Week 1

Collecting & Exploring Data

What is data and why do we need it?

Information that can be used to recognize patterns, analyze outcomes, make decisions, and understand the world around us.

What is the study of statistics?

Statistics teaches us how to make informed judgements or discussions in the presense of uncertainty + variation

Motivating Example (Scenario 1)

You have been hired as a consultant by Benton County to study local power usage. Your task is to collect information about and then summarize household power consumption, measured in BTUs. There are approximately 37,000 households in Benton County.

↓ populations

Data, Individuals, and Variables

An individual/observational unit is an object described by data. Ex. single house hold

A characteristic of an individual is a variable. eg. consumption

A data set describes a set of individuals and their variable values.

Variable types:

Categorical: based on a quality asigned to a category

Quantitative: numerical, measured quantity

Populations, Samples, and Inference

The population is the complete collection of subjects or things in which we are interested.

ex: All 37,000 households in Benton County

A sample is a subset of the population.

eg. 90 selected households from our populations

Statistical inference is the process of using known sampled information to form a conclusion about unknown population characteristics.

Parameters & Statistics

Parameters → Population

- A parameter is: characteristic of population

  eg. Average power consumption for all in households in Benton County in the month of April

- Parameters are often unknown

- Parameters are estimated from a statistic.

Statistics → Sample

- A statistic is: characteristic of the sample

  eg: Average power consumption for the 90 sampled households in the months of April

- Used to estimate parameter values.

Observational Studies

An observational study observes and collects information on units but does not attempt to change or influence the units. Information is collected based on an established factor. Data represents a finite observable population.

An observational study cannot establish a causal relationship.

An unaccounted for confounding variables, associated with both factor and response variables, may be the underlying cause of the relationship.

Random Sampling

- Using a random mechanism is the best way to reduce bias in a sample.

- Bias is the tendency to systematically favor certain parts of the population over others.

Random Sampling Designs [SRS]

n represents sample size

Simple Random Sampling

Every combination of n individuals has an equal likelihood of being chosen for the sample.

Scenario 1: $n = 90$,

Assign a number to each household in Benton County and then use a random number generator to select 90 households

Stratified Random Sampling

Population is broken into groups or strata.

based on a shared characteristic within each strata, perform a simple random sample

Scenario 1: Suppose there 10 zip codes in Benton Conty, Randomly select 9 households from each zip code so that $n = 90$

Systematic Random Sampling

All units in population are listed. Then randomly generate a starting point on the list.

Then choose every Kth element on the list so that the desired sample size is obtained.

Scenario 1:

Crate a list of 37,000 hoseholds organized alphabetically by address. Randomly generate a starting point (eg. # 4241) then select every 411th household on the usr to reach $n = 90$

Bias & Poor Sampling Design

- A convenience sample is a sample in which sampled observations are the easiest to obtain.

- A voluntary sample is special case of convenience sampling. Participants are not selected, instead they volunteer their response, i.e. an online poll.

Data from poor sampling designs are most likely biased and should not be used to make inference. They may lead to overestimating or underestimating certain characteristics of the population.

Designed Experiments

How does an experiment differ from an observational study?

Individuals are randomly assigned to the treatments

Motivating Example (Scenario 2)

A pharmaceutical company is interested in assessing whether taking daily aspirin reduces the risk of heart attack. 1,500 individuals over the age of 55 have agreed to participate in the company's study. Of the 1,500 participants, 550 report being at-risk for heart disease based on family medical history.

Principles of Experimental Design

- Controlling

control for differences among treatmart groups
eg. all patients take pill with some amout of water even the non-asprin group

- Randomization

participants are randomly assigned to treatment groups
eg: patients are randomly assigned to receive either asprin or no asprin

- Replication

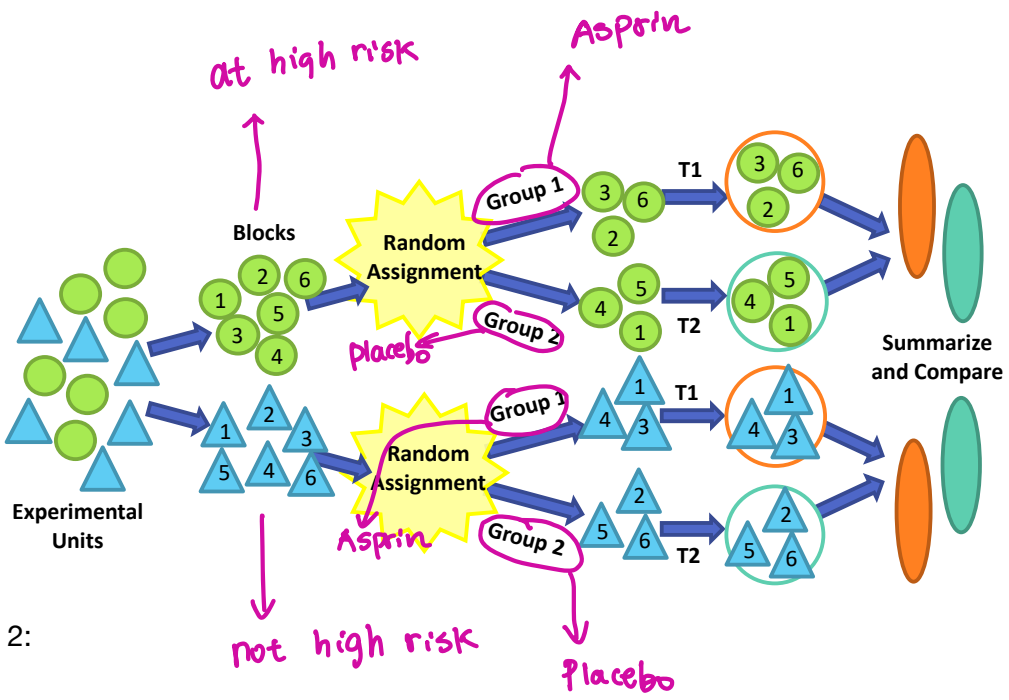
multiple individuals receive the same treatments to get a more accurate understandy of treatment effects
eg. multiple patients in each treatment groups

- Blocking

known variables may affect the response so we can break individuals into blocks then happy randomization

ex: Block by whether patient is at high risk for nearest disease

Completely Randomized Block Design



Scenario 2:

Visualizing Quantitative Data

Quantitative Variables: Histograms

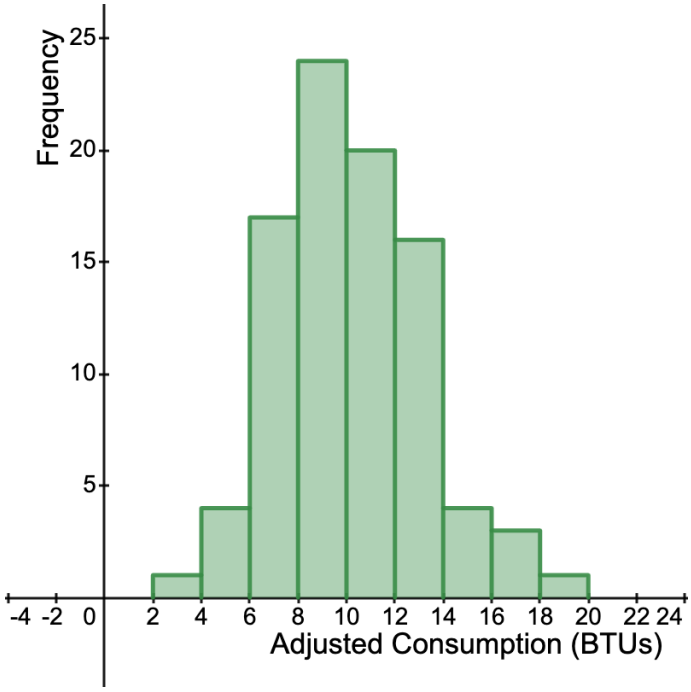
- Bins divide data into groups.
- Number of observations in each bin represented by a bar.
- Bar height is on vertical axis and variable range is on horizontal axis.

Scenario 1: A simple random sample of 90 gas-heated homes in Benton County yielded the following adjusted consumption data.

Adjusted consumption = consumption during given period/(weather in degree days x house area)

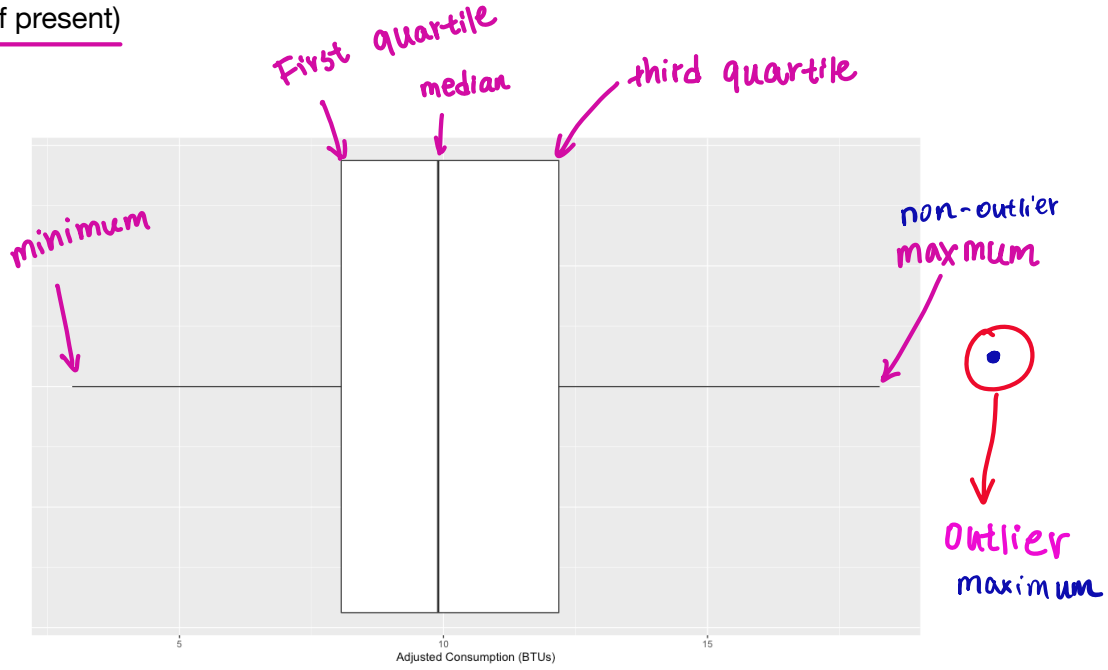
2.97	4.00	5.20	5.56	5.98	6.35	6.62	6.72	6.78	6.80
6.85	6.94	7.15	7.16	7.23	7.29	7.62	7.62	7.69	7.73
7.87	7.93	8.00	8.26	8.29	8.37	8.47	8.54	8.58	8.61
8.67	8.69	8.81	9.07	9.27	9.37	9.43	9.52	9.58	9.60
9.76	9.82	9.83	9.83	9.84	9.96	10.04	10.21	10.28	10.28
10.30	10.35	10.36	10.40	10.49	10.50	10.64	10.95	11.09	11.12
11.21	11.29	11.43	11.62	11.70	11.70	12.16	12.19	12.28	12.31
12.62	12.69	12.71	12.91	12.92	13.11	13.38	13.42	13.43	13.47
13.60	13.96	14.24	14.35	15.12	15.24	16.06	16.90	16.95	18.26

Bins	Frequency
[2,4)	1
[4,6)	4
[6,8)	17
[8,10)	24
[10,12)	20
[12,14)	16
[14,16)	4
[16,18)	3
[18,20)	1



Quantitative Variables: Boxplots

Displays:
- Minimum
- First quartile
- Median
- Third quartile
- Maximum
- Outliers (if present)



Describing Distributions

Shape: overall pattern of data
- symmetric or skewed
- unimodal, bimodal, multimodal

Center:

mean & and median

Spread:

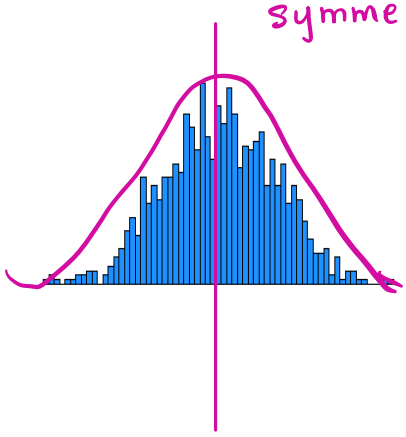
Standard deviation
range or Interquatile range

Outliers:

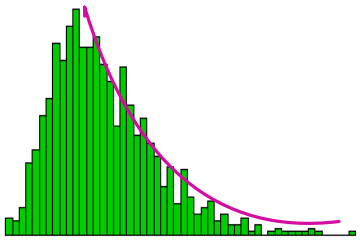
observation that fall outside of pattern of the data

extreme values that fall far from the center

Skewed vs. Symmetric



have two maxmums

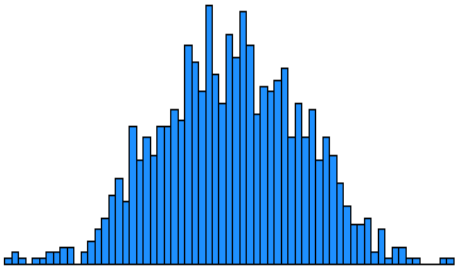


right-skewed

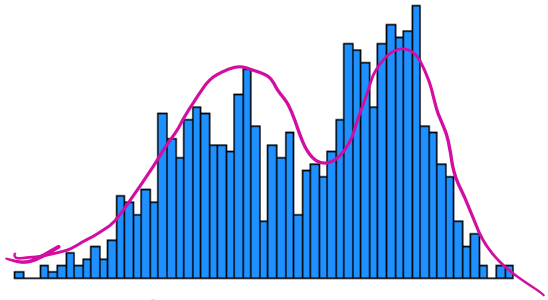
Unimodal, Bimodal, Multimodal

have multiple maxmums

have one maximum



Unimodal



Bimodal

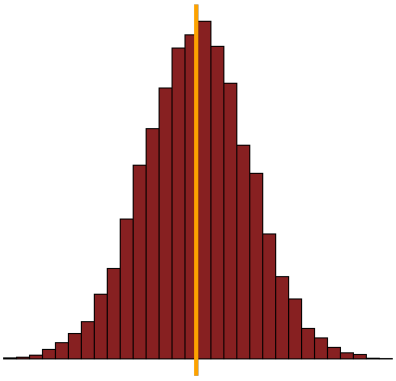
Summarizing Quantitative Data

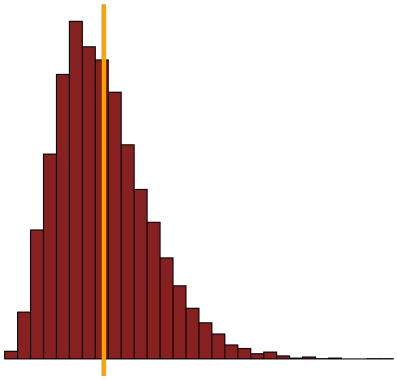
Center: Mean / Average

The mean is the average or expected value of a quantitative characteristic of interest. The mean equally weights each observation value in the set.

$$\bar{x} = \frac{x_1 + x_2 + x_3 + \cdots + x_n}{n} = \frac{\sum_{i=1}^{n} x_i}{n}$$

$\bar{x}$: sample mean





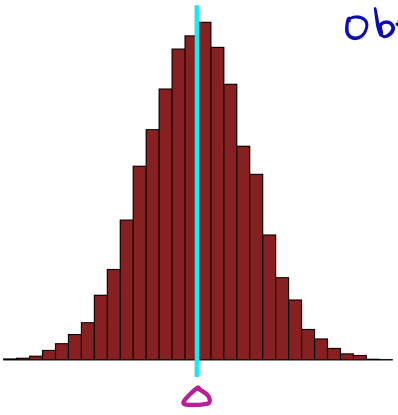
Center: Median

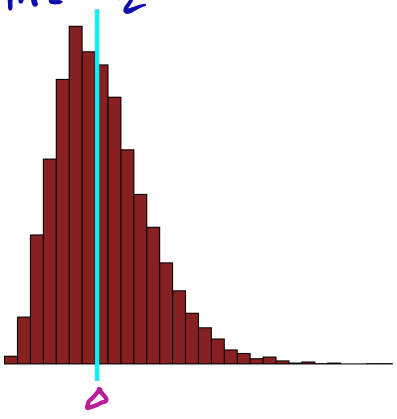
The median is the value in which approximately 50% of observations fall below.

The sample median denoted by M represents the middle observation when data is ordered from smallest to largest.

If n is odd: $\left(\frac{n+1}{2}\right)$ th observation

If n is even: Average of $\left(\frac{n}{2}\right)^{th}$ and $\left(\frac{n}{2}+1\right)^{th}$ observations





Comparing Mean & Median

When data is skewed or has outliers the mean will be influenced by the outliers or extreme observations.

The Median is a better representation of a typical value for skewed data.

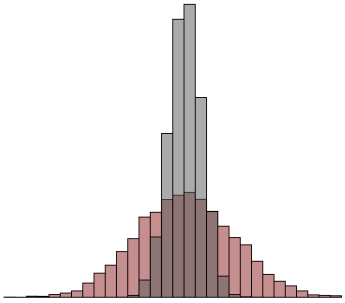
Spread: Standard Deviation

Sort of

The standard deviation is a measurement for the amount of spread or variability of a quantitative random variable.

The sample standard deviation, s, describes the typical deviation from $\bar{x}$ in a sample.

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



The standard deviation is highly influenced by skewed data and extreme observations.

Spread: Quartiles & IQR

median

Q_1 ↓

The values that cut the data into four quarters are the quartiles which are the 25th, 50th, and 75th (Q_3) percentiles.

The 1st quartile, Q1, and 3rd quartile, Q3, are the respective medians of the first and second half of the data.

The Interquartile Range (IQR) is a measurement for variability that is resistant to extreme observations or outliers.

The IQR is the range for the middle 50% of the data.

IQR = $Q_3 - Q_1$

Determining Outliers

An outlier is an extreme observation that falls far from the center of the data.
The $1.5 \times$ IQR rule for outliers determines that any point outside the cutoff is an outlier.

lower cutoff $= Q_1 - 1.5 \times IQR$ Any values less than this cutoff are outliers

upper cutoff $= Q_3 + 1.5 \times IQR$ Any values greater than this cutoff are outliers.

Summarizing Data: Scenario 1

2.97	4.00	5.20	5.56	5.98	6.35	6.62	6.72	6.78	6.80
6.85	6.94	7.15	7.16	7.23	7.29	7.62	7.62	7.69	7.73
7.87	7.93	8.00	8.26	8.29	8.37	8.47	8.54	8.58	8.61
8.67	8.69	8.81	9.07	9.27	9.37	9.43	9.52	9.58	9.60
9.76	9.82	9.83	9.83	9.84	9.96	10.04	10.21	10.28	10.28
10.30	10.35	10.36	10.40	10.49	10.50	10.64	10.95	11.09	11.12
11.21	11.29	11.43	11.62	11.70	11.70	12.16	12.19	12.28	12.31
12.62	12.69	12.71	12.91	12.92	13.11	13.38	13.42	13.43	13.47
13.60	13.96	14.24	14.35	15.12	15.24	16.06	16.90	16.95	18.26

R Code - create a vector of numerical values

```
consumption <-
c(2.97,4.00,5.20,5.56,5.98,6.35,6.62,6.72,6.78,6.80,6.85,6.94,7.15,7.16,7.23,7.29,7.62
,7.62,7.69,7.73,7.87,7.93,8.00,8.26,8.29,8.37,8.47,8.54,8.58,8.61,8.67,8.69,8.81,9.07,
9.27,9.37,9.43,9.52,9.58,9.60,9.76,9.82,9.83,9.83,9.84,9.96,10.04,10.21,10.28,10.28,10
.30,10.35,10.36,10.40,10.49,10.50,10.64,10.95,11.09,11.12,11.21,11.29,11.43,11.62,11.7
0,11.70,12.16,12.19,12.28,12.31,12.62,12.69,12.71,12.91,12.92,13.11,13.38,13.42,13.43,
13.47,13.60,13.96,14.24,14.35,15.12,15.24,16.06,16.90,16.95,18.26)
```

Mean: $\bar{x} = 10.141$ R Code: mean( )

Median: $M = 9.9$ median( )

Standard Deviation: $S = 2.925$ sd( )

Quartiles and IQR: $Q_1 = 8$, $Q_3 = 12.19$, $IQR = 4.19$ IQR(vector, type=1)

Outliers:

lower cutoff = $8 - 1.5(4.19) = 1.715$

upper cutoff = $12.19 + 1.5(4.19) = 18.475$

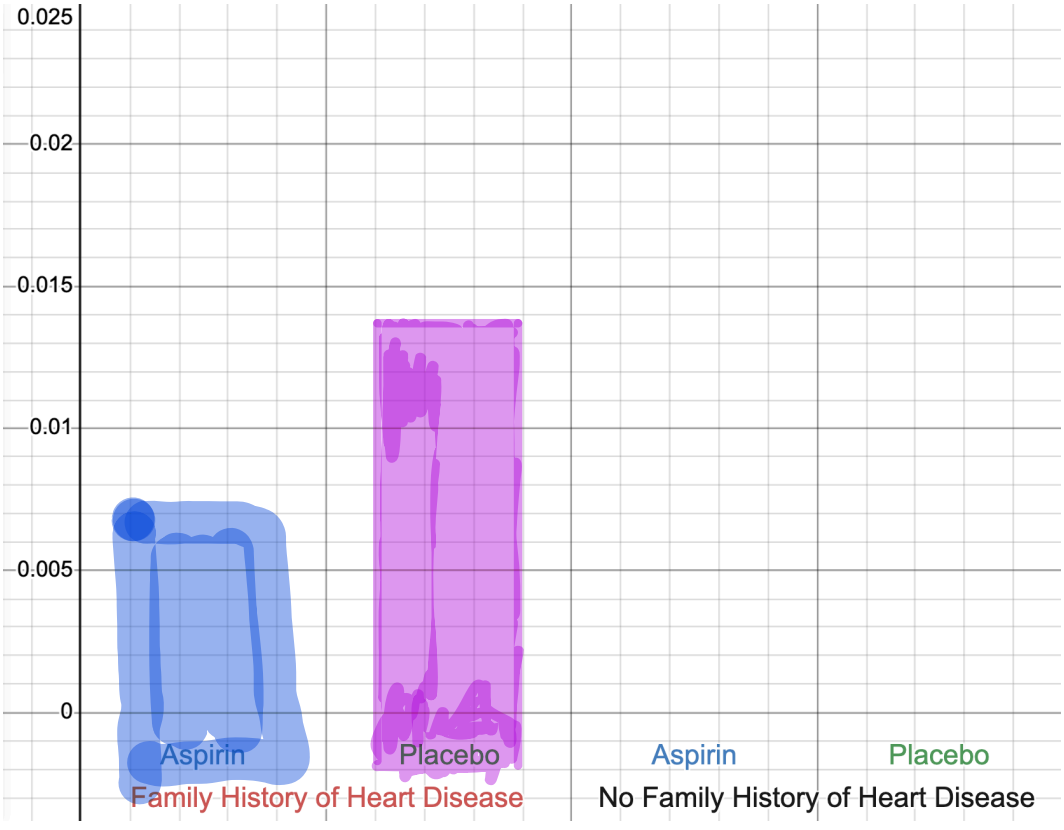
} no outliers

Categorical Variables: Proportions & Frequency Tables

- Categorical variables are numerically summarized using:
  proportions, counts, percentages
- A sample proportion ($\hat{p}$) represents the proportion of individuals in a sample within a category and is a special case of a sample mean.
  $\hat{p}$ = # of individuals in given category

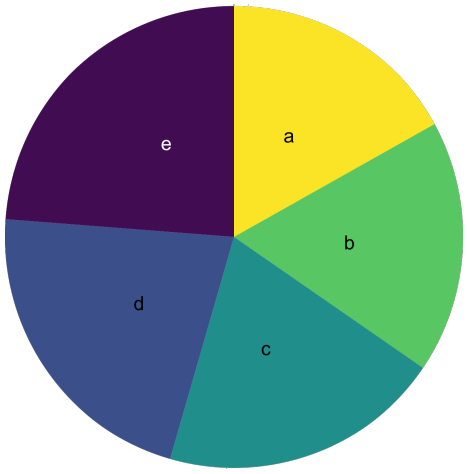
Scenario 2: In a study of 1,500 participants over 55 years of age, 550 individuals were identified as having a predisposition to heart disease based on family history. The remaining 950 participants had no known predisposition. Within the two family history groups, participants were randomly assigned to receive either daily Aspirin or a daily placebo pill throughout the study period. The table below shows the proportion of individuals within each group that suffered a heart attack during the study period.

	Aspirin	Placebo	Total
Family History of Heart Disease	2/275 = 0.0073	5/275 = 0.0182	7/550
NO Family History of Heart Disease	4/475	10/475	14/950
Total	6/750	15/750	21/1500



Categorical Variables: Pie Charts

Order the groups (a, b, c, d, e) from largest to smallest.



Week 2 Probability

OpenIntro 3.1 Defining Probability

- Probability
- Law of Large Numbers
- Disjoint (mutually exclusive)
- Addition Rule of Disjoint Outcomes
- Event
- General Addition Rule
- Probability Distribution
- Rules for a Probability Distribution
- Sample Space
- Complement
- Independent Events
- Multiplication Rule for Independent Events

ST 314

Class #4

OpenIntro 3.2 Conditional Probability

- Marginal probability
- Joint probability (intersection)
- Conditional probability
- General Multiplication Rule
- Sum of Conditional Probabilities
- Tree Diagram
- Bayes' Theorem

Probability

Introducing Ideas in Probability

Suppose we roll two fair six-sided dice.

1. How many possible outcomes are there for rolling the two dice? What are the outcomes?

36 outcomes

2. How many ways can we roll a double? the same number

6

3. What is the chance that the dice dots sum to seven?

$\frac{6}{36} = \frac{1}{6}$

Definitions

Sample space, S

set of all possible outcomes, $S = \{2, 3, 3, 4, 4, 4 \dots 12\}$

Event

↳ subset of sample space for which we are intrested in calculating the probability A = dice dots sum 7

Probability of event A

$P(A) = \frac{\text{size of event A}}{\text{size of sample space}}$

Basic Probability Models

$P(A) = \frac{6}{36} = \frac{1}{6}$

A table, function, or graph consisting of all events along with probabilities for each outcome is a probability model / distribution

X	2	3	4	5	6	7	8	9	10	11	12
P(X)	0.028	0.056	0.083	0.111	0.139	0.167	0.139	0.111	0.083	0.056	0.028

ST 314

Rules for Probability Distributions

Rule #1

Outcomes must be disjoint

Rule #2

each probability must with in interval [0,1]

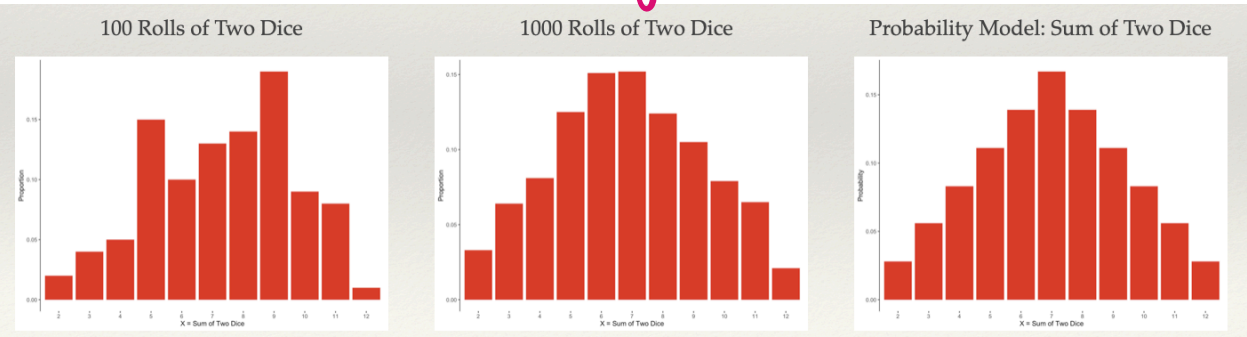
Rule #3

Probabilities of all outcomes sum to 1

Law of Large Numbers

The law of large numbers tells us As more observation are collected. the portion of occurrence of an outcomes will converge to the true probability of that outcome.

theoretical probability



theoretical prob model

Motivating Example (Scenario 3)

There are 9,250 undergraduates in the College of Engineering at OSU. Here is some information about those students:

- 336 have declared a civil engineering degree
- 4,297 have declared a computer science degree
- 549 have declared an electrical and computer engineering degree
- 1,265 have declared a mechanical engineering degree
- The remaining 2,803 have declared a different engineering degree or are undeclared
- 25.98% of engineering undergraduates identify as female
- 858 students identify as female and have declared a computer engineering degree (this last bullet point was made up by Erin)

Science

ST 314

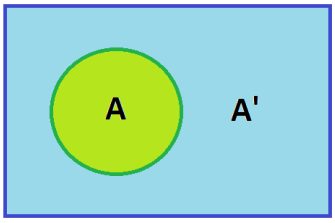
Example Problem

Suppose a couple tells you they have three children. What is the probability that exactly two of them are boys? (Assume the probabilities of having a boy and a girl are equally likely). A = exactly 2 of the 3 children boys

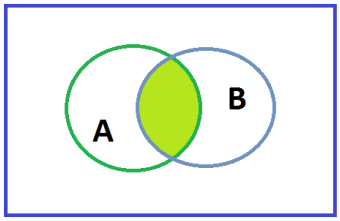
$P(A) = \frac{3}{8}$

B B B
B B G
B G B
G B B
G G B
B G G
G B G
G G G

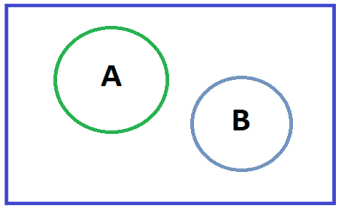
The Venn Diagram



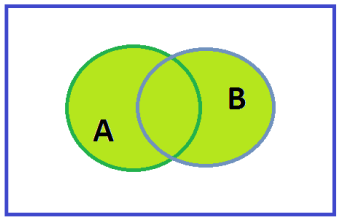
An event is the subset of all the possible outcomes in a random experiment or population called a sample space. The complement of an event A, denoted A', is the set of all outcomes in the sample space that are not contained in A.



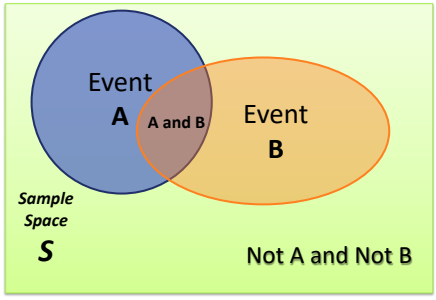
The Intersection of events A and B, denoted $A \cap B$ is the event consisting of all outcomes that are in both A and B. "A" and "B"



If A and B have no common outcomes, the events are said to be disjoint or mutually exclusive. Therefore, the intersection is an empty set, $A \cap B = \phi$.

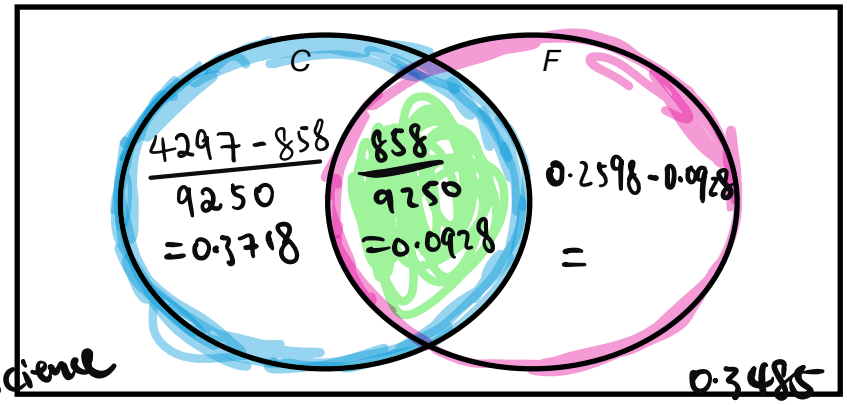


The union of the two events A and B, denoted $A \cup B$ is the event consisting of the elements in A, B, and both A and B. "A" or "B"



Example Problem (Scenario 3)
Suppose one engineering student is randomly selected from the population of 9,250 OSU engineering students.

Let C represent the event that a computer science student is selected. Let F represent the event that a student that identifies as female is selected. Draw a Venn diagram that represents the scenario.



Marginal prob of selecting a compter science major

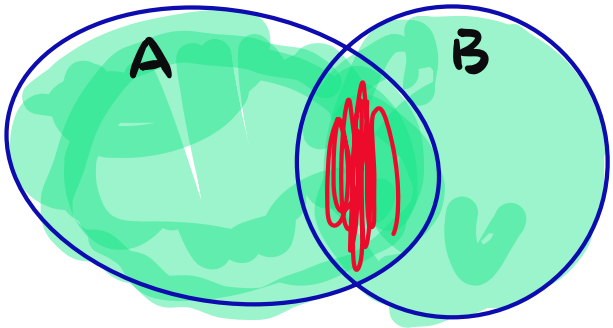
- $P(C) =$ $\frac{4297}{9250} = 0.4645$
- $P(F) =$ 0.2598 (covert 25.98% to porpotion)
- $P(C \cap F) =$ $\frac{859}{9250} = 0.0928$
- $P(C' \cap F') =$ $1 - (0.0928 + 0.3718 + 0.1470) = 0.3485$

The Addition Rule

The General Addition Rule calculates the probability of the union between two events A and B, the chance that either A or B or both occur.

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

If and only if A and B are disjoint, $P(A \cap B) = 0$, the rule simplifies to



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

Example Problem (Scenario 3)
Suppose an engineering student is randomly selected. What is the probability the student has declared a computer science degree or identifies as female?

That is, calculate $P(C \cup F)$ CS or female or both

$$P(C \cup F) = P(C) + P(F) - P(C \cap F)$$
$$= 0.4645 + 0.2598 - 0.0928$$
$$= 0.4315$$

Conditional and Independent Probabilities

You select a card from a standard deck of 52 cards. It's a queen. You decide to select a second card without replacing the first back to the deck.

a. What is the probability the second card is a queen?

3/51

b. Are the events described above disjoint?

NO

c. Are the events described above independent?

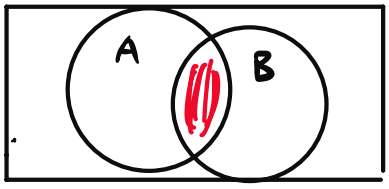
No → first draw would impact the second

Conditional Probability

For any two events, A and B, with $P(B) > 0$, the conditional probability of A given B has occurred is defined by:

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

"A given B"



Conditional probabilities redefine the sample space to include only events that are in the conditional subset B.

$P(C) = 0.4645$ $P(C \cap F) = 0.0928$

$P(F) = 0.2598$

ST 314

Example Problem (Scenario 3)
Suppose a student is randomly selected from the population of engineering students at OSU.

1. Given that the student identifies as female, what is the probability that the student has declared a computer science degree? That is, calculate $P(C \mid F)$.

$$P(C|F) = \frac{P(C \cap F)}{P(F)} = \frac{0.0928}{0.2598} = 0.3572$$

2. Given that the student has declared a computer science degree, what is the probability that the student identifies as female? That is, calculate $P(F \mid C)$.

$$P(F|C) = \frac{P(E \cap F)}{P(C)} = \frac{0.0928}{0.4645} = 0.1998$$

Definition of conditional: $P(A|B) = \frac{P(A \cap B)}{P(B)}$

Multiplication Rules

General multiplication rule:

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

Multiplication rule for independent events:

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

if and only if A and B are independent

A and B are ind $P(A|B) = P(A)$ $P(B|A) = P(B)$

Example Problems
a. You select one card from a deck of 52 cards, look at it, then replace it to the deck. You select another card and look at it. What is the probability that both cards selected are queens?

$$P(Q_1 \cap Q_2) = P(Q_1)P(Q_2) = \left(\frac{4}{52}\right)\left(\frac{4}{52}\right) = 0.0059$$

b. You select two cards from the deck at the same time. What is the probability that both cards are queens?

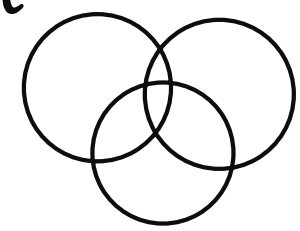
$$P(Q_1 \cap Q_2) = P(Q_1)P(Q_2|Q_1) = \left(\frac{4}{52}\right)\left(\frac{3}{52}\right) = 0.0045$$

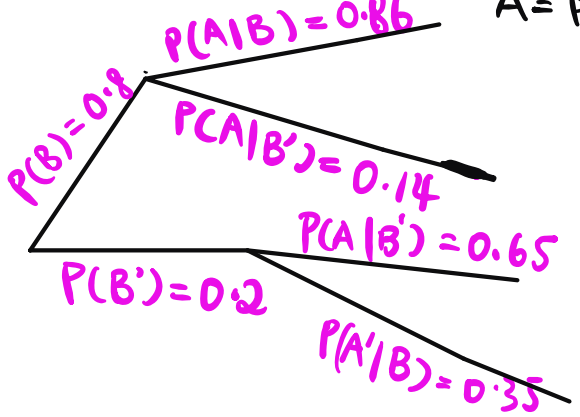
c. After an introductory statistics course, 80% of students can successfully construct a box plot. Of those who can construct box plots, 86% passed, while only 65% of those students who could not construct a box plot passed.

Draw a tree diagram of this scenario.

B= can construct box plot

A= passed class

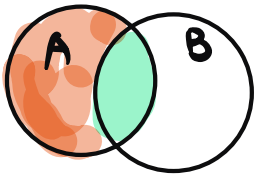




Law of Total Probability & Bayes Theorem

The law of total probability finds the probability that event A occurs is the sum of all the mutually exclusive events that contain A:

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



Bays theorem can be used to find a conditional probability when only other conditional probabilities are known. The formula is a combination of the Law of Total Probability and the Multiplication Rule.

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

(general multiplication rule; definition of conditional probability)

Example Problem (Exercise 3.19)
After an introductory statistics course, 80% of students can successfully construct a box plot. Of those who can construct box plots, 86% passed, while only 65% of those students who could not construct a box plot passed.

Calculate the probability that a student is able to construct a box plot if it is known that they passed the class.

$$P(B|A) = \frac{P(A|B)\,P(B)}{P(A|B)\,P(B) + P(A|B')\,P(B')}$$

$$= \frac{0.84(0.8)}{0.84(0.8) + 0.45(0.2)} = 0.841$$

Week 3 Random variables

OpenIntro 3.4 Random Variables

- Random variable
- Expected value
- Expected value of a discrete random variable
- Variance of a discrete random variable
- Linear combination
- Expectation of a linear combination
- Variance of a linear combination
- Probability Mass Function (not mentioned in OpenIntro text - see additional resources in this week's Canvas module)
- Probability Distribution Function (not mentioned in OpenIntro text - see additional resources in this week's Canvas module)

OpenIntro 4.3 Binomial Distribution

- Binomial distribution
- Success
- Probability of success
- Trials
- Probability mass function of a Poisson random variable

$P(x) = \binom{n}{x} p^x q^{n-x}$ → number of times for a specific outcomes within n trials; → number of trials; → prob of succ; → prob of failure on a single trial

---

OpenIntro 4.5 Poisson Distribution

- Poisson distribution
- Rate
- Probability mass function of a Poisson random variable

$P(X = x) = \frac{\lambda^x e^{-\lambda}}{x!}$ → 2.71828; → $\mu = \lambda$; (for $x = 1, 2, 3, \ldots$)

~~OpenIntro 3.5 Continuous Distributions~~

It is recommended that you read section 3.5 of the OpenIntro text; however, our discussion of continuous distributions will be at a slightly higher level than that presented in the text. Please see the additional resources in this week's module on canvas for detailed definitions and examples of the concepts below.

- Probability Density Function
- Cumulative Density Function
- Expectation of Continuous Random Variable
- Variance of Continuous Random Variable

OpenIntro 4.1 Normal Distribution → continuous D.

- Normal distribution
- Parameters of the Normal distribution
- Standard Normal Distribution
- Z-score
- Using R to find the tail area (to the right or left) of a z-score (*Note, the text mentions using probability tables to find the area under the normal curve. These tables are available in the book, but we will **never** use these tables in class. They are outdated and we can find more exact probability values using R.)*

$$P(a \le X \le b) = \int_a^b \frac{e^{-(x-\mu)^2/2\sigma^2}}{\sigma\sqrt{2\pi}}\,dx = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

μ → mean

Discrete Random Variables

Introduction to Random Variables

A random variable is a quantity that takes on a real number values for event in a sample space. Each value of a random variable has a likelihood of occurring. Letters (e.g. X and Y) are often used to denote RVs.

Types of Random Variables

Discrete
A random variable X is discrete if it has a countable set of values. X may include as many elements as there are whole numbers.

countable → whole number

There must be a gap between the values the variable can take on

Continuous
A random variable X is continuous if it takes on all values over an interval [redacted]. X has an infinite number of values between any two values

Discrete Random Variables

Probability Mass Function

The Probability Mass Function (PMF) is a probability distribution for a discrete random variable. The function assigns a probability to each possible value of *x* where

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

The following properties must hold:

$$0 \leq P(x) \leq 1 \qquad \sum P(x) = 1$$

↓ probability of all outcomes in the sample space must sum to 1

Motivating Example (Scenario 4)

Six lots of components are ready to be shipped by a certain supplier. The number of defective components in each lot is as follows:

Lot	1	2	3	4	5	6
Number of defectives	0	2	0	1	2	0

One of these lots is to be randomly selected for shipment to a particular customer. The customer will receive a lot with no defectives, one defective, or two defectives.

PMF

x = number of defective components received

x	0	1	2
p(x)	$P(0) = \frac{3}{6} = \frac{1}{2}$ $P(X=0)$	$P(1) = \frac{1}{6}$ $P(X=1)$	$P(2) = \frac{2}{4} = \frac{1}{3}$ $P(X=2)$

Cumulative Distribution Function

The Cumulative Distribution Function (CDF) for values of the discrete random variable X, F(X), is the probability that the random variable will be less than or equal to some real number x.

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

Example (Scenario 4)

Fill in the CDF table for scenario 4.

x	0	1	2
F(x)	$F(0) = \frac{1}{2}$ $F(X \le 0) = P(X=0)$	$F(1) = \frac{4}{6} = \frac{2}{3}$ $(X \le 1)$	$F(2) = 1 = \frac{1}{2} + \frac{1}{4} + \frac{1}{3}$ $F(X \le 2)$

Expectation

Let X be a discrete random variable with D possible outcomes and a PMF p(x).
The expected value is the value we would expect X to take on. This is the average or mean value, denoted by E(X) or μ_x.

$$E(X) = \sum_{i=1}^{D} x_i p(x_i)$$

Calculate the expected number of defectives in the shipment for the example in scenario 4.

$$E(x) = 0\left(\frac{1}{2}\right) + 1\left(\frac{1}{6}\right) + 2\left(\frac{1}{3}\right) = 0.8333$$

In the randomly received shipment, we expect 0.8333 defective components

Variance and Standard Deviation

The variance σ_x^2 is a measurement for variability or dispersion of X.

$$Var(x) = \sigma_x^2 = \sum_{i=1}^{n} (x_i - \mu_x)^2 p(x_i)$$

The standard deviation σ_x is the size of a typical deviation away from μ_x. The standard deviation has the same units as X and μ_x.

$$SD(x) = \sigma_x = \sqrt{Var(x)}$$

Calculate the standard deviation of the number of defectives in the shipment for the example in scenario 4.

$$\sqrt{(0-0.8333)^2\left(\tfrac{1}{2}\right) + (1-0.8333)^2\left(\tfrac{1}{4}\right) + (2-0.8333)^2\left(\tfrac{1}{3}\right)} = 0.8975$$

Linear Combinations

If W is a linear combination of X such that $W = aX + b$ where a and b are constants then the expectation and variance of W is:

$\mu_w = E(W) = E(aX + b) =$

$\sigma_w^2 = Var(W) = Var(aX + b) =$

Binomial Distribution (discrete)

only two possible outcomes in the trial

Models the number of successful outcomes out of n Bernoulli trials with a probability of success p.

- A RV that is modeled with a binomial distribution will have come from an experiment consisting of _n_ independent Bernoulli trials.

n → one of the parameters of Binomial distribution

- The outcome of interest is referred to as a success and the other outcome a failure.
- The probability of a "success" on a single trial is denoted by p ← the other parameter
- p is a constant that does not change from trial to trial.

Suppose we flip three coins, where X is the number of times we flip a heads. The random variable X is binomially distributed!

$n = 3 \quad p = 0.5$

A random variable X with a binomial distribution has probability mass function:

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

	TTT	HTT THT TTH	HHT HTH THH	HHH
x	0	1	2	3
$p(X = x)$	1/8	3/8	3/8	1/8

$p(x) = \binom{3}{x} 0.5^x (1-0.5)^{3-x}$

A combination calculates the number of ways x elements can be chosen from a sample of n total elements, order does not matter and elements cannot be repeated.

$$\binom{n}{x} = \frac{n!}{x!\,(n-x)!}$$

Binomial Expectation and Variance

A random variable X with a binomial distribution has expected value:

$$E(x) = np$$ → success

A random variable X with a binomial distribution has variance:

$$Var(x) = np(1-p)$$

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

Binomial Example

On a twenty question quiz, X is the number of questions answered correctly. Each question on the quiz is a multiple choice question with four possible answers. The chance a student guesses the correct answer is $1/4$. Each question is a Bernoulli trial.

A. What are the parameter values for the binomial distribution?

B. What is probability mass function of X?

C. How many questions should the student expect to get correct?

D. What is the standard deviation of X?

E. How likely is it that the student will get exactly 10 questions correct?

F. How likely is it that the student will get less than 20% of the questions correct?

Poisson Distribution Discrete

Models the number of occurrences of an event over an interval of space or time.
A RV that is modeled with a poisson distribution is defined by the rate parameter λ.

Example: Suppose X is the number of people that arrive at an emergency room in an hour. On average, 5 people arrive at the emergency room per hour.

$$P(x) = P(X=x) = \frac{\lambda^x e^{-\lambda}}{x!} \qquad x = 0, 1, 2, 3, \ldots$$

A random variable X with a Poisson distribution has probability mass function:

The constant e is defined by $e = \lim_{n\to\infty} \left(1 + \frac{1}{n}\right)^n \approx 2.7182.$

Poisson Expectation and Variance

A random variable X with a Poisson distribution has expected value:

$$E(x) = \lambda$$

A random variable X with a Poisson distribution has variance:

$$Var(X) = \lambda$$

$$SD(X) = \sqrt{\lambda}$$

Poisson Example

Suppose X is the number of people that arrive at an emergency room in an hour. On average, 5 people arrive at the emergency room per hour.

A. What is the parameter value for the Poisson distribution?

$\lambda = 5$ because "on average, 5 people arrive at the ER per hour"

B. What is probability mass function of X?

$$P(x) = P(X=x) = \frac{5^x e^{-5}}{x!}$$

what is probability that 10 people arrive at ER in an hour?

$$P(X=10) = P(10) = \frac{5^{10} e^{-5}}{10!}$$

R: = dpois(10,5)

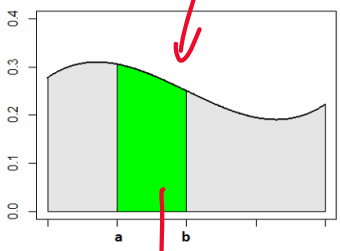
C. How many arrivals should the emergency room expect in an hour?

5

D. What is the standard deviation of X?

$$SD(X) = \sqrt{5}$$

Continuous Random Variables



For a continuous random variable X, the Probability Density Function (PDF), denoted by f(X), is a mathematical expression for the shape of the distribution of X. Denoted by f(X) we use the PDF to calculate probabilities between two values of X and expected values of X.

Probability Density Function (PDF)

The probability density function is a continuous function of the random variable X such that

$$P(a \le x \le b) = \int_a^b \underbrace{f(x)}_{\text{PDF}}\, dx$$

$f(x) \ge 0$

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

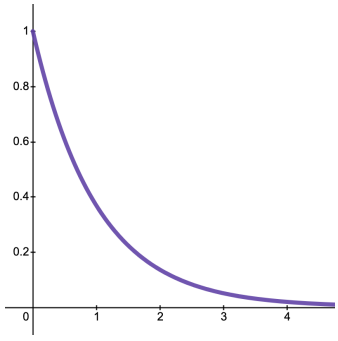
can change to the lower and upper bounds of the support of x

support: set of values for which the random variable is defined

Let X be the random variable that represents the number of minutes between customers who enter the drive thru of a fast food restaurant. Let's model the time between customers with the following probability density function. **Does the area under $f(x)$ equal 1?**

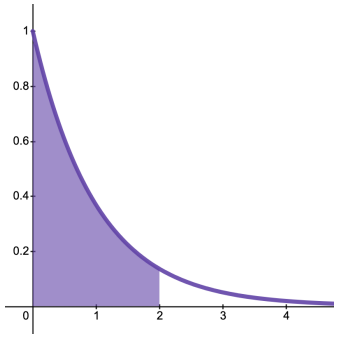
$f(x) = e^{-x}$ for $0 \le x$ (support)

$$f(x) = \begin{cases} e^{-x}, & x \ge 0 \\ 0, & x < 0 \end{cases}$$



$$\int_0^\infty e^{-x}dx = -e^{-x}\Big|_0^\infty = 0-(-1) = 1$$

A customer has just entered the drive thru. Use the PDF to determine the probability that the next customer will arrive within 2 minutes.



$$P(0 \le x \le 2) = P(0 < x < 2)$$

$$\int_0^2 e^{-x}dx = -e^{-x}\Big|_0^2 = -e^{-2}+1 = 0.8647$$

Cumulative Distribution Function (CDF)

For a continuous random variable X, the Cumulative Density Function (CDF) is used to calculate the probability the random variable X will be less than or equal to some real number x.

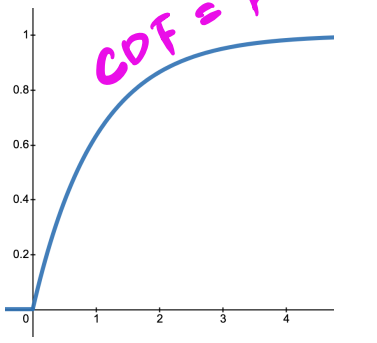
$$F(x) = P(X \le x) = \int_{-\infty}^{x} f(t)\,dt$$

$-\infty$ → can change to the lower bound of the support of x

Properties

- $0 \le F(x) \le 1$
- $P(a \le x \le b) = F(b) - F(a)$
- $P(x \le a) = P(x < a)$ because for a continuous RV, $P(x=a)=0$
- $\frac{dF(x)}{dx} = F(x)$

The random variable *X* has the follow probability density function. Determine its CDF.



$f(x) = e^{-x}$ for $0 \le x$

$$F(x) = \int_0^x e^{-t}\,dt = -e^{-t}\Big|_0^x$$

$$= -e^{-x} + 1 = 1 - e^{-x}$$

A customer has just entered the drive thru. Use the CDF to determine the probability that the next customer will arrive within 2 minutes.

CDF: $F(x) = 1 - e^{-x}$

$$P(x \le 2) = F(2) = 1 - e^{-2} = 0.8647$$

Expectation

Let X be a continuous random variable with PDF and possible outcomes from $-\infty$ to ∞.

The **expected value** is the value we would expect X to take on. This is the average or mean value, denoted by E(X) or μ_x.

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

change the lower and upper bounds of the support.

Variance

The **variance**, σ_x^2, is a measurement of how much spread there is in the distribution of X.

$$Var(X) = \int_{-\infty}^{\infty} (x - E(x))^2 f(x) = E(X^2) - (E(x))^2$$

$$E(X^2) = \int_{-\infty}^{\infty} x^2 f(x)\,dx$$

The **standard deviation**, σ_x, is the size of a typical deviation away from μ_x and has the same units as X and μ_x.

$$SD(x) = \sqrt{Var(x)} = \sigma_x$$

Expectation & Variance Example

Suppose X is a continuous random variable defined on the interval [0,1] with the following probability density function: $f(x) = 3x^2$.

support

Find the expectation, E(X), and the variance, Var(X), of X.

$$E(x) = \int_0^1 x \cdot 3x^2 dx = \int_0^1 3x^3 dx = \frac{3}{4}x^4 \Big|_0^1 = \frac{3}{4}$$

$$E(x^2) = \int_0^1 x^2 \cdot 3x^2 \cdot dx = \int_0^1 3x^4 dx = \frac{3}{5}x^5 \Big|_0^1 = \frac{3}{5}$$

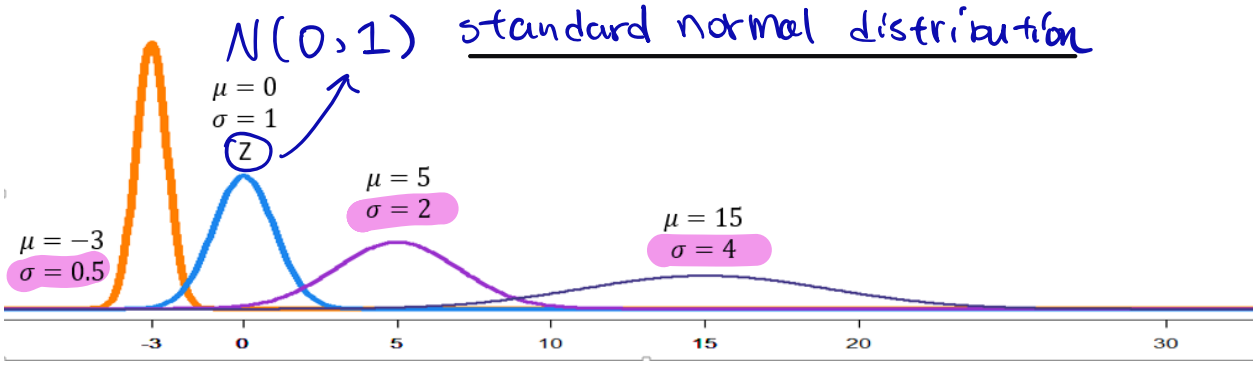
$$Var(x) = E(x^2) - (E(x))^2 = \frac{3}{5} - \left(\frac{3}{4}\right)^2 = 0.0375$$

$$SD(x) = \sqrt{0.0375}$$

The Normal Distribution

- Defined by two parameters: μ = mean | center
  σ = standard deviation $X \sim N(\mu, \sigma)$
- Symmetric, single-peaked (uni-modal), bell-shaped

- Models many naturally occurring random variables



Normal PDF

A random variable X that follows a Normal distribution has a probability density function:

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \text{ for } -\infty < x < \infty$$ Support

$E(X) =$ μ

$Var(X) =$ σ^2 $\quad SD(X) = \sigma$

Standardizing Normal Distributions

For a Normal random variable X, a z-score represents the number of standard deviations any observation x is from the mean:

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

The Standard Normal Distribution

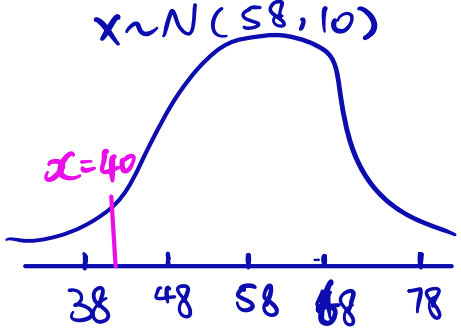
- Any normally distributed random variable can be "standardized" or transformed into a **standard normal distribution**.
- The standard normal random variable, denoted by Z, has $\mu = 0$ and $\sigma = 1$.

Normal Random Variable Example

For a particular bridge, recorded vehicle speeds are normally distributed with a mean of 58 mph and a standard deviation of 10 mph. Suppose a randomly chosen vehicle is going 40 miles per hour.

X = Vehicle speed

How many standard deviations away from the mean is 40 mph?



$$Z = \frac{40-58}{10} = \frac{-18}{10} = -1.8$$

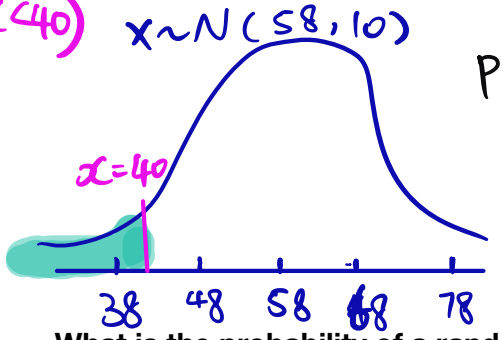
below

Normal Random Variable Example

For a particular bridge, recorded vehicle speeds are normally distributed with a mean of 58 mph and a standard deviation of 10 mph. Suppose a randomly chosen vehicle is going 40 miles per hour.

What is the probability of a randomly selecting a vehicle going less than 40 mph?

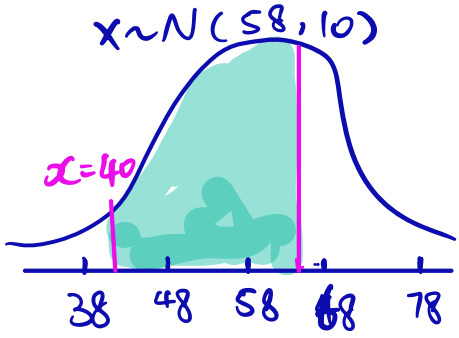
$P(x<40)$



$$P(x<40) = \int_{-\infty}^{40} \frac{1}{\sqrt{2\pi(10^2}} e^{\frac{-(x-58)^2}{2(10)^2}} dx$$

We can't solve this integrated Intergral analiticaly, we need to use R: pnorm(40, 58, 10) = 0.0359

What is the probability of a randomly selecting a vehicle going between 40 and 65 mph?



$$\int_{40}^{65} \frac{1}{\sqrt{2\pi(10)^2}} e^{\frac{-(x-58)^2}{2(10)^2}} dx$$

OR

$$P(40<x<65) = P(x<45) - P(x<40)$$

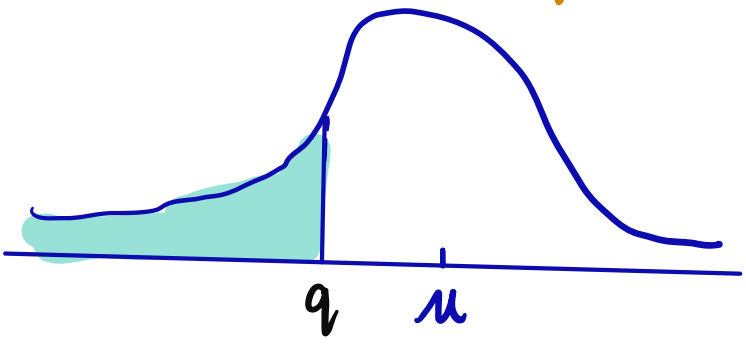
$$= pnorm(45, 58, 10) - pnorm(40, 58, 10)$$

$$= 0.7221$$

Using R to solve probabilities for various distributions

Distribution	R Function	Function Values	What does it do?
Binomial	`dbinom(q,size,prob)`	q = value of interest size = number of trials, n prob = probability of success, p	This is the probability mass function (PMF) for a binomial distribution
Binomial	`pbinom(q,size,prob)`	q = value of interest size = number of trials, n prob = probability of success, p	This is the cumulative distribution function (CDF) for a binomial distribution
Poisson	`dpois(q,lambda)`	q = value of interest lambda = rate parameter	This is the probability mass function (PMF) for a Poisson distribution
Poisson	`ppois(q, lambda)`	q = value of interest lambda = rate parameter	This is the cumulative distribution function (CDF) for a Poisson distribution
Normal	`pnorm(q, mu, sigma)`	q = value of interest mean = mean sd = standard deviation	This is the cumulative distribution function (CDF) for a Normal distribution

$P(x \le q) = pnorm(q, \mu, \sigma)$



$P(x \ge q) = 1 - pnorm(q, \mu, \sigma)$

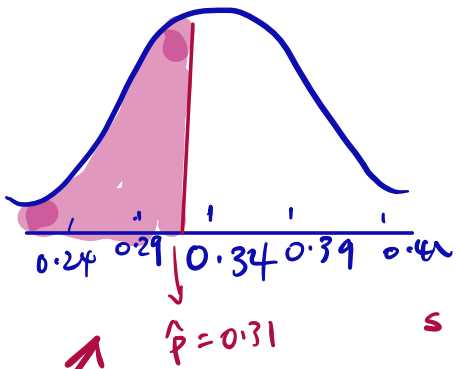
Week

The center limit theorem states :

size $n = 100$

$$\hat{P} \sim N\left(P, \sqrt{\frac{P(1-P)}{n}}\right)$$

$$N(0.34, 0.05)$$



$$P(\hat{P} \leq 0.31) =$$

$$Pnorm(0.31, 0.34, 0.05) = 0.274$$

If you take a Random sample of 100 people there is a 27.4% chance that the sample proportion of individual that support the phase out would be less or equal to 0.31

Week 4. Sampling & The Central Limit theorem

Sampling Variability & The Central Limit Theorem

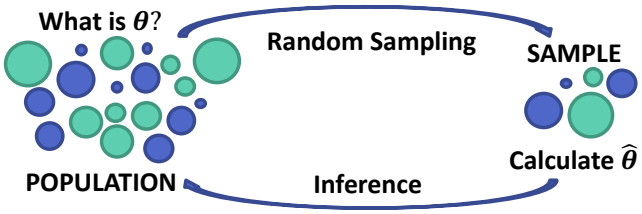
Inferential Statistics

Recall that inferential statistics use information from a sample to estimate or test characteristics from a population of interest.

How good is the sampled statistic $\hat{\theta}$ at estimating the population parameter θ?

general notation for sample statistic

general notation for population parameter



Things to consider:

- method of data collection
- sampling variability
- sample size

Point Estimates

As part of a quality control process for computer chips, an engineer a factory randomly samples 212 chips during a week of production to test the current rate of chips with severe defects. She finds that 27 of the chips are defective.

(a) What population is under consideration in the data set?

All computer chips manufactured at this factory this week.

(b) What is the parameter being estimated?

p = population propotion of defective chips

(c) Based on the sample what is the point estimate for the parameter?

$\hat{p} = \frac{27}{212} = 0.127$ sample proportion = "point estimate"

Sampling Variability

Suppose the study previously described was repeated by two other engineers in the same week. The following table gives the results from each of the three studies.

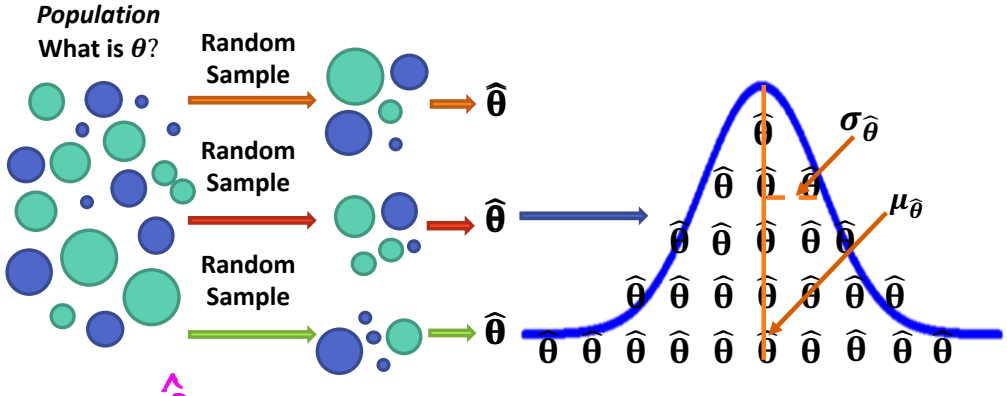
Study	Number of Chips Sampled	Number of Defective Chips	Sampled proportion
1	212	27	0.127
2	212	19	0.090
3	212	23	0.108

Compare the point estimates from the three studies. What do you notice? Which point estimate is the "best"?

Sampling Distributions

The probability distribution of a statistic, $\hat{\theta}$, is the Sampling Distribution. The sampling distribution defines the variability of the point estimate (or sample statistics) and quantifies the chance occurrence of specific values

Statistics are random variables! If N is the number of units in the population and n is the sample size, there are $\binom{N}{n}$ possible sample combinations.



$\hat{p}$ = sample proportion

$\bar{x}$ = sample mean

S^2 = Sample variance

Unbiased Estimators

Common statistics, such as $\hat{p}, \bar{x},$ and s^2 are unbiased.

The expected values of these estimators are equal to their parameters:

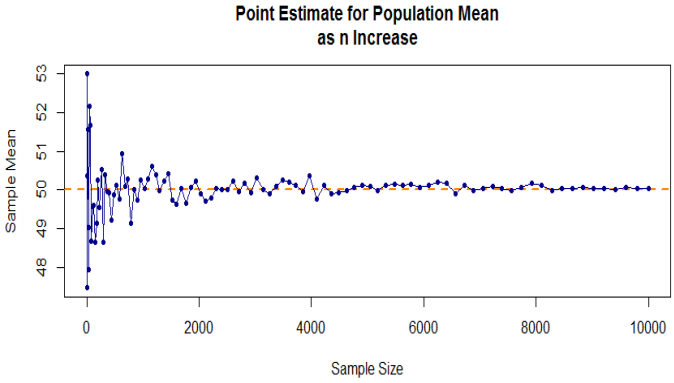
$E(\hat{p}) = p$ → population proportion

$E(\bar{x}) = \mu$ → population mean

$E(S^2) = \sigma^2$ → population variance

Law of Large Numbers

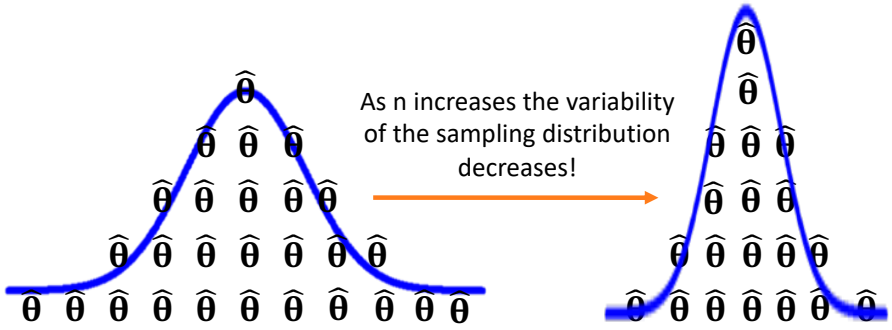
The Law of Large Numbers states that as n increases, the statistic will approach the true population parameter.



More formally, $\hat{\theta}$ converges in probability to θ. This implies that $\hat{\theta}$ is a consistent estimator.

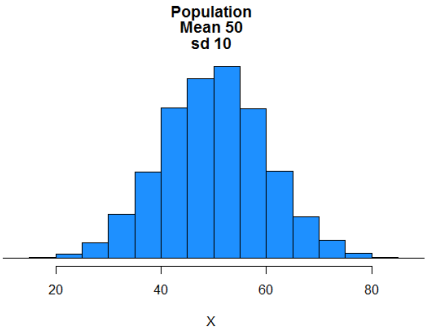
$\bar{x}$, $\hat{p}$, and s^2 are all consistent estimator

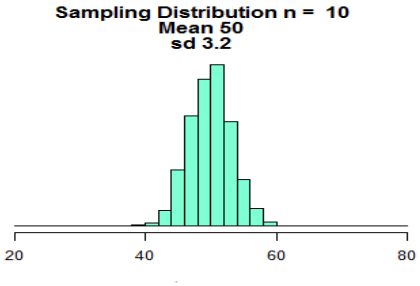
Sample Size and Sampling Variability

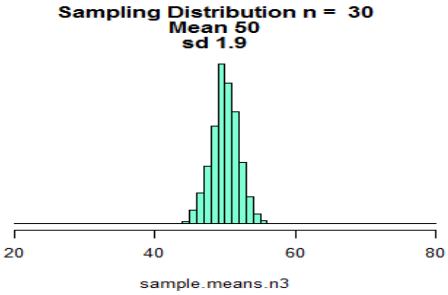


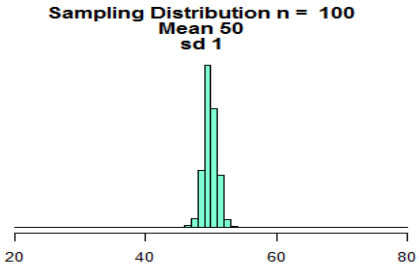
The variability of the sampling distribution of $\hat{\theta}$ is referred to as the standard error, denoted by $SE_{\hat{\theta}}$ or $\sigma_{\hat{\theta}}$. The standard error describes the typical error or uncertainty of the statistic. It is the standard deviation of the statistic.

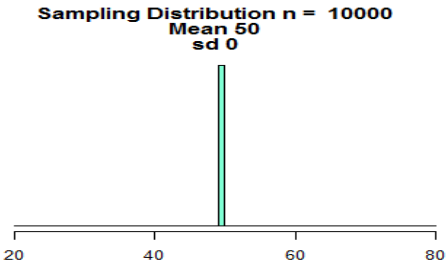
A random variable X is simulated from a normal distribution with population parameters μ_x and σ_x.





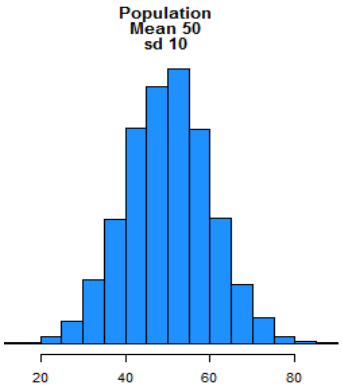


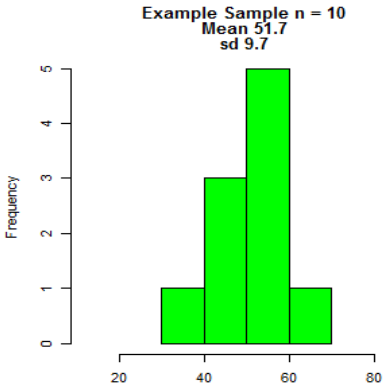


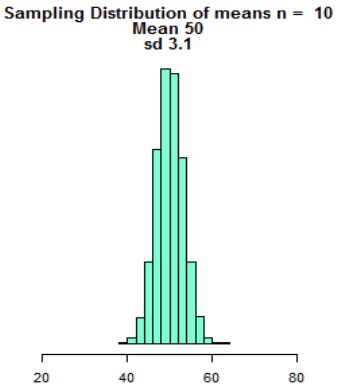


As n increases, the variability of the sampling distribution decreases.

Distributions of Inference







Population Distribution

Distribution of the entire collection of interest.

Sampled Distribution

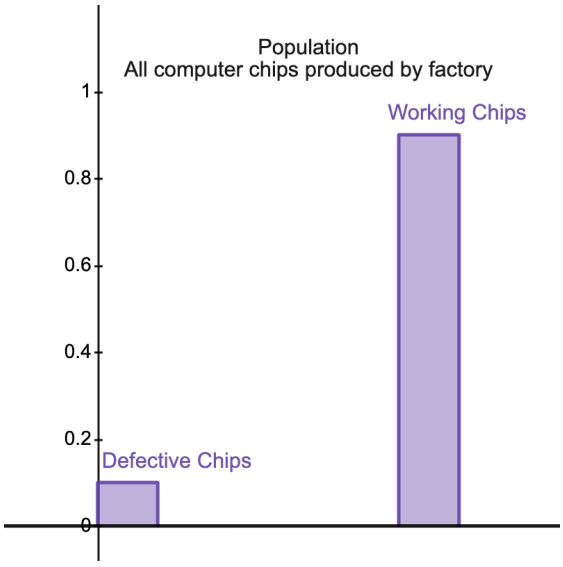
Distribution of observations obtained from a single sample

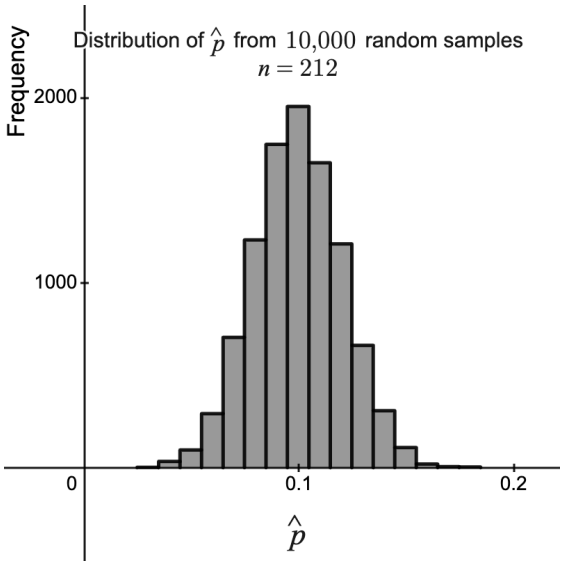
Sampling Distribution

Distribution of a sample statistic from repeated samples all of size n

In practice (real research study) we only observe the sampled distribution

Sampling Distribution of the Sample Proportion



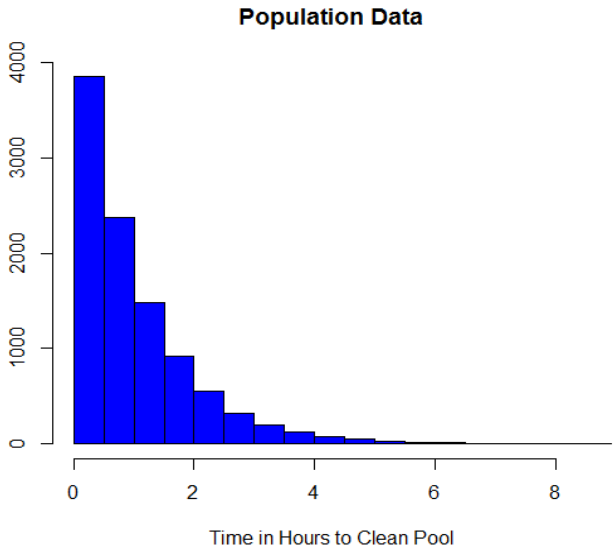


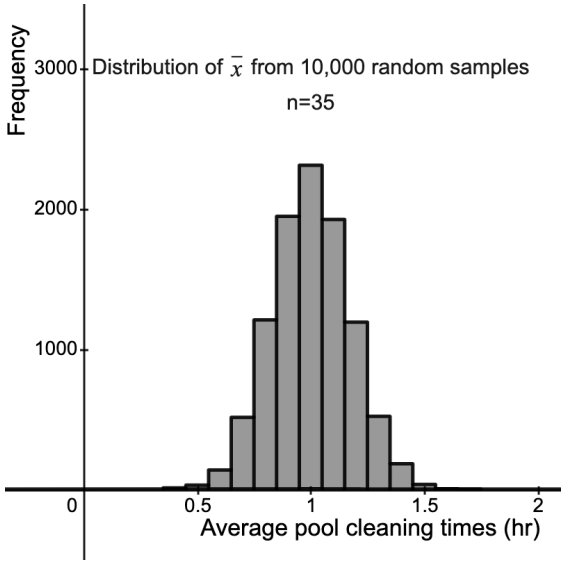
If n is sufficiently large, then the Central Limit Theorem states the sampling distribution of the statistic $\hat{p}$ is:

$\hat{p} \sim N\left(p, \sqrt{\frac{p(1-p)}{n}}\right)$ standard error

standard deviation

sampling distribution of sample mean:





If n is sufficiently large, then the Central Limit Theorem states the sampling distribution of the statistic $\bar{x}$ is:

$\bar{X} \sim N(\mu, \sigma/\sqrt{n})$

population SD

sample size

population mean

Sufficiently Large Sample Size, *n*

To determine if the sample size is sufficiently large enough for the Central Limit Theorem to apply, first you need to identify the parameter of interest!

When considering a population proportion, p, we need to assess the **success-failure conditions**:

- $np \geq 10$
- $n(1-p) \geq 0$

} sample size conditions for a proportion

When considering a population mean, μ, we need to identify the sample size used and observe the shape the the sample to determine if the sample size is sufficiently large.

$n \geq 30$: DOSn't matter if sampled distribution suggest skewness in the population. the sampling will be approximately normal (we can apply CLT)

$n < 30$

need to look at sampled distribution if the sampled distribution suggests it came from a skewed population, we should not apply the CLT

Sampling Distributions

Sampling distributions are never observed, but we keep them in mind!

Week 5 Confidence Intervals

Using only a point estimate is like fishing in a murky lake with a spear: If we only report estimate $\hat{p}$, we probably will not hit the exact population proportion.

A confidence interval is like casting a net: If we report a range of plausible values, we have a good shot at capturing the parameter.

Constructing a 95% Confidence interval

In the normal distribution, 95% of the observations are within 1.96 standard deviation of the distribution's center:

Point estimate $\pm$ 1.96 $\times$ SE

95% — 1.96
99% — 2.58
confidence level selected

$$\hat{p} \pm 1.96 \times \sqrt{\frac{p(1-p)}{n}}$$

Point estimate $\pm$ z^* $\times$ SE

a/ 95% confidence interval for the parameter p.

eg. Pew Research found that 88.7% of sample of 1000 US adults supported expanding the use of solar energy.

- check conditions,
- The standar error of the point estimate is 0.01

$$\hat{p} \pm 1.96 \times SE_{\hat{p}} \rightarrow 0.887 \pm 1.96 \times 0.010$$

$$\Rightarrow (0.8674, 0.9066)$$

We are 95% confident that the actual proportion of American adults who support expanding solar power is 6. ∴ — ∴ of American

eg2. In the Pew Research poll about solar energy, they also inquired about other forms of energy, and 84.8% of the 1000 respondents supported the use of wind turbines

Is is reasonable to model the proportion of us adults who support expanding wind turbines using a normal distribution?

Independence → Simple random sample

Success-failure condition:

0.848×1000 and $0.152 \times 1000 = 152$ (number of "failures") are both greater than 10

eg3. Create a 99% confidence Interval ....

The point estimate is: 0.848

SE: $SE_{\hat{p}} = \sqrt{\frac{0.848(1-0.848)}{1000}} = 0.0114$

Construct the Interval: $0.848 \pm 2.58 \times 0.0114 \Rightarrow (0.8186, 0.8774)$

Confidence Intervals for a Proportion

* If we want to estimate the population proportion, then we would use sample proportion as point estimate

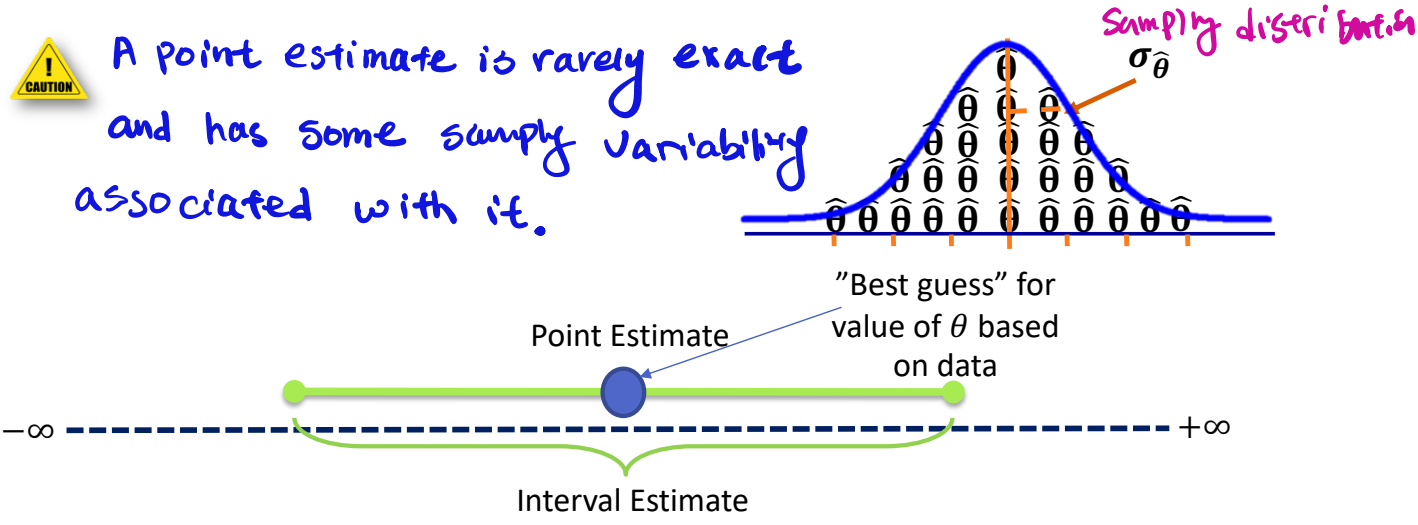
Estimation

A point estimate is the single value consider to be the "best guess" for the parameter of interest

An unbiased point estimate is one in which the expected value is equal to the parameter: $E(\hat{\theta}) = \theta$.

It is desirable to use estimates that are unbiased and have small standard error. is standard deviation of Sampling Proportion

Estimation and Sampling Variability



A confidence interval can increase the likelihood an estimation procedure will capture the population parameter. An interval estimate provides a

range of plausible values for the population parameter where the point estimate is the center of the interval.

Constructing a 95% Confidence Interval

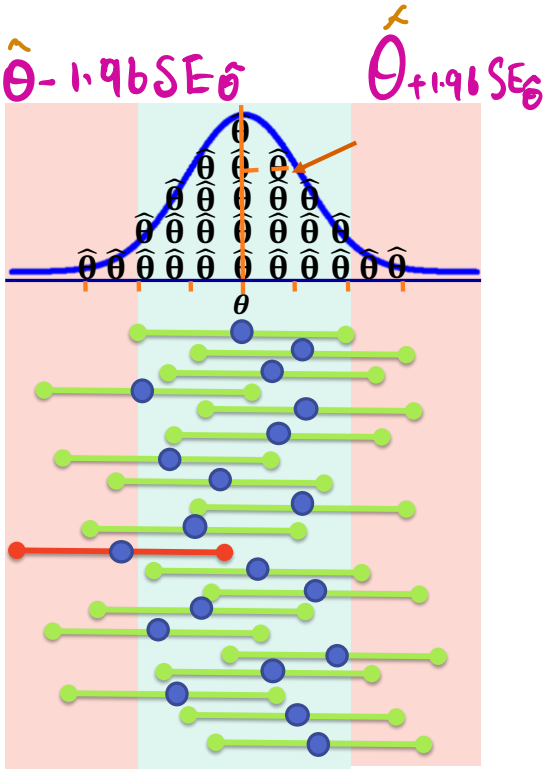
Point estimate / sample statistics

If the sampling distribution of $\hat{\theta}$ is normal, then 95% of all possible estimates will be within 1.96 standard errors of the population parameter, θ.

If a point estimate within 1.96 standard errors is selected, then the interval $\hat{\theta} \pm 1.96\,SE_{\hat{\theta}}$ will capture the parameter, θ.

We can be 95% confident the parameter is captured by the above interval estimate.

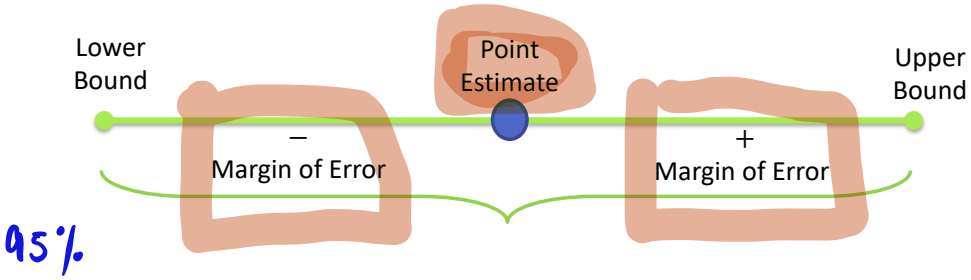
5% of intervals won't capture θ. Since we don't know the value of θ, we won't know if the interval we constructed is one of the "good ones".



Constructing a Confidence Interval

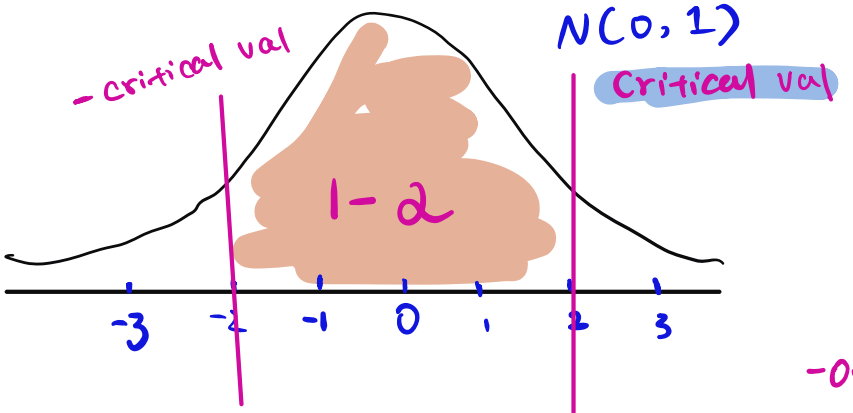
All confidence intervals have the same general form: Point estimate ± margin of error

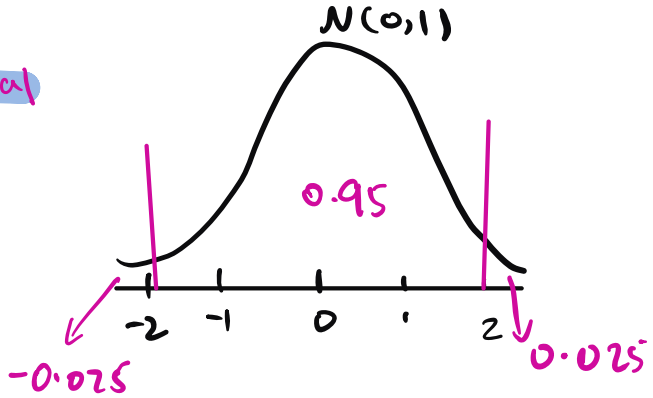
The margin of error is the product of a critical value and standard error of $\hat{\theta}$.



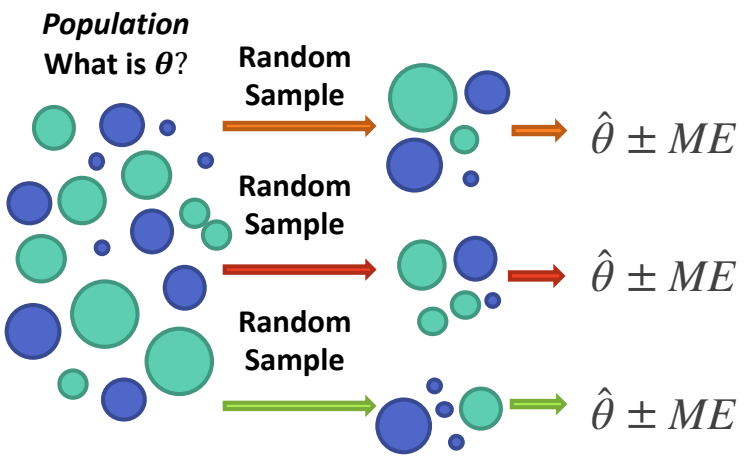
For a $100(1-\alpha)\,\%$ confidence interval, the critical value is the $1-\frac{\alpha}{2}$ percentile in the standardized sampling distribution.

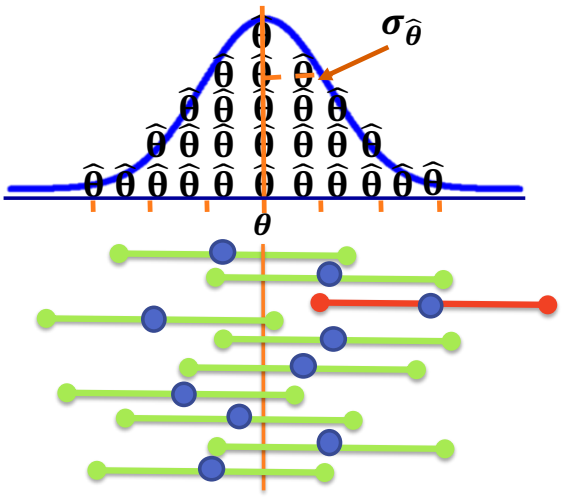
α is the error of the interval and is determined by the desired level of confidence.





"Confidence" in Confidence Intervals





The term "confidence" comes from theory based on repeated sampling.

If a 90% CI is calculated from every random sample, then on average 9 out of 10 intervals will capture the parameter, where as 10% of all intervals will fail.

Therefore, we are 90% confident our interval will capture the parameter.

"Interpret the Interval"

How to Describe a Confidence Interval

confidence level → parameter of Intrest in contat

Template: We are _____% confident that the _____ is between _____ and _____, with a point estimate of _____.

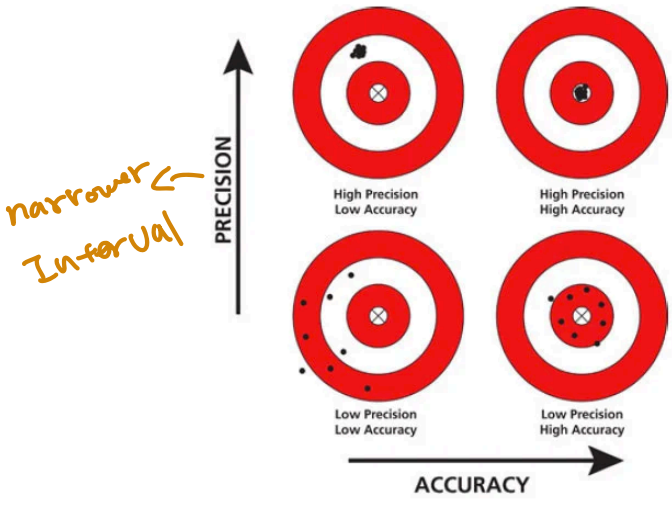
point estimate

lower bound

upper bound

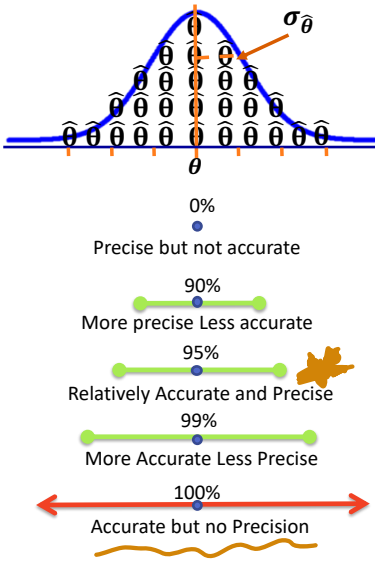
Choosing a Confidence Level

The confidence level is chosen based on whether precision or accuracy is more desirable.



narrower Interval

wider Interval



distribution

Central Limit Theorem for a Proportion

Recall that for a sample of size n, if n is sufficiently large, then the Central Limit Theorem states that $\hat{p}$ follows an approximately Normal distribution with a mean equal to the population proportion, p, and a standard deviation of $\sqrt{\frac{p(1-p)}{n}}$.

$\hat{P} \sim N\left(P, \sqrt{\frac{P(1-P)}{n}}\right)$

Confidence interval for p

When to use: want to estimate a population proportion, P, from a single population

The confidence interval for a population proportion is:

critical value

$\hat{P} \pm z^* \sqrt{\frac{P(1-P)}{n}}$ estimate for the standard error

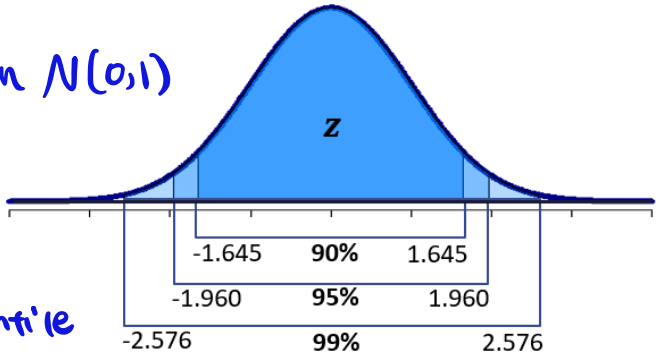
The interval above relies on the Central Limit Theorem. Therefore, we need a sufficiently large sample size. Since we don't know p, we need to check the following:

is $n\hat{p} \geq 10$?

is $n(1-\hat{p}) \geq 10$?

Critical Values for z Confidence Intervals

Confidence Level	Critical Value
90%	The 95th percentile on N(0,1) 1.645
95%	The 97.5th percentile on N(0,1)
99%	The 99.5th percentile on N(0,1) 2.574



If you are asked to construct a confidence interval for the population proportion, you can use the table above to find z^*, the critical value.

Confidence Interval Example

A 2013 survey of 434 adults found that 52% of U.S. adult Twitter users get at some news on Twitter. Construct a 99% confidence interval for the proportion of U.S. adult Twitter users who get some news on Twitter and interpret the interval in context.

t-distributions and Confidence Intervals for a Mean

Conditions for Inference About a Mean

A car manufacturer wants to estimate the average annual mileage of drivers of their top of the line sports car. They would like to be 95% confident in the estimate and assume the population to be normal. A random sample of 10 drivers yields the following data:

11501	8987	12166	9247	10143
8230	3111	13009	7891	10392

According to the central limit theorem, what is the sampling distribution of the sample mean?

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

"approximately normally distributed"

n: sample size

σ: population SD → usually unknown

μ: population mean

t-distribution

- The sampling distribution is normal when standardized with σ_x.

$$\frac{\bar{x} - \mu}{\sigma_x / \sqrt{n}} \sim N(0, 1)$$

- If σ_x is unknown, can we use the sample standard deviation s_x instead?

yes, but we can't use the normal distribution

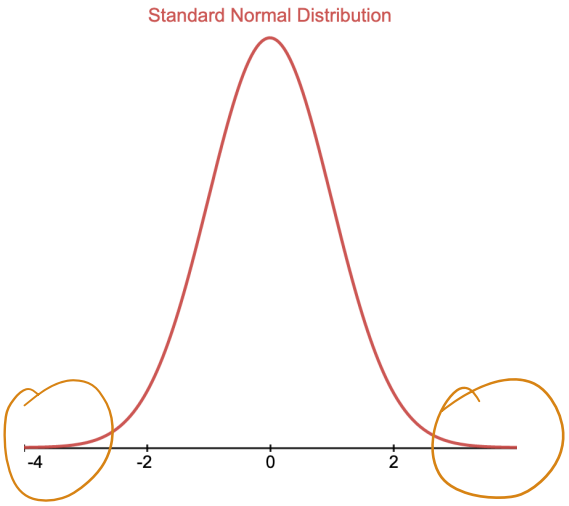
- The distribution is not normal when we replace σ_x with s_x.

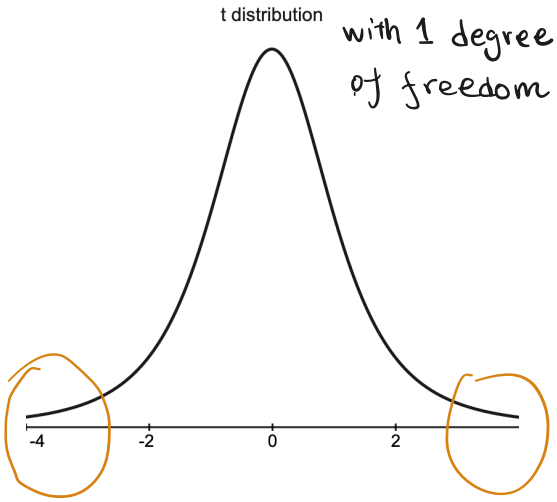
$$\frac{\bar{x} - \mu}{s_x / \sqrt{n}}$$

- When we standardize using s_x, our sampling distribution comes from a

t-distribution.

$$t = \frac{\bar{x} - \mu}{s_x / \sqrt{n}}$$





- Symmetric and centered at 0.
- Distribution is defined by its degree of freedom.
- As the degrees of freedom increases towards ∞, the t distribution approaches the standard normal distribution
- The t distribution has heavier area in the tails when compared to the standard normal distribution.
- Degrees of freedom are based on the sample size and number of unknown parameters.

For sufficiently large n, the distribution of the standardized sample mean follows a t-distribution with $n-1$ degrees of freedom.

$$\frac{\bar{x}-\mu}{s/\sqrt{n}} \sim t_{n-1}$$

∴ follows t distribution with n-1 degrees of freedom"

- check if $n \geq 30$
  If yes. good to go
  If no, check the sampled distribution for symmetry

Confidence Interval for μ when σ is unknown

μ → pop mean
σ → pop SD

When to use:

Want to estimate the population mean but we don't the population standard deviation

point estimate ± margin of error

margin of error: critical value × standard error estimate

The confidence interval for a population mean is:

$$\bar{x} \pm t^* \left(\frac{s}{\sqrt{n}}\right)$$

sample SD

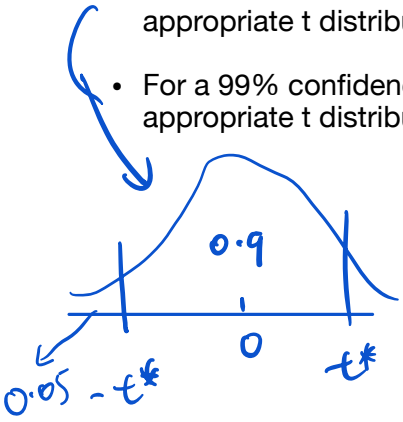
sample mean

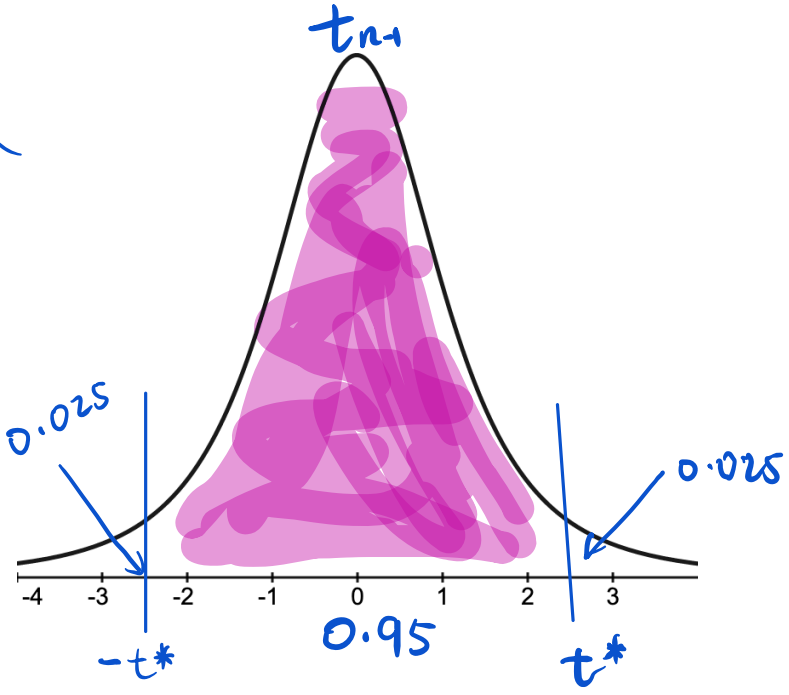
critical value from t-distribution

Finding *t* Critical Values

Recall that for a confidence interval, the critical value is the $\left(1 - \frac{\alpha}{2}\right)^{th}$ percentile.

- For a 95% confidence interval, we need to find the 97.5 th percentile from the appropriate t distribution. t-distribution with n-1 degrees of freedom
- For a 90% confidence interval, we need to find the 95 th percentile from the appropriate t distribution.
- For a 99% confidence interval, we need to find the 99.5th percentile from the appropriate t distribution.





Finding *t* Critical Values Using qt( )

Function	Function Values	What does it do?
`qt(p,df)`	`p` = area under the curve to the left of x_p `df` = degrees of freedom	This is the inverse cumulative distribution function. Finds percentiles for the t distribution. That is, finds x_p for the expression $P(X \le x_p) = p$

degree of freedom

area under the curve to the left of t^*

Ex: 95% CI for μ $n=40$ → R Code: qt(0.95, 39) ← n-1

Confidence Interval Example #1

The FDA's webpage provides some data on mercury content of fish. Based on a sample of 15 croaker white fish (Pacific), a sample mean and standard deviation were computed as 0.287 and 0.069 ppm (parts per million), respectively. From a visualization of the 15 points, we do not find any signs of strong skewness or extreme outliers. We will assume these observations are independent. Calculate the 90% confidence interval for the average mercury content of croaker white fish in the pacific.

Even though $n < 30$, since the sampled distribution does not suggest that the population distribution is skewed, we are able to construct the confidence interval

$n = 15$, $\bar{x} = 0.287$, $s = 0.049$, conf. level $= 90\% \rightarrow t^*$

$t = qt(0.95, 14) = 1.761$

Confidence Interval: $0.287 \pm 1.761\left(\frac{0.1049}{\sqrt{15}}\right) = (0.27, 0.318)$

(point estimate)

we are 90% confident that the average mercury content of croaker white fish in the pacific is between 0.257 PPM and 0.318 PPM with a point estimate of 0.287 PPM

Confidence Interval Example #2

Researchers interested in lead exposure due to car exhaust sampled the blood of 52 police officers subjected to constant inhalation of automobile exhaust fumes while working traffic enforcement in a primarily urban environment. The blood samples of these officers had an average lead concentration of 124.32 µg/l and a SD of 37.74 µg/l. Construct the 95% confidence interval for average lead concentration in the blood of traffic enforcing police officers in this urban environment.

Hypothesis Testing

Estimation vs. Hypothesis Testing

Estimation	Want to know more about a population characteristic.	Goal is to _____ the population characteristic (unknown parameter).	Point estimates and confidence intervals
Hypothesis Testing	Have a claim or guess about a population characteristic.	Goal is to use sampled data to _____ the validity of the claim.	Hypotheses, test statistics, and p-values

Examples of when to use a hypothesis test:
- Test the average train ride time against the advertised time
- Test the proportion of defective parts in a manufacturing process against a specified standard
- Compare whether the average tensile strength of rubber seals is different between machines

Hypothesis Test Example

A high-speed train between Florence and Rome, Italy is advertised to take 94 minutes. A frequent rider is consistently late and is convinced the average ride time is longer. She decided to record the time of 40 random rides. Her sampled average is 99 minutes. The sample has a standard deviation of 10 minutes.

If the advertised time of 94 minutes is the actual average, what is the chance the mean from a sample of 40 rides would exceed 99 minutes?

ST 314

Steps in Performing a Hypothesis Test

1.

2.

3.

4.

5.

6.

The Hypotheses

The Null Hypothesis	The Alternative Hypothesis

Hypothesis Test Example

A high-speed train between Florence and Rome, Italy is advertised to take 94 minutes. A frequent rider is consistently late and is convinced the average ride time is longer. She decided to record the time of 40 random rides. Her sampled average is 99 minutes. The sample has a standard deviation of 10 minutes.

Determine the null and alternative hypotheses for this test based on the train site's claim that rides take 94 minutes.

ST 314

One- and Two-sided Alternatives

Lower one-sided
Question of interest: is the parameter less than the claimed value?

Upper one-sided
Question of interest: is the parameter greater than the claimed value?

Two-sided
Question of interest: does the parameter differ from the claimed value?

One- or two-sided practice

Consider the following alternative hypothesis examples. Which are one-sided? Which are two-sided?

1. The average time to finish a maze is not equal to 16 minutes.

2. The average running pace is less than 7 mph.

3. The average IQ score of OSU students is greater than 119.

4. The average exam score is different than 75%.

5. The average lunch break is less than 30 minutes.

Test Statistics and p-values

Similar to “standardizing”, a test statistic compares a _____ to a

_____ while accounting for the variability of the statistic.

The general form of a test statistic is:

The Null Distribution

$H_0 : \mu = \mu_0$

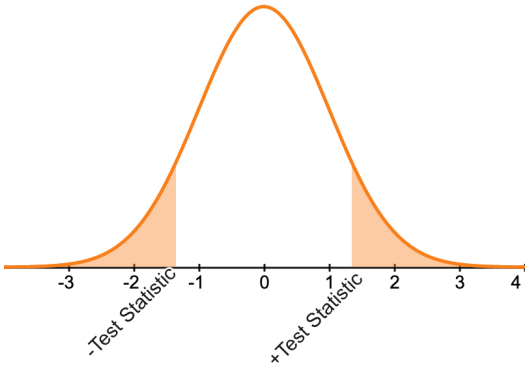
$H_0 : p = p_0$

p-values

The p-value is a _____. It is the area under the tail(s) of the curve.

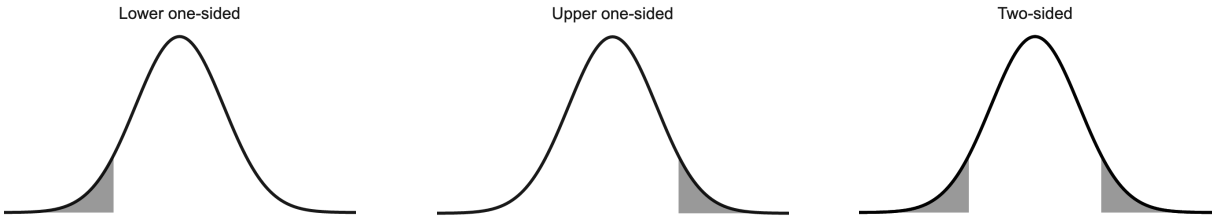
Assuming the hypothesized parameter value is true, the _____ the sample statistic would take on a value as or more extreme than the one we actually observed is the

_____.



ST 314

One- and two-sided p-values



t-test for a Mean

When to use:

Conditions required for inference:

Null & Alternative hypotheses:

Test statistic:

t-test for a Mean Example

A high-speed train between Florence and Rome, Italy is advertised to take 94 minutes. A frequent rider is consistently late and is convinced the average ride time is longer. She decided to record the time of 40 random rides. Her sampled average is 99 minutes. The sample has a standard deviation of 10 minutes.
We want to perform a hypothesis test to test whether the average ride time is greater than 94 minutes. To do so, first check the conditions. If they are met, proceed to calculating the test statistic.

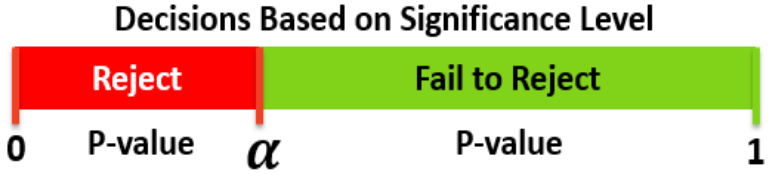
Now that we've calculated the test statistic, we know how many standard deviations away from the hypothesized mean our sample mean is. We can use this value to calculate the probability we would have observed a sample mean as or more extreme than 99 minutes if we assume the true population mean to be 94 minutes (this calculation is the p-value). Calculate the p-value.

Decisions based on p-values

The significance level, denoted by α, is a predetermined cut off to which we will reject the null hypothesis. Common values for α are:

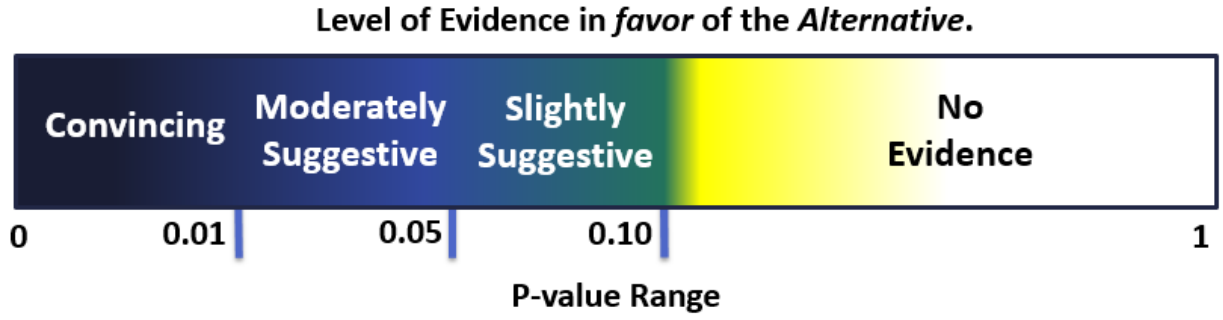
A _____ α requires a more extreme test statistic to reject the null. Decreasing α will reject less often, but will decrease the chance of error.

Never say the null is "false". Instead, we say "the evidence is very strong to reject the null".



Statements in terms of the alternative

- Using only terms like “reject” and “fail to reject” in conclusions may confuse novice readers.
- We’ll provide a more complete conclusion by providing a statement of evidence in terms of the alternative hypothesis that reflects the question of interest.



Four-part Conclusion

A good conclusion must have _____ and should include the following four important pieces.

- Statement for the _____ for the alternative hypothesis.
- Whether to _____ or _____ the null hypothesis based on α.
- The _____ estimate for the parameter of interest.
- A $(1-\alpha)100\,\%$ _____ estimate for the parameter of interest.

Hypothesis Test Conclusion Example

A high-speed train between Florence and Rome, Italy is advertised to take 94 minutes. A frequent rider is consistently late and is convinced the average ride time is longer. She decided to record the time of 40 random rides. Her sampled average is 99 minutes. The sample has a standard deviation of 10 minutes.

Write a four-part conclusion for this hypothesis test.

Errors in Hypothesis Testing

	H_0 True	H_0 False
Reject H_0	Type I Error	Correct Conclusion
Fail to reject H_0	Correct Conclusion	Type II Error

- When a test performs as it should, the value is the probability of making a _____ error.

- When the consequences are high, reduce the chance of making a Type I error by choosing a _____ significance level.

Type I and Type II are related. Decreasing the chance of one error will increase the chance of the other.

z-test for a Proportion

When to use:

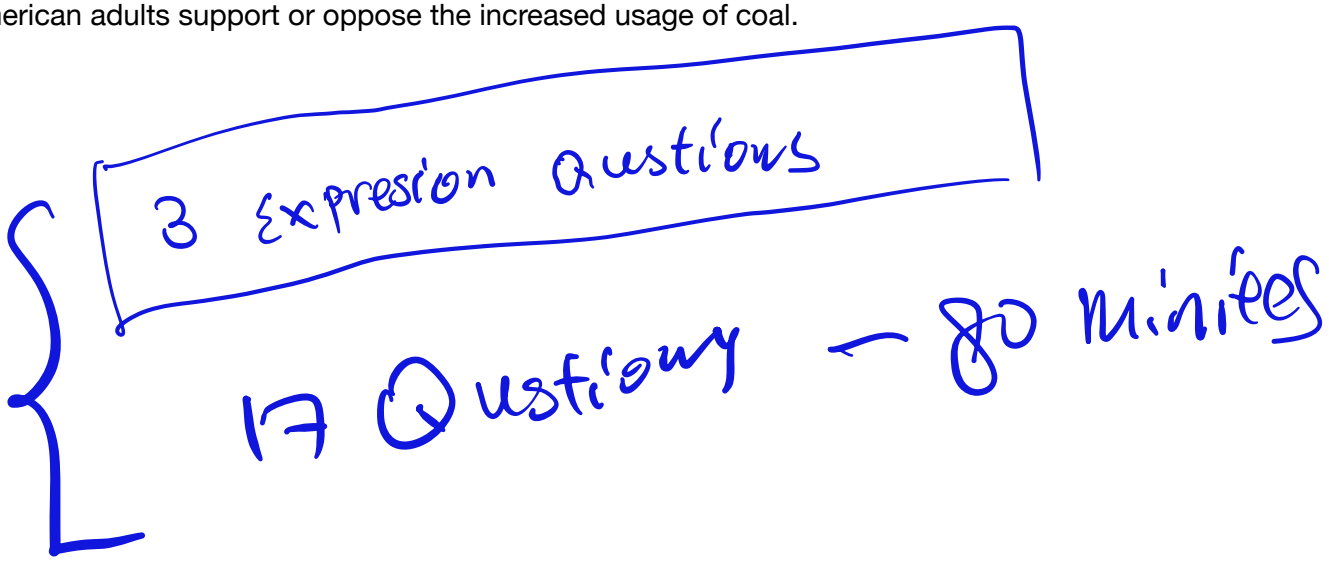
Conditions required for inference:

Null & Alternative hypotheses:

Test statistic:

z-test Example

Pew Research asked a random sample of 1000 American adults whether they supported the increased usage of coal to produce energy. Set up hypotheses to evaluate whether a majority of American adults support or oppose the increased usage of coal.



n: sample size

ST 314 Practice Midterm Exam

The following questions are meant to give you sense of what the midterm exam might look like. This should not be your only source of study material! See Canvas for suggestions of other resources you can use to prepare for the exam.

True or False: Questions 1-5

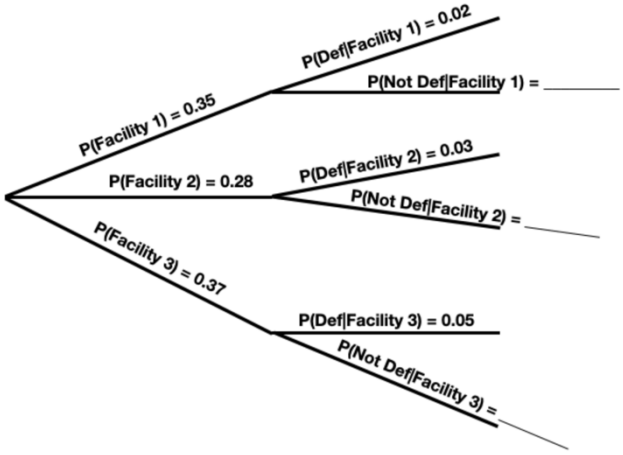
1. Under a standard normal density curve, 50% of the distribution falls below the mean.
    a. True
    b. False
2. For data that are right skewed, the mean is less than the median.
    a. True
    b. False
3. For any two events, A and B, the probability of the intersection of the two events can be found by computing the product of the probabilities of each individual event. That is, $P(A \cap B) = P(A)P(B)$.
    a. True
    b. False
4. An observational study imposes treatments onto subjects or units for the purpose of measuring a response.
    a. True
    b. False
5. A confidence interval becomes less precise as the sample size, n, increases.
    a. True
    b. False

Use the following information for questions 6-7

A computer manufacturer has three different facilities they use to manufacture monitors for their desktop computers. From each of the facilities, the monitors are inspected for defects. The tree diagram depicts this scenario, Def = defective or Monitor has defects.

6. Given a randomly selected monitor was manufactured at facility 2, what is the probability it has no defects?

7. Which of the following expressions correctly represents the probability that a randomly selected monitored was



manufactured at facility 3 and is defective?

a. $0.37 + 0.05$

b. 0.37×0.05

c. $0.37 - 0.95$

d. 0.37×0.95

Use the following information for questions 8-9

The random variable X is described by the following pdf:

8. What is the **cumulative** density function?

   a. $F(x) = 1$ for $x > 1$

   b. $F(x) = -x^{-3}$ for $x > 1$

   c. $F(x) = 1 - x^{-3}$ for $x > 1$

   d. $F(x) = 3x^{-4}$ for $x > 1$

9. What is the expected value of X?

10. An injection molding process for making detergent bottles uses three different machines the table gives the probability mass function for the number of machines operating at any given time.

y_i	0	1	2	3
$p(y_i)$	0.020	0.030	0.200	0.750

Which of the following expressions correctly represents that probability that two or fewer machines are running at the same time?

a. $P(Y = 2) = 0.2$

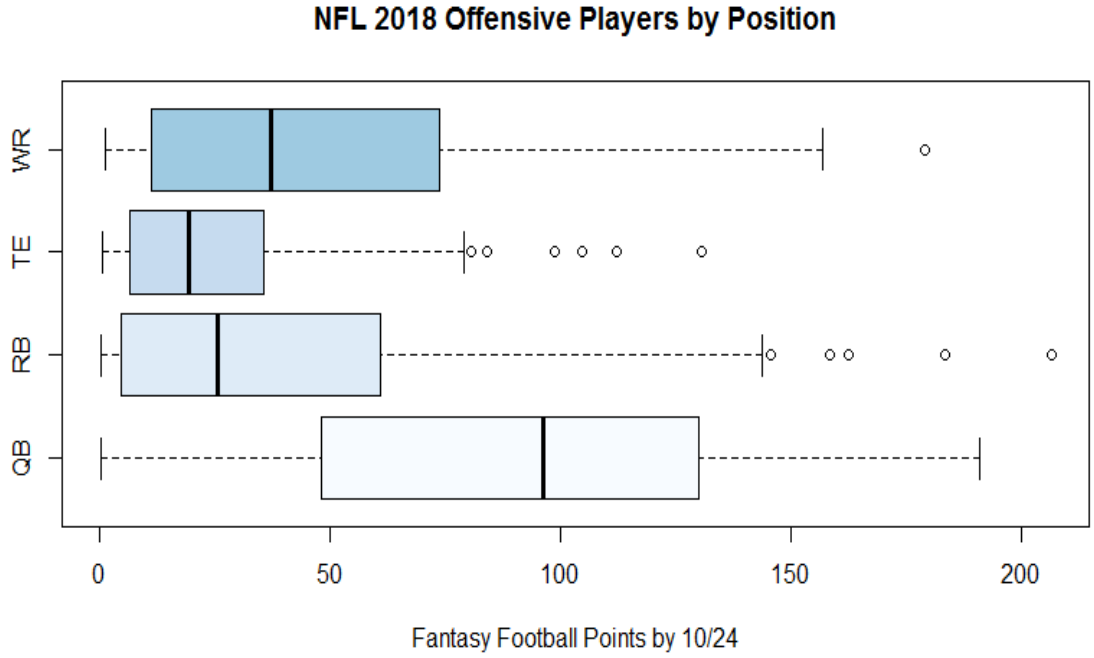
b. $P(Y \leq 2) = 0.2$

c. $P(Y \leq 2) = 0.2 + 0.03 + 0.02$

d. $P(Y < 2) = 0.2 + 0.03$

11. Which of the following statements is FALSE with respect to the standard deviation?
    a. The standard deviation is the same as the variance.
    b. The standard deviation measures the typical deviation from the mean.
    c. The standard deviation is influenced by outliers in the data.
    d. The standard deviation has the same units as the mean.

12. The graph below depicts total fantasy football points earned by offensive players by the end of week 7 in the 2018 NFL Season by position.



From the plot which of the following statements is FALSE?
    a. The minimum value is approximately the same for all positions.
    b. There are no outliers in the quarterback (QB) position.
    c. The median for running backs (RB) is approximately 25 total fantasy points.
    d. Based on the distributions for each position it is safe to assume the average total number of points will be the same as the median total number of points.

13. The number of cars entering a roadway at a specific exit during the morning rush hour traffic (6AM to 9AM, M-F) is a random variable modeled by a Poisson distribution, with an average of 6 cars per minute. Which of following distributions would be appropriate to model the number of cars entering the roadway?
    a. The Normal distribution
    b. The Binomial distribution
    c. The Poisson distribution
    d. The random variable cannot be modeled with any distribution.

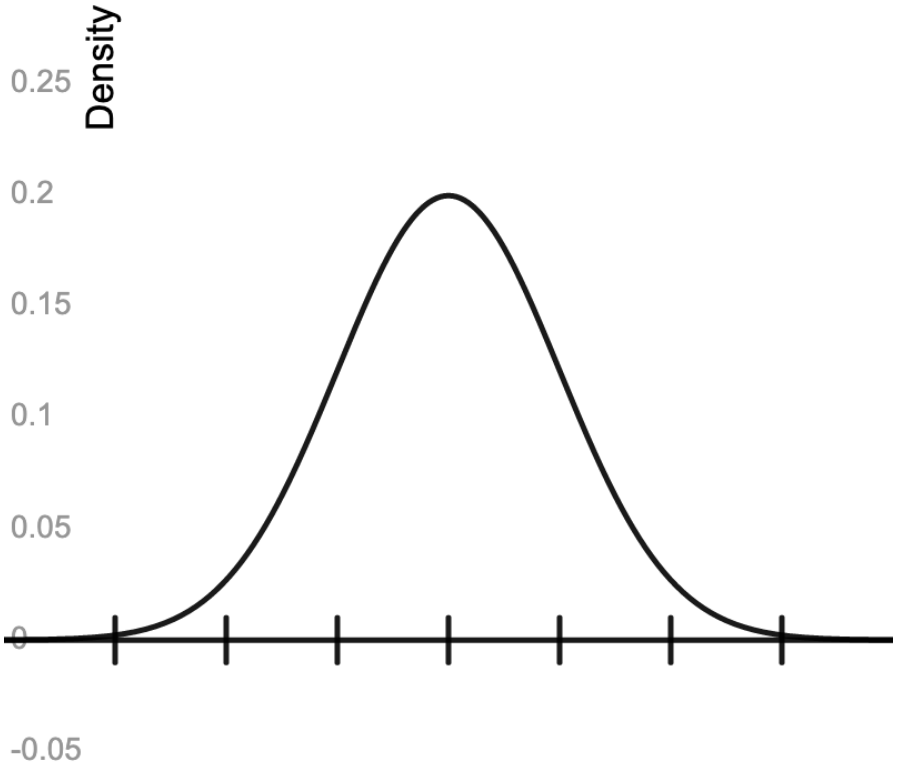
Use the following information for questions 14-16

The time it takes to travel between two campuses of a university in a city via shuttle bus takes an average 24 minutes. The distribution of all times is right skewed with a standard deviation of 12. Suppose a random sample of 36 transport times are recorded during a given week.

14. Which of the following describes the approximate sampling distribution of the sample mean from a random sample of size $n = 36$?

    a. The sampling distribution will be approximately normal with a mean of 24 minutes and standard deviation of 12 minutes.
    b. The sampling distribution will be approximately normal with a mean of 24 minutes and standard deviation of 2 minutes.
    c. The sampling distribution will be right skewed with a mean of 24 minutes and standard deviation of 12 minutes.
    d. The sampling distribution will follow a binomial distribution with a mean of 24 minutes and standard deviation of 2 minutes.

15. Using the outline of the sampling distribution curve, illustrate the probability of selecting a random sample of 36 transport times with a mean travel time less than 22 minutes. Include numerical values on the tick marks representing $\pm 1, \pm 2, \pm 3$ standard deviations above and below the center, as well as shading the region that



represents the probability of interest.

16. Suppose all possible combinations of 36 observations were collected from the population. That is, all possible samples of size $n = 36$ are observed. Which of the following statements regarding the sample means from these repeated samples is true?
    a. Approximately 95% of sample means will be between $24 - 1.96 \times 2 = 20.08$ minutes and $24 + 1.96 \times 2 = 27.92$ minutes.
    b. Approximately 95% of sample means will be between $24 - 1.96 \times 12 = 0.48$ minutes and $24 + 1.96 \times 12 = 47.52$ minutes.
    c. Approximately 90% of sample means will be between $24 - 1.96 \times 2 = 20.08$ minutes and $24 + 1.96 \times 2 = 27.92$ minutes.
    d. Approximately 90% of sample means will be between $24 - 1.96 \times 12 = 0.48$ minutes and $24 + 1.96 \times 12 = 47.52$ minutes.

17. The **sampling distribution** of a statistic is
    a. the probability that we obtain the statistic in repeated random samples.
    b. the probability distribution of the randomly sampled statistic.
    c. the mechanism that determines whether randomization was effective.
    d. the extent to which the sample results differ systematically from the truth.

Use the following information for questions 18-19

A store randomly samples 603 shoppers over the course of a year and finds that 142 of them made their visit because of a coupon they received in the mail.

18. Construct the **90%** confidence interval for the proportion of shoppers that visited the store because they received a coupon in the mail. *You do not need to do any of the calculations – just set up the construction of the confidence interval.*

19. The **95%** confidence interval is (0.202, 0.269). Interpret this confidence interval in the context of the problem.