Fall 2013

Northwestern University

STAT 202-0 Section 23

Introduction to Statistics

Final Exam

Name: _____

December 13, 2013

1. Please do not leave blank for any question.
2. There are 15 questions, each question is 10 points.
3. You have 2 hours for this exam.

Formulas

Mean (average): $\mu, \overline{X}$
Standard deviation (SD): σ, s

$\sum_{i=1}^{n} X_i = X_1 + X_2 + \cdots + X_n$

$\overline{X} = \frac{1}{n} \sum_{i=1}^{n} X_i$

$Z = \frac{X-\mu}{\sigma}$

$X = \mu + \sigma X$

$s_X = \sqrt{\frac{\sum_{i=1}^{n} \left(X_i - \overline{X}\right)^2}{n-1}}$

$r = \frac{1}{n-1} \sum_{i=1}^{n} \left(\frac{X_i - \overline{X}}{s_X}\right) \left(\frac{Y_i - \overline{Y}}{s_Y}\right)$

$r = \frac{\sum_{i=1}^{n} \left(X_i - \overline{X}\right)\left(Y_i - \overline{Y}\right)}{(n-1) s_X s_Y}$

$Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}}$

$t = \frac{\bar{X} - \mu}{s/\sqrt{n}}$

$n = \left(\frac{z^* \sigma}{M.E.}\right)^2$

Under Equal Population SD Assumption

$$t = \frac{\overline{X}_1 - \overline{X}_2}{s_{12}\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}},$$

$$s_{12} = \sqrt{\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}},$$

and

$$d.f. = n_1 + n_2 - 2.$$

Under Unequal Population SD Assumption

$$t = \frac{\overline{X}_1 - \overline{X}_2}{s_{1-2}},$$

$$s_{1-2} = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}},$$

and

$$d.f. = \left\lfloor \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\frac{\left(\frac{s_1^2}{n_1}\right)^2}{n_1 - 1} + \frac{\left(\frac{s_2^2}{n_2}\right)^2}{n_2 - 1}} \right\rfloor.$$

Linear regression with the intercept term

$$y = \alpha + \beta x,$$

$$\hat{\beta} = r_{xy}\frac{s_y}{s_x},$$

$$\hat{\alpha} = \bar{y} - \hat{\beta}\bar{x}.$$

$$\hat{y}_i = \hat{\alpha} + \hat{\beta}x_i$$

$$\varepsilon_i = y_i - \hat{y}_i$$

T-test
for slope

$$t = \frac{\hat{\beta} - \beta}{s_\beta},$$

$$s_\beta = \sqrt{\frac{\frac{1}{n-2}\sum_{i=1}^{n}\hat{\varepsilon}_i^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}},$$

$$d.f. = n - 2.$$

for intercept

$$t = \frac{\hat{\alpha} - \alpha}{s_\alpha},$$

$$s_\alpha = s_\beta\sqrt{\frac{1}{n}\sum_{i=1}^{n}x_i^2},$$

$$d.f. = n - 2.$$

1. Sampling

   Please randomly draw three students from a group of seven: Jacob, Sophia, Mason, Emma, Ethan, Isabella, Noah. You may use Table 1 of random digits. Briefly describe the sampling procedure you use.

Table 1: Random Digits

1	2	9	0	0	1	8	3	5	3
6	9	0	9	9	4	7	1	9	7
3	5	3	9	0	7	2	8	4	1
1	1	6	0	4	7	2	1	8	6
2	1	1	6	0	8	9	3	2	8
9	0	1	1	7	4	2	1	3	6

2. Experimental Design

   An experiment compares three approaches. Here are the names of 6 subjects.

   - William
   - Emily
   - Michael
   - Alexander
   - Madison
   - Elizabeth

   Please assign 2 subjects at random to each of the three groups. You may use Table 2 of random digits. Briefly describe the allocation method you use.

Table 2: Random Digits

4	3	1	7	9	3	7	7	1	2
7	6	8	8	9	5	2	2	3	6
0	9	6	7	4	5	0	7	5	8
4	9	3	4	1	4	2	0	7	4
8	1	3	1	4	6	7	5	1	4
2	9	7	4	9	6	9	9	6	0

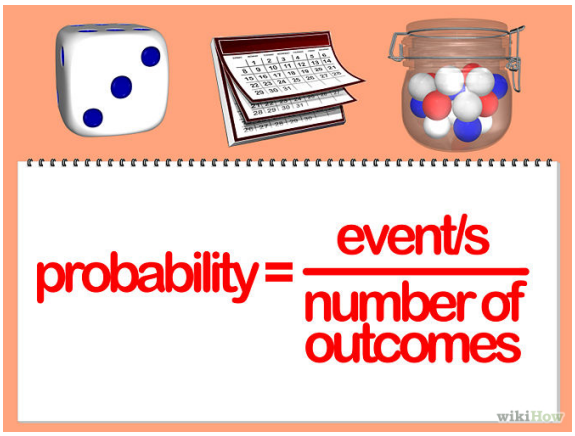
3. Data Ethics

   "During World War II, President Franklin Roosevelt created an Office of Scientific Research and Development to combat diseases such as dysentery, influenza and malaria, diseases that commonly affect soldiers. One of the research teams created a potential vaccine for dysentery. To test it the researchers used orphans and mentally retarded individuals in institutions. The orphans developed dangerously high fevers, thus proving that the vaccine did not work. Another research team purposefully gave psychotic patients at then Illinois State Hospital with malaria, in order to test a cure. Penicillin, the wonder drug of the century, was tested on prisoners to find the most effective dosage. " (Jessica Kiefer)

   Briefly state your opinion.

4. Probability

   There are 27 white balls, 5 red balls and 18 blue balls in a bucket. If you randomly draw a ball from this bucket (every ball has the same chance to be chosen), what is the probability that you get a color ball?



5. Sample Space

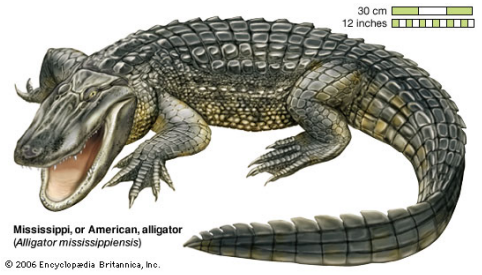
   What is the sample space for choosing 1 letter at random from the word STATISTICS?

6. Confidence Intervals

   Biologists measured 8 American crocodiles (male, adult), the average length of this sample is 14.6 feet. From past experience we know that the standard deviation of the lengths of American crocodiles (male, adult) is 0.74 feet. Find a 99% confidence interval of the mean length for all American crocodiles (male, adult). What is the margin of error?

Table 3: Common Confidence Levels

Confidence level	90%	95%	99%
Critical value z^*	1.645	1.96	2.576



7. Sample Size

   A Company claims its program will allow your computer to download movies quickly. We'll test the free evaluation copy by downloading a movie several times, hoping to estimate the mean download time with a margin of error of only 2 minutes. We think the standard deviation of download times is about 2.5 minutes. How many trial download must we run if we want 90% confidence in our estimate with a margin of error of 2 minutes?

Table 4: Common Confidence Levels

Confidence level	90%	95%	99%
Critical value z^*	1.645	1.96	2.576

8. Hypothesis Test

   Mirex is a chlorinated hydrocarbon that was commercialized as an insecticide and later banned because of its impact on the environment.

   Researchers tested 12 farm-raised salmon for organic contaminants. They found the mean concentration of the carcinogenic insecticide mirex to be 0.0913 parts per million, with sample standard deviation $s = 0.0195$ ppm. As a safety recommendation to recreational fishers, the Environmental Protection Agency's (EPA) recommended "screening value" for mirex is 0.08 ppm. Population distribution is assumed to be normal.

   Are farmed salmon contaminated beyond the level permitted by the EPA? Use the significance level $\alpha = 0.05$.

9. Hypothesis Test

We want to compare ground versus air-based temperature sensors to determine the earths temperature, which is important for agricultural modelling, etc. Ground-based sensors are expensive, and air-based (from satellites or air-planes) of infrared wavelengths may be biased. We collected temperature data from ground and air-based sensors at ten locations, and we want to test if they are different.

Table 5: Temperature

Location	Ground Temperature	Air Temperature
1	46.9	47.3
2	45.4	48.1
3	36.3	37.9
4	31	32.7
5	24.7	26.2
6	22.3	23.3
7	49.8	50.2
8	40.5	42.6
9	37.7	39.4
10	35.5	37.9

State the null hypothesis and alternative hypothesis. Explain what type of test procedure should be applied here. Write down the corresponding formula. (You do not need to compute.)

10. Hypothesis Test

Can you tell how much you are eating from how full you are? Or do you need visual cues? Researchers constructed a table with two ordinary 18 oz soup bowls and two identical-looking bowls that had been modified to slowly, imperceptibly, refill as they were emptied. They assigned experiment participants to the bowls randomly and served them tomato soup. those eating from the ordinary bowls had their bowls refilled by ladle whenever they were one-quarter full. If people judge their portions by internal cues, they should eat about the same amount. How big a difference was there in the amount soup consumed? The table summarizes their results.

Table 6: Summary

	Ordinary bowl	Refilling bowl
n	9	12
$\bar{x}$	8.5 oz	14.7 oz
s	6.1 oz	8.4 oz

Test whether refilling bowls makes participants eat more. Use the significance level $\alpha = 0.05$. Population distribution is assumed to be normal.

11. Hypothesis Test

In the last question (ordinary bowl and refilling bowl), which t-test did you use?

- Two-sample t-test with assumption of equal population SD
- Two-sample t-test with assumption of unequal population SD

Explain why you choose this type of test.

Now use the other t-test and do this question again. What is the p-value by following the other t-test? Which p-value is larger? Can you explain it?

12. Hypothesis Test

Table 7 shows a set of bi-variate data, the fitted linear equation is

$$\hat{y} = 4.2090 + 4.6454x,$$

with correlation $r = 0.9443$.

Table 7: Data

Name	x	y	$x - \bar{x}$	$y - \bar{y}$	$(x - \bar{x})^2$	$(y - \bar{y})^2$	$(x - \bar{x})(y - \bar{y})$	x^2	y^2
Data	1.2	15	-3.62	-11.6	13.10	134.56	41.99	1.44	225
	3.4	12	-1.42	-14.6	2.02	213.16	20.73	11.56	144
	4.6	24	-0.22	-2.6	0.05	6.76	0.57	21.16	576
	5.1	31	0.28	4.4	0.08	19.36	1.23	26.01	961
	9.8	51	4.98	24.4	24.80	595.36	121.51	96.04	2601
Sum	24.1	133	0	0	40.05	969.20	186.04	156.21	4507

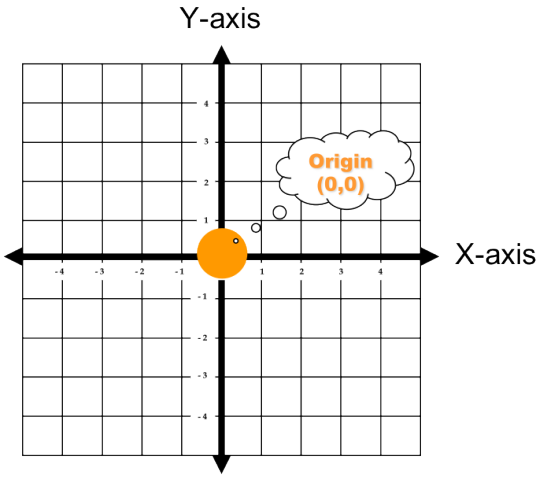
Table 8: Residuals

Name	x	y	ε	ε^2
Data	1.2	15	5.216	27.211
	3.4	12	-8.003	64.056
	4.6	24	-1.578	2.490
	5.1	31	3.099	9.606
	9.8	51	1.266	1.602
Sum	24.1	133	0	104.965

Is the slope significantly different from 0 at the $\alpha = 0.05$ level? State the hypothesis and draw the conclusion.

13. Hypothesis Test

By using the same data set as shown in Table 7, can you conclude that if the regression line passes the origin $(x, y) = (0, 0)$? Use the significance level $\alpha = 0.05$.

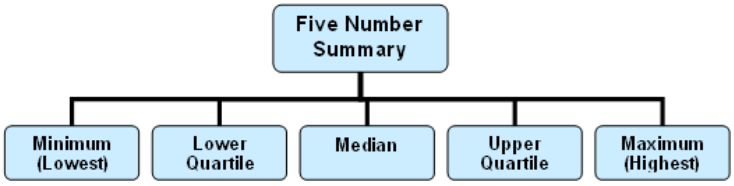


14. Statistical plots

    Based on the summary of Sales, roughly draw a histogram. The sample size is 100. Note: You only need to draw it roughly.

Table 9: Summary pf Sales

median	100
min	95
max	140
first quartile	97
third quartile	108



15. Liner regression

A bi-variate data set is shown in Table 10. The equation of simple linear equation is

$$\hat{y} = 3 + 3x.$$

Table 10: Data and Residuals

x	y	Residual
0	3	
1	7	
2	9	
3	9	
4	17	

Compute residuals. What is the sum of residuals?

End of the Final Exam of Stat 202-0-23