STA 570 — Spring 2012 — Dr. Charnigo

Written Assignment 1 Solutions

1a. Among adults on active treatment, the sample mean and sample standard deviation are 112.6 and 34.2, respectively. [Details for pencil-and-paper calculations: We have $n = 49$, $\sum_{i=1}^{n} x_i = 5519$, $\sum_{i=1}^{n} x_i^2 = 677903$, and $(\sum_{i=1}^{n} x_i)^2 = 5519^2 = 30459361$. Hence, we find that $\bar{x} = 5519/49 = 112.6$, $s^2 = (677903 - 30459361/49)/48 = 1172.6$, and $s = \sqrt{1172.6} = 34.2$.]

Among adults on placebo treatment, the sample mean and sample standard deviation are 136.5 and 29.8, respectively. [Details for pencil-and-paper calculations: We have $n = 50$, $\sum_{i=1}^{n} x_i = 6827$, $\sum_{i=1}^{n} x_i^2 = 975643$, and $(\sum_{i=1}^{n} x_i)^2 = 6827^2 = 46607929$. Hence, we find that $\bar{x} = 6827/50 = 136.5$, $s^2 = (975643 - 46607929/50)/49 = 887.4$, and $s = \sqrt{887.4} = 29.8$.]

1b. Among adults on active treatment, the sample median and sample interquartile range are 104 and 133-89=44, respectively. [Details for pencil-and-paper calculations: Since $n = 49$ is odd, the median is the 25th ordered observation; this is 104. To find the 75^{th} percentile, note that $p = 75$ and that $np/100 = 36.75$ is not an integer. The greatest integer less than or equal to 36.75 is $k = 36$, so the 75^{th} percentile is the 37th ordered observation; this is 133. To find the 25^{th} percentile, note that $p = 25$ and that $np/100 = 12.25$ is not an integer. The greatest integer less than or equal to 12.25 is $k = 12$, so the 25^{th} percentile is the 13th ordered observation; this is 89. The interquartile range is the difference between the 75^{th} and 25^{th} percentiles; this is 44.]

Among adults on placebo treatment, the sample median and sample interquartile range are 132.5 and 165-114 = 51, respectively. [Details for pencil-and-paper calculations: Since $n = 50$ is even, the median is the average of the 25th and 26th ordered observations; this is $(131 + 134)/2 = 132.5$. To find the 75^{th} percentile, note that $p = 75$ and that $np/100 = 37.5$ is not an integer. The greatest integer less than or equal to 37.5 is $k = 37$, so the 75^{th} percentile is the 38th ordered observation; this is 165. To find the 25^{th} percentile, note that $p = 25$ and that $np/100 = 12.5$ is not an integer. The greatest integer less than or equal to 12.5 is $k = 12$, so the 25^{th} percentile is the 13th ordered observation; this is 114. The interquartile range is the difference between the 75^{th} and 25^{th} percentiles; this is 51.]

1c. [You should display a printout of the side-by-side box plots.] The sample distribution of low density lipoprotein among adults on placebo treatment is slightly skewed to the right. The distance from the 75^{th} percentile to the 50^{th} percentile is slightly greater than the distance from the 50^{th} percentile to the 25^{th} percentile, and the upper "whisker" is slightly longer than the lower "whisker".

Likewise, the sample distribution of low density lipoprotein among adults on active treatment is slightly skewed to the right.

Regarding central tendency, there is moderate distinction between the two sample distributions: larger values of low density lipoprotein occur more often among adults on placebo treatment, as evidenced by the 25^{th} percentile among those on placebo treatment exceeding the median among those on active treatment. Regarding variability, there is little distinction between the two sample distributions: the interquartile range (i.e., length of the box) is slightly smaller among adults on active treatment, but the range (i.e., length of the box plus whiskers) is slightly larger.

1d. The population percentage of adults on active treatment who have low density lipoprotein measurements between 130 and 160 is $P(130 \leq X \leq 160)$, where X is normal with mean 112.6 and standard deviation 34.2. By standardization, the requested probability is $P\left(\frac{130-112.6}{34.2} \leq Z \leq \frac{160-112.6}{34.2}\right)$, where Z is standard normal. Using Table 3 or SAS, we find that $\Phi\left(\frac{160-112.6}{34.2}\right)-\Phi\left(\frac{130-112.6}{34.2}\right)=\Phi(1.386)-\Phi(0.509)=0.917-0.695=0.222=22.2\%$.

The population percentage of adults on active treatment who have low density lipoprotein measurements above 160 is $P(X>160)=1-P(X\leq 160)=1-P\left(Z\leq\frac{160-112.6}{34.2}\right)=1-\Phi(1.386)=0.083=8.3\%$.

1e. The low density lipoprotein measurement defining the boundary between the top 20% and the bottom 80% in the population of adults on active treatment is the number c such that $P(X\leq c)=0.800$, where X is normal with mean 112.6 and standard deviation 34.2. By standardization, we have $P\left(Z\leq\frac{c-112.6}{34.2}\right)=\Phi\left(\frac{c-112.6}{34.2}\right)=0.800$. On the other hand, Table 3 or SAS shows that $\Phi(0.842)=0.800$, so $0.842=(c-112.6)/34.2$. Solving for c yields 141.4.

2a. Let A denote the event that an adult smokes, and let B denote the event that the adult experiences myocardial infarction within the next 10 years. We are told that $P(A)=0.25$, $P(B|A)=0.10$, and $P(B|\bar{A})=0.05$. Note that we must have $P(\bar{A})=1-0.25=0.75$.

To find $P(B)$, we apply the law of total probability,

$$P(B)=P(B|A)P(A)+P(B|\bar{A})P(\bar{A})=0.10\times 0.25+0.05\times 0.75=0.025+0.0375=0.0625.$$

Thus, 6.25% of adults will experience myocardial infarction.

2b. Our first task is to find $P(A\cap B)$. This equals $P(B|A)P(A)=0.10\times 0.25=0.025$. Thus, 2.5% of adults both are smokers and will experience myocardial infarction.

Our second task is to find $P(\bar{A}\cap\bar{B})$. Noting that $P(\bar{B}|\bar{A})=1-0.05=0.95$, we find that this equals $P(\bar{B}|\bar{A})P(\bar{A})=0.95\times 0.75=0.7125$. So, 71.25% of adults neither are smokers nor will experience myocardial infarction.

2c. To find $P(A|B)$, we apply Bayes' Theorem,

$$P(A|B)=\frac{P(B|A)P(A)}{P(B|A)P(A)+P(B|\bar{A})P(\bar{A})}=\frac{0.025}{0.0625}=0.40.$$

Among adults who will experience myocardial infarction, 40% are smokers.

2d. Let X be the number of smoking adults, among 10 randomly selected, who will experience myocardial infarction. Then X is a binomial random variable with $n=10$ and $p=0.10$. Our task is to find $P(X\geq 2)$. Noting that this equals $1-P(X=0)-P(X=1)$, we can employ SAS or the formula for the probability mass function of a binomial distribution. Either way, we find that $P(X=0)=0.3487$ and $P(X=1)=0.3874$. From this we conclude that $P(X\geq 2)=1-0.3487-0.3874=0.2639$.

Let Y be the number of smoking adults, among 100 randomly selected, who will experience myocardial infarction. Then Y is a binomial random variable with $n=100$ and $p=0.10$. Our task is to approximate $P(Y\geq 20)=P(20\leq Y\leq 100)$. Since $np(1-p)=9\geq 5$, we may invoke the Central Limit Theorem (although perhaps with some misgivings since $9<10$) to obtain the approximate probability. Noting that $\Phi(x)\approx 1$ for $x\geq 4$, we obtain

$$\begin{aligned} P(20\leq Y\leq 100) &\approx \Phi\left(\frac{100.5/100-0.10}{\sqrt{0.10(0.90)/100}}\right)-\Phi\left(\frac{19.5/100-0.10}{\sqrt{0.10(0.90)/100}}\right)\\ &= \Phi(30.17)-\Phi(3.17)\end{aligned}$$

$$
\begin{aligned}
&\approx 1 - 0.9992 \\
&= 0.0008.
\end{aligned}
$$

2e. Let Y denote the number among 4000 randomly selected adults who experience a myocardial infarction. Let us model Y as a binomial random variable with $n = 4000$ and $p = 0.0625$. The expected value of Y is $4000 \times 0.0625 = 250$, the variance of Y is $4000 \times 0.0625 \times 0.9375 = 234.4$, and the standard deviation of Y is $\sqrt{234.4} = 15.3$.

So, if 500 out of 4000 adults in a community experience myocardial infarction, that is far more than the 250 we would expect. Actually, the 500 is more than 16 standard deviations above the 250 we would expect. Hence, this finding is quite surprising. [The probability that a normal random variable is more than 16 standard deviations away from its mean is essentially zero. The same is true for a binomial random variable when n is large enough for us to invoke the Central Limit Theorem.]