Outline

Review of Previous Lecture
  Confidence Interval

Significance Test
  Significance Test
  One-Sample Test for the Mean
  SAS PROC TTEST
Part I

Review of Previous Lecture
Confidence Intervals for the Mean

- **Known \( \sigma \) (population standard deviation):**
  \[
  \bar{x} \pm z^* \frac{\sigma}{\sqrt{n}}
  \]
  - Small \( n \), normal population.
  - Large \( n \), any population!

- **Unknown \( \sigma \):**
  \[
  \bar{x} \pm t^* \frac{s}{\sqrt{n}}
  \]
  - Small \( n \), normal population.
  - Large \( n \), near-normal population.
Example

- The mean birth weight of 78 SIDS (Sudden Infant Death Syndrome) cases in King County, WA, 1976-77, was 2994 g. The population standard deviation is assumed to be 800 g.
- The standard error is $800 / \sqrt{78} \approx 91$, so the 95% CI is

  $$(2994 - 1.96 \times 91, 2994 + 1.96 \times 91).$$

- **NOW**: Suppose that in the general population, the true mean birth weight is about 3300 g. Is the sample mean of 2994 consistent with this value?
Motivating example: SIDS Data

Note that:

- The sample mean is more than three standard errors away from the population average.
- The sample mean has an approximate normal distribution.

Conclusion: Either

- We have seen a very unusual event, OR
- Our collection of SIDS cases is not a random sample from the general population.
Hypothesis Testing

Definition
Often we wish to test whether the data provide *significant* evidence *against* some statement about the population (model) from which the data are sampled. The statement in question is called the *null hypothesis*, and is often denoted by $H_0$.

Example
The null hypothesis in the SIDS example is:

The population of SIDS cases has mean 3300, 

or, in statistical notation, 

$$H_0 : \mu = 3300.$$
About $H_0$, the Null Hypothesis

- $H_0$ is often formulated as an equality: “no difference”, “no effect”, etc.
- $H_0$ is always formulated with respect to population parameters.
- We are not trying to prove $H_0$, rather, we formulate $H_0$ such that we can disprove it.
- If there is “significant” evidence against $H_0$, we reject the null hypothesis, otherwise we fail to reject the null hypothesis.
- It is an oddity of classical hypothesis testing that one can never “accept” the null – turns out this is another reason to be Bayesian instead!
Alternative Hypothesis

Definition
When testing a null hypothesis, we also specify an *alternative hypothesis* ($H_1$ by convention) that in some sense contradicts the null hypothesis.

Example
Possible alternative hypotheses for the SIDS example:

- $H_1 : \mu \neq 3300$
- $H_1 : \mu < 3300$
- $H_1 : \mu > 3300$
About $H_1$, the Alternative Hypothesis

- There are often many ways of specifying the alternative.
- $H_1$ is not necessarily the complement of $H_0$. In other words, when $H_0$ is false, $H_1$ is not necessarily true (e.g., the “one-sided” alternatives $H_1 : \mu < 3300$ and $H_1 : \mu > 3300$).
- Again, we cannot prove the alternative hypothesis is true.
- Often $H_1$ is the statement that we hope is true.
Definition
When \( H_0 \) is a point null hypothesis (e.g., \( \mu = \mu_0 \)), the alternative can be either one-sided (e.g., \( \mu > \mu_0 \)) or two-sided (e.g., \( \mu \neq \mu_0 \)).

Tip: Always use a two-sided alternative unless you have a very good scientific reason not to.

Reason: We do not want to fail to reject \( H_0 \) when there is significant evidence against it, but in the opposite direction from the alternative hypothesis (say, \( H_1 : \mu < \mu_0 \) but \( \bar{X} \) very very large)!
Intuition Behind Hypothesis Testing

▶ We construct a measure of the compatibility between the data and $H_0$.
▶ Compatibility is measured in terms of probability.
▶ We reject the null if the compatibility is very low.
Definition

A test statistic measures the compatibility between the null hypothesis and the data. It is a random variable with a known distribution (since it gets to assume $H_0$ is true).
Example

In SIDS example, for $H_0 : \mu = \mu_0 = 3300$, we measure compatibility in terms of number of standard errors separating $\bar{X}$ from $\mu_0$:

$$\frac{\bar{X} - \mu_0}{se(\bar{X})} = \frac{2994 - 3300}{91} \approx -3.4.$$ 

So the data mean is 3.4 standard errors away from (below) the hypothesized population mean – very surprising!
For testing the null hypothesis $H_0 : \mu = \mu_0$ when the population standard deviation $\sigma$ is known, we use the z statistic:

$$z = \frac{\bar{x} - \mu_0}{\sigma / \sqrt{n}},$$

which has approximately a standard normal distribution (when $n$ is large).
Significance Test

One-Sample Test for the Mean
SAS PROC TTEST

$t$ statistic

For testing the null hypothesis $H_0 : \mu = \mu_0$ when population standard deviation $\sigma$ is not known, we use the $t$ statistic:

$$t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}},$$

which has approximately a $t_{n-1}$ distribution (when $n$ is large and the population is not too far from normal). As usual, $s$ is the sample standard deviation.
Intuition Behind Hypothesis Testing

- We construct a measure of the compatibility between the data and $H_0$.
- Compatibility is measured in terms of probability.
- We reject the null if the compatibility is very low.
Definition
Assuming $H_0$ is true, the probability that the test statistic would take a value as extreme or more extreme than that observed is called the p-value. The smaller the p-value, the stronger the evidence against $H_0$ provided by the data. Here, “extremeness” is defined with respect to the alternative hypothesis.

Brad’s definition: A p-value is the probability of getting something as surprising as you got, or more so.
Example

The test statistic is \( z = -3.4 \). Thus for the one-sided alternative \( H_1 : \mu < 3300 \), the \( p \)-value is:

\[
p = P(Z < -3.4) \approx 0.0003 .
\]

For the two-sided \( H_1 : \mu \neq 3300 \), the \( p \)-value becomes

\[
p = P(|Z| > |-3.4|) = P(Z < -3.4 \text{ or } Z > 3.4) = 2 \times P(z < -3.4) = 0.0006 .
\]
Intuition Behind Hypothesis Testing

- We construct a measure of the compatibility between the data and $H_0$.
- Compatibility is measured in terms of probability.
- We reject the null if the compatibility is very low.
Significance of a Test

Definition
We often pre-specify a significance level $\alpha$, e.g., $\alpha = 0.05$. When $p < \alpha$ we say that the test is statistically significant at level $\alpha$.

So:

\[ p < \alpha \implies \text{Reject } H_0 \]
\[ p > \alpha \implies \text{Fail to reject } H_0 \]

BUT: Despite this, $p$ is NOT “the probability that $H_0$ is true.” It is merely the (long-run) probability of getting something as surprising as you got, or more so.

– analogy with the long-run interpretation of CI’s!
Caution about Reporting

- The prespecified significance level (typically $\alpha = 0.05$) is somewhat arbitrary. Do not interpret the binary answer (significant or not) too strictly (e.g., $p = 0.049$ and 0.051 are not really that different)!

- Always report the exact $p$-value (e.g., $p = 0.0003$). Do not use “*”, “**”, “***”, etc. to indicate significant results.

- Whenever possible, also report a confidence interval based on the test statistic (e.g., $(2700, 3200)$).
Confidence Intervals vs. Significance Tests

Fact

A level-$\alpha$ two-sided significance test rejects a null hypothesis $H_0 : \mu = \mu_0$ if and only if the value $\mu_0$ falls outside of a level $1 - \alpha$ confidence interval for $\mu$.

Proof of this fact is on the next slide...
Confidence Intervals vs. Significance Tests

Fail to reject $H_0$ at level $\alpha$ if and only if

$$\Pr \left( |Z| > \left| \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}} \right| \right) > \alpha \iff 2 \Pr \left( Z > \left| \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}} \right| \right) > \alpha$$

$$\iff \Pr \left( Z < \left| \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}} \right| \right) < 1 - \alpha/2$$

$$\iff \frac{|\bar{x} - \mu_0|}{\sigma/\sqrt{n}} < z_{1-\alpha/2}$$

$$\iff -z_{1-\alpha/2} < \frac{(\bar{x} - \mu_0)}{\sigma/\sqrt{n}} < z_{1-\alpha/2},$$

which means $\bar{x} - z_{1-\alpha/2}(\sigma/\sqrt{n}) < \mu_0 < \bar{x} + z_{1-\alpha/2}(\sigma/\sqrt{n}),$

or that $\mu_0$ is in the CI.
Summary: Confidence and Testing

100(1 − \(\alpha\))% CI for the mean:

\[
\left( \bar{x} - z^* \frac{\sigma}{\sqrt{n}}, \bar{x} - z^* \frac{\sigma}{\sqrt{n}} \right)
\]

where recall \(z^* = z_{1-\alpha/2}\) = the upper \(\alpha/2\) point of a \(N(0, 1)\) random variable.
Summary: One Sample Inference for the Population Mean

We have drawn an SRS of size $n$ from a population with unknown mean $\mu$. We want to test:

$$H_0 : \mu = \mu_0 .$$

The alternative hypothesis can be either one-sided:

$$H_1 : \mu < \mu_0 \quad \text{or} \quad H_1 : \mu > \mu_0 ,$$

or two-sided:

$$H_1 : \mu \neq \mu_0 .$$
Test Statistics

The test is based on $\bar{x}$. When $\sigma$ is known, we use the $z$-statistic:

$$z = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}}.$$

When $\sigma$ is not known, we replace it with $s$ and use the $t$-statistic:

$$t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}}.$$
Under the null hypothesis (i.e., the population mean $\mu = \mu_0$), when

(a) the population is normal, or

(b) $n$ is large and the population is not too different from normal,

then $z$ has a $\mathcal{N}(0, 1)$ distribution and $t$ has a $t_{n-1}$ distribution.
Computing the \( p \)-value

- For \( H_1 : \mu \neq \mu_0 \): \( p = P(|T| > |t|) = 2P(T > t) \).
- For \( H_1 : \mu > \mu_0 \): \( p = P(T > t) \).
- For \( H_1 : \mu < \mu_0 \): \( p = P(T < -t) \).

This procedure is often called one-sample *t*-test or one-sample *z*-test, depending on whether we’re using *z* or *t*.
data radon;
    input reading @@;
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 = radon h0=105 alpha=0.05;;
    var reading;
run;
PROC TTEST (V8.0+)

Statistics

<table>
<thead>
<tr>
<th>Variable</th>
<th>N</th>
<th>Lower CL Mean</th>
<th>Upper CL Mean</th>
<th>Lower CL Std Dev</th>
<th>Upper CL Std Dev</th>
</tr>
</thead>
<tbody>
<tr>
<td>reading</td>
<td>12</td>
<td>98.162</td>
<td>104.13</td>
<td>6.6571</td>
<td>9.3974</td>
</tr>
</tbody>
</table>

Statistics

<table>
<thead>
<tr>
<th>Variable</th>
<th>Upper CL Std Dev</th>
<th>Std Err</th>
<th>Minimum</th>
<th>Maximum</th>
</tr>
</thead>
<tbody>
<tr>
<td>reading</td>
<td>15.956</td>
<td>2.7128</td>
<td>91.9</td>
<td>122.3</td>
</tr>
</tbody>
</table>
T-Tests

| Variable | DF | t Value | Pr > |t| |
|----------|----|---------|------|---|
| reading  | 11 | -0.32   | 0.7554 |
Let’s Do an Example or Two

MM pp.455-456!