Stats Lab

Data Mining

1. State the central limit theorem, and make sure you understand the program that simulates it by generating the sampling distribution of the sample means.

2. Assume you have a population that follows the exponential distribution with rate $\lambda = 0.1$. (Note: this means that the probability density function is $f(x) = 0.1e^{-0.1x}$.) Use the simulation program to find the empirical probability of drawing a random sample of size $n = 9$ from this population and getting a sample mean less than 7.

2a. As it turns out, the mean and standard deviation of this distribution are both equal to $\frac{1}{\lambda} = 10$. Can we use this information to compute the theoretical probability of getting a sample mean smaller than 7? Explain.

3. Repeat Problems 2 and 2a using a sample size of $n = 64$.

4. A 95% confidence interval for the population mean μ uses a sample mean (usually called $\overline{x}$, but sometimes $\hat{\mu}$) to generate a range of plausible values for the population mean. This idea can be quantified as follows: there is a 95% probability that the confidence interval we generated will capture the true population mean. One subtlety here is that it is not quite correct to say ”the population mean is in our interval with probability 95%.” This statement is wrong because it implies that the population mean is subject to randomness, which it isn't. The population mean is fixed and we're trying to estimate it. All the randomness in the process is due to sampling. If we were to generate 100 samples, we'd get 100 different sample means, and hence, 100 different confidence intervals. We expect about 95 of these 100 confidence intervals to catch μ, and a few to miss it. Let's do some experiments in R to see some of these ideas in action.

Lets start with an (unrealistic) scenario where we know the population mean and standard deviation, say a normal distribution with mean $\mu = 50$ and $\sigma = 12$.

4a. Write an R program that generates 1000 95% confidence intervals for the mean using sample size $n = 9$, and stores them in a data frame that has a column named l.end (the left endpoint of the confidence interval) and a column named r.end (the right endpoint). **4b.** What percentage of the intervals in your data frame do, in fact, catch the true population mean?

Notice that your computation of confidence intervals required you to use the standard deviation of the sampling distribution, $\sigma_{\overline{x}}$. We know from the central limit theorem that $\sigma_{\overline{x}} = \frac{\sigma}{\sqrt{n}}$, where n is the sample size.

5. Now imagine a more realistic scenario where you are taking a single sample from an unknown population. The central limit theorem still tells us that, for large enough sample size, our sample comes from an approximately normal distribution. But what is the standard deviation of this distribution?

It's more than likely that we do not know the population standard deviation σ, and therefore, we also do not know the standard deviation of the sampling distribution $\sigma_{\overline{x}}$. So how can we generate confidence intervals?

Since we are using our sample mean $\overline{x}$ to estimate the population mean μ, it seems only reasonable to use the standard deviation of our sample, s, to approximate σ. Unfortunately, we run into a problem here. While

$$z = \frac{\overline{x} - \mu}{\sigma/\sqrt{n}} = \frac{\overline{x} - \mu}{\sigma_{\overline{x}}}$$

follows a standard normal distribution, our new statistic

$$t = \frac{\overline{x} - \mu}{s/\sqrt{n}}$$

does not! The the use of s to estimate σ introduces more variability. Fortunately, if we assume that the original population is approximately normal (and even that assumption is not completely essential) the distribution of t is well known and is called the Student's t-distribution. The t-distribution is still very much like the standard normal: it's symmetric, bell shaped, and has mean 0. Now, however, the distribution of t will depend on the sample size n, and will have thicker tails. That is, observations taken from a t-distribution are more likely to live further away from the mean. The quantity

$$SE_{\overline{x}} = \frac{s}{\sqrt{n}}$$

is called the standard error of the mean.

Let's take a look at a few t-distributions for various values of the sample size n. If the sample size is n, the t-distribution is said to have $n-1$ degrees of freedom.

5a. Try out the following code which will plot the t-distribution with $n-1$ degrees of freedom, and the standard normal distribution, on the same axis for several values of n.

```
range <- seq(-4,4,.01)
layout(matrix(1:6, ncol=3))
for (i in c(1,2,3,5,20,50)){
  plot(range,dnorm(range),lty=1,col="orange")
  lines(range,dt(range,df=i),lty=2,col="blue")
  mtext(paste("df=",i),cex=1.2)
}
```

As you can see, the difference between the standard normal and t-distributions is most pronounced for small sample size. For large sample size the difference becomes very small.

5b. Use R to find $P(t_{\nu=5} < -1)$ and compare it to $P(Z < -1)$ where Z is the standard normal random variable. Note that $t_{\nu=5}$ is a t-distribution with

5 degrees of freedom. For what value of t^* is $P(-t^* < t_{\nu=5} < t^*) = 0.95$? Repeat these questions with $\nu = 10$ and $\nu = 100$.

Now suppose we had a sample of size 10:

```
samp <- rnorm(10,50,9)
```

5c. Generate a 95% confidence interval using the t distribution as follows:

```
t.test(x)$conf.int
```

5d. Compare this confidence interval to the same one you would get if you knew that $\sigma = 9$. How and why are they different? Now do the same thing for a larger sample size of $n = 100$. Comment on the difference between the z and t intervals. **5e.** How would the confidence interval change if we raise the confidence level to 99%? (Make sure you understand how to explain the difference between a 95% CI and a 99% CI). Try this in R using:

```
t.test(x,conf.level=.99)$conf.int
```

6. Modify the confidence interval code you wrote above to generate confidence intervals using the t-distribution and verify that the results (the percentage of intervals that catch the true mean) still holds.

7. Write a program that verifies empirically that the sample variance is in fact a good approximation to the population variance assuming a normally distributed population. Does this depend on sample size? Comment on the shapes of any distributions you consider.