Week 3: Inferential Statistics

Inferential statistics is the process of making conclusions regarding to the population from a given sample. In other words, a sample from a population is investigated and then the inference is made about the population. Inference statistics can be used to examine data for their differences, relationships and associations. The variable types are used to determine what inference statistics method could be employed. There are two types of data: non-parametric (nominal and ordinal data) and parametric (interval and ratio data).

The parametric test assumes that the data is normality (come from the normal distribution population), whereas there is no assumption regarding the distribution for non-parametric tests. Nevertheless, random sampling is the requirement for both types of tests.

Here is the review of the important terms for analyzing the results from hypothesis testing
Significance level (α) or type I error is the probability of rejecting the null hypothesis when the null hypothesis is true. In simple word, this is the probability of making the wrong decision.

P-value is the probability of obtaining your sample data if the null hypothesis was true. The higher the P-value, it is more likely that the null hypothesis is true. On the contrary, if the P-value is very low, there is a small chance that the null hypothesis was actually true.

Confidence Interval is the likely range of the population parameter (e.g. mean)
Some important properties [Minitab Blog]:
1) Confidence level + alpha = 1
   For example, if the confidence level is 0.95, then the alpha is 0.05. If the confidence level is 0.99, then the alpha is 0.01.
2) Confidence interval and p-value will result in the same conclusion
   When p-value is less than alpha, the confidence interval will not contain the hypothesis mean. Whereas, when the p-value is greater than alpha, the confidence interval will include the hypothesized mean.
3) If the p-value is low, the null must go (reject the null hypothesis)
   When p-value is less than alpha, it is possible that you will make the wrong decision. Therefore, you reject the null hypothesis or "it is significant". For example, the alpha is set to 0.05 and the calculated p-value is 0.003, the null hypothesis is rejected. On the contrary, "It is not significant", if we cannot reject the null hypothesis.

The most basic and commonly used inference statistics tests are: one and two sample t-tests, Chi-Square tests, general linear model, regression analysis, ANOVA, and Bayesian, etc.
- ***One and two sample t-tests***
One sample t-test is used to examine whether the sample mean is different from the hypothesized mean (or an already known mean). Whereas, the two sample t-test is used to test whether the sample means of two groups are different. The two sample t-test can

be divided into a dependent sample t-test and an independent sample t-test. The required assumption is that the population distribution is normal or the sample size is large enough to apply the central limit theorem. T-tests use parametric data.

Different types of t-tests:

a) The one sample t-test is used to test a single sample group with a hypothetical mean

b) The dependent sample t-test (or paired sample t-test) is suitable to test the mean of two samples where they are **related**. For example, the samples taken from the same person twice. (e.g. compare blood pressure before and after taking a prescription drug on the same persons)

c) The independent sample t-test (or two-sample t-test) is used to examine whether the means of the population, where the samples drawn are the same. Two samples are assumed to be **unrelated** and the populations are approximately normal distribution.

Note that analysis of variance (ANOVA) is used when there are more than two groups.

a) One sample T-Test

An ice-cream company programs an automatic machine to fill the ice-cream containers to a volume of 56-ounces. It is suspected that the machine is under-filling. So, you collect 20 samples of ice-cream containers and measure the volume.

One sample t-test is used to determine whether the ice-cream containers are under filling. The null hypothesis is the mean of the ice-cream volume, which is equal to 56 oz. The alternative hypothesis is the mean of filling volume less than 56 oz. The significance level (alpha) is assumed to be 0.01.

One sample t-test is performed in R using the "*t.test*" function. Suppose that the volumes of 20 sample containers are:

```
> vol =
c(50.3,55.4,53.5,51.6,59.9,54.2,52.4,53.1,56.0,55.7,52.4,55.1,54.7,52.8,57.3,55.4,56.1,54.9,54.4,55.9)
>  t.test(vol, mu=56, alternative="less", conf.level=0.99)
```

Arguments:

mu denotes the sample mean

alternative can be set as "greater" or "less" for the one-tailed test. A two-tailed test is the default if an alternative is not specified. One-tailed tests are suitable if you have a directional hypothesis. Two-tailed tests are appropriate if you are not sure of the alternative hypothesis direction (either positive or negative direction).

conf.level is the confidence interval.

The outputs are:

One Sample t-test

data: vol
t = -3.0002, df = 19, p-value = 0.003679
alternative hypothesis: true mean is less than 56
99 percent confidence interval:
  -Inf 55.7781
sample estimates:
mean of x
  54.555

It can be seen that the mean of the sample ice-cream containers is 54.555 oz. It is likely that mean of the filling volume is less than 55.78 oz. with 99% confidence interval.
The p-value of 0.003679 tells us that the probability for obtaining the sample is about 0.3%, if the mean of the ice-cream volume is 56 oz. is true.
Since the p-value is less than the significance level of 0.01, we can reject the null hypothesis that the mean volume is 56 oz. This means that, the ice-cream containers are under filled.

As a reminder, the confidence level is the probability of a Type I error. It is the probability of rejecting H_0 when it is actually true. In this case, it is set to 0.01 level. The power of the test is the probability that the test reject H_0. Here, it is set to 0.99.

Power
Power is the chance that H_0 is rejected when H_0 is **false**. In other words, it is the correct conclusion that there is a treatment effect when there actually is a treatment effect.

A power function "***power.t.test***" is used to calculate the power of the t-test or determine parameters to obtain the specific power. Here, we assume that the mean of filling volume is overestimated by 5 oz. and the standard deviation of the population is 8.4.

> power.t.test(n=20, delta=5,sd=8.4,sig.level=0.05,type="one.sample",
alternative="one.sided")

#--- One-sample t test power calculation ---

        n = 20
      delta = 5
        sd = 8.4
   sig.level = 0.05
      power = 0.8213376
   alternative = one.sided
#--

The power of the T-Test is 0.82. If we want to find what is the sample size to give the 95% power, we can use the following command.

> power.t.test(delta=5,sd=8.4,power=0.95,type="one.sample",alternative="one.sided")

#--- One-sample t test power calculation ---

> *n = 31.95149*
> *delta = 5*
> *sd = 8.4*
> *sig.level = 0.05*
> *power = 0.95*
> *alternative = one.sided*

#--

The result tells us that a sample size of at least 32 is needed to reject H_0 under the alternative hypothesis H_a to have a power of 0.95.

b) Dependent sample t-test

Example: A new learning technique is proposed to use for elementary school students. The test scores before (scoreBF) and test scores after (scoreAF) using the new learning technique are collected from a group of 10 students (random sample) as follows:
> scoreBF=c(78,89,56,90,86,83,77,81,75,62)
> scoreAF=c(82,84,60,85,88,88,80,74,70,60)
> t.test(scoreBF,scoreAF,**paired=TRUE**)

 # --- Paired t-test ---

data: scoreBF and scoreAF
t = 0.4082, df = 9, p-value = 0.6926
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
-2.724678 3.924678
sample estimates:
mean of the differences
> *0.6*

#-----------------------

Here, p-value is not less than alpha (0.05) so we cannot reject the hypothesis. This means that there is no difference in the testing score before and after using the new learning technique.

The power analysis function (power.t.test) can be applied for both dependent and independent sample t-tests.

c) Independent sample t-test (two-sample t-test)

For example: To study the effectiveness of the Z vaccine, 10 subjects are given the Z vaccine (treatment group) and another 10 subjects are given placebo (control group). The means of the treatment and control groups are compared using a two-sample t-test. Suppose that μ_1 is the mean of the population getting Z vaccine and μ_2 is the means of the population getting placebo.

$H_0: \mu_1 - \mu_2 = 0$
$H_a: \mu_1 - \mu_2 < 0$

```
> control <- c(45,67,72,59,64,81,57,60,49,61)
> trmt <- c(87,78,91,85,74,69,71,83,66,77)
> t.test(control,trmt, alternative="less",var.equal=TRUE) # for equal standard deviation
> t.test(control,trmt, alternative="less") # for not equal standard deviation
```

• *Chi-Square Tests*

The Chi-Square test is used as a frequencies test for percentages and proportions. The variable types should be nominal data. There are two-types of Chi-Square: test of independence and goodness of fit.

a) Test of independence is used to test the association of variables in a two-way table. A two-way table of frequencies is also known as a contingency table.

For example, a random sample of 380 persons is classified according to their income levels (low, middle, high) and loan defaults (yes, no) as in this contingency table:

	Low	Middle	High
Yes	75	15	85
No	50	40	115

H_0: income levels and loan defaults are independent
H_a: income levels and loan defaults are not independent (associated)

Here shows an example of creating a table directly using R

```
> default <- matrix(c(75,50,15,40,85,115),ncol=3,nrow=2)
> colnames(default) <-c("Low","Middle","High")
> rownames(default) <-c("Yes","No")
> default <- as.table(default)
> chisq.test(default)
```

--- Pearson's Chi-squared test ---

```
 data:  default
 X-squared = 18.6112, df = 2, p-value = 9.091e-05
 # ----------------------------------------
```

Here, the p-value is less than the default value α (0.05) therefore, the result is significant, we conclude that income levels and loan defaults are associated.

b) Goodness of fit test (or test of homogeneity) examines how close the observed values to the expected values under the fitted model.

For example, a lottery game draws 5 numbers from a set of 10 numbers (0 to 9). Suppose the numbers were drawn 250 times, the observed frequencies of each number from 0 to 9 respectively are as follows:
> obsv_freq = c(24,31,26,18,23,17,27,28,35,21)
Does each number have equal chance to be selected?
H_0: $p_1 = p_2 = p_3 = p_4 = p_5 = p_6 = p_7 = p_8 = p_9$
H_a: p_i not equal $\frac{1}{10}$ for some i

To do Chi-square test in R uses ***chisq.test*** function.
> chisq.test(obsv_freq)

\# --- Chi-squared test for given probabilities ---

data: obsv_freq
X-squared = 11.36, df = 9, p-value = 0.2518
\# --

From the result, there is not enough evidence to reject the null hypothesis so we can conclude that each number has equal chance to be selected.

- ***Regression model***

The regression model is used to show how the response variable (or dependent variable) relates to the predictor variable (also known as an independent, explanatory variable). The response variable is usually a continuous value.

A simple linear regression assumes the linear relationship between one response (*y*) and one predictor variable (*x*). Note that there is only one predictor variable involved in this model. Whereas, a multiple linear regression models the relationship between one response variable (*y*) and two or more predictors $(x_1, x_2, ..., x_n)$ by the linear equation.

Simple Linear regression describes the linear relationship between variable *x* and *y*. The linear regression model is expressed as:

$$y = \alpha + \beta x + \epsilon$$

When α, β are parameters (from populations), and ϵ is the error term

The parameters α, β are chosen so that they minimize the sum squares error (ϵ).
Therefore, an estimated simple regression can be written as:

$$\hat{y} = a + bx$$

Typically, the model is used to predict the response variable from the predictor value that we know.

Example: Estimate the next eruption duration of Old Faithful. Assume that the waiting time is 75 minutes from the last eruption. This data set is called *faithful* from the R library *MASS*. The function *lm* is used to fit linear models.

$>$ library(MASS)
$>$ eruption = lm(eruptions~waiting,data=faithful)
$>$ coeffs = coefficients(eruption)
$>$ coeffs
(Intercept)   waiting
-1.87401599 0.07562795
$>$ waiting=75
$>$ duration=coeffs[1]+coeffs[2]*waiting
$>$ duration
(Intercept)
   3.79808

From the result, if the waiting time is 75 minutes from the last eruption then the next eruption should last 3.79808 minutes.

Alternative solution:
$>$ data1 = data.frame(waiting=75)
$>$ predict (eruption, data1)
       1
3.79808

Applying a model to estimate the value outside the original data is called *extrapolation*.

- **Correlation coefficient (r) and Coefficient of Determination (R^2)**

Correlation is the test for strength of a relationship between two variables. It examines whether one variable varies with another variable. The frequent used correlation tests are:
- Pearson's (r) is used for parametric data (i.e. how study time and score correlate)
- Spearman's rho and Kendall's Tau are used when one of the variable is non-parametric (i.e. ordinal)

The correlation coefficient varies from 1 to -1. It has direction and can be either positive or negative. Positive correlation means that one variable varies in the same direction of the other variable (i.e. X increase, Y increase). Negative correlation means that one variable is varied in the opposite direction of the other variable (e.g. X increase, Y decrease). The correlation is close to +1, if the relationship is strong and positive. The correlation is close to -1 if it strong and negative.
- When r= 1, it is a perfect *positive* relationship between two variables
- When r= 0, there is no relationship between two variables
- When r= -1, it is a perfect *negative* relationship between two variables

It is important to note that ***correlation describes the strength of the relation and does not necessarily mean causation***.

For example, the correlation between eruptions period and waiting time since last eruption is computed using:
> cor (faithful$eruptions, faithful$waiting)

r is beneficial as an initial exploratory tool, especially when there are several variables considered. The low value of r does not always indicate absence in the relationship. This may happen because of the non-linear relationship. So, it is very important to plot data to see the trend. As a reminder, exploratory data analysis is the crucial step for data analysis.

As mention earlier, the correlation r is used to determine the strength and direction of the linear relationship between two variables. However, the strength of a linear fit is commonly represented with the coefficient of determination (R^2). In other words, it is a measure of how close the fitted (estimated) regression line is to the data. R^2 is the ratio of the explained variation to the total variation. The value of R^2 is between 0 and 1. For example, $R^2 = 0.75$ can be interpreted as 75% of the total variation that can be explained by the linear relationship between x and y, whereas the other 25% of the total are left unexplained.
We can extract the value of R^2 from the summary of the linear model function.

> eruption = lm(eruptions ~ waiting, data=faithful)
> summary(eruption)$r.squared

Reference:

Introduction to SAS (2014). UCLA: Statistical Consulting Group. Retrieved from:
> http://www.ats.ucla.edu/stat/r/dae/t_test_power.htm

Kang, H. (2012). Statistical Inference. Wharton School, University of Pennsylvania.
> Retrieved from: http://stat.wharton.upenn.edu/~khyuns/stat431/#Lecture

King, B.W. (2014). *R Tutorials*. Coastal Carolina University. Retrieved from:
> http://ww2.coastal.edu/kingw/statistics/R-tutorials/

One and Two-sample T-tests (2014), Statistics, Columbia University. Retrieved from:
> http://www.stat.columbia.edu/~martin/W2024/R2.pdf

Stynes, J.D. (2003) *Statistics-Summary*. Michigan State University. Retrieved from:
> http://www.msu.edu/course/prr/844/stat97.doc

Stowell, S. (Dec 29, 2012). *Performing a one-sample t-test in R*. Retrieved from:
> http://www.instantr.com/2012/12/29/performing-a-one-sample-t-test-in-r/

The Minitab Blog (2014), *Alphas, P-Values, and Confidence Intervals*. Retrieved from:
> http://blog.minitab.com/blog/michelle-paret/alphas-p-values-confidence-intervals-oh-my

Waller, L.J. (2010). *Statistical Analysis – The First Steps*. Medical College of Georgia.
> Retrieved from: http://analytics.ncsu.edu/sesug/2010/HOW01.Waller.pdf