1. (20 points) A case-control study was designed to examine risk factors for cervical dysplasia (Becker et al., 1994); the women in the study were patients were aged 18 to 40. There were \( n_1 = 175 \) cases with cervical dysplasia and \( n_2 = 308 \) controls without. Each woman was tested for human papilloma virus (HPV) and either positive (HPV+) or negative (HPV-):

<table>
<thead>
<tr>
<th></th>
<th>dysplasia</th>
<th>no dysplasia</th>
</tr>
</thead>
<tbody>
<tr>
<td>HPV+</td>
<td>164</td>
<td>130</td>
</tr>
<tr>
<td>HPV−</td>
<td>11</td>
<td>178</td>
</tr>
<tr>
<td>total</td>
<td>175</td>
<td>308</td>
</tr>
</tbody>
</table>

Let \( p_1 \) be the probability of HPV+ among the cases and \( p_2 \) be the probability of HPV+ among the controls.

(a) Estimate the difference \( p_1 - p_2 \), the relative risk \( p_1/p_2 \) and the odds ratio \( \theta = p_1/(1-p_1) \)

\[ \hat{p}_1 = \frac{164}{175} = 0.937, \quad \hat{p}_2 = \frac{130}{308} = 0.422, \]
\[ \hat{p}_1 - \hat{p}_2 = 0.515, \quad \frac{\hat{p}_1}{\hat{p}_2} = 2.22, \quad \hat{\theta} = \frac{\hat{p}_1/(1-\hat{p}_1)}{\hat{p}_2/(1-\hat{p}_2)} = 20.4. \]

The probability of being HPV+ increases by one half when dysplasia is present. One is twice as likely to have HPV when dysplasia is present. The odds of being HPV+ increase by a factor of 20 when dysplasia is present.

(b) Find a 95% CI for the difference \( p_1 - p_2 \), and test \( H_0 : p_1 = p_2 \) at the 5% level.

\[ \hat{p}_1 = \frac{164 + 1}{175 + 2} = 0.932, \quad \hat{p}_2 = \frac{130 + 1}{308 + 2} = 0.423, \]
\[ SE_{\hat{p}_1 - \hat{p}_2} = \sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1 + 2} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2 + 2}} = 0.0338. \]
\[ \hat{p}_1 - \hat{p}_2 \pm 1.96SE_{\hat{p}_1 - \hat{p}_2} = (0.443, 0.576). \]

We reject that \( H_0 : p_1 = p_2 \) at the 5% level because the 95% confidence interval does not include zero.

(c) Obtain the 95% CI for the odds ratio \( \theta \) and interpret it in the context of these data. Do you reject \( H_0 : \theta = 1 \) at the 5% level? Why or why not? Is dysplasia associated with HPV?

\[ \log \hat{\theta} = \log \frac{164 \times 178}{130 \times 11} = \log 20.4 = 3.02. \]
\[ SE_{\log \theta} = \sqrt{\frac{1}{y_1} + \frac{1}{n_1 - y_1} + \frac{1}{y_2} + \frac{1}{n_2 - y_2}} \]
\[ = \sqrt{\frac{1}{164} + \frac{1}{11} + \frac{1}{130} + \frac{1}{178}} \]
\[ = 0.332. \]

The 95% confidence interval for \( \log \theta \) is given by

\[ 3.02 \pm 1.96 \times 0.332 = (2.37, 3.67). \]

The 95% confidence interval for \( \theta \) is given by

\[ (e^{2.37}, e^{3.67}) = (10.6, 39.1). \]

The odds of dysplasia increase by 10 to 40 times for HPV+ versus HPV−. Also, the odds of HPV+ increase 10 to 40 times for those with dysplasia versus those without. We reject \( H_0 : \theta = 1 \) at the 5% level because the 95% confidence interval does not include one. Therefore there is a significant association between dysplasia and HPV.
2. (20 points) Dental measurements were made on \( n_1 = 11 \) girls and \( n_2 = 16 \) boys, all 16 years old. Each measurement is the distance (mm) from the center of the pituitary to the pterygomaxillary fissure. We want to test whether the distributions of these measurements are the same for boys and girls. Here are the sorted data:

<table>
<thead>
<tr>
<th>Girls</th>
<th>Boys</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>19.5</td>
</tr>
<tr>
<td>0</td>
<td>21.5</td>
</tr>
<tr>
<td>0</td>
<td>22.5</td>
</tr>
<tr>
<td>0</td>
<td>23.0</td>
</tr>
<tr>
<td>0</td>
<td>23.5</td>
</tr>
<tr>
<td>0</td>
<td>24.0</td>
</tr>
<tr>
<td>1</td>
<td>25.0</td>
</tr>
<tr>
<td>2.5</td>
<td>25.5</td>
</tr>
<tr>
<td>4.5</td>
<td>26.0</td>
</tr>
<tr>
<td>7</td>
<td>26.5</td>
</tr>
<tr>
<td>10.5</td>
<td>28.0</td>
</tr>
<tr>
<td></td>
<td>28.5</td>
</tr>
<tr>
<td></td>
<td>29.5</td>
</tr>
<tr>
<td></td>
<td>30.0</td>
</tr>
<tr>
<td></td>
<td>31.0</td>
</tr>
<tr>
<td></td>
<td>31.5</td>
</tr>
<tr>
<td>25.5</td>
<td>150.5</td>
</tr>
</tbody>
</table>

(a) Compute the Wilcoxon-Mann-Whitney counts \( K_1 \) and \( K_2 \), and the test statistic \( U_s \).

Answer: Remember to add 0.5 for tied values. \( K_1 = 25.5 \) and \( K_2 = 150.5 \). \( U_s = 150.5 \), the bigger of the two.

(b) Check that \( K_1 + K_2 = n_1 \times n_2 \).

Answer:

\[
K_1 + K_2 = 25.5 + 150.5 = 176 = 11 \times 16 = n_1 \times n_2.
\]

(c) Compute the p–value for the test, perhaps by bounding using Table 6 (pp. 680–684). Use a two-sided alternative.

Answer:

\[
0.001 < \text{p-value} < 0.002,
\]

from Table 6 page 682.

(d) Do you reject that the distributions are the same at the 5% significance level? Why or why not?

Answer: Yes, we reject that the distributions are the same at the 5% significance level because the p-value is less than 0.05.
3. (20 points) A criminologist studying the relationship between level of education and crime rate in medium-sized U.S. counties collected data from a random sample of \( n = 84 \) counties. \( Y \) is the crime rate (crimes per 100 people) and \( X \) is the percentage of individuals in the county having at least a high-school diploma. Here’s a scatterplot:

![Scatterplot](image)

For these data,

\[
\begin{align*}
\bar{y} &= 7.111 \\
\bar{x} &= 78.60 \\
\sum (x_i - \bar{x})(y_i - \bar{y}) &= -547.9 \\
\sum (x_i - \bar{x})^2 &= 3212 \\
SS(resid) &= 455.3
\end{align*}
\]

(a) Compute the least squares estimates \( b_0 \) and \( b_1 \) for regressing crime rate on percentage of high-school graduates.

**Answer:**

\[
b_1 = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2} = \frac{-547.9}{3212} = -0.171.
\]

\[
b_0 = \bar{y} - b_1\bar{x} = 7.111 - (-0.171)78.60 = 20.5.
\]

(b) Draw the regression line onto the scatterplot; describe the relationship.

**Answer:** Roughly linear, crime rates tend to decrease with more education.

(c) For Richland County, \( X = 85.2\% \). Predict Richland’s crime rate.

**Answer:**

\[
20.5 - 0.171(85.2) = 5.9\%.
\]
(d) Compute a 95% confidence interval for $\beta_1$.

Answer:

$$SE_{b_1} = \frac{s_{y|x}}{\sqrt{\sum(x_i - \bar{x})^2}}$$

$$= \frac{\sqrt{SS(resid)}/(n-2)}{\sqrt{\sum(x_i - \bar{x})^2}}$$

$$= \frac{\sqrt{455.3}/(84-2)}{\sqrt{3212}} = 0.0416.$$  

$$b_1 \pm t_{0.975}SE_{b_1} = -0.171 \pm 1.99 \times 0.0416 = (-0.25, -0.088).$$

Note that $df = 84 - 2 = 82$ for the $t$ distribution; I rounded down to $df = 80$ to use the table in the back of the book.

(e) Do you reject $H_0: \beta_1 = 0$ at the 5% significance level? Why or why not? What does this tell you about the relationship between crime and education?

Answer: We reject $H_0: \beta_1 = 0$ at the 5% significance level because a 95% confidence interval for $\beta_1$ does not include zero. There is a significant, linear association between crime rate and graduating high school.
4. (20 points) For the dental data in problem 2, assume that a $t$-test is appropriate and that we want to test $H_0 : \mu_1 = \mu_2$ versus $H_A : \mu_1 < \mu_2$ where $\mu_1$ is the mean for girls and $\mu_2$ is the mean for boys. Formula (7.1) gives $df = 19.3$.

<table>
<thead>
<tr>
<th></th>
<th>Girls</th>
<th>Boys</th>
</tr>
</thead>
<tbody>
<tr>
<td>$n$</td>
<td>11</td>
<td>16</td>
</tr>
<tr>
<td>$\bar{y}$</td>
<td>24.09</td>
<td>27.47</td>
</tr>
<tr>
<td>$s$</td>
<td>2.44</td>
<td>2.09</td>
</tr>
</tbody>
</table>

(a) Compute $SE_{\bar{y}_1 - \bar{y}_2}$.

Answer:

$$SE_{\bar{y}_1 - \bar{y}_2} = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} = \sqrt{\frac{2.44^2}{11} + \frac{2.09^2}{16}} = 0.902.$$  

(b) Find a 95% confidence interval for $\mu_1 - \mu_2$.

Answer:

$$\bar{y}_1 - \bar{y}_2 \pm t_{0.975}SE_{\bar{y}_1 - \bar{y}_2} = 24.09 - 27.47 \pm 2.093(0.902) = (-5.27, -1.49).$$  

I used $df = 19$ in the $t$ table in the back of the book.

(c) Compute the test statistic for testing $H_0 : \mu_1 = \mu_2$.

Answer:

$$t_s = \frac{\bar{y}_1 - \bar{y}_2}{SE_{\bar{y}_1 - \bar{y}_2}} = \frac{24.09 - 27.47}{0.902} = -3.75.$$  

(d) Find the p–value for the alternative $H_A : \mu_1 < \mu_2$.

Answer: Using the $df = 19$ row in the back of the book,

$$0.0005 < p\text{-value} < 0.005.$$  

(e) Do you reject $H_0 : \mu_1 = \mu_2$ in favor of $H_A : \mu_1 < \mu_2$ at the 5% level? Why or why not?

Answer: We reject $H_0 : \mu_1 = \mu_2$ in favor of $H_A : \mu_1 < \mu_2$ at the 5% significance level because the p-value is less than 0.05.
5. TRUE or FALSE. Each question is worth 2 points.

(a) TRUE or [FALSE]: A 95% confidence interval for \( \mu_1 - \mu_2 \) contains 0 means we reject \( H_0: \mu_1 = \mu_2 \) at the 5% level.

(b) TRUE or [FALSE]: The population median is always within one standard deviation from the population mean.

(c) [TRUE] or FALSE: For rare events, the odds ratio is approximately equal to the relative risk.

(d) [TRUE] or FALSE: p-values are smaller for one-sided alternatives compared to two-sided alternatives.

(e) TRUE or [FALSE]: The central limit theorem tells us that sample means will be approximately distributed binomial as the sample size increases.

(f) TRUE or [FALSE]: The sample proportion \( \hat{p} \) always equals the true proportion \( p \) under random sampling.

(g) [TRUE] or FALSE: In a box-plot, the box contains 50% of the sample values.

(h) TRUE or [FALSE]: A and B are independent if and only if \( \Pr(A \text{ and } B) = \Pr(A) + \Pr(B) \).

(i) TRUE or [FALSE]: Alice rejects \( H_0 \). The only kind of error Alice could have made is a Type II error.

(j) [TRUE] or FALSE: The least squares regression line minimizes the residual sum of squares.

(k) [TRUE] or FALSE: (Extra credit) The t test has its origins in the Guinness brewery.

6. (Extra credit): A study was performed to determine the effect of a carcinogen on the survival of rats. Thirty rats were injected with varying levels of a carcinogen \( x \) (mg). For rats who survived one week, \( Y_i = 1 \) was recorded; for rats who died within one week, \( Y_i = 0 \) was recorded. A logistic regression was fit, and computer output is shown below:

<table>
<thead>
<tr>
<th>Parameter</th>
<th>DF</th>
<th>Estimate</th>
<th>Standard Error</th>
<th>Chi-Square</th>
<th>Pr &gt; ChiSq</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>1</td>
<td>6.6656</td>
<td>2.3119</td>
<td>8.3129</td>
<td>0.0039</td>
</tr>
<tr>
<td>x</td>
<td>1</td>
<td>-0.2506</td>
<td>0.0826</td>
<td>9.2070</td>
<td>0.0024</td>
</tr>
</tbody>
</table>

(a) (2 points) Does the carcinogen increase or decrease the odds of survival? Is the effect significant? Explain.

Answer: The regression coefficient is negative, so increasing the carcinogen decreases the odds of survival. This makes intuitive sense. The effect is significant, we reject \( H_0: \beta_1 = 0 \) at the 5% level because 0.0024 < 0.05.

(b) (3 points) Find a 95% confidence interval for how the odds of survival changes when the amount of carcinogen is increased by one milligram.

Answer: A 95% confidence interval for the log odds ratio is

\[
 b_1 \pm 1.96 \times SE_{b_1} = -0.2506 \pm 1.96 \times 0.0826 = (-0.412, -0.089).
\]

Exponentiating gives the 95% confidence interval for how the odds change when increasing the carcinogen by 1 mg:

\[
(e^{-0.412}, e^{-0.089}) = (0.66, 0.91).
\]