I.A: Levels of Data, Types of Samples

I.A.1. Categorize the following data according to level: nominal, ordinal, interval, or ratio.

(a) Time of first class.
(b) Length of time to complete an exam.
(c) Course evaluation scale: poor, acceptable, good.
(d) The length of time it takes for someone to run a marathon.
(e) The time of day the marathon starts.
(f) The name of the city in which the marathon is run.

**Answer.**
(a) Interval—differences in time are meaningful, but ratios are not. For example, a first class at 3:00pm is not 1.5 times later than a first class at 2:00pm.
(b) Ratio—differences in time make sense as do ratios. For example, if Student A takes 50 minutes and Student B takes 100 minutes, it makes sense to say that Student B took twice as long as Student A.
(c) Ordinal—the categories can be ranked, but differences between ranks do not make sense.
(d) Ratio—similar to (b).
(e) Interval—similar to (a)
(f) Nominal—categories cannot even be ranked.

I.A.2. To estimate the average GPA of all La Sierra Students, President Geraty computed the average GPA obtained in his Advanced Hebrew Grammar class.

(a) Identify the variable?
(b) What is the implied population?
(c) What is the sample?
(d) What type of sample was this?

**Answer.**
(a) GPA
(b) The GPA’s of all La Sierra University Students.
(c) The GPA’s of all students in Dr. Geraty’s Advanced Hebrew Grammar class.
(d) This is a sample of convenience.

I.A.3. (a) If your instructor were to compute the class mean of this test when it is graded, and use it to estimate the average for all tests taken by this class this quarter, would this be an example of descriptive or inferential statistics? Explain.
(b) A study on attitudes about smoking is conducted at a college. The students are divided by class, and then a random sample is selected from each class. What type of sampling technique is this (e.g. simple random, convenient, stratified, systematic, cluster)? Explain why this type of sample is not a simple random sample.

(c) A politician wishes to determine the reading level of 5th graders in her State. She does not have funding to test all 5th graders in her state, so she randomly selects some of the schools in her state and tests every 5th grader in those schools. What type of a sample is this?

Answer. (a) Inferential: using a sample mean (one test) to estimate the population (all tests, quizzes and assignments) mean.

(b) Stratified: a random sample is taken from each class (strata). This cannot be a simple random sample because it requires elements from each class. A simple random sample need not have representation from each class.

(c) Cluster: the population is divided into groups, some of the groups are selected randomly and everyone in those groups is tested.

I.A.4. To compute the average amount of medical insurance its patients had in 2006, a hospital considered taking the following types of samples. For each case classify the sample as simple random, stratified, systematic, cluster, convenient.

(a) The hospital numbered its complete list of patients from 2006 and used a random number generator to select 100 patients from the list from which to collect the data.

(b) The hospital decided to collect the data from the first 50 patients admitted on July 4, 2006.

(c) The hospital randomly chose a patient, and collected data from that patient and every 200th patient admitted thereafter.

Answer. (a) Simple random; (b) convenient; (c) systematic.

I.A.5. Explain how you could use the table of random numbers in your text to help design a true false test of 10 questions so that the pattern of answers is random.

Answer. Randomly select a starting spot in the table. If the digit is odd, make a question with a false answer, if the digit is even make a question with a true answer. Proceed along the row for 10 such digits. For example, if the starting point had been the beginning of the 3rd row, the digits are: 59654 71966 which leads to answers of F F T F T F F F T T

I.A.6. (a) Describe the difference between population data and sample data.

(b) Describe the difference between inferential and descriptive statistics.

(c) Describe the difference between quantitative and qualitative variables.

Answer. See text.

I.B: Organizing and Presenting Data

I.B.1. Consider the data (which are systolic blood pressures of 50 subjects):

(a) What class width should be chosen if you would like to have 6 classes.

(b) Suppose you don’t want a class width of 19, but would like a class width of 15 irrespective of how many classes that would give you. Complete the following table for this data given the first class has limits 100 – 114.

(c) Draw a frequency histogram using the table in (b).

(d) Draw a frequency polygon using the table in (b).

(e) Draw an Ogive using the table in (b).

(f) Draw a relative frequency histogram for the table in (b).

(g) Describe in words how to construct each of the above-mentioned types of graphs.

**Answer.**  (a) $\frac{(208 - 100)}{6} = 18$. Go to the next higher whole number to ensure that all of the data is covered. Thus a class width of 19 would be suitable.

(b)

<table>
<thead>
<tr>
<th>Lower Limit</th>
<th>Upper Limit</th>
<th>Lower Boundary</th>
<th>Upper Boundary</th>
<th>Midpoint</th>
<th>Frequency</th>
<th>Cumulative Frequency</th>
<th>Relative Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>100</td>
<td>114</td>
<td>99.5</td>
<td>114.5</td>
<td>107</td>
<td>10</td>
<td>10</td>
<td>.20</td>
</tr>
<tr>
<td>115</td>
<td>129</td>
<td>114.5</td>
<td>129.5</td>
<td>122</td>
<td>15</td>
<td>25</td>
<td>.30</td>
</tr>
<tr>
<td>130</td>
<td>144</td>
<td>129.5</td>
<td>144.5</td>
<td>137</td>
<td>14</td>
<td>39</td>
<td>.28</td>
</tr>
<tr>
<td>145</td>
<td>159</td>
<td>144.5</td>
<td>159.5</td>
<td>152</td>
<td>6</td>
<td>45</td>
<td>.12</td>
</tr>
<tr>
<td>160</td>
<td>174</td>
<td>159.5</td>
<td>174.5</td>
<td>167</td>
<td>1</td>
<td>46</td>
<td>.02</td>
</tr>
<tr>
<td>175</td>
<td>189</td>
<td>174.5</td>
<td>189.5</td>
<td>182</td>
<td>0</td>
<td>46</td>
<td>.00</td>
</tr>
<tr>
<td>190</td>
<td>204</td>
<td>189.5</td>
<td>204.5</td>
<td>197</td>
<td>2</td>
<td>48</td>
<td>.04</td>
</tr>
<tr>
<td>205</td>
<td>219</td>
<td>204.5</td>
<td>219.5</td>
<td>212</td>
<td>2</td>
<td>50</td>
<td>.04</td>
</tr>
</tbody>
</table>

(c) — (g): For (c) — (f), see answers to In-class exercises, April 6 in Spring 2007. For (g), see your text for the precise descriptions.

**I.B.2** Consider the collection of 30 data

2 3 4 4 5 6 8 9 11 12 14 15 18 19 21
23 23 25 27 34 40 45 51 56 58 62 73 83 91 98

Complete the following table given that the first class has limits 1 – 20.

**Answer.**
### I.B.3 Consider the following table

<table>
<thead>
<tr>
<th>Lower Limit</th>
<th>Upper Limit</th>
<th>Lower Boundary</th>
<th>Upper Boundary</th>
<th>Midpoint</th>
<th>Frequency</th>
<th>Cumulative Frequency</th>
<th>Relative Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>20</td>
<td>.5</td>
<td>20.5</td>
<td>10.5</td>
<td>14</td>
<td>14</td>
<td>14/30 ≈ .467</td>
</tr>
<tr>
<td>21</td>
<td>40</td>
<td>20.5</td>
<td>40.5</td>
<td>30.5</td>
<td>7</td>
<td>21</td>
<td>7/30 ≈ .233</td>
</tr>
<tr>
<td>41</td>
<td>60</td>
<td>40.5</td>
<td>60.5</td>
<td>50.5</td>
<td>4</td>
<td>25</td>
<td>4/30 ≈ .133</td>
</tr>
<tr>
<td>61</td>
<td>80</td>
<td>60.5</td>
<td>80.5</td>
<td>70.5</td>
<td>2</td>
<td>27</td>
<td>2/30 ≈ .067</td>
</tr>
<tr>
<td>81</td>
<td>100</td>
<td>80.5</td>
<td>100.5</td>
<td>90.5</td>
<td>3</td>
<td>30</td>
<td>3/30 = .100</td>
</tr>
</tbody>
</table>

(a) Construct a relative frequency histogram using the information given in the table above.
(b) Construct an ogive using the information given in the table above.
(c) Construct a frequency histogram using the information given in the table above.
(d) Construct a frequency polygon using the information given in the table above.

**Answer.** For (a) and (b) see answers to Question 4, Test I, Winter 2007. For (c), the difference between a relative frequency histogram and a frequency histogram is that for the latter the heights of the bars are the relative frequencies of the class.

### I.B.4. Make a stem and leaf display for the following data.

58 52 68 86 72 66 97 89 84 91 91 92 66 68 87 86
73 61 70 75 72 73 85 84 90 57 77 76 84 93 58 47

**Answer.**

```
<table>
<thead>
<tr>
<th>Stem</th>
<th>Leaf</th>
</tr>
</thead>
<tbody>
<tr>
<td>4</td>
<td>7</td>
</tr>
<tr>
<td>5</td>
<td>2 7 8 8</td>
</tr>
<tr>
<td>6</td>
<td>1 6 6 8 8</td>
</tr>
<tr>
<td>7</td>
<td>0 2 2 3 3 5 6 7</td>
</tr>
<tr>
<td>8</td>
<td>4 4 4 5 6 6 7 9</td>
</tr>
<tr>
<td>9</td>
<td>0 1 1 2 3 7</td>
</tr>
</tbody>
</table>
```

### I.C Numerical Representations of Data
I.C.1. Consider the following sample consisting of 20 numbers.

\[
\begin{array}{cccccccccccc}
20 & 23 & 24 & 24 & 25 & 26 & 28 & 29 & 31 & 32 \\
33 & 34 & 35 & 37 & 41 & 46 & 52 & 53 & 58 & 98 \\
\end{array}
\]

(a) Given that \( \sum x = 749 \) and \( \sum x^2 = 34109 \) find the mean, variance and standard deviation for this sample.

(b) Find the \( Q_1, Q_2, Q_3 \) and the IQR for the data and then construct a box-and-whisker plot for the data.

**Answer.** (a) The mean is \( \bar{x} = \frac{749}{20} = 37.45 \). Because this is a sample, the variances is

\[
s^2 = \frac{34109 - \frac{749^2}{20}}{19} \approx 318.892
\]

and the standard deviations is \( s = \sqrt{s^2} \approx 17.858 \).

(b) The quartiles are \( Q_1 = 25.5 \), \( Q_2 = 32.5 \), and \( Q_3 = 43.5 \) and the inter quartile range is \( IQR = Q_3 - Q_1 = 18 \). For a sketch of the box plot see answers to Test 1, Winter 2007, Question 5.

I.C.2. 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;  (b) the sample variance  (c) the population variance

(d) the sample standard deviation (e) the median  (f) the 50th percentile (g) the mode

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

(b) the sample variance is \( s^2 = \frac{5206.917}{11} = 473.36 \).

(c) the population variance is \( \sigma^2 = \frac{5206.917}{12} = 433.90915 \).

(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.

I.C.3 The depth of ground water is given in the following grouped data table.

<table>
<thead>
<tr>
<th>Distance from ground to water level (ft), ( x )</th>
<th>15 – 19</th>
<th>20 – 24</th>
<th>25 – 29</th>
<th>30 – 34</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of wells, ( f )</td>
<td>3</td>
<td>5</td>
<td>8</td>
<td>4</td>
</tr>
</tbody>
</table>

(a) Estimate the mean depth of the ground water.

(b) Estimate the sample standard deviation for the depth of the ground water.

(c) Estimate the coefficient of variation for this data.
Answer.  (a) \( \bar{x} \approx \frac{\sum x f}{n} \), where \( n = \sum f \) and so
\[
\bar{x} \approx \frac{(17)(3) + (22)(5) + (27)(8) + (32)(4)}{20} = \frac{505}{20} = 25.25.
\]
(b) \( s \approx \sqrt{\frac{\sum x^2 f - (\sum x f)^2/n}{n-1}} \), and \( \sum x^2 f = (17^2)(3) + (22^2)(5) + (27^2)(8) + (32^2)(4) = 13215 \), and so
\[
s \approx \sqrt{\frac{13215 - \frac{505^2}{20}}{19}} \approx \sqrt{24.4079} \approx 4.940
\]
(c) C.V. = \( \frac{s}{\bar{x}} \cdot 100\% \approx \frac{4.940}{25.25} \cdot 100\% \approx 19.57\% \)

I.C.4 (Short answer general) (a) Given a collection of data with lowest number 4 and highest number 100. What class width should be chosen if 6 classes are desired?
(b) Given a collection of ordered data with 65 numbers, in what position is the median?
(c) What percentile is the third quartile \( Q_3 \)?
(d) In a set of data, approximately what percentage of the data lie at or above the 63rd percentile?
(e) If you are among 8000 people that took a test and you scored at the 79th percentile, approximately how many people scored at your score or lower? Approximately how many people scored at your score or higher?
(f) If you are among 4000 students taking the MCAT, and you wish to score at least the 95th percentile, what is the maximum number of students that can score at least as well or better than you?

Answer.  (a) \( (100 - 4)/6 = 16 \), and we use the next higher whole number, so the class width should be 17.
(b) The median is in the 33rd position.
(c) The 75th percentile.
(d) 37% of the data.
(e) Approximately \( (.79)(8000) = 6320 \) scores were less than or equal to your score, and approximately 1680 were greater than or equal to your score.
(f) Not more than 5% of 4000, or 200 can score as well or better than you if you are to achieve the 95th percentile, or better.

I.C.5.  A population is known to have a mean of 50 and a standard deviation of 6.
(a) Use Chebyshev’s theorem to find an interval that contains at least 75% of the data.
(b) Use Chebyshev’s theorem to find an interval that contains at least 24/25 of the data.
(c) At least what portion of data is contained in the interval from 26 to 74?

Answer.  (a) This is the data within two standard deviations of the mean, hence the interval 50 ± 2(6), i.e. from 38 to 62.
(b) This is the data within 5 standard deviations of the mean, hence the interval from 20 to 80.

(c) The interval from 26 to 74 includes data that is within 4 standard deviations of the mean. According to Chebyshev’s theorem, this must include at least \(1 - \frac{1}{4^2} = 15/16\) of the data, that is, at least 93.75% of the data.

I.C.6. Consider the following data of 26 numbers.

\[
8, 35, 47, 48, 51, 57, 60, 64, 65, 66, 70, 72, 76, 78, 80, 82, 84, 85, 89, 90, 90, 93, 94, 96, 111
\]

(a) Find the median of the data.

(b) Find Q1, Q3 and the IQR. Construct a box and whisker plot for the data.

(c) Compute the interval \((Q1 - 1.5 \cdot IQR, Q3 + 1.5 \cdot IQR)\). Data outside of this interval are identified as suspected outliers. Are there any suspected outliers in the above data?

**Answer.**

(a) Because 26 is even, the median is the average of the \(26/2 = 13\)th place and the 14th place, therefore the median is \((72 + 76)/2 = 74\)

(b) The first quartile is the median of the 13 numbers below the median 74 of the entire set. Hence the first quartile is \(Q_1 = 60\) and the third quartile is the median of the 13 numbers above 74. Therefore the third quartile is \(Q_3 = 89\).

The IQR is the inter quartile range, \(IQR = Q_3 - Q_1 = 8960 = 29\).

See section 3.4 in the text for construction of the box and whisker plot. Note the lower whisker will go down to 33, the bottom of the box will start at 60, the line in the box will be at 74, the top of the box will be at 89, the upper whisker will go to 97.

(c) The interval is \((16.6, 132.5)\), so 8 is a suspected outlier.

I.C.7. (General Questions on Means) (a) A student receives grades of A, A, B, C, C and is surprised to receive a GPA of 3.5 because the average grade of A, A, B, C, C is a B since

\[
\frac{4 + 4 + 3 + 2 + 2}{5} = \frac{15}{5} = 3.
\]

Explain how the GPA could be 3.5.

(b) Bob is pleased to learn from his boss that his annual salary is in the top 10 percent of all salaries in the company. A month later Bob learns that his salary is less than half of the mean salary in the company and suspects his boss lied to him. Explain how Bob’s boss could have told the truth.

(c) An army crosses a river whose average depth is 1 foot. Explain how several soldiers could have drowned crossing the river because they cannot swim.

(d) Ken had an average of 50% on exams and 90% on assignments in his class, so he computed that his average should be 70% (a C). Why was he shocked when he saw that his grade was a D?

**Answer.**

(a) GPA’s are weighted averages. For example if the A’s were received in 5 unit classes and the B was received in a 4 unit class and the C’s were received in 1 unit classes, then the GPA is computed as follows.

\[
P = \frac{\sum xw}{\sum w} = \frac{4(5) + 4(5) + 3(4) + 2(1) + 2(1)}{5 + 5 + 4 + 1 + 1} = \frac{56}{16} = 3.5.
\]
(b) For example consider a company with 20 employees whose annual salaries in thousands are
\[
\begin{align*}
20 &
20 &
25 &
25 &
30 &
30 &
30 &
30 &
40 &
40 &
40 &
40 &
40 &
40 &
45 &
45 &
45 &
45 &
50 &
50 &
2000
\end{align*}
\]
where Bob’s salary is 60K per year. The average annual salary is
\[
\frac{\sum x}{20} = \frac{2680}{20} = 134K
\]
which is more than double Bob’s salary. Notice the difference if you computed a 5% trimmed mean.

(c) (Hint) Think of a wide river that is 6 inches deep for the majority of the width and then has a deep but rather narrow trench. (An unlikely possibility is that the river is 1 foot deep all the way across and the soldiers that could not swim were not in good shape and ended up falling face first in the water because of tiredness.)

(d) (Hint) The exams must have a higher weighting. For example, what is his average if the exams are worth 80% of his overall grade, and the assignments are worth 20%? (An unlikely possibility is that Ken wasn’t paying attention, and in that class 70% is a D.)

I.C.8. (General Questions on Measures of Data) Consider a data set of 25 distinct measurements with mean A, median B, and range R.

(a) If the largest number is increased by 250, what is the new mean?
(b) If the largest number is increased by 250, what is the new median?
(c) If the largest number is increased by 250, what is the new range?
(d) If every number is increased by 10, what is the effect on the mean, median, range and standard deviation?
(e) In what position of the ordered data is the median?
(f) Explain why the data does not have a mode.
(g) Is it possible for the mean and median to be equal?

Answer. (a) the mean is increased by 10.
(b) The median remains the same.
(c) The range increases by 250.
(d) The mean and median increase by 10, then range and standard deviation remain the same.
(e) The median is in the 13th place.
(f) All numbers are distinct, so there is no most frequently occurring number.
(g) Yes, for example let the numbers be 1, 2, 3, \ldots, 24, 25. Then the median and mean are both 13.

I.C.9. (Weighed Average) Suppose the weightings in a class are such that the tests are worth 60% of the grade, the assignments are worth 10% of the grade, the quizzes are worth 5% of the grade, and the final is worth 25% of the grade.

(a) Calculate the overall grade of a student that has 90% on assignments, 95% on quizzes, 75% on tests and 80% on the final exam.
(b) Is the answer in (a) higher or lower than if you had taken the average of 90%, 95%, 75% and 80%? Explain why in this example you would expect the weighted average to be lower.

(c) Suppose a student has 90% on assignments, 95% on quizzes, 75% on tests going into the final exam. What percentage is needed on the final exam for the student to finish with an overall grade of 82% in the class?

**Answer.** (a) \[0.1(90) + 0.05(95) + 0.6(75) + 0.25(80) = 78.75\% \] (Note in this case the weights were used as 0.1, 0.05, 0.60 and 0.25 so they sum to 1).

(b) The weighted average is lower because the lower scores were in the categories that were weighted more heavily.

(c) Let \( x \) be the needed percentage; then solve

\[0.1(90) + 0.05(95) + 0.6(75) + 0.25x = 82\]

to get \(58.75 + 0.25x = 82\), and so \(0.25x = 82 - 58.75 = 23.25\), or \(x = 4(23.25) = 93\). Thus, 93% is needed on the final exam.

**I.C.10 On-time percentages are given for two airlines in Phoenix, Los Angeles and Seattle for 2006.**

<table>
<thead>
<tr>
<th>Crashcade Airlines</th>
<th>Los Angeles</th>
<th>Phoenix</th>
<th>Seattle</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of Fights</td>
<td>1000</td>
<td>500</td>
<td>3500</td>
</tr>
<tr>
<td>On time %</td>
<td>90</td>
<td>95</td>
<td>85</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Pacific Worst Airlines</th>
<th>Los Angeles</th>
<th>Phoenix</th>
<th>Seattle</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of Fights</td>
<td>250</td>
<td>4500</td>
<td>250</td>
</tr>
<tr>
<td>On time %</td>
<td>85</td>
<td>90</td>
<td>80</td>
</tr>
</tbody>
</table>

(a) Calculate the on-time percentage average for these three cities for each airline. Do this as a weighted average where the weight for each airline and city is the number of flights.

(b) Given that the on-time percentage for Crashcade Airlines is 5% higher in each city, does the answer in (a) surprise you? Why or why not?

**Answer.** (a) For Crashcade we compute

\[
\frac{\sum xw}{\sum w} = \frac{(1000)(90\%) + (500)(95\%) + (3500)(85\%)}{1000 + 500 + 3500} = \frac{435,000}{5000} = 87\%
\]

For Pacific Worst we compute

\[
\frac{\sum xw}{\sum w} = \frac{(250)(85\%) + (4500)(90\%) + (250)(80\%)}{250 + 4500 + 250} = \frac{446,250}{5000} = 89.25\%
\]

You should get the same answer if you computed this as follows (we illustrate with Crashcade): Out of Los Angeles, 90% of 1000 flights = 900 flights were on time, out of Phoenix, 95% of 500 = 475 flights were on-time, out of Seattle 85% of 3500 = 2975 flights were on time. Therefore, Crashcade had a total of 900 + 475 + 2975 = 4350 out of 5000 flights on time, which is 87%.

(b) On the surface it is very surprising that Pacific Worst has a better overall on-time percentage. However, this happens because Pacific Worst’s schedule is heavily weighted to flights in Phoenix
where they have their best on-time percentage, whereas Crashcade’s flights heavily weighted in Seattle where they have their worst on-time percentage.

**I.C.11.** (See the Histogram in figure 3-4, p. 133 of Text) The following data is for hours of sleep of a random sample of 200 subjects. Estimate the mean hours of sleep, standard deviation hours of sleep and coefficient of variation.

<table>
<thead>
<tr>
<th>Hours of Sleep</th>
<th>Number of Subjects</th>
</tr>
</thead>
<tbody>
<tr>
<td>3.5</td>
<td>2</td>
</tr>
<tr>
<td>4.5</td>
<td>2</td>
</tr>
<tr>
<td>5.5</td>
<td>4</td>
</tr>
<tr>
<td>6.5</td>
<td>22</td>
</tr>
<tr>
<td>7.5</td>
<td>64</td>
</tr>
<tr>
<td>8.5</td>
<td>90</td>
</tr>
<tr>
<td>9.5</td>
<td>14</td>
</tr>
<tr>
<td>10.5</td>
<td>2</td>
</tr>
</tbody>
</table>

**Answer.**

\[
\bar{x} = \frac{\sum f x}{\sum f} = \frac{3.5 \cdot 2 + 4.5 \cdot 2 + \ldots + 9.5 \cdot 14 + 10.5 \cdot 2}{200} = 7.9 \\
\sqrt{\frac{\sum x^2 f}{\sum f} - \left(\frac{\sum f x}{\sum f}\right)^2} = \sqrt{\frac{12702 - \frac{1580^2}{200}}{199}} \approx 1.05
\]

Thus the mean hours of sleep is 7.9 with a standard deviation of 1.05 hours. The coefficient of variation is \(C.V. = \frac{1.05}{7.9} \cdot 100\% = 13.29\%\)

**I.C.12.** Consider the following ordered data of 21 numbers

34 36 40 46 46 48 51 54 57 58 59
60 61 62 63 64 66 70 78 85 101

Find \(Q_1\), \(Q_2\), \(Q_3\), the IQR and then construct a box and whisker plot for the data.

**Answer.** \(Q_1\) is the median of the first 10 data (all data below the position of the median of the entire set, which is 11th position), so \(Q_1 = 47\); \(Q_2\) is the median, so \(Q_2 = 59\) and \(Q_3\) is the median of the last 10 data (all data above the position of the median of the entire set), so \(Q_3 = 65\).

\[
\text{IQR} = Q_3 - Q_1 = 65 - 47 = 18
\]

The box plot then has the lower whisker starting at 34 and ending at the bottom edge of the box which is at 47, the line in the box is at the median, i.e. 59 and the upper edge of the box is at 65. The upper whisker starts there and ends at 101. See text for further examples.
II.A Probability

II.A.1. The following is from a sample of 500 bikers who attended the annual rally in Sturgis South Dakota last August.

<table>
<thead>
<tr>
<th>Beard</th>
<th>No Beard</th>
<th>Row Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Female</td>
<td>0</td>
<td>200</td>
</tr>
<tr>
<td>Male</td>
<td>120</td>
<td>180</td>
</tr>
<tr>
<td>Column Total</td>
<td>120</td>
<td>380</td>
</tr>
</tbody>
</table>

Suppose that a biker is selected at random from the 500 bikers. Denote the events as follows: $B =$ has a beard, $N =$ does not have a beard, $F =$ is Female and $M =$ is male.

(a) Compute $P(B)$.
(b) Compute $P(B|F)$.
(c) Compute $P(M)$
(d) Compute $P(B|M)$.
(e) Compute $P(B$ and $M)$
(f) Are the events $B$ and $F$ independent?
(g) Are the events $B$ and $F$ mutually exclusive?

**Answer.** (a) $120/500 = .240$

(b) $0/200 = 0$

(c) $300/500 = .600$

(d) $120/300 = .400$

(e) $120/500 = .240$

(f) No, because $P(B|F) \neq P(F)$—the probability of someone having a beard is dependent on gender.

(g) Yes, because they cannot occur together, they are therefore mutually exclusive.

II.A.2 The following represents the outcomes of a flu vaccine study.

<table>
<thead>
<tr>
<th></th>
<th>Got the Flu</th>
<th>Did not get Flu</th>
<th>Row Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>No Flu Shot</td>
<td>223</td>
<td>777</td>
<td>1000</td>
</tr>
<tr>
<td>Given Flu Shot</td>
<td>446</td>
<td>1554</td>
<td>2000</td>
</tr>
<tr>
<td>Column Total</td>
<td>669</td>
<td>2331</td>
<td>3000</td>
</tr>
</tbody>
</table>

Let $F$ represent the event the person caught the flu, let $V$ represent the even the person was vaccinated, let $H$ represent the event the person remained healthy (didnt catch the flu), and let $N$ represent the even that the person was not vaccinated.

(a) Compute: $P(F), P(V), P(H), P(N), P(F$ given $V)$ and $P(V$ given $F), P(V$ and $F), P(V$ or $F)$. 
(b) Are the events V and F mutually exclusive? Are the events V and F independent? Explain your answers.

**Answer.** (a) \( P(F) = .223, \ P(V) = 2/3, \ P(H) = .777, \ P(N) = 1/3, \ P(F \text{ given } V) = .223, \ P(V \text{ given } F) = 446/669 = 2/3, \ P(V \text{ and } F) = 446/3000 \approx .149 \) or, by the multiplication rule we get the same answer: \( P(V \text{ and } F) = P(F) \cdot P(V \text{ given } F) = (.223)(2/3) \approx .149 \)

\[ P(V \text{ or } F) = \frac{223 + 446 + 1554}{3000} = .741, \] or by the addition rule: \( P(V \text{ or } F) = P(V) + P(F) - P(V \text{ and } F) \approx .667 + .223 - .149 = .741 \)

(b) \( V \text{ and } F \) are not mutually exclusive because they can both occur together. Another way of saying this is they are not mutually exclusive because \( P(V \text{ and } F) > 0 \). (As a contrast, \( F \text{ and } H \) are mutually exclusive because they cannot occur together.)

\( V \text{ and } F \) are independent because \( P(V \text{ given } F) = P(V) \), and \( P(F \text{ given } V) = P(F) \). What this says is in this hypothetical study, the vaccine had absolute no effect.

**II.A.3.** Class records at Rockwood College indicate that a student selected at random has a probability 0.77 of passing French 101. For the student who passes French 101, the probability is 0.90 that he or she will pass French 102. What is the probability that a student selected at random will pass both French 101 and French 102?

**Answer.** Let A be the even the student passes French 101 and let B be the event the student passes French 102. We need to calculate \( P(A \text{ and } B) \):

\[ P(A \text{ and } B) = P(A) \cdot P(B \text{ given } A) = (.90)(.77) = .693. \]

Thus, 69.3% of all students pass both French 101 and French 102.

**II.A.4** The following data is for U.S. family size

<table>
<thead>
<tr>
<th>x</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7 or more</th>
</tr>
</thead>
<tbody>
<tr>
<td>P(x)</td>
<td>?</td>
<td>.23</td>
<td>.21</td>
<td>.10</td>
<td>.03</td>
<td>.01</td>
</tr>
</tbody>
</table>

(a) What is the probability a family has 5 or more members?

(b) Determine \( P(2) \), the probability a family has 2 members.

**Answer.** (a) \( P(x \geq 5) = .10 + .03 + .01 = .14 \)

(b) \( P(2) = 1 - (.23 + .21 + .10 + .03 + .01) = 1 - .58 = .42 \)

**II.A.5.** At Litchfield College of Nursing, 85% of incoming freshmen nursing students are female and 15% are male. Recent records indicate that 70% of the entering female students will graduate with a BSN degree, while 90% of the male students will obtain a BSN degree. In an incoming nursing student is selected at random, find

(a) \( P(\text{student will graduate, given student is female}) \)

(b) \( P(\text{student will graduate, and student is female}) \)

(c) \( P(\text{student will graduate, given student is male}) \)

(d) \( P(\text{student will graduate, and student is male}) \)
(e) \(P(\text{student will graduate})\)

(f) \(P(\text{student will graduate, or student is female})\)

**Answer.** Let \(F = \text{event student is female}, M = \text{event student is male}, G = \text{event student will graduate, and } N = \text{event student does not graduate.}\)

(a) \(P(G \text{ given } F) = 0.70\)

(b) \(P(F \text{ and } G) = P(F) \cdot P(G \text{ given } F) = (0.85)(0.7) = 0.595\)

(c) \(P(G \text{ given } M) = 0.90\)

(d) \(P(G \text{ and } M) = P(M) \cdot P(G \text{ given } M) = (0.15)(0.90) = 0.135\)

(e) \(P(G) = P(G \text{ given } F) + P(G \text{ given } M) = 0.595 + 0.135 = 0.73\) (since a grad is either male or female)

(f) \(P(G \text{ or } F) = P(G) + P(F) - P(G \text{ and } F) = 0.73 + 0.85 - 0.595 = 0.985\)

II.A.6. Suppose a 30km bicycle race has 28 entrants. In how many ways can the gold, silver and bronze medals be awarded.

**Answer.** This is a permutation problem, order of finish makes a difference: \(P_{28,3} = 28 \cdot 27 \cdot 26 = 19,656.\)

II.A.7 President Geraty has recently received permission to excavate the site of an ancient palace.

(a) In how many ways can he choose 10 of the 48 graduate students in the School of Religion to join him?

(b) Of the 48 graduate students, 25 are female and 23 are male. In how many ways can President Geraty select a group of 10 that consists of 6 females and 4 males?

(c) What is the probability that President Geraty would randomly select a group of 10 consisting of 6 females and 4 males?

**Answer.** (a) \(C_{48,10} = \frac{48!}{38!10!} = 6,540,715,896\)

(b) \(C_{25,6} \cdot C_{23,4} = (177,100)(8855) = 1,568,220,500.\)

(c) The probability is the answer in (b) divided by the answer in (a) which is approximately .2398


(a) In how many ways can the teacher choose 5 of the 15 problems to grade.

(b) Suppose Nicole had the flu and was only able to complete 13 of the 15 problems. What is the probability that she completed all 5 of the problems that the teacher will randomly choose to grade?

**Answer.** (a) \(C_{15,5} = \frac{15!}{5! \cdot 10!} = 3003\)

(b) The number of combinations of 5 problems Nicole has done is \(C_{13,5} = 1287.\) Thus the probability that she completed all 5 of the problems the teacher will randomly choose to grade is

\[
\frac{C_{13,5}}{C_{15,5}} = \frac{1287}{3003} \approx .4286
\]
II.A.9. A local pizza shop advertises a different pizza for every day of your life. They offer 3 choices of crust style (pan, thin or crispy), 20 toppings of which each pizza must have 4, and 5 choices of cheese of which each pizza must have 1. Is their claim valid?

**Answer.** The number of ways of choosing 4 toppings from 20 is \( C_{20,4} = \frac{20!}{16!4!} = 4845 \), thus the total number of choices of pizza is: \( 3 \cdot 4845 \cdot 5 = 72,675 \) which would give a different pizza each day for over 199 years.

II.A.10 (a) How many different license plates can be made in the form xzz-zzz where x is a digit from 1 to 9, and z is a digit from 0 to 9 or a letter A through Z?

(b) What is the probability that a randomly selected license plate will end with the number 00? That is the license plate looks like xzz-z00?

**Answer.** (a) The number of license plates is \( 9 \cdot 36 \cdot 36 \cdot 36 \cdot 36 = 544,195,584 \).

(b) The number of license plates of the form xzz-z00 is \( 9 \cdot 36 \cdot 36 \cdot 36 = 419,904 \). Thus the probability a randomly selected license plate is of this form is

\[
\frac{419,904}{544,195,584} \approx .0007716
\]

II.A.11 Determine whether the following statements are True or False.

(a) Mutually exclusive events must be independent because they can never occur at the same time.

(b) Independent events must be mutually exclusive because they are independent from one another.

(c) If A and B are mutually exclusive, then \( P(A \text{ and } B) = P(A) + P(B) \).

(d) If A and B are independent, then \( P(A \text{ and } B) = P(A) \cdot P(B) \).

**Answer.** (a) False: mutually exclusive events cannot occur together, but they may not be independent.

(b) False: there are independent events that are not mutually exclusive.

(c) False: however it is true that \( P(A \text{ or } B) = P(A) + P(B) \) because the formula \( P(A \text{ or } B) = P(A)+P(B)−P(A \text{ and } B) \) simplifies to \( P(A \text{ or } B) = P(A) + P(B) \) owing to the fact \( P(A \text{ and } B) = 0 \).

(d) True: \( P(B \text{ given } A) = P(B) \) for independent events, so the formula \( P(A \text{ and } B) = P(A) \cdot P(B \text{ given } A) \) simplifies to \( P(A \text{ and } B) = P(A) \cdot P(B) \).

II.B. Random Variables

II.B.1 Classify the following random variables as discrete or continuous.

(a) Speed of an airplane

(b) Age of a college professor chosen at random.

(c) Number of books in the college bookstore.
(d) Weight of a football player chose at random.
(e) Number of lightning strikes in the United States in a given year.
(f) The length of time it takes for someone to run a marathon.
(g) The number of cars on La Sierra campus at a given time.

**Answer.** (a), (b), (d) and (f) are continuous; (c), (e) and (g) are discrete.

**II.B.2.** (a) A 5th grade class holds a raffle in which it sells 5000 tickets at $10 a piece. They will give 1 prize of $1000, 2 prizes of $500, and 7 prizes of $100, and 10 prizes of $50. Make a probability distribution for the net expected winnings $x$ given that 1 ticket is purchased; note the net winnings for the prize of $1000 is $990 because the ticket price is subtracted, and so on all the way down to the $−$10 for a ticket that wins no prize.

(b) What are the expected net earnings of one ticket?

**Answer.** (a) The distribution is

<table>
<thead>
<tr>
<th>$x$</th>
<th>-10</th>
<th>40</th>
<th>90</th>
<th>490</th>
<th>990</th>
</tr>
</thead>
<tbody>
<tr>
<td>$P(x)$</td>
<td>.9960</td>
<td>.0020</td>
<td>.0014</td>
<td>.0004</td>
<td>.0002</td>
</tr>
</tbody>
</table>

(b) $E(x) = 990(.0002) + 490(.0004) + 90(.0014) + 40(.002) − 10(.996) = −9.36$ This means on average, ticket purchasers will lose $9.36 per ticket.

**II.B.3.** The number of computers per household in a small town in 1993 was given by

<table>
<thead>
<tr>
<th>Computers</th>
<th>0</th>
<th>1</th>
<th>2</th>
<th>3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Households</td>
<td>300</td>
<td>280</td>
<td>95</td>
<td>20</td>
</tr>
</tbody>
</table>

(a) Make a probability distribution for $x$ where $x$ represents the number of computers per household in this small town.

(b) Find the mean and standard deviation for the random variable in (i)

(c) What is the average number of computers per household in that small town? Explain what you mean by average.

**Answer.** (a) The probability distribution is

<table>
<thead>
<tr>
<th>Computers ($x$)</th>
<th>0</th>
<th>1</th>
<th>2</th>
<th>3</th>
</tr>
</thead>
<tbody>
<tr>
<td>$P(x)$</td>
<td>.4317</td>
<td>.4029</td>
<td>.1367</td>
<td>.0288</td>
</tr>
</tbody>
</table>

(b) Mean: $\mu = 0(.4317) + 1(.4029) + 2(.1367) + 3(.0288) = .7627$

Variance: $\sigma^2 = 1^2(.4029) + 2^2(.1367) + 3^2(.0288) − .7627^2 = .6272$

Standard Deviation: $\sigma = \sqrt{.6272} = .79195$

(c) On average, there are .7627 computers per household. This is the number one would get if they took the total number of computers in the town and divided by the total number of households. (In actuality, it is off by a little because of rounding in the probability distribution.)

**II.B.4.** Compute the expected value and standard deviation for the following discrete random variable.

<table>
<thead>
<tr>
<th>$x$</th>
<th>2</th>
<th>4</th>
<th>7</th>
</tr>
</thead>
<tbody>
<tr>
<td>$P(x)$</td>
<td>.4</td>
<td>.25</td>
<td>.35</td>
</tr>
</tbody>
</table>
Answer. Expected Value: $E(x) = 2(.4) + 4(.25) + 7(.35) = 4.25$

Standard Deviation: $\sigma^2 = 2^2(.4) + 4^2(.25) + 7^2(.35) - 4.25^2 = 4.6875$ and so the standard deviation is $\sigma = \sqrt{4.6875} \approx 2.165$

II.B.5. In the following table, $x = \text{family size}$ with the corresponding percentage of families that size.

<table>
<thead>
<tr>
<th>$x$</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7 or More</th>
</tr>
</thead>
<tbody>
<tr>
<td>Percentage</td>
<td>42%</td>
<td>23%</td>
<td>21%</td>
<td>10%</td>
<td>3%</td>
<td>1%</td>
</tr>
</tbody>
</table>

(a) Convert the percentages to probabilities and make a probability distribution histogram.

(b) Find the mean and standard deviation for family size (treat the 7 or More category as 7).

(c) What is the probability that a randomly selected family will have: (i) 4 or fewer members; (ii) from 4 to 6 members; (iii) exactly 6 members; (iv) more than 2 members?

Answer. (a) The probability distribution is

<table>
<thead>
<tr>
<th>$x$</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7 or More</th>
</tr>
</thead>
<tbody>
<tr>
<td>$P(x)$</td>
<td>.42</td>
<td>.23</td>
<td>.21</td>
<td>.10</td>
<td>.03</td>
<td>.01</td>
</tr>
</tbody>
</table>

See your text for constructions of histograms; note the bars have width 1 and are centered on $x$ and have height $P(x)$.

(b) Mean: $\mu = 2(.42) + 3(.23) + 4(.21) + 5(.10) + 6(.03) + 7(.01) = 3.12$ Variance: $\sigma^2 = 4(.42) + 9(.23) + 16(.21) + 25(.10) + 36(.02) + 49(.01) = 1.4456$; therefore the standard deviation is $s = \sqrt{1.4456} \approx 1.20233$

(c) (i) $P(x \leq 4) = .42 + .23 + .21 = .86$; (ii) $P(4 \leq x \leq 6) = .21 + .10 + .03 = .34$ (iii) $P(x = 6) = .03$ (iv) $P(x > 2) = 1 - .42 = .58$

II.B.7 Combinations of Random Variables. Norb and Gary entered in a local golf tournament. Both have played the local course many times. Their scores are random variables with the following means and standard deviations.

Norb, $x_1$: $\mu_1 = 115$; $\sigma_1 = 12$ Gary, $x_2$: $\mu_2 = 100$; $\sigma_2 = 8$

Assume that Norbs and Garys scores vary independently of each other.

(a) The difference between their scores is $W = x_1 - x_2$. Compute the mean, variance and standard deviation for the random variable $W$.

(b) The average of their scores is $A = 0.5x_1 + 0.5x_2$. Compute the mean, variance and standard deviation for the random variable $A$.

(c) Norb has a handicap formula $L = 0.8x_1^2$. Compute the mean, variance and standard deviation for the random variable $L$.

Answer. (a) Mean: $\mu = 115 - 100 = 15$ Variance: $\sigma^2 = 12^2 + 8^2 = 144 + 64 = 208$. Standard Deviation: $\sigma \approx 14.42$.

(b) Mean: $\mu = (115 + 100)/2 = 107.5$ Variance: $\sigma^2 = (0.5)^212^2 + (0.5)^28^2 = 208/4 = 52$. Standard Deviation: $s \approx 7.21$.

(c) Mean: $\mu = 0.8(115) - 2 = 90$ Variance: $\sigma^2 = (0.8)^212^2 = 92.16$. Standard Deviation: $\sigma \approx 9.6$. 
II.C. Binomial Probabilities.

II.C.1 Suppose the probability that a (very good) baseball hitter will get a hit at an at bat is .330; use binomial probabilities to compute the probability that the hitter will get

(a) no hits in his next 5 at bats.

(b) exactly one hit in his next 5 at bats.

(c) at least two hits in his next 5 at bats.

Answer. (a) $(.67)^5 \approx .1350$

(b) $C_{5,1}(.33)(.67)^4 \approx .3325$

(c) $1 - (.1350 - .3325) = .5325$

II.C.2 Suppose a certain type of laser eye surgery has a 97% success rate. Suppose that this surgery is performed on 25 patients and the results are independent of one another.

(a) What is the probability that all 25 of the surgeries will be successful?

(b) What is the probability that exactly 24 of the surgeries will be successful?

(c) What is the probability that 24 or fewer of the surgeries will be successful?

Answer. (a) $(.97)^{25} \approx .4669747$

(b) $C_{25,24}(.97)^24(.03)^1 \approx .3610629$

(c) The probability is $1 - .4669747 = .5330253$ (This is the complement of all 25 surgeries being successful)

II.C.3 Suppose that the success rate of a laser eye surgery is 97%, and the surgeries satisfy the properties of a binomial experiment as in the previous question.

(a) Find the mean and standard deviation for the number of successes if the surgery is performed 300 times.

(b) Suppose a given surgeon has performed the surgery 300 times, and has had 280 successes. Find the z-score for the 280 successes. Given that z-score, would you expect a surgeon with a 97% success rate to have 280 or fewer successes out of 300? Explain.

Answer. (a) Mean: $\mu = (300)(.97) = 291$; standard deviation: $\sigma = \sqrt{(300)(.97)(.03)} \approx 2.9546573$.

(b) The z-score is: $z = \frac{280.291}{2.9546573} = -3.722$

Because z-scores below 3 are uncommon in any distribution (compare Chebyshev's theorem), you would not expect to see a surgeon having 280 or fewer successes out of 300 if the success rate is 97% (because fewer than 280 successes would lead to z-scores even lower than 3.722).

II.C.4 Suppose a certain type of laser eye surgery has a 96% success rate. Suppose that this surgery is performed on 17 patients and the results are independent of one another.

(a) What is the probability that all 17 of the surgeries will be successful?
(b) What is the probability that exactly 16 of the surgeries will be successful?
(c) What is the probability that 15 or fewer of the surgeries will be successful?
(d) Find the mean and standard deviation for the expected number of successes if the surgery is performed 50 times?

**Answer.**
(a) \((0.96)^{17} \approx 0.495868\)
(b) \(C_{17,16}(0.96)^{16}(0.04)^1 \approx 0.3539\)
(c) \(1 - 0.4996 - 0.3539 = 0.1465\) (This is the complementary event of 16 or 17 successes)
(d) Mean: \(\mu = (50)(0.96) = 48\); Standard Deviation: \(\sigma = \sqrt{(50)(0.96)(0.04)} \approx 1.386\)

**II.C.5** Consider a binomial random variable with \(n\) trials with probability of success on each trial given as \(p\). Find formulas for: (a) The probability of \(n\) successes; (b) the probability of \(n\) failures; (c) the variance.

**Answer.**
(a) The probability of \(n\) successes is \(p^n\). (b) The probability of no successes is \(nq = n(1-p)\).
(c) The variance is \(\sigma^2 = npq = np(1-p)\).

**II.D: Normal Distributions**

**II.D.1.** Miscellaneous Questions Regarding Normal Distributions.

(a) Find \(z\) so that 85% of the standard normal curve lies to the right of \(z\).
(b) Find \(z\) so that 61% of the standard normal curve lies to the left of \(z\).
(c) Find the \(z\) value so that 90% of the normal curve lies between \(-z\) and \(z\).
(d) Suppose \(x\) is a normal random variable with \(\mu = 50\) and \(\sigma = 13\).
(i) Convert the interval \(37 < x < 48\) to a \(z\) interval.
(ii) Convert the interval \(x > 71\) to a \(z\) interval.
(iii) Convert the interval \(z < -1\) to an \(x\) interval.
(iv) Convert the interval \(1 < z < 3\) to an \(x\) interval.
(e) Let \(b > 0\). If we know that \(P(-b < z < b) = d\), find \(P(z > b)\) in terms of \(d\).
(f) Let \(a\) be any number. If \(P(z < a) = d\), find \(P(z > a)\) in terms of \(d\).
(g) Let \(b > 0\). If we know that \(P(z < -b) = d\), what is \(P(z < -b)z > b)\)?

**Answer.**
(a) Find a \(z\)-value corresponding to an area of 0.15 to the left of it: \(z \approx -1.04\)
(b) Find \(z\) value corresponding to an area of 0.61 to left of it: \(z \approx 0.28\)
(c) Find \(z\) value so that only 5% of normal curve lies above \(z\) (then 5\% will lie below \(z\) be symmetry and 90\% will be between \(z\) and \(z\)). Thus find the \(z\) values so that the area to the left of it is .95: \(z \approx 1.65\)
(d) (i) \(-1 < z < -1.154\); (ii) \(z < 1.615\); (iii) \(x < 37\) (iv) \(63 < x < 89\)
(e) The area of the curve not between \(-b\) and \(b\) is \(1 - d\), because of symmetry half of that will be above \(b\), so \(P(z > b) = (1 - d)/2\)

(f) \(1 - d\)

(g) \(P(z < -b \text{ or } z > b) = P(z < -b) + P(z > b) = d + d = 2d\)

II.D.2 Suppose the distribution of heights of 11-year-old girls is normally distributed with a mean of 62 inches and a standard deviation of 2.5 inches.

(a) What height has a percentile rank of 70?
(b) What proportion of 11 year-old-girls are between 60 and 65 inches tall?
(c) In a group of 800 randomly selected 11-year-old girls, how many would you expect find that are (i) less than 60 inches tall; and (ii) more than 65 inches tall?

Answer. (a) \(z \approx .52\), and so \(x = 62 + .5225 \approx 63.3\), or approximately 63.3 inches.

(b) \(z = \frac{60 - 62}{2.5} = -.8\) and \(z = \frac{65 - 62}{2.5} = 1.2\). Now
\[
P(-.8 < z < 1.2) = .8849 - .2119 = .673
\]
Thus the answer is approximately .673 or 67.3%.

(c) (i) \(P(z < -.8) = .2119\), and \(.2119)(800) = 169.52\); thus, on average we would expect to find about 169.5 girls that are less than 60 inches tall.

(ii) \(P(z > 1.2) = 1 - .8849 = .1151\); and \(.1151)(800) \approx 92\); thus, on average, we would expect to find about 92 girls that are more than 65 inches tall

II.D.3. The distribution of weights of a type of salmon is normal with a mean of 21 lbs and standard deviation of 3 lbs.

(a) What weight has a percentile rank of 30?
(b) What proportion of salmon weigh between 15 lbs and 25 lbs?
(c) What proportion of salmon weigh less than 15 lbs?
(d) What proportion of salmon weight more than 20 lbs?
(e) What proportion of salmon weigh more than 22 lbs?
(f) What proportion of salmon weigh less than 22 lbs?

Answer. (a) \(z \approx -.52\), thus \((x - 21)/3 \approx -.52\) implies \(x \approx 19.44\). So the 30th percentile is approximately 19.44 lbs.

(b) \(P(-2 < z < 1.33) = .9082 - .0228 = .8854\)

(c) \(P(z < -2) = .0228\)

(d) \(P(z > -1/3) = 1 - .3707 = .6293\)

(e) \(P(z > 1/3) = 1 - .6293 = .3707\)

(f) \(P(z < 1/3) = .6293\)
II.D.4 A company determines that the life of the laser beam device in their compact disc player is normally distributed with mean 5000 hours and standard deviation 450 hours. If you wish to make a guarantee so that no more than 5% of the laser beam devices fail during the guarantee period, how many playing hours should the guarantee period cover?

**Answer.** To answer this, find a $z$-value so that 5% of the standard normal curve is to the left of this $z$-value, and then convert it to an $x$-value. Thus an appropriate $z$-value is $z \approx -1.645$, this converts to $x = 5000 \cdot 1.645(450) = 4259.75$. So we would guarantee 4260 hours of use.

II.D.5. Let $z$ be the standard normal random variable.

(a) What are the mean and the standard deviation of $z$?

(b) Find $P(z > 1.69)$.

(c) Find $P(-1.3 < z \leq - .5)$.

(d) Find $P(z < 2.33)$

(e) Find the $z$ value so that 85% of the normal curve lies to the right of $z$.

**Answer.** (a) The means of the standard normal distribution is $\mu = 0$, and the standard deviation is $\sigma = 1$. 

(b) $P(z > 1.69) = 1 - P(z < 1.69) = 1 - .9545 = .0455$

(c) $P(-1.3 < z \leq - .5) = P(z < - .5) - P(z < -1.3) = .3085 - .0968 = .2117$

(d) $P(z < 2.33) = .9901$

(e) $z \approx -1.04$

II.D.6. Which of the following are true about normal random variables $x_1$ and $x_2$ where $x_1$ has a larger standard deviation and larger mean than $x_2$.

(a) There is a higher proportion of the normal curve for $x_1$ that is within two standard deviations of its mean than for $x_2$ because $x_1$ has a larger standard deviation.

(b) Because the mean of $x_2$ is smaller, a smaller proportion of its normal curve is below its mean.

(c) The distribution for $x_1$ is flatter and more spread out than the distribution for $x_2$.

(d) The distributions for $x_1$ and $x_2$ are symmetric about their means.

**Answer.** C and D are true, while A and B are false. Note for (a) the proportions are equal for any normal distributions, and approximately 0.95. Note for (b), this proportion is always 0.5 for all normal distributions, so the proportions are equal.

II.D.7. Suppose that the weights of adult female English Springer Spaniel dogs are normally distributed with a mean of 42 pounds and a standard deviation of 5 pounds. Find

(a) the probability that a randomly selected adult female English Springer Spaniel weighs between 40 and 50 pounds.

(b) the probability that a randomly selected adult female English Springer Spaniel weighs more than 50 pounds.
(c) weight of an adult female English Springer Spaniel whose weight is at the 90th percentile.

**Answer.** (a) \( P(-.4 < z \leq 1.6) = P(z < 1.6) - P(z < -.4) = .9452 - .3446 = .6006. \)
(b) \( P(z > 1.6) = 1 - P(z < 1.6) = 1 - .9452 = .0548 \)
(c) \( z \approx 1.28 \) so \( x \approx 42 + 1.28(5) = 48.4 \), thus approximately 48.4 pounds.

II.E. Normal Approximation to Binomial Distribution

II.E.1. Blood type AB is found in only 3% of the population. If 250 people are chosen at random, what is the probability that
(a) 5 or more will have this blood type?
(b) between 5 and 10 (including 5 and 10) will have this blood type?

**Answer.** (a) For this question we can use the normal approximation to the binomial distribution because \( np = (250)(.03) = 7.5 > 5 \) and \( nq = (250)(.97) = 242.5 > 5 \)
We approximate the binomial distribution with the normal random variable with \( \mu = (250)(.03) = 7.5 \) and \( \sigma = \sqrt{(250)(.03)(.97)} \approx 2.69722 \) using the continuity correction \( P(r \geq 5) = P(x > 4.5) \), so we compute
\[
P(x > 4.5) = P\left(z > \frac{4.575}{2.69722}\right) = P(z > -1.11) = 1 - .1335 = .8665
\]
We can check this using the binomial distribution formula by computing
\[
1 - (P(0) + P(1) + P(2) + P(3) + P(4)) \approx 1 - (0.00049 + .003812 + .014681 + .03753 + .08347) \approx .86
\]
So the normal approximation worked quite well, with an error of .0065.
(b) As in (a) we can use the normal approximation to the binomial distribution. Using the continuity correction \( P(5 \leq r \leq 10) = P(4.5 < x < 10.5) \), so we compute
\[
P(4.5 < x < 10.5) = P\left(-1.11 < z < \frac{10.575}{2.69722}\right) = P(-1.11 < z < 1.11) = .8665 - .1335 = 0.7330
\]

II.E.2. Alaska Airlines has found that 94% of people with tickets will show up for their Friday afternoon flight from Seattle to Ontario. Suppose that there are 128 passengers holding tickets for this flight, and the jet can carry 120 passengers, and that the decisions of passengers to show up are independent of one another.

(a) Verify that the normal approximation of the binomial distribution can be used for this problem.
(b) What is the probability that more than 120 passengers will show up for the flight (i.e., not everyone with a ticket will get a seat on the flight)?

**Answer.** (a) We check that \( np = (128)(.94) = 120.32 > 5 \) and \( nq = (128)(.06) = 7.68 > 5. \)
(b) We approximate the binomial distribution with the normal random variable with \( \mu = 120.32 \) and \( \sigma = \sqrt{(128)(.94)(.06)} \approx 2.687 \). Using the continuity correction \( P(r \geq 121) = P(x > 120.5) \), so we compute

\[
P(x > 120.5) = P \left( z > \frac{120.5120.32}{2.687} \right) = P(z > .07) = 1 - .5279 = .4721
\]

That is, there is a 47.21% chance that the airline will have to bump passengers.

II.E.3. An airline determines that there is a 95% chance that a passenger with a ticket will show up for a given flight. Suppose that an airline has sold tickets to 330 passengers for a flight with 320 seats.

(a) Find the mean and standard deviation for the number of passengers that will show up for the flight.

(b) Explain why the normal approximation to the binomial distribution can be used in this situation.

(c) Use the normal approximation to the binomial distribution to compute the probability that 320 or less of the 330 people holding tickets will show up for the flight.

**Answer.** (a) Mean: \( \mu = np = 330(.95) = 313.5 \)

Standard Deviation: \( \sigma = \sqrt{npq} = \sqrt{330(.95)(.05)} \approx 3.959 \)

(b) Because \( np = 313.5 > 5 \) and \( nq = 16.5 > 5 \).

(c) \( z = \frac{320.5 - 313.5}{3.959} \approx 1.77 \) and \( P(z < 1.77) = .9616 \)
III.A: The Central Limit Theorem and Sampling Distributions.

III.A.1 The heights of 18-year-old men are approximately normally distributed with mean 68 inches and standard deviation 3 inches.

(a) What is the probability that a randomly selected 18-year-old man is between 67 and 69 inches tall.

(b) If a random sample of nine 18-year-old men is selected, what is the probability that the mean height $\bar{x}$ is between 67 and 69 inches?

(c) Is the probability in (b) higher? Why would you expect this?

**Answer.**
(a) Convert $x$ to $z$ using $z = \frac{x - \mu}{\sigma}$. Then

\[
P(67 \leq x \leq 69) = P(-.33 \leq z \leq .33) = .6293 - .3707 = .2586.
\]

(b) The sample size is $n = 9$, and so $\sigma_{\bar{x}} = \sigma/\sqrt{n} = 3/3 = 1$. Then convert $\bar{x}$ to $z$ using $z = \frac{x - \mu}{\sigma_{\bar{x}}}$. Therefore,

\[
P(67 \leq \bar{x} \leq 69) = P(-1 \leq z \leq 1) = .6826.
\]

(c) Yes. One would expect averages of groups to have a much higher probability of being close to the mean, than an individual measurement. Mathematically, this is true because $\sigma_{\bar{x}} < \sigma$.

III.A.2 Suppose the taxi and takeoff time for commercial jets is a random variable $x$ with a mean of 8.5 minutes and a standard deviation of 2.5 minutes. What is the probability that for 36 jets on a given runway total taxi and takeoff time will be

(a) less than 320 minutes?

(b) more than 275 minutes?

(c) between 275 and 320 minutes?

**Answer.** Because $n \geq 30$ we may apply the central limit theorem. For this, $\sigma_{\bar{x}} = \sigma/\sqrt{n} = 2.5/6 \approx .4167$.

(a) Convert 320 to an average, that is 320/36 = 8.89. Then

\[
P(\bar{x} \leq 8.89) = P \left( z \leq \frac{8.89 - 8.5}{.4167} \right) = P(z \leq .93) = .8238.
\]

(b) Convert 275 to an average: 275/36 = 7.638. Then

\[
P(\bar{x} \geq 7.638) = P \left( z \leq \frac{7.638 - 8.5}{.4167} \right) = P(z \geq -2.07) = 1 - .0192 = .9808.
\]

(c) $P(-2.07 \leq z \leq .93) = .8328 -.0192 = .8046$
### III.A.3.

(a) A produce company claims that the mean weight of peaches in a large shipment is 6.0 oz with a standard deviation of 1.0 oz. Assuming this claim is true, what is the probability that a random sample of 1000 of these peaches would have a mean weight of 5.9 oz or less?

(b) If the store manager randomly selected 1000 peaches and found that the mean weight of those 1000 peaches was 5.9 oz, should she be suspicious of the produce company’s claim that the mean weight of peaches in the shipment is 6.0 oz? Explain.

**Answer.**

(a) Use the central limit theorem:

\[
    z = \frac{5.9 - 6.0}{1.0 / \sqrt{1000}} \approx -3.16
\]

\[
    P(z < -3.16) = .5 - .4992 = .0008
\]

Thus, there is approximately a .0008 probability of obtaining such a sample assuming that the mean is 6.0 oz.

(b) Yes, she is about 99.9% sure that the true mean weight is less than 6.0 oz. Because if the mean were 6.0 oz, there is only a .08% chance of obtaining a random sample like the store manager’s.

### III.A.4.

Suppose the mean life span of English Springer Spaniel dogs is normally distributed with a mean of 13 years and a standard deviation of 1.5 years.

(a) What is the probability that a randomly selected English Springer Spaniel will live to be 14 years or older?

(b) What is the probability that a randomly selected sample of 25 English Springer Spaniels will have a mean life span of 14 years or more?

(c) What is the probability that a randomly selected sample of 25 English Springer Spaniels will have a mean life span between 12.5 and 14 years?

**Answer.**

(a) \[ P(x > 14) = P\left( z > \frac{14 - 13}{1.5} \right) = P(z > .67) = 1 - .7486 = .2514 \]

(b) \[ P(\bar{x} > 14) = P\left( z > \frac{14 - 13}{1.5/\sqrt{25}} \right) = P(z > 3.33) = 1 - .9996 = .0004 \]

(c) \[ P(11.5 < \bar{x} < 14) = P\left( \frac{12.5 - 13}{1.5/\sqrt{25}} < z < 3.33 \right) = P(-1.67 < z < 3.33) = .9996 - .0475 = .9521 \]

### III.B. Confidence Intervals on Means

#### III.B.1.

Find \( z_c \) for (a) \( c = .99 \), (b) \( c = .95 \) and (c) \( c = .92 \).

**Answer.**

(a) \( z_{.99} = 2.58 \), (b) \( z_{.95} = 1.96 \).

(c) To find \( z_{.92} \), look for the \( z \) value so that \(.5 + .92/2 = .96 \) of the \( z \)-values on the normal curve are to the left of \( z \). Therefore, \( z_{.92} \approx 1.75 \)

#### III.B.2.

A random sample of 40 farmers gave a sample mean of \( \bar{x} = $6.88 \) received by farmers per 100 pounds of watermelon. Assume that \( \sigma \) is known to be $1.92 per 100 pounds.

(a) Find a 90% confidence interval for the population mean price (per 100lbs) that farmers get for their watermelon crop. What is the margin of error? Repeat this question, except with a 99% confidence interval. Do you expect a longer or shorter interval for a higher level of confidence?
(b) Find the 90% confidence interval for the mean amount of money farmers will receive for 15 tons of watermelon. What is the margin of error?

**Answer.** (a) First for \( c = 0.90 \), then \( z_{90} = 1.645 \) and then the margin of error is \( E = 1.645 \frac{1.92}{\sqrt{40}} \approx 0.50. \)

Therefore, the 90% confidence interval is \( 6.38 < z < 7.38 \)

For \( c = 0.99 \), then \( z_{99} = 2.58 \), and then the margin of error is \( E = 2.58 \frac{1.92}{\sqrt{40}} \approx 0.7832. \) Then the 99% interval is \( 6.097 < z < 7.663 \).

A longer confidence interval should be expected for a higher confidence, because the longer the interval, the more likely it will contain the true population mean.

(b) For this case, multiply answers in (a) by 300, because 15 tons is 300 times 100lbs. Thus, for \( c = 0.90 \), the margin of error is 150, and the confidence interval is \( 1914 < z < 2214 \).

**III.B.3.** Suppose you wish to find the average annual salary for school teachers in the United States. Suppose you took a random sample and found that \( \bar{x} = 53,145 \) and you estimated \( \sigma = 16,451 \).

Find a 95% confidence interval for the mean salary, given that your sample had size (a) \( n = 36 \), (b) \( n = 64 \), (c) \( n = 144 \).

(d) Does the margin of error increase or decrease as \( n \) gets larger. What effect did quadrupling \( n \) have on the margin of error?

(e) Does the confidence interval get longer or shorter as the sample size increases? Is that what you would expect?

**Answer.** (a) \( E = 1.96 \frac{16451}{\sqrt{36}} \approx 5374, \) thus the 95% confidence interval is \( 47771 < z < 58519 \).

(b) \( E = 1.96 \frac{16451}{\sqrt{64}} \approx 4031, \) thus the 95% confidence interval is \( 49114 < z < 57176 \).

(c) \( E = 1.96 \frac{16451}{\sqrt{144}} \approx 2687, \) thus the 95% confidence interval is \( 50458 < z < 55832 \).

(d) The margin of error decreases as \( n \) gets larger. If \( n \) is quadrupled, the margin of error is cut by a factor of 2 (since 2 is the square root of 4).

(e) Confidence intervals get shorter as sample sizes increase. One should expect this, because the larger the sample size, the better its chance of more accurately representing the population.

**III.B.4.** Price to earnings ratios of stock from a random selection of 51 large company stocks are given (see your text). Using the formulas

\[
x = \frac{\sum x}{n} \quad \text{and} \quad s = \sqrt{\frac{\sum x^2 - (\sum x)^2}{n - 1}}
\]

it was computed that \( \bar{x} = 25.2 \) and \( s \approx 15.5 \).

(a) Find a 90% confidence interval for the P/E population mean for all large U.S. companies.

(b) Repeat (a) with a 99% confidence level.

(c) How do companies with P/E’s of 12, 72 and 24 compare to the population mean?

(d) Is it necessary to assume that \( x \) is approximately normal for this case?
Answer. (a) For this case \( \bar{x} = 25.2, \ s \approx 15.5, \) and \( n = 51. \) Therefore, \( d.f. = 50, \) and using the table we find that \( t_{0.05} = 1.676. \) The endpoints of the confidence interval are

\[
25.2 \pm 1.676 \frac{15.5}{\sqrt{51}}
\]

and so the interval is 21.56 to 28.84.

(b) The only thing that changes is \( t_c = 2.678 \) in this case, so the interval is 19.39 to 31.01.

(c) 24 well within the 99% confidence interval, so it is probably close to the mean, but neither 72 nor 12 are in the 99% confidence interval, so we are very certain that those P/E ratios are higher and lower than the mean respectively.

(d) No, the central limit theorem says that the sampling distribution is approximately normal for sufficiently large \( n. \)

II.B.5. Total calcium. A total calcium level below 6mg/dl is related to severe muscle spasms. Recently, the patients total calcium tests gave the following readings (in mg/dl).

\[
\begin{align*}
9.3 & \quad 8.8 & \quad 10.1 & \quad 8.9 & \quad 9.4 & \quad 9.8 & \quad 10.0 & \quad 9.9 & \quad 11.2 & \quad 12.1
\end{align*}
\]

(a) Verify that \( \bar{x} = 9.95 \) and \( s \approx 1.02. \) What formulas did you use?

(b) Find a 99.9% confidence interval for the population mean of total calcium in this patients blood.

(c) Do you think the patient still has a calcium deficiency?

(d) What properties should the distribution possess for the confidence interval in (b) to be valid?

Answer. (a) Use the formulas

\[
\bar{x} = \frac{\sum x}{n} \quad \text{and} \quad s = \sqrt{\frac{\sum x^2 - (\sum x)^2}{n - 1}}
\]
as you did in chapter 3.

(b) For this, \( n = 10, \ \bar{x} = 9.95, \ s \approx 1.02, \ d.f. = 9, \) and so \( t_{0.001} = 4.781 \) and the confidence interval has endpoints

\[
9.95 \pm 4.781 \frac{1.02}{\sqrt{10}}
\]

which means the 99.9% interval for the population mean is from 8.41 to 11.49.

(c) No, the lowest point of the confidence interval is 8.41, so we are very certain (99.9%) that the mean is above 8.41 (and below 11.49).

(d) Because the sample size is small (less than 30), the original distribution must be normal, or at least approximately normal.

III.B.6. The Roman Arches is an Italian restaurant. The manager wants to estimate the average amount a customer sends on lunch Monday through Friday. A random sample of 115 customers’ lunch tabs gave a mean of \( \bar{x} = $9.74 \) with a standard deviation \( s = $2.93. \)

(a) Find a 95% confidence interval for the average amount spent on lunch by all customers.
(b) For a day when the Roman Arches has 115 lunch customers, use part (a) to estimate the range of dollar values for the total lunch income that day.

**Answer.** (a) The sample size is \( n = 115 \geq 30 \), but the population standard deviation is not known so we will use the \( t \)-distribution with \( d.f. = 114 \), from our table we estimate \( t_c = 1.984 \) (since we go to \( d.f. = 100 \)). Thus

\[
E = 1.984 \cdot \frac{2.93}{\sqrt{115}} \approx 0.542 \text{ so the interval is } (9.20, 10.28)
\]

So we are 95% confident the mean amount is from $9.20 to $10.28.

(b) We are 95% confident that total would be in the range \((115 \cdot 9.20, 115 \cdot 10.28)\), that is \((1058, 1182)\).

### III.B.7.

The number of calories for 3 ounces of french fries at eight popular fast food chains are as follows.

<table>
<thead>
<tr>
<th></th>
<th>222</th>
<th>255</th>
<th>254</th>
<th>230</th>
<th>249</th>
<th>222</th>
<th>237</th>
<th>287</th>
</tr>
</thead>
</table>

Use these data to find a 99% confidence interval for the mean calorie count in 3 ounces of french fries obtained from fast-food restaurants.

**Answer.** First, one needs to compute the sample mean and sample standard deviation for the 8 numbers above. For this, one finds

\[
\sum x = 1956 \quad \sum x^2 = 481548 \quad SS_x = 481548 - \frac{1956^2}{8} = 3306.
\]

Thus \( \bar{x} = \frac{1956}{8} = 244.5 \) and \( s = \sqrt{\frac{3306}{7}} = 21.73 \).

This is a small sample \((n < 30)\) with unknown standard deviation, and we’ll assume the population is normal, or nearly so. We use the \( t \)-distribution with \( 7 \) degrees of freedom, thus \( t_{.99} = 3.499 \), and the endpoints of the confidence interval are given by \( \bar{x} \pm t_c \frac{s}{\sqrt{n}} \). Thus the 99% confidence interval has endpoints \( 244.5 \pm 3.499 \frac{21.73}{\sqrt{8}} \) which yields the interval \((217.6, 271.4)\).

### III.C. Confidence Intervals on Proportions

#### III.C.1.

A recent poll reported that 23% of adult Americans surveyed approve of the way the United States has handled the war in Iraq. Moreover, the polling organization reported their methods were as follows.

“These results are based on telephone interviews with a randomly selected national sample of 1,006 adults, aged 18 and older, conducted Feb. 24-26, 2007. For results based on this sample, one can say with 95% confidence that the maximum error attributable to sampling and other random effects is ±3 percentage points. In addition to sampling error, question wording and practical difficulties in conducting surveys can introduce error or bias into the findings of public opinion polls.”

(a) What is the confidence interval that polling organization is suggesting for the proportion of adult Americans favoring the way the U.S. has handled war in Iraq?
(b) Find a 99% confidence interval for the proportion of adult Americans that approve of the way the United States has handled the war in Iraq.

(c) Explain (or show) why the conditions necessary for constructing a confidence interval on a proportion are satisfied in this case.

Answer. (a) Interval: $0.20 < p < 0.26$ Level of Confidence: $c = 0.95$

(b) First, $z_c = 2.58$ and $E = 2.58\sqrt{\frac{(0.23)(0.77)}{1006}} \approx 0.0342$. The endpoints of the interval are then $0.23 \pm 0.0342$. Therefore, the confidence interval is $0.1958 < p < 0.2642$ with 99% confidence.

(c) Because both $np \approx (1006)(0.23) > 5$ and $nq \approx (1006)(0.77) > 5$.

III.C.3. In a 2003 Gallup poll it was reported that 47% of adult Americans surveyed approve of the way the United States has handled the war in Iraq. Moreover, Gallup reported their methods were as follows:

“These results are based on telephone interviews with a randomly selected national sample of 1,006 adults, aged 18 and older, conducted Oct. 24-26, 2003. For results based on this sample, one can say with 95% confidence that the maximum error attributable to sampling and other random effects is ±3 percentage points. In addition to sampling error, question wording and practical difficulties in conducting surveys can introduce error or bias into the findings of public opinion polls.”

(a) What is the confidence interval that Gallup is suggesting for the proportion of adult Americans favoring the way the U.S. has handled war in Iraq?

(b) Find a 99% confidence interval for the proportion of adult Americans that approve of the way the United States has handled the war in Iraq.

(c) What sample size would be needed to estimate the proportion of adult Americans that approve of the way the United States has handled the war in Iraq to within ±0.01 with 95% confidence?

Answer. (a) Gallup is claiming that the 95% confidence interval for the proportion is $(0.44, 0.50)$.

(b) We are given $\hat{p} = 0.47$, $n = 1006$. Now, $\hat{q} = 0.53$ and clearly $n\hat{p} > 5$ and $n\hat{q} > 5$ so we can use the formulas $\hat{p} \pm z_c \sigma_{\hat{p}}$ where $\sigma_{\hat{p}} = \sqrt{\frac{pq}{n}}$. Using the table we find $z_{99} = 2.58$. Therefore, the 99% confidence interval is $0.47 \pm 2.58\sqrt{\frac{(0.47)(0.53)}{1006}}$ which yields the interval $(0.4294, 0.5106)$.

(c) If we use $p = 0.47$ as an estimate for $p$ in the formula $n = pq\left(\frac{z_c}{E}\right)^2$ we get $n = (0.47)(0.53)\left(\frac{1.96}{0.01}\right)^2 = 9569.43$ and so we would use $n = 9570$. If we don’t assume an initial estimate for $p$, we get $n = \frac{1}{4}\left(\frac{z_c}{E}\right)^2 = 9605$.

III.C.4. A recent Gallup poll of 1006 adult Americans found that 37% of those asked oppose cloning of human organs.

(a) Find a 98% confidence interval for the proportion of adult Americans that oppose cloning of human organs.
(b) Based on the answer to (a), would it be appropriate for the Gallup organization to state that less than 38% of adult Americans oppose cloning of human organs? Explain.

**Answer.** (a) First $z_{.98} = 2.33$ (use table), $\hat{p} = .37$ and $n = 1006$. Now $n\hat{p} > 5$ and $n\hat{q} > 5$ and so we can use the confidence interval formula for proportions since the sampling distribution will be approximately normal. Therefore, the 98% confidence interval has endpoints 

$$.37 \pm 2.33 \sqrt{\frac{(.37)(.63)}{1006}}$$

That gives a confidence interval of (.3345, .4055).

(b) Because the upper end of the confidence interval is 40.55% which is well above 38%, it would not be appropriate to state that less that 38% of all adult Americans oppose cloning of human organs using this confidence interval.

### III.C.5.

On June 7, 1999 a poll on the USA Today website showed that out of 2000 respondents, 71% felt that Andre Agassi deserved to be ranked among the greatest tennis players ever.

(a) Assuming that the 2000 respondents form a random sample of the population of tennis fans, construct a 95% confidence interval for the proportion of all tennis fans who feel that Andre Agassi should be ranked among the greatest tennis players ever.

(b) Based on (a), would you be comfortable in saying that the poll is accurate to within plus or minus 2 percent 19 times out of 20? Explain.

(c) In actuality, the survey was based on voluntary responses from readers of the USA Today sports website. Do you think the 2000 respondents actually formed random sample? Explain.

**Answer.** (a) The confidence interval is (.6901, .7299). To find this confidence interval we found 

$$z_{.95} = 1.96, \hat{p} = .71, \hat{q} = .29 \text{ and } n = 2000.$$ Clearly $n\hat{p} > 5$ and $n\hat{q} > 5$ and so we computed:

$$.71 \pm 1.96 \sqrt{\frac{(.71)(.29)}{2000}} = .71 \pm .01999 \text{ to find the endpoints of the interval.}$$

(b) Yes, the 95% confidence interval is .71 ± .01999, hence intervals based on this size of random sample with the given proportion should have an accuracy of ±2% on average 19 times out of 20.

(c) No – the readership of the website is limited to those who have access to the site and choose to visit it; moreover, the survey was not based on a random selection of even those users of the website, but on those who chose to respond to the poll.

### III.C.6.

(Section 8.3#2) In a sample of 519 judges it was found that 285 were introverts.

(a) Let $p$ represent the proportion of all judges who are introverts. Find a point estimate for $p$.

(b) Find a 99% confidence interval for $p$, and explain what it means.

(c) Do you think $np > 5$ and $nq > 5$ for this problem? Why, and why is it important?

**Answer.** (a) $\hat{p} = \frac{285}{519} \approx .5491$.

(b) $z_c = 2.58$, $n = 519$, $\hat{p} = .5491$, $\hat{q} = .4509$ and so 

$$E = 2.58 \sqrt{\frac{(.549)(.451)}{519}} \approx .05635,$$ and so $.4927 < p < .6054$ with 99% confidence
(c) Yes, because there are far more than 5 successes (285) and far more than 5 failures (234) in the sample. This is important, because the confidence interval in (b) relies on the fact that the distribution is approximately normal.

III.C.7. (Section 8.3 #17) In a survey of 1000 large corporations, 250 said given a choice between a job candidate who smokes and an equally qualified nonsmoker, the nonsmoker would get the job (USA Today).

(a) Let \( p \) represent the proportion of all corporations preferring a nonsmoking candidate. Find a point estimate for \( p \).

(b) Find a 95% confidence interval for \( p \).

(c) How would a news writer report these results? What was the margin of error for the 95% confidence interval?

Answer.  (a) \( \hat{p} = \frac{250}{1000} = .250 \)

(b) \( z_c = 1.96, \ n = 1000, \ \hat{p} = .250, \ \text{and} \ \hat{q} = .750. \) Therefore we compute \( E \) and then \( \hat{p} \pm E \) will be the endpoints of the interval.

\[
E = 1.96 \sqrt{\frac{(.25)(.75)}{1000}} \approx .02684 \ \text{and} \ \ .2231 < p < .2768 \ \text{with} \ 95\% \ \text{confidence.}
\]

(c) A survey of 1000 large corporations has shown that 25% would intentionally hire a nonsmoker over an equally qualified smoker. The poll is accurate to within ±2.7 percent 19 times out of 20.

III.D. Sample Sizes for Confidence Intervals

III.D.1. A sample of 37 adult male desert bighorn sheep indicated the standard deviation of the sheep weights to be 15.8 lb (Source: The Bighorn of Death Valley...) How many adult male sheep should be included in a sample to be 90% sure that the sample mean weight \( \bar{\mathbf{x}} \) is within 2.5 lb of the population mean weight \( \mu \) for all such bighorn in the region.

(b) Repeat the question in (a), but given you found an error in the original calculation and the standard deviation was 25.8 lb.

(c) Repeat the question in (a), but you double checked, and the standard deviation of 15.8 lb was correct, but you decided you need a 99% confidence interval.

Answer.  (a) For this question \( z_c = 1.645, \ \sigma = 15.8 \) and the desired \( E = 2.5. \) Therefore,

\[
n = \left( \frac{z_c \sigma}{E} \right)^2 = \left( \frac{(1.645)(15.8)}{2.5} \right)^2 \approx 108.09 \ \text{thus choose a sample size of} \ n = 109.
\]

(b) Same as (a), except \( \sigma = 25.8, \) and so \( n \approx 288.2 \) and so we would use a sample size of \( n = 289. \)

(c) Same as (a), except \( z_c = 2.58, \) and so \( n = \left( \frac{(2.58)(15.8)}{2.5} \right)^2 \approx 265.9 \) and so we would use a sample size of \( n = 266. \)

III.D.2 (Sample Sizes for Proportions) Suppose president Bush decides to conduct a survey concerning the support for sending additional troops to Iraq.
(a) What sample size would he need to estimate the percentage within ±3% with 95% confidence? (assume he doesn’t have an initial estimate for the true proportion)

(b) Same question as (a), but within ±1% with 95% confidence.

(c) Same question as (a), but within ±3% with 99% confidence.

(d) Repeat questions (a) through (c), but assume to start with that a news poll estimated that 27% of Americans support sending additional troops, and that the president used this as a good starting estimate for p.

(e) Approximately what sample size do you think the Gallup Organization uses to create polls that are accurate to ±3 percent 19 times out of 20?

**Answer.** (a) For this part, \( z_c = 1.96 \), the desired \( E = .03 \), and there is no initial estimate for \( p \), so

\[
 n = \frac{1}{4} \left( \frac{z_c}{E} \right)^2 = \frac{1}{4} \left( \frac{1.96}{.03} \right)^2 \approx 1067.1
\]

This means he should use a sample size of 1068.

(b) In this case \( E = .01 \), and so

\[
 n = \frac{1}{4} \left( \frac{z_c}{E} \right)^2 = \frac{1}{4} \left( \frac{1.96}{.01} \right)^2 = 9604.
\]

Notice the huge increase in sample size from (a) to 9604 to get within 1 percent.

(c) Same as (a), except \( z_c = 2.58 \) and so \( n = 1849 \) using the same formula.

(d) Now use the formula \( n = pq \left( \frac{z_c}{E} \right)^2 \) where we use .27 as an estimate for \( p \), and .73 as an estimate for \( q \). To get an accuracy of .03 with 95 percent confidence, then

\[
 n = pq \left( \frac{z_c}{E} \right)^2 \approx (.27)(.73) \left( \frac{1.96}{.03} \right)^2 \approx 841.3
\]

and so a sample size of \( n = 842 \) would be sufficient. Use the same formula for the other parts, but with \( E = .01 \) in the next calculation, and \( z_c = 2.58 \) in the final calculation.

(e) They will never need a random sample larger than 1068 (see (a)). If there is a good estimate for \( p \), they may be able to use a smaller sample (see (d)). If you look at the fine print of some polls, you will see that often a sample size of approximately 1000 is used.

**III.D.3.** (a) What sample size would be needed to estimate the mean from a population with standard deviation of 100 with a maximum error of 5 with 95% confidence?

(b) What sample size from the same population would be needed to estimate the mean with a maximum error of 10 with 95% confidence?

(c) In general, what effect would increasing a sample size by a factor of 16 have on the maximum error \( E \)?

(d) In general, how much must a sample size be increased to cut the maximum error by one-half?

**Answer.** (a) We use the formula \( n = \left( \frac{z_c \sigma}{E} \right)^2 \), and so

\[
 n = \left( \frac{1.96 \cdot 100}{5} \right)^2 = 1536.64 \text{ and so we use } n = 1537.
\]
(b) We use the formula \( n = \left( \frac{z_c \sigma}{E} \right)^2 \), and so \( n = \left( \frac{1.96 \cdot 100}{10} \right)^2 = 384.16 \) (which is one-fourth of the answer in (a)). Going up to the next whole number, we get \( n = 385 \).

(c) Multiplying the sample size by 16 reduces the error to one-fourth of what it was with the original sample size. This is verified with the following calculation. If \( E = z_c \frac{\sigma}{\sqrt{n}} \) the error with the original sample size, then the error with a sample 16 times the size of the original is

\[
z_c \frac{\sigma}{\sqrt{16n}} = \frac{1}{4} z_c \frac{\sigma}{\sqrt{n}} = \frac{1}{4} E.
\]

(d) The sample size must be quadrupled. This is verified with the following calculation. If \( n = \left( \frac{z_c \sigma}{E} \right)^2 \) is the original sample size, then sample size needed to halve \( E \) is

\[
\left( \frac{z_c \sigma}{E} \right)^2 = 4 \left( \frac{z_c \sigma}{E} \right)^2 = 4n.
\]

III.E. Confidence Intervals on Differences of means/proportions

III.E.1. A government official wishes to determine if there has been “grade inflation” for graduating seniors in her state’s high schools over the last 10 years. So she took random sample of 900 graduating seniors GPA’s in 1996 and found the sample to have a mean GPA of 3.13 and a standard deviation of .56, and she found a sample of 1225 graduating seniors in 2006 had a mean GPA of 3.29 with a standard deviation of .53.

(a) Help this official by constructing a 99% confidence interval for the difference of the population means, use the 1996 GPA’s as population 1. Assume the standard deviations given are also the population standard deviations.

(b) Describe in words (like a news reporter) what the interval in (a) means.

(c) Based on your answer in (a), would you be convinced that there has been grade inflation in that state? Explain.

Answer. (a) \( E = 2.58 \sqrt{\frac{(0.56)^2}{900} + \frac{(0.53)^2}{1225}} \approx .062 \). The endpoints of the interval are \( \bar{x}_1 - \bar{x}_2 \pm E \) and so

\[-.222 < \mu_1 - \mu_2 < -.098 \] with 99% confidence.

(b) We are 99% confident that the average GPA of high school graduates in the state in 1996 were .222 to .098 lower than the average GPA of high school graduates in the state in 2006.

(c) Yes, we are 99% certain that the GPA’s are least .098 higher in 2006 than they were in 1996.

III.E.2. In a 1999 survey of 80 Computer Science graduates and 110 Electrical Engineering graduates, it was found that the Computer Science graduates had a mean starting salary of $48,100 with a standard deviation of $7,200, while the Electrical Engineering graduates had a mean starting salary of $52,900 with a standard deviation of $5,300.

(a) Find a point estimate for the difference in average starting salaries for Computer Science and Electrical Engineering graduates.
(b) Let \( \mu_1 \) be the population mean starting salary for the Computer Science graduates and let \( \mu_2 \) be the population mean salary for the Electrical Engineering graduates. Find a 96% confidence interval for \( \mu_1 - \mu_2 \). Assume the standard deviations given are the population standard deviations.

(c) Based on the interval in (b), would you be comfortable saying that the mean starting salary for Computer Science graduates is less than that for Electrical Engineering graduates? Explain.

(d) How would your answer change if the standard deviations given are sample standard deviations, and you were asked to find a 95% confidence interval?

**Answer.** (a) The reasonable point estimate is \( \bar{x}_1 - \bar{x}_2 = 48,100 - 52,900 = -4800 \).

(b) Since each sample size is at least 30, we use the formula:

\[
\bar{x}_1 - \bar{x}_2 \pm z_c \sigma_{\bar{x}_1-\bar{x}_2} \quad \text{where} \quad \sigma_{\bar{x}_1-\bar{x}_2} = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}
\]

For this, \( c = .96 \) and so \( z_{.96} = 2.05 \) (which is found from the normal table as the \( z \)-value for which \( P(z < z_c) = .98 \) since \( .98 = .5 + .96/2 \)). Now compute

\[
\sigma_{\bar{x}_1-\bar{x}_2} = \sqrt{\frac{7,200^2}{80} + \frac{5,300^2}{110}} = 950.45
\]

and so we compute \(-4800 \pm 2.05 \cdot 950.45 \) which yields the interval \((-6,748.43, -2,851.57) \). This means that we are 96% confident that the true difference in salaries is between \(-6,748.43 \) and \(-2,851.57 \), i.e. we are very confident that the Computer Science graduates’ starting mean salary is from $2,851.57 to $6,748.43 less than Electrical Engineering graduates’ mean starting salaries.

(c) Yes, we are very confident that it is at least $2,851.57 less than the mean for Electrical Engineering graduates.

(d) Use the \( t \)-distribution with d.f. = 79, which is not on our table, so we use the next lower number with is 70. For this \( t_{.95} = 1.994 \) in place of the \( z_{.96} = 2.05 \)

III.E.3 (a) A Gallup News Release (see http://www.gallup.com/poll/releases/pr031030.asp) reported that in July 1996, 69% of those surveyed supported assisted suicide for terminally ill, while in May 2003, 72% of those surveyed supported assisted suicide for the terminally ill. Assume the 1996 poll surveyed 1103 adult Americans, while the May 2003 poll surveyed 1009 adult Americans. Find a 99% confidence interval for \( p_1 - p_2 \) where \( p_1 \) is the proportion of adult Americans supporting assisted suicide in July 1996, and \( p_2 \) is the proportion of adult Americans supporting assisted suicide in May 2003.

(b) Based on your interval in (a) would you be comfortable in saying that the proportion of adult Americans supporting assisted suicide was higher in May 2003 than in July 1996? Explain.

**Answer.** (a) First, because \( n_1 \hat{p}_1 = (1103)(.69) > 5 \), \( n_1 \hat{q}_1 = (1103)(.31) > 5 \), \( n_2 \hat{p}_2 = (1009)(.72) > 5 \) and \( n_2 \hat{q}_2 = (1009)(.28) > 5 \), we can use the formula

\[
\hat{p}_2 - \hat{p}_1 \pm z_c \sigma_{\hat{p}_1-\hat{p}_2} \quad \text{where} \quad \sigma_{\hat{p}_1-\hat{p}_2} = \sqrt{\frac{\hat{p}_1 \hat{q}_1}{n_1} + \frac{\hat{p}_2 \hat{q}_2}{n_2}}
\]

With \( c = .99 \) we find \( z_{.99} = 2.58 \) from the table, and we compute

\[
\sigma_{\hat{p}_1-\hat{p}_2} = \sqrt{\frac{.69 \cdot .31}{1103} + \frac{.72 \cdot .28}{1009}} = 0.0198426
\]
The 99% confidence interval has endpoints \((.69 - .72) \pm 2.58 \cdot .0198436\), thus the interval is \((- .0812, .0212)\). We are 99% confident in July 1996 the percentage of adult Americans that supported assisted suicide was 8.12% less to 2.12% more than the percentage in May 2003.

(b) No, I could not say that with 99% confidence because the interval found in (a) included the possibility that the support was up to 2.12% higher in July 1996 than in May 2003.

III.F. Hypothesis Tests on Means

III.F.1 (Length of time to complete a form) You wish to test claim that the average time to complete a form is 15 minutes. You believe it is longer. For this, you surveyed a random sample of 45 people and found that it took them an average of 17 minutes to complete the form with a standard deviation of 5 minutes which you have reason to believe is the population standard deviation. Conduct your test with a level of significance of \(\alpha = .05\).

**Answer.** The hypotheses are: \(H_0: \mu = 15; \quad H_1: \mu > 15\).

The test statistic is \(z = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}} = \frac{17 - 15}{5/\sqrt{45}} = 2.68\).

The P-value is \(P(z \geq 2.68) = .037\). Because .037 < .05, you reject the null hypothesis. The average time is longer than 15 minutes.

III.F.2 For healthy adults, the mean blood pH is \(\mu = 7.4\). It is suspected a new arthritis drug changes blood pH (either higher or lower). Test this at a 5% significance level. To do this, a sample of 31 patients on the new drug were tested, and the sample had a mean \(\bar{x} = 8.1\) and a standard deviation \(s = 1.9\).

**Answer.** The hypotheses are: \(H_0: \mu = 7.4; \quad H_1: \mu \neq 7.4\), and \(\alpha = .05\).

The test statistic is \(t = \frac{\bar{x} - \mu}{s/\sqrt{n}} = \frac{8.1 - 7.4}{1.9/\sqrt{31}} = 2.05\).

Then \(2.042 < t < 2.457\), so the table for \(d.f. = 30\) and a two tailed test shows \(.02 < P\text{-value} < .05\). Because the P-value is less than \(\alpha\), we reject \(H_0\). That is we are quite certain that the drug does change blood pH.

III.F.3 It was reported that the average life span of people in Hawaii is 77 years. Conduct an hypothesis test to see if the life span of people in Honolulu is less than 77 years. Use a level of significance of \(\alpha = .05\). For this test, a random sample of 20 obituary notices of people in Honolulu had a mean \(\bar{x} = 71.4\) years, and \(s \approx 20.65\).

**Answer.** The hypotheses are: \(H_0: \mu = 77; \quad H_1: \mu < 77, \alpha = .05\).

The test statistic is \(t = \frac{\bar{x} - \mu}{s/\sqrt{n}} = \frac{71.4 - 77}{20.65/\sqrt{20}} = -1.21\).

Using the \(t\)-distribution with \(d.f. = 19\), we find \(.10 < P\text{-value} < .125\). Because the P-value is larger than \(\alpha\), we do not reject \(H_0\). The evidence does not suggest that the life span in Honolulu is shorter than 77 years.

**Note.** the data is leaning toward suggesting a shorter life span than 77 years, so if you had the resources to take a larger random sample, you may get a more conclusive result.
III.F.4 (Do SDA’s live Longer?) Suppose the average life span of an American is 76 years with a standard deviation of 14 years. Test at a level of significance of $\alpha = 0.01$ whether the average American SDA life span is longer. For this, suppose that a random sample of 400 recently recorded SDA deaths had an average life span of 78 years, and that the standard deviation was also 14 years. (Do the test using P-value method, and then using the critical region method)

**Answer.** (P-Value Method) The hypotheses are: $H_0: \mu = 76$; $H_1: \mu > 76$, $\alpha = 0.01$. The test statistic is

$$z = \frac{78 - 76}{14/\sqrt{400}} = 2.86$$

The P-value is $P(z > 2.86) = 0.0021$. Because the P-value is less than .01, we reject $H_0$, that is, the evidence does show that on average SDA’s live longer than 76 years.

(Critical Region) For $\alpha = 0.01$ on a right-tailed test, the critical region is $z \geq 2.33$, because the $z = 2.86 > 2.33$ is in the critical region, as with the $P$-value, we reject $H_0$, the evidence suggests that on average SDA’s live longer than 76 years.

III.F.5. The 13 Feb 2007 AAA survey of gasoline prices found that the average price for regular gas in Riverside was 2.602 (from AAA website). Suppose the Riverside Chamber of Commerce thought this price was not accurate, and wanted to prove that the gas price is actually lower. Suppose their survey of 80 randomly selected gas stations in Riverside had a sample mean $\bar{x} = 2.576$ with a standard deviation of $s = 0.152$. Conduct an hypothesis test to determine if the mean is lower than 2.602 using a 5% level of significance.

**Answer.** The hypotheses are: $H_0 : \mu = 2.602$ and $H_1 : \mu < 2.602$, and $\alpha = 0.05$. Given that $\sigma$ is not known, we will use the $t$-distribution with $d.f. = 79$, and the test statistic is

$$t = \frac{2.576 - 2.602}{0.152/\sqrt{80}} \approx -1.53$$

On the table we go to $d.f. = 70$ (since 79 is not on the table) and find that this $t$-value means $0.050 < P$-value $< 0.075$. Because the P-value is larger than $\alpha$, there is not sufficient evidence at the 5% significance level to show that the mean gas price in Riverside is less than 2.602.

III.F.6. (a) Find the critical region if a two-tailed test on a mean is conducted using a large sample at a level of significance of 0.01.

(b) Explain what type I and type II errors are in hypothesis tests.

(c) What is the probability with which we are willing to risk a type I error called?

**Answer.** (a) With the help of the normal table in the front cover, we find that the critical region is $z \leq -2.58$ or $z \geq 2.58$.

(b) See text Section 9.1.

(c) The level of significance, which is denoted by $\alpha$.

III.F.7. (a) A developer wishes to test whether the mean depth of water below the surface in a large development tract was less than 500 feet. The sample data was as follows: $n = 32$ test holes, the sample mean was 486 feet, and the standard deviation was 53 feet, assume that this is the population standard deviation $\sigma = 53$. Complete the test by computing the P-value, and report the conclusion for a 1% level of significance.
(b) What would the conclusion of the test be for a level of significance of \( \alpha = .05 \).

(c) What type of error was possibly made in (b)?

**Answer.**  \( H_0: \mu = 500 \)
\( H_1: \mu < 500 \)

Using the sample data, we compute

\[
    z = \frac{486 - 500}{\frac{53}{\sqrt{32}}} = -1.49
\]

Thus the P-value is \( P(z < -1.49) = .0681 \) We would not reject the null hypothesis at a level of significance of .01, because the P-value is larger than 0.01. Thus the data does not show at the 5% significance level that the average water depth is less than 500 feet.

(b) Do not reject \( H_0 \) since the P-value is bigger than .05.

(c) A type II error—we did not reject the null hypothesis when it could possibly be false.

**III.F.8.** (Connecting Hypothesis Tests and Confidence Intervals) A vendor was concerned that a soft drink machine was not dispensing 6 ounces per cup, on average. A sample size of 40 gave a mean amount per cup of 5.95 ounces and a standard deviation of .15 ounce, again assume that we know \( \sigma = .15 \)

(a) Find the P-value for an hypothesis test to determine if the mean is different from 6 ounces.

(b) For which of the following levels of significance would the null hypothesis be rejected?

   (i) \( \alpha = .10 \)  
   (ii) \( \alpha = .05 \)  
   (iii) \( \alpha = .04 \)  
   (iv) \( \alpha = .01 \)

(c) For each case in part (b), what type of error has possibly been committed?

(d) Find a 96% confidence interval for the mean amount of soda dispensed per cup.

(e) Is your interval in (d) consistent with the test conclusion in (b)(iii)? Explain.

(f) Based on your answer to (b)(iv) would you expect a 99% confidence interval to contain 6?

(g) Suppose that the population standard deviation is \( \sigma = .15 \), what sample size would be needed so that the maximum error in a 96% confidence interval is \( E = .01 \)?

**Answer.**  (a) This is a two-tailed test with

Null Hypothesis: \( \mu = 6 \)

Alternative Hypothesis: \( \mu \neq 6 \).

Using the data, we compute

\[
    z = \frac{5.95 - 6}{\frac{.15}{\sqrt{40}}} = -2.11
\]

The P-value is: \( P(z < -2.11) + P(z > 2.11) = 2 \cdot P(z < -2.11) = 2(.0174) = .0348 \)

(b) Reject the null hypothesis in (i), (ii) and (iii) since the P-value is smaller than \( \alpha \); do not reject the null hypothesis in (iv).

(c) Possible Type I error may occur in (i), (ii) and (iii) while a Type II error may occur in (iv).
(d) For $c = .96$, $z_c = 2.05$ (approximately), look at $z$ value corresponding to an area of .98 on table. Thus the confidence interval, using the large sample method ($n$ is at least 30) yields endpoints:

$$5.95 \pm 2.05 \cdot \frac{.15}{\sqrt{40}}$$

and, so the confidence interval is: $(5.901, 5.999)$. 

(e) Yes. In (b)(iii) we have a confidence level of (at least) $c = 1 - \alpha = .96$ that the mean is different from 6, while in (d) we had a 96% confidence interval that did not contain 6, and so we were (at least) 96% certain that the mean is different from 6. Notice also the confidence interval comes very close to containing 6, this is reflected in the Pvalue being very close to 0.04 in the hypothesis test.

(f) Yes, because we were not 99% confident that the mean was different from 6, we would expect the corresponding confidence interval to contain 6.

(g) The formula to use is: $n = \left(\frac{z_c \sigma}{E}\right)^2$. So we compute

$$n = \left(\frac{2.05 \cdot .15}{.01}\right)^2 = 945.5625,$$

thus we should use a sample size of $n = 946$.

III.F.9. The recent survey of gasoline prices found that the average price for regular gas in Riverside was $2.838$ per gallon. However, we have reason to suspect the gas price in Riverside is higher than this. Conduct an hypothesis test to determine whether the average price for regular gasoline in Riverside is more than $2.838$ per gallon, and test at $\alpha = .05$

To do this hypothesis test, prices at 34 randomly selected gas stations were computed to have a sample mean of $2.891$ with a sample standard deviation of $.19$.

(a) State the null and alternative hypothesis.

(b) Report the P-value of the test

(c) Should you reject or not reject the null hypothesis? Explain basis for your decision.

(d) Interpret your conclusion in (c) in ordinary language.

Answer. (a) Null Hypothesis: $H_0 : \mu = 2.838$  
Alternative Hypothesis: $H_0 : \mu > 2.838$

(b) Because the population standard deviation is not known, we will use the $t$-distribution with $d.f. = 33$. First,

$$t = \frac{2.891 - 2.838}{.19/\sqrt{34}} = 1.627.$$ 

Thus, from the table $.050 < \text{P-value} < .075$.

(c) You should not reject the null hypothesis because the P-value is larger than the level of significance of .05.

(d) The data is not strong enough to suggest that true average price for gasoline in Riverside is more than $2.838$ per gallon (at the 5% level of significance). In other words, we are not 95% sure that the average gas price in Riverside is more than $2.838$ per gallon.

III.G. Hypothesis test on Proportions
III.G.1 (a) Suppose that a February Gallup poll of 1200 randomly selected voters found that 53 percent support George W. Bush’s energy policy. Conduct an hypothesis test at a level of significance of $\alpha = .01$ to test whether the true voter population support for George W. Bush’s energy policy in February was less than 56 percent (use critical region method).

(b) Report the P-value of the test in (a) and give a practical interpretation of it.

Answer. The null hypothesis is $H_0 : p = .56$, and the alternative hypothesis is $H_1 : p < .56$. The critical region is $z \leq -2.33$. Since $n = 1200$ and $p = .56$ and $q = .44$, we clearly have $np > 5$ and $nq > 5$. Thus we compute

$$z = \frac{.53 - .56}{\sqrt{(\frac{.56)(.44)}{1200}}} \approx -2.09$$

Because $-2.09$ does not fall in the critical region, we conclude there is not sufficient evidence to reject the null hypothesis at a level of significance of 1%.

(b) The P-value is $P(z < -2.09) = .0183$. Thus we would not reject $H_0$, however, we are roughly 98% certain that less than 56% of all voters at the time of the poll supported President G.W. Bush’s energy policy.

III.G.2. (a) Suppose that a February Gallup poll of 1200 randomly selected voters found that 53 percent support George W. Bush’s energy policy. Conduct an hypothesis test at a level of significance of .05 to test whether the true voter population support for George W. Bush’s energy policy in February was less than 56 percent.

(b) Explain what the P-value of the test in (a) means.

(c) Verify that the necessary conditions on the random sample were satisfied in order for us to conduct this test.

Answer. (a) The null hypothesis is $H_0 : p = .56$, and the alternative hypothesis is $H_1 : p < .56$. The test statistic is

$$z = \frac{.53 - .56}{\sqrt{(\frac{.56)(.44)}{1200}}} \approx -2.09$$

The P-value for the test is $P(z < -2.09) = .0183$. Because the P-value is less than $\alpha (.0183 < .05)$, we reject the null hypothesis. That is, at the 5% significance level, the evidence shows that the February level of support for the President’s energy policy was less than 56 percent.

(b) The P-value is $P(z < -2.09) = .0183$. Thus we are roughly 98% certain that less than 56% of all voters at the time of the poll supported President G.W. Bush’s energy policy.

(c) Because $np \approx (.53)(1200) > 5$ and $nq \approx (.47)(1200) > 5$.

III.H. Hypothesis tests on differences of means/proportions.

III.H.1. A reading test is given to both a control group and an experimental group (which received special tutoring). The average score for the 30 subjects in the control group was 349.2 with a sample standard deviation of 56.6. The average score for the 30 subjects in the experimental group was 368.4 with a sample standard deviation of 39.5. Use a 4% level of significance to test the claim that the experimental group performed better than the control group.

Answer. We wish to conduct the test $H_0 : \mu_1 = \mu_2$ versus $H_1 : \mu_1 > \mu_2$ where $\mu_1$ is the population mean test score for all people who received the special tutoring. We use the $t$-distribution with
d.f. = 29 because the population standard deviations are unknown, and the test statistic is
\[
t = \frac{\bar{x}_1 - \bar{x}_2}{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} = \frac{(368.4 - 349.2) - 0}{12.6013} = 1.52.
\]
The P-value is \(P(t > 1.52)\) with \(d.f = 29\). The table shows \(.05 < \text{P-value} < .075\). Because the
P-value is larger than \(\alpha = .04\), we do not reject the null hypothesis at a 4% level of significance.
There is not sufficient evidence to show that the tutoring increases the mean score.

**III.H.2.** Nationally about 28% of the population believes that NAFTA benefits America. A random
sample of 48 interstate truck drivers showed that 19 believe NAFTA benefits America. Conduct an
hypothesis test to determine whether the population proportion of interstate truckers who believe
NAFTA benefits America is higher than 28%. Test at a level of significance of \(\alpha = .05\).

(a) State the null and alternative hypotheses. Is this a right-tailed, left-tailed or two-tailed test?

(b) Find the P-value for the test, and report your conclusion for the test.

(c) Would you reject the null hypothesis at a level of significance of \(\alpha = .04\)?

(d) Would you reject the null hypothesis at a level of significance of \(\alpha = .01\)?

(e) Do you think the proportion of interstate truckers who believe NAFTA benefits America is higher
than 28%? Explain your answer.

**Answer.** (a) The null hypothesis is \(H_0 : p = .28\), the alternative hypothesis is \(H_1 : p > .28\). This
is a right-tailed test.

(b) For this data, we have \(n = 48\), \(p = .28\) and \(q = .72\). Clearly \(np > 5\) and \(nq > 5\) and so we compute \(\hat{p} = \frac{19}{48} = .39583\); then
\[
z = \frac{.39583 - .28}{\sqrt{\frac{(0.28)(0.72)}{48}}} \approx 1.79
\]
Therefore, the P-value is \(P(z > 1.79) = 1 - .9633 = .0367\). Because the P-value is less than \(\alpha = .05\),
we reject \(H_0\). The results are statistically significant; they indicate that the proportion of interstate
truck drivers that believe NAFTA benefits America is higher than .28.

(c) Yes, because the P-value is less than 0.04.

(d) No, because the P-value is more than 0.01.

(e) Yes, and I’m quite sure of this, in fact, I’m roughly 96% sure of it.

**III.H.3.** A random sample of \(n_1 = 288\) voters registered in the state of California showed that 141
voted in the last general election. A random sample of \(n_2 = 216\) registered voters in Colorado showed
that 125 voted in the last general election. Do these data indicate that the population proportion
of voter turnout in Colorado is higher than that in California? Use a 5% level of significance.

**Answer.** Let \(p_1\) and \(p_2\) represent the population proportions of voter turnout in California and
Colorado respectively. We will test \(H_0 : p_1 = p_2\) versus \(H_1 : p_1 < p_2\). Since we are assuming \(p_1 = p_2\),
we use the pooled estimate for proportions, that is
\[
\bar{p} = \frac{141 + 125}{288 + 216} = .5278.
\]
Then we use this for an estimate of $p_1$ and $p_2$ in the formula for $\sigma_{\hat{p}_1 - \hat{p}_2}$, so we obtain

$$\sigma_{\hat{p}_1 - \hat{p}_2} = \sqrt{\frac{(0.5278)(0.4722)}{288} + \frac{(0.5278)(0.4722)}{216}} = 0.04494.$$  

From this, we compute

$$z = \frac{(\hat{p}_1 - \hat{p}_2) - (p_1 - p_2)}{\sigma_{\hat{p}_1 - \hat{p}_2}} = \frac{(.4896 - .5787) - 0}{0.04494} = -1.98.$$  

The P-value is $P(z < -1.98) = 0.0239$. Because the P-value is less than .05, we reject $H_0$ at the 5% level of significance.

### III.J Short Answer Questions from topics in III.

#### III.J.1
Explain how you decide when to use normal distribution or $t$-distribution when

(a) finding confidence intervals for a mean;

(b) finding confidence intervals for a difference of two means;

(c) conducting and hypothesis test on a population mean.

In each case, what formulas do you use, if using the $t$-distribution how many degrees of freedom do you use, what assumptions are necessary on the distribution(s) and or sample sizes?

**Answer.** See your Text!

#### III.J.2
(a) What is the significance of the Central Limit Theorem, and what does it say?
(b) Let $x$ be a random variable from any population with mean $\mu$ and standard deviation $\sigma$. What type of distribution will the distribution for the mean from samples of size $n$ with $n$ large approximate? What are the mean and standard deviation for the sampling distribution of sample size $n$?

**Answer.** (a) Reasons for its importance is that we do not need to assume a population is normal when finding confidence intervals for the mean or when conducting hypothesis tests on a mean when sample sizes are large. For the rest of (a) and (b) see your text.

#### III.J.3
Find the mean and standard deviation for the distribution for $\bar{x}$ based on a random sample of size 64 from a population with mean 37 and standard deviation 24.

**Answer.** $\mu_{\bar{x}} = 37$ and $\sigma_{\bar{x}} = \frac{24}{\sqrt{64}} = 3$.

#### III.J.4
(a) What conditions are necessary in order to find a confidence interval for a proportion?
(b) Same question, but for a confidence interval for the difference of two proportions?
(c) Same question, but for an hypothesis test on a proportion.
(d) In each of the above cases, what formulas do you use?

**Answer.** See your text.
III.J.5 If a population has a standard deviation of 5, what sample size would be necessary in order for a 99% confidence interval to estimate the population mean within 2?

**Answer.** We use the formula \( n = \left( \frac{z_c \sigma}{E} \right)^2 \) and we find that \( n = \left( \frac{2.58 \cdot 5}{2} \right)^2 = 41.6025 \). The sample size must be the next larger whole number which is \( n = 42 \).

III.J.6. What size of random sample is needed by the Gallup organization to estimate a population proportion within ±.02 with 90% confidence? In your calculation assume that there is no preliminary estimate for \( p \).

**Answer.** We use the formula \( n = \frac{1}{4} \left( \frac{z_c}{E} \right)^2 = \frac{1}{4} \left( \frac{1.645}{.02} \right)^2 = 1691.265 \) and so we use a sample size at the next larger whole number which is \( n = 1692 \).

III.J.7. Repeat III.J.6., but assume that there is a preliminary estimate that \( p = .33 \).

**Answer.** We use the formula \( n = pq \left( \frac{z_c}{E} \right)^2 = (.33)(.67) \left( \frac{1.645}{.02} \right)^2 = 1495.76 \) and so we use a sample size at the next larger whole number which is \( n = 1496 \).

III.J.8. Suppose you are to construct a 95% confidence interval for the mean using a sample of size \( n=19 \) from a normal population with unknown standard deviation. What value of \( t_c \) would you use in your confidence interval formula?

**Answer.** We use a \( t \)-distribution with \( n - 1 = 18 \) degrees of freedom, and look on the table in the back cover to find \( t_{.95} = 2.101 \)

III.J.9. Suppose you are to construct a 92% confidence interval for the mean using a large sample from a population with known standard deviation, what value of \( z_c \) would you use?

**Answer.** We find \( z_{.92} = 1.75 \) (use the table for areas of the standard normal, find the \( z \)-value corresponding to an area of \( .5 + .92/2 = .96 \)).

III.J.10. Which distribution (normal or \( t \) with how many \( d.f. \)) would you use to find a confidence interval from a normal population with known standard deviation given a sample size of 15?

**Answer.** Normal distribution because the population standard deviation is known.

III.J.11. Explain what Type I and Type II errors are in hypothesis tests.

**Answer.** See your text.

III.J.12. Explain what the level of significance of an hypothesis test is in terms of Type I errors.

**Answer.** In short, the probability you are willing to risk making a Type I error. See text for more.

III.J.13. (a) If you did a right-tailed hypothesis test on a mean, and found that \( t = 1.9641 \) with \( d.f. = 16 \), what \( P \)-value would you report.
(b) Same question, but with a two-tailed test.
(c) Same question, but with a left-tailed test.
(d) Same question, for left-tailed test, but $t = -1.9641$.

**Answer.**  Note $1.746 < 1.9641 < 2.583$. Thus
(a) $0.025 < \text{P-value} < 0.050$
(b) $0.050 < \text{P-value} < 0.100$
(c) Huge: The P-value is larger than 0.500 because $t > 0$, in fact, $0.95 < \text{P-value} < 0.975$.
(d) $0.025 < \text{P-value} < 0.050$

**III.J.14.** (a) If you did a right-tailed hypothesis test on a mean, and found that $z = 2.13$, what P-value would you report.

(b) Same question, but with a two-tailed test.
(c) Same question, but with a left-tailed test.
(d) Same question, for left-tailed test, but $z = -2.13$.

**Answer.** (a) $P(z > 2.13) = 1 - 0.9834 = 0.0166$
(b) $2(0.0166) = 0.0332$
(c) Huge: $P(z < 2.13) = 0.9834$.
(d) $P(z < -2.13) = 0.0166$.

**III.J.15.** Explain how you use the level of significance and P-value to determine when you (a) reject $H_0$, and (b) do not reject $H_0$

**Answer.** Reject the null hypothesis when the P-value $\leq \alpha$; do not reject the null hypothesis when the P-value $> \alpha$.

**III.J.16.** (a) Explain what the level of significance $\alpha$ is for an hypothesis test.
(b) If a two tailed test reports a test statistic of $z = 2.23$, what is the P-value for the test?
(c) What value of $z_c$ should be used for a $94\%$ confidence interval?
(d) What sample size should be used in a population with a standard deviation of $\sigma = 7.2$ to estimate the population mean to within $\pm 1$ in a $95\%$ confidence interval?
(e) What is the standard deviation for the sampling distribution of $\bar{x}$ based on samples of size $n = 64$ for a population with a standard deviation of 18?
(f) What sample size should the Gallup organization use if it wishes to estimate the percentage of Americans who support troop escalation in Iraq to an accuracy of plus or minus 3 percent 19 times out of 20?

**Answer.** (a) The level of significance the probability at which the tester is willing to risk making a Type I error.
(b) \( P(z < -2.23) + P(z > 2.23) = 2(.0129) = .0258 \)
(c) Find \( z \)-value so that 50% + 94/2% = 97% of the normal curve is to the left of \( z \). Thus \( z_{.94} = 1.88 \)
(d) \( n = \left( \frac{ z_c \sigma }{ E } \right)^2 = \left( \frac{ 1.96 \cdot 7.2 }{ 1 } \right)^2 = 199.14. \) Thus use \( n = 200. \)
(e) \( \sigma_{\bar{x}} = \frac{ \sigma }{ \sqrt{n} } = \frac{ 18 }{ \sqrt{64} } = \frac{ 18 }{ 8 } = \frac{ 9 }{ 4 } = 2.25 \)
(f) \( n = \frac{ 1 }{ 4 } \left( \frac{ z_c }{ E } \right)^2 = \frac{ 1 }{ 4 } \left( \frac{ 1.96 }{ .03 } \right)^2 = 1067.1. \) Thus use \( n = 1068. \)

**III.J.17.**

(a) If a population has a standard deviation of 80, what sample size would be necessary in order for a 95% confidence interval to estimate the population mean within 10?

(b) What size of sample is needed by the Gallup organization to estimate a population proportion within ±.02 with 95% confidence. In your calculation assume that there is no preliminary estimate for \( p. \)

(c) Suppose you are to construct a 99% confidence interval for the mean using a sample of size \( n = 12 \) from a normal population with unknown standard deviation. What value of \( t_c \) would you use in the formula \( \bar{x} \pm t_c \frac{ s }{ \sqrt{n} }? \)

**Answer.**

(a) We use the formula \( n = \left( \frac{ z_c \sigma }{ E } \right)^2 \) and we find that \( n = \left( \frac{ 1.96 \cdot 80 }{ 10 } \right)^2 = 245.86. \) The sample size must be the next larger whole number which is \( n = 246. \)

(b) We use the formula \( n = \frac{ 1 }{ 4 } \left( \frac{ z_c }{ E } \right)^2 = \frac{ 1 }{ 4 } \left( \frac{ 1.96 }{ .02 } \right)^2 \) which gives us a sample size of 2401.

(c) We use a \( t \)-distribution with \( n - 1 = 11 \) degrees of freedom, and look on the table in the back cover to find \( t_{.99} = 3.106 \)
IV. Questions from Chapters 10 and 11

IV.A. Least Squares and Correlation

IV.A.1. Draw scatter plots that exhibit: (a) no linear correlation; (b) moderate to good positive linear correlation; (c) moderate to good negative linear correlation (d) perfect positive linear correlation (e) perfect negative linear correlation.

Answer. See Text, Section 10.1.

IV.A.2. Approximately what value would you expect for the correlation coefficient when a scatter plot exhibits: (a) no linear correlation; (b) good positive linear correlation; (c) good negative linear correlation (d) perfect positive linear correlation (e) perfect negative linear correlation?

(f) What are the possible values that a correlation coefficient can take?

Answer. (a) Close to 0; (b) close to, but less than 1; (c) close to, but more than, −1. (d) −1; (e) 1; (f) −1 ≤ r ≤ 1.

IV.A.3. The following sample data concerns the number of years a student studied German in school versus their score on a proficiency test.

<table>
<thead>
<tr>
<th>Years (x)</th>
<th>3</th>
<th>4</th>
<th>4</th>
<th>2</th>
<th>5</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>3</th>
<th>2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Test Score (y)</td>
<td>57</td>
<td>78</td>
<td>72</td>
<td>58</td>
<td>89</td>
<td>63</td>
<td>73</td>
<td>84</td>
<td>75</td>
<td>48</td>
</tr>
</tbody>
</table>

The sums are: \( \sum x = 35 \) \( \sum y = 697 \) \( \sum x^2 = 133 \) \( \sum y^2 = 50085 \) \( \sum xy = 2554 \)

(a) Find the equation of the least squares line for this data.

(b) Use your line from (a) to predict the score on the proficiency test of a person who had 3.5 years of German.

(c) Use the regression line in (a) to predict the number of years of German required to achieve a proficiency score of 75.

(d) Compute the correlation coefficient \( r \) for this data. What does this coefficient suggest about a linear relationship between number of years German was studied in school and test scores for this sample? That is, determine whether it is a good fit, and whether it indicates a positive or negative linear relationship.

(e) Compute the coefficient of determination, and interpret what it means.

Answer. (a) The slope of the line is \( b = \frac{n \sum xy - (\sum x)(\sum y)}{n \sum x^2 - (\sum x)^2} = 10.90476 \). The \( y \)-intercept is then \( a = \bar{y} - b\bar{x} = 69.7 - (10.90476)(3.5) = 31.533333 \).

Thus the equation of the line is \( y = 10.90476x + 31.533333 \)

(b) \( y = 10.90476(3.5) + 31.533333 = 69.7 \), the predicted score is 69.7.

(c) \( x = \frac{75 - 31.533333}{10.90476} = 3.99 \) years.
(d) \[ r = \frac{n \sum xy - (\sum x)(\sum y)}{\sqrt{n \sum x^2 - (\sum x)^2}\sqrt{n \sum y^2 - (\sum y)^2}} = .91111. \] This value is reasonably close to 1, which means it represents a good linear relation with positive slope (i.e. as x increases so does y). Remember, the closer r is to 1 (or to -1) the better the linear fit will be.

(e) The coefficient of determination is \( r^2 = .8301. \) This means that 83% of the variation of the observed y’s from \( \hat{y} \) is explained by the regression line, while 17% is not explained.

**IV.A.4.** 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) Find the equation of the least squares regression line.

(b) Does the data appear to be positively or negatively correlated? Explain.

(c) Use the regression line to find: (i) How many absences per year would be expected from an employee that makes $38,000 per year? (ii) The expected salary for someone with 2 absences per year?

(d) Given that the correlation coefficient is \(-.945\), determine how well the data fits the line of best fit. Explain your answer.

**Answer.** (a) Therefore, the slope is \( b = \frac{n \sum xy - (\sum x)(\sum y)}{n \sum x^2 - (\sum x)^2} = -.09054, \) 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) Negatively correlated because as the x’s increase, the y’s generally decrease.

(c) (i) Solve \( y = -.09054(38) + 4.4195 = .98 \) absences per year. (ii) Solve \( 2 = -.09054x + 4.4195 \) and so \( x = 26.723, \) or approximately $26,723 per year.

(d) 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 -.945 is quite close to -1, the fit should be quite good.

**IV.B. Goodness of Fit**

**IV.B.1 (a)** A local radio station claims that 15 percent of all people in Riverside say it is their favorite station, 65 percent of all people in Riverside listen to it occasionally, while 20 percent never listen to it. Suppose you surveyed 200 randomly selected people in Riverside and found that of those 200 people, 20 claimed it was their favorite station, 131 said they listen to it occasionally, while 49 never listen to it. Conduct an hypothesis test at a level of significance of .05 to determine whether the station’s claim concerning the distribution of listeners is correct.
(b) Explain why the sample size was sufficiently large for this test to be valid.

**Answer.** (a) This is goodness of fit problem for which we use the $\chi^2$ distribution (see Section 11.2). There are three categories: Favorite, Occasional and Never. Thus we have $3 - 1 = 2$ degrees of freedom. We test the hypotheses:

$H_0 : p_1 = .15 \quad p_2 = .65 \quad p_3 = .20$

$H_1 :$ the proportions are not as claimed in $H_0$.

In the Favorite category, the expected value is is 15% of 200 which is 30, while the observed value was 20. In the Occasional category, the expected value is 65% of 200 which is 130 while the observed value was 131. In the Never category the expected value is 20% of 200 which is 40 while the observed value was 49. We now compute

$$\chi^2 = \sum \frac{(O - E)^2}{E} = \frac{(20 - 30)^2}{30} + \frac{(131 - 130)^2}{130} + \frac{(49 - 40)^2}{40} = 5.366.$$  

Using the $\chi^2$ table with $d.f. = 2$, we find $4.61 < 5.366 < 5.994$ which means $.050 < P-value < .100$. Because the P-value is bigger than .05, there is not enough evidence to show that the distribution is different from what was claimed in the null hypothesis.

(b) The expected values in the categories are (.15)(200) = 30, (.65)(200) = 130 and (.20)(200) = 40 which are all bigger than 5.

**IV.B.2.** 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, P-value 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, $E = \frac{1}{6}(120) = 20$ in each category, and we use the formula

$$\chi^2 = \sum \frac{(O - E)^2}{E},$$

where there are $n - 1 = 6 - 1 = 5$ degrees of freedom. The sample test statistic is

$$\chi^2 = \frac{(24 - 20)^2 + (18 - 20)^2 + (15 - 20)^2 + (26 - 20)^2 + (21 - 20)^2 + (16 - 20)^2}{20} = \frac{98}{20} = 4.9.$$  

Now with $d.f. = 5$ on the $\chi^2$ distribution, we see $1.61 < 4.9 < 9.24$, which means $.10 < P-value < .90$. Because the P-value $> .05$ we do not reject the null hypothesis, that is, there is not sufficient evidence to conclude that the die is not fair.