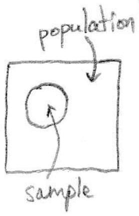
Lesson 1 - Solution

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 population is the collection of Xs for all possible objects/people of interest.
b) A sample is a subset of the population of Xs.
c) Explain parts (a) and (b) with a picture.
d) A simple random sample has characteristics similar to the population and is useful when we want to make inferences.
e) List some advantages of a **random sample.**

Also, measurements are independent

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

Quantitative & Qualitative

Type. List the data type for each measurement.

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

Desks. Seat and desk purchases for a new CSU classroom building depend on student height X.

a) What is the population of interest? All CSU students
b) Why not measure the height of each individual in the population? Too many students!
c) Why is the basketball team not a representative sample? Most likely, they are significantly taller than the rest of the students
d) Why would a simple random sample of students be a good choice?
Gives all students the same chance of being selected.

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?

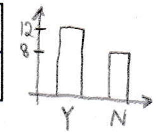
There are differences in each replicate. Whether these differences are caused by the physicist or not, this "error" impacts our results each time.

Lesson 2 - Solution

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

Y	N
12	8

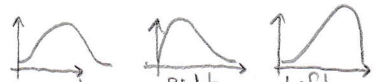
Y	Y	N	Y	N	N	Y	Y	Y	N
Y	N	N	N	Y	Y	Y	N	Y	Y



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

C = center, U = unusual points, S = Shape, S = Spread

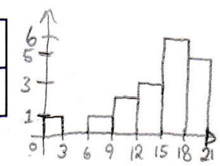
3. Sketch examples of the common distribution shapes.



Symmetric Right Skewed Left Skewed

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



Left Skewed

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

1 | 24677
2 | 023569
3 | 2378
4 | 15668

Roughly Symmetric

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**; $\bar{X} = 3$
b) $S_{xx} = \sum(X_i - \bar{X})^2 = 42$
c) Sample **variance**; $S^2 = \frac{S_{xx}}{n-1} = \frac{42}{3} = 14$
d) Sample **standard deviation**; $S = \sqrt{14}$
e) Sample **coefficient of variation (CV)**

$\hat{CV} = \frac{S}{\bar{X}} = \frac{\sqrt{14}}{3}$

f) Discuss the units of each statistic.

$\bar{X}$: unit S_{xx}: unit^2 S^2: unit^2 S: unit $\hat{CV}$: none

Centered. Prove the sum of $\underbrace{\text{centered}}_{X_i - \bar{X}}$ data values is always zero.

$\sum(X_i - \bar{X}) = \sum X_i - \bar{X} \cdot \sum 1 = \sum X_i - n\bar{X} = \sum X_i - n(\frac{1}{n}\sum X_i) = \sum X_i - \sum X_i = 0$

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. Bar Graph

Yes, she will be happy because 68% of surveyed people use the brand at least some of the time.

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. No, it is skewed right. There are higher counts in the classes with lower values. A histogram will also reveal this.

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 $\underbrace{\text{variance}}_{S^2}$.

$\hat{CV} = \frac{S}{\bar{X}} \Rightarrow .15 = \frac{S}{40} \Rightarrow S = 6 \Rightarrow S^2 = 36$

Lesson 3 – Solution

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

$S_{xx} = \Sigma(x_i - \bar{x})^2 = \Sigma(x_i^2 - 2\bar{x}x_i + \bar{x}^2) = \Sigma x_i^2 - 2\bar{x}\Sigma x_i + \Sigma \bar{x}^2 = \Sigma x_i^2 - 2\bar{x}\cdot n\bar{x} + n\bar{x}^2 = \Sigma x_i^2 - n\bar{x}^2$

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. $\bar{X} = \frac{170}{17} = 10$; $s^2 = 1844 - 17(10)^2 = 144$
   b) Compute and interpret the **Z-score** for the data value X = 14.8. $z = \frac{14.8-10}{144} = 1/30$
   c) Compute the "within 2 standard deviations of the mean" boundaries.

   14.8 is $\frac{1}{30}$ of s above $\bar{x}$

   $\bar{X} - 2s = -14$, $\bar{X} + 2s = 34$

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. 23
   b) Compute the **five-number summary**. 4, 7, 12, 19, 29
   c) Create the **box plot**.
   d) Compute the **range** and **interquartile range (IQR)**. range = 29 − 4 = 25, IQR = 19 − 7 = 12
   e) Compute the outlier fences and check for **outliers** (i.e. data values much smaller or larger than the other values in the set).

Lower: $Q_1 - 1.5(IQR) = -11$
Upper: $Q_3 + 1.5(IQR) = 37$
; Data set is all between these two ⇒ No outliers

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.

$Y = aX + b$; $\bar{Y} = a\bar{X} + b$; $S_Y^2 = a^2 \cdot S_X^2$

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? 1,800 · 1000 = $1.8 million
   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) Median!
   c) Does the data set contain any outliers? Provide evidence.

   IQR = 34; Lower = −14, Upper = 122 ⇒ 1800 is an outlier

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

$\bar{X} = 77.6$

Lesson 4 – Solutions

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. $S = \{1,2,3,4,5,6\}$
   b) Let A be the event the outcome of a single roll is an odd number. Compute the **probability** event A occurs, P(A). $P(A) = \frac{\text{\# of odd rolls possible}}{\text{Total \# of outcomes}} = \frac{3}{6} = \frac{1}{2}$
2. For an arbitrary event A, express numerical boundaries for P(A). $0 \le P(A) \le 1$

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

Event	Probability notation	Explanation
a) **Complement** of A	A^c	A does not occur
b) **Union** of A, B	$A \cup B$	All elements in both A & B
c) **Intersection** of A, B	$A \cap B$	Elements common to both A&B
d) **Null/empty**	$\emptyset$	The Empty Set

4. A few rules:
   a) State the **complement rule** for probabilities. $P(A) = 1 - P(A^c)$
   b) State the **addition rule** for probabilities.
   c) State **DeMorgan's laws.** $P(A) + P(B) - P(A \cap B) = P(A \cup B)$

$(A \cap B)^c = A^c \cup B^c$ & $(A \cup B)^c = A^c \cap B^c$

See Next Page

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.

See Next Page

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.

See Next Page

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.

Maze

a) $S = \{FF, FL, FR, LF, LL, LR, RF, RL, RR\}$; Left letter is first mouse's movement. Right letter is 2^{nd} mouse's movement.

b) $B = \{FF, LL, RR\}$

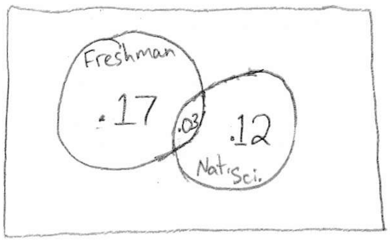
c) $C = \{FF, FL, FR, LF, RF\}$

d) $C^c = \{LL, LR, RL, RR\}$

e) $B \cap C = \{FF\} \neq \emptyset \Rightarrow$ Not mutually exclusive

f) $B \cup C = \{FF, FL, FR, LF, LL, RF, RR\}$

Wossamotta U Use F for Freshmen & NS for Natural Sciences

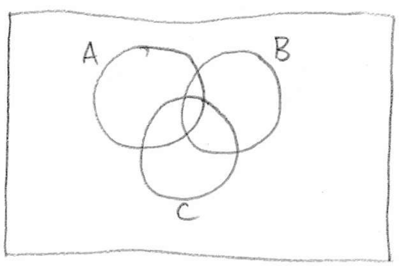


a) $\mathbb{P}(F \cup NS) = .17 + .03 + .12 = .32$

b) $\mathbb{P}(F^c \cap NS^c) = \mathbb{P}((F \cup NS)^c)$
$= 1 - \mathbb{P}(F \cup NS) = 1 - .32 = .68$

c) $\mathbb{P}(F \cap NS^c) = .17$

Triple



$\mathbb{P}(A) = \mathbb{P}(B) = \mathbb{P}(C) = .4$

$\mathbb{P}(A \cup B) = \mathbb{P}(A \cup C) = \mathbb{P}(B \cup C) = .64$

$\mathbb{P}(A \cup B \cup C) = .82$

Inclusion-Exclusion Principle:

$$\mathbb{P}(A \cup B \cup C) = \mathbb{P}(A) + \mathbb{P}(B) + \mathbb{P}(C) - \mathbb{P}(A \cap B) - \mathbb{P}(A \cap C) - \mathbb{P}(B \cap C) + \mathbb{P}(A \cap B \cap C)$$

$$\Rightarrow .82 = .4(3) - (\mathbb{P}(A) + \mathbb{P}(B) - \mathbb{P}(A \cup B)) - (\mathbb{P}(A) + \mathbb{P}(C) - \mathbb{P}(A \cup C)) - (\mathbb{P}(B) + \mathbb{P}(C) - \mathbb{P}(B \cup C)) + \mathbb{P}(A \cap B \cap C)$$

$$\Rightarrow .82 = 1.2 - (.4(2) - .64) - (.4(2) - .64) - (.4(2) - .64) + \mathbb{P}(A \cap B \cap C)$$

$$\Rightarrow -.38 = \mathbb{P}(A \cap B \cap C) - 3(.16)$$

$$\Rightarrow \mathbb{P}(A \cap B \cap C) = -.38 + 3(.16)$$
$$= .10$$

Lesson 5 - Solutions

1. Consider two events with $P(A) = 0.4$, $P(B) = 0.5$, and $P(A \cap B) = 0.2$.
   a) Compute $P(A \cup B)$. $= P(A) + P(B) - P(A \cap B) = .4 + .5 - .2 = .7$
   b) Explain **conditional probability** and the **multiplication rule**.
   c) Compute the conditional probabilities $P(A \mid B)$ and $P(B \mid A)$. $\rightarrow P(B|A) = \frac{P(A \cap B)}{P(A)}$

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

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

   $P(A) = P(A|B) \cdot P(B) + P(A|B^c) \cdot P(B^c)$

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

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

$P(B^c) = P(B^c|A)P(A) + P(B^c|A^c)P(A^c) = [1 - P(B|A)]P(A) + [1 - P(B|A^c)]P(A^c) = (.2)(.4) + (.7)(.6) = .5$

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%

See Next Page

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?

a) $P(2 \text{ reds}) = P(\text{Red First} \cap \text{Red Second}) = P(\text{Red Second}|\text{Red First}) \cdot P(\text{Red First})$

$= \left(\frac{5}{6}\right)\left(\frac{3}{5}\right) = \frac{3}{6} = 1/2$

b) $P(\text{Same Dist.}) = P(\text{Same Color Each Time}) = P(2 \text{ reds}) + P(2 \text{ greens})$

$= (.5) + P(\text{Green Second}|\text{Green First}) \cdot P(\text{Green First})$

$= .5 + \left(\frac{2}{6}\right)\left(\frac{2}{5}\right) = .5 + \frac{4}{30} = 19/30 \approx .6333$

Fishing

a) $\mathbb{P}(A) = .14 + .20 = .34$

b) $\mathbb{P}(\text{Female}) = .2 + .1 + .1 = .40$

c) $\mathbb{P}(A \mid \text{Female}) = \dfrac{\mathbb{P}(A \cap \text{Female})}{\mathbb{P}(\text{Female})} = \dfrac{.2}{.4} = .5$

d) $\mathbb{P}(\text{Female} \mid A) = \dfrac{\mathbb{P}(A \cap \text{Female})}{\mathbb{P}(A)} = \dfrac{.2}{.34} \approx .5882$

Lost

We are given:

$\mathbb{P}(\text{Located}) = .8$, $\mathbb{P}(\text{Beacon} \mid \text{Located}) = .7$, $\mathbb{P}(\text{Beacon}^c \mid \text{Located}^c) = .9$

b) $\mathbb{P}(\text{Beacon}) = \mathbb{P}(\text{Beacon} \mid \text{Located})\mathbb{P}(\text{Located}) + \mathbb{P}(\text{Beacon} \mid \text{Located}^c)\mathbb{P}(\text{Located}^c)$

$= (.7)(.8) + (.1)(.2) = .56 + .02 = .58$

c) $$\mathbb{P}(\text{Located}^c \mid \text{Beacon}) = \frac{\mathbb{P}(\text{Located}^c \cap \text{Beacon})}{\mathbb{P}(\text{Beacon})}$$

$$= \frac{\mathbb{P}(\text{Beacon} \mid \text{Located}^c) \cdot \mathbb{P}(\text{Located}^c)}{\mathbb{P}(\text{Beacon} \mid \text{Located}) \cdot \mathbb{P}(\text{Loc.}) + \mathbb{P}(\text{Beacon} \mid \text{Located}^c)\mathbb{P}(\text{Loc.}^c)}$$

$$= \frac{(.1)(.2)}{(.7)(.8) + (.1)(.2)} \approx .0345$$

d) $$\mathbb{P}(\text{Located} \mid \text{Beacon}^c) = \frac{\mathbb{P}(\text{Located} \cap \text{Beacon}^c)}{\mathbb{P}(\text{Beacon}^c)}$$

$$= \frac{\mathbb{P}(\text{Beacon}^c \mid \text{Located}) \cdot \mathbb{P}(\text{Located})}{1 - \mathbb{P}(\text{Beacon})}$$

$$= \frac{(.3)(.8)}{1 - ((.7)(.8) + (.1)(.2))} \approx .5714$$

Lesson 6 - Solutions

See Next Page

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. $P(C) \cdot P(D) = (.3)(.6) = .18 \neq .24 = P(C \cap D)$
   b) Compute $P(C \mid D)$ and $P(D \mid C)$ and compare to $P(C)$ and $P(D)$.
   $\Rightarrow$ C & D Dependent

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	$X = 0, 1, \ldots, 50$	Discrete
b)	Fill a 16 oz. bottle	# of ounces	$0 \leq X \leq 16$	Continuous
c)	ATM operation	Time between use	$0 \leq X$	Continuous
d)	Operate a business for one day	# of customers	$X = 0, 1, \ldots$	Discrete

See Next Page

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?

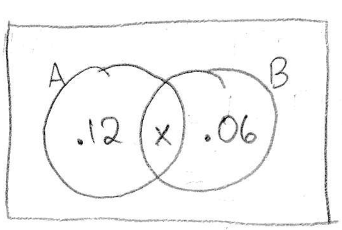
See Next Page

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

See Next Page

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?



By independence, $P(A \cap B) = P(A) \cdot P(B)$

$\Rightarrow X = (.12 + x)(.06 + x)$

$\Rightarrow X = .0072 + .18x + x^2$

$\Rightarrow x^2 - .82x + .0072 = 0 \Rightarrow$ $\cancel{X \approx .8111}$

$X \approx .0089$ b/c most projects succeed

1| a) $\mathbb{P}(A)\cdot\mathbb{P}(B) = (.4)(.5) = .2 = \mathbb{P}(A \cap B) \Rightarrow$ A & B are independent
(written $A \perp\!\!\!\perp B$)

c) $\mathbb{P}(B|A) = \dfrac{\mathbb{P}(B \cap A)}{\mathbb{P}(A)} = \dfrac{\mathbb{P}(B)\cdot\mathbb{P}(A)}{\mathbb{P}(A)} = \mathbb{P}(B)$

$\mathbb{P}(A|B) = \dfrac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)} = \dfrac{\mathbb{P}(A)\cdot\mathbb{P}(B)}{\mathbb{P}(B)} = \mathbb{P}(A)$

Parts|

a) $\mathbb{P}(A)\cdot\mathbb{P}(D) = (.4)(.06) = .024 \neq .02 = \mathbb{P}(A \cap D) \Rightarrow$ A & D dependent

b) $\mathbb{P}(D|A) = \dfrac{\mathbb{P}(D \cap A)}{\mathbb{P}(A)} = \dfrac{.02}{.4} = .05 \neq .06 = \mathbb{P}(D)$

$\Rightarrow$ Yes, knowing it comes from A reduces the chances of a defect.

c) $\mathbb{P}(B)\cdot\mathbb{P}(D) = (.2)(.06) = .012 = \mathbb{P}(B \cap D) \Rightarrow B \perp\!\!\!\perp D$

d) $\mathbb{P}(D|B) = \mathbb{P}(D)$ b/c $B \perp\!\!\!\perp D \Rightarrow$ No, knowing it comes from B has no effect on the chances of a defect.

Independent|

$A \perp\!\!\!\perp B \Rightarrow \mathbb{P}(A \cap B) = \mathbb{P}(A)\cdot\mathbb{P}(B)$. Hint: $\mathbb{P}(B) = \mathbb{P}(B \cap A) + \mathbb{P}(B \cap A^c)$

WWTS $\mathbb{P}(A^c \cap B) = \mathbb{P}(A^c)\cdot\mathbb{P}(B)$

$\mathbb{P}(A^c)\cdot\mathbb{P}(B) = (1-\mathbb{P}(A))\mathbb{P}(B) = \mathbb{P}(B) - \mathbb{P}(A)\cdot\mathbb{P}(B) = \mathbb{P}(B) - \mathbb{P}(A \cap B) = \mathbb{P}(B \cap A^c)$

Events| Let A, B be events s.t. $\mathbb{P}(A) \neq 0$ and $\mathbb{P}(B) \neq 0$.

- Suppose $A \perp\!\!\!\perp B$. Then $\mathbb{P}(A)\cdot\mathbb{P}(B) = \mathbb{P}(A \cap B)$. Since neither are zero, we must have $\mathbb{P}(A)\cdot\mathbb{P}(B) \neq 0$. Thus, $\mathbb{P}(A \cap B) \neq 0$, and A & B cannot be mutually exclusive.
- Suppose A & B are mutually exclusive. Then $\mathbb{P}(A \cap B) = 0$. Since neither are zero, we must have $\mathbb{P}(A)\cdot\mathbb{P}(B) \neq 0$. Thus, $\mathbb{P}(A)\cdot\mathbb{P}(B) \neq \mathbb{P}(A \cap B)$, and A & B cannot be independent.

Lesson 7 - Solutions

1. a) This is a valid pmf because:

   - $\sum p(x) = .05 + .25 + .60 + .10 = 1$
   - All $p(x) \geq 0$.

b) $\mathbb{P}(X=4) = p(4) = .6$; 60% chance that $X=4$

c) $F(4) = \mathbb{P}(X \leq 4) = \mathbb{P}(X=2 \cup X=3 \cup X=4) = \mathbb{P}(X=2) + \mathbb{P}(X=3) + \mathbb{P}(X=4)$
$= .05 + .25 + .60 = .90$

90% chance that X is less than or equal to 4.

d) $\mathbb{E}[X] = \sum x \cdot p(x) = 2(.05) + 3(.25) + 4(.6) + 5(.1) = 3.75$

e) $\mathbb{E}[X^2] = \sum x^2 \cdot p(x) = 2^2(.05) + 3^2(.25) + 4^2(.6) + 5^2(.1) = 14.55$

F) $Var[X] = \mathbb{E}[(X - \mathbb{E}[X])^2] = \mathbb{E}[X^2 - 2X\mathbb{E}[X] + \mathbb{E}[X]^2]$
$= \mathbb{E}[X^2] - 2\mathbb{E}[X]^2 + \mathbb{E}[X]^2 = \mathbb{E}[X^2] - \mathbb{E}[X]^2$
$= 14.55 - 3.75^2 = .4875$

g) $SD[X] = \sqrt{Var[X]} = \sqrt{.4875} \approx .6982$

2. $Y = a + bX$; $a, b \in \mathbb{R}$

a) $\mathbb{E}[Y] = \mathbb{E}[a + bX] = \sum (a+bX)p(x) = \sum a \cdot p(x) + \sum bx p(x) = a \cdot \sum p(x) + b \sum x p(x)$
$= a + b \cdot \mathbb{E}[X]$

b) $Var[Y] = Var[a+bX] = \mathbb{E}[((a+bX) - \mathbb{E}[a+bX])^2] = \mathbb{E}[(a+bX-a-b\mathbb{E}[X])^2]$
$= \mathbb{E}[(bX - b\mathbb{E}[X])^2] = \mathbb{E}[b^2(X - \mathbb{E}[X])^2] = b^2 \mathbb{E}[(X - \mathbb{E}[X])^2]$
$= b^2 \cdot Var[X]$

3. Table of Notation:

Quantity of Interest → / Characteristic ↓	Sample Data $\{X_1, X_2, \ldots, X_n\}$	Random Variable X	Population $\{X_1, \ldots, X_N\}$
Mean	$\bar{X}$	$\mathbb{E}[X]$	μ_X
Variance	S^2	$Var[X]$	σ_X^2
St. Deviation	S	$SD[X]$	σ_X

Cereal

a)

k	20	23	26
$\mathbb{P}(X=k)$	.20	.50	.30

b) $\mathbb{E}[X] = 20(.2) + 23(.5) + 26(.3)$
$= 23.3$

c) $\mathbb{E}[X^2] = 20^2(.2) + 23^2(.5) + 26^2(.3) = 547.3$

$\Rightarrow Var[X] = \mathbb{E}[X^2] - \mathbb{E}[X]^2 = 547.3 - 23.3^2 = 4.41$

$\Rightarrow SD[X] = \sqrt{4.41} = 2.1$

d) $\mathbb{E}[Y] = \mathbb{E}[2.4 + .1 \cdot X] = 2.4 + (.10)\mathbb{E}[X] = 2.4 + (.10)(23.3) = 4.73$

e) $Var[Y] = Var[2.4 + (.1)X] = .1^2 \cdot Var[X] = .01 \cdot 4.41 = .0441$

$\Rightarrow SD[Y] = \sqrt{.0441} = .21$

f) $\mathbb{E}[W] = \mathbb{E}[X - .001X^2] = \mathbb{E}[X] - .001\mathbb{E}[X^2] = 23.3 - .001(547.3)$
$= 22.7527$

Paperwork

We are told that $\mathbb{P}(X=k) \propto k$, so $\mathbb{P}(X=k) = c \cdot k$ for some $c \in \mathbb{R}$.

Since we must have a valid pdf, the following must hold:

$$\sum p(x) = 1 \Rightarrow 2c + 3c + 4c + 5c = 1$$

$$\Rightarrow 14c = 1 \Rightarrow c = 1/14$$

So, our pmf is:

k	2	3	4	5
$\mathbb{P}(X=k)$	$2/14$	$3/14$	$4/14$	$5/14$

Lesson 8 - Solutions

1. a) • Anything that falls into a "success or failure" structure.
   • As customers come into a store, they will either make a purchase or they will not.
   • Students in a course will either pass or fail.

b) $X \sim \text{Bernoulli}(p)$ means "X follows a Bernoulli distribution with probability of success p.

c) $\mathbb{P}(X=k) = p^k(1-p)^{1-k}$; $k \in \{0,1\}$

That is, $\mathbb{P}(X=1)=p$ and $\mathbb{P}(X=0)=1-p$

d) $\mathbb{E}[X] = 1 \cdot p + 0 \cdot (1-p) = p$

e) $\mathbb{E}[X^2] = 1^2 \cdot p + 0^2 \cdot (1-p) = p$

$\Rightarrow Var[X] = p - p^2 = p(1-p) \Rightarrow SD[X] = \sqrt{p(1-p)}$

2. a) $\mathbb{P}(X=k) = \binom{n}{k} p^k (1-p)^{n-k}$; $k \in \{0,1,2,\dots,n\}$

b) $\mathbb{E}[X] = np$

$Var[X] = np(1-p)$

SLEEP) a) $X \sim \text{Bernoulli}(.25)$

b) $\mathbb{E}[X] = .25$, $Var[X] = .25(1-.25) = .1875$, $SD[X] \approx .433$

c) $\mathbb{E}[X^{79}] = 1^{79}(.25) + 0^{79}(.75) = .25$

Sleep 2

a) $Y \sim \text{Binom}(10, .25)$

b) $\mathbb{P}(Y=4) = \binom{10}{4}(.25)^4(.75)^6 \approx .146$

c) $\mathbb{E}[Y] = (10)(.25) = 2.5$

$\text{Var}[Y] = (10)(.25)(.75) = 1.875 \Rightarrow SD[X] \approx 1.3693$

Medication

a) X = # of people that recover in the sample of 12

$X \sim \text{Binom}(12, .5)$

$\mathbb{P}(X \geq 11) \approx .0032$

b) Since we observed an event that would be very unlikely if the drug had no effect, this could be evidence that the new treatment is effective.

Lesson 9 - Solutions

1. a) $\mathbb{P}(X=k) = (1-p)^{k-1}p \;;\; k \in \mathbb{N}$

   b) $\mathbb{E}[X] = 1/p \;,\; Var[X] = \frac{1-p}{p^2}$

2. a) $\mathbb{P}(X=2) = .16$

   b) $\mathbb{P}(X>2) = 1 - \mathbb{P}(X \leq 2) = .04$

3. a) $\mathbb{P}(X=k) = e^{-\mu} \cdot \frac{\mu^k}{k!} \;;\; k \in \{0,1,2,3,\ldots\}$

   b) $\mathbb{E}[X] = Var[X] = \mu$

   c) pmf: dpois(input, μ)
      cdf: ppois(input, μ)

4. a) $\mathbb{P}(X=5) \approx .1165$

   b) $Var[X] = 3.24 \Rightarrow SD[X] = 1.8$

Stop| $X \sim Geom(.4)$

a) $\mathbb{P}(X=5) = .05184$

b) $\mathbb{P}(X=1) + \mathbb{P}(X=2) = .64$

c) $\mathbb{E}[X] = 1/.4 = 2.5$

d) $SD[X] = \sqrt{\frac{1-.4}{.4^2}} \approx 1.9365$

Lesson 10 - Solutions

3. a) median = 4 (given in 5 # summary)

b) Upper = $Q_3 + 1.5(IQR) = 8 + (1.5)(8-3) = 15.5$

Lower = $Q_1 - 1.5(IQR) = 3 - (1.5)(8-3) = {}^-4.5$

c) Upper = $\bar{x} + 2s = 5.3 + 2\cdot\sqrt{11.6} \approx 12.112$

Lower = $\bar{x} - 2s = 5.3 - 2\cdot\sqrt{11.6} \approx {}^-1.512$

4. a) All values of $p(x)$ are non-negative ✓

$\Sigma p(x) = .15 + .15 + .2 + .3 + .2 = 1$ ✓

b) $\mathbb{E}[X] = (^-2)(.15) + 0(.15) + 1(.2) + 3(.3) + 5(.2) = 1.8$

$\mathbb{E}[X^2] = (^-2)^2(.15) + 0^2(.15) + 1^2(.2) + 3^2(.3) + 5^2(.2) = 8.5$

$\Rightarrow Var[X] = \mathbb{E}[X^2] - \mathbb{E}[X]^2 = 8.5 - (1.8)^2 = 5.26$

$\Rightarrow SD[X] \approx 2.293$

Phones

a) A; 2 phones in use

b) A^c; 0, 1, 3, 4, 5, or 6 phones

c) B; 4 phones in use

Switches

In the following, think of each string of 4 digits as indicating what position the 4 switches are in. For example, 0110 is the event where switches 1 & 4 are off, while switches 2 & 3 are on.

$$S = \{0000, 0001, 0010, 0100, 1000,$$
$$0011, 0101, 1001, 1010, 1100,$$
$$0110, 0111, 1011, 1101, 1110, 1111\}$$

Quiz

a) $X \sim \text{Binom}(18, .3)$

b) dbinom(k, 18, .3)

e) $\mathbb{P}(X \geq 3) = 1 - \mathbb{P}(X < 3) = 1 - \mathbb{P}(X \leq 2)$

$= 1 -$ sum(dbinom(0:2, 18, .3))

$= 1 - (.0016 + .0126 + .0458) = .94$

Baseball

Let A = player A gets a hit, & B = player B gets a hit

a) $\mathbb{P}(A^c | B) = \dfrac{\mathbb{P}(A^c \cap B)}{\mathbb{P}(B)} = \dfrac{\mathbb{P}(B) - \mathbb{P}(A \cap B)}{\mathbb{P}(B)} = \dfrac{.76 - .64}{.76} \approx .1579$

b) $\mathbb{P}(A^c | B^c) = \dfrac{\mathbb{P}(A^c \cap B^c)}{\mathbb{P}(B^c)} = \dfrac{\mathbb{P}(A^c) + \mathbb{P}(B^c) - \mathbb{P}(A^c \cup B^c)}{1 - \mathbb{P}(B)}$

$= \dfrac{(1 - \mathbb{P}(A)) + (1 - \mathbb{P}(B)) - \mathbb{P}((A \cap B)^c)}{1 - \mathbb{P}(B)}$

$= \dfrac{1 - \mathbb{P}(A) + 1 - \mathbb{P}(B) - (1 - \mathbb{P}(A \cap B))}{1 - \mathbb{P}(B)} = \dfrac{1 - .7 + 1 - .76 - 1 + .64}{1 - .76} \approx .0432$

Lesson 11 - Solutions

1. See Notes

2. $\int_0^3 k(9-x^2)\,dx = \left(9kx - \frac{1}{3}kx^3\right)\Big|_0^3 = (9k\cdot 3 - \frac{1}{3}k\cdot 27) = 1$

$\Rightarrow 27k - 9k = 1 \Rightarrow 18k = 1 \Rightarrow k = 1/18$

3. a) $(0,3)$

b) See R Code

c) $F(1) = \mathbb{P}(X \le 1) = \int_0^1 \frac{1}{18}(9-x^2)\,dx = \left(\frac{1}{2}x - \frac{1}{54}x^3\right)\Big|_0^1 = \frac{1}{2} - \frac{1}{54} = \frac{26}{54} = 13/27 \approx .4815$

4. a) Area = Length · Height $\Rightarrow 1 = (22-12)\cdot$ Height $\Rightarrow$ Height $= 1/10$

b) Area $= (19-15)\cdot(1/10) = 4/10 = 2/5$

5. a) $X \sim Unif(a,b)$

$\Rightarrow$ "X has a uniform distribution on the interval (a,b)"

b) $f(x) = \frac{1}{b-a}$; $x \in (a,b)$

$\boxed{\text{Fluid}}$

a) $X \sim Unif(7.5, 20)$; $f(x) = \frac{1}{20-7.5} = \frac{1}{12.5} = \frac{1}{25/2} = 2/25$; $x \in (7.5, 20)$

b) $\mathbb{P}(X \le 10) = \int_{7.5}^{10} \frac{2}{25}\,dx = \left(\frac{2}{25}x\right)\Big|_{7.5}^{10} = \frac{2}{25}(10-7.5) = \frac{2}{25}(2.5) = \frac{5}{25} = 1/5 = .2$

c) $\mathbb{P}(k < X < k+2) = \int_k^{k+2} \frac{2}{25}\,dx = \left(\frac{2}{25}x\right)\Big|_k^{k+2} = \frac{2}{25}(k+2-k) = 4/25 = .16$

d) Zero! The only scores we see are between 7.5 & 20

e) One! All possible scores are less than 25.

Washers)

a) $(2,6)$

b) See R code

c) $P(X<4) = \int_2^4 \frac{3}{80}x(6-x)\,dx = \int_2^4 \frac{18}{80}x - \frac{3}{80}x^2\,dx = \left(\frac{9}{80}x^2 - \frac{1}{80}x^3\right)\Big|_2^4$

$= \frac{1}{80}\left[(9\cdot16 - 4^3) - (9\cdot4 - 8)\right] = \frac{1}{80}\left[(9\cdot16 - 4\cdot16) - (9\cdot4 - 2\cdot4)\right]$

$= \frac{1}{80}\left[5\cdot16 - 7\cdot4\right] = \frac{1}{80}(80 - 28) = 1 - \frac{28}{80} = 1 - \frac{14}{40} = 1 - \frac{7}{20} = 13/20$

$= .65$

Lesson 12 - Solutions

1. a) $\mathbb{E}[X] = \int_{\text{support}} x \cdot f(x)\,dx = \int_0^3 x \cdot \frac{1}{18}(9-x^2)\,dx = \int_0^3 \frac{1}{2}x - \frac{1}{18}x^3\,dx$

$$= \left(\frac{1}{4}x^2 - \frac{1}{72}x^4\right)\Big|_0^3 = \frac{9}{4} - \frac{81}{72} = \frac{9}{4} - \frac{9}{8} = \frac{18-9}{8} = 9/8 = 1.125$$

b) $\mathbb{E}[g(X)] = \int_{\text{support}} g(x) \cdot f(x)\,dx$

$$\mathbb{E}[X^2] = \int_0^3 x^2 \cdot \frac{1}{18}(9-x^2)\,dx = \int_0^3 \frac{1}{2}x^2 - \frac{1}{18}x^4\,dx = \left(\frac{1}{6}x^3 - \frac{1}{90}x^5\right)\Big|_0^3$$

$$= \frac{27}{6} - \frac{3^5}{90} = \frac{9}{2} - \frac{3^3}{10} = \frac{45-27}{10} = \frac{18}{10} = 1.8$$

c) $Var[X] = \mathbb{E}[(X-\mathbb{E}[X])^2] = \mathbb{E}[X^2] - \mathbb{E}[X]^2$

$$= 1.8 - (1.125)^2 \approx .5344$$

d) $SD[X] = \sqrt{Var[X]} \approx .731$

e) $\int_0^k \frac{1}{18}(9-x^2)\,dx = .60 \Rightarrow \frac{1}{18}\left(9x - \frac{1}{3}x^3\right)\Big|_0^k = .60$

$$\Rightarrow \frac{1}{18}\left(9k - \frac{1}{3}k^3\right) = .6 \Rightarrow 9k - \frac{1}{3}k^3 = 10.8 \Rightarrow$$ $k \approx 1.28$ (from R)

2. $\mathbb{E}[X] = \int_a^b x \cdot \frac{1}{b-a}\,dx = \frac{1}{b-a}\left(\frac{1}{2}x^2\right)\Big|_a^b = \frac{1}{b-a}\left(\frac{1}{2}(b^2-a^2)\right)$

$$= \frac{1}{\cancel{b-a}}\left(\frac{(\cancel{b-a})(b+a)}{2}\right) = \frac{a+b}{2}$$

$$\mathbb{E}[X^2] = \int_a^b x^2 \cdot \frac{1}{b-a}\,dx = \frac{1}{b-a}\left(\frac{1}{3}x^3\right)\Big|_a^b = \frac{1}{b-a}\left(\frac{1}{3}(b^3-a^3)\right)$$

$$= \frac{1}{3(b-a)}(b-a)(b^2+ab+a^2) = \frac{1}{3}(b^2+ab+a^2)$$

$$\Rightarrow \mathrm{Var}[X] = \frac{1}{3}(b^2+ab+a^2) - \left(\frac{a+b}{2}\right)^2$$

$$= \frac{1}{3}(b^2+ab+a^2) - \frac{1}{4}(b^2+2ab+b^2)$$

$$= \frac{1}{12}(b^2-2ab+a^2) = \frac{1}{12}(b-a)^2$$

Fluid 2 $X \sim \mathrm{Unif}(7.5, 20)$; $f(x) = \frac{2}{25}$; $x \in (7.5, 20)$

a) $F_X(x) = \mathbb{P}(X \le x) = \int_{7.5}^{x} \frac{2}{25}\,dt = \left(\frac{2}{25}t\right)\Big|_{7.5}^{x} = \frac{2}{25}(x-7.5)$; $x \in (7.5, 20)$

$$\Rightarrow F_X(x) = \begin{cases} 0 & , x \le 7.5 \\ \frac{2}{25}(x-7.5) & , x \in (7.5, 20) \\ 1 & , x \ge 20 \end{cases}$$

b) $\mathbb{E}[X] = \frac{a+b}{2} = \frac{7.5+20}{2} = \frac{27.5}{2} = 13.75$

$\mathrm{Var}[X] = \frac{(b-a)^2}{12} = \frac{(20-7.5)^2}{12} = \frac{12.5^2}{12} = 13.02083$

c) $SD[X] \approx 3.61$

$\mathbb{P}(13.75 - 3.61 \le X \le 13.75 + 3.61) = \mathbb{P}(10.14 \le X \le 17.36)$

$$= \int_{10.14}^{17.36} \frac{2}{25}\,dx = \frac{2}{25}(17.36 - 10.14) = .5776$$

Washers 2

a) $\int_2^k \frac{3}{80}x(6-x)\,dx = .5 \Rightarrow \int_2^k \frac{18}{80}x - \frac{3}{80}x^2\,dx = .5$

$\Rightarrow \left(\frac{9}{80}x^2 - \frac{1}{80}x^3\right)\Big|_2^k = .5 \Rightarrow \frac{1}{80}\left[(9k^2 - k^3) - (9\cdot 4 - 8)\right] = .5$

$\Rightarrow \frac{1}{80}\left[9k^2 - k^3 - 28\right] = .5 \Rightarrow k \approx 3.524$

b) $\mathbb{E}[X] = \int_2^6 x \cdot \frac{3}{80}x(6-x)\,dx = \frac{3}{80}\int_2^6 6x^2 - x^3\,dx = \frac{3}{80}\left(2x^3 - \frac{1}{4}x^4\right)\Big|_2^6$

$= \frac{3}{80}\left(\left(2\cdot 6^3 - \frac{1}{4}6^4\right) - \left(2\cdot 2^3 - \frac{1}{4}\cdot 2^4\right)\right)$

$= \frac{3}{80}\left(432 - 324 - 16 + 4\right) = \frac{3}{80}(96) = 3.6$

c) $\mathbb{E}[X^2] = \int_2^6 \frac{3}{80}(6x^3 - x^4)\,dx = \frac{3}{80}\left(\frac{6}{4}x^4 - \frac{1}{5}x^5\right)\Big|_2^6 = 13.92$

$\Rightarrow Var[X] = 13.92 - 3.6^2 = 0.96$

$\Rightarrow SD[X] \approx .98$

Lesson 13 - Solutions

1. $X \sim N(\mu, \sigma^2) \Rightarrow f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2\sigma^2}(x-\mu)^2}\ ;\ x \in \mathbb{R}$

2. a) $\Phi(x) = F_Z(x) = \mathbb{P}(Z \leq x) = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}z^2}\, dz$ This has no closed form!

   b) $X \sim N(\mu, \sigma^2) \Rightarrow Z = \frac{X-\mu}{\sigma} \sim N(0,1)$

3. $X \sim Exp(\lambda)$

   a) $f(x) = \lambda e^{-\lambda x}\ ;\ x > 0$

   b) $\mathbb{E}[X] = \int_0^\infty x \cdot \lambda e^{-\lambda x}\, dx \quad \begin{pmatrix} u = x & dv = \lambda e^{-\lambda x} dx \\ du = dx & v = -e^{-\lambda x} \end{pmatrix}$

$$= \left(-xe^{-\lambda x}\right)\Big|_0^\infty + \int_0^\infty e^{-\lambda x}\, dx = (0+0) + \left(-\tfrac{1}{\lambda} e^{-\lambda x}\right)\Big|_0^\infty$$

$$= 0 - \left(-\tfrac{1}{\lambda} e^0\right) = 1/\lambda$$

$\mathbb{E}[X^2] = \int_0^\infty x^2 \cdot \lambda e^{-\lambda x}\, dx \quad \begin{pmatrix} u = x^2 & dv = \lambda e^{-\lambda x} dx \\ du = 2x dx & v = -e^{-\lambda x} \end{pmatrix}$

$$= \left(-x^2 e^{-\lambda x}\right)\Big|_0^\infty + \int_0^\infty 2x e^{-\lambda x}\, dx = 0 + \tfrac{2}{\lambda}\int_0^\infty x \cdot \lambda e^{-\lambda x}\, dx$$

$$= \tfrac{2}{\lambda}\left(\tfrac{1}{\lambda}\right) = 2/\lambda^2$$

$$\Rightarrow Var[X] = \tfrac{2}{\lambda^2} - \tfrac{1}{\lambda^2} = 1/\lambda^2$$

   c) $F(x) = \int_0^x \lambda e^{-\lambda z}\, dz = \left(-e^{-\lambda z}\right)\Big|_0^x = -e^{-\lambda x} + 1 = 1 - e^{-\lambda x}\ ;\ x > 0$

4. $X \sim Exp(2)$

a) $\mathbb{E}[X] = 1/2$

b) $F(3) = 1 - e^{-2 \cdot 3} = 1 - e^{-6}$

Reactor $X \sim N(30, 36)$

a) $\mathbb{P}(X > 25) = 1 - \mathbb{P}(X \le 25) = 1 - \text{pnorm}(25, 30, 6) \approx .7977$

b) $\mathbb{P}(|X - 30| < 8) = \mathbb{P}(-8 < X - 30 < 8) = \mathbb{P}(22 < X < 38)$

$= \mathbb{P}(X < 38) - \mathbb{P}(X < 22)$

$= \text{pnorm}(38, 30, 6) - \text{pnorm}(22, 30, 6) \approx .8176$

c) $\text{qnorm}(.85, 30, 6) \approx 36.2186$

Airdrop

Let X = Opening Altitude

So $X \sim N(220, 40^2)$

Let Y = # of payloads with damage in the sample of 5.

$\Rightarrow Y \sim \text{Binom}(5, \mathbb{P}(\text{Damage}))$

$\mathbb{P}(\text{Damage}) = \mathbb{P}(X < 120) = \text{pnorm}(120, 220, 40) \approx .0062$

So $Y \sim \text{Binom}(5, .0062)$, and we have:

$\mathbb{P}(Y \ge 1) = 1 - \mathbb{P}(Y < 1) = 1 - \mathbb{P}(Y = 0)$

$= 1 - \text{dbinom}(0, 5, .0062) \approx .0307$

Lesson 14 - Solutions

1. See Notes

2. See Notes

3. $p_X(x) = \sum_y p(x,y)$; $p_Y(y) = \sum_x p(x,y)$

X	$\mathbb{P}(X=x)$
0	.7
2	.3

Y	$\mathbb{P}(Y=y)$
-1	.4
1	.6

X & Y are not independent because

$p(0,-1) = .3 \neq (.7)(.4) = p_X(0) \cdot p_Y(-1)$

4. The support of X is $\{0,1\}$
The support of Y is $\{0,1,2\}$
& X and Y are independent, so $p(x,y) = p_X(x) \cdot p_Y(y)$

a)

X \ Y	0	1	2
0	$(.8)(.2) = .16$	$(.8)(.3) = .24$	$(.8)(.5) = .40$
1	$(.2)(.2) = .04$	$(.2)(.3) = .06$	$(.2)(.5) = .10$

b) $\mathbb{P}(X+Y \leq 1) = p(0,0) + p(1,0) + p(0,1) = .16 + .04 + .24 = .44$

5. $f(x,y) = k(x+y)$; $x,y \in (0,2)$

a) $\int_0^2 \int_0^2 k(x+y)\,dx\,dy = 1 \Rightarrow k \cdot \int_0^2 \left(\frac{1}{2}x^2 + xy\right)\Big|_0^2 dy = 1$

$\Rightarrow k \cdot \int_0^2 2 + 2y\,dy = 1$

$\Rightarrow k(2y + y^2)\Big|_0^2 = 1$

$\rightarrow k(4+4) = 1 \Rightarrow k = 1/8$

b) $f_X(x) = \int_0^2 \frac{1}{8}(x+y)\,dy = \frac{1}{8}\left(xy + \frac{1}{2}y^2\right)\Big|_0^2 = \frac{1}{8}(2x+2)$; $x \in (0,2)$

Since $f(x,y) = f(y,x)$, we have $f_Y(y) = \frac{1}{8}(2y+2)$; $y \in (0,2)$

c) No, because $f_X(x) \cdot f_Y(y) = \frac{1}{64}(2x+2)(2y+2) \neq \frac{1}{8}(x+y) = f(x,y)$

d) $P(X > 1 \cap Y > 1) = \int_1^2 \int_1^2 \frac{1}{8}(x+y)\,dx\,dy$

$= \frac{1}{8}\int_1^2 \left(\frac{1}{2}x^2 + xy\right)\Big|_1^2 dy = \frac{1}{8}\int_1^2 (2+2y) - \left(\frac{1}{2}+y\right) dy$

$= \frac{1}{8}\int_1^2 \frac{3}{2} + y\,dy = \frac{1}{8}\left(\frac{3}{2}y + \frac{1}{2}y^2\right)\Big|_1^2$

$= \frac{1}{8}\left[(3+2) - \left(\frac{3}{2}+\frac{1}{2}\right)\right] = \frac{1}{8}(5-2) = 3/8$

Lesson 15 - Solutions

1. See Notes

2. a) $\text{Cov}[X,Y] = \mathbb{E}[(X-\mathbb{E}[X])(Y-\mathbb{E}[Y])] = \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y]$

$$\mathbb{E}[X] = 0(.7) + 2(.3) = .6$$
$$\mathbb{E}[Y] = -1(.4) + 1(.6) = .2$$
$$\mathbb{E}[XY] = (-1\cdot 0)(.3) + (-1\cdot 2)(.1) + (1\cdot 0)(.4) + (1\cdot 2)(.2) = -.2 + .4 = .2$$
$$\Rightarrow \text{Cov}[X,Y] = .2 - (.2)(.6) = .08$$

b) $\text{Cov}[X,X] = \mathbb{E}[(X-\mathbb{E}[X])(X-\mathbb{E}[X])] = \mathbb{E}[(X-\mathbb{E}[X])^2] = \text{Var}[X]$

c) $\text{Corr}[X,Y] = \dfrac{\text{Cov}[X,Y]}{SD[X]\cdot SD[Y]}$

$$\mathbb{E}[X^2] = 0^2(.7) + 2^2(.3) = 1.2 \Rightarrow \text{Var}[X] = 1.2 - (.6)^2 = .84$$
$$\mathbb{E}[Y^2] = (-1)^2(.4) + 1^2(.6) = 1 \Rightarrow \text{Var}[Y] = 1 - (.2)^2 = .96$$
$$\Rightarrow \text{Corr}[X,Y] = \frac{.08}{\sqrt{(.84)(.96)}} \approx .0891$$

d)
- $-1 \le \text{Corr}[X,Y] \le 1$
- X, Y independent $\Rightarrow \text{Corr}[X,Y] = 0$

3. a) $\text{Cov}[X,Y] = \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y] \underset{\text{by independence}}{=} \mathbb{E}[X]\mathbb{E}[Y] - \mathbb{E}[X]\mathbb{E}[Y] = 0$

b) $\text{Corr}[X,Y] = \dfrac{\text{Cov}[X,Y]}{\sqrt{\text{Var}[X]\text{Var}[Y]}} = \dfrac{0}{\sqrt{\text{Var}[X]\text{Var}[Y]}} = 0$

4. From Lesson 14, we know the marginal pdf's:

$$f_X(x) = \frac{1}{8}(2+2x) \;;\; x \in (0,2)$$

$$f_Y(y) = \frac{1}{8}(2+2y) \;;\; y \in (0,2)$$

So, we can find everything we'll need:

$$E[X] = \int_0^2 \frac{1}{8}(2x+2x^2)\,dx = \frac{1}{8}\left(x^2+\frac{2}{3}x^3\right)\Big|_0^2 = \frac{1}{8}\left(4+\frac{2}{3}(8)\right) = \frac{1}{2}+\frac{2}{3} = 7/6$$

$$E[X^2] = \int_0^2 \frac{1}{8}(2x^2+2x^3)\,dx = \frac{1}{8}\left(\frac{2}{3}x^3+\frac{1}{2}x^4\right)\Big|_0^2 = \frac{1}{8}\left(\frac{2}{3}(8)+\frac{1}{2}(16)\right) = \frac{2}{3}+1 = 5/3$$

$$\mathrm{Var}[X] = \frac{5}{3}-\left(\frac{7}{6}\right)^2 = \frac{5}{3}-\frac{49}{36} = 11/36$$

Since Y has the same pdf as X, these are also the values of $E[Y]$, $E[Y^2]$, and $\mathrm{Var}[Y]$.

a) $\mathrm{Cov}[X,Y] = E[XY] - E[X]E[Y]$

$$= \int_0^2\int_0^2 xy\cdot\frac{1}{8}(x+y)\,dx\,dy - \left(\frac{7}{6}\right)\left(\frac{7}{6}\right)$$

$$= \int_0^2\int_0^2 \frac{1}{8}(x^2y+xy^2)\,dx\,dy - 49/36 = \frac{1}{8}\int_0^2\left(\frac{1}{3}x^3y+\frac{1}{2}x^2y^2\right)\Big|_0^2 dy - \frac{49}{36}$$

$$= \frac{1}{8}\int_0^2 \frac{8}{3}y+2y^2\,dy - \frac{49}{36} = \frac{1}{8}\left(\frac{4}{3}y^2+\frac{2}{3}y^3\right)\Big|_0^2 - \frac{49}{36}$$

$$= \frac{1}{8}\left(\frac{16}{3}+\frac{16}{3}\right) - \frac{49}{36} = \frac{4}{3}-\frac{49}{36} = \frac{-1}{36}$$

b) $$\mathrm{Corr}[X,Y] = \frac{-1/36}{\sqrt{(11/36)(11/36)}} = \frac{-1/36}{11/36} = -\frac{1}{11}$$

5. See Notes

6.

a) X_1 and X_2 could be either 1's or 4's.

$W = X_1 + 2X_2$

$$\left.\begin{array}{l} X_1=1, X_2=1 \Rightarrow W=3 \\ X_1=1, X_2=4 \Rightarrow W=9 \\ X_1=4, X_2=1 \Rightarrow W=6 \\ X_1=4, X_2=4 \Rightarrow W=12 \end{array}\right\}$$

Thus, the support of W is $\{3, 6, 9, 12\}$

b) $\mathbb{P}(W=3) = \mathbb{P}(X_1=1 \cap X_2=1) = \mathbb{P}(X_1=1)\mathbb{P}(X_2=1) = (.7)(.7) = .49$

$\mathbb{P}(W=6) = \mathbb{P}(X_1=4 \cap X_2=1) = \mathbb{P}(X_1=4)\mathbb{P}(X_2=1) = (.3)(.7) = .21$

$\mathbb{P}(W=9) = \mathbb{P}(X_1=1 \cap X_2=4) = \mathbb{P}(X_1=1)\mathbb{P}(X_2=4) = (.7)(.3) = .21$

$\mathbb{P}(W=12) = \mathbb{P}(X_1=4 \cap X_2=4) = \mathbb{P}(X_1=4)\mathbb{P}(X_2=4) = (.3)(.3) = .09$

c) $\mathbb{E}[W] = 3(.49) + 6(.21) + 9(.21) + 12(.09) = 5.7$

$\mathbb{E}[W^2] = 3^2(.49) + 6^2(.21) + 9^2(.21) + 12^2(.09) = 41.94$

$Var[W] = 41.94 - (5.7)^2 = 9.45$

Lesson 16 - Solutions

1. a) Let $X_1, \ldots, X_n$ be n different RVs.

   Let $a_1, \ldots, a_n$ be n different constants.

   A linear combination of the X_i's is $a_1X_1 + a_2X_2 + \ldots + a_nX_n = \sum_{i=1}^{n} a_iX_i$

b) $\mathbb{E}\left[\sum_{i=1}^{n} a_iX_i\right] = \sum_{i=1}^{n} a_i\mathbb{E}[X_i]$

c) $Var\left[\sum_{i=1}^{n} a_iX_i\right] = \sum_{i=1}^{n} a_i^2 Var[X_i] + 2\sum_{i<j} a_ia_j Cov[X_i, X_j]$

d) If $X_1, \ldots, X_n$ are independent:

   $Var\left[\sum_{i=1}^{n} a_iX_i\right] = \sum_{i=1}^{n} a_i^2 Var[X_i]$

e) If each $X_i \sim Normal(\mu_i, \sigma_i)$ and they are independent, then:

   $\sum_{i=1}^{n} a_iX_i \sim Normal\left(\sum_{i=1}^{n} a_i\mu_i, \sqrt{\sum_{i=1}^{n} a_i^2\sigma_i^2}\right)$

2. a) $\mathbb{E}[3X+2Y] = 3\mathbb{E}[X] + 2\mathbb{E}[Y]$

b) $\mathbb{E}[3X-2Y] = 3\mathbb{E}[X] - 2\mathbb{E}[Y]$

c) $\mathbb{E}[X+Y+W] = \mathbb{E}[X] + \mathbb{E}[Y] + \mathbb{E}[W]$

d) $Var[3X+2Y] = 9\,Var[X] + 4\,Var[Y] + 6\,Cov[X,Y]$

e) $Var[3X-2Y] = 9\,Var[X] + 4\,Var[Y] - 6\,Cov[X,Y]$

f) $Var[X+Y+W] = Var[X] + Var[Y] + Var[W] + 2Cov[X,Y] + 2Cov[X,W] + 2Cov[Y,W]$

g) Drop all covariance terms.

Shipping

Let T be the total weight shipped.

So, $T = 5S + 7M + 10L$

a) $\mathbb{E}[T] = 5\mathbb{E}[S] + 7\mathbb{E}[M] + 10\mathbb{E}[L] = 5(100) + 7(200) + 10(150) = 3400$

$Var[T] = 25\,Var[S] + 49\,Var[M] + 100\,Var[L] = 25(36) + 49(64) + 100(16) = 5636$

$SD[T] = \sqrt{5636} \approx 75.07$

b) The expected value would be correct, but the variance would not be because we did not account for covariances.

Ethanol

a) We do not know, because these are random!

b) $\mathbb{P}(X_1 > X_4) = \mathbb{P}(X_1 - X_4 > 0)$

Since X_1 & X_4 are normal and independent, we know:

$X_1 - X_4 \sim Normal(24-22, \sqrt{16+9})$

$\Rightarrow X_1 - X_4 \sim Normal(2, 5)$

So, $\mathbb{P}(X_1 - X_4 > 0) = 1 - \mathbb{P}(X_1 - X_4 \leq 0) = 1 - pnorm(0, 2, 5) \approx .6554$

c) $\mathbb{P}(\frac{1}{3}(X_1 + X_2 + X_3) > X_4) = \mathbb{P}(\frac{1}{3}X_1 + \frac{1}{3}X_2 + \frac{1}{3}X_3 - X_4 > 0)$

$= 1 - pnorm(0, \frac{1}{3}(24) + \frac{1}{3}(24) + \frac{1}{3}(24) - 22, \sqrt{\frac{1}{9}(16) + \frac{1}{9}(16) + \frac{1}{9}(16) + 9})$

$= 1 - pnorm(0, 2, \sqrt{\frac{43}{3}}) \approx .7013$

Lesson 17 - Solutions

1. $X_1, X_2, X_3 \overset{iid}{\sim} N(4,5)$

a) $Y = X_1 + X_2 + X_3 \sim N(4+4+4, \sqrt{25+25+25}) \Rightarrow Y \sim N(12, \sqrt{75})$

b) $W = \frac{1}{3}(X_1+X_2+X_3) = \frac{1}{3}Y \sim N(\frac{1}{3}(12), \sqrt{\frac{75}{9}}) \Rightarrow W \sim N(4, \frac{5}{\sqrt{3}})$

c) Let $Y = \sum_{i=1}^{n} X_i$, where each $X_i \overset{iid}{\sim} N(\mu, \sigma)$. Then $Y \sim N(n\mu, \sqrt{n\sigma^2})$

Let $W = \frac{1}{n}\sum_{i=1}^{n} X_i$, where each $X_i \overset{iid}{\sim} N(\mu, \sigma)$. Then $W \sim N(\mu, \frac{\sigma}{\sqrt{n}})$

2. Standard error of $\sum_{i=1}^{n} X_i$ is $\sqrt{n\sigma^2}$

Standard error of $\frac{1}{n}\sum_{i=1}^{n} X_i$ is $\sigma/\sqrt{n}$

3. Central Limit Theorem: Let $X_1, \ldots, X_n$ be independent and identically distributed with mean μ and standard deviation σ. Then $\overline{X}$ is approximately $Normal(\mu, \frac{\sigma}{\sqrt{n}})$, and this approximation is extremely accurate for $n > 30$.

4. a) $\overline{X}_n \sim Normal(7, 5/3)$, $SE[\overline{X}_n] = 5/3$

b) $\overline{X}_n \sim Normal(7, 5/6)$, $SE[\overline{X}_n] = 5/6$

c) As the sample size is small (< 30), we cannot state anything about this distribution.

d) $\overline{X}_n \approx Normal(7, 5/6)$, $SE[\overline{X}_n] = 5/6$

Rivets $X \sim (10000, 500)$

a) $\overline{X}_{40} \approx \text{Normal}(10000, \frac{500}{\sqrt{40}})$

b) $SE[\overline{X}_{40}] = 500/\sqrt{40}$

c) $\mathbb{P}(9900 < \overline{X}_{40} < 10{,}200) = \mathbb{P}(\overline{X}_{40} < 10200) - \mathbb{P}(\overline{X}_{40} < 9900)$

$= \text{pnorm}(10200, 10000, 500/\sqrt{40}) - \text{pnorm}(9900, 10000, 500/\sqrt{40})$

$\approx .8913$

d) No! This sample size would be too small.

Uniform $X \sim \text{Unif}(10, 20)$

a) $E[X] = \frac{10+20}{2} = 15$, $Var[X] = \frac{(20-10)^2}{12} = \frac{100}{12} = 50/6 = 25/3$

b) $\overline{X}_{50} \approx \text{Normal}(15, \frac{\sqrt{25/3}}{\sqrt{50}}) \Rightarrow \overline{X}_{50} \approx \text{Normal}(15, 1/\sqrt{6})$

$SE[\overline{X}_{50}] = \frac{1}{\sqrt{6}}$

c) $\mathbb{P}(\overline{X}_{50} < 14) = \text{pnorm}(14, 15, 1/\sqrt{6}) \approx .0072$

Lesson 18 - Solutions

1. a) $X_i \overset{iid}{\sim} \text{Bernoulli}(p)$

   b) $\mathbb{E}[X_i] = p$; $Var[X_i] = p(1-p)$

   c) $\hat{p} = \frac{1}{n}\sum_{i=1}^{n} X_i$

   $$\mathbb{E}[\hat{p}] = \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n} X_i\right] = \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}[X_i] = \frac{1}{n}\sum_{i=1}^{n} p = \frac{1}{n}(np) = p$$

   $$Var[\hat{p}] = Var\left[\frac{1}{n}\sum_{i=1}^{n} X_i\right] = \frac{1}{n^2}\sum_{i=1}^{n} Var[X_i] = \frac{1}{n^2}\sum_{i=1}^{n} p(1-p) = \frac{1}{n^2}\cdot np(1-p) = \frac{p(1-p)}{n}$$

   d) $\mathbb{E}[\hat{p}] = p$, so $\hat{p}$ is unbiased

   e) $\hat{p} \sim \text{Normal}\left(p, \sqrt{\frac{p(1-p)}{n}}\right)$

   approximately

   f) $SE[\hat{p}] = \sqrt{\frac{p(1-p)}{n}}$

2. An interval estimator for a parameter θ is a range of numbers that we think of as a set of statistically reasonable values for θ.

   That is, any number in our interval estimator is a "good guess" for θ.

3. See Notes

4. Let $z_{1-\frac{1-\alpha}{2}}$ be the appropriate crit. value of the standard normal distribution.

   $$MoE = z_{1-\frac{1-\alpha}{2}}\frac{\sigma}{\sqrt{n}} \Rightarrow n = \left(\frac{z_{1-\frac{1-\alpha}{2}}\cdot\sigma}{MoE}\right)^2$$

   In this case:

   $$n = \left(\frac{1.96\cdot\sqrt{76}}{3}\right)^2 = 32.44 \Rightarrow 33 \text{ subjects}$$

Particles $\hat{\theta} = 6\overline{X}$; $X_i \overset{iid}{\sim} f(x) = \frac{1}{4}(2+\theta x)$; $-1<x<1$

$$\mathbb{E}[\hat{\theta}] = \mathbb{E}[6\overline{X}] = 6\mathbb{E}[\overline{X}] = 6\mathbb{E}[X]$$

$$= 6\int_{-1}^{1} x \cdot \frac{1}{4}(2+\theta x)\,dx = \frac{6}{4}\int_{-1}^{1} 2x + \theta x^2\,dx = \frac{3}{2}\left(x^2 + \frac{\theta}{3}x^3\right)\Big|_{-1}^{1}$$

$$= \frac{3}{2}\left[\left(1+\frac{\theta}{3}\right) - \left(1-\frac{\theta}{3}\right)\right] = \frac{3}{2}\left(\frac{2}{3}\theta\right) = \theta$$

$\Rightarrow \hat{\theta}$ is unbiased

Blades $\sigma = 20$

a) $\overline{X} = 402$ $\quad z_{.975} = 1.96$ $\quad \overline{X} \pm z_{.975} \cdot \frac{\sigma}{\sqrt{n}} = 402 \pm (1.96) \cdot \frac{20}{\sqrt{25}}$

$n = 25$ $\quad = 402 \pm 7.84$

$cl = .95$ $\quad = (394.16, 409.84)$

b) $\overline{X} = 402$ $\quad z_{.91} = 1.34$ $\quad \overline{X} \pm z_{.91} \cdot \frac{\sigma}{\sqrt{n}} = 402 \pm (1.34) \cdot \frac{20}{\sqrt{81}}$

$n = 81$ $\quad = 402 \pm 2.978$

$cl = .82$ $\quad = (399.02, 404.98)$

We are 82% confident that the true mean power is between 399.02 kW and 404.98 kW.

c) $$n = \left(\frac{z_{.95} \cdot 20}{2.5}\right)^2 = \left(\frac{1.645 \cdot 20}{2.5}\right)^2 = 173.1856 \Rightarrow 174 \text{ subjects}$$

Note: We are told that we want the entire width of the CI to be 5 kW. The MoE is half the width, hence the 2.5 in the denominator above.

Lesson 19 - Solutions

1. See Notes

2. See Notes

3. See Following page

4. a) $t_{.975,24} = 2.064$

$$\bar{X} \pm t_{.975,24} \cdot \frac{s}{\sqrt{n}} = 16 \pm 2.064\left(\frac{3}{\sqrt{25}}\right) = (14.7616, 17.2384)$$

b) See Notes

c) $\bar{X} \pm t_{.975,24} \cdot S \cdot \sqrt{\frac{n+1}{n}} = 16 \pm 2.064(3)\sqrt{\frac{26}{25}} = (9.6854, 22.3146)$

$\boxed{\text{Irisin}}$ See Notes

$\boxed{\text{Screws}}$ Population is Normal, σ known $\Rightarrow$ z critical values

$\sigma^2 = .0009 \Rightarrow \sigma = .03$

$\bar{X} = 12.6$

$n = 12$

$cl = .95 \Rightarrow z_{.975} = 1.96$

a) $12.6 \pm 1.96\left(\frac{.03}{\sqrt{12}}\right) = (12.583, 12.617)$

b) $12.6 \pm 1.96(.03)\sqrt{\frac{13}{12}} = (12.5388, 12.6612)$

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	$\sigma = 2$	z
b)	$X \sim N(\mu, \sigma^2)$	any	s = 4.1	z
c)	$X \sim (\mu, 9)$, unkn. shape	36	$\sigma = 3$	z
d)	$X \sim (\mu, \sigma^2)$, unkn. shape	44	s = 11	z or t
e)	$X \sim (\mu, 7^2)$, unkn. shape	22	$\sigma = 7$	z
f)	$X \sim (\mu, \sigma^2)$, unkn. shape	17	s = 8	t

t-crit value for small sample sizes

negligible difference

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.

$\bar{X}$

   a) Compute the 90% confidence interval for the true mean irisin level under these experimental conditions. μ
   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 - Solutions

1. See Notes

2. a) Null - $H_0: \mu \leq 260$

   Alternative - $H_a: \mu > 260$

   b) $H_0: \mu \geq 707$

   $H_a: \mu < 707$

   c) $H_0: \mu = 345$

   $H_a: \mu \neq 345$

   d) $H_0: \mu = 35$

   $H_a: \mu \neq 35$

3. Type 1 Error - rejecting a true hypothesis

   Type 2 Error - Failing to reject a false hypothesis

Roads

a) $H_0: \mu \leq 25$

$H_a: \mu > 25$

b) Concluding that the mean safety index is greater than 25 when it is actually less than or equal to 25.

c) Concluding that the true mean safety index is less than or equal to 25 when it is actually greater than 25.

Lesson 23 - Solutions

1. See Notes

2. See Notes

3. a) We reject H_0 at the .05 significance level and conclude that we have evidence for the true mean being different from 7.

   b) We reject H_0 at the .05 significance level and conclude that we have evidence for the true mean being greater than 3.

   c) We fail to reject H_0 at the .05 significance level and conclude that we do not have evidence for the true mean being less than 4.1

   d) We fail to reject H_0 at the .05 significance level and conclude that we do not have evidence for the true mean being different from 5.6

$\boxed{\text{Reaction}}$

N $H_0: \mu \le 260$

A $H_a: \mu > 260$

T $$z_{test} = \frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}} = \frac{270 - 260}{40/\sqrt{20}} \approx 1.118 \overset{D}{\sim} N(0,1)$$

$$\text{p-value} = P(Z > z_{test}) = P(Z > 1.118)$$
$$= 1 - \text{pnorm}(1.118) \approx .1318 \overset{R}{>} .05$$

C Therefore, we fail to reject H_0 at the .05 significance level and conclude that we do not have evidence for the true mean reaction time being greater than 260 ms.

Horsepower

a) N $H_0: \mu \geq 707$

A $H_a: \mu < 707$

T $t_{test} = \frac{\bar{X} - \mu_0}{S/\sqrt{n}} = \frac{702 - 707}{30/\sqrt{35}} \approx -.986 \overset{D}{\sim} t_{34}$

p-value $= \mathbb{P}(t_{34} < -.986) = pt(-.986, 34) \approx .1655 \overset{R}{>} .05$

C Therefore, we fail to reject H_0 at the .05 significance level and conclude that we do not have evidence for the true mean horsepower being less than 707 hp.

b) Since we failed to reject H_0, it is possible we have made a type 2 error.

Inflation

N $H_0: \mu = 35$

A $H_a: \mu \neq 35$

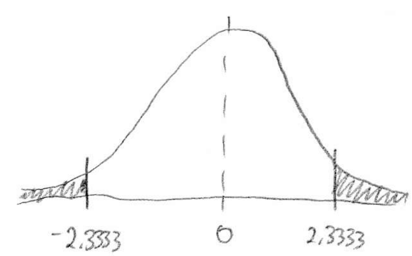
T $t_{test} = \frac{\bar{X} - \mu_0}{S/\sqrt{n}} = \frac{36.2 - 35}{2.3/\sqrt{20}} \approx 2.3333 \overset{D}{\sim} t_{19}$

p-value $= \mathbb{P}(t_{19} > 2.3333 \cup t_{19} < -2.3333)$

$= \mathbb{P}(t_{19} > 2.3333) + \mathbb{P}(t_{19} < -2.3333)$

$= 2 \cdot \mathbb{P}(t_{19} < -2.3333)$ (by symmetry)

$= 2 \cdot pt(-2.3333, 19) \approx .0308 \overset{R}{<} .05$



C Therefore, we reject H_0 at the .05 significance level and conclude that we have evidence for the true mean tire inflation being different from 35 psi.

Lesson 24 - Solutions

1. a) $t_{.975,39} = 2.0227$ $\quad \bar{X} \pm t_{.975,39} \cdot \frac{s}{\sqrt{n}} = 20 \pm 2.0227\left(\frac{4}{\sqrt{40}}\right)$

$$= (18.7207, 21.2793)$$

b) Since 22 is not in our 95% CI, our hypothesis test at the .05 significance level would reject $H_0: \mu = 22$ in favor of $H_a: \mu \neq 22$.

We show this by carrying out the test:

Ⓝ $H_0: \mu = 22$
Ⓐ $H_a: \mu \neq 22$

Ⓣ $t_{test} = \frac{20-22}{4/\sqrt{40}} \approx -3.1623$ Ⓓ $\sim t_{39}$

$\text{p-value} = 2 \cdot \mathbb{P}(t_{39} < -3.1623) \approx .003$ Ⓡ $< .05$

Ⓒ Therefore, we reject H_0 at the .05 significance level and conclude that we have evidence for μ being different from 22.

2. Ⓝ $H_0: p = .35$
Ⓐ $H_a: p \neq .35$

Ⓣ $z_{test} = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}} = \frac{.3 - .35}{\sqrt{\frac{(.35)(.65)}{100}}} \approx -1.0483$ Ⓓ $\sim N(0,1)$

$\text{p-value} = 2 \cdot \mathbb{P}(Z < -1.0483) = 2 \cdot \text{pnorm}(-1.0483) \approx .2945$ Ⓡ $> .05$

Ⓒ Therefore, we fail to reject H_0 at the .05 significance level and conclude that we do not have evidence for the proportion of green items being different from .35.

Kissing

Ⓝ $H_0: p = .67$

Ⓐ $H_a: p \neq .67$

Ⓣ $$z_{test} = \frac{\frac{80}{124} - .67}{\sqrt{\frac{.67(1-.67)}{124}}} \approx -.5882 \sim N(0,1) \text{ Ⓓ}$$

$$\text{p-value} = 2 \cdot \mathbb{P}(Z < -.5882) \approx .5564 \overset{Ⓡ}{>} .05$$

Ⓒ We fail to reject H_0 at the .05 significance level and conclude that we do not have evidence for the proportion of right-leaning kissing behaviored couples being different from 67%.

ADHD

a) $cl = .99 \Rightarrow z_{.995} = 2.576$

$$\hat{p} \pm z_{.995}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} = \frac{42}{1494} \pm 2.576\sqrt{\frac{\frac{42}{1494} \cdot \frac{1452}{1494}}{1494}} = (.0171, .0391)$$

b) $H_0: p = .03$

$H_a: p \neq .03$

As .03 is in our 99% CI, we fail to reject H_0 at the .01 sig. level and conclude that we do not have evidence for the proportion of the Netherland's elderly with ADHD being different from 3%.

Robots

Ⓝ $H_0: p \geq .05$

Ⓐ $H_a: p < .05$

Ⓣ $$z_{test} = \frac{.046 - .05}{\sqrt{\frac{.05(.95)}{500}}} \approx -.4104 \sim N(0,1) \text{ Ⓓ}$$

$$\text{p-value} = \mathbb{P}(Z < -.4104) \approx .3408 \overset{Ⓡ}{>} .05$$

Ⓒ We fail to reject H_0 at the .05 sig. level and conclude that we do not have evidence for the true proportion of defects being less than 5%.

Lesson 25 - Solutions

1. a) $\overline{X}_1 \sim Normal(10, \frac{4}{\sqrt{25}}) \sim Normal(10, 4/5)$

$\overline{X}_2 \sim Normal(7, \frac{2}{\sqrt{25}}) \sim Normal(7, 2/5)$

b) $\overline{X}_1 - \overline{X}_2 \sim Normal(10-7, \sqrt{\frac{16}{25}+\frac{4}{25}}) \sim Normal(3, \frac{2}{\sqrt{5}})$

c) $SE[\overline{X}_1 - \overline{X}_2] = 2/\sqrt{5}$

d) Suppose $X \sim Normal(\mu_1, \sigma_1)$ and $Y \sim Normal(\mu_2, \sigma_2)$ are independent. We collect a random sample of size n_1 from X, and one from Y of size n_2. Then:

$$\overline{X} - \overline{Y} \sim Normal\left(\mu_1 - \mu_2, \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}\right)$$

2. When σ's are known, use a z-critical value/Normal dist.:

CI: $(\overline{x} - \overline{y}) \pm z_{1-\frac{1-cl}{2}} \cdot \sqrt{\frac{\sigma_x^2}{n_1} + \frac{\sigma_y^2}{n_2}}$

Test-stat: $\dfrac{(\overline{x}-\overline{y}) - (\mu_1 - \mu_2)_0}{\sqrt{\frac{\sigma_x^2}{n_1} + \frac{\sigma_y^2}{n_2}}} \sim N(0,1)$

When σ's are unknown, use the t-distribution:

CI: $(\overline{x} - \overline{y}) \pm t_{1-\frac{1-cl}{2}, df} \sqrt{\frac{S_x^2}{n_1} + \frac{S_y^2}{n_2}}$

Test-stat: $\dfrac{(\overline{x}-\overline{y}) - (\mu_1 - \mu_2)_0}{\sqrt{\frac{S_x^2}{n_1} + \frac{S_y^2}{n_2}}} \sim t_{df}$

$$df = \underbrace{\left\lfloor \frac{\left(\frac{S_x^2}{n_1} + \frac{S_y^2}{n_2}\right)^2}{\frac{1}{n_1-1}\left(\frac{S_x^2}{n_1}\right)^2 + \frac{1}{n_2-1}\left(\frac{S_y^2}{n_2}\right)} \right\rfloor}_{\text{round down!}}$$ (Welch-Satterthwaite Approximation)

★ Conservative $df = \min(n_1 - 1, n_2 - 1)$

Traffic

$cl = .95 \Rightarrow z_{.975} = 1.96$

$$(\bar{x}-\bar{y}) \pm z_{.975}\sqrt{\frac{\sigma_x^2}{n_1}+\frac{\sigma_y^2}{n_2}} = (39.9-15.3) \pm (1.96)\sqrt{\frac{1.7^2}{20}+\frac{.8^2}{20}}$$

$$= 24.6 \pm .8234$$

$$= (23.7766, 25.4234)$$

We are 95% confident that the true difference between the mean number of cars in the spaced arrangement and the bunched arrangement is between 23.7766 and 25.4234.

Tools ★ Assume σ's are known ★

$$(16.2-13.9) \pm (1.96)\sqrt{(.6)^2+(.9)^2} = 2.3 \pm 2.1201 = (.1799, 4.4201)$$

We are 95% confident that the true difference in mean lifetimes between grade A and grade B is between .1799 and 4.4201 minutes.

Stopping

b) $$df = \left\lfloor \frac{\left(\left(\frac{5.4^2}{8}\right)+\left(\frac{5^2}{10}\right)\right)^2}{\frac{1}{8-1}\left(\frac{5.4^2}{8}\right)^2+\frac{1}{10-1}\left(\frac{5^2}{10}\right)^2} \right\rfloor = 14 \Rightarrow t_{.975,14} = qt(.975,14) = 2.1448$$

$$(126.8-119.3) \pm (2.1448)\sqrt{\frac{5.4^2}{8}+\frac{5^2}{10}} = 7.5 \pm 5.3168 = (2.1832, 12.8168)$$

c) $H_0: \mu_1 - \mu_2 = 0$

$H_a: \mu_1 - \mu_2 \neq 0$

As 0 is not in our 95% CI, we reject H_0 at the .05 sig-level and conclude that we have evidence that the braking times for the two systems are different.

Lesson 26 - Comparing Two Parameters continued

$H_0: \mu_1 - \mu_2 = 0$
$H_a: \mu_1 - \mu_2 \neq 0$
Are the two means the same?

In Lesson 25, we assumed our samples are independent.

In Lesson 26, we look at "matched pairs" comparison
↳ every X_i has a corresponding Y_i

Transform the data into differences:

$$d_i = X_i - Y_i$$

point estimate: $\bar{d}$
standard error: $\frac{S_d}{\sqrt{n}}$

value from hypotheses

CI: $\bar{d} \pm (\text{c.v.}) \cdot \frac{S_d}{\sqrt{n}}$
often t_{cv}

test-statistic $= \frac{\bar{d} - (\mu_1 - \mu_2)_0}{S_d/\sqrt{n}}$

1.a) Independent Samples CI $= (-46.9568, 66.9568)$
$t_{.975,7} = 2.365$

b) $\bar{d} = 10$, $S_d = 15.8114$, $n = 5$

Matched Pairs CI

$$\bar{d} \pm t_{.975,4} \cdot \frac{S_d}{\sqrt{n}} = 10 \pm 2.7764\left(\frac{15.8114}{\sqrt{5}}\right) = (-9.6321, 29.6321)$$

In matched pairs, there is dependence between $\bar{X}$ & $\bar{Y}$

$Var[\bar{X} - \bar{Y}] = Var[\bar{X}] + Var[\bar{Y}] - 2\,Cov[\bar{X}, \bar{Y}]$ < Variance in independent samples.

- Difference in proportions

☆ Always assume the two sample proportions are independent.

CI: $(\hat{p}_1 - \hat{p}_2) \pm Z_{1-\frac{1-CL}{2}} \sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}$

$$\text{Test-statistic} = \frac{(\hat{p}_1 - \hat{p}_2) - (p_1 - p_2)_0}{\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}} \sim N(0,1)$$

$\boxed{\text{Corruption}}$

p_u = proportion of people in the UK that believe the government is corrupt

p_c = " " " " Canada " " " " " "

$\hat{p}_u = \frac{230}{500}$; $\hat{p}_c = \frac{264}{600}$

$$\left(\frac{230}{500} - \frac{264}{600}\right) \pm (1.96)\sqrt{\frac{\left(\frac{230}{500}\right)\left(\frac{270}{500}\right)}{500} + \frac{\left(\frac{264}{600}\right)\left(\frac{336}{600}\right)}{600}} = (-.0390, .0790)$$

Ⓝ $H_0: p_u - p_c = 0$

→Ⓐ $H_a: p_u - p_c \neq 0$

As 0 is in the 95% CI for $p_u - p_c$, we fail to reject H_0 at the .05 sig. level and conclude that we do not have evidence for a difference in the true proportions

Lesson 27 - Solutions

1. See Notes

2. See Notes

3. a) $H_0: \mu_1 = \mu_2 = \dots = \mu_k$

H_a: At least one μ_i is different

c) $SSTR = \sum_{i=1}^{k} n_i(\bar{X}_{i.} - \bar{X}_{..})^2$, $df = k-1$

$SSE = \sum_{i=1}^{k} (n_i - 1)s_i^2$, $df = N-k$

$MSTR = \dfrac{SSTR}{k-1}$, $MSE = \dfrac{SSE}{N-k}$

$F_{test} = \dfrac{MSTR}{MSE}$

k = # of groups
N = overall sample size
n_i = sample size of group i
$\bar{X}_{..}$ = mean of overall data set
$\bar{X}_{i.}$ = mean of the i^{th} group
S_i^2 = variance of i^{th} group
X_{ij} = j^{th} data value from i^{th} group

4. $$\bar{X}_{..} = \frac{\sum_{i,j} X_{ij}}{6 \cdot 4} = \frac{1}{4}\left(\frac{1}{6}\sum_{i,j} X_{ij}\right) = \frac{1}{4}\left[\frac{1}{6}\sum_{i=1}^{4}\sum_{j=1}^{6} X_{ij}\right] = \frac{1}{4}\left[\sum_{i=1}^{4}\left(\frac{1}{6}\sum_{j=1}^{6} X_{ij}\right)\right] = \frac{1}{4}\sum_{i=1}^{4}\bar{X}_{i.}$$

$$= \frac{1}{4}(15 + 24 + 12 + 21) = 18$$

$$SSTR = \sum_{i=1}^{4} 6(\bar{X}_{i.} - 18)^2 = 6\left[(15-18)^2 + (24-18)^2 + (12-18)^2 + (21-18)^2\right] = 540, \quad df = 4 - 1 = 3$$

$$SSE = \sum_{i=1}^{4}(6-1)s_i^2 = 5(2^2 + 3^2 + 3^2 + 2^2) = 130, \quad df = 24 - 4 = 20$$

$$MSE = \frac{130}{20} = 13/2 = 6.5, \quad MSTR = \frac{540}{3} = 180$$

$$F_{test} = \frac{180}{6.5} = 27.69231$$

5. See next page for the completed table.

Ⓝ $H_0: \mu_1 = \mu_2 = \mu_3$

Ⓐ H_a: At least one group mean differs

Ⓣ $F_{test} = 10 \sim$ Ⓓ $F_{2,12} \Rightarrow$ p-value $= \mathbb{P}(F_{2,12} > 10) = 1 - pf(10, 2, 12) \approx .0028 <$ Ⓡ $.05$

Ⓒ We reject H_0 at the .05 significance level and conclude that we have evidence that there is a difference among the three means.

$\boxed{\text{Cells}}$

Source	df	SS	MS	F_{test}
Treatment	2	1.476	.738	.9969
Error	11	8.143	.7403	—
Total	13	9.619	—	—

Ⓝ $H_0: \mu_A = \mu_B = \mu_C$

Ⓐ H_a: At least one mean differs

Ⓣ $F_{test} = .9969 \sim$ Ⓓ $F_{2,11}$

p-value $= 1 - pf(.9969, 2, 11) \approx .40 >$ Ⓡ $.05$

Ⓒ We fail to reject H_0 at the .05 significance level and conclude that we have evidence for the three mean cell concentrations are the same.

Lesson 27 – Solutions

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	2	60	30	10
Error (residual)	12	36	3	xx
Total	14	96	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 - Solutions

1. The Tukey-Kramer CI for comparing the means of groups i & j:

$$(\bar{X}_{i\cdot} - \bar{X}_{j\cdot}) \pm q \cdot \sqrt{\frac{MSE}{2}\left(\frac{1}{n_i} + \frac{1}{n_j}\right)}$$

where q is a critical value from the Q-distribution.

$$q = \text{qtukey}\Big(cl, \underbrace{k}_{\#\text{ of groups}}, \sum_{i=1}^{k}(n_i - 1)\Big)$$

Flicker

We have p-value $= .023 < .05$, so H_0 is rejected. Our Tukey-Kramer CI's are:

$$(\bar{X}_{1\cdot} - \bar{X}_{2\cdot}) \pm q\sqrt{\frac{MSE}{2}\left(\frac{1}{n_1} + \frac{1}{n_2}\right)} = (25.6 - 26.9) \pm (3.6491)\sqrt{\frac{2.394}{2}\left(\frac{1}{8} + \frac{1}{5}\right)} = (^-3.5760, 0.9760)$$

$$(\bar{X}_{1\cdot} - \bar{X}_{3\cdot}) \pm q\sqrt{\frac{MSE}{2}\left(\frac{1}{n_1} + \frac{1}{n_3}\right)} = (25.6 - 28.2) \pm (3.6491)\sqrt{\frac{2.394}{2}\left(\frac{1}{8} + \frac{1}{6}\right)} = (^-4.7561, ^-.4439)$$

$$(\bar{X}_{2\cdot} - \bar{X}_{3\cdot}) \pm q\sqrt{\frac{MSE}{2}\left(\frac{1}{n_2} + \frac{1}{n_3}\right)} = (26.9 - 28.2) \pm (3.6491)\sqrt{\frac{2.394}{2}\left(\frac{1}{5} + \frac{1}{6}\right)} = (^-3.7175, 1.1175)$$

As zero is not in the CI for $\mu_1 - \mu_3$, we have evidence that $\mu_1 \neq \mu_3$.

Lesson 30 - Solutions

1. See R Code
2. See R Code
3. See R Code
4. See Notes
5. See Notes
6. See Notes
7. $Y = 3 + 5X + \varepsilon \; ; \quad \varepsilon \sim N(0,4)$

a) $Y = 3 + 5X$

b) $\beta_1 = 5$: We expect Y to increase by 5 units when X increases by 1 uni

c) $\varepsilon \sim N(0,4) \Rightarrow Y = 3+5X+\varepsilon \sim N(3+5X, 4)$

$\Rightarrow \mathbb{E}[Y] = 3+5X \quad @\ X=6, \ \mathbb{E}[Y] = 3+5(6) = 3+30 = 33$

d) $\mathbb{P}(Y<30 \mid X=6) = \mathbb{P}(N(33,4)<30) =$ pnorm$(30,33,4) \approx .2266$

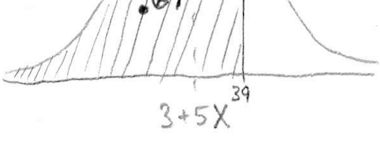
e) Let Y_1 be the value of Y when $X=8 \Rightarrow Y_1 \sim N(43,4)$
Let Y_2 " " " " " when $X=6 \Rightarrow Y_2 \sim N(33,4)$ } Independent

$\mathbb{P}(Y_1 > Y_2) = \mathbb{P}(Y_2 - Y_1 < 0) \; ; \; Y_1 - Y_2 \sim N(43-33, \sqrt{4^2+4^2}) \sim N(10, \sqrt{32})$

$=$ pnorm$(0, 10, \sqrt{32}) \approx .0385$

f) See R Code; $z_{.67} \approx .44$

g) $\mathbb{P}(Y<39) = .67 \Rightarrow \mathbb{P}\left(\frac{Y-(3+5X)}{4} < \frac{39-(3+5X)}{4}\right) = .67 \Rightarrow \mathbb{P}\left(Z < \frac{36+5X}{4}\right) = .67$



$\Rightarrow \frac{36-5X}{4} = .44 \Rightarrow 36-5X = 1.76 \Rightarrow {}^{-}5X = {}^{-}34.24$

$\Rightarrow X = 6.848$

Lesson 31 – Solutions

1. $S_{xy} = \sum(x_i-\bar{x})(y_i-\bar{y}) = \sum x_i y_i - \frac{1}{n}(\sum x_i)(\sum y_i)$

$\boxed{2 \& 3}$

We want to use our data to fit $\hat{y} = b_0 + b_1 x$; that is, we need to estimate b_0 & b_1. To do this, we minimize our "residuals"

$$\ell(b_0,b_1) = \sum(y_i-\hat{y}_i)^2 = \sum(y_i-(b_0+b_1x_i))^2 = \sum(y_i-b_0-b_1x_i)^2$$

$$\Rightarrow \frac{\partial}{\partial b_0}\ell = \sum[2(y_i-b_0-b_1x_i)(-1)] = 0$$

$$\frac{\partial}{\partial b_1}\ell = \sum[2(y_i-b_0-b_1x_i)(-x_i)] = 0$$

$$\Rightarrow \left.\begin{array}{l} nb_0 + b_1(\sum x_i) = \sum y_i \\ b_0(\sum x_i) + b_1(\sum x_i^2) = \sum x_i y_i \end{array}\right\} \text{Normal Equations}$$

$$\Rightarrow b_0 = \frac{\sum y_i - b_1\sum x_i}{n} = \bar{y} - b_1\bar{x}$$

$$\Rightarrow (\bar{y}-b_1\bar{x})(\sum x_i) + b_1(\sum x_i^2) = \sum x_i y_i \Rightarrow b_1\left(\sum x_i^2 - \frac{1}{n}(\sum x_i)^2\right) + \frac{1}{n}(\sum x_i)(\sum y_i) = \sum x_i y_i$$

$$\Rightarrow b_1 = \frac{\sum x_i y_i - \frac{1}{n}(\sum x_i)(\sum y_i)}{\sum x_i^2 - \frac{1}{n}(\sum x_i)^2}$$

$$\Rightarrow b_1 = \frac{S_{xy}}{S_{xx}}$$

In all: $b_1 = \frac{S_{xy}}{S_{xx}}$, $b_0 = \bar{y} - b_1\bar{x}$

4. See R Code

5. a) residual $= \hat{e} =$ (observed y) − (predicted y)
$= y - \hat{y}$

b) See R Code

c) & d)
$SSE = \sum(y_i - \hat{y}_i)^2 = \sum(y_i - b_0 + b_1 x_i)^2 = \sum y_i^2 - b_0 \sum y_i - b_1 \sum x_i y_i$

$SSR = \sum(\hat{y}_i - \bar{y})^2$

$SST = \sum(y_i - \bar{y})^2$

$\underbrace{SST}_{\text{Total Sum of Squares}} = \underbrace{SSR}_{\text{Regression Sum of squares}} + \underbrace{SSE}_{\text{Error Sum of squares}}$

e)

Source	df	SS	MS	F_{test}	f)
Regression	p	SSR	$\frac{SSR}{p} = MSR$	$\frac{MSR}{MSE}$	$\sim F_{p,n-p-1}$
Error (Residual)	$n-p-1$	SSE	$\frac{SSE}{n-p-1} = MSE$	×	
Total	$n-1$	SST	×	×	

$H_0: \beta_1 = 0$ $\quad$ p = # of explanatory variables
$H_a: \beta_1 \neq 0$ $\quad$ n = # of observations

g & h) $\hat{\sigma}^2 = s^2 = MSE \Rightarrow S = \sqrt{MSE}$

6. a) $\hat{y} = 3.1 + 2x$

b) We expect a 2 unit increase in y for a 1 unit increase in x.

c) On Sheet, $F_{test} = \frac{156.72}{1.304} \approx 120.184$

d) $n = (df_1) + (df_2) + 1 = 1 + 3 + 1 = 5$

e) $MSE = 3.91$

f) $S = \sqrt{3.91} \approx 1.9774$

Lesson 33 - Solutions

1. See R Code & Notes - answer is <u>no</u>, but it can happen

2. SLR Model utility test = F-test (ANOVA) from Lesson 31

$H_0: \beta_1 = 0$ (Model is not useful)

$H_a: \beta_1 \neq 0$ (Model is useful)

3. $$\hat{\beta}_1 = b_1 = \frac{S_{xy}}{S_{xx}} = \frac{1}{S_{xx}}\sum(x_i-\bar{x})(y_i-\bar{y}) = \frac{1}{S_{xx}}\Big[\sum(x_i-\bar{x})y_i - \underbrace{\sum(x_i-\bar{x})\bar{y}}\Big]$$

$$= \bar{y}\cdot\sum(x_i-\bar{x}) = \bar{y}(\sum x_i - n\bar{x}) = \bar{y}(\sum x_i - \sum x_i) = 0$$

$$= \frac{1}{S_{xx}}\Big[\sum(x_i-\bar{x})y_i\Big] = \sum\left(\frac{x_i-\bar{x}}{S_{xx}}\right)y_i$$ Linear combination of normals!

$$y_i \sim N(\beta_0+\beta_1 x_i, \sigma) \Rightarrow \sum\left(\frac{x_i-\bar{x}}{S_{xx}}\right)y_i = b_1 \text{ is Normal!}$$

$$E[b_1] = E\left[\sum\left(\frac{x_i-\bar{x}}{S_{xx}}\right)y_i\right] = \sum\left(\frac{x_i-\bar{x}}{S_{xx}}\right)E[y_i] = \sum\left(\frac{x_i-\bar{x}}{S_{xx}}\right)(\beta_0+\beta_1 x_i)$$

$$= \frac{1}{S_{xx}}\Big[\beta_0\cdot\sum(x_i-\bar{x}) + \beta_1\cdot\sum(x_i-\bar{x})x_i\Big] = \frac{1}{S_{xx}}\Big[\beta_1\cdot\sum(x_i^2 - x_i\bar{x})\Big]$$

$$= \frac{1}{S_{xx}}\cdot\beta_1\left(\sum x_i^2 - \bar{x}\sum x_i\right) = \frac{1}{S_{xx}}\beta_1\left(\sum x_i^2 - \frac{1}{n}(\sum x_i)^2\right) = \frac{1}{S_{xx}}\cdot\beta_1\cdot S_{xx} = \beta_1$$

$$Var[b_1] = Var\left[\sum\left(\frac{x_i-\bar{x}}{S_{xx}}\right)y_i\right] = \sum\left(\frac{x_i-\bar{x}}{S_{xx}}\right)^2\cdot Var[y_i] \quad \text{(by independence)}$$

$$= \frac{1}{S_{xx}^2}\sum(x_i-\bar{x})^2\cdot\sigma^2 = \frac{S_{xx}}{S_{xx}^2}\cdot\sigma^2 = \frac{\sigma^2}{S_{xx}}$$

$$\Rightarrow b_1 \sim \text{Normal}\left(\beta_1, \frac{\sigma}{\sqrt{S_{xx}}}\right)$$

4. CL% CI for β_1: $$b_1 \pm t_{1-\frac{1-CL}{2}, df}\cdot\frac{s}{\sqrt{S_{xx}}} \quad (df = n-2)$$

$s = \sqrt{MSE}$

5. a) $r^2 = 1 - \frac{SSE}{SST} = \frac{SSR}{SST}$

b) $r = \sqrt{r^2} = \sqrt{\frac{SSR}{SST}}$ is the (linear) correlation coefficient from Lesson 15!
(we just called it Corr[X,Y])

$\boxed{\text{Search}}$ $n=14$, $\bar{X} = 24.8$, $S_{xx} = 1450.4$, $SSR = 536.58$, $SSE = 25$, $b_0 = .32$, $b_1 = .61$

a) $\hat{y} = b_0 + b_1 x \Rightarrow \hat{y} = .32 + .61x$

b)

Source	df	SS	MS	F_{test}
Regression	1	536.58	536.58	257.5625
Residuals	12	25	2.0833	×
Total	13	561.58	×	×

c) Ⓝ $H_0: \beta_1 = 0$ Ⓣ $F_{test} = 257.5625 \sim$ Ⓓ $F_{1,12}$

Ⓐ $H_a: \beta_1 \neq 0$ $p\text{-value} = 1 - pf(257.5625, 1, 12) \approx 0 <$ Ⓡ $.05$

Ⓒ We reject H_0 at the .05 sig. level and conclude that we have evidence for $\beta_1 \neq 0$, indicating that our SLR model is useful.

d) $(.5274, .6926) \longrightarrow$ See R Code for calculation

Yes! Zero is not in our CI, so reject $H_0: \beta_1 = 0$.

e) $t_{test} = \frac{b_1 - 0}{s/\sqrt{S_{xx}}} = \frac{.61}{\sqrt{2.0833}/\sqrt{1450.4}} \approx 16.09525$

f) $r^2 = \frac{SSR}{SST} = \frac{536.58}{561.58} \approx .9555 \Rightarrow$ About 95.5% of the variation

g) $r = \sqrt{r^2} \approx .9775 \Rightarrow$ strong, positive, linear relationship

Lesson 34 - Solutions

1. a) CI for $\mathbb{E}[Y|X^*]$: $\hat{y} \pm t_{1-\frac{1-cl}{2},df} \cdot S\sqrt{\frac{1}{n}+\frac{(x^*-\bar{x})^2}{S_{xx}}}$ $(df=n-2)$

   b) PI for $(Y|X^*)$: $\hat{y} \pm t_{1-\frac{1-cl}{2},df} \cdot S\sqrt{1+\frac{1}{n}+\frac{(x^*-\bar{x})^2}{S_{xx}}}$ $(df=n-2)$

   $S_{xx} = \Sigma(x_i-\bar{x})^2$

2. a) The bold 1.15 is the intercept estimate, b_0
   The bold 3.92 is the slope estimate, b_1
   The bold 0.764 is the standard error of b_1
   The bold 5.13 is the test statistic for the test of $H_0: \beta_1=0$ vs. $H_a: \beta_1 \neq 0$.

   b) $\hat{y} = 1.15 + 3.92x$

   c) Since the model utility test considers $H_0: \beta_1=0$ vs. $H_a: \beta_1 \neq 0$, the p-value is the probability corresponding to the test statistic of 5.13. This is .001, as in the table.

   d)

Source	df	SS	MS	F
Regression	1	1228	1228	26.2885
Error	8	373.7	46.7125	×
Total	9	1601.7	×	×

   e) There's no reason to think our model will represent the relationship between the variables outside of the data we observed!

f) Since we only observed x-values from 2 to 10, estimating y at 16 would be extrapolation.

g) $\hat{y}(5) = 1.15 + 3.92(5) = 20.75$

$\bar{X} = \frac{60}{10} = 6$, $S_{xx} = 440 - \frac{1}{10}(60)^2 = 80$, $s = \sqrt{MSE} = \sqrt{46.7125} \approx 6.835$

95% PI for $(Y|5)$: $20.75 \pm t_{.975,8} \cdot 6.835\sqrt{\frac{1}{10} + \frac{(5-6)^2}{80}}$

$= 20.75 \pm 2.306(2.2925)$

$= 20.75 \pm 5.2866 = (15.4634, 26.0366)$

$\boxed{\text{Search 2}}$ $s = \sqrt{MSE} = \sqrt{\frac{SSE}{n-p-1}} = \sqrt{\frac{25}{14-1-1}} = \sqrt{\frac{25}{12}} = \frac{5}{2\sqrt{3}}$; $t_{.975,12} = 2.1788$

a) 95% CI for $E[Y|30]$: $(.32 + .61(30)) \pm (2.1788)\left(\frac{5}{2\sqrt{3}}\right)\sqrt{\frac{1}{14} + \frac{(30-24.8)^2}{1450.4}}$

$= 18.62 \pm .9438$

$= (17.6762, 19.5638)$

b) 95% PI for $(Y|30)$: $18.62 \pm (2.1788)\left(\frac{5}{2\sqrt{3}}\right)\sqrt{1 + \frac{1}{14} + \frac{(30-24.8)^2}{1450.4}}$

$= 18.62 \pm 3.2834$

$= (15.3366, 21.9034)$

Lesson 35 - Solutions

1. • Linearity
   • Normality (of error term)
   • Homoskedasticity (constant variance of error term)

2. See Notes

3. See attached pictures

4. Our assumptions seem reasonable:
   ▫ Scatterplot is fairly linear - Linearity ✓
   ▫ QQ-plot is fairly linear; there is a slight S-shape, but not enough to violate normality - Normality ✓
   ▫ Residual plot is not patterned and seems to have a constant spread over the x-values - Linearity + Homoskedasticity ✓

5. Our assumptions are <u>NOT</u> met - scatterplot is not linear, QQ-plot is curved, and residual plot is patterned.

a) Quadratic Regression Model: $Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \varepsilon\ ;\ \varepsilon \sim N(0,\sigma)$
   Quadratic Regression Line: $\hat{y} = b_0 + b_1 x + b_2 x^2$

b) $\hat{y} = 5.4 + 4.1x - x^2$

c) $x=8 \Rightarrow \hat{y} = 5.4 + 4.1(8) - 8^2 = {}^-25.8 \Rightarrow \text{resid} = {}^-30 - ({}^-25.8) = {}^-4.2$

d) $R^2 = 1 - \frac{SSE}{SST} = \frac{SSR}{SST}$

   Adjusted $R^2 = R_a^2 = 1 - \left(\frac{n-1}{n-k-1}\right)\left(\frac{SSE}{SST}\right)$

   $R^2 = \frac{17256}{17256 + 409.9} \approx .9768$

   $R_a^2 \approx .9758$

   Approximately 97.5% of the variability in Y is explained by our model.

e) No, this would be extrapolation!

f) $\hat{y} = -x^2 + 4.1x + 5.4$

vertex (x-value) of a quadratic is at $x = \frac{-b}{2a} = \frac{-4.1}{2(-1)} = 2.05$

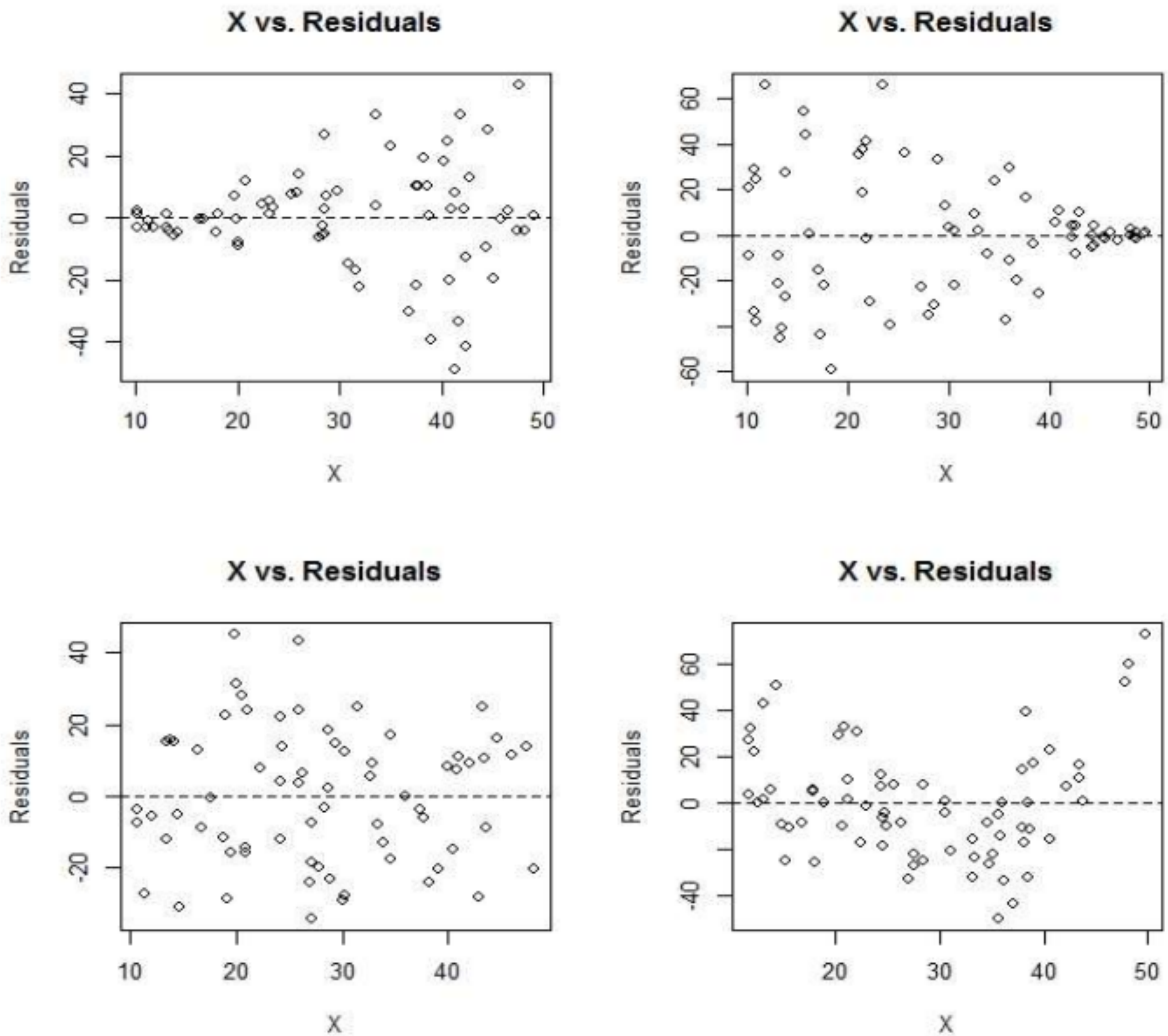
g) 95% CI for β_1: $b_1 \pm t_{.975, 50-3} \cdot SE[b_1]$

$= 4.1 \pm (2.0117)(.55)$

$= 4.1 \pm 1.106435$

$\approx (2.99, 5.21)$

Examples of residual plots



Lesson 36 - Solutions

1. In multiple linear regression, we wish to relate Y to $X_1, X_2, \ldots, X_p$ (p predictor variables) through a linear function. That is:

   Model: $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_p X_p + \varepsilon$; $\varepsilon \sim \text{Normal}(0, \sigma)$

   Fitted Plane: $\hat{y} = b_0 + b_1 x_1 + b_2 x_2 + \ldots + b_p x_p$

2. a) $y = 3 + 5x_1 - x_2 + 2x_3$

   b) β_1: We expect a 5 unit increase in y when x_1 increases by one, holding x_2 and x_3 constant.

   β_2: We expect a one unit decrease in y when x_2 increases by one, controlling for x_1 and x_3.

   c) $\mathbb{E}[Y \mid X_1 = 6, X_2 = 10, X_3 = 3] = 3 + 5(6) - 10 + 2(3) = 29$

   d) $(Y \mid X_1 = 6, X_2 = 10, X_3 = 3) \sim \text{Normal}(29, 14)$

   $\mathbb{P}(Y < 30 \mid X_1 = 6, X_2 = 10, X_3 = 3) = \text{pnorm}(30, 29, 14) \approx .5285$

   e) $(Y \mid X_1 = 7, X_2 = 20, X_3 = 2) \sim \text{Normal}(22, 14)$

   $$(Y \mid X_1 = 7, X_2 = 20, X_3 = 2) - (Y \mid X_1 = 6, X_2 = 10, X_3 = 3) \sim \text{Normal}(22 - 29, \sqrt{14^2 + 14^2})$$
   $$\sim \text{Normal}(-7, \sqrt{392})$$

   $\text{pnorm}(0, -7, \sqrt{392}) \approx .6382$

3. a) $\hat{y} = 2.89 + 2.18x_1 + .85x_2$

b) $\hat{y} = 2.89 + 2.18(2) + .85(5) = 11.5 \Rightarrow \text{resid} = y - \hat{y} = 14 - 11.5 = 2.5$

c) We expect a .85 increase in y when x_2 increases by one, controlling for x_1.

d)

Source	df	SS	MS	F
Regression	$p = 2$	1948.77	974.385	34.8904
Error	$n-p-1 = 37$	1033.3	27.927	×
Total	$n-1 = 39$	2982.07	×	×

e) $s^2 = MSE = 27.927$

f) Ⓝ $H_0: \beta_1 = \beta_2 = 0$

Ⓐ H_a: At least one is not zero

Ⓣ $F_{test} = 34.8904 \overset{Ⓓ}{\sim} F_{2,37}$

$\text{p-value} = 1 - pf(34.8904, 2, 37) \approx 3 \cdot 10^{-9} \overset{Ⓡ}{<} .05$

Ⓒ We reject H_0 at the .05 significance level and conclude that we have evidence for at least one of our model's slopes being non-zero.

g) $\hat{y} = 2.89 + 2.18(3) + .85(8) = 16.23$

h) $R_a^2 = 1 - \left(\frac{n-1}{n-k-1}\right)\left(\frac{SSE}{SST}\right) = 1 - \left(\frac{39}{36}\right)\left(\frac{1033.3}{2982.07}\right) \approx .6246$

Approximately 62.46% of the variability in y is explained by the variability in x_1 & x_2.

Lesson 37 - Solutions

1. An indicator variable is one that is equal to 1 if a specified condition is met, and 0 if it is not met. Including this in a regression model allows for a different intercept under certain cases.

   An interaction effect (or variable) is the product of predictor variables, often one usual predictor and an indicator variable. Including this in a regression model allows for a different slope under certain cases.

2. a) $\hat{y} = 3 + 2x_1 + 5x_2 - 4x_1x_2$

   b) $\hat{y}_A = 3 + 2x_1 + 5(0) - 4x_1(0) \Rightarrow \hat{y}_A = 3 + 2x_1$

   c) $\hat{y}_B = 3 + 2x_1 + 5(1) - 4x_1(1) \Rightarrow \hat{y}_B = 8 - 2x_1$

   d) OMIT

   e) OMIT

3. OMIT

$\boxed{\text{Hurricane}}$ OMIT

$\boxed{\text{Sonic}}$

a) $\hat{y} = 9.9 + 5x_1 + 9.6x_2$

b) $\hat{y}_{screw} = 9.9 + 5x_1 + 9.6(0) = 9.9 + 5x_1$

$\hat{y}_{blast} = 9.9 + 5x_1 + 9.6(1) = 19.5 + 5x_1$