SOLUTIONS TO BIOSTATISTICS PRACTICE PROBLEMS

BIOSTATISTICS
DESCRIBING DATA, THE NORMAL DISTRIBUTION

SOLUTIONS

1.

a. To calculate the mean, we just add up all 7 values, and divide by 7. In fancy statistical notation, $\overline{X} = \frac{\sum_{i=1}^{7} X_i}{7} =$

$\frac{12.0 + 9.5 + 13.5 + 7.2 + 10.5 + 6.3 + 12.5}{7} = 10.2$ years.

b. To calculate the sample median, first rank the values from lowest to highest:

6.3 7.2 9.5 10.5 12.0 12.5 13.5

Since there are 7 values, an odd number, we can simply select the middle value, 10.5, to calculate the sample median.

b. It's a good thing we have calculated the sample mean- we ned this to calculate the sample standard deviation! Recall the formula for SD:

$$SD = \sqrt{\frac{\sum_{i=1}^{7}(X_i - \overline{X})^2}{7-1}} = \sqrt{\frac{(12.0-10.2)^2 + (9.5-10.2)^2 + ...(12.5-10.2)^2}{6}}$$

= 2.71 years

d. 1. sample mean – Would decrease, as the lowest value gets lower, pulling down the mean.

2. sample median – Would remain the same since the middle value is still 10.5 By replacing the 6.3 with 1.5, the rank of the 7 values is not affected.

3. sample standard deviation – Would increase. Because our minimum value has now gotten smaller, while the rest of the data points remain unchanged, the spread or variability in our data has increased; since SD is a measure of spread, it too will increase (prove it to yourself!).

e. While the sample mean and sample standard deviations of the 14

observation will likely be different than the respective quantities from the sample with seven observations, it is not possible to predict how the values will differ (at least without seeing the data!) as neither the sample mean nor the sample mean values are linked explicitly to sample size. Recall, these sample quantities are estimating the same underlying population parameters whether they are computed from a sample of size 7, 14, or 1,000.

In this example, the sample mean of the 14 observations is 9.9 years, smaller than the sample mean of 10.5 years for the original seven observations. The sample standard deviation of the 14 observations is 3.1 years, larger than the sample standard deviation of 2.7 years for the original seven observations.

2.

This question is really about is calculating standard normal scores. Recall,

$$Z = \frac{\text{Observed} - \text{Mean}}{\text{SD}}$$

a. The boy who is 170 cm tall is above average by $\frac{170-146}{8} = \frac{24}{8} = 3$ SDs.

b. The boy who is 148 cm tall is above average by $\frac{148-146}{8} = \frac{2}{8} = .25$ SDs.

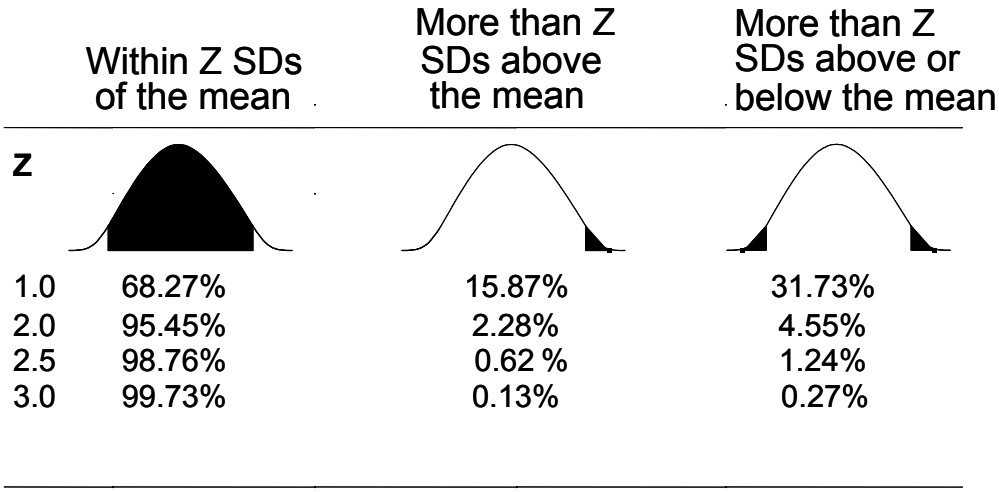
c. A third boy was 1.5 SDs below the average height. He was 146 – 1.5*8 = 146-12 = 134 cm tall.

d. If a boy was within 2.25 SD's of average height, the shortest he could be is 146 – 2.25*8 = 128 cm tall, and the tallest he could be is 146 + 2.25*8 = 164 cm tall.

e.
1. 150 cm – about average (.5 SDs above mean)
2. 130 cm - unusually short (2 SDs below mean)
3. 165 cm –unusually tall (2.4 SDs above mean)
4. 140 cm – about average (.75 SDs below mean)

3.

These questions refer to the table relating normal scores to area (percent population) under the density curve.



z	Within Z SDs of the mean	More than Z SDs above the mean	More than Z SDs above or below the mean
1.0	68.27%	15.87%	31.73%
2.0	95.45%	2.28%	4.55%
2.5	98.76%	0.62 %	1.24%
3.0	99.73%	0.13%	0.27%

a. If individuals considered "abnormal" have glucose levels outside of 1 standard deviation of the mean (above or below) , then approximately 32% (31.73 to be exact) of the individuals will need to be retested. The "normal range" of glucose level would range from (90 – 38) mg/dL to (90 + 38) mg/dL, or from 52 mg/dL to 128 mg/dL.

b. If individuals considered "abnormal" have glucose levels outside of 2 standard deviations of the mean (above or below) , then approximately 5% (4.55 to be exact of the individuals will need to be retested. The "normal range" of glucose levels would range from (90 – 2*38) mg/dL to (90 + 2*38) mg/dL, or from 14 mg/dL to 166 mg/dL.

4. **A** is the correct answer. Remember, in order to calculate the median, you must first order the values in the sample from lowest to highest. Doing so yields:

110 116 124 132 168

This sample is of size 5, and odd number, so the middle value of 124 is the sample median.

5. **C** is the correct answer. Here the sample mean, $\overline{X}$ = 64 inches, and the SD = 5 inches. Since we are given that the distribution of heights in 12 year old boys is normal, we know that 2 SDs above or below the sample mean will give us an interval containing approximately 95% of the heights in the sample. This interval would run from 64 – 2*5 to 64 + 2*5, or 54 inches to 74 inches.

8. **D** is the correct answer. Remember, whether we calculate sample SD from a sample of 1,000 or a sample of 3,000, both are estimating the same quantity- the population standard deviation. These two estimates should be about the same, and we cannot predict which will be larger.

BIOSTATISTICS
SAMPLING DISTRIBUTIONS, CONFIDENCE INTERVALS

SOLUTIONS

QUESTION 1.

a. It can not be determined which researcher will get the bigger standard deviation – both sample SDs from the sample with n = 100, and with n = 1,000 are estimating the same quantity – the population standard deviation. Therefore, the two estimates should be similar, and it is not possible to tell which will be larger , prior to calculating the values. Standard deviation does not depend on sample size, but will vary from random sample to random sample.

b. Standard error does depends on sample size, however; the larger the sample size, the smaller the standard error of the mean (SEM). Therefore, the SEM calculated from the sample with n = 1,000 will likely be smaller the SEM calculated from the sample with n = 100.

c. Extreme values are more likely in larger samples – therefore, the investigator with the sample of n = 1,000 is more likely to have the tallest man.

d. Extreme values are more likely in larger samples – therefore, the investigator with the sample of n = 1,000 is more likely to have the shortest man.

QUESTION 2.

a. In this study of 60 year old women with glaucoma, n = 200, $\overline{X}$=140 mmHg, and SD = 25mm Hg. Since n is large, we can use the Central Limit Theorem to aid us in constructing a 95% confidence interval for the population mean blood pressure, μ. Its "business as usual" via the formula:

$$\overline{X} \pm 2*(\text{SEM}), \text{ where SEM} = \frac{SD}{\sqrt{n}} = \frac{25}{\sqrt{200}} = 1.77 \text{ mm Hg}$$

Plugging in our sample values gives us:

$140 \pm 2*(1.77) \rightarrow$ (136.5 mm Hg, 143.5 mmHg)

b. If a second study yielded the same sample statistic values, but were done with 100 women, what would happen to the width of the 95% confidence interval? Well, we know since this sample is smaller than the previous example, the SEM will be larger, leading to a wider confidence interval. In non-mathematical terms, our sample contains less information than a sample of 200 women, and therefore will yield a less precise (more uncertain) estimate of the population mean. The proof is as follows:

$$\bar{X} \pm 2*(\text{SEM}), \text{ where SEM} = \frac{SD}{\sqrt{n}} = \frac{25}{\sqrt{100}} = 2.5 \text{ mm Hg}$$

Plugging in our sample values gives us:

$140 \pm 2*(2.5) \rightarrow$(135 mm Hg, 145 mm Hg)

3. **A** is the correct answer. Here the sample is of size n = 500, which is large enough to ensure that the Central Limit Theorem kicks in . By the Central Limit theorem, the sampling distribution the of the sample mean from a sample of 500 will be normally distributed.

4. **D** is the correct answer. No general statement can be made as we do not know whether or not the sample of 200 women who agreed to participate from the original random sample of 300 was still representative of all 18 year old females. If these 200 women are inherently different from the other 100 non-participants, the results shown are biased.

5. **B** is the correct answer. The more confident we want to be, the wider our confidence interval. Ninety-nine percent confidence is higher than ninety-five percent confidence; therefore the 95% confidence interval is not so wide as the 99% confidence interval.

6. **C** is the correct answer. The sample is random, i.e. representative – therefore, the sample distribution should mimic the larger population distribtion, which is right-skewed.

7. **B** is the correct answer. We would expect the two samples to have SD values that are similar. but, recall that the standard error (SE) is the standard deviation divided by the square-root of the sample size. Because Sample B is much larger (N=2000) than Sample A (N=100), we would then expect the SE of Sample B to be smaller than the SE of Sample A.

8. **A** is the correct answer. This question is asking about the shape of the sampling distribution of the sample mean, based on samples of size 100: As the sample size is large (n=100) the Central Limit Theorem applies and the sampling distribution should be normal: hence a histogram based on the sample means of 3,000 random samples should be approximately normal : note it is not the number of samples that determines whether the Central Limit Theorem "kicks in " but the size of each of the samples.

9. **B** is the correct answer. A very straightforward application of the formula $\bar{x} \pm 2SE(\bar{x})$ - you are given sample s.d. of 25 ounces, and know that the sample size is 100 – the estimated standard error of the sample mean is $s/\sqrt{n} = 25/\sqrt{100} = \frac{25}{10} = 2.5$. all you need do is plug in:

$$\bar{x} \pm 2SE(\bar{x}) = 120 \pm 2(2.5) = 120 \pm 5 = (115, 125).$$

10. The correct answer is **C**. In this sample, $\hat{p}$, the estimated proportion of Baltimoreans with health insurance, is $\frac{650}{1000} = .65$, or 59%. As 1000*.65*(1-.65)≈228, we can use the normal approximation for the 95% CI for a population proportion, given info from a random sample. The standard error of this estimate is $\sqrt{\frac{(.65)(1-.65)}{1000}} \approx .015$ Applying the formula $\hat{p} \pm 2SE(\hat{p})$, yields as 95% confidence interval of (.62,.68), or 62% to 68% for the proportions of Baltimoreans with health insurance.

BIOSTATISTICS
HYPOTHESIS TESTING

SOLUTIONS

QUESTION 1. (answers will vary, of course)

A sample of 107 patients with one-vessel coronary artery disease was given percutaneous transluminal coronary angioplasty (PTCA). Patients were given exercise tests at baseline and after 6 months of follow up. Exercise tests were performed up to maximal effort until symptoms, such as angina, were present. A paired t-test was used to assess whether there was a significant change in duration of exercise after 6 months of PTCA treatment, and a 95% confidence interval was constructed for the mean difference (after - before) in exercise duration.

Exercise duration increased 2.1 minutes (95% CI 1.5 – 2.7 minutes) on average after the PTCA treatment. There was evidence that exercise was significantly higher after exposure to 6 months of PTCA treatment ($p < .001$). As there was no comparison group of individuals not receiving PTCA, we cannot prove PTCA as the cause of this increase in exercise duration. It is not known whether there would have been a similar 6-month change without PTCA.

QUESTION 2.(answers will vary, of course)

A sample of 171 women between 75 and 80 years old were classified into one of two groups based on whether the subject took Vitamin E supplements at the time of enrollment. Each woman was subsequently given a test to measure cognitive ability. Higher scores on this test indicate better cognition. A two sample t-test was used to compare mean cognition test scores between the two groups of women, and a 95% confidence interval for the difference was constructed.

The average test score amongst the women taking vitamin E was 27 (sd = 6.9) as compared to a mean score of 24 (sd = 6.2) for women not taking the supplements. The average score difference between the two groups is 3.0 points (95% CI 1.0 – 5.0 points). The cognition scores were statistically significantly the women taking Vitamin E supplements. ($p < .01$) As the women were randomized to the vitamin and placebo groups, the results of this study strongly suggest a positive relationship between Vitamin E consumption and increased cognition amongst elderly women.

If the study was not randomized, and women self-selected to be in the Vitamin E group, the statistical comparisons could still be made, but the scientific conclusions would be harder to make without further analyses (the type of which is coming in 612!). The write-up of the results would be similar, but the last sentence in the second paragraph in part 1 would change to something like the following paragraph:

"However, because women were not randomized to take the vitamin supplements but were self-selected into the vitamin exposure groups, it is not possible to attribute the higher scores to Vitamin E. It is possible that the women taking Vitamin E differed on multiple factors when compared to the women who were not taking the supplement. The difference in test scores could be attributable, at least partially, to some of these other factors."

3. The correct answer is **C**. Because the 95% confidence interval does not include zero, we would reject null hypothesis of a true mean difference of zero at the $\alpha = .05$ level. Testing

$$\text{Ho: } \mu_2 - \mu_1 = 0$$

is equivalent to testing Ho: $\mu_2 = \mu_1$, the equality of the two means.

4. The correct answer is **A**. The data collected in this example is paired data, and a p-value would be obtained from the paired t-test. The test statistics would be:

$$t = \frac{observed\_mean\_difference}{standard\_error\_of\_mean\_difference} = \frac{\overline{X}_{diff}}{se(\overline{X}_{diff})}$$

where $\overline{X} = 15$, and $se(\overline{X}) = \frac{40}{\sqrt{100}} = \frac{40}{10} = 4$.

So Z = 15/4 = 3.75. Since t > 2, we know p < .05

5. The correct answer is **B**. The standard error of a statistic is a measure of the variability of that statistic across different sample sizes – the variability of the sampling distribution. Therefore, the standard error of a statistic is the standard deviation of the sampling distribution.

6. B is the correct answer. Despite the fact that we are computing before/after differences we ultimately are comparing these differences between two independent groups: those randomized to the diet program, and those randomized to exercise. Since we are making a comparison of mean changes between two independent groups, the appropriate test is the 2 sample unpaired t-test.

7. The correct answer is **E**. This is a hard, but important question : choice a is just flat-out incorrect, based on the definition of the p-value, and choices b-c are impossible to ascertain from just a p-value, as it imparts no information about the direction/magnitude and clinical or scientific significance of the results of a study.

8. **A** is the correct answer. As the 95% confidence interval for the mean difference does not include, the resulting p-value would be less than .05.

9. **C** is the correct answer. The chi-squared is the correct statistical test for comparing two population proportions based on information from two (large) samples – both the sample meet the "large sample" criteria.

BIOSTATISTICS
PROPORTIONS

SOLUTIONS

Question 1.

(a) To estimate the 95% confidence interval for each group, we need to know the estimated proportion in each group (150/262 = 0.57 in the vaccine group and 83/134 = 0.62 in the placebo group), and their standard errors. Recall that the formula for the standard error of a proportion is $\sqrt{\frac{p(1-p)}{N}}$, so that the standard error of estimate p in the vaccine group is 0.030 and in the placebo group is 0.042. Now we can implement the formula for the confidence interval for a proportion (for large N): $\hat{p} \pm 2se(\hat{p})$

Plugging into this equation for the vaccine group, we have:

$$(0.57 - 2*0.03,\ 0.587 + 2*0.03) = (0.51,\ 0.63)$$

Plugging into this equation for the placebo group, we have:

$$(0.62 - 2*0.04,\ 0.62 + 2*0.04) = (0.54,\ 0.70)$$

These confidence intervals do overlap in the range of 0.0.54 to 0.63, which seems to be a large fraction of the intervals.

(b) To compute the 95% confidence interval for the difference in proportions, we use the general formula for the confidence interval, where we use the standard error for the difference of proportions provided:

$$(\hat{p}_1 - \hat{p}_2) \pm 2se(\hat{p}_1 - \hat{p}_2)$$

Plugging into this equation:

$$(0.57 - 0.62 - 2*0.05,\ 0.57 - 0.62 + 2*0.05) = (-0.15,\ 0.05).$$

The interpretation of this 95% CI basically suggests that the results from our samples indicates that the vaccine could be associated with a decrease in the proportion of children experiencing at least one episode of AOM of at most 15%, but could also be associated with an **increase** as large as 5%.

(c) The null hypothesis would be that the underlying true proportions of children experiencing at least one episode of AOM are the same for the vaccinated and non-vaccinated children : in other words, there is no relationship between the

influenza vaccine and occurrence of AOM in children. The alternative hypothesis is the underlying true proportions of children experiencing at least one episode of AOM in the follow-up period are different for the vaccinated and non-vaccinated children : in other words, there is a relationship between the influenza vaccine and occurrence of AOM in children.

The p-value for testing this is greater than 0.05. We know this because the 95% CI for the difference in proportions between the two groups includes 0.

(d) This is a randomized study (it says so right in the question!): because randomization helps to equalize the two groups of children in terms of other characteristics (age, health, etc…) it makes it "easier" to attribute any differences found in AOM episodes to the vaccine, or attribute non-difference to the lack of vaccine efficacy in affecting AOM (as is this case with these study results). In other words, the randomized study design indicates that the lack of association found between the flu vaccine and episodes of AOM is not because the vaccine is really associated with AOM and this association is being "hidden" by other characteristics of the children that differ between the treatment and placebo group.

2. (a) Using the same approach as in question 1, the large sample approximation for the 95% confidence intervals for the proportions in the two groups are (0.58, 0.86) in the pet group and (0.88, 1.01) in the non-pet group. But, we see a problem here! One of these confidence intervals overlaps 1! This is impossible because proportions must take values between 0 and 1. So, in thinking about the large sample approximation again, perhaps it isn't such a good idea: the sample sizes in the two groups are only 39 and 53. An 'exact' approach is needed in this case.

(c) Most likely, it was not possible for the researchers to randomize these 92 patients to a pet ownership group (for practical reasons, and ethical issues for both the patients and the animal!). Ergo, it is not possible to attribute the increase survival in the pet-ownership group to owning a pet. The pet ownership-survival relationship may be fueled by other differences existing between the pet owners and those without pet: for example, differences in level and depression status. Further analysis would be necessary to help control for some of these potential difference when estimating the pet ownership/survival relationship.

3. The best answer is **E**. Statistically speaking, the question of interest reduces to testing for a difference in the proportion of individuals who quit smoking on program A as compared to program B. This limits our possible choices to (b) and (e). Because of the small sample sizes (the best answer is (e), Fisher's Exact Test.

4. The correct answer is **C**. In order to complete the story you would also need to have estimates of the standard deviations of the birth weight measurements in each of the two groups of infants being compared.

BIOSTATISTICS
LINEAR REGRESSION

SOLUTIONS

1. The correct answer is **D**. The coefficient for weight is 0.10, indicating that the expected difference in SBP for two children of the same age who differ by one ounce in birth weight is 0.10 mmHg, the heavier child compared to the lighter child. So if we are comparing a child who weighed 120 ounces at birth to a child who weighed 90 ounces at birth, and both children were the same age, the estimated expected (mean) difference in SBP is 30*.10mmHG = 3.0 mmHg.

2. The correct answer is **B**. Well, the coefficient for age is an estimate of the difference in SBP between 2 infants with the same birthweight who differ by one day in age: the older compared to the younger will have SBP of 4 mmHg higher, one average (95% CI: 4±2*.6 = (2.8, 5.3)). To get the corresponding CI for the difference in SBP for equally weighted infants who differ by 2 days In age, we can just double the endpoints for the previously computed CI.

3. **C** is the correct answer. All that's being changed is the units in which the weight is measured – the measurements themselves are not being altered, just the units in which the values are expressed – ergo, the correlation between SBP and a child's age and weight should not be altered.

4. The correct answer is D. Recall, r tells us something about both the strength and the direction of a relationship. It is the appropriately signed value of $\sqrt{R^2}$. Since the slope is negative, we know r must be negative: hence it is $-\sqrt{0.76}$ = -.87.

5. D is the correct answer. This model relates wage as a function of a subject's sex, union membership status, and years of education via the following equation - $y = -0.3 + -1.9 * sex + 1.9 * union\_member + 0.76 * years\_education$. Male, non-union workers with 12 years of education have the following predictor values: sex = 0, union_member = 0, years_education = 12, so the resulting predicted value is $y = -0.3 + -1.9 * 0 + 1.9 * 0 + 0.76 * 12 = -0.3 + 9.12 = 8.82\_dollars/hr$.

6. A is the correct answer. What this is asking for in more user friendly terms is the 95% confidence interval for the coefficient of union_member in a model that also includes sex and years_education: recall the interpretation of this coefficient is that it estimates the adjusted mean hourly wage for union members compared to non-union members of the same sex and same years of education (ie: adjusted for sex and years of education). So this estimated coefficient is 1.9, and its standard error is 0.5: as we have a large sample, we can just employ the $\hat{b}_1 \pm 2SE(\hat{b}_1) = 1.9 \pm 2 * 0.5$ method to get the 95% CI of (.90,

2.90), or $0.90 to $2.90 per hour.

7. (a)

Two possible phrasings:

- a 1 year increase on age is associated with an .02 liter increase in FEV, on average
- In two groups of men who differ by one year of age, the older groups will have average FEV of .02 liters higher than the younger group

(b) Since we have a sample of 200 men, we need not fuss with pesky t-corrections, and can just employ the general formula $\hat{b}_1 \pm 2SE(\hat{b}_1)$, which gives a 95% CI of 0.02±2*(.005), giving a 95% CI of (.01, .03). So based on this sample of 200 men, the true increase associated with an 1 year increase in age is between .01 liters and .03 liters. (with 95% confidence, etc..)

(c) The strength of the linear association can not be assessed without viewing a scatterplot and seeing an estimated correlation coefficient.

(d) To find the difference between 60 and 50 year old men, we simply multiply the coefficient for age (representing a 1 year difference) by 10: 0.02*10 = 0.2.

(e) No – these results are based on information from a sample of men aged 20 – 60: The results are not necessarily applicable to men outside this age range.

BIOSTATISTICS
SURVIVAL ANALYSIS

SOLUTIONS

1. Survival analysis would be used. The outcome variable is "time to AIDS", where some of the times are censored. When we have time to event data, the best choice is to use survival, and we could more specifically use the Kaplan-Meier approach to estimate the survival curve and the median time to AIDS.

2. The correct answer is **D**. The median time (i.e. the time at which $S(t) = 0.50$) is not shown on the plot. We see that S(t) only ranges from 0.90 to 1.00, meaning that the median time is not within the 180 days.

3. The correct answer is **B**. At 100 days, the height of the survival curve, S(t), is approximately 0.94.

4. **C** is the correct answer. By taking the average, we are treating the *censored* times as *observed* times of death. But, when an observation is censored, we know that the true time of death must be after the censored time of death. So, the censored times are underestimates of the true survival times. As a result, taking the mean of both the observed time of death and censored times of death, we get an underestimate of the true mean survival time.