STA 580 — Spring 2009 — Dr. Charnigo

Written Assignment 1 Solutions

1a. Among smoking adolescents, the sample mean and sample standard deviation are 3.389 and 0.755, respectively. [Details for pencil-and-paper calculations: We have $n = 23$, $\sum_{i=1}^n x_i = 77.955$, $\sum_{i=1}^n x_i^2 = 276.749$, and $(\sum_{i=1}^n x_i)^2 = 77.955^2 = 6076.982$. Hence, we find that $\bar{x} = 77.955/23 = 3.389$, $s^2 = (276.749 - 6076.982/23)/22 = 0.570$, and $s = \sqrt{0.570} = 0.755$.]

Among non-smoking adolescents, the sample mean and sample standard deviation are 3.994 and 0.902, respectively. [Details for pencil-and-paper calculations: We have $n = 26$, $\sum_{i=1}^n x_i = 103.842$, $\sum_{i=1}^n x_i^2 = 435.074$, and $(\sum_{i=1}^n x_i)^2 = 103.842^2 = 10783.161$. Hence, we find that $\bar{x} = 103.842/26 = 3.994$, $s^2 = (435.074 - 10783.161/26)/25 = 0.813$, and $s = \sqrt{0.813} = 0.902$.]

1b. Among smoking adolescents, the sample median and sample interquartile range are 3.345 and 4.070-2.795 = 1.275, respectively. [Details for pencil-and-paper calculations: Since $n = 23$ is odd, the median is the 12th ordered observation; this is 3.345. To find the 75^{th} percentile, note that $p = 75$ and that $np/100 = 17.25$ is not an integer. The greatest integer less than or equal to 17.25 is $k = 17$, so the 75^{th} percentile is the 18th ordered observation; this is 4.070. To find the 25^{th} percentile, note that $p = 25$ and that $np/100 = 5.75$ is not an integer. The greatest integer less than or equal to 5.75 is $k = 5$, so the 25^{th} percentile is the 6th ordered observation; this is 2.795. The interquartile range is the difference between the 75^{th} and 25^{th} percentiles; this is 1.275.]

Among non-smoking adolescents, the sample median and sample interquartile range are 3.9725 and 4.500-3.211 = 1.289, respectively. [Details for pencil-and-paper calculations: Since $n = 26$ is even, the median is the average of the 13th and 14th ordered observations; this is $(3.96 + 3.985)/2 = 3.9725$. To find the 75^{th} percentile, note that $p = 75$ and that $np/100 = 19.5$ is not an integer. The greatest integer less than or equal to 19.5 is $k = 19$, so the 75^{th} percentile is the 20th ordered observation; this is 4.500. To find the 25^{th} percentile, note that $p = 25$ and that $np/100 = 6.5$ is not an integer. The greatest integer less than or equal to 6.5 is $k = 6$, so the 25^{th} percentile is the 7th ordered observation; this is 3.211. The interquartile range is the difference between the 75^{th} and 25^{th} percentiles; this is 1.289.]

1c. [You should display a printout of the side-by-side box plots.] The sample distribution of forced expiratory volume among non-smoking adolescents is roughly symmetric, arguably a little right skewed. The upper "whisker" is more than twice as long as the lower "whisker", but on the other hand the distance from the 25^{th} percentile to the 50^{th} percentile is slightly greater than the distance from the 50^{th} percentile to the 75^{th} percentile.

The sample distribution of forced expiratory volume among smoking adolescents is roughly symmetric, arguably a little right skewed. The upper "whisker" is slightly longer than the lower "whisker", and the distance from the 50^{th} percentile to the 75^{th} percentile is slightly greater than the distance from the 25^{th} percentile to the 50^{th} percentile.

Regarding central tendency, there is moderate distinction between the two sample distributions: larger values of forced expiratory volume occur more often among non-smoking adolescents. Regarding variability, there is little distinction between the two sample distributions.

1d. The population percentage of smoking adolescents who have forced expiratory volume measurements between 3.500 and 4.000 is $P(3.500 \leq X \leq 4.000)$, where X is normal with mean 3.389 and standard deviation 0.755. By standardization, the requested probability is $P\left(\frac{3.500-3.389}{0.755} \leq Z \leq \frac{4.000-3.389}{0.755}\right)$, where Z is standard normal. Using Table 3 or SAS, we find that $\Phi\left(\frac{4.000-3.389}{0.755}\right) - \Phi\left(\frac{3.500-3.389}{0.755}\right) = \Phi(0.809) - \Phi(0.147) = 0.791 - 0.558 = 0.233 = 23.3\%$.

The population percentage of smoking adolescents who have forced expiratory volume measurements above 4.000 is $P(X > 4.000) = 1 - P(X \leq 4.000) = 1 - P\left(Z \leq \frac{4.000-3.389}{0.755}\right) = 1 - \Phi(0.809) = 0.209 = 20.9\%$.

1e. The forced expiratory volume measurement defining the boundary between the top 20 percent and bottom 80 percent in the population of smoking adolescents is the number c_{80} such that $P(X \leq c_{80}) = 0.80$, where X is normal with mean 3.389 and standard deviation 0.755. By standardization, we have $P\left(Z \leq \frac{c_{80}-3.389}{0.755}\right) = \Phi\left(\frac{c_{80}-3.389}{0.755}\right) = 0.80$. On the other hand, Table 3 or SAS shows that $\Phi(0.842) = 0.80$, so $0.842 = (c_{80} - 3.389)/0.755$. Solving for c_{80} yields 4.025.

The forced expiratory volume measurement defining the boundary between the top 10 percent and bottom 90 percent in the population of smoking adolescents is the number c_{90} such that $P(X \leq c_{90}) = 0.90$. We have $P\left(Z \leq \frac{c_{90}-3.389}{0.755}\right) = \Phi\left(\frac{c_{90}-3.389}{0.755}\right) = 0.90$. On the other hand, Table 3 or SAS shows that $\Phi(1.282) = 0.90$, so $1.282 = (c_{90} - 3.389)/0.755$. Solving for c_{90} yields 4.357.

2a. Let A denote the event that a man aged 50-59 has diastolic blood pressure greater than 100, and let B denote the event that the man will be alive in ten years. We are told that $P(A) = 0.20$, $P(B|A) = 0.80$, and $P(B|\bar{A}) = 0.90$. Note immediately that we must have $P(\bar{A}) = 1 - 0.20 = 0.80$, $P(\bar{B}|A) = 1 - 0.80 = 0.20$, and $P(\bar{B}|\bar{A}) = 1 - 0.90 = 0.10$.

Our first task is to find $P(B)$. We apply the law of total probability,

$$P(B) = P(B|A)P(A) + P(B|\bar{A})P(\bar{A}) = 0.80 \times 0.20 + 0.90 \times 0.80 = 0.16 + 0.72 = 0.88.$$

Thus, 88% of men aged 50-59 will be alive in ten years.

Our second task is to find $P(\bar{B})$. This equals $1 - P(B) = 1 - 0.88 = 0.12$. So, 12% of men aged 50-59 will not be alive in ten years.

2b. Our first task is to find $P(\bar{A} \cap B)$. This equals $P(B|\bar{A})P(\bar{A}) = 0.90 \times 0.80 = 0.72$. Thus, 72% of men aged 50-59 satisfy both conditions (i) and (ii).

Our second task is to find $P(\bar{A} \cap \bar{B})$. This equals $P(\bar{B}|\bar{A})P(\bar{A}) = 0.10 \times 0.80 = 0.08$. Thus, 8% of men aged 50-59 satisfy both conditions (iii) and (iv).

2c. Our first task is to find $P(A|B)$. We apply Bayes' Theorem,

$$P(A|B) = \frac{P(B|A)P(A)}{P(B|A)P(A) + P(B|\bar{A})P(\bar{A})} = \frac{0.16}{0.88} = 0.182.$$

Among men aged 50-59 who will be alive in ten years, 18.2% have diastolic blood pressure greater than 100.

Our second task is to find $P(A|\bar{B})$. We apply Bayes' Theorem,

$$P(A|\bar{B}) = \frac{P(\bar{B}|A)P(A)}{P(\bar{B}|A)P(A) + P(\bar{B}|\bar{A})P(\bar{A})} = \frac{0.20 \times 0.20}{0.20 \times 0.20 + 0.10 \times 0.80} = 0.333.$$

Among men aged 50-59 who will not be alive in ten years, 33.3% have diastolic blood pressure greater than 100.

2d. Let X be the number of men aged 50-59, among 8 randomly selected, with diastolic blood pressure greater than 100. Then X is a binomial random variable with $n = 8$ and $p = 0.20$. Our task is to find

$P(X \geq 2)$. Noting that this equals $1 - P(X = 0) - P(X = 1)$, we can employ Table 1 of Rosner. Since $P(X = 0) = 0.1678$ and $P(X = 1) = 0.3355$, we conclude that $P(X \geq 2) = 1 - 0.1678 - 0.3355 = 0.4967$.

Let Y be the number of men aged 50-59, among 80 randomly selected, with diastolic blood pressure greater than 100. Then Y is a binomial random variable with $n = 80$ and $p = 0.20$. Our task is to approximate $P(Y \geq 20) = P(20 \leq Y \leq 80)$. Since $np(1-p) = 12.8 \geq 10$, we are comfortable invoking the Central Limit Theorem to obtain the approximate probability. Noting that $\Phi(x) \approx 1$ for $x \geq 4$, we obtain

$$\begin{aligned}
P(20 \leq Y \leq 80) &\approx \Phi\left(\frac{80.5/80 - 0.20}{\sqrt{0.20(0.80)/80}}\right) - \Phi\left(\frac{19.5/80 - 0.20}{\sqrt{0.20(0.80)/80}}\right) \\
&= \Phi(18.03) - \Phi(0.978) \\
&\approx 1 - 0.836 \\
&= 0.164.
\end{aligned}$$

2e. Let Y denote the number among 1000 randomly selected men aged 50-59 who have diastolic blood pressure greater than 100. Then Y is a binomial random variable with $n = 1000$ and $p = 0.20$. The expected value of Y is $1000 \times 0.20 = 200$, the variance of Y is $1000 \times 0.20 \times 0.80 = 160$, and the standard deviation of Y is $\sqrt{160} = 12.65$.

So, if 300 out of 1000 men aged 50-59 in a community have diastolic blood pressure above 100, that is more than the 200 we would expect. Actually, the 300 is more than seven standard deviations greater than the 200 we would expect, so this finding is quite surprising. [The probability that a normal random variable is more than seven standard deviations away from its expected value 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.]