Statistics - OR 155, Section 2, Final Examination
Tuesday, May 8, 2005

Name: ____________________________Key____________________________

Pledge: I have neither given nor received aid on this examination.

Signature: ____

Instructions: Do not do any actual numerical calculations. Answers in a form that you would type into an Excel field, such as “=28*SQRT(74)^2”, with a *working* answer, are expected).

1. [20] For each of the statistical settings, (a) – (d):

a. Neurological arguments suggest that piano lessons improve reasoning. Reasoning tests are given to 24 students both before and after piano lessons. iii, iv

b. A physician tests a drug for controlling shakiness in older people, by tossing a coin to randomly separate a set of patients into groups that get the drug, and that get a placebo instead (and the patients don’t know which they received). She then does a careful diagnosis of each patient, and compares results. iii, iv, v

c. To find the preferred treatment for breast cancer, between mastectomy and radiation, medical records of patients from 25 hospitals were searched for the survival times of a large number of patients of each type. ii

d. To investigate connections between cell phones and brain cancer, 358 brain cancer patients were each paired with a person who has no brain cancer, of the same sex, age, race and size. Both were asked about frequency of cell phone use. ii, iii

Attach of the following labels that apply:

i. Anecdotal evidence

ii. Observational study

iii. Designed experiment

iv. Controlled experiment

v. Blind experiment

vi. Double blind experiment

2. [30] A company makes 40% of its cars at factory A, 30% at Factory B, and the rest at Factory C. Factory A produces 25% lemons, Factory B produces 35% lemons, and Factory C produces 15% lemons. A car is chosen at random. What is the probability that:

a. It came from Factory A or from Factory B?

P{A or B} = P{A} + P{B} – P{A & B} = 0.4 + 0.3 – 0 = 0.7

b. It came from Factory C?

P{C} = 1 – P{not C} = 1 – P{A or B} = 1 – (0.4 + 0.3 – 0) = 1 – 0.7 = 0.3

c. It is a lemon if it came from Factory C?

P{L|C} = 0.15

d. It is a lemon from Factory C?

P{L & C} = P{L|C} P{C} = 0.15 * 0.3 = 0.045

e. It is a lemon?

P{L} = P{(L & A) or (L & B) or (L & C)} = P{L & A} + P{L & B} + P{L & C} =
= P{L|A} P{A} + P{L|B} P{B} + P{L|C} P{C} =
= 0.25 * 0.4 + 0.35 * 0.3 + 0.15 * 0.3

f. It came from Factory C, if it is a lemon?

P{C|L} = P{L & C} / P{L} =
= P{L|C} P{C} / (P{L|A} P{A} + P{L|B} P{B} + P{L|C} P{C}) =
= (0.15 * 0.3) / (0.25 * 0.4 + 0.35 * 0.3 + 0.15 * 0.4)

3. [25] A survey of 1000 student loan borrowers found that 300 had loans totaling $30k or more.

a. Why does it make sense to use the Normal distribution in this context?

n * p ≈ n * phat = 1000 * (300 / 1000) = 300 > 10
n * (1 – p) p ≈ n * (1 – phat) = 1000 * (1 - 300 / 1000) = 700 > 10

b. Give a 98% conservative Confidence Interval for the proportion of all student loans that are $30k or more.

Left: = (300 / 1000) - NORMINV(0.99,0,1)*SQRT(0.25/1000) =
= 0.3 - NORMINV(0.99,0,1)*0.5 / SQRT(1000)
= (300 / 1000) - CONFIDENCE(0.02,0.5,1000)

Right: : = (300 / 1000) + NORMINV(0.99,0,1)*SQRT(0.25/1000)
= 0.3 + NORMINV(0.99,0,1)*0.5 / SQRT(1000)
= (300 / 1000) +- CONFIDENCE(0.02,0.5,1000)

c. Will the 98% best guess Confidence Interval be longer or shorter than that in (b)?

Shorter, since conservative is longer.

$$n = \left(\frac{NORMINV(0.975,0,1)}{m} \right)^2 p(1-p)$$

d. Give the p-value for testing whether the overall population proportion is significantly different from 0.4.

H0: p = 0.4 HA: p ≠ 0.4
P{phat = 0.3 or m.c. | Bdry} = P{|phat – 0.4| > 0.1 | p = 0.4} =
= P{|Z| > 0.1 / sqrt(0.4 * 0.6 / 1000)} =
= 2 * NORMDIST(-0.1 / sqrt(0.4 * 0.6 / 1000),0,1,true)
= 2 * NORMDIST(-0.1,0, sqrt(0.4 * 0.6 / 1000),true)

e. What sample size will be needed to be 90 percent sure that the sample proportion will be within 0.01 of the true population proportion, in the best guess sense?

n = (NORMINV(0.95,0,1) / m)^2 * phat * (1 - phat) =
= (NORMINV(0.95,0,1) / 0.01)^2 * 0.3 * (1 – 0.3)

4. [25] In a telephone poll of 1425 randomly selected adults, 38% said that pro football was their favorite TV sport.

a. Give the 90% “best guess” margin of error for the true percent that say pro football is their favorite.

m = NORMINV(0.95,0,1) * SQRT(0.38 * (1 – 0.38) / 1425)
= CONFIDENCE(0.10,SQRT(0.38 * (1 – 0.38)),1425)

b. Give the 98% “conservative” Confidence Interval for the true percent that say pro football is their favorite.

Left: = 0.38 - NORMINV(0.99,0,1)*SQRT(0.25/1425) =
= 0.38 - NORMINV(0.99,0,1)*0.5 / SQRT(1425)
= 0.38 - CONFIDENCE(0.02,0.5,1425)

Right: = 0.38 + NORMINV(0.99,0,1)*SQRT(0.25/1425)
= 0.38 + NORMINV(0.99,0,1)*0.5 / SQRT(1425)
= 0.38 +- CONFIDENCE(0.02,0.5,1425)

c. Explain, in 10 words or less, why we cannot just say “38% of Americans say pro football is their favorite TV sport”.

Because this was based on a random subsample.

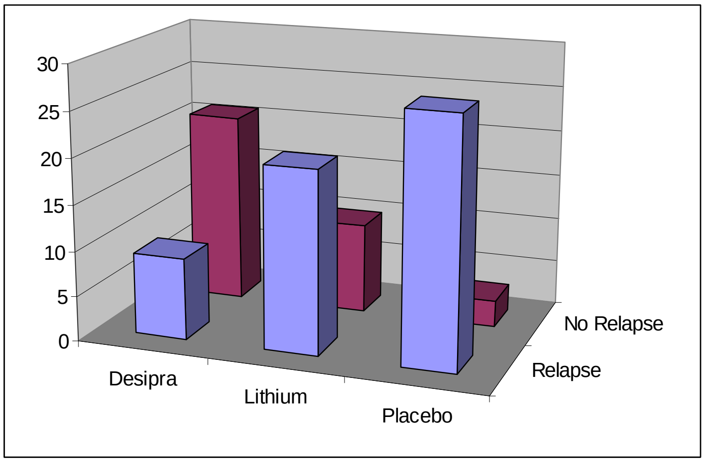
d. Explain, in 12 words or less, the interpretation of the 90% confidence interval in (a).

90% of repeated sampling intervals will cover true value.

e. Give a conservative estimate of the sample size required to give a sample proportion that is within 0.03 of the true proportion, with probability 0.99.

n = (NORMINV(0.995,0,1) / m)^2 / 4 =
= (NORMINV(0.995,0,1) / 0.03)^2 * 0.25

5. [25] Desipra and Lithium are two treatments for cocaine addiction. They were compared with each other, and with a Placebo for control purposes, by giving each to random samples of 30 addicts. The resulting counts of relapse versus non-relapse are shown here:



Answer the following True-False Questions:

a. T [✓ F] The Desipra treatment does not seem any better than the Placebo at preventing relapse.

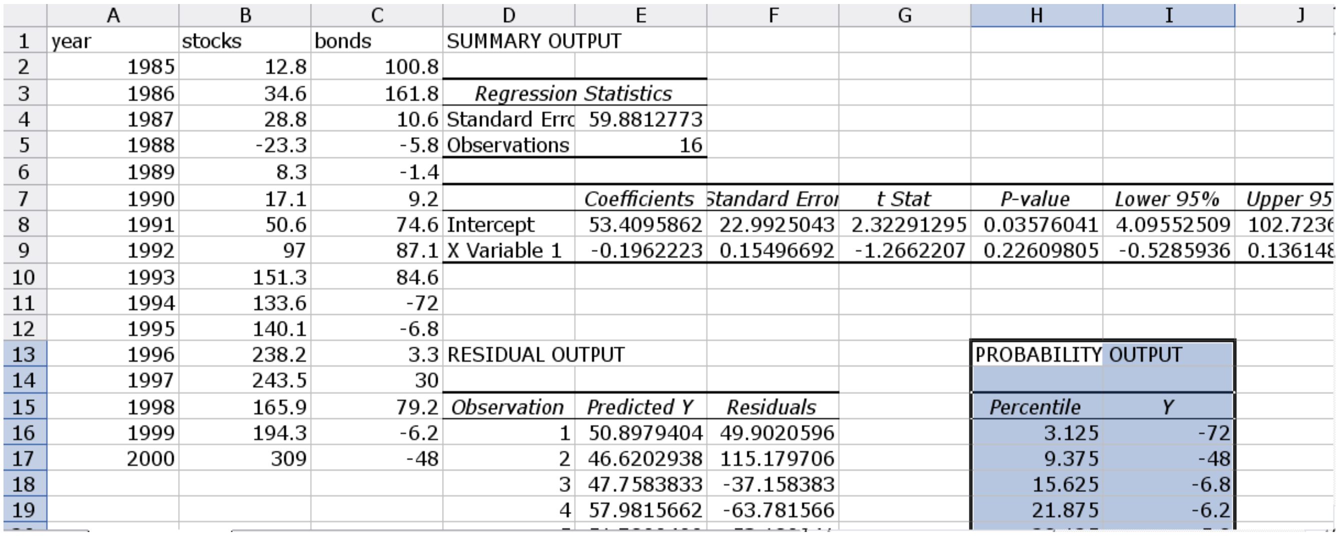
b. [✓ T] F The Lithium treatment appears to be somewhat better than the Placebo at preventing relapse.

c. T [✓ F] Overall there was less relapse, than non-relapse.

d. [✓ T] F Despira appears to be the best overall treatment at preventing relapse of cocaine addiction.

e. [✓ T] F The Chi Square Test of Independence of Relapse and Treatment type appears likely to reject.

6. [25] Here are data on money flow into stocks and bonds for 1985 – 2000 (in $billions). Also shown are part of the results of using the Excel Regression Tool, treating stocks as the explanatory variable. Use this to answer the following questions.



	A	B	C	D	E	F	G	H	I	J
1	year	stocks	bonds	SUMMARY OUTPUT						
2	1985	12.8	100.8							
3	1986	34.6	161.8	*Regression Statistics*						
4	1987	28.8	10.6	Standard Errc	59.8812773					
5	1988	-23.3	-5.8	Observations	16					
6	1989	8.3	-1.4							
7	1990	17.1	9.2		*Coefficients*	*Standard Erroi*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95*
8	1991	50.6	74.6	Intercept	53.4095862	22.9925043	2.32291295	0.03576041	4.09552509	102.7236
9	1992	97	87.1	X Variable 1	-0.1962223	0.15496692	-1.2662207	0.22609805	-0.5285936	0.136148
10	1993	151.3	84.6							
11	1994	133.6	-72							
12	1995	140.1	-6.8							
13	1996	238.2	3.3	RESIDUAL OUTPUT				PROBABILITY OUTPUT		
14	1997	243.5	30							
15	1998	165.9	79.2	*Observation*	*Predicted Y*	*Residuals*		*Percentile*	*Y*	
16	1999	194.3	-6.2	1	50.8979404	49.9020596		3.125	-72	
17	2000	309	-48	2	46.6202938	115.179706		9.375	-48	
18				3	47.7583833	-37.158383		15.625	-6.8	
19				4	57.9815662	-63.781566		21.875	-6.2	

a. Formulate hypotheses, in terms of population parameters, for whether or not there is a statistically significant relationship between stocks and bonds.

Let a = true slope, H0: a = 0, HA: a ≠ 0

b. Give a p-value to indicate the observed significance of the results in (a), and give a yes-no interpretation of the result.

= H9 = 0.226, Accept H0, no strong evidence of linear relationship.

c. Give the p-value for testing whether the Y-intercept of the regression line is significantly different from 0, and give a gray-level interpretation.

= H8 = 0.0358, Rather strong evidence.

d. Give a 95% Confidence Interval for the Y-intercept of the least squares fit line.

Left: = I8 = 4.10
Right: = J8 = 103

e. Why are the answers to (c) and (d) consistent with each other?

CI does not contain 0, which is equivalent to rejecting H0

f. Write the equation of the least squares fit line, using numbers from the table of Regression Tool results.

y = -0.196 x + 53.4

g. Explain how (f) can be answered using only Excel commands applied to the data.

-0.196 = SLOPE(C2:C17,B2:B17)

53.4 = INTERCEPT(C2:C17,B2:B17)

h. What is the predicted Bond flow for a new Stock flow of 80?

= -0.196 * 80 + 53.4 =
= FORECAST(80,C2:C17,B2:B17) =
= TREND(C2:C17,B2:B17,80,TRUE)

i. Give an Excel command to find a 98% confidence interval for the mean Bond flow for a new Stock flow of 80.

```
Left:   (-0.196 * 80 + 53.4) - TINV(1-0.98,16-2) * E4 *
                SQRT((1/E5) + (B87-AVERAGE(B2:B17))^2 / ((E5-1)*VAR(B2:B17)))

Right:  (-0.196 * 80 + 53.4) + TINV(1-0.98,16-2) * E4 *
                SQRT((1/E5) + (B87-AVERAGE(B2:B17))^2 / ((E5-1)*VAR(B2:B17)))
```

j. Give an Excel command to find a 90% prediction interval for the value of the Bond flow for a new Stock flow of 80.

```
Left:   (-0.196 * 80 + 53.4) - TINV(1-0.98,16-2) * E4 *
                SQRT(1 + (1/E5) + (B87-AVERAGE(B2:B17))^2 / ((E5-1)*VAR(B2:B17)))

Right:  (-0.196 * 80 + 53.4) + TINV(1-0.98,16-2) * E4 *
                SQRT(1 + (1/E5) + (B87-AVERAGE(B2:B17))^2 / ((E5-1)*VAR(B2:B17)))
```

7. [25] Human proofreaders tend to catch, on average, 80% of word errors in text. In an essay with 25 deliberately constructed errors, a student who catches errors at this typical rate, is asked to search for errors.

a. What is the distribution (with numerical values of the parameters) of the number of errors caught?

#Caught ~ Binomial(25, 0.8), #Missed ~ Binomial(25, 0.2)

b. What is the exact probability that the proofreader misses 8 or more out of the 25?

P{#Missed >= 8} = 1 – P{#Missed <= 7} = 1 – BINOMDIST(7,25,0.2,TRUE)

P{#Missed >= 8} = P{#Caught <= 17} = BINOMDIST(17,25,0.8,TRUE)

c. What is the mean number of errors missed?

E#Missed = 25 * 0.2 = 5

E#Missed = n – E#Caught = 25 - 25 * 0.8 = 25 * 0.2 = 5

d. What is the standard deviation of the number of errors missed?

sd(#Missed) = sd(n - #Caught) = sd(#Caught)= SQRT(25 * 0.8 * (1 – 0.8))

e. Give a normal approximation, with continuity correction, to the answer of part (b)

P{#Missed >= 8} = 1 – P{#Missed < 8} = 1 – P{#Missed < 7.5} ≈
≈ 1 – NORMDIST(7.5,25 * 0.2,SQRT(25 * 0.2 * (1 – 0.2))) =
= 1 – NORMDIST(7.5,5,2)
P{#Missed >= 8} = P{#Caught <= 17} = P{#Caught <= 17.5}
≈ NORMDIST(17.5,25 * 0.8,SQRT(25 * 0.2 * (1 – 0.2))) =
= NORMDIST(17.5,20,2)