Week 7 - Review

Monday, February 19, 2018 4:46 AM

Business Statistics - 10.1 -10.5
Basic Statistics - Pg. 184-193 and Pg. 201-202

Key Terms from Business Statistics

- ***Independent samples***
  - Two or more samples in which the selected items are related only by chance
  - Items or people in each group are in no way related to those in the other group
- ***Dependent samples***
  - Two or more samples selected in such a way as to be dependent or related; each item or person in one sample has a corresponding matched or related item in the other samples. (a.k.a. related samples).
- ***F Test***
  - The F test of two population variances is extremely sensitive to violations of assumption that the populations are normally distributed.
- ***F Value***
  - The ratio of two sample variances, used to reach statistical conclusions regarding the null hypothesis; in ANOVA, the ratio of the treatment variance to the error variance.
- ***F Distribution***
  - A distribution based on the ratio of two random variances; used in testing two variances and in analysis of variance.
- ***Matched-Pairs test***
  - A *t* test to test the difference in two related or matched samples; sometimes called the *t* test for related measures or the correlated *t* test.
- ***Related Measures***
  - Another name for matched pairs or paired data in which measurements are taken from pairs of items or persons match on some characteristic or from a before-and-after design and then separated into different samples.
- ***Analysis of variance (ANOVA)***
  - A technique for statistically analyzing the data from a completely randomized design; uses the F test to determine where there is a significant difference between two or more independent groups.
  - The test statistic in a one-way analysis of variance is not the p-value.
    - It is the F test statistic used for hypothesis testing
- ***Two-way analysis of variance (ANOVA)***
  - The process used to statistically test the effects of variables in factorial designs with two independent variables.

What is a p-value?

- A p-value is the computed probability of obtaining a test statistic value, or a more extreme value, using the assumptions stated for the null hypothesis.
- P-Value is one way to evaluate a test statistic value, but you must understand how it is determined in order to understand how to place it into context.
- ***Small p-value means = look further into the data and context of the study***
- Is a small p-value by itself conclusive evidence?

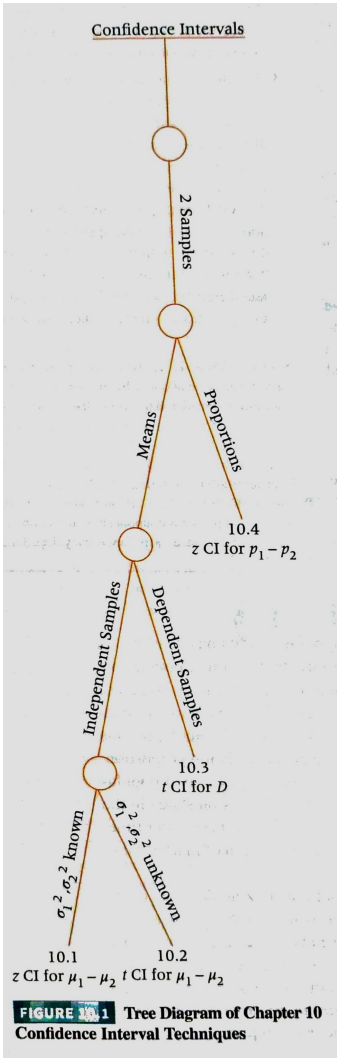


FIGURE 10.1 Tree Diagram of Chapter 10 Confidence Interval Techniques

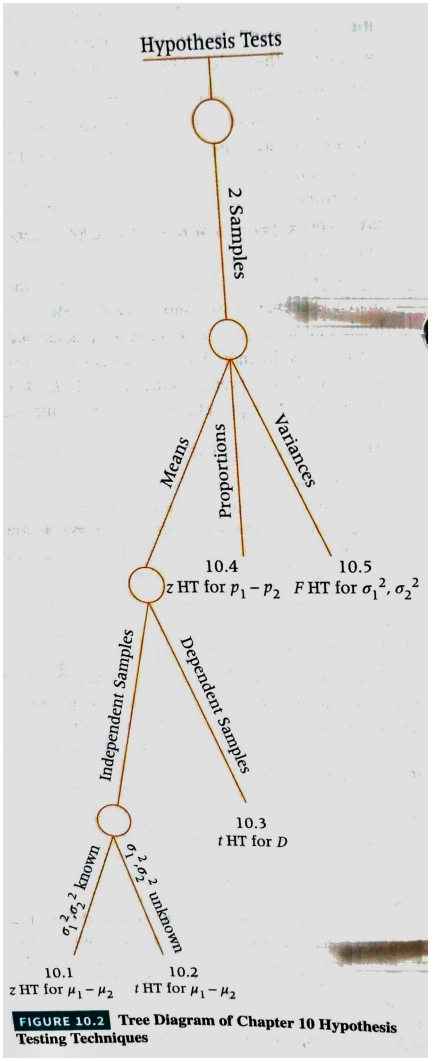


FIGURE 10.2 Tree Diagram of Chapter 10 Hypothesis Testing Techniques

z Formula for the Difference in Two Sample Means (Independent Samples and Population Variances Known)

$$z = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\quad}} \qquad (10.1)$$

determined in order to understand how to place it into context.

- ***Small p-value means = look further into the data and context of the study***
- Is a small p-value by itself conclusive evidence?
  - A small p-value casts doubt on or provides evidence against a null hypothesis or the underlying assumptions.
  - ***A p-value by itself without context or other evidence provides limited information and may mislead.***
  - You cannot just look at a p-value by itself - ***No single index should be substitute for sound reasoning.***
  - P-Hacking: data dredging, cherry picking, significance test manipulation, ignoring assumptions violations that have impact on results.
    - The number of publications which report results that cannot be replicated about 1/3 of the time.
- A p-value less than 0.05 may be statistically significant, but it does not provide conclusive proof the alternative hypothesis is true.
- A p-value greater than 0.05 does not mean the null hypothesis is true.
  - The null hypothesis is never accepted.
  - The correct terminology is that is has not been rejected.

Permutation Tests

- A permutation test is a nonparametric approach for hypothesis testing. No assumption is made regarding the underlying distribution shape of the data.
- No normality required!
- Perform them with a relatively small sample size
- A permutation test makes no assumption regarding the underlying distribution shape of the data.
- Permutation test can be used to construct confidence intervals.

Bootstrapping and Permutation Tests

- Allow hypotheses to be tested and confidence intervals to be formed without reference to a known theoretical distribution
- These are valuable when:
  - Serious outliers
  - Small sample sizes
  - No existing parametric method that is applicable
- Analysis is performed on the original scale of measure which differs from analyses which uses transformations
- Computer intensive
- **Cannot turn bad data into good data**
- *"It is perfectly proper to use both classical and robust methods routinely, and only worry when they differ enough to matter. But when they differ, you should think hard."* John Tukey (1975)

Two-sample Student's t-test

- This is the classic and best-known method for comparing the mean of two independent groups.
- Violation assumptions
  - There are two conditions where the assumption of normality, or equal variances, can be violated and yet Student's t appears to continue to perform well in terms of Type 1 errors and accurate confidence intervals.

Population Variances Known)

$$z = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \quad (10.1)$$

where

μ_1 = the mean of population 1
μ_2 = the mean of population 2
n_1 = size of sample 1
n_2 = size of sample 2
σ_1^2 = the variance of population 1
σ_2^2 = the variance of population 2

***t* Formula to Test the Difference in Means Assuming σ_1^2, σ_2^2, are Equal**

$$t = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{s_1^2(n_1 - 1) + s_2^2(n_2 - 1)}{n_1 + n_2 - 2}}\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \quad (10.3)$$

$$\text{df} = n_1 + n_2 - 2$$

***t* Formula to Test the Difference in Means**

$$t = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} \qquad \text{df} = \frac{\left[\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right]^2}{\frac{\left(\frac{s_1^2}{n_1}\right)^2}{n_1 - 1} + \frac{\left(\frac{s_2^2}{n_2}\right)^2}{n_2 - 1}}$$

Confidence Interval to Estimate $\mu_1 - \mu_2$ Assuming the Population Variances are Unknown and Equal

$$(\bar{x}_1 - \bar{x}_2) - t\sqrt{\frac{s_1^2(n_1 - 1) + s_2^2(n_2 - 1)}{n_1 + n_2 - 2}}\sqrt{\frac{1}{n_1} + \frac{1}{n_2}} \le \mu_1 - \mu_2 \le (\bar{x}_1 - \bar{x}_2) + t\sqrt{\frac{s_1^2(n_1 - 1) + s_2^2(n_2 - 1)}{n_1 + n_2 - 2}}\sqrt{\frac{1}{n_1} + \frac{1}{n_2}} \quad (10.4)$$

$$\text{df} = n_1 + n_2 - 2$$

***t* Formula to Test the Difference in Two Dependent Populations**

$$t = \frac{\bar{d} - D}{\frac{s_d}{\sqrt{n}}} \quad (10.5)$$

- There are two conditions where the assumption of normality, or equal variances, can be violated and yet Student's t appears to continue to perform well in terms of Type 1 errors and accurate confidence intervals.

- Conditions where it performs poorly
  - Student's t can be unsatisfactory in terms of Type I errors and accurate confidence interval when sampling from normal distributions with unequal sample sizes and unequal variances.
- Dealing with unequal variances: Welch's test (pg 191)
  - Welch's method seems to perform reasonably well compared to other techniques that have been derived when attention is restricted to comparing means.

T-Test Situation and Pairing

- Whenever you have a positive correlation between measurement on a per subject basis, its far better off basing statistical analysis on a paired t-test foundation.
  - This will be much more efficient and powerful because you are eliminating variation of total population and focusing on the variation between the two measurements by subject.
- Two basic design methods
  - Randomized - to balance the groups being compared and the outcomes that you are trying to compare show up regularly in each of the two groups.
  - Block - block on factors that are associated with the subjects where the factors are known to materially impact the variability of the subjects.

Power Calculations

- Power - the probability of rejecting a false null hypothesis
- Power Curve
  - A graph that plots the power values against various values of the alternative hypothesis.
  - Plotting the power values (1-beta) against the various values of the alternative hypotheses.
  - The power increases as the alternative mean moves away from the values of mu in the null hypotheses.
  - As the alternative mean moves farther and farther away from the null hypothesized mean, a correct decision to reject the null hypothesis becomes more likely.
- R function
  - Power.t.test <- this allows you to calculate the power of a t test under a set of particular parameters.

Randomized Block Designs

- The experimenter divides subjects into subgroups called 'blocks', so that the variability within these subgroups is less than the variability between the blocks.
- Subjects within each block are randomly assigned to treatment conditions.
- The design reduces variability within treatment conditions and potential producing better estimates of treatment effects.

Analysis of Variance (ANOVA)

- We are concerned about means between different groups.
- We are looking at the variability of the means to reach a conclusion as to whether the groups differ or not.
  - One way to do this is using the F test
- The ANOVA uses the F tests extensively.

$$t = \frac{}{\frac{s_d}{\sqrt{n}}} \qquad (10.5)$$

$$\text{df} = n - 1$$

where

n = number of pairs
d = sample difference in pairs
D = mean population difference
s_d = standard deviation of sample difference
$\bar{d}$ = mean sample difference

Formulas for $\bar{d}$ and s_d

$$\bar{d} = \frac{\Sigma d}{n}$$

$$s_d = \sqrt{\frac{\Sigma(d - \bar{d})^2}{n-1}} = \sqrt{\frac{\Sigma d^2 - \frac{(\Sigma d)^2}{n}}{n-1}} \qquad (10.6 \text{ and } 10.7)$$

Confidence Interval Formula to Estimate the Difference in Related Populations, D

$$\bar{d} - t\frac{s_d}{\sqrt{n}} \le D \le \bar{d} + t\frac{s_d}{\sqrt{n}} \qquad (10.8)$$

$$\text{df} = n - 1$$

z Formula for the Difference in Two Population Proportions

$$z = \frac{(\hat{p}_1 - \hat{p}_2) - (p_1 - p_2)}{\sqrt{\frac{p_1 \cdot q_1}{n_1} + \frac{p_2 \cdot q_2}{n_2}}} \qquad (10.9)$$

where

$\hat{p}_1$ = proportion from sample 1
$\hat{p}_2$ = proportion from sample 2
n_1 = size of sample 1
n_2 = size of sample 2
p_1 = proportion from population 1
p_2 = proportion from population 2
$q_1 = 1 - p_1$
$q_2 = 1 - p_2$

z Formula to Test the Difference in Population Proportions

$$z = \frac{(\hat{p}_1 - \hat{p}_2) - (p_1 - p_2)}{\sqrt{(\bar{p} \cdot \bar{q})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} \qquad (10.10)$$

- We are looking at the variability of the means to reach a conclusion as to whether the groups differ or not.
    - One way to do this is using the F test
- The ANOVA uses the F tests extensively.
- Sample variances must be determined for hypotheses to be tested.
- R Function
    - aov
    - TukeyHSD

Student's *t* Statistic Performance

- **Symmetric Distributions**
    - *Outliers are rare*
        - **Use** Student's t
        - May need n >= 30
    - *Outliers are common*
        - **May use** Student's t
        - Large sample size needed
        - Consider alternatives such as
            - Bootstrap-t method
            - Trimmed mean with Winsorized standard deviation
            - A non-parametric procedure
- **Asymmetric Distributions**
    - *Outliers are rare*
        - **May use** Student's t
        - May need n >= 200
    - *Outliers are common*
        - **May use** Student's t
        - May need n >= 300
        - Consider alternatives such as
            - Bootstrap-t method
            - Trimmed mean with Winsorized standard deviation
            - A non-parametric procedure

$$\sqrt{(\bar{p}\cdot\bar{q})\left(\frac{1}{n_1}+\frac{1}{n_2}\right)} \qquad (10.10)$$

where $\bar{p}=\dfrac{x_1+x_2}{n_1+n_2}=\dfrac{n_1\hat{p}_1+n_2\hat{p}_2}{n_1+n_2}$ and $\bar{q}=1-\bar{p}$

Confidence Interval to Estimate p_1-p_2

$$(\hat{p}_1-\hat{p}_2)-z\sqrt{\frac{\hat{p}_1\cdot\hat{q}_1}{n_1}+\frac{\hat{p}_2\cdot\hat{q}_2}{n_2}}\le p_1-p_2\le(\hat{p}_1-\hat{p}_2)+z\sqrt{\frac{\hat{p}_1\cdot\hat{q}_1}{n_1}+\frac{\hat{p}_2\cdot\hat{q}_2}{n_2}} \qquad (10.11)$$

Formula for Determining the Critical Value for the Lower-Tail *F*

$$F_{1-\alpha,\nu_2,\nu_1}=\frac{1}{F_{\alpha,\nu_1,\nu_2}} \qquad (10.13)$$

***F* Test for Two Population Variances**

$$F=\frac{s_1^2}{s_2^2} \qquad (10.12)$$

$$\mathrm{df}_{\text{numerator}}=\nu_1=n_1-1$$

$$\mathrm{df}_{\text{denominator}}=\nu_2=n_2-1$$

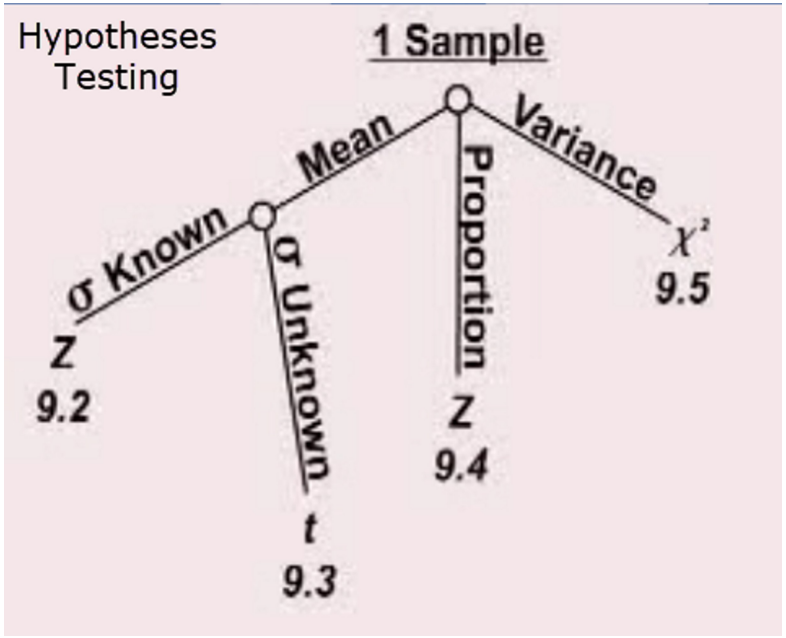
Formulas

z test for the difference in two independent sample means

$$z=\frac{(\bar{x}_1-\bar{x}_2)-(\mu_1-\mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}}}$$

Confidence interval for estimating the difference in two independent population means using z

$$(\bar{x}_1-\bar{x}_2)-z\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}}\le\mu_1-\mu_2\le(\bar{x}_1-\bar{x}_2)+z\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}}$$



$$W = \frac{(\bar{X}_1 - \bar{X}_2)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}},$$

t test for two independent sample means, and population variances unknown but assumed to be equal (assume also that the two populations are normally distributed)

$$t = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{s_1^2(n_1 - 1) + s_2^2(n_2 - 1)}{n_1 + n_2 - 2}}\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

$$\text{df} = n_1 + n_2 - 2$$

Confidence interval for estimating the difference in two independent means, and population variances unknown but assumed to be equal (assume also that two populations are normally distributed)

$$(\bar{x}_1 - \bar{x}_2) \pm t\sqrt{\frac{s_1^2(n_1 - 1) + s_2^2(n_2 - 1)}{n_1 + n_2 - 2}}\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$

$$\text{df} = n_1 + n_2 - 2$$

t test for the difference in two related samples (the differences are normally distributed in the population)

$$t = \frac{\bar{d} - D}{\frac{s_d}{\sqrt{n}}}$$

$$\text{df} = n - 1$$

Formulas for $\bar{d}$ and s_d

$$\bar{d} = \frac{\Sigma d}{n}$$

$$s_d = \sqrt{\frac{\Sigma(d - \bar{d})^2}{n - 1}} = \sqrt{\frac{\Sigma d^2 - \frac{(\Sigma d)^2}{n}}{n - 1}}$$

Confidence interval formula for estimating the difference in related samples (the differences are normally distributed in the population)

$$\bar{d} - t\frac{s_d}{\sqrt{n}} \le D \le \bar{d} + t\frac{s_d}{\sqrt{n}}$$

$$\text{df} = n - 1$$

z formula for testing the difference in population proportions

$$z = \frac{(\hat{p}_1 - \hat{p}_2) - (p_1 - p_2)}{\sqrt{(\bar{p} \cdot \bar{q})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

where $\bar{p} = \frac{x_1 + x_2}{n_1 + n_2} = \frac{n_1\hat{p}_1 + n_2\hat{p}_2}{n_1 + n_2}$ and $\bar{q} = 1 - \bar{p}$

Confidence interval to estimate $p_1 - p_2$

$$(\hat{p}_1 - \hat{p}_2) - z\sqrt{\frac{\hat{p}_1 \cdot \hat{q}_1}{n_1} + \frac{\hat{p}_2 \cdot \hat{q}_2}{n_2}} \le p_1 - p_2$$

$$\le (\hat{p}_1 - \hat{p}_2) + z\sqrt{\frac{\hat{p}_1 \cdot \hat{q}_1}{n_1} + \frac{\hat{p}_2 \cdot \hat{q}_2}{n_2}}$$

F test for two population variances (assume the two populations are normally distributed)

$$F = \frac{s_1^2}{s_2^2}$$

$$\text{df}_{\text{numerator}} = v_1 = n_1 - 1$$

$$\text{df}_{\text{denominator}} = v_2 = n_2 - 1$$

Formula for determining the critical value for the lower-tail *F*

$$F_{1-\alpha, v_2, v_1} = \frac{1}{F_{\alpha, v_1, v_2}}$$