K300 (4392) Statistical Techniques (Fall 2007)

Midterm Exam (Due October 17)

Instructor: Hun Myoung Park
kucc625@indiana.edu, (317) 274-0573

You must read questions carefully to get exactly what you are supposed to do.

1. (10 points) True/false questions. *In case of a false statement, correct the statement*.

1) A random variable is always normally distributed.
2) Surprisingly, I got a conditional probability of -.0001.
3) "My sample mean is equal to 3.5" is a relevant null hypothesis.
4) A standard error is nothing but the standard deviation of a sample mean.
5) The classical approach to hypothesis testing compares a test statistic with a critical value.
6) In the p-value approach, you need to reject the null hypothesis when the p-value is large than the significance level.
7) The 95 percent confidence interval can be interpreted as "I am 95 percent sure that the population mean exists somewhere within the confidence interval."
8) "Since the test statistic is smaller than the critical value, I do not accept the null hypothesis at the 5 percent significant level." is a proper expression.
9) A p-value means the probability that you would like to commit the type I error.
10) A critical value and a rejection area are computed from the sample, so they are objective criteria.
11) (bonus 1 point) a sample mean is a random variable.
12) (bonus 1 point) Mutually exclusive events are statistically independent.

2. (5 points) Suppose you have a question in a questionnaire asking a respondent to choose one of the following range of annual personal income. 1) $0-$10K, 2) $10K-$30K, 3) $30K-$50K, 4) $50-$70K, 5) $70K-$100K, 6) $100K-$150K, 7) $150K-$200K, 8) more than $200K. Note that you assigned numbers 1 through 8 to corresponding categories. Choose only one level of measurement that is most likely in each question

1) What level of measurement, in general, are you using?
2) If your boss asks you to assign the midpoint of a range of personal instead of the serial numbers in 1), what level of measurement does your boss have in mind? Note that midpoints are $5K, $20K, $40K and so on.

3. (5 points) You may earn up to 100 points from this midterm exam. I will record your raw score (e.g., 95.5, 89.0, and so on) in a SPSS file and analyze its distribution. Choose only one level of measurement that is most likely in each question.

1) What level of measurement am I using?
2) You final grade is based on eight assignments and two exams. The sum of these weighted raw scores will be transformed into a letter grade (e.g., A+ for 98-100, A for 93-97, A- for 90-92, and so on) Which level of measurement is this letter grade?

3) Again, the letter grades that you earn this semester are used to compute your GPA. A+ and A are treated as 4.0, A- as 3.7, B+ as 3.3, and so on. Which level of measurement is your GPA?

4. (15 points) The following data points are a part of your GPA (manipulated somehow) last semester. Do not forge to sort data first.

3.1, 3.8, 3.6, 2.9, 3.2, 2.89, 2.9, 3.5, 2.8, 3.3, 3.8, 3.5, 3.2, 3.8, 3.2, 2.5, 3.5, 2.5, 3.4, 2.8

1) Report N, minimum, Q1, Q2, Q3, maximum, interquartile range (IQR), which is Q3-Q1 , mean, variance, and standard deviation. You MUST show how you get these statistics; do not simply report single numbers.
2) Draw a box plot with the mean indicated with "*". You have to use a real scale; do not simply draw an arbitrary box without meaningful difference between particular two data points. This is an interval (or ratio) variable.
3) Do you think that this GPA is normally distributed? How do you know that? You have to imagine how the histogram or stem-and-leaf plot looks like by looking at the box plot you draw. Do not try to conduct formal tests such as Shapiro-Wilk W test and Jarque-Bera test (Skewness-Krutosis test). Just eyeballing and imagination will do.

5. (10 points) According to the 2006 Post-election survey of the Pew Internet and American Life Project, about 66 percent of respondents (1,682) use the Internet at least occasionally and about 61 percent (1,570) send or receive emails at least occasionally. A total of 2,559 respondents answered both questions. See the following table for detailed information. Determine whether using the Internet (let us call this event "I") and using emails ("E") are statistically independent.

	Use Emails (Yes)	Never Use Emails
Use the Internet (Yes)	1,526	156
Never Use the Internet	44	833

1) Compute P(I="yes"), P(E="yes"), and P(I="yes" and E="yes").
2) Get both P(I|E) and P(E|I). Which one is large? Are these two events statistically independent?
3) Interpret the result substantively. In other word, use the narrative definition of "statistical independence."

```
P(I)=1682/2559=.6573
P(E)=1570/2559=.6135
P(I and E)=1526/2559=.5963
P(E|I)=.5963/.6573=.9073
P(I|E)=.5963/.6135=.9720

P(I|E) should be 1. There are some problems in responds. Email uses the Internet.

. di 1570/2559
.61352091

. di 1682/2559
.657288
```

```
. di 1526/2559
.59632669

. di 1526/2559/(1682/2559)
.90725327

. di 1526/2559/(1570/2559)
.97197452
```

6. (5 points) What the central limit theorem? You may answer using either narrative explanation or formal definition. Can you conduct z-test and t-test without the central limit theorem?

7. (10 points) Since his debut in 1998, Peyton Manning records a successful passing rate of .6413 (64.13 percent) for the past 9 years. Let us assume that his rate does not change this season regardless of opponent, place, receiver, weather, and the like. Yes, this is a strong assumption that is not realistic. Indianapolis Colts will have a game with Jacksonville Jaguars on October 22. This is an away-game, but that does not influence his rate (by assumption). Think about his first 10 passes (of course, we assume he will play as a quarterback as usual) in the game. We do not know (even Peyton himself does not know either) exactly when, where, and how his pass will occur. So the number of successful passes can be considered as a random variable generated from the Bernoulli process.

1) What is the probability that Peyton makes five successes out of 10 passes? That is, P(x=5). You may take advantage of $.6413^4$=.1691 and $.3587^4$=.0166 to save time.

2) Suppose he fails to pass 8 times out of 10 trials. Do you think Peyton is in a normal condition? In order word, is the rate .6413 still valid? In order to answer, you need to compute P(x=2) for two successful passes, P(x=1) for only one successful pass, and P(x=0) for none of successful pass. If the sum of these three probabilities is less than .01 (significance level), you may think that he is not in a good condition and conclude that Peyton looks like a totally different person who has a poor successful passing rate (reject the null hypothesis of .6413 in a statistical sense). Would you like to forgive his unusual mistakes or throw a can of beer at him (But NEVER, EVER DO that in the real world)?

```
1) p=64.13, q=35.87, N=10
```

$P(x=5)={}_{10}C_5\ (.6413)^5(.3587)^5=252*.1085*.0059=.1613$

```
. di .6413^5
.10846914

. di .3587^5
.00593823
```

$P(x=0)={}_{10}C_0\ (.6413)^0(.3587)^{10}=.000035$
$P(x=1)={}_{10}C_1\ (.6413)^1(.3587)^9=.00063$
$P(x=1)={}_{10}C_2\ (.6413)^2(.3587)^8=.00507$

```
Sum of these three is .0057

> factorial(10)/factorial(0)/factorial(10)*.3587^10
[1] 3.526255e-05
> factorial(10)/factorial(1)/factorial(9)*.6413^1*.3587^9
```

```
[1] 0.0006304398
> factorial(10)/factorial(2)/factorial(8)*.6413^2*.3587^8
[1] 0.005072079
> .00003526255+.0006304398+.005072079
[1] 0.005737781
```

8. (20 points) GPA of IUPUI/SPEA is known to be normally distributed with mean 3.0 last semester. Suppose 21 of you were randomly selected to evaluate whether the hypothesized mean is reasonable. The sample mean and standard deviation were 3.1852 and .4142, respectively. Test the null hypothesis that population mean is 3.0 using three approaches at the .05 level. You must follow all five steps (procedures) as shown in the lecture note 2.

1) Use the test statistic (or test value) approach. Report the degress of freedom and then determine the critical value.

2) Use the p-value approach. Determine if the p-value is smaller than the significance level by looking at the *t* distribution table. (Hint: *You may not get the accurate p-value from the table. Compare the test statistic with t values for the .10, .05, .02, and .01 level.*)

3) Use the confidence interval approach. You have to construct the 95 percent confidence interval.

4) Did three approaches end up with the same conclusion? If not, tell me why?

```
Df=20=21-1, Critical value is 2.086 at the .05 level
```

$$t_{\bar{x}} = \frac{\bar{x}-\mu}{s/\sqrt{n}} = \frac{3.1852-3.0}{.4142/\sqrt{21}} = 2.0490 \sim t(21-1)$$

```
> (3.1852-3)/.4142*sqrt(21)
[1] 2.048993

. di ttail(20, 2.048993)
.02690611

. di 3.1852+2.086*.4142/sqrt(21)
3.3737449

. di 3.1852-2.086*.4142/sqrt(21)
2.9966551

. di 2.086*.4142/sqrt(21)
.18854488

. di ttail(20,2.0490)
.02690573

. di ttail(20,2.0490)*2
.05381146
```

$$\bar{x} \pm t_{\alpha/2}\frac{s}{\sqrt{n}} = 3.1852 \pm 2.086 \times \frac{.4142}{\sqrt{21}} = 3.1852 \pm .1885\text{, [2.9967, 3.3737]}$$

9. (10 points) Noise levels at various area urban hospitals were measured in decibels. The mean noise level of 24 randomly selected hospitals was 41.6 decibels and standard deviation was 7.5. Examine if the population noise level (true mean) is less than or equal to 45 at the .05 level using the p-value approach only (ignore the test statistic approach and the confidence interval approach). You have to follow all five steps.

1) Report N, μ, sample mean, test size (α), and the degrees of freedom.
2) Conduct the null hypothesis using the p-value approach.
3) How to substantively interpret the p-value you compute? What does that mean? Remember the lecture note that compares three approaches of hypothesis testing.

1) 24, 45, 41.6, 7.5, .05, 23 (=24-1)
2) 41.6±2.0685*7.5/sqrt(24).

$$t_{\bar{x}} = \frac{\bar{x} - \mu}{s/\sqrt{n}} = \frac{41.6 - 45}{7.5/\sqrt{24}} = -2.2209 \sim t(24-1)$$

```
. di (41.6-45)/7.5*sqrt(24)
-2.2208707

. di 1-ttail(23,-2.2208707)
.01824417

p-value is .0182
```

10. (10 points) According to *HIV/AIDS Surveillance Report: Cases of HIV Infection and AIDS in the United States, 2005* by National Center for HIV, STD and TB Prevention, Centers for Disease Control and Prevention, Department of Health and Human Services, 2006 (http://www.cdc.gov), the cumulative reported number of AIDS cases of Indiana in 2005 was 7,963. Suppose the Indiana population in 2005 is 6,313,520 (2006 estimation by U.S. Census Bureau).

1) Find the probability (empirical probability) that a person is reported as an AIDS case.
2) One of your classmates decided to examine whether the probability is likely. He or she surveyed 100 randomly selected subjects (residents) and found that .5 percent (.005) was reported (x=1). This is a sample proportion of the binomial distribution. Report the population probability (p), sample probability ($\hat{p}$), and N. What are the expected value (mean) and standard deviation (not variance) of x?
3) (7 points) Now, evaluate that Indiana rate (probability) of AIDS case is really what you got in 1). Conduct the z-test at the .05 significance level using the classical test statistic approach. Use the critical value of 1.96 at the .05 level. You have to follow all five steps. Do not forget to explicitly state the null and alternative hypothesis (step 1) and interpret the result substantively (step 5).

```
1)
. di 7963/6313520
.00126126
2) p=.0013, p hat=.005, N=100
```

3) $z_{\hat{p}} = \frac{\hat{p} - p}{\sqrt{pq/n}} = \frac{.005 - .0013}{\sqrt{.0013 * .9987/100}} = 1.0269$

Bonus (3 point) How do you interpret the result of question 10 if the sample was not randomly drawn (e.g., he or she went to a Kroger nearby at 6:00 P.M. and surveyed 100 people passing by. This is an example of *convenience sampling*). Can the result support the null hypothesis?