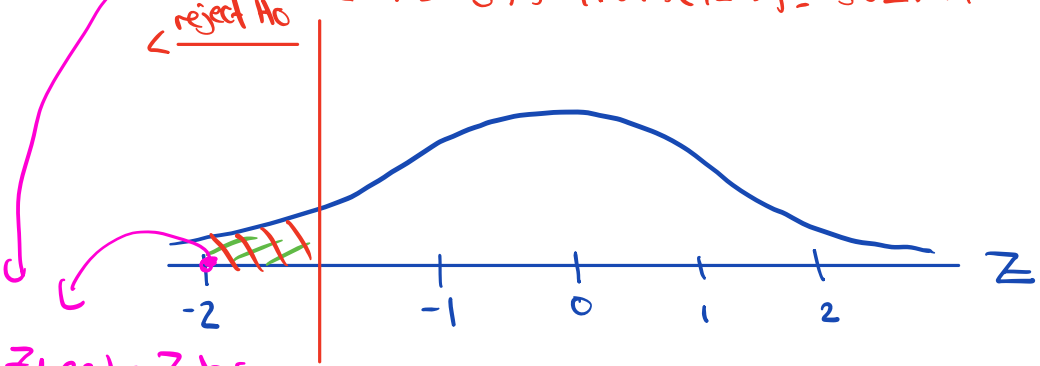
$H_0: \mu = 375 \Rightarrow$ nothing wrong $\Rightarrow$ not guilty

$H_a: \mu < 375 \Rightarrow$ something wrong $\Rightarrow$ guilty

1) Assume Ho is true $\Rightarrow$ $\mu = 375$ ← Assume the factory is working properly.

$\bar{X} \sim N(375, \frac{100}{\sqrt{64}})$

$\bar{x}_{crit} = 375 - 1.645(12.5) = 352.44$



$Z_{test} = Z_{obs}$

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

$= \frac{350 - 375}{12.5} = -2$

$Z_{crit} = -1.645$

what-if

$\bar{x} = 200$ guilty

$\bar{x} = 250$

$\bar{x} = 300$

$\bar{x} = 325$

$\bar{x} = 350$

$\bar{x} = 370$

is 350 small enough to statistically reject what we have assumed to be true under the null hypothesis. and conclude that there is something wrong w/ the factory. (Machines in the factory are producing light bulbs that aren't up to specifications.

To figure this out we needed a decision line. (critical value)

we know that whatever happens to $\bar{x}$ has a corresponding parallel universe in z.

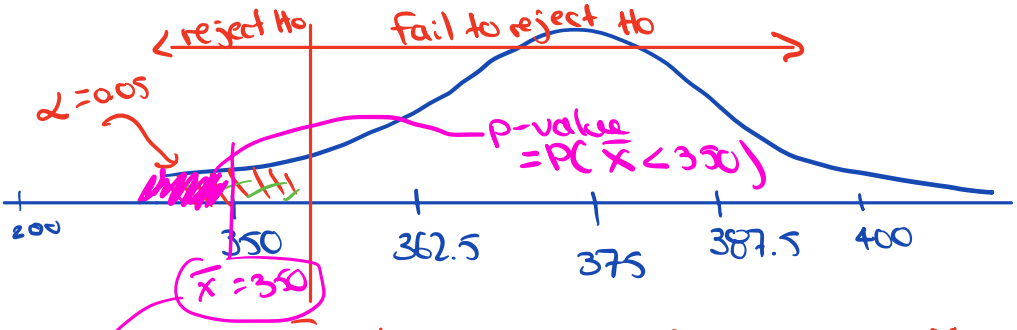
The way we found the critical region was through α (significance lvl) in this case 5%.

The critical region in the z-world is easy to find we have a table where we can look this up in our standard normal table. In the z world the critical z-value is determined by $\alpha = 5\%$.

Then we determined our test statistic corresponding to our evidence of 350 (z_{test} or z_{obs}).

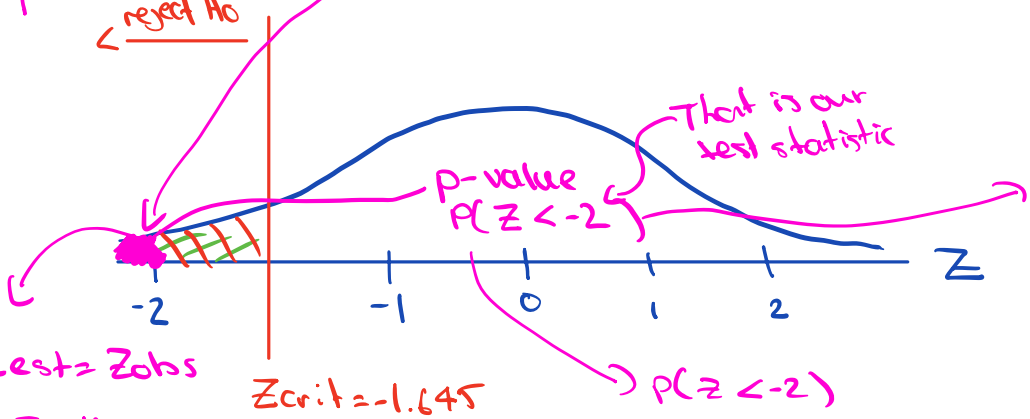
- we saw that a z test of -2 fell in the rejection region.

The p-value in this question is the probability of observing something even more extreme than the one we observed, in the one sample available to us.



$\bar{x}_{crit} = 375 - 1.645(12.5) = 352.44$

When we say more extreme it is the probability of observing a piece of evidence that is more in favour of the alternate hypothesis (H_a) than the one we have already observed.



This probability we can look up on our standard normal table.

test = Zobs

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

$\frac{350 - 375}{12.5} = -2$

$Z_{crit} = -1.645$

$P(Z < -2) = 0.0228$

The pink area is less than the red shaded region. The red shaded region on the left tail of this normal distribution is our α of 5%. The pink area is our p-value

whenever p-value $< \alpha$ (significance value), then we reject H_0

You can do the test using the p-value method without ever needing to compute or look up the critical value. You can just look up the p-value and compare it against α; if its less than α, then we reject H_0.
Ie Null has to go.

Lets go back to our what if scenario, and we observed an $\bar{x}=200$ how would that p-value associated with an $\bar{x}=200$ differ from the p-value we just found?

Ans: Our p-value would significantly decrease

A p-value for $\bar{x}=250$ would be bigger than the p-value associated with $\bar{x}=200$. P-value associated with $\bar{x}=300$ would be bigger than a p-value associated with $\bar{x}=250$

what the p-value does, is it helps us quantify how strong our evidence is. The smaller the p-value is the stronger our evidence is. The stronger our evidence is (i.e. $\bar{x}=200$, very strong evidence against the factory working up to spec)

what the p-value does is it not only helps you reject H_0, but it actually helps you quantify how strong your piece of evidence is.

$\bar{x} = 350$

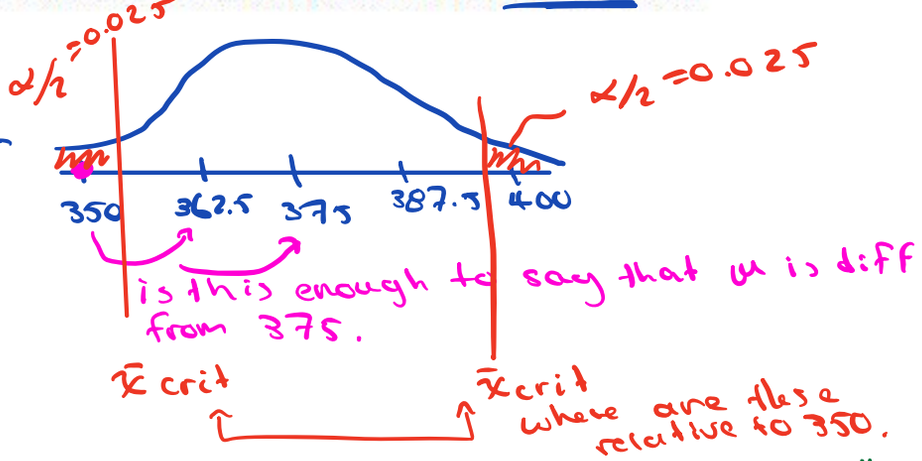
b) At the 5% level of significance is there evidence that the mean life is different from 375 hours?

$H_0: \mu = 375$

$H_a: \mu \neq 375$

① Assume H_0 is true

if H_0 is true we get this distribution



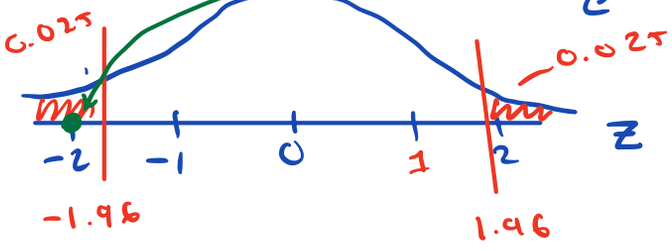
whenever we have ≠ this is called a 2 tail test rejection on both tails of dist

$\bar{x}_{obs} = 350$

$z_{obs} = z_{test} = \frac{\bar{x}_{obs} - \mu_0}{\sigma/\sqrt{n}} = -2$

our test statistic lies in the rejection region

c) Set up a 99% confidence interval estimate of the population mean life of the light bulbs.



∴ we reject H_0 in favour of H_a.

There's evidence showing that μ is not equal 375.

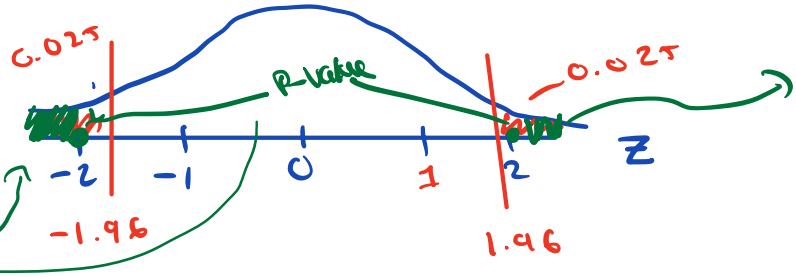
Part 2 (1:06:33)

Now lets compute the p-value for part b)

Because we have $H_a \neq$ this indicates that we will use a 2 tail test

- If you observe an $\bar{x} = 600$ than you have extreme evidence against the null hypothesis in favour of the alternate hypothesis.
- like wise if we observe evidence that $\bar{x} = 200$, again we have extreme evidence against the null hypothesis in favour of the alternate

It helps to write these what if scenarios and think through intuitively even b4 you do the test whether its something that is extreme enough for you to reject the null in favour of the alternate.



everything to the right of the 2 as well.

The p-value here is the probability of observing something more extreme than the one we observed in our one sample. which is where the arrow is showing. In the case of a 2 tailed test we also take its mirror image.

Together this area will be our p-value
(ie $2 \cdot P(Z < -2)$)

due to symmetry these areas are equal.

$2 \cdot 0.0228 = 0.0456 = \text{P-value}$

Since our p-value < 0.05
we reject H_0.

c) Set up a 95% confidence interval estimate of the population mean life of the light bulbs.

$$\bar{x} \pm z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}$$

$$350 \pm 1.96 \cdot (12.5)$$

$$= 350 \pm 24.5 = (325.5, 374.5).$$

we are 95% that μ is in there

Q2. A manufacturer of salad dressings uses machines to dispense liquid ingredients into bottles that move along a filling line. The machine that dispenses dressings is working properly when 250 millilitres are dispensed. The manufacturer is particularly concerned that the bottles do not have an average of less than 250 millilitres. Assume we know the population standard deviation of the process to be 4 millilitres. → σ (std)

A sample of 16 bottles is selected periodically, and the filling line is stopped if there is evidence that the mean amount dispensed is actually more than 250 millilitres.

On a particular sample of 16, we observe a sample mean of 252 ml. ~~Suppose that an average amount dispensed in a particular sample of 50 bottles is 249.5 millilitres.~~

At the 5% level of significance, using is there evidence that the mean amount dispensed is less than 250 millilitres?

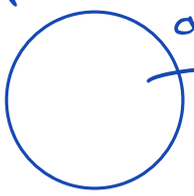
n

$\bar{x}$ (related to the 16 bottles)

more (change)

(is there evidence that the mean dispensed amount ($\mu < 250$)

Population



all bottles of dressing

$n = 16$

$\bar{x} = 252$mL

In hypothesis testing whatever you want to show evidence for, always goes in the alternate hypothesis.

The null hypothesis is either always out of equality or the complement of the alternate hypothesis is.

so here $H_a: \mu < 250$:

$H_0: \mu = 250$ or $\mu \geq 250$ } ⟹ if your H_0 is simply that of equality or what you can do is say $H_0: \mu \geq 250$ (either ans is fine).

we want to see if we have evidence in our data of 16 bottles that supports the alternative hypothesis.

Before we conduct any statistical test if you got an $\bar{x} = 252$mL does that support the alternate hypothesis (H_a)

Is this strong enough evidence to conclude that there is something wrong in the factory.

what if

much stronger evidence in favour of the H_a than the $\bar{x} = 252$ml than we have observed → $\bar{x} = 220$mL, $\bar{x} = 230$mL

Lets conduct a statisical test to identify the critical region.

1) Lets go to the z- world and compute the test statistic.

$$Z_{obs} = Z_{test} = \frac{\bar{x}_{obs} - \mu_0}{\sigma/\sqrt{n}} \Rightarrow \sigma_{\bar{x}} \text{ (also denoted by)}$$

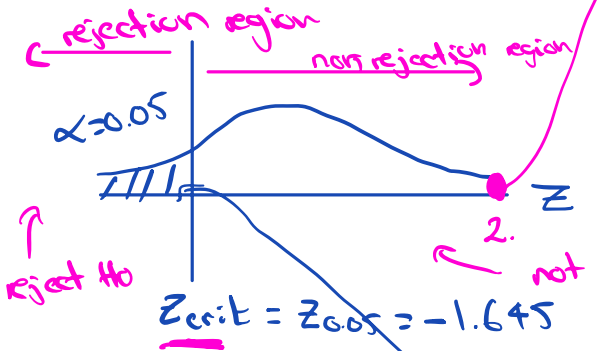
($\bar{x}_{obs}$ ← the one observed from our sample; μ_0 ← (The hypothesized mean from Ho))

$x \sim N$, $\bar{x} \sim N$ } Central Limit Theorem

They share the same mean but their standard deviation, are related. The standard deviation of $\bar{x}$ is related to the standard deviation of x by σ but is smaller by a factor of $\sqrt{n}$.

$$Z_{test} = \frac{252 - 250}{4/\sqrt{16}} = 2$$

2) The next thing we are going to do is identify our critical region in the z- world.



$Z_{crit} = Z_{0.05} = -1.645$

The rejection region as a rule of thumb always point in the direction of whatever that is in your alternate hypothesis

$H_a: \mu < 250$

here if you look @ this less than as an arrow our rejection region will be on the left side of the standard normal distribution

The rejection region is everything to the left of our critical value however our test statistic lies on the opposite side in the non-rejection side

Since we are testing @ a 5% critical value our $Z_{crit.}$ is determined s.t this left area is equal to our α of 5%. (i.e $\alpha = 0.05$)

Is this a 1 tailed test or 2 tailed test?

- This is a 1 tailed test. because we have a less than in our alternate hypothesis.

∴ there is insufficient evidence against the Ho. There is insufficient evidence suggesting that the mean amount dispensed is actually less than 250 mL.

* Since σ is given the size of n doesn't matter so we can use the z-statistic.
* if σ is not given and n is small we have to use the t test statistic.

Z-test (test statistic) helps us do the test (compute this).

Zcrit (z critical value) helps us determine the threshold rejection region and the non-rejection region. (look this up)

In this question if σ was unknown, then what we need to do is use the sample standard deviation (s). Then instead of using the z-statistic we would use the t test statistic.

- if σ is known use the z-tables (rule of thumb)
- if σ is unknown and you have an s and your sample size is small (ie. $n < 30$) then you have to use the t test static.

Let's do this question by computing the p-value.

Why do we bother with the p-value?

- There's two ways of doing the test 1 way is by doing the p test and the other way is using the critical value test.
- But what a p-value can help us with that a critical doesn't is that it helps us quantify how strong our evidence is against the null hypothesis in favour of the alternate.
  ↳ quantify the strength of our evidence.
- Both methods will yield the same conclusion.
  ↳ if they don't you made a mistake somewhere.

Solving the question again First due to a typo.

if there's more than 250ml of salad dressing this is money lost for the salad dressing company.

what we want to show is there's evidence

$H_a: \mu > 250$

$H_0: \mu = 250$ or $\mu \leq 250$

(The null is always out of equality or its complement).

MC: a) $H_a = \ldots$ $H_0 = \ldots$

b)

c)

d)

choose the right one.

$n = 16 \Rightarrow \bar{x} = 252\,mL$

we only have this evidence, with this evidence can we statistically conclude that there is something wrong with the factory.

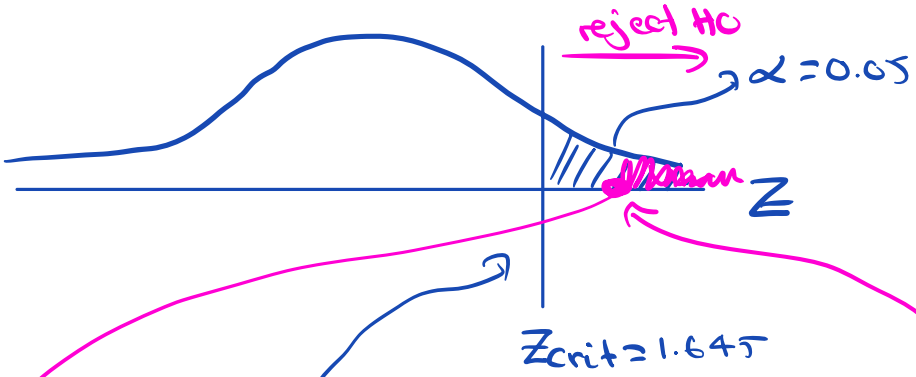
your first reaction is, it does seem to support H_a. It does seem to agree w/ the alternate hypothesis.

The question is, is it extreme enough, is it big enough to statistically conclude that there is something wrong with the factory.

what if

stronger evidence than 252:
- $\bar{x} = 260\,mL$
- $\bar{x} = 280\,mL$
- $\bar{x} = 350\,mL$

dressing probably spilled all over the factory.



The critical region always points in the direction of the arrow: $H_a > 250$

arrow

The threshold value that will help us determine whether we reject or fail to reject.

$$z_{test} = z_{obs} = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}} = \frac{252 - 250}{4/\sqrt{16}} = 2$$

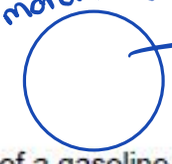
what's hypothesized to be true

Now that we found our critical region we now want to find and compute our test statistic and we want to see whether our test statistic falls in the rejection region or not

The p-value here is everything that is more extreme than the test statistic we observed.

$$p\text{-value} = P(z > 2) = 0.0228 < \alpha \quad (\alpha = 0.05)$$

$\therefore$ reject H_0.



(sample 60 of them and computed a sample mean and sample standard deviation.

Q3. The owner of a gasoline station wants to study gasoline purchasing habits by motorists at his station. A random sample of 60 motorists during a certain week is selected with the following results:

- Amount purchased: $\bar{x} = 42.8$ litres, $s = 6.7$ litres

n=60 (sample size)

At the 5% level of significance, is there evidence that the mean purchase is different from 40 litres?

$\bar{x} = 42.8$
$s = 6.7$
$n = 60$

off the bat we know: $n > 30$ σ is unknown but we are given s.
- what we want to show evidence for always goes in the alternate hypothesis.

$H_0 = \mu = 40$
$H_a: \mu \neq 40$

is this a one tailed test or a 2 tailed test?
This is a 2 tailed test because we have a $\neq$, which means that we will have 2 rejection areas. (one is on the left and one is on the right)

↳ what this means

There are 2 areas where you reject here. If you get a really small $\bar{x}$ you end up rejecting and if you get a really large $\bar{x}$ you also end up rejecting.

what-if
$\bar{x} = 20\ell$
Strong evidence against the null and in favour of the alternate

$\bar{x} = 65\ell$
This also constitutes strong evidence against the null in favour of the alternate.

S: Sample st deviation
σ: population std dev.
} They both capture the same thing.

if you square these they become the sample variance & population variance.

Soln: step 1 we compute the test statistic.

Test statistic:

There is 2 forms of the test statistic:

$Z_{test} =$ $t_{test} =$

In this case because our sample size is big (ie $n > 30$) even though σ is unknown we use the Z_{test} because our sample size is sufficiently big enough.

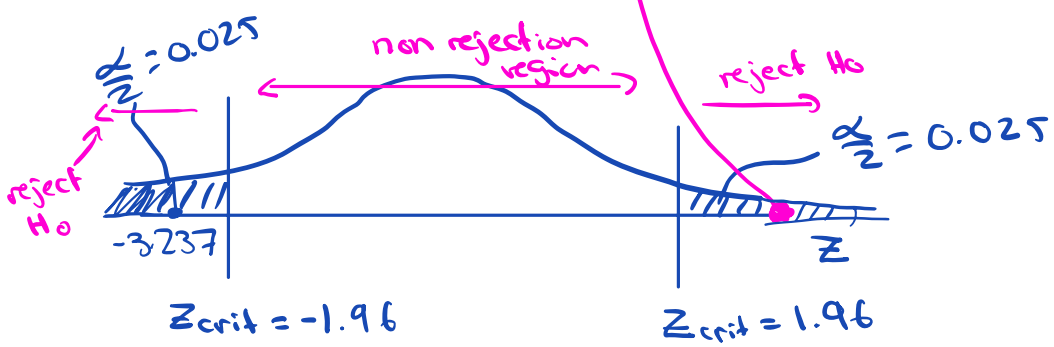
$$Z_{test} = \frac{\bar{x} - \mu_0}{s/\sqrt{n}}$$

we use s because σ is not available to us.

$$= \frac{42.8 - 40}{6.7/\sqrt{60}} = 3.237 \text{ (test statistic)}$$

2) Lets identify the critical region

2 areas of rejection



Z_{test} is deep in the rejection region

$\therefore$ Reject H_0 in favour of H_a

p-value is the area to the right of our test statistic

p-value $= 2P(Z > 3.237)$ in a 2 tail test

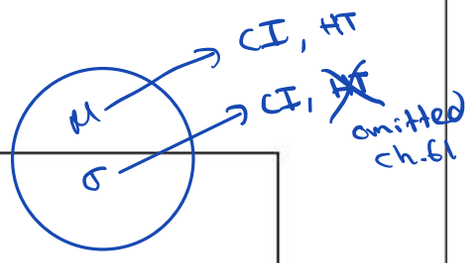
Q4. An auditor for an insurance company that provides health-care plans is assigned the task of evaluating reimbursement for office visits to chiropractors. The audit is to be conducted for all such payments in a particular geographic area during a certain month. Suppose the audit is conducted on a sample of 75 of the reimbursements, with the following result:

- The mean amount of reimbursement was $93.70 and the standard deviation was $34.55

At the 5% level of significance, is there evidence that the average reimbursement is less than $100?

Q5. A company that makes cola drinks states that the mean caffeine content per one 12-ounce bottle of cola is 40 mg. You work as a quality control manager and are asked to test this claim. During your tests, you find that a random sample of thirty 12-ounce bottles of cola has a mean caffeine content of 39.2 mg with a standard deviation of 7.5 mg. At a level of significance of 1%, can you reject the company's claim?

4.5 Inference for Variance/Standard Deviation



The test statistic for testing the hypothesis

Case I: Two Tail Test

Hypothesis

$$H_0 : \sigma^2 = \sigma_0^2$$
$$H_a : \sigma^2 \neq \sigma_0^2$$

not needed

Test Statistic

$$\chi_0^2 = \frac{(n-1)s^2}{\sigma_0^2} \sim \chi_{n-1}^2$$

Rejection Region

$$\chi_0^2 > \chi_{\alpha/2,n-1}^2 \quad \text{or} \quad \chi_0^2 < \chi_{1-\alpha/2,n-1}^2$$

Case II: One Tail Test

Hypothesis

$$H_0 : \sigma^2 = \sigma_0^2 \quad \text{or} \quad H_0 : \sigma^2 = \sigma_0^2$$
$$H_0 : \sigma^2 > \sigma_0^2 \quad \text{or} \quad H_0 : \sigma^2 < \sigma_0^2$$

not needed

Test Statistic

$$\chi_0^2 = \frac{(n-1)s^2}{\sigma_0^2} \sim \chi_{n-1}^2$$

Rejection Region

$$\chi_0^2 > \chi_{\alpha,n-1}^2$$
$$\chi_0^2 < \chi_{1-\alpha,n-1}^2$$

Confidence Interval for Variance

$$\frac{(n-1)s^2}{\chi_{\alpha/2n-1}^2} \leq \sigma^2 \leq \frac{(n-1)s^2}{\chi_{1-\alpha/2n-1}^2}$$

Ch 5.8

↳ our goal is to say we are 95% confident that σ^2 is between some number and some number.

$? < \sigma^2 < ?$

we are trying to do an inference on σ

4.6 Practice Questions

no need to do

Q1. A random sample of 20 bottles were selected from a salad dressing filling line the sample standard deviation of the fill volume was calculated to be $s^2 = 0.0169(mL)^2$. If the variance of fill volume exceeds $0.01(mL)^2$, too many bottles will be either underfilled or overfilled. Based on our sample, is there evidence of too much under/overfilling in the manufacturing line? Test at 5% significance level.

Q2. In a paper plant, the weight for 15 sheets of paper were measured. Their sample standard deviation was found to be 0.0084g.

a) Does the standard deviation differ from 0.01 at $\alpha = 0.05$? Find the p-value of the test.

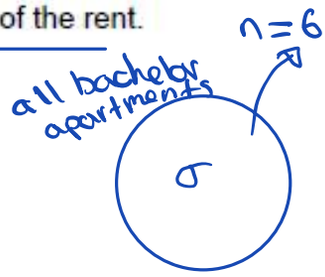
b) Construct a 95% confidence interval for σ. What do you notice?

σ σ^2

standard deviation and variance both measure the same thing, ie how much dispersion is in your data. Measures the all over the placeness of your data.

Q3. You are looking for a bachelor apartment and see 6 listings on Craigslist. The rents are $700, $690, $750, $800, $900 and $500.
Find the 99% confidence interval for the standard deviation of the rent.

99% CI for σ



The equation of intrest for σ

$$\frac{(n-1)s^2}{\chi^2_{n-1,\alpha/2}} < \sigma^2 < \frac{(n-1)s^2}{\chi^2_{n-1,1-\alpha/2}}$$

$\chi^2_{n-1,1-\alpha/2}$ ↳ (χ^2 value)

$\bar{x} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \mu < \bar{x} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$ analogus

These are values we look up on the χ^2 distribution table. We just need to understand how to look it up.

Here we have our n but we don't have our sampled standard deviation and variance.

So here we actually have to go and and compute s

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

sum of all the values of x and how they deviate from $\bar{x}$. Square that and divide it by n-1.

$$= \sqrt{\frac{\Sigma x^2 - \frac{(\Sigma x)^2}{n}}{n-1}} = 133.67$$

when you compute a C.I for σ^2 the only distribution table of intrest is the χ^2 (χ^2) table. so t distribution does not apply when you do inference on σ or σ^2. when you build a C.I for μ however this depends on your sample size. when doing this for σ^2 you don't care about your sample size.

here we are interested in how all over the place the rent for a bachelor apartment is? Usually bigger cities have more variance

$$\frac{(n-1)s^2}{\chi^2_{n-1,\alpha/2}} < \sigma^2 < \frac{(n-1)s^2}{\chi^2_{n-1,1-\alpha}} \qquad \alpha = 0.01 \qquad \frac{\alpha}{2} = 0.005$$

$\chi^2_{5,\,0.005} = 16.750$ $\qquad$ $\chi^2_{5,\,0.995} = 0.4$

5 degrees of freedom

This can be found in table A.7.

5 degrees of freedom gives us the row of interest

The 1st row of the χ^2 (chi^2) distribution table. gives us the right tail of the ki^2 distribution.

The CI for the variance is

$$\frac{5(133.67)^2}{16.75} < \sigma^2 < \frac{5(133.67)^2}{0.41}$$

$$5,333.33 < \sigma^2 < 217,886.18 \ (\$)^{\text{units}}$$

$$73.03 < \sigma < 466.8 \ (\$)^{\text{units}}$$

what does the answer? (σ) signify interms of rent?

- Based on the 6 data points we collected we want to understand, how much dispersion there is in our rent.

$\sigma \uparrow$ a high sigma $\Rightarrow$ that rent is all over the place.

- we are 99% confident that sigma is a number between 73$ an 466$

For new york we might see

$$600 < \sigma < 1000$$

we are 99% confident that σ is a much higher # than the city of calgary.

4.7 Types of Errors

No matter which hypothesis represents the claim, you always begin a hypothesis test by assuming that the equality condition in the null hypothesis is true. So, when you perform a hypothesis test, you make one of two decisions:

1. reject the null hypothesis or
2. fail to reject the null hypothesis

Because your decision is based on a sample rather than the entire population, there is always the possibility you will make the wrong decision.

There are two possible errors that may result:

1. A **type I error** occurs if the null hypothesis is rejected in spite of it being true.
2. A **type II error** occurs if the null hypothesis is not rejected (accepted) in spite of it being false.

The following table shows the four possible outcomes of a hypothesis test.

	Truth of H_0	
Decision	**H_0 is true**	**H_0 is false**
Do not reject H_0	Correct decision	Type II error
Reject H_0	Type I error	Correct decision

In a hypothesis test, the **level of significance** α is your maximum allowable probability of making a type I error.

The probability of a type II error is denoted by β.

Computing β for a one tail test.

The following is a case for a left tail test where

$H_0 : \mu = \mu_0$

$H_1 : \mu < \mu_0$

You MUST be given the true value of μ, call it μ_{truth}

Identify the critical $\bar{x}$, call it $\bar{x}_{critical}$

Identify region where we fail to reject H_0 (ie. $\bar{X} > \bar{x}_{critical}$)

Method 1

$$P(Type\ II\ error)$$
$$= P(fail\ to\ reject\ a\ false\ H_0)$$
$$= P(\bar{X} > \bar{x}_{critical} \mid \mu = \mu_{true})$$
$$= P(Z > \frac{\bar{x} - \mu_{true}}{\sigma/\sqrt{n}})$$

Method 2

$\beta = P\left(Z < z_\alpha - \frac{\delta}{\sigma/\sqrt{n}}\right)$ **where** $\delta = \mu_{true} - \mu_0$ **and** $z_\alpha = z_{critical}$

Computing β for a two tail test.

The following is a case for a two tail test where

$H_0 : \mu = \mu_0$

$H_1 : \mu \neq \mu_0$

You MUST be given the true value of μ, call it μ_{truth}

Identify the critical $\bar{x}$'s (for a two tail test, there will be two such values), call them $\bar{x}_{left-crit}$ and $\bar{x}_{right-crit}$

Identify region where we fail to reject H_0 (ie. $\bar{x}_{left-crit} < \bar{X} < \bar{x}_{right-crit}$)

Method 1

$$P(Type\ II\ error)$$
$$= P(fail\ to\ reject\ a\ false\ H_0)$$
$$= P(\bar{x}_{left-crit} < \bar{X} < \bar{x}_{right-crit} \mid \mu = \mu_{true})$$
$$= P\left(\frac{\bar{x}_{left-crit} - \mu_{true}}{\sigma/\sqrt{n}} < Z < \frac{\bar{x}_{right-crit} - \mu_{true}}{\sigma/\sqrt{n}}\right)$$

Method 2

$\beta = P\left(Z < z_{\alpha/2} - \frac{\delta}{\sigma/\sqrt{n}}\right) - P\left(Z < -z_{\alpha/2} - \frac{\delta}{\sigma/\sqrt{n}}\right)$ where $\delta = \mu_{true} - \mu_0$ and $z_{z_{\alpha/2}}$ are the two critical z-scores

The **power of a test** is the probability that a fixed level significance test will reject the null hypothesis when a particular alternative value of the alternative is true. The power is equal to $(1 - \beta)$ for the particular alternative considered.

To calculate the power of significance test consists of three steps:

1. State H_o, H_a, the particular alternative we want to detect, and the significance level α.
2. Find the values of $\bar{x}$ that will lead us to reject H_o.
3. Calculate the probability of observing these values of $\bar{x}$ when the alternative is true.

Beta (β) decreases or Power (1-β) increases when the following happens

1) alpha (α) increases
2) $|\mu_{true} - \mu_o|$ increases
3) Sample size (n) increases
4) Standard deviation (σ) decreases

Sample Size Formulas

We can determine the sample size needed to obtain a particular value of β for a given value of δ and α.

Sample Size for a Two-Sided Test on the Mean, Variance Known

$$n = \frac{(z_{\alpha/2} + z_\beta)^2 \sigma^2}{\delta^2} \quad where \quad \delta = \mu - \mu_O$$

Sample Size for a One-Sided Test on the Mean, Variance Known

$$n = \frac{(z_\alpha + z_\beta)^2 \sigma^2}{\delta^2} \quad where \quad \delta = \mu - \mu_O$$

textbook 6.13

4.8 Practice Questions

Q1. Let's go back to Question Q1 part (a) in the previous section about light bulbs. The quality-control manager was interested in whether the average life of light bulbs 375 or not. So he had collected a sample of 64 light bulbs.

a) i) At significance level of 5%, what is the critical z value?
ii) For what values of sample means would we reject the null hypothesis in favour of the alternate?
iii) For what values of sample means would we fail to reject the null hypothesis?

b) i) What is the Probability of a Type I error?
ii) If the true mean was 350 ($\mu = 350$). What is the Probability of a Type II error? What is the power of the test?

c) Now if the test was conducted at a significance level of 1%, how would the answers to part (a) and (b) change?

d) What factors decrease the Probability of Type II error? Which of these can we actually control?

e) What are the 4 possible decision outcomes?

f) Interpret Type I error and Type II error within the context of the problem.

Q1. Light bulb

$H_0: \mu = 375$ or $\mu \geq 375$ ✗

$H_a: \mu < 375$ ✓

If true have we made the correct decision ✓ ✗

$\mu = 375?$ $\sigma = 100$ → $n = 64$ → $\bar{x} = 350$

yes we have because truth is 350

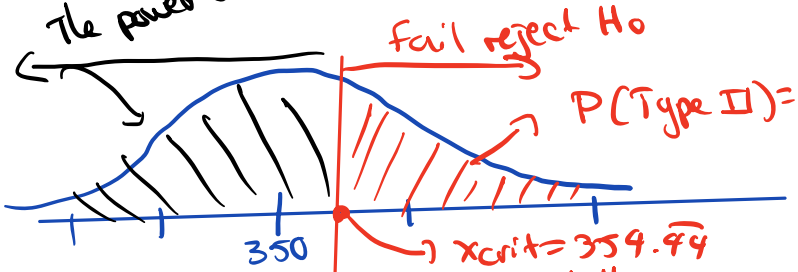
Truth
$\mu = 375$ ~~375~~ 350

our sample mean agrees with the alternate hypothesis

But does it statistically agree with it.

And can we establish statistical significance

The power of the test



$\bar{x} \sim N\left(350, \frac{100}{\sqrt{64}}\right)$ (12.5)

what-if

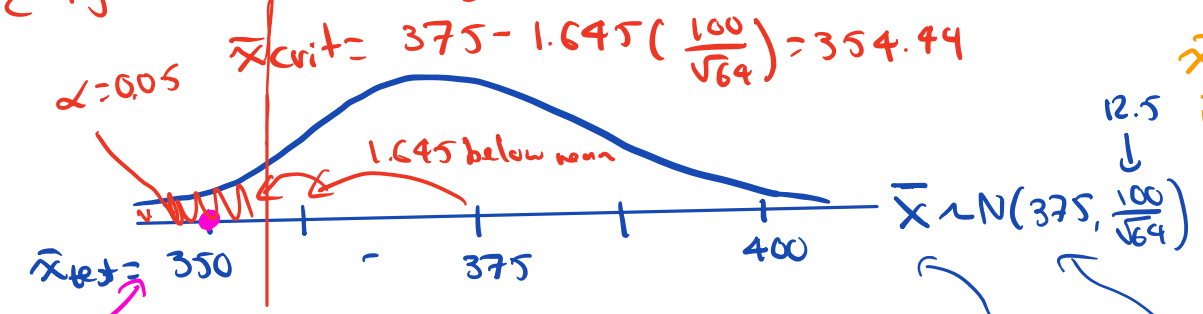
$\bar{x} = 200$

$\bar{x} = 250$

$\bar{x} = 300$

$\bar{x} = 373$

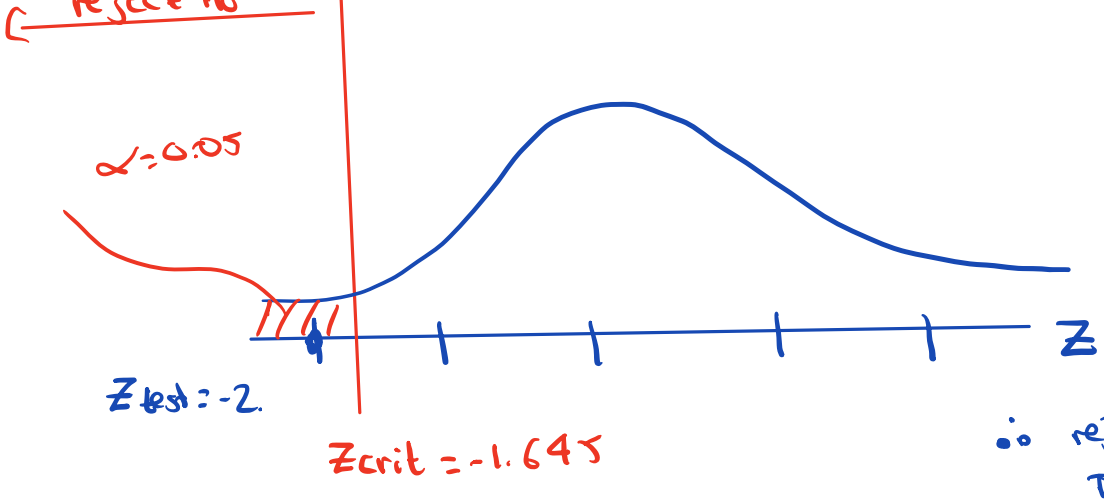
$\bar{x} = 350$ is this sufficiently small enough to conclude against the null in favour of the alternate.



Based on this observation of 64 light bulbs can we conclude that we have evidence against $H_0: 375$

hypothesized sampling distribution of $\bar{x}$.

assuming H_0 is true



∴ reject H_0

The avg lifetime of lightbulbs is below standards purpose.

We concluded that there is something wrong in the factory could we have made an error. Is it possible that in truth everything is cool in the factory is it indeed producing parts that are up to spec. Could we have incorrectly rejected something that we shouldn't have? Maybe.

The way our test is designed is that we reject H_0 when $\bar{x} < \bar{x}_{crit.}(354.44)$ if we knew the truth we could end making an incorrect decision w/ the test. when we reject an H_0 that happens to be true that's a type I error.

$$P(\text{type I error}) = \alpha = P(\bar{x} < 354.44) = \alpha = 0.05$$

we designed the tests (they aren't perfect), because its not perfect we can still make a mistake if the truth was indeed $\mu = 377$ and we sample 64 light bulbs and are 64 light bulbs yield an $\bar{x} < 354.44$ then we rejected something that we should not have rejected. And as such we have commited a type I error and the probability of commiting a type I error is α.

- Type I error rejecting something when the H_0 is true.
- A Type II error is commited when we fail to reject something that is false we failed to reject something that we should've rejected. we haven't detected a problem when there was one.
- To calculate a type II error we have to draw the true distribution of $\bar{x}$ and the true distribution of $\bar{x}$ is centered @ 350.

$$P(\text{Type II}) = P(\bar{x} \geq 354.44) \text{ when } X \sim N\left(350, \frac{100}{\sqrt{64}}\right)$$

$$= P(\bar{x} > 354.44 \mid \mu = 350)$$

$$= P\left(Z > \frac{354.44 - 350}{100/\sqrt{64}}\right)$$

$$= P(Z > 0.36)$$

$$= 1 - 0.6406 = 0.3594$$

The power of a test is

$$\text{Power} = 1 - P(\text{Type II}) = 0.6406$$

The power of the test is the probability of us rejecting something that should be rejected. The probability of us (to) correctly rejecting something that is supposed to be rejected.
A well designed test has a high power associated to it.

✓ Probably on the exam

b) if you want a power of 90%, when $\mu = 350$, how big of a sample do you need?

redo this problem.

as n↑, power↑

$n = 64$

power $= P(\bar{x} < "354.94" \mid \mu = 350)$

$\alpha = 1 - 0.9$
$= 0.10$

$n = ?$

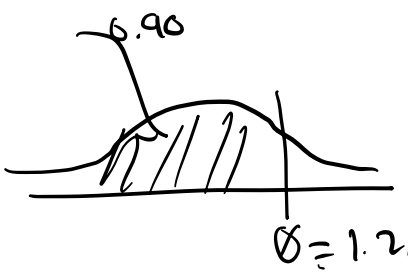
power $= 0.90$

$\bar{x} \sim N(350, \frac{100}{\sqrt{n}})$

$= P(\bar{x} < 375 - 1.645(\frac{100}{\sqrt{n}}) = 0.90$

standardize

$= P\left(Z < \frac{375 - 1.645(\frac{100}{\sqrt{n}}) - 350}{100/\sqrt{n}}\right) = 0.90$



$\frac{25}{\frac{100}{\sqrt{n}}} - 1.645 = 1.28 \rightarrow \sqrt{n} = 11.7$

$n = 136.89$↑

$n = 137$

2:44:46

Q2. Let's go back to Question 2 in the previous section about salad dressing.
We were concerned about whether the average amount of salad dressing poured was more than 250.
And evidence of this was collected by sampling 16 bottles.

a) -At significance level of 5%, what is the critical z value?
-For what values of sample means would we reject the null hypothesis in favour of the alternate?
-For what values of sample means would we fail to reject the null hypothesis?

b) -What is the Probability of a Type I error?
-If the true mean was 252 ($\mu = 252$). What is the Probability of a Type II error? What is the power of the test?

c) Now if the test was conducted at a significance level of 1%, how would the answers to part (a) and (b) change?

5. Two Sample Inference

This section concerns itself with the methods used to compare two sets of sample data to determine if there is a difference between population proportions or means.

5.1 Comparing Two Independent Means

Just as with the situation involving one sample, there are a number of cases to consider when testing claims involving two samples.

Case I: Large Sample Sizes

If two samples drawn from two independent samples that are both normally distributed, then the confidence interval for $(\mu_1 - \mu_2)$ is

$$(\bar{x}_1 - \bar{x}_2) \pm z_{\alpha/2}\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

the CI for $\mu_1 - \mu_2$ is given by this eqn.

The two-sample Z-test for $(\mu_1 - \mu_2)$ is as follows:

One-Tailed Test	Two-Tailed Test
H_0: $(\mu_1 - \mu_2) = D_0$	H_0: $(\mu_1 - \mu_2) = D_0$
H_a: $(\mu_1 - \mu_2) < D_0$ or H_a: $(\mu_1 - \mu_2) > D_0$	H_a: $(\mu_1 - \mu_2) \neq D_0$

where D_0 is the hypothesized difference between the means (often hypothesized to be 0).

Test statistic:

Z-test $$z = \frac{(\bar{x}_1 - \bar{x}_2) - D_0}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

when the population is large and σ_1 and σ_2 are known

*In both the confidence interval and the Z-test, if the population standard deviations are unknown, the sample standard deviations can be used.

When σ is not known, we can use the *t*-test to construct confidence intervals and perform hypothesis tests. In order to do so, the sampled populations must have distributions that are approximately normal.

There are two cases to consider when dealing with the *t*-distribution: one with the assumption/knowledge that the population variances are equal and the other when they are not equal.

Case II: Small Sample Sizes (Assuming $\sigma_1^2 = \sigma_2^2$)

The confidence interval for $(\mu_1 - \mu_2)$ is

$$(\bar{x}_1 - \bar{x}_2) \pm t_{\alpha/2}\sqrt{s_p^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}$$

where $t_{\alpha/2}$ is based on $(n_1 + n_2 - 2)$ degrees of freedom.

The pooled two-sample *t*-test for $(\mu_1 - \mu_2)$ is

One-Tailed Test	Two-Tailed Test
H_0: $(\mu_1 - \mu_2) = D_0$	H_0: $(\mu_1 - \mu_2) = D_0$
H_a: $(\mu_1 - \mu_2) < D_0$ or H_a: $(\mu_1 - \mu_2) > D_0$	H_a: $(\mu_1 - \mu_2) \neq D_0$

Test statistic:

$$t = \frac{(\bar{x}_1 - \bar{x}_2) - D_0}{\sqrt{s_p^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

← you use this when you can assume $\sigma_1^2 = \sigma_2^2$

For the case with the population variances being equal (or assumed equal) we need to calculate the pooled sample variance s_p^2.

$$s_p^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}$$

Case III: Small Sample Sizes (Assuming $\sigma_1^2 \neq \sigma_2^2$)

when we can't assume that $\sigma_1^2 = \sigma_2^2$

The confidence interval for $(\mu_1 - \mu_2)$ is

CI ⇒ $$(\bar{x}_1 - \bar{x}_2) \pm t_{\alpha/2}\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

where $t_{\alpha/2}$ is based on the degrees of freedom that is found by

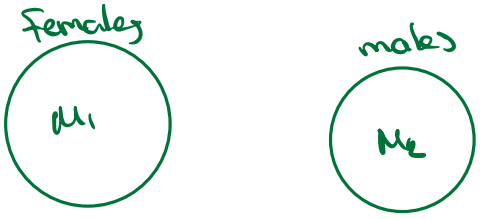
$$df = \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\frac{1}{n_1 - 1}\left(\frac{s_1^2}{n_1}\right)^2 + \frac{1}{n_2 - 1}\left(\frac{s_2^2}{n_2}\right)^2}$$

The two sample *t*-test of hypothesis for $(\mu_1 - \mu_2)$ has test statistic

Test statistic ⇒ $$t = \frac{(\bar{x}_1 - \bar{x}_2) - D_0}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

To motivate the idea of two sample inference:

we are now interested in comparing 2 populations



X ↙ random variable : marks in stats.

one of your friends has a theory: she believes that girls are better than boys in stats. You are like okay do you have any data, or statistical evidence to back it up, and your friend says yes I do have some data.

μ_1 represents the average stats mark of all female students from all universities in Canada.

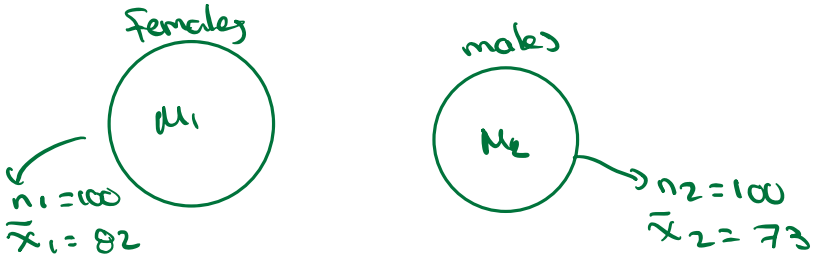
μ_2 represents the average stats mark of all male students from all universities in Canada.

Your friend is say look: $\mu_1 > \mu_2$.

↳ This is the statement and claim that your friend is making.

To prove a bold statement like this you should be able to back it with data.

↳ To do this



measure and write down all their final marks, on the final

Data 1 suggests that yes, females tend to do better than males, in statics

The question is this difference based on 100 samples and 100 females & males statistically significant for us to conclude and make this claim:

$H_a: \mu_1 > \mu_2 \Rightarrow \mu_1 - \mu_2 = 0$

$H_0: \mu_1 = \mu_2 \Rightarrow \mu_1 - \mu_2 > 0$

To test this we will need a test statistic using our evidence of one # that's going to allow us to do a test and compare that against a critical value

5.2 Practice Questions

Q1. A study examined whether the perception of service quality at five-star hotels in Jamaica differed by gender. Hotel guests were randomly selected from the lobby and restaurant areas and asked to rate ten service-related items. Each item was rated on a 5-point scale (1 = "much worse than I expected," 5 = "much better than I expected") and the sum of the items for each guest was determined. A summary of the guest scores are provided in the table.

Gender	Sample Size	Mean Score	Standard Deviation
Males	127 (large)	39.08 $\bar{x}_1$	6.73 s_1
Females	114 (large)	38.79 $\bar{x}_2$	6.94 s_2

you are given statistics on 2 populations.

a) Construct a 90% confidence interval for the difference between the population mean service-rating scores given by male and female guests at Jamaican 5-star hotels.

Just by eyeballing the data the males have a tendancy to give higher scores than the females

We want to build a 90% CI for $\mu_1 - \mu_2$

σ_1 & σ_2 are unknown we can't assume that $\sigma_1 = \sigma_2$

∴ The CI we will use is case 3. ⇒ used for small samples.

$$\bar{x}_1 - \bar{x}_2 \pm t_{\alpha/2, n-2}\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

because $n_1, n_2 > 30$ This t value will be very close to its corresponding z-value.

$90 = 100(1-\alpha)$

$0.9 = (1-\alpha)$

$\alpha = 0.1$

$\frac{\alpha}{2} = 0.05$

b) Use the interval to make an inference about whether the perception of service quality at five-star hotels in Jamaica differs by gender.

$$\approx \bar{x}_1 - \bar{x}_2 \pm z_{\alpha/2}\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

(we can make this approximation because $n_1, n_2 > 30$)

↳ This z-value will closely approximate this t-value.

$$= (39.08 - 38.79) \pm 1.645\sqrt{\frac{(6.73)^2}{127} + \frac{(6.94)^2}{114}} = 0.29 \pm 1.452$$

$\sqrt{\quad} = 0.8827$

$= (-1.16, 1.74)$

The male populations and female populations are independent of eachother.

The reason we can use the z dist inplace of t (when n_1 and n_2 are big) because as your sample size gets bigger and bigger your t dist starts to look more and more like your z distribution.

b) Use the interval to make an inference about whether the perception of service quality at five-star hotels in Jamaica differs by gender.

what does this interval tell us about $\mu_1 - \mu_2$

if the CI (3, 4)

we are 90% confident that the true difference between μ_1 and μ_2 lies in between 3 & 4 which is evidence that males give higher ratings than females.

for our CI (-1.16, 1.74) since it includes 0 there is a pretty high likelyhood that there is no difference between μ_1 & μ_2.
or there is no statistical difference between μ_1 and μ_2.

Q2. The operations manager at a light bulb factory wants to determine whether there is any difference in the average life expectancy of bulbs manufactured on two different types of machines. The population standard deviation of Machine I is known to be 20 hours and of Machine II is 25 hours. A random sample of 64 light bulbs obtained from Machine I indicates a sample mean of 375 hours, and a similar sample of 64 from Machine II indicates a sample mean of 362 hours. Using the 5% level of significance, is there any evidence of a difference in the average life of bulbs produced by the two types of machines?

Q3. A carpet manufacturer is studying difference between two of its major outlet stores. The company is particular interested in the time it takes customers to receive carpeting that has been ordered from the plant. Data concerning a sample of delivery times for the most popular type of carpet are summarized as follows:

	Store A	**Store B**
$\overline{x}$	34.3 days	43.7 days
s	2.4 days	3.1 days
n	41	31

a) Assuming that the population variances from both stores are equal, is there evidence of a difference in the average delivery time between the two outlet stores? (Use a 1% level of significance.)

b) Set up a 99% confidence interval estimate of the difference between the population means between Store A and Store B.

2nd half of exam will look like this. (a,b,c,d)

Q4. Medical researchers are comparing two treatments for migraine headaches. They wish to perform a double-blind experiment to assess if Treatment 2 (the new treatment) is significantly better than Treatment 1 (the standard treatment) using a 5% level of significance. Twenty subjects were available for the study and were randomized to one of the two treatment groups. Treatment 1 was administered to 10 subjects, while Treatment 2 was administered to the remaining 10. Each subject was instructed to take the medication at the onset of a migraine headache and to record the time that elapsed until relief, defined as a reduction in throbbing.

This is what we want to show evidence for.

↳ This is how we record effectiveness for a treatment.
(This is the random variable that we are measuring on each of the 10 subjects.)

a) State the appropriate hypotheses to be tested.

$H_0: \mu_1 - \mu_2 = 0$ or $\mu_1 - \mu_2 \leq 0$

H_a: new treatment is better than standard treatment.
(2) (1)

↑ This is what we want to show. We want to look for evidence

$\mu_2 < \mu_1$ (The avg time for relief is smaller for the new treatment than the time for relief for the standard treatment)

$\Rightarrow \mu_1 - \mu_2 > 0$ A better treatment is one that lowers the avg time.

b) The mean time to relief for Treatment 1 subjects was 22.6 minutes, with a standard deviation of 5.2 minutes. The mean time to relief for the Treatment 2 group was 19.4 minutes, with a standard deviation of 4.9 minutes. Assuming that the variances in the two treatments are equal, calculate the pooled sample standard deviation.

$n_1 = 10$ $n_2 = 10$

$\Rightarrow \bar{x}_1 = 22.6$ $\bar{x}_2 = 19.4$

$s_1 = 5.2$ $s_2 = 4.9$

How do we know that these are samples statistics

Because it is the mean time to relief for treatment #1 subjects. And those 10 subjects showed a standard deviation of 5.2 min.

Just by eyeballing the data (the 2nd row) based on 10 subjects that are treated with each treatment the data seems to agree w/ the alternate hypothesis, that $\bar{x}_1 > \bar{x}_2$

↳ we now want to see if there is statistical evidence that this is true. and do a formal hypothesis test.

Assume $\sigma_1 = \sigma_2$ or $\sigma_1^2 = \sigma_2^2$

Based on this assumption calculate the pooled standard deviation.

$$s_p^2 = \frac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{n_1+n_2-2} = 25.525$$

$s_p = 5.052$.

what if $\bar{x}_1 = 22.6$ $\bar{x}_2 = 9.4$

The 2nd treatment yielded significantly more effective results. than the standard treatment

(22.6 - 9.4) → pushed deeper into the rejection region.

c) Compute the test statistic.

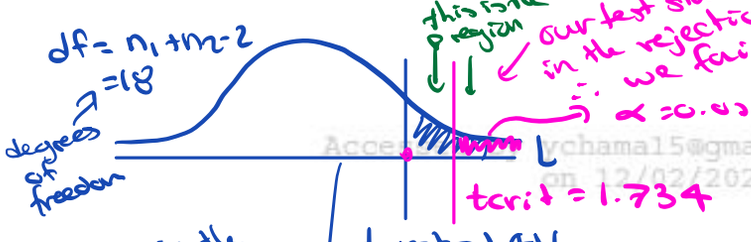
case #2 in the course booklet

the bigger this difference the more evidence we have in favour of the alternate hypothesis.

$$t_{test} = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{s_p^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} = \frac{22.6 - 19.4}{\sqrt{25.525\left(\frac{1}{10} + \frac{1}{10}\right)}} = 1.416$$

The bigger the numerator, the bigger our test statistic. The bigger our test statistic, the more extreme to the tails of our t distribution we will be. The more extreme we are pushed to either the left or right side of a t distribution then we have more evidence against the null in favour of the alternate.

d) Find the corresponding *p*-value.



H0: $\mu_1 - \mu_2 = 0$

Ha: $\mu_1 - \mu_2 > 0$

(1-tailed test because the arrow lies in the alternate hypothesis)

The rejection region always points in the direction of this arrow.

The p-value is the probability of observing more extreme evidence than the one we observed in our 2 samples.

e) State the decision and conclusion using a 5% level of significance.

$0.05 < p\text{-value} = P(t > 1.416) < 0.10$

all we care about since we are doing a statistical test is that our p-value > 0.05 (α)

∴ we fail to reject to reject H0.

unlike the z table the t table does not give you an exact probability.

You can't look up an exact probability associated with a t-value.

what we can find is the bounds ie upper bounds, lower bounds.

Thats sufficient because we care if this is less than α or greater than α.

High p-value ⇒ H0 is not rejected.

low p-value ⇒ reject H0

Q5. Researchers are evaluating a new drug for the treatment of Type II diabetes. They randomly assign 20 volunteers, each of whom has a history of poor glucose control levels, to 2 groups. One group gets the experimental drug on trial, the other receives the standard treatment regimen. At the end of the trial each participant has their A1C -a measure of long term blood glucose levels- tested and the results are summarized below.

A1C

X: glucose lvl

												mean	s.deviation
n=10	Standard	14.2	15.3	16.1	14.9	13.7	15.3	14.8	13.5	14.3	13.8	14.59 $\bar{x}_1$	0.79 s_1
n=10	Trial	10.6	12.5	13.8	13.5	10.2	12.4	13.2	11.6	11.7	11.1	12.06 $\bar{x}_2$	1.16 s_2

Researchers feel that it will be worthwhile to continue to fund research into the trial drug if the new regime lowers true A1C by one full point (i.e μ_{Trial} is at least 1.0 less than $\mu_{Standard}$). Test at a significance level of 5% whether of not the research should proceed.
Assume the variances of the two treatments are equal.

↓ glucose lvl the more effective the treatment is.

From 1st glance the 10 patients that were given the trial there does seem to be some evidence that the A1C lvl is lower

1) we need an alternate hypothesis here.

H_0: $\mu_1 - \mu_2 = 1$ or $\mu_1 - \mu_2 \leq 1$

H_a: μ_2 is at least 1.0 less than μ_1 (if μ_2 is low this is a good thing, it means that it helps lower the glucose lvl.)

$\Rightarrow \mu_1 - \mu_2 > 1$

our eyeballing of the data, the sample statistic that will help us test this is

$\bar{x}_1 - \bar{x}_2 = 14.59 - 12.06$
$= 2.53$

this is where you 1st develop an intuition it seems like the subjects treated by the new trial do seem to exhibit an average a lower A1C by 2.53. So it is atleast 1. Now the question is, is this low enough to statistically justify that the population avg differences are @ least 1.

Assume $\sigma_1 = \sigma_2$

Pooled version

hypothesized difference

$$t_{obs} = t_{test} = \frac{\bar{x}_1 - \bar{x}_2 - \Delta_0}{\sqrt{s_p^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

Its a pooled variance because we are told to assume this.

$$s_p^2 = \frac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{n_1+n_2-2}$$

$s_p = 1.378$

hypothesized difference

$$t_{test} = \frac{2.53-1}{\sqrt{(1.378)\left(\frac{1}{10}+\frac{1}{10}\right)}}$$

$= 2.914$

The bigger this is relative to the hypothesized difference the more evidence we have in favour of the alternate.

if there is a bigger difference this should indicate stronger evidence in favour of the alternate hypothesis.

what is the interpretation of a pooled variance?

A pooled variance can be thought of as this:

- our 1st sample standard deviation was 0.79, our 2nd sample standard deviation for the 2nd sample was 1.16. the pooled sample standard deviation can almost be seen as an average std of the 20 ppl in our sample.

we used this because we assume $\sigma_1 = \sigma_2$.

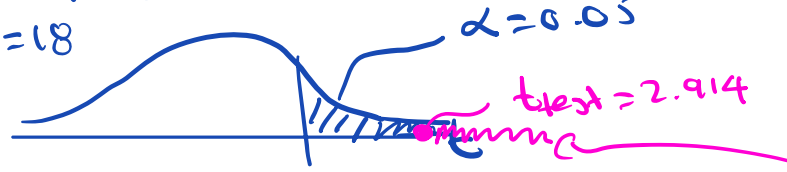
It helps us simplify the calculation

To conduct a test we can either use the p-value method or the critical value method

- This is a 1 tailed test because our Ha has either a > or <
- the rejection region will be to the right.

$dof = n_1 + n_2 - 2$

$= 18$



$\therefore$ we reject Ho.

$t_{crit} = t_{0.05,18} = 1.734$

p-value is the probability that t_{test} > than the one observed in our sample.

p-value = $P(t > 2.914)$

Graphically we can see that this p-value < 0.05

$\therefore$ we reject H

b) Provide a 95% Confidence Interval for the true difference in A1C means.

95% CI for $\mu_1 - \mu_2$

$$\bar{x}_1 + \bar{x}_2 \pm t_{\alpha/2, n_1+n_2-2} \sqrt{sp^2 \left(\frac{1}{n_1} + \frac{1}{n_2}\right)}$$

$$= (4.99 - 12.06) \pm 2.101 \left(1.174\left(\frac{1}{10} + \frac{1}{10}\right)\right)$$

$$= 2.53 \pm 1.103)$$

$$= (1.427, 3.633)$$

$\alpha = 1 - 0.95 = 0.05$

$\frac{\alpha}{2} = 0.025$

$t_{\frac{\alpha}{2}, n_1+n_2-2}$

$= t_{0.025, 18} = 2.101$

↑ we are 95% confident that the value of $\mu_1 - \mu_2$ is in this range.

It has one interesting characteristic in relation to the hypothesis test that we just did.

Notice that (1.427, 3.633) → excludes 1

In fact this is evidence that the difference between μ_1 and μ_2 is in between 1.427 & 3.633. which in turn suggest that $\mu_1 - \mu_2 > 1$

It looks like the evidence here suggests that the trial treatment suggests that the new trial is effective at reducing the glucose lvl amongst diabetics.

why don't we use $\bar{d}$ ie the paired test?

The reason why we don't use $\bar{d}$ is because these are independent samples. If these guys were the same person and you took 2 measurements per person we say the samples are dependent samples.

Q6. A teacher claims that use of *discovery* improves performance on the standard test given to third-grade students. She decides to compare the scores of 16 students in one of her sections (taught by her using *discovery*) with the scores of 16 students in a second section who have been taught by her without the *discovery* process. The data, summarized in the following table, are assumed to be independent random samples from two normal distributions:

	Sample Size	**Sample Mean**	**Sample Standard Dev.**
With Discovery	16	600	90
Without Discovery	16	550	100

Construct the 90% confidence interval for the difference in the population mean scores. Based on this confidence interval, does there appear to be an improvement with the *discovery* process? Assume the population standard deviations are unequal.

$\sigma_1 \neq \sigma_2$

$\bar{x}_1 - \bar{x}_2 \pm t_{\alpha/2,\nu} \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$

everything else is standard except we use t since $n < 30$.

we can't use the pooled sample variance because of this.

The one big headache is calculating this degree of freedom. (It has a super complicated eqn)

5.3 Inference on Means for Paired Data

Experiments in which observations are paired and the difference are analyzed are called **matched pairs experiments**. The idea is to compare population means by comparing the difference between pairs of experimental units that were very similar prior to the experiment.

Matched Pairs Confidence Interval $(\mu_D = \mu_1 - \mu_2)$

$$(\bar{x}_D) \pm t_{\alpha/2} \frac{s_D}{\sqrt{n_D}}$$

where $t_{\alpha/2}$ is based on (nD – 1) degrees of freedom.

Matched Pairs Test of Hypothesis $(\mu_D = \mu_1 - \mu_2)$

One-Tailed Test

H_0: $\mu_D = D_0$

H_a: $\mu_D < D_0$ or H_a: $\mu_D > D_0$

Two-Tailed Test

H_0: $\mu_D = D_0$

H_a: $\mu_D \neq D_0$

Test statistic:

$$t = \frac{\bar{x}_D - D_0}{\frac{s_D}{\sqrt{n}}}$$

5.4 Practice Questions

Q1. It is claimed that an industrial safety program is effective in reducing the loss of working hours due to factory accidents. The following data are collected concerning the weekly loss of working hours due to accidents in six plans both before and after the safety program is instituted:

Plant Number	1	2	3	4	5	6
Before program	12	29	16	37	28	15
After program	10	28	17	35	25	16
Difference	2	1	-1	2	3	-1

We are interested in analyaizng the difference, which were computed as before minus after. In particular, assess if there has been a significant reduction in the loss of hours on average. State the hypotheses and perform the test using a 10% significance level.

μ_D = weigh After quitting − B4 quitting

μ_D ← the average of differences in weight. b4 and after smoking.

Q2. A study was conducted to estimate the mean weight change of a female adult who quits smoking. The weights of 16 female adults before they stopped smoking and five weeks after they stopped smoking were recorded. The differences were computed as "after" – "before". The difference in weight is assumed to follow a normal distribution.

(Paired data)

Subject	1	2	3	4	5	6	7	8
After	154	181	151	120	137	134	121	128
Before	148	176	153	116	129	128	120	132
Difference	6	5	-2	4	8	6	1	-4

Subject	9	10	11	12	13	14	15	16
After	140	130	139	127	131	120	131	140
Before	131	112	143	116	129	123	130	132
Difference	9	18	-4	11	2	-3	1	8

$\mu_D = 0$?

$\mu_D > 0$?

$\mu_D < 0$?

is there a weight gain or weight loss.

a) Find the sample mean difference and the sample standard deviation of the differences.

Sample Differences is denoted by $\bar{D}$. Take the differences row and average them

$$\bar{D} = \frac{\sum x}{n} = 4.125$$

$$S_D = \sqrt{\frac{\sum x^2 - \frac{(\sum x)^2}{n}}{n-1}} = 6.032$$

b) Compute a 95% confidence interval estimate for the mean difference.

μ_D mean difference in weight After quitting smoking and b4.

The 95% CI is given by

$n < 30$, $n = 16$.

$$\bar{D} \pm t_{n-1,\alpha/2}\frac{S_D}{\sqrt{n}} \quad \text{or} \quad \bar{D} + Z_{\alpha/2}\frac{\sigma_D}{\sqrt{n}}$$

$t_{n-1,\alpha/2} = 2.131$

$95 = 100(1-\alpha)$

$0.95 = 1-\alpha$

$\alpha = 0.05$

$\frac{\alpha}{2} = 0.025$

$\alpha = 1 - 0.95$

$t_{0.025,15} = 2.131$

$$= 4.125 \pm 2.131\left(\frac{6.032}{\sqrt{16}}\right)$$

$$= 4.125 \pm 3.215$$

$$= (0.91, 7.34)$$

c) Interpret the interval.

We are 95% confident that μ_D is in between 0.91 & 7.34
Since it excludes 0, we can conclude at $\alpha = 5\%$ that there is a weight gain after quitting

$\bar{D} =$

d) Is there evidence of weight gain? Test at a 5% level of significance.

What if $\bar{D} = 25$ lbs → this would put us deep in the rejection region.

$H_0: \mu_D = 0$ or $\mu_D \leq 0$.

$H_a: \mu_D > 0$

Here it is a t test statistic (because $n < 30$)

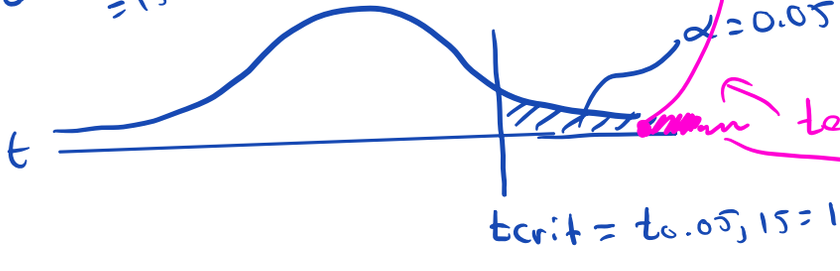
$$t_{test} = \frac{\bar{D} - (\mu_D)_0}{s_D/\sqrt{n}} = \frac{4.125 - 0}{6.032/} = 2.735$$

on avg there was a weight gain of 4 lbs in our sample of 16 females.

The bigger this diff the stronger your evidence in favor of H_0.

Bigger t test statistic the deeper it pushes you into the rejection region.

$df = \nu = n - 1 = 15$



test statistic lies in the rejection region.

$t_{crit} = t_{0.05,15} = 1.753$

p-value magenta area is smaller than α of 0.05%.

There is a weight gain happening after quitting smoking.

textbook Ch 6.11

5.5 Comparing Two Population Variances

The common statistical procedure to compare population variances σ_1^2 and σ_2^2, makes an inference about the ratio $\frac{\sigma_1^2}{\sigma_2^2}$.

The sampling distribution of $\frac{s_1^2}{s_2^2}$ is based on the assumptions:

1. The two sampled populations are normally distributed.
2. The samples are randomly and independenty selected from their respective populations.

F-Test for Equal Population Variances

One-Tailed Test	Two-Tailed Test
H_0: $\sigma_1^2 = \sigma_2^2$	H_0: $\sigma_1^2 = \sigma_2^2$
H_a: $\sigma_1^2 < \sigma_2^2$ or H_a: $\sigma_1^2 > \sigma_2^2$	H_a: $\sigma_1^2 \neq \sigma_2^2$

Test statistic:

$F = \frac{s_2^2}{s_1^2}$ $\quad F = \frac{s_1^2}{s_2^2}$ (or when H_a: $\sigma_1^2 > \sigma_2^2$)

* the larger sample variance is always on the top

where F is based on v_1 = numerator degrees of freedom
and v_2 = denominator degrees of freedom (both are $n - 1$)

5.6 Practice Questions

Q1. A professor in the accounting department of a business school claims that there is much more variability in the final exam scores of students taking the introductory accounting course as a requirement than in scores of students taking the course as part of a major in accounting. Random samples of 13 nonaccounting majors (Group 1) and 10 accounting majors (Group 2) are taken from the professor's class roster in his large lecture, and the following results are computed based on the final exam scores:

$n_1 = 13$ $s_1^2 = 210.2$ $n_2 = 10$ $s_2^2 = 36.5$

At the 5% level of significance, is there evidence to support the professor's claim?

Q2. A study was conducted to compare the mathematics achievement test scores of male and female students. The researchers hypothesized that the distribution of test scores for males is more variable than the corresponding distribution for females. Use the summary information reproduced in the table to test this claim at a 1% level of significance.

	Males	Females
Sample Size	1764	1739
Mean	48.9	48.4
Standard Deviation	12.96	11.85

5.7 Tests with Categorical Data

Chi-Squared Goodness of Fit Test

The chi-square goodness of fit test is used to measure how well observed data fit what would be expected under specified conditions. To test the null hypothesis that the observed frequency equals the expected frequency, compute the chi-square statistic:

$$\chi^2_{obs} = \sum \frac{(O_i - E_i)^2}{E_i}$$

where O_i = the observed frequency
E_i = the expected frequency
$df = k - 1$ where k is the number of categories/cells specified under H_o

Decision Rule: If $\chi^2_{obs} > \chi^2_{\alpha,k-1}$, reject H_o

Note: When performing the chi-square goodness of fit test, we combine any expected frequencies into groups such that no expected frequencies are less than 5. In this case the degrees of freedom is calculated as follows:

df = # categories after combining - # parameters being estimated – 1

Chi-Squared Test of Independence

The chi-square test of independence is used to determine whether the row and column variables in a two-way contingency table are independent or related. The chi-square statistics is a measure of how much the observed cell counts diverge from the expected cell counts. To test the hypothesis

H_o: no association between row and column variables

compute the chi-square statistic

$$\chi^2_{obs} = \sum \frac{(O_i - E_i)^2}{E_i}$$

where the expected cell count, f_e, is computed by:

n = total number of observations in the table
df = (r – 1)(c – 1) where r = number of rows and c = number of columns

Decision Rule: If $\chi^2_{obs} > \chi^2_{\alpha,df}$, reject H_o

5.8 Practice Questions

Q1. You roll a die 120 times and observe the following frequencies

	1	2	3	4	5	6
Observed Frequencies	26	23	14	29	15	13

Is there evidence that an unfair die is being used? Test at a significance level of 5%

Q2. A plant manager takes a random sample of 100 sick days. He divided his employees into age groups to see examine if age was a factor in the number of sick days taken. The table below shows the number of sick days out of the 100 that were taken by each group and their size.

Age Group	Proportion of Labour Force	Number of Sick Days
20 – 29	30%	26
30 – 39	40%	37
40 – 49	20%	24
50 and over	10%	13

Using the χ^2 goodness of fit test at the 5% level of significance, determine what the manager should conclude about the hypothesis that age is not a factor in taking sick days.

Q3. A random sample of 37 workers above the age of 65 in a town gave the results indicated by the contingency table below. Test at the 10% level of significance the hypothesis that the number of male and female workers in the 66 to 70 and 71-plus age groups in the town are independent of sex.

Age Group	Male	Female	Total
66 – 70	17	9	26
> 70	3	8	11
Total	20	17	37

Q4. The table shows the way in which a randomly chosen group intend to vote in the next election.

	Age of Voter		
	18 to 35	35 to 59	60+
Party A	85	95	131
Party B	168	197	173

Test at a 5% level whether there is any association between the age of a voter and the party they wish to vote for.

6. Linear Regression

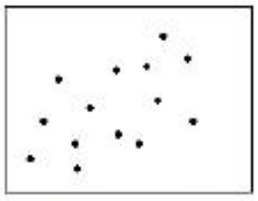
6.1 Correlation

A **correlation** is a relationship between two variables. The data can be represented by ordered pairs (x, y) where x is the independent variable and y is the dependent variable.

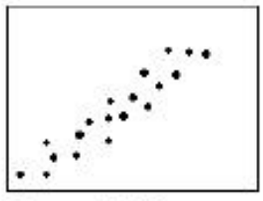
The graph of ordered pairs (x, y) is called a scatter plot. A scatter plot can be used to determine whether a linear (straight line) correlation exists between two variables. Furthermore, we can examine the strength and direction of the correlation.

If the dependent variable tends to ***increase*** as the independent variable increases, we say that there is a **positive** correlation between the variables. If the dependent variable tends to ***decrease*** as the independent variable increases, we say that there is a **negative** correlation between the variables.

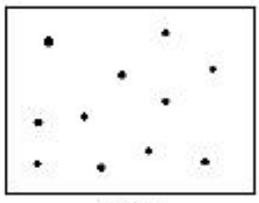
The strength of a linear relationship gives an indication of ***how closely*** the points in the scatter plot form a straight line.



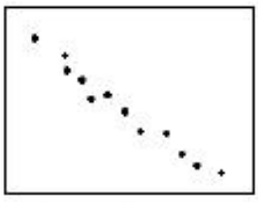
Weak Positive



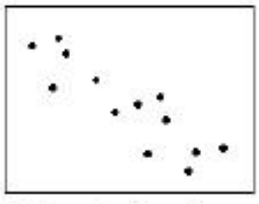
Strong Positive



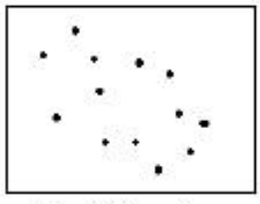
None



Strong Negative

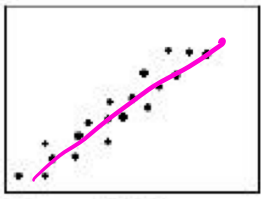


Moderate Negative



Weak Negative

y: final mark



Strong Positive

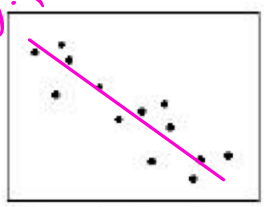
x: hrs studied

if you see a relationship like this between your x and y where there is a tendency for y to increase as x increases we say that x and y exhibit a +ve correlation

In general if you don't study for your final, your final mark will suffer.

if you put more time into studying then you should see your mark go up.

y: Final mark.

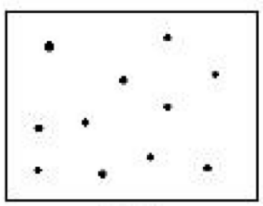


Moderate Negative

-ve correlation

As x increases y has a tendency to decrease

x: # of hrs watching tv



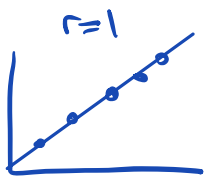
None

- or there can be no correlation between x and y.

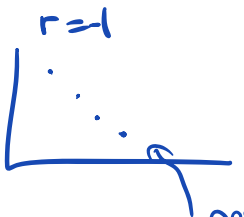
Theres different scenarios here and there is one index that captures how strong correlation is which is called to correlation coefficient denoted by

$$-1 \leq r \leq 1$$

where 1 =>



a perfect +ve correlation there is no randomness and no noise. Every single data pt is on a straight line



perfect -ve correlation between x and y.

r=0

There is no correlation.

- r is an index that measures how strong the correlation is.

- r is a sample statistic, meaning if, lets say each graph has 10 data pts

$r = f(x, y)$ (r is a function of your observations in x and y).

that function is given by

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

The correlation coefficient quantifies the correlation between x and y:

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

These measurements are used to compute your correlation coefficient.

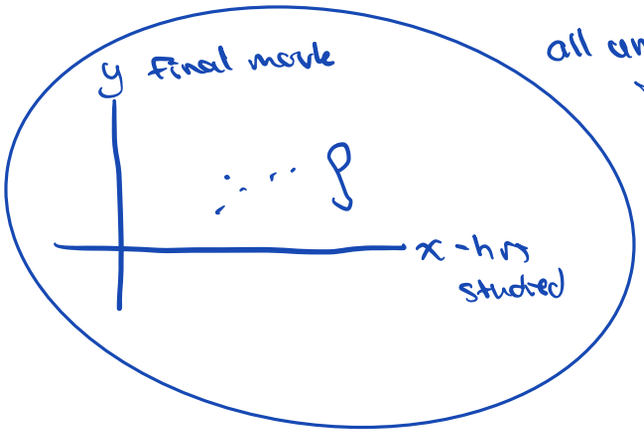
Where
$(-1 < r < 1)$

Properties of the Correlation Coefficient:
1. The correlation coefficient can take on any values between -1 and 1 inclusive.
2. Positive values indicate a positive correlation; negative values indicate a negative correlation.
3. Values closer to 1 and -1 indicate a strong correlation; values closer to 0 indicate a weak correlation

if you look closely @ your function it uses your data.

x_1	y_1
x_n	y_n

If I were to be interested in what the true correlation is between x hrs studied & y final mark



all university students in canada.

There's 100k of them no one knows the true correlation between your x and y But we know there is a population correlation, which is denoted by ro. (ρ)

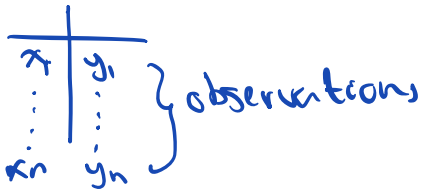
To infer about this take a sample of say 100 and compute a sample statistic say r. sample statistics.

$n = 100$
r

Population parameters
$\rho \to r$ (ro is to r)
$\mu \to \bar{x}$ as μ is to $\bar{x}$.
$\sigma \to s$ as σ is to s
$\mu_1 - \mu_2 \to \bar{x}_1 - \bar{x}_2$
$\mu_d \to \bar{D}$.

for correlation just like your μ & σ there is a sample correlation and a population correlation. (ie correlation coefficient)

- we probably want to do an inference on ρ using the observations at our disposle.

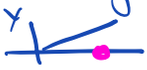


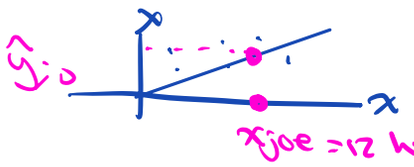
Taking advantage of correlation to predict stuff... let's discuss

Say Y captures the mark of a student in ENDG319.

If we sample 50 students and write down their marks, based on this data, what would be the best predictor?

Now from the same sample, say we also collected X: hours studied. Can we improve our prediction?

After the correlation coefficient, the nice thing about correlated data sets if we can establish x and y  are tvely correlated then we can take advantage of this linear relationship and build a linear regression model, to help use predict the final mark of joe, if we know x_{joe} has studied 12 hrs for an exam what we can do is fit a line of best fit using our 100 observations which excludes joe and after we fit this data to a model, now we can use this model to predict what joes mark could potentially be. The predicted mark on the linear regression line is denoted by $\hat{y}$.



The idea here in linear regression is lets say Y captures the mark of students in Endg 314

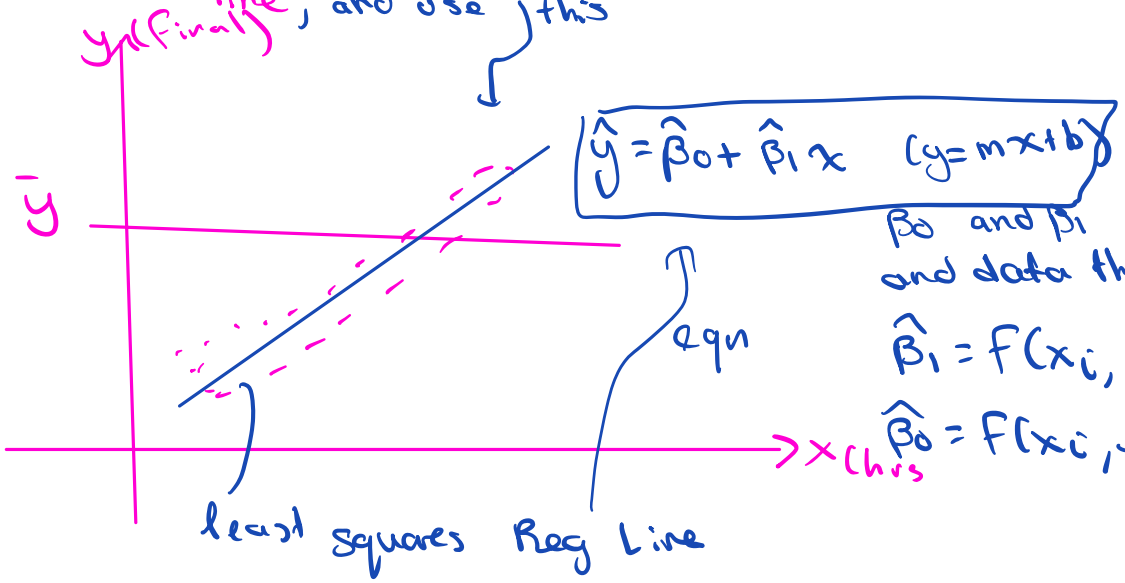
- if we sample 50 students and write down their marks based on this data what will be the best predictor.

$y_1 \dots y_{50}$ we have 50 sets of marks we don't know how much they have studied.

$\Downarrow$ and using these data points whats the best predictor to predict joes mark.

anything that is normally distributed tends to congregate around this measure of central tendency so chances are $\bar{y}$ is the most sensible prediction for his mark.

But if we know there is a correlation between x and y, what we can do instead of using $\bar{y}$ as a predictor for joe we can build a smarter prediction model than this flat line, and use this



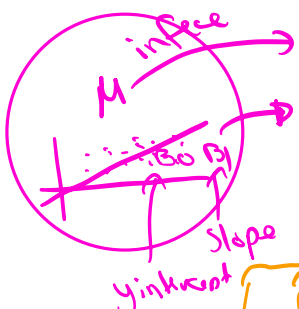
$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x \quad (y = mx + b)$$

β_0 and β_1 depend on the observations and data that you have collected

$$\hat{\beta}_1 = f(x_i, y_i)$$

$$\hat{\beta}_0 = f(x_i, y_i)$$

we argued if we know how many hours a particular student has studied and the final mark that a student gets. If we sample 50 students from past yrs that have taken Engy 319. And we draw a scatter plot of all their data pts and we see that they exhibit a true correlation between the two what we can do is look to fit a line of best fit which we call the least squares regression lines.



$\bar{x} \Rightarrow$ CI, HT $\bar{x} \sim N(\mu, \frac{\sigma}{\sqrt{n}})$

$\hat{\beta}_0 = f(x_i, y_i)$

$\hat{\beta}_1 = f(x_i, y_i)$

$\hat{\beta}_1 \sim N(\beta_1, s_{\hat{\beta}_1})$

Random variable

$\hat{\beta}_0 \sim N(\beta_0, s_{\hat{\beta}_0})$

using these sampling distributions we can build CI for $\hat{\beta}_{0,1}$ and also conduct HT for these.

As is to ← to infer on pop param we use sample statistics.

p	r
μ	$\bar{x}$
σ	s
$\mu_1 - \mu_2$	$\bar{x}_1 - \bar{x}_2$
β_1	$\hat{\beta}_1$
β_0	$\hat{\beta}_0$

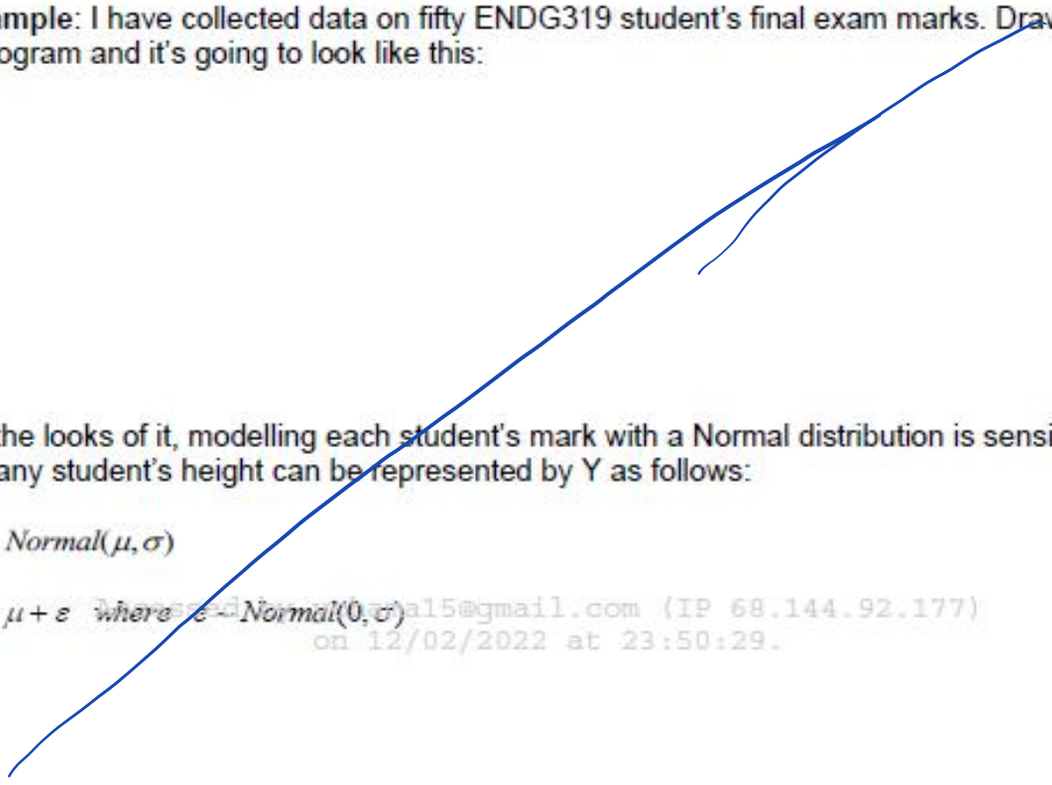
?

population parameters

they pertain to a big population. (nobody knows)

all these values are functions of your observations or data in your samples
f(data)

Example: I have collected data on fifty ENDG319 student's final exam marks. Draw a histogram and it's going to look like this:



By the looks of it, modelling each student's mark with a Normal distribution is sensible. So any student's height can be represented by Y as follows:

$$Y \sim Normal(\mu, \sigma)$$

$$Y = \mu + \varepsilon \quad where\ \varepsilon \sim Normal(0, \sigma)$$

What is the Distribution of a random Variable Y when we know that it is related to x? For example, if Y is ice cream sales, you would expect to see more sales when temperature (x) is higher. Y however is still a random variable with some unknown mean and a variance.

6.2 The Regression Model

The mean depends on values of x. And we make the following assumption (that the mean is linearly related to x):

$$\mu_{y_i} = \beta_0 + \beta_1 x_i$$

So for a given x value, a particular y value will be somewhere close to the mean. So the y-value is modeled as so:

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$

Where ε (epsilon) captures the deviation (or error) from the mean.
And we assume that ε has a normal distribution with a mean of zero and a standard deviation of sigma.

$$\varepsilon \sim N(0, \sigma)$$

If we can estimate β_0 and β_1, we can do all kinds of inference on μ at different values of x and also predict values of y for different values of x.

Estimating β_0 and β_1 (result of Least Squares Regression)

The estimated regression model is

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$$

where $\hat{y}$ = the predicted value of the response variable
$\hat{\beta}_0$: the estimate of the *y*-intercept β_0
$\hat{\beta}_1$: the estimate of the slope β_1

$\hat{\beta}_1 =$

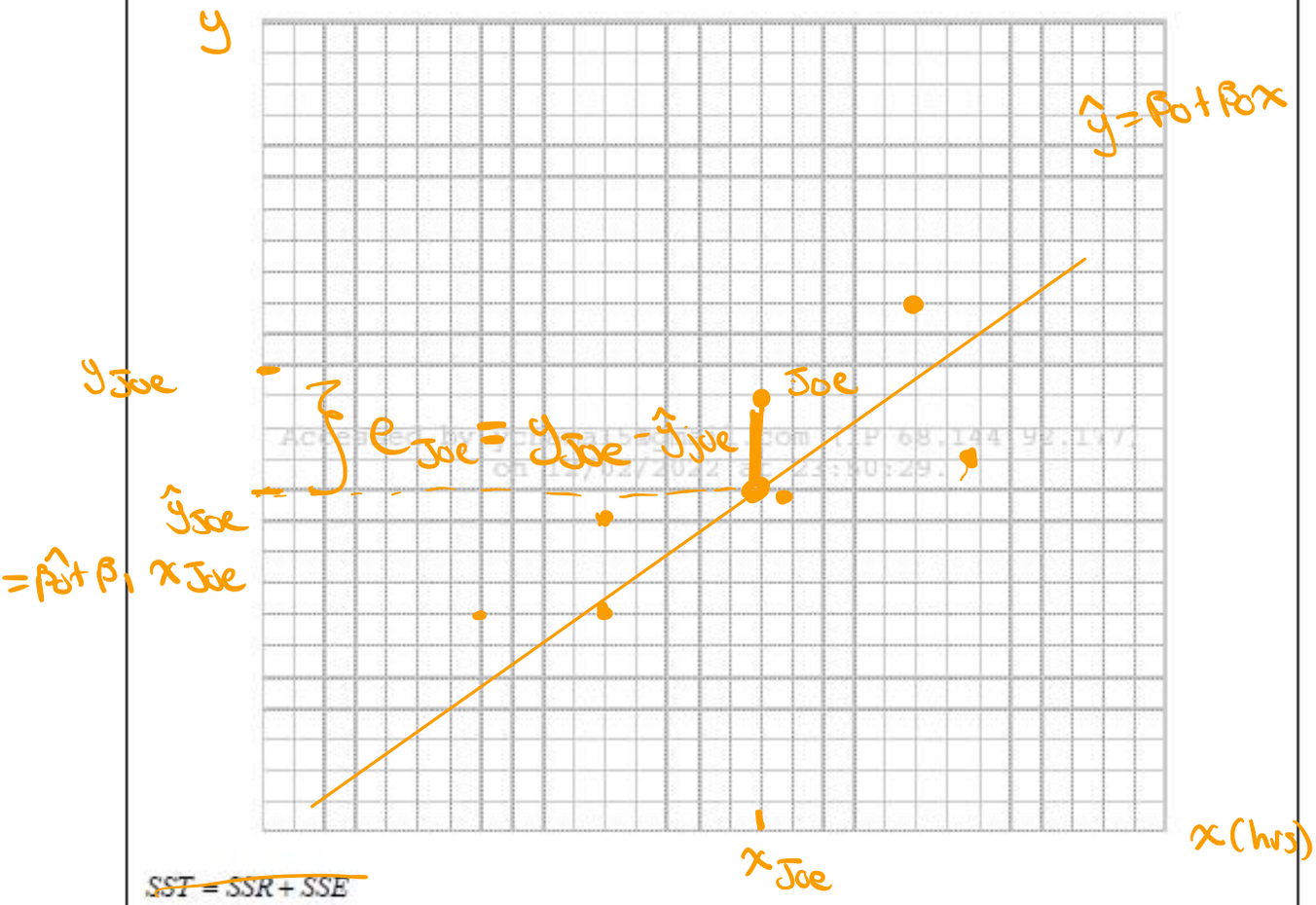
$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} = \frac{\sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y}}{\sum_{i=1}^{n} x_i^2 - n\bar{x}^2}$$

$\hat{\beta}_0 =$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

So how good is our model?

Let's first draw a scatterplot with a line fitted to it and try to analyze the different types of deviations.



~~$SST = SSR + SSE$~~

~~$\sum (y_i - \bar{y})^2 = \sum (\hat{y}_i - \bar{y})^2 + \sum (y_i - \hat{y}_i)^2$~~

So what do you think the ratio $R^2 = \frac{SSR}{SST}$ captures? If you can see what SSR and SST capture, you should be able to come up with your own logical interpretation of this ratio. In fact R-squared is close related to the correlation between x and y.

lets say we didn't know Joes mark, but we did know the # of hrs that joe studied for the exam. the predicted vale will be $\hat{y}_{Joe}$ (cause we have fitted Joes value)

our prediction won't be perfect then will be an error in our prediction.

↳ the diff between Joes actual mark and his predicted mark is called the residual.

Residual and SSE (Sum of Squared Errors)
These residuals will allow us to estimate the standard deviation of ε (ie. σ). We also call this the standard deviation of the model because it captures how much each datum deviates from the regression line.

Residual: $e_i = y_i - \hat{y}_i$

The estimate of error standard deviation σ is:

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

Standard Errors of slope and intercept
s^2 allows us to find the standard error of $\hat{\beta}_0$ and $\hat{\beta}_1$ (ie. their standard deviations).

The standard deviations of $\hat{\beta}_0$ and $\hat{\beta}_1$ are estimated with

$$s_{\hat{\beta}_0} = s\sqrt{\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^{n}(x_i-\bar{x})^2}} = s\sqrt{\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^{n} x_i^2 - n\bar{x}^2}}$$

$$s_{\hat{\beta}_1} = \frac{s}{\sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2}} = \frac{s}{\sqrt{\sum_{i=1}^{n} x_i^2 - n\bar{x}^2}}$$

We need these values to perform hypothesis tests and create confidence intervals on the Betas.

Equipped with these standard errors, we can now do inference on β_0 and β_1 just like we did inference on μ and p with their respective standard errors. Remember?!

The quantities $\frac{\hat{\beta}_0-\beta_0}{s_{\hat{\beta}_0}}$ and $\frac{\hat{\beta}_1-\beta_1}{s_{\hat{\beta}_1}}$ have student's *t*-distribution with $n-2$ degrees of freedom. These look familiar?

6.3 Inference on Betas

Confidence Intervals

Level $100(1-\alpha)\%$ confidence intervals for β_0 and β_1 are given by

$$\hat{\beta}_0 \pm t_{n-2,\alpha/2} \cdot s_{\hat{\beta}_0} \text{ and } \hat{\beta}_1 \pm t_{n-2,\alpha/2} \cdot s_{\hat{\beta}_1}$$

A level $100(1-\alpha)\%$ confidence interval for the quantity $\beta_0 + \beta_1 x$ is given by

$$\hat{\beta}_0 + \hat{\beta}_1 x_i \pm t_{n-2,\alpha/2} \cdot s_{\hat{y}}$$

where,

$$s_{\hat{y}} = s\sqrt{\frac{1}{n} + \frac{(x-\bar{x})^2}{\sum_{i=1}^{n}(x_i-\bar{x})^2}} = s\sqrt{\frac{1}{n} + \frac{(x-\bar{x})^2}{\sum_{i=1}^{n}(x_i-\bar{x})^2}}$$

6.4 Inference on Correlation

Inference on Correlation

When there is no relationship between the two variables, $\rho = 0$.

So we test the hypotheses

$H_0: \ \rho = 0$

$H_1: \ \rho \neq 0$.

The test statistic is $t = r\sqrt{\frac{n-2}{1-r^2}}$

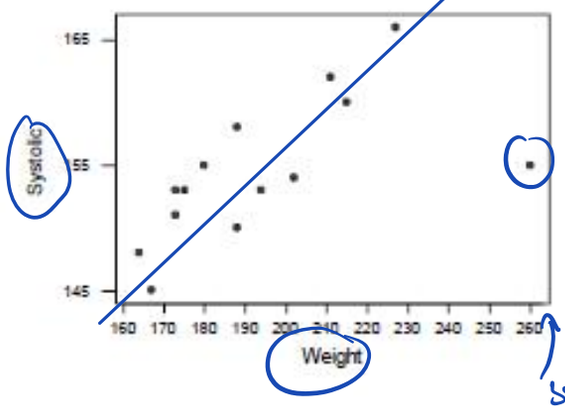
The test statistic is Student t distributed with $n - 2$ degrees of freedom provided that the variables are bivariate normally distributed.

6.5 Practice Questions

Q1. Systolic blood pressure readings of individuals are thought to be related to age and weight. For a sample of 14 men (of similar height), the following data were obtained:

Row	Systolic (y)	Weight (x)	xy	y^2	x^2	(y-y_bar)^2	(x-x_bar)^2
1	148	164	24,272	21,904	26,896	42.25	904.29
2	153	194	29,682	23,409	37,636	2.25	0.01
3	162	211	34,182	26,244	44,521	56.25	286.58
4	154	202	31,108	23,716	40,804	0.25	62.86
5	160	215	34,400	25,600	46,225	30.25	438.01
6	150	188	28,200	22,500	35,344	20.25	36.86
7	158	188	29,704	24,964	35,344	12.25	36.86
8	166	227	37,682	27,556	51,529	132.25	1084.29
9	151	173	26,123	22,801	29,929	12.25	444.01
10	155	180	27,900	24,025	32,400	0.25	198.01
11	145	167	24,215	21,025	27,889	90.25	732.86
12	153	173	26,469	23,409	29,929	2.25	444.01
13	155	260	40,300	24,025	67,600	0.25	4346.58
14	153	175	26,775	23,409	30,625	2.25	363.72
Total	**2163**	**2717**	**421,012**	**334,587**	**536,671**	**404**	9,379

y_bar	x_bar
154.5	194.07



linear relationship

$\hat{\beta}_0$ $\hat{\beta}_1$, r

$F(x_i, y_i)$

outlier.

super heavy guy that does not have high blood pressure.

a) Calculate the correlation coefficient between systolic blood pressure and weight.

$$r = \frac{\sum xy - n\bar{x}\bar{y}}{\sqrt{\sum x^2 - n\bar{x}^2}\sqrt{\sum y^2 - ny^2}}$$

0.635 > 0 (there is linear tve correlation)

$$= \frac{421,012 - (14)(194.07)(154.5)}{\sqrt{536,671 - 14(194.07)}\sqrt{339,587 - 14(154.5)^2}} = \frac{1239.59}{731.2 \cdot 576}$$

730 420,946.108

$$= 0.635$$

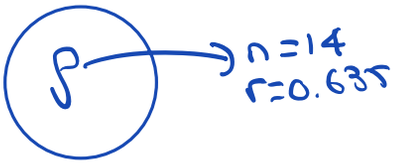
b) Calculate the proportion of the variance in systolic blood pressure that is explained by the regression model. (y)

r^2 = coefficient of determination

↳ it captures the amount of variance in y that is explained by x.

$r^2 = (0.635)^2 = 0.403$

The prof likes doing CI and hypothesis testing on ρ



c) Find the 95% Confidence Interval for the correlation between systolic blood pressure and weight.

chapt 7.1

we want to build a 95% CI for ρ

step 1: Compute w

$$w = \frac{1}{2}\ln\left(\frac{1+r}{1-r}\right) = \frac{1}{2}\ln\left(\frac{1+0.635}{1-0.635}\right) = 0.75$$

1.635

0.365

$\alpha = 1 - 0.95$

$\alpha = 0.05$

$\frac{\alpha}{2} = 0.025$

step2: Compute at 95 CI μ_w is $w \pm z_{\alpha/2}\sigma_w$

$$\sigma_w = \sqrt{\frac{1}{n-3}} = \sqrt{\frac{1}{11}} = 0.3015$$

$0.75 \pm 1.96(0.3015)$ $\therefore 0.159 < \mu_w < 1.341$

d) Is there evidence of a positive linear relationship between systolic blood pressure and weight? (ie. Can you conclude that $\rho > 0$)? Test at 5% level of significance.

step 3:

$$\text{CI for } \rho = \frac{e^{2\mu_w} - 1}{e^{2\mu_w} + 1}$$

1.3793

$$\frac{e^{2(0.159)} - 1}{e^{2(0.159)} + 1} < \rho < \frac{e^{2(1.341)} - 1}{e^{2(1.341)} - 1}$$

$$\frac{0.374}{2.379} < \rho < \frac{13.61}{15.61}$$

$$0.157 < \rho < 0.87$$

d) Is there evidence of a positive linear relationship between systolic blood pressure and weight? (ie. Can you conclude that ρ>0)? Test at 5% level of significance.

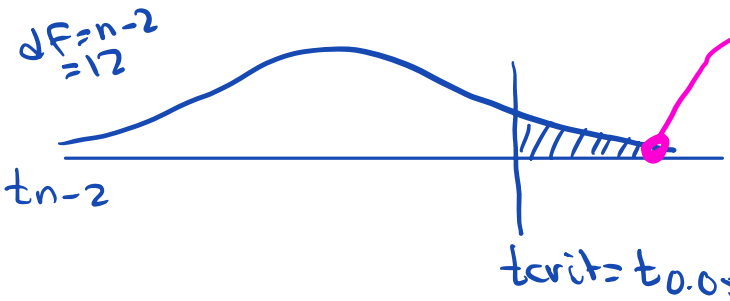
1) $H_0: \rho = 0$ or $\rho \leq 0$

$H_a: \rho > 0$ ⟶ positive relationship between x & y

2) calculate a test statistic and the test statistic is given by

$$t_{Test} = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}} = \frac{0.635\sqrt{14-2}}{\sqrt{1-(0.635)^2}} = 2.847$$

our test



3) compare that with the critical value of t distribution table.

$t_{crit} = t_{0.05,12} = 1.782$

Our test static lies in the rejection region ∴ we reject the null in and conclude that there is the relation between blood pressure and weight.

∴ we reject H_0.

when you do a test for ρ, where your test is now to see if $\rho > 0$ we have another test.

e) Can you conclude that $\rho > 0.5$? Test at 5% level of significance.

f) Compute the least-squares regression line for predicting systolic blood pressure (y) from weight (x).

Here it is asking you to compute what is your linear regression line

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$$

$$\hat{\beta}_1 = \frac{\sum xy - n\bar{x}\bar{y}}{\sum x^2 - n\bar{x}^2} = 0.1317 = \frac{421,012 - (14)(194.07)(154.5)}{(536,671) - (14)(194.07)^2}$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} = 128.93$$

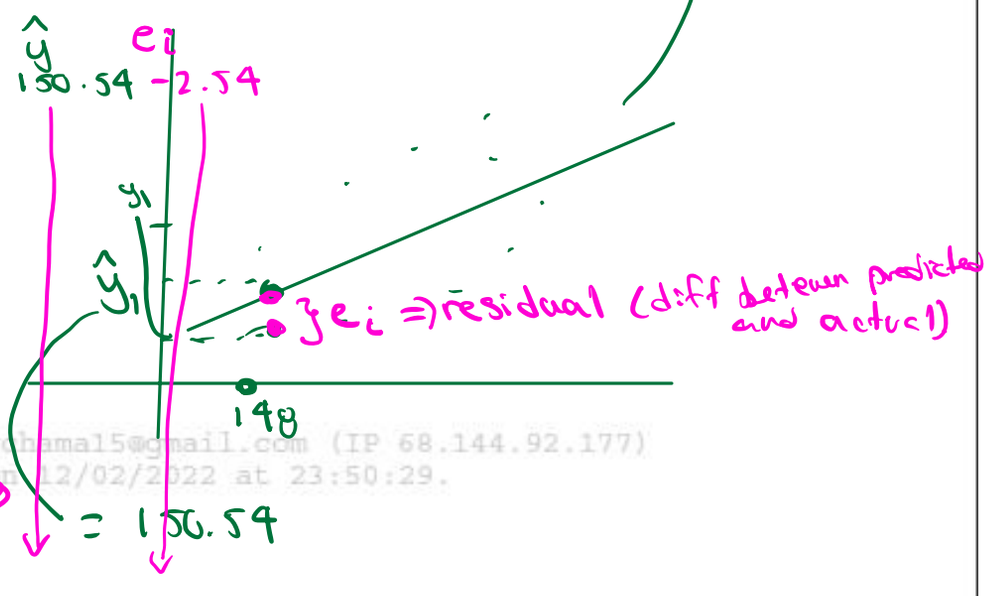
$$\frac{1238.59}{9386.69} = 0.13$$

$$\hat{y} = 128.93 + 0.1317x$$

Predicted values

g) Compute the fitted values and the residuals. Which point has the biggest residual?

Row	Systolic (y)	Weight (x)
1	148	164
2	153	194
3	162	211
4	154	202
5	160	215
6	150	188
7	158	188
8	166	227
9	151	173
10	155	180
11	145	167
12	153	173
13	155	260
14	153	175
Total	2163	2717



$e_{13} = -8.18 \Rightarrow$ biggest residual.

CI & HT on β; β_0, β_1

e) Can you conclude that ρ >0.5? Test at 5% level of significance.

$H_0: \rho = 0.5$ or $\rho \leq 0.5$ ⇒ "no strong relationship between x & y"

$H_a: \rho > 0.5$ ⇒ "strong relationship between x & y"

what we are interested in is whether ρ is a # that is greater than 0.5.

step 1) Assume H_0

step 2) compute the test statistic: $\boxed{r = 0.635}$

$$Z_{test} = \frac{W - \mu_W}{\sigma_W} \quad \text{where} \quad W \sim N(\mu_W, \sigma_W)$$

Here we need a value for W, μ_W & σ_W

$$W = \frac{1}{2}\ln\left(\frac{1+r}{1-r}\right) = \frac{1}{2}\ln\left(\frac{1+0.635}{1-0.635}\right) = 0.75$$

$$\mu_W = \frac{1}{2}\ln\left(\frac{1+\rho_0}{1-\rho_0}\right) = \frac{1}{2}\ln\left(\frac{1+0.5}{1-0.5}\right) = 0.5493$$

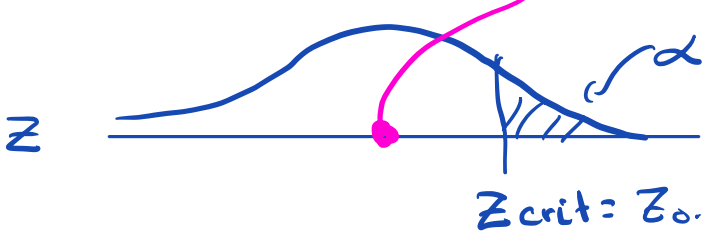
↑ That's the value of ρ_0 that was hypothesized to be true

$$\sigma_W = \sqrt{\frac{1}{n-3}} = 0.3015$$

$$Z_{test} = \frac{0.75 - 0.5493}{0.3015} = 0.67$$

∴ Fail to reject H_0

1 tail test H_a pts ⟶ right



$Z_{crit} = Z_{0.05} = 1.645$

There is insufficient evidence suggesting that $\rho_0 > 0.5$

$H_0: \rho = 0$

if $\rho > \rho_0$, then Z test CI for ρ.

if this $\rho > 0$, then t test ← used ∪

h) Compute the error standard deviation estimate s.

$$s = \sqrt{\frac{\sum e_i^2}{n-2}}$$ ← added up all the residuals square it then find average it out.

This captures on average how big your residuals is.

$$= \sqrt{\frac{(1-r^2)(\sum y_i - \bar{y})^2}{n-2}}$$

4.48

i) Construct a 95% confidence interval for the slope.

95% CI for β

$$\hat{\beta}_1 \pm t_{\alpha/2, n-2} \cdot s_{\hat{\beta}_1}$$ ← standard error.

$\hat{\beta}_1 = 0.1317$

$t_{0.025, 12} = 2.179$

$$= \frac{s}{\sqrt{\sum (x_i - \bar{x})^2}} = \sqrt{\frac{s}{\sum x^2 - n\bar{x}^2}} = 0.0322$$

$= 0.1317 \pm 2.179(0.0322)$

$= (0.062, 0.202)$ ⇒ excludes $\beta_1 = 0$ → we have evidence that β_1 is a true value > 0 and if $\beta_1 > 1$ ⇒ that there is a true linear relationship between weight and blood pressure

j) Find a 95% confidence interval for the mean systolic blood pressure for someone who weighs 180lbs.

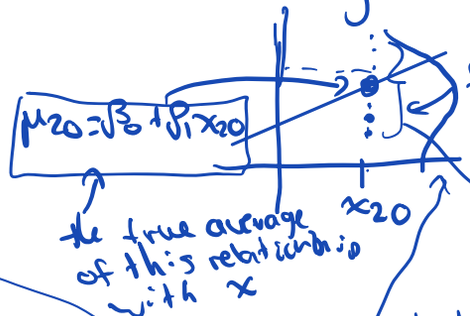
$\beta_0 + \beta_1 x$

This eqn allows us to find the CI for the avg blood pressure for everyone who has a weight of x

$\beta_0 + \hat{\beta}_1 \bar{x} \pm t_{n-2,\alpha/2} \cdot s\hat{y}$

we want to find a CI not for everybody but those ppl who weigh 180lbs

$\mu_{180} = \beta_0 + \beta_1(180)$



the true average of this relationship with x

for everyone who studies 20 hrs they can get a grade here There will be dispersion

what we assume in a linear regression model everyone who studied 20 hrs there is actually a Normal distribution

k) A nurse believes that β_1 is greater than 0. Can we conclude her claim is true? Test at 5% level of significance.

95% CI for $\mu_{180} = \beta_0 + \beta_1(180)$

we want 95% CI for this population parameter.

$\hat{\beta}_0 = 129.93$

$\hat{\beta}_1 = 0.1317$

$\hat{\beta}_0 + \hat{\beta}_1(180) \pm t_{\alpha/2, n-2} \cdot s\hat{y}$

$t_{\alpha/2, n-2} = 2.179$

$s\hat{y} = s\sqrt{\frac{1}{n} + \frac{(x-\bar{x})^2}{\sum(x-\bar{x})^2}} = 1.3628$

$= 136.64 \pm 2.179(1.3628)$

$= (149.67, 155.61)$

what we have done here is constructed a CI for Blood pressure but by taking advantage of the linear relationship between x and y we built a CI for just those ppl who weigh 180

$\mu \Rightarrow \mu_{180}$. we can do inference on those who weigh 180.

k) A nurse believes that β_1 is greater than 0. Can we conclude her claim is true? Test at 5% level of significance.

H_0 $\beta_1 = 0$ or $\beta_1 \leq 0$

H_a: $\beta_1 > 0$

1) Test statistic

$$t_{test} = \frac{\hat{\beta}_1 - \beta_1}{s_{\hat{\beta}_1}}$$

$$= \frac{0.1317 - 0}{0.0322} = 4.09$$

df = n − 2 = 12

$\alpha = 0.05$

$t_{0.05,12} = 1.782$

∴ we reject H_0

we have concluded using the CI that $\beta_1 > 0$

The data agrees that ppl who are heavier have higher blood pressure.

Aside

$$Z_{test} = \frac{\bar{x} - \mu_0}{\sigma_{\bar{x}}}$$

$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$

$$t_{test} = \frac{\bar{x} - \mu_0}{s_{\bar{x}}}$$

$s_{\bar{x}} = \frac{s}{\sqrt{n}}$

$$Z_{test} = \frac{(\bar{x}_1 - \bar{x}_2) - \Delta_0}{\sigma_{\bar{x}_1 - \bar{x}_2}}$$

$\sigma_{\bar{x}_1 - \bar{x}_2} = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$

t distribution is a wider version of the n distribution

as n gets bigger the orange begins to look like the blue