GRADESAVERS

ADMS 2320
Test 1 Crash Course
Chapters 4, & 6

CHAPTER 4

Measures Of Central Location

In a given set of data, what value does the data cluster around?
3 types of measures for central location: mean, median, mode

Arithmetic Mean ("Average")

Population mean:

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

Sample mean:

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

Legend
$\sum$ - addition / sum
x_i - the i^{th} data point
n - the total data points in the sample
N - the total data points in the population

Example
A sample of 5 students have the following heights in inches:
72 56 60 65 70

What is the mean height?

Example
A random sample of 7 students chose the following random numbers.
-5 3 0 6.5 -2.4 -0.5 52

What is the mean random number?

Median

The median is the middle data point. You must arrange the numbers in increasing or decreasing order and then find the number in the middle. If there are two numbers in the middle, the average of both must be taken to find the median.

When data is arranged in order, the position of the median can be found using the following formula:

Position of median = $\frac{n+1}{2}$
n is the total number of data points.

Example
Find the median of the following dataset
-5 3 0 6.5 -2.4 -0.5 52

Example
Find the median of the following dataset
2 8 5 3 9 4

Mode

The most frequently occurring observation in a dataset. It is possible to have more than one mode. In general, it helps to have the numbers arranged in increasing or decreasing order to see which number occurs the most often.
Note: If every number occurs the same number of times, there is NO MODE.

Example
Find the mode of the following dataset
1 9 6 3 4 2 3 5 9 9

Example
Find the mode of the following dataset
1 5 -1 5 0 -1

Example
Find the mode of the following dataset
1 5 -1 5 0 -1 1 0

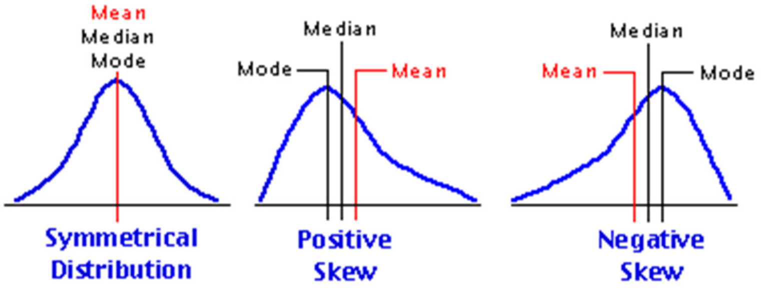
Which method to use: mean, median, or mode?

For interval data, any of them can be used, but usually it is the mean or median. If there are extreme observations in the dataset (some numbers way smaller or way larger), then usually the median is the better measure. <u>The median is not affected by extreme values while the mean is.</u>

For ordinal data, use median.

For nominal data, use mode. Note: For nominal data, finding the mode doesn't technically give us a 'central' location since it's not actually possible to find a center for nominal data.

Shape of Distribution Based on Mean/Median/Mode
Symmetric distribution: Mean = Mode = Median
Positively Skewed Distribution: Mode < Median < Mean
Negatively Skewed Distribution: Mode > Median >Mean



Geometric Mean

Used for finding average growth rate of a variable **over time.**

$$R_g = \sqrt[n]{(1+R_1)*(1+R_2)*\ldots*(1+R_n)} - 1$$
$$= [(1+R_1)*(1+R_2)*\ldots*(1+R_n)]^{1/n} - 1$$

R_g- Geometric mean
$R_\#$ - Rate of return in year 1, 2, 3, etc

Example

A $50 investment grows by 100% in year one to $100. In year 2, there is a loss of 50% bringing the investment value down to $50.

What is the geometric mean?

$R_1 = 100\% = 1 \qquad R_2 = -50\% = -0.5$

$R_g = \sqrt[2]{(1+1)*(1-.5)} - 1 = 0$

Therefore average growth **over 2 years** is 0%

What is the arithmetic mean?

$$\bar{R} = \frac{1+(-0.5)}{2} = 0.25$$

Therefore average growth rate per year is 25%

Example

An investment earned 5.2% the first year, earned 9.6% the second year, and lost 3.2% the third year. What is the average return over the 3 year period?

Measures of Variability

In a given set of data, how much variation is there between the data points?

Types of measures for variability:
-Range, Variance, Standard Deviation, Mean Absolute Deviation (and more later)

Range

Range = Largest value – Smallest Value
Note: Range is not a very useful measure for variability, but it is quick to calculate.

Variance & Standard Deviation & Mean Absolute Deviation

Measures of Variability	**Variance**	**Standard Deviation**	**Mean Absolute Deviation**
Population	$\sigma^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2$	$\sigma = \sqrt{\sigma^2}$	MAD = $\frac{1}{N}\sum_{i=1}^{N}\lvert x_i - \bar{x}\rvert$
Sample	$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$ Shortcut Calculation Formula: $s^2 = \frac{1}{n-1}\left[\sum_{i=1}^{n} x_i^2 - \frac{(\sum_{i=1}^{n} x_i)^2}{n}\right]$	$s = \sqrt{s^2}$	MAD = $\frac{1}{n}\sum_{i=1}^{n}\lvert x_i - \bar{x}\rvert$

Note:
If you know standard deviation, then you can calculate variance, and vice versa.

$Variance = SD^2$

$SD = \sqrt{Variance}$

Mean Absolute Deviation

Take the absolute difference (i.e. make negative numbers positive), between a data point and the mean. Find the average of these absolute differences to find the Mean Absolute Deviation.

Example

Given the following dataset of a sample measurement, find the following

A) What is the range

B) What is the variance and standard deviation?

C) What is the mean absolute deviation?

Dataset: 1.2 -5 6 9 4.4

x	$x-\bar{x}$	$(x-\bar{x})^2$	$\lvert x-\bar{x}\rvert$
1.2			
-5			
6			
9			
4.4			

x	x^2
1.2	
-5	
6	
9	
4.4	

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

$$s^2 = \frac{1}{n-1}\left[\sum_{i=1}^{n} x_i^2 - \frac{\left(\sum_{i=1}^{n} x_i\right)^2}{n}\right]$$

$$s = \sqrt{s^2}$$

$$MAD = \frac{1}{N}\sum_{i=1}^{N}\lvert x_i - \bar{x}\rvert$$

Example

Given the following dataset of a sample measurement, find the following

A) What is the range
B) What is the variance and standard deviation?
C) What is the mean absolute deviation?

Dataset: 10 2 12

x	$x-\bar{x}$	$(x-\bar{x})^2$	$\lvert x-\bar{x}\rvert$
100			
20			
120			

x	x^2
10	
2	
12	

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

$$s^2 = \frac{1}{n-1}\left[\sum_{i=1}^{n} x_i^2 - \frac{(\sum_{i=1}^{n} x_i)^2}{n}\right]$$

$$s = \sqrt{s^2}$$

$$MAD = \frac{1}{N}\sum_{i=1}^{N}\lvert x_i - \bar{x}\rvert$$

Percentile

Location of Pth percentile: $L_p = (n+1) * \frac{P}{100}$

1st Quartile (Q1) = 25th percentile
Median = 50th percentile
3rd Quartile (Q3) = 75th percentile

IQR = Interquartile Range = Q3-Q1

STEPS

Step 1) Make sure your date is arranged smallest to largest, and after arranging, write down location position by each number.
Step 2) Find Location of Pth percentile using $L_p = (n+1) * \frac{P}{100}$
Step 3) Find the Pth percentile

if L_p is a whole number, Pth percentile = the number at location L_p

if L_p is not a whole number, Pth percentile =$LL_p + d(UL_p - LL_p)$
LL_p is the number at L_p rounded down
UL_p is the number at L_p rounded up
d is the decimal portion of L_p

Example

Dataset: 1.2 -5 6 9 4.4 5.8 0 -15.2 -7 12 9

A) What is the 27th Percentile?
B) What is the 88th percentile?
C) What is the 50th percentile?
D) What is the IQR?

Interpreting Standard Deviation and Variance

Both numbers represent how much variability is in the data. The units for variance will be the original units squared, while the units for standard deviation are the units for the raw measurement. For example, if we were finding the variance of data collected on employee salaries, the variance would have units $\$^2$, but the standard deviation would have units of $.

Standard deviation tells us how far away a given data point is from the mean, on average.

Coefficient Of Variation

This term allows us to see how much relative variability there is between different datasets.
For example, a standard deviation of 10 when the data set contains numbers in the 1000's is not large. A standard deviation of 10 when the data set contains numbers less than 25 is large.

Population coefficient of variation: $CV = \frac{\sigma}{\mu}$

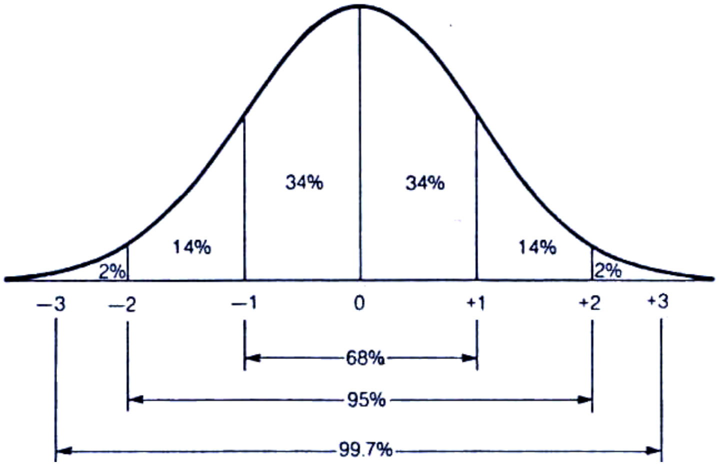
Sample coefficient of variation: $cv = \frac{s}{\bar{x}}$

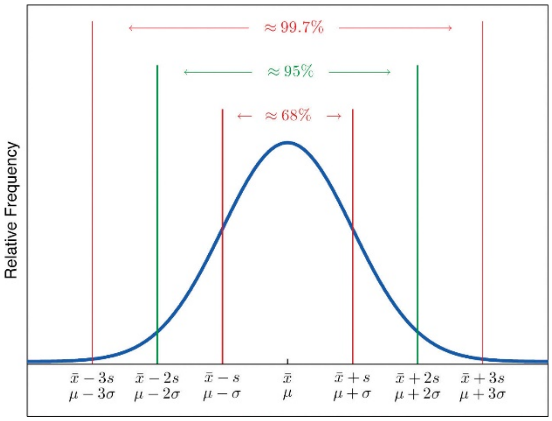
Example
Go back and calculate CV

Empirical Rule

Assuming the data distribution is bell-shaped:

68% of data lies with ± 1 Standard Deviation of the mean $(\bar{x} - s, \bar{x} + s)$ or $(\bar{x} - \sigma, \bar{x} + \sigma)$
95% of data lies with ± 2 Standard Deviation of the mean $(\bar{x} - 2s, \bar{x} + 2s)$ or $(\bar{x} - 2\sigma, \bar{x} + 2\sigma)$
99.7% of data lies with ± 3 Standard Deviation of the mean $(\bar{x} - 3s, \bar{x} + 3s)$ or $(\bar{x} - 3\sigma, \bar{x} + 3\sigma)$





Chebysheff's Theorem

For data that would **not fit** a bell-shaped distribution, the empirical rule cannot be used. Instead, we make use of the following theorem:

The minimum proportion of samples that lie within k standard deviations is : $1 - \frac{1}{k^2}$ for k > 1

k=2 → atleast 3/4 (75%) of data lies within 2 standard deviations
k=3 → atleast 8/9 (89%) of data lies within 3 standard deviations

Chebysheff's Theorem gives us a lower bound on how much data lies within a given standard deviation. The empirical rule gives pretty accurate approximations of the total amount of data that will lie within a given standard deviation.

Typically, any sort of skewed data would require the use of this theorem since it wouldn't be bell-shaped.

Linear Relationships

When a scatterplot is drawn with 2 sets of interval data, there may be a positive or negative linear relationship that exists between the data. Two types of measurements that tell us about linear relationships are covariance and coefficient of correlation.

Covariance

Population Covariance $\sigma_{xy} = \frac{\sum_{i=1}^{N}(x_i-\mu_x)(y_i-\mu_y)}{N}$

Sample Covariance $s_{xy} = \frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{n-1}$

Sample Covariance Shortcut: $s_{xy} = \frac{1}{n-1}\left[\sum_{i=1}^{n} x_i y_i - \frac{\sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{n}\right]$

Legend:
X and Y represent the 2 different variables, and x_i and y_i are the corresponding data points.n and N represent the number of data point pairs in the sample or population

Interpretation:
If covariance is a large positive number, then as x increases so does y, or as x decreases so does y. If the covariance is a large negative number, then as x increases y decreases or vice versa. If the covariance is small, there is no general pattern between x and y

“large” and “small” are relative terms that are hard to define. To determine how strong the linear relationship is, we can use the **coefficient of correlation.**

Coefficient of Correlation

Population Coefficient of Correlation $\rho = \frac{\sigma_{xy}}{\sigma_x \sigma_y}$

Sample Coefficient of Correlation $r = \frac{s_{xy}}{s_x s_y}$

The coefficient of correlation is essentially the covariance divided by the individual standard deviations of x and y.

In general, the coefficient of correlation must be between -1 and 1.

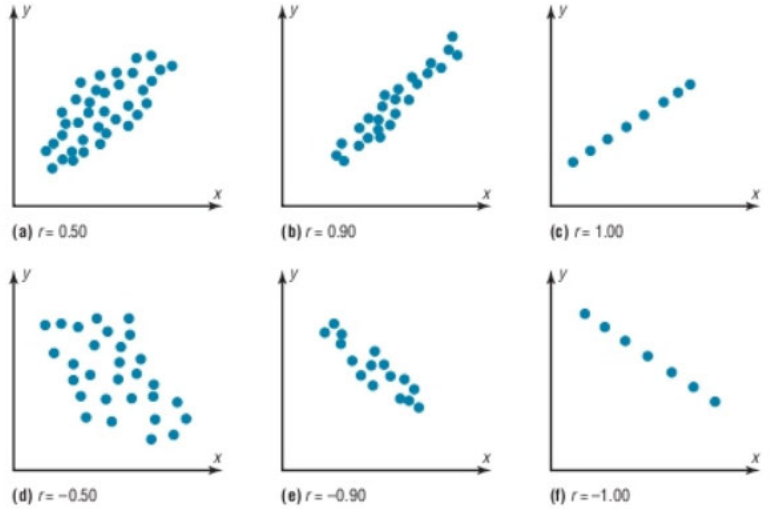
$-1 \leq \rho \leq 1 \qquad -1 \leq r \leq 1$

Interpretation:
0 - No linear relation
0.1 – 0.3 – weak linear relationship
0.4 - 0.6 – moderate linear relationship
0.7 – 0.9 – strong linear relationship
1 – Perfect linear relationship

Direction
+ for positive linear relation (as one variable increases, the other increases)
- for negative linear relation (as one variable increases, the other decreases)

-1.00 → Perfect negative correlation
+1.00 → Perfect positive correlation
0.00 → No correlation, i.e. no linear relationship between the variables

Scatterplots can help visually identify the approximate correlation



(a) $r = 0.50$ (b) $r = 0.90$ (c) $r = 1.00$

(d) $r = -0.50$ (e) $r = -0.90$ (f) $r = -1.00$

Example

A) Determine covariance for the following dataset

B) Determine the coefficient of correlation

Data Set 1	Data Set 2
-2	1
4	3
7	8

Solution:

Data Set 1 (x)	Data Set 2 (y)	x^2	y^2	xy
-2	1	4	1	-2
4	3	16	9	12
7	8	49	64	56
$\sum \boldsymbol{x}$**= 9**	$\sum \boldsymbol{y}$**= 12**	$\sum \boldsymbol{x}^2$ **= 69**	$\sum \boldsymbol{y}^2$ **= 74**	$\sum \boldsymbol{xy}$**= 66**

$\boldsymbol{\overline{x} = 3}$

$\boldsymbol{\overline{y} = 12}$

$\boldsymbol{s_x^2 = 21}$

$\boldsymbol{s_y^2 = 13}$

Sample Covariance Shortcut: $s_{xy} = \frac{1}{n-1}\left[\sum_{i=1}^{n} x_i y_i - \frac{\sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{n}\right]$

Sample Coefficient of Correlation $r = \frac{s_{xy}}{s_x s_y}$

CHAPTER 6

Probability

Probability - The likelihood of an event occurring.

Conditions for Probabilities

1) The probability for any given event must be a number between 0 and 1 inclusive. $0 \leq P(event) \leq 1$
2) The sum of probabilities for all events in the sample space must be 1.

<u>Three Approaches to Probability</u>

Classical Approach – Assume each outcome is equally likely and to determine probability
-i.e. if there are only 2 outcomes possible, each has a probability of 0.5
-i.e. if there are only 5 outcomes possible, each has a probability of 0.2

Relative Frequency Approach – Determine relative frequency over a number of trials. The relative frequency is the probability of an event occurring. For example, if you flip a coin and you get 25 heads in 100 flips, you would say the probability of heads is 0.25.

Subjective Approach – Analyze the situation and approximate what you think the probability should be. I.e. what is the probability of doing well on the stats midterm? Well it would depend on how much you studied, how early you studied, if you attended lectures, etc

<u>Terminology</u>

Intersection ("AND")

An intersection of events A & B is when both events A and B occur simultaneously.

Union ("OR")

Given 2 events A & B, the union of A and B is when one or the other, or both, occur.
Note: P(A or B) = P(B or A) (order is not important)

Joint Probability ("AND") - $\boldsymbol{P(A\ and\ B)}$

Joint probability refers to the probability of an intersection ("and"), that is, the probability of two events occurring at the same time.
Note: P(A and B) = P(B and A) (order is not important)

Marginal Probability $\boldsymbol{P(A)}$

The probability of a specific event occurring, usually calculated by summing the probabilities in a given row or column. Since the number would be calculated in the margin, it is called a 'marginal' probability.

Conditional Probability

The probability of an event happening **given** that another event has occurred.

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

"The probability of event A given that event B has occurred"

Note: the keyword "given" is not always used, so be sure to read the question carefully!

Independent Events

If the probability of one event does not affect the probability of another event occurring, then the events are independent.

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

Mutually Exclusive Events

Two events A & B are mutually exclusive if they cannot occur simultaneously.

$P(A\ and\ B) = 0$ $P(A\ or\ B) = P(A) + P(B)$

Note: Independent and mutually exclusive are completely different concepts!

Complement

Given an event A, the complement is referred to as A^c. The complement can be thought of as 'NOT event A'.

Complement Rule

$P(A^c) = 1 - P(A)$

Multiplication Rule ("AND" , Joint Probability)

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

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

However, if the events are independent:

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

Note: When you have a table given, it is easy to find a joint probability ("and" scenario). Simply find the intersection. You do not have to use the multiplication formula.

Addition Rule ("OR")

$P(A\ or\ B) = P(A) + P(B) - P(A\ and\ B)$

However, if the events are mutually exclusive:

$P(A\ or\ B) = P(A) + P(B)$ Because $P(A\ and\ B) = 0$

Important Note: Sometimes a question uses the word 'proportion' or 'percentage' instead of probability. As far as we're concerned, these words mean the same thing as finding the probability. Probability can be expressed as a decimal number or a percentage.

Example

The following table represents the probability of a randomly selected student passing or failing ADMS 2320 based on if a student reviewed after each lecture or crammed before each test.

	Pass	**Fail**	**Total**
Reviewed After Each Lecture	0.65	0.05	
Crammed Before Tests	0.1	0.2	
Total			

What is the probability that a randomly selected student passes the course?

What is the probability that a randomly selected student reviews after each lecture?

What is the probability that a randomly selected student crammed before a test and passed the course?

What is the probability that a randomly selected student failed the course and reviewed after each lecture?

What is the probability that a randomly selected student failed the course or reviewed after each lecture?

What is the probability that a randomly selected student failed the coursed if they had reviewed after each lecture?

Are the method of studying and the outcome (pass/fail) mutually exclusive events?

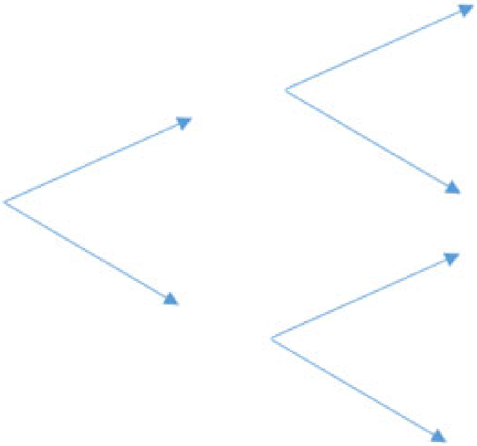
Are the method of studying and the outcome (pass/fail) independent events?

Probability Tree's & Bayes Theorem

When finding a joint probability, **multiply** along a branch to the outcome.
When finding a marginal probability, **add** the joint probabilities together.
When finding a conditional probability, use $P(A|B) = \frac{P(A \text{ and } B)}{P(B)}$

Example

There are 500 students taking ADMS 2320 this semester. 50 Students will go to Gradesavers for tutoring. For the students that go to Gradesavers, there is a 90% chance of passing the course, while for students that don't go to Gradesavers, there is a 50% chance of passing the course.



A) What is the probability of passing the course?

B) What is the probability of failing the course?

C) If a student passed the course, what is the probability they went to Gradesavers?

D) If a student went to Gradesavers, what is the probability that they passed?

Steps for answering chapter 6 questions.

If required, draw a tree diagram to picture the scenario. Fill in as much of the tree as possible. Create a table and fill in the joint and marginal probabilities. Joint probabilities are AND probabilities! From a tree diagram, they are calculated by multiplying along a branch

Always write down what the question is asking for in P notation.

1 event -> Marginal Probability -> Use number from the margin

2 events -> AND, OR, Conditional?

AND -> Use Joint probability (a number in the center of the contingency table)

OR -> Convert using $P(A\ or\ B) = P(A) + P(B) - P(A\ and\ B)$

Conditional -> Convert using $P(A|B) = \frac{P(A\ and\ B)}{P(B)}$

*Some key words for conditional probability: IF, GIVEN

If it's an OR or CONDITIONAL probability, we use formulas to convert to JOINT (AND) and MARGINAL probabilities as these are numbers we can read right off the table.