Outline

Review of Previous Lecture
  Review of Previous Lecture

One Sample Inference
  Critical Value and Rejection Region
  Paired t Test

Two-Sample Inference
  Two-sample problems.
  Equal Variances
  Unequal Variances
Part I

Reviews
Confidence Interval for the Mean

- **Known \( \sigma \) (population standard deviation):**
  \[
  \bar{x} \pm z_{1-\alpha/2} \frac{\sigma}{\sqrt{n}}
  \]
  - Small \( n \), normal population.
  - Large \( n \), any population.

- **Unknown \( \sigma \):**
  \[
  \bar{x} \pm t_{n-1,1-\alpha/2} \frac{s}{\sqrt{n}}
  \]
  - Small \( n \), normal population.
  - Large \( n \), near-normal population.
One Sample Test: $H_0 : \mu = \mu_0$

- **Known $\sigma$ (population standard deviation),** the $z$-statistic
  
  \[
  z \equiv \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}},
  \]
  
  has a $\mathcal{N}(0, 1)$ distribution if the population is normal, or if $n$ is large.

- **Unknown $\sigma$, the $t$-statistic**
  
  \[
  t \equiv \frac{\bar{x} - \mu_0}{s/\sqrt{n}},
  \]
  
  has a $t_{n-1}$ distribution if the population is normal, or if $n$ is large and the population is nearly normal.
$p$-values

- $H_1 : \mu \neq \mu_0$: $p = P(|Z| > |z|)$ or $P(|T| > |t|)$.
- $H_1 : \mu > \mu_0$: $p = P(Z > z)$ or $P(T > t)$.
- $H_1 : \mu < \mu_0$: $p = P(Z < z)$ or $P(T < t)$. 
The *long-run* probability that the 95% CI for the mean covers the true population mean is 95%.

If the population mean is truly $\mu_0$, the probability of observing a $z$ or $t$ statistic as extreme or more extreme than yours is the $p$-value.

In both cases, the probabilities refer to relative frequencies in an *infinite* number of repetitions of the experiment.
Define the Hypothesis First

- If the hypothesis depends on the data ("data snooping"), then the requirement for identical experiments is violated, and the interpretation of the $p$-value becomes difficult/impossible.

- That is, the $p$-value will tend to overstate the evidence against a null that was generated by the data itself.
Misinterpretation of the CI

For a 95% CI \((a, b)\), can we say:

- “The probability that the true population parameter is within \((a, b)\) is 0.95.”
- NO NO NO NO NO NO NO NO NO NO NO NO NO NO NO!

- We can only talk about the long-run probability of \((a, b)\) trapping the (fixed) true parameter.
Misinterpretations of the $p$-value

If $t = 2.7$ with d.f. $= 18$, and $p = 0.01$,

- The probability that the null hypothesis is true is 0.01.
- The probability of observing $z$ when null hypothesis is true is 0.01.
- The probability of observing such a difference due to chance is 0.01.
- The probability of finding a significant result in a replicate experiment is 0.99.
- NO NO NO NO NO none of these are true!

Again, all we can say is that the $p$-value is the long-run probability of getting something as extreme as we got, or more so.
Part II

One Sample Inference
For $H_0 : \mu = \mu_0$, and a two-sided $H_1$, we reject the null when $p = \Pr(|T| > |t|) < 0.05$, where $t$ is the observed $t$-statistic.

The bigger $t$ is, the smaller the $p$-value.

Alternatively, we can find $t_0 > 0$ such that

$$P(|T| > t_0) = 0.05$$

and then reject the null if $|t| > t_0$.

$t_0 = t_{n-1,1-\alpha/2}$ (what we called $t^*$ in the CI context)

Here, $t_0$ is called the critical value of the test.
We reject $H_0$ at significance level $\alpha$ if and only if

$$t = \frac{|\bar{x} - \mu_0|}{s/\sqrt{n}} > t_{n-1,1-\alpha/2}$$

$$\iff |\bar{x} - \mu_0| > t_{n-1,1-\alpha/2} \times \frac{s}{\sqrt{n}}$$

$$\iff \bar{x} > \mu_0 + t_{n-1,1-\alpha/2} \times \frac{s}{\sqrt{n}} \quad \text{OR} \quad \bar{x} < \mu_0 - t_{n-1,1-\alpha/2} \times \frac{s}{\sqrt{n}}$$

**Definition**

The *rejection region* is the range of values of $\bar{x}$ (or whatever sample statistic we are using) for which $H_0$ is rejected.
For $H_0 : \mu = \mu_0$, two-sided $H_1$, and significance level $\alpha$, 

- We do not reject $H_0$ if 

$$ \bar{x} \in \left( \mu_0 - t_{n-1,1-\frac{\alpha}{2}} \times \frac{s}{\sqrt{n}}, \mu_0 + t_{n-1,1-\frac{\alpha}{2}} \times \frac{s}{\sqrt{n}} \right). $$

- Alternatively, we do not reject $H_0$ if 

$$ \mu_0 \in \left( \bar{x} - t_{n-1,1-\frac{\alpha}{2}} \times \frac{s}{\sqrt{n}}, \bar{x} + t_{n-1,1-\frac{\alpha}{2}} \times \frac{s}{\sqrt{n}} \right), $$

i.e., if the null mean value lies inside the $100(1 - \alpha)\%$ CI for $\mu$. 

PubH 6450 Fall 2006
In many situations, confidence intervals and two-sided hypothesis tests are equivalent.

The confidence interval is often the non-rejection region of the corresponding hypothesis test.

Always report the sample size, relevant sample statistic, and its CI.

Report the $p$-value (and the sampling distribution for the test statistic) when appropriate, rather than just “reject” or “fail to reject” at some level $\alpha$. 
Paired Samples

- Sometimes we want to compare two groups, but only within matched pairs, e.g., measurements of the same subjects before and after they receive some treatment.
- While this initially appears to be a two-sample problem, the pairing and the lack of independence between the samples means that the best approach here is to first convert the problem into a one-sample problem by taking the differences of the pairs.
- See M&M Example 7.7!
Paired Samples: CI and $t$ test

- If $(X_i, Y_i)$ are the paired samples, define $D_i = X_i - Y_i$, and construct a CI for the true mean difference $\mu_D = \mu_X - \mu_Y$ using our usual one-sample CI techniques.

- Or, test the hypothesis

$$H_0 : \mu_X = \mu_Y \iff \mu_D = 0$$

using the usual one-sample test techniques!

- For paired samples, the sample size is the number of pairs, not the total number of data points.

- A common mistake is treating paired two-sample problems as independent two-sample problems – a big difference!
Paired samples in SAS: PROC TTEST using PAIRED statement

```sas
data radon2;
    input before after @@;
    cards;
    91.9 97.8 111.4 122.3 105.4 95.0
    103.8 99.6 96.6 119.3 104.8 101.7
;
    proc ttest data = radon2 alpha = 0.05;
        paired before * after;
    run;
```
### PROC TTEST (V8.0+)

#### Statistics

<table>
<thead>
<tr>
<th>Difference</th>
<th>N</th>
<th>Lower CL Mean</th>
<th>Upper CL Mean</th>
<th>Lower CL Std Dev</th>
</tr>
</thead>
<tbody>
<tr>
<td>before-after</td>
<td>6</td>
<td>-16.27</td>
<td>-3.633</td>
<td>7.517</td>
</tr>
</tbody>
</table>

#### Statistics

<table>
<thead>
<tr>
<th>Difference</th>
<th>Std Dev</th>
<th>Std Dev</th>
<th>Std Err</th>
<th>Minimum</th>
<th>Maximum</th>
</tr>
</thead>
<tbody>
<tr>
<td>before-after</td>
<td>12.043</td>
<td>29.536</td>
<td>4.9163</td>
<td>-22.7</td>
<td>10.4</td>
</tr>
</tbody>
</table>
## T-Tests

| Difference       | DF | t Value | Pr > |t| |
|------------------|----|---------|------|---|
| before - after   | 5  | -0.74   | 0.4931 |

**PROC TTEST (V8.0+)**
Define the differences yourself and then use the ordinary one-sample $t$ procedure:

```sas
data radondiff;
  set radon2;
  diff = after - before;
proc ttest data = radondiff h0 = 0.0
  alpha = 0.05;
  var diff;
run;
```
Part III

Two-Sample Inference
Two-sample problems

- The goal of inference is to compare the responses in two groups.
- Each group is considered to be a sample from a distinct distribution.
- The responses in each group are independent of those in the other group.
Two-sample z-statistic

**Definition**

If a sample of size $n_1$ is drawn from $\mathcal{N}(\mu_1, \sigma^2_1)$ and a sample of size $n_2$ is drawn from $\mathcal{N}(\mu_2, \sigma^2_2)$, suppose $\bar{x}_1$ and $\bar{x}_2$ are the sample means for each sample. Then the *two-sample z statistic*

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

has the standard normal $\mathcal{N}(0, 1)$ distribution.
Unknown but Equal Variances

- When the population variances $\sigma_1^2$ and $\sigma_2^2$ are not known, the situation is a bit more complex.

- If the two populations can be assumed to have the same variance (i.e., $\sigma_1^2 = \sigma_2^2 = \sigma^2$), then we can use the pooled estimator of $\sigma^2$,

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

$$= \frac{\sum_{i=1}^{n_1}(x_{1i} - \bar{x}_1)^2 + \sum_{i=1}^{n_2}(x_{2i} - \bar{x}_2)^2}{n_1 + n_2 - 2}.$$
We can use the $F$-statistic,

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

to test for the equality of the two population variances, i.e.,

$$H_0 : \frac{\sigma_1^2}{\sigma_2^2} = 1.$$ 
The alternative is

$$H_1 : \frac{\sigma_1^2}{\sigma_2^2} \neq 1.$$ 

$F$ has an $F$-distribution with $(n_1 - 1, n_2 - 1)$ degrees of freedom.

Stat software is available to give $p$-values for this $F$-test...

BUT the $F$ is so nonrobust (easily influenced by departures from normality) that it is rarely used (see M&M p.556)!

Instead, pool when $s_{\text{larger}}^2/s_{\text{smaller}}^2 < 4$. 
Pooled-sample $t$-test

Definition
When the two populations have the same variance, we can test $H_0 : \mu_1 = \mu_2$ using the *pooled-sample $t$-statistic*

$$t = \frac{(\bar{x}_1 - \bar{x}_2)}{s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

where $s_p = \sqrt{s_p^2}$, the pooled standard error estimate defined above.
Under $H_0 : \mu_1 = \mu_2$, the pooled sample $t$ statistic has a $t$ distribution with $n_1 + n_2 - 2$ degrees of freedom. For a two-sided alternative $H_1 : \mu_1 \neq \mu_2$, the $p$-value is once again

$$p = P(|T| > |t|).$$

See M&M Example 7.20
Confidence Interval

Based on similar argument as in one-sample case, we can also define a $100(1 - \alpha)\%$ CI for the difference of the two population means:

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

where $t^* = t_{n_1+n_2-2, 1-\alpha/2}$ and $s_p$ is again the pooled standard error estimate.

See M&M Example 7.21
Unknown and Unequal Variances

For the more general case when the variances are not equal, the two-sample $t$-statistic is defined as:

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

which does not quite have a $t$-distribution under $H_0: \mu_1 = \mu_2$...
Unequal Variances

- Do you still want to test for $\mu_1 = \mu_2$ when the variances are unequal?
- If you still do, we can use a $t$ distribution to approximate the actual sampling distribution of the two-sample $t$-statistic.
- The degrees of freedom is not $n_1 + n_2 - 2$. It is smaller, and there are several possible formulae for calculating it. Typically we use either
  - the smaller of $n_1 - 1$ and $n_2 - 1$, or
  - the ludicrous formula on M&M p.536
- Note that smaller df generally corresponds to wider confidence interval or bigger critical value (i.e., harder to reject $H_0$).
Unequal Variances

M&M Examples 7.14, 7.18!