Chapter 10: Inference Based on Two Samples

One population $\rightarrow$ Two populations: a natural generalization

$\mu \rightarrow$ Comparisons between μ_1 and μ_2.

1. Confidence interval for :
   - $\mu_1 - \mu_2$;
   - $p_1 - p_2$
2. Test about :
   - $\mu_1 - \mu_2$;
   - $p_1 - p_2$

Real data examples:

1. Comparison of Learning methods for GRE;
2. Comparison of Treatments for cancers ;
3. Comparison of MPG between Hybrid (e.g. 2013 Ford Fusion) and Traditional cars.
4. Comparison of Working Stress between industry and academia;

Outline of Chapter 10: Inference Based on Two Samples

1. z Tests and Confidence Intervals for a Difference Between Two Population Means
   - Normal sample
   - General sample, but large sample size
   - Confidence Interval for $\mu_1 - \mu_2$
2. The Two-Sample t Test and Confidence Interval
   - Non-pooled t Procedures
   - Pooled t Procedures
3. Analysis of Paired Data
   - Paired Versus Unpaired Experiments.
   - The paired t test for μ_D.
   - Confidence interval for μ_D.
4. Inference About Two Population Proportions
   - Large-sample Test Procedure
   - Large-sample Test Confidence interval for $p_1 - p_2$.

Keep in mind

In this chapter, statistical Settings:

1. $X_1, \ldots, X_m$ is a random sample from Population A with mean μ_1 and variance σ_1^2;
2. $Y_1, \ldots, Y_n$ is a random sample from Population B with mean μ_2 and variance σ_2^2;

And the tests and C.I.'s depend on

1. Whether the two samples are independent with each other;
2. Whether the sample sizes are large;
3. Whether we assume Normal distribution on the two samples;
4. When (2) is true, whether we know the values of σ_1^2 and σ_2^2.

Chapter 10 - Lecture 1
Tests and Confidence Intervals for a Difference between two population means

Yuan Huang

March 15, 2013

1. Case I: normal sample with known variance
   - Statistical Setting
   - Confidence Interval of $\mu_1 - \mu_2$
   - Testing $H_o : \mu_1 - \mu_2 = c_0$
   - Type II error

Case I: normal sample with known variance

Statistical Settings: This is the simplest case where the two samples are independent and we assume normal distribution on the two samples, with variance known.

1. $X_1, \ldots, X_m$ i.i.d $N(\mu_1, \sigma_1^2)$ with σ_1^2 known;
2. $Y_1, \ldots, Y_n$ i.i.d $N(\mu_2, \sigma_2^2)$ with σ_2^2 known;
3. The two samples are independent.

Then we have the probability distribution:

$$\frac{\bar{X}_m - \bar{Y}_n - (\mu_1 - \mu_2)}{\sqrt{\sigma_1^2/m + \sigma_2^2/n}} \sim N(0, 1).$$

Confidence Interval of $\mu_1 - \mu_2$

In this case, a $100(1-\alpha)\%$ Confidence Interval of $\mu_1 - \mu_2$ is:

$$(\bar{x}_m - \bar{y}_n - z_{\alpha/2}\sqrt{\frac{\sigma_1^2}{m} + \frac{\sigma_2^2}{n}}, \bar{x}_m - \bar{y}_n + z_{\alpha/2}\sqrt{\frac{\sigma_1^2}{m} + \frac{\sigma_2^2}{n}}).$$

Similarly we can get one-sided intervals.

Testing $H_o : \mu_1 - \mu_2 = c_0$

In this case the test statistic $H_o : \mu_1 - \mu_2 = c_0$ is:

$$Z = \frac{\bar{X}_m - \bar{Y}_n - c_0}{\sqrt{\sigma_1^2/m + \sigma_2^2/n}}$$

1. For $H_1 : \mu_1 - \mu_2 > c_0$, we reject H_o when $z > z_\alpha$;
2. For $H_1 : \mu_1 - \mu_2 < c_0$, we reject H_o when $z < -z_\alpha$;
3. For $H_1 : \mu_1 - \mu_2 \neq c_0$, we reject H_o when $|z| > z_{\alpha/2}$.

Recall: p-value approach

Comments:

1. If we reject $H_0 : \mu_1 = \mu_2$ with $H_1 : \mu_1 > \mu_2$, then we say μ_1 is significantly greater than μ_2;
2. If we reject $H_0 : \mu_1 = \mu_2$ with $H_1 : \mu_1 < \mu_2$, then we say μ_1 is significantly less than μ_2;
3. If we reject $H_0 : \mu_1 = \mu_2$ with $H_1 : \mu_1 \neq \mu_2$, then we say μ_1 is significantly different from μ_2;

Relationship between the confidence interval and the hypothesis test (for any number of populations) based on the same probability distribution:

1. We reject H_o in two-tailed test if the null value is not included in the two-tailed C.I;
2. We reject H_o in upper-tailed test if the null value is not included in the lower-tailed C.I;
3. We reject H_o in lower-tailed test if the null value is not included in the upper-tailed C.I.

Calculating Type II error probabilities

Let

$$\sigma = \sqrt{\frac{\sigma_1^2}{m} + \frac{\sigma_2^2}{n}}$$

1. $H_A : \mu_1 - \mu_2 > \Delta_0$,

$$\beta(\Delta') = \Phi\left(z_\alpha - \frac{\Delta' - \Delta_0}{\sigma}\right)$$

2. $H_A : \mu_1 - \mu_2 < \Delta_0$

$$\beta(\Delta') = 1 - \Phi\left(-z_\alpha - \frac{\Delta' - \Delta_0}{\sigma}\right)$$

3. $H_A : \mu_1 - \mu_2 \neq \Delta_0$

$$\beta(\Delta') = \Phi\left(z_{\alpha/2} - \frac{\Delta' - \Delta_0}{\sigma}\right) - \Phi\left(-z_{\alpha/2} - \frac{\Delta' - \Delta_0}{\sigma}\right)$$

Example 10.1

Each students in a class of 21 responded to a questionnaire that requested their grade point average (GPA) and the number of hours each week that they studied. Assume normality for GAP and the standard deviation is 0.6.

- 11 students studied for > 10 h/week: the sample mean is 3.06
- 10 students studied for < 10 h/week: the sample mean is 2.97

Treating the two samples as random, is there evidence that true average GPA differs for the two study times?