Sociology 592 - Research Statistics I
Exam 1 ANSWER KEY
September 27, 2002

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 $P(-1.96 \le Z \le 1.96) = .95$, 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. (5 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 $P(Z \le 0) = .7$ is false. To make it correct, don't just say $P(Z \le 0) <> .7$, instead say $P(Z \le 0) = .5$ or $P(Z \le .525) = .7$.

A. V(5X) = 5 * V(X)

False. We know $V(aX) = a^2V(X)$, so V(5X) = 25 * V(X)

B. $P(Z \ge .84) = .79954586$.

False. $P(Z \le .84) = .79954586$ or else $P(Z \ge .84) = 1 - .7995486 = .2004514$.

C. If $E(X^2) = 30$ and $E(X) = -5$, then $V(X) = 55$.

False. $E(X^2) – E(X)^2 = 30 – (-5)^2 = 30 – 25 = 5$

D. If a fair coin is tossed 5 times, there is a 50% of getting exactly 3 heads.

False. There is a 50% chance of getting 3 heads or more. Or, P(3 heads) =

$$\binom{N}{r} p^r q^{N-r} = \binom{5}{3} .5^3 .5^2 = \frac{5!}{3!2!} .03125 = 10 * .03125 = .3125$$

You can also look at Table 2 in Appx E, N = 5, p = .5, r = 3.

2. (10 points each, 30 points total) Answer <u>three</u> 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. $\overline{X} = 7$, N = 14. Determine the 95% confidence interval when

a. $\hat{\sigma} = 20$

For s = 20, look at Appx E, Table 3, v = 13, 2Q = .05:

$$\bar{x} \pm (t_{\alpha/2,v} * s/\sqrt{N}), i.e.$$
$$7 - (2.16 * 20/\sqrt{14}) \le \mu \le 7 + (2.16 * 20/\sqrt{14}), i.e.$$
$$-4.546 \le \mu \le 18.546$$

Stata double-check:

```
. cii 14 7 20

    Variable |        Obs        Mean    Std. Err.       [95% Conf. Interval]
-------------+---------------------------------------------------------------
             |         14           7    5.345225       -4.547656    18.54766
```

b. σ = 15

For σ = 15:

$$\bar{x} \pm (z_{\alpha/2} * \sigma/\sqrt{N}), \text{ i.e.,}$$
$$7 - (1.96 * 15/\sqrt{14}) \le \mu \le 7 + (1.96 * 15/\sqrt{14}), \text{ i.e.,}$$
$$-.857 \le \mu \le 14.857$$

Stata double-check:

```
. ztesti 7 15 0 14, level(95)

                                                      Number of obs =       14

------------------------------------------------------------------------------
Variable |      Mean    Std. Err.       z     P>|z|       [95% Conf. Interval]
---------+--------------------------------------------------------------------
       x |         7    4.008919   1.74611   0.0808      -.8573361    14.85734
------------------------------------------------------------------------------
```

B. Here are the results from a previous cohort's first exam in statistics. Compute the mean and variance of the scores. There were 10 Sociology Students in the class.

Score
74
83
84
92
93
95
97
98
104
106

Let’s expand the table as follows:

	Score	$(X_i - \mu)^2$	X_i^2
	74	345.96	5476
	83	92.16	6889
	84	73.96	7056
	92	0.36	8464
	93	0.16	8649
	95	5.76	9025
	97	19.36	9409
	98	29.16	9604
	104	129.96	10816
	106	179.56	11236
Σ	926	876.4	86624

So, μ = 926/10 = 92.6. σ^2 = 876.4/10 = 87.64, σ = 9.36.
Or, if you prefer, $\sigma^2 = E(X^2) - E(X)^2 = 86624/10 - 92.6^2 = 87.64$, σ = 9.36.

C. A company has been deluged with dozens of job applications. Only those candidates scoring in the top third on an aptitude test will be hired. If Scores ~ $N(80, 10^2)$, how high does your score have to be for you to get hired?

Z = .43, since F(.43) = .6664. So, the corresponding X score is $X = Z\sigma + \mu$ = .43*10 + 80 = 84.3.

Confirming with Stata,

```
. display invnorm(2/3) * 10 + 80

84.307273
```

D. It is October 26, 2002. The undefeated Notre Dame football team has continued its amazing success story, winning each of its last three games with 50 yard field goals in the closing seconds. Today, however, it faces its toughest challenge of the season: Florida State University.

The Irish coaches estimate that, if they can defeat Florida State, there is a 90% chance that Notre Dame will get to play in the Bowl Championship Series (BCS). But, if they lose, ND’s chances of being in the BCS drop to 50%. They further estimate that there is a 15% chance that they will lose to Florida State and also play in the BCS.

What is the probability that Notre Dame will beat Florida State? What is the probability that Notre Dame will play in the BCS?

We are told
P(BCS | Beat FSU) = .90
P(BCS | Lose to FSU) = .50
P(BCS ∩ Lose to FSU) = .15.

This implies P(Lose to FSU) = .30, since
P(BCS ∩ Lose to FSU) = P(Lose to FSU) * P(BCS | Lose to FSU) =
P(Lose to FSU) *.50 = .15.

So, P(Beat FSU) = $\boxed{.70}$ (Complements rule).

P(BCS) = P(Losing) * P(BCS|Lose to FSU) + P(Winning) * P(BCS|Beat FSU) =
(.30 * .50) + (.70 * .90) = .$\boxed{78}$

E. A polling firm reports that President Bush's foreign policy is supported by 70% of the American public. The firm further reports that the (approximate) 99% confidence interval for Bush's approval is $.582 \le p \le .818$. What was the sample size used in the study?

We know the formula for the lower and upper bounds of the approximate c.i. We also know that the upper c.i. was .818, the sample p was .7, and the critical value for Z is 2.58. So, all we have to do is algebraically solve for N.

$.818 = .7 + 2.58 * \sqrt{\frac{\hat{p}\hat{q}}{N}} = .7 + 2.58\sqrt{\frac{.21}{N}}$	Formula for upper bound of c.i.
$.118 = 2.58\sqrt{\frac{.21}{N}}$	Subtract .7 from both sides
$.045736 = \sqrt{\frac{.21}{N}}$	Divide both sides by 2.58
$21.8644 = \sqrt{\frac{N}{.21}}$	Take reciprocals
$478.05 = \frac{N}{.21}$	Square both sides
$100 = N$	Multiply both sides by .21

3. (25 points) A recent study done in Finland suggests that stressful events have a bigger impact on men's health than they do on women's. According to the synopsis of the report printed on the WebMD web pages (http://content.health.msn.com/content/article/1685.53485)

"The study found men who suffered a stressful life event were more likely than women to miss work due to illness in the following months… Interpersonal problems, financial difficulties, and violence among men were linked to psychological problems, such as anxiety, mental distress, and lack of coherence. Financial difficulties and violence were also associated with heightened use of cigarettes and alcohol, which was thought to lead to sick days. For women, none of these events increased the likelihood of a sick leave."

A researcher in the United States has decided to replicate this study. Data are collected from 1000 men and 1000 women who had experienced major stressful life events in the past six months. Each respondent is coded as either having missed work because of illness or not having missed work. In addition, the "support networks" (number of people the respondent talks to about their problems) is measured. Support networks are classified as "large" or "small". She finds that 60% of men, but only 40% of women, became ill and missed work after experiencing stressful events. She also finds that 70% of men have small support networks, while 60% of all women have large support networks. Finally, she found that 200 men and 200 women had small support networks and did not miss work.

a. (10 pts) Complete the following table, Remember, there were 1000 men and 1000 women in this study.

	Male			Female		
Income/GPA	Small support network	Large support network	Σ	Small support network	Large support network	Σ
Became ill and missed work						
Did not miss work						
Σ			1000			1000

Items in italics can be inferred from the information given, and the rest of the table can easily be deduced from there:

	Male			Female		
Income/GPA	Small support network	Large support network	Σ	Small support network	Large support network	Σ
Became ill and missed work	500	100	*600*	200	200	*400*
Did not miss work	*200*	200	400	*200*	400	600
Σ	700	300	1000	400	600	1000

b. (5 pts) What percentage of those who missed work had small support networks?

1,000 missed work (600 men and 400 women), and 700 of those (500 men, 200 women, a total of 70%) had small support networks.

c. (10 pts) As these figures show, men tend to have smaller support networks than do women, and those with smaller support networks are more likely to get ill after stressful events and miss work. Suppose that men had the same support network distribution as women did, i.e. 60% of men had large support networks. Suppose further that it continued to be the case that men maintained their network-specific illness rates. What percentage of men would then become ill and miss work after stressful events? Based on these results, do you think differences in networks explain much of the difference in stress-related illness between men and women, or does it explain relatively little?

For women, P(Small) = .4, P(Large) = .6.
For men, P(ILL|Small) = 5/7, P(ILL|Large) = 1/3.

So, with female composition and male illness rates,

$P(ILL) = P(Small)^W * P(ILL|Small)^H + P(Large)^W * P(ILL|Large)^M =$
$(.4 * 5/7) + (.6 * 1/3) = .486$

About 486 men would still miss work, compared to the original 600. Support networks account for more than half (57%) of the male/female differences, but differences still remain.

4. (25 points) A stock broker company claims that, despite recent rough economic conditions, only 30% of its customers lost money last year. Angry investors suspect otherwise. A random sample of 70 investors is taken, and it is discovered that 30 lost money last year. Test the company's claim at the .01 level of significance. Be sure to indicate:

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

H_0: p = .30 or else E(X) = 21

H_A: p > .30 or else E(X) > 21

(b) The appropriate test statistic

The appropriate test statistic is

$$z = \frac{\#\ of\ losers \pm CC - Np_0}{\sqrt{Np_0 q_0}} = \frac{x \pm CC - (70 * .3)}{\sqrt{70 * .3 * .7}} = \frac{x \pm CC - 21}{3.834}$$

For the correction for continuity, we will subtract .5 if there are more than 21 losers, we will add .5 if there are less than 21 losers.

(c) The critical region

For the critical region, we will reject H_0 if $Z_c > 2.33$

(d) The computed value of the test statistic

$$z = \frac{\#\ of\ losers \pm CC - Np_0}{\sqrt{Np_0 q_0}} = \frac{x \pm CC - (70 * .3)}{\sqrt{70 * .3 * .7}} = \frac{x \pm CC - 21}{3.834} = \frac{30 - .5 - 21}{3.834} = \frac{8.5}{3.834} = 2.217$$

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

Do not reject the null. The computed test statistic falls just barely within the acceptance region.

Confirming with the Stata `bintesti` command,

```
. bintesti 70 29.5 .30 , normal

  Variable |       Obs  Proportion   Std. Error
  ---------+-----------------------------------
         x |       70    .4214286    .0590189

           Ho:      p = .3
                    z = 2.22

. display 1-norm(2.22)
.01320938
```

So, using the normal approximation to the binomial with correction for continuity, we see that there is a 1.3% chance of getting this many investors losing money if the company is telling the truth. Do not reject if using the .01 level of significance.

Using the more precise `bitesti` command,

```
. bitesti 70 30 .30

        N   Observed k   Expected k   Assumed p   Observed p
------------------------------------------------------------
       70        30           21       0.30000      0.42857

  Pr(k >= 30)            = 0.015236  (one-sided test)
  Pr(k <= 30)            = 0.992020  (one-sided test)
  Pr(k <= 12 or k >= 30) = 0.025739  (two-sided test)
```

So, there is about a 1.5% chance that you would get this many investors losing money if the company is telling the truth. Do not reject if using the .01 level of significance.