Outline

Review of Previous Lectures
  Techniques and Principles
  Significance
  Overlapping Confidence Intervals
  Paired vs Independent Samples
  To pool or not to pool

Error, Power and Sample Size
  Type I Error
  Power
  Sample Size Calculation
Part I

Reviews
Confidence Interval

- For a population parameter $\mu$, first find an estimator $\hat{\mu}$.
- Find the sampling distribution of $\hat{\mu}$.
- 95% CI $(a, b)$ is defined such that $P_{\hat{\mu}}(a \leq \mu \leq b) = 0.95$.
- Interpretation: the probability refers to $a$ and $b$ (which are random), not $\mu$ (which is fixed).
Hypothesis Testing

- Define the null and (two-sided) alternative hypotheses; e.g.,
  \( H_0 : \mu = \mu_0 \) and \( H_1 : \mu \neq \mu_0 \).

- Define a test statistic \( T \) that also involves \( \mu_0 \); e.g., the \( t \)-statistic

\[
T = \frac{\bar{X} - \mu_0}{s/\sqrt{n}}.
\]

- Find the distribution of \( T \) assuming the null is true.

- Compute the \( p \)-value: \( P_0(|T| \geq |t|) \).

- Interpretation: \( p \) is computed under the assumption that the null is true.
How significant is the test result?

- Recall for a one-sample \( t \)-test, the observed value of the test statistic is:
  \[
  t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}},
  \]
  and \( p = P_{\mu_0}(|T| > |t|) \).

- For given \( H_0 \), the \( p \)-value is small if \(|\bar{x} - \mu_0|\) is large or the sample size \( n \) is large.

- Conversely, if \( p \) is large (> 0.05 say), it means the effect is small or the sample size is small.

- Moral: Always consider both the \( p \)-value and the effect size.
Overlapping Confidence Intervals

- **Counterintuitive result:** Confidence intervals associated with statistics can overlap as much as 29% and the statistics can still be significantly different.

- Consider the means of two independent samples, 10 and 22, with equal standard errors of 4. The margin of error is thus $1.96(4) = 7.84$, so 95% confidence intervals for the two statistics are $(2.16, 17.84)$ and $(14.16, 29.84)$. Are the two means significantly different?

- Surprisingly, **yes:** The $z$-statistic is:

  $$z = \frac{22 - 10}{\sqrt{4^2 + 4^2}} = 2.12.$$
Longitudinal Study

Hypertension and Oral Contraceptives (OC)

- Identify a group of non-pregnant, pre-menopausal women between age 16-49 who are not current OC users and measure their blood pressure (BP) (“baseline”).

- One year later, rescreen these women to find a subgroup who have remained non-pregnant and have become OC users. This subgroup is the study sample. Measure their BP (“follow-up”).

- Compare the baseline and follow-up BP of the women in the study sample.

- Paired (same women, baseline and follow-up)
Hypertension and Oral Contraceptives (OC)

- Identify both a group of OC users and a group of non-users among non-pregnant, pre-menopausal women between age 16-49 and measure their BP.
- Compare the BP of the OC users and nonusers.
- Not paired (no explicit pairing of users and nonusers)
Hypertension and Oral Contraceptives (OC)

- Identify a group of OC users among non-pregnant, pre-menopausal women between age 16-49.
- For each subject in the OC users group, find a non-pregnant, pre-menopausal woman of the same age, race, occupation, and education who is not an OC user.
- Measure the BP of both groups.
- Compare the BP of the OC users and nonusers.
- \[\Rightarrow\text{Paired}\] (explicit matching of cases and controls)
Paired-Sample vs Two-Sample

- Two samples are said to be *paired* when each data point of the first sample is matched and related to a unique data point of the second sample.
- Two samples are *independent* when the data points in one sample are unrelated to the data point in the second sample.
- To test the difference between samples:
  - Paired samples: use a one-sample procedure on the mean difference of the pairs, \( \bar{D} = \frac{1}{n} \sum_{i=1}^{n} D_i \) where \( D_i = X_i - Y_i \).
  - Independent samples: Use a two-sample procedure on the difference of the means, \( \bar{X}_1 - \bar{X}_2 = \frac{1}{n_1} \sum_{i=1}^{n_1} X_{1i} - \frac{1}{n_2} \sum_{i=1}^{n_2} X_{2i} \).
When the two populations have the same variance $\sigma^2$, it can be estimated using the pooled (all mixed together) sample.

The pooled sample $t$-statistic (which uses the pooled variance estimate $s_p^2$) has a $t$-distribution with $n_1 + n_2 - 2$ degrees of freedom (when $n_1$ and $n_2$ are large or the two populations are normal).
Unpooled-Sample t Test

- Used when the two population variances $\sigma_1^2$ and $\sigma_2^2$ are both unknown and unequal.
- When $n_1$ and $n_2$ are large or the two populations are normal, the unpooled sample $t$-statistic $t$ has an *approximate* $t$-distribution.
- Here we take df equal to either the min of $n_1 - 1$ and $n_2 - 1$, or the df formula on p.536 (due to Welch and Satterthwaite; thank goodness the computer will do this for us).
To pool or not to pool

- It is generally safer to use the unpooled sample $t$-test (with its more conservative $p$-value).
- When the sample sizes are equal, $n_1 = n_2$, the pooled-sample $t$-test is fairly robust against unequal variances.
- $F$-test for equal variances is almost useless (to be valid, it requires the two population distributions to be normal).
- We will pool when $s_{\text{larger}}^2/s_{\text{smaller}}^2 < 4$, BUT...
- When the two variances are very different, rethink whether you still want to test for a difference in the means (perhaps you will get the correct answer to the wrong question?...
Part II

Errors and Sample Size
Type I Error

Definition

- A \textit{type I error} occurs when the null hypothesis ($H_0$) is true but is rejected (i.e., a false positive).
- When $H_0$ is rejected at significance level $\alpha$, $\alpha$ is itself the (nominal) type I error rate.
- \textit{The p-value is not the type I error rate.}
Example

- For two tests with $p_1 = 0.04$ and $p_2 = 0.0001$, in both cases we reject $H_0$ at level $\alpha = 0.05$. The type I error rates are 0.05 for both tests.

- We can set the significance level at 0.01 and still reject the $H_0$ for the second test, thus achieving a smaller type I error rate. However, $\alpha$ has to be set before the data is collected, since error rates are long-run frequency properties (can’t “snoop the data”).

- This is another reason CIs are often preferred to significance tests (you set a confidence level and report whether or not the corresponding null is rejected).
Type II Error and Power

Definitions

- **A type II error** occurs when the null hypothesis is false, but is not rejected (i.e., a false negative).

- The probability of a type II error is often written as $\beta$. Note that unlike $\alpha$, we do not “set” $\beta$; it is just a characteristic of a level-$\alpha$ test.

- The **power** of a test is the probability of rejecting the null hypothesis when it is false.

- The power of a test against a particular alternative (e.g., $H_1 : \mu = \mu_1$) is $1 - \beta$, where $\beta$ is the probability of a type II error for that alternative.
Power Calculation

▶ For a test statistic \( z \), suppose the critical value is \( z_0 \). That is, we reject \( H_0 \) if \( |z| > z_0 \), where \( P_{H_0}(|z| > z_0) = \alpha \).

▶ The power is \( P_{H_1}(|z| > z_0) \), the probability of rejection when \( H_1 \) is true.

▶ Usually this means we need to consider a specific alternative; say, \( H_1 : \mu = 1 \) instead of merely \( H_1 : \mu > 0 \).

▶ More generally, for \( H_1 : \mu = m \), the power \( P_m(|z| > z_0) \) is a function of \( m \), and is called the \textit{power function}. 
Example 6.17

- **Fail to reject $H_0$**
- **Reject $H_0$**
- **$\alpha = 0.05$**
- **Power = 0.80**

Distribution of $\bar{x}$ when $\mu = 0$

Distribution of $\bar{x}$ when $\mu = 1$
Power/Type I Error Trade-Off

- If we reject the null no matter what the test statistic is, we have perfect power, but (potentially) very high type I error probability.

- Conversely, if we do not reject the null regardless of the test statistic, we incur no type I error, but also have no power!
Controlling for Type I Error

- A common practice is to choose a fixed type I error rate (say, $\alpha = .05$), and this then determines the type II error rate $\beta$, hence the power.
- Different testing procedures can be compared based on their power for a fixed $\alpha$ and particular alternative(s) $H_1$.
- For a fixed type I error rate ($\alpha$), increasing the sample size will increase the power $(1 - \beta)$. 
Sample Size Calculation

- To obtain a CI with a given margin of error (say, \( m = 0.01 \)), we have seen the sample size needed is

\[
    n = \left( \frac{z_0 \sigma}{m} \right)^2.
\]

- How to apply this idea to delivering a particular power in testing?
Sample Size and Testing

- For a given $\alpha$, we can find the critical value $z_0$ such that we reject $H_0 : \mu = \mu_0 = 0$ if $|z| > z_0$.
- For a given $H_1 : \mu = \mu_1$, the power is

$$P_{H_1}(|Z| > z_0) = P_{H_1} \left( \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} > z_0 \right).$$

- Therefore for a given power $1 - \beta$ and a particular alternative $H_1$, we can solve for the $n$ that will detect this alternative value with this power!
- If we use an even bigger alternative value $\mu_2 > \mu_1$, the power is even larger; can you see why?
Sample Size Calculation: Requirements

1. Distribution of the test statistic under the alternative (Gaussian for two-sample t-tests.)
2. Type I error rate, usually $\alpha = 0.05$.
3. The (minimal) power, often $1 - \beta = 0.8$.
4. The (minimal) magnitude of the effect $\mu_1 - \mu_0$ to be detected.
5. Variability: $\sigma^2$ (in the equal variance case)