1. (30 p) A political science professor is interested in comparing the characteristics of students who do and do not vote in national elections. For a random sample of 114 students who claimed to have voted in the last presidential election, she found a mean grade point of average of 2.71 and a standard deviation of 0.64. For an independent random sample of 123 students who did not vote, the mean grade point of average was 2.79 and the standard deviation was 0.56. Test, at the 5% significance level the null hypothesis that the population means are equal, against a two-sided alternative that population means are different.

2. (35 p) A random sample of 12 college baseball players participated in a special weight-training program in an attempt to improve their batting averages. The program lasted for 20 weeks immediately prior to the start of the baseball season. The average number of hours per week ($X$) and the change in their batting averages ($Y$) from the preceding season as follows:

<table>
<thead>
<tr>
<th>$X$</th>
<th>8.0</th>
<th>20.0</th>
<th>5.4</th>
<th>12.4</th>
<th>9.2</th>
<th>15.0</th>
<th>6.0</th>
<th>8.0</th>
<th>18.0</th>
<th>25.0</th>
<th>10.0</th>
<th>5.0</th>
</tr>
</thead>
<tbody>
<tr>
<td>$Y$</td>
<td>10</td>
<td>100</td>
<td>-10</td>
<td>79</td>
<td>50</td>
<td>89</td>
<td>34</td>
<td>30</td>
<td>68</td>
<td>110</td>
<td>34</td>
<td>10</td>
</tr>
</tbody>
</table>

a) Prepare a scatter plot of the data  
b) Compute covariance and correlation coefficient  
c) Compute and interpret $b_1$ and $b_0$  
d) What the points in their batting averages would you expect to have if the weight-training were 21 hours?

3. (35 p) A public ‘Research Center for Nutrition Policy and Promotion’ uses the Healthy Eating Index to monitor diet quality of the population. Data collected on a random sample of individuals are given as contingency table according to the activity level and gender of the participants as follows:

<table>
<thead>
<tr>
<th>Activity Level</th>
<th>Gender</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Male</td>
</tr>
<tr>
<td>Sedentary</td>
<td>957</td>
</tr>
<tr>
<td>Active</td>
<td>340</td>
</tr>
<tr>
<td>Very active</td>
<td>842</td>
</tr>
</tbody>
</table>

Determine if there is an association between activity level and gender at the 5% significance level. 

Recall:

$$S_X = \sqrt{\frac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{n-1}}; \quad S_Y = \sqrt{\frac{\sum_{i=1}^{n}(Y_i - \bar{Y})^2}{n-1}}; \quad S_{XY} = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{n-1};$$

$$r_{XY} = \frac{S_{XY}}{S_X S_Y}; \quad b_1 = \frac{\text{Cov}(X,Y)}{S_X^2}; \quad b_0 = \bar{Y} - b_1 \bar{X}; \quad \bar{X} = \frac{\sum_{i=1}^{n}X_i}{n}; \quad \bar{Y} = \frac{\sum_{i=1}^{n}Y_i}{n};$$

$$\chi^2 = \sum_{i=1}^{r} \sum_{j=1}^{c} \frac{(O_{ij} - E_{ij})^2}{E_{ij}}; \quad E_{ij} = \frac{R_i C_j}{n}.$$
**STAT 204 STATISTICS II SUMMER SCHOOL 13-14 GRADUATION MAKEUP EXAM SOLUTIONS**

1. Let 1- students who vote; 2- students who do not vote
First identify the null and alternative hypothesis

\[ H_0: \mu_1 - \mu_2 = 0 \]
\[ H_A: \mu_1 - \mu_2 \neq 0 \]

Given: \( \bar{x}_1 = 2.71; \bar{x}_2 = 2.79; s_1 = 0.64; s_2 = 0.56 \) (\( \sigma_1 \) and \( \sigma_2 \) unknown)
\( n_1 = 114; n_2 = 123 \)

To test \( H_0: \mu_1 - \mu_2 = 0 \) against the alternative \( H_A: \mu_1 - \mu_2 \neq 0 \) the decision rule is to reject \( H_0 \) and accept \( H_A \) if

\[ t > t_{a/2} \text{ or } t < -t_{a/2} \text{ with } d.f = v = n_1 + n_2 - 2 = 114 + 123 - 2 = 235 \]
(two - tail test, From Appendix Guidelines Diagram E - 2)

\[ t = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} < -t_{a/2} \text{ or } > t_{a/2} \text{(Guidelines E - 4)} \]

\[ s_p = \sqrt{\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}} = \sqrt{\frac{(114 - 1)(0.64)^2 + (123 - 1)(0.56)^2}{114 + 123 - 2}} = 0.5998 \]

\[ t = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} = -1.0259 \]

\[ \alpha = 5\%, t_{a/2} = t_{235,0.025} \cong t_{\infty,0.025} = z_{0.025} = 1.96 \text{ (Read from Appendix Table 8)} \]

Compare the t-test statistic value to \( t_{\infty,0.025} \)

\[ \text{If } t < -t_{\infty,0.025}, \text{ Reject } H_0 \]

\[-1.0259 > -1.96 \]

\[ p - \text{value} = 2[1 - F(z = 1.03)] = 2[1 - 0.85] = 0.30 > \alpha \]

Do not reject \( H_0 \) at the 5% significance level: \( \mu_1 - \mu_2 = 0 \)

Conclusion: There is sufficient evidence that the two population means are equal.

2.a) Scatter plot- \( X \) vs. \( Y \)

---

**Scatterplot of Weight Training Hours vs. Batting Avg. Change**

![Scatterplot of Weight Training Hours vs. Batting Avg. Change](image-url)
b) Use tabulated method to compute covariance and correlation coefficient as follows:

<table>
<thead>
<tr>
<th>(i)</th>
<th>(X_i)</th>
<th>(Y_i)</th>
<th>((X_i - \bar{X}))</th>
<th>((X_i - \bar{X})^2)</th>
<th>((Y_i - \bar{Y}))</th>
<th>((Y_i - \bar{Y})^2)</th>
<th>((X_i - \bar{X})(Y_i - \bar{Y}))</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>8.0</td>
<td>10</td>
<td>-3.83</td>
<td>14.6689</td>
<td>-40.33</td>
<td>1626.5089</td>
<td>154.4639</td>
</tr>
<tr>
<td>2</td>
<td>20.0</td>
<td>100</td>
<td>8.17</td>
<td>66.7489</td>
<td>49.67</td>
<td>2467.1089</td>
<td>405.8039</td>
</tr>
<tr>
<td>3</td>
<td>5.4</td>
<td>-10</td>
<td>-6.43</td>
<td>41.3449</td>
<td>-60.33</td>
<td>3639.7089</td>
<td>387.9219</td>
</tr>
<tr>
<td>4</td>
<td>12.4</td>
<td>79</td>
<td>0.57</td>
<td>0.3249</td>
<td>28.67</td>
<td>821.9689</td>
<td>16.3419</td>
</tr>
<tr>
<td>5</td>
<td>9.2</td>
<td>50</td>
<td>-2.63</td>
<td>6.9169</td>
<td>-0.33</td>
<td>0.1089</td>
<td>0.8679</td>
</tr>
<tr>
<td>6</td>
<td>15.0</td>
<td>89</td>
<td>3.17</td>
<td>10.0489</td>
<td>38.67</td>
<td>1495.3689</td>
<td>122.5839</td>
</tr>
<tr>
<td>7</td>
<td>6.0</td>
<td>34</td>
<td>-5.83</td>
<td>33.9889</td>
<td>-16.33</td>
<td>266.6689</td>
<td>95.2039</td>
</tr>
<tr>
<td>8</td>
<td>8.0</td>
<td>30</td>
<td>-3.83</td>
<td>14.6689</td>
<td>-20.33</td>
<td>413.3089</td>
<td>77.8639</td>
</tr>
<tr>
<td>9</td>
<td>18.0</td>
<td>68</td>
<td>6.17</td>
<td>38.0689</td>
<td>17.67</td>
<td>312.2289</td>
<td>109.0239</td>
</tr>
<tr>
<td>10</td>
<td>25.0</td>
<td>110</td>
<td>13.17</td>
<td>173.4489</td>
<td>59.67</td>
<td>3560.5089</td>
<td>785.8539</td>
</tr>
<tr>
<td>11</td>
<td>10.0</td>
<td>34</td>
<td>-1.83</td>
<td>3.3489</td>
<td>-16.33</td>
<td>266.6689</td>
<td>29.8839</td>
</tr>
<tr>
<td>12</td>
<td>5.0</td>
<td>10</td>
<td>-6.83</td>
<td>46.6489</td>
<td>-40.33</td>
<td>1626.5089</td>
<td>275.4539</td>
</tr>
</tbody>
</table>

\[
\bar{X} = \frac{\sum_{i=1}^{n} X_i}{n} = \frac{142}{12} = 11.83; \bar{Y} = \frac{\sum_{i=1}^{n} Y_i}{n} = \frac{604}{12} = 50.33
\]

\[
S_{XY} = \text{Cov}(X, Y) = \frac{\sum_{i=1}^{n} (X_i - \bar{X})(Y_i - \bar{Y})}{n - 1} = \frac{2461.2668}{12 - 1} = 223.75153
\]

\[
S_X = \sqrt{\frac{\sum_{i=1}^{n} (X_i - \bar{X})^2}{n - 1}} = \sqrt{\frac{450.2268}{12 - 1}} = 6.397633; S_X^2 = 40.9297
\]

\[
S_Y = \sqrt{\frac{\sum_{i=1}^{n} (Y_i - \bar{Y})^2}{n - 1}} = \sqrt{\frac{16496.67}{12 - 1}} = 38.72593; S_Y^2 = 1499.697
\]

\[
r_{XY} = \frac{S_{XY}}{S_X S_Y} = \frac{223.75153}{(6.397633)(38.72593)} = 0.90312
\]

c)

\[
b_1 = \frac{\text{Cov}(X, Y)}{S_X^2} = \frac{223.75153}{40.9297} = 5.4667
\]

Interpretation of \(b_1\) : Each additional hour of the weight training program yields an expected improvement in batting average of 5.47 points.

\[
b_0 = \bar{Y} - b_1 \bar{X} = 50.33 - (5.4667)(11.83) = -14.3411
\]

Interpretation of \(b_0\) : If \(X\) were 0 hour, we would expect to have -14.34 points in \(Y\). Interpret with caution—note that we are extrapolating the results beyond the observed data.

d) Prediction (Estimation) from the least-square linear correlation \(Value of r_{XY} = 0.90312\) shows good correlation of \(Y\) with \(X\) positively.

\[
\hat{Y}_{n+1} = b_0 + b_1 X_{n+1} = -14.3411 + (5.4667)(21) = 100.5 \text{ points}
\]

3. Complete the contingency table (3x2 matrix) given to calculate expected counts \((E_{ij})\) from the corresponding formula. These expected counts \((E_{ij})\) were written below observed counts \((O_{ij})\) as boldface and italic with the row and column total as follows:
<table>
<thead>
<tr>
<th>Activity Level</th>
<th>Male</th>
<th>Female</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sedentary</td>
<td>957</td>
<td>1226</td>
<td>2183</td>
</tr>
<tr>
<td></td>
<td>1046.96</td>
<td>1136.04</td>
<td></td>
</tr>
<tr>
<td>Active</td>
<td>340</td>
<td>417</td>
<td>757</td>
</tr>
<tr>
<td></td>
<td>363.05</td>
<td>393.95</td>
<td></td>
</tr>
<tr>
<td>Very active</td>
<td>842</td>
<td>678</td>
<td>1520</td>
</tr>
<tr>
<td></td>
<td>728.99</td>
<td>791.01</td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>2139</td>
<td>2321</td>
<td>4460</td>
</tr>
</tbody>
</table>

$H_0$: No association exists between activity level and gender.

$H_1$: Association exists between activity level and gender.

$$E_{ij} = \frac{R_i C_j}{n} \rightarrow E_{11} = \frac{R_1 C_1}{n} = \frac{(2183)(2139)}{4460} = 1046.96$$

$$E_{12} = \frac{R_1 C_2}{n} = \frac{4460}{(2183)(2321)} = 1136.04$$

$$E_{21} = \frac{R_2 C_1}{n} = \frac{4460}{(757)(2139)} = 363.05$$

$$E_{22} = \frac{R_2 C_2}{n} = \frac{4460}{(757)(2321)} = 393.95$$

$$E_{31} = \frac{R_3 C_1}{n} = \frac{(1520)(2139)}{4460} = 728.99$$

$$E_{32} = \frac{R_3 C_2}{n} = \frac{(1520)(2321)}{4460} = 791.01$$

$$\chi^2 = \sum_{i=1}^{r} \sum_{j=1}^{c} \left( \frac{O_{ij} - E_{ij}}{E_{ij}} \right)^2$$

$$\chi^2 = \frac{(957 - 1046.96)^2}{1046.96} + \frac{(1226 - 1136.04)^2}{1136.04} + \frac{(340 - 363.05)^2}{363.05} + \frac{(417 - 393.95)^2}{393.95}$$

$$\chi^2 = 7.730 + 7.124 + 1.464 + 1.349 + 17.519 + 16.146 = 51.332$$

Decision rule:

$\chi^2 = \sum_{i=1}^{r} \sum_{j=1}^{c} \left( \frac{O_{ij} - E_{ij}}{E_{ij}} \right)^2 > \chi^2_{(r-1)(c-1), \alpha}$

$\chi^2 = 51.332 > \chi^2_{(3-1)(2-1), 0.001}$

$\chi^2_{2, 0.05} = 5.99$ (Read from Appendix Table 7)

$51.332 > 5.99$

Conclusion: Reject $H_0$ of no association at the 5% significance level ($p-value = P(\chi^2_{2} > 51.332) \ll 0.05$)