<u>CREDIT</u>: The questions on this document were written by Erik Packard, PhD, Associate Professor of Mathematics at Colorado Mesa University.

1. Here are sample data about their hit from the first 34 games for CMU women's softball team this season (they are 34-0) for two players, Brooke Hodgson and Kaila Jacobi.

	Singles	**Doubles**	**Triples**	**Homeruns**
Brooke Hodgson	35	18	2	12
Kaila Jacobi	14	15	1	5

A) (7 pts) Give the distribution in percentages of the different type of hits for Brooke Hodgson.

Total Hits $= 35 + 18 + 2 + 12 = 67 \quad \therefore \quad \frac{35}{67} \times 100 = 52.2\% \qquad \frac{18}{67} \times 100 = 26.9\%$

$\frac{2}{67} \times 100 = 3.0\% \qquad \frac{12}{67} \times 100 = 17.9\%$

	Singles	**Doubles**	**Triples**	**Homeruns**
Percentages	52.2%	26.9%	3.0%	17.9%

B) (6 pts) Which player had a higher percent of their hits as doubles?

For Kaila Jacobi, Total Hits $= 14 + 15 + 1 + 5 = 35 \quad \therefore \quad \frac{15}{35} \times 100 = 42.9\% > 26.9\%$

Therefore, Kaila Jacobi had a higher percent of their hits as doubles.

2. (14 pts) A candy comes in 4 colors, RED, GREEN, BLUE, YELLOW. The manufacturer claims that 30% are RED, 30% are GREEN, 20% are BLUE, and 20% are YELLOW. Can we conclude at the 5% level of significance the distribution is different than stated? DATA: A random sample of candies had 58 RED, 50 GREEN, 25 BLUE, and 17 YELLOW. Make sure to give the critical value(s) from the table and the test statistic from the data as well as the Yes or No answer.

H_0 : The candy's color follows a distribution of...

Color	**Percentage**
RED	30%
GREEN	30%
BLUE	20%
YELLOW	20%

H_a : The candy's color follows a different distribution.

Given that $\alpha = 0.05$ and $\text{df} = 4 - 1 = 3$, $\left(\chi^2\right)_\alpha = 7.81$

For RED, $E = (0.30)(150) = 45$

For GREEN, $E = (0.30)(150) = 45$

For BLUE, $E = (0.20)(150) = 30$

For YELLOW, $E = (0.20)(150) = 30$

Color:	RED	GREEN	BLUE	YELLOW	**Total**:
Observed:	58	50	25	17	150
Expected:	45	45	30	30	150

NOTE: The **Total** for both the Observed and Expected values must be the same!

$$\left(\chi^2\right)_{\text{Data}} = \sum\left[\frac{(O-E)^2}{E}\right] = \frac{(58-45)^2}{45} + \frac{(50-45)^2}{45} + \frac{(25-30)^2}{30} + \frac{(17-30)^2}{30} = 10.778$$

Since $\left(\chi^2\right)_{\text{Data}} > \left(\chi^2\right)_{\alpha}$, we can reject the null hypothesis (Yes).

3. (5 pts) Suppose somebody else collects data for #2. What is the chance they will conclude the distribution differs from what the manufacturer states when in fact the manufacturer is correct?

   5% or less

4. (5 pts) Suppose somebody else collects data for #2. What is the chance they will not conclude the distribution differs from what the manufacturer states when in fact the manufacturer is wrong?

   Unknown

5. (14 pts) Can we conclude at the 1% significance level there is any relationship between gender and favorite color of ice cream (chosen from vanilla, chocolate, and strawberry)? Data in table gives results from a random sample. Make sure to give the critical value(s) from the table and the test statistic from the data as well as the Yes or No answer.

   The problem at hand is a contingency table problem. Therefore,

   H_0 : There is no relationship between gender and favorite flavor of ice cream.

   H_a : There is a relationship between gender and favorite flavor of ice cream.

   Given that $\alpha = 0.01$ and df = (# of rows – 1)(# of columns – 1) = (2 – 1)(3 – 1) = 2,

   $\left(\chi^2\right)_{\alpha} = 9.21$

Observed Values Table:

	Vanilla	**Chocolate**	**Strawberry**	Total:
Men	35	65	20	120
Women	70	30	80	180
Total:	105	95	100	300

The expected value for each gender-ice cream flavor pair is $E = \dfrac{\text{(Row Total)(Column Total)}}{\text{(Grand Total)}}$

Expected Values Table:

	Vanilla	**Chocolate**	**Strawberry**	Total:
Men	42	38	40	120
Women	63	57	60	180
Total:	105	95	100	300

NOTE: Notice how the **Total** values in both the Observed and Expected tables match!

$$\left(\chi^2\right)_{\text{Data}} = \sum\left[\frac{(O-E)^2}{E}\right]$$

$$= \frac{(35-42)^2}{42} + \frac{(65-38)^2}{38} + \frac{(20-40)^2}{40} + \frac{(70-63)^2}{63} + \frac{(30-57)^2}{57} + \frac{(80-60)^2}{60} = 50.585$$

Since $\left(\chi^2\right)_{\text{Data}} > \left(\chi^2\right)_{\alpha}$, we can reject the null hypothesis (Yes).

6. (10 pts) Suppose someone else collects data for #5 and they got $\left(\chi^2\right)_{\text{Data}} = 5.991$. Give their *p*-value and its meaning in everyday terms.

In this problem, we are *given* a test statistic. All we have to do is determine the *p*-value for the *given* test statistic and explain it in everyday terms.

If $\left(\chi^2\right)_{\text{Data}} = 5.991$, then the *p*-value is equal to the tail area that lies to right of the test statistic, which is 0.05.

What this means is if there's no relationship (H_0) between gender and favorite flavor of ice cream, then the chance of finding as strong or stronger evidence suggesting that there is a relationship (H_a) between gender and favorite flavor of ice cream is 0.05 (*p*-value).

7. (14 pts) Speeds of cars are measured along I-70 near Loma, Colorado. Assume the data is random from normal populations with equal variances. Can we conclude at the 5% significance level that there is any difference in the average speed of cars with license plates from Colorado, Utah, and California? Make sure to give the critical value(s) from the table and the test statistic from the data as well as the Yes or No answer.

Colorado	**Utah**	**California**
$n = 92$	$n = 21$	$n = 10$
$\overline{x} = 76$	$\overline{x} = 79$	$\overline{x} = 82$
$s = 3.4$	$s = 4.5$	$s = 6.1$

Since we're comparing multiple means, the problem at hand is an ANOVA hypothesis test. Therefore,

$\text{H}_0 : \mu_{\text{CO}} = \mu_{\text{UT}} = \mu_{\text{CA}}$

H_a : At least one of the three states differs from the other two.

To find the critical value, let's start off by finding the $\text{df}_{\text{Factor}}$ and df_{Error} values.

$$\text{df}_{\text{Factor}} = S - 1 = 3 - 1 = 2 \qquad \text{df}_{\text{Error}} = \sum n_i - S = 123 - 3 = 120$$

NOTE: S stands for # of sources.

We will now use Dr. Packard's F table. On his F table, df_{Error} = Vertical Column, and $\text{df}_{\text{Factor}}$ = Horizontal Row $\therefore$ $F_\alpha = 3.0178$.

Next, we must calculate F_{Data}, but first we must find $(s^2)_{\text{Factor}}$ and $(s^2)_{\text{Error}}$.

To find $\left(s^2\right)_{\text{Factor}}$ ···

$$\left(s^2\right)_{\text{Factor}} = \frac{\sum \left[n_i(\overline{x}_i - \overline{x})^2\right]}{\text{df}_{\text{Factor}}}, \text{ where } \overline{x} = \frac{\sum \left[n_i \overline{x}_i\right]}{n} = \frac{\sum \left[n_i \overline{x}_i\right]}{\sum n_i} \quad \therefore$$

$$\overline{x} = \frac{(92)(76) + (21)(79) + (10)(82)}{(92 + 21 + 10)} = 77 \quad \therefore$$

$$\left(s^2\right)_{\text{Factor}} = \frac{(92)(76-77)^2 + (21)(79-77)^2 + (10)(82-77)^2}{2} = \frac{92 + 84 + 250}{2} = 213$$

To find $\left(s^2\right)_{\text{Error}}$ ...

$$\left(s^2\right)_{\text{Error}} = \frac{\sum\left[(n_i - 1)\left(s^2\right)_i\right]}{\text{df}_{\text{Error}}} = \frac{(92-1)(3.4)^2 + (21-1)(4.5)^2 + (10-1)(6.1)^2}{120} = 14.93 \quad \therefore$$

$$F_{\text{Data}} = \frac{\left(s^2\right)_{\text{Factor}}}{\left(s^2\right)_{\text{Error}}} = \frac{213}{14.93} = 14.27$$

Since $F_{\text{Data}} > F_{\alpha}$ $\quad\therefore\quad$ we can reject the null hypothesis (Yes).

8. The number of wins college football teams (FW) and college basketball teams (BW) had last season is given for 8 teams picked at random.

	Wisconsin	**Texas**	**Missouri**	**Colorado St.**
x = FW	13	7	7	7
y = BW	15	19	20	11

	Arkansas St.	**Southern Miss.**	**Florida Int.**	**Idaho**
x = FW	7	8	8	4
y = BW	11	16	14	22

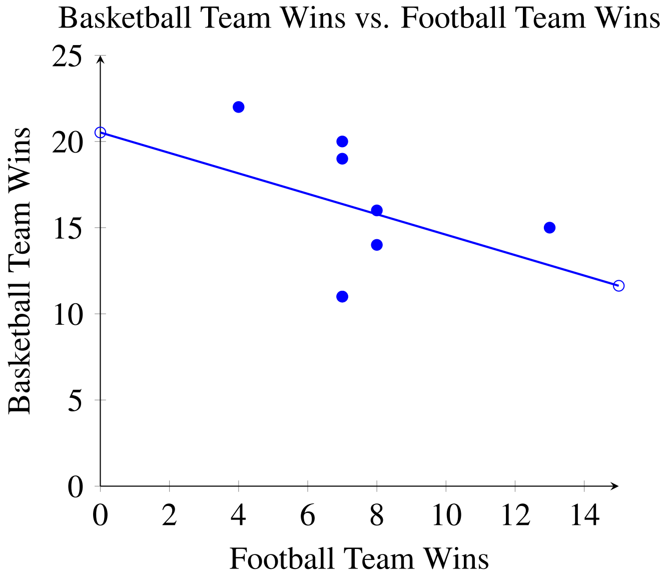
$$\sum x = 61 \qquad \sum y = 128 \qquad \sum\left(x^2\right) = 509 \qquad \sum\left(y^2\right) = 2164 \qquad \sum(xy) = 950$$

$$\sum\left(x^2\right) - \frac{\left(\sum x\right)^2}{n} = 43.875 \qquad \sum\left(y^2\right) - \frac{\left(\sum y\right)^2}{n} = 116 \qquad \sum(xy) - \frac{\left(\sum x\right)\left(\sum y\right)}{n} = -26$$

$$\sqrt{\frac{\sum\left(y^2\right) - b\sum y - m\sum(xy)}{n-2}} = 4.095 \qquad \sqrt{1 + \frac{1}{8} + \frac{(1-7.625)^2}{\sum\left(x^2\right) - \frac{\left(\sum x\right)^2}{n}}} = 1.458$$

$$\sqrt{\frac{1}{8} + \frac{(1-7.625)^2}{\sum\left(x^2\right) - \frac{\left(\sum x\right)^2}{n}}} = 1.061$$

A) (4 pts) Give a scatterplot.



B) (5 pts) Find r, the linear correlation coefficient.

$$r = \frac{\sum(xy) - \frac{\left(\sum x\right)\left(\sum y\right)}{n}}{\sqrt{\left(\sum(x^2) - \frac{\left(\sum x\right)^2}{n}\right)\left(\sum(y^2) - \frac{\left(\sum y\right)^2}{n}\right)}} = \frac{-26}{\sqrt{(43.875)(116)}} = -0.364$$

C) (6 pts) Find the line of best fit and graph it on the scatterplot, by plotting two points.

$$y = mx + b \quad \therefore \quad m = \frac{\sum(xy) - \frac{\left(\sum x\right)\left(\sum y\right)}{n}}{\sum(x^2) - \frac{\left(\sum x\right)^2}{n}} = \frac{-26}{43.875} = -0.593$$

$$b = \frac{\sum y - m \sum x}{n} = \frac{128 - (-0.593)(61)}{8} = 20.522 \quad \therefore \quad y = -0.593x + 20.522$$

This equation represents the BLUE line on the scatterplot.

D) (4 pts) Use the line of best fit to estimate the BW (basketball wins) for a university that had 1 FW (football wins).

$y = -0.593(1\text{ FW}) + 20.522 =$ 19.929 BW

E) (3 pts) What do you think of the answer in D), explain.

I think the answer in part D) is unrealistic due to the weak correlation and extrapolation.

F) (4 pts) Interpret the slope in everyday terms.

The slope tells us how many basketball team wins are lost for every football team win.

In this case, there are −0.593 basketball wins lost for every football team win.

G) (4 pts) What percent of the differences in BW can be explained by the regression line on FW?

$r^2 = (-0.364)^2 = 0.132$

Therefore, 13.2% of the differences can be explained by the regression line on FW.

H) (6 pts) Give a 95% CI for the BW for a university that had 1 FW.

$$\text{CI} = y_0 \pm \left(t_{\alpha/2}\right) \sqrt{\frac{\sum\left(y^2\right) - b\sum y - m\sum(xy)}{n-2}} \sqrt{1 + \frac{1}{n} + \frac{\left(x_0 - \overline{x}\right)^2}{\sum\left(x^2\right) - \frac{\left(\sum x\right)^2}{n}}}$$

When $x_0 = 1$ FW $\quad y_0 = -0.593(1 \text{ FW}) + 20.522 = 19.929$ BW $\quad \therefore$

95% CI $= 19.929 \pm (2.447)(1.458)(4.095) =$ 19.929 ± 14.610

I) (6 pts) Is there good evidence at the 10% significance level that ρ, the population linear correlation coefficient is not 0? Make sure to give the critical value(s) from the table and the test statistic from the data as well as the Yes or No answer.

$H_0 : \rho = 0$

$H_a : \rho \neq 0$

Given that $\alpha = 0.10$, $\text{df} = n - 2 = 8 - 2 = 6$, and realizing that we're carrying out a two-tailed test, $t_{\alpha/2} = 1.943$

$$t_{\text{Data}} = \sqrt{\frac{n-2}{1-r^2}} = \sqrt{\frac{6}{1-(-0.364)^2}} = 2.630$$

Since $t_{\text{Data}} > t_{\alpha/2}$, we can reject the null hypothesis (Yes).

The rest are each worth 1 pt.

9. What does the least squares line minimize?

   The least squares line minimizes the sum of the squares of the vertical distances from the points to the line.

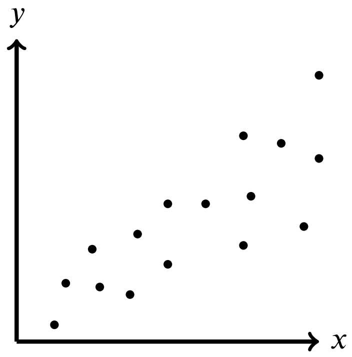
10. If there is a negative relationship, then r will be negative in part because bigger than average x's will correspond to smaller than average y's making $\left(\frac{x-\bar{x}}{s_x}\right)\left(\frac{y-\bar{y}}{s_y}\right)$ the product of a positive and a negative, which is negative.

11. Name three other relationships besides linear. Such as exponential, logistic, & quadratic.

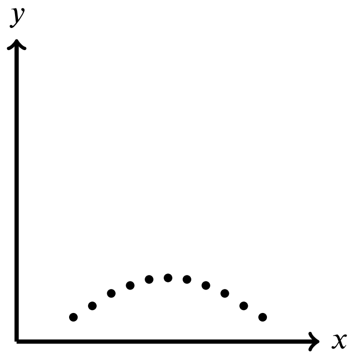
12. Is r sensitive to outliers? Yes.

13. Which scatter plot shows a stronger relationship? Scatter Plot B)

A)



B)



14. Give an example in which there is a strong association between x and y, but there is no cause and effect.

    Ice cream sales and sunburns.

15. Do you think that people with an agenda will still try to show x affects y even if the setting is too complex with many variables interacting?

    Yes.

16. There is a strong correlation between education and wealth. Give a possible lurking variable that could explain this without having education have a cause and effect on wealth.

    A lurking variable in this case would be how motivated a person is.

17. There is a strong correlation between education and wealth. Give a possible lurking variable that could explain this without having education have a cause and effect on wealth.

    A lurking variable in this case would be how motivated a person is.

18. If a person is motivated they are likely to become wealthy and also educated. Do you think that motivation explains part of the association between education and wealth, so in fact there is no cause and effect?

    No.

19. If a person is motivated they are likely to become wealthy and also educated. Do you think that motivation explains part of the association between education and wealth, so in fact the cause and effect still exists, but it's not as strong as many people might think?

    Yes.

20. Give an example of Simpson's Paradox.

    Helicopters versus the road for accident victims, helicopters do better in serious and less serious cases, but when combined the road did better.

21. The 4 assumptions for ANOVA are normal populations, equal variance, independent samples and simple random samples (SRS's).

22. The 4 assumptions for ANOVA are normal populations, equal variance, independent samples and simple random samples (SRS's).

23. If you do a good job of collecting data from different sources, the data will vary for only two reasons, those are source and experimentation.

24. Variance due to factor is a weighted variance of the sample means.

25. Variance due to error is a weighted mean of the sample variances.