Math 251, Final Exam

Instructions: Do 11 of the following 12 questions. Each question is worth 10 pts. The first two questions each consist of several multiple choice, true/false or short answer questions. Please do your best by justifying all of your problem solutions with appropriate work.

1. (i) (2 pts) Given a normal probability distribution, which of the following is true?
   
   (a) The median is to the right of the mean.
   
   (b) The mean is to the right of the median.
   
   (c) The mean is equal to the median.
   
   (d) Any of the above can occur depending on whether the normal distribution is symmetric, negatively or positively skewed.

   **Answer.** C, normal distributions are always symmetric about their means.

   (ii) (2 pts) According to Chebyshev’s theorem, at least what proportion of data in any distribution lies within 5 standard deviations of the mean?

   **Answer.** \(1 - \frac{1}{5^2} = .96\).

   (iii) (1 pt) (True or False) Two events in a sample space are independent if they never occur at the same time.

   **Answer.** False, the above statement describes mutually exclusive events.

   (iv) (1 pt) (True or False) In an hypothesis test on the mean, a Type I error occurs if the null hypothesis is rejected when it is true.

   **Answer.** True.

   (v) (2 pts) For events \(A\) and \(B\) in a sample space \(S\), we are told \(P(A) = .6\) and \(P(B) = .3\) and \(P(A \text{ and } B) = .2\). Which of the following is true?

   (a) \(P(A|B) = 2/3\).
   
   (b) \(P(A \text{ or } B) = .9\)
   
   (c) \(A\) and \(B\) are mutually exclusive.
   
   (d) None of the above are true.

   **Answer.** A, \(P(A|B) = \frac{P(A \text{ and } B)}{P(B)} = \frac{.2}{.3}\).

   (vi) (2 pts) Which of the following is true about a binomial random variable for 10 trials with probability of success on each trial given as .1?

   (a) The probability of no successes is \((.1)^{10}\)
   
   (b) The 10 trials must be identical and independent from one another.
(c) The probability of 10 successes is \((.9)^{10}\)

(d) All of the above.

**Answer.** B

2. (i) (2 pts) Suppose that the maximum error \(E = 4\) for a 95% confidence interval for the mean using a sample of size \(n = 100\). What sample size would you expect to use from the same population for a 95% confidence interval of the mean to have a maximum error \(E = 1\)?

**Answer.** To cut the error by a factor of 4, the sample size must be multiplied by \(4^2\). Therefore, we need a sample size \(n = 1600\).

(ii) (2 pts) According to the empirical rule for bell shaped distributions, approximately what percentage of data lies within one standard deviation of the mean?

**Answer.** 68%

(iii) (2 pts) Suppose a weight of 50 pounds is at the 67th percentile for Springer Spaniel dogs. In a collection of 500 randomly selected Springer Spaniel dogs, how many would you expect to weigh more than 50 pounds?

**Answer.** Approximately 33% of 500 which is 165.

(iv) (1 pt) What z-score would you assign to a number that lies 3.5 standard deviations below the mean?

**Answer.** \(-3.5\).

(v) (1 pt) (True or False) In hypothesis testing, the null hypothesis should be rejected if the level of significance is smaller than the \(P\)-value.

**Answer.** False, it should be rejected when the level of significance is smaller than the \(P\) value.

(vi) (2 pts) Unaware that 33% of the 8,000 voters in his district still support him, a politician decides to estimate his political strength. A sample of 900 voters shows that 36% support him. Which of the following is true?

(a) The parameter of interest is 36%.

(b) The statistic of interest is 33%.

(c) The sample size is 8000.

(d) None of the above.
3. (a) (2 pts) In how many ways can the state of Nevada make license plates of the form yxx-aaa where the first character is a number 1 - 9, the second and third characters are numbers 0 - 9, and the last three characters are letters A - Z (all uppercase)?

**Answer.** \( 9 \cdot 10 \cdot 10 \cdot 26 \cdot 26 \cdot 26 = 15,818,400 \text{ ways.} \)

(b) (2 pts) Given that the probability that a newborn child is female is 48% and 52% that the child is male. What is the probability that exactly 11 of the first 20 babies born in 2003 were female?

**Answer.** \( P(11) = C_{20,11} \cdot (0.48)^{11} \cdot 0.52^{(20-11)} = 0.1455. \)

(c) (2 pts) In how many ways can three medals (gold, silver and bronze) be awarded in a 100m race that contains 9 runners?

**Answer.** \( P_{9,3} = 9 \cdot 8 \cdot 7 = 504 \text{ ways.} \)

(d) Consider the random variable \( x \) whose probability distribution is given by the following table.

\[
\begin{array}{c|cccc}
 x & 2 & 4 & 6 & 8 \\
p(x) & .3 & ? & .3 & .1 \\
\end{array}
\]

(i) (1 pt) Find \( P(x=4) \). Answer = \( 1 - .7 = 0.3 \)

(ii) (3 pts) Find the expected value of \( x \) and the standard deviation of \( x \). The expected value is

\[
\mu = E(x) = \sum x \cdot p(x) = 2(.3) + 4(.3) + 6(.3) + 8(.1) = 4.4.
\]

The variance is

\[
\sigma^2 = \left( \sum x^2 \cdot p(x) \right) - \mu^2 = 4(.3) + 16(.3) + 36(.3) + 64(.1) - 4.4^2 = 3.84
\]

Therefore, the standard deviation is \( \sigma = \sqrt{3.84} = 1.959592. \)

4. Consider the sample of 30 numbers

\[
\begin{array}{cccccccccccccccccccc}
31 & 37 & 44 & 46 & 47 & 51 & 52 & 65 & 68 & 68 & 69 & 70 & 71 & 72 & 73 \\
75 & 75 & 76 & 76 & 77 & 78 & 79 & 80 & 81 & 82 & 83 & 85 & 88 & 89 \\
\end{array}
\]

(a) (3 pts) Construct a stem and leaf plot for this data with stems 3,4,5,6,7,8

(b) (7 pts) Construct a relative frequency histogram for the data where the first class has limits 30 – 39. State the class width, and list the limits, boundaries and relative frequencies for all classes.
5. (10 pts) Given the data 9,12,15,17,18,19,23,45,52,61,63,63. One has \( \sum x = 397 \) and \( \sum (x - \mu)^2 \approx 5206.917 \). Find:

(a) the mean is \( \frac{397}{12} = 33.08 \).

(b) the sample variance is \( s^2 = \frac{5205.917}{11} = 473.36 \).

(c) the population variance is \( \sigma^2 = \frac{5205.917}{12} = 433.90915 \).

(d) the sample standard deviation is \( s = \sqrt{473.36} = 21.76 \).

(e) the median is \( \frac{19 + 23}{2} = 21 \).

(f) the 50th percentile of the data is 21 (the same as the median).

(g) the mode is 63 the most common data.

6. The following data is for investigating the relation between salary in thousands (x) and average number of absences per year (y).

<table>
<thead>
<tr>
<th>Salary (x)</th>
<th>20</th>
<th>23</th>
<th>28</th>
<th>30</th>
<th>33</th>
<th>35</th>
<th>37</th>
<th>40</th>
<th>42</th>
</tr>
</thead>
<tbody>
<tr>
<td>Absences (y)</td>
<td>2.4</td>
<td>2.2</td>
<td>1.9</td>
<td>2.1</td>
<td>1.5</td>
<td>1.4</td>
<td>1.3</td>
<td>0.5</td>
<td>0.4</td>
</tr>
</tbody>
</table>

For this data: \( \sum x = 288 \), \( \sum x^2 = 9660 \), \( \sum y = 13.7 \), \( \sum y^2 = 24.93 \), \( \sum xy = 398.2 \).

(a) (4 pts) Find the equation of the least squares regression line.

**Answer.** First, \( SS_x = 9660 - \frac{288^2}{9} = 444 \) and \( SS_{xy} = 398.2 - \frac{(288)(13.7)}{9} = -40.2 \).

Therefore, the slope is \( b = \frac{SS_{xy}}{SS_x} = -.09054 \), and the y-intercept is \( a = \frac{13.7}{9} - (-.09054) \left( \frac{288}{9} \right) = 4.4195 \).

Thus the equation of the line is \( y = -.09054x + 4.4195 \).
(b) (2 pts) Does the data appear to be positively or negatively correlated? Explain.

Answer. Negatively correlated because as the $x$'s increase, the $y$'s generally decrease.

(c) (2 pts) How many absences per year would you expect from an employee that makes $38,000 per year?

Answer. Solve $y = y = -0.09054(38) + 4.4195 = 0.98$ absences per year.

(d) (2 pts) Given that the correlation coefficient is $-0.945$, determine how well the data fits the line of best fit. Explain your answer.

Answer. A correlation coefficient of $r = -1$ would mean the data is on a line of negative slope, values close to $-1$ indicate good fits to a line of negative slope. Since $-0.945$ is quite close to $-1$, the fit should be quite good.

7. A business employs 300 men and 700 women. Of these employees, 60 men and 210 women have been working there for more than 10 years. Let $A$ be the event the employee is a woman, and $B$ be the event the employee has been employed for more than 10 years.

(a) (3 pts) Find $P(A)$ and $P(B)$.

Answer. $P(A) = 0.7$ and $P(B) = 0.27$.

(b) (3 pts) Find $P(A \text{ and } B)$ and $P(A \text{ or } B)$.

Answer. $P(A \text{ and } B) = P(A)P(B|A) = (0.7) \left( \frac{210}{700} \right) = 0.21$.

$P(A \text{ or } B) = P(A) + P(B) - P(A \text{ and } B) = 0.76$.

(c) (2 pts) Are the events $A$ and $B$ independent? Explain.

Answer. No, $P(A \text{ and } B) \neq P(A) \cdot P(B)$ since $0.21 \neq (0.7)(0.27)$.

(d) (2 pts) Find the probability an employee is female, given that the length of employment is more than 10 years.

Answer. $\frac{210}{270} = 0.777778$.

8. In a recent survey of 180 randomly selected engineering graduates starting work in the United States, it was found that the average annual starting salary for the sample was $43,200 with a standard deviation of $6,000.

(a) (7 pts) Conduct an hypothesis test to determine whether the average starting salary for engineers is more than $42,500 per year using a significance level of $\alpha = 0.05$. Make sure to state $H_0$, $H_1$, the critical region, and your conclusion.

Answer. $H_0 : \mu = 42,500$, $H_1 : \mu > 42,500$. This is a right-tailed test with critical region $z \geq 1.645$. The sample test statistic is

$$z = \frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}}} = \frac{43,200 - 42,500}{\frac{6000}{\sqrt{180}}} = 1.565.$$
Since 1.565 is not in the critical region, we fail to reject $H_0$; there is not enough evidence to show that the true mean starting salary is more than $42,500 at the 5% level of significance.

(b) (3 pts) If the true mean starting salary for engineers in the United States is actually $43,000 with standard deviation $6,000, what is the probability that a random sample of 180 starting engineers will have a sample mean starting salary less than $44,000?

**Answer.** $P(\bar{x} < 44,000) = P\left(z < \frac{44,000 - 43,000}{\frac{6000}{\sqrt{180}}}\right) = P(z < 2.24) = .9875.$

9. A “Gallop” poll of 1100 randomly selected Americans found that 47.3 percent think the Elian Gonzales story was the top news event of 2000.

(a) (5 pts) Conduct an hypothesis test to determine if the true population proportion of Americans who think the Elian Gonzales story was the top news event of 2000 is less than .50. Report the P-value of your test and conclusion at a level of significance of $\alpha = .05$.

**Answer.** $H_0 : p = 0.5$ and $H_1 : p < 0.5$. This is a left-tailed test at a level of significance of 0.05. We compute the sample statistic

$$z = \frac{\hat{p} - p}{\sigma_{\hat{p}}} = \frac{0.473 - 0.5}{\sqrt{\frac{(0.5)(0.5)}{1100}}} = -1.79.$$ 

The P-value is $P(z < -1.79) = 0.367$. Because the P-value is less than $\alpha$ we reject the null hypothesis.

(b) (2 pts) Based on this information, would you be comfortable saying that less than 50 percent of Americans think the Elian Gonzalez story was the top news event in 2000. Explain.

**Answer.** Yes, based on the data we are about 96% sure that the population percentage is less than 50 percent.

(c) (3 pts) What sample size would be needed so that you could estimate the true proportion in a population within ±.02 with 99 percent confidence?

**Answer.** If we do not assume an initial estimate for $p$, we use the formula $n = \frac{1}{4} \left(\frac{z_c}{E}\right)^2 = \frac{1}{4} \left(\frac{2.58}{.02}\right)^2 = 4160.25$, therefore, we go to the next larger whole number and select a sample size of $n = 4161$.

(If you assume an intial estimate of .473 for $p$, you will find a slightly smaller sample size).

10. In a 1993 survey of 150 Education graduates and 200 Social Science graduates, the following data were obtained for their average starting salaries.

<table>
<thead>
<tr>
<th>Major</th>
<th>Mean</th>
<th>SD</th>
</tr>
</thead>
<tbody>
<tr>
<td>Education</td>
<td>24,500</td>
<td>2600</td>
</tr>
<tr>
<td>Social Sciences</td>
<td>21,500</td>
<td>2400</td>
</tr>
</tbody>
</table>
(a) (5 pts) Conduct an hypothesis test to determine whether $\mu_1 - \mu_2 < 3800$ using a 5% level of significance where $\mu_1$ is the population mean salary for the Education graduates and $\mu_2$ is the population mean salary for the Social Science graduates. Make sure to state $H_0$, $H_a$, the rejection region, and your conclusion.

**Answer.** $H_0 : \mu_1 - \mu_2 = 3800$ and $H_1 : \mu_1 - \mu_2 < 3800$. For this, we compute $z = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sigma_{\bar{x}_1-\bar{x}_2}}$ where $\sigma_{\bar{x}_1-\bar{x}_2} = \sqrt{\frac{\sigma^2_1}{n_1} + \frac{\sigma^2_2}{n_2}}$. First,

$$
\sigma_{\bar{x}_1-\bar{x}_2} = \sqrt{\frac{2600^2}{150} + \frac{2400^2}{200}} = 271.78.
$$

Therefore,

$$
z = \frac{(24,500 - 21,500) - (3800)}{271.78} = -2.94
$$

Thus the Pvalue is $P(z < -2.94) = .0016$, and so we reject the null hypothesis at the 5% level, the difference is less than $3,800$.

(c) (5 pts) Find a 99 percent confidence interval for $\mu_1$, the mean salary of Education graduates. Does the confidence interval suggest that $\mu_1 > 24,000$? Explain.

**Answer.** From the table, $z_{.99} = 2.58$, thus the 99% confidence interval has endpoints $24,500 \pm 2.58 \cdot \frac{2600}{\sqrt{150}}$ which gives us the interval $(23,953.36, 25046.64)$.

**Answer.** Because 24,000 is in the confidence interval, there is a possibility that the mean is $24,000$ or less; however, the possibility looks to be quite small, and performing an hypothesis test with a Pvalue would provide more information to this question. However, as is, the confidence interval does not rule out at the 99% level of confidence that the mean is different from $24,000$.

11. (a) (2 pts) Find the value $z_c$ needed in the formula for constructing a 94% confidence interval from a large sample.

$z_c = 1.88$ (look on normal table corresponding to an area to the left of $.5 + .94/2 = .97$).

(b) (2 pts) What assumptions must be made on the population and/or sample when constructing a confidence interval for the mean using a large sample?

**Answer.** The sample size is large, usually $n \geq 30$ is a good rule of thumb for most distributions.

(c) (2 pts) Find the rejection region for a two-tailed hypothesis test at level of significance $\alpha = .05$ on a mean using a sample of size $n = 12$ from an approximately normal population with unknown standard deviation?

**Answer.** This is a small sample mean test. Use the $t$-distribution with $d.f. = n - 1 = 11$ and find that the critical region is $t \leq -2.201$ or $t \geq 2.201$. (Not covered Fall 2003)

(d) (2 pts) What assumptions must be made on the population and or sample to do a hypothesis test on a proportion?
Answer. The number of successes and the number of failures in the sample are both larger than 5.

(e) (2 pts) Find the rejection region for a two-tailed test on a population proportion at a level of significance of $\alpha = .06$ assuming the appropriate conditions as requested in (d) are satisfied.

Answer. The critical region is $z \leq -1.88$ or $z \geq 1.88$. These values are chosen so that 3% of the normal curve is to the right of $z = 1.88$ and 3% is to the left of $z = -1.88$.

12. (a) A company wishes to check whether the mean weekly production levels at their five factories are equal using the method of analysis of variance.

(i) (2 pts) State what conditions should be satisfied by the populations and samples in order to use the method of analysis of variance.

Answer. Production levels at each factor have a normal distribution. Samples are randomly chosen and are independent from one another. The standard deviations for all the populations are (approximately) the same.

(ii) (1 pt) State the null and alternative hypotheses.

Answer. $H_0: \mu_1 = \mu_2 = \mu_3 = \mu_4 = \mu_5$, $H_1$: at least two of the means $\mu_1, \mu_2, \mu_3, \mu_4, \mu_5$ are different.

(iii) (2 pts) Given that random samples of size 4 were obtained from each factory. What is the rejection region for the hypothesis test conducted at a level of significance of $\alpha = .05$.

Answer. $N = 20$ is the total number of data, and $k = 5$ is the total number of populations. Therefore, $d.f.\text{bet} = k - 1 = 5 - 1 = 4$, thus d.f. numerator is 4, while $d.f.\text{within} = N - k = 20 - 5 = 15$, thus d.f. denominator. Looking at the $F$-distribution table, we find that $F \geq 3.06$.

(b) (5 pts) Suppose that in an attempt to determine whether a die is fair, you tossed it 120 times with the following results:

<table>
<thead>
<tr>
<th>Outcome</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
</tr>
</thead>
<tbody>
<tr>
<td>Frequency</td>
<td>24</td>
<td>18</td>
<td>15</td>
<td>26</td>
<td>21</td>
<td>16</td>
</tr>
</tbody>
</table>

Does this provide sufficient evidence at a level of significance of $\alpha = .05$ to conclude that the die is not fair? Be sure to find the test statistic, rejection region, and state the conclusion.

Answer. We test $H_0 : p_1 = p_2 = p_3 = p_4 = p_5 = p_6 = \frac{1}{6}$ versus $H_1 :$ not all of the proportions are $\frac{1}{6}$.

Since there are 120 tosses, $E = 20$ in each category, and we use the formula

$$\chi^2 = \sum \frac{(0 - E)^2}{E},$$
where there are $n - 1 = 6 - 1 = 5$ degrees of freedom, so the critical region is $\chi^2 \geq 11.07$. The sample test statistic is

$$\chi^2 = \frac{(24 - 20)^2 + (18 - 20)^2 + (15 - 20)^2 + (26 - 20)^2 + (21 - 20)^2 + (16 - 20)^2}{20}$$

$$= \frac{98}{20} = 4.9.$$  

Because this does not fall in the critical region, we do not reject the null hypothesis, that is, there is not sufficient evidence to conclude that the die is not fair.