

Stat 139: Midterm

Thursday, October 16, 2014

NAME: _____

This exam consists of 24 pages, including this cover page, exam questions (pp. 2-21), distribution tables (pp. 22-23), and a grading sheet for the teaching staff.

Please write responses to all questions in the space provided. There are three (3) exam questions, with multiple parts in each. Use your time wisely and try to answer as many parts of each question as you can. If you need more space, use the back of the page. However, long explanations will not earn extra points.

You may not collaborate and discuss your answers during the exam; the answers you submit have to be the results of your own efforts. You may not use the textbook, computer, or a cell phone during the exam. You are allowed to have a calculator and two 2-sided pages of notes.

The point values for each question are indicated. There are 100 total points possible. For a full credit you should show your work as well as the final answer.

Feel free to ask the teaching staff if you have any questions, and good luck!

1. (22 points total) This problem has four **unrelated** parts, (a) through (d).

- (a) (3 points) In comparing 10 groups, you notice that \bar{Y}_7 is the largest and \bar{Y}_3 is the smallest, and proceed to test the hypothesis that $\mu_3 - \mu_7 = 0$. Why should a multiple comparison procedure be used even though there is only one comparison being made? Which multiple comparison method seems to be most appropriate?

Any time we look at data, we should use
multiple comparisons

- (b) Suppose a researcher wants to know whether a new experimental drug relieves symptoms attributable to the common cold. The response variable is the time until the cold symptoms go away. She randomized 6 subjects between two groups and observed the following times:

Drug: 36, 55, and 70 hours;

Placebo: 39, 60, and 73 hours.

Her alternative hypothesis is that cold symptoms go away in less time for those who took the drug than for those who took placebo.

Table 1 **on the next page** lists all 20 possible ways to randomize six subjects in two groups of three and specifies corresponding differences in means, medians, and sums of ranks. The six's line with the asterisk is the observed data.

(The question continues on page 4)

Combination number	Drug (Ranks)	Placebo (Ranks)	Difference between means	Difference between medians	Sum of Ranks (for Drug group)
1	36 39 55 (1 2 3)	60 70 73 (4 5 6)	-24.33	-31	6
2	36 39 60 (1 2 4)	55 70 73 (3 5 6)	-21	-31	7
3	36 39 70 (1 2 5)	55 60 73 (3 4 6)	-14.33	-21	8
4	36 39 73 (1 2 6)	55 60 70 (3 4 5)	-12.33	-21	9
5	36 55 60 (1 3 4)	39 70 73 (2 5 6)	-10.33	-15	8
6*	36 55 70 (1 3 5)	39 60 73 (2 4 6)	-3.67	-5	9
7	36 55 73 (1 3 6)	39 60 73 (2 4 5)	-1.67	-5	10
8	36 60 70 (1 4 5)	39 55 73 (2 3 6)	-0.33	5	10
9	36 60 73 (1 4 6)	39 55 70 (2 3 5)	1.67	5	11
10	36 70 73 (1 5 6)	39 55 60 (2 3 4)	8.33	15	12
11	39 55 60 (2 3 4)	36 70 73 (1 5 6)	-8.33	-15	9
12	39 55 70 (2 3 5)	36 60 73 (1 4 6)	-1.67	-5	10
13	39 55 73 (2 3 6)	36 60 70 (1 4 5)	0.33	-5	11
14	39 60 70 (2 4 5)	36 55 73 (1 3 6)	1.67	5	11
15	39 60 73 (2 4 6)	36 55 70 (1 3 5)	3.67	5	12
16	39 70 73 (2 5 6)	36 55 60 (1 3 4)	10.33	15	13
17	55 60 70 (3 4 5)	36 39 73 (1 2 6)	12.33	21	12
18	55 60 73 (3 4 6)	36 39 70 (1 2 5)	14.33	21	13
19	55 70 73 (3 5 6)	36 39 60 (1 2 4)	21	31	14
20	60 70 73 (4 5 6)	36 39 55 (1 2 3)	24.33	31	15

Table 1: List of combinations and corresponding statistics.

(9 points) Perform a randomization test on the difference in means, a randomization test on the difference in medians, and a rank-sum test for these data. Report **exact one-sided** p -values and specify the null and alternative hypotheses for each test. Do the tests agree in their conclusions? Use $\alpha = 0.05$.

All tests are randomized == "is treatment effect = 0

Text

(c) (6 points total) You are using a two-sample t -test to compare average blood pressure in two groups of patients treated with different drugs. After a careful examination of the data you conclude that the assumptions of the t -test are satisfied, including the normality of both populations and equality of their variances, and there are no obvious reasons to suspect dependence or heterogeneity. The t -test rejects the null hypothesis of no difference in average blood pressure.

- i. (3 points) If you run a rank-sum test on the same data, will it also reject the null? Explain.

Must mention power

Rank sum test is least powerful
two-sample t -test is most powerful

The answer: Maybe, because rank sum test has lower power

- ii. (3 points) Suppose that the rank-sum test also rejected the null hypothesis, would you expect a permutation test (using the difference in sample means as a test statistic) to reject the null as well? Explain.

Probably will be rejected by rank sum — b/c it has more power than rank sum.

- (d) (4 points) You were given two samples of the same size. The samples were collected independently. However, you mistakenly treated them as paired data and ran a paired t -test. Which of the following was (were) affected by this mistake: difference in sample means, its standard error, significance level of the test, or reference distribution? Would you have understated or overstated the p -value? Explain.

The statistic for each of the test will be the same (difference in sample means).

Sig level stays the same.

Standard error is the denominator of the t -statistic — the SE will be very close!

Reference distribution: two-sample = $t(2n-2)$ <- has thinner tails

paired = $t(n-1)$ <- has thicker tails

Overestimate: p -value for paired will be greater than p -value for two-sample

2. (62 points total) In this question we will explore various aspects of the real estate market in the greater Boston area. The data set consists of a random sample of home sale records between 10/2013 and 10/2014 in the Boston-Cambridge-Somerville metropolitan area (BCS area). In addition to initial list price and final sale price, each data point contains home characteristics such as living space size (ranging from 2,000 to 2,200 sq.ft), home type (*condo*, *single-family home*, or *townhouse*), and whether it was a short sale (this term will be defined later as it becomes relevant). Table 2 below shows the first five rows of the dataset.

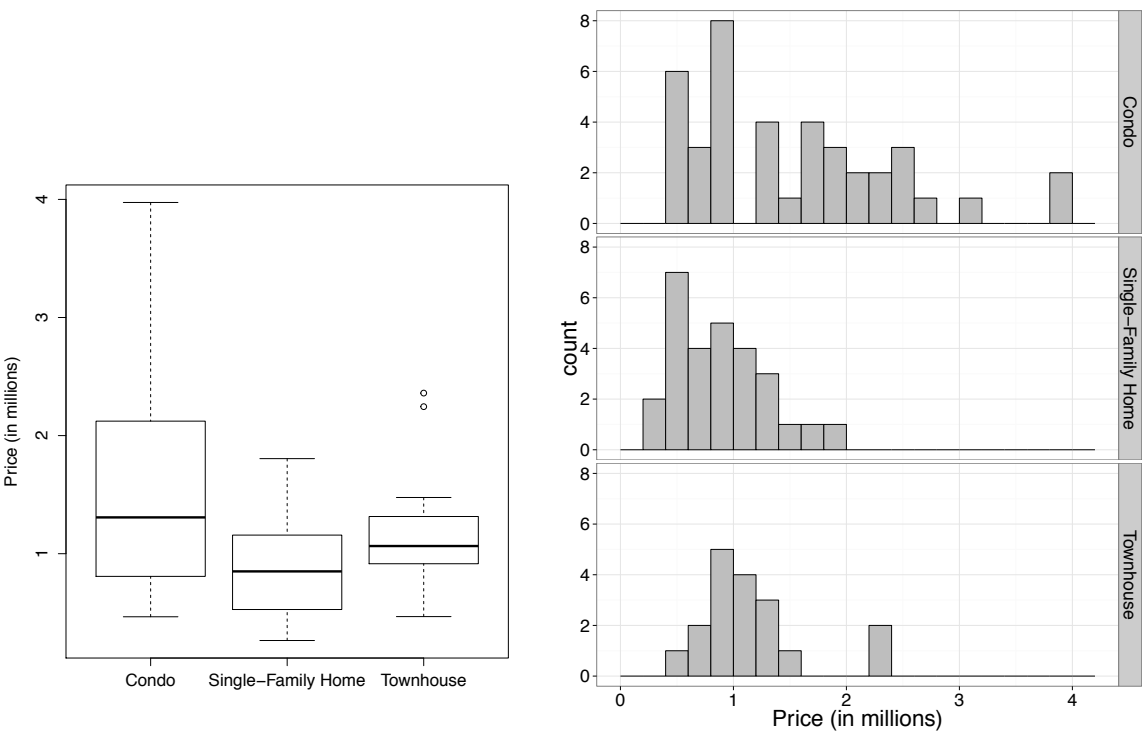
ID #	Type	Sq. ft.	List price, \$	Sale price, \$	Is Short Sale
1	Single-Family Home	2,091	548,500	610,000	No
2	Townhouse	2,051	918,500	915,000	No
3	Condo	2,142	NA	564,000	Yes
4	Townhouse	2,014	895,000	915,000	No
5	Condo	2,145	799,000	795,000	No
...

Table 2: First five rows of the data.

(a) (2 points) What is the study unit?

Home sales

- (b) (15 points total) Suppose your client is deciding between a *townhouse* and a *single-family home* with 2,000 to 2,200 sq.ft. in the BCS area, and he would like to know which option is the most cost-effective.
- i. Below are various graphical summaries and sample statistics, including sample sizes, sample averages, and standard deviations, for sale prices of three type of homes.



i	Group	n_i	\bar{y}_i , \$ million	s_i , \$ million
1	Condo	40	1.54	0.92
2	Single-Family Home	28	0.87	0.40
3	Townhouse	18	1.17	0.46

Table 3: Summary statistics for home sale prices.

(The question continues on the next page.)

(5 points) Use this information to answer the question whether there is a difference in average sale prices between *townhouses* and *single-family homes* with 2,000 to 2,200 sq.ft. located in the BCS area.

Name the statistical test that you are performing and specify the hypotheses that you are testing, defining and using symbols to represent population parameters. Show your calculation of a test statistic and specify a corresponding (approximate) p -value.

Cannot pool variances, b/c they are different b/w groups.

Just pool two variance that are quite similar

- ii. (3 points) Find a 95% confidence interval based on the test performed in part 2(b)i. Would you reject a hypothesis that the center of the distribution of sale prices for *townhouses* is greater than the one for *single-family homes* by \$400,000 (i.e., \$0.4 million)?

- iii. (4 points) Specify and comment on all assumptions of the test performed in part 2(b)i.

Independence and homogeneity are issues.

- iv. (3 points) Based on your results in part 2(b)i, state your conclusion in the context of your client's question of interest. Use $\alpha = 0.05$.

Is it a buyer's or seller's market?

- (c) (20 points total) Your colleague approached you with a request for statistical assistance. Her client is interested in learning whether there is evidence that distributions of prices on all three types of homes, i.e., condos, townhouses and single-family homes between 2,000 to 2,200 sq.ft., are the same or not. The colleague is working with the same data, however, she remembers learning in her statistics classes that it is recommended to log-transform cost data, so she applied the **natural log transformation** to all sale prices.
- i. (2 points) Help your colleague formulate the null and alternative hypotheses on the original scale of the data.

Wrong: $H_0: \ln(\mu_1) = \ln(\mu_2) = \ln(\mu_3)$

Correct: $H_0: \text{median}(\text{group1}) = \text{median}(\text{group2}) = \text{median}(\text{group3})$

log transformation preserves the median : median on log scale = log of median on original scale

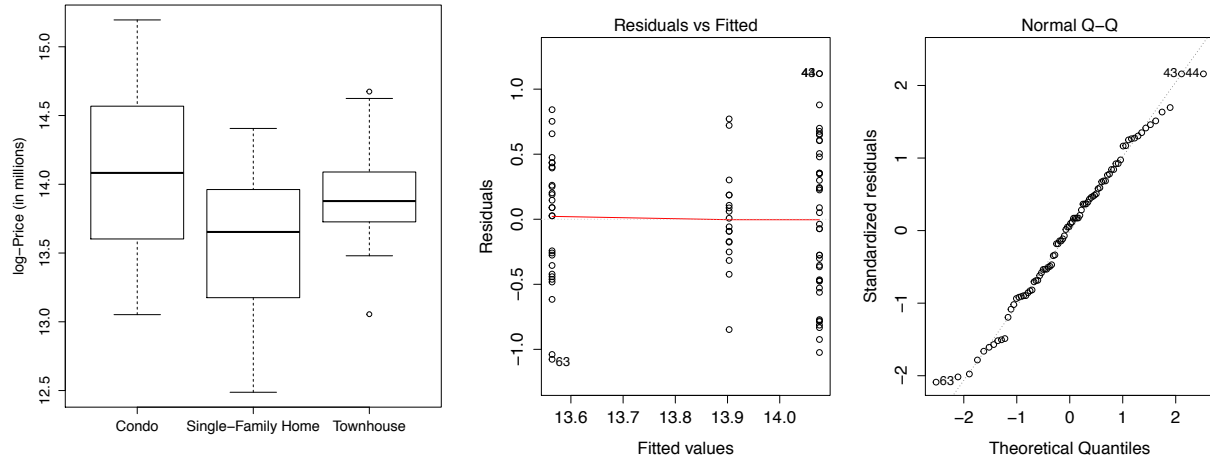
- ii. (4 points) Complete the ANOVA table below. Do you reject the null hypothesis? Use $\alpha = 0.05$.

SSR type	SSR value	<i>d.f.</i>	Mean Square	F-statistic	<i>p</i> -value
<i>SSB</i>					
<i>SSW</i>	22.86			—	—
<i>SST</i>			0.32	—	—

iii. (4 points) Suppose you rejected the null hypothesis in part 2(c)ii. Formulate the conclusion that your colleague can present to her client. Comment on the scope of inference.

iv. (4 points) Specify all the assumptions required for the test in part 2(c)ii. (You may state them for the log-transformed data).

- v. (4 points) Comment on whether the assumptions are plausible using graphical summaries below as well as the information about the study design on page 6. Compare the graphical summaries to those on page 7: Is the log-transformation justified?



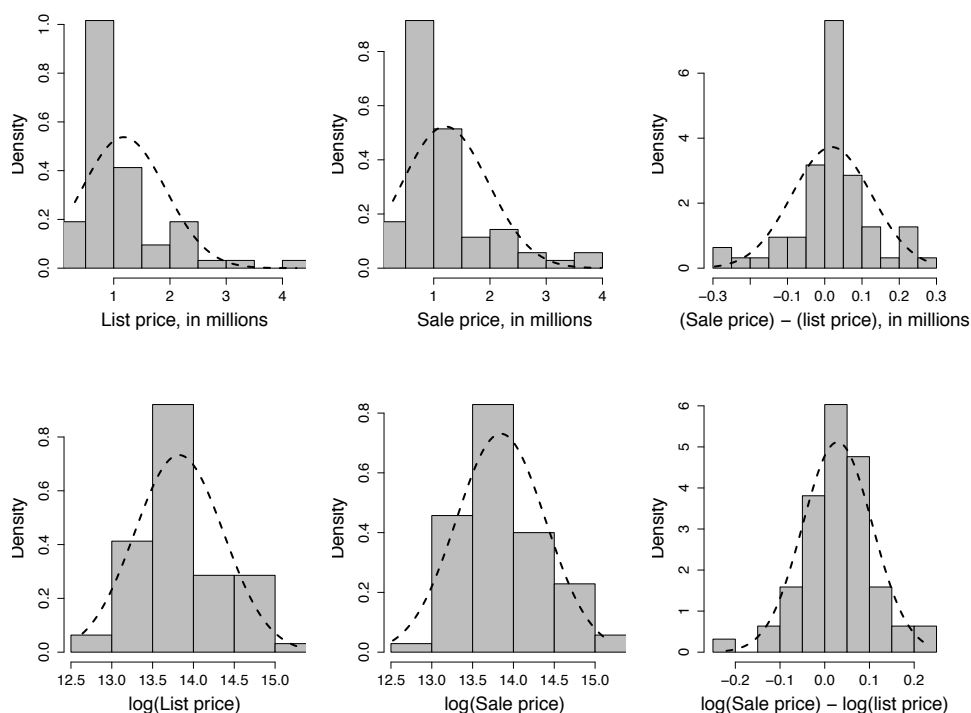
log transformation is justified.

- vi. (2 points) Name an alternative test that your colleague can run to answer her question of interest that does not assume any particular distribution for the data.

Kruskal-wallis Test

- (d) (25 points total) Another question that your client is interested in is whether he is in a “seller’s” or a “buyer’s” market. One way to answer it is to compare original list prices with final sale prices and see if buyers are offering more (seller’s market) or less (buyer’s market) than the asking price. You asked your assistant to explore this question, however, he came back with various graphical summaries of the data and multiple R outputs of statistical tests, unsure which one is the most appropriate to use.

Choose test with transformation & do paired test.



i	Group	n_i	\bar{y}_i , \$ million	\bar{y}_i (log scale)	s_i , \$ million	s_i (log scale)
1	List price	62	1.18	13.82	0.74	0.54
2	Sale price	62	1.21	13.85	0.76	0.55

Output (I):

Two Sample t-test

data: SALE.PRICE and LIST.PRICE

t = 0.2084, df = 131, p-value = 0.8353

alternative hypothesis: true difference in means is not equal to 0

95 percent confidence interval:

-0.2315060 0.2860204

sample estimates:

mean of x mean of y

1.205284 1.178027

Output (II)

One Sample t-test

```
data: SALE.PRICE - LIST.PRICE
t = 1.3945, df = 62, p-value = 0.1682
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
 -0.008155375  0.045780486
sample estimates:
mean of x
0.01881256
```

Output (III):

Welch Two Sample t-test

```
data: SALE.PRICE and LIST.PRICE
t = 0.2087, df = 130.186, p-value = 0.835
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -0.2311491  0.2856634
sample estimates:
mean of x mean of y
1.205284  1.178027
```

Output (IV):

Two Sample t-test

```
data: log(SALE.PRICE) and log(LIST.PRICE)
t = 0.2327, df = 131, p-value = 0.8163
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -0.1652468  0.2093159
sample estimates:
mean of x mean of y
13.84613  13.82409
```

Output (V)

One Sample t-test

```
data: log(SALE.PRICE) - log(LIST.PRICE)
t = 2.9177, df = 62, p-value = 0.004907
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
 0.00902699 0.04831010
sample estimates:
mean of x
0.02866854
```

Output (VI):

Welch Two Sample t-test

```
data: log(SALE.PRICE) and log(LIST.PRICE)
t = 0.2328, df = 129.608, p-value = 0.8163
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -0.1652397 0.2093087
sample estimates:
mean of x mean of y
13.84613 13.82409
```

- i. (5 points) Decide which R output is the most appropriate for this question and indicate its number. Specify which test is being performed and explain your choice.

- ii. (4 points) Write down the hypotheses that are being tested in the test chosen in part 2(d)i in terms of parameters defined for distributions of list and sale prices. Specify the parameter that is being estimated and report the corresponding estimate. Also, report the test statistic and the p -value.

Original Scale:

$$D_i = X_i - Y_i, D_i \sim N(\mu, \sigma^2)$$

$$H_0: \mu = 0,$$

$$H_a: \mu \neq 0$$

Log Transformed Scale (correct):

$$D_i = \log(X_i) - \log(Y_i), \text{ where } X_i \text{ is sale price, } Y_i \text{ is list price}$$

$$D_i = \log(X_i/Y_i)$$

$$H_0: \text{median}(X_i/Y_i) = 1$$

- iii. (3 points) State your conclusion in the context of the original question of interest (stated on page 13). Use $\alpha = 0.05$.

- iv. (4 points) What is the 95% confidence interval for the parameter estimated in part 2(d)ii? Interpret the interval in the context of the original question of interest.

Back-transform to get correct scale.

- v. (4 points) List all the assumptions required for the test that you used in part 2(d)i.

Don't need equal variances.

- vi. (2 points) Name at least two alternative tests that your assistant can run to answer the question of interest, that do not assume any particular distribution for the data.

- vii. (3 points) You noticed that the sample size in your assistant's analysis was smaller than the original one. Further investigation revealed that all records on short sales (i.e., homes sold out of necessity and, usually, for less than they are worth) are missing list prices and, therefore, were not included in the analysis. Comment on whether excluding these sales affected the analysis and, if so, how. Think of a situation when it may be acceptable to exclude them.

3. (16 points total) Let

$$\begin{aligned} Y_{1j} &\stackrel{\text{i.i.d.}}{\sim} N(\mu_1, \sigma^2) \text{ for } j = 1, 2, \dots, n_1, \\ Y_{2j} &\stackrel{\text{i.i.d.}}{\sim} N(\mu_2, \sigma^2) \text{ for } j = 1, 2, \dots, n_2, \text{ and} \\ Y_{3j} &\stackrel{\text{i.i.d.}}{\sim} N(\mu_3, \sigma^2) \text{ for } j = 1, 2, \dots, n_3 \end{aligned}$$

be independent samples from the specified Normal distributions.

(a) (2 points) What is the sampling distribution of the sample variance for the i 'th sample,

$$S_i^2 = \frac{\sum_{j=1}^{n_i} (Y_{ij} - \bar{Y}_i)^2}{n_i - 1}, i = 1, 2, \text{ or } 3,$$

where \bar{Y}_i is the sample mean for the i 'th sample?

Shown in homework.

(b) (2 points) Write down the estimator of the pooled variance, S_p^2 , that uses all three samples to estimate σ^2 .

(c) (4 points) Using parts 3a and 3b, derive the sampling distribution of S_p^2 .

(d) (2 points) What is the expected value of S_p^2 ?

- (e) (4 points) In class we learned that the sum of squared residuals for a separate-means model,

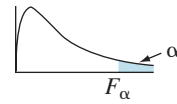
$$SSR = \sum_{i=1}^3 \sum_{j=1}^{n_i} (Y_{ij} - \bar{Y}_i)^2,$$

can be expressed in terms of the pooled sample variance as follows,

$$SSR = (n_1 + n_2 + n_3 - 3)S_p^2.$$

Using definitions of S_p^2 and S_i^2 , $i = 1, 2, \text{ and } 3$, show that this equality holds.

- (f) (2 points) Use the equality in the previous part and the result in part 3c to determine the sampling distribution of SSR.

**TABLE 8**Percentage points of the F distribution (df_2 at least 40)

df_2	α	df_1									
		1	2	3	4	5	6	7	8	9	10
40	.25	1.36	1.44	1.42	1.40	1.39	1.37	1.36	1.35	1.34	1.33
	.10	2.84	2.44	2.23	2.09	2.00	1.93	1.87	1.83	1.79	1.76
	.05	4.08	3.23	2.84	2.61	2.45	2.34	2.25	2.18	2.12	2.08
	.025	5.42	4.05	3.46	3.13	2.90	2.74	2.62	2.53	2.45	2.39
	.01	7.31	5.18	4.31	3.83	3.51	3.29	3.12	2.99	2.89	2.80
	.005	8.83	6.07	4.98	4.37	3.99	3.71	3.51	3.35	3.22	3.12
	.001	12.61	8.25	6.59	5.70	5.13	4.73	4.44	4.21	4.02	3.87
60	.25	1.35	1.42	1.41	1.38	1.37	1.35	1.33	1.32	1.31	1.30
	.10	2.79	2.39	2.18	2.04	1.95	1.87	1.82	1.77	1.74	1.71
	.05	4.00	3.15	2.76	2.53	2.37	2.25	2.17	2.10	2.04	1.99
	.025	5.29	3.93	3.34	3.01	2.79	2.63	2.51	2.41	2.33	2.27
	.01	7.08	4.98	4.13	3.65	3.34	3.12	2.95	2.82	2.72	2.63
	.005	8.49	5.79	4.73	4.14	3.76	3.49	3.29	3.13	3.01	2.90
	.001	11.97	7.77	6.17	5.31	4.76	4.37	4.09	3.86	3.69	3.54
90	.25	1.34	1.41	1.39	1.37	1.35	1.33	1.32	1.31	1.30	1.29
	.10	2.76	2.36	2.15	2.01	1.91	1.84	1.78	1.74	1.70	1.67
	.05	3.95	3.10	2.71	2.47	2.32	2.20	2.11	2.04	1.99	1.94
	.025	5.20	3.84	3.26	2.93	2.71	2.55	2.43	2.34	2.26	2.19
	.01	6.93	4.85	4.01	3.53	3.23	3.01	2.84	2.72	2.61	2.52
	.005	8.28	5.62	4.57	3.99	3.62	3.35	3.15	3.00	2.87	2.77
	.001	11.57	7.47	5.91	5.06	4.53	4.15	3.87	3.65	3.48	3.34
120	.25	1.34	1.40	1.39	1.37	1.35	1.33	1.31	1.30	1.29	1.28
	.10	2.75	2.35	2.13	1.99	1.90	1.82	1.77	1.72	1.68	1.65
	.05	3.92	3.07	2.68	2.45	2.29	2.18	2.09	2.02	1.96	1.91
	.025	5.15	3.80	3.23	2.89	2.67	2.52	2.39	2.30	2.22	2.16
	.01	6.85	4.79	3.95	3.48	3.17	2.96	2.79	2.66	2.56	2.47
	.005	8.18	5.54	4.50	3.92	3.55	3.28	3.09	2.93	2.81	2.71
	.001	11.38	7.32	5.78	4.95	4.42	4.04	3.77	3.55	3.38	3.24
240	.25	1.33	1.39	1.38	1.36	1.34	1.32	1.30	1.29	1.27	1.27
	.10	2.73	2.32	2.10	1.97	1.87	1.80	1.74	1.70	1.65	1.63
	.05	3.88	3.03	2.64	2.41	2.25	2.14	2.04	1.98	1.92	1.87
	.025	5.09	3.75	3.17	2.84	2.62	2.46	2.34	2.25	2.17	2.10
	.01	6.74	4.69	3.86	3.40	3.09	2.88	2.71	2.59	2.48	2.40
	.005	8.03	5.42	4.38	3.82	3.45	3.19	2.99	2.84	2.71	2.61
	.001	11.10	7.11	5.60	4.78	4.25	3.89	3.62	3.41	3.24	3.09
inf.	.25	1.32	1.39	1.37	1.35	1.33	1.31	1.29	1.28	1.27	1.25
	.10	2.71	2.30	2.08	1.94	1.85	1.77	1.72	1.67	1.63	1.60
	.05	3.84	3.00	2.60	2.37	2.21	2.10	2.01	1.94	1.88	1.83
	.025	5.02	3.69	3.12	2.79	2.57	2.41	2.29	2.19	2.11	2.05
	.01	6.63	4.61	3.78	3.32	3.02	2.80	2.64	2.51	2.41	2.32
	.005	7.88	5.30	4.28	3.72	3.35	3.09	2.90	2.74	2.62	2.52
	.001	10.83	6.91	5.42	4.62	4.10	3.74	3.47	3.27	3.10	2.96

Teaching Staff Only

Problem	Points	Possible
1		22
2		62
3		16
Total		100