1. **[20 points] Testing for Diabetes.**

   (a) **[3 points]** Give estimates for the sensitivity of Test I and of Test II.
   
   **Solution:** 156 patients out of total 223 patients were tested positive by Test I. Hence the estimated sensitivity of Test I is \( \frac{156}{223} = 0.6996 \).
   Similarly, the estimated sensitivity of Test II is \( \frac{200}{223} = 0.8969 \).

   (b) **[10 points]** Are the results of the two tests independent? Use a \( \chi^2 \)-test to test for independence.
   
   **Solution:** We want to test
   
   \[ H_0 : \text{Test I and Test II are independent} \]
   vs.
   
   \[ H_a : \text{The two tests are not independent.} \]
   
   The expected counts for the four cells under the null hypothesis are as follows,

<table>
<thead>
<tr>
<th></th>
<th>Test I</th>
<th>Test II</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>positive</td>
<td>negative</td>
</tr>
<tr>
<td>positive</td>
<td>139.9103</td>
<td>16.0897</td>
</tr>
<tr>
<td>negative</td>
<td>60.0897</td>
<td>6.9103</td>
</tr>
</tbody>
</table>

   The value of the \( \chi^2 \) test statistic for the above hypothesis is
   
   \[ X^2 = \left( \frac{(142 - 139.9103)^2}{139.9103} + \frac{(14 - 16.0897)^2}{16.0897} + \frac{(58 - 60.0897)^2}{60.0897} + \frac{(9 - 6.9103)^2}{6.9103} \right) \]
   
   \[ = 1.0072. \]
   
   and the degree of freedom is \((2 - 1)(2 - 1) = 1\). From the \( \chi^2 \) table we have the
   
   P-value for this test is more than 0.25. Hence we fail to reject the null hypothesis that the two tests are independent at any reasonable level.

   (c) **[12 points]** We assume that Test I has sensitivity, 72%, and specificity, 80%. We also assume that 5% of our population has diabetes.

   i. **[4 points]** What is the probability that a randomly selected individual from the population will be tested positive by Test I?
   
   **Solution:**
   
   \[
P \left[ \text{A randomly selected individual will be tested positive by Test I} \right]
   = P \left[ \text{positive result by Test I} \mid \text{the person has diabetes} \right] P \left[ \text{the person has diabetes} \right] 
   + P \left[ \text{positive result by Test I} \mid \text{the person doesn’t have diabetes} \right] P \left[ \text{The person doesn’t have diabetes} \right]
   \]
   
   \[= \frac{72}{100} \times \frac{5}{100} + \left(1 - \frac{80}{100} \right) \times \frac{95}{10000} = \frac{72 \times 5 + 20 \times 95}{10000} = 0.226.\]
ii. [4 points] What is the probability that a random selected individual from the population has diabetes given his/her Test I result is positive?

Solution:

\[ P \left[ \text{A randomly selected individual has Diabetes and his/her Test I result is positive} \right] = P \left[ \text{Positive result by Test I | The person has diabetes} \right] \times P \left[ \text{The person has diabetes} \right] \]

\[ = \frac{72}{100} \times \frac{5}{100} = 0.036. \]

Hence, the conditional probability that a random selected individual from the population has diabetes given his/her Test I result is positive is \( 0.036 + 0.1593. \)

iii. [4 points] What is the probability that a random selected individual from the population will be tested positive be Test I given he/she has diabetes?

Solution: The conditional probability that a random selected individual from the population will be tested positive be Test I given he/she has diabetes is the sensitivity of Test I which is 0.72.

2. [10 points]

(a) [5 points] Find \( \alpha = \) the probability of a Type I error, that is, the probability that \( H_0 \) is rejected when actually \( H_0 \) is true.

Solution: \( \alpha = P(\bar{x} > 0 \mid \mu = 0) = 0.5 \) since \( \bar{x} \) follows normal distribution with mean \( \mu \) and variance \( \sigma^2/n. \)

(b) [5 points] Find the power of this test when \( \mu = 0.2. \)

Solution: Power of this test when \( \mu = 0.2 \) is

\[ P(\bar{x} > 0 \mid \mu = 0.2) = P \left( \frac{\bar{x} - 0.2}{\sigma/\sqrt{n}} > -\frac{0.2}{\sigma/\sqrt{n}} \right) = P \left( Z > -\frac{0.2}{1/\sqrt{16}} \right) = P(Z > -0.8) = 0.7881. \]


(a) [5 points] Let \( p \) be the probability that the answer of a surveyed voter is “yes”. What is the expression of \( p \) in terms of \( q \)?

Solution:

\[ p = P(\text{the answer of a surveyed voter is “yes”}) \]
\[ = P[\text{the answer of a surveyed voter is “yes” | the voter actually voted}] \times P[\text{a surveyed voter actually voted}] \]
\[ + P[\text{the answer of a surveyed voter is “yes” | the voter didn’t vote}] \times P[\text{a surveyed voter didn’t vote}] \]
\[ = 1 \times 0.66 + q \times 0.34 = 0.66 + 0.34q. \]
(b) [5 points] 
Solution: We know that for \( \hat{p} = X/n \),
\[
E(\hat{p}) = p \quad \text{and} \quad \text{Var}(\hat{p}) = \frac{p(1-p)}{n}.
\]
Hence
\[
E(\hat{q}) = E(-1.94 + 2.94\hat{p}) = -1.94 + 2.94E(\hat{p}) = -1.94 + 2.94p = q
\]
and
\[
\text{Std}(\hat{q}) = \text{Std}(-1.94 + 2.94\hat{p}) = 2.94 \times \text{Std}(\hat{p}) = 2.94 \times \sqrt{\frac{p(1-p)}{n}}
\]

Now using the fact that \( q = -1.94 + 2.94q \) we have \( p = \frac{q + 1.94}{2.94} \). Hence,
\[
2.94 \times \sqrt{\frac{p(1-p)}{n}} = 2.94 \times \sqrt{\frac{(q + 1.94)(2.94q - 1.94)}{(2.94)^2 \times n}} = \sqrt{\frac{(1.94 + q)(1-q)}{n}}.
\]
Hence standard deviation of \( \hat{q} \) is \( \sqrt{(1.94 + q)(1-q)/n} \).

(c) [5 points] Suppose \( X = 80 \) was observed with \( n = 100 \). What is your estimate for \( q \)? Give an approximate 99% confidence interval for \( q \).
Solution: Here \( \hat{p} = \frac{X}{n} = \frac{80}{100} = 0.8 \). So the estimate for \( q \) is
\[
\hat{q} = -1.94 + 2.94\hat{p} = -1.94 + 2.94 \times 0.8 = 0.412.
\]
The estimated standard deviation of \( \hat{q} \) is
\[
\text{SE}_\hat{q} = 2.94 \times \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} = 2.94 \times \sqrt{\frac{0.8 \times 0.2}{100}} = 0.1176.
\]
Now \( \frac{\hat{q}}{\text{SE}_\hat{q}} \) approximately follows N(0, 1) distribution. Hence an 99% confidence interval for \( q \) is \( (\hat{q} \pm z^*\text{SE}_\hat{q}) \) where \( z^* \) is the 99% normal cutoff point 2.576. So the 99% C.I. is
\[
(0.412 \pm 2.576 \times 0.1176) = (0.412 \pm 0.303) = (0.109, 0.715).
\]

(d) [5 points] 
Solution: The margin of error in a 99% confidence interval for \( q \) is \( 2.576 \times \text{SE}_\hat{q} = 2.576 \times 2.94 \times \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \). Using the fact that \( p \geq 0.66 \) we have margin of error is less than \( 2.576 \times 2.94 \times \sqrt{\frac{0.66 \times 0.34}{n}} = \sqrt{\frac{12.871}{n}} \). Hence
\[
\sqrt{\frac{12.871}{n}} \leq 0.05 \quad \Rightarrow \quad n \geq \frac{12.871}{0.05^2} = 5148.4
\]
\[
\Rightarrow \quad n \geq 5149.
\]
4. [15 points] Plates for Glass

(a) [6 points]

Solution: An unbiased estimate for $\mu_X - \mu_Y$ is $\bar{X} - \bar{Y} = 469 - 463 = 6$. Here our assumptions are

- The $X$ and $Y$ samples are independent.
- The population variances are equal.

So an unbiased estimate for the common variance of $X$ and $Y$ is the pooled variance

$$s_p^2 = \frac{(5 - 1) \times s_x^2 + (5 - 1) \times s_y^2}{5 + 5 - 2} = \frac{4 \times 839.5 + 4 \times 916.5}{8} = 878.$$

i.e. $s_p = 29.63$. So estimated standard error of $\bar{X} - \bar{Y}$ is $SE_{\bar{X} - \bar{Y}} = s_p \sqrt{\frac{1}{5} + \frac{1}{5}} = 29.63 \times \sqrt{0.4} = 18.740$. Now

$$\frac{(\bar{X} - \bar{Y}) - (\mu_X - \mu_Y)}{SE_{\bar{X} - \bar{Y}}}$$

follows t-distribution with $5 + 5 - 2 = 8$ degrees of freedom. 95% cutoff point for t-distribution with 8 d.f. is $t^* = 2.307$. Hence an 95% confidence interval for $\mu_X - \mu_Y$ is

$$\left[ (\bar{X} - \bar{Y}) \pm t^* \times SE_{\bar{X} - \bar{Y}} \right] = \left[ 6 \pm 2.307 \times 18.74 \right] = \left[ -37.233, 49.233 \right].$$

Since 0 is in the 95% confidence interval of $\mu_X - \mu_Y$, we fail to reject the null hypothesis that “the averages for the two processes are same” at 5% significance level.

(b) [6 points]

Solution: Here an unbiased estimate for $\mu_X - \mu_Y$ is again $\bar{X} - \bar{Y} = 469 - 463 = 6$. But the assumptions are

- The $X$ and $Y$ samples are paired. So the pairs $(X_i, Y_i)$ are independent.

So estimated standard error of $\bar{X} - \bar{Y}$ is $SE_{\bar{X} - \bar{Y}} = \frac{\text{difference}}{\sqrt{n}} = \sqrt{\frac{21.5}{5}} = 2.074$. Now

$$\frac{(\bar{X} - \bar{Y}) - (\mu_X - \mu_Y)}{SE_{\bar{X} - \bar{Y}}}$$

follows t-distribution with $5 - 1 = 4$ degrees of freedom. 95% cutoff point for t-distribution with 4 d.f. is $t^* = 2.777$. Hence an 95% confidence interval for $\mu_X - \mu_Y$ is

$$\left[ (\bar{X} - \bar{Y}) \pm t^* \times SE_{\bar{X} - \bar{Y}} \right] = \left[ 6 \pm 2.777 \times 2.074 \right] = \left[ 0.241, 11.759 \right].$$

Since 0 is not in the 95% confidence interval of $\mu_X - \mu_Y$, we reject the null hypothesis that “the averages for the two processes are same” at 5% significance level.
(c) [3 points]
Solution: (Less variance due to positive correlation between the pairs. Removal of the effect of lurking variables.)

5. [20 points] Prediction of ozone level

(a) [3 points] Write down the multiple regression equation.
Solution: The estimated regression equation is
\[ \text{OZONE} = 388.4121 - 0.1957033 \times \text{YEAR} + 0.0342877 \times \text{RAIN}. \]

(b) [7 points]
Solution: The missing value for the t-statistic for RAIN is
\[ \frac{b_{\text{RAIN}}}{se_{b_{\text{RAIN}}}} = \frac{0.0342877}{0.0096548} = 3.5514. \]
The error degrees of freedom is \( n - p - 1 = 13 - 2 - 1 = 10. \) So
\[ \frac{b_{\text{RAIN}} - \beta_{\text{RAIN}}}{se_{b_{\text{RAIN}}}} \]
follows \( t \)-distribution with 10 degrees of freedom. Now 95% cutoff point for \( t \)-distribution with 10 d.f. is \( t^* = 2.229. \) Hence a 95% confidence interval for the regression parameter for rain (\( \beta_{\text{RAIN}} \)) is
\[ [b_{\text{RAIN}} \pm t^* \times se_{b_{\text{RAIN}}}] = [0.0342877 \pm 2.229 \times 0.0096548] = [0.012767, 0.055808]. \]

(c) [4 points]
Solution: For the regression model the degrees of freedom is \( p = 2 \) and the regression mean square is \( \text{MSR}=10.3680841/2 = 5.18404205. \)
The error degrees of freedom is \( 12 - 2 = 10 \) and the mean square error is \( \text{MSE}=1.03960755/10 = 0.103960755. \)

(d) [6 points]
Solution: The value of the F-statistic used for this test is
\[ F = \frac{\text{MSR}}{\text{MSE}} = \frac{5.18404205}{0.103960755} = 49.8654. \]
The degrees of freedom for the F-statistic is \( (p, n - 1 - p) = (2, 10). \)

6. [10 points] For the items below, select True or False.

(a) [2 points] If the correlation between two random variables \( x \) and \( Y \) is negative, then \( \text{Var}(X + Y) < \text{Var}(X - Y). \)
Solution: TRUE. Note that \( \text{Var}(X \pm Y) = \sigma_X^2 + \sigma_Y^2 \pm 2\rho\sigma_X\sigma_Y. \) Hence if the correlation \( \rho \) between \( X \) and \( Y \) is negative we have \( \text{Var}(X+Y) - \text{Var}(X-Y) = 4\rho\sigma_X\sigma_Y < 0. \)
(b) [2 points] For two random variables \( X \) and \( Y \), if \( \mathbb{E}(X - Y) = \mathbb{E}(X + Y) \), then \( \mathbb{E}(Y) \) must be equal to 0.

Solution: TRUE. \( \mathbb{E}(X - Y) = \mathbb{E}(X + Y) \) implies \( \mathbb{E}(X) - \mathbb{E}(Y) = \mathbb{E}(X) + \mathbb{E}(Y) \), so \( \mathbb{E}(Y) = 0 \).

(c) [2 points] If we fail to reject a null hypothesis \( H_0 \) at the 0.05 significant level, then there is a 95% probability that \( H_0 \) is true.

Solution: FALSE. Probability of \( H_0 \) being TRUE is 0 or 1.

(d) [2 points] For a specified sample size \( n \), the margin of error for a confidence interval for a population mean \( \mu \) increases as the confidence level increases.

Solution: TRUE. Note that margin of error is \( \frac{z^* \sigma}{\sqrt{n}} \) for known variance and \( \frac{t^* s}{\sqrt{n}} \) for unknown variance. And the cutoff \( z^* \) or \( t^* \) increases as the confidence level increases.

(e) [2 points] In order to calculate a P-value, you must know the distribution of the test statistic under the alternative hypothesis \( H_a \).

Solution: FALSE. We must know the distribution of the test statistic under the null hypothesis \( H_0 \).