Homework 2, Due in Class 10/12

Due 10-12 in Class

October 17, 2014

1 Remind me that you know probability

1. Given independent random variables X & Y, X distributed Normal with μ = 3 and σ = 1, and Y distributed Normal with μ = -4 and σ = 2.

   (a) Calculate the mean and variance by hand of:

      i. X + Y

      ii. X - Y

      iii. 3X - 2Y

   (b) Write, in R, a function that you could use to appropriately simulate this addition and subtraction. If you like, you may write a function for each of these different sub-problems, however, style points if you write something that is sufficiently general to handle all three by passing different arguments to the function. Double style points if you write it so that I can manipulate the scalars that are transforming the RVs.

2. Let X $\sim$ N(2,7). For parts (a) and (b) please: draw a picture (by hand), of this region, then use R to find the probability.

   (a) What is P(x < 4)?

   (b) What is $P(0 \leq x \leq 2.5)$?

   (c) What value on this distribution has 30% of the distribution below this number?

   (d) Find the high-bound, h to set if you would like 40% of the distribution to fall on the range $[1, h]$.

3. Lou asked a good question in class when he asked, "At what point do we think the central limit theorem 'starts to work.' " Test the performance of the Central Limit Theorem for the following random variables. Take many samples of size n, calculate a mean of the sample, save it, [repeat], then examine the distribution, mean, and variance of these sample means. For each of the samples, does the **CLT** work? Why, or why not? How many samples do you need before it starts to work? Style points if you can make a graph that shows increasing performance (or not-increasing performance) as the number of iterations increases.

   (a) $Y \sim B(n, p)$ (Binomial) for $p = 0.9$

   (b) $Y \sim E(1)$ (Exponential) with rate parameter 1.

   (c) $Y = |X|^x$, where $X \sim N(0, 1)$.

   (d) $Y \sim \chi_5^2$ (Chi-squared with parameter $k = 5$).

   (e) $Y \sim C(0, 1)$ where C is the Cauchy distribution with location parameter 0 and scale parameter 1.

2 Test some questions fake data

1. The Gubenatorial election is coming up. What is the support for Governor Brown?

   (a) We take a sample of 200 Californias and find that 120 support the governor and 80 do not support the governor.

      i. Report a confidence interval for the true proportion of Californians that support Brown.

      ii. Test the hypothesis that **less** than a majority support Brown. Follow the steps from lecture (short-form is fine).

      iii. How many survey respondents would we need to sample so that our Margin of Error were $\pm 2\%$?

   (b) You've just won the NSF and are excited to run more surveys. So, you field a new survey of San Diegans as they come back from a surf at Black's Beach (most are, in fact, wearing clothes). You speak to 200 men (110 support Brown) and 400 women (190 support Brown).

      i. Report a confidence interval for the gender gap.

      ii. Test the hypothesis that men and women have different support levels. Follow the steps from lecture.

iii. Write a function that will calculate the test-statistic for this difference that takes as arguments to the function nM, sM, nW, sW (number of men, support of men, number of women, support of women; respectively).

2. Do faculty members at ivy league schools earn more than faculty at public institutions? We know that salaries at public institutions average 90k per year (which we get from public information requests, and is the population mean). Private schools aren't required to disclose this, so we put out a survey. We sample 100 ivy-league faculty and find that the average salary is 96k with a sample standard deviation (s) of 20k.

   (a) Report a confidence interval for the true average salary earned by ivy-league faculty.

   (b) Test the hypothesis that there is no difference in salaries.

   (c) Write a function that will compare simulated draws of 100 salary members. If s is assumed to be constant, what sample size will give us enough *power* to capture a difference of just 2k in expected earnings 80% of the time?

3. Test for the independence of Partisanship and Region in the US. First, do it by hand. Then, write a program that will calculate the chi-square test-statistic and associated p-value for **any** 3x3 contingency table.

Region	D	R	NPP
Deep South	123	25	101
North East	89	40	100
South	74	65	25

3 Test some questions using real data

1. Get study #7039 from ICPSR.

   (a) Conduct a chi-square test for independence between occupation and the most important problem of the coutnry (see p. 11 of the codebook).

   (b) Try a two-sample proportion test to test whether there are different rates of party membership for men and women (see p. 10 & 12 of the codebook).

2. Get study #8475 from the ICPSR.

(a) Do men and women have the same or different opinions about Hillary Clinton? Test this using variables VCF0471 and VCF0104.

(b) Is there more diversity of opinion about Hillary Clinton for men or for women? How would you test this (hint: it is in our slides)? On this question, be specific about the test statistic that you are using.

3. Read the following vectors. Unfortunately, I corrupted a file. . . and you’ve got to fix it before you can perform the test. All of the 3’s were changed to underscores (_) in the first column, decimal points all were turned into “\\” in the second column, and 4s all turned into spaces. Fix these problems, then use a KS-test to test for the probability that they are drawn from the same distribution. Can you distinguish between either of these vectors and a normal with mean = 100 sd = 10? What about a chi-square with 99 df?