1. Please state the Central Limit Theorem as discussed in class.

Select a sample of size n from a population with unknown distribution mean μ and std. dev. σ. Then the sample mean \( \bar{x} \) is normal with mean \( μ \) and std. dev. \( \frac{σ}{\sqrt{n}} \).

2. Please state, in your own words, what the following terms mean

- Contingency Table
- Chi-Square Test
- Least-Square Regression
- Confidence Interval
- p-value of a Chi-square Test
- Correlation Coefficient r
- Scatter Plot
- Least Square Regression line
- P(event) = 0
- P(z < 1), where z has a standard normal distribution
- Degree of freedom df for a t-distribution
- When to use a t distribution and when to use a normal distribution
- P(x > 55) if x is N(50, 12)
- Standard error

3. Please decide if the following statements are true or false.

- If the p-value of a Chi-Square test is close to one, the association between two variables is strong.
- Both the p-value of a Chi-Square test and the correlation coefficient r tell you whether two variables are related, but the correlation coefficient r carries even more information.
- A Chi-Square test is appropriate for categorical variables, a regression analysis is appropriate for two numeric variables.
- The expected value in a cell of a contingency table tells you how many items would fall in that cell if the two variables were independent of one another.
- If an expected value in any cell of a contingency table is less than 5, then the two variables are dependent.
- Suppose you compute the equation of a least-square regression line as \( y = -2x + 3 \) and the correlation coefficient r = 0.8, could that be possible?
- If r = 0.8, it means that two variables are strongly related in such a way that as x gets larger, the corresponding y gets smaller.
- P(z < 2) = 0.02211
- If \( X \) is N(10, 2) and \( X = 11 \), then the corresponding z-value is 2.1.
- If \( X \) is N(95, 10) then P(X > 105) = 0.25
- A 95% confidence interval means that you are 95% certain that the true population mean is contained in the computed interval.
- A 99% confidence interval is smaller than a 90% confidence interval.
- If a sample of size 100 has mean 123 and std. dev. 12, then the standard error is 1.2

4. Compute the following probabilities:
• In tossing one coin twice, find P(HH) or P(exactly one head) or P(no head) or P(at least one head)

\[ P(HH) = \frac{1}{4}, \quad P(\text{exactly one head}) = \frac{1}{2}, \quad P(\text{no head}) = \frac{1}{4}, \quad P(\text{at least one head}) = \frac{3}{4} \]

• In throwing two dice, find P(sum is 4) or P(sum = 1) or P(sum is 4 or more)

\[
P(\text{sum is 4}) = \frac{3}{36}\]
\[
P(\text{sum = 1}) = 0\]
\[
P(\text{sum is 4 or more}) = \frac{10}{36} = \frac{5}{18}\]

• In drawing one card randomly from a standard 52-card deck, find P(card is Ace)

\[
P(\text{card is Ace}) = \frac{4}{52} = \frac{1}{13}\]

5. A (hypothetical) frequency distribution for the age of people in a survey, the categories have the following probabilities:

<table>
<thead>
<tr>
<th>Category</th>
<th>Probability</th>
</tr>
</thead>
<tbody>
<tr>
<td>0 – 18</td>
<td>0.15</td>
</tr>
<tr>
<td>19-40</td>
<td>0.25</td>
</tr>
<tr>
<td>41-65</td>
<td>0.15</td>
</tr>
<tr>
<td>65 and older</td>
<td>0.3</td>
</tr>
</tbody>
</table>

One number is missing – what is that number? 
0.3

What is the chance that a randomly selected person is 40 years or younger?

\[
P(\text{\leq 40 years}) = 0.25 + 0.15 + 0.15 = 0.55
\]

6. Please consider the following data:

- Father’s schooling (years): 8, 11, 5, 12
- Respondent’s schooling (years): 8, 12, 5, 15

a) find the mean for both variables

\[
\bar{x} = \frac{8 + 11 + 5 + 12}{4} = 8.75
\]
\[
\bar{y} = \frac{8 + 12 + 5 + 15}{4} = 10
\]

\[
y = 1.367x - 2.3
\]

Recall the corresponding formulas (or use StatCrunch)
\[
\begin{align*}
\rho &= \frac{S_{xy}}{S_{xx} S_{yy}} \\
\text{slope} &= \frac{S_{xy}}{S_{xx}} \\
\text{y - intercept} &= \bar{Y} - \text{slope} \cdot \bar{X}
\end{align*}
\]

7. The following scores were obtained as part of a sample with mean 10 and standard deviation 2. For each score, find the appropriate z-score: \(X = 10, X = 14, X = 6, X = -1\). Then, for each z-score found, find the probabilities of obtaining a score less than or equal to the computed z-score. Note: in mathematical notation this means that we want to find \(P(z \leq z_0)\), where \(z_0\) is the computed z-score.

\[
\begin{align*}
&P(10 - 10 = 0) = 0.5 \\
&P(14 - 10 = 2) = 0.9772 \\
&P(6 - 10 = -2) = 0.0228 \\
&P(-1 - 10 = -6) = 0
\end{align*}
\]

8. Each score listed below comes from a sample with the indicated mean and standard deviation. Find the indicated probability (in percent).

- \(X\) is normal with mean 3, standard deviation 1.5, find \(P(x \leq 6)\) \(= 0.9772\)
- \(X\) is normal with mean 3, standard deviation 3, find \(P(x \geq 9)\) \(= 0.0224\)
- \(X\) is normal with mean 0, standard deviation 2, find \(P(1 < x < 2)\) \(= P(x \leq 2) - P(x \leq 1)\) \(= 0.7413 - 0.0539\)
- \(X\) is normal with mean 3, standard deviation 1, find \(P(x \geq 2)\) \(= 0.8413\)

9. Consider the following sample data, selected at random from some population:

\[12, 16, 5, 19\]

a) What is your best guess for the unknown population mean? The sample mean \(\frac{12 + 16 + 5 + 19}{4} = \frac{52}{4} = 13\)

b) Find the standard error for the sample mean.

\[
\text{std error} = \frac{\sqrt{\frac{52}{4}}}{\sqrt{4}} = 3.025
\]

c) Do you need to use the t distribution to find the multiplier? \(\text{Yes, because it is a small sample (n < 30)}\)
d) Find a 95% confidence interval for the unknown population mean. 

\[ \bar{x} = 3.12, \quad \text{SE} = 1.8, \quad 95\% CI: 13 \pm 9.60 \]

\[ df = n-1 = 3 \]

e) Find a 99% confidence interval for the population mean. Explain why this interval differs from the previous one.

\[ \bar{x} = 5.14, \quad \text{SE} = 5.84, \quad 99\% CI: 13 \pm 7.60 \]

f) If you were to compute a 90% confidence interval, would it be wider or narrower than the previous two?

10. Find the following probabilities, assuming \( z \) is \( N(0, 1) \).

\[ P(z < 1.1) = 0.8643 \]

\[ P(z > -1.2) = 0.0949 \]

\[ P(z > 1.3) = 0.0906 \]

\[ P(z < -1.6) = 0.0543 \]

\[ P(1.2 < z < 2.1) = P(z < 2.1) - P(z < 1.2) = 0.9841 - 0.9525 = 0.0316 \]

\[ P(-2.1 < z < -1.2) = P(z < -1.2) - P(z < -2.1) = 0.0150 - 0.0414 = 0.0264 \]

\[ P(-1.2 < z < 2.1) = P(z < 2.1) - P(z < -1.2) = 0.9841 - 0.0543 = 0.9305 \]

11. When using StatCrunch to draw a scatter plot, it comes up with the following picture:

a) Draw a “best-fit” line through this data.

b) Use the line to estimate the y-intercept and slope of the equation of the least-square regression line

\[ y - \bar{y} = b(x - \bar{x}), \quad \text{slope} = \frac{\bar{y} - \bar{y}'}{x - \bar{x}}, \quad \approx y = (x + 210 \text{ approx.}) \]

c) Look at the data and your line and estimate whether \( r \) would be close to -1, close to 0, or close to 1

\[ r \approx \text{very close to -1} \]

12. Please consider the following results on a quiz, measuring scores before and after a certain lecture.

**Before lecture:** 5, 6, 7, 9, 3

**After lecture:** 6, 7, 9, 9, 4

- Create a scatter plot representing this data, including a best-fit line for the data.
• Find the exact equation of the least-square regression line

\[ y = 0.9x + 1.6 \]  

• Compute the correlation coefficient \( r \)

\[ r = 0.941 \]

• Predict the “after lecture” score for a “before lecture” score of 8.

\[ y = 0.9(8) + 1.6 = 8.5 \]

13. When using StatCrunch for a linear regression analysis of pre-test versus post-test scores, it computes the output:

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Estimate</th>
<th>Std. Err.</th>
<th>Alternative</th>
<th>DF</th>
<th>T-Stat</th>
<th>P-Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>2.5</td>
<td>10.712480</td>
<td>= 0</td>
<td>3</td>
<td>0.23338082</td>
<td>0.0305</td>
</tr>
<tr>
<td>Slope</td>
<td>0.95454244</td>
<td>0.12605364</td>
<td>≠ 0</td>
<td>2</td>
<td>7</td>
<td>0.008</td>
</tr>
</tbody>
</table>

a) Find the equation of the least-square regression line

\[ y = 0.9545x + 2.5 \]

b) What is the correlation coefficient, and what does it mean

\[ r = 0.97 \text{ is close to 1} \]

c) Predict the post-test score of someone with a pretest score of 77.

\[ y = 0.9545(77) + 2.5 = 75.99 \]

d) Do you think your prediction is accurate? Justify your answer using the correlation coefficient

\[ r \text{ in close to 1, so I think it's an accurate prediction} \]

14. The table below shows a contingency table for the variables “DEGREE” by “RACE”. Each cell lists three numbers: the count, the row, and the column percentage.
15. Consider the contingency table for religious preference versus political opinion, using our GSS survey below.

a) Compute the row percentage in the "Liberal and Catholic" cell, as well as column percentage and expected value.

b) Using StatCrunch, we conducted a Chi-Square test with the output as follows:

<table>
<thead>
<tr>
<th>Statistic</th>
<th>Value</th>
<th>P-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Chi-square</td>
<td>72.15</td>
<td>0.0001</td>
</tr>
</tbody>
</table>

What is your conclusion?

\[ P < 0.05 \text{ \rightarrow \text{variables are related!} } \]

c) Why should you compute and double-check all expected values in that table before finalizing your conclusion?

If an expected value is less than 5, test is not valid.
16. To investigate whether a relation exists between affiliation with a particular political party and the opinion on gun permits we used StatCrunch to create the following contingency table.

<table>
<thead>
<tr>
<th></th>
<th>FAVOR</th>
<th>OPPOSE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Democrat</td>
<td>35.7%</td>
<td>23.4%</td>
</tr>
<tr>
<td>Independent</td>
<td>36.5%</td>
<td>100.0%</td>
</tr>
<tr>
<td>Republican</td>
<td>26.5%</td>
<td>34.7%</td>
</tr>
<tr>
<td>Other</td>
<td>1.3%</td>
<td>2.4%</td>
</tr>
<tr>
<td>Total</td>
<td>100.0%</td>
<td>100.0%</td>
</tr>
</tbody>
</table>

a) Based on the table, do you think there is strong evidence that the two variables associated, using common sense?

   *No, everything seems evenly distributed*

b) Based on your analysis in part (a), what do you think might be the p-value of a Chi-Square test for this data?

   \[ p > 0.05 \]

17. Use the GSS survey data to find the average number of siblings for people in the US in 1996 with reasonable accuracy.

Note: Using StatCrunch we found that the descriptive statistics for the variable ‘sibs’ is as follows:

- N: 2897
- Mean: 3.86
- Standard Deviation: 3.52

\[
\begin{align*}
\bar{x} & = \frac{\sum x}{n} = \frac{0.096}{0.096} = 1.05 \\
\sigma & = \sqrt{\frac{\sum (x-\bar{x})^2}{n-1}} = \sqrt{0.1291} \\
\end{align*}
\]

\[
2.38 \pm 0.1291 = 95\% \text{ confidence interval}
\]

18. When using StatCrunch to draw a “scatter plot, it comes up with the following picture:

   ![Scatter plot image]

a) Draw a “best-fit” line through this data.

b) Use the line to estimate the y-intercept and slope of the equation of the least-square regression line

\[
y = \frac{48-40}{40-20} x - \bar{y} = 1.9x - 3.9
\]

c) Look at the data and your line and estimate whether r would be close to -1, close to 0, or close to 1.

   \[ r \approx 0.8 \text{ close to 1} \]

19. Suppose you were asked to compute a 95% confidence interval. The resulting interval, however, turned out to be too large to be of use to your client. What could you do to achieve a smaller confidence interval?

   - Redo survey with a larger sample, or compute 90% - conf. interval, which will be smaller.
20. The lifetimes (in years) of ten automobile batteries of a particular brand are:

\[
\begin{array}{cccccc}
2.4 & 1.9 & 2.0 & 2.1 & 1.8 \\
2.3 & 2.1 & 2.3 & 1.7 & 2.0 \\
\end{array}
\]

Estimate the mean lifetime for all batteries, using a 95% confidence interval.

\[
\text{Sample mean: 2.06} \\
\text{Standard error: 0.211} \\
\text{DF = 9} \\
\text{From } 1.995 \text{ to } 2.222
\]

21. A test was conducted to determine the length of time required for a student to read a specified amount of material. All students were instructed to read at the maximum speed at which they could still comprehend the material. Sixteen students took the test, with the following results (in minutes):

25, 18, 27, 29, 20, 19, 25, 24, 32, 21, 24, 19, 23, 28, 31, 22

Estimate the mean length of time required for all students to read the material, using a 95% confidence interval.

\[
\text{Sample mean: 24.191} \\
\text{Std. error: 1.0809} \\
\text{From } 21.91 \text{ to } 26.49
\]

22. The caffeine content of a random sample of 90 cups of black coffee dispensed by a new machine is measured. The mean and standard deviation for the sample are 110 mg and 6.1 mg, respectively.

a) Compute a 90% confidence interval for the true population mean caffeine content per cup dispensed by the machine.

\[
X \pm 1.645 \cdot \frac{S}{\sqrt{N}} = 110 \pm 1.059
\]

b) If you would compute a 99% confidence interval for the true population, would it be wider or narrower than the 90% confidence interval? (You do not actually have to compute this interval to answer the question).

wider

c) Another person selected a random sample of 900 instead of 90 cups, and the mean and standard deviation of this larger sample turned out to be 110 mg and 6.1 mg as well. That person uses her data to compute a 90% confidence interval. Would the 90% confidence interval for the larger sample size be wider or narrower than the 90% confidence interval for the smaller sample size?

90% - in general, the larger sample would be narrower (smaller)