

Math 214

Confidence Intervals

Laboratory Project #5

Due: Monday March 25

Lets look at confidence intervals and hypothesis tests using some ‘real’ data. Actually, we can simulate some data and then use R to take samples, compute means and standard deviations, etc.

For example, let’s reconsider a question concerning the amount of soda in some soda bottles.

The Pepsi Bottling Plant in Astoria Queens produces a huge number of half gallon, plastic Diet Pepsi bottles. The company claims that, due to manufacturing upgrades, the mean amount of soda in each bottle they ship has a median volume of 64.00 ounces.

Now suppose we had **ALL** the data on Pepsi’s Astoria soda bottles. This is the same as assuming we know everything about the **population**. We will use R to simulate a population of amounts of soda in 10,000 soda bottles. We assume some distribution function and randomly create 10,000 pieces of data in a variable called `pop`.

```
> pop = 65-rexp(10000,0.65)
> hist(pop)
```

Look at the distribution, you’ve produced. It’s definitely not normal.

To Do:

- Describe, in words, what the histogram says about the data. What are the chances of finding a soda bottle with much less than 64 oz in it? What about much more than 64 oz?
- Make a boxplot of the data. Is it what you expect? Why?
- Find the population mean and standard deviation.

Let’s look into the **Central Limit Theorem** using this population. First, the central limit theorem says that no matter what the shape of the population distribution, the shape of the distribution of sample means (for large enough sample size) is approximately **normal**. Suppose we take many samples of size $n = 40$. We can do this in R with the `sample` command. Lets take 500 samples of size 40 in `xsamp` and look at the sample distribution.

```
> for (i in 1:500) samp_mean[i] = mean(sample(pop,40, rep=T))
> hist(samp_mean)
```

Well, the distribution of sample means *looks* more like a normal distribution than the distribution of volumes in the population. How normal is the distribution of 500 sample means? One way to quantify this is to make the ‘normal probability’ plot. The basic idea is to plot the quantiles of your distribution against the quantiles of a truly normal distribution. If the plot is a straight line, then the data is distributed just like the normal data. Try this in R.

```
> qqnorm(samp_mean)
> qqline(samp_mean)
```

Pretty close to a straight line, at least in the center of the distribution.

To Do:

- Change the sample size and compare the qqnorm plots. Try sample sizes $n = 64$, $n = 100$ and $n = 256$.
- What trend do you see? Explain.
- Is the sample mean *more Normal* for the larger values of n ? Support your statement with graphs.

We want to compute some Confidence Intervals from our samples. In this (artificial) case, we know the population parameters: μ (the population mean) and σ , the population standard deviation. Use R to calculate these.

Now lets return to our samples of size $n = 40$, and take 50 such samples.

```
> samp_mean = numeric(0); # Clear out the old values
> for (i in 1:50) samp_mean[i] = mean(sample(pop,40, rep=T))
> hist(samp_mean)
```

We know how to calculate the standard deviation of the sample mean:

$$\sigma_{\bar{x}} = \sigma / \sqrt{(n)}$$

Do this given your population standard deviation. If we pick a confidence interval, say $\alpha = 0.1$ (90% confidence), we can compute a confidence interval for our measure of the population mean for each one of our samples. Now lets compute and plot the confidence intervals for the 50 samples:

```
> m = 50; n = 40; mu = mean(pop); sigma = sd(pop);
> SE = sigma/sqrt(n) # Standard error in mean
> alpha = 0.10 ; zstar = qnorm(1-alpha/2); # Find z for 90% confidence
> matplot(rbin( samp_mean - zstar*SE, samp_mean + zstar*SE),rbind(1:m,1:m), type="l", lty=1);
> abline(v=mu)
```

Voila. What is (or should be) plotted is 50 confidence intervals for our 50 estimates of the population mean at 90% confidence. How many of these 50 intervals actually enclose the true value of the population mean? If we change: (a) the sample size or (b) the confidence level, how will this picture change?

To Do:

- Change the sample size, n . Try sample sizes $n = 64$, $n = 100$ and $n = 256$.
- What happens to the standard error in the mean as n increases?
- Replot the 50, 90% confidence intervals. How many of the 50 actually bracket the true mean?
- Set the sample size, $n = 40$ and change the confidence interval. For $\alpha = 0.05$, make a plot of 100 confidence intervals. How many do expect will NOT bracket the true population mean? Is this what you see?

To Do: For the following problems, compute the %80,%90,%95 confidence intervals for the estimate of the population mean. You can use R, or do it by hand.

- A population has a standard deviation of 101. You take a sample of size $n = 100$ and measure the sample mean, $\bar{x} = 12$.
- For the same population, you take a sample of size $n = 20$ and measure the sample mean, $\bar{x} = 12$.