STA 580 — Fall 2008 — Dr. Charnigo

Written Assignment 1 Solutions

1a. Among non-smokers, the sample mean and sample standard deviation are 130.16 and 14.84, respectively. [Details for pencil-and-paper calculations: We have $n = 32$, $\sum_{i=1}^{n} x_i = 4165$, $\sum_{i=1}^{n} x_i^2 = 548925$, and $(\sum_{i=1}^{n} x_i)^2 = 4165^2 = 17347225$. Hence, we find that $\bar{x} = 4165/32 = 130.16$, $s^2 = (548925 - 17347225/32)/31 = 220.14$, and $s = \sqrt{220.14} = 14.84$.]

Among smokers, the sample mean and sample standard deviation are 138.72 and 15.29, respectively. [Details for pencil-and-paper calculations: We have $n = 25$, $\sum_{i=1}^{n} x_i = 3468$, $\sum_{i=1}^{n} x_i^2 = 486690$, and $(\sum_{i=1}^{n} x_i)^2 = 3468^2 = 12027024$. Hence, we find that $\bar{x} = 3468/25 = 138.72$, $s^2 = (486690 - 12027024/25)/24 = 233.71$, and $s = \sqrt{233.71} = 15.29$.]

1b. Among non-smokers, the sample median and sample interquartile range are 131 and 139.5-119 = 20.5, respectively. [Details for pencil-and-paper calculations: Since $n = 32$ is even, the median is the average of the 16th and 17th ordered observations; this is 131. To find the 75^{th} percentile, note that $p = 75$ and that $np/100 = 24$ is an integer. The greatest integer less than or equal to 24 is $k = 24$, so the 75^{th} percentile is the average of the 24th and 25th ordered observations; this is 139.5. To find the 25^{th} percentile, note that $p = 25$ and that $np/100 = 8$ is an integer. The greatest integer less than or equal to 8 is $k = 8$, so the 25^{th} percentile is the average of the 8th and 9th ordered observations; this is 119. The interquartile range is the difference between the 75^{th} and 25^{th} percentiles; this is 20.5.]

Among smokers, the sample median and sample interquartile range are 135 and 150-130 = 20, respectively. [Details for pencil-and-paper calculations: Since $n = 25$ is odd, the median is the 13th ordered observation; this is 135. To find the 75^{th} percentile, note that $p = 75$ and that $np/100 = 18.75$ is not an integer. The greatest integer less than or equal to 18.75 is $k = 18$, so the 75^{th} percentile is the 19th ordered observation; this is 150. To find the 25^{th} percentile, note that $p = 25$ and that $np/100 = 6.25$ is not an integer. The greatest integer less than or equal to 6.25 is $k = 6$, so the 25^{th} percentile is the 7th ordered observation; this is 130. The interquartile range is the difference between the 75^{th} and 25^{th} percentiles; this is 20.]

1c. [You should display a printout of the side-by-side box plots.] The sample distribution of systolic blood pressure among non-smokers appears roughly symmetric: the distance from the 25^{th} percentile to the 50^{th} percentile is similar to the distance from the 50^{th} percentile to the 75^{th} percentile, and the upper "whisker" is of comparable length to the lower "whisker". The sample distribution of systolic blood pressure among smokers appears slightly but not markedly right skewed: the distance from the 25^{th} percentile to the 50^{th} percentile is less than the distance from the 50^{th} percentile to the $75th$ percentile, but the upper "whisker" is of comparable length to the lower "whisker".

With regard to central tendency, there is a clear visual distinction between the two sample distributions: larger values of systolic blood pressure occur more often among smokers (as we would anticipate). With regard to variability, there is little if any visual distinction between the two sample distributions.

1d. The population percentage of non-smokers who have systolic blood pressure measurements between 140 and 160 is $P(140 \leq X \leq 160)$, where X is normal with mean 130.16 and standard deviation 14.84. By standardization, the requested probability is $P\left(\frac{140-130.16}{14.84} \leq Z \leq \frac{160-130.16}{14.84}\right)$, where Z is standard normal. Using Table 3 or SAS, we find that $\Phi\left(\frac{160-130.16}{14.84}\right) - \Phi\left(\frac{140-130.16}{14.84}\right) = \Phi(2.011) - \Phi(0.663) = 0.978 - 0.746 = 0.232 = 23.2\%$.

The population percentage of non-smokers who have systolic blood pressure measurements above 160 is $P(X > 160) = 1 - P(X \le 160) = 1 - P\left(Z \le \frac{160-130.16}{14.84}\right) = 1 - \Phi(2.011) = 0.022 = 2.2\%$.

1e. The systolic blood pressure measurement defining the boundary between the top 15 percent and bottom 85 percent in the population of non-smokers is the number c_{85} such that $P(X \le c_{85}) = 0.85$, where X is normal with mean 130.16 and standard deviation 14.84. By standardization, we have $P\left(Z \le \frac{c_{85}-130.16}{14.84}\right) = \Phi\left(\frac{c_{85}-130.16}{14.84}\right) = 0.85$. On the other hand, Table 3 or SAS shows that $\Phi(1.036) = 0.85$, so $1.036 = (c_{85} - 130.16)/14.84$. Solving for c_{85} yields 145.5.

The systolic blood pressure measurement defining the boundary between the top 5 percent and bottom 95 percent in the population of non-smokers is the number c_{95} such that $P(X \le c_{95}) = 0.95$. We have $P\left(Z \le \frac{c_{95}-130.16}{14.84}\right) = \Phi\left(\frac{c_{95}-130.16}{14.84}\right) = 0.95$. On the other hand, Table 3 or SAS shows that $\Phi(1.645) = 0.95$, so $1.645 = (c_{95} - 130.16)/14.84$. Solving for c_{95} yields 154.6.

2a. Let A denote the event that a person exercises, and let B denote the event that a person develops coronary heart disease. We are told that $P(B|A) = 0.15$, $P(B|\bar{A}) = 0.25$, and $P(A) = 0.30$. The first order of business is to find $P(B)$. We can use the law of total probability: $P(B) = P(B|A)P(A) + P(B|\bar{A})P(\bar{A}) = 0.15 \times 0.30 + 0.25 \times (1 - 0.30) = 0.045 + 0.175 = 0.22$. So, 22% of adults develop coronary heart disease.

Coronary heart disease and exercise are not independent since $0.22 = P(B) \neq P(B|A) = 0.15$. In words, the knowledge that a person exercises causes us to revise downward the probability that the person develops coronary heart disease, from 22% to 15%. [An alternative solution is to calculate $P(A \cap B)$, which is requested in item b, and show that it does not equal $P(A)P(B) = 0.30 \times 0.22 = 0.066$.]

2b. To answer the first question, we must find $P(A \cap B)$. Since we know $P(B|A)$ and $P(A)$, the easiest way to find the requested probability is to employ the definition of conditional probability: $P(B|A) = P(A \cap B)/P(A)$, implying that $P(A \cap B) = P(A)P(B|A) = 0.30 \times 0.15 = 0.045 = 4.5\%$.

To answer the second question, we must find $P(A \cap \bar{B})$. Since we know $P(\bar{B}|A) = 1 - P(B|A)$ and $P(A)$, the easiest way to find the requested probability is again to employ the definition of conditional probability: $P(\bar{B}|A) = P(A \cap \bar{B})/P(A)$, implying that $P(A \cap \bar{B}) = P(\bar{B}|A)P(A) = (1 - 0.15) \times 0.30 = 0.255 = 25.5\%$. [An alternative solution is to note that A is the union of two mutually exclusive events, $A \cap B$ and $A \cap \bar{B}$. Thus, $0.30 = P(A) = P(A \cap B) + P(A \cap \bar{B}) = 0.045 + P(A \cap \bar{B})$, from which one can solve for $P(A \cap \bar{B})$.]

2c. To answer the first question, we must find $P(\bar{A}|B) = 1 - P(A|B)$. Since we know $P(A \cap B)$ and $P(B)$, the easiest way to find the requested probability is to employ the definition of conditional probability: $P(A|B) = P(A \cap B)/P(B) = 0.045/0.22 = 0.205 = 20.5\%$, whence $P(\bar{A}|B) = 1 - P(A|B) = 0.795 = 79.5\%$. [An alternative solution is to employ Bayes' Theorem, $P(\bar{A}|B) = \{P(B|\bar{A})P(\bar{A})\}/\{P(B|\bar{A})P(\bar{A}) + P(B|A)P(A)\} = \{0.25 \times 0.70\}/\{0.25 \times 0.70 + 0.15 \times 0.30\} = 0.175/0.22 = 0.795 = 79.5\%$.]

To answer the second question, we must find $P(\bar{A}|\bar{B}) = 1 - P(A|\bar{B})$. Since we know $P(A \cap \bar{B})$ and $P(\bar{B})$, the easiest way to find the requested probability is to employ the definition of conditional probability: $P(A|\bar{B}) = P(A \cap \bar{B})/P(\bar{B}) = 0.255/0.78 = 0.327 = 32.7\%$, whence $P(\bar{A}|\bar{B}) = 1 - P(A|\bar{B}) = 0.673 = 67.3\%$. [An alternative solution is to employ Bayes' Theorem, $P(\bar{A}|\bar{B}) = \{P(\bar{B}|\bar{A})P(\bar{A})\}/\{P(\bar{B}|\bar{A})P(\bar{A}) + P(\bar{B}|A)P(A)\} = \{0.75 \times 0.70\}/\{0.75 \times 0.70 + 0.85 \times 0.30\} = 0.525/0.78 = 0.673 = 67.3\%$.]

2d. Let X denote the number who exercise in a randomly selected group of 13 adults. Since 30% of adults exercise, we may regard X as binomial with $n = 13$ trials and success probability $p = 0.30$. We need to find $P(X \ge 5)$. This can be done in SAS. If not using SAS, the next best option is Table 1 in the back of your textbook. Begin by noting that $P(X \ge 5) = P(X = 5) + P(X = 6) + P(X = 7) + \cdots + P(X = 13)$, so all you have to do is look up each of the individual probabilities in Table 1 and then add them. For instance, $P(X = 5) = 0.1803$ and $P(X = 6) = 0.1030$. The final answer is 0.3458. A convenient shortcut is to write $P(X \geq 5) = 1 - P(X \leq 4) = 1 - P(X = 0) - P(X = 1) - \cdots - P(X = 4)$, so that you only have to look up five numbers (instead of nine) in Table 1. If not using SAS or Table 1, you can start with the fact that $P(X \geq 5) = P(X = 5) + P(X = 6) + P(X = 7) + \cdots + P(X = 13)$ and then calculate each of the individual probabilities for yourself using the probability mass function. For example, $P(X = 5) = \frac{13!}{5! \times 8!} \times 0.3^5 \times 0.7^8 = 1287 \times 0.00243 \times 0.05765 = 0.1803$.

Now let Y denote the number who exercise in a randomly selected group of 130 adults. We may regard Y as binomial with $n = 130$ trials and success probability $p = 0.30$. We want to find $P(Y \geq 50)$, but Table 1 will not help and n is too large to proceed using the probability mass function. While we could use SAS, the word "approximate" signals us that invocation of the Central Limit Theorem is permissible. So, we proceed as indicated on page 10 of Lecture 3. Note that Y can be written in the form $X_1 + X_2 + \cdots + X_{130}$, where $X_i = 1$ if the i^{th} person in our sample exercises and $X_i = 0$ otherwise. We have $E[X_1] = E[X_2] = \cdots = E[X_{130}] = 0.3$ and $Var[X_1] = Var[X_2] \cdots = Var[X_{130}] = 0.21$. Let $\mu := 0.3$ and $\sigma^2 := 0.21$. Then the Central Limit Theorem says that $\bar{X} = (X_1 + X_2 + \cdots + X_{130})/130 = Y/130$ is approximately normally distributed with mean $\mu = 0.3$ and variance $\sigma^2/n = 0.21/130 = 0.001615$. Thus, we have

$$
\begin{aligned}
P(Y \geq 50) &= P(50 \leq Y \leq 130) \\
&= P(49.5 \leq Y \leq 130.5) \\
&= P(49.5/130 \leq \bar{X} \leq 130.5/130) \\
&\approx P(49.5/130 \leq \text{Normal random variable with mean 0.3 and variance 0.001615} \leq 130.5/130) \\
&= P\left(\frac{49.5/130 - 0.3}{\sqrt{0.001615}} \leq \text{Standard normal random variable} \leq \frac{130.5/130 - 0.3}{\sqrt{0.001615}}\right) \\
&= \Phi\left(\frac{130.5/130 - 0.3}{\sqrt{0.001615}}\right) - \Phi\left(\frac{49.5/130 - 0.3}{\sqrt{0.001615}}\right) \\
&= \Phi(17.514) - \Phi(2.010) \\
&\approx 1 - 0.978 \\
&= 0.022.
\end{aligned}
$$

The second equality uses the "continuity correction". The fourth approximate equality represents invocation of the Central Limit Theorem. The fifth equality represents standardization. The eighth approximate equality uses the fact that $\Phi(x) \approx 1$ when $x \geq 4$. As a final remark, you do not have to go through all of the steps I did above each time you solve this kind of problem. Indeed, page 10 of Lecture 3 shows that you can begin with $\Phi\left(\frac{130.5/130 - 0.3}{\sqrt{0.001615}}\right) - \Phi\left(\frac{49.5/130 - 0.3}{\sqrt{0.001615}}\right)$, noting that $n = 130$, $p = 0.3$, $a = 50$, and $b = 130$ in the present problem.

2e. Let Y denote the number among 1600 randomly selected non-exercising adults who develop coronary heart disease. Then Y is binomial with $n = 1600$ trials and success probability $p = 0.25$. The expected value of Y is $1600 \times 0.25 = 400$, the variance of Y is $1600 \times 0.25 \times 0.75 = 300$, and the standard deviation of Y is $\sqrt{300} = 17.32$.

So, if 385 non-exercising adults in a community develop coronary heart disease, this is less than what we expect (400). However, the 385 is within one standard deviation of what we expect, so the difference between 385 and what we expect is unremarkable.