STAT 305 D Final Exam

Show all your work.

1. (20 points) Find the following:

   a. (4 points) $P((Z-2, Z+2) \text{ contains } 0)$, $Z \sim N(0,1)$

$$\begin{aligned} P((Z-2, Z+2) \text{ contains } 0) &= P(Z-2 < 0 < Z+2) \\ &= P(-2 < Z < 2) \\ &= \Phi(2) - \Phi(-2) \\ &= 0.9772 - 0.0227 \qquad = 0.9545 \end{aligned}$$

   b. (4 points) $P((X-5, X+5) \text{ contains } 8)$, $X \sim N(3,4)$

$$\begin{aligned} P((X-5, X+5) \text{ contains } 8) &= P(X-5 < 8 < X+5) \\ &= P(X-13 < 0 < X-3) \\ &= P(3 < X < 13) \\ &= P(\frac{3-3}{2} < \frac{X-3}{2} < \frac{5-3}{2}) \\ &= P(0 < Z < 1) \\ &= \Phi(1) - \Phi(0) \\ &= 0.8413 - 0.5 \\ &= 0.3413 \end{aligned}$$

   c. (4 points) $P((X-2, X+2) \text{ contains } 0)$, $X \sim N(1, 0.4)$

$$\begin{aligned} P((X-2, X+2) \text{ contains } 0) &= P(X-2 < 0 < X+2) \\ &= P(-2 < X < 2) \\ &= P(\frac{-2-1}{\sqrt{0.4}} < \frac{X-1}{\sqrt{0.4}} < \frac{2-1}{\sqrt{0.4}}) \\ &= P(-4.74 < Z < 1.58) \\ &\approx P(Z < 1.58) \\ &= \Phi(1.58) \\ &= 0.9429 \end{aligned}$$

d. (4 points) $P((X-5, X+5) \text{ contains } 8),\ X \sim t_7$

$$\begin{aligned} P((X-5, X+5) \text{ contains } 8) &= P(X-5 < 8 < X+5) \\ &= P(3 < X < 13) \\ &= P(t_7 < 13) - P(t_7 < 3) \\ &= 1 - 0.099 \\ &= 0.0100 \end{aligned}$$

e. (4 points) $P((X - 1.5 \cdot \sigma, X + 1.5 \cdot \sigma) \text{ contains } \mu),\ X \sim N(\mu, \sigma^2)$

$$\begin{aligned} P((X - 1.5 \cdot \sigma, X + 1.5 \cdot \sigma) \text{ contains } \mu) &= P(X - 1.5 \cdot \sigma < \mu < X + 1.5 \cdot \sigma) \\ &= P(-1.5 < \frac{X-\mu}{\sigma} < 1.5) \\ &= P(-1.5 < Z < 1.5) \\ &= \Phi(1.5) - \Phi(-1.5) \\ &= 0.9331 - 0.0668 \\ &= 0.8664 \end{aligned}$$

2. (20 points) Every year, your eccentric neighbor carves an excessive number of pumpkins for Halloween and leaves them outside his front door through November. Being eccentric yourself, you weigh all the pumpkins and record their combined weight each year. Here are the data from the last six years:

Year	2007	2008	2009	2010	2011	2012
Weight (lb)	311.0	271.1	320.7	274.3	332.4	304.4

a. (10 points) Conduct a hypothesis test at $\alpha = 0.05$ to find out if the true mean combined weight of your neighbor's pumpkins exceeds 290 lb.

First, I calculate the mean weight to be:

$$\overline{x} = \frac{1}{6}(311.0 + 271.1 + \cdots + 304.4) = 302.3$$

and the standard deviation is:

$$s = \frac{1}{6-1}[(311.0 - 302.3)^2 + (271.1 - 302.3)^2 + \cdots + (304.4 - 302.3)^2] = 24.8$$

And here is the hypothesis test:

1. $H_0 : \mu = 290$, $H_a : \mu > 290$, where μ is the true mean combined pumpkin weight.
2. $\alpha = 0.05$
3. I use the test statistic:

$$K = \frac{\overline{x} - 290}{s/\sqrt{n}}$$

I assume the combined pumpkin weights are iid $N(\mu, \sigma^2)$. Under the additional assumption that H_0 is true, $K \sim t_{n-1} = t_5$. I will reject H_0 if $K > t_{n-1,1-\alpha} = t_{5,0.95} = 2.02$

4. The moment of truth:

$$K = \frac{302.3 - 290}{24.8/\sqrt{6}} = 1.21$$

5. With a test statistic of $K = 1.21 < 2.02 = t_{n-1,1-\alpha}$, I fail to reject H_0.
6. There is not enough evidence to conclude that your neighbor's true mean combined pumpkin weight (per year) exceeds 290 lb.

b. (10 points) Attempt to convince your neighbor of his absurdity by constructing and interpreting a lower 99% confidence interval for the true mean pumpkin weight.

Since $n < 25$ and σ^2 is unknown, I use the confidence interval:

$$\begin{aligned}
&(\overline{x} - t_{n-1,\ 1-\alpha} \cdot \frac{s}{\sqrt{n}},\ \infty) \\
&= (302.4 - t_{5,0.995} \cdot 24.8\sqrt{6},\ \infty) \\
&= (302.4 - 4.03 \cdot 10.12) \\
&= 261.62
\end{aligned}$$

With 99% confidence, the true mean annual combined pumpkin weight exceeds 261.62 lb. That's TOO MUCH PUMPKIN MATTER for individual use.

3. (20 points) Not to be outdone, you have been carving and displaying absurd levels of pumpkin matter this entire time. Here are *your* combined pumpkin weights for the last six years.

Year	2007	2008	2009	2010	2011	2012
Weight (lb)	317.7	368.0	341.8	367.1	318.1	292.1

a. (10 points) In a typical year, do you exceed your neighbor in pumpkin weight? Conduct the appropriate hypothesis at $\alpha = 0.05$ to find out.

First, compute the differences in pumpkin mass

Year	2007	2008	2009	2010	2011	2012
Your pumpkin weight (lb)	317.7	368.0	341.8	367.1	318.1	292.1
Neighbor's pumpkin weight (lb)	311.0	271.1	320.7	274.3	332.4	304.4
Difference (lb)	6.7	96.9	21.1	92.8	-14.3	-12.3

I compute the mean of the differences (yours - your neighbor's) to be $\overline{d} = 31.8$ and the standard deviation to be $s_d = 50.5$.

Here is the hypothesis test:

1. $H_0 : \mu_d = 0$, $H_a : \mu_d > 0$
2. $\alpha = 0.05$
3. I use the test statistic:

$$K = \frac{\overline{d}}{s_d/\sqrt{n}}$$

   I assume the combined pumpkin weight differences are iid $N(\mu_d, \sigma_d^2)$. Under the additional assumption that H_0 is true, $K \sim t_{n-1} = t_5$. I will reject H_0 if $K > t_{n-1,1-\alpha} = t_{5,0.95} = 2.02$
4. The moment of truth:

$$K = \frac{31.8}{50.5/\sqrt{6}} = 1.54$$

5. With a test statistic of $K = 1.54 < 2.02 = t_{n-1,1-\alpha}$, I fail to reject H_0.
6. There is not enough evidence to conclude that you exceed your neighbor in combined pumpkin weight in a typical year. Your efforts have been in vain. And now you have TOO MANY SEEDS!

b. (10 points) Construct and interpret a 2-sided 95% confidence interval for the true mean difference in pumpkin mass between the two of you.

I use the confidence interval:

$$\begin{aligned}
&(\overline{d} - t_{n-1,\ 1-\alpha/2} \cdot \frac{s_d}{\sqrt{n}},\ \overline{d} + t_{n-1,\ 1-\alpha/2} \cdot \frac{s_d}{\sqrt{n}}) \\
&= (31.8 - t_{5,0.975} \cdot \frac{50.5}{\sqrt{6}},\ 31.8 + t_{5,0.975} \cdot \frac{50.5}{\sqrt{6}}) \\
&= (31.8 - 2.57 \cdot 20.6, 31.8 + 2.57 \cdot 20.6) \\
&= (-21.1,\ 84.7)
\end{aligned}$$

I am 95% confidence interval that the true mean difference between your annual combined pumpkin weight and your neighbor's is between -21.1 lb and 84.7 lb.

4. (20 points) Now that you have bought and carved too many pumpkins, you have too many pumpkin seeds. You also have your neighbor's seeds because he is too eccentric for any kind of seed. You have thousands and thousands of seeds. Naturally, you decide to measure the length of each seed individually, painstakingly, carefully, painstakingly, and individually. Pain! Your 11201 seeds (sample 1) have a mean length of 1.11 cm and a standard deviation of 0.25 cm. Your neighbor's 6732 seeds (sample 2) have a mean length of 1.23 cm and a standard deviation of .25 cm.

   a. (10 points) Use the appropriate hypothesis test at $\alpha = 0.01$ to test the claim that, on average, your seeds are a different length than your neighbor's.

   I will need the pooled standard deviation, s_p:

$$\begin{aligned} s_p &= \frac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{n_1+n_2-2} \\ &= \frac{(11201-1)(.25)^2 + (6732-1)(.25)^2}{11201+6732-2} \\ &= 0.25 \end{aligned}$$

   1. $H_0 : \mu_1 - \mu_2 = 0$, $H_a : \mu_1 - \mu_2 \neq 0$
   2. $\alpha = 0.01$
   3. I use the test statistic:

$$K = \frac{\overline{x}_1 - \overline{x}_2}{s_p\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}}$$

      I assume your pumpkin seed lengths are iid (μ_1, σ^2) and your neighbor's pumpkin seed lengths are iid (μ_2, σ^2). Under the additional assumption that H_0 is true, $K \sim N(0,1)$ since the sample sizes are huge. I will reject H_0 if $|K| > |z_{1-\alpha/2}| = z_{0.995} = 2.58$
   4. The moment of truth:

$$K = \frac{1.11 - 1.23}{0.25\sqrt{\frac{1}{11201}+\frac{1}{6732}}} = -31.13$$

   5. With $|K| = 31.13 > 2.58$, I reject H_0 and conclude H_a.
   6. There is overwhelming evidence that your pumpkin seed lengths are different from your neighbor's.

   b. (10 points) Compute and interpret a two-sided 99% confidence interval for the difference in true mean pumpkin seed lengths (yours - your neighbor's).

First I compute:

$$0.25\sqrt{\frac{1}{11201}+\frac{1}{6732}} = 0.00386$$

Next, I note that $z_{0.995} = 2.58$, and the difference in means is 1.11 - 1.23 = 0.12. Now, I compute the confidence interval as

$$(0.12 - 2.58 \cdot 0.00386, 0.12 + 2.58 \cdot 0.00386)$$
$$= (0.11cm, 0.13cm)$$

With 99% confidence, your pumpkin seed lengths exceed those of your neighbor by anywhere from 0.11 cm to 0.13 cm.

5. (20 points) Return to the New York rivers data:

```
## % latex table generated in R 3.0.0 by xtable 1.7-1 package
## % Fri May 10 12:38:08 2013
## \begin{table}[ht]
## \centering
## \begin{tabular}{rlrrrrr}
##   \hline
##  & Name & X1 & X2 & X3 & X4 & Y \\
##   \hline
## 1 & Olean &  26 &  63 & 1.20 & 0.29 & 1.10 \\
##   2 & Cassadaga &  29 &  57 & 0.70 & 0.09 & 1.01 \\
##   3 & Oatka &  54 &  26 & 1.80 & 0.58 & 1.90 \\
##   4 & Neversink &   2 &  84 & 1.90 & 1.98 & 1.00 \\
##   5 & Hackensack &   3 &  27 & 29.40 & 3.11 & 1.99 \\
##   6 & Wappinger &  19 &  61 & 3.40 & 0.56 & 1.42 \\
##   7 & Fishkill &  16 &  60 & 5.60 & 1.11 & 2.04 \\
##   8 & Honeoye &  40 &  43 & 1.30 & 0.24 & 1.65 \\
##   9 & Susquehanna &  28 &  62 & 1.10 & 0.15 & 1.01 \\
##   10 & Chenango &  26 &  60 & 0.90 & 0.23 & 1.21 \\
##   11 & Tioughnioga &  26 &  53 & 0.90 & 0.18 & 1.33 \\
##   12 & West\_Canada &  15 &  75 & 0.70 & 0.16 & 0.75 \\
##   13 & East\_Canada &   6 &  84 & 0.50 & 0.12 & 0.73 \\
##   14 & Saranac &   3 &  81 & 0.80 & 0.35 & 0.80 \\
##   15 & Ausable &   2 &  89 & 0.70 & 0.35 & 0.76 \\
##   16 & Black &   6 &  82 & 0.50 & 0.15 & 0.87 \\
##   17 & Schoharie &  22 &  70 & 0.90 & 0.22 & 0.80 \\
##   18 & Raquette &   4 &  75 & 0.40 & 0.18 & 0.87 \\
##   19 & Oswegatchie &  21 &  56 & 0.50 & 0.13 & 0.66 \\
```

```
##   20 & Cohocton &  40 &  49 & 1.10 & 0.13 & 1.25 \\
##    \hline
## \end{tabular}
## \end{table}
```

Remember:

- Y is the mean nitrogen content (mg/liter) of the river based on samples based on regular intervals taken in the spring, summer, and fall months.
- $X1$ is the percentage of surrounding land used in agriculture.
- $X2$ is the % surrounding forested land.
- $X3$ is the % surrounding residential land.
- $X4$ is the % surrounding commercial/industrial land.

I fit the model:

$$Y_i = \beta_0 + \beta_1 X_{1,i} + \beta_2 X_{2,i} + \beta_3 X_{3,i} + \beta_4 X_{4,i} + \varepsilon_i$$

And here is part of the JMP output:

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	1.7222135	1.234082	1.40	0.1832
X1	0.0058091	0.015034	0.39	0.7046
X2	-0.012968	0.013931	-0.93	0.3667
X3	-0.007227	0.03383	-0.21	0.8337
X4	0.3050278	0.163817	1.86	0.0823

a. (4 points) State the assumptions on the ε_i's that we need in order to do inference.

$\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_{20} \sim$ iid $N(0, \sigma^2)$.

b. (4 points) Does % agricultural land affect the nitrogen content in the rivers? Conduct the appropriate hypothesis test at $\alpha = 0.05$ to find out.

1. $H_0 : \beta_1 = 0$, $H_a : \beta_1 \neq 0$
2. $\alpha = 0.05$
3. I use the test statistic, $K = b_1 / \widehat{SD}(b_1)$, which has a $t_{n-p} = t_{20-5} = t_{15}$ distribution under H_0, along with the assumptions that the model is correct and the error terms behave as in part (a).

4. The moment of truth: the p-value is already computed in the table as 0.7046.
5. With a p-value $> \alpha$, we fail to reject H_0.
6. There is not enough evidence to conclude that river pollution varies with the % of surrounding land used in agriculture.

c. (4 points) Test $H_0 : \beta_4 = 0.1$ vs. $H_a : \beta_4 > 0.1$ at $\alpha = 0.01$.

1. $H_0 : \beta_4 = 1$, $H_a : \beta_4 > 1$
2. $\alpha = 0.01$
3. I will use the test statistic:

$$K = \frac{b_4 - 0.1}{\widehat{SD}(b_4)}$$

which has a t_{15} distribution under the assumptions stated in part (b). I will reject H_0 if $K > t_{15,1-\alpha} = t_{15,0.99} = 2.60$.

4. The moment of truth:

$$K = \frac{0.305 - 0.1}{0.164} = 1.25$$

5. With $K = 1.25 < 2.60$, we fail to reject H_0.
6. There is not enough evidence to conclude that for each percentage increase in commercial/industrial land use, the true mean nitrogen content in the rivers increases by over 0.1 mg/L. In this situation, it doesn't seem that high % commercial/industrial land is associated with high pollution, although that may be because there is such a low percentage of commercial/industrial land around each river. Also, a small percentage of industrial/commercial land might be enough to catastrophically raise the pollution to a certain level for each river. Also, the covariates are probably highly correlated, which increases the standard errors of the estimates.

d. (4 points) Construct and interpret a 2-sided 95% confidence interval for the intercept in the model.

In the margin of error, I will use $t_{15,0.975} = 2.13$. The margin of error is $2.13 \cdot 1.234 = 2.63$. The confidence interval is $(1.72 - 2.63, 1.72 + 2.63) = (-0.91, 4.35)$. We're 95% confident that the true pollution level with no surrounding land used at all is somewhere between -0.91 mg/L and 4.35 mg/L.

e. (4 points) What does the model intercept represent? Is there anything problematic in your interpretation on a practical level? The intercept represents the pollution level in rivers with no surrounding land at all. But we never observed such rivers: all the rivers in the data had a high degree of developed land. This is a case of extrapolation.

6. EXTRA CREDIT (10 points). Suppose I fit a different model with just X2 and X3:

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	2.1193202	0.282635	7.50	<.0001*
X2	-0.016004	0.004117	-3.89	0.0012*
X3	0.0162203	0.011483	1.41	0.1758

Why is the coefficient on X2 significantly different from 0 here (p-value = 0.0012) even though it was not significantly different from 0 before (p-value = 0.3667)?

The covariates are highly correlated. For example, we expect more residential land when there is also more industrial land. This correlation in the covariates tends to raise the standard errors of the estimated slopes, which decreases the sensitivity of significance tests.