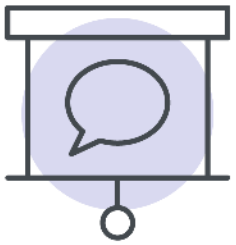
U of A

STAT151
FINAL EXAM
STUDY GUIDE



Lecture Notes

Ch. 9 – 11 – Gathering Data

Def'n: An observational study is a study where a researcher observes characteristics of subjects in samples from populations of interest.

A retrospective study is an observational study in which subjects are selected and then their previous conditions/behaviours are determined.

A prospective study is an observational study in which subjects are followed to observe future outcomes. No treatments are deliberately applied.

An experiment is a study where a researcher applies different treatments to different subjects and observes the outcomes. A controlled clinical trial is a type of experiment.

Drawing conclusions:

1. Infer to a larger population (Population)
2. Factor causes change in response (Causal)

		Random sampling? Yes	Random sampling? No
Random Assignment?	Yes	Both inferences are possible	Causal inferences
	No	Population inferences	Neither is possible

Both types of study allow for population inferences, but only a properly designed (and *randomized*) experiment allows for causal inferences to be valid. Experiments are not always feasible.

Sampling

Def'n: A sampling frame is the list of subjects in the population from which the sample is taken.

A random sample is a sample drawn in such a way that each element of the population has a chance of being selected. If chances are all the same → SRS of size *n*

Ex9.1) A deck of cards: picking a card is a simple random sample. Moreover, placing the card back in the deck is a sample *with replacement*. Otherwise, there is sampling *without replacement*.

Other sampling methods:

- *stratified random sample*: divide population into strata, SRS from each stratum.
- *cluster random sample*: divide population into large # of clusters, select SRS of clusters.
- *systematic sample*: select individuals systematically from a sampling frame.
- *convenience sample*: select individuals who are conveniently available.
- *voluntary response sample*: collect data from individuals who volunteer their responses.

Since the first two use simple random sampling at different points of the data collection process, it is best to compare SRS to stratified and cluster random sampling.

SRS: Advantage: sample tends to be a good reflection of population
Disadv.: sample may not reflect well if sample size is not large enough
Stratified RS: Adv.: ensures enough subjects in each group to compare
Disadv.: must have a sampling frame, must know how data separates into strata, more costly since each stratum must be used
Cluster RS: Adv.: do not need a sampling frame, less expensive to implement
Disadv.: still same problems as SRS if size not large enough

Bias:
- undercoverage: samples differ due to systematic exclusion of part of the population
Ex9.2) excluding French citizens with a phone survey in English
- response bias: samples differ because of method of observation
Ex9.3) if 4 of 5 dentists recommend Trident, is gum bad for you?
- nonresponse bias: samples differ due to unobtainable data
Ex9.4) sending surveys to Terrans & extra-terrestrials (who don't send the surveys back in non-Glorpian style)

More terms to know
Def'n: A control group is an experimental group that receives no treatment.
A placebo is identical to a treatment but definitively has no effect.
A single-blind experiment is where the subjects are unaware of which treatment is received, but investigator knows. Conversely, also possible for investigator to be unaware while subject knows.
A double-blind experiment has both subjects and investigators unaware of which treatment is received.
A design of an experiment is the overall plan for conducting the experiment.
A factor is a categorical explanatory variable. "Extraneous" factors may exist.

Four principles of experimental design:
Direct Control: hold or fix these factors at a constant level
Randomization: random assignment of treatments to remove effect of particular condition
Replication: repeating treatment enough for adequate sample size and more confidence in results
Blocking: arrange groups by these factors, apply all treatments inside each block

Chapter 1

Def'n: Statistics:

1) are commonly known as numerical facts
2) is a field of discipline or study

Here, statistics is about variation.

3 main aspects of statistics:

1) Design ("Think"): Planning how to obtain data to answer questions.
2) Description ("Show"): Summarizing the obtained data.
3) Inference ("Tell"): Making decisions and predictions based on data.

Def'n: A population consists of all elements whose characteristics are being studied.

Ex1.1)

A sample is a portion of the population selected for study.

Ex1.2)

A parameter is a summary measure calculated for population data.

A statistic is a summary measure calculated for sample data.

Types of statistics:

Descriptive: methods to view a given dataset.

→

Inferential: methods using sample results to infer conclusions about a larger population.

→

Def'n: A variable is any characteristic that is recorded for subjects in a study.

- Qualitative (categorical): cannot assume a numerical value but classifiable into 2 or more non-numeric categories. →
- Quantitative (numerical): measured numerically.
  - Discrete: only certain values with no intermediate values. →
  - Continuous: any numerical value over a certain interval or intervals.

→

Chapter 2 – Categorical Data Graphs

Def'n: A frequency table (for qualitative data) is a listing of possible values for a variable, together with the # of observations for each value.

$$Relative\ frequency = \frac{f}{\sum f} \qquad Percentage = Relative\ frequency \times 100\%$$

Table 2X0

Faculty	M	F	Frequency (*f*)	Relative frequency	Percentage (%)
Science					
Arts					
Business					
Phys. Ed. & Rec.					
Other					

Ex2.1) In percentage, how many students are in science? How many students are female? How many students are female and in science? How many science students are female? How many female students are in science?

Graphical Summaries

Def'n: A bar chart is a graph of bars whose heights represent the (relative) frequencies of respective categories.

Look for: frequently and infrequently occurring categories.

A pie chart is a circle divided into portions that represent (relative) frequency belonging to different categories.

Look for: categories that form large and small proportions of the data set.

A segmented bar chart uses a rectangular bar divided into segments that represent frequency or relative freq. of different categories.

Ex2.2) Draw appropriate graphs for data from Table 2X0.

Chapter 3 – Numerical Variable Graphs

Def'n: A stem-and-leaf display has each value divided into two portions: a stem and a leaf. The leaves for each stem are shown separately. (Values should be ranked.)

Look for: - typical values and corresponding spread
- gaps in the data or outliers
- presence of symmetry in the distribution
- number and location of peaks

Ex3.1)

Note: *Dotplots* also exist (see p. 48 in textbook), but "replace" the values with dots.

Def'n: A histogram, like a bar graph, graphically shows a frequency distribution. The data here, however, is quantitative.

Look for: - central or typical value and corresponding spread
- gaps in the data or outliers
- presence of symmetry in the distribution
- number and location of peaks

The data divide into intervals (normally of equal width).

Cumulative Relative Frequency = (Cumul. freq. of a class) / (Total obs'ns in dataset)

Table 3X0 – Total earnings as of Sep. 1/2017

Worldwide Box Office (in millions)	Number of movies *f*	Relative Frequency	Cumulative rel. freq.
200 to 599			
600 to 999			
1000 to 1399			
1400 to 1799			
1800 to 2199			
2200 to 2599			
2600 to 3000			

Ex3.2)

NOTE: Dot and S-and-L plots are good for small data sets because data values are retained. Histograms are better for large data sets to condense the data.

Histogram shapes/traits: (corresponding figures drawn in class)
1. Modes (unimodal, bimodal, multimodal, uniform)
2. Skewness (symmetric, left-skewed & right-skewed) → term refers to "TAIL"
3. Tail weight (normal, heavy-tailed, light-tailed)

Def'n: A timeplot is a graph of data collected over time (or a *time series*).

Look for: - a *trend* over time, denoting a decrease or increase.

- a pattern repeating at regular intervals (a *cycle* or *seasonal variation*)

Ex3.3)

Chapters 3/4 – Summary measures (and one more graph)

Measures of Center

Def'n: An outlier is an obs'n that falls well above or below the overall bulk of the data.

Population mean: $\mu = \frac{\sum y_i}{N}$ Sample mean: $\bar{y} = \frac{y_1 + y_2 + ... + y_n}{n} = \frac{\sum y_i}{n}$

The median is the value of the midpoint of a data set that has been ranked in order, increasing or decreasing. If dataset has an even # of observations, use the average of the middle 2 values.

Note: median resistant to outliers, mean uses all observations.

Table 4X0 – Estimated provincial populations circa 2016 (in millions)

ON	QC	BC	AB	MB	SK	NS	NB	NL	PEI
13.983	8.326	4.752	4.253	1.318	1.151	0.950	0.757	0.530	0.149

Ex4.1)

Mathematical Characteristics of the Mean

Adding or subtracting a constant to all scores will alter the value of the mean by the value of the constant.

$$\frac{\sum (y_i + c)}{n} = \frac{\sum y_i + \sum c}{n} = \frac{\sum y_i}{n} + \frac{\sum c}{n} = \bar{y} + \frac{nc}{n} = \bar{y} + c$$

Multiplying or dividing all scores by a constant will alter the value of the mean by the value of the constant.

$$\frac{\sum (cy_i)}{n} = \frac{cy_1 + cy_2 + ... + cy_n}{n} = \frac{c(y_1 + y_2 + ... + y_n)}{n} = c\left(\frac{\sum y_i}{n}\right) = c\bar{y}$$

Ex4.2) Use data from Ex4.1 to see these work with the first two averages by a) adding 1 million to each province, and b) multiplying each province population by 1.05.

a) Avg. pop'n of all provinces: Avg. pop'n from sample of 3 provinces:

b) Avg. pop'n of all provinces: Avg. pop'n from sample of 3 provinces:

Comparing Mean and Median: (corresponding figures drawn in class)

1. Symmetric curve & histogram

2. Right-skewed: Median < Mean

3. Left-skewed: Mean < Median

Def'n: The <u>mode</u> is the most frequent value in a data set.

Ex4.3) Provinces → Movies →

Measures of Spread

Def'n: Range = largest value – smallest value = max – min

Ex4.4) (from Table 4X0) range =

Deviations from the Mean:

Ex4.5) 1, 2, 4, 3

y_i	$y_i - \bar{y}$
1	$1 - 2.5 =$
2	$2 - 2.5 =$
4	$4 - 2.5 =$
3	$3 - 2.5 =$
	$\sum(y_i - \bar{y}) =$

Note that $\sum(y_i - \mu)$ and $\sum(y_i - \bar{y})$, or deviation of x from the mean, both equal zero.

Variance and Standard Deviation:
The most common measure of spread is standard deviation. Informally interpreted as the size of a "typical" deviation from the mean. Variance, however, must be calculated first. The basic formulas for variance are:

$$\sigma^2 = \frac{\sum (y_i - \mu)^2}{N} \qquad s^2 = \frac{\sum (y_i - \bar{y})^2}{n-1}$$

where σ^2 is the population variance and s^2 the sample variance.

Since $\sum (y_i - \bar{y})^2 = \sum y_i^{\,2} - \frac{(\sum y_i)^2}{n}$, the variance formulas become

$$\sigma^2 = \frac{1}{N}\left[\sum y_i^{\,2} - \frac{(\sum y_i)^2}{N}\right] \text{ and } s^2 = \frac{1}{n-1}\left[\sum y_i^{\,2} - \frac{(\sum y_i)^2}{n}\right]$$

Finding the standard deviation only requires taking the *positive* square root of the variance.

Population: $\sigma = \sqrt{\sigma^2}$ Sample: $s = \sqrt{s^2}$

Ex4.6) 1, 2, 4, 3

Important notes:
1. Standard deviation measures spread *only* about the mean (i.e. not the median).
2. Values of variance and std. dev. are never negative. (Equals zero only if *no spread*.)
3. The measurement units of variance are always the square of the units of the original data.
4. Standard deviation, like the mean, is not resistant to outliers.
5. Consider the sample variance s^2 to have $n-1$ degrees of freedom. There are n observations, and n deviations from the mean. Since the total always sums to zero, $n-1$ of these quantities determines the remaining one. Thus, only $n-1$ of the n deviations, $y_i - \bar{y}$, are freely determined. (Degrees of freedom apply only to samples.)

Mathematical Characteristics of the Standard Deviation
Adding or subtracting a constant to all scores will NOT alter the standard deviation.

Multiplying or dividing all scores by a constant will alter the standard deviation by the value of the constant.

Ex4.7) Compute standard deviation (population & sample) for 10, 20, 40, 30.

Measures of Position

Def'n: The p^{th} percentile is a value such that p percent of the observations fall below or at that value. Three useful percentiles are the quartiles. The *first quartile* has $p = 25$, the *second quartile* (the median) has $p = 50$, and the *third quartile* has $p = 75$.

Note: For odd n, EXCLUDE the median in each half when calculating Q_1 and Q_3.

The five-number summary consists of the min, Q_1, median, Q_3, and the max.

Def'n: The interquartile range (IQR) is the difference between the first and third quartiles. IQR = $Q_3 - Q_1$ (IQR is actually a measure of *spread*)

Ex4.8) 7.9 9.1 9.2 9.3 9.4 9.4 9.5 9.6 9.6 9.7

Boxplots:
Def'n: A boxplot shows the center, spread, and skewness of a data set.
To construct it:
Step 1: Rank the data in increasing order and find the median, Q_1, Q_3, and IQR.
Step 2: Find the points beyond the boundaries: 1.5*IQR below Q_1 and 1.5*IQR above Q_3, known as the lower & upper inner fences, respectively. These points are outliers.
Ex4.9) 1.5*IQR =
Lower i.f. = Upper i.f. =
Step 3: Determine smallest & largest values within the respective inner fences.
small = large =
Step 4: Draw linear scale containing entire range of data.
Step 5: Draw perpendicular lines to the scale to indicate Q_1 and Q_3. Connect ends of both lines. Box width = IQR
Step 6: Draw another line perpendicular to the scale to indicate the median inside the box.
Step 7: Draw two smaller lines perpendicular to the scale for the values from Step 3. Join their centers to the box to make whiskers.

What to do with outliers?
Consider lower & upper outer fences at 3.0*IQR below Q_1 and 3.0*IQR above Q_3.
Ex4.10) 3.0*IQR =
Lower o.f. = Upper o.f. =
A (*mild) outlier* is outside an inner fence but inside the outer fence.
A *far (or extreme) outlier* is outside either outer fence.
All textbooks are different for distinguishing outliers. Our textbook uses open circles for mild and asterisks, '*', for far outliers. Overall, classifying outliers is important whereas drawing them a certain way is subjective.
Whiskers extend on each end to the most extreme observations that are *not* outliers.
Ex4.11)

Looking at center, spread, and skewness:
Approx. value of the center? Width of IQR? Symmetric or skewed?

Boxplot vs. Histogram: Each graph highlights different features of a data set (layers of skewness and skewness/modality, respectively), so it's always better to construct both.

Chapter 5 – Standard Deviation as a Ruler

Empirical Rule applies only to a bell-shaped distribution.
1. About 68% of observations lie within 1σ of the mean.
2. About 95% of observations lie within 2σ of the mean.
3. About 99.7% of observations lie within 3σ of the mean.

Suppose we go further…, say, 6σ. Software produces a value of 99.999 999 803%, which means far less chance for "error" (the observations beyond 6σ from the mean).

Extra Measure of Position/Potential Outlier Identifier
z-score = (observation – mean) / (std. dev.)
- *z*-score tells us how many standard deviations the value is from the mean, positive OR negative
- more useful when distribution approximately normal.
- a potential outlier is more than 3σ from the mean.

Ex5.1) $\mu = 31.6$, $\sigma = 26.4$, $y = 50$

Ch. 12 – From Randomness to Probability

Def'n: An experiment is a process that, when performed, results in one and only one of many observations (or outcomes).

Probability is a numerical measure of likelihood that a specific outcome occurs.

3 Conceptual Approaches to Probability:

1) Classical probability

- equally likely outcomes exist when two or more outcomes have the same probability of occurrence
- *classical probability rule*:

$P(A)$ = (# of outcomes favourable to A) / (total # of outcomes for experiment)

2) Relative frequency concept of probability

- experiment repeated n times to simulate probability
- relative frequencies are NOT probabilities, they only approximate them.
- *Law of Large Numbers*: If an experiment is repeated again and again, the prob. of an event obtained from the relative frequency approaches the actual or theoretical prob.

3) Personal (or subjective) probability

- personal probability is the degree of belief that an outcome will occur, based on the available information

Calculating Probability

Def'n: A sample space (S) is the set of all *elementary* outcomes of an experiment.

An event (A) is a set of some of the elementary outcomes; $A \subset S$.

→ $P(A)$ = probability that A occurs

- A *union* of 2 events (A, B, **or** both happen) is denoted by A or B (or $A \cup B$).
- An *intersection* of 2 events (A **and** B happen together) is by A and B (or $A \cap B$).
- A *complement* of an event (event does **not** happen) is denoted by A^C.

A Venn diagram is a picture that depicts S (events above drawn in class).

Experiment	**Outcomes**	**Sample Space**
Toss a coin		
Toss 2-headed coin		
Toss a $5 bill		
Pick a suit		

Properties for calculating probabilities:
1. $0 \le P(A) \le 1$
2. $P(A)$ is the sum of probabilities of all elementary outcomes comprising A.
3. $P(S) = 1$

Ch. 13 – Probability Rules!

Basic Rules for Finding the Probability of a Pair of Events:

Table 13X0 – 2-way table of responses

	Like Hockey (A)	Indifferent (B)	Dislike Hockey (C)	**Total**
Male (M)				
Female (F)				
Total				

Def'n: Marginal probability is the probability of a single event without consideration of any other event.

Ex13.1) $P(M) =$ $P(F) =$

$P(A) =$ $P(B) =$ $P(C) =$

Conditional probability is the probability that an event will occur given that another event has already occurred. If A and B are 2 events, then the conditional probability of A given B is written as $P(A \mid B)$. Keywords: **given**, **if**, **of**

$$P(A \mid B) = \frac{P(A \cap B)}{P(B)} \quad \text{and} \quad P(B \mid A) = \frac{P(B \cap A)}{P(A)}$$

such that $P(A) \neq 0$ and $P(B) \neq 0$.

Ex13.2) a) If you are male in this class, what is the probability that you like hockey?

b) What is the probability of being female in this class, given that you are indifferent to hockey?

Two events are independent if the occurrence of one does not affect the probability of the occurrence of the other. In other words,

$P(A \mid B) = P(A)$ OR $P(B \mid A) = P(B)$

Ex13.3) From Table 13X0, $P(F) =$ $P(F \mid B) =$

Ex13.4) deck of cards: $P(Black) =$ $P(Black \mid Face) =$

Disjoint (or mutually exclusive) events are events that cannot occur together.

Ex13.5) deck of cards
- R = get red suit →
- B = get black suit →
- F = get face card →
- Which pairs are disjoint?

Ex13.6) a single die
- E = even =
- O = odd =
- Pr = prime =

Note: Two events are either disjoint or independent, but not both (unless one has zero probability). How to differentiate between disjoint, independent, and dependent events?

Complement Rule: $P(A) + P(A^C) = 1$, so
$P(A) = 1 - P(A^C)$ and $P(A^C) = 1 - P(A)$

Ex13.7) From Table 13X0, $P(Female^C) = P(F^C) = 1 - P(F) =$

Ex13.8) deck of cards: $P(Face^C) = P(F^C) = 1 - P(F) =$

Note: $P(A^C \mid B) = 1 - P(A \mid B)$ Does $P(A \mid B^C) = 1 - P(A \mid B)$? Not necessarily.

Ex13.9) deck of cards:
$P(Face \mid Black) =$ $P(Face^C \mid Black) =$
$P(Face \mid Black^C) =$

Ex13.10) deck of cards: $P(Heart \mid Red) =$

Multiplication Rule: $P(A \cap B) = P(A) \times P(B \mid A) = P(B) \times P(A \mid B)$
- If A and B are two *independent* events, $P(A \cap B) = P(A) \times P(B)$.
- If A and B are two *disjoint* events, $P(A \cap B) = 0$.

Ex13.11) From Table 13X0, what is the probability of being male and liking hockey? Being indifferent to hockey and female?

Ex13.12) deck of cards: What is the probability of drawing a black face card? Black and red card off single draw?

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

- If A and B are two *disjoint* events, $P(A \cup B) = P(A) + P(B)$.

Ex13.13) From Table 13X0, what is the probability of being male or liking hockey? Being indifferent to hockey or female?

Ex13.14) deck of cards: What is the probability of drawing a black card or a red card? Face card or ace? Black card or face card?

With 3 or more independent events, the multiplication rule becomes

$$P(A_1 \cap A_2 \cap ... \cap A_k) = P(A_1) \times P(A_2) \times ... \times P(A_k)$$

With 3 or more disjoint events, the addition rule becomes

$$P(A_1 \cup A_2 \cup ... \cup A_k) = P(A_1) + P(A_2) + ... + P(A_k)$$

Total Probability Rule (for two events): → diagram drawn in class

$$P(A) = P(A \cap B) + P(A \cap B^C) \qquad \text{OR} \qquad P(B) = P(A \cap B) + P(A^C \cap B)$$

Overall examples:

Ex13.15) Suppose the probability of liking Gretzky is 0.86, the probability of liking Crosby is 0.79, and the probability of liking both is 0.71.

a) What is the probability of liking neither Gretzky nor Crosby? The probability of liking Gretzky but not Crosby?

b) What is the probability of liking Gretzky or Crosby?

c) What is the probability of liking Gretzky or not liking Crosby?

d) What is the probability of liking Crosby, given you like Gretzky?

Ex13.16) Suppose 30% of calls to an Oilers ticket phone line result in a sale being made. Assume all calls are independent. Suppose an operator handles 10 calls.

a) What is the probability that none of the 10 calls results in a sale?

b) What is the probability that at least one call results in a sale being made?

Ex13.17) Three friends play tennis (call them A, B, and C). The probability that A beats B is 0.7, the probability that A beats C is 0.8 and the probability that B beats C is 0.6. Assume all events are independent and that each player plays another at most once.

a) What is the probability that A wins both of its games?

b) What is the probability that A loses both of its games?

c) What is the probability that everyone wins a game?

Ex13.18) Assume that 70% of students who take the midterm next month have studied for the test. Of those who study for the midterm, 95% pass; of those who do not study for the test, 60% pass. What is the probability that a student did not study for the midterm, given that they pass the midterm?

Ex13.19) Bob and Mark regularly play a simple darts game. Each play consists of one throw each at a target. A player wins if they hit the target and the other doesn't. Each time they play, they flip a coin to determine who throws first. History has shown that Bob hits the target 30% of the time and Mark hits the target 37% of the time. Also, there is a 9% chance that both will hit the target on any random play.

(a) Are Bob and Mark's throws independent?

(b) On a single play, what is the probability that neither hits the target?

(c) Suppose that on a single play, somebody wins. What is the probability that it was Bob?

Ch. 14 – Random Variables

Def'n: A random variable is a numerical measurement of the outcome of a random phenomenon.

A discrete random variable is a random variable that assumes separate values.

→ # of people who think stats is dry

The probability distribution of a discrete random variable lists all possible values that the random variable can assume and their corresponding probabilities.

Notation: X = random variable; x = particular value;
$P(X = x)$ denotes probability that X equals the value x.

Ex14.1) Toss a coin 3 times. Let X be the number of heads. What is the prob. dist'n?

Table 14X0

x	$P(X = x)$

Two noticeable characteristics for discrete probability distribution:
1. $0 \leq P(X = x) \leq 1$ for each value of x
2. $\sum P(X = x) = 1$

Ex14.2) Find the probabilities of the following events:

"no heads":

"at least one head":

"less than 2 heads":

Ex14.3) Refer back to Ex13.19. Suppose Bob and Mark play this game twice and that each play is independent of the other. Define X as the total number of hits on two independent plays. Define Y_i as the total number of hits on play i. Find the probability distribution of X.

Ex14.4) Suppose you roll two dice. If you roll 7 or 11, you win $20. Otherwise, you win nothing. a) Let X be your winnings. Find the probability distribution of X. b) Suppose you pay $10 to play the game. Let Y be your net profit. Find the probability distribution of Y.

The population mean μ of a discrete random variable is a measure of the center of its distribution. It can be seen as a long-run average under replication. More precisely,

$$\mu = \sum x_i P(X = x_i)$$

Sometimes referred to as $\mu = E(X)$ = the expected value of X.
Keep in mind that μ is not necessarily a "typical" value of X (it's not the mode).

Ex14.5) Find the mean for Ex14.1).

Ex14.6) Toss an unfair coin 3 times (hypothetical). Let X be as in previous example.

x	$P(X = x)$
0	0.10
1	0.05
2	0.20
3	0.65

As 2nd example shows, interpretation of μ as a measure of center of a distribution is more useful when the distribution is roughly symmetric, less useful when the distribution is highly skewed.

Ex14.7) Using Ex14.3), what is the expected total number of hits on two independent plays?

Ex14.8) Using Ex14.4), what are the expected winnings? The expected net profit? How much would you pay to play this game?

The population standard deviation σ of a discrete random variable is a measure of variability of its distribution. As before, the standard deviation is defined as the square root of the population variance σ^2, given by

$$\sigma^2 = \sum (x_i - \mu)^2 P(X = x_i) = \sum x_i^2 P(X = x_i) - \mu^2$$

Ex14.9) Find the standard deviation for X in Ex14.1).

Ex14.10) Using Ex14.3), find the standard deviation for the total number of hits on two independent plays.

*Continuous Distribution*s

Def'n: A continuous random variable assumes any value contained in one or more intervals.
→ average alcohol intake by a student, average alcohol outtake by a student
The probability distribution of a continuous r.v. is specified by a curve.

Two noticeable characteristics for continuous probability distribution (sans calculus):
1. The probability that X assumes a value in any interval lies in the range 0 to 1.
2. The interval containing all possible values has probability equal to 1, so the total area under the curve equals 1.

Using probability symbols, point 1 is denoted by

$$P(a \le X \le b) = \text{Area under the curve from } a \text{ to } b$$

The probability that a continuous random variable X assumes a single value is always zero. This is because the area of a line, which represents a single point, is zero.

In general, if a and b are two of the values that X can assume, then

$$P(a) = 0 \qquad \text{and } P(b) = 0$$

Hence, $P(a \le X \le b) = P(a < X < b)$. For a continuous probability distribution, the probability is always calculated for an interval, such as $P(X > b)$ or $P(X \le a)$.

Ex14.11) Suppose we have a "uniform" distribution where obtaining each value between two endpoints has equal probability. Suppose the endpoints are 0 and 2.

a) What is the probability of $P(X < 1.5)$?
b) What is the probability of $P(X \leq 1.5)$?
c) What is the probability of $P(0.5 < X < 2.5)$?

The Normal Distribution

- most widely used and most important of all (continuous) probability distributions
- the normal distribution has 2 *parameters*: μ and σ
- the density function is:

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

- the *normal (distribution) curve*, when plotted, gives a bell-shaped curve such that
  1. The total area under the curve is 1.0.
  2. The curve is symmetric about the mean (or bell-shaped).
  3. The two tails of the curve extend indefinitely.

- there is not just one normal curve, but a *family* of normal curves. Each different set of μ and σ gives a different curve. μ determines the center of the distribution and σ gives the spread of the curve.

Standard Normal Distribution

Def'n: The standard normal distribution is the normal distribution with $\mu = 0$ and $\sigma = 1$. It is the distribution of normal z-scores.

Recall Empirical Rule.
$\mu \pm \sigma$ gives middle 68.26% of the data. In terms of z-scores, this is the interval (-1.0, 1.0).
$\mu \pm 2\sigma$ gives middle 95.44% of the data; z-score interval of (-2.0, 2.0).
$\mu \pm 3\sigma$ gives middle 99.74% of the data; z-score interval of (-3.0, 3.0).

Using Table of Standard Normal Curve Areas:
For any number z between -3.90 and 3.90 and rounded to 2 decimal places, Table Z gives
(area under curve to the left of z) = $P(Z < z) = P(Z \leq z)$ $Z \sim N(0, 1)$

Helpful tips:
- diagrams are helpful
- Complement: $P(Z \geq z) = P(Z > z) = 1 - P(Z \leq z)$
- Symmetry: $P(Z \geq z) = P(Z \leq -z)$
- $P(a \leq Z \leq b) = P(Z \leq b) - P(Z \leq a)$
- If $z > 0$, then $P(-z \leq Z \leq z) = 1 - 2P(Z \leq -z)$

Ex14.12) *Examples with z-scores (finding prob.):*
a) $P(Z < -3.14) =$
b) $P(Z > 1.44) =$
OR $P(Z > 1.44) =$
c) $P(-3.14 \leq Z \leq 1.44) =$
d) $P(-2.00 \leq Z \leq 2.00) =$

Standardizing a normal distribution:
$X \sim N(\mu, \sigma)$ and $Z \sim N(0, 1)$. What is Z? $Z = \frac{X - \mu}{\sigma}$, $z = \frac{x - \mu}{\sigma}$

$P(X \leq x) \rightarrow P\left(\frac{X - \mu}{\sigma} \leq \frac{x - \mu}{\sigma}\right) = P(Z \leq z)$

Ex14.13) *Examples with z-scores (standardizing):*
Find the following probabilities for $X \sim N(75, 6.5)$.

a) What is the probability of getting a value greater than 94.5?

b) What is the probability of getting a value between 71.75 and 84?

Identifying values:
Using the area under the curve, you can find appropriate z values; so, what are the corresponding x values?
$x = \mu + z\sigma$

Ex14.14) *Examples with z-scores (finding values):*
Use the same X as in Ex14.13) to answer the following:

a) What value denotes the top 5%?

b) What values bound the middle 70% of the data?

Combinations and Functions of Random Variables
For any constants a and b,

Means:	*Variances*:
1. $E(a) = a$	1. $V(a) = 0$
2. $E(aX) = aE(X)$	2. $V(aX) = a^2V(X)$
3. $E(aX + b) = aE(X) + b$	3. $V(aX + b) = a^2V(X)$
4. $E(aX \pm bY) = aE(X) \pm bE(Y)$	4. $V(aX \pm bY) = a^2V(X) + b^2V(Y) \pm$ ~~$2ab\text{cov}(X, Y)$~~

Rule 4 for variance eliminates the last component only if X and Y are independent.

Ex14.15) Let X be the temperature in Edmonton on a random day in March and Y be the temperature on a random day in February. Suppose all days are independent and that
$E(X) = 4$, $V(X) = 3$ $E(Y) = -3$, $V(Y) = 1$

a) Find the mean and standard deviation of $W = 4X - 3Y - \pi$.

b) Find the mean and standard deviation for the total of two random days in March.

c) Find the mean and standard deviation for the difference between the total of two random days in March and the total of three random days in February.

d) Find the mean and standard deviation of the average of two random days in March and one random day in February.

Ch. 15 – Sampling Distributions

Expanded def'n: A parameter is: - a numerical value describing some aspect of a pop'n
- usually regarded as constant
- usually unknown

A statistic is: - a numerical value describing some aspect of a sample
- regarded as random before sample is selected
- observed after sample is selected

The observed value depends on the particular sample selected from the population; typically, it varies from sample to sample. This variability is called sampling variability. The distribution of all the values of a statistic is called its sampling distribution.

Def'n: $\hat{p}$ = proportion of ppl with a specific characteristic in a random sample of size n

p = population proportion of ppl with a specific characteristic

The estimate of the standard deviation of a sampling distribution is called a standard error.

General Properties of the Sampling Distribution of $\hat{p}$:

Let $\hat{p}$ and p be as above. Also, $\mu_{\hat{p}}$ and $\sigma_{\hat{p}}$ are the mean and standard deviation for the distribution of $\hat{p}$. Then the following rules hold:

Rule 1: $\mu_{\hat{p}} = p$. (Textbook uses $\mu(\hat{p})$)

Rule 2: $\sigma_{\hat{p}} = \sqrt{\frac{p(1-p)}{n}} = \sqrt{\frac{pq}{n}}$. (standard error → $\hat{\sigma}_{\hat{p}}$)

Ex15.1) Suppose the population proportion is 0.5.

a) What is the standard deviation of $\hat{p}$ for a sample size of 4?

b) What is the smallest that n can be so that the sample proportion has a standard deviation of at most 0.125?

Rule 3: When n is large and p is not too near 0 or 1, the sampling distribution of $\hat{p}$ is approximately normal. The farther from $p = 0.5$, the larger n must be for accurate normal approximation of $\hat{p}$. Thus, if np and $n(1-p)$ are both sufficiently large (≥ 15), then it is safe to use a normal approximation.

Further assumptions: the sample should always be random and, if sampling without replacement, the sample should be less than 10% of the population.

Using all 3 rules, the distribution of $\hat{p}$ is approximately normal.

$$Z = \frac{\hat{p} - p}{\sqrt{\frac{p(1-p)}{n}}} \sim N(0,1)$$

Ex15.2) Suppose that the true proportion of people who have heard of Sidney Crosby is 0.87 and that a new random sample consists of 158 people.

a) Find the mean and standard deviation of $\hat{p}$.

b) What can you say about the shape of the distribution of $\hat{p}$?

c) What is the probability of getting a sample proportion greater than 0.94?

d) What is the probability of less than 140 people hearing of Sidney Crosby in the sample?

Sampling Distribution of Mean
How does the sampling distribution of the sample mean compare with the distribution of a single observation (which comes from a population)?

Ex15.3) An epically gigantic jar contains a large number of balls, each labeled 1, 2, or 3, with the same proportion for each value.

Let Y be the label on a randomly selected ball. Find μ_Y and σ_Y.

Let $\{Y_1, Y_2\}$ be a random sample of size $n = 2$. Find the sampling distribution of the sample mean $\bar{Y}$. Calculate $\mu_{\bar{y}}$ and $\sigma_{\bar{y}}$.

There are ___ possible samples:

$\bar{y}$					
$P(\bar{Y} = \bar{y})$					

Progressing further with inference, we can now discuss the following properties.

General Properties of the Sampling Distribution of $\bar{y}$ (or $\bar{x}$):
Let $\bar{y}$ denote the mean of the observations in a random sample of size n from a population having mean μ and standard deviation σ. Also, $\mu_{\bar{y}}$ and $\sigma_{\bar{y}}$ are the mean and standard deviation for the distribution of $\bar{y}$. Then the following rules hold:

Rule 1: $\mu_{\bar{y}} = \mu$.

Rule 2: $\sigma_{\bar{y}} = \frac{\sigma}{\sqrt{n}}$.

Note also that:
1. The spread of the sampling dist'n of $\bar{y}$ is smaller than the spread of the pop'n dist'n.
2. As n increases, $\sigma_{\bar{y}}$ decreases.

Ex15.4) Suppose the population standard deviation is 10.

a) What is the std. dev. of the sample mean for some of the following sample sizes?
n = 1, 2, 4, 9, 16, 25, 100

b) What is the smallest that n can be so that the sample mean has a standard deviation of at most 2?

Rule 3: When the population distribution is normal, the sampling distribution of $\bar{y}$ is also normal for any sample size n.

Combining the 3 rules, if the population distribution is $N(\mu, \sigma)$, then $\bar{Y}$ is $N(\mu, \sigma/\sqrt{n})$.

Rule 4 (**Central Limit Theorem**): When n is sufficiently large, the sampling distribution of $\bar{y}$ is well approximated by a normal curve, even when the population distribution is not itself normal. The Central Limit Theorem can safely be applied if n exceeds 30.

Using all 4 rules, if n is large and/or the population is normal, then the sampling distribution of $\bar{Y}$ is approximately normal.

$$Z = \frac{\bar{Y} - \mu}{\sigma / \sqrt{n}} \dot{\sim} N(0,1)$$

Ex15.5) Suppose the mean length of all episodes of a (formerly) hilarious series is 20.834 minutes, whereas the standard deviation is 0.593 minutes. Let $\bar{Y}$ be the average length for a random sample of 100 episodes.

a) Find the mean and standard deviation of $\bar{Y}$.

b) What can you say about the shape of the distribution of $\bar{Y}$?

c) What is the probability of getting a sample mean between 20.7 and 21 minutes?

d) Can you find $P(20.7 \leq Y \leq 21)$, where Y is the length of a single randomly selected episode? How would this value compare with the one in part c)?