Statistics 2215 **Exam # 1** **Spring 2009**

Name: Solutions .

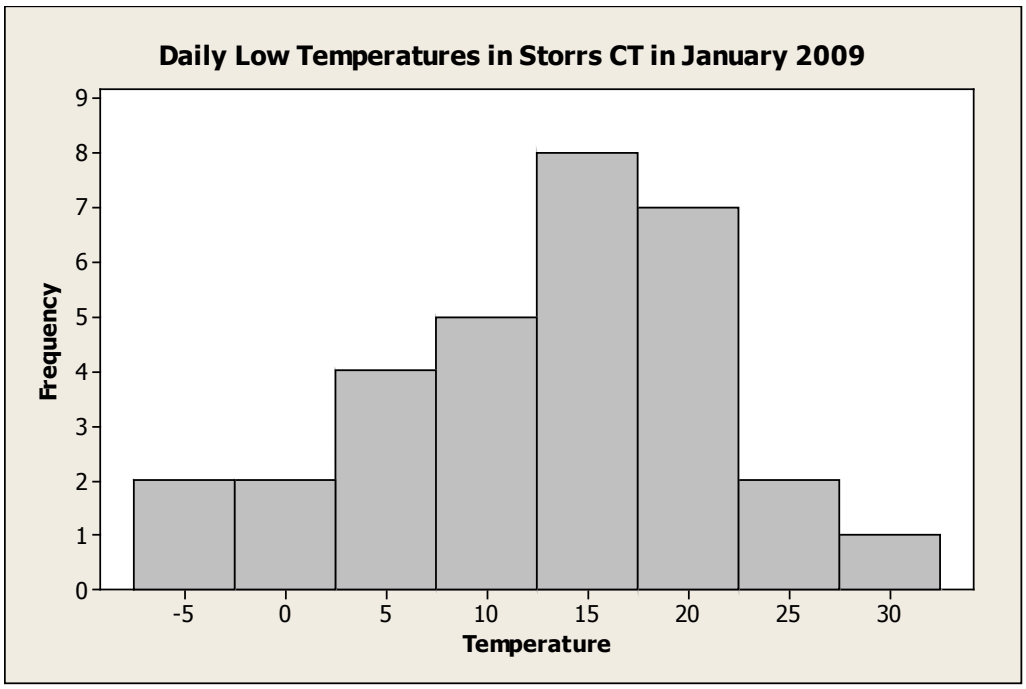
Please answer the following questions. There are some short answer questions and some computational questions. Partial credit will be given, so showing your work is a good idea. Raise your hand if you have any questions, and I will be by to assist you.

Note that the questions have different point values. Use your time wisely.

The *t*-table is attached in case you need it.

Good luck!!

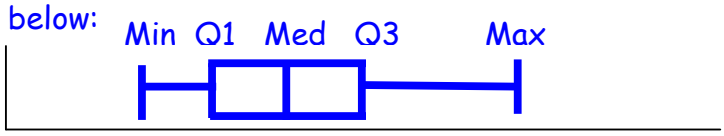
Question 1: (6 pts) A histogram of the daily low temperatures in Storrs for January 2009 is given below. Describe the distribution of the temperatures.



The low temperatures have a distribution that looks somewhat normal, or possibly skewed to the left. We have a single-peaked, or unimodal, distribution that is centered at about 12 degrees, with a spread from -5 to 30 degrees.

Question 2: (7 pts) Describe how a boxplot is constructed, and sketch an example.

A boxplot relies on the five number summary: min, Q1, median, Q3, max. The plot only plots these five points with short lines, and connects the lines from Q1 to Q3 to make a box. A sketch is shown below:



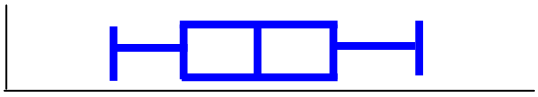
Question 3: (5 pts) Give an example of some data that would be skewed to the right.

Some data that would be right-skewed are the salaries of a major league baseball teams. Most of the players earn about the same amount, but the few superstars on a team earn much more, skewing the distribution to the right.

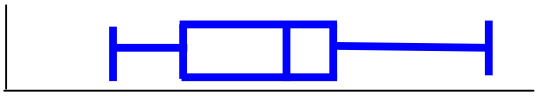
Question 4: (6 pts) What general features are evident in a box plot of data from a normal distribution? How do these features differ when the data come from a skewed distribution?

A boxplot for normal data will have whiskers that are about the same length, and the median line will lie in the center of the box. For skewed data, the whiskers will be uneven, and the median may no longer be centered in the box. Examples are sketched below:

NORMAL



SKEWED



Question 5: (6 pts) A study found that individuals who lived in houses with more than two bathrooms tended to have higher blood pressure than individuals who lived in houses with two or fewer bathrooms.

Can cause-and-effect be determined from this? (Please justify) If not, list a possible confounding variable that might explain this result.

This is an observational study, so cause-and-effect cannot be established. We may establish and *association* between number of bathrooms and blood pressure, but it is probably not true. People with larger homes probably have larger families, or have more stressful, higher paying jobs that are the true cause of the higher blood pressure.

Question 6: (6 pts) Botanists observed 30 bristlecone pines and estimated their ages. A 95% confidence interval for the mean age of bristlecone pines was calculated to be (1775 , 4225) years.

In addition the botanists wanted to do a hypothesis test. Let μ be the mean age of bristlecone pines. The botanists want to test $H_0 : \mu = 4000$ versus $H_a : \mu \neq 4000$. They plan to use a significance level of $\alpha = 0.05$.

Based on the information given, what can you say about the *p*-value for such a hypothesis test?

Notice that the null hypothesis value of 4000 years lies within the confidence interval (1775 , 4225). The mean age is not significantly different from 4000. In a hypothesis test, this means that we would not have rejected the null hypothesis. In such a situation, the *p*-value would have to be larger than *α*, or bigger than 0.05.

Question 7: (6 pts) Suppose the following statement is made in a statistical summary:

"A comparison of breathing capacities of individuals in households with low nitrogen dioxide levels and individuals in households with high nitrogen dioxide levels indicated that there is no difference in the means (two-sided *p*-value = 0.04)." What is wrong with this statement?

The writers have made the wrong conclusion! With a *p*-value of 0.04, they should have concluded that there WAS a difference in the means.

Question 8: (4 pts each) Which *t*-test would you use in each of the following situations? Options are: 1-sample t-test, matched pairs *t*-test, two-sample *t*-test.

- You are comparing the job placement success of UCONN Business School graduates with those of Yale. You randomly sample 25 UCONN graduates and 15 Yale graduates, and record the starting salary of each graduate. What test could you use to determine whether the starting salary of UCONN Business School graduates is less than that of Yale graduates?

  Two-sample *t*-test .

- To report a mileage estimate to the EPA for a new sedan, a car manufacturer randomly selects 18 cars from their production line. They test each car under identical conditions, and record the mileage per gallon for each car. They wish to test the claim that the mean mileage for this sedan is over 30 mpg. What test would they use to do this?

  One-sample *t*-test .

- Drug companies do a lot of clinical trials while researching their products. Early in drug development, these companies conduct trials of their drug on normal, healthy individuals. In a "crossover design," each sample person receives both the drug and a placebo. Measurements are made each time to record information such as blood pressure when using the placebo and blood pressure when using the drug. The company wishes to claim that their drug lowers blood pressure. What test could be used to test such a statement?

  Matched pairs *t*-test .

- Are Idaho's "famous potatoes" really better? The Idaho Potato Growers Association wants to find out. They randomly sample 50 people. From this sample, 25 people are randomly selected to sample a baked Idaho potato and rank their satisfaction from 1 to 10. The remaining 25 people sample a baked Maine potato and rank their satisfaction on the same scale. The Association would like to claim that the mean satisfaction rating of Idaho potatoes is higher than that of Maine potatoes. Which *t*-test could be used to test this claim?

  Two-sample *t*-test .

Use the following scenario to answer Question 9.

In 1964 there was a study that contrasted cholesterol levels between urban and rural Guatemalans. The data along with some summary statistics and graphs of the data are shown below.

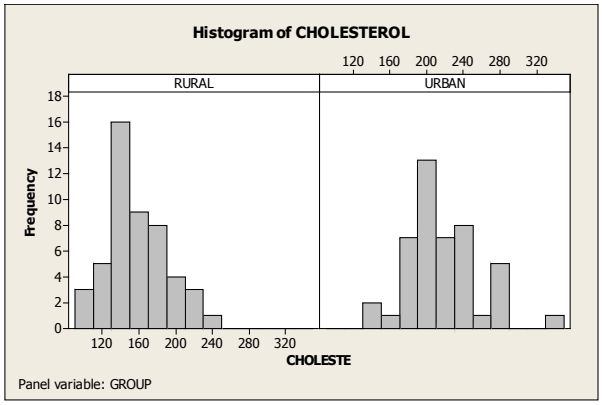
Cholesterol levels in urban and rural guatemalans

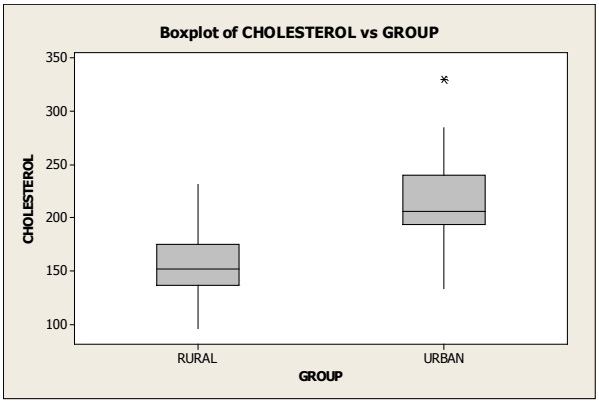
Serum total cholesterol (mg/l) levels among *urban* residents (n=45)
133 134 155 170 175 179 181 184 188 189 190 196 197 199 200 200 201 201 204 205
205 205 206 214 217 222 222 227 227 228 234 234 236 239 241 242 244 249 252 273
279 284 284 284 330

Serum total cholesterol (mg/l) levels among *rural* residents (n=49)
95 108 108 114 115 124 129 129 131 131 135 136 136 139 140 142 142 143 143 144
144 145 145 148 152 152 155 157 158 158 162 165 166 171 172 173 174 175 180 181
189 192 194 197 204 220 223 226 231

Descriptive Statistics: CHOLESTEROL

Variable	GROUP	n	Mean	StDev
CHOLESTEROL	RURAL	49	157.00	31.76
	URBAN	45	216.87	39.92





Question 9: (10 pts) Researchers wanted to show that the mean cholesterol level for urban Guatemalans was higher than that of rural Guatemalans. Use a statistical method to establish whether this is the case. You can use either a hypothesis test or a confidence interval, but you should justify which procedure you choose.

Calculate the test or confidence interval. Be sure to state your hypotheses for a hypothesis test. If you need it, you may use the fact that $S_p = 35.8948$ without calculating it. What do you conclude?

This analysis requires a two-sample *t*-test or interval, because there are two groups. The two standard deviations are very similar $\left(\frac{39.92}{31.76} = 1.26 < 1.7\right)$. Therefore, the equal variances *t*-test or interval should be used.

Two-sample t-test (equal variances)

$H_0 : \mu_{urban} = \mu_{rural}$ vs $H_a : \mu_{urban} > \mu_{rural}$

By hand

Test statistic:

$$t = \frac{\bar{y}_2 - \bar{y}_1}{s_p\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}} = \frac{216.87 - 157}{35.8948\sqrt{\frac{1}{49}+\frac{1}{45}}}$$

$= 8.08$, with $n_1 + n_2 - 2 = 92$ d.f.

p-value and conclusion:
Looking on row 90 of the t-Table, this is off the chart! We know the p-value < 0.005.

By TI-83/84

p-value $= 1.24 \times 10^{-12} \approx 0$

We reject the null hypothesis, and conclude that the mean cholesterol for urban Guatemalans is higher than for rural ones.

95% Confidence Interval

By hand

We have $n_1 + n_2 - 2 = 92$ degrees of freedom. Looking at row 90 of the t-table, the critical t-value for a 95% CI is $t = 1.987$.

$$\left(\bar{y}_2 - \bar{y}_1\right) \pm t \bullet s_p\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}$$

$$= \left(216.87 - 157\right) \pm 1.987\left(35.8948\right)\sqrt{\frac{1}{49}+\frac{1}{45}}$$

$= 59.87 \pm 14.7261$

$= (45.1439 , 74.5961)$

By TI-83/84

Interval is (45.151 , 74.589)

We are 95% confident that the true mean difference in cholesterol is between 45.151 and 74.589. Since zero is not in this interval, we conclude that the mean cholesterol for urban Guatemalans is higher than for rural ones.

Use the following scenario and Minitab output to answer Questions 10 – 14.

A group of scientists was interested in studying air pollution. One component of air pollution is airborne particulate matter such as dust and smoke. To measure particulate pollution, a vacuum motor draws air through a filter for 24 hours. The filter is weighed at the beginning and at the end of the period. The weight gained over the 24 hour period is a measure of the concentration of particles in the air. This study made measurements in the center of a small city and at a rural location 10 miles southwest of the city. The data are shown below:

Location	Particulate Level (grams)
Rural	67, 42, 33, 46, 43, 54, 38, 88, 108, 57, 70, 42, 43, 39, 52, 48, 56, 44, 51, 21, 74, 48, 84, 51, 43, 45, 41, 47, 35
City	39, 68, 42, 34, 48, 82, 45, 60, 57, 39, 123, 59, 71, 41, 42, 38, 57, 50, 58, 45, 69, 23, 72, 49, 86, 51, 42, 46, 44, 42

The alternative hypothesis used in this analysis is the 2-sided (not equal) hypothesis. Equal variances for the two populations were assumed. Notice that two pieces of information, the degrees of freedom (df) and the T-Value (the observed *t* statistic), have been left blank.

Two-Sample T-Test and CI: Rural, City

```
Two-sample T for Rural vs City

        N  Mean  StDev  SE Mean
Rural  29  52.1   18.2      3.4
City   30  54.1   19.4      3.5


Difference = mu (Rural) - mu (City)
Estimate for difference:  -1.99770
95% CI for difference:  (-11.81540, 7.82000)
T-Test of difference = 0 (vs not =): T-Value = 0.4079  P-Value = 0.685  DF = 57
Both use Pooled StDev = 18.8269
```

S_p

Question 10: (4 pts) How many degrees of freedom are associated with the *t* statistic for this problem?

$$n_1 + n_2 - 2 = 29 + 30 - 3 = 57 \text{ d.f.}$$

Question 11: (6 pts) Write down the appropriate formula for the *t* statistic value for this analysis, and calculate the *t* value based on the information provided in the Minitab output.

We're using a two-sample *t*-test with equal variances:

$$t = \frac{\bar{Y}_2 - \bar{Y}_1}{S_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} = \frac{54.1 - 52.1}{18.8269\sqrt{\frac{1}{29} + \frac{1}{30}}} = \frac{2}{4.9028} = 0.4079$$

The TI-83/84 gives $t = 0.4081$.

Question 12: (6 pts) What would you conclude about the difference in particulate pollution between the rural and city locations? Please justify the reason for your conclusion.

Since the *p*-value is 0.685, we do not reject H_0. There is no difference in mean particulate pollution for rural versus city locations.

(Continued from the previous page)

Question 13: (6 pts) Name two of the assumptions the data must satisfy in order for the conclusions based on the *t*-test to be valid.

Assumption #1- The data must follow a normal distribution.

Assumption #2- The two groups must have equal variances

Assumption #3- The two groups must be independent

Question 14: (10 pts) Additional graphical output is given below. Discuss the validity of the assumptions you listed in Question 10 on the basis of the graphical output. Remember to cite the graph number you are referring to.

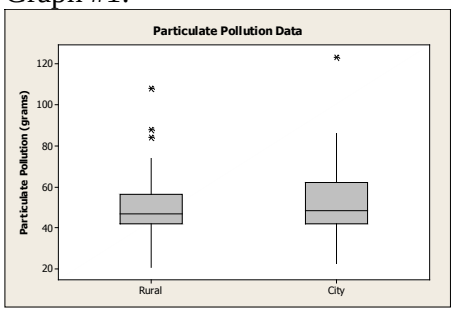
Based on your discussion of the validity of the assumptions, would you still conclude that the two-sample *t*-test with equal variances is appropriate?

All four graphs indicate there are problems with the normality assumptions. The Q-Q plots are not very linear, and the histograms look skewed to the right. The variances are pretty similar $\left(\frac{3.5}{3.4} = 1.03 < 1.7\right)$, though. Also, the boxplot (Graph #1) indicates that there are some outliers in both groups. Use of the *t*-procedure might be questionable. However, here we have:

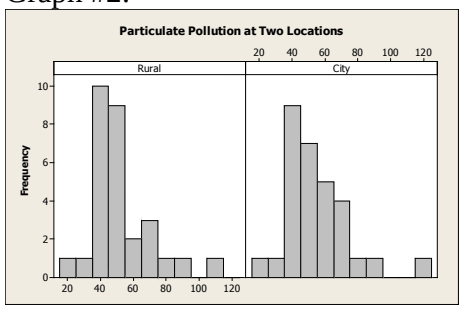
- $n_1 \approx n_2$
- $s_1 \approx s_2$
- Skewness in the same direction.

The *t*-procedure is actually still moderately reliable in this situation. The two-sample t-test with equal variances may still be appropriate.

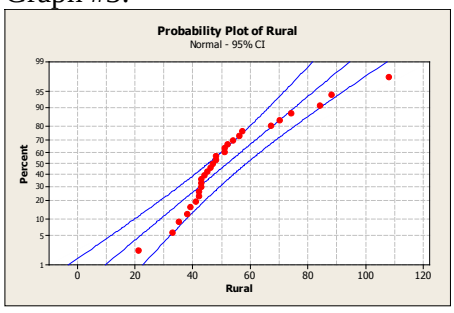
Graph #1:



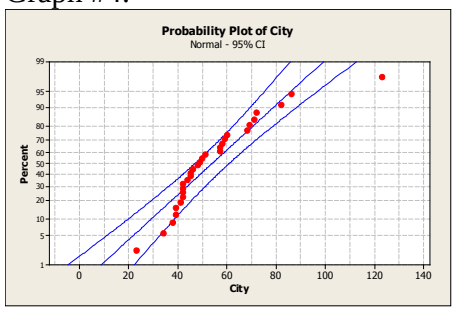
Graph #2:



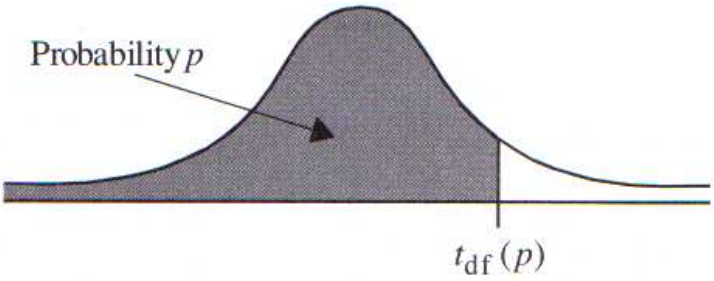
Graph #3:



Graph #4:



Selected Percentiles of *t*-Distributions



Tabled values are $t_{df}(p)$

	Probability p											
d.f.	*.75*	*.80*	*.85*	*.90*	*.95*	*.975*	*.98*	*.99*	*.995*	*.9975*	*.999*	*.9995*
1	1.000	1.376	1.963	3.078	6.314	12.71	15.89	31.82	63.67	127.3	318.3	636.6
2	0.816	1.061	1.386	1.886	2.920	4.303	4.849	6.965	9.925	14.09	22.32	31.60
3	0.765	0.978	1.250	1.638	2.353	3.182	3.482	4.541	5.841	7.453	1.215	12.92
4	0.741	0.941	1.190	1.533	2.132	2.776	2.999	3.747	4.604	5.598	7.173	8.610
5	0.727	0.920	1.156	1.476	2.015	2.571	2.757	3.365	4.032	4.773	5.893	6.869
6	0.718	0.906	1.134	1.440	1.943	2.447	2.612	3.143	3.707	4.317	5.208	5.959
7	0.711	0.896	1.119	1.415	1.895	2.365	2.517	2.998	3.499	4.029	4.785	5.408
8	0.706	0.889	1.108	1.397	1.860	2.306	2.449	2.896	3.355	3.833	4.501	5.041
9	0.703	0.883	1.100	1.383	1.833	2.262	2.398	2.821	3.250	3.690	4.297	4.781
10	0.700	0.879	1.093	1.372	1.812	2.228	2.359	2.764	3.169	3.581	4.144	4.587
11	0.697	0.876	1.088	1.363	1.796	2.201	2.328	2.718	3.106	3.497	4.025	4.437
12	0.695	0.873	1.083	1.356	1.782	2.179	2.303	2.681	3.055	3.428	3.930	4.318
13	0.694	0.870	1.079	1.350	1.771	2.160	2.282	2.650	3.012	3.372	3.852	4.221
14	0.692	0.868	1.076	1.345	1.761	2.145	2.264	2.624	2.977	3.326	3.787	4.140
15	0.691	0.866	1.074	1.341	1.753	2.131	2.249	2.602	2.947	3.286	3.733	4.073
16	0.690	0.865	1.071	1.337	1.746	2.120	2.235	2.583	2.921	3.252	3.686	4.015
17	0.689	0.863	1.069	1.333	1.740	2.110	2.224	2.567	2.898	3.222	3.646	3.965
18	0.688	0.862	1.067	1.330	1.734	2.101	2.214	2.552	2.878	3.197	3.610	3.922
19	0.688	0.861	1.066	1.328	1.729	2.093	2.205	2.539	2.861	3.174	3.579	3.883
20	0.687	0.860	1.064	1.325	1.725	2.086	2.197	2.528	2.845	3.153	3.552	3.850
21	0.686	0.859	1.063	1.323	1.721	2.080	2.189	2.518	2.831	3.135	3.527	3.819
22	0.686	0.858	1.061	1.321	1.717	2.074	2.183	2.508	2.819	3.119	3.505	3.792
23	0.685	0.858	1.060	1.319	1.714	2.069	2.177	2.500	2.807	3.104	3.485	3.768
24	0.685	0.857	1.059	1.318	1.711	2.064	2.172	2.492	2.797	3.091	3.467	3.745
25	0.684	0.856	1.058	1.316	1.708	2.060	2.167	2.485	2.787	3.078	3.450	3.725
26	0.684	0.856	1.058	1.315	1.706	2.056	2.162	2.479	2.779	3.067	3.435	3.707
27	0.684	0.855	1.057	1.314	1.703	2.052	2.158	2.473	2.771	3.057	3.421	3.690
28	0.683	0.855	1.056	1.313	1.701	2.048	2.154	2.467	2.763	3.047	3.408	3.674
29	0.683	0.854	1.055	1.311	1.699	2.045	2.150	2.462	2.756	3.038	3.396	3.659
30	0.683	0.854	1.055	1.310	1.697	2.042	2.147	2.457	2.750	3.030	3.385	3.646
40	0.681	0.851	1.050	1.303	1.684	2.021	2.123	2.423	2.704	2.971	3.307	3.551
50	0.679	0.849	1.047	1.299	1.676	2.009	2.109	2.403	2.678	2.937	3.261	3.496
60	0.679	0.848	1.045	1.296	1.671	2.000	2.099	2.390	2.660	2.915	3.232	3.460
70	0.678	0.847	1.044	1.294	1.667	1.994	2.093	2.381	2.648	2.899	3.211	3.435
80	0.678	0.846	1.043	1.292	1.664	1.990	2.088	2.374	2.639	2.887	3.195	3.416
90	0.677	0.846	1.042	1.291	1.662	1.987	2.084	2.368	2.632	2.878	3.183	3.402
100	0.677	0.845	1.042	1.290	1.660	1.984	2.081	2.364	2.626	2.871	3.174	3.390
500	0.675	0.842	1.038	1.283	1.648	1.965	2.059	2.334	2.586	2.820	3.107	3.310
1000	0.675	0.842	1.037	1.282	1.646	1.962	2.056	2.330	2.581	2.813	3.098	3.300
∞	0.674	0.842	1.036	1.282	1.645	1.960	2.054	2.326	2.576	2.807	3.090	3.291