Lesson 1

1. Explain summation notation, its linearity property, and compute the sums.

$n = 10$	$\Sigma x_i = 2$	$\Sigma y_i = 5$	$\Sigma x_i^2 = 160$

a) $\Sigma(7y_i + 3)$ b) $\Sigma(x_i - 1)^2$

2. Statistics is the science of **data**. We collect data by measuring objects (or people) and we use this evidence to solve problems and make decisions. Let X represent a specific measurement.
   a) A _____ is the collection of Xs for all possible objects/people of interest.
   b) A _____ is a subset of the population of Xs.
   c) Explain parts (a) and (b) with a picture.
   d) A _____ sample has characteristics similar to the population and is useful when we want to make inferences.
   e) List some advantages of a **random sample**.

3. List two possible **data types** for a specific measurement X.

Type. List the data type for each measurement.

	Field	Measurement, X	Data type
a)	Electronics	Switch: open/closed	
b)	Physics	Mass	
c)	Engineering	Type of material	
d)	Construction	Job completion time	

Desks. Seat and desk purchases for a new CSU classroom building depend on student height X.
   a) What is the population of interest?
   b) Why not measure the height of each individual in the population?
   c) Why is the basketball team not a representative sample?
   d) Why would a simple random sample of students be a good choice?

Speed. A physicist conducts 50 replications of an experiment to measure the speed of light based on travel time between two locations. Why are the sample measurements not identical?

Lesson 2

1. Summarize the sample **categorical data** by creating a **frequency table** and **bar graph**.

Y	Y	N	Y	N	N	Y	Y	Y	N
Y	N	N	N	Y	Y	Y	N	Y	Y

2. Explain sample **quantitative data** summary using CUSS.

3. Sketch examples of the common distribution shapes.

4. Create the **histogram** of the frequency table for the quantitative data and describe its shape.

Class	[0, 3)	[3, 6)	[6, 9)	[9, 12)	[12, 15)	[15, 18)	[18, 21)
Frequency	1	0	1	2	3	6	5

5. Create the **stem-and-leaf plot** and describe its shape.

12	14	16	17	17	20	22	23	25	26
29	32	33	37	38	41	45	46	46	48

6. Provide the formulas and compute the statistics for the X values.

–1	2	3	8

a) Sample **mean**
b) S_{xx}
c) Sample **variance**
d) Sample **standard deviation**
e) Sample **coefficient of variation (CV)**
f) Discuss the units of each statistic.

Centered. Prove the sum of centered data values is always zero.

Guac. Wholly Guacamole executives survey customers to determine how often their brand is purchased. In a random sample of guacamole customers, 75 reply "Always," 27 reply "Sometimes," and 48 reply "Never." Summarize the categorical data. A manager wants 60% of guacamole customers to use Wholly Guacamole brand at least some of the time. Will she be happy with the survey evidence? Explain.

Exoplanets. Astronomers are actively searching for planets beyond our solar system. A frequency table for a sample of planetary mass values, relative to Earth, is provided. One astronomer wants to know if the distribution of masses is bell-shaped. Provide an answer and supporting evidence.

Class	[0, 2)	[2, 4)	[4, 6)	[6, 8)	[8, 10)	[10, 12)	[12, 14)	[14, 16)
Frequency	5	6	7	5	3	2	0	1

CV. The CV for a certain data set is 0.15 and the sample mean is 40. Compute the sample variance.

Lesson 3

1. State and verify the S_{xx} "shortcut" formula.

2. Consider the summary statistics for a random sample of quantitative data.

n = 17	$\Sigma x_i = 170$	$\Sigma x_i^2 = 1844$

   a) Compute the sample mean and standard deviation.
   b) Compute and interpret the **Z-score** for the data value X = 14.8.
   c) Compute the "within 2 standard deviations of the mean" boundaries.

3. Describe **percentiles** for **ordered data** (i.e. data listed from smallest to largest).
   The percentile rule using locator value L is:
   1) integer L: average the data values in positions L and L + 1
   2) non-integer L: round L to the next larger integer and use the data value at that position

4. Consider the ordered data set consisting of 10 values.

Data, X	4	6	7	10	11	13	17	19	23	29
Position	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*

   a) Compute the 90th percentile.
   b) Compute the **five-number summary**.
   c) Create the **box plot**.
   d) Compute the **range** and **interquartile range (IQR)**.
   e) Compute the outlier fences and check for **outliers** (i.e. data values much smaller or larger than the other values in the set).

Conversion. In some unit conversion problems, Y is a linear function of the data X. Find the sample mean and variance of Y in terms of the sample mean and variance of X.

Salary. We have an ordered random sample of fourteen Wossamotta U employee salaries rounded to the nearest thousand dollars.

26	29	35	37	38	41	44
54	57	61	71	87	112	1800

   a) The 1,800 represents the head football coach's salary. What is the dollar amount?
   b) Is the sample mean or median more representative of a typical salary at Wossamotta U? (Hints: $\Sigma x_i = 2492$ and five-number summary: 26, 37, 49, 71, 1800)
   c) Does the data set contain any outliers? Provide evidence.

Yohimbine. Unregulated supplements containing an ingredient from the African yohimbe tree are banned in many countries. Ten random supplements contain the following relative amounts of yohimbine (i.e. 100 is the labeled amount). What value would you report as a typical relative amount of yohimbine in a randomly selected supplement?

23	36	46	47	62	70	99	104	142	147

Lesson 4

1. A **chance experiment** is a process whose outcome is subject to uncertainty. For example, rolling a fair six-sided die.
   a) List the possible outcomes (i.e. **sample space**) for the experiment.
   b) Let A be the event the outcome of a single roll is an odd number. Compute the **probability** event A occurs, P(A).

2. For an arbitrary event A, express numerical boundaries for P(A).

3. Complete the table for arbitrary events A and B.

Event	Probability notation	Explanation
a) **Complement** of A		
b) **Union** of A, B		
c) **Intersection** of A, B		
d) **Null/empty**		

4. A few rules:
   a) State the **complement rule** for probabilities.
   b) State the **addition rule** for probabilities.
   c) State **DeMorgan's laws**.

Maze. Behavioral psychologists study mice in a lab maze. At the entry point a mouse can move forward (F), left (L), or right (R). Consider observing the direction selected by each of two consecutive mice. List the outcomes in:
   a) the sample space S.
   b) event B, that both mice make the same selection.
   c) event C, that at least one mouse moved forward.
   d) the complement of C.
   e) the event that B and C both occur. Two events are **mutually exclusive (m.e.)** if they have no outcomes in common. Are B and C mutually exclusive?
   f) the event that B or C, or both, occur.

Wossamotta U. At Wossamotta University, 20% of the students are freshmen, 15% are in the Natural Sciences college, and 3% meet both criteria. Draw the **Venn diagram** and compute the probability a randomly selected student:
   a) is either a freshman or in the Natural Sciences college.
   b) is neither a freshman nor in the Natural Sciences college.
   c) is a freshman but not in the Natural Sciences college.

Triple. Consider three events with P(A) = P(B) = P(C) = .40 and the probability of the union of any pair of events is .64. If the probability of the union of all three events is .82, use the inclusion/exclusion property to compute the probability of the intersection of the three events. Draw the Venn diagram.

Lesson 5

1. Consider two events with $P(A) = 0.4$, $P(B) = 0.5$, and $P(A \cap B) = 0.2$.
   a) Compute $P(A \cup B)$.
   b) Explain **conditional probability** and the **multiplication rule**.
   c) Compute the conditional probabilities $P(A \mid B)$ and $P(B \mid A)$.

2. Explain the **law of total probability**.

3. Create a **chance tree** and compute $P(B^C)$.

$P(A) = 0.4$	$P(B \mid A) = 0.8$	$P(B \mid A^C) = 0.3$

Fishing. A certain pond is stocked with three types of fish (A, B, C) according to the following percentages . Males and females of each fish type are equally aggressive towards food sources.

	A	B	C
Male	14%	20%	26%
Female	20%	10%	10%

Your friend just caught a fish. Compute the probability the fish is:
   a) Type A.
   b) female.
   c) Type A, if we know it is female.
   d) female, if we know it is Type A.

Lost. A small plane has disappeared and we know there is an 80 percent chance this type of plane will be located. Historically, 70% of the located planes have an emergency beacon, and 90% of the unlocated planes have no emergency beacon.
   a) Create a chance tree for these conditions.
   b) What percentage of missing planes have emergency beacons?
   c) If the missing plane has an emergency beacon, what is the chance it will not be located?
   d) If the missing plane has no emergency beacon, what is the chance it will be located?

Red-green. One box contains three red balls and two green balls, and a second box contains four red balls and one green ball. A ball is randomly chosen from the first box and placed in the second box. Then a ball is randomly selected from the second box and placed in the first.
   a) Compute the probability that a red ball is selected each time.
   b) At the conclusion of the two selections, what is the chance that the boxes have the same distributions as at the start?

Lesson 6

1. Consider two events with $P(A) = 0.4$, $P(B) = 0.5$, and $P(A \cap B) = 0.2$.
   a) Events A and B are **independent** if the known occurrence of one event does not affect the probability of the other. Or, by definition: $P(A \cap B) = P(A) \cdot P(B)$. Are A and B independent?
   b) Compute the conditional probabilities $P(A \mid B)$ and $P(B \mid A)$ and compare to $P(A)$ and $P(B)$.
   c) Verify for independent events A and B: $P(A) = P(A \mid B)$.

2. Consider two events with $P(C) = 0.3$, $P(D) = 0.6$, and $P(C \cap D) = 0.24$.
   a) Are C and D independent? Justify your answer.
   b) Compute $P(C \mid D)$ and $P(D \mid C)$ and compare to $P(C)$ and $P(D)$.

3. The value of a **random variable (RV)** is assigned by a chance experiment. The possible assigned values are called the **support** of the RV. In each example, indicate if the random variable is **discrete** (i.e. countable support) or **continuous** (i.e. interval support).

	Chance experiment	RV	Support
a)	Inspect a shipment of 50 objects	# defective	$X = 0, 1, \ldots, 50$
b)	Fill a 16 oz. bottle	# of ounces	$0 \le X \le 16$
c)	ATM operation	Time between use	$0 \le X$
d)	Operate a business for one day	# of customers	$X = 0, 1, \ldots$

Parts. Consider the selection of a part from a collection parts produced at three companies. Let A be the event the part is from company A, B be the event the part is from company B, and D be the event the part is defective.

$P(A) = 0.4$	$P(B) = 0.2$	$P(D) = 0.06$	$P(A \cap D) = 0.02$	$P(B \cap D) = 0.012$

   a) Are events A and D independent? Justify your answer.
   b) Compare $P(D)$ and $P(D \mid A)$. Does the probability of selecting a defective depend on my knowledge about the part coming from Company A?
   c) Are events B and D independent? Justify your answer.
   d) Compare $P(D)$ and $P(D \mid B)$. Does the probability of selecting a defective depend on my knowledge about the part coming from Company B?

Independent. For independent events A and B, verify: B and A^C are independent. Hint: $P(B) = P(B \cap A) + P(B \cap A^C)$.

Events. Show that two non-null events cannot be mutually exclusive and independent.

Project. Two parts for a project are machined independently and the project fails if both parts are defective. There is a 12% chance only Part A is defective and a 6% chance only Part B is defective. If most projects succeed, what is the probability this project will fail?

Lesson 7

1. Consider the **probability mass function (pmf)** for the discrete random variable X.

x	2	3	4	5
p(x)	0.05	0.25	0.60	0.10

   a) Graph the pmf and explain why the pmf is valid.
   b) The pmf describes a population of X values. Compute and interpret P(X = 4).
   c) The **cumulative distribution function (cdf)** is the cumulative probability from the left end of the support, i.e. $F(c) = P(X \leq c)$. Compute and interpret F(4).
   d) The **expected value E[X]** is the center of the pmf. Write the formula and compute E[X].
   e) The **law of the unconscious statistician (LOTUS)** allows calculation of expected values of functions of a random variable. Write the formula and compute $E[X^2]$.
   f) The **variance Var(X)** is the spread of the pmf. Write the formula and its shortcut formula and compute Var(X).
   g) Explain the relationship between **standard deviation SD(X)** and Var(X). Compute SD(X).

2. Let Y be a linear function of random variable X.
   a) Find E[Y] in terms of E[X].
   b) Find Var(Y) in terms of Var(X).

3. Draw a picture to distinguish the different notations for mean, variance, and standard deviation when working with sample data and random variables.

Cereal. A certain cereal comes in three box sizes. At a Sally's Supermarket, 20% of the sales are 20 oz. boxes, 50% are 23 oz. boxes, and 30% are 26 oz. boxes. Let X be the number of ounces in the next box of this cereal purchased.

   a) Write the pmf of X.
   b) Compute E[X].
   c) Compute $E[X^2]$, Var(X), and SD(X).
   d) The price of an X ounce box is $Y = 2.4 + 0.1X$. Compute the expected price paid for the next box purchased.
   e) Compute the variance and standard deviation of the price paid for the next box.
   f) Suppose the filling machine at the cereal factory is incorrectly adjusted and the actual fill weight is $W = X - 0.001X^2$. Compute the expected actual fill weight of the next box purchased.

Paperwork. A certain document requires two, three, four, or five signatures for validation. The probability that X signatures are required is proportional to X. Write the pmf of X.

***** Bring R printouts next time! *****

Lesson 8

1. A **Bernoulli random variable** X is assigned the outcome of a single Yes(1)/No(0) experiment.
   a) List several applications for this random variable.
   b) Explain the notation: $X \sim$ Bernoulli(p)
   c) The pmf describes a population of X values. Write the pmf for X.
   d) Compute E[X].
   e) Compute $E[X^2]$, Var(X), and SD(X).

2. A **binomial random variable** X counts the number of 1's in n independent Bernoulli(p) trials. This RV notation contains two parameters: $X \sim$ binom(n, p).
   a) Write the pmf for X (include the support).
   b) Write the formulas for E[X] and Var(X).

Sleep. Twenty-five percent of Americans suffer from chronic insomnia. Randomly select one American and let X equal one if she suffers from chronic insomnia and zero if not.
   a) What specific type of random variable is X?
   b) Compute E[X], Var(X), and SD(X).
   c) Compute $E[X^{79}]$.

Sleep 2. Twenty-five percent of Americans suffer from chronic insomnia. Consider a random sample of ten Americans and let Y be the number who suffer from chronic insomnia.
   a) What specific type of random variable is Y?
   b) Compute P(Y = 4).
   c) Compute the mean, variance, and standard deviation of Y.

Download R. To download a free version of R, click on "Download R" at the R website. Next, scroll down to USA and click on Berkeley's http site. Depending on your operating system (Mac or Windows), select the appropriate download file.
   a) Explain the "R basics" printout.
   b) Explain the "Binomial distribution" printout.

Bolts. Bolts manufactured for a certain application must be scrapped if they are defective. At a certain manufacturing facility, 5% of all bolts are defective. Let X be the number of defective bolts in a random sample of 25. Use R to compute the probabilities in parts a, b, and c.
   a) Exactly two defective bolts.
   b) At most four defective bolts.
   c) More than four defective bolts.
   d) Compute the mean and standard deviation of X.

Medication. Half of all people with a certain illness recover naturally. Twelve random people with the illness are treated with a new medication and 11 recover shortly thereafter.
   a) Assume the medication has no effect. What is the probability that at least 11 of the 12 people receiving an ineffective medication for this illness will recover?
   b) Does the computation in part (a) suggest the medication has no effect?

Lesson 9

1. A **geometric random variable** X is assigned the count of the first independent Bernoulli(p) trial on which a 1 (Yes) occurs. The RV notation is: $X \sim geom(p)$.
   a) Write the pmf for X (include the support).
   b Write the formulas for E[X] and Var(X).

2. Consider random variable $X \sim geom(0.8)$.
   a) Compute the probability the first Yes occurs on the second trial.
   b) Use the complement rule to compute $P(X > 2)$.

3. A **Poisson random variable** X is often used to count the number of randomly occurring events on an interval of space or time. The RV notation is: $X \sim Pois(\mu)$.
   a) Write the pmf for X (include the support).
   b) Write the formulas for E[X] and Var(X).
   c) Write the R commands used to evaluate the pmf and cdf.

4. Consider random variable $X \sim Pois(3.24)$.
   a) Compute $P(X = 5)$.
   b) Compute Var(X) and SD(X).

Stop. A commuter encounters a single traffic light on her way to work. Each morning there is a 40% chance she will have to stop and wait at the light.
   a) If we start counting on a random morning, what is the chance she will stop at the light for the first time on the fifth day? Clearly define an appropriate random variable X.
   b) What is the chance her first stop will be on one of the first two days?
   c) What is the expected day of her first stop?
   d) Compute SD(X).

Aim. Rocket bombs launched toward London by Germany near the end of World War II were rumored to possess sophisticated aiming devices. Allied military intelligence tested these claims by dividing a map of London into 576 regions of 0.25 square kilometers and recording the number of actual bomb hits in each region. Because the Poisson distribution is based on the random occurrence of independent events, a close match between the distribution of actual counts and the predictions based on the Poisson distribution would provide strong evidence that the bombs were landing at random locations (i.e. no aiming device). The observed data:

Bomb hits	0	1	2	3	4	5
# of regions	229	211	93	35	7	1

   a) Compute the mean number of bomb hits per region.
   b) Use R to help determine if the Allies found evidence of a German aiming device.

Arrivals. Aircraft arrive at a certain airport according to a Poisson process at the rate of twenty per hour, so the mean number of arrivals during t hours is 20t. Use R to compute the probabilities
   a) What is the chance that exactly 22 aircraft arrive during a one-hour period?
   b) What is the chance that at least 22 aircraft arrive during a 90 minute period?

Lesson 10 – Exam one review.

1. **Exam details.**

2. **Topics.**

3. Consider a sample of 55 values with mean 5.3 and variance 11.6. Also, the five-number summary is: 0, 3, 4, 8, 12.
   a) Compute the median of the data.
   b) Compute the outlier fences.
   c) Compute the "within two standard deviations of the mean" boundaries.

4. Consider the pmf for the random variable X.

x	–2	0	1	3	5
p(x)	0.15	0.15	0.20	0.30	0.20

   a) Verify the validity of the pmf.
   b) Compute the mean and standard deviation of X.

Phones. Six phone lines are available for use at a call center. Let A be the event that exactly two phones are in use. How many phones are in use for each of the events?
   a) A
   b) A^C
   c) B = exactly 2 phones are not in use

Switches. Each of four independent, ordered switches are randomly in the off (0) or on (1) position at any given time. Consider the chance experiment that records the four switch positions at an arbitrary time. Carefully write the sample space for this experiment.

Quiz. A student correctly solves quiz problems 30% of the time and each different problem is an independent event. Let X represent the number of correct answers for this student on an 18 problem quiz.
   a) Use conventional notation to complete the distribution statement: X ~
   b) Carefully write the R code that would provide the probability that the student correctly solves exactly k problems.
   c) Use relevant R output to compute the probability the student correctly solves at least three problems.

```
> dbinom(0:9,18,.3)
 [1] 0.0016 0.0126 0.0458 0.1046 0.1681 0.2017 0.1873 0.1376 0.0811 0.0386
> dpois(0:9,.3)
 [1] 0.7408 0.2222 0.0333 0.0033 0.0003 0.0000 0.0000 0.0000 0.0000 0.0000
> dpois(0:9,1/.3)
 [1] 0.0357 0.1189 0.1982 0.2202 0.1835 0.1223 0.0680 0.0324 0.0135 0.0050
```

Baseball. In a randomly selected game, there is a 70% chance player A gets at least one hit, a 76% chance player B gets at least one hit, and a 64% chance both of them get at least one hit.
   a) What is the probability player A goes hitless if we know player B gets at least one hit?
   b) What is the probability player A goes hitless if we know player B goes hitless?

Lesson 11

1. Probabilities for **continuous random variables** are equivalent to areas. Explain this with a graph for random variable X with valid f(x).

2. Graph the function $f(x) = (9 - x^2)$ on $0 < x < 3$. Next, find k so that

$$f(x) = k(9 - x^2) \text{ on } 0 < x < 3$$

has area = 1 (i.e. 100% of the probability). The value k is called a normalizing constant.

3. Consider the **probability density function (pdf)** for the continuous random variable X.

$$f(x) = (1/18)(9 - x^2) \text{ on } 0 < x < 3$$

   a) What is the support of X?
   b) Graph this distribution.
   c) The cumulative distribution function (cdf) is the cumulative probability from the left end of the support, i.e. $F(c) = P(X \leq c)$. Compute F(1).

4. Consider a rectangle sitting on the interval from 12 to 22.
   a) Find the height of the rectangle with area = 1.
   b) For the rectangle in part (a) find the area between 15 and 19.

5. A **uniform random variable** X is a specific type of continuous RV. Its probability distribution has a rectangular shape.
   a) Explain the notation: $X \sim unif(A, B)$
   b) Write the pdf for X.

Fluid. Neuroscientists studying brain functional connectivity profiles are interested in a measure called fluid intelligence (gF). For a certain population of individuals, gF scores follow a uniform distribution on the interval from 7.5 to 20. Let X be the gF score for an individual.
   a) Determine the pdf of X and sketch the density curve.
   b) What is the probability an observed gF score is at most 10?
   c) For any k such that $7.5 < k < k + 2 < 20$, what is the probability an observed gF score is between k and k + 2?
   d) What percentage of gF scores are greater than 25?
   e) Compute the probability that an observed gF score is less than 25.

Washers. Production quality managers know that thicknesses of rubber washers vary between 2 and 6 mm. Let X be washer thickness with the pdf: $f(x) = (3/80)(x)(6 - x)$ on $2 < x < 6$.
   a) What is the support of X?
   b) Create a graph of this distribution.
   c) Compute the percentage of washers with thicknesses less than 4 mm.

Lesson 12

1. Consider the pdf for the random variable X.

$$f(x) = (1/18)(9 - x^2) \text{ on } 0 < x < 3$$

   a) Write the expected value formula and compute E[X].
   b) Write the LOTUS formula and compute $E[X^2]$.
   c) Write the variance formula and its shortcut formula. Compute Var(X).
   d) Compute the standard deviation SD(X).
   e) Compute the 60th **percentile** of X.

2. Consider random variable $X \sim unif(A, B)$. Develop formulas for E[X] and Var(X).

Fluid 2. Neuroscientists studying brain functional connectivity profiles are interested in a measure called fluid intelligence (gF). For a certain population of individuals, gF scores follow a uniform distribution on the interval from 7.5 to 20. Let X be the gF score for an individual.

   a) What is the cdf of X?
   b) What are the mean and variance of the gF scores?
   c) What percentage of gF scores are within one standard deviation of the mean?

Washers 2. Production quality managers know that thicknesses of rubber washers vary between 2 and 6 mm. Let X be washer thickness with the pdf: $f(x) = (3/80)(x)(6 - x)$ on $2 < x < 6$.

   a) Compute the median of X.
   b) Compute E[X].
   c) Compute $E[X^2]$, Var(X), and SD(X).

Lesson 13

1. A **normal (Gaussian) random variable** X is a specific type of continuous RV. Its probability distribution is bell-shaped and the RV notation is: $X \sim N(\mu, \sigma^2)$. Many populations of X values are assumed to be Gaussian. Write the pdf for X.

2. A **standard normal random variable** is a specific normal RV with notation: $Z \sim N(0, 1)$.
   a) Explain Φ notation for the cdf of Z.
   b) Write the **Z-score** formula that converts a non-standard normal random variable X into a standard normal random variable Z.
   c) Draw shaded distribution pictures for each situation on the "Standard normal distribution" page of the R printout.

3. An **exponential random variable** X is a specific type of continuous RV. It is commonly used to measure distances or times between randomly occurring events in space or time and its probability distribution has a positive skew shape. The RV notation is: $X \sim expon(\lambda)$.
   a) Write the pdf for X.
   b) Write the formulas for E[X] and Var(X).
   c) Develop the cdf formula for X.

4. Consider random variable $X \sim expon(2)$.
   a) Compute E[X].
   b) Compute F(3).

Reactor. Safety inspectors monitor a certain waste reactor's influent concentration. Assume concentration values X are Gaussian with mean 30 and standard deviation 6.
   a) What is the probability the concentration exceeds 25?
   b) Compute $P(|X - 30| < 8)$.
   c) Compute the 85th percentile of concentration values.

Airdrop. Certain humanitarian aid parachutes are designed to automatically open at 220 meters above ground. Model the actual opening altitude as a normal random variable with mean 220 m and standard deviation 40 m. Delivery damage occurs if the parachute opens at an altitude less than 120 m. What is the probability of damage to at least one payload in five independently dropped parachutes?

Lesson 14

1. Consider the simultaneous probability distribution for two random variables X and Y. For the discrete case the **joint pmf is p(x,y)** and for the continuous case the **joint pdf is f(x,y)**.
   a) Sketch examples of p(x,y) and f(x,y) with the support in the XY-plane.
   b) Explain how to find the marginal pmf of the single variable X from p(x,y).
   c) Explain how to find the marginal pdf of the single variable Y from f(x,y).
   d) Explain the importance of joint distributions using an example.

2. Two random variables are **independent** if information about the observed value of one random variable provides no information about the value of the other random variable.
   a) Write the definition for independent discrete random variables X and Y.
   b) Write the definition for independent continuous random variables X and Y.

3. Compute the marginal pmfs of X and Y and provide evidence that X and Y are not independent.

	Y = –1	Y = 1
X = 0	0.3	0.4
X = 2	0.1	0.2

4. Consider independent RVs X and Y with marginal distributions: $p_X(0) = 0.8$, $p_X(1) = 0.2$, and $p_Y(0) = 0.2$, $p_Y(1) = 0.3$, $p_Y(2) = 0.5$.
   a) Display the joint pmf of (X,Y) in a joint probability table.
   b) Compute the probability that the sum X + Y is at most one.

5. Consider the joint pdf: $f(x,y) = k(x + y)$ on $0 < x < 2$ and $0 < y < 2$.
   a) Find k that makes this a valid joint pdf.
   b) Compute the marginal pdfs of X and Y. (Hint: use symmetry)
   c) Are X and Y independent?
   d) What is the probability X and Y are both greater than one?

Lesson 15

1. Write the LOTUS formulas for joint distributions.

2. Consider the joint pmf.

	Y = –1	Y = 1
X = 0	0.3	0.4
X = 2	0.1	0.2

   a) **Covariance Cov(X, Y)** measures the association between random variables X and Y. Write the Cov(X, Y) formula and shortcut. Compute Cov(X, Y) for this situation.
   b) Verify: Cov(X, X) = Var(X)
   c) **Correlation Corr(X, Y)** measures the strength of the linear relationship between X and Y. Write the Corr(X, Y) formula and compute Corr(X, Y) for this situation.
   d) List two useful correlation properties.

3. Consider independent RVs X and Y with marginal distributions: $p_X(0) = 0.8$, $p_X(1) = 0.2$, and $p_Y(0) = 0.2$, $p_Y(1) = 0.3$, $p_Y(2) = 0.5$.
   a) Compute the covariance between X and Y.
   b) Compute the correlation coefficient for X and Y.

4. Consider the joint pdf: $f(x,y) = (1/8)(x + y)$ on $0 < x < 2$ and $0 < y < 2$.
   a) Compute the covariance between X and Y.
   b) Compute the correlation coefficient for X and Y.

5. A **random sample** is a collection of n independent RVs, each with the same pmf or pdf. Explain common notation for a random sample.

6. Consider a random sample of size two from the following distribution.

x	1	4
p(x)	0.7	0.3

   a) A **statistic** is any quantity whose value can be calculated from sample data. Because its value varies from sample to sample, a statistic is a random variable. Consider the statistic: $W = X_1 + 2X_2$. What is the support of the new random variable W?
   b) The pmf or pdf of a statistic is called a **sampling distribution**. Use the joint pmf of X_1 and X_2 to compute the sampling distribution of W. (i.e. find W's pmf)
   c) Compute the mean and variance of W.

Lesson 16

1. Many common statistics are **linear combinations** of random variables.
   a) Define a linear combination of n random variables.
   b) State the expected value property for a linear combination of RVs.
   c) State the variance property for a linear combination of RVs.
   d) How does the part (c) formula simplify if the RVs are independent?
   e) State the property for a linear combination of normal, independent RVs.

2. Write equivalent expressions for the following expected values and variances of linear combinations of random variables X, Y, and W. Do not assume independence.
   a) E[3X + 2Y]
   b) E[3X – 2Y]
   c) E[X + Y + W]
   d) Var(3X + 2Y)
   e) Var(3X – 2Y)
   f) Var(X + Y + W)
   g) How do the answers to parts (d), (e), and (f) change if X, Y, and W are independent?

Shipping. A company ships three sizes of a certain product and wants to predict its weekly shipping weight. The small size weighs 5 pounds, the medium size weighs 7 pounds, and the large size weighs 10 pounds. Let S, M, and L denote the number of each size shipped during a random week.

	S	M	L
E[·]	100	200	150
SD(·)	6	8	4

   a) Assume S, M, and L are independent and calculate the expected value, variance, and standard deviation of the total weight shipped.
   b) Will the calculations be correct if S, M, and L are dependent? Explain.

Ethanol. How does ethanol in gasoline affect fuel economy? Four automobiles of the same type are driven on the same long-distance route and mpg X is recorded for each car. Three cars use pure gasoline and the fourth car uses an ethanol-blend gasoline. Assume the RVs are independent and normal with $X_1, X_2, X_3 \sim N(24, 4)$ and $X_4 \sim N(22, 3)$.
   a) Will Car 1 (pure gas) have a higher mpg than Car 4 (ethanol-blend)?
   b) Compute the probability that the mpg of Car 1 is greater than the mpg of Car 4.
   c) Compute the probability that the average mpg of the three pure gasoline cars is greater than the mpg of the one ethanol-blend car.

Lesson 17

1. As explained in Lesson 15, each statistic calculated from a sample creates a new population called a **sampling distribution**. Consider a random sample of size n = 3 from a N(4,5) distribution.
   a) Describe the sampling distribution for the statistic $Y = X_1 + X_2 + X_3$.
   b) Describe the sampling distribution for the statistic $W = (X_1 + X_2 + X_3)/3$.
   c) Generalize the results of parts (a) and (b) for the **sample sum** and **mean** statistics.

2. The standard deviation of a statistic's sampling distribution is called the **standard error of the statistic**. Write the standard errors of the sample sum and mean statistics.

3. The **central limit theorem (CLT)** is useful for establishing the approximate normality of the sampling distribution of the sample mean if the original population distribution is unknown and the sample size n > 30. Do we need the CLT for the sample mean statistic if the original population distribution is normal?

4. For each random sample, write the sampling distribution of the sample mean statistic and compute the standard error.

	Original population	n	(X-bar) ~	SE(X-bar)
a)	X ~ N(7, 5)	9		
b)	X ~ N(7, 5)	36		
c)	X ~ (7, 5), unkn. shape	9		
d)	X ~ (7, 5), unkn. shape	36		

Rivets. Two parts of a vehicle are attached by a seam of rivets. An individual rivet's breaking strength is a random variable with mean 10,000 psi and standard deviation 500 psi.
   a) Write the sampling distribution for the sample mean of a random sample of 40 rivets.
   b) The standard deviation of a statistic is called the standard error of the statistic. What is the standard error of the statistic in part (a)?
   c) Compute the probability that the sample mean of a random sample of 40 rivets is between 9900 and 10,200 psi.
   d) If the sample size had been 15 rather than 40, could we have calculated the part (c) probability from the given information? Explain.

Uniform. Consider a random sample of size 50 from a uniform(10,20) distribution.
   a) What are the mean and variance of the parent population?
   b) Write the sampling distribution and standard error of the sample mean statistic.
   c) What is the chance that the sample mean is less than 14?

Lesson 18

1. Consider a random sample of size n from an X ~ Bernoulli(p) distribution.
   a) Express this sample using common notation.
   b) Write the mean and variance of an individual X.
   c) Let p-hat be the sample mean statistic. Compute its mean and variance.
   d) The **point estimator** θ-hat for unknown parameter θ is **unbiased** if E[θ-hat] = θ.
      Is p-hat an unbiased estimator of the parameter p? Explain.
   e) The CLT applies in this situation if $np \geq 10$ and $n(1 - p) \geq 10$.
      Assuming these conditions are met, what is p-hat's sampling distribution?
   f) What is the standard error of p-hat?

2. A point estimator is a statistic whose single value represents our "best guess" about an unknown parameter. What is an **interval estimator**?

3. Develop the **confidence interval (CI)** for an unknown population mean for a Gaussian distribution.
   a) **Critical values** are a function of the CI's **confidence level** (i.e. the success rate of the CI formation method). Use R and the standard normal distribution to find the Z critical values for a confidence level of 95%.
   b) Use part (a) to derive the 95% CI formula.
   c) Generalize the part (b) CI result to any confidence level.
   d) Why not use a 100% confidence level?

   * In this class, **use 95% as the default confidence level** if no value is provided.

4. What **sample size** is required if we want a 95% CI for a population mean with a margin of error (m.o.e.) equal to 3? Assume the population variance is 76.

Particles. To estimate the unknown parameter θ, a particle physicist will select a random sample of size n from the pdf: $f(x) = 0.25(2 + \theta x)$ for $-1 < x < 1$ where $-1 < \theta < 1$. Show that 6 times the sample mean is an unbiased estimator of θ.

Blades. Energy scientists suspect that insects caught on the leading edges of wind-turbine blades can drastically reduce power under certain conditions. They model these power values using a normal distribution with population standard deviation 20 kW.
   a) Compute a 95% confidence interval for the true mean power under these conditions if n = 25 and the sample mean is 402 kW.
   b) Compute an 82% confidence interval for the true mean power under these conditions if n = 81 and the sample mean is 402 kW. Interpret the result.
   c) What sample size is required if we want a 90% CI for μ that is 5 kW wide?

Lesson 19

1. **Student's t distribution** is a pdf commonly used when the sample standard deviation s estimates the population standard deviation σ. A particular t distribution is identified by its **degrees of freedom (df)** value.
   a) Draw the graph of t(df).
   b) Write the R commands used to evaluate the cdf and find percentiles.

2. Explain how the confidence interval for an unknown population mean changes when the population standard deviation is also unknown.

3. For each random sample, indicate which **critical value** would be used when creating a CI for an unknown population mean.

	Original population	n	std. dev.	z_{cv} or t_{cv} ?
a)	$X \sim N(\mu, 4)$	any	$\sigma = 2$	
b)	$X \sim N(\mu, \sigma^2)$	any	s = 4.1	
c)	$X \sim (\mu, 9)$, unkn. shape	36	$\sigma = 3$	
d)	$X \sim (\mu, \sigma^2)$, unkn. shape	44	s = 11	
e)	$X \sim (\mu, 7^2)$, unkn. shape	22	$\sigma = 7$	
f)	$X \sim (\mu, \sigma^2)$, unkn. shape	17	s = 8	

4. Consider the summary information for a random sample from a normal population.

n = 25	X-bar = 16	s = 3

   a) Compute the 95% CI for the unknown population mean.
   b) A **prediction interval (PI)** for X_{new} is an interval estimate for a new individual X value randomly selected from a Gaussian distribution. Explain the two variability components and write the PI formulas.
   c) Compute the 95% PI for a new X value.

Irisin. Increased levels of the hormone irisin have been found in the blood of exercising humans. Under certain experimental conditions, a random sample of six subjects has a mean irisin level of 4.3 mg/ml with standard deviation 0.4 mg/ml. The population of measurements is Gaussian.
   a) Compute the 90% confidence interval for the true mean irisin level under these experimental conditions.
   b) Compute the 90% prediction interval for the irisin level of a single new subject under these experimental conditions.

Screws. A machinist is examining a population of cap screw body diameters. The values are Gaussian with variance 0.0009 mm^2. The sample mean of 12 screw diameters is 12.60 mm.
   a) Compute an interval estimate for the true mean cap screw diameter..
   b) Compute an interval estimate for the diameter of the next cap screw selected.

Lesson 22

1. A confidence interval is one type of statistical inference. Another type of inference is the **hypothesis test**. Use the six-step **NATDRC** process for all hypothesis test questions on homework and exams in this class.

N. Null hypothesis
A. Alternative hypothesis
T. Test statistic
D. Distribution
R. Result
C. Conclusion

a) Explain the basic ideas behind a hypothesis test for an unknown population mean.
b) Explain null and alternative hypotheses.

2. Show **hypothesis** steps NA of NATDRC for each conjecture.
   a) The true mean reaction time is more than 260 ms.
   b) The true mean horsepower is less than 707.
   c) The true mean cost differs from $345.
   d) The true mean inflation pressure is 35 psi.

3. In the **conclusion** step C we: **Reject H_0** or **Fail to reject (FTR) H_0.** We also use the word "suggests" instead of "proves" because wrong conclusions are possible, even when we apply correct methods.
   a) Explain **Type I** and **Type II errors**.
   b) Explain the significance level of a NATDRC test.

   * In this class, **use 0.05 as the default significance level** if no value is provided.

Roads. A certain state's department of transportation has a road-safety index based on visibility and road surface conditions. As long as the true mean safety index for an interstate highway segment remains above 25, the road stays open. Suppose a winter storm is currently passing through the state. Should the interstate remain open?
   a) Write the appropriate hypotheses for this situation.
   b) Describe a Type I error in the context of this problem.
   c) Describe a Type II error in the context of this problem.

Lesson 23

1. Steps TD provide evidence for the NATDRC test. The **test statistic** (step T) is computed from sample data. Explain step D.

2. A **p-value** is the probability of obtaining a test statistic at least as extreme as ours, if H_0 is true. (Know this definition.) Explain p-value use in steps RC of NATDRC.

3. Write careful **conclusions** (step C) for each situation, using a 0.05 significance level.
   1) Reject H_0; evidence suggests … [Ha in context]
   2) FTR H_0; evidence suggests … [μ_0 in context] … is plausible

	Step A	Step R	Step C
a)	H_a: $\mu \neq 7$	p-value = 0.04 < 0.05	
b)	H_a: $\mu > 13$	p-value = 0.03 < 0.05	
c)	H_a: $\mu < 4.1$	p-value = 0.12 > 0.05	
d)	H_a: $\mu \neq 5.6$	p-value = 0.09 > 0.05	

Reaction. Scientists believe the true mean reaction time to a certain stimulus is more than 260 ms. The mean of a random sample of 20 reaction times is 270 ms. Assume reaction times are normally distributed with population standard deviation 40 ms. Test the reaction time conjecture using NATDRC.

Horsepower. The Dodge Challenger Hellcat's engine is reported to generate 707 hp. A consumer group believes the true horsepower is less than the listed value and conducts its own analysis using a dynamometer. A random sample of 35 measurements has a mean of 702 hp with standard deviation 30 hp.

a) Use NATDRC to test the consumer group's claim.
b) Based on our conclusion, what type of error might we have made?

Inflation. New tires of a certain type have a recommended inflation pressure of 35 psi. A random sample of 20 tires has a mean inflation of 36.2 psi with standard deviation 2.3 psi. Assume the tire inflation values follow a normal distribution and use NATDRC to determine if the true mean inflation pressure differs from 35 psi.

Lesson 24

1. For a two-sided hypothesis test, confidence interval information can replace steps TDR. Consider a random sample with n = 40, X-bar = 20, and s = 4.
   a) Compute the 95% confidence interval for the population mean.
   b) Use the **CI and NATDRC** to test if the population mean differs from 22.

2. Consider a random sample in which 30 of 100 items are green. Does the evidence suggest the population proportion of green items differs from 0.35? Use **NATDRC for p**.

Kissing. Roughly 67% of all humans have a dominant right foot, ear, or eye. Do couples also have a right-sided dominance in kissing behavior? A 2003 article reports that both people in 80 of 124 random couples lean more to the right when kissing. Does the result of the experiment suggest that the 67% figure is plausible for kissing behavior? Use NATDRC.

ADHD. Attention deficit hyperactivity disorder (ADHD) is now being studied in the elderly as a possible cause of increased forgetfulness. In the Netherlands, 1494 adults over 60 years-old are screened with an ADHD questionnaire and a portion of these adults also undergo a structured diagnostic interview. A total of 42 individuals are assessed with syndromatic ADHD.
   a) Compute the 99% confidence interval for the true proportion of elderly Dutch with syndromatic ADHD.
   b) Does this suggest that the true percentage differs from 3%? Use part (a) and NATDRC.

Robots. A Chinese cell-phone manufacturer exclusively uses robots in one of its factories. A random test run results in 23 defective assemblies out of 500. Does this data suggest that the true percentage of defectives is less than 5%? Use NATDRC.

Lesson 25

1. Consider independent random samples of size 25 from $X_1 \sim N(10, 4)$ and $X_2 \sim N(7, 2)$.
   a) What is the sampling distribution of the sample mean for each population?
   b) Write the sampling distribution of the difference between the two sample mean statistics.
   c) What is the standard error of the new statistic in part (b)?
   d) Write the general formula for the part (b) result.

2. List three cases for comparing **two unknown population means** and write corresponding confidence interval formulas.

Traffic. Common acceleration allows more cars through an intersection after a red light if the cars are adequately spaced. Traffic engineers design a study and collect car count data for two red light spacing arrangements. Assume normality for both populations of values and create a 95% confidence interval for the difference in true mean number of cars through an intersection for the two arrangements.

	n	Sample mean	σ
Spaced	20	39.9	1.7
Bunched	20	15.3	.8

Tools. Machinists want to know which grade of carbide has a longer lifetime in minutes for a certain tool component. Assume normality for both populations of tool lifetimes. Estimate the difference in true mean lifetimes for the two types of carbide inserts in a way that conveys information about precision and reliability.

	n	Sample mean	SE(mean)
Grade A	7	16.2	0.6
Grade B	7	13.9	0.9

Stopping. Two car braking systems are being tested for stopping distances under similar driving conditions. Assume both stopping distance distributions are normally distributed.

	n	Sample mean	s
System 1	8	126.8	5.4
System 2	10	119.3	5.0

   a) What is the Welch-Satterthwaite degrees of freedom formula and why do we need it here?
   b) Compute a 95% confidence interval for the difference between true mean stopping distance for System 1 and System 2.
   c) Use the result of part (b) and NATDRC to determine if the two population mean stopping distances differ.

Lesson 26

1. Compute the differences d_i for each ordered pair, then find the sample mean and standard deviation for each row.

	1	*2*	*3*	*4*	*5*	Mean	Std. dev.
Sample 1	80	50	160	100	60		
Sample 2	60	50	130	90	70		
d_i							

a) Assume independent samples from two populations and compute the 95% CI for the difference between population means. Use df = 7.
b) Assume dependent samples (**matched pairs**) and compute the 95% CI for the difference between population means. What df is used here?
c) Explain why matched pairs reduced the confidence interval width.

2. Consider independent random samples of 100 objects each from two populations. Forty objects in the first sample are purple and 52 objects in the second sample are purple. Compute the 95% confidence interval for the **difference in population proportions** of purple objects.

Corruption. When asked if they believe corruption is widespread in their government, 230 of 500 United Kingdom citizens say yes, compared with 264 of 600 Canadian citizens who provide the same answer. Compute the 95% confidence interval for the difference in true proportions.

Ratio. In some medical investigations the ratio of proportions is useful.
a) Explain the basic formulation of a large-sample confidence interval for this situation.
b) Apply the result of part (a) to the following shingles vaccine study. There are 642 shingles cases among the 19,247 individuals in the control (placebo) group, and 315 cases among the 19,254 who receive the actual vaccine. Does the evidence suggest the vaccine is useful?

Lesson 27

1. **Snedecor's F distribution** is a pdf commonly used for variance ratios. A particular F distribution is identified by its numerator df and denominator df values.
   a) Draw the graph of $F(df_1, df_2)$.
   b) Write the R command used to evaluate the cdf.

2. A **treatment** is a specific condition applied to objects in a study.
   a) Provide an example of each type of **comparative study**.
      1) **experimental**: the investigator controls treatment assignment
      2) **observational**: treatment groups exist or are self-selected
   b) Which type of carefully designed study can establish cause-and-effect?

3. Single-factor **analysis of variance (ANOVA)** checks for equality of k > 2 population (treatment) means. We assume normal populations, each with the same variance σ^2. Next, collect samples from each population and compute two estimates for σ^2 (MSTR and MSE).
   a) Write the appropriate NA steps for NATDRC.
   b) Explain why F = MSTR/MSE is the appropriate test statistic.
   c) Write the MSTR and MSE formulas for equal sample sizes.

4. Compute the ANOVA F statistic for the data. Common sample size, n = 6.

Treatment	1	2	3	4
Sample mean	15	24	12	21
Sample std. dev.	2	3	3	2

5. In many ANOVA situations (including situations with differing sample sizes), we use computer generated **ANOVA tables** to find the F statistic. Complete the analysis of variance (ANOVA) table with k = 3 and n_T = 15. Also, conduct the corresponding NATDRC test.

Source	df	SS	MS	F
Treatments		60		
Error (residual)		36		xx
Total			xx	xx

Cells. Regenerative biologists want to know if endothelial cell concentration (ECC) is the same in three different regions (A, B, C) of embryonic zebrafish. They study 14 distinct embryos for each region. We know SSTR = 1.476 and SST = 9.619. Create the ANOVA table. Are the population mean ECC values the same in each region? Provide evidence.

Lesson 28

1. When we Reject H_0 in the single-factor ANOVA F test, we usually want to make pairwise comparisons of population means. Explain why the **Tukey-Kramer (T-K) method** is needed. Also, provide the T-K confidence interval formula.

Flicker. Above a certain frequency, a flickering light source appears to be continuous. A 1973 study examined the critical flicker frequency (cff) in cycles/sec for subjects with one of three iris colors. Use NATDRC to test the equality of the true mean cff for the three eye colors. If relevant, use the Tukey-Kramer method to investigate the differences.

Color	1. Brown	2. Green	3. Blue
Sample size	8	5	6
Sample mean	25.6	26.9	28.2

```
          Df Sum Sq Mean Sq F value  Pr(>F)
trt        2  23.00  11.499   4.802   0.023
Residuals 16  38.31   2.394
```

Groups. Consider T-K method application for k = 5 population means with ordered sample means (by population index): 2, 3, 1, 5, 4. The underscore scheme shows lines under the groups: (2), (3, 1, 5), and (5,4). Write this report using standard notation and explain the CI results.

Lesson 30

1. Predict the value of a new Y based on the sample data.

Y	5	11	10	14

2. Next, suppose each sampled Y value has a corresponding X value. Create the **scatter plot** and use it to predict the value of a new Y when X = 50. Compare this value to the predicted Y in the previous problem.

X	20	30	40	50
Y	5	11	10	14

3. Sketch representative scatter plots for each X, Y relationship.
   a) Positive linear
   b) Negative linear
   c) Nonlinear
   d) No relationship - random points

4. **Regression analysis** examines the probabilistic relationship between two or more variables. Briefly explain the difference between **deterministic** and **probabilistic models**.

5. Describe the **simple linear regression (SLR)** model.

6. Describe the **true regression function** for the SLR model.

7. Assume the SLR model: $Y = 3 + 5X + \varepsilon$ with $\sigma = 4$.
   a) Write the true regression function.
   b) Interpret β_1.
   c) Compute the expected value of Y when X = 6.
   d) What percentage of Y values are less than 30 when X = 6.
   e) Consider two independent Y values observed at $X_1 = 8$ and $X_2 = 6$. Compute the probability the Y value at $X_1 = 8$ exceeds the Y value at $X_2 = 6$.
   f) Compute the 67th percentile of the standard normal distribution.
   g) For what X is the 67th percentile of Y's equal to 39?

Lesson 31

1. Define S_{xy} and provide the shortcut formula.

2. Explain the **estimated SLR equation** using **least squares (LS)**.

3. Explain how the **normal equations** lead to the LS estimators.

4. Compute the estimated SLR equation. Hint: X-bar = 35 and Y-bar = 10.

X	20	30	40	50
Y	5	11	10	14

5. We often use SLR to predict Y at a specific X. Before we can safely make predictions at arbitrary values of X, we first need to assess the model's usefulness by assessing the variation of the observed Y values.
   a) A **residual** is the vertical distance from an observed point to the estimated regression line. Write this as a formula and draw a picture.
   b) Compute residuals for the data points in the previous problem.
   c) Explain: **SST = SSR + SSE** as it relates to the variation of the observed Y values.
   d) Define and compute SST, SSR, and SSE.
   e) Complete the **SLR ANOVA table**.

Source	df	SS	MS	F
Regression				
Error (residual)				xx
Total			xx	xx

   f) Indicate the distribution for the F test statistic.
   g) What is the estimated variance of the random error term?
   h) Compute the **estimated residual standard error, s**.

6. Consider the computer output for a SLR model.

```
            Estimate Std. Error
(Intercept)     3.1       1.20
x               2.0       0.18

          Df  Sum Sq Mean Sq
x          1  156.72  156.72
Residuals  3    3.91   1.304
```

   a) Write the estimated SLR equation.
   b) Interpret b_1.
   c) Create the SLR ANOVA table.
   d) What is the sample size n?
   e) What is the estimated variance for the random error term?
   f) Compute the estimated residual standard error.

Lesson 33

1. Suppose X and Y are linearly related with $\beta_1 \neq 0$. Do we expect the same Y value for two different X values? Explain with a graph.

2. Explain the **SLR model utility test**.

3. Describe the **sampling distribution of β_1-hat**.

4. Write the **confidence interval (CI) for β_1**.

5. The **coefficient of determination (r^2)** is the proportion of Y variation accounted for by its assumed relationship with X. This is one measure of how well the estimated regression equation fits the data.
   a) Write the formula for r^2.
   b) Explain the **sample correlation coefficient (r)**.

Search. An Army psychologist studies the visual search process by noting the time Y in seconds required to locate a critical symbol placed on line X of a page of text. We have data from 14 trials where the sample mean line placement is 24.8 and S_{xx} = 1450.4. Assume the SLR model. We also know b_0 = 0.32, b_1 = 0.61, SSR = 536.58, and SSE = 25.0.
   a) Write the estimated SLR equation.
   b) Create the complete SLR ANOVA table.
   c) Is the SLR model useful? Provide evidence.
   d) Compute the 95% confidence interval for β_1. Could we use this result to conduct a model utility test?
   e) Compute the t test statistic for the NATDRC model utility test.
   f) What percentage of variation in time values is accounted for by the assumed relationship with critical symbol line placement?
   g) Compute the sample correlation coefficient.

Lesson 34

1. If we Reject H_0 in the SLR model utility test, we often want to predict Y at a given X*. The star indicates the X value can differ from the observed X values. The Y-hat value is a point estimate.
   a) Explain the **confidence interval (CI) for E[Y] at X***.
   b) Explain the **prediction interval (PI) for Y_{new} at X***.

2. Consider the summary information and computer output for a SLR model. The observed X values range from 2 to 10.

n = 10	$\Sigma x_i = 60$	$\Sigma x_i^2 = 440$	SSR = 1228.0	SSE = 373.7

```
            Estimate Std. Error t value Pr(>|t|)
(Intercept)    1.15      5.068    0.23    0.826
x              3.92      0.764    5.13    0.001
```

   a) Explain what each **bold** number represents.
   b) Write the estimated SLR equation.
   c) What is the p-value for the SLR model utility test? Hint: t test.
   d) Create the SLR ANOVA table.
   e) **Extrapolation** occurs when we predict a Y value for an X* outside the range of observed X values. Explain why we should avoid extrapolation.
   f) If reasonable, compute an interval estimate for Y for the next value of X equal to 16.
   g) If reasonable, compute an interval estimate for Y for the next value of X equal to 5.

Search 2. An Army psychologist studies the visual search process by noting the time Y in seconds required to locate a critical symbol placed on line X of a page of text. We have data from 14 trials where the sample mean line placement is 24.8 and $S_{xx} = 1450.4$. Assume the SLR model. We also know $b_0 = 0.32$, $b_1 = 0.61$, SSR = 536.58, and SSE = 25.0.
   a) Compute an interval estimate for the mean time required to locate a critical symbol for all trials when the symbol is placed on line 30.
   b) Compute an interval estimate for the time required for the next individual to locate the critical symbol when it is placed on line 30.

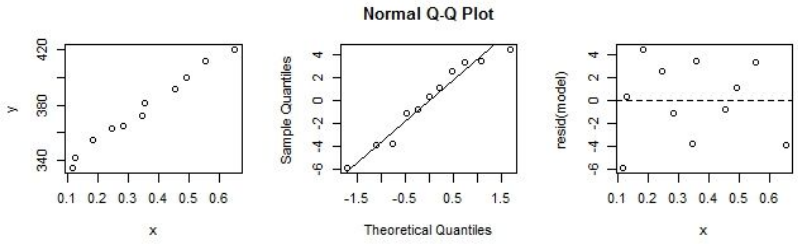
Lesson 35

1. **Regression diagnostics** are part of the SLR analysis process. Specifically, we use the residuals to check assumptions made about the random error term. List the assumptions to check.

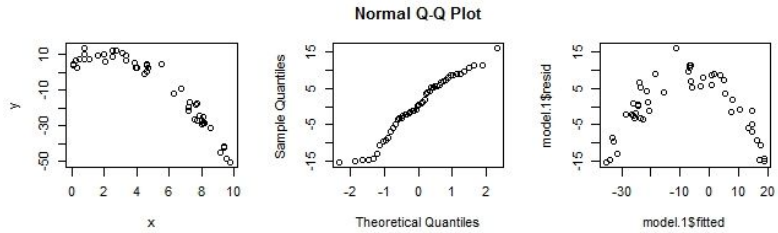
2. Many studies assume sample data from a normal distribution. To evaluate this assumption we pair the ordered sample data values with "ideal" data values from a standard normal distribution to create a **normal probability plot**. What pattern indicates the normal distribution assumption seems reasonable?

3. We assume our random error term has mean zero and constant variance for Y distributions at different X values. Provide representative **residual plots** for the following situations:
   a) E[ε] = 0 and constant variance assumptions seem reasonable.
   b) Evidence that a different model should be used.
   c) Evidence of non-constant error term variance.

4. Assess the reasonableness of the error term assumptions based on the residual plot and the normal probability plot of the residuals.



5. Assess the reasonableness of the error term assumptions based on the residual plot and the normal probability plot of the residuals, under the SLR model.



   a) The SLR model utility test p-value < 0.05, but the plots suggest a different model. Explain the **polynomial (quadratic) regression** model and assumptions. For the selected model, n = 50, SSR = 17256.0, SSE = 409.9, and the model utility test p-value < 0.05.

```
            Estimate Std. Error
(Intercept)     5.4       1.10
x               4.1       0.55
x.sq           -1.0       0.06
```

   b) Write the estimated regression equation.
   c) Compute the residual for the ordered pair (8, −30).
   d) Compute and interpret the **coefficient of multiple determination**.
   e) Should we predict Y when X = 14? Explain.
   f) What X is related to the largest estimated mean Y?
   g) Compute the 95% CI for β_1.

Lesson 36

1. Describe the **general linear multiple regression (MR)** model.

2. Assume the MR model: $Y = 3 + 5X_1 - X_2 + 2X_3 + \boldsymbol{\varepsilon}$ with $\boldsymbol{\sigma} = 14$.
   a) Write the true regression function.
   b) Interpret β_1 and β_2.
   c) Compute the expected value of Y when $X_1 = 6$, $X_2 = 10$, and $X_3 = 3$.
   d) What percentage of Y values are less than 30 when $X_1 = 6$, $X_2 = 10$, and $X_3 = 3$.
   e) Consider two independent Y values observed at $(X_1, X_2, X_3) = (6, 10, 3)$ and $(7, 20, 2)$. Compute the probability the Y value at (6, 10, 3) exceeds the Y value at (7, 20, 2).

3. Assume the model: $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon$. The sample size is 40.

```
            Estimate Std. Error
(Intercept)    2.89      2.242
x1             2.18      0.300
x2             0.85      0.302
```

   a) Write the estimated regression equation.
   b) Compute the residual for the sample point $(Y, X_1, X_2) = (14, 2, 5)$
   c) Interpret b_2.
   d) Complete the **MR ANOVA table**.

Source	df	SS	MS	F
Regression		1948.77		
Error (residual)		1033.30		xx
Total			xx	xx

   e) What is the estimated variance of the random error term?
   f) Conduct a NATDRC model utility test.
   g) Predict Y when $X_1 = 3$ and $X_2 = 8$.
   h) Compute and interpret the coefficient of multiple determination.

Lesson 37

1. Describe the use of **interaction** and **indicator variables** to fit simultaneous linear models.

2. Assume the model: $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 X_2 + \varepsilon$, where $X_2 = 0$ for population A and $X_2 = 1$ for population B. We know $b_0 = 3$, $b_1 = 2$, $b_2 = 5$, and $b_3 = -4$.
   a) Write the estimated regression equation.
   b) Write the estimated regression equation for population A.
   c) Write the estimated regression equation for population B.
   d) Graph the equations from parts (b) and (c) on $0 < X_1 < 3$.
   e) For what value of X_1 are the Y-hat values the same for each population?

3. Explain the difference between SLR and **simple logistic regression** using graphs. Provide some basic details about simple logistic regression.

Hurricane. A safety engineer uses simple logistic regression to study the relationship between occupancy years X and evacuation decision Y, under a major hurricane forecast. The value Y = 0 if the occupant stays and Y = 1 if the occupant evacuates. The model utility test p-value < 0.05.

```
            Estimate Std. Error
(Intercept)     2.83      0.986
x              -0.24      0.066
```

   a) Write the estimated logistic response function and compute the probability that a homeowner of 10 years will evacuate.
   b) Compute and interpret the estimated **odds** that a homeowner of 10 years will evacuate.
   c) Develop a formula for the estimated **odds ratio**. Compute and interpret the estimated odds ratio for this model.

Sonic. A mechanical engineer at a certain sonic power tool company studies the relationship between a heat index Y and the operation setting X_1 for two extremely useful tools. Assume the model: $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon$, where $X_2 = 0$ for the sonic screwdriver and $X_2 = 1$ for the sonic blaster. The model utility test p-value < 0.05.

```
             Estimate Std. Error
(Intercept)       9.9      0.687
x1                5.0      0.119
x2                9.6      0.495
```

   a) Write the estimated regression equation.
   b) Compute separate equations for the estimated mean heat index for each sonic tool.