Exploring Confidence Intervals and Simple Hypothesis Testing

ACTIVITY 35

A confidence interval describes the amount of uncertainty associated with a sample estimate of a population parameter. 95% confidence level means that 95% of the interval estimates are expected to include the population parameter. A simple hypothesis testing is an assumption which specifies the population distribution completely. It examines a random sample from a population to see if the sample data are consistent with statistical hypothesis. If not, the null hypothesis is rejected. This activity focuses on the steps of processing confidence intervals and performing statistical hypothesis testing in R.

Purpose

The purpose of this activity is give you the opportunity to learn how to compute confidence intervals and perform simple hypothesis testing in R

Learning Objective

Learn statistical function in R and diagnose the difference of two sample datasets in plots.

Required Resources

- R, R Studio
- Sample script file from class web site

Time Allocated

90 minutes in class
40 minutes out-of-class

Tasks

This activity begins assuming R Studio is open.
If not, please start R Studio and open the sample script for the activity (if provided).

Browse the metadata for the bike performance data to familiarize yourself with the dataset. Use phpPgadmin to browse through *bicycle.performance* table. Connect to database via RODBC and extract the full file that will be using for data analysis set later. Extract the average speed `'set1'` and `'set2'` of this data set grouped by levels of grade (0 = no grade; 1 = grade).

Ultimately we are going to compare the means and attempt to answer the following question: Does grade have an effect on average observed speed? And if so, what is it?

A. Diagnostic Plots

In my experience, the best thing to do before running any statistical tests or computing confidence intervals is to do summary diagnostic plots. In this exercise, four plots in different distribution functions and two statistical measures will be applied for making contrast and comparison of *set1* and *set2*.

Fortunately you now have a lot of easy tools you can use. Follow along in the sample R script provided for the today's activity. A helpful diagnostic function to produce summaries is:

```
summary(set1,na.rm=TRUE)
```

where `na.rm=TRUE` is for missing data. If `na.rm=TRUE` then missing values are removed before computation proceeds. But this doesn't tell the whole story.

1. Review the plots which follow (Figures 57 and 58) and make a short note describing your interpretation of the plots, including your "hunch" about the effect of grade on average speed.

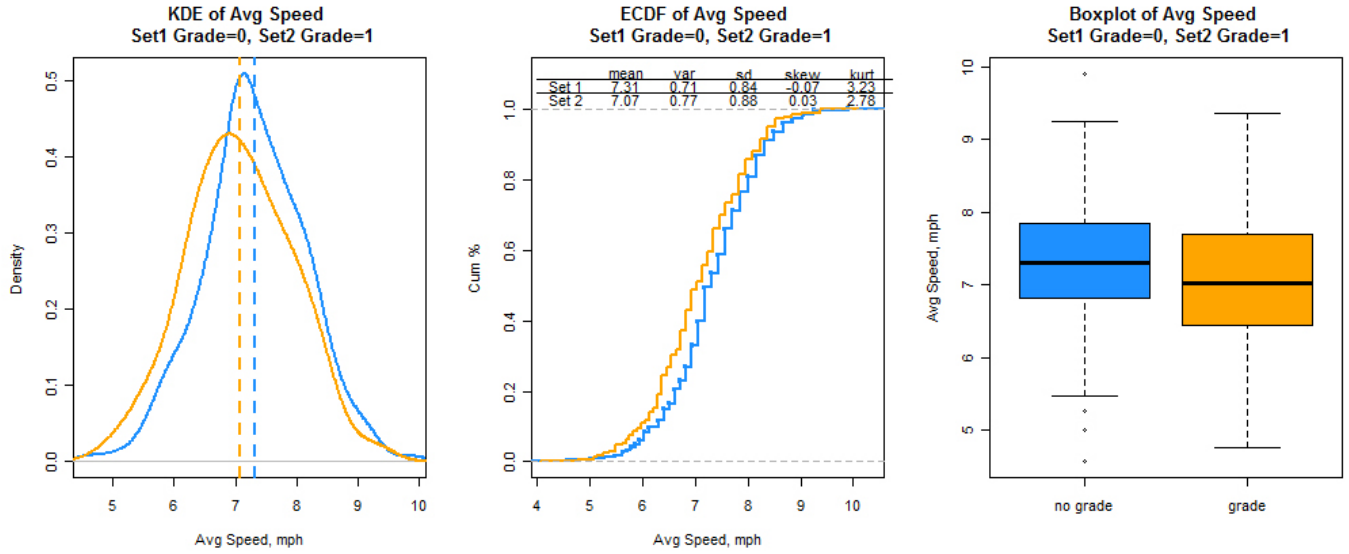

Figure 57 Summary Plots Comparing the Distributions

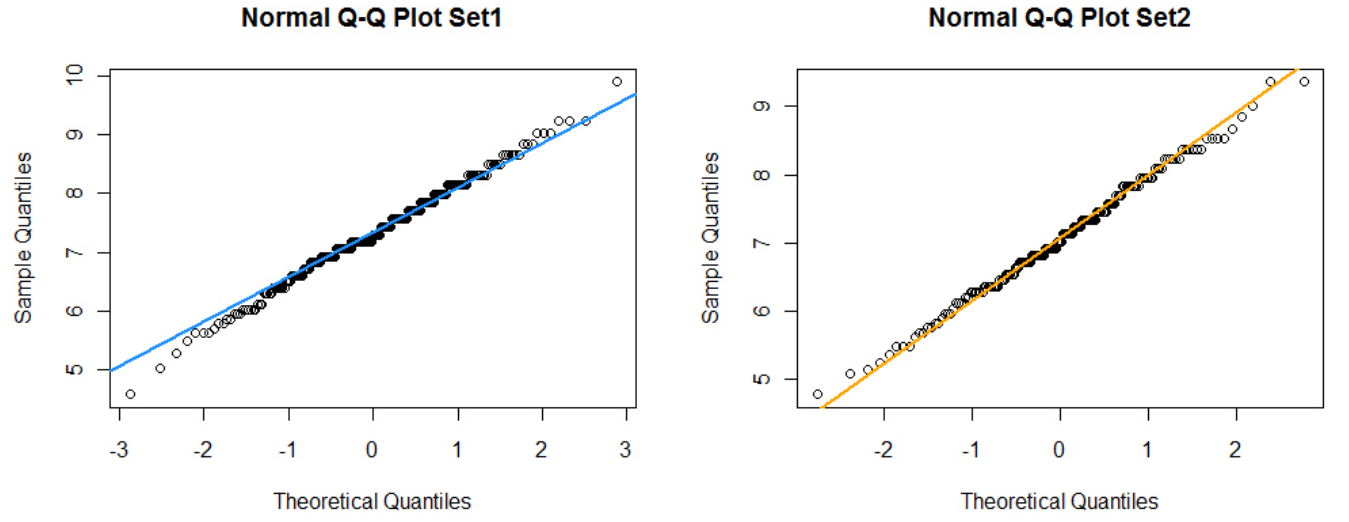

Figure 58 Q-Q Plots of the Distributions

B. Calculate Confidence Intervals Around the Mean

When average speed data are assumed in normal distribution, quantiles of normal distribution, the sample mean and standard error of the mean can be used to calculate approximate confidence intervals for the mean.

`qt` quantile function for t distribution with `df` degrees of freedom;

`qnorm` quantile function for the standard normal distribution (Z), with mean=0 and sd=1

We can do either the Z test or *t* distributions to determine our confidence interval of the mean. For samples this large, it makes little difference (as you will see later). But, let's follow Dalgaard and use the student's *t*.

The standard error of the mean is equal to standard deviation divided by square root of sample size.

$$SEM = (Standard\ deviation) / \sqrt{n}$$

To estimate the two-tailed confidence interval:

$$Upper\ confidence\ limit = \bar{x} + \left(SEM * t_{\alpha/2} \right)$$

$$Lower\ confidence\ limit = \bar{x} - \left(SEM * t_{\alpha/2} \right)$$

Where x bar is the mean and α is the chosen confidence interval by the analyst (you!). To determine the t statistic we also need the degrees of freedom, which here is equal to (n − 1).

Certainly the most common confidence level is the 95th percentile. By selecting this we are estimating a confidence interval for the mean that includes with 95% probability the true mean. We are accepting a small probability (5%) that true mean is not in this interval. Since the distribution is symmetrical, this 5% can be on either tail (as seen in the figure below). For the 95th percentile, α = 0.05 and α/2=0.025. Thus we need the cumulative probability that corresponds to the value of 1.00 − 0.025 = 0.975. To return $t_{0.975}$ for the degrees of freedom then use the `quantile()` function to return the value that captures 97.5% of the cumulative probability. Use this R syntax;

```
qt(0.975, n-1)
```

The sample R script generates the probability density function (PDF) and cumulative density function (CDF) of the t-distribution and shows the quantile of the t-distribution that corresponds to the given alpha. The figures below show that for set 1 with 249 degrees of freedom, 95% of the cumulative probability is between −1.969576 and 1.969576 (area under the curve in the PDF) and there is a 5% chance (2.5% on each tail) that the true mean is outside these bounds. This is shown in the CDF as 97.5% of the cumulative probability (since we are excluding the left tail we add 95% + 2.5% = 97.5%.

Change the alpha level in the script and notice what happens to the location of the t-statistic and the tails of the distribution. Make a connection between the alpha level and results shown in these 2 plots (see Figure 59).

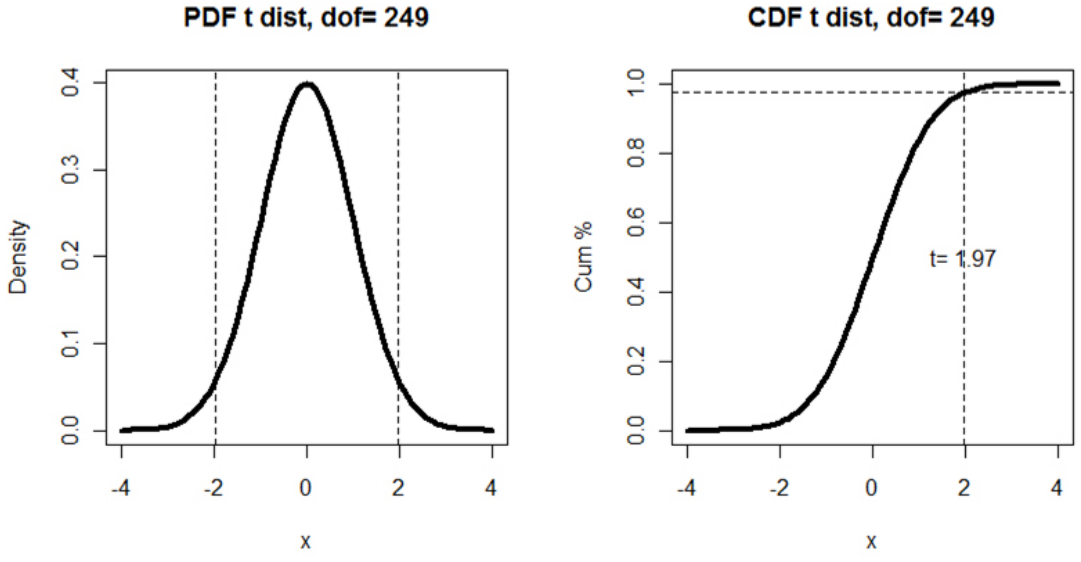

Figure 59 Plots of PDF and CDF for t-distribution with dof=249

Now, we can calculate the upper and lower confidence intervals with:

```
# use t statistic
stat1 <- qt(1-alpha/2, n-1)  # use t for set 1
stat2 <- qt(1-alpha/2, n2-1) # use t for set 2
#Calculate the bars
error <- stat1*stddev/sqrt(n)
lower <- mu-error
upper <- mu+error
error2 <- stat2*stddev2/sqrt(n2)
lower2 <- mu2-error2
upper2 <- mu2+error2
```

These are shown graphically in Figure 60: *set1* statistics are shown in blue; *set2* statistics are in orange. "+" is the upper bound of confidence interval; "-"is the lower bound of confidence interval; dot is the mean.

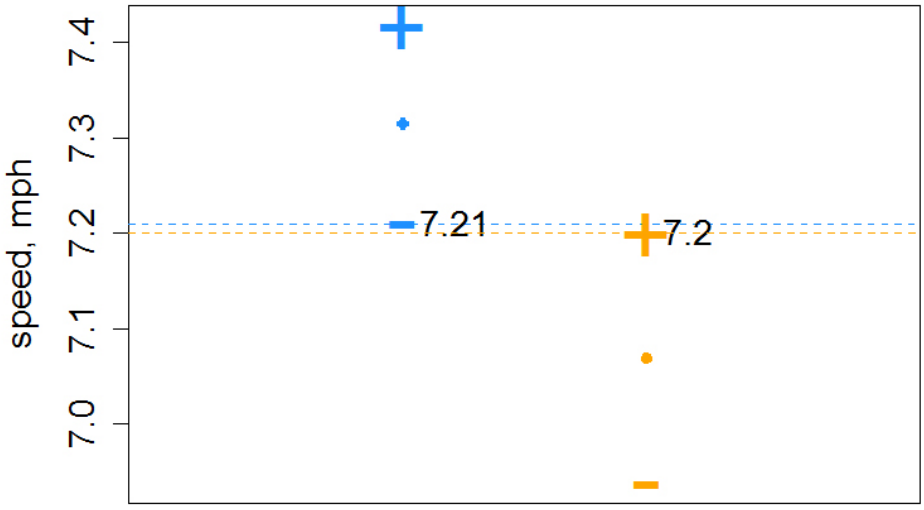

Figure 60 Plot of Mean and 95th Percentile Confidence Interval

2. Now, make 3 similar plots assuming a 85, 90, 95 and 97% confidence intervals and plot them in a `c(1,4)` arrangement. Below these 4 plots, include the CDF distribution plot as shown in Figure 60. Write a show description of your interpretation. Can you explain the differences in the confidence intervals that you see in the plots?

C. t-test

Of course the next step is to test whether there is a difference between the mean values of these sets of speed data. The test requires 2 assumptions:

- that the data being tested are normally distributed and
- the variance of the sample data being tested are equal.

The t-test is robust to some departures in the assumption 1 and there is a version of the test if you are not willing to assume that the variances are equal. First, comparing the Q-Q plots reveals that these data are reasonably normally distributed.

The variance assumption is more difficult to establish, but we can use the F test for testing underlying assumption of homogeneity of variances. Let's take *set1* and *set2* as an example to test equality of variance.

```
> var.test (set1, set2)
   F test to compare two variances
data:  set1 and set2
F = 0.9193, num df = 248, denom df = 172, p-value = 0.5429
alternative hypothesis: true ratio of variances is not equal to 1
95 percent confidence interval:
 0.6952998 1.2071398
sample estimates:
ratio of variances
         0.9193456
```

The alternative hypothesis is that the true ratio of variances is not equal to 1. The p-value of the test is 0.5429. Since this is greater than our selected alpha of 0.05, there is insufficient evidence to reject the null hypothesis that the ratios of the variance are equal to 1. Thus we conclude that the variances are equal. Alternatively, we could run the t-test with the Welch correction (see Dalgaard).

Now, let's perform the t test. The output is printed below for both variance assumptions:

```
> t.test (set2, set1, var.equal=TRUE)
   Two Sample t-test
data:  set2 and set1
t = -2.8951, df = 420, p-value = 0.003988
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -0.41188434 -0.07876422
sample estimates:
mean of x mean of y
 7.069296  7.314620
> t.test (set2, set1, var.equal=FALSE)
   Welch Two Sample t-test
data:  set2 and set1
t = -2.8733, df = 360.028, p-value = 0.004303
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -0.41323353 -0.07741503
sample estimates:
mean of x mean of y
 7.069296  7.314620
```

Both tests have very small p-values, leading us to reject the null hypothesis that the means are equal and accept the alternative hypothesis "`true difference in means is not equal to 0`". You can check your interpretation by inspecting the 95% confidence interval of the difference between the means. Note that this does not include zero.

2. Now, subset the data by MALE and FEMALE and conduct a t-test of the difference in means.

D. When does the t distribution looks like Z distribution

You often see testing for equality of means or the calculation of confidence intervals based on either distribution. The Z distribution is the "standard normal" that we have seen so much already. Use of the Z distribution in hypothesis testing and confidence interval calculations requires a large sample size and a known ***population*** standard deviation. Since we are nearly always estimating the population standard deviation from the ***sample***, this second assumption introduces additional uncertainty. The t-distribution has larger tails and helps account for this uncertainty especially with small sample sizes. But what is a small sample? A rule of thumb is that a sample size larger than 20 is sufficiently large so that use of the Z distribution is acceptable. Can we confirm this rule of thumb in R? The parameter of the t-distribution is the degrees of freedom. A loop was set up for plotting t distribution under different degree of freedoms. The standard normal distribution is shown in heavy black line.

```
for (i in 2:length(dof)) {
t <-  pt (x, dof[i])
lines (x, z-t, col=colorvector[i])
}
```

We see that the higher degree of freedom, t distribution is more closed to z distribution as shown in Figure 61. A simple subtraction of the densities is shown in the right panel, confirming that as the degrees of freedom approaches 20, the t and Z distributions are converging.

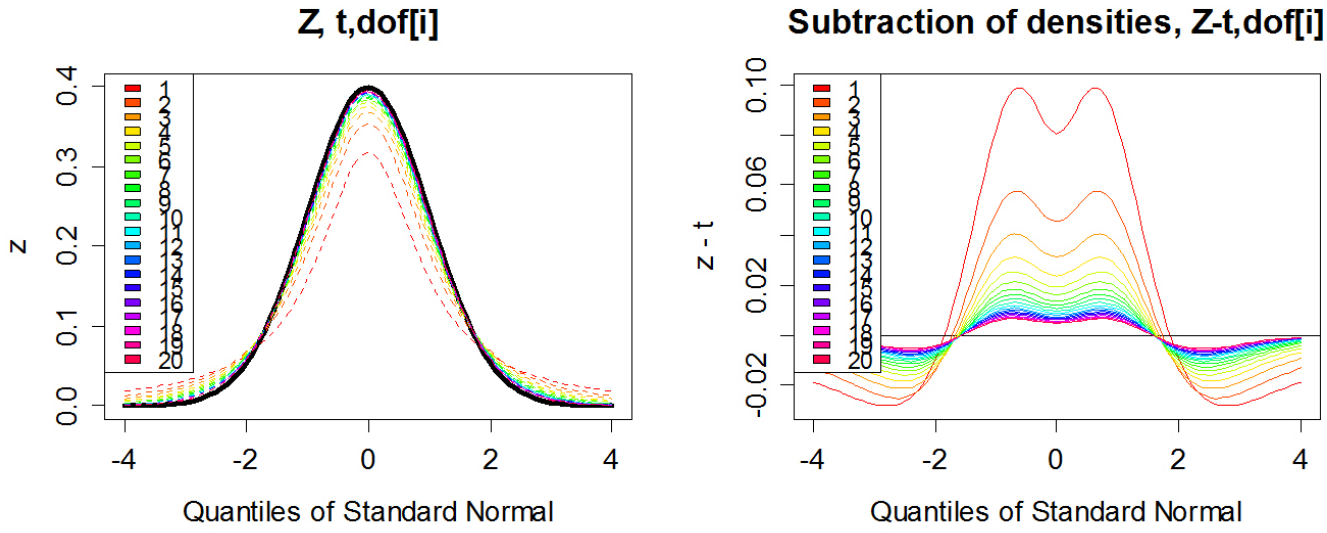

Figure 61 Z and t distributions with varying degrees of freedom

Now, should I be confused? The purpose of this is to show you the convergence of the 2 common distributions. The best advice is to use the t-distribution approaches whenever possible (note R does not have a Z-test function in the standard build) since it is more conservative. Also, when interpreting the outputs of linear regression or other outputs, pay close attention to whether the parameters are estimated by either Z or t distributions.

E. Wilcoxson Non-parametric Test

There are tests that do not require the assumption of distribution to be applied. These are called non-parametric, since the underlying distribution is not specified.

3. Use Dalgaard as a reference and run the Wilcoxson test of means on the male/female and grade/no grade data sets.

Deliverable

Summarize the results of all the numbered questions in this activity. If instructor assigns "BONUS" question in the script file, complete as well. Submit an electronic PDF of your write up to the dropbox folder on the course management system.

Assessment

This activity is a short response activity. The score that you will receive will be based on the quality and depth of discussion. The response expected differs by question as described in the following rubric:

Activity 35 Grading Rubric

	Excellent (10)	Good (8)	Poor (6)	NONE
Discussion	Insightful discussion or commentary relating to the question at hand demonstrating student understanding of the task.	Discussion was competent regarding the lessons, but lacked in insightful discussion.	The discussion did not address the lessons or applicability of the activity.	Did not submit
Graphic(s)	Graphic correctly demonstrated the distribution of the data set.	Graphic did not accurately demonstrated the distribution of the data.	No graphics were present in the discussion.	Did not submit
Quality	Document is typed, formatted, contains appropriate grammar and language.	Document has minor grammatical errors or inappropriate language.	Document was unorganized, contained inappropriate language and/or grammatical errors.	Did not submit

Student Notes