

Sociology 592 - Research Statistics I
Exam 1 Answer Key - DRAFT
September 24, 2004

Where appropriate, show your work - partial credit may be given. (On the other hand, don't waste a lot of time on excess verbiage.) Do not spend too much time on any one problem. It is legitimate (and probably essential) to refer to results that have previously been proven in class or homework, without re-proving them - for example, you wouldn't need to prove that $P(-1.96 \leq Z \leq 1.96) = .95$, since we have already shown that in class. Likewise, you are free to refer to anything that was demonstrated in the homework or handouts.

1. (4 points each, 20 points total). Indicate whether the following statements are true or false. If you think the statement is false, indicate how the statement could be corrected. For false statements, do not just say that you could substitute not equals for equals. For example, the statement $P(Z \leq 0) = .7$ is false. To make it correct, don't just say $P(Z \leq 0) < .7$, instead say $P(Z \leq 0) = .5$ or $P(Z \leq .525) = .7$.

A. If X has a binomial distribution, then if N is large \bar{X} will also have a binomial distribution.

False. The central limit theorem tells us that the distribution of \bar{X} will be approximately normal if N is large.

B. $P(Z \geq 1.96) = .05$

False. Substitute $P(Z \geq 1.96) = .025$ or $P(Z \geq 1.65) = .05$

C. When two events are mutually exclusive, the occurrence of one event in no way affects the occurrence of the other.

False. If events are mutually exclusive, the occurrence of one makes the occurrence of the other impossible. Substitute "independent" for "mutually exclusive."

D. $V(8 + X) = 64 + V(X)$

False. $V(8 + X) = V(X)$, i.e. adding a constant does not change the variance.

E. When N is small, the T distribution and the $N(0, 1)$ distribution are almost identical.

False. It is when N is large that the 2 distributions become virtually identical; when N is small they can be quite a bit different.

2. (10 points each, 30 points total) Answer three of the following. The answers to most of these are fairly straightforward, so do not spend a great deal of time on any one problem. NOTE: I will give up to 5 points extra credit for each additional problem you do correctly.

A. $\bar{X} = 30, N = 25$. Determine the 95% confidence interval when

a. $\hat{\sigma} = 15$

This falls under confidence intervals, case III, σ unknown. For $V = 24$ and $2Q = .05$, the critical value of t is 2.064 (see Appendix E, Table 3).

$$\bar{x} \pm (t_{\alpha/2, v} * s / \sqrt{N}), i.e.,$$

$$30 - (2.064 * 15 / \sqrt{25}) \leq \mu \leq 30 + (2.064 * 15 / \sqrt{25}), i.e.,$$

$$23.808 \leq \mu \leq 36.192$$

Using Stata to double-check,

```
. ci 25 30 15
```

Variable	Obs	Mean	Std. Err.	[95% Conf. Interval]
	25	30	3	23.8083 36.1917

b. $\sigma = 10$

This falls under confidence intervals, case I, σ known. The critical value for Z is 1.96.

$$\bar{x} \pm (z_{\alpha/2} * \sigma / \sqrt{N}), i.e.,$$

$$30 - (1.96 * 10 / \sqrt{25}) \leq \mu \leq 30 + (1.96 * 10 / \sqrt{25}), i.e.,$$

$$26.08 \leq \mu \leq 33.92$$

Using Stata to double-check,

```
. ztesti 30 10 0 25, level(95)
```

Number of obs = 25

Variable	Mean	Std. Err.	z	P > z	[95% Conf. Interval]
x	30	2	15	0.0000	26.08007 33.91993

B. Here is a random sample of 7 scores from a previous cohort's first exam in statistics. Compute the sample mean and the sample standard deviation.

Score

- 76
- 86
- 90
- 91
- 93
- 99
- 105

Expand the table as follows:

	Score	(Score – E[Score]) ²
	76	238.04
	86	29.47
	90	2.04
	91	.18
	93	2.47
	99	57.33
	105	184.18
Sum	640	513.71

$$\bar{X} = 640/7 = 91.43$$

$$s^2 = 513.71/6 = 85.62, \quad s = 9.25$$

Remember, this is a sample, not the population, so you use N-1 instead of N when computing the variance.

Using Stata to double-check,

```
. list score, sep(7)
```

```

+-----+
|      score      |
+-----+
1. |      76      |
2. |      86      |
3. |      90      |
4. |      91      |
5. |      93      |
6. |      99      |
7. |     105      |
+-----+

```

```
. sum score
```

```

-----+-----+-----+-----+-----+-----+
Variable |      Obs      |      Mean      |      Std. Dev.      |      Min      |      Max      |
-----+-----+-----+-----+-----+-----+
score    |      7      |    91.42857    |    9.253056         |      76      |     105      |

```

C. A school district is concerned about the practice of “social promotion,” i.e. the practice of moving students on to the next grade whether they are ready or not. It therefore requires that all 8th graders take a standardized test that measures how much they have learned in grade school. All students scoring in the bottom 10% of the test will be held back and made to take 8th grade again. If test scores $\sim N(90, 10)$, how high does your score have to be to avoid getting held back a year?

The critical value for Z is -1.28. So, $x = z\sigma + \mu = -1.28 * \sqrt{10} + 90 = 85.95$

Using Stata to double-check,

```
. display invnorm(.10) * sqrt(10) + 90
```

```
85.947378
```

D. It is election day, 2004. Once again, the American Presidential race is too close to call – and, just like in 2000, it appears that the state of Florida could play a critical role in determining the next President.

John Kerry estimates that if he wins Florida, there is an 80% chance that he will win the Presidency. But, if he loses Florida, he has only a 24% chance of becoming president. Finally, he estimates that he and George Bush both have a 50% chance of carrying the state.

What is the probability that John Kerry will be elected President? What is the probability that he will lose both the State of Florida and the Presidency?

We are told $P(\text{Wins Presidency} \mid \text{Wins Florida}) = .80$

$P(\text{Wins Presidency} \mid \text{Loses Florida}) = .24$, implying $P(\text{Loses Presidency} \mid \text{Loses Florida}) = .76$

$P(\text{Wins Florida}) = .50$. Ergo,

$P(\text{Wins Presidency}) =$

$P(\text{Wins Florida}) * P(\text{Wins Presidency} \mid \text{Wins Florida}) +$

$P(\text{Loses Florida}) * P(\text{Wins Presidency} \mid \text{Loses Florida}) =$

$(.5 * .80) + (.5 * .24) = .52$

$P(\text{Loses Presidency and Loses Florida}) =$

$P(\text{Loses Florida}) * P(\text{Loses Presidency} \mid \text{Loses Florida}) =$

$.50 * .76 = .38$

E. The population variance of x is known to be 4. Stata's `ztesti` command produces the following output:

Variable	Mean	Std. Err.	z	P> z	[95% Conf. Interval]
x	5	1	5	0.0000	3.040036 6.959964

What was the sample size? How big would the sample need to be if you wanted the true standard error of the mean to be only 0.1?

Note that the true standard error of the mean is reported and equals 1. Further, true standard error of the mean $= 1 = \frac{\sigma}{\sqrt{N}} = \frac{2}{\sqrt{N}}$, so $N = 4$. To have the True s.e.

equal .1, N would have to be 400 (you don't even need the printout to compute this, all you need to know is that $\sigma = 2$).

It is more tedious, but you could also work this by solving for N in the confidence interval:

$6.96 = 5 + 1.96 * \frac{\sigma}{\sqrt{N}} = 5 + 1.96 * \frac{2}{\sqrt{N}}$	Formula for upper bound of c.i.
$1.96 = 1.96 * \frac{2}{\sqrt{N}}$	Subtract 5 from both sides

$1 = \frac{2}{\sqrt{N}}$	Divide both sides by 1.96
$1 = \frac{\sqrt{N}}{2}$	Take reciprocals
$1 = \frac{N}{4}$	Square both sides
$4 = N$	Multiply both sides by 4

3. (25 points) A large company is being forced to lay off several thousand of its employees. It has agreed to hire an outside firm that will help its ex-employees find new jobs. Two employment placement services are being considered for the task. Each has provided information on the last 1,000 individuals who used their services. A review of the two companies reveals that

- Company A placed 840 of its clients in new jobs in three months or less. Company B, on the other hand, placed only 600 of its clients that quickly.
- 70% of Company A's clients previously had jobs in management, while the rest were manual laborers. For Company B, only 40% of its clients had been in management.
- For company A, 630 of those placed within 3 months were previously in management. For Company B, 300 of those placed within 3 months were previously managers.

a. (10 pts) Complete the following table

	Company A			Company B		
	Placed within 3 months	Not Placed within 3 months	Σ	Placed within 3 months	Not Placed within 3 months	Σ
Management						
Labor						
Σ			1000			1000

The numbers in bold are based on the information given, and the rest of the numbers can easily be determined from there.

	Company A			Company B		
	Placed within 3 months	Not Placed within 3 months	Σ	Placed within 3 months	Not Placed within 3 months	Σ
Management	630	70	700	300	100	400
Labor	210	90	300	300	300	600
Σ	840	160	1000	600	400	1000

b. (10 pts) Company A was much more successful at finding jobs for its clients within 3 months. However, Company A also had many more clients who were previously in management, and those who were in management were much more likely to get new jobs quickly. Suppose that Company B had had Company A's clients (i.e. it had 700 managers and 300 laborers). How many of those individuals would have been placed in new jobs within 3 months, assuming that Company B's placement rates for managers and for laborers each stayed the same?

Company B placed 75% (300 of 400) of its managers within 3 months. It placed 50% of its Laborers (300 of 600) within 3 months. Therefore, if it had 700 managers and 300 laborers, it would place $(700 * .75) + (300 * .50) = 675$ individuals within 3 months. i.e. it would place 75 more individuals than it currently does if it had the same number of managers and laborers as Company A does.

More formally,

For Company A, $P(\text{Management}) = .7$, $P(\text{Labor}) = .3$

For Company B, $P(\text{Placement}|\text{Management}) = .75$, $P(\text{Placement}|\text{Labor}) = .50$

$P(\text{Placement}) =$

$$P(\text{Management})^A * P(\text{Placement} | \text{Management})^B + P(\text{Labor})^A * P(\text{Placement} | \text{Labor})^B =$$

$$(.70 * .75) + (.30 * .50) = .675$$

c. (5 pts) Based on these results, would you recommend that Company A or Company B be hired to provide placement services? Explain your reasoning.

I would go with Company A. There is a 240 person gap in the number of people placed, and only 75 of that (less than a third) is due to differences in the types of clients. Further, we see that Company A placed 90% of its managers (630 of 700) within 3 months compared to Company B's 75% (300 of 400). Further, Company A placed 70% of its laborers within 3 months (210 of 300) compared to only 50% for company B (300 of 600). So, for both managers and laborers, Company A has a better placement rate.

4. (25 points) Supporters of Republican Congressman Chris Chocola claim that at least 60% of the voters in the Congressional District support his re-election. Skeptical Democrats immediately dismiss these claims. A random sample of 200

registered voters is drawn, 107 of whom say they will vote for Chocola. Using the .05 level of significance, test whether the Chocola supporters' claim is supported. Be sure to indicate:

- (a) The null and alternative hypotheses - and whether a one-tailed or two-tailed test is called for.

$$H_0: X = 120$$

$$H_A: X < 120$$

The alternative is one-tailed because critics claim that Chocola's support is less than 60%.

- (b) The appropriate test statistic

The appropriate test statistic is

$$z = \frac{\# \text{ of Chocola Supporters} \pm CC - Np_0}{\sqrt{Np_0q_0}} = \frac{x \pm CC - (200 * .6)}{\sqrt{200 * .4 * .6}} = \frac{x \pm CC - 120}{\sqrt{48}}$$

For the correction for continuity, we will subtract .5 if there are more than 120 supporters, and we will add .5 if there are less than 120 supporters.

- (c) The critical region

For the critical region, we will reject H_0 if $Z_c < -1.645$

- (d) The computed value of the test statistic

$$z = \frac{\# \text{ of Chocola Supporters} \pm CC - Np_0}{\sqrt{Np_0q_0}} = \frac{x \pm CC - 120}{\sqrt{48}} = \frac{107 + .5 - 120}{\sqrt{48}} = \frac{-12.5}{\sqrt{48}} = -1.80$$

- (e) Your decision - should the null hypothesis be rejected or not be rejected? Why?

Reject. The test statistic falls in the critical region. Using the .05 level of significance, Chocola's support is significantly below .6. There is only about a 3.6% chance ($F[-1.8] = .036$) that Chocola could have the support of 60% of the vote and get this few supporters in the sample.

- (f) Would your decision change if you used the .01 level of significance instead? Why or why not?

Do not reject. For the .01 level of significance, you reject if $Z_c < -2.33$. The computed test statistic of -1.80 is greater than that. Ergo, if we use a more stringent standard, we would not conclude that Chocola's support is below .6. Either way, the Democrats can't be too happy, because the poll still shows that Chocola is supported by a majority of the voters.

Double-checking with Stata, we can use the `bintesti` command to do the normal approximation to the binomial with correction for continuity:

```
. bintesti 200 107.5 .60, normal

Variable |      Obs  Proportion  Std. Error
-----+-----
       x |      200      .5375    .0352558

      Ho:      p =
            z = -1.80

. display norm(-1.80)
.03593032
```

Hence, there is about a 3.6% chance Chocola could have the support of 60% of the voters and have this few supporters show up in the sample. This is less than 5% but more than 1%. So, reject the null if using the .05 level of significance, but do not reject if using the .01 level.

Or, using the more exact `bitesti` command,

```
. bitesti 200 107 .60

      N   Observed k   Expected k   Assumed p   Observed p
-----+-----
      200         107         120         0.60000         0.53500

Pr(k >= 107)           = 0.973665 (one-sided test)
Pr(k <= 107)           = 0.036306 (one-sided test)
Pr(k <= 107 or k >= 133) = 0.070900 (two-sided test)
```

Again, there is about a 3.6% chance Chocola could have the support of 60% of the voters and have this few supporters show up in the sample. This is less than 5% but more than 1%. So, reject the null if using the .05 level of significance, but do not reject if using the .01 level.