Introduction to Hypothesis Testing
(Extensions of the on-screen presentation of 11/19/03. Also minor corrections.)

Suppose we are using a pH meter to test acidity/alkalinity of a solution used in a manufacturing process.

- The pH of pure water is 7.00.
- For samples in the pH range from about 5 to 9 the meter gives normally distributed readings with
  - \( \mu \) equal to the true pH of the sample and
  - standard deviation \( \sigma = 0.20 \).
- The meter is tested each morning to make sure it is operating properly: \( n = 4 \) samples of pure water are tested.

If the average of the four readings is "sufficiently near" 7.00 the meter is deemed to be in proper adjustment, otherwise it is serviced before further use.

The issue to be decided on a particular day is whether the meter is

- OK (giving proper readings) or
- NOT OK (needs adjustment before use).

Setting up the hypotheses to settle the issue:

- Our null hypothesis is \( H_0: \mu = 7.00 \) (meter is properly adjusted) and
- Our alternative hypothesis is \( H_a: \mu \neq 7.00 \) (meter needs service).

NOTES:

1. The null and alternative hypotheses involve the parameter \( \mu \) (not the statistic \( \bar{X} \)).
2. The null hypothesis \( H_0 \) always contains an equal sign.
3. The null hypothesis \( H_0 \) specifies a population distribution—here \( N(7.00, 0.20) \).
4. Knowing whether \( H_0 \) is true tells us whether to adjust the meter.
Performing the significance test:

We will use the mean $\bar{X}$ of the four trial measurements with pure water as the estimate of $\mu$.

(In this situation there is no need to estimate $\sigma$ because we know that $\sigma = 0.20$ is a property of the meter.)

Using standardized values, we judge the meter to be in proper adjustment if

$$Z = \frac{\bar{X} - \mu_0}{\sigma / \sqrt{n}} = \frac{\bar{X} - 7.00}{0.20 / \sqrt{4}} = \frac{\bar{X} - 7.00}{0.10}$$

is near 0.

If $H_0$ is true, then $Z$ has a standard normal distribution:

- $Z$ should "usually" be between ±1.96 (probability 95%), and
- $Z$ should "rarely" fall outside this interval, less than −1.96 or greater than +1.96 (probability $\alpha = 5\%$).

Thus we

- "Reject" $H_0$ (and service the meter) if $|Z| \geq 1.96$,
- "Accept" $H_0$ (assume the meter to be well enough adjusted) if $|Z| < 1.96$.

We say that the null hypothesis is tested at the 5% level because that is the probability of rejecting a true $H_0$ (uselessly servicing a properly adjusted meter).
**Day 1:** Suppose that, in four runs with pure water, the meter gives readings

6.9, 7.1, 7.4, 7.0.

Is there evidence that the meter is out of adjustment?

For these data \( \bar{X} = 7.10 \), so

\[
Z = \frac{\bar{X} - \mu_0}{\sigma / \sqrt{n}} = \frac{7.10 - 7.00}{0.20 / \sqrt{4}} = \frac{0.10}{0.10} = 1.0
\]

and \(|Z| = 1.0 < 1.96.\)

\( H_0 \) is accepted at level \( \alpha = 5\% \).

We will not service the meter.

In statistical language we say that \( \bar{X} \) does not differ "significantly" from 7.0.

(This does not mean that we believe \( \mu = 7.0 \) *exactly*, only that \( \mu \) does not detectably differ from 7.0 based on available information.)
**Day 2:** On another day, suppose the readings with pure water are 6.8, 6.6, 7.0, 6.6, giving \( \bar{X} = 6.75 \). Then

\[
Z = \frac{\bar{X} - \mu_0}{\sigma / \sqrt{n}} = \frac{6.75 - 7.00}{0.20 / \sqrt{4}} = \frac{-0.25}{0.10} = -2.50
\]

so that

\[ |Z| = |-2.50| = 2.50 > 1.96. \]

- We reject \( H_0 \) at the 5% level,
- Say that \( \bar{X} \) differs significantly from 7.0, and
- Service the meter.

**Exercise:** Suppose we want to test at level \( \alpha = 1\% \). Then show that the "critical" value on the \( Z \) scale (the value that separates accepting and rejecting) would have to be 2.576 instead of 1.96. In plain English, we would be less likely to service a properly adjusted meter. On Day 2, would \( H_0 \) be rejected at the 1% level?

(Why might we use level \( \alpha = 1\% \) instead of 5%? If it is really expensive to service the meter, we might want to reduce the probability of servicing the meter needlessly. But what bad consequence might come from this change to a lower significance level?)
**P-values.**

Assuming H₀ to be true, a P-value is the probability of a more extreme result than the one obtained.

In our case this means:

- A value of \( \bar{X} \) farther from 7.0 than the observed value of \( \bar{X} \) observed in the experiment or
- A value of |Z| farther from 0 than the computed value of the Z statistic.

On Day 1: The P-value is

\[
P\left\{ \left| \frac{(\bar{X} - 7.00)}{0.10} \right| > 1.0 \right\} = P\{|Z| > 1.0\} = P\{Z < -1.0\} + P\{Z < 1.0\} = 2(0.158655) = 0.3173.
\]

(The tail area 0.158655 was obtained from statistical software; you can obtain a similar value from normal tables.)

On Day 2: From a similar computation the P-value is

\[
P\{|Z| > 2.50\} = 0.0124.
\]

If you know the P-value, then you can test at any desired level.

At the level \( \alpha = 5\% = 0.05 \):

- The P-value on **Day 1** exceeds 5%, so we **accept** H₀, but
- The P-value on **Day 2** is smaller than 5%, so we **reject** H₀.

- At level \( \alpha = 1\% = 0.01 \):
  The P-values on both days (0.3173 and 0.1124) exceed 1%, so we **accept** H₀ in both cases.
**Minitab and P-values**

Most statistical software, including Minitab show P-values in their output instead of declaring whether to Accept or Reject $H_0$.

Below we put the pH readings from Day 1 into c1 and the results from Day 2 into c2. Here are Minitab printouts based on these data.

```
MTB > set c1
DATA> 6.9, 7.1, 7.4, 7.0.
DATA> end
MTB > OneZ c1;
SUBC> sigma .2;
SUBC> test 7.00.
```

**One-Sample Z: Day1**

Test of mu = 7 vs mu not = 7
The assumed sigma = 0.2

<table>
<thead>
<tr>
<th>Variable</th>
<th>N</th>
<th>Mean</th>
<th>StDev</th>
<th>SE Mean</th>
</tr>
</thead>
<tbody>
<tr>
<td>Day1</td>
<td>4</td>
<td>7.100</td>
<td>0.216</td>
<td>0.100</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Variable</th>
<th>95.0% CI</th>
<th>Z</th>
<th>P</th>
</tr>
</thead>
<tbody>
<tr>
<td>Day1</td>
<td>(6.904, 7.296)</td>
<td>1.00</td>
<td>0.317</td>
</tr>
</tbody>
</table>

Notice that:

- The values $n = 4$, $\bar{X} = 7.10$, and $Z = 1.0$ are the same as calculated above. $H_0$: $\mu = 7.00$ is accepted at level $\alpha = 5\%$ because $|Z| = |1.00| = 1.00 < 1.96$.
- The $P$-value also agrees with what we obtained above (rounded here to three places). $H_0$: $\mu = 7.00$ is accepted at level $\alpha = 5\%$ because the $P$-value $0.317$ exceeds $5\% = 0.05$.
- The 95% confidence interval is a span of "believable values" of the population mean, or a span of "acceptable values $\mu_0$."
  Thus $H_0$: $\mu = 7.00$ is accepted because 7.00 lies in the interval.
  (Caution: This use of confidence intervals works only for two-sided alternatives such as $H_a$: $\mu \neq 7.00$.)
- $H_0$ is also accepted at any level lower than 5%, for example 1%. 


Now for Day 2.

MTB > set c2
DATA> 6.8, 6.6, 7.0, 6.6.
DATA> end
MTB > OneZ c2;
SUBC> sigma .2;
SUBC> test 7.00.

One-Sample Z: Day2

Test of mu = 7 vs mu not = 7
The assumed sigma = 0.2

<table>
<thead>
<tr>
<th>Variable</th>
<th>N</th>
<th>Mean</th>
<th>StDev</th>
<th>SE Mean</th>
</tr>
</thead>
<tbody>
<tr>
<td>Day2</td>
<td>4</td>
<td>6.750</td>
<td>0.191</td>
<td>0.100</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Variable</th>
<th>95.0% CI</th>
<th>Z</th>
<th>P</th>
</tr>
</thead>
<tbody>
<tr>
<td>Day2</td>
<td>( 6.554, 6.946)</td>
<td>-2.50</td>
<td>0.012</td>
</tr>
</tbody>
</table>

Notice that:

- The values $n = 4$, $\bar{X} = 6.75$, and $Z = -1.0$ are the same as calculated above.
  - $H_0: \mu = 7.00$ is rejected at level $\alpha = 5\%$ because $|Z| = |-2.50| = 2.50 > 1.96$.
  - $H_0: \mu = 7.00$ is accepted at level $\alpha = 1\%$ because $|Z| = |-2.50| = 2.50 < 2.576$.
- The $P$-value also agrees with what we obtained above (rounded in the printout to three places).
  - $H_0: \mu = 7.00$ is rejected at level $\alpha = 5\%$ because the $P$-value 0.012 is smaller than 5% = 0.05.
  - $H_0: \mu = 7.00$ is accepted at level $\alpha = 1\%$ because the $P$-value 0.012 exceeds 1% = 0.01.
- The 95% confidence interval is a collection of "believable values" of the population mean, or a collection of "acceptable values $\mu_0$" at the 5% level. Thus $H_0: \mu = 7.00$ is rejected at 5% because 7.00 lies outside the interval. (Because this is a 95% confidence interval, we cannot use it to test $H_0$ at level $\alpha = 1\%$).
**Student's t distribution** (to be covered later for significance tests).

In most applications, the population standard deviation $\sigma$ is unknown. Then we use the sample standard deviation $s$ to estimate $\sigma$, and compute the test statistic

$$t = \frac{\bar{X} - \mu_0}{s/\sqrt{n}} = \frac{\bar{X} - 7.00}{s/\sqrt{4}}.$$

With $n = 4$, $df = n - 1 = 3$, and the critical value $t^* = 3.182$ for a 5% level test (see tables: which row? which column?). The *critical value* is the value that separates accepting from rejecting.

- On Day 2, $s = 0.1915$ and $|t| = |–2.61| = 2.61$ (verify these values for yourself on a calculator).
- Thus, if $\sigma$ were unknown, we would accept $H_0$: $\mu = 7.0$ when tested against the alternative $H_1$: $\mu \neq 7.00$ at level $\alpha = 5\%$ on Day 2 because $|t| = |–2.61| = 2.61 < 3.182$.

Minitab output for this situation is shown below.

```plaintext
MTB > OneT c2;
SUBC> test 7.00.

One-Sample T: Day2

Test of mu = 7 vs mu not = 7

<table>
<thead>
<tr>
<th>Variable</th>
<th>N</th>
<th>Mean</th>
<th>StDev</th>
<th>SE Mean</th>
</tr>
</thead>
<tbody>
<tr>
<td>Day2</td>
<td>4</td>
<td>6.7500</td>
<td>0.1915</td>
<td>0.0957</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Variable</th>
<th>95.0% CI</th>
<th>T</th>
<th>P</th>
</tr>
</thead>
<tbody>
<tr>
<td>Day2</td>
<td>( 6.4453, 7.0547)</td>
<td>-2.61</td>
<td>0.080</td>
</tr>
</tbody>
</table>
```

Note: It is not possible to find the $P$-value for this $t$ test from the tables in your text. However, Minitab computes the $P$-value as $0.080 > 5\%$, so we know to accept $H_0$ based on the Minitab printout. (Also, notice that the 95% CI covers the value $\mu_0 = 7.0$. You should verify the CI for yourself.)
### Types of Error

<table>
<thead>
<tr>
<th>TRUTH</th>
<th>OUR DECISION</th>
<th>H₀ True</th>
<th>Hₐ True</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Accept H₀</td>
<td>Correct</td>
<td>Type II</td>
</tr>
<tr>
<td></td>
<td>Decision</td>
<td></td>
<td>Error</td>
</tr>
<tr>
<td></td>
<td>Reject H₀</td>
<td>Type I</td>
<td>Correct</td>
</tr>
<tr>
<td></td>
<td>Error</td>
<td></td>
<td>Decision</td>
</tr>
</tbody>
</table>

**Type I Error** = Reject H₀ when H₀ is True

= Service a pH meter that is OK.

**Type II Error** = Accept H₀ when Hₐ is True

= Use a meter that needs service.

\[ \alpha = P(\text{Type I Error}) = P(\text{Rej } H₀ \mid H₀ \text{ True}) \]

= Significance level

**Power** = 1 - P(\text{Type II Error})

---

*Power (Advanced, Optional)*: This discussion assumes that \( \sigma = 0.2 \) and refers to the z-test (not the \( t \)-test). Above, let \( \alpha = P\{\text{Reject } H₀ | \mu = 7.0\} = P(\text{Adjust meter | meter OK}) = P\{\text{Type I Error}\} = 5\% \). We want \( \alpha \) to be small to keep from making a Type I Error too often. However, another kind of error to be avoided is to use a meter that is out of adjustment, that is \( \mu = \mu_a \neq 7.0 \):

\[
\beta(\mu_a) = P\{\text{Accept } H₀ | \mu = \mu_a\} = P(\text{Use meter | meter not OK}) = P\{\text{Type II Error for } \mu_a\}
\]

Using \( \alpha = 5\% \) and recalling that \( \mu = 7.0 \) if H₀ is true, we accept when \( |Z| > 1.96 \). In terms of \( \bar{X} \) the "acceptance region" is \( 6.804 < \bar{X} < 7.196 \). If H₀ is not true and instead \( \mu = 7.5 \), then

\[
\beta(7.5) = P\{6.804 < \bar{X} < 7.196\} = P\{(6.804 - 7.5)/0.1 < Z < (7.196 - 7.5)/0.1\}
\]

\[
= P\{|-6.96 < Z < -3.04|\} = 0.0012.
\]

The power of a test for alternative \( \mu_a \) is defined as \( \pi(\mu_a) = 1 - \beta(\mu_a) \). So, \( \pi(7.5) = 0.9988 \). Thus if the meter is so badly out of adjustment that \( \mu_a = 7.5 \), then we are almost sure to reject H₀ and service the meter.