Lab 4: Confidence Intervals

ENVIRON 710: Applied Statistical Modeling*

The goal of this lab is to understand and practice testing for normality and calculating confidence intervals. We start again by expanding our knowledge of R through the use of for-loops, which you will employ to test data for normality and to compute confidence intervals.

After this lab, you should be able to:

- Test data for assumptions of normality
- Compute and interpret confidence intervals for small and large samples
- Use for-loops to run calculations or produce figures multiple times.

At the end of the lab, there are two problems to solve. Please show your code, type your answers to each problem, and plot graphs using R Markdown. *Submit your answers and R-code to the class Sakai site under the Assignments folder.*

More functions in R

In this lab, we introduce a few new R commands.

- `qqnorm()` - produces a normal QQ plot of the values in a vector
- `qqline()` - adds a theoretical line to the QQ plot based on a normal distribution
- `geom_abline()` - adds a straight line to an existing ggplot plot
- `geom_density()` - creates a density plot
- `runif()` - selects a random variable from a uniform distribution, defined by minimum and maximum values

Frequently in statistical analysis, it is important to be able to run calculations or functions multiple times. One intuitive way to do this is by using a for-loop. For-loops are a little bit controversial because running loops is slower than other methods (for example, using the `apply` family of functions). However, for-loops are intuitive and a good place to start before advancing into more efficient commands. The basic set-up looks like this:

```
for (counter in vector) {commands}
```

For a very simple demonstration, run this for-loop:

```
for(i in 1: 10){print(i)}
```

Now, let's illustrate the Central Limit Theorem (CLT) using a for-loop to pick random numbers from a uniform distribution. The CLT states that if you take repeated samples from a population with finite variance and calculate their averages, then the sampling distribution of the sample means will be normally distributed even if the underlying distribution is not normally distributed.

```
# Create a vector to store results
```

*Created by John Poulsen with edits from TAs.

```r
umeans <- numeric(10000)
set.seed(1001)

# Run the loop 10,000 times

 for (i in 1:10000){

# Take the mean of a sample of 5 uniformly distributed numbers
# and store it in the vector, umeans

  umeans[i] <- mean(runif(5, min = 0, max = 10))
 }

# Make a histogram of the 10,000 means
gg1 <- ggplot(data.frame(umeans), aes(x=umeans)) +
        geom_histogram(colour = "white", fill = "blue") +
         ylab("Frequency") + xlab("Sample means") +
          theme_bw()

# Give it a gradient of colors for fun
gg2 <- ggplot(data.frame(umeans), aes(x=umeans, fill = ..count..)) +
        geom_histogram(colour = "white") +
         ylab("") + xlab("Sample means") +
          scale_fill_gradient(low = "green", high = "blue") + theme_bw() +
           theme(legend.position="none")

ggpubr::ggarrange(gg1, gg2, ncol = 2, labels = "auto")
```

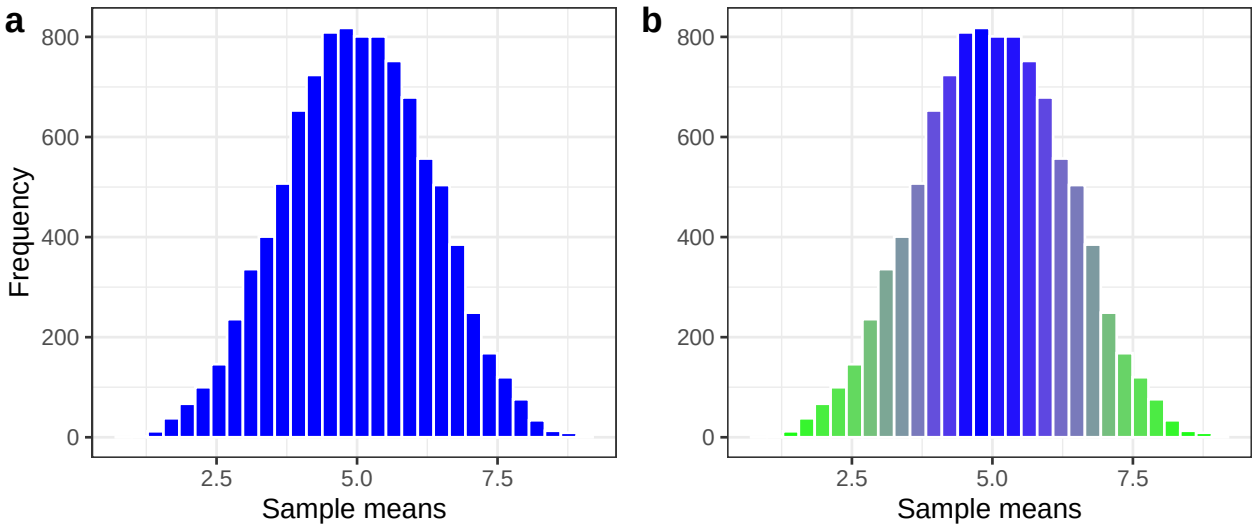

Figure 1: Histogram of 10000 sample means from $X \sim U(0, 10)$. Plots are the same, just colored differently.

We will use this example below, but for now you should pay attention to a few things. First, notice that we had to create a vector, `umeans`, to store stuff prior to running the loop. Second, the number of loops is determined by the `for` command. In this example, it runs from 1 to 10,000: i starts at 1 and counts up to 10,000 as the loops are completed. The start and end numbers can be defined by either a number or a function that returns a number (e.g. for `(i in 1:length(umeans)){...}`). Third, note that the new vector has to be subscripted with square brackets, `umeans[i]`, which defines each of the i values of means.

If the above is not crystal clear to you, then re-run the code so that the loop runs from 1 to 10. Make sure you know what each step of the code does.

To Do

Create a for-loop to take the square root of every element of a vector from 1 to 20.

Testing for Normality

Now let's look at a few ways to assess whether data are normally distributed. Let's start with the data we just created, and then look at a more typical dataset. By the way, the histogram of means from above looks like a normal distribution of means even though the data came from a uniform distribution. But is it?

One way to assess normality is to draw a normal distribution with the same parameters on top of the histogram. A couple of things to notice here. First, `umeans` is a numerical vector, but `ggplot()` requires that the data be in the form of a dataframe. This is why the above plots the following plot include `data.frame(umeans)` for the data. An alternative would be to define a dataframe, `um <- data.frame(umeans)` that is used in the `ggplot()` functions. Second, our previous histogram graphed the "Frequency" of observations on the y-axis, which ranged from 0 to 1500. If we fit a probability density curve on top of it, the curve will look like a flat line because the probability density function has an integral of 1.0 (the area under the curve). In other words, the scales of the histogram and the pdf would not correspond. Therefore, we need to tell `ggplot()` to express the histogram y-axis in terms of density.

```
ggplot(data.frame(umeans), aes(x=umeans)) + geom_histogram(aes(y = ..density..),
                                  colour="white", fill = "blue", alpha = 0.7) +
  geom_density(size = 1) + xlab("Sample means") + ylab("Density") +
   theme_bw()
```

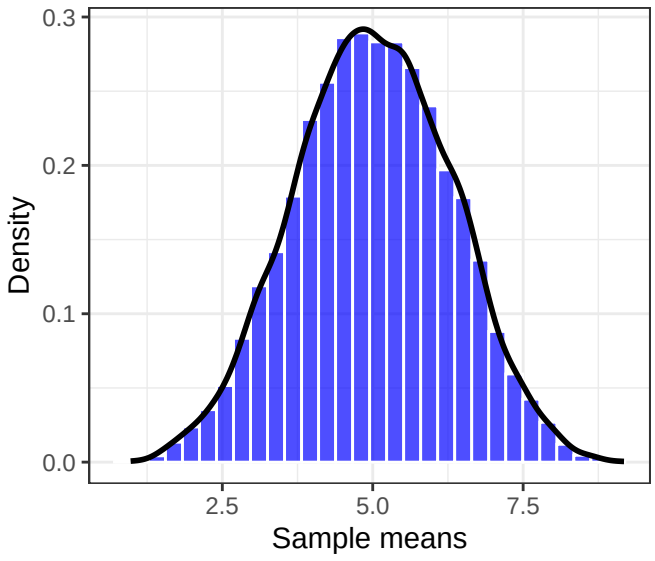

Figure 2: Histogram of the sampling distribution from 10,000 samples of a uniform distribution (as above) with a density curve overlaid.

This looks very nice, but is a rather subjective way to evaluate the normality of the data. A more objective method is to use q-q plots (quantile-quantile plots). The q-q plot is a graphical device used to check the validity of a distributional assumption for a data set. The basic idea is to compute the theoretically expected value for each data point. If the data indeed follow the assumed distribution, then the points on the q-q plot will fall approximately on a straight one-to-one line. The q-q plot provides a visual comparison of the sample quantiles to the corresponding theoretical quantiles.

This is quickly done in R, with `qqnorm` plotting the quantiles of our data and `qqline` fitting a line from a standard normal distribution as comparison.

```
qqnorm(umeans, las=1, main ="")
qqline(umeans)
```

Alternatively, in `ggplot()`:

```
p2 <- ggplot(data.frame(umeans), aes(sample=umeans)) +
  stat_qq() + stat_qq_line() +
   ylab("Sample quantiles") + xlab("Theoretical quantiles") + theme_bw()

plot_grid(p1, p2, labels = 'auto')
```

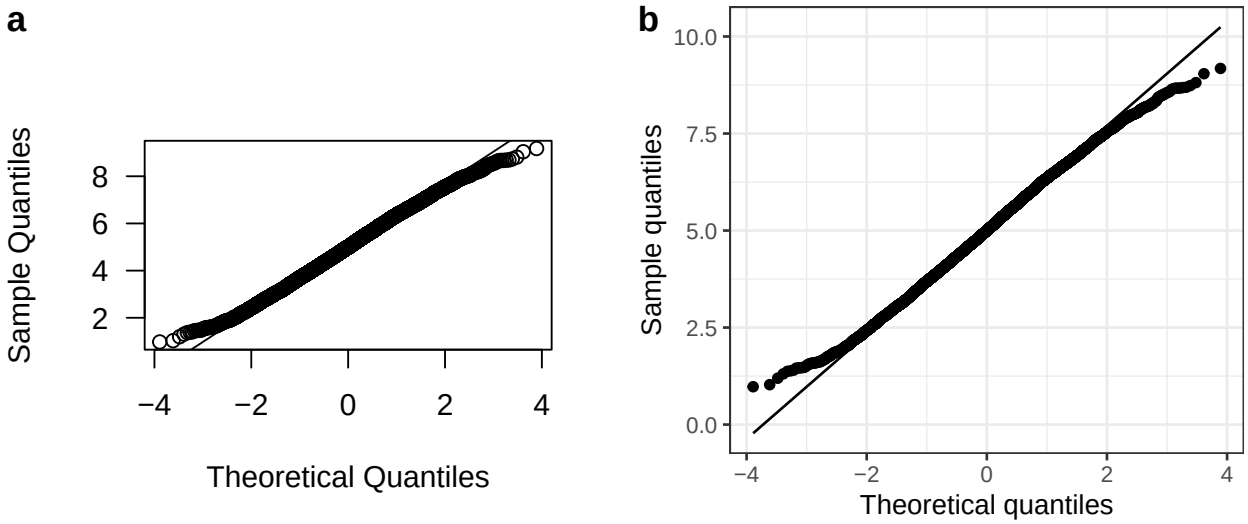

Figure 3: Same qqplots made with (a) `qqnorm()` and (b) `ggplot()`.

What is R doing to compute these lines? Let's dig in deeper using the Africa Plots dataset (`AfrPlots.csv`). As a reminder, the `AfrPlots` database consists of data from 30 1-ha forest plots. The diameters of the trees in the plots were measured once in 2005 (Census 1) and again in 2009 (Census 2). The database includes information on the aboveground biomass of each plot, `ChaveMoist`, the number of trees, `Trees`, the number of dead trees, `Dead`, and the number of new trees from Census 1 to Census 2, `Recruits`.

Start by making a histogram of the dead trees, removing the data from the first census, `CensusNo`, so that you are only using data from the second census period (`CensusNo == 2`). Does it look like these data are normally distributed? Now write your own q-q plot function, `myqqplot`, by completing the following steps. (This is Problem 1 below.)

1. *Sort the data from the lowest to highest.* Hint: `sort(Dead)`
2. *Convert observations to percentiles.* Let n be the number of observations. The lowest observation, denoted as $x(1)$, is the (1/n)th quantile of the data. A quantile times 100 is the percentile, so $x(1)$ is also the (1/n) x 100 percentile of the data. With this convention, however, the largest observation becomes

the 100 percentile of the data, which presents a problem as the 100 percentile of a normal distribution is infinity. To avoid this problem, compute `(i - 0.5)/n` to define the ith largest observation, $x(i)$, as the `(i-0.5)/(n)`th. Hint: `p <- (1:n - 0.5)/n`

3. *Calculate the expected percentiles.* The next step is to determine for each observation the corresponding quantile of the standard normal distribution, using `qnorm()`. This value is the "expected" quantile if the data come from a normal distribution. In other words, $x(i)$ should be close to $x'(i)$ if the distribution follows a normal distribution. Hint: `qZ <- qnorm(p, mean = 0, sd = 1)`
4. *Plot oberserved percentiles vs. expected percentiles.* Plot the values (ordered data, z-score) with the measurement scale (observations converted to percentiles) along the y-axis and the z-scale along the x-axis. This is a normal probability plot – a scatter plot of the data vs. the expected quantiles.

Does your q-q plot look the same as that produced by `qqnorm`?

If you are really ambitious, add a line to the q-q plot like that done by `qqline`. To do so, figure out the slope, m, of the line and, b, the y-intercept, and then use `geom_abline()` or `abline()` to draw the line. The line should pass through the 25% and 75% quantiles.

Understanding Confidence Intervals

There are two types of estimates for each population parameter: a point estimate and confidence interval (CI) estimates. We first compute the point estimate from a sample. Recall that a sample mean is an unbiased estimate of the corresponding population mean. The confidence interval is a range of likely values for the population parameter based on:

- the point estimate, e.g., the sample mean;
- the desired level of confidence (most commonly 95%, but any level between 0-100% can be selected);
- and the sampling variability, or the standard error of the point estimate.

Strictly speaking a 95% confidence interval means that if we were to take 100 different samples and compute a 95% confidence interval for each sample, then approximately 95 of the 100 confidence intervals would contain the true mean value, μ. Let's see this in action. The below code simulates 100 samples of 15 numbers from a normal distribution with a mean of 0. It then plots the CIs, highlighting those that do not include the true mean. This code could be easily turned into a function, but try running this code multiple times to see what happens such as the number of CIs that do not capture the population mean.

```
# 1st, set the parameters & variables

mu <- 0
sd <- 1
n <- 15
runs <- 100
ci <- matrix(nrow=runs, ncol=4)

# 2nd, sample from a normal distribution, calculate the mean and CI and add
# columns for the run number and whether the CI overlaps the mean or not

 for (i in 1:runs){
     samp <- rnorm(n=n, mu, sd)
          sx <- mean(samp)
     conf <- qnorm(p = 0.975, mean = 0, sd = 1)

     ci[i,1:3] <- c(sx-conf*(sd(samp)/sqrt(n)), sx+conf*(sd(samp)/sqrt(n)),
                    i)
     ci[i,4] <- ifelse(ci[i,1] > mu | ci[i,2] < mu, 2, 1)}

# 3rd, turn matrix into data frame and define the colors, x-axis limits, and position
```

```
# for the count of CIs not capturing the population mean

ci_df <- data.frame(ci)
fun_cols <- c("black", "red")
xlms <- c(mu - 2*sd, mu + 2*sd)
ypos <- max(ci_df$X3)

# 4th, make the graph of CIs with a vertical line at the population mean
# The CIs are colored red if they do not capture the population (overlap with 0)
# A lot of the code here makes the graph pretty without a y-axis, etc.

gg_ci <- ggplot(data = ci_df, aes(x=x)) +
       geom_segment(aes(x=X1, y = X3, xend = X2, yend = X3,
                        colour = fun_cols[X4])) + scale_colour_identity() +
        geom_vline(xintercept = mu, linetype = "dashed") +
         xlab("CI's of samples") + ylab("") + xlim(xlms) +
         theme_bw() +
         theme(legend.position="none",
               axis.text.y = element_blank(),
               axis.ticks.y = element_blank(),
               panel.border = element_blank(),
                axis.line.x = element_line(linewidth = 0.5,
                              linetype = "solid", colour = "black"),
               panel.grid.major = element_blank(),
               panel.grid.minor = element_blank())

# 5th, add text showing how many CIs didn't capture the population mean
cnt <- length(ci_df$X4[ci_df$X4 > 1])
gg_ci + annotate("text", x = max(xlms)*0.92, y=ypos*0.9,
                 label = paste0("Count = \n", cnt))
```

In practice, however, we select one random sample and generate one confidence interval, which may or may not contain the true mean. The observed interval may over- or underestimate μ. Another way of thinking about a confidence interval is that it is the range of likely values of the parameter with a specified level of confidence (which is similar, but not the same as a probability).

As you see in the code above, a confidence interval for the mean is calculated as the mean $\pm$ the margin of error, where the margin of error is the critical value, z_c, multiplied by the sample standard deviation. For a 95% confidence interval, the area in each tail is equal to $p = \frac{(1-C)}{2} = 0.05/2 = 0.025$. z_c represents the point(s) on the standard normal density curve such that the probability of observing a value greater or equal to z is equal to p. For example, if $p = 0.025$, we need to find the value z_c such that $P(Z \geq z_c) = 0.025$ and $P(Z \leq z_c) = 0.025$. We can do this with:

```
qnorm(p = 0.025, mean = 0, sd = 1)
## [1] -1.959964
```

Note that this gives you a negative value for z_c. It is easier to work with a positive value, and therefore the statement for the critical values is written to find the upper z_c, i.e., `qnorm(p = 0.975, mean = 0, sd = 1)`.

For a population with unknown mean, μ, and unknown standard deviation, σ, a confidence interval for the population mean, based on a simple random sample of size n, is $\overline{x} \pm z_c \cdot \frac{s}{\sqrt{n}}$. This code below calculates the 95% CI for a normally distributed sample ($\bar{x} = 5$, $s = 2$, $n = 50$):

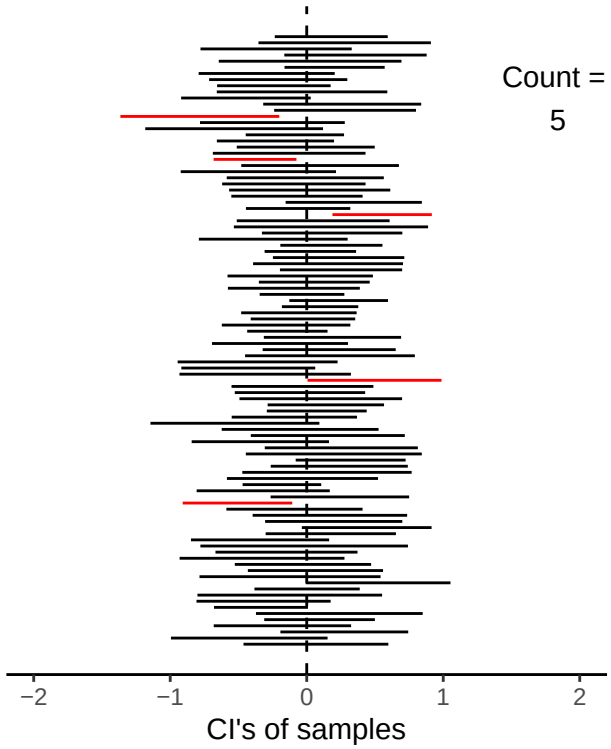

Figure 4: Simulation of 95% CIs from 100 random samples of 15 observations from a standard normal distribution. Red bars do not overlap 0 and failed to capture the population mean.

```
set.seed(999)
 sample_x <- rnorm(n = 50, mean = 5, sd = 2)
  mean_x <- mean(sample_x)
   z_c <- qnorm(0.975, mean = 0, sd = 1)
    low_x <- mean_x - z_c*sd(sample_x)/sqrt(length(sample_x))
    up_x <- mean_x + z_c*sd(sample_x)/sqrt(length(sample_x))

data.frame(mean = mean_x, LCI = low_x, UCI = up_x)
##       mean      LCI      UCI
## 1 4.484663 3.935983 5.033342
```

To Do

Calculate the 95% confidence interval for the brain measurements in the datafile, `mammals.csv`.

Problem 1

Turn in the code for your q-q plot function, `myqqplot`. Include the following: (a) a graph of the q-q plot from the `Dead` data (from `AfrPlots`) using `myqqplot`, (b) a graph of the q-q plot from the `Death` data using `qqnorm` and `qqline`. Write a short paragraph describing how q-q plots are created in your own words.

Problem 2

Using the datafile, *epa2012.csv*, evaluate the data for highway gas mileage. Assess whether highway gas mileage is normally distributed, doing the following: (a) plot a histogram of the data with a density curve reflecting the expected normal distribution of the data, (b) plot the q-q plot with the one-to-one line, (c) create a boxplot of the data, and (d) calculate the skewness and kurtosis of the data. Do this for both the raw data and log-transformed data (e.g. `log(data)`). Which looks more normally distributed? Calculate the mean and confidence interval of the dataset that you think is mostly likely normally distributed.