Contents

1 Lecture 1 Notes: Population and Sample

Definitions:

- **Population** is defined as a complete collection of all elements that are of interest.
- A **census** is defined as a list of data from every member of the population. A census is not always possible, imagine studying all individuals in the United There are people that are homeless or are here momentarily.
- A **sample** defined as a sub-collection of members from part of the population. Normally collected using a random method. Extra: Think of ways of generating a random sample.

Information:

- **Data** is defined as observations, measurements, of elements of the sample that is of interest
- **Statistics** is collection of methods for planning experiments, obtaining data; then organizing, summarizing, analyzing, interpreting, presenting and drawing conclusions based on data.

Information from the Population and Sample:

- **Parameter** is a numerical measurement describing some characteristic of the population **Statistic** is a numerical measurement describing some characteristic of sample that is a proxy of the parameter. Since we cannot always know the parameter we use the statistic

Table 1: My caption

Information	Population	Sample
Mean	μ	$\bar{x}$
Standard Deviation	σ	s
Variance	σ^2	s^2
Proportion	p	$\hat{p}$

Types of Information:

- Qualitative (categorical) - name or a description
  - Nominal Level - Categorical data only. Data can not be arranged in order. (i.e., YES/NO/NA or iPhone/Android/NA)
  - Ordinal Level - Data can be arranged in order. Differences are meaningless. (i.e., A/B/C/D or Like it/Love it/Got to Have it)
- Quantitative (numerical) - Counts or Measurements (i.e., # of red cars or weight)

- Interval Level - Like Ordinal but differences are meaningful. There is no natural zero starting point.(i.e., 1988/1989)
- Ratio Level - Similar to interval level with a zero starting point. The zero starting point makes ratios meaningful. (i.e., weight, height)

1.1 Collection of Data is done various ways

Various ways of collecting data:

1. Experiments - apply a treatment to determine if it has an effect on individuals
2. Observational Study - observe and measure characteristics of a sample, we do nothing to the sample

1.2 Design of Experiment:

1. Replication is to repeat an experiment with many individuals to measure a particular effect (i.e., multiple individuals receive treatment and multiple individuals don't)
2. Blinding the individual does not know treatment type. This is a way to remove the placebo effect. Double blind is when both the subject and researcher does not know the treatment type (i.e., give treatment but patient does not know which drug was given)
3. Randomization individuals are randomly assigned to treatment type, this is important to remove confounding factors (i.e., assign random number between 1 and number of individuals, use random number generator to select individuals to be part of each treatment)
4. Confounding when a effect is seen but we do not know why it is present (i.e, every one at the the Dominican Hospital is older than 75)

1.3 Observational Studies:

1. Cross-Sectional Study is when data observed, measured, and collected at one time (i.e., analyze medical records and record tonometry tomorrow)
2. Retrospective (or Case-Control) Study is when data is collected from a past time period (i.e., analyze medical records 10 years ago)
3. Prospective (or Longitudinal or Cohort) Study is when data is collected in the future (i.e., design analysis for medical records 10 years from now)

1.4 Sampling Techniques:

1. Simple Random Sample is when everyone in the population has the same chance of being in the the sample
2. Stratified Sampling is when g natural groups (strata) are in the population and a SRS is selected from each group (strata)
3. Systematic Sampling is when a list of the population is created and every k^{th} individual is selected to be in sample
4. Convenience Sampling is when the closest/easiest to obtain group from the population is the sample
5. Cluster Sampling is when the population is divided into clusters and randomly selected in which the entire population is selected to be in the sample
6. Volunteering Sampling is when individuals volunteer to be a part of the sample

2 Lecture 2 Notes: Graphical Representation & Central Tendency of Data

Important Distinction:

1. **Descriptive Statistics**: The objective is to summarize or describe the data
2. **Inferential Statistics**:The objective is to make inference of the population from the sample

2.1 Summarizing the Data

1. **Frequency**: - the number or count a number appears

2. **Frequency Distribution**: - shows how data is broken up into classes (bins) and number the number of occurrences that appear within each bin based on data

Example 1: Frequency distribution of cotinine (a metabolite of nicotine) level of smokers. A sample of 40 smokers and their cotinine level) in ng/ml (1st edition)

1	0	131	173	265	210	44	277	32	3
35	112	477	289	227	103	222	149	313	491
130	234	164	198	17	253	87	121	266	290
120	167	250	245	48	86	284	1	208	173

2.2 Procedure for Constructing a Frequency Distribution

1. Select number of bins (between 5-20), lets choose 5
2. Calculate Width:

   $$\text{Class Width} = \frac{Max – Min}{\text{\# of bins}} = \frac{491–0}{5} = 98.2 \approx 100$$

   Round up to make life easier.
3. **Find the Lower limits (LL)** for each bin. Choose the lowest number in the data set and add the Class Width
4. **Find Upper limit (UL)** Use the Lower Limit of the next bin to find the UL
5. Make a list of the LL and UL, as follows:

6. Go through the data and determine the occurrences within each bin:

7. Determine Relative Frequency

8. Determine Cumulative Frequency

LL	UL	Frequency	Relative Frequency	Cumulative Frequency
0	99	11	11/40=0.275	11
100	199	12	12/40=0.3	23
200	299	14	14/40=0.35	37
300	399	1	1/40=0.025	38
400	499	2	2/40=0.05	40

2.3 Types of Plots (purposes)

1. Histograms – visually displays the shape of the distribution of the data, shows location of the center, spread, if there are outliers (i.e., gas prices)

2. Frequency Polygons – uses line segments connected to points located directly above class midpont values for each bin (i.e., IOP)

$$\text{Mid Point} = \frac{UL - LL}{2}$$

3. Bar Graphs & Bar Plot - used of equal width to show frequencies of categories (i.e., Political Party)

4. Pareto Charts - bar graph for categorical data, bars are arranged in descending order per frequencies, decrease left to right (i.e., Favorite Ice Cream)

5. Scatter Plots - shows the relationship between two variables (i.e., study hours vs. gpa)

6. Time Serie Plots - data collected at different time points (i.e., weather, finances, blood pressure)

7. Others: Dot Plots, Stem-and-Leaf Plots, and Pie Charts

2.4 Central Tendency (Measures of the Center)

New Notation
N : Population Size
n : Sample Size
x_i i^{th} observation within population/sample
$\sum$:

1. Mean - the central or typical value in a set of data,

$$\mu = \sum_{i=1}^{N} = \frac{x_i}{N}$$

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

2. Median - Is the middle value of the original data values when they are arranged in increasing order.

   - - Case n is odd: The median is exactly the center value

   - - Case n is even: The median is the average of the two middle values

3. Mode - Value that occurs most frequently

4. Midrange – maximum value plus minimum value divided by two

$$\text{Midrange} = \frac{\text{Max} + \text{Min}}{2}$$

5. Weighted Mean - Each value has a different level of importance:

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

Problem 1 5.40,1.10, 0.42, 0.73, 0.48, 1.10

1. Mean 1.538
2. Median 0.915
3. Mode 1.10

Problem 2 27, 27, 27, 55, 55, 55, 88, 88, 99

1. Mean 57.89
2. Median 55
3. Mode 27, 55

Problem 3 1, 2, 3, 4, 5, 6, 7, 8, 9, 10

1. Mean 5.5
2. Median 5.5
3. Mode NA

3 Lecture 3 Notes: Measures of Variation. The Boxplot. Definition of Probability

3.1 Week 1 Review

Creativity is more than just being different. Anybody can plan weird; that's easy. What's hard is to be as simple as Bach. Making the simple, awesomely simple, that's creativity. Charles Mingus

1. Population Parameters and Sample Statistics
2. Interval/Ratio/Ordinal/Nominal
3. Graphing Data
4. Central Tendency
5. Skewness (Mean and Median Relationship)

3.2 Measures Variation

Measure of spread/dispersion/variation:

1. Range: Max - Min
2. Variance: The average of the squared differences from the mean
$$s^2 = \sum_{i=1}^{n} \frac{(x_i - \bar{x})^2}{n-1} = \frac{n(\sum_{i=1}^{n} x_i^2) - \sum_{i=1}^{n} x_i^2}{n(n-1)} \tag{1}$$
3. Standard Deviation: Measure of the variation of observations about the mean.
$$s = \sqrt{\sum_{i=1}^{n} \frac{(x_i - \bar{x})^2}{n-1}} \tag{2}$$
4. Coefficient of Variation: Ratio of the standard deviation to the mean (normalized measure of dispersion)
$$CV = \frac{\sigma}{\mu} * 100\% : \text{Population} \tag{3}$$
$$CV = \frac{s}{\bar{x}} * 100\% : \text{Sample} \tag{4}$$
5. Interquartile Range: (Soon)

Procedure Through Example: Image we have 5 numbers $(5, 9, 1, 7, 3)$ and you have to find the first measures of spread.

1. **Find Range**: Max - Min $9 - 1 = 8$
2. **Find Variance**:
   - Find Mean: 5 (verify this)
   - Square the difference between each number and the mean and add them
     - $(5-5)^2 + (9-5)^2 + (1-5)^2 + (7-5)^2 + (3-5)^2 = 40$
   - Take sum and divide it by the sample size minus 1 $(n - 1)$
     - $s^2 = \frac{40}{4} = 10$
3. **Find Standard Deviation**:
   - $s = \sqrt{s^2} = \sqrt{10} = 3.1623$
4. **Coefficient of Variation:**:
   - $\frac{s}{\bar{x}} * 100\% = 63.246\%$

3.3 Empirical Rule

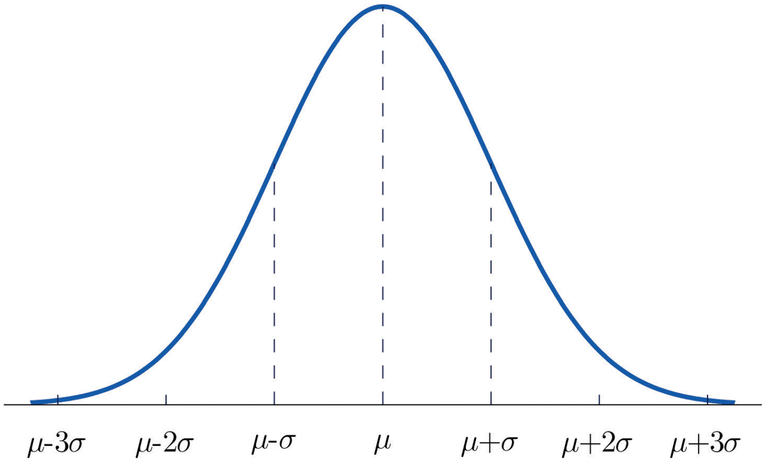
Rules for the Bell Shaped Distribution (Show picture and fill)

1. 68 %: data falls falls with 1 standard deviation of the mean
2. 95 %: data falls falls with 2 standard deviation of the mean
3. 99.7 %: data falls falls with 3 standard deviation of the mean

Example 2: Heights of women have a bell-shaped distribution with mean of 163 cm and a standard deviation of 6 cm. What percentage of women have heights between 151 and 175 cm?

3.4 Random Variable

X denotes a random a number that can be any number within a population



3.5 Z-Score

How far is the random variable away from the mean? Use the Z-score. Its units are standard deviations.

$$Z = \frac{X - \mu}{\sigma} \quad \text{Population} \tag{5}$$

$$Z = \frac{X - \bar{x}}{s} \quad \text{Sample} \tag{6}$$

First Definition of an **Outlier**:

- Ordinary values: $-3 \leq Z \leq 3$
- Outlier: $Z \leq -3$ **or** $Z \geq 3$

Example 3: $X = 36$ inches (Radius of RedWood Tree) and the average is $\bar{x} = 33.25$ and standard deviation $s = 1.71$. Find the z score and determine if it is an outlier.

3.6 Quartiles

Procedure For finding Quartiles

1. Sort Data
2. Q_2 (Second Quartile, Median): 50% of observations above it and 50% of observations below it.
3. Q_1 (First Quartile): Value with 75% of observations above it, and 25% below it. (same rules for even number of observations)
4. Q_3 (Third Quartile): 75% of observations are below it and 25% above it. (same rules for even number of observations)
5. Interquartile Range (IQR): $Q_3 - Q_1$

Second Definition of an **Outlier**:
Lower Fence $LF = Q_1 - 1.5 * IQR$

Upper Fence $UF = Q_3 + 1.5 * IQR$

Outlier if $X < LF$ **or** $X > UF$

3.7 Percentiles

Percentile of X is defined as:

$$\frac{\text{\# values less than X}}{\text{total \# of values}} \tag{7}$$

0	1	1	3	17	32	35	44	48	86
87	103	112	120	121	130	131	149	164	167
173	173	198	208	210	222	227	234	245	250
253	265	266	277	284	289	290	313	477	491

Example 4: Find Min Q_1, Q_2, and Q_3, Max. What percentile is 17? Is 17 an outlier?

- Min: 0
- Q_1: 86.75
- Q_2: 170
- Q_3: 250.8
- Max: 491

The percentile of 17 is $\frac{4}{40} = 0.1$.

$UF = 250.8 + 1.5 * (250.8 - 86.75) = 496.875$

$LF = 86.75 - 1.5 * (250.8 - 86.75) = -159.325$

So 17 is not an outlier.

3.8 Important Notions

Important Characteristics of a data set:

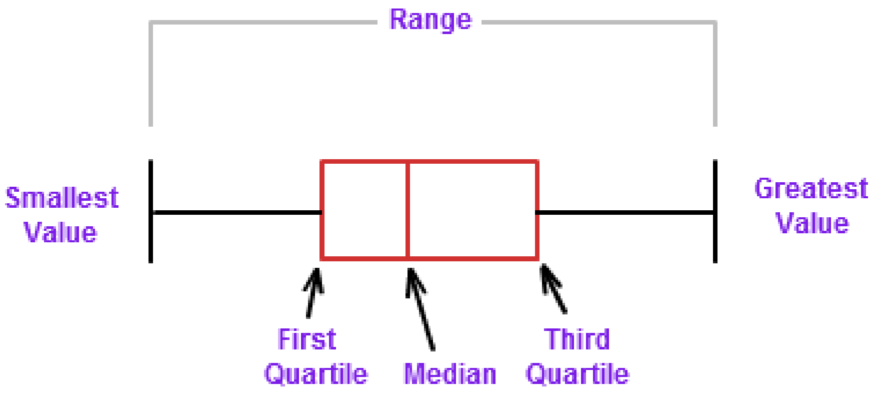
1. Center - mean/mode/median
2. Spread - variance/standard deviation/Range
3. Distribution - Symmetric, skewed right, or skewed left, bimodal

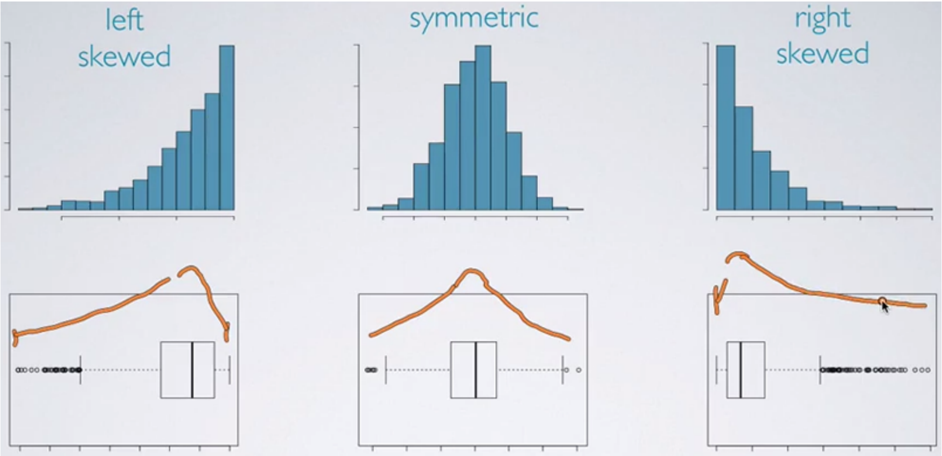
3.9 Boxplots

Boxplot are used to show a five number summary:

1. Min
2. Q_1
3. Q_2
4. Q_3
5. Max

3.9.1 Boxplots and Distributions





3.10 Probability

Probability: Underlying foundation of inferential statistics
Definitions:

- **An Event:** Any collection of results or outcomes of a procedure.
  Example: Tossing 1 die (a procedure) and getting even numbers:
  $A = \{2, 4, 6\}$
- **A Simple Event:** It is an outcome or event that can not be further broken down into simple pieces.
  Example: Outcomes when you roll a die: $\{1\}$ or $\{2\}$ or $\{3\}$ or $\{4\}$ or $\{5\}$ or $\{6\}$
- **Sample Space:** All possible simple events for a procedure.
  Example: Tossing a die. Possible outcomes are $S = \{1, 2, 3, 4, 5, 6\}$

Notation:

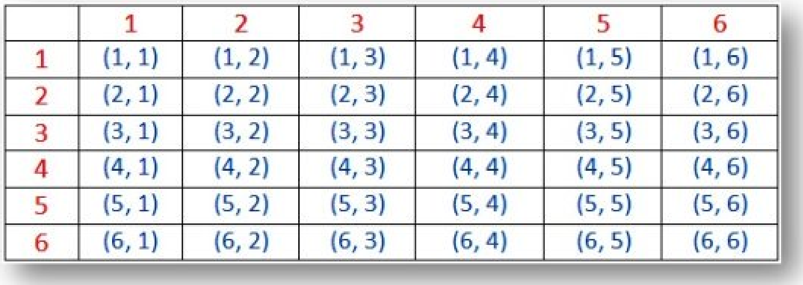
- P denotes Probability
- A, B, C denote specific events
- $P(A)$ denotes the probability of event A occurring

Example 5 Procedure: Rolling 1 die. Simple Event: $\{1\}$
Sample Space: $S = \{1, 2, 3, 4, 5, 6\}$

1. Find the Probability of a Particular Event A defined as rolling a 1

$$P(A) = \frac{\text{Number of Event A}}{\text{Total number of Events}} = \frac{1}{6}$$

Example 6 Procedure: Rolling two dice. Simple Events, were $\{\text{1st die, 2nd die}\}$



	1	2	3	4	5	6
1	(1, 1)	(1, 2)	(1, 3)	(1, 4)	(1, 5)	(1, 6)
2	(2, 1)	(2, 2)	(2, 3)	(2, 4)	(2, 5)	(2, 6)
3	(3, 1)	(3, 2)	(3, 3)	(3, 4)	(3, 5)	(3, 6)
4	(4, 1)	(4, 2)	(4, 3)	(4, 4)	(4, 5)	(4, 6)
5	(5, 1)	(5, 2)	(5, 3)	(5, 4)	(5, 5)	(5, 6)
6	(6, 1)	(6, 2)	(6, 3)	(6, 4)	(6, 5)	(6, 6)

1. Find the Probability of a Particular Event B defined as rolling two dice equal to one of the pairs below

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

$$P(B) = \frac{\text{Number of Event B}}{\text{Total number of Evens}} = \frac{6}{36} = \frac{1}{6}$$

Practice Problems Events
$A = \{\{1,1\}, \{2,1\}, \{3,1\}, \{4,1\}, \{5,1\}, \{6,1\}\}$.
$B = \{\{1,1\}, \{2,2\}, \{3,3\}\}$.
$C = \{\{1,3\}, \{2,6\}, \{3,6\}, \{4,1\}\}$.

1. $P(A) = \frac{6}{36} = \frac{1}{6}$
2. $P(B) = \frac{3}{36} = \frac{1}{12}$
3. $P(C) = \frac{4}{36} = \frac{1}{9}$

4 Lecture 4 & 5 Notes: Introduction to Probability. Probability Rules. Independence and Conditional Probability. Bayes Theorem. Risk and Odds Ratio

Wrong is right. Thelonious Monk

4.1 Three Definitions of Probability

Definition 1: Relative Frequency Approximation

$$P(A) = \frac{\text{Number of times A occurred}}{\text{Number of times a trial was repeated}}$$

Example 1: P(A new born infant will live to see his or her first birthday in any given year and location)

Definition 2: Subjective Probability. P(A) is estimated by using previous knowledge.
Example 2: P(Candidate A wins an election)

Definition 3: Classical approach (requires equally likely outcomes). If event A can occur in s of n ways, then:

$$P(A) = \frac{\text{Number of ways A can occur}}{\text{Number of different simple events}}$$

Example 3: P(Getting a 4) when you roll a balanced die= 1/6

Law of Large Numbers: When a procedure is repeated again and again, the relative frequency probability tends to approach the actual probability.

4.2 Complement of Probability

Complement of a an event:
Notation: $\bar{A}$ or A^c
Consists of all outcomes in which event A does not occur.

Example 4: Event A is rolling a die and getting a 5
$P(\text{Getting a 5}) = P(A) = 1/6$
$P(\text{Not Getting a 5}) = P(\bar{A}) = 1 - 1/6 = 5/6$

4.3 Compound Event

A compound event: Is any event combining two or more simple events

Example 5: Getting an even number when rolling a die=$\{2, 4, 6\}$
Remember a simple event is just a single event, for example $\{3\}$

4.4 Rules of Probability

The Rules: For any event A

1. $0 \leq P(A) \leq 1$
2. If $P(A) = 1$, A always occurs and $P(\bar{A}) = 0$
3. If $P(A) = 0$, A never occurs and $P(\bar{A}) = 1$
4. $P(A) + P(\bar{A}) = 1 \rightarrow P(\bar{A}) = 1 - P(A)$

4.5 Union, Intersection, and Disjoint Events

Example 6: Say you have a die and a coin

1. Event A = Getting a head when a coin is tossed
2. Event B= Getting a 5 when a single die is rolled

Union of Events: The union of two sets is a new set that contains all of the elements that are in both sets.

$$\begin{aligned} P(A \cup B) &= P(A \text{ or } B) \\ &= P(\text{Event A occurs or event B occurs or they both occur}) \\ &= P(\text{Getting a head or getting a 5}) \end{aligned}$$

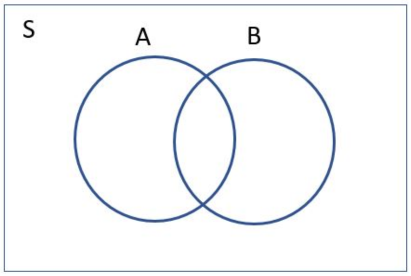
Intersection OF of Events: The intersection of two sets is a new set that contains the shared elements that are in both sets.

$$\begin{aligned} P(A \cap B) &= P(A \text{ and } B) \\ &= P(\text{Event A occurs and event B occurs simultaneously}) \\ &= P(\text{Getting a head and getting a 5}) \end{aligned}$$

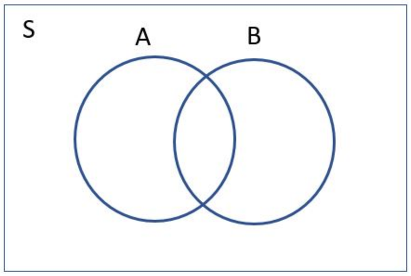
Disjoint Events (or mutually exclusive): They can not occur simultaneously

4.6 Venn Diagram

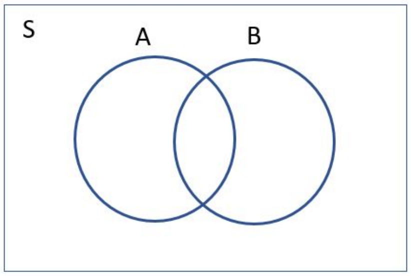
S: Sample Space
Union of A or B can be seen as $P(A \cup B)$



Intersection of A and B can seen as $P(A \cap B)$



Complement of A, $P(\bar{A})$



4.7 Probability Rules

Very important rule when calculating probabilities:

1. Addition Rule
2. Conditional Probability
3. Independence Events
4. Multiplication

Addition Rule
$P(A \cup B) = P(A) + P(B) - P(A \cap B)$
Same thing $P(A \cap B) = P(A) + P(B) - P(A \cup B)$

If the events are disjoints $P(A \cap B) = 0$
In this case $P(A \cup B) = P(A) + P(B)$

Example 7:
A die is rolled. What is the probability of getting a 1 or a 6?. $A = \{1\}$; $B = \{6\}$. A and B are mutually exclusive.

$$\begin{aligned} P(A \cup B) &= P(A) + P(B) - P(A \cap B) \\ &= P(A) + P(B) \\ &= 1/6 + 1/6 \\ &= 1/3 \end{aligned}$$

Conditional Probability: Probability of an event B given A occurs

$$\begin{aligned} P(B|A) &= P(\text{event B occurs given (after) event A has already occurred}) \\ &= P(B \text{ given } A) \end{aligned}$$

Independence Events Two events A and B are independent if the occurrence of one does not affect the probability of occurrence of the other. This means $P(B|A) = P(B)$ (Run to venn diagrams)

Example 8: When tossing a coin twice

$$P(\text{Head in a 2nd}|\text{Tail in a 1st}) = P(\text{Head in the 2nd}) = 1/2$$

Multiplication Rule Probability of an event A and B can be expressed as the probability of A multiplied by the the probability of B given A this can be expressed

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

If A and B are independents:

$P(A \cap B) = P(A \text{ and } B) = P(A)P(B) \quad (\text{because } P(B|A) = P(B))$

Example 9: *A* coin is tossed and a die is rolled.
P(Getting a head and Getting a 5) = P(H and 5)

$$\begin{aligned} P(A \cap B) &= P(A)P(B|A) \\ &= P(A)P(B) \quad A \text{ and } B \text{ are independent} \\ &= (\frac{1}{2})(\frac{1}{6}) \\ &= \frac{1}{12} \end{aligned}$$

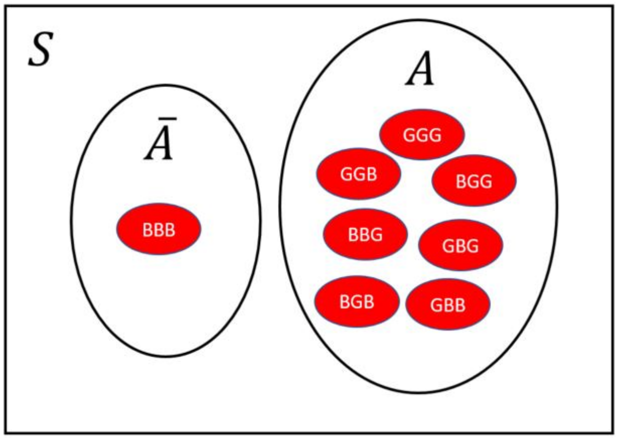
Many ways to the top of the mountain, choose one.... ***Many ways to the top of the mountain, know many.*** Immanuel Williams

4.8 A Probability Notion:

Probability of "at least 1": This statement means one or more.

It is **EASIER** to find the complement of one or more.

Example 10: If a two parents had 3 children, what is the probability that at least 1 one the children is a girl (assuming boys and girls are equally likely).



Event A: Getting at least 1 girl among 3 children.
Event $\bar{A}$: No girls among the 3 children. All children are Boys.

$$\begin{aligned} P(\bar{A}) &= P(\text{Boy} \cap \text{Boy} \cap \text{Boy}) \\ &= P(\text{Boy})P(\text{Boy})P(\text{Boy}) \text{ Independence} \\ &= \frac{1}{2} \text{ x } \frac{1}{2} \text{ x } \frac{1}{2} \\ &= \frac{1}{8} \end{aligned}$$

P(Getting at least 1 girl among 3 children)= $P(A) = 1 - P(\bar{A}) = 1 - \frac{1}{8} = \frac{7}{8}$

Example 11: In a garden of 50 flowers, the probability that a flower does not bloom is 0.5% = 0.005. What is the probability of getting at least one flower that does not bloom in the garden of 50 flowers?

$P(1 \text{ flower not blooming}) = 0.005$
$P(1 \text{ flower blooming}) = 0.995$

Event A: At least 1 flower not blooming.
Event $\bar{A}$: All flowers blooming.

$$\begin{aligned} P(A) &= 1 - P(\bar{A}) \\ &= 1 - 0.995 \text{ x } 0.995 \text{ x } 0.995 \text{ x } 0.995 ... \text{ x } 0.995 \\ &= 1 - 0.995^{50} \\ &= 1 - 0.7783 \\ &= 0.2217 \end{aligned}$$

4.9 Conditional Probability & Bayes Theorem

Recall **Conditional Probability**:

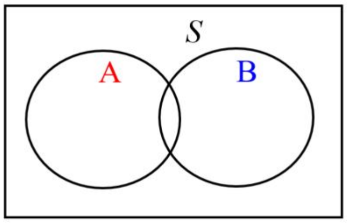
- $P(A \text{ given } B) = P(A|B)$
- $P(B \text{ given } A) = P(B|A)$

These statements are equivalent:

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

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

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



Bayes Theorem:

$$\begin{aligned}
P(A|B) &= \frac{P(A \cap B)}{P(B)} \tag{8}\\
&= \frac{P(B|A)P(A)}{P(A \cap B) + P(\bar{A} \cap B)} \tag{9}\\
&= \frac{P(B|A)P(A)}{P(A)P(B|A) + P(\bar{A})P(B|\bar{A})} \tag{10}\\
& \tag{11}
\end{aligned}$$

4.10 Risk & Odds

Imagine a prospective study.

	Disease	No Disease	
Treatment	a=100	b=400	a+b=100+400
Placebo	c=250	d=250	c+d=250+250

Risk: is given as a probability value.

$p_t = P(\text{Disease}|\text{treatment}) = \frac{100}{100+400} = \frac{1}{5}$

p_t : Proportion (or incidence rate) of some characteristic in the treatment group.
Example: Proportion of heart attacks in the treatment group

$p_c = P(\text{Disease}|\text{control}) = \frac{250}{250+250} = \frac{1}{2}$

p_c : Proportion (or incidence rate) of some characteristic in the control group.
Example: Proportion of heart attacks in the control group

Absolute Risk Reduction:

$$\begin{aligned} |P(\text{event in treatment}) - P(\text{event in control})| &= |\frac{100}{100+400} - \frac{250}{250+250}| \\ &= |p_t - p_c| \end{aligned}$$

Relative Risk:

$$\text{Relative Risk} = \frac{p_t}{p_c} = \frac{\frac{1}{5}}{\frac{1}{2}} \text{ Rick Ratio}$$

Odds Ratio:

$$\text{Odds of Event} A = \frac{P(A)}{P(\bar{A})}$$

$$\text{Odds against Event} A = \frac{P(\bar{A})}{P(A)}$$

Relative Odds:

$$\text{Odds Ratio} = \frac{\text{odds in favor of treatment group}}{\text{odds in favor of control group}}$$

$$\text{Odds Ratio} = \frac{ad}{bc}$$

Use relative risk and/or odds ratio in Prospective studies.

Use odds ratio only in Retrospective studies. In this case the incidence rate of an event might be the result of the study design and not the true incidence rate.

5 Lecture 6: Discrete Distributions

Wish you would learn to love people and use things, and not the other way around. Aubrey Graham
Friday's Session

- Fri Apr 21 2017
- 10:40AM - 1:30PM
- Space Assignment(s):Rachel Carson Acad 252

5.1 Random Variables

A **random variable** is a variable (denoted by X or x) that has a single event, determined by chance, can be any outcome of interest in the sample space.

- X Random variable
- x Observed variable

Example 1: Rolling a die, the outcomes can be $1, 2, 3, 4, 5, 6$. X can be any of these values.

Example 2: Tossing a coin, the outcomes can be H or T. X can be H or T.

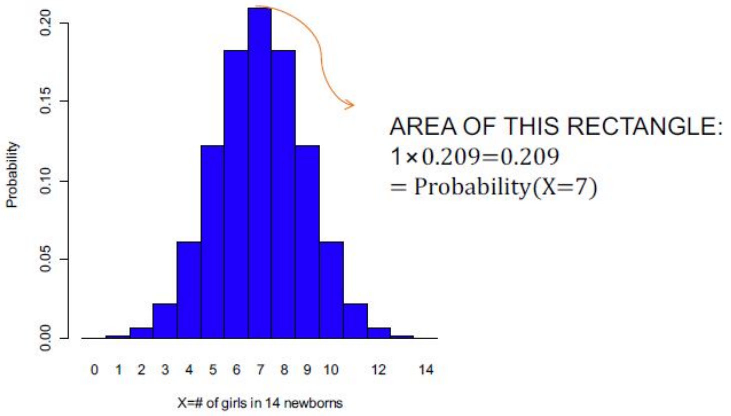
All events in the sample space has an associated probability.

Example 3: Rolling a die.

X	$P(X)$
1	1/6
2	1/6
3	1/6
4	1/6
5	1/6
6	1/6

This table is a **probability distribution** which assigns a probability to each of the random variables possible outcomes. This can also be used within a graph or a formula.

Example 4: The following is a probability distribution for number of new born girls.



What does the x-axis represent?

What does the y-axis represent?

5.2 Discrete and Continuous Distributions.

These random variables can be one of two types of variables.

1. Discrete Variables - has either a finite number of values or a countable number of values (countable means, an event that can be counted, can be infinity)
   - number of dogs owned
   - number of friends on Facebook
   - number of texts a day (countable)
2. Continuous Variables - has infinitely many values, and those values can be associated with measurements on continuous scale without gaps
   - GPA
   - velocity of a car
   - weight

Two Requirements for a Probability Distribution

1. $\sum P(X) = 1$ X is all values in the sample space
2. $0 \leq P(X) \leq 1$ probability of an event X will alway be between 0 and 1 (the probability of X can equal 0 or 1)

Important. Important. Important.
Finding the mean, variance, and standard deviation from a distribution is done differently.

Mean, Expected Value

$$E = \mu = \sum XP(X)$$

Variance

$$\begin{aligned} \sigma^2 &= \sum[(X-\mu)^2 P(X)] \\ &= \sum[(X^2 P(X))] - \mu^2 \end{aligned}$$

Standard Deviation

$$\sigma = \sqrt{\sum[(X^2 P(X))] - \mu^2}$$

Review of **Outliers** You have outliers or unusual values if a value goes beyond

1. Maximum Value: $\mu + 3\sigma$
2. Minimum Value: $\mu - 3\sigma$

Example 5: X is the number of girls from 14 babies

X	$P(X)$	$XP(X)$	X^2	$X^2P(X)$
0	0.000	0.000	0	0.000
1	0.001	0.001	1	0.001
2	0.006	0.012	2	0.024
3	0.022	0.066	9	0.198
4	0.061	0.244	16	0.976
5	0.122	0.610	25	3.050
6	0.183	1.098	36	6.588
7	0.209	1.463	49	10.241
8	0.183	1.464	64	11.712
9	0.122	1.098	81	9.882
10	0.061	0.610	100	6.100
11	0.022	0.242	121	2.662
12	0.006	0.072	144	0.864
13	0.001	0.013	169	0.169
14	0.000	0.000	196	0.000

Find μ, σ^2, and σ. Procedure:

1. Find $\mu = \sum XP(X) = 6.993$
2. Find $\sum X^2P(X) = 52.467$

Mean, Expected Value

$$\mu = \sum XP(X) = 6.993 \approx 7$$

Which value for X has the highest probability?

It is expected to have 7 girls among 14 newborn babies.

Variance

$$\sigma^2 = \sum 52.467 - 6.993^2 = 3.564951 \approx 3.6 \text{ girls}^2$$

Standard Deviation

$$\sigma = \sqrt{\sum[(X^2P(X))] - \mu^2} = \sqrt{3.564951} = 1.9 \text{ girls}$$

Usual Values Maximum usual value: μ + $3\sigma = 7.0 + 3(1.9) = 12.7$ girls

Minimum usual value: μ - $3\sigma = 7.0 - 3(1.9) = 1.3$ girls

Extreme Values

For 14 randomly selected babies, the number of girls usually falls between 1.3 and 12.7. The probability of unusual events $P($ 13 or more girls$) = P(X \geq 13) = P(X = 13) + P(X = 14)$. This is $0.001 + 0.000 = 0.001$ (LOW value). This implies it is unusual to get 13 girls or more. This event would not happen by chance.

5.3 The Binomial Distribution

Introduction:
Tossing one coin follows a Bernoulli Distribution
The Random Variable X is Heads $X = 1$ or Tails $X = 0$
$P(X = \text{ Heads}) = P(X = 1) = 1/2$ $P(X = \text{ Tails}) = P(X = 0) = 1/2$

Binomial Distribution: Requirements

- Suppose a fixed number of trials (Ex. Flip a coin a n of times)
- The trials must be independent (Ex. flips do not affect each other)
- Each trial must have all outcomes classified into 2 categories (Ex. Tail or Head, disjoint)
- The probabilities must remain constant for each trial (Ex. P(head)=1/2, and this does not change)

Random Variable: X Meaning of X: Number of successes in n trials

Examples 6:

1. Getting 6 heads in 10 tosses of a coin. In a fair coin probability of head is 1/2.
2. Getting 3 correct answers in a multiple choice 5 question exam (student is unprepared. Each question has 5 possibilities (a,b,c,d,e)). Probability of getting a correct answer at any question is 1/5.
3. Hospital records show that of patients suffering from a certain disease, 75% die of it. Of 6 randomly selected patients, 4 will recover. P(recovering)= 1 – P(No recovering)= 1-0.75=0.25

Notation:

1. n = Number of trials
2. X = number of successes in n trials
3. p = Denotes the probability of success
4. q = Probability of failure $= 1 - p$

Note: Success does not necessarily mean something good!!!!!

Examples 6 CONTD:

1. Getting 6 heads in 10 tosses of a coin. In a fair coin probability of head is 1/2.
    - $n = 10$
    - $X = 6$
    - $p = 0.5$
    - $q = 0.5$
2. Getting 3 correct answers in a multiple choice 5 question exam (student is unprepared. Each question has 5 possibilities (a,b,c,d,e)). Probability of getting a correct answer at any question is 1/5.
    - $n = 5$
    - $X = 3$
    - $p = 0.2$
    - $q = 0.8$
3. Hospital records show that of patients suffering from a certain disease, 75% die of it. Of 6 randomly selected patients, 4 will recover. P(recovering)= 1 − P(No recovering)= 1-0.75=0.25
    - $n = 6$
    - $X = 4$
    - $p = 0.25$
    - $q = 0.75$

To find probabilities we must use the **binomial probability distribution**, which can be seen as

$$P(X = x) = \frac{n!}{(n-x)!x!} p^x q^{(n-x)} \tag{12}$$

where $x = 0, 1, 2, ..., n$

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

Examples 6 CONTD:
Getting 6 heads in 10 tosses of a coin. In a fair coin probability of head is 1/2. P(6 heads from 10 flips of a coin)

$$
\begin{aligned}
P(X=6) &= \binom{10}{6} 0.5^6 (1-0.5)^{(10-6)} && (13)\\
&= \frac{10!}{6!4!} 0.5^6 0.5^4 && (14)\\
&= 210(0.5)^1 0 && (15)\\
&= 0.205 && (16)
\end{aligned}
$$

Getting 3 correct answers in a multiple choice 5 question exam (the student is unprepared. Each question has 5 possibilities (a,b,c,d,e)). Probability of getting a correct answer at any question is 1/5. P(3 correct answers in a 5 question exam)

$$
\begin{aligned}
P(X=6) &= \binom{5}{3} \left(\frac{1}{5}\right)^3 (1-\frac{1}{5})^{(5-3)} && (17)\\
&= \frac{5!}{3!2!} 0.2^3 0.8^2 && (18)\\
&= 10(0.2)^3(0.8)^2 && (19)\\
&= 0.0512 && (20)
\end{aligned}
$$

Probability of getting at least 3 correct answers out of five. This equivalent to find the P(3 correct answers) + P(4 correct answers) + P(5 correct answers) = $P(X=3)+P(X=4)+P(X=5)=0.051+0.006+0=0.057$

You can find the mean, variance, standard deviation, maximum usual value and minimum usual value for the binomial distribution with special formulas

Mean, Expected Value

$$E = \mu = \sum XP(X) = np$$

Variance

$$\sigma^2 = \sum[(X^2P(X))] - \mu^2 = npq$$

Standard Deviation

$$\sigma = \sqrt{\sum[(X^2P(X))] - \mu^2} = \sqrt{npq}$$

Outliers

1. Maximum Value: $np + 3\sqrt{npq}$
2. Minimum Value: $np - 3\sqrt{npq}$

Example 7: A study shows that 10% of Americans adults are left-handed. A statistics discussion has 25 students in attendance. What is the probability 3 people are left-handed.
Part 1. P(3 people are left-handed)
Information:

- X is the number of left-handed people in class
- $n = 25$
- $X = 3$
- $p = 0.1$
- $q = 0.9$

$$\begin{aligned} P(X=3) &= \binom{25}{3}\left(\frac{1}{10}\right)^3(1-\frac{9}{10})^{(25-3)} \tag{21}\\ &= \frac{25!}{3!22!}0.1^3 0.9^{22} \tag{22}\\ &= 10(0.2)^3(0.8)^2 \tag{23}\\ &= 0.226(Rounded) \tag{24} \end{aligned}$$

Part 2. Find the mean and standard deviation of left handed students in the discussion.

1. $\mu = np = 25(0.1) = 2.5$ left handed students
2. $\sigma = \sqrt{npq} = \sqrt{25(0.1)(0.9)} = 1.5$ left handed students

Part 3. Would it be unusual to find a discussion of 25 students with 5 left-handed students?

1. Maximum Value: $np + 3\sqrt{npq} = 2.5 + 3(1.5) = 7$
2. Minimum Value: $np - 3\sqrt{npq} = 2.5 - 3(1.5) = -2$

5 is an usual value because it is between the max and min.

"The worst thing you can do about a situation is nothing." O'Shea Jackson Sr.

5.4 The Poisson Distribution.

Description of the Poisson Distribution

- Discrete probability distribution.
- The random variable is the number of occurrences (counts) of an event in an interval
- The interval can be: time, distance, area, volume, or some similar unit.

EXAMPLES:

- Number of earthquakes (at least 6.0 on the Richter scale) in the last 100 years
- Number of patients arriving at the Emergency Room on Fridays between 10:00 pm and 11:00 pm
- Number of buses that pass a bus stop within an hour

Poisson Distribution: Requirements

- Random variable X is the number of occurrences of an event over some interval
- The occurrences must be random
- The occurrences must be independent

To find probabilities we must use the Poisson probability distribution, which can be seen as

$$P(X = x) = \frac{\mu^x \exp{-\mu}}{x!} \tag{25}$$

where $x = 0, 1, 2, 3, 4, ...$ and $e \approx 2.71828$ (Euler's number) The Poisson distribution only depends on μ (the mean of the process).

You can find the mean, variance, standard deviation, maximum usual value and minimum usual value for the Poisson distribution with special formulas

Mean, Expected Value

$$E = \mu = \sum XP(X) = \mu(\text{\# occurrences within interval})$$

Variance

$$\sigma^2 = \sum[(X^2P(X))] - \mu^2 = \mu(\text{Variance is equal to the Mean})$$

Standard Deviation

$$\sigma = \sqrt{\sum[(X^2P(X))] - \mu^2} = \sqrt{\mu}((\text{Standard deviation is the square root of the mean})$$

Example 8: In a recent year, there were 4500 births at NYU Langone Medical Center. Assume that the the number of births each day is about the same, and assume that the Poisson Distribution is a suitable model.

1. Find μ, the mean number of births per day.

   $$\mu = \frac{\text{Number of births}}{\text{Number of days}} = \frac{4500}{365} = 12.3288$$

   On average, 12 babies per day

2. Find the probability that on a randomly selected day, there are exactly 8 births, $P(X = 8)$. **Use unrounded value for mean.**

   $$P(X = 8) = \frac{\mu^x e^{-\mu}}{x!} = \frac{12.3288^8(e^{-12.3288})}{8!} = 0.0585$$

   The probability of obtaining 8 babies in 1 day us 0.0585.

6 Lecture 7: Normal Distribution

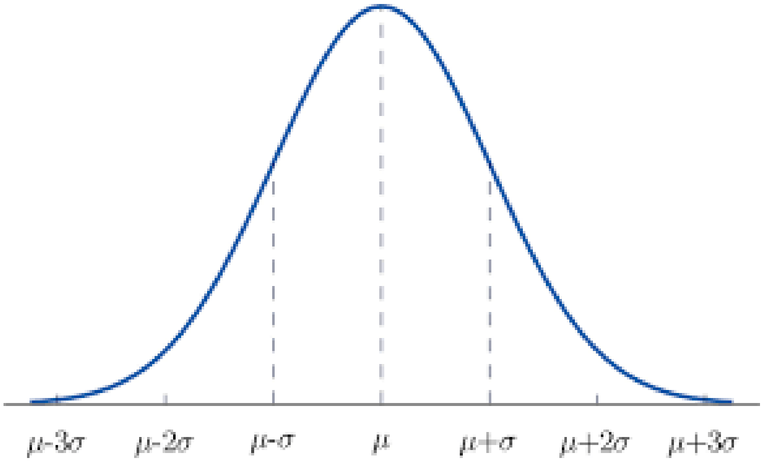
Random Variable X is continuous

- Random Variable X is continuous
- Most experiments use this to model events
- **An assumption about the sample mean uses this model (BIG)**
- Bell-shaped: Curve is symmetric around the mean μ
- Distribution is determined by two parameters: the mean μ and standard deviation σ

$$P(X = x) = \frac{1}{\sqrt{(2\pi\sigma^2)}} \exp^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

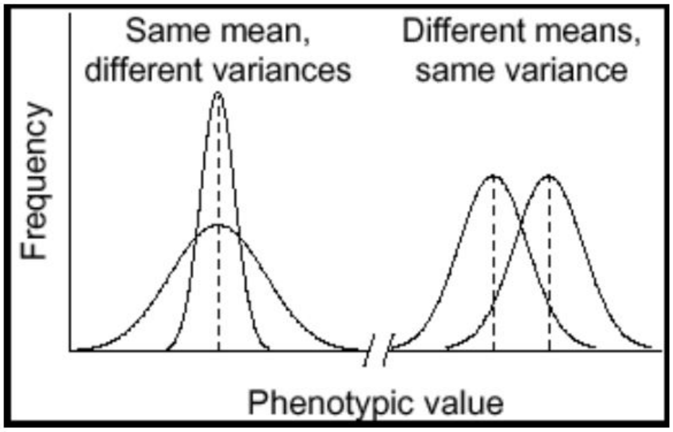
Properties for a continuous probability distribution:

1. The total area under the curve must be equal to 1
2. $0 \leq P(X) \leq 1$
3. Graphs is called a density curve



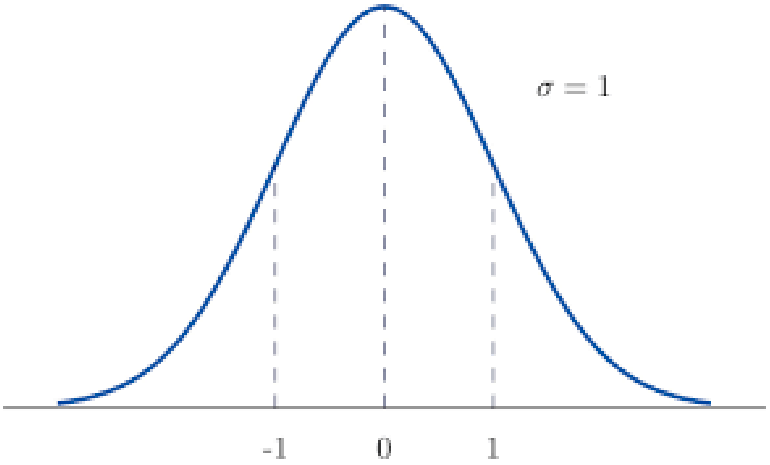
Understanding Mean and Variance

- Same mean, different variances
- Different means, same variances



When $\mu = 0$ and $\sigma = 1$, the Normal Distribution becomes a Standard Normal Distribution

$$P(X = x) = \frac{1}{\sqrt{(2\pi)}} \exp^{-\frac{x^2}{2}}$$



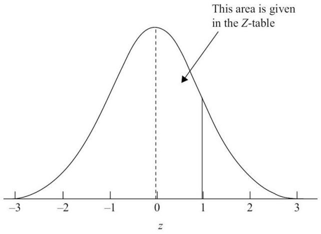
This distribution is used to calculate the probability of a random event given that the data follows a normal distribution. In order to use this distribution, we must be given the following:

1. x: Observed Variable(s)

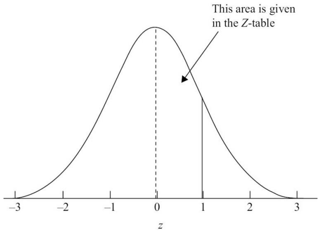
2. μ or $\bar{x}$: Population or Sample Mean
3. σ or s: Population or Sample Standard Deviation

Calculate Z-Score and Probability

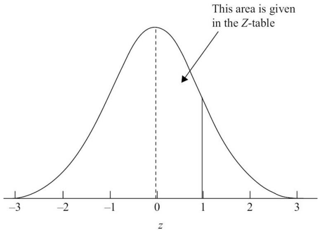
Example 1: It is assumed that the weight of the 49s football team follows a normal distribution. The population mean is $\mu = 259.62$ and $\sigma = 15.25$, find the following: $P(X \leq 243.2)$



$P(X \geq 265.33) = 1$ - $P(X \leq 265.33)$



$P(253.66 \leq X \leq 264.2) = P(X \leq 264.2) - P(X \leq 253.66)$



Example 2: If the mean quiz score was 15.5 and standard deviation was 1.1, you were the 95th percentile, what was your score? Reverse engineer the Z-score equation and use table.

$X = \mu + \sigma Z$

Question how do you find Z?

7 Lecture 8: Sampling Distribution, CLT, Binomial Approximation from Normal Distribution

Reality is wrong. Dreams are for real. – Tupac Shakur

7.1 Sampling Distribution

Example 1 It is known that the GPA **mean** at UCSC is $\mu = 3.22$ and **standard deviation** $\sigma = 0.15$. Imagine I ask 10 people their GPA and I record the mean and standard deviation from these 10 individuals. Imagine I do this a lot of times:

Groups	$\bar{x}$	s
Group 1	3.22	0.22
Group 2	3.89	0.15
Group 3	2.22	0.45
$\vdots$	$\vdots$	$\vdots$
Group 100,000	3.6	0.25

The means from each sample are random variables. Which makes sense, if we ask keep asking different samples we should get different means and standard deviations.

Which implies that the sample means can be model as a distribution. Which distribution do you think we will use?

What do you think the mean of this distribution will be?

BIG Concepts

1. The variability between the means will be smaller than the variability within each sample.
2. The sampling distributions of the mean allows us to determine the **probability** of a sample mean. One of the biggest reasons why we learned about probability is to discuss this idea.
   - Is the sample mean probable?
   - Is the sample mean not probable?

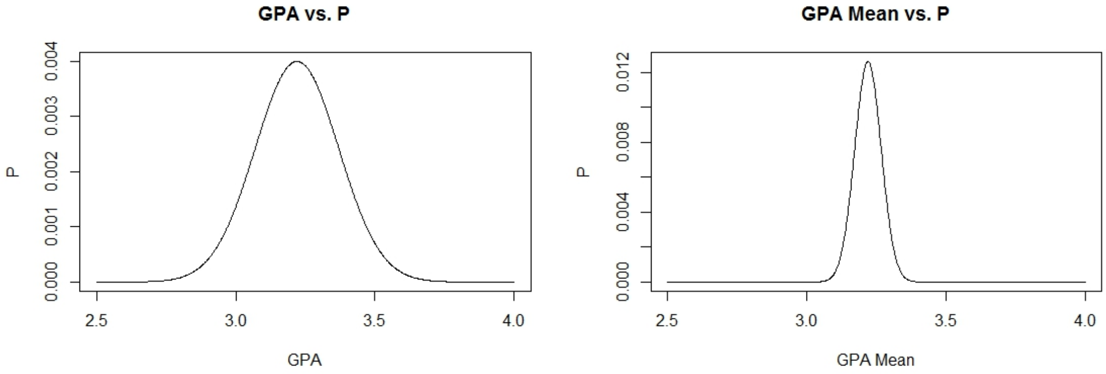


Figure 1: **Left Panel:** A random sample's distribution $(X_1, X_2, \cdots, X_n)$ **Right Panel:** The sample means' distribution $(\bar{x}_1, \bar{x}_2, \cdots, \bar{x}_n)$

When the sample size increases, the sample distribution of the sample mean becomes a normal distribution and becomes more and more narrower.

The larger the sample the closer the sample mean will resemble the population mean.

The sample mean will not vary as much as the sample size increases. Matter of fact, it well get closer to the true mean μ.

Here we are talking about the sample mean, but there are many other statistics

- **Sample Mean**
- Sample Median
- Sample Variance
- Sample Standard Deviation
- **Sample Proportions**

7.2 Central Limit Theorem

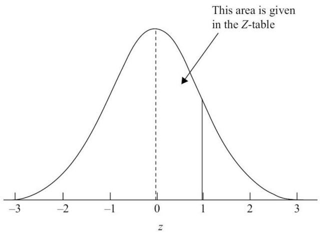
Based on these results we can conclude:

$$\bar{x} \sim N(\mu_{\bar{x}}, \sigma_{\bar{x}})$$

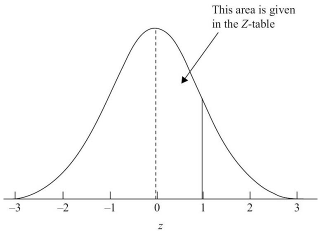
where $\mu_{\bar{x}} = \mu$ and $\sigma_{\bar{x}} = \sigma/\sqrt{n}$

Example 1: Implementing the Sampling Distribution and CLT
Part 1: What is the probability that a random individual's GPA is less than 3.15? Part 2: What is the probability that a random sample of 10 mean GPA is less than 3.15? Part 3: What is the probability that a random sample of 100 mean GPA is less than 3.15?

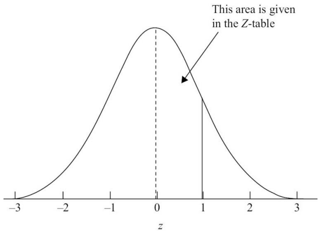
NOW LOOK
Random Individual $P(X \leq 3.15)$



Random Sample of size $n = 10$ $P(\bar{x} \leq 3.15)$



Random Sample of size $n = 100$ $P(\bar{x} \leq 3.15)$



7.3 Normal Approximation To the Binomial Distribution

It is the probability distribution of sample proportions, with all samples having the same sample size n.

Example 2: Use a random sample from the population, for example, $1,000$ people, to estimate the proportion who will vote for Bernie Sanders in the previous primary elections (using Polls).

IMPORTANT RESULT: Under certain conditions, the distribution of sample proportions approximates a normal distribution.

When working with a binomial distribution if $np \geq 5$ and $nq \geq 5$ the binomial random variable has a probability distribution that can be approximated by a normal distribution, with mean and standard deviation:

$$\mu = np$$
$$\sigma = \sqrt{npq}$$

We must verify that it is reasonable to approximate the binomial distribution by the normal distribution

Using the same equation as the Central Limit Theorem

$$\hat{p} \sim N(\mu_{\hat{p}}, \sigma_{\hat{p}})$$

where $\mu_{\hat{p}} = np$ and $\sigma_{\hat{p}} = \sqrt{npq}$

Example 3: Let X denote the people out of the sample of 200 people who do not eat avocado.

Each person is independent within this sample, each person can either eat avocado or not, and the probability of eating avocado is $p = 0.5$ (will given to you directly or *indirectly on exam*) for everyone in the sample.

We can use the binomial distribution to model the number of individuals, X, who eat avocado from a sample size of 200.

Find the probability that less than 120 people like avocado.
Some Calculations preliminary = calculations that we can do

1. With $n = 200, p = 0.5$
    - $q = 1 - p = 0.5$
    - $np = 200(0.5) = 100 \quad (np \geq 5)$
    - $nq = 200(0.5) = 100 \quad (nq \geq 5)$
2. Then we can calculate:
    - $\mu = np = 200(0.5) = 100$
    - $\sigma = \sqrt{npq} = \sqrt{200(0.5)(0.5)} = 7.07$

The probability that less than 120 people like avocado $\implies P(X < 120)$