STAT 509 2017 Summer HW12
Instructor: Shiwen Shen
Lecture Day: June 1

---

1. A textile fiber manufacturer is investigating a new drapery yarn, which the company claims has a mean thread elongation of 12 kilograms with a standard deviation of 0.5 kilograms. The company wishes to test the hypothesis

$$H_0 : \mu = 12$$
$$H_a : \mu \neq 12$$

using a random sample of four specimens. Suppose the random sample is from a normal population. *(Hint: notice in this question, the population variance is assumed to be known with $\sigma = 0.5$)*

(a) Given the sample mean $\overline{y}$ is 11.3 and the confidence level is 95%, follow the 4-step procedure to conduct a hypothesis test. What is your conclusion?

(b) Using the confidence interval approach to calculate a 95% two-sided confidence interval for μ. Does the confidence interval cover 12? Is the results of confidence interval consistent to the testing conclusion?

(c) What is margin of error and the length of interval in (b)? If we want to control the length of the confidence interval to be 0.6, how many observations do we need in the sample?

2. A manufacturing firm is interested in the mean batteries hours used in their electronic games. To investigate mean batteries life in hours, say μ. The following data are collected

```
20,25,21,28,21,30,23,27,26,26,28,31,26,32,33,35
```

(Hint: the population variance is not given, therefore we assume it is not known)

(a) Is it reasonable to assume that the sample data has come from a normal distribution? The R code is given below (*Hint: use fat pencil test in R.*)

```
battery<-c(20,25,21,28,21,30,23,27,26,26,28,31,26,32,33,35)
qqnorm(battery)
qqline(battery)
```

(b) Suppose it is reasonable to assume the data has come from a normal distribution, construct a 99% two-sided confidence interval for μ. The quantile can be found via R or t-table. The sample mean and the sample standard deviation can be computed via the following command:

```
mean(battery)
sd(battery)
```

(c) **Using R** to test the following hypothesis with the level of significance $\alpha = 0.01$:

$$H_0 : \mu = 24$$
$$H_a : \mu \neq 24$$

You need to print out both your R code and testing results.

3. Inexperienced data analysts often erroneously place too much faith in qq plots when assessing whether a distribution adequately represents a data set (especially when the sample size is small). The purpose of this problem is to illustrate to you the dangers that can arise. In this problem, you will use R to simulate the process of drawing repeated random samples from a given population distribution and then creating normal probability plots (Q-Q plots). Follow the code provided

   (a) Generate your own data and create a qq plot for each sample using this R code:

   ```
   # create 2 by 2 figure
   par(mfrow = c(2,2))
   B = 4
   n = 10
   # create matrix to hold all data
   data = matrix(round(rnorm(n*B,0,1),4), nrow = B, ncol = n)
   # this creates a qq plot for each sample of data
   for (i in 1:B){
      qqnorm(data[i,],pch=16,main="")
      qqline(data[i,])
   }
   ```

   mark the qq plot that appears to violate the normal assumption the most. Note: In theory, all of these plots should display perfect linearity! Why? Because we are generating the data from a normal distribution! **Therefore, even when we create normal qq plots with normally distributed data, we can get plots that don't look perfectly linear.** This is a byproduct of sampling variability. This is why you don't want to rush to discount a distribution as being plausible based on a single plot, especially when the sample size n is small (e.g. $n = 10$).

   (b) Increase your sample size to $n = 100$ and repeat. What happens? What if $n = 1000$? Just change n in the R code on the last page and re-run.

   (c) Take $n = 100$, replace

   ```
   data = matrix(round(rnorm(n*B,0,1),4), nrow = B, ncol = n)
   ```

   with

   ```
   data = matrix(round(rexp(n*B,1),4), nrow = B, ncol = n)
   ```

   and re-run. By doing this, you are changing the underlying population distribution from $\mathcal{N}(0,1)$ to exponential(1). What do these normal qq plots look like? Are you surprised?

4. Two machines are used for filling plastic bottles with a net volume of 16.0 ounces. The fill volume can be assumed to be normal with standard deviation $\sigma_1 = 0.020$ and $\sigma_2 = 0.025$ ounces. A member of the quality engineering staff suspects that both machines fill to the same mean net volume, whether or not this volume is 16.0 ounces. A random sample of 10 bottles is taken from the output of each machine.

   ```
   Machine 1:  16.03, 16.04, 16.05, 16.05, 16.02, 16.01, 15.96, 15.98, 16.02, 15.99
   Machine 2:  16.02, 15.97, 15.96, 16.01, 15.99, 16.03, 16.04, 16.02, 16.01, 16.00
   ```

   (a) Do you think the engineer is correct? Conduct a formal 4-step procedure with $\alpha = 0.05$. What is your conclusion? *(Hint: Sample means can be computed using R.)*

   (b) Calculate a 95% confidence interval on the difference in population means. Provide a practical interpretation of this interval.

5. Data on pH for 16 random batches of low and high volt electrolyte were collected. The data are given by

```
Low volt:  7.78, 5.77, 7.08, 6.75, 7.09, 8.27, 6.5, 5.16, 6.81, 7.28, 7.88, 7.87,
7.2, 5.95, 6.58, 6.99
high volt:  4.54, 5.04, 5.07, 6.18, 8.62, 6.28, 7.41, 6.17, 6.25, 4.25, 6.08, 7.23,
4.68, 6.19, 5.85, 5.83
```

(a) Population variances σ_1^2 and σ_2^2 are unknown. Do you think it is reasonable to assume $\sigma_1^2 = \sigma_2^2$? Let's figure it out! First, draw a side-by-side boxplot in R. Based on the boxplot, do you believe $\sigma_1^2 = \sigma_2^2$?

```
low <- c(7.78,5.77,7.08,6.75,7.09,8.27,6.5,5.16,6.81,7.28,7.88,7.87,7.2,
         5.95,6.58,6.99)
high <- c(4.54,5.04,5.07,6.18,8.62,6.28,7.41,6.17,6.25,4.25,6.08,7.23,
          4.68,6.19,5.85,5.83)
boxplot(low,high,names=c("low","high"),col="grey")
```

(b) Now, let's conduct a formal test of

$$H_0 : \sigma_1^2/\sigma_2^2 = 1$$
$$H_a : \sigma_1^2/\sigma_2^2 \neq 1$$

using the following R code.

```
var.test(low, high)
```

What is the **p-value** of the testing result? With significance level 0.05, do you reject H_0 or fail to reject H_0? Is the result consist to the one you get from (a)?

(c) Assuming the two samples are independent. The engineer want to test that the low volt average pH is greater than the high volt average pH. Let μ_L be the average pH of low volt electrolyte and μ_H be the average pH of high volt electrolyte. State the null and alternative hypotheses.

(d) Calculate the appropriate test statistic for the test. The sample means and sample variances can be computed using R.

(e) Use R to calculate the p-value of the test. *(Hint: $P(T_{n_1+n_2-2} > t_0)$ can be calculated using R with command* `1 - pt(`t_0`, `$n_1 + n_2 - 2$`))`*.*

(f) Make decision and state your conclusion at a 0.05 level of significance.

(g) Use `t.test` in R to check your work.

```
t.test(low,high,alternative="greater",paired = FALSE, var.equal = TRUE)
```

(h) Construct a two-sided 95% confidence interval of $\mu_L - \mu_H$ by hand. Provide a practical interpretation of this interval.