MATH 145 – Elementary Statistics

(Instructor: Dr. Toribio)

Statistics

- Is the science of learning from data.
- Is a science that deals with the collection, analysis, interpretation, and presentation of data.
- Is a bunch of methods used for the collection, analysis, interpretation, and presentation of data.

Two kinds of Statistics:

1. *Descriptive Statistics* - Methods for organizing and summarizing information.
2. *Inferential Statistics* - Methods for drawing and measuring the reliability of conclusions about a *population* based on information obtained from a *sample* of the population.

- **Population** - A set of units (people, objects, transactions, or events) that we are interested in studying.
  1. Concrete populations
  2. Conceptual or hypothetical populations
- **Sample -** A part of the population from which information is collected.

Some examples of statistical problems:

1. The president of the Student Council of UWL wants to determine the proportion of the student population who are in favor of making the UWL campus a non-smoking zone.
2. A politician wants to know his chance of winning in the coming election.
3. Economists want to estimate the average cost of a 4-year college education in the Midwest.
4. A city engineer wants to estimate the average weekly water consumption for single-family dwellings units in the city.
5. Union leaders want to know if people in this state are earning less or more than those living in other states?
6. An environmentalist group wants to determine the number of deer living in a certain region or the number of fish in a lake.

   ***Capture-Recapture Method*:**

   Step 1. Capture a sample of the population, mark them, and release back to the population.

   Step 2. After a certain period of time, capture another sample from the same population.

   Step 3. Compute the proportion of marked individuals, and use it to estimate the population size.

7. Market researchers want to check if the type of background music in a store affects the purchasing behavior of customers.
8. Pharmaceutical companies want to prove that their drugs are effective and better than existing drugs.
9. Credit card companies want to know the likelihood that a person with certain known characteristics will be able to pay a loan.
10. Medical doctors want to determine if physical fitness level prior to knee surgery has an effect on the length of rehabilitation.

Methods of Acquiring Information:

I. **Public Source**: The data set of interest has already been collected and is available to the public.

1. ***Statistical Abstract of the United States*** – A comprehensive summary of statistics on the social, political, and economic organization of the United States (yearly).
2. ***Survey of Current Business*** – Data on the economy of the United States (monthly).
3. ***The Wall Street Journal*** – Financial data.
4. ***The Sporting News*** – Sports information.
5. ***The Internet***

II. **Census** – Information is obtained from the whole population.

III. **Sampling** – Information is obtained from a small group (***sample***) of objects/individuals taken from the ***population***. The sample should be a ***representative sample***, that is, it should reflect as closely as possible the relevant characteristics of the population under consideration.

Simple Random Sampling – is a sampling procedure for which each possible sample of a given size is equally likely to be the one obtained. A sample obtained in this way is called a ***Simple Random Sample*** (SRS).

- ***Observational Study*** – researchers simply observe characteristics and take measurements, as in a sample survey. Can only establish **association**.
- ***Designed Experiment*** – researchers impose treatments and controls and then observe characteristics and take measurements. Can establish **causal link**.

Other Common Sampling Designs:

1. **Systematic Random Sampling**.
   Step 1. Divide the population size by the sample size and round the result down to the nearest whole number, m.
   Step 2. Use a random-number generator (table, computer or any similar device) to obtain a number k, between 1 and m.
   Step 3. Select for the sample those numbers of the population that are numbered k, k+m, k+2m, ...
2. **Cluster Sampling**.
   Step 1. Divide the population into groups (clusters).
   Step 2. Obtain a simple random sample of the clusters.
   Step 3. Use all the members of the clusters in Step 2 as the sample.
3. **Stratified Random Sampling with Proportional Allocation**.
   Step 1. Divide the population into subpopulations (strata)
   Step 2. From each stratum, obtain a simple random sample of size proportional to the size of the stratum; that is, the sample size for a stratum equals the total sample size times the stratum size divided by the population size.
   Step 3. Use all the members obtained in Step 2 as the sample.

Homework: Read Chapter 1.

Statistics

is the science of collecting, analyzing, interpreting, and presenting data.

Two kinds of Statistics:

1. Descriptive Statistics.
2. Inferential Statistics.

A *statistical inference* is an estimate, prediction, or some other generalization about a *population* based on information contained in the *sample.*

→ Use a *representative* sample.

Sampling Designs

1. Simple Random Sampling.
2. Systematic Random Sampling.
3. Cluster Sampling.
4. Stratified Random Sampling with Proportional Allocation.

Simple Random Sampling

- A sampling procedure for which each possible sample of a given size has the same chance of being selected.
- Population of 5 objects: {A, B, C, D, E}
- Take a sample of size 2.
- Possible samples: {(A,B), (A,C), (A,D), (A,E), (B,C), (B,D), (B,E), (C,D), (C,E), (D,E)}
- Random number generators

Systematic Random Sampling

- Step 1. Divide the population size by the sample size and round the result down to the nearest number, m.
- Step 2. Use a random-number generator to obtain a number k, between 1 and m.
- Step 3. Select for the sample those numbers of the population that are numbered k, k+m, k +2m, …
  - Expected number of customers = 1000
  - Sample size of 30 → m = 1000/30 = 33.33 ≈ 33
  - Suppose k = 5. Then select {5, 5+33, 5+66, …}

Cluster Sampling

- Step 1. Divide the population into groups (clusters).
- Step 2. Obtain a simple random sample of clusters.
- Step 3. Use all the members of the clusters in step 2 as the sample.

Stratified Random Sampling with Proportional Allocation

- Step 1. Divide the population into subpopulations (strata).
- Step 2. From each stratum, obtain a simple random sample of size proportional to the size of the stratum.
- Step 3. Use all the members obtained in Step 2 as the sample.
  - Population of 10,000 with 60% females and 40% males
  - Sample of size 80.

  → 48 females (from 6,000) and 32 males (from 4,000).

Homework: Answer # 1, 2, 5, 7, and 10 on page 18.

Math 145 – Elementary Statistics **Handout #3**

I. Consider the population of students in Dr. Toribio's Elementary Statistic class.

	Name	Year level	Gender	Random #
1	Buck	3	F	
2	Diehl	1	F	
3	Feasel	1	F	
4	Harbaugh	2	F	
5	Howard	2	F	
6	Kaitschuck	4	F	
7	Killmer	1	F	
8	Leung	1	F	
9	Little	3	M	
10	Norbeck	4	F	
11	Stiriz	3	F	
12	Veres	4	F	
13	Verlei	3	M	
14	Vitale	2	F	

a) Assign 3-digit random numbers to each person. *Use the empty column*.

b) Using the simple random sampling, draw a random sample of size 4. Explain how you chose your sample.

Sample: { ___________, ___________, ___________, ___________}

c) Using the systematic random sampling, draw a sample of size 4. Explain how you chose your sample.

Sample: { ___________, ___________, ___________, ___________}

d) Using the 4 year-levels as clusters, use cluster sampling to draw a sample composed of 2 clusters. Explain how you chose your sample.

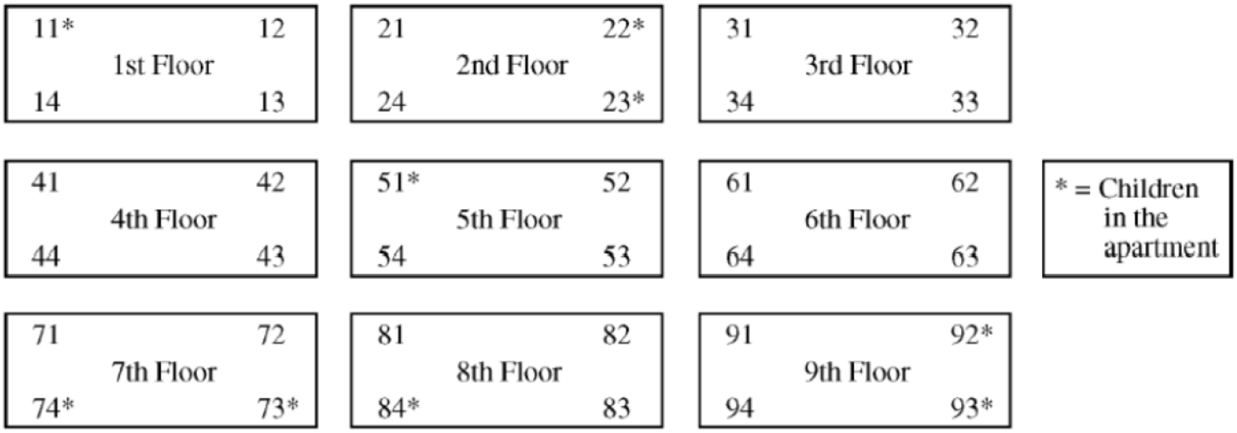
Sample:

e) Using gender as strata, draw a sample of 7 students using the stratified random sampling with proportionate allocation.

Sample: { ___________, ___________, ___________, ___________,
___________, ___________, ___________}

II. An apartment building has nine floors and each floor has four apartments. The building owner wants to install new carpeting in eight apartments to see how well it wears before she decides whether to replace the carpet in the entire building.

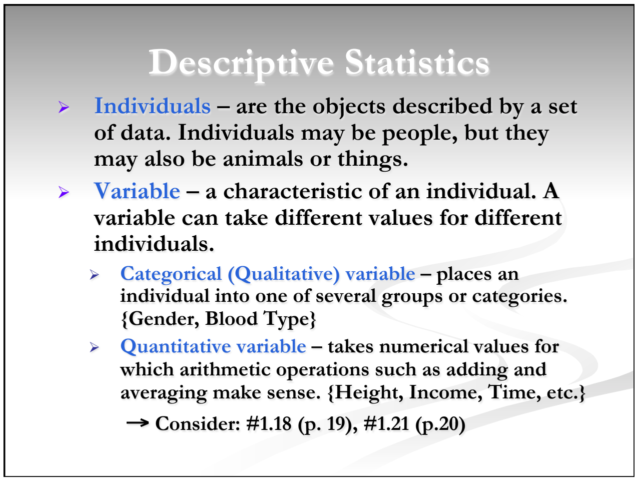
The figure below shows the floors of apartments in the building with their apartment numbers. Only the nine apartments indicated with an asterisk (*) have children in the apartment.



a) Describe a process for randomly selecting eight different apartments using the *simple random sampling* procedure.

b) For convenience, the apartment building owner wants to use the *cluster sampling* procedure, in which the floors are clusters, to select the eight apartments. Describe a process for randomly selecting eight different apartments using this method.

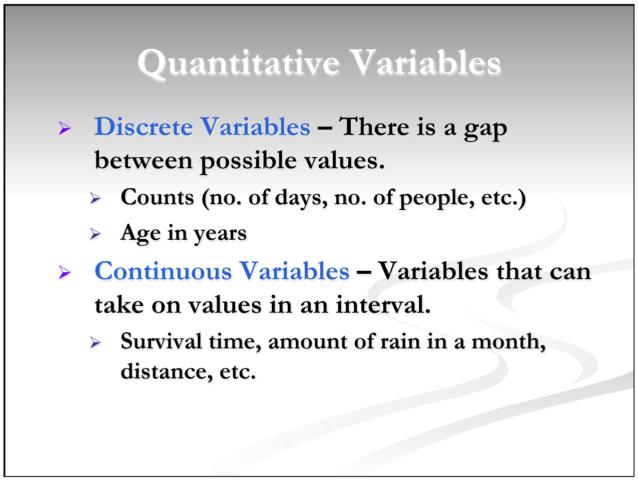
c) An alternative sampling procedure would be to select a *stratified random sample* of eight apartments, where the strata are apartments with children and apartments with no children. Describe a process for randomly selecting eight different apartments using this method. Give one statistical advantage of selecting such a stratified sample as opposed to a cluster sample of eight apartments using the floors as clusters.



Descriptive Statistics

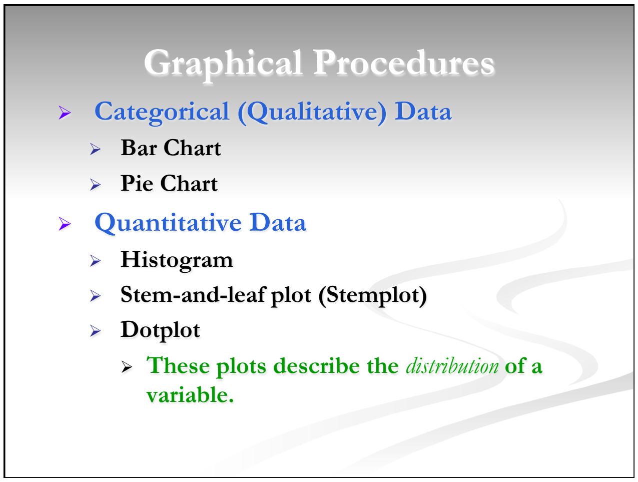
- Individuals – are the objects described by a set of data. Individuals may be people, but they may also be animals or things.
- Variable – a characteristic of an individual. A variable can take different values for different individuals.
  - Categorical (Qualitative) variable – places an individual into one of several groups or categories. {Gender, Blood Type}
  - Quantitative variable – takes numerical values for which arithmetic operations such as adding and averaging make sense. {Height, Income, Time, etc.}

    → Consider: #1.18 (p. 19), #1.21 (p.20)



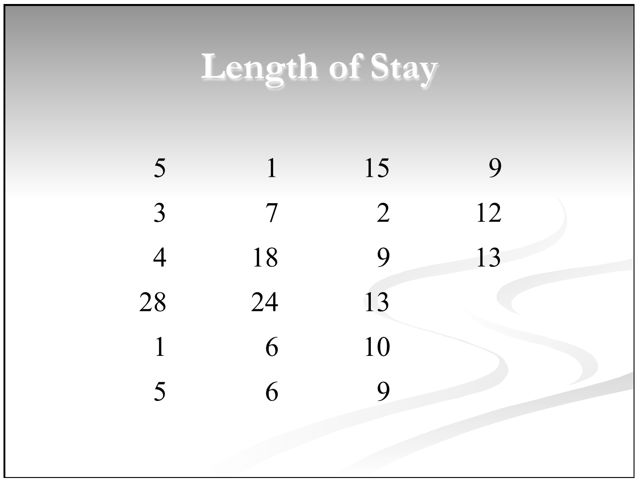
Quantitative Variables

- Discrete Variables – There is a gap between possible values.
  - Counts (no. of days, no. of people, etc.)
  - Age in years
- Continuous Variables – Variables that can take on values in an interval.
  - Survival time, amount of rain in a month, distance, etc.



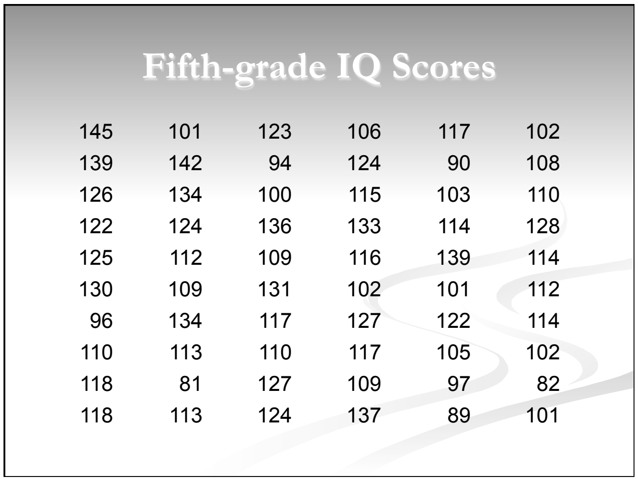
Graphical Procedures

- Categorical (Qualitative) Data
  - Bar Chart
  - Pie Chart
- Quantitative Data
  - Histogram
  - Stem-and-leaf plot (Stemplot)
  - Dotplot
    - These plots describe the *distribution* of a variable.



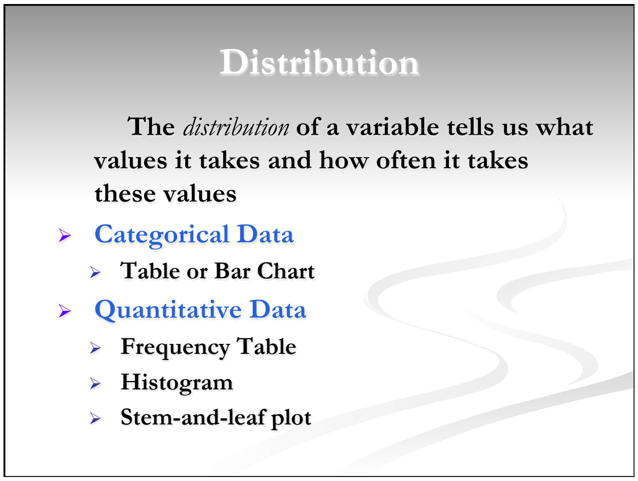
Length of Stay

5	1	15	9
3	7	2	12
4	18	9	13
28	24	13	
1	6	10	
5	6	9	



Fifth-grade IQ Scores

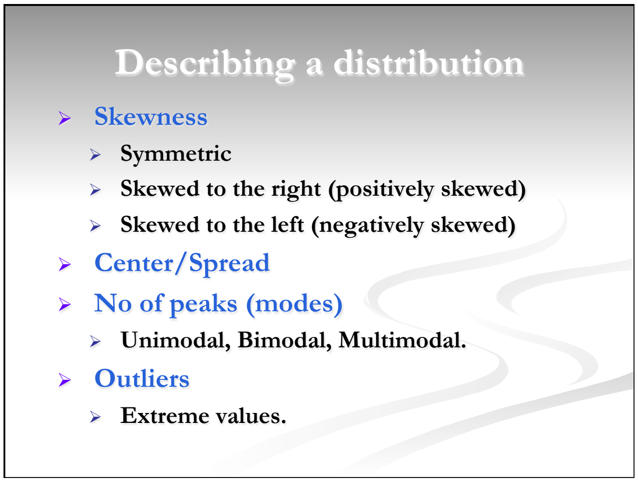
145	101	123	106	117	102
139	142	94	124	90	108
126	134	100	115	103	110
122	124	136	133	114	128
125	112	109	116	139	114
130	109	131	102	101	112
96	134	117	127	122	114
110	113	110	117	105	102
118	81	127	109	97	82
118	113	124	137	89	101



Distribution

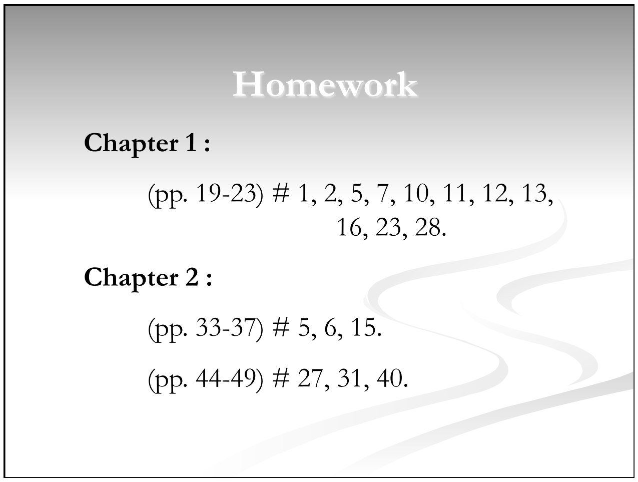
The *distribution* of a variable tells us what values it takes and how often it takes these values

- Categorical Data
  - Table or Bar Chart
- Quantitative Data
  - Frequency Table
  - Histogram
  - Stem-and-leaf plot



Describing a distribution

- Skewness
  - Symmetric
  - Skewed to the right (positively skewed)
  - Skewed to the left (negatively skewed)
- Center/Spread
- No of peaks (modes)
  - Unimodal, Bimodal, Multimodal.
- Outliers
  - Extreme values.



Homework

Chapter 1 :

(pp. 19-23) # 1, 2, 5, 7, 10, 11, 12, 13, 16, 23, 28.

Chapter 2 :

(pp. 33-37) # 5, 6, 15.

(pp. 44-49) # 27, 31, 40.

MATH 145 - ELEMENTARY STATISTICS HANDOUT #5

Graphical Procedures

- The U.S. National Center for Health Statistics compiles data on the length of stay by patients in short-term hospitals and publishes its findings in Vital and Health Statistics. A random sample of 21 patients yielded the following data on length of stays, in days.

5	28	1	24	15	13	9
3	1	7	6	2	10	12
4	5	18	6	9	9	13

1. Construct a frequency and relative frequency tables using the following class intervals: $0 < 5, 5 < 10, \ldots, 25 < 30$.

2. Construct a frequency histogram.

3. Construct a split stem-and-leaf plot(stemplot).

- **Practice**: Consider the IQ test scores for 60 randomly chosen fifth-grade students.

145	139	126	122	125	130	96	110	118	118
101	142	134	124	112	109	134	113	81	113
123	94	100	136	109	131	117	110	127	124
106	124	115	133	116	102	127	117	109	137
117	90	103	114	139	101	122	105	97	89
102	108	110	128	114	112	114	102	82	101

1. Construct a frequency and relative frequency tables using the following class intervals: $80 < 90, 90 < 100, \ldots, 140 < 150$. Then construct a frequency histogram.

2. Construct a stem-and-leaf plot(stemplot).

Measures of Center and Dispersion

- **Measures of Center**

1. Mean ($\mu, \bar{x}$) - average (equal to the sum of the values divided by the number of values).

a. Population mean, $\mu = \dfrac{1}{N}\displaystyle\sum_{i=1}^{N} X_i$, where N is the population size.

b. Sample mean, $\bar{X} = \dfrac{1}{n}\displaystyle\sum_{i=1}^{n} X_i$, where n is the sample size.

Example 1: Consider a random sample of 12 monthly salaries (in thousands of dollars):

$$\{2.5, 3.2, 3.2, 3.4, 3.5, 3.6, 3.9, 3.9, 3.9, 4, 4.2, 4.2\}$$

Find the sample mean.

Practice 1: Consider the random sample of 21 length of stay in a hospital (in days):

$$\{1, 1, 2, 3, 4, 5, 5, 6, 6, 7, 9, 9, 9, 10, 12, 13, 13, 15, 18, 24, 28\}$$

Determine the sample mean.

Consider the problem in Example 1. Suppose a 13th person was sampled and the person earns $50K per month (or $600K per year). Compute the sample mean. Is the new sample mean a good representative of the sample?

Remark 1: Means are *sensitive* to outliers.

2. Median ($\tilde{\mu}, \tilde{x}$) - middle value (when values are arrangement from lowest to highest). If there are an odd number of values, there will be one value right in the middle. If there are an even number of values, then the median is the average of the middle pair.

Example 2: Using the random sample of 12 monthly salaries, find the sample median.

Practice 2: Using the random sample of 21 lengths of stay, find the sample median.

Consider the problem in Example 2. Again, suppose a 13th person was sampled and the person earns $50K per month (or $600K per year). Compute the sample median. What can you say about the magnitude of the effect of the outlying value to the sample median?

Remark 2: Medians are *robust* against outliers.

3. Mode - most frequent occurring value.

Example 3: Using the random sample of 12 monthly salaries, determine the mode.

Practice 3: Using the random sample of 21 lengths of stay, determine the sample mode.

Example 4: Compute the mean, median, and mode for the following data sets:

a. $S_1 = \{1, 2, 5, 5, 8, 9\}$
b. $S_2 = \{3, 4, 5, 5, 6, 7\}$
c. $S_3 = \{5, 5, 5, 5, 5, 5\}$

Remark 3: Locating the center of the data is not enough. We also need to measure the spread of the values.

- **Measures of Dispersion (Spread)**

1. Range = Maximum value - Minimum value

2. Variance

a. Population variance, $\sigma^2 = \dfrac{1}{N}\displaystyle\sum_{i=1}^{N}(x_i - \mu)^2$

b. Sample variance, $S^2 = \dfrac{1}{n-1}\displaystyle\sum_{i=1}^{n}(x_i - \bar{x})^2$

Computing formula: $S^2 = \dfrac{1}{n-1}\left(\displaystyle\sum_{i=1}^{n} x_i^2 - \frac{1}{n}(\sum_{i=1}^{n} x_i)^2\right)$

3. Standard Deviation $= \sqrt{\text{Variance}}$

Homework.

Section 2.4: (pp. 56 - 60) # 2.50, 2.53, 2.55, 2.57, 2.58(a), 2.66, 2.70.
Section 2.5: (pp. 64 - 66) # 2.74, 2.75, 2.76, 2.77, 2.85.

Empirical Rule

- The **Empirical Rule** is a rule of thumb that applies to data sets with frequency distributions that are mound shaped and symmetric.

  1. Approximately 68% of the measurements will fall within 1 standard deviation of the mean, i.e., within the interval $(\bar{x} - s, \bar{x} + s)$ for samples and $(\mu - \sigma, \mu + \sigma)$ for populations.

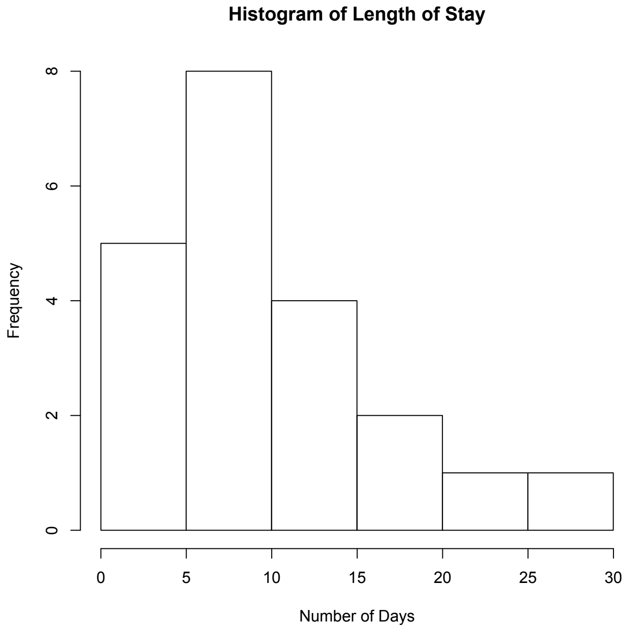
  2. Approximately 95% of the measurements will fall within 2 standard deviations of the mean, i.e., within the interval $(\bar{x} - 2s, \bar{x} + 2s)$ for samples and $(\mu - 2\sigma, \mu + 2\sigma)$ for populations.

  3. Approximately 99.7% of the measurements will fall within 3 standard deviation of the mean, i.e., within the interval $(\bar{x} - 3s, \bar{x} + 3s)$ for samples and $(\mu - 3\sigma, \mu + 3\sigma)$ for populations.

- Consider the random sample of 21 patients that yielded the following data on length of stays, in days. Find $\bar{x}$, s^2, and s. Note that $\sum x_i = 200$ and $\sum x_i^2 = 2936$.

5	28	1	24	15	13	9
3	1	7	6	2	10	12
4	5	18	6	9	9	13

$$Sorted: \{1, 1, 2, 3, 4, 5, 5, 6, 6, 7, 9, 9, 9, 10, 12, 13, 13, 15, 18, 24, 28\}$$



1. Determine the proportion of values between $(\bar{x} - s, \bar{x} + s)$.

2. Determine the proportion of values between $(\bar{x} - 2s, \bar{x} + 2s)$.

3. Determine the proportion of values between $(\bar{x} - 3s, \bar{x} + 3s)$.

Z-scores

- **Definition of $Z-$scores**

  1. The *sample z-score* for a measurement x is $z = \dfrac{x-\bar{x}}{s}$.

  2. The *population z-score* for a measurement x is $z = \dfrac{x-\mu}{\sigma}$.

- **Application of $Z-$scores.** Suppose we look at the December temperature information at the following 3 cities: [Note: $C = \frac{5}{9}(F-32)$]

	μ	σ	x
Juneau, Alaska	$10\ ^oF$	$3\ ^oF$	$7\ ^oF$
La Crosse, Wisconsin	$25\ ^oF$	$5\ ^oF$	$18\ ^oF$
Manila, Philippines	$27\ ^oC$	$2\ ^oC$	$24\ ^oC$

  1. If x represents the temperature on a certain day in December, compute the $z-$score for the temperature in La Crosse, WI. Interpret the value of the $z-$score.

  2. Which of the three cities experienced the most extreme temperature relative to their normal temperature?

- **Recentering and rescaling of data**

  Consider the previous random sample of 21 length of stays of patients.

  1. If 5 is subtracted from each observation, find

  a. the sample mean of the new set of values.

  b. the standard deviation (s) of the new set of values.

  Rule 1: If $Y = X + a$, then $\mu_Y = \mu_X + a$ or $\bar{Y} = \bar{X} + a$, and $\sigma_Y = \sigma_X$ or $S_Y = S_X$

  2. If each observation is doubled, find

  a. the sample mean of the new set of values.

  b. the standard deviation (s) of the new set of values.

  Rule 2: If $Y = b * X$, then $\mu_Y = b * \mu_X$ or $\bar{Y} = b * \bar{X}$, and $\sigma_Y = b * \sigma_X$ or $S_Y = b * S_X$

  General Rule: If $Y = bX + a$, then

  a. $\mu_Y = b * \mu_X + a$ or $\bar{Y} = b * \bar{X} + a$

  b. $\sigma_Y = b * \sigma_X$ or $S_Y = b * S_X$

- **Mean and Standard Deviation of $Z-$scores**

  Let $z_i = \dfrac{x_i - \bar{x}}{s_x}$, find $\bar{z}$ and s_z.

Homework.

Section 2.6: (pp. 71 - 73) # 2.94 (a) and (b), 2.95.
Section 2.7: (pp. 76 - 77) # 2.111, 2.113, 2.117, 2.123.

Rules for Means and Standard Deviations

- If $Y = bX + a$, then

  1. $\mu_Y = b\mu_X + a$ or $\bar{Y} = b\bar{X} + a$

  2. $\sigma_Y = b\sigma_X$ or $S_Y = bS_X$

- Example 1: The mean daily high temperature in December in Melbourne, Australia is 23.5^oC with a standard deviation of 0.25^oC. Determine the mean and standard deviation of the daily high temperatures in December in Melbourne, Australia in ^oF. [note: $F = \frac{9}{5}C + 32$] [2]

- Example 2: The mean salary of workers in a large retail store is \$8.25 per hour with a standard deviation of \$1.75. As an incentive for outstanding performance, the retail store decided to give a bonus of \$10 per day (8-hour work day) for the employee of the month. That is, if a worker's regular pay is x dollars per hour, then when this person is the employee of the month, he/she will receive a total of $T = 8x + 10$ dollars per day. Determine the mean and standard deviation of T.

- Example 3: Linda is a sales associate at a large auto dealership. At her commission rate of 25% of profit on each vehicle she sells, Linda expects to earn \$350 for each car sold. On average, she is able to sell one car on a typical Saturday. If she has a fixed salary of \$50 per day, compute Linda's average income on a typical Saturday.

- Example 4: Based on past data, an insurance company pays on average \$500 per customer per year on car accident claims, with a standard deviation of \$400. If the company spends about 30 million dollars for overhead expenses, determine the mean of the total annual cost for this company to insure 600,000 customers.

Boxplot

- Review: Consider the random sample of 21 patients that yielded the following data on length of stays, in days.

5	28	1	24	15	13	9
3	1	7	6	2	10	12
4	5	18	6	9	9	13

{1,1,2,3,4,5,5,6,6,7,9,9,9,10,12,13,13,15,18,24,28}

- **Five-number summary**

  1. Minimum - $X_{(1)}$

  2. First Quartile - Q_1

  3. Median - $\tilde{X}$

  4. Third Quartile - Q_3

  5. Maximum - $X_{(n)}$

- **Practice.**The following data set records 40 yearly charitable contributions (in dollars) to the United Fund for a group of employees at a public university. *Note that the data are arranged in increasing order.*

24	28	41	??	47	51	56	59	63	63
63	65	67	69	70	71	75	77	78	79
79	80	80	80	80	81	81	??	83	84
90	92	93	99	101	103	103	112	115	125

Obtain the 5-no. summary and construct the (modified) boxplot for this data set.

Homework.

1. Section 2.8: (pp. 85 - 87) # 2.126, 2.131, 2.132.

Review - Basic Set Theory

Consider the universal set $U = \{1, 2, 3, 4, 5, 6, 7, 8, 9, 10\}$.

- **Subset.** A set A is said to be a *subset* of U, if all elements (or entries) of A are in U. In this case, we write $A \subseteq U$.

- **Complement.** The *complement* of set A, written as A^c, is the set containing all elements of U that are <u>not</u> in A.

- **Union.** The *union* of A and B, written as $A \cup B$, is the set of elements that belong to either A or B or both. That is,

$$A \cup B = \{x | x \text{ is in } A \text{ or } x \text{ is in } B\}.$$

- **Intersection.** The *intersection* of A and B, written as $A \cap B$, is the set of elements that belong to both A and B. That is,

$$A \cap B = \{x | x \text{ is in } A \text{ and } x \text{ is in } B\}.$$

- **Properties:**

1. Commutativity

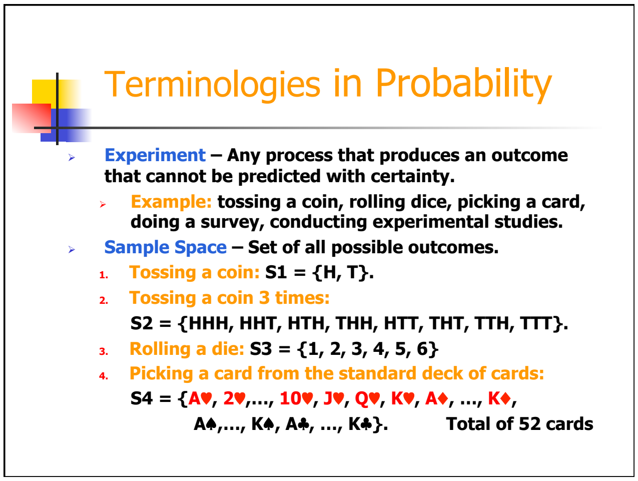
a. $A \cup B = B \cup A$

b. $A \cap B = B \cap A$

2. De Morgan's Laws

a. $(A \cup B)^c = A^c \cap B^c$

b. $(A \cap B)^c = A^c \cup B^c$

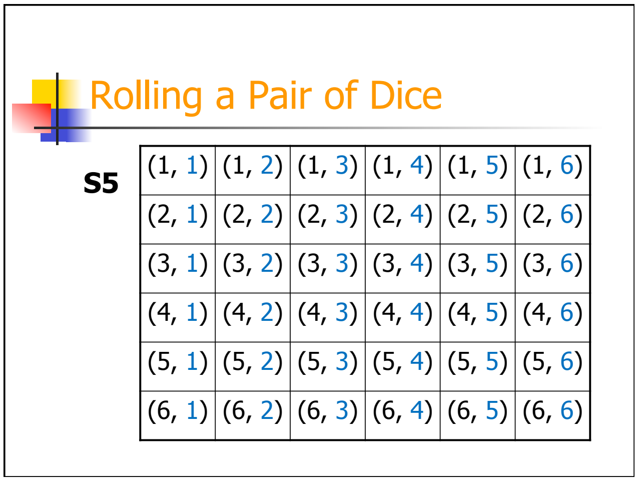


Terminologies in Probability

- **Experiment – Any process that produces an outcome that cannot be predicted with certainty.**
  - **Example: tossing a coin, rolling dice, picking a card, doing a survey, conducting experimental studies.**
- **Sample Space – Set of all possible outcomes.**
  1. **Tossing a coin: S1 = {H, T}.**
  2. **Tossing a coin 3 times:**

     S2 = {HHH, HHT, HTH, THH, HTT, THT, TTH, TTT}.
  3. **Rolling a die: S3 = {1, 2, 3, 4, 5, 6}**
  4. **Picking a card from the standard deck of cards:**

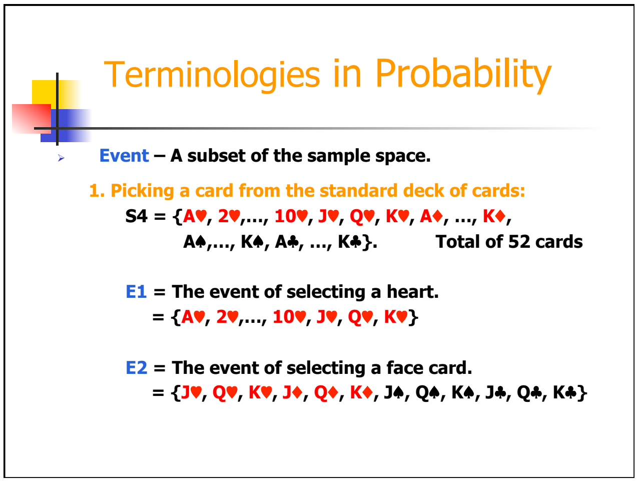
     S4 = {A♥, 2♥,…, 10♥, J♥, Q♥, K♥, A♦, …, K♦, A♠,…, K♠, A♣, …, K♣}. Total of 52 cards



Rolling a Pair of Dice

S5

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



Terminologies in Probability

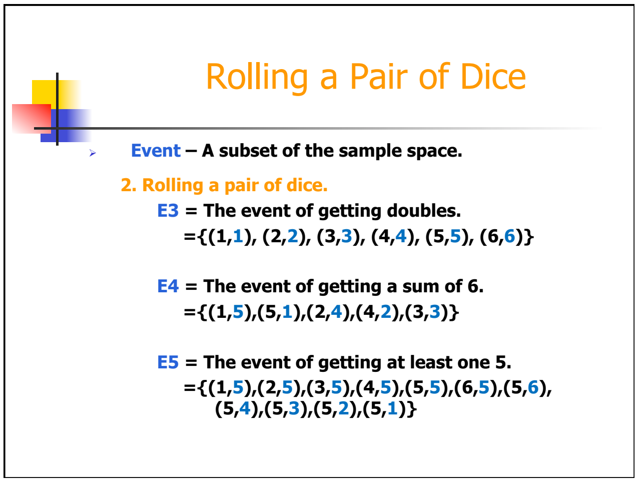
- **Event – A subset of the sample space.**

1. Picking a card from the standard deck of cards:

S4 = {A♥, 2♥,…, 10♥, J♥, Q♥, K♥, A♦, …, K♦, A♠,…, K♠, A♣, …, K♣}. Total of 52 cards

E1 = The event of selecting a heart.
= {A♥, 2♥,…, 10♥, J♥, Q♥, K♥}

E2 = The event of selecting a face card.
= {J♥, Q♥, K♥, J♦, Q♦, K♦, J♠, Q♠, K♠, J♣, Q♣, K♣}



Rolling a Pair of Dice

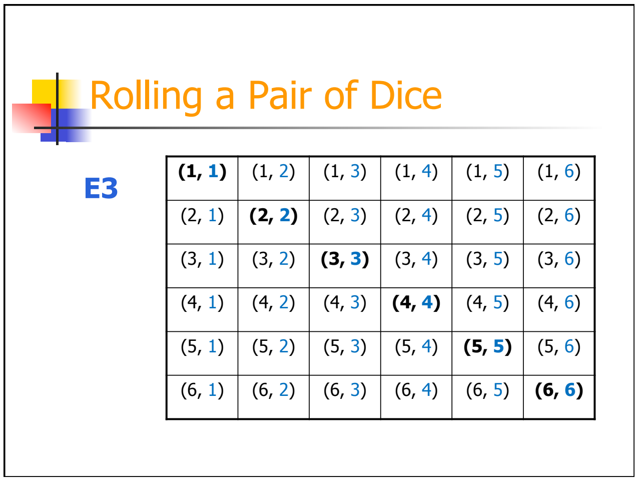
- **Event – A subset of the sample space.**

2. Rolling a pair of dice.

E3 = The event of getting doubles.
={(1,1), (2,2), (3,3), (4,4), (5,5), (6,6)}

E4 = The event of getting a sum of 6.
={(1,5),(5,1),(2,4),(4,2),(3,3)}

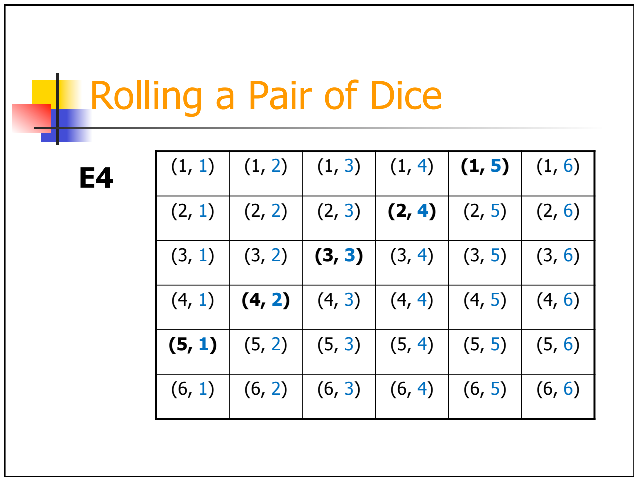
E5 = The event of getting at least one 5.
={(1,5),(2,5),(3,5),(4,5),(5,5),(6,5),(5,6), (5,4),(5,3),(5,2),(5,1)}



Rolling a Pair of Dice

E3

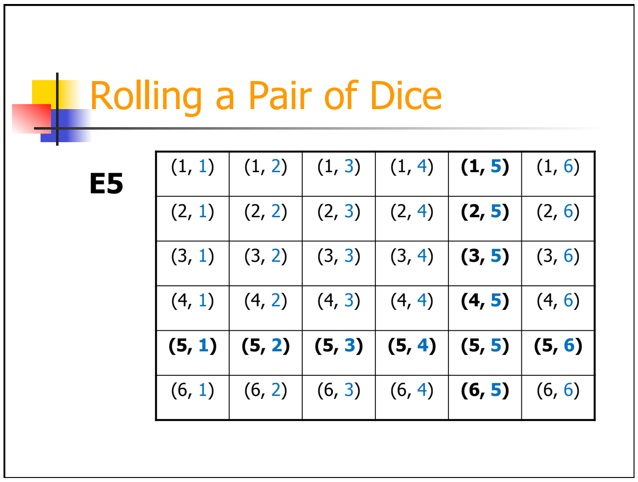
(1, 1)	(1, 2)	(1, 3)	(1, 4)	(1, 5)	(1, 6)
(2, 1)	**(2, 2)**	(2, 3)	(2, 4)	(2, 5)	(2, 6)
(3, 1)	(3, 2)	**(3, 3)**	(3, 4)	(3, 5)	(3, 6)
(4, 1)	(4, 2)	(4, 3)	**(4, 4)**	(4, 5)	(4, 6)
(5, 1)	(5, 2)	(5, 3)	(5, 4)	**(5, 5)**	(5, 6)
(6, 1)	(6, 2)	(6, 3)	(6, 4)	(6, 5)	**(6, 6)**



Rolling a Pair of Dice

E4

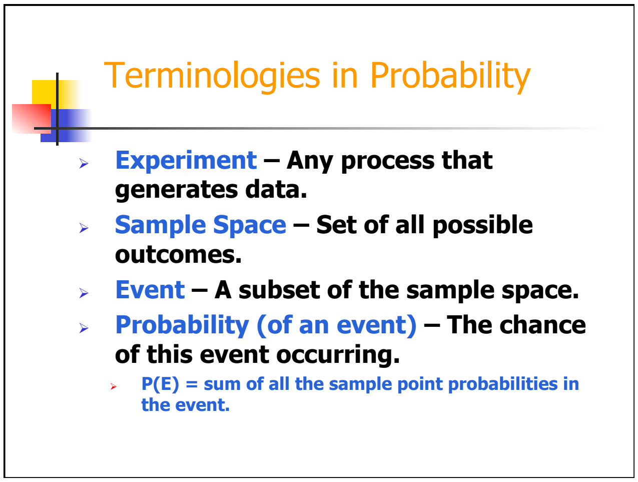
(1, 1)	(1, 2)	(1, 3)	(1, 4)	**(1, 5)**	(1, 6)
(2, 1)	(2, 2)	(2, 3)	**(2, 4)**	(2, 5)	(2, 6)
(3, 1)	(3, 2)	**(3, 3)**	(3, 4)	(3, 5)	(3, 6)
(4, 1)	**(4, 2)**	(4, 3)	(4, 4)	(4, 5)	(4, 6)
(5, 1)	(5, 2)	(5, 3)	(5, 4)	(5, 5)	(5, 6)
(6, 1)	(6, 2)	(6, 3)	(6, 4)	(6, 5)	(6, 6)



Rolling a Pair of Dice

E5

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



Terminologies in Probability

- **Experiment – Any process that generates data.**
- **Sample Space – Set of all possible outcomes.**
- **Event – A subset of the sample space.**
- **Probability (of an event) – The chance of this event occurring.**
  - **P(E) = sum of all the sample point probabilities in the event.**

Chapter 3 - Probability

1. Definitions

- *Experiment* - Any process of observation that leads to a single outcome that cannot be predicted with certainty.
- *Sample Space* - Set of all possible outcomes.
- *Event* - A subset of the sample space.
- *Probability (of an event)* - The chance of this event occurring. It is equal to the sum of sample point probabilities in the event. Under the *Equally Likely* model (also known as the *Uniform* model), it reduces to

$$P(E) = \frac{\text{no. of favorable outcomes}}{\text{no. of possible outcomes}}$$

2. Properties

a. $0 \leq P(E) \leq 1$; $P(S) = 1$; and $P(\phi) = 0$

b. *Probability of the Complement*: $P(E^c) = 1 - P(E)$

c. *Probability of the Union*:

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

If A and B are *mutually exclusive*(*disjoint*) events (they don't intersect), then

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

3. Practice:

a. Suppose all 30 students in a class were asked for their preferences for fast food. The results obtained is given below

Gender	McDonalds	Taco Bell	Subway	Total
Male	2	6	10	
Female	3	2	7	
Total				

If a student is selected at random, what is the probability that

i. the student prefers Taco Bell?
ii. the student is a female?
iii. the student does not prefer Subway?
iv. the student prefer either McDonalds or Subway?
v. the student is female who prefers Subway?

b. In a small town of 2000 people, there are 800 males, 700 of whom are employed. If a total 250 people are unemployed in this town, find the probability that a randomly selected person is

i. a male?
ii. an unemployed?
iii. a male and unemployed (an unemployed male)?
iv. male or unemployed?
v. female and employed (an employed female)?
vi. female or employed?

4. Homework.

Sec 3.4; (pp. 131-134) 42, 45, 47, 51, 53, 55, 57, 61.

Counting Techniques

- **Equally Likely Model.** If an experiment has n different possible outcomes each of which has the same chance of occurring, then the probability that a specific outcome will occur is $1/n$. If an event occurs in k out of these n outcomes, then the probability of the event A, $P(A)$, is

$$P(A) = \frac{\text{no. of favorable outcomes}}{\text{no. of possible outcomes}} = \frac{k}{n}$$

  Examples:

  1. If a card is drawn from a well-shuffled standard deck of cards, what is the probability of getting

  a. a heart?

  b. a face card?

  c. a heart or a face card?

  2. If two cards are drawn, what is the probability of getting a pair (two of a kind)?

  3. If five cards are drawn, what is the probability of getting a flush (all cards of the same suit)?

- **Fundamental Principle of Counting (Product Rule).** Suppose two tasks are to be performed in succession. If the first task can be done in exactly n_1 different ways, and the second task can be done independently in n_2 different ways, then the sequence of things can be done in $n_1 \times n_2$ different ways.

  Examples:

  1. If you roll a six-sided die and then pick a card from the standard deck of cards, how many different possible outcomes are there?

  2. In a certain town there are 3 male and 2 female candidates for mayor and 4 male and 2 female candidates for vice-mayor. If each candidate has the same chance of winning the election, what is the probability that after the election they will have a female mayor and a male vice-mayor?

- **General Fundamental Principle of Counting.** Suppose k tasks are to be performed in succession. If the first task can be done in exactly n_1 different ways, the second task can be done independently in n_2 different ways, the third task can be done independently in n_3 different ways, and so forth, then this sequence of k tasks can be done in $n_1 n_2 n_3 \cdots n_k$ different ways.

  Examples:

  1. In a statistical study, an individual is classified according to gender, income bracket (upper, middle, or lower class) and highest level of educational attained (elementary, high school, or college). Find the number of ways in which an individual can be classified.

  2. Eight elementary students are to be lined up to board a bus. How many different possible arrangements are there?

  3. In the Mathematics department of UW-L, there are 3 male and 2 female professors, 7 male and 2 female associate professors, and 4 male and 2 female assistant professors. A committee consisting of a professor, an associate professor, and an assistant professor is to be set up to review the current math curriculum of the department. Assuming that each faculty member has the same chance of being selected for the committee work, what is the probability that the committee will be composed of all female teachers?

4. Practice: Do #3.132 (page 163)

5. In a local swimming competition, there are 20 contestants. The first, second, and third placer will be awarded with gold, silver, and bronze medals, respectively. How many different possible competition results are there?

6. Three girls and nine boys, including Mark and Jim, are to be lined up to get in a bus.
   - **a.** In how many ways can this be done?
   - **b.** In how many ways can this be done if all the girls insist to be together?
   - **c.** In how many ways can this be done if Mark and Jim refused to be together?

- **Combination.** Given a set of n distinct objects, any unordered subset of size k of the objects is called a *combination.* The number of combinations of size k that can be formed from n distinct objects will be denoted by $\binom{n}{k}$.

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

Examples:

1. Compute $\binom{5}{2}$, $\binom{5}{3}$, $\binom{5}{5}$.

2. If you draw 5 cards from a standard deck of cards, how many different possible outcomes are there?

3. If 5 cards are drawn from a standard deck of cards, in how many ways can you get
   - **a.** a flush (all of the same suit).
   - **b.** a full house (a trio and a pair).

4. A jar contains 6 white and 4 red marbles. If 3 marbles are randomly selected, what is the probability of selecting
   - **a.** two white and one red marbles?
   - **b.** all white marbles?
   - **c.** all of the same colors?

5. There are 14 male and 6 female professors in the Mathematics department at UW-L. A committee of 5 members is to be formed to review the current math curriculum of the department. If each professor has equal chance of being selected for this committee, what is the probability that the committee will have a female majority?

- **Homework.**

Sec 3.1: (pp. 119-123) 9, 10, 13, 15, 17, 21, 27, 33.
Sec 3.8; (pp. 161-164) 119, 121, 123, 125, 129, 135.

Conditional Probability

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

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

Example 1. In a small town of 2000 people, there are 800 males, 700 of whom are employed. If a total 250 people are unemployed in this town, find the probability that a randomly selected person is

1. a male?

2. employed?

3. an employed male?

4. employed given he is a male?

5. male given the person is employed?

Example 2. Suppose that of all individuals buying a certain digital camera, 60% include an optional memory card in their purchase, 40% include an extra battery, and 30% include both a card and battery.

1. Given that the selected individual purchased an extra battery, what is the probability that an optional memory card was also purchased?

2. What is the probability that an extra battery is included in the purchase given that the person got the optional memory card?

- **Independence of Events.** Events A and B are said to be *independent* if one of the following is true:

1. $P(A|B) = P(A)$

2. $P(B|A) = P(B)$

3. $P(A \cap B) = P(A) * P(B)$

- **The Multiplication Rule.**

$$P(A \cap B) = P(B) \cdot P(A|B) \quad \text{or} \quad P(A \cap B) = P(A) \cdot P(B|A)$$

Example 3. A chain of video stores sells three different brands of DVDs. Of its DVDs sales, 50% are brand 1 (the least expensive), 30% are brand 2, and 20% are brand 3. Each manufacturer offers a 1-year warranty on parts and labor. It is known that 25% of brand 1's DVDs require warranty repair work, whereas the corresponding percentages for brands 2 and 3 are 20% and 10%, respectively.

1. What is the probability that a randomly selected purchaser bought a brand 1 DVD that will need repair while under warranty?

2. What is the probability that a randomly selected purchaser bought a DVD that will need repair while under warranty?

3. If a customer returns to the store with a DVD that needs warranty repair work, what is the probability that it is a brand 1 DVD?

Example 4. Online chat rooms are dominated by the young. Teens are the biggest users. If we look only at adult internet users (age 18 and over), 47% of the 18 to 29 age group chat, as do 21% of those aged 30 to 49 and just 7% of those 50 and over. It is known that 29% of adult internet users are age 18 to 29, another 47% are 30 to 49 years old, and the remaining 24% are age 50 and over. If an adult internet user is randomly selected, what is the probability that

1. the person is at least 50 years old?

2. the person chats online given he/she is at least 50 years old?

3. the person is at least 50 years old and chats online?

4. the person chats online?

5. the person is at least 50 years old given that the person chats online?

- **Total Probability.** Let $A_1, \ldots, A_k$ be mutually exclusive and complementary events (That is, $A_1, \ldots, A_k$ form a partition of the sample space). Then for any other event B,

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

- **Bayes' Rule.** Let $A_1, \ldots, A_k$ be a collection of k mutually exclusive and complementary events with $P(A_i) > 0$ for $i = 1, \ldots, k$. Then for any other event B for which $P(B) > 0$,

$$P(A_j|B) = \frac{P(A_j \cap B)}{P(B)} = \frac{P(B|A_j)P(A_j)}{\sum_{i=1}^{k} P(B|A_i)P(A_i)}$$

Practice.

1. A company uses three different assembly lines $-A_1, A_2, A_3-$ to manufacture a particular component. Of those manufactured by A_1, 5% need rework to remedy a defect, whereas 8% of A_2's components need rework and 10% of A_3's need rework. Suppose that 50% of all components are produced by line A_1, 30% are produced by line A_2, and 20% come from line A_3. If a component is randomly selected, what is the probability that

a. it needs rework?

b. it came from line A_1 given that it requires rework?

c. it came from line A_2 given that it requires rework?

2. Do #3.142 on page 166.

3. Do #3.148 on page 167.

- **Homework problems:**
Sec 3.5/3.6; (pp. 145-150) # 67, 69, 71, 77, 84, 85, 95.
Sec 3.9; (pp. 166-168) # 141, 145, 147

Math 145 - Elementary Statistics — Exam I Review

- **Chapter 1:** Terminologies

    1. Two types of statistical methods - Descriptive and Inferential Statistics.

    2. Two types of data - Quantitative and Categorical (Qualitative).

    3. Two types of quantitative variables - Continuous and Discrete.

    4. Sampling Designs

    - **a.** Simple Random Sampling (SRS)
    - **b.** Systematic Sampling
    - **c.** Cluster Sampling
    - **d.** Stratified Sampling (with proportional allocation)

- **Chapter 2:** Descriptive Statistics

    1. Graphical Procedures - Bar graph, Pie chart, Histograms, Stem-and-leaf plot, Dotplot, Boxplot.

    2. Frequency and Relative Frequency Tables.

    3. Numerical Procedures

    - **a.** Measures of Center - Mean, Median, Mode.
    - **b.** Measures of Spread - Range, Mean Deviation, Variance, Standard Deviation.
    - **c.** Five-number summary.

    4. Empirical Rule (for mound-shaped and symmetric distributions: 68%, 95%, 99.7%).

    5. $z-$score $= \frac{x-\mu_x}{\sigma_x}$. [Note: $\mu_z = 0$ and $\sigma_z = 1$].

    6. Rules for means and standard deviations: If $Y = b * X + a$, then

    - **a.** $\mu_Y = b * \mu_X + a$ or $\bar{Y} = b * \bar{X} + a$
    - **b.** $\sigma_Y = b * \sigma_X$ or $S_Y = b * S_X$

- **Chapter 3:** Probability

    1. *Experiment* - Any process of observation that leads to a single outcome that cannot be predicted with certainty.

    2. *Sample Space* - Set of all possible outcomes of an experiment.

    3. *Event* - A subset of the sample space.

    4. *Probability (of an event)* - The chance of this event occurring. It is equal to the sum of sample point probabilities in the event. Under the *Equally Likely* model (also known as the *Uniform* model), it reduces to

    $$P(E) = \frac{\text{no. of favorable outcomes}}{\text{no. of possible outcomes}}$$

- **Properties of probabilities:**

    1. $0 \leq P(E) \leq 1$; $P(S) = 1$; and $P(\phi) = 0$.

    2. *Probability of the Complement*: $P(E^c) = 1 - P(E) \quad \Rightarrow P(E) = 1 - P(E^c)$.

    3. *Probability of the Union*: $P(A \text{ or } B) = P(A \cup B) = P(A) + P(B) - P(A \cap B)$.

- Events A and B are *mutually exclusive* if they do not intersect. That is, $P(A \cap B) = 0$.

- **Counting Techniques:** Multiplication Principle, Permutation, and Combination.

- **Conditional Probability:** When $P(B) > 0$, the conditional probability of A given B is

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

- **Independence:** Events A and B are said to be *independent* if one of the following is true:

    $$(i)\ P(A|B) = P(A), \quad (ii)\ P(B|A) = P(B), \quad (iii)\ P(A \cap B) = P(A)P(B).$$

- **Total Probability:** Using Tree Diagrams.

- **Bayes Rule:** Using the Tree Diagram or the conditional probability formula.

Additional Exercises on Probability

1. If 4 girls and 6 boys are to be seated in a row, what is the probability that

a. all the girls will be sitting together?

b. all the boys will be sitting together?

c. a particular girl (Rosy) and a particular boy (Rex) will not be together?

2. In the previous problem, suppose you want to form a dance group of five. What is the probability that the dance group will be composed of

a. all boys?

b. 3 boys and 2 girls?

c. all girls?

d. at least 1 girl?

3. Suppose that in a sample of 200 students, 120 are taking an English course, 110 are take a Mathematics course, and 60 are taking both Math and English. If a student is randomly selected from this sample, what is the probability that

a. the student in not taking any Math course?

b. the student is taking English but not Math?

c. the student is taking Math but not English?

d. the student is taking at least one of these two courses?

e. the student is NOT taking any of these two courses?

f. the student is taking English given that he/she is taking Math?

g. the student is taking Math given that he/she is taking English?

4. The probability the a certain bank will setup a new branch in La Crosse is .80. The probability that this bank will setup a new branch in Onalaska is .70. The probability that the bank will setup a new branch in both cities is .60. What is the probability that this bank

a. will NOT setup a new branch here in La Crosse?

b. will setup a new branch in at least one of these two cities?

c. will NOT setup a new branch in either cities?

d. will setup a new branch in Onalaska given
that it already setup a new branch in La Crosse?

e. will setup a new branch in La Crosse given
that it already setup a new branch in Onalaska?

5. Police plans to enforce speed limits by using radar traps at 4 different locations within the city limits. The radar traps at each of the locations L_1, L_2, L_3, and L_4 are operated 30%, 40%, 50%, and 20% of the time, respectively. A person who is speeding on his way to work has probabilities of 0.2, 0.1, 0.5, and 0.2, respectively, of passing through these locations. Construct the appropriate tree diagram to answer the questions below. What is the probability that

a. he will pass through L_2?

b. he will receive a speeding ticket given that he passed through L_2?

c. he will pass through L_2 and will receive a speeding ticket?

d. he will receive a speeding ticket?

e. he passed through L_2 given that he received a speeding ticket?

Random Variable

– A random variable is a variable whose value is a numerical outcome of a random phenomenon.

– A random variable is a function or a rule that assigns a numerical value to each possible outcome of a statistical experiment.

Two Types:

1. **Discrete Random Variable** – A *discrete random variable* has a countable number of possible values (There is a gap between possible values).
2. **Continuous Random Variable** – A *continuous random variable* takes all values in an interval of numbers.

Examples

Tossing a coin 3 times:

Sample Space

= {HHH, HHT, HTH, THH, HTT, THT, TTH, TTT}.

Random variables :

X_1 = The number of heads.

= {3, 2, 2, 2, 1, 1, 1, 0}

X_2 = The number of tails.

= {0, 1, 1, 1, 2, 2, 2, 3}

Rolling a Pair of Dice

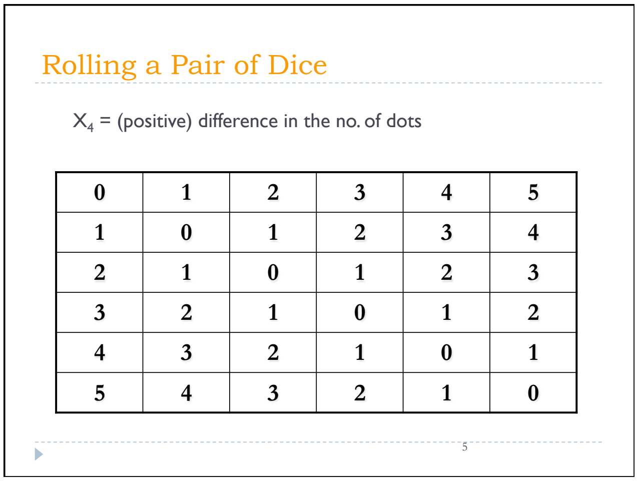
Sample Space:

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

Rolling a Pair of Dice

Random variable: X_3 = Total no. of dots

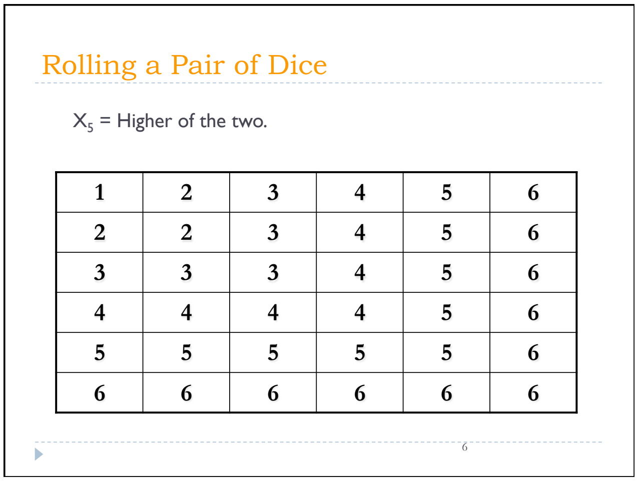
2	3	4	5	6	7
3	4	5	6	7	8
4	5	6	7	8	9
5	6	7	8	9	10
6	7	8	9	10	11
7	8	9	10	11	12



Rolling a Pair of Dice

X_4 = (positive) difference in the no. of dots

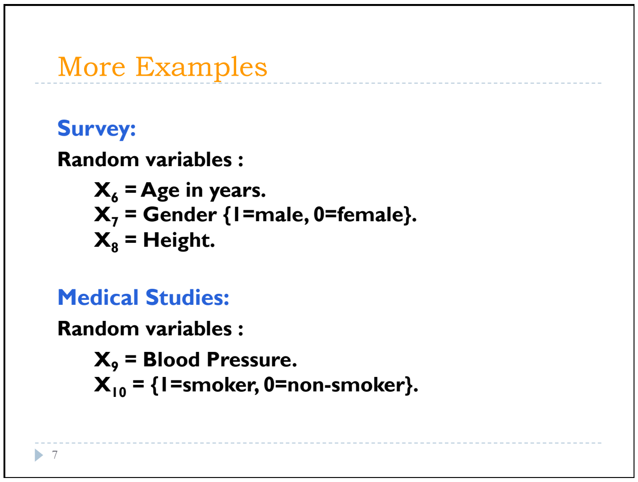
0	1	2	3	4	5
1	0	1	2	3	4
2	1	0	1	2	3
3	2	1	0	1	2
4	3	2	1	0	1
5	4	3	2	1	0



Rolling a Pair of Dice

X_5 = Higher of the two.

1	2	3	4	5	6
2	2	3	4	5	6
3	3	3	4	5	6
4	4	4	4	5	6
5	5	5	5	5	6
6	6	6	6	6	6



More Examples

Survey:

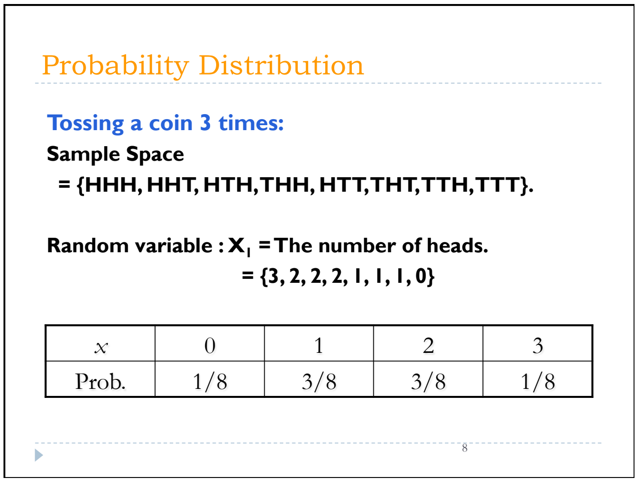
Random variables :

X_6 = Age in years.
X_7 = Gender {1=male, 0=female}.
X_8 = Height.

Medical Studies:

Random variables :

X_9 = Blood Pressure.
X_{10} = {1=smoker, 0=non-smoker}.



Probability Distribution

Tossing a coin 3 times:

Sample Space
= {HHH, HHT, HTH, THH, HTT, THT, TTH, TTT}.

Random variable : X_1 = The number of heads.
= {3, 2, 2, 2, 1, 1, 1, 0}

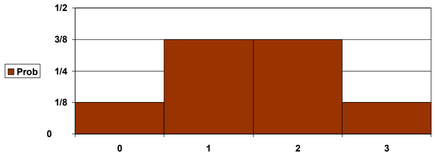
x	0	1	2	3
Prob.	1/8	3/8	3/8	1/8

Probability Histogram

Tossing a coin 3 times:

Random variable : X_1 = The number of heads.

X	0	1	2	3
Prob.	1/8	3/8	3/8	1/8



Rolling a Pair of Dice

Sample Space:

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

Rolling a Pair of Dice

Random variable: X_3 = Total no. of dots

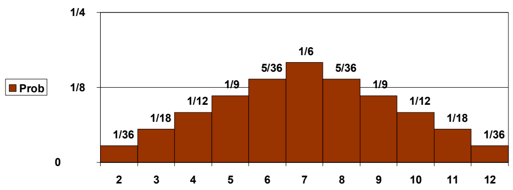
2	3	4	5	6	7
3	4	5	6	7	8
4	5	6	7	8	9
5	6	7	8	9	10
6	7	8	9	10	11
7	8	9	10	11	12

x	2	3	4	5	6	7	8	9	10	11	12
P	1/36	2/36	3/36	4/36	5/36	6/36	5/36	4/36	3/36	2/36	1/36

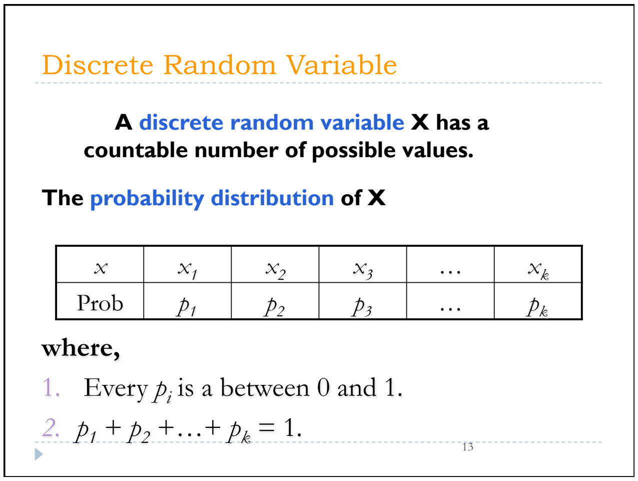
Rolling a Pair of Dice

Random variable: X_3 = Total no. of dots

x	2	3	4	5	6	7	8	9	10	11	12
P	1/36	2/36	3/36	4/36	5/36	6/36	5/36	4/36	3/36	2/36	1/36



1. $\Pr(X_3<5)=$ 2. $\Pr(3<X_3<12)=$



Discrete Random Variable

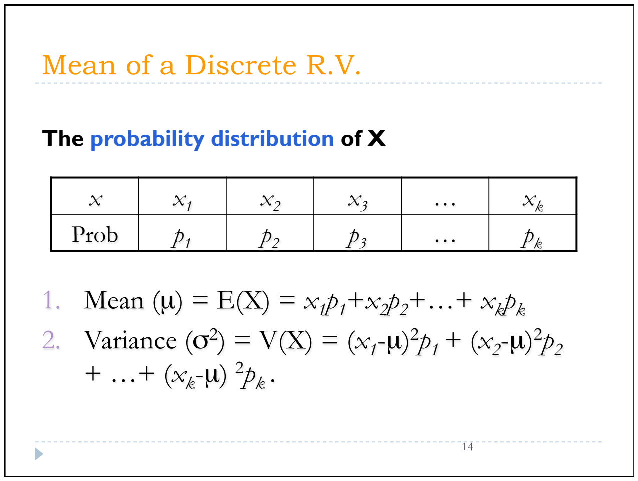
A discrete random variable X has a countable number of possible values.

The probability distribution of X

x	x_1	x_2	x_3	…	x_k
Prob	p_1	p_2	p_3	…	p_k

where,

1. Every p_i is a between 0 and 1.
2. $p_1 + p_2 + \ldots + p_k = 1$.

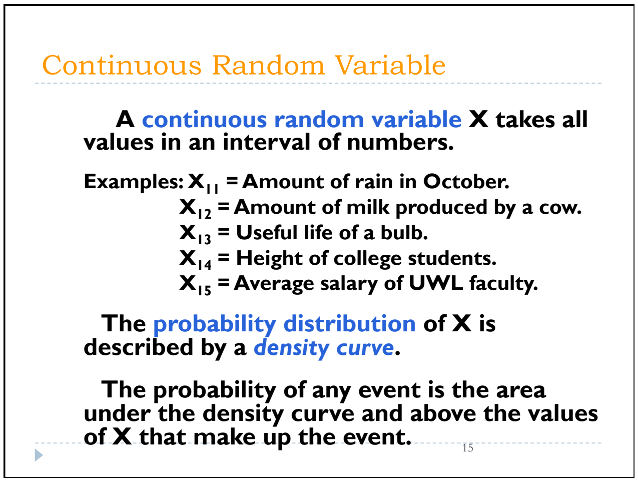


Mean of a Discrete R.V.

The probability distribution of X

x	x_1	x_2	x_3	…	x_k
Prob	p_1	p_2	p_3	…	p_k

1. Mean (μ) = E(X) = $x_1p_1+x_2p_2+\ldots+ x_kp_k$
2. Variance (σ^2) = V(X) = $(x_1-\mu)^2p_1 + (x_2-\mu)^2p_2 + \ldots + (x_k-\mu)^2p_k$.



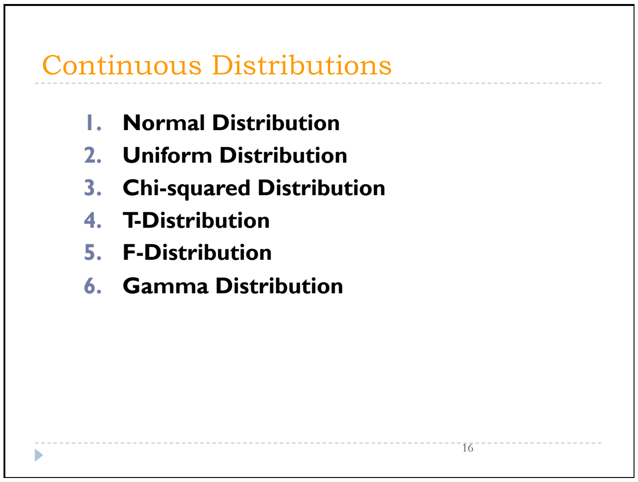
Continuous Random Variable

A continuous random variable X takes all values in an interval of numbers.

Examples: X_{11} = Amount of rain in October.
X_{12} = Amount of milk produced by a cow.
X_{13} = Useful life of a bulb.
X_{14} = Height of college students.
X_{15} = Average salary of UWL faculty.

The probability distribution of X is described by a *density curve*.

The probability of any event is the area under the density curve and above the values of X that make up the event.



Continuous Distributions

1. **Normal Distribution**
2. **Uniform Distribution**
3. **Chi-squared Distribution**
4. **T-Distribution**
5. **F-Distribution**
6. **Gamma Distribution**

Math 145 - Elementary Statistics **Summer 2015 #16**

Random Variables

- **Definition:** A *random variable* is a variable that assumes values associated with the random outcomes of an experiment, where one (and only one) numerical value is assigned to each sample point.

- **Definition:** The *probability distribution* of a discrete random variable is a graph, table, or formula that specifies the probability associated with each possible value the random variable can assume.

- **Examples:**

  1. Toss a fair coin three times and let X denote the number of heads.

  2. Roll a pair of dice (one red and one green) and let Y denote the total number of dots on the top faces.

	1	2	3	4	5	6
1						
2						
3						
4						
5						
6						

y											
$P(Y = y)$											

  3. Rolling a fair six-sided die until you get a 5. Let X denote the number of rolls needed to get the first 5.

- **Practice:**

  1. In the experiment of rolling a pair of six-sided dice, let the random variable Y be the absolute value of the difference in the number of dots. Construct the probability distribution table for Y and the corresponding probability histogram.

y						
$P(Y = y)$						

2. In the experiment of tossing a fair coin 4 times, let the random variable W be the number of heads out of the 4 tosses. Construct the probability distribution table for W and the corresponding probability histogram.

w					
$P(W = w)$					

3. In a survey study of 200 college students, list 5 meaningful random variables that you can define.

a.

b.

c.

d.

e.

- **Two Types of Random Variable.**

1. Discrete Random Variable: A *discrete random variable* has a *countable* number of possible values (i.e., there is a gap between possible values). Consider the discrete r.v. X with only k possible values, then the probability distribution of X can be expressed in a table form:

Values of X	x_1	x_2	x_3	$\cdots$	x_k
Probability	p_1	p_2	p_3	$\cdots$	p_k

where,

a. Every probability p_i is a number between 0 and 1.
b. $p_1 + p_2 + \cdots + p_k = 1$

Practice: Do # 18 and # 20 (p. 187)

2. Continuous Random Variable: A *continuous random variable* X takes all values in an interval of numbers. The *probability distribution* of X is described by a density curve. The probability of any event is the area under the density curve and above the values of X that make up the event. (*This is the topic for chapter 5.*)

Practice: Do # 4 (p. 183)

- **Examples of Common Discrete Probability Distribution:**

1. Uniform Distribution
2. Binomial Distribution
3. Geometric Distribution
4. Poisson Distribution
5. Hypergeometric Distribution

- **Homework.**

Section 4.1: page 183; # 3 - 5.
Section 4.2: pp. 187-190; # 17, 18, 19, 23, 27.

Means and Variances of Discrete R.V.

- **Discrete Random Variable:** A *discrete random variable* X has a countable number of possible values (That implies that there is a gap between possible values).
- **Discrete Probability Distribution:** Consider a discrete random variable X with only k possible outcomes. The table below is a discrete probability distribution of X if the following two requirements are satisfied:

  1. Every probability p_i is a number between 0 and 1.

  2. $p_1 + p_2 + \cdots + p_k = 1$

x	x_1	x_2	x_3	$\cdots$	x_k
$\Pr(X = x)$	p_1	p_2	p_3	$\cdots$	p_k

- **Mean and Variance of a Discrete Probability Distribution:**
  - The **Mean** (or **expected** value) of X, $\mu = E(X) = x_1p_1 + x_2p_2 + \cdots + x_kp_k = \sum_{i=1}^{k} x_ip_i$
  - The **Variance** of X,

$$\sigma^2 = (x_1 - \mu_X)^2p_1 + (x_2 - \mu_X)^2p_2 + \cdots + (x_k - \mu_X)^2p_k = \sum_{i=1}^{k}(x_i - \mu_X)^2p_i$$

- Example 1: Let X be the number of cars that Linda is able to sell on a typical Saturday. The probability distribution for X is given below:

Car sold (x)	0	1	2	3
Probability	0.3	0.4	0.2	0.1

Find the mean and variance of X.

1. $\mu =$

2. $\sigma^2 =$

- Example 2: In the experiment of rolling a die, let Z be the number of dots on top of the die. Construct the probability distribution table and compute the mean and variance of Z.

- Example 3: A large insurance company estimates that the probability that a \$20,000 car insurance policy will result in a claim in the amount of X dollars is given in the table below:

x	\$0	\$1,000	\$5,000	\$10,000	\$20,000
$p(x)$	0.894	0.09	0.01	0.005	0.001

1. Find the mean or expected value of X.

2. Find the standard deviation of X.

- **Law of Large Numbers:** Draw n independent observations at random from any population with finite mean μ. As the number of observations drawn increases, the sample average $\bar{x}$ of the observed values eventually approaches the mean μ of the population as closely as you wish and then stays that close.

- **Rules for Means:**

  1. If X is a random variable and a and b are fixed numbers, then

$$\mu_{a+b*X} = a + b * \mu_X \quad \text{or} \quad E(a + b * X) = a + b * E(X)$$

  2. If X and Y are random variables, then

$$\mu_{X \pm Y} = \mu_X \pm \mu_Y \quad \text{or} \quad E(X \pm Y) = E(X) \pm E(Y)$$

- **Example 4:** Linda is a sales associate at a large auto dealership. At her commission rate of 25% of profit on each vehicle she sells, Linda expects to earn \$350 for each car sold. The probability distributions for the number of cars (X) that she is able to sell on a typical Saturday is given below. If she has a fixed salary of \$50 per day, compute Linda's expected earnings on a typical Saturday.

Car sold (x)	0	1	2	3
Probability	0.3	0.4	0.2	0.1

- **Example 5:** In the previous example, suppose Linda also receives \$400 in commision for each truck or SUV sold. The probability distributions for the number of SUVs (Y) that she is able to sell on a typical Saturday is given below. Compute Linda's total expected earnings on such a day.

Vehicles sold (y)	0	1	2
Probability	0.4	0.5	0.1

- **Rules for Variances:**

  1. If X is a random variable and a and b are fixed numbers, then

$$Var(a + bX) = b^2 Var(X)$$

  2. If X and Y are *independent* random variables, then

$$Var(X \pm Y) = Var(X) + Var(Y)$$

- **Homework problems:**
  Section 4.3: pp. 193-195; # 35, 37, 39, 43, 47.

Binomial Distribution

- **Bernoulli Trial:** A *Bernoulli Trial* is a trial that has only two (2) possible outcomes − a success "S" or a failure "F".
- **Bernoulli Distribution:** Let the random variable $X = 1$ when a success is observed from a Bernoulli trial and $X = 0$ when a failure is observed. If the probability of getting a success is denoted by p, then the probability distribution of X is given below

x	1	0
$\Pr(X = x)$	p	$1-p$

- **Binomial Distribution:** Suppose a *Bernoulli* trial is repeated n times. The random variable X, that counts the number of successes out of n, follows the *Binomial* probability distribution. The probability for each possible value of X is given by the following formula

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

  - Mean: $\mu = np$
  - Variance: $\sigma^2 = np * (1-p)$

- A couple plans to have exactly 5 children. Let the random variable X be the number of girls that they will have. Assume that the probability of a male birth is 0.4 and the probability of a female birth is 0.6 and that gender of any successive child is unaffected by previous brothers or sisters.

  1. Find the probability distribution of X.

  2. Find the mean and standard deviation of X.

- Suppose that 20% of all copies of a particular textbook fail a certain binding strength test. If 15 of these textbooks are randomly selected, what is the

  1. probability that exactly 4 textbooks will fail this binding strength test?

  2. probability that at most 4 textbooks will fail this binding strength test?

  3. probability that at least 4 textbooks will fail this binding strength test?

  4. probability that between 4 to 7 (inclusive) textbooks will fail this binding strength test?

  5. mean, variance, and standard deviation of this distribution?

- **Practice problems.**

  1. Suppose that each time you take a free throw shot, you have a 60% chance of making it. If you take 15 shots,

  a. What is the probability of making exactly 8 of them?

  b. What is the probability of making fewer than 5 shots?

  c. What is the probability of making at least 10 of them?

  d. What is the probability of making between 6 to 10 (inclusive) shots?

  2. In a 20-question multiple choice exam (5 choices per item), what is the probability that a student who purely guesses his answers will get

  a. exactly 10 questions right?

  b. between 3 to 6 (inclusive) questions right?

  c. at least half of the questions right?

  3. The probability that a student will catch a cold from another student in the classroom is 0.20. A student with a cold is sitting in the classroom with 7 fellow students. What is the probability that

  a. exactly 2 of the fellow students will catch the cold?

  b. at least 2 of the fellow students will catch the cold?

  4. If 80 percent of the mortgage loan applications received by a savings and loan association are approved, what is the probability that among 25 loan applications

  a. less than 18 will be approved?

  b. at least 18 will be approved?

  c. at least 20 will be approved?

  d. between 20 and 23 (inclusive) will be approved?

- **Homework problems:**
  Section 4.4: pp. 204-207; # 55, 57, 59, 61, 65, 67, 71, 73.

Continuous Distributions

- Let X be a continuous random variable with density function $f(x)$ in the interval $[c, d]$.

  1. $P(X = a) \approx 0$, for any a in $[c, d]$.

  2. $P(a \leq X \leq b)$ is the area under the density curve between a to b.

  3. The total area under the density curve is 1.

- **Uniform Distribution**. $X \sim \text{Unif}[c, d]$.

$$f(x) = \frac{1}{d-c} \qquad (c \leq x \leq d)$$

  1. Mean: $\mu = \dfrac{c+d}{2}$.

  2. Variance: $\sigma^2 = \dfrac{(d-c)^2}{12}$.

Example: Too much cholesterol in the blood increases the risk of heart disease. Young women are generally less afflicted with high cholesterol than other groups. The cholesterol levels of women aged 20 to 34 follow an approximately uniform distribution from 120 mg/dl to 250 mg/dl.

1. What percent of young women have cholesterol levels below 245 mg/dl? *Sketch an appropriate density curve and shade the area under the curve that corresponds to this question.*

2. What percent of young women have cholesterol levels above 245 mg/dl? *Sketch an appropriate density curve and shade the area under the curve that corresponds to this question.*

3. What percent of young women have cholesterol levels between 210 and 245 mg/dl? *Sketch an appropriate density curve and shade the area under the curve that corresponds to this question.*

4. If 220 mg/dl is the threshold cholesterol level, what is the probability that a randomly selected young woman has a cholesterol level that is higher than this threshold value?

Homework: (pp. 229-231) # 3, 5, 9, 13, 17.

- **Normal Distribution**. $X \sim N(\mu, \sigma^2)$. The mean is μ and the variance is σ^2.

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

Example: Suppose the cholesterol levels of women aged 20 to 34 follow an approximately normal distribution with mean 185 milligrams per deciliter (mg/dl) and standard deviation 39 mg/dl.

1. What percent of young women have cholesterol levels below 245 mg/dl? *Sketch an appropriate normal curve and shade the area under the curve that corresponds to this question.*

2. What percent of young women have cholesterol levels above 245 mg/dl? *Sketch an appropriate normal curve and shade the area under the curve that corresponds to this question.*

3. What percent of young women have cholesterol levels between 210 and 245 mg/dl? *Sketch an appropriate normal curve and shade the area under the curve that corresponds to this question.*

4. If 220 mg/dl is the threshold cholesterol level, what is the probability that a randomly selected young woman has a cholesterol level that is higher than this threshold value?

Homework: (pp. 241-244) # 21, 25, 29, 31, 33, 35, 37, 41, 45.

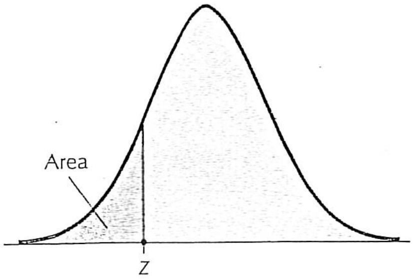


Table C Standard normal distribution

z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
–3.4	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0002
–3.3	0.0005	0.0005	0.0005	0.0004	0.0004	0.0004	0.0004	0.0004	0.0004	0.0003
–3.2	0.0007	0.0007	0.0006	0.0006	0.0006	0.0006	0.0006	0.0005	0.0005	0.0005
–3.1	0.0010	0.0009	0.0009	0.0009	0.0008	0.0008	0.0008	0.0008	0.0007	0.0007
–3.0	0.0013	0.0013	0.0013	0.0012	0.0012	0.0011	0.0011	0.0011	0.0010	0.0010
–2.9	0.0019	0.0018	0.0018	0.0017	0.0016	0.0016	0.0015	0.0015	0.0014	0.0014
–2.8	0.0026	0.0025	0.0024	0.0023	0.0023	0.0022	0.0021	0.0021	0.0020	0.0019
–2.7	0.0035	0.0034	0.0033	0.0032	0.0031	0.0030	0.0029	0.0028	0.0027	0.0026
–2.6	0.0047	0.0045	0.0044	0.0043	0.0041	0.0040	0.0039	0.0038	0.0037	0.0036
–2.5	0.0062	0.0060	0.0059	0.0057	0.0055	0.0054	0.0052	0.0051	0.0049	0.0048
–2.4	0.0082	0.0080	0.0078	0.0075	0.0073	0.0071	0.0069	0.0068	0.0066	0.0064
–2.3	0.0107	0.0104	0.0102	0.0099	0.0096	0.0094	0.0091	0.0089	0.0087	0.0084
–2.2	0.0139	0.0136	0.0132	0.0129	0.0125	0.0122	0.0119	0.0116	0.0113	0.0110
–2.1	0.0179	0.0174	0.0170	0.0166	0.0162	0.0158	0.0154	0.0150	0.0146	0.0143
–2.0	0.0228	0.0222	0.0217	0.0212	0.0207	0.0202	0.0197	0.0192	0.0188	0.0183
–1.9	0.0287	0.0281	0.0274	0.0268	0.0262	0.0256	0.0250	0.0244	0.0239	0.0233
–1.8	0.0359	0.0351	0.0344	0.0336	0.0329	0.0322	0.0314	0.0307	0.0301	0.0294
–1.7	0.0446	0.0436	0.0427	0.0418	0.0409	0.0401	0.0392	0.0384	0.0375	0.0367
–1.6	0.0548	0.0537	0.0526	0.0516	0.0505	0.0495	0.0485	0.0475	0.0465	0.0455
–1.5	0.0668	0.0655	0.0643	0.0630	0.0618	0.0606	0.0594	0.0582	0.0571	0.0559
–1.4	0.0808	0.0793	0.0778	0.0764	0.0749	0.0735	0.0721	0.0708	0.0694	0.0681
–1.3	0.0968	0.0951	0.0934	0.0918	0.0901	0.0885	0.0869	0.0853	0.0838	0.0823
–1.2	0.1151	0.1131	0.1112	0.1093	0.1075	0.1056	0.1038	0.1020	0.1003	0.0985
–1.1	0.1357	0.1335	0.1314	0.1292	0.1271	0.1251	0.1230	0.1210	0.1190	0.1170
–1.0	0.1587	0.1562	0.1539	0.1515	0.1492	0.1469	0.1446	0.1423	0.1401	0.1379
–0.9	0.1841	0.1814	0.1788	0.1762	0.1736	0.1711	0.1685	0.1660	0.1635	0.1611
–0.8	0.2119	0.2090	0.2061	0.2033	0.2005	0.1977	0.1949	0.1922	0.1894	0.1867
–0.7	0.2420	0.2389	0.2358	0.2327	0.2296	0.2266	0.2236	0.2206	0.2177	0.2148
–0.6	0.2743	0.2709	0.2676	0.2643	0.2611	0.2578	0.2546	0.2514	0.2483	0.2451
–0.5	0.3085	0.3050	0.3015	0.2981	0.2946	0.2912	0.2877	0.2843	0.2810	0.2776
–0.4	0.3446	0.3409	0.3372	0.3336	0.3300	0.3264	0.3228	0.3192	0.3156	0.3121
–0.3	0.3821	0.3783	0.3745	0.3707	0.3669	0.3632	0.3594	0.3557	0.3520	0.3483
–0.2	0.4207	0.4168	0.4129	0.4090	0.4052	0.4013	0.3974	0.3936	0.3897	0.3859
–0.1	0.4602	0.4562	0.4522	0.4483	0.4443	0.4404	0.4364	0.4325	0.4286	0.4247
–0.0	0.5000	0.4960	0.4920	0.4880	0.4840	0.4801	0.4761	0.4721	0.4681	0.4641

(Continued)

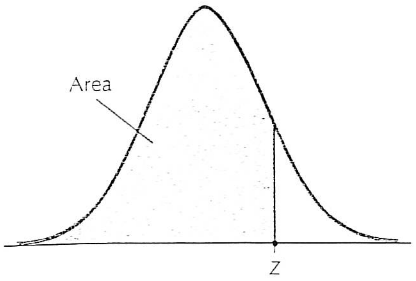


Table C Standard normal distribution (*continued*)

z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.5000	0.5040	0.5080	0.5120	0.5160	0.5199	0.5239	0.5279	0.5319	0.5359
0.1	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596	0.5636	0.5675	0.5714	0.5753
0.2	0.5793	0.5832	0.5871	0.5910	0.5948	0.5987	0.6026	0.6064	0.6103	0.6141
0.3	0.6179	0.6217	0.6255	0.6293	0.6331	0.6368	0.6406	0.6443	0.6480	0.6517
0.4	0.6554	0.6591	0.6628	0.6664	0.6700	0.6736	0.6772	0.6808	0.6844	0.6879
0.5	0.6915	0.6950	0.6985	0.7019	0.7054	0.7088	0.7123	0.7157	0.7190	0.7224
0.6	0.7257	0.7291	0.7324	0.7357	0.7389	0.7422	0.7454	0.7486	0.7517	0.7549
0.7	0.7580	0.7611	0.7642	0.7673	0.7704	0.7734	0.7764	0.7794	0.7823	0.7852
0.8	0.7881	0.7910	0.7939	0.7967	0.7995	0.8023	0.8051	0.8078	0.8106	0.8133
0.9	0.8159	0.8186	0.8212	0.8238	0.8264	0.8289	0.8315	0.8340	0.8365	0.8389
1.0	0.8413	0.8438	0.8461	0.8485	0.8508	0.8531	0.8554	0.8577	0.8599	0.8621
1.1	0.8643	0.8665	0.8686	0.8708	0.8729	0.8749	0.8770	0.8790	0.8810	0.8830
1.2	0.8849	0.8869	0.8888	0.8907	0.8925	0.8944	0.8962	0.8980	0.8997	0.9015
1.3	0.9032	0.9049	0.9066	0.9082	0.9099	0.9115	0.9131	0.9147	0.9162	0.9177
1.4	0.9192	0.9207	0.9222	0.9236	0.9251	0.9265	0.9279	0.9292	0.9306	0.9319
1.5	0.9332	0.9345	0.9357	0.9370	0.9382	0.9394	0.9406	0.9418	0.9429	0.9441
1.6	0.9452	0.9463	0.9474	0.9484	0.9495	0.9505	0.9515	0.9525	0.9535	0.9545
1.7	0.9554	0.9564	0.9573	0.9582	0.9591	0.9599	0.9608	0.9616	0.9625	0.9633
1.8	0.9641	0.9649	0.9656	0.9664	0.9671	0.9678	0.9686	0.9693	0.9699	0.9706
1.9	0.9713	0.9719	0.9726	0.9732	0.9738	0.9744	0.9750	0.9756	0.9761	0.9767
2.0	0.9772	0.9778	0.9783	0.9788	0.9793	0.9798	0.9803	0.9808	0.9812	0.9817
2.1	0.9821	0.9826	0.9830	0.9834	0.9838	0.9842	0.9846	0.9850	0.9854	0.9857
2.2	0.9861	0.9864	0.9868	0.9871	0.9875	0.9878	0.9881	0.9884	0.9887	0.9890
2.3	0.9893	0.9896	0.9898	0.9901	0.9904	0.9906	0.9909	0.9911	0.9913	0.9916
2.4	0.9918	0.9920	0.9922	0.9925	0.9927	0.9929	0.9931	0.9932	0.9934	0.9936
2.5	0.9938	0.9940	0.9941	0.9943	0.9945	0.9946	0.9948	0.9949	0.9951	0.9952
2.6	0.9953	0.9955	0.9956	0.9957	0.9959	0.9960	0.9961	0.9962	0.9963	0.9964
2.7	0.9965	0.9966	0.9967	0.9968	0.9969	0.9970	0.9971	0.9972	0.9973	0.9974
2.8	0.9974	0.9975	0.9976	0.9977	0.9977	0.9978	0.9979	0.9979	0.9980	0.9981
2.9	0.9981	0.9982	0.9982	0.9983	0.9984	0.9984	0.9985	0.9985	0.9986	0.9986
3.0	0.9987	0.9987	0.9987	0.9988	0.9988	0.9989	0.9989	0.9989	0.9990	0.9990
3.1	0.9990	0.9991	0.9991	0.9991	0.9992	0.9992	0.9992	0.9992	0.9993	0.9993
3.2	0.9993	0.9993	0.9994	0.9994	0.9994	0.9994	0.9994	0.9995	0.9995	0.9995
3.3	0.9995	0.9995	0.9995	0.9996	0.9996	0.9996	0.9996	0.9996	0.9996	0.9997
3.4	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9998

Normal Approximation To Binomial

- If $X \sim bin(n, p)$ with $np \geq 10$ and $n(1-p) \geq 10$, then $X \approx N(\mu = np, \sigma = \sqrt{np(1-p)})$.
- **Examples:**

1. Let $X \sim bin(n = 10, p = 0.4)$. Calculate $P(X \leq 5)$.

a. Using the exact binomial distribution.

b. Using the normal approximation without continuity correction.

c. Using the normal approximation with continuity correction.

2. An unnoticed mechanical failure has caused $\frac{1}{3}$ of a machine shop's production of rifle firing pins to be defective. If an inspector will check 90 random selected pins from this batch, what is the probability that the inspector will find

a. no more than 25 defective pins?

b. less than 25 defective pins?

c. at least 20 defective pins?

d. between 22 to 34 (inclusive) defective pins?

3. Suppose that 40% of all drivers in a certain state regularly wear a seat belt. A random sample of 500 drivers is selected. What is the probability that

a. fewer than 175 of those in the sample regularly wear a seat belt?

b. at least 220 of those in the sample regularly wear a seat belt?

c. between 180 and 230 (inclusive) of the drivers in the sample regularly wear a seat belt?

- **Homework problems:**
  Section 5.5: pp. 256-257; # 75, 77, 79, 81,83,91.

Central Limit Theorem

- **Parameter.** A *parameter* is a numerical descriptive measure of a population. *This quantity is usually unknown but it is constant at a given time.*

  Examples: $\mu, \sigma^2, \sigma, \rho$.

- **Statistic.** A *sample statistic* is a numerical descriptive measure of a sample. *Its value is calculated from the observations in the sample.*

  Examples: $\bar{x}, s^2, s, \hat{p}$.

- **Sampling Distribution.** The *sampling distribution* of a sample statistic calculated from a sample of n measurements is the probability distribution of the statistic.

- **Point Estimator.** A *point estimator* of a population parameter is a rule or formula that tells us how to use the sample data to calculate a single number that can be used as an *estimate* of the population parameter.

- **Unbiased Estimator.** If the sampling distribution of a sample statistic has a mean equal to the population parameter the statistic is intended to estimate, the statistic is said to be an *unbiased estimator* of the parameter. Otherwise, the statistic is said to be *biased estimator* of the parameter.

- **Sampling Distribution of a Sample Mean:** If a population has the $N(\mu, \sigma)$ distribution, then the sample mean $\bar{X}$ of n independent observations has the $N(\mu, \sigma/\sqrt{n})$ distribution.That is,

$$\bar{X} \sim N(\mu, \sigma/\sqrt{n}). \tag{1}$$

- **Central Limit Theorem:** Draw an **SRS** of size n from any population with mean μ and finite standard deviation σ. When n is large (*rule of thumb*: $n \geq 30$), the sampling distribution of the sample mean $\bar{X}$ is *approximately* normal with mean μ and standard deviation $\sigma/\sqrt{n}$. That is,

$$\bar{X} \approx N(\mu, \sigma/\sqrt{n}). \tag{2}$$

- **Practice:**

  1. A soft-drink machine is regulated so that it discharges an average of 200 ml per cup. If the amount of drink discharged is normally distributed with a standard deviation equal to 10 ml, what is the probability that a cup will contain less than 188 ml?

  2. In the previous problem, suppose we sample 25 cups from this machine and take their average content. What is the probability that, the average content of these 25 cups is less than 188 ml?

3. The mean time it takes a senior high school student to complete a certain achievement test is 46.2 minutes with standard deviation of 8 minutes. If a random sample of 100 senior high school students who took the test was selected, find the probability that the average time it takes the group to complete the test will be

a. less than 44 minutes.

b. more than 48 minutes.

c. between 44 to 48 minutes.

d. more than 50 minutes.

- **Homework.**

  Sec 6.1: (pp. 278-279) # 1, 3, 5.
  Sec 6.2: (pp. 282-283) # 11, 13, 15.
  Sec 6.3: (pp. 289-292) # 21, 23, 25, 27, 29, 31, 37, 41.
  Supplementary: (pp. 293-296) # 49, 51, 55, 61, 67, 73.

Math 145 - Elementary Statistics — Exam II Review

- **Chapter 4: Discrete Random Variables and Probability Distributions**

  1. ***Discrete Random Variable:*** Let X be a discrete random variable with probability distribution given below.

Values of X	x_1	x_2	x_3	$\cdots$	x_k
Probability	p_1	p_2	p_3	$\cdots$	p_k

  where,

  a. Every probability p_i is a number between 0 and 1.

  b. *Total probability*: $p_1 + p_2 + \cdots + p_k = 1$.

  c. *Cumulative Distribution*: $F(x) = P(X \leq x)$.

  d. *Mean* (or **expected** value) of X: $\mu = E(X) = x_1p_1 + x_2p_2 + \cdots + x_kp_k = \sum_{i=1}^{k} x_ip_i$

  e. *Variance* of X: $\sigma^2 = Var(X) = (x_1-\mu)^2p_1 + \cdots + (x_k-\mu)^2p_k = \sum_{i=1}^{k}(x_i-\mu)^2p_i$

  f. *Standard Deviation* of X: $\sigma = \sqrt{Var(X)}$

  g. *Rules for Means*: (i) $E(a + b * X) = a + b * E(X)$ and (ii) $E(X \pm Y) = E(X) \pm E(Y)$.

  h. *Rules for Variance*: $Var(a + b * X) = b^2 * Var(X) \qquad \Rightarrow SD(a + b * X) = b * SD(X)$.

  2. ***Binomial Distribution:*** Suppose $X \sim Bin(n, p)$, then

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

  a. *Mean*: $\mu = np$

  b. *Variance*: $\sigma^2 = np(1-p)$

  c. *Special Case*: When $n = 1$, $X \sim Ber(p)$ [***Bernoulli Distribution***]

- **Chapter 5: Continuous Random Variables and Density Functions**

  1. ***Properties:*** Let X be a continuous random variable.

  a. $P(X = a)$ is virtually 0.

  b. $P(a \leq X \leq b)$ is the area under the density curve between a and b.

  c. The total area under a density curve is 1.

  2. ***Uniform Distribution***. Let $X \sim \text{Unif}[c, d]$. Then, $f(x) = \frac{1}{d-c}$, when x is in $[c, d]$.

  a. Mean: $\mu = \dfrac{c+d}{2}$.

  b. Variance: $\sigma^2 = \dfrac{(d-c)^2}{12} \qquad \Rightarrow$ Standard Deviation: $\sigma = \dfrac{d-c}{\sqrt{12}}$.

  3. ***Normal Distribution***. Let $X \sim N(\mu, \sigma)$. The mean is μ and the standard deviation is σ.

  a. *Standard Normal*: $Z = \frac{X-\mu}{\sigma} \sim N(\mu = 0, \sigma = 1)$.

  b. $P(a < X < b) = P(\frac{a-\mu}{\sigma} < \frac{X-\mu}{\sigma} < \frac{b-\mu}{\sigma}) = P(z_a < Z < z_b) = F(z_b) - F(z_a)$.

  4. If $X \sim Bin(n, p)$ with $np \geq 10$ and $n(1-p) \geq 10$, then $X \approx N(\mu = np, \sigma = \sqrt{np(1-p)})$.
  [Apply ***Continuity Correction:***] $\Rightarrow P(X_{Binomial} \leq a) \approx P(X^*_{Normal} \leq a + 0.5)$.

- **Chapter 6: Sampling Distributions**

  1. If $E(X) = \mu$ and $SD(X) = \sigma$, then $E(\bar{X}) = \mu$ and $SD(\bar{X}) = \dfrac{\sigma}{\sqrt{n}}$.

  2. If X is normal, then $\bar{X}$ is also normal.

  3. ***Central Limit Theorem:*** When the sample size (n) is large (*rule of thumb*: $n \geq 30$), the sampling distribution of the sample mean $\bar{X}$ is *approximately normal.* That is,

$$\bar{X} \approx N(\mu_{\bar{X}} = \mu, \sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}})$$

Confidence Interval for a Population Mean (σ^2 is known)

1. Important Results:

a. If $X \sim N(\mu, \sigma)$, then $\bar{X} \sim N(\mu, \frac{\sigma}{\sqrt{n}}) \quad \Rightarrow \quad Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \sim N(0,1)$

b. If $n \geq 30$, then $\bar{X} \approx N(\mu, \frac{\sigma}{\sqrt{n}}) \quad \Rightarrow \quad Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \approx N(0,1)$

2. Notation. The value z_α is defined as the value of the standard normal random variable Z such that the area α will lie to its right. In other words, $P(Z > z_\alpha) = \alpha$.

3. The $(1-\alpha)100\%$ confidence interval for μ is $\left[\bar{X} - z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}, \bar{X} + z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}\right]$.

a. The 90% C.I. for μ is $\left[\bar{X} - 1.645\frac{\sigma}{\sqrt{n}}, \bar{X} + 1.645\frac{\sigma}{\sqrt{n}}\right]$.

b. The 95% C.I. for μ is $\left[\bar{X} - 1.96\frac{\sigma}{\sqrt{n}}, \bar{X} + 1.96\frac{\sigma}{\sqrt{n}}\right]$.

c. The 99% C.I. for μ is $\left[\bar{X} - 2.576\frac{\sigma}{\sqrt{n}}, \bar{X} + 2.576\frac{\sigma}{\sqrt{n}}\right]$.

4. Practice

a. A random sample of 90 observations were obtained from a population with a known standard deviation of $\sigma = 2.7$. The sample produced an average of $\bar{x} = 25.9$.

i. Find a 95% confidence interval for μ. What is the margin of error?

ii. Find a 90% confidence interval for μ. What is the margin of error?

iii. Find a 99% confidence interval for μ. What is the margin of error?

b. A random sample of 100 observations from a normal population with $\sigma = 6.4$ yielded a sample mean of 83.2.

i. Find a 95% confidence interval for μ.

ii. What does it mean when you say that the confidence level is 95%?

iii. Find a 99% confidence interval for μ.

iv. What happens to the width of a confidence interval as the value of the confidence level is increased while the sample size is held fixed?

v. Would your confidence intervals of parts (i) and (iii) be valid if the distribution of the original population was not normal? Explain.

c. A random sample of n measurements were taken from a population with $\sigma = 3.3$. The sample produced an average of 33.9.

i. Find a 95% confidence interval for μ if $n = 100$.

ii. Find a 95% confidence interval for μ if $n = 400$.

iii. What is the effect on the width of a confidence interval of quadrupling the sample size while holding the confidence coefficient fixed?

Homework problems:
Section 7.2: pp. 306-310; # 3, 7, 9, 11, 13, 15, 25.

Confidence Intervals for One Population

- **Estimating the population mean (μ) when σ is known.**

  1. Use $Z = \dfrac{\bar{X} - \mu}{\sigma/\sqrt{n}}$. Z follows the $N(0, 1)$.

  2. The $(1 - \alpha)100\%$ confidence interval for μ is $\left[\bar{X} - z_{\frac{\alpha}{2}}\dfrac{\sigma}{\sqrt{n}}, \bar{X} + z_{\frac{\alpha}{2}}\dfrac{\sigma}{\sqrt{n}}\right]$.

  a. The 90% C.I. for μ is $\left[\bar{X} - 1.645\dfrac{\sigma}{\sqrt{n}}, \bar{X} + 1.645\dfrac{\sigma}{\sqrt{n}}\right]$.

  b. The 95% C.I. for μ is $\left[\bar{X} - 1.96\dfrac{\sigma}{\sqrt{n}}, \bar{X} + 1.96\dfrac{\sigma}{\sqrt{n}}\right]$.

  c. The 99% C.I. for μ is $\left[\bar{X} - 2.576\dfrac{\sigma}{\sqrt{n}}, \bar{X} + 2.576\dfrac{\sigma}{\sqrt{n}}\right]$.

  3. The *Margin of Error*, $M = Z_{\frac{\alpha}{2}}\dfrac{\sigma}{\sqrt{n}}$.

  4. For a specified margin of error M, the required sample size is $n = \left(\dfrac{Z_{\frac{\alpha}{2}}\sigma}{M}\right)^2$.

- **Estimating the population mean (μ) when σ is unknown but the sample size $n \geq 30$.**
  *In this case, treat **s** as if it is **σ**, then use the first method. This is due to the fact that the sample size is large ($n \geq 30$) and hence, the value of **s** is very close to **σ**.*

- **Estimating the population mean (μ) when σ is unknown and the sample size $n < 30$.**

  1. Use $T = \dfrac{\bar{X} - \mu}{s/\sqrt{n}}$. T follows the student t-distribution with $n - 1$ degrees of freedom.

  2. The $(1 - \alpha)100\%$ confidence interval for μ is $\left[\bar{X} - (t_{\{\frac{\alpha}{2},(n-1)\}})\dfrac{s}{\sqrt{n}}, \bar{X} + (t_{\{\frac{\alpha}{2},(n-1)\}})\dfrac{s}{\sqrt{n}}\right]$.

  3. The *Margin of Error*, $M = (t_{\{\frac{\alpha}{2},(n-1)\}})\dfrac{s}{\sqrt{n}}$.

- **Estimating the population proportion (p) when the sample size is large.**

  1. The unbiased estimator of p is the sample proportion $\hat{p} = \dfrac{Y}{n}$, where Y is the number of successes in the sample.
  Recall that if $Y \sim \text{bin}(n, p)$, then $E(Y) = np$ and $Var(Y) = np(1 - p)$.
  Therefore, $E(\hat{p}) = p$ and $Var(\hat{p}) = \frac{p(1-p)}{n}$.

  2. The $(1 - \alpha)100\%$ confidence interval for p is $\left[\hat{p} - (z_{\frac{\alpha}{2}})\sqrt{\dfrac{\hat{p}(1 - \hat{p})}{n}}, \hat{p} + (z_{\frac{\alpha}{2}})\sqrt{\dfrac{\hat{p}(1 - \hat{p})}{n}}\right]$.

  3. The *Margin of Error*, $M = (z_{\frac{\alpha}{2}})\sqrt{\dfrac{\hat{p}(1 - \hat{p})}{n}} \quad \Rightarrow \quad n = \dfrac{z^2_{\alpha/2}\hat{p}(1 - \hat{p})}{M^2} \leq \dfrac{z^2_{\alpha/2}(0.25)}{M^2}$.

Homework problems:
Section 7.3: (pp. 316-319) # 27, 31, 33, 35, 37, 39. 43.
Section 7.4: (pp. 325-327) # 49, 51, 53, 55, 61.
Section 7.5: (pp. 332-334) # 69, 71, 73, 75, 81, 85, 87, 89.
Supplementary: (pp. 340-345) # 109, 111, 113, 115, 119, 127, 137.

- Practice problems.

  1. A soft-drink machine is regulated so that the amount of drink dispensed is approximately normally distributed with a standard deviation equal to 1.5 deciliters. A random sample of 36 drinks had an average content of 22.5 deciliters.

  a. Construct a 95% confidence interval estimate for the mean of all drinks dispensed by this machine.

  b. Give a practical interpretation for the interval estimate you obtained in part (**a**).

  c. Determine how large a sample is needed if we wish to be 95% confident that our estimate will be within 0.1 deciliters of the true mean?

  d. Construct a 99% confidence interval estimate for the mean of all drinks dispensed by this machine. Interpret your answer.

  2. The contents of 10 similar containers of a commercial soap are 10.2, 9.7, 10.1, 10.3, 10.1, 9.8, 9.9, 10.4, 10.3, and 9.8 liters. Assume that these values come from a normal population.

  a. Find a 95% confidence interval for the mean soap content of all such containers. Interpret your answer.

b. Find a 99% confidence interval for the mean soap content of all such containers. Interpret your answer.

3. In the Federal Trade Commission (FTC) "Price Check" study of electronic checkout scanners, the FTC inspected 1,669 scanners at retail stores and supermarkets by scanning a sample of items at each store and determining if the scanned price was accurate. The FTC gives a store a "passing grade" if 98% or more of the items are priced accurately. Of the 1,669 stores in the study, 1,185 passed inspection.

a. Find a 90% confidence interval for the true proportion of retail stores and supermarkets with electronic scanners that pass the FTC price-check test. Interpret your result.

b. Two years prior to the study, the FTC found that 45% of the stores passed inspection. Use the interval you obtained in part (**a**) to determine whether the proportion of stores that now pass inspection exceeds 45%.

c. Determine the sample size need to have a margin of error of at most 0.01.

Test of Hypothesis

- **Null Hypothesis:** The *null hypothesis*, denoted by H_0, represents the hypothesis that will be accepted unless the data provide convincing evidence that it is false. This usually represents the "status quo" or some claim about the population parameter that the researcher wants to test.

- **Alternative Hypothesis:** The *alternative hypothesis*, denoted by H_a or H_1, represents the hypothesis that will be accepted only if the data provide convincing evidence of its truth. This usually represents the values of a population parameter for which the researcher wants to gather evidence to support.

  Common Types:

  1. $H_0 : \mu = \mu_0$ vs. $H_1 : \mu > \mu_0$ (One-sided alternative)
  2. $H_0 : \mu = \mu_0$ vs. $H_1 : \mu < \mu_0$ (One-sided alternative)
  3. $H_0 : \mu = \mu_0$ vs. $H_1 : \mu \neq \mu_0$ (Two-sided alternative)

- **Test Statistic:** A *test statistic* measures compatibility between the null hypothesis and the observed data.

- **Z-Test for a Population Mean (σ is known):** To test the hypothesis $H_0 : \mu = \mu_0$ based on a SRS of size n from a population with unknown mean μ and known standard deviation σ, compute the test statistic

$$Z = \frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}} \sim N(0, 1). \qquad (3)$$

- **Steps to do Hypothesis Testing:**

  1. Formulate the Null hypothesis (H_0) and Alternative hypothesis (H_a).
  2. Specify the level of significance (Commonly used: $\alpha = 0.05$ or 0.01).
  3. Determine the appropriate *test statistic* to use. State the required assumptions.
  4. Define your rejection rule. (Or use, reject H_0 if p-value $< \alpha$).
  5. Compute the observed value of the test statistic (or compute the p-value).
  6. Write your conclusion.

- **Examples:**

  1. ETS, the company that administers the SAT exam, reports that the mean SAT mathematics score is 519. But some people think that this score overestimates the ability of typical high school seniors because only those who plan to attend college take the SAT. To check if this is true, a test was given to a SRS of 500 seniors from California. These students had an average score of $\bar{x} = 504$. Is this enough evidence to say that the mean for all California seniors is lower than 519? Use a level of significance equal to $\alpha = 0.05$. (Assume that $\sigma = 100$).

2. Do middle-aged male executives have different average blood pressure than the general population? The national Center for Health Statistics reports that the mean systolic blood pressure for males 35 to 44 years of age is 128 and the standard deviation in this population is 15. The medical director of a company looks at the medical records of 72 company executives in this age group and finds that the mean systolic blood pressure in this sample is $\bar{x} = 126.07$. Is this enough evidence that executive blood pressures differ from the national average? Use $\alpha = 0.05$.

- **Practice:**

1. The Federal Trade Commission (FTC) periodically conducts statistical studies designed to test the claims that manufacturers make about their products. For example, the label on a large can of Hilltop Coffee states that the can contains 3 pounds of coffee. To protect the consumers, FTC wants to make sure that the population mean amount of coffee in each can is at least 3 pounds. If a sample of 36 Hilltop coffee cans provides a sample mean of $\bar{x} = 2.92$ pounds and σ is known to be 0.18, is there enough evidence to conclude that the population mean is statistically lower than 3 pounds per can? Use a level of significance equal to $\alpha = 0.05$.

2. Reis, Inc., a New York real state research firm, tracks the cost of apartment rentals in the United States. In mid-2002, the nationwide mean apartment rental rate was \$895 per month (*The Wall Street Journal*, July 8, 2002). Assume that, based on the historical quarterly surveys, a population standard deviation $\sigma = \$225$ is reasonable. In a current study of apartment rental, a sample of 180 apartments nationwide provided a sample mean of \$915 per month. Do the sample data enable Reis to conclude that the population mean apartment rental rate now exceeds the level reported in 2002? $\alpha = 0.01$.

3. The Florida Department of Labor and Employment Security reported the state mean annual wage was \$26,133 (*The Naples Daily News*, Feb. 13, 1999). A hypothesis test of wages by county can be conducted to see whether the mean annual wage for a particular county differs from the state mean. A sample of 550 individuals from Collier County showed a sample mean annual wage of \$25,457. Assuming that $\sigma = \$7600$, is there enough evidence to conclude that mean annual wage of people from this county is different than the state mean? Use $\alpha = 0.05$.

4. The national mean sales price for new one-family homes is \$181,900 (*The New York Times Almanac 2000*). A sample of 40 one-family home sales in the south showed a sample mean of \$166,400. If $\sigma = \$33,500$, is there enough evidence to say that the population mean sales price for new one-family homes in the south is less than the national mean? Use $\alpha = 0.01$.

5. Individuals filing federal income tax returns prior to March 31 received an average refund of \$1056. Consider the population of "last minute" filers who mail their tax return during the last five days of the income tax period (typically April 10 to April 15).

a. A researcher suggests that a reason individuals wait until the last five days is that on average these individuals receive lower refunds than do early filers. Formulate appropriate hypotheses such that the rejection of H_0 will support the researcher's contention. Clearly define the parameter you used.

b. For a sample of 400 individuals who filed a tax return between April 10 to April 15, the sample mean refund was \$910. Based on prior experience a population standard deviation of $\sigma = \$1600$ may be assumed. Calculate the value of the appropriate test statistic.

c. At $\alpha = 0.05$, what is your conclusion?

d. Compute the $p-$value? What is your conclusion based on the value of the $p-$value?
The **observed significance level***, or* **p-value***, for a specific statistical test is the probability (assuming that H_0 is true) of observing a value of the test statistic that is at least as contradictory to the null hypothesis, and supportive of the alternative hypothesis, as the actual one computed from the sample data.*

Type I and Type II Errors:

1. *Type I error* is rejecting H_0 when it is true.
Note: Pr(Type I error)=α.

2. *Type II error* is not rejecting H_0 when it is false.
Note: Pr(Type II error)=β, and the *Power* of the test is $(1 - \beta)$.

Homework:
Sec 8.2: (pp. 360-361) # 1, 9, 11, 13.
Sec 8.3: (pp. 364-367) # 21, 23, 25, 27, 29, 31, 35.
Sec 8.4: (pp. 371-373) # 39, 41, 43, 45, 47, 49, 51.

Summary

1. One population (μ)

a. If σ is known or $n \geq 30$, $Z = \dfrac{\bar{X} - \mu}{\sigma/\sqrt{n}}$ follows the $N(0,1)$.

i. The $(1-\alpha)100\%$ confidence interval for μ is $[\bar{X} - Z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}, \bar{X} + Z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}]$.

ii. To test $H_0 : \mu = \mu_0$ vs. $H_1 : \mu > \mu_0$, reject the null if $Z_{\text{obs}} = \dfrac{\bar{X}_{\text{obs}} - \mu}{\sigma/\sqrt{n}} > Z_\alpha$, or if the $p-$value$= P(Z \geq Z_{\text{obs}})$ is $< \alpha$.

iii. To test $H_0 : \mu = \mu_0$ vs. $H_1 : \mu < \mu_0$, reject the null if $Z_{\text{obs}} < -Z_\alpha$, or if the $p-$value$= P(Z \leq Z_{\text{obs}})$ is $< \alpha$.

iv. To test $H_0 : \mu = \mu_0$ vs. $H_1 : \mu \neq \mu_0$, reject the null if $|Z_{\text{obs}}| > Z_{\frac{\alpha}{2}}$, or if the $p-$value$= 2 * P(Z \geq |Z_{\text{obs}}|)$ is $< \alpha$.

b. If σ is unknown and the $X_i' s$ come from a normal population, $t = \dfrac{\bar{X} - \mu}{s/\sqrt{n}}$ follows the $t-$distribution with $(n-1)$ degrees of freedom.

i. The $(1-\alpha)100\%$ confidence interval for μ is $[\bar{X} - t_{\frac{\alpha}{2}} \frac{s}{\sqrt{n}}, \bar{X} + t_{\frac{\alpha}{2}} \frac{s}{\sqrt{n}}]$.

ii. To test $H_0 : \mu = \mu_0$ vs. $H_1 : \mu > \mu_0$, reject the null if $t_{\text{obs}} = \dfrac{\bar{X}_{\text{obs}} - \mu}{s/\sqrt{n}} > t_\alpha$.

iii. To test $H_0 : \mu = \mu_0$ vs. $H_1 : \mu < \mu_0$, reject the null if $t_{\text{obs}} < -t_\alpha$.

iv. To test $H_0 : \mu = \mu_0$ vs. $H_1 : \mu \neq \mu_0$, reject the null if $|t_{\text{obs}}| > t_{\frac{\alpha}{2}}$.

2. One population (p)

a. If n is large ($n\hat{p} \geq 15$ and $n(1-\hat{p}) \geq 15$), $Z = \dfrac{\hat{p} - p}{\sqrt{\frac{p(1-p)}{n}}} \approx N(0,1)$.

i. The $(1-\alpha)100\%$ confidence interval for p is $[\hat{p} - Z_{\frac{\alpha}{2}} \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}, \hat{p} + Z_{\frac{\alpha}{2}} \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}]$.

ii. To test $H_0 : p = p_0$ vs. $H_1 : p > p_0$, reject the null if $Z_{\text{obs}} = \dfrac{\hat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}} > Z_\alpha$.

iii. To test $H_0 : p = p_0$ vs. $H_1 : p < p_0$, reject the null if $Z_{\text{obs}} < -Z_\alpha$.

iv. To test $H_0 : p = p_0$ vs. $H_1 : p \neq p_0$, reject the null if $|Z_{\text{obs}}| > Z_{\frac{\alpha}{2}}$.

3. One population (σ^2): When the $X_i' s$ come from a normal population, $\dfrac{(n-1)S^2}{\sigma^2}$ follows the χ^2 distribution with $n-1$ degrees of freedom.

a. To test $H_0 : \sigma^2 = \sigma_0^2$ vs. $H_1 : \sigma^2 > \sigma_0^2$, reject the null if $X^2_{\text{obs}} = \dfrac{(n-1)S^2_{\text{obs}}}{\sigma_0^2} > \chi^2_\alpha$.

b. To test $H_0 : \sigma^2 = \sigma_0^2$ vs. $H_1 : \sigma^2 < \sigma_0^2$, reject the null if $X^2_{\text{obs}} = \dfrac{(n-1)S^2_{\text{obs}}}{\sigma_0^2} < \chi^2_{1-\alpha}$.

c. To test $H_0 : \sigma^2 = \sigma_0^2$ vs. $H_1 : \sigma^2 \neq \sigma_0^2$, reject the null if $X^2_{\text{obs}} > \chi^2_{\frac{\alpha}{2}}$ or if $X^2_{\text{obs}} < \chi^2_{1-\frac{\alpha}{2}}$.

4. Two populations (μ_1, μ_2)

a. If σ_1 and σ_2 are known (or n_1 and $n_2 \geq 30$), $Z = \dfrac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$ follows the $N(0,1)$.

i. The $(1-\alpha)100\%$ confidence interval for $\mu_1 - \mu_2$ is $(\bar{X}_1 - \bar{X}_2) \pm Z_{\frac{\alpha}{2}}\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$.
ii. To test $H_0 : \mu_1 - \mu_2 = d_0$ vs. $H_1 : \mu_1 - \mu_2 > d_0$, reject the null if $Z_{\text{obs}} > Z_\alpha$.
iii. To test $H_0 : \mu_1 - \mu_2 = d_0$ vs. $H_1 : \mu_1 - \mu_2 < d_0$, reject the null if $Z_{\text{obs}} < -Z_\alpha$.
iv. To test $H_0 : \mu_1 - \mu_2 = d_0$ vs. $H_1 : \mu_1 - \mu_2 \neq d_0$, reject the null if $|Z_{\text{obs}}| > Z_{\frac{\alpha}{2}}$.

b. If σ_1 and σ_2 are unknown, $t = \dfrac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$ follows **approximately** $t-$distribution with k degrees of freedom, where k is approximated by

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

i. The $(1-\alpha)100\%$ confidence interval for $\mu_1 - \mu_2$ is $(\bar{X}_1 - \bar{X}_2) \pm t_{\frac{\alpha}{2}}\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$.
ii. To test $H_0 : \mu_1 - \mu_2 = d_0$ vs. $H_1 : \mu_1 - \mu_2 > d_0$, reject the null if $t_{\text{obs}} > t_\alpha$.
iii. To test $H_0 : \mu_1 - \mu_2 = d_0$ vs. $H_1 : \mu_1 - \mu_2 < d_0$, reject the null if $t_{\text{obs}} < -t_\alpha$.
iv. To test $H_0 : \mu_1 - \mu_2 = d_0$ vs. $H_1 : \mu_1 - \mu_2 \neq d_0$, reject the null if $|t_{\text{obs}}| > t_{\frac{\alpha}{2}}$.

c. If σ_1 and σ_2 are unknown **but can be assumed to be equal**, $t = \dfrac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{s_p^2}{n_1} + \frac{s_p^2}{n_2}}}$ follows $t-$distribution with $(n_1 + n_2 - 2)$ degrees of freedom, where $s_p^2 = \dfrac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{n_1 + n_2 - 2}$.

i. The $(1-\alpha)100\%$ confidence interval for $\mu_1 - \mu_2$ is $(\bar{X}_1 - \bar{X}_2) \pm t_{\frac{\alpha}{2}}\sqrt{\frac{s_p^2}{n_1} + \frac{s_p^2}{n_2}}$.
ii. To test $H_0 : \mu_1 - \mu_2 = d_0$ vs. $H_1 : \mu_1 - \mu_2 > d_0$, reject the null if $t_{\text{obs}} > t_\alpha$.
iii. To test $H_0 : \mu_1 - \mu_2 = d_0$ vs. $H_1 : \mu_1 - \mu_2 < d_0$, reject the null if $t_{\text{obs}} < -t_\alpha$.
iv. To test $H_0 : \mu_1 - \mu_2 = d_0$ vs. $H_1 : \mu_1 - \mu_2 \neq d_0$, reject the null if $|t_{\text{obs}}| > t_{\frac{\alpha}{2}}$.

d. For **paired** observations (the two samples are **NOT** independent), work with $d_i = x_i - y_i$.

i. To test $H_0 : \mu_D = d_0$ vs. $H_1 : \mu_D > d_0$, reject the null if $t_{\text{obs}} = \dfrac{\bar{d} - d_0}{S_D/\sqrt{n}} > t_{\alpha,(n-1)}$.
ii. To test $H_0 : \mu_D = d_0$ vs. $H_1 : \mu_D < d_0$, reject the null if $t_{\text{obs}} < -t_{\alpha,(n-1)}$.
iii. To test $H_0 : \mu_D = d_0$ vs. $H_1 : \mu_D \neq d_0$, reject the null if $|t_{\text{obs}}| > t_{\frac{\alpha}{2},(n-1)}$.

5. Two populations (p_1, p_2): If n_1 and n_2 are large ($n_i\hat{p}_i > 15$ and $n_i(1-\hat{p}_i) > 15$),

$$Z = \frac{\hat{p}_1 - \hat{p}_2 - (p_1 - p_2)}{\sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}} \approx N(0,1).$$

a. The $(1-\alpha)100\%$ confidence interval for $p_1 - p_2$ is $\left[(\hat{p}_1 - \hat{p}_2) \pm Z_{\frac{\alpha}{2}}\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}\right]$.

b. To test $H_0 : p_1 - p_2 = 0$ vs. $H_1 : p_1 - p_2 > 0$, reject the null if $Z_{\text{obs}} = \dfrac{(\hat{p}_1 - \hat{p}_2) - 0}{\sqrt{\frac{\hat{p}(1-\hat{p})}{n_1} + \frac{\hat{p}(1-\hat{p})}{n_2}}} > Z_\alpha$,

where, $\hat{p} = \dfrac{Y_1 + Y_2}{n_1 + n_2}$, the pooled sample proportion.

c. To test $H_0 : p_1 - p_2 = 0$ vs. $H_1 : p_1 - p_2 < 0$, reject the null if $Z_{\text{obs}} < -Z_\alpha$.

d. To test $H_0 : p_1 - p_2 = 0$ vs. $H_1 : p_1 - p_2 \neq 0$, reject the null if $|Z_{\text{obs}}| > Z_{\frac{\alpha}{2}}$.

Inferences Based on Two Samples

1. Exposure to dust at work can lead to lung disease later in life. One study measured the workplace exposure of tunnel construction workers. Part of the study compared 115 drill and blast workers with 220 outdoor concrete workers. Total dust exposure was measured in milligram years per cubic meter ($mg.y/m^3$). This part of the study aims to see if there is a difference in the dust exposure between these two groups of workers.

a. Formulate the appropriate null and alternative hypotheses.

b. Determine the appropriate test statistics to use and specify its distribution.

c. If the mean exposure for the drill and blast workers was 18.0 $mg.y/m^3$ with a standard deviation of 7.8 $mg.y/m^3$ and for the outdoor concrete workers, the corresponding values were 6.5 $mg.y/m^3$ and 3.4 $mg.y/m^3$, test the null hypothesis $H_0 : \mu_1 = \mu_2$ versus the alternative hypothesis you specified in part (b) using $\alpha = 0.01$

d. Construct a 99% confidence interval for $\mu_1 - \mu_2$.

2. Does increasing the amount of calcium in our diet reduce blood pressure? A randomized comparative experiment gave one group of 10 men a calcium supplement for 12 weeks. The control group of 11 men received a placebo that appeared identical to the calcium supplement. The experiment was double-blind. The table below gives the decrease in the blood pressure for each subject.

Calcium	7	-4	18	17	-3	-5	1	10	11	-2	
Placebo	-1	12	-1	-3	3	-5	5	2	-11	-1	-3

a. Determine the appropriate test statistics to use and specify its distribution.

b. Construct and interpret the 95% confidence interval for the true difference between the mean reduction in BP level for the Calcium group and that of the Placebo group ($\mu_1 - \mu_2$).

c. Formulate the appropriate null and alternative hypotheses.

d. Conduct a complete test of significance using $\alpha = 0.05$.

3. In the previous problem, suppose we can assume that $\sigma_1 = \sigma_2$.

a. Determine the appropriate test statistics to use and specify its distribution.

b. Construct and interpret the 95% confidence interval for the true difference between the mean reduction in BP level for the Calcium group and that of the Placebo group ($\mu_1 - \mu_2$).

c. Formulate the appropriate null and alternative hypotheses.

d. Conduct a complete test of significance using $\alpha = 0.05$.

4. To determine the effectiveness of an industrial safety program, the following data was collected (over a period of one year) on the average weekly loss of worker-hours due to accidents in 12 plants "before and after" the program was put into operations.

Plants	1	2	3	4	5	6	7	8	9	10	11	12
Before	37	45	12	72	54	34	26	13	39	125	79	26
After	28	46	18	59	43	29	24	15	35	120	75	24

a. Explain why the standard two-sample independent t-test is not appropriate for this study. How can we use this data? [4]

b. Formulate the most appropriate null and alternative hypotheses to test whether the safety program is effective. [3]

c. Specify the appropriate test statistic for this analysis. What is its distribution? [4]

d. Using a level of significance of $\alpha = 0.05$, define your rejection rule. [3]

e. Does the data provide sufficient evidence to say that the safety program is effective? Provide all necessary results to support your answer. [5]

5. In a winter of an epidemic flu, babies were surveyed by a well-known pharmaceutical company to determine if the company's new medicine was effective after two days. Among 120 babies who had the flu and were given the medicine, 29 were cured within two days. Among 280 babies who had the flu but were not given the medicine, 56 were cured within two days.

a. Formulate the most appropriate null and alternative hypotheses. Define clearly the parameters you use in your hypotheses. [4]

b. Specify the most appropriate test statistic and define your rejection rule using a level of significance of $\alpha = 0.05$. [4]

c. Test the company's claim of the effectiveness of the medicine. Write a practical conclusion. [9]

d. Compute the p-value of this sample. [5]

e. Construct and interpret a 99% confidence interval for $(p_1 - p_2)$. [9]

Homework problems:
Sec 9.2: (pp. 422-425) # 5, 7, 9, 11, 13, 15, 19.
Sec 9.3: (pp. 435-439) # 35, 37, 41, 47, 49.
Sec 9.4: (pp. 444-447) # 55, 57, 59, 61, 63, 65.
Supp.: (pp. 460-467) # 107, 109, 111, 113, 119, 121.

MATH 145 - ELEMENTARY STATISTICS HANDOUT #27

Review

1. Decide which inferential method is the most appropriate for each of the following practical research questions. Write the letter of the most appropriate statistical procedure next to each experiment or study question. Each procedure may be used more than once or not at all.

A) One-sample $t-$test for a mean.
B) One-sample $z-$test for a proportion.
C) One-sample χ^2-test for a variance.
D) Two-sample independent $t-$test for means.
E) Paired differences $t-$test.

F) Two-sample $z-$test for proportions.
G) C.I. for the difference of proportions.
H) C.I. for a mean.
I) C.I. for difference of means using independent samples.
J) C.I. for the mean of paired differences.

________1. On average, how much do lawyers earn annually?

________2. Is there statistical evidence that lawyers earn on average more than $100,000 annually?

________3. Is there statistical evidence lawyers earn on average more than accountants?

________4. Estimate the difference in the average annual income of lawyers and accountants.

________5. Policy makers want to know if the proportion of women in the United States who are in favor of legalizing marijuana use is more than 50%.

________6. Is there statistical evidence that there is a lower percentage of women than men who favor legalizing marijuana use?

________7. Is there evidence that men spend more than women for food?

________8. Estimate how much more or less men spend on average per month than women for food.

________9. Using only one group of men, test whether men spend more or less for the month of July than for the month of December.

________10. Using only one group of men, estimate how much more or less they spend for food for the month of July than for the month of December.

________11. Determine if the diameter of pipes produced by a machine is consistent enough to pass quality control.

________12. How much does a UWL student spend, on average, for the month of February?

________13. Using one set of students, estimate how much more or less they spend for the month of February than for the month of March.

________14. Is there evidence that male students spend more, on average, than female students for the month of February?

________15. Is there evidence that there is a higher percentage of female students than male students who are in favor of making UWL a none smoking zone?

________16. Inconsistency in product dimensions is a sign that the machine is not functioning properly. Is there evidence that a machine is not functioning properly?

1.H 2.A 3.D 4.I 5.B 6.F 7.D 8.I 9.E 10.J 11.C 12-16. *See fall 2012 sample exam key.*

Math 145 - Elementary Statistics Formula Sheet - Exam III

1. If $Z = \dfrac{\bar{X} - \mu}{\sigma/\sqrt{n}} \sim N(0,1)$.

a. The $(1-\alpha)100\%$ confidence interval for μ is $[\bar{X} - Z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}, \bar{X} + Z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}]$.

b. For a specified margin of error M, the required sample size is $n = \left(\dfrac{Z_{\frac{\alpha}{2}}\sigma}{M}\right)^2$.

c. To test $H_0 : \mu = \mu_0$ vs. $H_1 : \mu > \mu_0$, reject the null if $Z_{\text{obs}} = \dfrac{\bar{X}_{\text{obs}} - \mu}{\sigma/\sqrt{n}} > Z_\alpha$, or if the $p-$value$= P(Z \geq Z_{\text{obs}})$ is $\leq \alpha$.

d. To test $H_0 : \mu = \mu_0$ vs. $H_1 : \mu < \mu_0$, reject the null if $Z_{\text{obs}} < -Z_\alpha$, or if the $p-$value$= P(Z \leq Z_{\text{obs}})$ is $\leq \alpha$.

e. To test $H_0 : \mu = \mu_0$ vs. $H_1 : \mu \neq \mu_0$, reject the null if $|Z_{\text{obs}}| > Z_{\frac{\alpha}{2}}$, or if the $p-$value$= 2 * P(Z \geq |Z_{\text{obs}}|)$ is $\leq \alpha$.

2. If $t = \dfrac{\bar{X} - \mu}{s/\sqrt{n}} \sim t_{(n-1)}$

a. The $(1-\alpha)100\%$ confidence interval for μ is $[\bar{X} - t_{\frac{\alpha}{2}} \frac{s}{\sqrt{n}}, \bar{X} + t_{\frac{\alpha}{2}} \frac{s}{\sqrt{n}}]$.

b. To test $H_0 : \mu = \mu_0$ vs. $H_1 : \mu > \mu_0$, reject the null if $t_{\text{obs}} = \dfrac{\bar{X}_{\text{obs}} - \mu}{s/\sqrt{n}} > t_\alpha$.

c. To test $H_0 : \mu = \mu_0$ vs. $H_1 : \mu < \mu_0$, reject the null if $t_{\text{obs}} < -t_\alpha$.

d. To test $H_0 : \mu = \mu_0$ vs. $H_1 : \mu \neq \mu_0$, reject the null if $|t_{\text{obs}}| > t_{\frac{\alpha}{2}}$.

3. If $Z = \dfrac{\hat{p} - p}{\sqrt{\frac{p(1-p)}{n}}} \approx N(0,1)$.

a. The $(1-\alpha)100\%$ confidence interval for p is $[\hat{p} - Z_{\frac{\alpha}{2}}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}, \hat{p} + Z_{\frac{\alpha}{2}}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}]$.

b. For a specified margin of error M, the required sample size is $n = \dfrac{Z^2_{\alpha/2}[p(1-p)]}{M^2} \leq \dfrac{Z^2_{\alpha/2}(0.25)}{M^2}$.

c. To test $H_0 : p = p_0$ vs. $H_1 : p > p_0$, reject the null if $Z_{\text{obs}} = \dfrac{\hat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}} > Z_\alpha$.

d. To test $H_0 : p = p_0$ vs. $H_1 : p < p_0$, reject the null if $Z_{\text{obs}} < -Z_\alpha$.

e. To test $H_0 : p = p_0$ vs. $H_1 : p \neq p_0$, reject the null if $|Z_{\text{obs}}| > Z_{\frac{\alpha}{2}}$.

4. If $\dfrac{(n-1)S^2}{\sigma^2} \sim \chi^2_{(n-1)}$.

a. To test $H_0 : \sigma^2 = \sigma_0^2$ vs. $H_1 : \sigma^2 > \sigma_0^2$, reject the null if $X^2_{\text{obs}} = \dfrac{(n-1)S^2_{\text{obs}}}{\sigma_0^2} > \chi^2_\alpha$.

b. To test $H_0 : \sigma^2 = \sigma_0^2$ vs. $H_1 : \sigma^2 < \sigma_0^2$, reject the null if $X^2_{\text{obs}} = \dfrac{(n-1)S^2_{\text{obs}}}{\sigma_0^2} < \chi^2_{1-\alpha}$.

c. To test $H_0 : \sigma^2 = \sigma_0^2$ vs. $H_1 : \sigma^2 \neq \sigma_0^2$, reject the null if $X^2_{\text{obs}} > \chi^2_{\frac{\alpha}{2}}$ or if $X^2_{\text{obs}} < \chi^2_{1-\frac{\alpha}{2}}$.

5. If $Z = \dfrac{(\bar{X_1} - \bar{X_2}) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \sim N(0,1)$.

- **a.** The $(1-\alpha)100\%$ confidence interval for $\mu_1 - \mu_2$ is $(\bar{X_1} - \bar{X_2}) \pm Z_{\frac{\alpha}{2}} \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$.
- **b.** To test $H_0 : \mu_1 - \mu_2 = d_0$ vs. $H_1 : \mu_1 - \mu_2 > d_0$, reject the null if $Z_{\text{obs}} > Z_\alpha$.
- **c.** To test $H_0 : \mu_1 - \mu_2 = d_0$ vs. $H_1 : \mu_1 - \mu_2 < d_0$, reject the null if $Z_{\text{obs}} < -Z_\alpha$.
- **d.** To test $H_0 : \mu_1 - \mu_2 = d_0$ vs. $H_1 : \mu_1 - \mu_2 \neq d_0$, reject the null if $|Z_{\text{obs}}| > Z_{\frac{\alpha}{2}}$.

6. If $t = \dfrac{(\bar{X_1} - \bar{X_2}) - (\mu_1 - \mu_2)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} \approx t_{(k^*)}$, where $k^* \approx \dfrac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\frac{1}{n_1-1}\left(\frac{s_1^2}{n_1}\right)^2 + \frac{1}{n_2-1}\left(\frac{s_2^2}{n_2}\right)^2}$.

- **a.** The $(1-\alpha)100\%$ confidence interval for $\mu_1 - \mu_2$ is $(\bar{X_1} - \bar{X_2}) \pm t_{\frac{\alpha}{2}} \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$.
- **b.** To test $H_0 : \mu_1 - \mu_2 = d_0$ vs. $H_1 : \mu_1 - \mu_2 > d_0$, reject the null if $t_{\text{obs}} > t_\alpha$.
- **c.** To test $H_0 : \mu_1 - \mu_2 = d_0$ vs. $H_1 : \mu_1 - \mu_2 < d_0$, reject the null if $t_{\text{obs}} < -t_\alpha$.
- **d.** To test $H_0 : \mu_1 - \mu_2 = d_0$ vs. $H_1 : \mu_1 - \mu_2 \neq d_0$, reject the null if $|t_{\text{obs}}| > t_{\frac{\alpha}{2}}$.

7. If $t = \dfrac{(\bar{X_1} - \bar{X_2}) - (\mu_1 - \mu_2)}{s_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \sim t_{(n_1+n_2-2)}$, where $s_p = \sqrt{\dfrac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{n_1+n_2-2}}$.

- **a.** The $(1-\alpha)100\%$ confidence interval for $\mu_1 - \mu_2$ is $(\bar{X_1} - \bar{X_2}) \pm t_{\frac{\alpha}{2}} * s_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$.
- **b.** To test $H_0 : \mu_1 - \mu_2 = d_0$ vs. $H_1 : \mu_1 - \mu_2 > d_0$, reject the null if $t_{\text{obs}} > t_\alpha$.
- **c.** To test $H_0 : \mu_1 - \mu_2 = d_0$ vs. $H_1 : \mu_1 - \mu_2 < d_0$, reject the null if $t_{\text{obs}} < -t_\alpha$.
- **d.** To test $H_0 : \mu_1 - \mu_2 = d_0$ vs. $H_1 : \mu_1 - \mu_2 \neq d_0$, reject the null if $|t_{\text{obs}}| > t_{\frac{\alpha}{2}}$.

8. If $Z = \dfrac{\hat{p}_1 - \hat{p}_2 - (p_1 - p_2)}{\sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}} \approx N(0,1)$.

- **a.** The $(1-\alpha)100\%$ confidence interval for $p_1 - p_2$ is $\left[(\hat{p}_1 - \hat{p}_2) \pm Z_{\frac{\alpha}{2}} \sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}\right]$.
- **b.** To test $H_0 : p_1 - p_2 = 0$ vs. $H_1 : p_1 - p_2 > 0$, reject the null if $Z_{\text{obs}} = \dfrac{(\hat{p}_1 - \hat{p}_2) - 0}{\sqrt{\frac{\hat{p}(1-\hat{p})}{n_1} + \frac{\hat{p}(1-\hat{p})}{n_2}}} > Z_\alpha$,

  where, $\hat{p} = \dfrac{Y_1 + Y_2}{n_1 + n_2}$, the pooled sample proportion.
- **c.** To test $H_0 : p_1 - p_2 = 0$ vs. $H_1 : p_1 - p_2 < 0$, reject the null if $Z_{\text{obs}} < -Z_\alpha$.
- **d.** To test $H_0 : p_1 - p_2 = 0$ vs. $H_1 : p_1 - p_2 \neq 0$, reject the null if $|Z_{\text{obs}}| > Z_{\frac{\alpha}{2}}$.

Chi-square (χ^2) Tests

- **Common Uses of the χ^2−test.**

  1. Testing Goodness-of-fit.

  2. Testing Equality of Several Proportions.

  3. Homogeneity Test.

  4. Testing Independence.

- **Testing Goodness-of-fit of Data from Multinomial Experiment.**

  Properties of the Multinomial Experiment.

  1. The experiment consists of n identical trials.

  2. There are k possible outcomes to each trial. These possible outcomes are sometimes called *classes* or *categories.*

  3. The probabilities of the k outcomes, denoted by $p_1, p_2, \ldots, p_k$, remain the same from trial to trial, and $p_1 + p_2 + \cdots + p_k = 1$.

  4. The trials are independent.

  5. The random variables of interest are the *cell counts*, $n_1, n_2, \ldots, n_k$, of the number of observations that fall in each of the k categories. (note: $n_1 + n_2 + \cdots + n_k = n$).

	Cat 1	Cat 2	$\cdots$	Cat k
Observed	n_1	n_2	$\cdots$	n_k
Expected	$n \cdot p_1$	$n \cdot p_2$	$\cdots$	$n \cdot p_k$

- **Chi-Square Statistic**. The *Chi-square statistic* is a measure of how much the observed cell counts diverge from the expected cell counts. In this case, we use the one-dimensional χ^2−statistic because only one factor is being investigated. The formula for this statistic is

$$\begin{aligned} X^2 &= \sum_{i=1}^{k} \frac{(\textit{observed count} - \textit{expected count})^2}{\textit{expected count}} \qquad (4) \\ &= \sum_{i=1}^{k} \frac{(O-E)^2}{E} \qquad (5) \\ &= \sum_{i=1}^{k} \frac{(n_i - n \cdot p_i)^2}{n \cdot p_i} \sim \chi^2_{(k-1)}. \qquad (6) \end{aligned}$$

  This X^2 statistic follows **approximately** the χ^2 distribution with $k-1$ degrees of freedom.

- **Conditions Required for χ^2−Test to be valid:**

  1. *A multinomial experiment has been conducted. This is generally satisfied by taking a random sample from the population of interest.*

  2. *The sample size n is large enough so that the expected count for each cell is at least 5. Sometimes two or more categories are collapsed together to form a bigger category with higher expected count.*

- **Null Hypothesis and Alternative Hypothesis.**

  1. Null Hypothesis: $H_o : p_1 = p_{o1}, p_2 = p_{o2}, \ldots, p_k = p_{ok}$ [Note: $p_{o1} + p_{o2} + \cdots + p_{ok} = 1$]

  2. Alternative Hypothesis: H_1 : The null hypothesis is not true.

- **Examples:**

1. *Nature* (Sept. 1993) reported on a study of animal and plant species "hotspots" in Great Britain. A hotspot is defined as a 10-km^2 area that is species-rich, that is, is heavily populated by the species of interest. Analogously, a coldspot is a 10-km^2 area that is species-poor. The table below gives the number of butterfly hotspots and the number of butterfly coldspots in a sample of 2,588 10-km^2 areas. In theory, 5% should be butterfly hotspots and 5% coldspots, while the remaining areas (90%) are neutral. Test the theory using $\alpha = 0.01$.

	Hotspots	Coldspots	Neutral Areas	Total
Observed (Expected)	123 ( )	147 ( )	2318 ( )	2588

2. **Car Crashes and Age Brackets.** Among drivers who have had a car crash in the last year, 88 are randomly selected and categorized by age, with the results listed in the accompanying table. If all ages have the same crash rate, we would expect (because of the age distribution of licensed drivers) the given categories to have 16%, 44%, 27%, and 13% of the subjects, respectively. At the 0.05 level of significance, test the claim that the distribution of crashes conforms to the distribution of ages. Does any age group appear to have a disproportionate number of crashes?

Age	Under 25	25 - 44	45 - 64	Over 64
No. of Crashes	36 ( )	21 ( )	12 ( )	19 ( )

3. **Checking Normality.** The table below shows the number of values in a sample of size $n = 200$ observations falling in each category. At the 0.05 level of significance, test whether the sample can be reasonably assumed to have come from the Standard Normal distribution.

$z-$values	$z < -1$	$-1 \leq z < -.5$	$-.5 \leq z < 0$	$0 \leq z < .5$	$.5 \leq z < 1$	$z > 1$
No. of Obs.	30	38	43	38	20	31

- Testing Equality of Several Proportions.

  1. Null Hypothesis: $H_0 : p_1 = p_2 = \cdots = p_k$

  2. Alternative Hypothesis: H_1 : Not all are equal.

  3. Test Statistic: The *Two-Factor Chi-square Statistic* (X^2) for Two-Way ($r \times c$) Contingency Table. The formula for the statistic is very similar to the one we have before and is given below:

$$X^2 = \sum_{\text{all cells}} \frac{(\textit{observed count} - \textit{expected count})^2}{\textit{expected count}} \tag{7}$$

$$= \sum_{i=1}^{r}\sum_{j=1}^{c} \frac{(O_{ij} - E_{ij})^2}{E_{ij}} = \sum_{i=1}^{r}\sum_{j=1}^{c} \frac{(n_{ij} - r_i * \hat{p}_j)^2}{r_i * \hat{p}_j} \tag{8}$$

$$= \sum_{i=1}^{r}\sum_{j=1}^{c} \frac{(n_{ij} - (r_i * c_j)/n)^2}{(r_i * c_j)/n} \sim \chi^2_{(r-1)(c-1)}. \tag{9}$$

  where, r_i and c_j are the total of the ith row and jth column, respectively. This X^2 statistic follows **approximately** the χ^2 distribution with $(r-1)(c-1)$ degrees of freedom.

  4. Conditions Required for χ^2−Test to be valid:

  a. *The n observed counts are a random sample from the population of interest.*

  b. *The sample size n is large enough so that the expected count for each cell is at least 5. Sometimes two or more categories are collapsed together to form a bigger category with higher expected count.*

- Examples.

  1. Quitting Smoking. The accompanying table summarizes successes and failures when subjects used different methods in trying to stop smoking. The determination of smoking or not smoking was made five months after the treatment was begun, and the data are based on results from the Centers for Disease Control and Prevention. Use a 0.05 level of significance to test the claim that success is independent of the method used. If someone wants to stop smoking, does the choice of the method make a difference?

	Nicotine Gum	Nicotine Patch	Nicotine Inhaler
Smoking	191	263	95
No smoking	59	57	27

2. **Survey Refusals and Age Bracket.** A study of people who refused to answer survey questions provided the randomly selected sample data shown in the table. At the 0.01 significance level, test the claim that the cooperation of the subject (response or refusal) is independent of the age category. Does any particular age group appear to be particularly uncooperative?

Age	18-21	22-29	30-39	40-49	50-59	60 and over
Responded	73	255	245	136	138	202
Refused	11	20	33	16	27	49

- Testing for Homogeneity (In a Two-Way Contingency Tables).

  1. Null Hypothesis: $H_0 : p_{1j} = p_{2j} = \cdots = p_{kj}$, $j = 1, 2, \ldots, J$
  2. Alternative Hypothesis: $H_1 : H_0$ is not true.
  3. Test Statistic: The *chi-square statistic* X^2.

- Example.

  1. A company packages a particular product in cans of three different sizes, each one using a different production line. Most cans conform to specifications, but a quality control engineer has identified the following reasons for non-conformance: Blemish on can, crack in can, improper pull tab location, pull tab missing, and other. A sample of non-conforming units is selected from each of the three lines, and each unit is categorized according to reason for nonconformity, resulting in the contingency table below. Using $\alpha = 0.05$, test the hypothesis that the distributions of non-conformance of the three production lines are homogeneous.

Production	Blemish	Crack	Location	Missing	Other	Total
Line 1	34	65	17	21	13	150
Line 2	23	52	25	19	6	125
Line 3	32	28	16	14	10	100
Total	89	145	58	54	29	375

- Testing for Independence (In a Two-Way Contingency Tables).

  1. Null Hypothesis: Factor A is independent of factor B.

  2. Alternative Hypothesis: Factors A and B are not independent.

  3. Test Statistic: The same *chi-square statistic* X^2.

- Example.

  1. Background Music. Market researchers know that background music can influence the mood and purchasing behavior of customers. One study in a supermarket in Northern Ireland compared three treatments: no music, French accordion music, and Italian string music. Under each condition, the researchers recorded the numbers of bottles of French, Italian, and other wine purchased. The two-way table that summarizes the data are given below. Test the hypothesis that background music and purchasing behavior (type of wine purchased) are independent. Use $\alpha = 0.05$.

Wine	French music	Italian music	No music
French	39	30	30
Italian	1	19	11
Other	35	35	43

Homework problems:

Section 13.2: pp. 733-736; # 5, 7, 9, 11, 17.
Section 13.3: pp. 745-751; # 23, 25, 27, 29, 31.
Supplement: pp. 753-758; # 45, 47, 49, 51.

Analysis of Variance (ANOVA)

Two Types of Study:

1. **Observational Study – observes individuals and measures variables of interest but does not attempt to influence the responses.**
   - This type of study can also establish association between a factor and the response variable.
2. **Designed Experiments – deliberately imposes some treatment on individuals in order to observe their responses.**
   - This type of study can also establish cause and effect (between a factor and the respon

Do # 10.5 on page 480.

Examples of Designed Experiment

Example 1 : Consider the problem of comparing the effectiveness of 3 kinds of diets (A, B, C). Forty males and 80 females were included in the study and were randomly divided into 3 groups of 40 people each. Then a different diet is assigned to each group. The body weights of these 120 people were measured before and after the study period of 8 weeks and the differences were computed.

Example 2 : In a classic study, described by F. Yates in the *The Design and Analysis of Factorial Experiments*, the effect on oat yield was compared for three different varieties of oats (A, B, C) and four different concentrations of manure (0, 0.2, 0.4, and 0.6 cwt per acre).

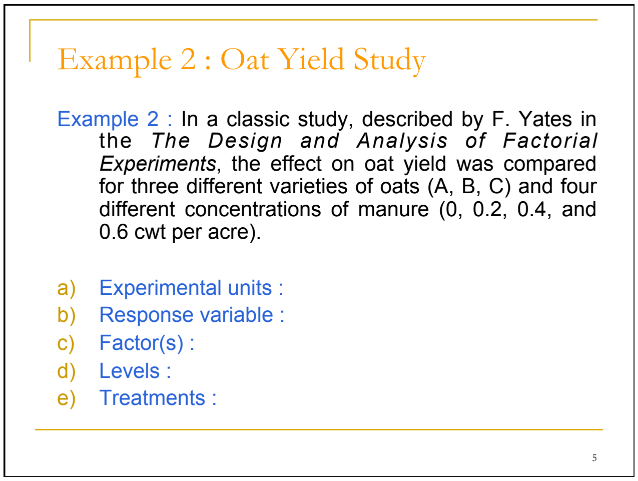
Terminologies in Experiments

- Experimental Units – These are the individuals on which the experiment is done.
  - Subjects – human beings.
- Response variables – Measurement of interest.
- Factors – Things that might affect the response variable (explanatory variables). {new drug}
- Levels of a factor – {different concentration of the new drug; no drug, 10 mg, 25 mg, etc.}
- Treatment – A combination of levels of factors.
- Repetition – putting more than one experimental units in a treatment.

Example 1 : Diet Study

Example 1 : Consider the problem of comparing the effectiveness of 3 kinds of diets (A, B, C). Forty males and 80 females were included in the study and were randomly divided into 3 groups of 40 people each. Then a different diet is assigned to each group. The body weights of these 120 people were measured before and after the study period of 8 weeks and the differences were computed.

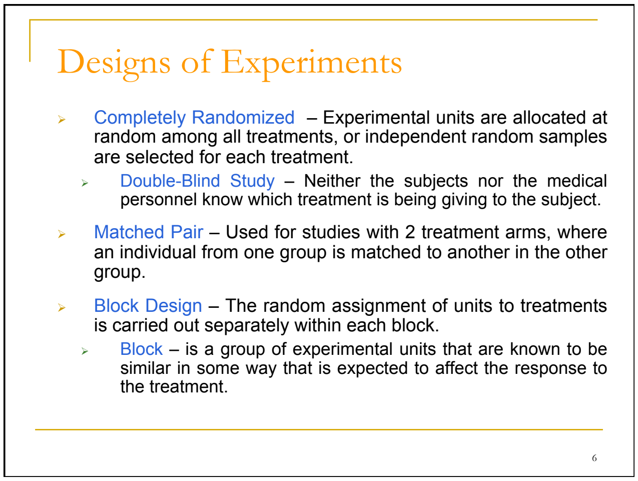
a) Experimental units :
b) Response variable :
c) Factor(s) :
d) Levels :
e) Treatments :



Example 2 : Oat Yield Study

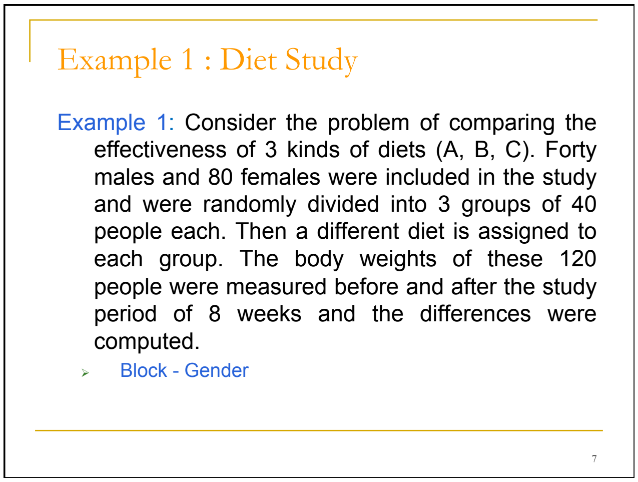
Example 2 : In a classic study, described by F. Yates in the *The Design and Analysis of Factorial Experiments*, the effect on oat yield was compared for three different varieties of oats (A, B, C) and four different concentrations of manure (0, 0.2, 0.4, and 0.6 cwt per acre).

a) Experimental units :
b) Response variable :
c) Factor(s) :
d) Levels :
e) Treatments :



Designs of Experiments

- Completely Randomized – Experimental units are allocated at random among all treatments, or independent random samples are selected for each treatment.
  - Double-Blind Study – Neither the subjects nor the medical personnel know which treatment is being giving to the subject.
- Matched Pair – Used for studies with 2 treatment arms, where an individual from one group is matched to another in the other group.
- Block Design – The random assignment of units to treatments is carried out separately within each block.
  - Block – is a group of experimental units that are known to be similar in some way that is expected to affect the response to the treatment.



Example 1 : Diet Study

Example 1: Consider the problem of comparing the effectiveness of 3 kinds of diets (A, B, C). Forty males and 80 females were included in the study and were randomly divided into 3 groups of 40 people each. Then a different diet is assigned to each group. The body weights of these 120 people were measured before and after the study period of 8 weeks and the differences were computed.

- Block - Gender

Example of Studies

1. A company investigated the effects of selling price on sales volume of one of its products. Three selling prices (55 cents, 60 cents, 65 cents) were studied. Twelve communities throughout the United States, of approximately equal size and similar socioeconomic characteristics, were selected and the treatments were assigned to them at random, such that each treatment was given to four communities.

a. Is this an example of *experimental* or *observational* study? ______________________________

b. What study design was employed in this study? ______________________________

c. Identify the response variable. ______________________________

d. Identify the Experimental units. ______________________________

e. Identify the Factor and the levels. ______________________________

f. Identify the treatments. ______________________________

2. In the previous example, suppose the company investigated the effects of selling price and type of promotional campaign on sales volume of one of its products. Three selling prices (55 cents, 60 cents, 65 cents) were studied, as were two types of promotional campaigns (radio advertising, newspaper advertising). Twelve communities throughout the United States, of approximately equal size and similar socioeconomic characteristics, were selected and the treatments were assigned to them at random, such that each treatment was given to two communities.

a. What study design was employed in this study? ______________________________

b. Identify the response variable. ______________________________

c. Identify the Factors and their levels. ______________________________

d. How many treatments are there? ______________________________

3. An analyst studied the effects of family income (under \$15,000, \$15,000 – \$29,999, \$30,000 – \$49,999, \$50,000 and more) and stage in the life cycle of the family (stages 1, 2, 3, 4) on amount spent on appliance purchases in the last 5 years. The analyst selected 20 families with the required income and life-cycle characteristics for each of the "treatment" classes for this study, yielding 320 families for the entire study.

a. Is this an example of *experimental* or *observational* study? ______________________________

b. Identify the response variable. ______________________________

c. Identify the Experimental units. ______________________________

d. Identify the Factors and their levels. ______________________________

e. How many treatments are there? ______________________________

4. A medical investigator studied the relationship between the response to three blood pressure lowering drug types for hypertensive males and females. The investigator selected 30 adult males and 30 adult females and randomly assigned 10 males and 10 females to each of the three drug types, yielding 60 total subjects.

a. Is this an example of *experimental* or *observational* study? ______________________________

b. What study design was employed in this study? ______________________________

c. Identify the response variable. ______________________________

d. Identify the Experimental units. ______________________________

e. Identify the Factors and their levels. ______________________________

f. How many treatments are there? ______________________________

Analysis of Variance - ANOVA

I. Testing Equality of Several Means.

$H_0 : \mu_1 = \mu_2 = \cdots = \mu_k$ vs. H_1 : Not all are equal

ANOVA Table

Source	DF	Sum of Squares	Mean Square	F
Treatment	$k-1$	$SSTR = \sum_{i=1}^{k} n_i(\bar{x}_i - \bar{x})^2$	$MSTR = \dfrac{SSTR}{k-1}$	$F_{\text{obs}} = \dfrac{MSTR}{MSE}$
Error	$n-k$	$SSE = \sum_{i=1}^{k} (n_i - 1)s_i^2$	$MSE = \dfrac{SSE}{n-k}$	
Total	$n-1$	$SSTO = \sum_{i=1}^{k}\sum_{j=1}^{n_i}(x_{ij} - \bar{x})^2$		

Note: $SSTO = SSTR + SSE$. Under H_0, $F_{\text{obs}} \sim f_{(k-1,n-k)}$.

II. Assumptions.

1. The samples are independent SRS from each population.

2. The populations are assumed to be normal.

3. The standard deviations are equal. [Rule of thumb: *The largest standard deviation should not be more than twice the smallest standard deviation.*]

III. Idea. ANOVA is based on separating the total variation observed in the data into two parts: variation *among the group means* and variation *within groups*. If the variation among groups is large relative to the variation within groups, we have evidence against the null hypothesis.

IV. Examples.

1. **Productivity Improvement.** An economist compiled data on productivity improvements last year for a sample of firms producing electronic computing equipment. The firms were classified according to the level of their average expenditures for research and development in the past three years (low, moderate, high). The results of the study is given in the table below (productivity improvement is measured on a scale from 0 to 100). Assuming that ANOVA model is appropriate, test the hypothesis that the level of expenditures for research and development has no effect on the productivity improvements. Use $\alpha = 0.05$.

	1	2	3	4	5	6	7	8	9	10	11	12
Low	7.6	8.2	6.8	5.8	6.9	6.6	6.3	7.7	6.0			
Moderate	6.7	8.1	9.4	8.6	7.8	7.7	8.9	7.9	8.3	8.7	7.1	8.4
High	8.5	9.7	10.1	7.8	9.6	9.5						

2. Rehabilitation Therapy. A rehabilitation center researcher was interested in examining the relationship between physical fitness prior to surgery of persons undergoing corrective knee surgery and time required in physical therapy until successful rehabilitation. Patient records in the rehabilitation center were examined, and 24 male subjects ranging in age from 18 to 30 years who had undergone similar corrective knee surgery during the past year were selected for the study. The number of days required for successful completion of physical therapy and the prior physical fitness status (below average, average, above average) for each patient follow.

	1	2	3	4	5	6	7	8	9	10
Below Average	29	42	38	40	43	40	30	42		
Average	30	35	39	28	31	31	29	35	29	33
Above Average	26	32	21	20	23	22				

a. Construct the ANOVA table.

b. Test whether or not the mean number of days required for successful rehabilitation is that same for the three fitness groups. Use $\alpha = 0.05$ level of significance.

3. Analysis of variance methods are often used in clinical trials where the goal is to assess the effectiveness of one or more treatments for a particular medical condition. One such study compared three treatments for dandruff and a placebo. The treatments were 1% pyrithione zinc shampoo (PyrI), the same shampoo but with instructions to shampoo two times (PyrII), 2% ketoconazole shampoo (Keto), and a placebo shampoo (Placebo). After six weeks of treatment, eight sections of the scalp were examined and given a score that measured the amount of scalp flaking on a 0 to 10 scale. The response variable was the sum of these eight scores. An analysis of the baseline flaking measurements indicated that randomization of patients to treatments was successful in that no differences were found between the groups. At baseline there were 112 subjects in each of the three treatment groups and 28 subjects in the Placebo group. During the clinical trial, 3 dropped out from the PyrII group and 6 from the Keto group. No patients dropped out of the other two groups. A summary of the data is given below. Using $\alpha = 0.01$, test the hypothesis that there is no difference in the effectiveness of the four treatments.

Treatments	n	$\bar{x}$	s
Pyr I	112	17.39	1.142
Pyr II	109	17.20	1.352
Keto	106	16.03	0.931
Placebo	28	29.39	1.595

Homework problems:
Section 10.1: pp. 480-481; # 1,3, 5, 7, 9, 11, 13.
Section 10.2: pp. 492-497; # 17, 19, 21, 29, 35.

Correlation and Simple Linear Regression

- Regression analysis is a statistical tool that utilizes the relation between two or more quantitative variables so that one variable can be predicted from the other, or others.

- **Some Examples:**

  1. Waistline and Weight.

  2. SAT score and First year college GPA.

  3. Number of customers and Revenue.

  4. Family income and Family expenditures.

- **Functional Relation vs. Statistical Relation between two variables.**

  – A *functional relation* between two variables is expressed by a mathematical formula. If X is the *independent variable* and Y the *dependent variable*, a functional relation is of the form:

  $$Y = f(X).$$

  That is, given a particular value of X, we get only one corresponding value Y.

  1. For example, let x denote the number of printer cartridges that you order over the internet. Suppose each cartridge costs $40 and there is a fixed shipping fee of $10, determine the total cost y of ordering x cartridges.

  – A *statistical relation*, unlike a functional relation, is not a perfect one. If X is the *independent variable* and Y the *dependent variable*, a statistical relation is of the form:

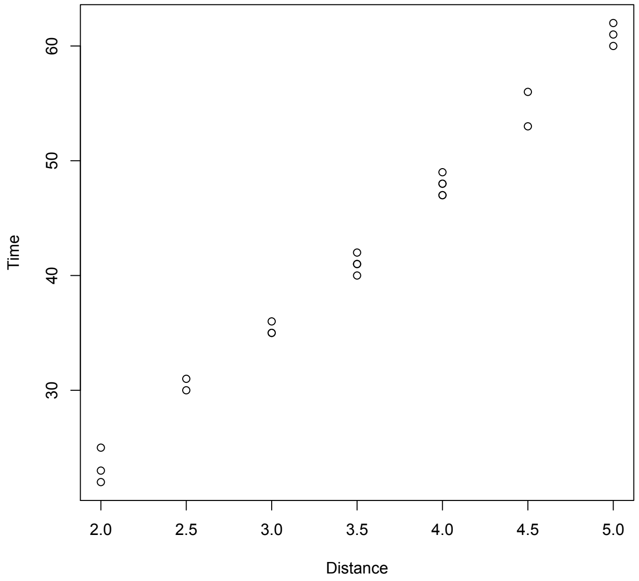
  $$Y = f(x) + \epsilon.$$

  In such cases, we call X an *explanatory variable* and Y a *response variable.*

  1. For example, let x denote the distance that a person plans to jog and y the time that it will take this person to finish it. Consider his 22 jogging distances and times from last month shown in the table below. If this person plans to jog for 5.5 miles tomorrow, predict how long it will take him to finish the run.

	1	2	3	4	5	6	7	8	9	10	11
Distance (x)	2	2	3	3	2	2.5	2.5	3	3.5	3.5	4
Time (y)	25	22	35	36	23	30	31	35	41	40	49
	12	13	14	15	16	17	18	19	20	21	22
Distance (x)	4	4	4	4.5	4.5	5	5	5	3.5	3.5	4
Time (y)	47	48	48	56	53	62	60	61	42	41	47

- **Scatterplots.** A *scatterplot* (or *scatter diagram*) is a graph in which the paired (x, y) sample data are plotted with a horizontal $x-$axis and a vertical $y-$axis. Each individual (x, y) pair is plotted as a single point. Scatterplots are useful as they usually display the relationship between two quantitative variables.
  - Always plot the explanatory variable on the $x-$axis, while the response variable on the $y-$axis.
  - In examining a scatterplot, look for an overall pattern showing the **form**, **direction**, and **strength** of the relationship, and then for outliers or other deviations from this pattern.
    - Form - linear or not.
    - Direction - positive or negative association.
    - Strength - how close the points lie to the general pattern (usually a line).



- **Correlation.** A *correlation* exists between two variables when one of them is related to the other in some way.
- **Linear Correlation.** The *linear correlation coefficient* r measures the strength of the linear relationship between the paired $x-$ and $y-$quantitative values in a sample.

$$\begin{aligned} r &= \frac{\sum(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum(x_i-\bar{x})^2\sum(y_i-\bar{y})^2}} && (10) \\ &= \frac{n\sum x_i y_i - (\sum x_i)(\sum y_i)}{\sqrt{n\sum x_i^2-(\sum x_i)^2}\sqrt{n\sum y_i^2-(\sum y_i)^2}} && (11) \\ &= \frac{1}{n-1}\sum_{i=1}^{n}\left(\frac{x_i-\bar{x}}{s_x}\right)\left(\frac{y_i-\bar{y}}{s_y}\right) && (12) \\ &= \frac{SS_{xy}}{\sqrt{SS_{xx}SS_{yy}}} && (13) \end{aligned}$$

where,

$$\begin{aligned} SS_{xx} &= \Sigma x^2 - \frac{1}{n}(\Sigma x)^2 = (n-1)s_x^2 && (14) \\ SS_{yy} &= \Sigma y^2 - \frac{1}{n}(\Sigma y)^2 = (n-1)s_y^2 && (15) \\ SS_{xy} &= \Sigma xy - \frac{1}{n}(\Sigma x)(\Sigma y) && (16) \end{aligned}$$

- **Tree Circumference and Height.** Listed below are the circumferences (in feet) and the heights (in feet) of trees in Marshall, Minnesota (based on data from "Tree Measurements" by Stanley Rice, *American Biology Teacher*.

x (circ)	1.8	1.9	1.8	2.4	5.1	3.1	5.5
y (height)	21.0	33.5	24.6	40.7	73.2	24.9	40.4
x (circ)	5.1	8.3	13.7	5.3	4.9	3.7	3.8
y (height)	45.3	53.5	93.8	64.0	62.7	47.2	44.3

1. Determine the values of SS_{xx}, SS_{yy}, and SS_{xy}.

2. Determine the correlation coefficient r.

3. What can you say about the linear relationship of x and y? Is it a strong linear relationship.

- The table below displays data on age (in years) and price (in $100)for a sample of 11 cars.

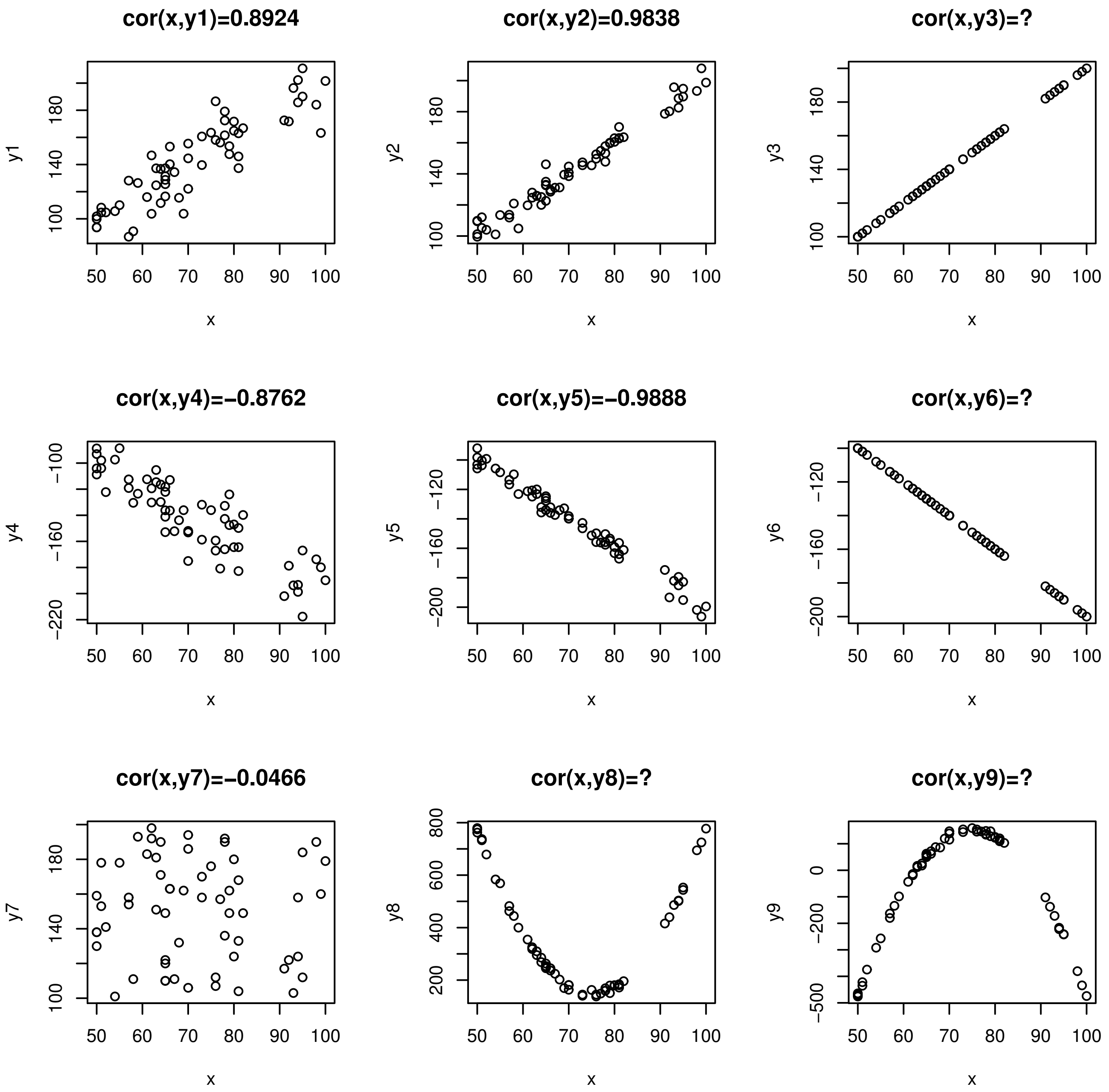
Age (x)	5	4	6	6	5	5	6	6	2	7	7
Price (y)	85	102	70	80	89	98	66	90	169	68	50

1. Determine the values of SS_{xx}, SS_{yy}, and SS_{xy}.

2. Determine the correlation coefficient r.

3. What can you say about the linear relationship of x and y? Is it a strong linear relationship.

- Sample plots with correlation values



- Some guidelines in interpreting r.

Value of $\|r\|$	Strength of linear relationship
If $\|r\| \geq .95$	Very Strong
If $.85 \leq \|r\| < .95$	Strong
If $.65 \leq \|r\| < .85$	Moderately to Strong
If $.45 \leq \|r\| < .65$	Moderate
If $.25 \leq \|r\| < .45$	Weak
If $\|r\| < .25$	Very weak/Close to none

- **Recommended problems:**
  Section 11.5: pp. 585-588; # 67, 69, 71.

Simple Linear Regression

- If X is the *independent variable* and Y the *dependent variable*, a statistical relation is of the form:

$$Y = f(X) + \epsilon.$$

  In such cases, we call X an *explanatory variable* and Y a *response variable.*

- In a *simple linear regression* model, the response variable Y is linearly related to one *explanatory* variable X. That is,

$$Y_i = (a + bx_i) + \epsilon_i. \qquad i = 1, 2, \ldots, n.$$

  Assumptions:

  1. The mean of ϵ_i is 0 and the variance of ϵ_i is σ^2.

  2. The random errors ϵ_i are uncorrelated.

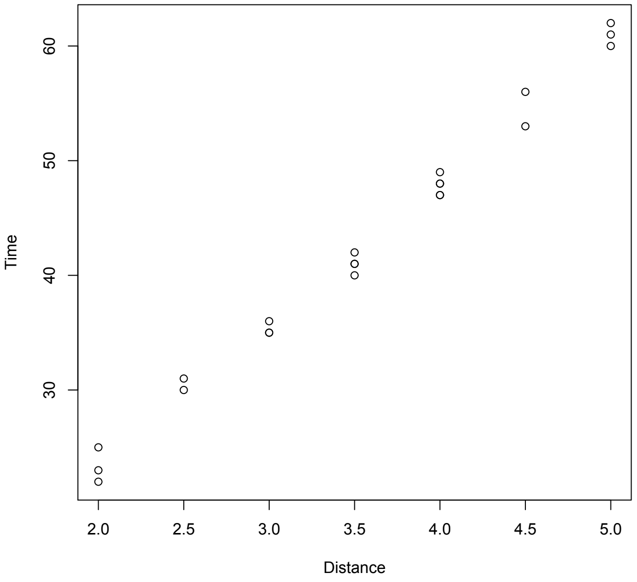
  3. a and b are parameters.

  4. x_i is a known constant.

- For example, let x denote the distance that a person plans to jog and y the time that it will take this person to finish it. Consider his 22 jogging distances and times from last month shown in the table below. If this person plans to jog for 5.5 miles tomorrow, predict how long it will take him to finish the run.

	1	2	3	4	5	6	7	8	9	10	11
Distance (x)	2	2	3	3	2	2.5	2.5	3	3.5	3.5	4
Time (y)	25	22	35	36	23	30	31	35	41	40	49

	12	13	14	15	16	17	18	19	20	21	22
Distance (x)	4	4	4	4.5	4.5	5	5	5	3.5	3.5	4
Time (y)	47	48	48	56	53	62	60	61	42	41	47



- **Equation of the Least-Squares Regression Line .** Suppose we have data on an explanatory variable x and a response variable y for n individuals. The means and standard deviations of the sample data are $\bar{x}$ and s_x for x and $\bar{y}$ and s_y for y, and the correlation between x and y is r. The equation of the least-squares regression line of y on x is

$$\hat{y} = \hat{a} + \hat{b}x$$

with *slope*

$$\hat{b} = \frac{SS_{xy}}{SS_{xx}} = \frac{(\Sigma xy) - \frac{1}{n}(\Sigma x)(\Sigma y)}{(\Sigma x^2) - \frac{1}{n}(\Sigma x)^2} = r\frac{s_y}{s_x} \tag{17}$$

and *intercept*

$$\hat{a} = \bar{y} - \hat{b}\bar{x} \tag{18}$$

- **Practice.**

1. The table below displays data on age (in years) and price (in $100) for a sample of 11 cars.

Age (x)	5	4	6	6	5	5	6	6	2	7	7
Price (y)	85	102	70	80	89	98	66	90	169	68	50

a. Determine the values of SS_{xx}, SS_{yy}, and SS_{xy}.

b. Determine the correlation coefficient r.

c. What can you say about the linear relationship of x and y? Is it a strong linear relationship.

d. Determine the regression line.

e. Estimate the expected value of a car that is 3 years old.

2. Tree Circumference and Height. Listed below are the circumferences (in feet) and the heights (in feet) of trees in Marshall, Minnesota (based on data from "Tree Measurements" by Stanley Rice, *American Biology Teacher*.

x (circ)	1.8	1.9	1.8	2.4	5.1	3.1	5.5	5.1	8.3	13.7	5.3	4.9	3.7	3.8
y (height)	21.0	33.5	24.6	40.7	73.2	24.9	40.4	45.3	53.5	93.8	64.0	62.7	47.2	44.3

a. Determine the correlation coefficient r.

b. What can you say about the linear relationship of x and y? Is it a strong linear relationship.

c. Determine the regression line.

d. Estimate the expected height of a tree that has a circumference of 10 feet.

3. A criminologist studying the relationship between population density and robbery rate in medium-sized U.S. cities collected the following data for a random sample of 16 cities; X is the population density of the city (number of people per unit area), and Y is the robbery rate last year (number of robberies per 100,000 people). Assume that the simple linear regression model is appropriate.

i	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
X_i	59	49	75	54	78	56	60	82	69	83	88	94	47	65	89	70
Y_i	209	180	195	192	215	197	208	189	213	201	214	212	205	186	200	204

a. Determine the correlation coefficient r.

b. What can you say about the linear relationship of x and y? Is it a strong linear relationship.

c. Determine the regression line.

d. Estimate the expected robbery rate (no. of robberies per 100,000 people) of a city with population density of $x = 90$.

4. To study the relationship between age x (in years) and body fat y, 18 adults (with ages from 33 to 48) were randomly selected. A summary of the data obtained is given below:

$$SS_{xx} = 2970,\ SS_{xy} = 1998,\ SS_{yy} = 1683,\ \sum x_i = 834 \text{ and } \sum y_i = 499$$

a. Determine the correlation coefficient.

b. Determine the regression line.

c. Using the regression line that you obtained in part (b), estimate the body fat of a 50-year-old person.

d. Based on your result in part (b), what can you expect will happen to someone's body fat as he/she ages by one year?

e. Based on your result in part (a), what can you say about the degree of linear relationship of age and body fat? Explain your answer.

f. Do you think it was appropriate to use linear regression to predict the body fat of a 50-year-old person? Explain your answer.

- **Recommended problems:**
  Section 11.1: (pp. 553-554) # 5, 7.
  Section 11.2: (pp. 561-566) # 11, 13, 15, 17.

Simple Linear Regression

- The response variable Y is linearly related to one explanatory variable X. That is,

$$y_i = (a + bx_i) + \epsilon_i. \qquad i = 1, 2, \ldots, n.$$

Assumptions:

1. The mean of ϵ_i is 0 and the variance of ϵ_i is σ^2.

2. The random errors ϵ_i are uncorrelated.

3. a and b are parameters.

4. x_i is a known constant.

- **Equation of the Least-Squares Regression Line .** Suppose we have data on an explanatory variable x and a response variable y for n individuals. The means and standard deviations of the sample data are $\bar{x}$ and s_x for x and $\bar{y}$ and s_y for y, and the correlation between x and y is r. The equation of the least-squares regression line of y on x is

$$\hat{y} = \hat{a} + \hat{b}x$$

with *slope*

$$\hat{b} = \frac{SS_{xy}}{SS_{xx}} = \frac{(\Sigma xy) - \frac{1}{n}(\Sigma x)(\Sigma y)}{(\Sigma x^2) - \frac{1}{n}(\Sigma x)^2} = r\frac{s_y}{s_x} \tag{19}$$

and *intercept*

$$\hat{a} = \bar{y} - \hat{b}\bar{x} \tag{20}$$

- The **fitted** (or **predicted**) **values** $\hat{y}_i$'s are obtained by successively substituting the x_i's into the estimated regression line: $\hat{y} = \hat{a} + \hat{b}x_i$. The **residuals** are the vertical deviations, $e_i = y_i - \hat{y}_i$, from the estimated line.

- The **error sum of squares**, (equivalently, **residual sum of squares**) denoted by SSE, is

$$\begin{aligned} SSE = \sum e_i^2 = \sum (y_i - \hat{y}_i)^2 &= \sum [y_i - (\hat{a} + \hat{b}x_i)]^2 & (21) \\ = SS_{yy} - \hat{b}SS_{xy} &= \sum y_i^2 - \hat{a}\sum y_i - \hat{b}\sum x_i y_i & (22) \end{aligned}$$

and the estimate of σ^2 is

$$\hat{\sigma}^2 = s^2 = \frac{SSE}{n-2} = \frac{(n-1)s_y^2(1-r^2)}{n-2}. \tag{23}$$

- The **coefficient of determination**, denoted by r^2, is the amount of the variation in y that is explained by the regression line.

$$\begin{aligned} r^2 &= 1 - \frac{SSE}{SST}, \qquad \text{where, } SST = SS_{yy} = \sum (y_i - \bar{y})^2 & (24) \\ &= \frac{SST - SSE}{SST} = \frac{\text{explained variation}}{\text{total variation}} & (25) \end{aligned}$$

- **Inference for b.**

1. Test statistic:
$$\frac{\hat{b} - b}{SE_{\hat{b}}} \sim t_{(n-2)} \qquad SE_{\hat{b}} = \frac{\hat{b}\sqrt{1-r^2}}{r\sqrt{n-2}} = \frac{s}{s_x\sqrt{n-1}} = \frac{s}{\sqrt{SS_{xx}}}$$

2. Confidence Interval: $\hat{b} \pm t_{\alpha/2} SE_{\hat{b}}$

- **Mean Response of Y at a specified value x^*, $(\mu_{Y|x^*})$.**

  1. Point Estimate. For a specific value $x*$, the estimate of the **mean** value of Y is given by

  $$\hat{\mu}_{Y|x^*} = \hat{a} + \hat{b}x^*$$

  2. Confidence Interval. For a specific value $x*$, the $(1-\alpha)100\%$ confidence interval for $\mu_{Y|x^*}$ is given by

  $$\hat{\mu}_{Y|x^*} \pm t_{\alpha/2;(n-2)} SE_{\hat{\mu}}$$

  where, $SE_{\hat{\mu}} = s\sqrt{\dfrac{1}{n} + \dfrac{(x^* - \bar{x})^2}{(n-1)s_x^2}}$, and $s = s_y\sqrt{\dfrac{(n-1)(1-r^2)}{n-2}}$

- **Prediction of Y at a specified value x^*.**

  1. Point Estimate. For a specific value $x*$, the predicted value of Y is given by

  $$\hat{y} = \hat{a} + \hat{b}x^*$$

  2. Prediction Interval. For a specific value $x*$, the $(1-\alpha)100\%$ prediction interval is given by

  $$\hat{y} \pm t_{\alpha/2;(n-2)} SE_{\hat{y}}$$

  where, $SE_{\hat{y}} = s\sqrt{1 + \dfrac{1}{n} + \dfrac{(x^* - \bar{x})^2}{(n-1)s_x^2}}$, and $s = s_y\sqrt{\dfrac{(n-1)(1-r^2)}{n-2}}$

- **Blood Pressure Measurements.** To see if there is a linear relationship between the Systolic and Diastolic blood pressure of a person, the following measurements from 14 randomly selected individuals were recorded.

Person	1	2	3	4	5	6	7
Systolic (x)	138	130	135	140	120	125	120
Diastolic (y)	82	91	100	100	80	90	80

Person	8	9	10	11	12	13	14
Systolic (x)	130	130	144	143	140	130	150
Diastolic (y)	80	80	98	105	85	70	100

  1. Determine the correlation coefficient r.

  2. What can you say about the linear relationship of x and y? Is it a strong linear relationship.

  3. Determine the coefficient of determination r^2. Explain the meaning of this quantity in this context.

  4. Find the residual of the first person and determine if this residual is an outlier using the estimate of σ.

5. Determine the regression line.

6. Test $H_0 : b = 0$ vs. $H_1 : b \neq 0$

7. Construct a 95% confidence interval for b.

8. Predicted the diastolic blood pressure for a person with a systolic reading of 122.

9. Estimate to the mean diastolic blood pressure (μ_y) for people with a systolic reading of 122.

10. Construct a 95% confidence interval for μ_y for people with a systolic reading of 122.

- **Homework Problem :** Consider the following data of 10 production runs of a certain manufacturing company.

Production run	1	2	3	4	5	6	7	8	9	10
Lot size (x)	30	20	60	80	40	50	60	30	70	60
Man-Hours (y)	73	50	128	170	87	108	135	69	148	132

1. Determine the correlation coefficient r.

2. What can you say about the linear relationship of x and y? Is it a strong linear relationship.

3. Determine the coefficient of determination r^2. Explain the meaning of this quantity in this context.

4. Determine the regression line.

5. Using $\alpha = 0.01$, test $H_0 : b = 0$ vs. $H_1 : b \neq 0$

6. Construct and interpret a 99% confidence interval for b.

7. Find an estimate for the **mean** number of man-hours ($\hat{\mu}_{Y|x^*}$) required to produce a lot size 100.

8. Construct a 95% confidence interval for the **mean** number of man-hours ($\mu_{Y|x^*}$) required to produce a lot size 100.

9. Predict the number of man-hours ($\hat{y}$) required to produce a lot size 100.

10. Construct a 95% prediction interval for the number of man-hours ($\hat{y}$) required to produce a lot size 100.

- **Homework problems:**
  Section 11.3: (pp. 569-570) # 35, 37.
  Section 11.4: (pp. 575-578) # 47, 51, 53.
  Section 11.5: (pp. 586) # 79.
  Section 11.6: (pp. 593-596) # 93, 95.

Review

1. Decide which inferential method is the most appropriate for each of the following practical research questions. Write the letter of the most appropriate statistical procedure next to each experiment or study question. Each procedure may be used more than once or not at all.

A) Analysis of Variance (ANOVA).
B) Chi-square test.
C) Linear Regression.

________1. A study was done to determine if there is a difference in the mean income of people from five different religious affiliation.

________2. A study was done to determine if there is a relationship between the heights of married couples.

________3. A study was done to determine where marital status (divorced, married and never divorced) and religious affiliation (A,B,C,D,none) are independent.

________4. A farmer wants to know if the type of fertilizer (A, B, C) has an effect on the corn yield. He plans to measure the corn yield based on the weight (in kg) of the total harvest from each type of fertilizer.

________5. A company that produces a certain fertilizer for corn compiled data from 500 farmers who are using their product. For each farmer, the company recorded how much fertilizer (in kg) they use per acre of corn field and how much corn they harvest per acre (in kg). The company wants to see if there is a relationship between the amount of fertilizer used per acre and the corn yield per acre.

________6. Suppose in the previous problem, the company decided to divide the farmers into 3 groups according to the amount of fertilizer they use (small, moderate, high). The company wants to know if there is a difference in the corn yield per acre (in kg) between the three groups.

________7. Suppose in the previous problem, the company also divided the farmers into 3 groups according to their corn yield per acre (low, moderate, high). The company wants to know if there is a relationship between amount of fertilizer used and corn yield per acre.

________8. The school district of La Crosse wants to see if marital status of parents (single, married, divorced) has an effect on the child's ACT score.

________9. The Department of Public Instruction of Wisconsin wants to know if there is a relationship between poverty rate (percentage of households with total income below the poverty level in the school district) and the high school drop out rate (percentage of students who drop out of high school). One hundred randomly selected school districts in Wisconsin were selected and for each school district the poverty rate and high school drop out rate were recorded.

________10. Seven hundred fifty patients were randomly assigned to receive one of four types of "Frontier medicine" treatment: (1) prayer, (2) MIT (Musis, imagery, and touch), (3) prayer and MIT, and (4) standard care (no prayer and no MIT). After six months, the patients were evaluated and determined if they suffered a major adverse cardiovascular event (e.g. heart attack) or not.

1.A 2.C 3.B 4.A 5.C 6.A 7.B 8.A 9.C 10.B

- **Chi-Square Statistic**. The *chi-square statistic* is a measure of how much the observed cell counts diverge from the expected cell counts. The formula for the statistic is

  1. **Goodness-of-fit Test**.

$$X^2 = \sum_{i=1}^{k} \frac{(\textit{observed count} - \textit{expected count})^2}{\textit{expected count}} \tag{26}$$

$$= \sum_{i=1}^{k} \frac{(n_i - n * p_i)^2}{n * p_i} \sim \chi^2_{k-1}. \tag{27}$$

  This X^2 statistic follows **approximately** the χ^2 distribution with $k-1$ degrees of freedom.

  2. **Testing equality of several proportions, independence, and homogeneity**.

$$X^2 = \sum_{\text{all cells}} \frac{(\textit{observed count} - \textit{expected count})^2}{\textit{expected count}} \tag{28}$$

$$= \sum_{i=1}^{r}\sum_{j=1}^{c} \frac{(O_{ij} - E_{ij})^2}{E_{ij}} = \sum_{i=1}^{r}\sum_{j=1}^{c} \frac{(n_{ij} - r_i * \hat{p}_j)^2}{r_i * \hat{p}_j} \tag{29}$$

$$= \sum_{i=1}^{r}\sum_{j=1}^{c} \frac{(n_{ij} - r_i * c_j/n)^2}{r_i * c_j/n} \sim \chi^2_{(r-1)(c-1)}. \tag{30}$$

  where, r_i and c_j are the total of the ith row and jth column, respectively. This X^2 statistic follows **approximately** the χ^2 distribution with $(r-1)(c-1)$ degrees of freedom.

- **ANOVA Table**

Source	DF	Sum of Squares	Mean Square	F
Groups	$k-1$	$SSG = \sum_{i=1}^{k} n_i(\bar{x}_i - \bar{x})^2$	$MSG = \frac{SSG}{k-1}$	$F_{\text{obs}} = \frac{MSG}{MSE}$
Error	$n-k$	$SSE = \sum_{i=1}^{k}(n_i - 1)s_i^2$	$MSE = \frac{SSE}{n-k}$	
Total	$n-1$	$SST = \sum_{i=1}^{k}\sum_{j=1}^{n_i}(x_{ij} - \bar{x})^2$		

  Note: $SST = SSG + SSE$. Under H_0, $F_{\text{obs}} \sim f_{(k-1,n-k)}$.

- **Linear Correlation.** The *linear correlation coefficient* r measures the strength of the linear relationship between the paired $x-$ and $y-$quantitative values in a sample.

$$r = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum(x_i - \bar{x})^2 \sum(y_i - \bar{y})^2}} \tag{31}$$

$$= \frac{SS_{xy}}{\sqrt{SS_{xx}SS_{yy}}} \tag{32}$$

  where,

$$SS_{xx} = \Sigma x^2 - \frac{1}{n}(\Sigma x)^2 = (n-1)s_x^2 \tag{33}$$

$$SS_{yy} = \Sigma y^2 - \frac{1}{n}(\Sigma y)^2 = (n-1)s_y^2 \tag{34}$$

$$SS_{xy} = \Sigma xy - \frac{1}{n}(\Sigma x)(\Sigma y) \tag{35}$$

- **Equation of the Least-Squares Regression Line .** Suppose we have data on an explanatory variable x and a response variable y for n individuals. The means and standard deviations of the sample data are $\bar{x}$ and s_x for x and $\bar{y}$ and s_y for y, and the correlation between x and y is r. The equation of the least-squares regression line of y on x is

$$\hat{y} = \hat{a} + \hat{b}x$$

with *slope*

$$\hat{b} = \frac{SS_{xy}}{SS_{xx}} = \frac{(\Sigma xy) - \frac{1}{n}(\Sigma x)(\Sigma y)}{(\Sigma x^2) - \frac{1}{n}(\Sigma x)^2} = r\frac{s_y}{s_x} \tag{36}$$

and *intercept*

$$\hat{a} = \bar{y} - \hat{b}\bar{x} \tag{37}$$

- The **fitted** (or **predicted**) **values** $\hat{y}_i$'s are obtained by successively substituting the x_i's into the estimated regression line: $\hat{y} = \hat{a} + \hat{b}x_i$. The **residuals** are the vertical deviations, $e_i = y_i - \hat{y}_i$, from the estimated line.

- The **error sum of squares**, (equivalently, **residual sum of squares**) denoted by SSE, is

$$\begin{aligned} SSE &= \sum e_i^2 = \sum (y_i - \hat{y}_i)^2 = \sum [y_i - (\hat{a} + \hat{b}x_i)]^2 & (38) \\ &= SS_{yy} - \hat{b}SS_{xy} = \sum y_i^2 - \hat{a}\sum y_i - \hat{b}\sum x_i y_i & (39) \end{aligned}$$

and the estimate of σ^2 is

$$\hat{\sigma}^2 = s^2 = \frac{SSE}{n-2} = \frac{SS_{yy} - \hat{b}SS_{xy}}{n-2} = \frac{(n-1)s_y^2(1-r^2)}{n-2}. \tag{40}$$

- The **coefficient of determination**, denoted by r^2, is the amount of the variation in y that is explained by the regression line.

$$\begin{aligned} r^2 &= 1 - \frac{SSE}{SST}, \qquad \text{where, } SST = SS_{yy} = \sum (y_i - \bar{y})^2 & (41) \\ &= \frac{SST - SSE}{SST} = \frac{\text{explained variation}}{\text{total variation}} = (r)^2 & (42) \end{aligned}$$

- **Inference for b.**

  1. Test statistic:
  $$\frac{\hat{b} - b}{SE_{\hat{b}}} \sim t_{(n-2)} \qquad SE_{\hat{b}} = \frac{\hat{b}\sqrt{1-r^2}}{r\sqrt{n-2}} = \frac{s}{s_x\sqrt{n-1}} = \frac{s}{\sqrt{SS_{xx}}} \tag{43}$$

  2. Confidence Interval: $\hat{b} \pm t_{\alpha/2}SE_{\hat{b}}$

- **Mean Response of Y at a specified value x^*, $(\mu_{Y|x^*})$.**

  1. Point Estimate. For a specific value $x*$, the estimate of the **mean** value of Y is given by

$$\hat{\mu}_{Y|x^*} = \hat{a} + \hat{b}x^* \tag{44}$$

- **Prediction of Y at a specified value x^*.**

  1. Point Estimate. For a specific value $x*$, the predicted value of Y is given by

$$\hat{y} = \hat{a} + \hat{b}x^* \tag{45}$$

- **Some Properties of Probability**.

  1. $P(A \cup B) = P(A) + P(B) - P(A \cap B)$. **2.** $P(A|B) = \dfrac{P(A \cap B)}{P(B)}$.

- **Descriptive Statistics.** Let $\{x_1, x_2, \dots, x_n\}$ be a sample of size n, then

  1. $\bar{x} = \dfrac{1}{n} \displaystyle\sum_{i=1}^{n} x_i$. **2.** $s^2 = \dfrac{1}{n-1} \displaystyle\sum_{i=1}^{n} (x_i - \bar{x})^2 = \dfrac{\sum x^2 - \frac{1}{n}(\sum x)^2}{n-1} = \dfrac{SS_{xx}}{n-1}$.

- **Discrete Random Variable.**

  1. $\mu_X = E(X) = \displaystyle\sum_{i=1}^{k} x_i p_i = x_1 p_1 + x_2 p_1 + x_3 p_3 + \cdots + x_k p_k$.

  2. $\sigma_X^2 = Var(X) = \displaystyle\sum_{i=1}^{k} (x_i - \mu)^2 p_i = (x_1 - \mu)^2 p_1 + (x_2 - \mu)^2 p_2 + \cdots + (x_k - \mu)^2 p_k$.

- **Binomial Distribution.** $X \sim \text{bin}(n, p) \quad \Rightarrow \quad P(X = x) = \dbinom{n}{x} p^x (1-p)^{n-x}$, for $x = 0, 1, \dots, n$.

  1. $E(X) = np$ **2.** $V(X) = np(1-p)$.

- **(Continuous) Uniform Distribution.** $X \sim \text{unif}[c, d] \quad \Rightarrow \quad f(x) = \dfrac{1}{d-c}$, for $c \le x \le d$.

  1. $E(X) = \frac{1}{2}(c + d)$ **2.** $V(X) = \frac{1}{12}(d - c)^2$.

- If $Z = \dfrac{\bar{X} - \mu}{\sigma/\sqrt{n}} \sim N(0, 1)$.

  1. The $(1 - \alpha)100\%$ confidence interval for μ is $[\bar{X} - Z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}, \bar{X} + Z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}]$.

  2. For a specified margin of error M, the required sample size is $n = \left(\dfrac{Z_{\frac{\alpha}{2}} \sigma}{M} \right)^2$.

  3. To test $H_0 : \mu = \mu_0$ vs. $H_1 : \mu > \mu_0$, reject the null if $Z_{\text{obs}} = \dfrac{\bar{X}_{\text{obs}} - \mu}{\sigma/\sqrt{n}} > Z_\alpha$, or if the $p-$value$= P(Z \ge Z_{\text{obs}})$ is $\le \alpha$.

  4. To test $H_0 : \mu = \mu_0$ vs. $H_1 : \mu < \mu_0$, reject the null if $Z_{\text{obs}} < -Z_\alpha$, or if the $p-$value$= P(Z \le Z_{\text{obs}})$ is $\le \alpha$.

  5. To test $H_0 : \mu = \mu_0$ vs. $H_1 : \mu \ne \mu_0$, reject the null if $|Z_{\text{obs}}| > Z_{\frac{\alpha}{2}}$, or if the $p-$value$= 2 * P(Z \ge |Z_{\text{obs}}|)$ is $\le \alpha$.

- If $t = \dfrac{\bar{X} - \mu}{s/\sqrt{n}} \sim t_{(n-1)}$

  1. The $(1 - \alpha)100\%$ confidence interval for μ is $[\bar{X} - t_{\frac{\alpha}{2}} \frac{s}{\sqrt{n}}, \bar{X} + t_{\frac{\alpha}{2}} \frac{s}{\sqrt{n}}]$.

  2. To test $H_0 : \mu = \mu_0$ vs. $H_1 : \mu > \mu_0$, reject the null if $t_{\text{obs}} = \dfrac{\bar{X}_{\text{obs}} - \mu}{s/\sqrt{n}} > t_\alpha$.

  3. To test $H_0 : \mu = \mu_0$ vs. $H_1 : \mu < \mu_0$, reject the null if $t_{\text{obs}} < -t_\alpha$.

  4. To test $H_0 : \mu = \mu_0$ vs. $H_1 : \mu \ne \mu_0$, reject the null if $|t_{\text{obs}}| > t_{\frac{\alpha}{2}}$.

- If $\dfrac{(n-1)S^2}{\sigma^2} \sim \chi^2_{(n-1)}$.

  1. To test $H_0 : \sigma^2 = \sigma_0^2$ vs. $H_1 : \sigma^2 > \sigma_0^2$, reject the null if $X^2_{\text{obs}} = \dfrac{(n-1)S^2_{\text{obs}}}{\sigma_0^2} > \chi^2_\alpha$.

  2. To test $H_0 : \sigma^2 = \sigma_0^2$ vs. $H_1 : \sigma^2 < \sigma_0^2$, reject the null if $X^2_{\text{obs}} = \dfrac{(n-1)S^2_{\text{obs}}}{\sigma_0^2} < \chi^2_{1-\alpha}$.

  3. To test $H_0 : \sigma^2 = \sigma_0^2$ vs. $H_1 : \sigma^2 \ne \sigma_0^2$, reject the null if $X^2_{\text{obs}} > \chi^2_{\frac{\alpha}{2}}$ or if $X^2_{\text{obs}} < \chi^2_{1-\frac{\alpha}{2}}$.

- If $Z = \dfrac{\hat{p} - p}{\sqrt{\frac{p(1-p)}{n}}} \approx N(0,1)$.

  1. The $(1-\alpha)100\%$ confidence interval for p is $[\hat{p} - Z_{\frac{\alpha}{2}}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}, \hat{p} + Z_{\frac{\alpha}{2}}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}]$.

  2. For a specified margin of error M, the required sample size is $n = \dfrac{Z^2_{\alpha/2}[p(1-p)]}{M^2} \leq \dfrac{Z^2_{\alpha/2}(0.25)}{M^2}$.

  3. To test $H_0 : p = p_0$ vs. $H_1 : p > p_0$, reject the null if $Z_{\text{obs}} = \dfrac{\hat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}} > Z_\alpha$.

  4. To test $H_0 : p = p_0$ vs. $H_1 : p < p_0$, reject the null if $Z_{\text{obs}} < -Z_\alpha$.

  5. To test $H_0 : p = p_0$ vs. $H_1 : p \neq p_0$, reject the null if $|Z_{\text{obs}}| > Z_{\frac{\alpha}{2}}$.

- If $Z = \dfrac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \sim N(0,1)$.

  1. The $(1-\alpha)100\%$ confidence interval for $\mu_1 - \mu_2$ is $(\bar{X}_1 - \bar{X}_2) \pm Z_{\frac{\alpha}{2}}\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$.

  2. To test $H_0 : \mu_1 - \mu_2 = d_0$ vs. $H_1 : \mu_1 - \mu_2 > d_0$, reject the null if $Z_{\text{obs}} > Z_\alpha$.

  3. To test $H_0 : \mu_1 - \mu_2 = d_0$ vs. $H_1 : \mu_1 - \mu_2 < d_0$, reject the null if $Z_{\text{obs}} < -Z_\alpha$.

  4. To test $H_0 : \mu_1 - \mu_2 = d_0$ vs. $H_1 : \mu_1 - \mu_2 \neq d_0$, reject the null if $|Z_{\text{obs}}| > Z_{\frac{\alpha}{2}}$.

- If $t = \dfrac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} \approx t_{(k)}$, where $k \approx \dfrac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\frac{1}{n_1-1}\left(\frac{s_1^2}{n_1}\right)^2 + \frac{1}{n_2-1}\left(\frac{s_2^2}{n_2}\right)^2}$.

  1. The $(1-\alpha)100\%$ confidence interval for $\mu_1 - \mu_2$ is $(\bar{X}_1 - \bar{X}_2) \pm t_{\frac{\alpha}{2}}\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$.

  2. To test $H_0 : \mu_1 - \mu_2 = d_0$ vs. $H_1 : \mu_1 - \mu_2 > d_0$, reject the null if $t_{\text{obs}} > t_\alpha$.

  3. To test $H_0 : \mu_1 - \mu_2 = d_0$ vs. $H_1 : \mu_1 - \mu_2 < d_0$, reject the null if $t_{\text{obs}} < -t_\alpha$.

  4. To test $H_0 : \mu_1 - \mu_2 = d_0$ vs. $H_1 : \mu_1 - \mu_2 \neq d_0$, reject the null if $|t_{\text{obs}}| > t_{\frac{\alpha}{2}}$.

- If $t = \dfrac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{s_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \sim t_{(n_1+n_2-2)}$, where $s_p = \sqrt{\dfrac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{n_1 + n_2 - 2}}$.

  1. The $(1-\alpha)100\%$ C. I. for $\mu_1 - \mu_2$ is $(\bar{X}_1 - \bar{X}_2) \pm t_{\frac{\alpha}{2}} * s_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$.

  2. To test $H_0 : \mu_1 - \mu_2 = d_0$ vs. $H_1 : \mu_1 - \mu_2 > d_0$, reject the null if $t_{\text{obs}} > t_\alpha$.

  3. To test $H_0 : \mu_1 - \mu_2 = d_0$ vs. $H_1 : \mu_1 - \mu_2 < d_0$, reject the null if $t_{\text{obs}} < -t_\alpha$.

  4. To test $H_0 : \mu_1 - \mu_2 = d_0$ vs. $H_1 : \mu_1 - \mu_2 \neq d_0$, reject the null if $|t_{\text{obs}}| > t_{\frac{\alpha}{2}}$.

- If $Z = \dfrac{\hat{p}_1 - \hat{p}_2 - (p_1 - p_2)}{\sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}} \approx N(0,1)$.

  1. The $(1-\alpha)100\%$ confidence interval for $p_1 - p_2$ is $\left[(\hat{p}_1 - \hat{p}_2) \pm Z_{\frac{\alpha}{2}}\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}\right]$.

  2. To test $H_0 : p_1 - p_2 = 0$ vs. $H_1 : p_1 - p_2 > 0$, reject the null if $Z_{\text{obs}} = \dfrac{(\hat{p}_1 - \hat{p}_2) - 0}{\sqrt{\frac{\hat{p}(1-\hat{p})}{n_1} + \frac{\hat{p}(1-\hat{p})}{n_2}}} > Z_\alpha$,

  where, $\hat{p} = \dfrac{Y_1 + Y_2}{n_1 + n_2}$, the pooled sample proportion.

  3. To test $H_0 : p_1 - p_2 = 0$ vs. $H_1 : p_1 - p_2 < 0$, reject the null if $Z_{\text{obs}} < -Z_\alpha$.

  4. To test $H_0 : p_1 - p_2 = 0$ vs. $H_1 : p_1 - p_2 \neq 0$, reject the null if $|Z_{\text{obs}}| > Z_{\frac{\alpha}{2}}$.

- **Chi-Square Statistic**. The *chi-square statistic* is a measure of how much the observed cell counts diverge from the expected cell counts. The formula for the statistic is

  1. **Goodness-of-fit Test**.

$$X^2 = \sum_{i=1}^{k} \frac{(\textit{observed count} - \textit{expected count})^2}{\textit{expected count}} \tag{46}$$

$$= \sum_{i=1}^{k} \frac{(n_i - n * p_i)^2}{n * p_i} \sim \chi^2_{k-1}. \tag{47}$$

  This X^2 statistic follows **approximately** the χ^2 distribution with $k-1$ degrees of freedom.

  2. **Testing equality of several proportions, independence, and homogeneity**.

$$X^2 = \sum_{\text{all cells}} \frac{(\textit{observed count} - \textit{expected count})^2}{\textit{expected count}} \tag{48}$$

$$= \sum_{i=1}^{r}\sum_{j=1}^{c} \frac{(O_{ij} - E_{ij})^2}{E_{ij}} = \sum_{i=1}^{r}\sum_{j=1}^{c} \frac{(n_{ij} - r_i * \hat{p}_j)^2}{r_i * \hat{p}_j} \tag{49}$$

$$= \sum_{i=1}^{r}\sum_{j=1}^{c} \frac{(n_{ij} - r_i * c_j/n)^2}{r_i * c_j/n} \sim \chi^2_{(r-1)(c-1)}. \tag{50}$$

  where, r_i and c_j are the total of the ith row and jth column, respectively. This X^2 statistic follows **approximately** the χ^2 distribution with $(r-1)(c-1)$ degrees of freedom.

- **ANOVA Table**

Source	DF	Sum of Squares	Mean Square	F
Groups	$k-1$	$SSG = \sum_{i=1}^{k} n_i(\bar{x}_i - \bar{x})^2$	$MSG = \frac{SSG}{k-1}$	$F_{\text{obs}} = \frac{MSG}{MSE}$
Error	$n-k$	$SSE = \sum_{i=1}^{k} (n_i - 1)s_i^2$	$MSE = \frac{SSE}{n-k}$	
Total	$n-1$	$SST = \sum_{i=1}^{k}\sum_{j=1}^{n_i} (x_{ij} - \bar{x})^2$		

Note: $\bar{x} = (\sum n_i\bar{x}_i)/(\sum n_i)$ and $SST = SSG + SSE$. Under H_0, $F_{\text{obs}} \sim f_{(k-1,n-k)}$.

- **Linear Correlation.** The *linear correlation coefficient* r measures the strength of the linear relationship between the paired $x-$ and $y-$quantitative values in a sample.

$$r = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum(x_i - \bar{x})^2 \sum(y_i - \bar{y})^2}} \tag{51}$$

$$= \frac{SS_{xy}}{\sqrt{SS_{xx}SS_{yy}}} \tag{52}$$

  where,

$$SS_{xx} = \Sigma x^2 - \frac{1}{n}(\Sigma x)^2 = (n-1)s_x^2 \tag{53}$$

$$SS_{yy} = \Sigma y^2 - \frac{1}{n}(\Sigma y)^2 = (n-1)s_y^2 \tag{54}$$

$$SS_{xy} = \Sigma xy - \frac{1}{n}(\Sigma x)(\Sigma y) \tag{55}$$

- **Equation of the Least-Squares Regression Line .** Suppose we have data on an explanatory variable x and a response variable y for n individuals. The means and standard deviations of the sample data are $\bar{x}$ and s_x

for x and $\bar{y}$ and s_y for y, and the correlation between x and y is r. The equation of the least-squares regression line of y on x is

$$\hat{y} = \hat{a} + \hat{b}x$$

with *slope*

$$\hat{b} = \frac{SS_{xy}}{SS_{xx}} = \frac{(\Sigma xy) - \frac{1}{n}(\Sigma x)(\Sigma y)}{(\Sigma x^2) - \frac{1}{n}(\Sigma x)^2} = r\frac{s_y}{s_x} \tag{56}$$

and *intercept*

$$\hat{a} = \bar{y} - \hat{b}\bar{x} \tag{57}$$

- The **fitted** (or **predicted**) **values** $\hat{y}_i$'s are obtained by successively substituting the x_i's into the estimated regression line: $\hat{y} = \hat{a} + \hat{b}x_i$. The **residuals** are the vertical deviations, $e_i = y_i - \hat{y}_i$, from the estimated line.

- The **error sum of squares**, (equivalently, **residual sum of squares**) denoted by SSE, is

$$\begin{aligned} SSE &= \sum e_i^2 = \sum (y_i - \hat{y}_i)^2 = \sum [y_i - (\hat{a} + \hat{b}x_i)]^2 && (58) \\ &= SS_{yy} - \hat{b}SS_{xy} = \sum y_i^2 - \hat{a}\sum y_i - \hat{b}\sum x_i y_i && (59) \end{aligned}$$

and the estimate of σ^2 is

$$\hat{\sigma}^2 = s^2 = \frac{SSE}{n-2} = \frac{(n-1)s_y^2(1-r^2)}{n-2}. \tag{60}$$

- The **coefficient of determination**, denoted by r^2, is the amount of the variation in y that is explained by the regression line.

$$\begin{aligned} r^2 = (r)^2 &= 1 - \frac{SSE}{SST}, \quad \text{where, } SST = SS_{yy} = \sum (y_i - \bar{y})^2 && (61) \\ &= \frac{SST - SSE}{SST} = \frac{\text{explained variation}}{\text{total variation}} && (62) \end{aligned}$$

- **Inference for b.**

  1. Test statistic:

  $$\frac{\hat{b} - b}{SE_{\hat{b}}} \sim t_{(n-2)} \qquad \text{where, } SE_{\hat{b}} = \frac{\hat{b}\sqrt{1-r^2}}{r\sqrt{n-2}} = \frac{\hat{\sigma}}{\sqrt{SS_{xx}}} \tag{63}$$

  2. Confidence Interval: $\hat{b} \pm t_{\alpha/2} SE_{\hat{b}}$

- **Mean Response of Y at a specified value x^*, ($\mu_{Y|x^*}$).**

  1. Point Estimate. For a specific value $x*$, the estimate of the **mean** value of Y is given by

  $$\hat{\mu}_{Y|x^*} = \hat{a} + \hat{b}x^*$$

  2. Confidence Interval. For a specific value $x*$, the $(1-\alpha)100\%$ confidence interval for $\mu_{Y|x^*}$ is given by

  $$\hat{\mu}_{Y|x^*} \pm t_{\alpha/2;(n-2)} SE_{\hat{\mu}}$$

  where, $SE_{\hat{\mu}} = s\sqrt{\dfrac{1}{n} + \dfrac{(x^* - \bar{x})^2}{(n-1)s_x^2}}$, and $s = s_y\sqrt{\dfrac{(n-1)(1-r^2)}{n-2}}$

- **Prediction of Y at a specified value x^*.**

  1. Point Estimate. For a specific value $x*$, the predicted value of Y is given by

  $$\hat{y} = \hat{a} + \hat{b}x^*$$

  2. Prediction Interval. For a specific value $x*$, the $(1-\alpha)100\%$ prediction interval is given by

  $$\hat{y} \pm t_{\alpha/2;(n-2)} SE_{\hat{y}}$$

  where, $SE_{\hat{y}} = s\sqrt{1 + \dfrac{1}{n} + \dfrac{(x^* - \bar{x})^2}{(n-1)s_x^2}}$, and $s = s_y\sqrt{\dfrac{(n-1)(1-r^2)}{n-2}}$