MAT 2379 C, Final Examination (with Solutions)

December 15, 2014 **Professor Maryam Sohrabi**
Time: 3 hours

Student Number:_____ **Seat Number:** _____

Family Name: _____ **First Name:** _____

- **This is a closed book examination. Only faculty-approved calculators are permitted: TI 30, TI 34, Casio fx-260 and Casio fx-300.**
- **Record your answer to each question in the table below. Each question is worth 1 mark.**
- **At the end of the examination, hand in only this page.**

Question	Answer	Question	Answer
1	D	14	A
2	B	15	A
3	D	16	A
4	B	17	D
5	C	18	B
6	A	19	C
7	B	20	B
8	D	21	E
9	C	22	E
10	A	23	B
11	C	24	D
12	E	25	C
13	E		

1. The average length of human gestation is approximately 40.5 weeks. It is thought that maternal diabetes may influence the length of the gestation. In a study consisting of 20 diabetic pregnant women, it was found that the mean gestation period was 38.8 weeks with a standard deviation of 5 weeks. Is there enough evidence that the length of gestation in diabetic women is significantly different than the value of 40.5 weeks? Use an appropriate test of hypotheses of level $\alpha = 0.05$. Report the range of the p-value and the conclusion of the test.

   A) The p-value is between 0.01 and 0.025. The mean length of gestation for diabetic women is significantly different than 40.5.
   B) The p-value is between 0.025 and 0.05. The mean length of gestation for diabetic women is significantly different than 40.5.
   C) The p-value is between 0.05 and 0.10. There is not enough evidence that the mean length of gestation for diabetic women is significantly different than 40.5.
   D) The p-value is between 0.10 and 0.20. There is not enough evidence that the mean length of gestation for diabetic women is significantly different than 40.5.
   E) The p-value is between 0.20 and 0.40. There is not enough evidence that the mean length of gestation for diabetic women is significantly different than 40.5.

   Solution: We denote by μ the mean length of gestation for diabetic women. We would like to test $H_0 : \mu = 40.5$ against $H_1 : \mu \neq 40.5$. We know that $n = 20$, $\bar{x} = 38.8$ and $s = 5$. The observed value of the test statistic is:

$$t_0 = \frac{\bar{x} - 40.5}{s/\sqrt{n}} = \frac{38.8 - 40.5}{5/\sqrt{20}} = -1.52.$$

   The p-value of the test is:

$$p\text{-value} = 2P(T_{19} > 1.52)$$

   From Table 17.4 (row 19) we see that 1.52 is between the values 1.328 and 1.729, whose corresponding probabilities to the right are 0.10 and 0.05. Hence $P(T_{19} > 1.52)$ is between 0.05 and 0.10 and

$$0.10 < p\text{-value} < 0.20$$

Since p-value > 0.05, we fail to reject H_0. There is not enough evidence that the mean length of gestation for diabetic women is significantly different than 40.5. The answer is D.

2. The Bacillus Calmette-Guérin (BCG) vaccine for tuberculosis (TB) is mandatory for school-age children in many European countries. In Canada, before BCG vaccination, the patient is tested for TB using a tuberculin skin test, called the Mantoux test. People who have been BCG vaccinated will often have a positive Mantoux test result, although they many not have TB. Therefore, the Mantoux test is not a very efficient tool for detecting TB. In a recent study, 12% of the subjects had a positive Mantoux test result. Among those with a positive test result, only 10% had TB. On the other hand, 1% of the patients with a negative test result also had TB. What was the percentage of patients with TB in this study?

   A) 1.10% B) 2.08% C) 0.88% D) 1.20% E) 13.03%

   Solution: We denote by TB the event that a randomly selected person in this group has tuberculosis.

$$\begin{aligned} P(\text{TB}) &= P(\text{TB}|\text{Test}+)P(\text{Test }+) + P(\text{TB}|\text{Test}-)P(\text{Test}-) \\ &= (0.10)(0.12) + (0.01)(0.88) = 0.0208 \end{aligned}$$

   The answer is B.

3. A team of researchers have studied the effect of reminiscence therapy on older women suffering from depression. The researchers have measured the depression level (on a scale from 1 to 10) for 20 women of age 60+ who have stayed for at least 3 months in a long-term care facility. On this scale, a high depression level is interpreted as severe depression. For each of the 20 women, the depression level was measured before and after the reminiscence therapy. Using the data from this study, we created in R the variables `before` and `after`, which contain the depression levels for the 20 women, before and after the therapy, respectively. We also calculated the difference `d` between the depression level after the therapy and the depression level before the therapy, for each women. Below is the R output which gives the summary for these differences.

```
> d=after-before
> summary(d)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  -5.00   -3.00   -0.50   -1.35    0.00    1.00
> sd(d)
[1] 2.084403
```

Suppose that the difference between the depression level after the therapy and the depression level before the therapy is normally distributed. Using the above summary, calculate a 95% confidence interval (c.i.) for the mean difference μ_D between the depression level after the therapy and the depression level before the therapy. Based on this confidence interval, are we confident that on average, the depression is less severe after the therapy?

A) c.i.=$[0.37; 2.33]$. We are confident that the depression is more severe after the therapy, on average.
B) c.i.=$[-2.33; 0.37]$. There is not enough evidence that the depression is less severe after the therapy, on average.
C) c.i.=$[-2.53; -0.57]$. We are confident that the depression is less severe after the therapy, on average.
D) c.i.= $[-2.33; -0.37]$. We are confident that the depression is less severe after the therapy, on average.
E) c.i.=$[-2.33; -0.37]$. There is not enough evidence that the depression is less severe after the therapy, on average.

Solution: These are paired data. We define the difference

$D =$ depression level after therapy $-$ depression level before therapy.

We know that $n = 20, \overline{d} = -1.35$ and $s_d = 2.0844$. A 95% confidence interval for μ_D is

$$-1.35 \pm (2.093)\frac{2.0844}{\sqrt{20}} = -1.35 \pm 0.976 = [-2.33; -0.37],$$

where $t = 2.093$ is found in Table 17.4 such that $P(T \leq t) = 0.975$ and T has a T distribution with 19 degrees of freedom. Because all the values in the interval are negative, we are confident that $\mu_D < 0$, i.e. the depression level after the therapy is smaller than the depression level

before the therapy. So the depression is less severe after the therapy. The answer is D.

4. A study was conducted to estimate the sensitivity and specificity of a new procedure for detecting the presence of a kidney disease among patients suffering from hypertension. Among the 54 hypertensive patients who had the kidney disease, the procedure identified the disease for 45 subjects. Among the 83 hypertensive patients who did not have the kidney disease, the procedure identified the disease for 24 subjects. Consider a patient chosen from a certain hypertensive population in which the prevalence of this kidney disease is 8%. If the new procedure identifies the presence of the kidney disease for this patient, what is the probability that patient truly has the disease? Assume that the sensitivity and specificity of the procedure remain the same as in the study mentioned above.

   A) 0.8820 B) 0.2004 C) 0.3727 D) 0.5732 E) 0.0545

   Solution: Let $T+$ be the event that the new procedure identifies the presence of the disease and D the event that the patient has the disease. The fact that the prevalence of the disease is 8% means that $P(D) = 0.08$. The fact that the sensitivity and specificity of the procedure remain the same as in the study means that:

$$P(T+|D) = 45/54 \quad \text{and} \quad P(T+|D') = 24/83.$$

   By Bayes' rule,

$$\begin{aligned} P(D|T+) &= \frac{P(D \cap T+)}{P(T+)} = \frac{P(T+|D)P(D)}{P(T+)} \\ &= \frac{(45/54)(0.08)}{0.3327} = 0.2004 \end{aligned}$$

   where

$$\begin{aligned} P(T+) &= P(T+|D)P(D) + P(T+|D')P(D') \\ &= (45/54)(0.08) + (24/83)(1-0.08) \\ &= 0.3327. \end{aligned}$$

   The answer is B.

5. The intraocular pressure is the fluid pressure inside the eye. Glaucoma is an eye disease that is manifested by high intraocular pressure. The distribution of intraocular pressure in the general population is approximately normal with mean 16 mm Hg and standard deviation 3 mm Hg. The normal range for intraocular pressure is considered to be between 12 mm Hg and 20 mm Hg (including these values). Which one of the following commands in R gives the probability that a randomly chosen person has normal intraocular pressure? (Only one answer is correct.)

   A) `qnorm(20,16,3)-qnorm(12,16,3)`
   B) `pnorm(20,3,16)-pnorm(12,3,16)`
   C) `pnorm(20,16,3)-pnorm(12,16,3)`
   D) `pnorm(20,16,3)-pnorm(11,16,3)`
   E) `pnorm(20,16,9)-pnorm(12,16,9)`

   Solution: We wish to calculate $P(12 \leq X \leq 20)$, where X has a normal distribution with mean $\mu = 16$ and standard deviation $\sigma = 3$. This probability is:

$$P(12 \leq X \leq 20) = P(X \leq 20) - P(X < 12) = P(X \leq 20) - P(X \leq 12)$$
$$= \text{pnorm}(20, 16, 3) - \text{pnorm}(12, 16, 3)$$

   We used the fact that $P(X < 12) = P(X \leq 12)$, since X is a *continuous* random variable. The answer is C. (The incorrect answer D is obtained using $P(X < 12) = P(X \leq 11)$, which would be true if X was a *discrete* random variable.)

6. Aboriginal people in Canada have a higher risk of developing many chronic diseases compared with the rest of the population. In a particular Aboriginal community, 16% of the population has tuberculosis, 20% have diabetes and 8% have both diseases. What is the probability that a randomly selected individual in this community does not have either one of the two diseases?

   A) 0.72 B) 0.28 C) 0.64 D) 0.85 E) 0.90

   Solution: Let A be the event that the person has tuberculosis and B the event that the person has diabetes. We know that $P(A) = 0.16$, $P(B) = 0.20$ and $P(A \cap B) = 0.08$. By the addition rule,

$$P(A \cup B) = P(A) + P(B) - P(A \cap B) = 0.16 + 0.20 - 0.08 = 0.28.$$

The probability that the person does not have either one of the two diseases is:

$$P(A' \cap B') = 1 - P(A \cup B) = 1 - 0.28 = 0.72$$

The answer is A.

7. Ebola virus disease (EVD), formerly known as Ebola haemorrhagic fever, is a severe, often fatal illness in humans. It is thought that fruit bats of the Pteropodidae family are natural Ebola virus hosts. The virus is introduced into the human population through close contact with bodily fluids of infected animals. The incubation period (the time interval from infection with the virus to onset of symptoms) is between 2 to 21 days. The following data gives the incubation period (in days) for 16 patients infected with the Ebola virus:

   4 5 6 6 7 8 9 9 11 12 13 15 15 17 20 21

   Calculate the median ($\tilde{x}$), first quartile (q_1) and third quartile (q_3) for this data set. Give the values of the outliers (if they exist).

   A) $\tilde{x} = 10, q_1 = 5, q_3 = 15$; the value 21 is an outlier
   B) $\tilde{x} = 10, q_1 = 6.25, q_3 = 15$; there are no outliers
   C) $\tilde{x} = 11, q_1 = 6, q_3 = 17$; the value 4 is an outlier
   D) $\tilde{x} = 11, q_1 = 6.25, q_3 = 15$; the values 4 and 5 are outliers
   E) $\tilde{x} = 11, q_1 = 5.25, q_3 = 15.5$; there are no outliers

   Solution: Note that the data is already arranged in increasing order. Hence

   $y_1 = 4 \quad y_2 = 5 \quad y_3 = 6 \quad y_4 = 6 \quad y_5 = 7 \quad y_6 = 8 \quad y_7 = 9 \quad y_8 = 9 \quad y_9 = 11$

   $y_{10} = 12 \quad y_{11} = 13 \quad y_{12} = 15 \quad y_{13} = 15 \quad y_{14} = 17 \quad y_{15} = 20 \quad y_{16} = 21$

   For this dataset, $n = 16$ is even. Hence, the median is:

   $$\tilde{x} = \frac{y_8 + y_9}{2} = \frac{9 + 11}{2} = 10$$

   To compute the first quartile, we note that $(n + 1)/4 = 17/4 = 4.25$, which is between 4 and 5 (closer to 4). The first quartile is:

   $$q_1 = (0.75)y_4 + (0.25)y_5 = (0.75)(6) + (0.25)(7) = 6.25$$

To compute the third quartile, we note that $3(n+1)/4 = 51/4 = 12.75$, which is between 12 and 13 (closer to 13). The third quartile is:

$$q_3 = (0.25)y_{12} + (0.75)y_{13} = (0.25)(15) + (0.75)(15) = 15$$

To find the outliers, we need to find the location of the two fences. The inter-quartile range is $IQR = q_3 - q_1 = 15 - 6.25 = 8.75$. Hence

$$\text{Fence1} = q_1 - (1.5)\text{IQR} = 6.25 - (1.5)(8.75) = 6.25 - 13.125 = -6.875$$

$$\text{Fence2} = q_3 + (1.5)\text{IQR} = 15 + (1.5)(8.75) = 15 + 13.125 = 28.125$$

Since there are no data points outside the two fences, we conclude that there are no outliers. The answer is B.

8. In biochemistry and pharmacology, a receptor is a protein molecule usually found embedded within the plasma membrane surface of a cell that receives chemical signals from outside the cell. A sample of 9 cells was found to contain an average of 1203 fmol receptors per milligram of membrane protein, with an estimated standard error of the mean of 64 fmol. (An fmol is equal to 10^{-15} moles.) Using this data, give a 95% confidence interval for the average amount (in fmols) of receptors per milligram found in the membrane protein of these cells. Assume that the amount of receptors per milligram of membrane protein is normally distributed.

A) $[1077.3; 1329.7]$ B) $[1153.8; 1252.2]$ C) $[0; 1322.8]$
D) $[1055.4; 1350.6]$ E) $[1098.1; 1308.9]$

Solution: We denote by μ the average amount of receptors per milligram of membrane protein. The 95% confidence interval for μ is:

$$\bar{x} \pm t\frac{s}{\sqrt{n}}$$

where the value t is read in Table 17.4 such that $P(-t < T_8 < t) = 0.95$, i.e. $P(T_8 < t) = 0.975$. From Table 17.4 (row 8, column 0.975), we find the value $t = 2.306$. Recall that the estimated standard error of the mean is $s_{\bar{X}} = s/\sqrt{n}$. In our case, this value is known to be 64. Therefore, the 95% confidence interval for μ becomes:

$$1203 \pm 2.306(64) = 1203 \pm 147.584 = [1055.416;\ 1350.584].$$

The answer is D. The incorrect answer B is obtain using: $1203 \pm (2.306)(64/\sqrt{9}) = [1153.805;\ 1252.195]$.

9. The following R output summarizes the data representing the weights (in lb) of 50 premature births in a major hospital.

```
> summary(x)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  1.432   4.607   6.158   6.809   8.212  18.700
> sd(x)
[1] 3.420575
```

A logarithmic transformation is applied to this data set, using the following R command:

```
> y=log(x)
```

The following R output gives the summary for the transformed measurements:

```
> summary(y)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 0.3591  1.5276  1.8178  1.7990  2.1056  2.9285
> sd(y)
[1] 0.5064342
```

Calculate a =the geometric mean and b =the geometric standard deviation for the 50 weights of premature birth.
A) $a = 905.96$ and $b = 30.59$ B) $a = 6.809$ and $b = 3.42$
C) $a = 6.04$ and $b = 1.66$ D) $a = 1.79$ and $b = 0.50$
E) $a = 0.59$ and $b = 1.36$

Solution: The geometric mean and geometric standard deviation are: $e^{\bar{y}} = e^{1.799} = 6.04$ and $e^{s_y} = e^{0.5064342} = 1.66$. The answer is C. (The incorrect answer A is obtained by computing $a = e^{\bar{x}} = e^{6.809} = 905.96$ and $b = e^{s_x} = e^{3.420575} = 30.59$.)

10. The following data gives the birth weights (in ounces) for 6 consecutive deliveries at the Civic Hospital. Assuming that the birth weights follow

a normal distribution, find a 90% confidence interval for the average birth weight μ.

97 117 140 78 99 148

A) $[91.0; 135.4]$ B) $[84.8; 141.5]$ C) $[91.6; 134.8]$ D) $[95.0; 131.3]$ E) $[92.3; 133.6]$

Solution: The sample mean and sample standard deviation for this sample are:

$$\bar{x} = \frac{1}{6}\sum_{i=1}^{6} x_i = 113.1667,$$

and

$$s = \sqrt{\frac{1}{5}\sum_{i=1}^{6}(x_i - \bar{x})^2} = 27.00679.$$

A 90% confidence interval for μ is based on the T distribution with $6-1 = 5$ degrees of freedom. For this level of confidence, the probability at the right of the point t is 0.05. Table 17.4 gives the value $t = 2.015$. Therefore, the 90% confidence interval for μ is

$$113.1667 \pm 2.015\left(\frac{27.00679}{\sqrt{6}}\right) = [90.950;\ 135.383]$$

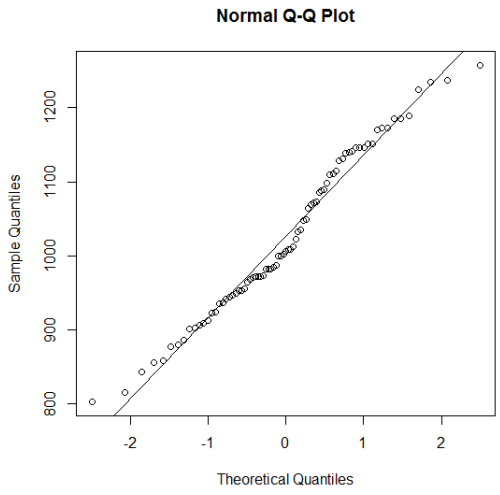
The answer is A. The incorrect answers B, C and D are obtained using the wrong values $t = 2.571$, $t = 1.96$, respectively $t = 1.645$.

11. The Younger Dryas Cold Event (or the "Big Freeze") was an abrupt cooling event of the Northern Hemisphere which occurred approximately 12,000 years ago, and might have resulted from a slowing of the Atlantic meridional overturning circulation (AMOC). The most common means of slowing the AMOC involves the reduction of oceanic surface water density via an increase in freshwater discharge to the North Atlantic. To predict if such an event might happen again, the density of the ocean water near surface is closely monitored. The following R output gives the summary for 79 measurements of the density of the Atlantic ocean water near surface (in kg/m^3), at a latitude of 45 degrees north:

```
> summary(x)
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  802.5   948.1  1006.0  1026.0  1122.0  1258.0
> var(x)
[1] 12013.36
> sd(x)
[1] 109.61
```

The picture below gives the QQ-plot for this data, together with the line of best fit, produced using R:




```
> qqnorm(x)
> abline(mean(x),sd(x))
```

If the line of best fit has the equation $y = a + bz$, where y is the sample quantile and z is the theoretical quantile, what are the values of a and b?

A) $a = 1006$ and $b = 109.61$
B) $a = 109.61$ and $b = 1026$
C) $a = 1026$ and $b = 109.61$
D) $a = 1026$ and $b = 173.90$
E) $a = 1006$ and $b = 12013.36$

Solution: The line of best fit has equation $y = \hat{\mu} + \hat{\sigma}z$ where $\hat{\mu} = \bar{x} = 1026$ and $\hat{\sigma} = s = 109.61$. Hence $a = 1026$ and $b = 109.61$. The answer is C.

12. The following data gives the number of deadly bear attacks in North America per decade, for the 9 decades between 1900 and 1989:

$$2,\ 1,\ 4,\ 8,\ 6,\ 9,\ 9,\ 19,\ 20.$$

Calculate the mean and standard deviation for the number of deadly bear attacks in North America per decade.

A) The mean is 8.667 and the standard deviation is 5.6505.
B) The mean is 8.0 and the standard deviation is 19.0.
C) The mean is 8.0 and the standard deviation is 5.0.
D) The mean is 8.667 and the standard deviation is 46.0.
E) The mean is 8.667 and the standard deviation is 6.7823.

Solution: The mean is

$$\overline{x} = \frac{1}{9}\sum_{i=1}^{9} x_i = \frac{78}{9} = 8.6667$$

and the standard deviation is:

$$s = \sqrt{\frac{(\sum_{i=1}^{9} x_i^2) - (\sum_{i=1}^{9} x_i)^2/9}{8}} = \sqrt{\frac{1044 - (78)^2/9}{9-1}} = \sqrt{46} = 6.7823.$$

The answer is E.

13. In the Unites States, the blood types have the following distribution: 41% O, 31% A, 22% B and 6% AB. It is known that O is a universal donor, A can donate only to A and AB, B can donate only to B and AB, and AB can donate only to AB. If a patient who needs a blood transfusion receives blood from a randomly selected donor, and the two persons are independent of each other, what is the probability that the transfusion is successful?

A) 0.6607 B) 0.3393 C) 0.4101 D) 0.7314 E) 0.5899

Solution: Let A_1, A_2, A_3, A_4 be the events that the donor's blood type are O, A, B, respectively AB. Let B_1, B_2, B_3, B_4 be the events that the blood type of the receiving individual are O, A, B, respectively AB. The event A_i is independent of B_j, for any $i = 1, 2, 3, 4$ and $j =$

$1, 2, 3, 4$. The event that the transfusion is successful can be written as the following union of disjoint events:

$$C = (A_1 \cap B_1) \cup (A_1 \cap B_2) \cup (A_1 \cap B_3) \cup (A_1 \cap B_4) \cup (A_2 \cap B_2) \cup (A_2 \cap B_4) \cup (A_3 \cap B_3) \cup (A_3 \cap B_4) \cup (A_4 \cap B_4).$$

Hence,

$$\begin{aligned} P(C) &= P(A_1)P(B_1) + P(A_1)P(B_2) + P(A_1)P(B_3) + P(A_1)P(B_4) + \\ &\quad P(A_2)P(B_2) + P(A_2)P(B_4) + P(A_3)P(B_3) + P(A_3)P(B_4) + \\ &\quad P(A_4)P(B_4) \\ &= (0.41)(0.41) + (0.41)(0.31) + (0.41)(0.22) + (0.41)(0.06) + \\ &\quad (0.31)(0.31) + (0.31)(0.06) + (0.22)(0.22) + (0.22)(0.06) + \\ &\quad (0.06)(0.06) \\ &= 0.5899 \end{aligned}$$

The answer is E.

14. 20% of the trees in a certain forest are maple trees. In this forest, 15% of the maple trees are mature trees, with age between 10 and 15 years. We select randomly a tree in this forest. What is the probability that this is a maple tree with age between 10 and 15 years?

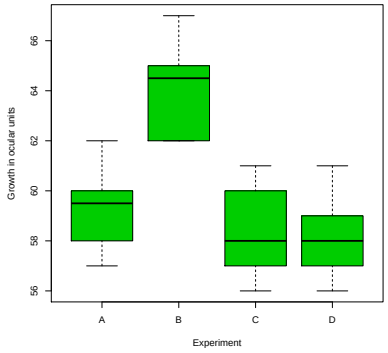
A) 0.03 B) 0.15 C) 0.20 D) 0.75 E) 0.175

Solution: We denote by A the event that the tree is a maple tree and B the event that the tree has an age between 10 and 15 years. We know that $P(A) = 0.2$ and $P(B|A) = 0.15$. By the multiplication rule,

$$P(A \cap B) = P(A)P(B|A) = (0.2)(0.15) = 0.03$$

The answer is A.

15. The boxplots below show the effects of different sugars on the growth of pea sections grown in tissue culture, measured in ocular units. (An ocular unit is 0.114 cm.) In experiment A, 2% of glucose was added to the culture. In experiment B, 2% of sucrose was added to the culture. In experiment C, 1% of glucose and 2% of fructose was added to the



culture. Finally, in experiment D, 1% of fructose was added to the culture.

Which one of the following statements is correct? (Only one statement is correct.)

A) The median growth in experiments C and D is the same.
B) The data in experiments A and C have the same inter-quartile range.
C) There are outliers in the data of experiments A, C and D, but not in experiment B.
D) The distribution of the data in experiment B is approximately symmetric.
E) Experiment B has produced the smallest growth.

Solution: The answer is A.

16. One of the objectives of a study is to describe the distribution of the body mass index (BMI) for women whose age is between 20 and 29 years. Suppose that women in this age group have an average BMI of 26.8 with a standard deviation of 7.42. Consider a random sample of 50 women in this age group. Give an approximation for the probability that the average BMI for these 50 women is greater than 29.

A) 0.0179 B) 0.9821 C) 0.6179 D) 0.3821 E) 0.0375

Solution: Let $\overline{X}$ be the mean of this sample. By the central limit

theorem, we know that the random variable

$$\frac{\overline{X} - 26.8}{7.42/\sqrt{50}}$$

has approximatively a standard normal distribution. Hence,

$$\begin{aligned} P(\overline{X} > 29) &= P\left(\frac{\overline{X} - 26.8}{7.42/\sqrt{50}} > \frac{29 - 26.8}{7.42/\sqrt{50}}\right) \approx P(Z > 2.10) \\ &= 1 - P(Z < 2.10) = 1 - 0.9821 = 0.0179. \end{aligned}$$

The answer is A.

17. A pharmaceutical company is testing a new analgesic (medication for pain relief) on a sample of 6 patients suffering from migraine. Among these, 4 patients reported that their migraines disappeared after using the drug. However, it is known that 20% of migraines disappear anyways without any treatment. What is the probability that in a sample of 6 patients suffering from migraine, the migraines will disappear without any treatment for exactly 4 them?

A) 0.0016 B) 0.2534 C) 0.3523 D) 0.0154 E) 0.9992

Solution: Let X be the number of patients for whom the migraine will disappear without any treatment, in a sample of 6 patients. Then X has a binomial distribution with $n = 6$ trials and probability $p = 0.2$ of success. The desired probability is

$$P(X = 4) = \binom{6}{4} (0.2)^4 (0.8)^2 = 0.01536$$

The answer is D.

18. Approximately 4% of men with age between 40 and 55 years will have a heart attack in a 5-year period. A new drug was developed to reduce the probability of having a heart attack for men in this age group. A 5-year study was conducted involving men in this age group who have been treated with the new drug. Among the 2046 participants in the study, 56 had a heart attack within the 5-year period. Let p be the proportion of men in the age group 40-55 using this drug who will have

a heart attack. Give a 95% confidence interval (c.i.) for p. Using this interval, can we conclude that the new drug is efficient in reducing the risk of having a heart attack for men in this age group?

A) c.i.=[0.020; 0.034]. The new drug is not efficient.
B) c.i.=[0.020; 0.034]. The new drug is efficient.
C) c.i.=[0.014; 0.040]. The new drug is efficient.
D) c.i.=[0.041; 0.052]. The new drug is not efficient.
E) c.i.=[0.020; 0.056]. Using this data, we cannot draw any conclusion about the efficiency of the new drug.

Solution: An estimate for p is $\hat{p} = 56/2046 = 0.02737$. The 95% confidence interval for p is:

$$0.02737 \pm 1.96\sqrt{\frac{(0.02737)(1-0.02737)}{2046}} = [0.020; 0.034].$$

Because all the values in the interval are smaller than 0.04, we are confident that p is smaller than 0.04. We conclude that the new drug is efficient in reducing the risk of a heart attack. The answer is B.

19. Continue with the situation in Problem 18. Formulate a null hypothesis H_0 and an alternative hypothesis H_1 which could be used for testing that the new drug is efficient in reducing the risk of having a heart attack for men in this age group. Calculate the p-value of this test.

A) $H_0 : p = 0.04$ against $H_1 : p > 0.04$. The p-value is 0.9982.
B) $H_0 : p = 0.04$ against $H_1 : p < 0.04$. The p-value is 0.0036.
C) $H_0 : p = 0.04$ against $H_1 : p < 0.04$. The p-value is 0.0018.
D) $H_0 : p = 0.04$ against $H_1 : p \neq 0.04$. The p-value is 0.0036.
E) $H_0 : p = 0.04$ against $H_1 : p < 0.04$. The p-value is 0.0154.

Solution: We would like to test $H_0 : p = 0.04$ against $H_1 : p < 0.04$. The observed value of the test statistic is:

$$z_0 = \frac{\hat{p} - 0.04}{\sqrt{(0.04)(0.96)/2046}} = -2.92,$$

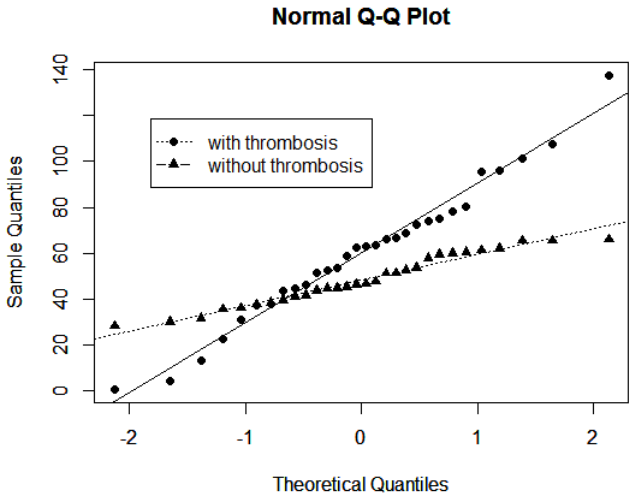
where $\hat{p} = 56/2046 = 0.02737$. This is a left-tailed test. Using Table 17.2, we see that the p-value of the test is given by

$$p\text{-value} = P(Z < -2.92) = 0.0018.$$

The answer is C.

20. The focus of a study was to determine if there were different concentrations of the anticardiolipin antibody IgG (in mg/dl) in subjects with and without thrombosis. In R, we assigned the concentration of IgG for the 30 subjects with thrombosis to the variable x, while the concentration of IgG for the 30 subjects without thrombosis was assigned to the variable y. We produced the summary statistics for both variables and the overlayed quantile-quantile plots.

```
> summary(x)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  0.527  43.850  62.680  60.240  74.840 137.600
> summary(y)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  28.01   39.81   46.48   48.14   58.94   66.06
```



We want to use the `t.test()` function to verify that the mean concentration of IgG for patients with thrombosis is different than the mean concentration of IgG for patients without thrombosis. Which one of the following statements is correct? (Only one statement is correct.)

A) The two samples obviously come from non-normal populations, so we should not use the `t.test` function in `R`.
B) It is reasonable to assume that we have normal populations and the correct `R` command is `t.test(x,y)`.
C) It is reasonable to assume that we have normal populations and the correct `R` command is `t.test(x,y,var.equal=TRUE)`.
D) It is reasonable to assume that we have normal populations and the correct `R` command is `t.test(x,y,alternative="greater")`.
E) It is reasonable to assume that we have normal populations and the correct `R` command is `t.test(x,y,paired="TRUE")`.

Solution: There are linear tendencies in both quantile-quantile plots, so it is reasonable to assume that the populations are normal. So we can use the `t.test()` function to compare the means. So A) is incorrect. The slopes of the lines in the plots are very different, so it is not reasonable to assume the equality of variances. So C) is incorrect. We do not have paired measurements, so E) is incorrect. We want to test the null hypothesis of equality of means against an alternative of differing means. So D) is incorrect. The answer is B).

21. An experimental cancer vaccine has been developed to reduce the tumor size. We would like to test the null hypothesis that the vaccine has no effect on the tumor size, against the alternative hypothesis that the vaccine is effective in reducing the tumor size. Denote by μ_D the mean difference between the tumor size before the vaccine and the tumor size 3 months after the vaccine. Set-up a test of hypotheses and explain when type I error or type II error occur by choosing the correct statement from the list below. (Only one statement is correct.)

A) $H_0 : \mu_D = 0$ versus $H_1 : \mu_D > 0$. Type II error occurs when we decide that the vaccine is effective in reducing the tumor size, when in fact it is not.
B) $H_0 : \mu_D = 0$ versus $H_1 : \mu_D < 0$. Type I error occurs when we decide that vaccine is effective in reducing the tumor size, when in fact it is not.
C) $H_0 : \mu_D = 0$ versus $H_1 : \mu_D > 0$. Type I error occurs when we decide that the vaccine is not effective in reducing the tumor size, when in fact it is.

D) $H_0 : \mu_D = 0$ versus $H_1 : \mu_D \neq 0$. Type I error occurs when we decide that the vaccine is effective in reducing the tumor size, when in fact it is not.
E) $H_0 : \mu_D = 0$ versus $H_1 : \mu_D > 0$. Type II error occurs when we decide that the vaccine is not effective in reducing the tumor size, when in fact it is.

Solution: The hypotheses to be tested are: $H_0 : \mu_D = 0$ versus $H_1 : \mu_D > 0$. Type I error occurs when we decide that the vaccine is effective in reducing the tumor size, when in fact it is not. Type II error occurs when we decide that the vaccine is not effective in reducing the tumor size, when in fact it is. The answer is E.

22. A study is conducted to investigate the relationship between the number X of hours of exercise per week and the systolic blood pressure Y for men of age 50. The following data was obtained on 10 individuals:

Number of hours x_i	Systolic blood pressure y_i	x_i^2	$x_i y_i$
4	120	16	480
10	110	100	1100
2	120	4	240
3	135	9	405
3	140	9	420
5	115	25	575
1	150	1	150
2	165	4	330
2	160	4	320
0	180	0	0

For this data, we have:

$$\sum_{i=1}^{10} x_i = 32, \quad \sum_{i=1}^{10} y_i = 1395, \quad \sum_{i=1}^{10} x_i^2 = 172, \quad \sum_{i=1}^{10} x_i y_i = 4020$$

Give the equation of the estimated regression line for this data and an estimate for the average systolic blood pressure of an individual who exercises 6 hours per week.

A) The estimated regression line is $\hat{y} = 181.19 - (8.60)x$. An estimate

for the average systolic blood pressure of an individual who exercises 6 hours per week is 130.
B) The estimated regression line is $\hat{y} = 167.31 - (7.12)x$. An estimate for the average systolic blood pressure of an individual who exercises 6 hours per week is 125.
C) The estimated regression line is $\hat{y} = 138.90 - (5.25)x$. An estimate for the average systolic blood pressure of an individual who exercises 6 hours per week is 107.
D) The estimated regression line is $\hat{y} = 145.76 - (3.42)x$. An estimate for the average systolic blood pressure of an individual who exercises 6 hours per week is 125.
E) The estimated regression line is $\hat{y} = 159.91 - (6.38)x$. An estimate for the average systolic blood pressure of an individual who exercises 6 hours per week is 122.

Solution: For calculating $\hat{\beta}$ and $\hat{\alpha}$, we use the formulas:

$$\hat{\beta} = \frac{n\sum_{i=1}^{n} x_i y_i - (\sum_{i=1}^{n} x_i)(\sum_{i=1}^{n} y_i)}{n\sum_{i=1}^{n} x_i^2 - (\sum_{i=1}^{n} x_i)^2} \quad \text{and} \quad \hat{\alpha} = \bar{y} - \hat{\beta}\bar{x}.$$

We obtain:

$$\hat{\beta} = \frac{10(4020) - (32)(1395)}{10(172) - (32)^2} = -6.37931, \quad \hat{\alpha} = \frac{1395}{10} + (6.37931)\frac{32}{10} = 159.9138$$

The estimated regression line is $\hat{y} = \hat{\alpha} + \hat{\beta}x$, which in our case becomes:

$$\hat{y} = 159.91 - (6.38)x$$

An estimate for the average systolic blood pressure of an individual who exercises 6 hours per week is

$$\hat{\mu}_{Y|x=6} = 159.91 - (6.38)(6) = 121.6.$$

The answer is E.

23. An important hypothesis in hypertension research is that sodium reduction may lower blood pressure. Since sodium restriction is difficult to maintain during a long period of time, dietary counseling in a group setting is sometimes used to achieve this goal. The data on urinary

sodium levels were obtained on 8 individuals enrolled in a sodium-restricted group. Data were collected at baseline (i.e. at the beginning of the study) and after one week of dietary counseling. The data is as follows:

```
> Week0=c(7.85,12.03,21.84,13.94,16.68,41.78,14.97,12.07)
> Week1=c(5.59,8.5,4.55,12.78,10.69,23.51,5.46,11.95)
```

Researcher 1 types into the R console

```
> t.test(Week0,Week1,alternative="greater")
```

and obtains the following output:

```
data:  Week0 and Week1
t = 1.6814, df = 11.267, p-value = 0.06008
alternative hypothesis: true difference in means is greater than 0
95 percent confidence interval:
 -0.4779779        Inf
sample estimates:
mean of x mean of y
 17.64500  10.37875
```

Researcher 2 types into the R console:

```
> t.test(Week0,Week1,paired=TRUE,alternative="greater")
```

and obtains the following output:

```
data:  Week0 and Week1
t = 2.8835, df = 7, p-value = 0.01177
alternative hypothesis: true difference in means is greater than 0
95 percent confidence interval:
 2.492072      Inf
sample estimates:
mean of the differences
                7.26625
```

Suppose that both researchers verified that the data is normally distributed before using the function `t.test`. Which one of the following statements is correct? (Only one statement is correct.)
A) Researcher 1 uses the correct command. When $\alpha = 0.05$, there is enough evidence that dietary counseling leads to a reduction in sodium levels.
B) Researcher 2 uses the correct command. When $\alpha = 0.05$, there is enough evidence that dietary counseling leads to a reduction in sodium levels.
C) Researcher 1 uses the correct command. When $\alpha = 0.05$, there is not enough evidence that dietary counseling leads to a reduction in sodium levels.
D) Researcher 2 uses the correct command. When $\alpha = 0.05$, there is not enough evidence that dietary counseling leads to a reduction in sodium levels.
E) Neither one of the researchers are using the correct command. The t-test should not be used in this situation.

Solution: These are paired data. Researcher 2 uses the correct command. Let μ_X be the average blood pressure at baseline, and μ_Y be the average blood pressure after one week. Since in the output of researcher 2, the p-value is smaller than 0.05, we reject $H_0 : \mu_X = \mu_Y$ in favor of $H_1 : \mu_X > \mu_Y$. We conclude that dietary counselling leads to a reduction in the sodium levels. The answer is B.

24. The plant-water relation plays an important role in plant physiology. We consider an experiment in which 16 seedlings of birch tree were flooded with water for one day and 13 other seedlings were kept as controls. At the end of the experiment, the roots of all plants were analyzed for the level of adenosine triphosphate (ATP), as a measure for the intracellular energy transfer. Below is the summary of the data:

	flooded plants	control plants
sample size	$n_1 = 16$	$n_2 = 13$
sample mean	$\bar{x}_1 = 1.17$	$\bar{x}_2 = 1.91$
sample standard deviation	$s_1 = 0.16$	$s_2 = 0.23$

Give a 90% confidence interval for the difference $\mu_1 - \mu_2$, where μ_1 is the average ATP level for the flooded plants and μ_2 is the average ATP

level for the controls. Based on this interval, can we conclude that flooding causes a decrease or an increase in the ATP level? (Assume that the ATP levels for flooded plants and controls are normally distributed with equal variances.)

A) [0.5673; 0.7614]; flooding causes an increase in the mean ATP level
B) [0.4532; 0.6719]; flooding causes an increase in the mean ATP level
C) [-0.6182; -0.4820]; flooding causes a decrease in the mean ATP level
D) [-0.8635; -0.6165]; flooding causes a decrease in the mean ATP level
E) [-0.0346; 0.3471]; we cannot conclude that flooding causes a decrease or an increase in the mean ATP level

Solution: The pooled sample variance is:

$$s_p^2 = \frac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{n_1+n_2-2} = \frac{(15)(0.16)^2 + (12)(0.23)^2}{16+13-2} = 0.03773$$

The 90% confidence interval for $\mu_1 - \mu_2$ is

$$\bar{x}_1 - \bar{x}_2 \pm t\sqrt{s_p^2(1/n_1 + 1/n_2)}$$

The value t is found in Table 17.4 such that $P(-t \leq T \leq t) = 0.90$, where T has a T distribution with 27 degrees of freedom. This means that $P(T \leq t) = 0.95$. In Table 17.4 (row 27, column 0.95) we find the value $t = 1.703$. The 90% confidence interval for $\mu_1 - \mu_2$ is:

$$1.17 - 1.91 \pm (1.703)\sqrt{(0.03773)(1/16 + 1/13)} =$$

$$-0.74 \pm 0.1235 = [-0.8635; -0.6165]$$

Since the interval contains only negative values, we infer that $\mu_1 < \mu_2$. We conclude that flooding causes a decrease in the ATP level. The answer is D.

25. The systolic blood pressure level in a certain population is approximately equal to the value 125 mm Hg. A topic of recent clinical interest is the fact that extensive use of oral contraceptive (OC) may cause a reduction in the systolic blood pressure under the value 125. A study is organized to test this hypothesis. The n women who participated in this study used OC for a period of 3 months. At the end of the study, their systolic blood pressure was measured. The summary of the data and the R output of the test are given below:

```
> mean(x)
[1] 120.4
> sd(x)
[1] 13.23
> t.test(x,mu=125,alternative="less")

        One Sample t-test

data:  x
t = -1.0998, df =-, p-value = 0.15
alternative hypothesis: true mean is less than 125
95 percent confidence interval:
    -Inf 128.067
sample estimates:
mean of x
    120.4
```

What was the number n of participants in this study?

A) 12 B) 40 C) 10 D) 32 E) 25

Solution: We would like to test $H_0 : \mu = 125$ against $H_1 : \mu < 125$. The observed value of test statistic is

$$t_0 = \frac{\bar{x} - 125}{s/\sqrt{n}} = \frac{120.4 - 125}{13.23/\sqrt{n}}.$$

From the R output we know that $t_0 = -1.0998$. We infer that

$$\frac{120.4 - 125}{13.23/\sqrt{n}} = -1.0998.$$

Therefore

$$n = \left(\frac{-1.0998 \times 13.23}{120.4 - 125}\right)^2 = 10.005$$

We conclude that the sample size was $n = 10$. The answer is C.