**Practice for Exam 1**

Econ 29000

Kevin R Foster, CCNY

Fall 2011

Not all of these questions are strictly relevant; some might require a bit of knowledge that we haven't covered this year, but they're a generally good guide.

1. A random variable is distributed as a standard normal. (You are encouraged to sketch the PDF in each case.)
   a. What is the probability that we could observe a value as far or farther than 1.3?
   b. What is the probability that we could observe a value nearer than 1.8?
   c. What value would leave 10% of the probability in the right-hand tail?
   d. What value would leave 25% in both the tails (together)?
2. You are given the following data on the number of people in the PUMS sample who live in each of the five boroughs of NYC and who commute in each specified manner (where 'other' includes walking, working from home, taking a taxi or ferry or rail).

| | Bronx | Manhattan | Staten Is | Brooklyn | Queens |
|---|---|---|---|---|---|
| car | 5788 | 2692 | 5526 | 10990 | 16905 |
| bus | 3132 | 2789 | 1871 | 4731 | 4636 |
| subway | 6481 | 13260 | 279 | 18951 | 14025 |
| other | 2748 | 10327 | 900 | 6587 | 4877 |

   a. Find the Joint Probability for drawing, from this sample, a person from Queens who commutes by bus. Find the Joint Probability of a person from the Bronx who commutes by subway.
   b. Find the Marginal Probability of drawing, from among the people who commute by subway, someone who lives in Brooklyn. Find the Marginal Probability, of people who commute by bus, someone who lives in the Bronx.
   c. Find the Marginal Probability of drawing, from among the people who live in Staten Island, someone who drives a car to work. Find the Marginal Probability, of people in Brooklyn, who commute by subway.
   d. Are these two choices (which borough to live in, how to commute) independent? Explain using the definition of statistical independence.
3. Download the PUMS data on people living in the 5 boroughs. Run a regression that models the variable, "GRPIP," "Gross Rent as Percent of Income," which tells how burdensome are housing costs for different people.
   a. What are the mean, median, 25th, and 75th percentiles for Rent as a fraction of income? Does this seem reasonable?
   b. What is the fraction spent on rent by households in Brooklyn? In Queens? Is the difference statistically significant? Between Brooklyn and the Bronx?
4. Using the NHANES 2007-09 data (not necessary to actually download), reporting a variety of socioeconomic variables as well as behavior choices such as the number of sexual partners reported (number_partners), we want to see if richer people have more sex than poor people. The following table is constructed, showing three categories of family income and 5 categories of number of sex partners:

| | number of sex partners | | | | | |
|---|---|---|---|---|---|---|
| family income | zero | 1 | 2 - 5 | 6 - 25 | >25 | Marginal: |
| < 20,000 | 11 | 63 | 236 | 255 | 92 | ______ |
| 20 - 45,000 | 7 | 117 | 323 | 308 | 117 | ______ |
| > 45,000 | 3 | 234 | 517 | 607 | 218 | ______ |
| Marginal: | ______ | ______ | ______ | ______ | ______ | |

   a. Where is the median, for number of sex partners, for poorer people? For middle-income people? For richer people?
   b. Conditional on a person being poorer, what is the likelihood that they report fewer than 6 partners? Conditional on being middle-income? Richer?

c. Conditional on reporting 2-5 sex partners, what is the likelihood that a person is poorer? Middle-income? Richer?

d. Explain why the average number of sex partners might not be as useful a measure as, for example, the data ranges above or the median or the 95%-trimmed mean.

5. A random variable is distributed as a standard normal. (You are encouraged to sketch the PDF in each case.)
   a. What is the probability that we could observe a value as far or farther than -0.9?
   b. What is the probability that we could observe a value nearer than 1.4?
   c. What value would leave 5% of the probability in the right-hand tail?
   d. What value would leave 5% in both the tails (together)?

6. You are in charge of polling for a political campaign. You have commissioned a poll of 300 likely voters. Since voters are divided into three distinct geographical groups, the poll is subdivided into three groups with 100 people each. The poll results are as follows:

| | total | A | B | C |
|---|---|---|---|---|
| number in favor of candidate | 170 | 58 | 57 | 55 |
| number total | 300 | 100 | 100 | 100 |
| std. dev. of poll | 0.4956 | 0.4936 | 0.4951 | 0.4975 |

Note that the standard deviation of the sample (not the standard error of the average) is given.

   a. Calculate a t-statistic, p-value, and a confidence interval for the main poll (with all of the people) and for each of the sub-groups.
   b. In simple language (less than 150 words), explain what the poll means and how much confidence the campaign can put in the numbers.
   c. Again in simple language (less than 150 words), answer the opposing candidate's complaint, "The biased media confidently says that I'll lose even though they admit that they can't be sure about any of the subgroups! That's neither fair nor accurate!"

7. Calculate the probability in the following areas under the Normal pdf with mean and standard deviation as given. You might usefully draw pictures as well as making the calculations. For the calculations you can use either a computer or a table.
   a. What is the probability, if the true distribution has mean -15 and standard deviation of 9.7, of seeing a deviation as large (in absolute value) as -1?
   b. What is the probability, if the true distribution has mean 0.35 and standard deviation of 0.16, of seeing a deviation as large (in absolute value) as 0.51?
   c. What is the probability, if the true distribution has mean -0.1 and standard deviation of 0.04, of seeing a deviation as large (in absolute value) as -0.16?

8. Using data from the NHIS, we find the fraction of children who are female, who are Hispanic, and who are African-American, for two separate groups: those with and those without health insurance. Compute tests of whether the differences in the means are significant; explain what the tests tell us. (Note that the numbers in parentheses are the standard deviations.)

| | with health insurance | without health insurance |
|---|---|---|
| female | 0.4905 | 0.4811 |
| | (0.49994) N=7865 | (0.49990) N=950 |
| Hispanic | 0.2587 | 0.5411 |
| | (0.43797) N=7865 | (0.49857) N=950 |
| African American | 0.1785 | 0.1516 |
| | (0.38297) N=7865 | (0.35880) N=950 |

9. Using the BRFSS 2009 data, the following table compares the reported health status of the respondent with whether or not they smoked (defined as having at least 100 cigarettes)

| | | SMOKED AT LEAST 100 CIGARETTES | | |
|---|---|---|---|---|
| | | Yes | No | Marginal |
| GENERAL HEALTH | Excellent | 27775 | 49199 | ____ |
| | Very good | 58629 | 77357 | ____ |
| | Good | 64237 | 67489 | ____ |
| | Fair | 31979 | 26069 | ____ |
| | Poor | 15680 | 9191 | ____ |

| | | |
|---|---|---|
| Marginal | ____ | ____ |

a. (10 points) What is the median health status for those who smoked? For non-smokers?
b. (10 points) Fill in the marginal probabilities – make sure they are probabilities.
c. (5 points) Explain what you might conclude from this data.

10. For a Standard Normal distribution (you are encouraged to sketch the PDF in each case),
    a. what is the area to the left of -1.5?
    b. what is the area to the right of 0.2?
    c. what is the area to the right of -1.6?
    d. what is the area to the left of -2.2?
    e. what is the area in both tails farther than 1.7?
    f. what is the area in both tails farther than -1.4?
    g. what distance from the mean (in absolute value) leaves 0.17 in both tails?
    h. what distance from the mean (in absolute value) leaves 0.29 in both tails?
11. For a Normal distribution(you are encouraged to sketch the PDF in each case),
    a. with mean 12 and standard deviation of 4, what is the area to the left of 20.4?
    b. with mean 7 and standard deviation of 4, what is the area to the right of -0.2?
    c. with mean -12 and standard deviation of 5, what is the area in both tails farther from the mean (in absolute value) than -3.5?
    d. with mean 13 and standard deviation of 2, what is the area in both tails farther from the mean (in absolute value) than 11.6?
    e. with mean -13 and standard deviation of 9 what values leave 0.09 in both tails?
    f. with mean -12 and standard deviation of 9 what values leave 0.97 in both tails?
12. With the ATUS dataset, people 20-50 years old with positive earnings were selected and then grouped into "low-earning" (people in families with earnings below the 25th percentile) and "high-earning" (people in families with earnings above the 75th percentile). The following statistics, the sample average and sample standard deviation, were calculated by SPSS:

| | | **hours shopping per day** | |
|---|---|---|---|
| | **N** | **Average** | **Std Dev** |
| **low earnings** | 9372 | 44.70 | 77.97 |
| **high earnings** | 9503 | 46.08 | 78.14 |

    a. What is the difference in average time spent shopping? For the null hypothesis of zero difference, form a hypothesis test and explain the result.
    b. What is a confidence interval for the difference?
    c. What is the p-value of the difference?

13. With the ATUS dataset, people 20-50 years old with positive earnings were selected and then grouped into "low-earning" (people in families with earnings below the 25th percentile) and "high-earning" (people in families with earnings above the 75th percentile). The following statistics, the sample average and sample standard deviation, were calculated by SPSS:

| | | **hours watching TV per day** | |
|---|---|---|---|
| | **N** | **Average** | **Std Dev** |
| **low earnings** | 9372 | 2.31 | 2.40 |
| **high earnings** | 9503 | 1.90 | 2.01 |

    a. What is the difference in average time spent watching TV? For the null hypothesis of zero difference, form a hypothesis test and explain the result.
    b. What is a confidence interval for the difference?
    c. What is the p-value of the difference?

14. SPSS produces the following output from the CPS data, a crosstab of income category with kids in the household. "Low family income" means that the person is in a family with income in the lowest quartile; middle means income in the next two quartiles; high is in the top quartile. Each household is classified with either no children, children under 6, or children under 18 but not under 6. (At 6 years old, children must start school.)

Count

| | | children in hh categories | | | |
|---|---|---|---|---|---|
| | | no kids | kids under 6 | kids older than 6 but less than 18 | Total |
| family income categories | low family income (less than 25th percentile) | 33782 | 10417 | 9712 | 53911 |
| | mid family income (25th - 75th percentile) | 41349 | 28450 | 33409 | 103208 |
| | high family income (more than 75th percentile) | 16964 | 13988 | 21731 | 52683 |
| Total | | 92095 | 52855 | 64852 | 209802 |

  a. What is the marginal probability for a household without children to have a low family income? What is the marginal probability for a household without children to have a high family income?
  b. What is the marginal probability for a household with a high family income to have children under 6 years old? What is the marginal probability for a household with low family income to have children under 6?

15. Using the ATUS dataset that we've been using in class (download from Blackboard), form a comparison of the mean amount of time spent on religious activity by two groups of people (you can define your own groups, based on any of race, ethnicity, gender, age, education, income, or other of your choice).
    a. What are the means for each group?
    b. What is the standard deviation of each mean? What is the standard error of each mean?
    c. What is a 95% confidence interval for each mean?
    d. Is the difference statistically significant? Explain carefully. What can be concluded from this?
16. Calculate the probability in the following areas under the Standard Normal pdf with mean of zero and standard deviation of one. You might usefully draw pictures as well as making the calculations. For the calculations you can use either a computer or a table.
    a. What is the probability, if the true distribution is a Standard Normal, of seeing a deviation from zero as large (in absolute value) as 1.9?
    b. What is the probability, if the true distribution is a Standard Normal, of seeing a deviation from zero as large (in absolute value) as -1.5?
    c. What is the probability, if the true distribution is a Standard Normal, of seeing a deviation as large (in absolute value) as 1.2?
17. Calculate the probability in the following areas under the Normal pdf with mean and standard deviation as given. You might usefully draw pictures as well as making the calculations. For the calculations you can use either a computer or a table.
    a. What is the probability, if the true distribution has mean -1 and standard deviation of 1.5, of seeing a deviation as large (in absolute value) as 2?
    b. What is the probability, if the true distribution has mean 50 and standard deviation of 30, of seeing a deviation as large (in absolute value) as 95?
    c. What is the probability, if the true distribution has mean 0.5 and standard deviation of 0.3, of seeing a deviation as large (in absolute value) as zero?
18. A paper by Chiappori, Levitt, and Groseclose (2002) looked at the strategies of penalty kickers and goalies in soccer. Because of the speed of the play, the kicker and goalie must make their decisions simultaneously (a Nash equilibrium in mixed strategies). For example, if the goalie moves to the left when the kick also goes to the left, the kick scores 63.2% of the time; if the goalie goes left while the kick goes right, then the kick scores 89.5% of the time. In the sample there were 117 occurrences when both players went to the left and 95 when the goalie went left while the kick went right. What is the p-value for a test that the probability of scoring is different? What advice, if any, would you give to kickers, based on these results? Why or why not?
19. A paper by Claudia Goldin and Cecelia Rouse (1997) discusses the fraction of men and women who are hired by major orchestras after auditions. Some orchestras had applicants perform from behind a screen (so that the gender of the applicant was unknown) while other orchestras did not use a screen and so were able to see the gender of the applicant. Their data show that, of 445 women who auditioned from behind a screen, a fraction 0.027 were "hired".

Of the 599 women who auditioned without a screen, 0.017 were hired. Assume that these are Bernoulli random variables. Is there a statistically significant difference between the two samples? What is the p-value? Explain the possible significance of this study.

20. Another paper, by Kristin Butcher and Anne Piehl (1998), compared the rates of institutionalization (in jail, prison, or mental hospitals) among immigrants and natives. In 1990, 7.54% of the institutionalized population (or 20,933 in the sample) were immigrants. The standard error of the fraction of institutionalized immigrants is 0.18. What is a 95% confidence interval for the fraction of the entire population who are immigrants? If you know that 10.63% of the general population at the time are immigrants, what conclusions can be made? Explain.
21. Calculate the probability in the following areas under the Standard Normal pdf with mean of zero and standard deviation of one. You might usefully draw pictures as well as making the calculations. For the calculations you can use either a computer or a table.
    a. What is the probability, if the true distribution is a Standard Normal, if seeing a value as large as 1.75?
    b. What is the probability, if the true distribution is a Standard Normal, if seeing a value as large as 2?
    c. If you observe a value of 1.3, what is the probability of observing such an extreme value, if the true distribution were Standard Normal ?
    d. If you observe a value of 2.1, what is the probability of observing such an extreme value, if the true distribution were Standard Normal ?
    e. What are the bounds within which 80% of the probability mass of the Standard Normal lies?
    f. What are the bounds within which 90% of the probability mass of the Standard Normal lies?
    g. What are the bounds within which 95% of the probability mass of the Standard Normal lies?
22. Consider a standard normal pdf with mean of zero and standard deviation of one.
    a. Find the area under the standard normal pdf between -1.75 and 0.
    b. Find the area under the standard normal pdf between 0 and 1.75.
    c. What is the probability of finding a value as large (in absolute value) as 1.75 or larger, if it truly has a standard normal distribution?
    d. What values form a symmetric 90% confidence interval for the standard normal (where symmetric means that the two tails have equal probability)? A 95% confidence interval?
23. Now consider a normal pdf with mean of 3 and standard deviation of 4.
    a. Find the area under the normal pdf between 3 and 7.
    b. Find the area under the normal pdf between 7 and 11.
    c. What is the probability of finding a value as far away from the mean as 7 if it truly has a normal distribution?
24. If a random variable is distributed normally with mean 2 and standard deviation of 3, what is the probability of finding a value as far from the mean as 6.5?
25. If a random variable is distributed normally with mean -2 and standard deviation of 4, what is the probability of finding a value as far from the mean as 0?
26. If a random variable is distributed normally with mean 2 and standard deviation of 3, what values form a symmetric 90% confidence interval?
27. If a random variable is distributed normally with mean 2 and standard deviation of 2, what is a symmetric 95% confidence interval? What is a symmetric 99% confidence interval?
28. A random variable is distributed as a standard normal. (You are encouraged to sketch the PDF in each case.)
    a. What is the probability that we could observe a value as far or farther than 1.7?
    b. What is the probability that we could observe a value nearer than 0.7?
    c. What is the probability that we could observe a value as far or farther than 1.6?
    d. What is the probability that we could observe a value nearer than 1.2?
    e. What value would leave 15% of the probability in the left tail?
    f. What value would leave 10% of the probability in the left tail?
29. A random variable is distributed with mean of 8 and standard deviation of 4. (You are encouraged to sketch the PDF in each case.)
    a. What is the probability that we could observe a value lower than 6?
    b. What is the probability that we could observe a value higher than 12?
    c. What is the probability that we'd observe a value between 6.5 and 7.5?
    d. What is the probability that we'd observe a value between 5.5 and 6.5?
    e. What is the probability that the standardized value lies between 0.5 and -0.5?
30. You know that a random variable has a normal distribution with standard deviation of 16. After 10 draws, the average is -12.
    a. What is the standard error of the average estimate?
    b. If the true mean were -11, what is the probability that we could observe a value between -10.5 and -11.5?

31. You know that a random variable has a normal distribution with standard deviation of 25. After 10 draws, the average is -10.
    a. What is the standard error of the average estimate?
    b. If the true mean were -10, what is the probability that we could observe a value between -10.5 and -9.5?
32. You are consulting for a polling organization. They want to know how many people they need to sample, when predicting the results of the gubernatorial election.
    a. If there were 100 people polled, and the candidates each had 50% of the vote, what is the standard error of the poll?
    b. If there were 200 people polled?
    c. If there were 400 people polled?
    d. If one candidate were ahead with 60% of the vote, what is the standard error of the poll?
    e. They want the poll to be 95% accurate within plus or minus 3 percentage points. How many people do they need to sample?