NYU STERN
NEW YORK UNIVERSITY LEONARD N. STERN SCHOOL OF BUSINESS

IOMS Department

Statistics and Data Analysis

Professor William Greene
Phone: 212.998.0876
Office: KMC 7-90
Home page: http://people.stern.nyu.edu/wgreene
Email: wgreene@stern.nyu.edu
Course web page: http://people.stern.nyu.edu/wgreene/Statistics/Outline.htm

Assignment 3

Note:
The data sets for this homework (and for the other problem sets for this course) are all stored on the home page for this course. You can find links to all of them on the course outline, at the bottom with the links to the problem sets themselves.

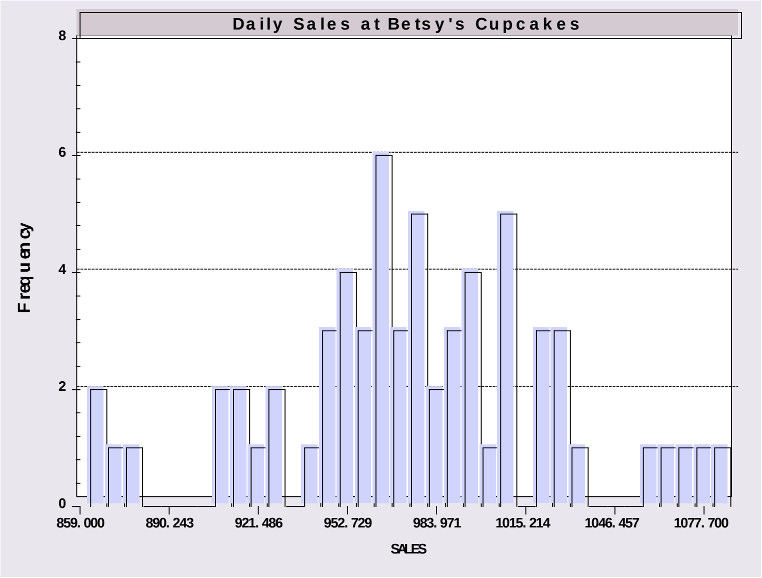
Part I. Cupcake Wars: Opening Salvo

Betsy has been selling cupcakes at her store at 30th and M Street for several years. The steady daily demand has been about 1000 cupcakes per day, with a standard deviation of 50. Now Allison has opened a competiting store 3 blocks away at Wisconsin and M Street and Betsy is concerned that her demand is being drawn away. It appears that her average daily demand is less than before, since Allison opened her store. If this is really true, Betsy has decided that she will have to lay off one of her 10 bakers (Julie). To find out, Betsy hires Peter, a consultant from Stern, to analyze the demand. Peter decides that it would be a good idea to study the sales pattern, so he watches the store for 64 days, and observes the following sales figures:

```
969  1023 986  939  1022 918  950  962  961  1000 925  859
967  907  1025 975  1019 992  989  990  907  1067 956  911
972  952  955  1000 968  868  973  971  937  1012 921  877
967  975  974  944  862  945  1014 999  942  961  1014 1002
986  988  1064 1001 1074 945  981  995  958  909  960  962
1052 1102 1055 948
```

Some summary statistics are shown below. What should Peter advise Betsy to do – let Julie go, or keep her in the kitchen? Show in detail how you reached your conclusion.

Assignment 3



```
Descriptive Statistics
--------+----------------------------------------------------------------
Variable|     Mean       Std.Dev.     Minimum      Maximum        Cases
--------+----------------------------------------------------------------
   SALES|  971.938     49.84989          859         1102           64
--------+----------------------------------------------------------------
```

The question is whether the mean is still at least 1,000 cupcakes per day. We can approach this as a hypothesis test:

H_0: $\mu \geq 1,000$
H_1: $\mu < 1,000$.

The rejection region is sample means far below 1,000. The sample mean is 971.938. The difference is 971.938 – 1,000 = -28.062. How far is this below 1,000. The standard error of the mean is s/sqr(N) = 49.84989/sqr(64) = 6.23. So, the sample mean of -28.062 is 4.5 standard errors below the mean. This is extremely far. The critical value for a one tail test at 95% significance would be, from the t table, with 63 degreess of freedom is -1.669. If you used the normal table, instead, it would be -1.645. Either way, -4.5 is far below the critical value. You should reject the null hypothesis. Unfortunately (for her), it looks like Betsy will decide to lay Julie off.

Assignment 3

Part II. Basic Interval Estimation and Hypothesis Tests

1. A supermarket chain computes for each of its stores the recovery ratio of redeemed deposit bottles divided by the number sold. For a sample of 15 of its stores, these values are the following (in percents):

82.4, 95.6, 90.4, 96.2, 88.4, 91.3, 102.4, 95.2, 96.1, 94.4, 91.7, 89.4, 93.8, 96.6, 101.2

The values over 100 mean that some stores redeemed more bottles than they sold. It is management's belief that the mean value of the population sampled should be 90, meaning 90%. At the 0.05 level of significance, test the hypothesis $H_0 : \mu = 90$ against the alternative $H_1 : \mu \neq 90$. State any assumptions that you make. By the way, $\bar{x} = 93.673$ and $s = 5.0259$ for these data.

Assume that the observed values come from a normal distribution. The sample is small, so we will want to use the t table to locate the critical value. The null hypothesis is $H_0:\mu = 90$. The sample mean is 93.673. The standard error of the mean is 5.0279/sqr(15) = 1.298. The distance of 93.673 from 90 is (93.673 – 90)/1.298 = 2.829. The critical value from the t table for a two sided test, with 14 degrees of freedom is 2.145. This is less than 2.829, so we will reject the null hypothesis. The P value would be the probability that $\bar{x}$ is at least 2.829 or -2.829 standard errors from 90, which is about 0.012. This is well under .05, leading to the same conclusion.

2. The data set MONET.mpj used at various points in the course will be used for this exercise. (a) Assuming that the data represent a sample from some well-defined population, provide a 95% confidence interval for the population mean of the surface area, Square Inches. This is column C17 in the data set, and it represents surface area.

Results for: Monet.mtp

Descriptive Statistics: Square Inches

```
Variable          N  N*   Mean  SE Mean  StDev  Minimum     Q1  Median      Q3  Maximum
Square Inches   430   0  940.8     32.6  676.5     26.1  637.2   816.6  1038.9   6193.7
```

The confidence interval would be 940.8 +/- 1.96(32.6) = [876.9 to 1004.69]

(b) For the variable Square Inches, test the null hypothesis $H_0 : \mu_{\text{Area}} = 900$ against the alternative $H_1 : \mu_{\text{Area}} \neq 900$ using the 0.05 level of significance.

The t statistic for the test would be (940.8 – 900)/32.6 = 1.25. The critical value would be 1.96, so we would not reject the null hypothesis.

3. Suppose that a sample of 200 accounts receivable entries at a large mail-order firm had a mean price of $846.20 and a standard deviation of $1,840.80. Give a 95% confidence interval for the population mean. State any assumptions that you use. (Note that the standard deviation is far larger than the mean. Given what you know about the Empirical Rule for sample data, what does this result suggest to you?)

The standard error of the mean will be 1,840.80/sqr(200) = 130.16.
The confidence interval will be 846.2 +/- 1.96(130.16) = [591.09 to 1101.31]

I use the 1.96 for the normal distribution based on the central limit theorem. The value from the t table for 199 degrees of freedom would be 1.96 anyway. The huge standard deviation compared to the mean suggests that the distribution is skewed. Theoretically, you can't say in which direction, except you know that accounts receivable would all be positive, so you can assert that the data are right skewed – there are some very large values in the sample.

4. On Tuesday, November 20, 2007, at 6:00 AM, CNN announced that Democratic presidential candidate Barack Obama had pulled ahead of candidate Hillary Clinton in the polling of Iowa voters. They announced that based on a sample of voters, 30% favored Obama and 26% favored Clinton.. They also cited a margin of error +/- 4%.
(a) Assume that they mean by the 4% that this is +/- two standard errors. Deduce (approximately) the sample size they used in the survey.

The MOE is approximately +/- 2 standard errors. The standard error is sqr[p(1-p)/N]. With 4%, the standard error is about 2%. The P is .3, so .02 is about sqr(.3(.7)/N) or .0004 is about .3(.7)/N. Solving for N, we get about .3(.7)/.0004 which is about 525.

(b) Form a 95% confidence interval for the true proportion of voters who favor Obama.

> Use .30 +/- the margin of error, since that is 2 standard errors
> .30 +/- .or, or .26 to .34

(c) Form a 95% confidence interval for the true proportion of voters who favor Clinton.

> Use the same logic as for Obama: The standard error would be sqr[.26(.74)/N]. We deduced N = 525, so the standard error is .019. The margin of error is 2 standard errors or .038, so the confidence interval is .26 +/- .038 or .222 to .298.

(d) Test the hypothesis that the two proportions are equal. (Note, in doing this test, you should assume that the two samples are independent. Strictly speaking, this cannot be true, since voters cannot (at least probably would not) choose more than one candidate. The assumption works for the purpose of this exercise, however.)

> Use sqr[PO(1-P)/N + PC(1-PC)/N] as the standard deviation of the difference. Sqr[.0004 + .00037] = .0277. The difference is only .3 - .26 = .04, which is onl .04/.0277 = 1.44 standard errors. We would not reject the hypothesis that the difference is 0.0.

Assignment 3

5 Of the 210 Sydney to Melbourne travelers sampled in the survey discussed in class, 152 chose a ground based mode of transport (train, bus or car). I divided those 152 individuals into low and high income families based on the median income. A cross tabulation of the mode of travel vs. the income level for these 152 travelers was as follows:

INCOME	TRAIN	BUS	CAR	Total
LOW	46	16	18	80
HIGH	17	14	41	72
Total	63	30	59	152

Do the data suggest that travel mode and income are related, or are they independent? Test the hypothesis that travel mode and income are independent.

Dividing the frequencies by 152 gives

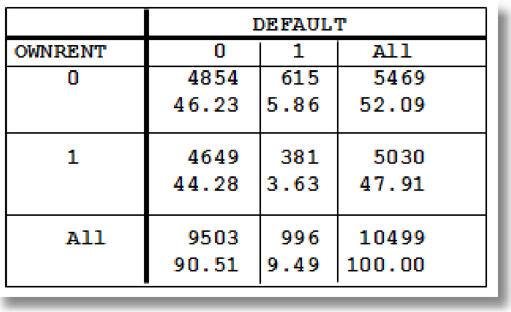
.303	.105	.118	(.526)
.112	.092	.270	(.474)
(.415)	(.197)	(.388)	(1.00)

Expected Frequencies are the products of the marginal

.415(.526) = .218   .197(.526) = .104   .388(.526) = .204
.415(.474) = .196   .197(.474) = .093   .388(.474) = .184

The chi-squared statistic is 152 times the sum of (Observed – Expected)^2/Expected.
Computing the for all 6 cells and then multiplying the sum by 152 gives 22.13.
The degrees of freedom for the test are (Rows – 1)(Columns – 1) = (2-1)(3-1) = 2.
The critical chi squared for 2 degrees of freedom is 5.99. The hypothesis of independence would be rejected.

6. Exercise:Are the default rates the same for owners and renters? The data for the 10,499 applicants who were accepted are in the table. Test the hypothesis that the two default rates are the same.



	DEFAULT		
OWNRENT	0	1	All
0	4854 46.23	615 5.86	5469 52.09
1	4649 44.28	381 3.63	5030 47.91
All	9503 90.51	996 9.49	10499 100.00

This is not the independence test. Rather, it is essentially the same as 4.d above.
The difference is Pr – Po = 615/5469 – 381/5030 = .1125 - .0758 = .0367.
The standard error of the difference would be sqr[Pr(1-Pr)/Nr + Po(1-Po)/No]
Using Pr and Po from above, and Nr = 5469 and No = 5030, the standard error is .00567.
The statistic for the test is .0367 / .00567 = 6.47. This is quite large. Much larger than 1.96.
So, we reject the hypothesis that the two proportions are the same.