Week 6 Live Session

w203 Instructional Team

Homework 5 Presentation

Sampling Distributions

What is the difference between the sampling distribution of a statistic and the population distribution of a variable?

Why do we want to know things about the sampling distrbution of a statistic?

Review of the Central Limit Theorem

In this exercise, you will recreate the demonstration of the CLT seen in the async. Instead of using the Old Faithful data, you are to take random draws from a Bernoulli distribution.

Recall that a Bernoulli random variable with parameter p takes on just two values: 1, with probability p; and 0, with probability $1 - p$. We choose this variable because (1) it's very simple, and (2) its distribution is distinctly non-normal.

It turns out that (base) R doesn't have a Bernoulli function. To simulate draws from a Bernoulli variable, you can either

a. Use the sample command to select values from {0,1}

```
n=3
p = 0.5
sample(c(0,1), 3, prob = c(1-p,p), replace = TRUE)
```

```
## [1] 0 0 0
```

b. Note that the Bernoulli distribution is a special case of the more general binomial distribution, with the binomial size parameter set to 1. R has an rbinom function that lets you draw from this distribution.

```
rbinom(3, size=1, prob=0.5)
```

```
## [1] 1 1 0
```

The Fair Coin

Using R, complete the following simulation exercise.

1. First, set p = 0.5 so your population distribution is symmeteric. Use a variable n to represent your sample size. Initially, set $n = 3$.
2. Simulate n draws from a Bernoulli variable with parameter p, then compute the sample mean.
3. Write code to replicate the above experiment 100,000 times, storing all of the resulting sample means. Create a histogram of your result. Compute the standard deviation of the result.
4. Increase n to 30. Mathematically, how should this change the standard deviation of the sampling distribution of the mean?

4) Experiment with different values of n and note the point at which the sampling distribution of the mean looks normal to you.

A Skewed Distribution

Now let p = 0.0001. This is now a highly skewed Bernoulli variable. We are especially interested to see how skewed the sampling distribution of the mean will be for different sample sizes.

For this activity, you can simply assess the skew of a distribution visually. If you prefer, you can also use the skewness command in the moments package. You may hear a rule of thumb that a skewness less than -1 or greater than 1 is considered substantially skewed.

```
library(moments)
skewness(rbinom(100000, size = 1, prob = .001))
```

```
## [1] 34.05571
```

1. Rerun the previous exercise for n = 3, p = 0.001 and note the shape of the sampling distribution.
2. Increase n to 30, and note the shape of the sampling distribution.
3. As before, experiment with different values of n and note the point at which the sampling distribution looks normal.

Discussion:

How does the skew of a distribution affect the applicability of the Central Limit Theorem?

Name a variable you would be interested in measuring that has a substatially skewed distribution.

A Cauchy Distribution is a well-known distribution with some interesting mathematical properties. In particular, it has "infinite" variance That is, the variance does not exist because the tails are too spread out. Would the previous exercise work if you took draws from a Cauchy distribution?

Generality of the Central Limit Theorem

Even though we stated the Central Limit Theorem for a sample mean, there are different versions for almost every statistic we will use in this course.

a. How could you apply the CLT to the product of random variables, $X_1 \cdot X_2 \cdot ... \cdot X_n$ where each X_i only takes on positive values? Hint: what transformation turns products into sums?

b. The sampling distribution of the variance is approximately normal for large samples.

c. We will use a version of the CLT when we study linear regression. It turns out that the sampling distribution of our slope coefficients will be approximately normally distributed in large samples.

At the same time, there may be some rare statistics for which the CLT does not hold.

a. Would today's exercise work if you used the sample median instead of the sample mean?