Statistics 31, Section 3, Final Examination, Solution
Tuesday, December 16, 2000

Name: ____

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(82)^2”, with a *working* answer, are expected).
[points per part]

1. Government data show that 30% of the labor force have at least 4 years of college, and that 10% work as machine operators.

   a. Can it be concluded that, because 0.30 + 0.10 = 0.40, about 40% of the labor force either have 4 years of college, or are machine operators? Why or why not?

[5]
No, P{C or MO} = P{C} + P{MO} – P{C and MO}
But C and MO are not mutually exclusive

   b. Can it be concluded that, because (0.30)(0.10) = 0.03, about 3% of the labor force have 4 years of college and are also machine operators? Why or why not?

[5]
No, This would need independence of C and MO, but MOs tend to have C less often.

2. A researcher looking for evidence of extrasensory perception (ESP) tests 500 subjects. Four of these subjects do significantly better (p-val < 0.01) than random guessing.

   a. Is it proper to conclude that these 4 people have ESP? Why or why not?

[5]
No, Just by chance expect around 1 in 100 (i.e. 5 in 500) to get scores this high, so thiscould easily by explained by chance occurrence even if completely random.

   b. What should the researcher now do to test whether any of these four subjects have ESP?

[5]
Rerun a newly randomized version of the experiment only on these people.

3. The SSHA is a test that measures motivation and study habits. Scores range from 0 to 200. The mean score for U. S. college students is about 120 and the standard deviation is 20. A teacher suspects that older students have different motivation and study habits. To investigate this idea the test was given to 25 older students, and their mean score was 128, with a standard deviation of 17.

a. Assume that the standard deviation of the population of older students is the same as the rest of the population. Give an Excel formula to calculate a p-value to assess the strength of the evidence in favor of the teacher's idea

[10]

H0: mu = 120   Ha: mu not = 120

p-val = P{Xbar = 128 or m.c. | B'dry} = P{|Xbar – 120| > 8 | mu = 120}

= 2*P{Xbar – 120 < -8} =2*NORMDIST(-8,0,20/SQRT(25),TRUE)

=2*(1-NORMDIST(8,0,20/SQRT(25),TRUE))

=2*NORMDIST(-8,0,4,TRUE)

=2*NORMDIST(-2,0,1,TRUE)

b. If the numerical answer to (a) is 0.0673, assess the p-value using the gray level point of view.

[5]

Some evidence, but not particularly strong.

c. Repeat (a), but assume that the standard deviation of the population of older students is likely to be completely different from the rest of the population.

[10]

this time have unknown standard deviation, so use T distribution, and SAMPLE standard deviation

=TDIST(8/(17/SQRT(25)),24,2)

d. If the numerical answer to (c) is 0.0418, assess the p-value using the yes-no point of view, with $\alpha = 0.01$.

[5]

0.0418 > 0.01, so "no strong evidence"

e. Using the assumption of (c) give an Excel formula for the 80% margin of error in estimating the mean of the score of older students.

[10]

=TINV(0.2,24)*17/SQRT(25)

4. In a manufacturing process, the random flame temperature X , varies according to this distribution, in degrees Celsius:

	A	B	C	D
18	Temperature:	520	525	530
19	Probability:	0.2	0.5	0.3

a. Find $P\{X \geq 525\}$

[5]
=C19+D19 = 0.8

b. Find $P\{X = 520 \mid X \leq 525\}$

[5]
=B19/(B19+C19) = 0.2 / 0.7

c. Write an Excel formula to calculate the expected value μ_X .

[5]
=B18*B19+C18*C19+D18*D19

d. Write an Excel formula to calculate the standard deviation, assuming the answer to (c) is 526.

[5]
=SQRT(B19*(B18-526)^2+C19*(C18-526)^2+D19*(D18-526)^2)

e. A manager wants these results in degrees Fahrenheit. The conversion formula is $Y = \frac{9}{5}X + 32$ What is the Fahrenheit mean and standard deviations, μ_Y and σ_Y ? Suppose that the answer to (c) is 526, and to (d) is 3.

[5]
mean: =(9/5)*526+32
s.d.: =(9/5)*3

5. A pollster asked a SRS of 2000 adults if they ate broccoli in the last month. 600 of them answered "Yes".

a. Give Excel formulas, and explain how to use them, to check that it is OK to use the Normal approximation.

[5]
=2000*(600/2000)
=2000*(1-600/2000)
check that these are bigger than 10

b. Give Excel formulas for the endpoints of the 99% best guess confidence intervals for the proportion of broccoli eaters in the whole population.

[10]
=600/2000-CONFIDENCE(0.01,SQRT((600/2000)*(1-600/2000)),2000)
=600/2000+CONFIDENCE(0.01,SQRT((600/2000)*(1-600/2000)),2000)

c. If the numerical answer to (b) is (0.27,0.31) then is there strong evidence that at least ¼ of the population are broccoli eaters? Why or why not?

[5]
Yes, 0.25 is outside 99% CI, so could reject 2 sided hypothesis test, even at level 0.01.

d. Give an Excel formula to give a conservative calculation of how large a sample size is required, to obtain a margin of error of ± 0.01 in a 95% Confidence Interval for the population proportion of broccoli eaters?

[10]
=(NORMINV(0.975,0,1)/0.01)^2*0.25

e. For each of the following variations on the design specifications, state whether the desired sample size will be higher, lower or the same: [5]

i. Use a 90% Confidence Interval. lower
ii. Change the allowable margin of error to 0.001. higher
iii. Use a planning value of $p = 0.5$. same
iv. Use a planning value of $p = 0.25$. lower
v. Do the same study on cauliflower. same

6. Match each statistical setting, (a) – (d) below, to all that apply, among:
[15]

- i. Anecdotal evidence
- ii. Observational study c
- iii. Designed experiment a, b
- iv. Controlled experiment b
- v. Blind experiment b
- vi. Double blind experiment

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

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. 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

7. An examination consists of multiple choice problems, having 4 possible answers. Linda estimates that she has conditional probability of 0.8 of knowing the answer to any question that may be asked. If she does not know the answer, she will guess, with conditional probability of ¼ of being correct.

a. What is the probability that Linda gives the correct answer to a question?
[10]

P{C} = P{(C and K) or (C and not K)} = P{C and K} + P{C and not K} - 0
= P{C | K} P{K} + P{C | not K} P{not K}
= 1 * 0.8 + 0.25 * (1 – 0.8) = 0.8 + 0.05 = 0.85

b. What is the conditional probability that Linda knows the answer, given that she supplies the correct answer?
[10]

P{K | C} = P{C and K} / P{C}
= 1 * 0.8 /(1 * 0.8 + 0.25 * (1 – 0.8)) = 0.8 / 0.85

8. Data was gathered to study the effect of precipitation on soil pH. The Excel regression tool was used to analyze the data, and the tabular output included:

	Coefficients	Standard Err	t Stat	P-value	Lower 95%	Upper 95%	Lower 99.0%	Upper 99.0%
Intercept	4.484252	0.832453	5.386795	0.012533	1.835014	7.133491	-0.37798	9.346482
X Variable	0.155514	0.127882	1.216073	0.310914	-0.25146	0.562492	-0.59143	0.902454

a. Write the equation of the least squares fit line.

[5]

y = 0.156 x + 4.48

b. Give the standard error of the slope of the least squares fit line.

[5]

0.128

c. Give the p-value for testing the hypothesis that the y-intercept is different from 0. Use a gray level interpretation, and relate the conclusion to precipitation and soil pH.

[5]

p-value = 0.0125, which is strong evidence that the Y-intercept is different from 0. Thus, when when there is no precipitation, the pH is significantly different from 0.

d. What are the endpoints of the 99% Confidence Interval for the slope?

[5]

-0.591, 0.902

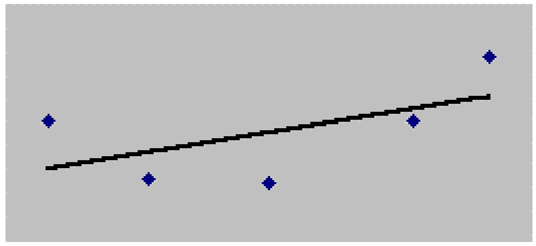
e. Is there strong evidence (yes-no interpretation) that the slope is different from 0? And what does his suggest about causation of soil pH by precipitation?

[5]

No, p-value for slope is very large, no strong evidence

Suggests no causation (at least under our assumption that the regression is linear)

Here is a scatterplot of the data, together with the least squares fit line:



f. Which of the following is most likely to be the sample correlation $\hat{\rho}$:

[5]

i. -0.27 slope not negative

ii. 0.03 this is "nearly independent"

iii. 0.57 X, only one with moderate positive correlation

iv. 0.98 points don't lie this close to a line

g. Does the scatterplot suggest that pH depends strongly on precipitation. Why or why not?

[5]

Yes, points are close to lying on a parabola, which is a very strong type of dependence.

h. The scatterplot suggests that which of the assumptions of linear regression are violated?

[5]

Y values do not seem to lie on a line, plus some random error. Instead lie on a parabola.