1 Solution to Problem 1.

(1) [25 points] *In a sample of $n = 81$ cockroaches the amount of sugar in their hindgut was measured 2 hours after being fed D-glucose. The sample mean was 44.44 micrograms and the sample standard deviation was 21.35 micrograms.*

(a) *Calculate a 95% confidence interval for the population mean sugar in the hindgut of cockroaches under these conditions.*

Solution: The sample size is large enough ($n = 81 > 30$) that we may use a z-interval, and the underlying distribution of the population doesn't matter. The estimated standard error of $\bar{Y}$ is

$$\hat{SE}(\bar{Y}) \;=\; s/\sqrt{n} \;=\; 21.35/9 \;=\; 2.37.$$

The desired confidence interval is

$$\bar{y} \pm z(.025)\hat{SE}(\bar{Y}) \;=\; 44.44 \pm 1.960 * 2.37 \;=\; 44.44 \pm 4.65 \;=\; (39.79, 49.09).$$

If we used the t distribution (based on the assumption of a normal population), then $z(.025) = 1.960$ in the previous calculation would have been replaced with $t(.025, 80) = 1.9901$. The result would have been

$$44.44 \pm 4.72 \;=\; (39.72, 49.16).$$

Obviously, it makes very little difference.

(b) *Construct a 95% prediction interval for a single new cockroach under the same conditions.*

Solution: Apply the result of Box 4.10, p. 161 with $m = 1$:

$$\bar{y} \pm t(.025, n-1) * s * \sqrt{1 + 1/n} \;=\; 44.44 \pm 1.9901 * 21.35 * \sqrt{1 + 1/81}$$

$$= 44.44 \pm 42.75 = (1.69, 87.19).$$

We should keep in mind that this is based on the assumption of a normal population.

One can also replace $t(.025, 80)$ with $z(.025)$ in the previous calculation (large sample size!). The result is

$$44.44 \pm 42.11 = (2.33, 86.55).$$

It wouldn't be a big mistake to forget the factor of $\sqrt{1 + 1/80}$ as this is close to 1 anyway. The result would be

$$44.44 \pm 41.85 = (2.59, 86.29).$$

Grading Notes:

-3 Multiplying confidence interval width in part (a) by $\sqrt{1 + 1/n}$.

-1 Forgot $\sqrt{\ldots}$ in $\hat{SE}$ computation.

(c) *In a previous experiment, the amount of sugar in the hindgut was measured 3 hours after being fed D-glucose in 45 cockroaches. The sample mean under these conditions was 33.72 and the standard deviation was 11.16. Is there a statistically significant difference in the mean amount of sugar at the two times after feeding? State an appropriate null and alternative (research) hypothesis, compute an appropriate test statistic and P-value, and determine if you can reject the null hypothesis at the* $\alpha = 0.05$ *level of significance.*

Solution: Before we can state hypotheses to be tested, we always have to define parameters. Let

$$\begin{aligned} \mu_1 &= \text{pop. mean of D glucose after 2 hrs.} \\ \mu_2 &= \text{pop. mean of D glucose after 3 hrs.} \\ \theta &= \mu_1 - \mu_2. \end{aligned}$$

The question asks, "Is there a statistically significant difference in the mean amount of sugar at the two times after feeding?" There is no indication of a direction – that one mean is expected to be larger or smaller than the other, just any difference. Therefore, we will test

$$H_0 : \theta = 0 \quad vs. \quad H_1 : \theta \neq 0.$$

Both sample sizes are large ($n_1 = 81 > 30$, $n_2 = 45 > 30$), so it is permissable to use a z-statistic (based on Central Limit Theorem) for the test (this has been discussed in class). The point estimate of θ is

$$\hat{\theta} \;=\; \bar{y}_1 - \bar{y}_2 \;=\; 44.44 - 33.72 \;=\; 44.44 - 33.72 \;=\; 10.72.$$

The estimated standard error is

$$\hat{SE}(\hat{\theta}) \;=\; \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} \;=\; \sqrt{\frac{21.35^2}{81} + \frac{11.16^2}{45}} \;=\; 2.90.$$

The z-statistic is

$$z \;=\; (\hat{\theta} - \theta_0)/\hat{SE}\hat{\theta} \;=\; (10.72 - 0)/2.90 \;=\; 3.70.$$

The P-value is

$$P[Z < -3.70 \text{ or } Z > 3.70] \;=\; 2 * P[Z > 3.70].$$

Note that this is a two sided test (two tailed region for rejecting H_0), so we need to consider probabilities on both sides, which is why the tail probability $P[Z > 3.70]$ is multiplied by 2. Table C.1 doesn't give us the probability that $Z > 3.70$, but it does give us that $P[Z > 3.09] = .0010$, so we can conclude that $P[Z > 3.70] < .0010$. Hence, all we can say for sure is

$$p - \text{value} \;<\; .0020.$$

Clearly the result is statistically significant at the usual 0.05 level, and we do conclude there is a difference between the mean D-glucose levels at the different times after feeding.

Grading Notes:

-2 Stating hypotheses about $\bar{y}_i$'s instead of parameters.

-1 Multiplying tail probability by 1 or 1/2 instead of 2.

-1 Not completing computation of test statistic.

2 Solution for Problem 2.

(2) [20 points] *Below are samples of size* 5 *observations from 6 different distributions:*

(1) *Binomial with parameters* $n = 8$ *and* $\pi = .8$*;*

(2) *Binomial with parameters* $n = 8$ *and* $\pi = .5$*;*

(3) *Poisson with parameter* $\lambda = 8$*;*

(4) *Normal with parameters* $\mu = 20$ *and* $\sigma^2 = 100$*;*

(5) *Normal with parameters* $\mu = 40$ *and* $\sigma^2 = 100$*;*

(6) *Normal with paramaters* $\mu = 40$ *and* $\sigma^2 = 1$*;*

Match the samples to the distributions.

The samples are:

(A) *44.81, 63.80, 26.35, 42.47, 56.13*

(B) *24.65, 24.56, 14.78, 34.13, 42.29*

(C) *11, 8, 6, 5, 12*

(D) *3, 5, 6, 6, 6*

(E) *6, 7, 8, 8, 5*

(F) *38.69, 40.64, 37.89, 38.18, 40.18*

Solution: Clearly samples A, B, and F are continuous, so they go with the normal distributions. The data in F are tightly clustered around 40, so they must belong to

distribution (6). The data in A tend to be larger than the data in B, so the data in A go with distribution 5 and the data in B with distribution 4. There are values in C bigger than 8, so it must be from the Poisson distribution 3. The values in E tend to be larger, so they go with distribution 1, and the data in D go with distribution 2. In summary:

Data	Distribution
A	5
B	4
C	3
D	2
E	1
F	6
E	1
D	2
C	3
B	4
A	5
F	6

Grading Notes: No one missed this problem!

3 Solution to Problem 3.

(3) [20 points] *True or False: Answer "T" or "F" according as the statement is True of False.*

T **F** *If the P-value is* 0.125*, then we reject the null hypothesis at the* $\alpha = 0.05$ *level of significance.*

Solution: FALSE We only reject if the p-value is less than the desired level of significance.

T **F** *When using the T distribution to make a test of equality of two populaiton means with small samples, the most important assumption is that the samples are independent random samples.*

Solution: TRUE This was stated in lecture.

T **F** *If* Y *is a random variable with a normal distribution, then* $Pr[Y \leq 1] = P[Y < 1]$.

Solution: TRUE Since $P[Y = 1] = 0$ for a continuous RV like the normal.

T **F** *The* χ^2 *distribution may not work well for constructing confidence intervals for a population variance if the observations come from a population which deviates even a little from a normal distribution.*

Solution: TRUE This was stated in lecture (see the "Robustness Study").

T **F** *In setting up a statistical test of hypotheses, we usually take the alternative hypothesis to be the statement we wish to prove because we control the probability of falsely accepting the alternative when setting the level of significance.*

Solution: TRUE We only reject the null hypothesis when there is strong evidence against it, so we are pretty sure we are doing the right thing when we accept the alternative hypothesis. However, if we accept the null hypothesis, we aren't confident we are making the correct decision without further analysis (e.g. computing the power or type II error probability).

4 Solution to Problem 4.

(4) [20 points] *In order to test the research hypothesis that taking Birth Control Pills (BCPs) increases cholesterol, a researcher selects a sample of birth control pill users. For each subject in the sample, another subject is found who is not a BCP user, but matches the birth control pill user within 3 years of age, 1 unit of body mass index (a measure of weight), smoking status, general diet, race, and exercise level. After deleting BCP users who were not matched to a nonuser, the sample sizes were 135 of each. Two statistical analyses are proposed:*

(1) *Compute the test statistic*

$$z = \frac{\bar{y}_u - \bar{y}_n}{\sqrt{(s_u^2 + s_n^2)/135}}$$

where $\bar{y}_u$ and s_u^2 are the sample mean and variance of cholesterol for the BCP users, and $\bar{y}_n$ and s_n^2 are the sample mean and variance for the BCP nonusers. If this test statistic is larger than z(.05), then we conclude there is significant evidence for an increase in cholesterol among the BCP users.

(2) *Within each matched pair, take the difference between the cholesterol values, subtracting the nonuser's cholesterol from the user's cholesterol. Compute the sample mean and variance of these differences (denoted $\bar{d}$ and s_d^2), and compute the test statistic*

$$z = \frac{\bar{d}}{s_d/\sqrt{135}}.$$

If this test statistic is larger than z(.05), then we conclude there is significant evidence for an increase in cholesterol among the BCP users.

Discuss both of these statistical analyses. Are either or both appropriate? If so, why or why not?

Solution: Let μ_1 be the mean cholesterol of BCP users, and μ_2 the mean cholesterol of non-BCP users. The parameter of interest is

$$\theta \ = \ \mu_1 - \mu_2.$$

Since the research hypothesis is that taking BCP increases cholesterol, we want to test

$$H_0 : \theta \leq 0 \quad vs. \quad H_1 : \theta \geq 0.$$

The description of the study indicates that it is a *matched case-control study*: the subjects taking the BCP are closely matched in several attributes with subjects not on BCP. Therefore, we must use the procedure in Box 4.5, page 143, which is for *paired data*. This is exactly the description of method (2): take differences within pairs and then θ is the population mean of these differences. Since the sample size is large (135 is much bigger than 30), we may use a z-test in place of a t-test, as for any one sample test. The z test statistic is computed exactly as in description (2). We reject for large values of z, names $z(.05)$ for the usual 0.05 level of significance.

Description 1 applies to the two sample z-test, which is based on the assumption of *independent random samples*. This is the most important assumption for this method. This is clearly inapplicable to this study design since within each matched pair there is *dependence*. For instance, if one of the members of a matched pair is old, overweight, and a smoker, then so is the other, and we expect both to have high cholesterol, irrespective of BCP use.

Grading Notes:

-12 for saying method 1 was appropriate but not method 2

-9 for saying both are appropriate

-8 for saying both are appropriate, but method 2 is better (method 1 is definitely not appropriate!)

-1 for saying that a t-distribution should have been used in method 2

5 Solution to Problem 5.

(5) [15 points] *A wildlife ecologist observes pumas in a certain area. Out of 49 pumas he has caught and tagged, only 14 are female. He believes this may indicate that there is some problem with the gender ratio in this area. He expected that about 50% of the pumas would be females.*

(a) *Formulate a statistical hypothesis testing problem appropriate to the ecologist's concerns. Compute a test statistic and P-value. Determine if the null hypothesis can be rejected.*

Solution: Let π be the true proportion of females in the area. Prior to taking the data, we had no idea that there may be a problem with too many or too few females. Therefore we want to test

$$H_0 : \pi = .5 \quad vs. \quad H_1 : \pi \neq .5.$$

Since the sample size is large ($n \geq 30$) and the null hypothesis is in the range where the normal approximation to the binomial applies ($n\pi = 24.5 = n(1-\pi) \geq 10$), we use a z test statistic. The test statistic is

$$z = \frac{\hat{\pi} - \pi_0}{SE_0(\hat{\pi})}$$

where

$$SE_0(\hat{\pi}) = \sqrt{\pi_0(1-\pi_0)/n}.$$

(Remark: this formula was given in class and corrects part 1 of Box 6.1, p. 206). For our data

$$\begin{aligned} n &= 49 \\ \hat{\pi} &= 14/49 = 0.2857 \end{aligned}$$

$$\begin{aligned} \pi_0 &= 0.5 \\ SE_0(\hat{\pi}) &= \sqrt{.5 * (1 - .5)/49} \ = \ 0.07143 \\ z &= (0.2857 - .5)/0.07143 \ = \ -3.000. \end{aligned}$$

The result is clearly statistically significant. The p-value is given by

$$p - \text{value} \ = \ P[Z < -3.00 \text{ or } Z > 3.00] \ = \ 2 * P[Z > 3.00] \ = \ 2 * 0.0013 \ = \ 0.0026.$$

Again, since the p-value is less than 0.05, we reject H_0 and conclude there is some imbalance in the gender ratio in this area.

Remark If you used the $\hat{SE}(\hat{\pi})$ as given below for the confidence interval, you will get

$$z \ = \ (0.2857 - .5)/0.06454 \ = \ -3.32.$$

You would conclude that the p-value is < 0.002.

(b) *Construct a 95% confidence interval for the proportion of females in the area under study.*

Solution: Now we use the estimated standard error

$$\hat{SE}(\hat{\pi}) \ = \ \sqrt{\hat{\pi}(1 - \hat{\pi})/n} \ = \ \sqrt{0.2857 * (1 - 0.2857)/49} \ = \ 0.06454.$$

The desired 95% confidence interval is

$$\hat{\pi} \pm z(.025) * \hat{SE}(\hat{\pi}) \ = \ 0.2857 \pm 1.960 * 0.06454 \ = \ 0.2857 \pm 0.1265 \ = \ (0.1592, 0.4122).$$

(c) *Comment on the validity of the statistical inferences in parts (a) and (b). Can we be sure that the statistical assumptions are valid?*

Solution: The most important assumptions are that we have a random sample from and infinite population. The random sampling assumption is clearly suspect here –

the investigator only observed the pumas he caught. It may be that male pumas are more bold and easier to see and so were caught easier, and the females are there but were not observed. Also, there is some doubt about the infinite population assumption since pumas are territorial and have somewhat large territories, so their population density is small. It did not say how big the study area was, so the population size may be not much bigger than the sample size.

Grading Notes:

-1 for using $\hat{SE}(\hat{\pi})$ in part (a) rather than $SE_0(\hat{\pi})$.

-1 dividing by 2 instead of multiplying by 2 in p-value calculation.

-2 Stating hypotheses about statistics ($\hat{\pi}$) instead of parameters.

-2 Using $H_1 : \pi < .5$.

-3 Some kind of strange z or t-statistic calculation instead of proportions, part (a).

-2 failing to multiply by 2 in p-value calculation.

-2 dividing by n instead of $\sqrt{n}$ in SE calculation.

-2 using SE_0 instead of $\hat{SE}$ in part (b).

-1 for conservative confidence interval ($\pm 1/\sqrt{n}$) in (b).

-4 Some strange result in part (b).

-3 for saying "the sample seems to be chosen randomly" in part (c). This is clearly a "convenience" sample.

6 Summary of Scores and Grades.

Basic descriptive statistics of the scores:

Min	72
1st Quart.	83.25
Mean	86
Median	85
3rd Quart.	90
Max	96
Total N	14
Std Dev.	6.40

Approximate grade assignments:

Grade	Score Range	Number
A	86-96	7
B	82-84	5
B-	72-79	2

Note: Of course, it is the numerical grade that counts. The letter grades are an indication of how grades will be assigned in the end (when there is a total numerical grade).