Online Review Course of
Undergraduate Probability and Statistics

Review Lecture 13
Inferences about a Mean

Chris A. Mack
Adjunct Associate Professor

Course Website: www.lithoguru.com/scientist/statistics/review.html

Making Inferences

- We wish to make inferences about the population based on data from a sample
  - Two complementary approaches: Confidence Intervals and Hypothesis Testing
  - Third approach: Bayesian (important, but we won't cover it here)
- Example: we measure the mean of a sample. What does it say about the mean of the population sampled?

Sampling Distribution of the Mean

- What is the sampling distribution of the mean?

$$\bar{X} = \frac{1}{n}\sum_{i=1,n} X_i \qquad E[\bar{X}] = \mu \qquad var[\bar{X}] = \frac{\sigma^2}{n}$$

(Assuming an infinite population)

- If population variance (σ^2) is finite, the central limit theorem can apply

$$Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \sim N(0,1) \quad \text{or} \quad Student's\ t = \frac{\bar{X} - \mu}{S/\sqrt{n}}$$

sample standard deviation

Confidence Interval for the Mean

- Confidence interval:

Margin of Error

$$\bar{x} - t_{\alpha/2}\frac{s}{\sqrt{n}} < \mu < \bar{x} + t_{\alpha/2}\frac{s}{\sqrt{n}}$$

- Example: large sample 95% confidence interval

$$\left(\bar{x} - 1.96\frac{s}{\sqrt{n}}, \bar{x} + 1.96\frac{s}{\sqrt{n}}\right)$$

(n > 30, use normal distribution rather than Student's t)

  - 95% of random samples will capture the true mean within an interval constructed in this way
  - Hypothesis testing: does the hypothesized mean fall within the confidence interval?

Comparing Two Sample Means

- We take samples from two populations (μ_1, σ_1) and (μ_2, σ_2). We wish to know if the populations have different means.
  - Compare two treatments, do they have different outcomes?
- Important sampling approaches:
  1. Independent samples
  2. Matched samples

Two Independent Sample Means

- Two independent samples ($\bar{X}_1$, S_1) and ($\bar{X}_2$, S_2). Can we infer that $\mu_1 - \mu_2 \neq 0$?

$$E[\bar{X}_1 - \bar{X}_2] = \mu_1 - \mu_2 \qquad var[\bar{X}_1 - \bar{X}_2] = \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}$$

- For large, independent samples,

$$Z = \frac{\bar{X}_1 - \bar{X}_2 - (\mu_1 - \mu_2)}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}} \sim N(0,1)$$

Pooled Samples

- Consider two populations with different means (potentially) but with the same variance
  - As before, we sample the two populations and use the sample means to make inferences about the population means
  - If we are confident enough that the population variances are the same, we can pool all of the sample data to make one estimate of the population variance:

$$s_p^2 = \frac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{(n_1-1)+(n_2-1)}$$

- For t-tests, we use $n_1 + n_2 - 2$ degrees of freedom

Pooled Samples

- The pooled estimate of the variance is

$$s_p^2 = \frac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{(n_1-1)+(n_2-1)}$$

- Thus, the variance for our estimator for $\mu_1 - \mu_2$ becomes

$$var[\overline{X_1} - \overline{X_2}] = \sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}} = S_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$

- For t-tests, we use $n_1 + n_2 - 2$ degrees of freedom

Two Matched Samples

- We sometimes compare some property before and after a treatment
  - We must use matched samples, where we look for a change in the property caused by the treatment
  - Measure before and after for the same test piece and calculate the difference in the measured property

$$D_i = X_i - Y_i \qquad i = 1, 2, \dots, n$$

(X_i and Y_i are not independent)

- For large samples, $Z = \frac{\overline{D} - \mu_D}{S_D/\sqrt{n}} \sim N(0,1)$

Getting the Samples Right

- All of these statistical tests and procedures assume random sampling
  - When comparing two treatments, every subject/test piece must have equal probability of getting each treatment
    - Equal sample sizes for each treatment produces the most powerful test
- Pairing (matched samples) can be used to eliminate the effect of an uncontrolled variable
- Larger samples always produce more powerful tests

Review #13: What have we learned?

- What assumptions go into the calculations of large-sample and small-sample confidence intervals for the sample mean?
- What are the two sampling approaches for comparing two sample means?
- When can we use a pooled sample variance?

