Statistical Inference II: Hypothesis Testing, ANOVA

BU609-4a-Testing

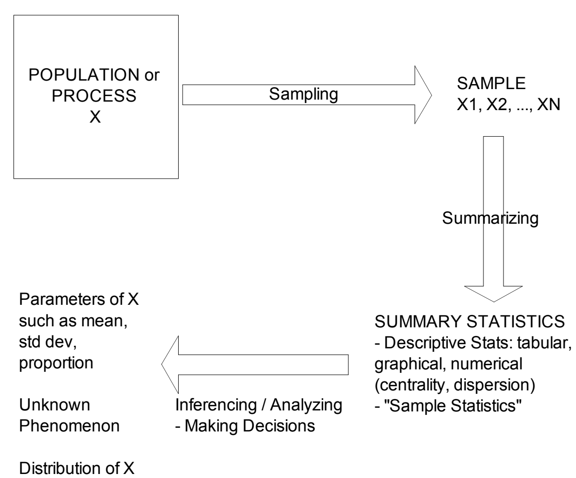
Objectives

- Understand and be able to apply hypothesis testing, as a method of statistical inference
- Be able to run Analysis of Variance test for multiple means, and interpret results using Regression

Hypothesis Testting - Intro

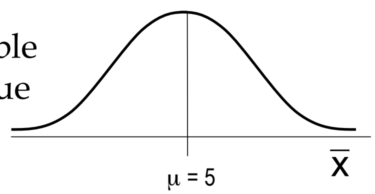
- Purpose of hypothesis testing is to determine whether there is sufficient **statistical evidence** in favor of a certain **belief** about a **parameter**
- Examples
  - Is there statistical evidence in a random sample to support hypothesis that > 20% of potential customers will purchase a new product?
  - Based on random sample, can manufacturer conclude manufacturing process is under control?
  - Is one type of longterm investment riskier (more variable returns) than another?
  - Does ad A result in more sales than ad B?

KEY STATISTICAL CONCEPTS



Concept of Hypothesis Testing

- Two hypotheses about pop. parameter(s)
  - H_0 - null hypothesis   [for example $\mu = 5$]
  - H_1 - alternative hypothesis [$\mu > 5$]   What we want to prove
  - Initially assume null hypothesis is true ($\mu = 5$)
  - Build a statistic related to parameter hypothesized
  - Pose question: How probable is it to obtain a statistic value at least as extreme as one observed from sample?



(Cont.)

- Make one of the following two decisions (based on test):
  - **Reject** null hypothesis in favor of alternative
  - **Do not reject** null hypothesis in favor of alternative hypothesis
- Two types of errors are possible when making decision whether to reject H_0
  - Type I error - reject H_0 when it is true.
  - Type II error - do not reject H_0 when it is false

Jury Illustration

	Guilty	**Innocent**
Find Innocent	Wrong	Correct Decision
Find Guilty	Correct Decision	Wrong

- Can jury be absolutely sure?
- Are errors equally serious? If not, which type is worse?

Possible Results

	H_0 is True	**H_0 is False**
Reject H_0	Type I Error P(Type I Error) = α	Correct Decision
Don't Reject H_0	Correct Decision	Type II Error P(Type II error) = β

- Easy to deal with α
- Difficult to deal with β (so often ignored!)
- Trade-off between these two

Testing Pop. Mean (σ known)

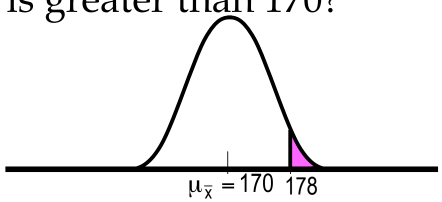
- Example
  - A new billing system for a department store will be cost- effective only if the mean monthly account is more than \$170
  - Sample of 400 monthly accounts has mean of \$178
  - If accounts are approx. normally distributed with σ = \$65, can we conclude that new system will be cost effective?

Solution

- Population of interest is credit accounts at store; parameter is mean
  - Want to show that the mean account for all customers is greater than \$170
    - » H_1: $\mu > 170$
  - Null hypothesis must specify a single value for parameter μ
    - » H_0: $\mu = 170$

(Solution cont.)

Is sample mean of 178 sufficiently greater than 170 to infer that population mean is greater than 170?



If μ is really equal to 170, then $\mu_{\bar{x}} = 170$. The distribution of the sample mean should look like this.

Is it likely to have $\bar{x} \geq 178$ under the null hypothesis (μ = 170)?

Recall: Sampling Dist. (σ known)

- Test statistic for μ

$$z = \frac{\bar{x} - \mu}{\sigma / \sqrt{n}}$$

- Confidence interval estimator of μ

$$\left[\bar{x} - z_{\alpha/2} \frac{\sigma}{\sqrt{n}}, \bar{x} + z_{\alpha/2} \frac{\sigma}{\sqrt{n}}\right]$$

Solution

Are we likely to see a sample mean of ≥ 178, if the population mean is 170?

- H_0: $\mu = 170$
- H_1: $\mu > 170$
- Test statistic:

$$z = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}} = \frac{178 - 170}{65/\sqrt{400}} = \frac{8}{3.25} = 2.4615$$

- p-value: p(z>2.46) = 1 – NORM**S**DIST(2.46) = 0.0069
- Conclusion: Extremely unlikely. Reject null hypothesis in favor of alternative

p-value (Prob of Type 1 Error)

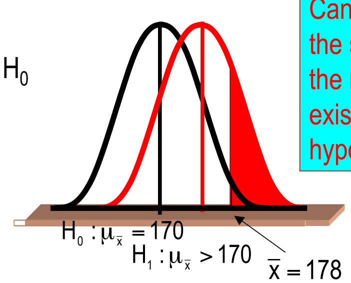
- p-value provides information about amount of statistical evidence that supports alternative hypothesis
- p-value of a test is probability of observing a test statistic at least as extreme as one computed, given that null hypothesis is true; P(Type 1 Error) = α
- NB: this p-value is not a proportion parameter or statistic!

Interpreting p-value

- Because probability sample mean will assume value of more than 178 when μ = 170 is so small (.0069), there are very strong reasons to believe that $\mu > 170$

Note how the event $\bar{x} \geq 178$ is rare under H_0 when $\mu_{\bar{x}} = 170$, but...

...it becomes more probable under H_a, when $\mu_{\bar{x}} > 170$



Can conclude that the smaller the p-value the more statistical evidence exists to support alternative hypothesis

Practical vs Statistical Significance

- Statistically significant results are those which produce sufficiently small p-values
  - support H_1: $\mu > 170$; However difference between H_0 and H_1 may not be of ***practical significance***
  - If two sample sizes are large enough, can almost always prove statistically significant differences between means, proportions, variances
- So, be wary of "statistical significance"

Economic Significance

- It is at least as important to know if results are **economically significant** - i.e., are they big?
- With lots of data, almost everything will be statistically significant; may not be economically significant
- With little data, little will be statistically significant; may still be economically significant

Read "Signifying Nothing?" in The Economist

http://www.economist.com/displaystory.cfm?story_id=2384590

"Rules of Thumb" re p-values

- If p-value is < 1%, ***overwhelming evidence*** that supports alternative hypothesis
- If 1% ≤ p-value ≤ 5%, there is ***strong evidence*** that supports alternative hypothesis
- If 5% < p-value ≤ 10%, there is ***weak evidence*** that supports alternative hypothesis
- If p-value > 10%, there is ***no evidence*** that supports alternative hypothesis

1-Tail vs 2-Tail Tests

- Our example is a "1-tail" test
  - Interested in only RHS of sampling dist.
  - Could also focus on LHS
- 2-tail tests involve **both** sides of sampling distribution (so use $\alpha/2$)
- Determined by alternative hypothesis
  - H_1: > or < , 1-tail test
  - H_1: ≠ , 2-tail test (< or >)
- Be careful when solving (<u>diagram helps</u>)

Conclusions of a Test of Hypothesis

- If null hypothesis rejected, conclude there is enough evidence to infer that alternative hypothesis is true
- If null hypothesis not rejected, conclude there is <u>not enough</u> statistical evidence to infer that alternative hypothesis is true (**<u>do not</u> accept null**)
- Alternative hypothesis is more important one; represents what we are investigating

Hypothesis Testing Summary

- Null hypothesis (H_0)
  - specifies a point value of the parameter
  - always **=**
- Alternative (H_1): what we want to test
  - 3 choices: **>** **,** **<** , or **≠** (1 or 2-tails)
- Test statistic: sample statistic upon which we reject or don't reject null hypothesis
  - use to calculate p-value
- Conclusion: if p-value small, reject null hypothesis; otherwise don't

Common Hypothesis Tests

- Means
  - 1 or 2 populations
  - paired difference (dependent samples)
- Variance
  - 1 or 2 populations
- Proportions
  - 1 or 2 populations

See Handout on website

Analysis of Variance (ANOVA) - Use Regression

Comparing means of 2 or more populations

Summary

- Analysis of variance helps compare **two or more** populations of quantitative data
- Specifically interested in relationships among population means (are they equal or not?)
- Procedure works by analyzing the **sample variance** (hence name)
- Restrict ourselves to single-factor (one-way) **ANOVA** with independent samples
- Use Dummy Variables for difference between means and Regression

ANOVA Basics

H_0: $\mu_1 = \mu_2 = \mu_3$
H_1: At least two means differ

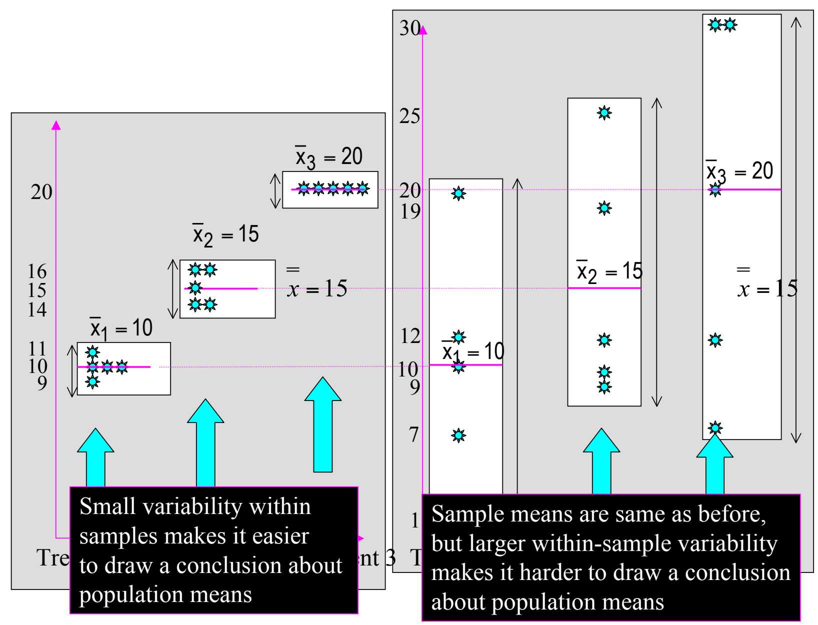
To perform analysis of variance, need to build an "F" statistic (compares 2 variances)

Use Regression!

Test Statistic

- Test stems from the following rationale:
  - If null hypothesis is true, would expect all sample means be close to one another (and as a result to grand mean)
  - If alternative hypothesis is true, at least some sample means would be different from one another (and from grand mean)
- Use **F-statistic**
  - Used when comparing two **variances** **(ratio)**
  - Based on **Sampling Distribution**

Why combine two types of variability, when testing for the equality of means?



Anova: Test of Hypothesis

H_0: $\mu_1 = \mu_2 = \ldots = \mu_k$
H_1: At least two means differ

Test statistic: $F = \dfrac{MST}{MSE} = \dfrac{s_1^2}{s_2^2}$

where MST = mean square treatments (variance of treatment means about grand mean);
MSE = mean square error (variance of observations about treatment means)

p-value: ?? Use p-value from regression of dummy variable representing difference in means

Anova (cont.)

H_o: $\mu_1 = \mu_2 = \mu_3$
H_1: At least two means differ

Test statistic: F = MST/MSE = 28,756.12 / 8,894.45 = 3.23

p-value = 0.047

Since p-value is < 0.05, there is sufficient evidence to reject H_o in favor of H_1, and argue that at least one of the mean sales is different from the others.