Unit II Summary Exercises

II.1 Let Y denote the fruit weight of a nectarine. Suppose Nancy wants to know how weights in her orchard compared from this season to the last. In particular, suppose she is interested in the averages μ_1 and μ_2. You may assume that Nancy has taken several statistics courses and knows a lot about statistics, including how to interpret confidence intervals and hypothesis tests. She usually chooses to limit her type I error rate to 0.05. You have random samples of fruit from each season and are to analyze the data and write a report. You plan to report to Nancy the two sample means, but you aren't sure what to say about how they compare. You seek advice from four persons:

Rudd says, "Conduct an $\alpha = 0.05$ test of H_0: $\mu_1 = \mu_2$ versus H_A: $\mu_1 \neq \mu_2$ and tell Nancy whether or not you reject H_0 at the $\alpha = 0.05$ level."

Linda says, "Report a 95% confidence interval for $\mu_1 - \mu_2$."

Steve says, "Conduct a test of H_0: $\mu_1 = \mu_2$ versus H_A: $\mu_1 \neq \mu_2$ and report to Nancy the P-value from the test."

Gloria says, "Compare $\bar{y}_1$ to $\bar{y}_2$. If $\bar{y}_1 > \bar{y}_2$ then test H_0: $\mu_1 = \mu_2$ versus H_A: $\mu_1 > \mu_2$ using $\alpha = 0.05$ and tell Nancy whether or not you reject H_0. If $\bar{y}_1 < \bar{y}_2$ then test H_0: $\mu_1 = \mu_2$ versus H_A: $\mu_1 < \mu_2$ using $\alpha = 0.05$ and tell Nancy whether or not you reject H_0."

Rank the four pieces of advice from worst to best and explain why you rank them as you do. That is, explain what makes one better than another.

II.2 Each of 41 students at a college was asked to calculate their "ecological footprint"—the number of hectares required to support their existence, taking into account such things as the land needed to produce the food they eat and so forth. For 27 women, the average was 6.59, and the standard deviation was 3.89 hectares. For 14 men, the average was 3.96, and the standard deviation was 1.06 hectares.

(a) Use these data to conduct a t test of the null hypothesis that there is no difference between the population means for men and women. Use a two-sided alternative and use $\alpha = 0.02$. (*Note:* There are approximately 32 degrees of freedom here for the t test.) Provide all steps: Construct the value of the test statistic, give bounds on the P-value, and state your conclusion regarding H_0.

(b) In nontechnical language, explain to a nonstatistician what this means about the ecological footprints of men and women at this college. Be specific.

(c) Suppose differences in ecological footprints smaller than 0.7 hectares are considered to be not ecologically important. Compute a confidence interval for the difference in mean global footprint between men and women and discuss whether the difference can be considered "ecologically important."

II.3 After seeing the sample means and SDs from Question II.2, Norman became concerned that a t test might not be appropriate here, so he wants Rebecca to do a different analysis.

(a) Why is Norman concerned? That is, what is it about the data that would make someone question whether a t test would be valid?

(b) If Norman is right, then what should Rebecca do to analyze her data?

(c) Alana says that Norman shouldn't be so worried and should let Rebecca go ahead with a t test. On what grounds can Alana look at the information in Question II.2 and justify using a t test (with the original data)?

II.4 Suppose someone constructs a 95% confidence interval for $\mu_1 - \mu_2$ and gets (1.3,12.7). If we were to test H_0: $\mu_1 = \mu_2$ versus H_A: $\mu_1 \neq \mu_2$ with $\alpha = 0.10$, would we reject H_0? Or can we not tell from the information given? Explain your reasoning.

II.5 Nitric oxide is sometimes given to newborns who experience respiratory failure. In one experiment, nitric oxide was given to 114 infants. This group was compared to a control group of 121 infants. The length of hospitalization (in days) was recorded for each of the 235 infants. The mean in the nitric oxide sample was $\bar{y}_1 = 36.4$; the mean in the control sample was $\bar{y}_2 = 29.5$. A 95% confidence interval for $\mu_1 - \mu_2$ is $(-2.3, 16.1)$, where μ_1 is the population mean length of hospitalization for infants who get nitric oxide and μ_2 is the mean length of hospitalization for infants in the control population For each of the following, say whether the statement is true or false and explain why.

(a) 95% of infants who experience respiratory failure would have their hospital stays altered by between a 2.3-day decrease to a 16.1-day increase if they received nitric oxide.

(b) We are 95% confident that nitric oxide has no effect on the length of hospitalization.

(c) We are 95% confident that if nitric oxide affects the length of hospitalization, its effect is less than 16.1 days, on average.

II.6 Researchers took random samples of subjects from two populations and applied a two-sample t test to the data using $\alpha = 0.10$; the P-value for the test, using a nondirectional alternative, was 0.06. For each of the following, say whether the statement is true or false and explain why.

(a) There is a 6% chance that the two population distributions actually are the same.

(b) If the two population distributions actually are the same, then a difference between the two samples as extreme as the difference that these researchers observed would only happen 6% of the time.

(c) If a new study were done that compared the two populations, there is a 6% probability that H_0 would be rejected again.

(d) If $\alpha = 0.05$ and a directional alternative were used, and the data departed from H_0 in the direction specified by the alternative hypothesis, then H_0 would be rejected.

II.7 Researchers measured blood levels of the hormone Androstenedione (Andro) in each of 24 women. Among 12 women who had recently fallen in love, the sample mean was 2.1 ng/ml; the SD of these data was 0.7. Among 12 women who had *not* recently fallen in love, the mean was 1.9, and the SD was 0.7. Formula (6.7.1) yields 22 degrees of freedom.

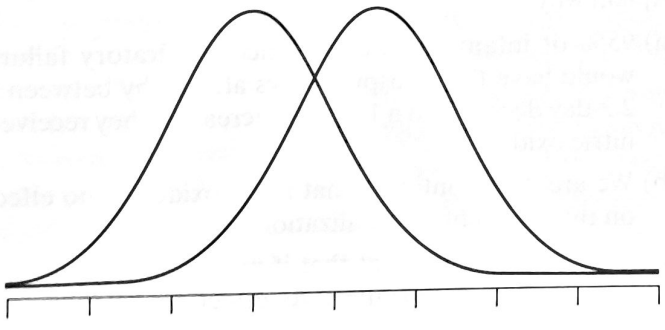
(a) Conduct a two-sample t test, with nondirectional alternative; show all steps, including bounds on the P-value. If $\alpha = 0.05$, do you reject H_0? Why or why not?

(b) A 90% confidence interval for the difference in mean Andro levels for the two populations is $(-0.39, 0.79)$ ng/ml. Suppose a change of 0.4 ng/ml is medically important. Are these results medically important? Briefly explain your reasoning.

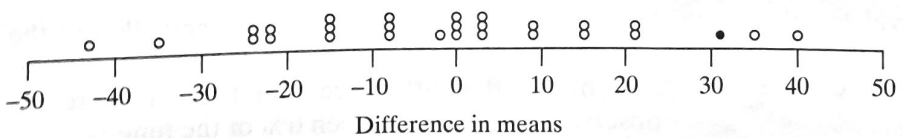
II.8 Consider the process of planning for adequate power when designing an experiment for which a two-sample t test will be used.

(a) Suppose the effect size is constant and the chosen alpha level is also constant (a) (e.g., at 0.05). If the sample sizes for the two groups go up from 12 to 18, how does this affect power? That is, does power go down, stay the same, or go up? Why? How is this related to the standard error?

(b) The graph below shows two normally distributed populations. Calculate the effect size here.



II.9 A researcher did a randomization test to compare men and women on a variable Y. The null hypothesis is that men and women are the same, and the alternative is that they are different. There are 28 possible randomizations of the data. The graph below shows the statistic "Difference in means" for each randomization. This includes the original assignment of observations to the male and female groups, which had a difference in means of 31.



(a) What is the randomization test P-value?

(b) Using $\alpha = 0.10$, is there statistically significant evidence that men and women differ with respect to the mean of variable Y?

(c) The difference in sample means between men and women was 31 $(\bar{y}_{men} - \bar{y}_{women})$. What is the randomization test P-value for the test that considers the alternative hypothesis $\mu_{men} > \mu_{women}$?

II.10 Consider studying whether drug A and drug B are equally effective in lowering blood pressure. For each of (a), (b), and (c) answer two questions (you do not need to explain your answers):

(I) What kind of paired design is this (i.e., What is the nature of the pairing)?

(II) True or false: "It is legitimate to conduct a paired t test analysis on these differences (assuming that the A−B differences follow a normal distribution)."

(a) We get a sample of 20 patients and record their ages, ranking them from 1 to 20. We match patient 1 with patient 2, patient 3 with patient 4, and so forth. In each pair, one patient gets drug A, and one gets drug B (randomly). Then we record change in blood pressure and look at the differences within each pair.

(b) For each of 10 consecutive weeks, we randomly select two patients, give one of them drug A and the other drug B (randomly), and record change in blood pressure. We then look at the differences within the pair from the first week, the pair from the second week, and so forth.

(c) We get a sample of 20 patients and give 10 of them drug A and the other 10 drug B (randomly). Then, we measure change in blood pressure. We match the patient with the greatest change on A with the patient with the greatest change on B, the second greatest change on A with the second greatest change on B, and so forth. Then, we look at the differences within each pair.

II.11 Fred collects data from two populations. Maria uses those data to construct a (two-sided) 90% confidence interval for $\mu_1 - \mu_2$, and gets $(-3.4, 23.7)$. At the same time Sam uses the data to test $H_0: \mu_1 = \mu_2$ versus $H_A: \mu_1 > \mu_2$ (which he chose to test before seeing the data) with $\alpha = 0.05$. Does Sam reject H_0? Or can we not tell from the information given? Explain your reasoning.

II.12 [Based on Koh, et al. (1997). Effects of hormone-replacement therapy on fibrinolysis in postmenopausal women. *New England Journal of Medicine* **336**, 683–690, which is the basis of Exercise 8.S.22.] Researchers wanted to test the effect of oral conjugated estrogen on plasminogen-activator inhibitor type 1 (PAI-1). They

took measurements on each of 30 women before taking conjugated estrogen and after taking conjugated estrogen. They then prepared to analyze the data from this paired design. One researcher, Smith, wanted to conduct a sign test. Another, Jones, advocated a paired t test on the grounds that the sign test uses only part of the information in the data (namely, whether the before–after difference is positive or negative) and thus is less powerful than the t test.

(a) What did Jones mean in saying that the sign test "is less powerful than the t test"?

(b) Assuming that Jones is correct, why would anyone ever conduct a sign test?

II.13 A random sample of 50 subjects received biofeedback training to reduce blood pressure. Researchers measured the decrease in blood pressure for each of them; the average decrease was 11.4, and the SD was 1.3. A second sample of 40 control subjects had an average decrease in blood pressure during the study of 5.0, with an SD of 1.4.

(a) Construct a 95% confidence interval for the population difference average in blood pressure decrease between treatment (biofeedback) and control patients.

(b) Consider just the treatment (biofeedback) group. True or false: We can estimate that roughly 95% of all persons given biofeedback training will experience a decrease in blood pressure in the range $11.4 \pm 2(1.3)$, that is, between 8.8 and 14.0. Explain your reasoning.

II.14 Suppose we want to test whether an experimental drug reduces blood pressure more than does a placebo. We are planning to administer the drug or placebo to some subjects and record how much their blood pressures are reduced. We have 20 subjects available.

(a) We could form 10 matched pairs, where we form a pair by matching subjects, as best we can, on the basis of age and sex. Briefly explain why using a matched pairs design might be a good idea.

(b) Briefly explain why a matched pairs design might *not* be a good idea. That is, how might such a design be inferior to a completely randomized design?

Background for II.15–II.19

Researchers wanted to assess whether or not training a cat can affect its acceptance of blood testing. Fourteen cats were selected to not be familiarized (trained) with having their blood drawn, while 17 other cats were selected to be trained by familiarizing them with being restrained and shaved for a blood draw. Blood cortisol levels (a stress hormone) was measured to compare the stress levels of the cats during their blood draw (μg/dL). The following table presents a summary of the data.[2]

	Cortisol (μg/dL)		
Group	Mean	SE	N
Trained	0.541	0.215	17
Untrained	2.324	0.239	14

II.15 Consider the research question: Can training *reduce* stress in cats when they are getting their blood drawn?

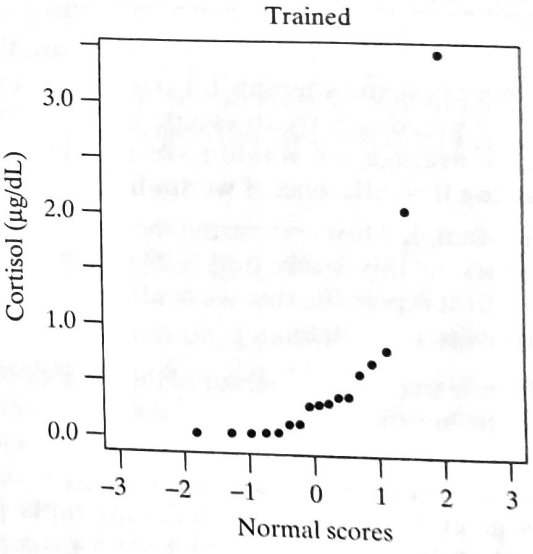
(a) In plain English, write the null and alternative hypotheses of interest to the researcher (regardless of whether or not this is truly testable with these data).

(b) Given the context of this study design, is this research question in (a) testable with these data? Briefly explain.

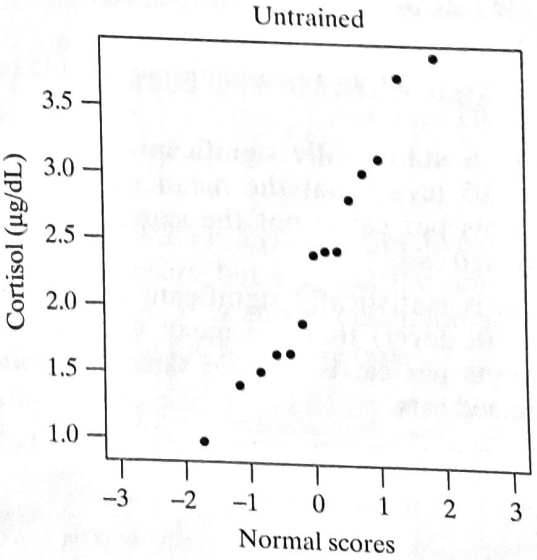
(c) Use an appropriate t test to test the hypothesis in part (a). You may use df ≈ 28 and $\alpha = 0.05$.

(d) Suppose that medically speaking, cortisol levels that differ by less 1.0 μg/dL are not treated as being different. Do you think we should regard the mean cortisol levels for trained and untrained cats as being different from a medical (i.e., practical) perspective? Use the data and an appropriate statistical procedure to support your answer.

II.16 Below are histograms and normal probability plots and Shapiro-Wilk test P-values of the cortisol levels of the 31 cats in the study.



Shapiro-Wilk P-value = 0.0001



Shapiro-Wilk P-value = 0.6686

(a) Do these data meet the normality requirements for the two-sample t test?

(b) If we obtained larger samples, would we expect the resulting plots to appear more normal? Briefly explain.

(c) In a sentence or two, what would be the primary argument for using the Wilcoxon-Mann-Whitney test to compare cortisol levels for trained and untrained cats?

(d) What is the primary drawback of using the Wilcoxon-Mann-Whitney test over the two-sample t test when you don't "need" to use it?

II.17 For each of the following statements say whether they are true or false and explain why.

(a) If training is truly unrelated to stress (cortisol) levels, a larger study would have more power than a smaller one.

(b) Changing the value of α will affect the P-value of the test.

(c) Consider a test to compare the number of escape attempts for trained and untrained cats. If training is really associated with fewer escape attempts, we would have a better chance of detecting this relationship by choosing a small α rather than a large one.

(d) As in (c) above, consider a test to compare the mean number of escape attempts for trained and untrained cats. If trained cats try to escape less than untrained cats, on average, we would have a better chance of detecting this difference if we studied more cats.

(e) A two-sample t test comparing the number of escape attempts in this study had a P-value of 0.022. If $\alpha = 0.01$ it is possible that we could be making a Type I error with these data.

II.18 The researchers also measured the number of escape attempts during the medical procedure for each cat. A 95% confidence interval for the difference in mean number of escape attempts (per cat) was $(\mu_{\text{trained}} - \mu_{\text{untrained}})$ and was given to be $(-1.29, -0.17)$ attempts per cat. Suppose a nondirectional two-sample t test was conducted on these same data. State whether the following statements are true or false or cannot be determined without further computations. Explain why.

(a) The P-value for the two-sample t test would be greater than 0.05.

(b) There is statistically significant evidence (at the $\alpha = 0.05$ level) that the mean number of escape attempts per cat is not the same for trained and untrained cats.

(c) There is statistically significant evidence (at the $\alpha = 0.01$ level) that the mean number of escape attempts per cat is not the same for trained and untrained cats.

(d) There is statistically significant evidence (at the $\alpha = 0.10$ level) that the mean number of escape attempts per cat is not the same for trained and untrained cats.

II.19 Write a sentence interpreting the confidence interval reported in II.18 in the context of the problem. That is, use the interval to carefully describe to a reader what the study reveals about how often trained cats try to escape versus untrained cats.

II.20 As part of a study to determine the effect of flaxseed on blood thiocyanate concentration in rats, the authors reported the amount of hydrogen cyanide (HCN) in the flaxseed fed to the rats. This value was reported to be 255 ± 8.3 mg/kg seed. Based on the authors' intent to communicate that flaxseed source is fairly consistent in its HCN content, do you think these values represent mean $\pm$ SD or mean $\pm$ SE?

II.21 For each of the following research studies, identify whether independent samples (e.g., two-sample t test, Wilcoxon-Mann-Whitney) or paired samples (e.g., paired t test, sign test) would most likely be appropriate and explain why.

(a) To investigate whether or not there are differences in turbidity (i.e., suspended sediment) at the surface of a stream versus near the bottom, 15 stream locations were sampled, and turbidity measurements were made at the surface and depth of 0.5 meters.

(b) To investigate if a marine protection policy off the coast of California improves the health of rockfish, researchers measured catch per unit effort (a surrogate measure for fish abundance) at eight rockfish habitat locations inside a marine protected area and eight additional nearby habitat locations outside the protected area.

(c) To compare the therapeutic value of different milk-based supplements in malnourished children, researchers randomly assigned one of two common supplements (10% milk or 25% milk) to approximately 1,900 malnourished children in rural Malawi and measured their weights at the start of the study, and again after 8 weeks of therapy.

II.22 Consider the study described in II.21(b).

(a) Considering conducting a study of the same size to answer the same question as described in II.21(b), describe in detail a study design that would use paired samples analysis.

(b) Which design (independent or paired) is likely more powerful and why?