Hypothesis Testing

LIR 832
Lecture #3
Sept, 2008

Topics of the Day

◆ A. Our Fundamental Problem Again: Learning About Populations from Samples
◆ B. Basic Hypothesis Testing: One Tailed Tests Using a Z Statistic
  \[ \mu \geq 25 \text{ or } \mu \leq -6 \]
◆ C. Probability and Critical Cutoff Approaches: Really the Same Thing
◆ D. How do we do hypothesis tests on small samples (n = 30 or less).
◆ E. How do we do hypothesis testing when we have information on population standard deviation? On sample standard deviations?
◆ F. How do we test a statement such as (two tailed test)?
  \[ \bar{X} = \mu \]
◆ G. How do we test for differences in means of two populations?
Hypothesis Testing

- **Fundamental Problem:** We want to know about a population which is not observed and want to use a sample to learn about the population. Our problem is that sampling variability makes sample an inexact estimator of the population. We need to come up with a method to learn about the population from samples which allows for sampling variability.

Hypothesis Testing: Example

- We are considering implementing a training program which purports to improve quality and reduce the number of defects. Currently, 10 out of every 1000 parts produced are not within spec. The standard deviation of defects is 8 parts (variance is 64). The typical employee produces 1,000 parts per day.
- The program costs $1,000 per employee and we have 10,000 production employees. We are unwilling to spend $10,000,000 for a pig in the poke.
- Instead we decide to run a pilot on 100 employees to determine whether the program is effective for our employees. The firm that does the training will do the program for free, so our only cost is lost production for the time during which employees are trained. *Note that the 100 employees are a pilot or, in our terminology, a sample.*
Hypothesis Testing: Example

• Employees are sent to the program and then given several days under instruction to apply what they have learned to their work. We run a one day test on the employees and find that they average 8 parts per thousand.

• Defects are down, but is this really an improvement or is it simply the result of sampling variation? Could we be reasonably certain if we trained a second group of employees, or ran this same test next week, that defects would also be down?

Hypothesis Testing: Example

• Abstractly, we are faced with the problem of distinguishing whether the program is effective or whether the improvement is reasonably explained by sampling variation (aka luck).
Let’s approach this as a statistical problem. We know that historically there have been 10 defects per 1000 with standard deviation of 8. So our question is, “How likely is it that we have pulled a sample of 100 employees with a mean defect rate of 8 if, in fact, the training program did not work (in other words, that the population rate of defects remains 10 per 1000)?

You can set this problem up as you have been already:

\[
P(\bar{x} < 8) \text{ or } P(\bar{x} < 8, \mu \geq 10) = P\left(\frac{\bar{x} - \mu}{\sigma / \sqrt{n}} < \frac{8 - 10}{8 / \sqrt{100}}\right) = P(z < -2.5)
\]

As the sample is larger than 30, we can use our z-table. The \(P(z < 2.5) = 0.62\%\), very small. There is so small a likelihood that the sample we have observed was drawn from a population with a mean of 10 that we reject the possibility that the training program was ineffective.
Hypothesis Testing: Example

- What do I mean when I say that the probability is 0.62%?
  Two possible interpretations:
  1. Suppose we set up a population with a mean of 10 and a standard deviation of 8 and draw samples of 100 from that population. Now imagine repeating this experiment 1000 times. We would expect that slightly over 6 of those samples would have a value of 8 or less.
  2. Alternatively, if the population had a mean of 10, standard deviation of 8, we would expect that 0.62% of the time we would draw a sample values 8 or less by chance.

- I find the first approach makes it easier to understand what I mean when I say there is a 0.62% probability that an event occurred by chance, but you may prefer the second.

Hypothesis Testing: Example

- Thinking about this graphically:
  - Two extreme possibilities:
    1. Training did little or nothing
    2. The training was highly effective (really useful in Thomas the Tank Engine terminology)
Hypothesis Testing: Example

- If the training did nothing, we would expect the distribution of defects post training to look a lot like the distribution of defects prior to training.

So we can use the old distribution as a standard against which to judge our sample results. If the sample looks a lot like the old distribution it would be reasonable to believe that the training did not work.

1. For convenience, we add cut points at ± 2 standard deviations (1.6 = 2 * .8)
Hypothesis Testing: Example

- When we pull our sample, we calculate a mean and compare it to the old distribution. So, for example, our sample returns a value of 8. This lies to the left of the low cut point and we conclude it is very unlikely that the training did nothing.

Hypothesis Testing: Example

- Let’s go behind the graphs to the underlying, if unobserved population.
Hypothesis Testing: Example

Distribution of Defects if Training is Not Effective (Mean is 9.8)

What If the Defect Rate is Now 7 per 1000
Hypothesis Testing: Example

Step 1: State our beliefs about the world clearly (hypotheses).

- Example: The consulting firms contention is that their training program reduces defect rates. Another possibility is that the training program is ineffective and that any changes in defect rates are simply the result of sampling variability (randomness).
Step 2: Formalize this into the alternative and null hypothesis:

- Alternative: \( H_A : \mu_{\text{post training}} < 10 \): the training program is effective
- Null: \( H_0 : \mu_{\text{post training}} \geq 10 \): the training program has no effect on defect rates.

More about Step 2:

- The two hypotheses are about the unobserved population. We will use samples to test these two hypotheses.
- Together, the null and alternative cover all possible outcomes.
- The null is about change being the result of sampling variability, not the result of systematic difference. So our null for the training program is that it is not effective. Our alternative is that the program has a positive effect.
- We always test to determine whether the null is reasonable (is sufficiently likely that we are unwilling to conclude it is false).
Hypothesis Testing: Formalizing the Steps

- Step 3: Set up the probability problem:

\[ P(\bar{x} < 8 | \mu_0 \geq 10) \]

or, more typically

\[ P(\bar{x} < 8) \]

- The sign of the test is taken from the alternative hypothesis. We ask, how likely are we to observe this extreme an outcome if, indeed, the null is true.

- Note that our standard deviation is the root of the population variance divided by sample size. This is the second part of the Central Limit Theorem and provides the predictive power to use of sample means in hypothesis testing.

- The result of doing the test is a probability of the null being true given the sample outcome. If that probability is sufficiently small, we conclude the null is not reasonable.
Hypothesis Testing: Formalizing the Steps

Step 5: Decide whether the probability of the null being true, given the sample outcome, is sufficiently low to allow you to reject (not believe) the null.

- The 0.62% likelihood is the probability that a population with a mean of 10 and a standard deviation of 8 would produce a sample of 100 observations with a mean of 8. It is very unlikely and we may well decide the null isn’t realistic.

A Graphical Approach
Hypothesis Testing: Example

- *A second example:* We are considering the absence behavior of a new group of workers and wish to determine if it is different from the behavior of our typical company employee.

- The company has hired a large group of human resource managers over the last three years. They are well versed in corporate policy and, as they do many of the unpleasant tasks in the firm, tend to have high stress levels. The Vice President of Human Resources has remarked that she is concerned about their stress levels and believes that our new HR managers they are taking more days off than most employees to compensate for their problems.

- We collect data on the 225 HR managers in the firm. We find they average 24 days of sick leave annually. The average managers in the firm has taken 22 days of sick leave annually over the last ten years. The standard deviation for the pool of typical managers is 30 days. Is our VP correct that our HR managers take more sick leave than other managers?
Step 1: State our views of the world.

- One view is that HR managers are taking more days of sick leave than has been typical of other managers.
- Another is that they are, as a group, no different than other managers.

Step 2: Formalize the hypotheses we are testing.

- Let $\mu_d$ be the population mean days off for the population of HR managers:
  - $H_0$: Null: HR managers are not different:
    - (NOTE: the null is about the population mean, not the sample)
    - $\mu \leq 22$
  - $H_A$: Alternative: HR managers take more sick time than other managers
    - (Similarly, the alternative is about the population)
    - $\mu > 22$
Step 3: Set up the probability problem.

- Note that our sample is 225, so we have central limit theorem results and can use a z distribution.

\[
P(\bar{x} > 24 | \mu_{H_0} \leq 22)
\]

or

\[
P(\bar{x} > 24)
\]

Step 4: Solve the probability problem.

\[
P(\bar{x} > 24) = P\left(\frac{\bar{x} - \mu}{\sigma / \sqrt{n}} > \frac{24 - 22}{\sqrt{900/225}}\right)
\]

\[
= (P\left(\frac{\bar{x} - \mu}{\sigma / \sqrt{n}} > \frac{2}{30/15}\right)
\]

\[
= P(z > 2 / 2)
\]

\[
= P(z > 1.00)
\]

\[
= 15.87\%
\]
Step 5: Determine whether the likelihood is sufficiently small to reject the null.

- Our calculations indicate that there is a 15.87 percent likelihood that a population with a mean of 22 would produce a sample of 225 individuals with a mean of 24. So, we would expect a sample of the type we got in about 1 in 6 experiments when HR managers were no different than other managers. That type of outcome just isn’t too unlikely.

There are, again, two ways of interpreting this result:

1. (Repeat experiment approach) If we ran 100 experiments in which we pulled random samples of 225 from a population in which the mean of 22 and a standard deviation of 30, between 15 and 16 of those samples would have a mean of 24 or greater.

2. (Probability approach) Alternatively, 15.87% of the time when we pull a sample of 225 from a population with a mean of 22 and standard deviation of 30, the mean will be 24 or greater.
Hypothesis Testing

- Now consider what we are doing abstractly…

- A little background. Employers are often very interested in employees views of their work and firm. There are many ways of measuring employee attitudes, a common method used in survey research are Likert scaled questions. For example, we might ask a question such as:
  - How would you rate this job? Would you rate it as
    - 1 - very bad
    - 2 - bad
    - 3 - good
    - 4 - very good

Hypothesis Testing

- A Somewhat Cranky Aside:
  - The response to this question is numeric, we get a number between 1 and 4, but it is ordinal data. We know the ordering of the response, a response of very good is stronger than a response of good; similarly a response of bad is weaker than a response of good.; but there is no arithmetic relationship between the responses. Two responses of “very bad” are not equal to one response of “bad”. Indeed, we cannot even be sure that the responses are equally spaced. The person who responds “bad” may be mildly unhappy relative to someone who responds good, but the person who responds “very bad” could be somewhat more unhappy than the “bad” or they could just about ready to “go postal”. Still the data is ordered.
  - Despite this limitation, it is common practice in behavior sciences to add the responses to form a mean response.
Back to our problem:

1. We begin with a “theory” or observation about the population. These are typically general statements such as:
   
   - a. Employees on evening or night shifts are less satisfied with their work than other workers.

2. Implicit in this statement is its opposite, the null hypothesis. This is what we are going to test
   
   - b. Null: Employees on evening or night shifts are no less satisfied with their work than employees on the day shift.

3. We turn these statements into operational hypotheses, statements which can be tested statistically.

   - c. We have been surveying employees for many years using Likert scaled surveys. We know from these surveys that day shift employees rate the firm as good (3) on a four point Likert scale (very bad, bad, good, very good). Standard deviation is also 3. Our null and alternative would be:
     
     - \( H_0: \) Null: \( \mu_{\text{evening or night}} \geq 3 \)
     - \( H_A: \) Alternative: \( \mu_{\text{evening or night}} < 3 \)

     *Note: The null and alternative exhaust all possibilities, no other outcome is possible.*
4. We pull a sample and test to determine the probability that the sample came from the population with the characteristics of the null.

- Suppose we pull a sample of 36 night employees with a sample mean of 2.3
- We ask:
  - What is the probability of pulling this particular sample if the mean response of our night shift employees really is 3.0 or greater?
    - $P(x\text{-bar} < 2.3 \text{ if } \mu_{\text{night}} \geq 3)$ or $P(x\text{-bar} < 2.3 \mid \mu_{\text{night}} \geq 3)$?
    - Note: the sign in the statement is taken from the alternative hypothesis.

5. We then calculate that probability:

\[
p(x < 2.3 \mid \mu_{\text{night}} \geq 3) = \Phi \left( \frac{\bar{x} - \mu_{\text{night}}}{\sigma / \sqrt{n}} \right) = \Phi \left( \frac{2.3 - 3}{\sigma / \sqrt{36}} \right) = \Phi \left( \frac{-0.7}{0.5} \right) = \Phi (-1.4) = 0.0808
\]
6. What does this 8.08% mean:
   - 1. Suppose we had a population with a mean value of 3 and standard deviation of 3 (this is exactly the population we build our null hypotheses around). Then suppose we ran an experiment in which we drew 36 individuals at random from that population and repeated the experiment 100 times. **We would expect that 8.0 of those samples would have a mean of 2.3 or less!**
   - 2. If our population has a mean and standard deviation of 3.0, we would expect that 8.08% of the time we would draw a sample of 36 individuals with a mean of 2.3 or less.

Criteria for rejecting or not rejecting the null:
   - We reject the null if it is sufficiently unlikely that the sample would be produced by the “null” population by chance.
   - We do not reject the null if the probability that the observed sample was pulled from the null population is reasonably large.
   - So our logic is that we only believe our alternative if the probability of the reverse (the null) is very low.
   - **We have not established how low one has to go (how low the probability has to be) before one rejects the null. We are getting to this.**
Hypothesis Testing: Example

- We are told by a consultant that their program will, by making workers more aware of the effects of their absence, reduce the use of sick leave and personal leave. We have no reason to disbelieve her, but given the price of the program $2,500 per employee, we want to run a pilot on 50 employees before taking it to the entire population of employees. Reviewing our records, we find that, on average, employees take 10 days of sick leave and personal leave a year. The standard deviation in leave time is 8.5 days per year.

Hypothesis Testing: Example

- Set up the null and alternative and test this problem.
  1. Let’s define μ as the number of absences due to sick leave or personal days.
  2. What is the null and alternative?
  3. You run the pilot and find that the mean number of sick and personal days taken was 8.7. Set up the hypothesis test.
  4. What is the z score from this test? What is the probability?
  5. Do you reject the null? Explain.
We use an examination for testing new employees. We need to calibrate our criteria for passing this test to the performance of our current employees with whom we are very satisfied.

The typical standard for hiring a person who has taken this test is a score of person hired after taking this test scores 500. We believe that our employees are superior and that, if we gave this test to our current employees, the mean of the test would be more than 500. The company which designs the test cannot help us directly as, although they know the typical performance of a test taker, they do not know the typical performance of our employees. They however, agree to administer the test to a sample of 24 of our employees. We are interested in using this sample to test the hypothesis that the mean score for our employees would be more than 500. The standard deviation of the sample, after allowing for sample size, is 4. Once we know this result, we can decide whether we can default to the typical performance standard or need to more closely determine our employees performance.

Suppose we want to test our belief about the population and are willing to be wrong by chance about 5% of the time. We know that 5% of the area of a normal is 1.645 standard deviations above the means, so we set up an upper cut point at +1.645 standard deviations (which is 506.58 = 1.645*4+ 500). What does this look like graphically?

What does this test look like?
Note that the limits for accepting the null are set around the hypothesized population.

Graphical Example
Graphical Example

Establishing Cut Points

- Establishing the Appropriate Cut-offs for rejecting the null hypothesis.
  - To this point, we have produced p-values from our hypothesis tests and then tried to judge whether a null was sufficiently unlikely that we were willing to reject it.
  - A more conventional approach is it use 10%, 5% and 1% as the standards
**Establishing Cut Points**

<table>
<thead>
<tr>
<th>P value or Level of Significance</th>
<th>z-value or critical value</th>
<th>Meaning</th>
</tr>
</thead>
<tbody>
<tr>
<td>10%</td>
<td>1.28 (to 1.644)</td>
<td>a population with a mean equal to that of the null would generate the observed result <em>by chance</em> in one of every ten samples:</td>
</tr>
<tr>
<td>5%</td>
<td>1.645 (to 2.32)</td>
<td>a population with a mean equal to that of the null would generate the observed result <em>by chance</em> in one of every twenty samples</td>
</tr>
<tr>
<td>1%</td>
<td>2.33 (and up)</td>
<td>a population with a mean equal to that of the null would generate the observed result <em>by chance</em> in one of every one hundred samples</td>
</tr>
</tbody>
</table>

Note that the 10%, 5% and 1% are referred to as the “level of significance.”

- Use of these percentile cut points is arbitrary but conventional.
  - *Example:* In court testimony we are allowed to reject the null if the likelihood of the observed result being generated by a population with the hypothesized mean is 5% or less.
Establishing Cut Points: Example

- Your firm has established that each department should have an interview to hire ratio of 5 to 1 or lower. Over the last two months, your shop has hired 49 new employees and have averaged 5.7 interviews per hire. As part of a quarterly review of human resource performance, you have been called on the carpet for having too high a ratio of interviews to hires. You have been told that your shop needs to rethink how you hire in order to meet the company target.

Establishing Cut Points: Example

- It is your view that your methods are good but that special circumstances required additional interviews and that, if you continue as you have been doing, your interview to hire ratio will return to 5 to 1 or lower. Before you argue this point, you would like to perform a hypothesis test to determine whether the argument, that your “population” interview to hire ratio is 5 to 1, is reasonable. The standard deviation of interview to hire ratios for the firm is 3.
Establishing Cut Points

Step 1: General issue
- Is our interview to hire ratio above the firms target ratio of 5 to 1, or is our 5.7 average reasonably explained by “chance”?

Step 2: Formalize the hypothesis test
- $H_A: \mu > 5$ aka: something is wrong
- $H_0: \mu \leq 5$ aka: it was just one of those two month periods
- $\sigma = 3$
- $n = 49$
Establishing Cut Points: Example

- Step 3: Set up the probability problem.

\[
P(x > 5.7 | \mu \leq 5) \\
\text{or} \\
P(x > 5.7)
\]

- Step 4: Solve the problem.

\[
P(x > 5.7) \\
P\left( \frac{x - x_{\text{HO}}}{\sigma / \sqrt{n}} > \frac{5.7 - 5.0}{3 / \sqrt{49}} \right) \\
= P(z > \frac{-7}{4.29}) \\
= P(z > 1.63) \\
P\text{ - value is } 5.16\%
\]
Establishing Cut Points: Example

- Step 5: See how things lie relative to the critical cut-offs
  - The 10% z critical value is 1.28
  - The 5% z critical value is 1.645
  - The 1% z critical is 2.33

- Our calculated z is 1.63 so we can reject in a 10% test, but cannot quite reject in a 5% test.

Figure 11: A Graphical Approach to Hypothesis Testing
Example (to be reviewed outside of class): We are concerned that one of our plants is not producing at its expected level. In the last two months output has been down and we are concerned this may reflect insufficient effort on the part of managers. Before going in and raising hell, we would like to check and see whether the downward fluctuation is reasonably explained as normal variance. Several years of production records indicate that our typical daily output is 100 units with a standard deviation of 10 units. Is it reasonable to believe that the last two months are explained as normal variance, or are they best explained as the result of something else?

Solution:
- $H_A: \mu < 100$ in last two months (something happened)
- $H_O: \mu \geq 100$ (It was just normal monthly variance in output)
- We check our records for the last two months and find that, in the 60 days the plant was operating (it was a leap year), we produced an average of 97 units per day.
Hypothesis Testing: Example

\[ P\left(\frac{x - \mu_{H_0}}{\sigma/\sqrt{n}} < \frac{97 - 100}{10/\sqrt{60}}\right) \]

\[ P(z < -3/1.29) \]

\[ P(z < -2.32) \]

= 1.02%

or, -2.32 falls between the critical point for 5% (1.645) and 1% (2.33) it's almost 1% but not quite.

So the likelihood that we would observe an average of 97 units if there wasn’t some exceptional condition is about 1%, fairly unlikely.
Additional Issues in Hypothesis Testing

- 1. How do we do hypothesis tests on small samples (n = 30 or less)?
- 2. How do we do hypothesis testing when we have information on population standard deviation? On sample standard deviations?
- 3. How do we test a statement such as $X = \mu$ rather than $X \geq \mu$ or $X \leq \mu$.
  - Alternatively, how do we do two tailed tests?
- 4. How do we test for differences in means of two populations?

Hypothesis Testing with Small Samples (n<30)

The Issue:
- A. CLT tells us that $x\bar{\text{-bar}} \sim N(\mu, \sigma^2/n)$ if
  - a. $n \geq 31$ or
  - b. $x \sim N(\mu, \sigma^2)$
- B. What do we do with samples of under 31?
  - a. $x$-bar is no longer $N(\mu, \sigma^2/n)$, rather the distribution is more spread out.
  - b. The shape of the distribution now depends on the number of observations. We are going to need the equivalent of a z-table for each size of sample 30 or under.
- C. Also need a new concept: degrees of freedom. For the type of test we are doing, this is the number of observations in the sample minus one.
  - DF = n - 1
Hypothesis Testing with Small Samples ($n \leq 30$)

As illustrated, the $t$ distribution is flatter than the $z$, there is more area in the tails. The shape of the $t$ also depends on the size of the sample. As $n$ gets to 30, the $t$ distribution looks very similar to the $z$ distribution. As $n$ gets close to 0, the $t$ distribution becomes very flat.

\[
\frac{x - \mu_H}{\sqrt{\sigma^2 / n}} \sim t(n - 1)
\]

where

$n$ is the size of the sample

and

$n - 1$ are the degrees of freedom
Using the t Distribution

- Example: We are trying to determine the starting wage to pay workers in our fast food restaurant. We have discussed this with several other owners of restaurants, and have been lead to believe that the typical worker is paid $7.00 per hour but we are beginning to suspect it is more. So we want to do a small survey, conduct a hypothesis test and figure out what we should be paying.
The following hypothesis is given:

- $H_A: \mu > 7.00$
- $H_O: \mu \leq 7.00$

We take a sample of 10 observations and get an x-bar of 9 with sample standard deviation of 3 (variance of 9)

Note, we will only reject the null if x-bar is sufficiently far above the mean (in the upper tail)

Using the t Distribution

$$P\left(\frac{\bar{x} - \mu_{H_O}}{\sqrt{s^2/n}} \geq \frac{9 - 7}{\sqrt{9/10}}\right)$$

$$= P(t(10 - 1) \geq 2.11)$$

$t(9)$ for a 10% one tailed test is 1.383

$t(9)$ for a 5% one tailed test is 1.833

$t(9)$ for a 1% one tailed test is 2.821

reject $H_O$ at the 10% and 5% level
Hypothesis Testing with Small Samples ($n \leq 30$)

- Lets look at how we work with small samples:
  a. We set up our hypothesis the same way as with large samples.
  b. We set up our problem the same way as we do with large samples but substitute $t(n - 1)$ for $z$.
  c. We get our statistic through the $z$ transform the same way as we did with large samples.
  d. We however have to use a $t$-table for a $t$-statistic rather than our $z$-table.
    i. Our $t$-table is set up in terms of critical points for $x\%$ tests. So once we have calculated our $t$, we go to the table to determine whether the null can be rejected at the $10\%$, $5\%$, $2.5\%$, $1\%$, $0.5\%$ or more rigorous test.
    ii. Need to be careful about degrees of freedom. ($n - 1$)
  e. Note how similar $t$ and $z$ statistics are once we have 31 observations. Differences are not very great so we might as well use a $z$ statistic.

Example (to be reviewed outside class): The magazine *Technical Analysis of Stocks and Commodities* is filled with advertisements for trading software. Each software package promises “untold profits” and “money making opportunities.” Suppose one package promises 35% profits on options trading. Following are four observation of the trading grains that would have resulted had the software been used:
- 53%  28%  34%  19%

- Can we accept the claim that use of the software will result in an average return of 35% per annum or better? We believe that the claim is false.
Hypothesis Testing with Small Samples (n<30)

- Step 1: State the alternative: $H_A: \mu < 35$
- Step 2: State the null: $H_0: \mu \geq 35$
- Step 3: Have to use the t because the sample is small (n = 4)
- Step 4: Calculate the t-statistic:
  - $x-bar = 33.5\%$
  - $s = 14.39\% = \sqrt{\frac{1}{4-1} \times ((53 - 33.5)^2 + (28 - 33.5)^2 + (34 - 33.5)^2 + (19 - 33.5)^2)}$

<table>
<thead>
<tr>
<th>Observation</th>
<th>diff</th>
<th>diff^2</th>
</tr>
</thead>
<tbody>
<tr>
<td>53%</td>
<td>14.5</td>
<td>200.25</td>
</tr>
<tr>
<td>28%</td>
<td>-5.5</td>
<td>30.25</td>
</tr>
<tr>
<td>34%</td>
<td>0.5</td>
<td>0.25</td>
</tr>
<tr>
<td>19%</td>
<td>-14.5</td>
<td>210.25</td>
</tr>
<tr>
<td>mean = 33.5%</td>
<td></td>
<td>sum = 621</td>
</tr>
<tr>
<td></td>
<td></td>
<td>divide by 2 = 310.5</td>
</tr>
<tr>
<td></td>
<td></td>
<td>sq root = 14.39</td>
</tr>
</tbody>
</table>

$P(t < \frac{(33.5 - 30.0)}{14.39/\sqrt{4}} - t(2))$
A problem for you to work on:

- Our firm requires that projects return *more than* 23% on their investment annually to be considered for funding. You are interested in a program which is intended to improve employees practical problem solving skills and so their productivity. The consultant provides information on the last 16 programs which she has run which indicate that the programs have an average return on investment of 26% with sample standard deviation of 5%. Can you reject the hypothesis that the ROI for this training is 23% or less?
Hypothesis Testing with Small Samples ($n \leq 30$)

Solution:

$H_0: \mu \leq 23$

$H_1: \mu > 23$

$\frac{(26.0 - 23.0)}{2.33} > \frac{7.06}{2.33}$

Critical points with 15 degrees of freedom are

<table>
<thead>
<tr>
<th>10%</th>
<th>1.34</th>
</tr>
</thead>
<tbody>
<tr>
<td>5%</td>
<td>1.75</td>
</tr>
<tr>
<td>1%</td>
<td>2.60</td>
</tr>
</tbody>
</table>

Populations vs. Samples

- 1. We have pretended to this point that we do not know the population mean, but we know the population variance. Is this realistic?
- 2. Our most typical situation is that we have to calculate variance and standard deviation from the sample. Our formula for sample and standard deviation is on the following slide.
- 3. The difference between the population and sample calculations is that the sample calculations divide by (n-1) rather than n.
## Populations vs. Samples

Population Variance
\[ \sigma^2 = \frac{\sum_{i=1}^{N} (X_i - \mu)^2}{N} \]

Population Standard Deviation
\[ \sigma = \sqrt{\sigma^2} \]

Sample Variance
\[ s^2 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{(n - 1)} \]

Sample Std. Deviation
\[ s = \sqrt{s^2} \]

where
- \( N \) is the individuals in the population
- \( n \) is the observations in the sample
- \( s^2 \) is the symbol for sample variance
- \( s \) is the symbol for sample std dev

### Formulas for Populations and Samples

<table>
<thead>
<tr>
<th>Terms</th>
<th>Symbol</th>
<th>Formula</th>
</tr>
</thead>
<tbody>
<tr>
<td>Population</td>
<td>( \mu )</td>
<td>[ \mu = \frac{1}{N} \sum_{i=1}^{N} X_i ]</td>
</tr>
<tr>
<td>Sample</td>
<td>( \bar{x} )</td>
<td>[ \bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i ]</td>
</tr>
<tr>
<td>Mean</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Variance</td>
<td>( \sigma^2 )</td>
<td>[ \sigma^2 = \frac{1}{N} \sum_{i=1}^{N} (X_i - \mu)^2 ]</td>
</tr>
<tr>
<td></td>
<td>( s^2 )</td>
<td>[ s^2 = \frac{1}{(n - 1)} \sum_{i=1}^{n} (x_i - \bar{x})^2 ]</td>
</tr>
<tr>
<td>Standard Deviation</td>
<td>( \sigma )</td>
<td>[ \sigma = \sqrt{\sigma^2} ]</td>
</tr>
<tr>
<td></td>
<td>( s )</td>
<td>[ s = \sqrt{s^2} ]</td>
</tr>
</tbody>
</table>
Populations vs. Samples

Q: How does this alter our calculations of probabilities?
A: Very little. So long as sample size is greater than 30, our calculations are identical.

Note that the use of $\sigma$ or $s$ in the z transform (and its division by $\sqrt{n}$) is distinct from the calculation of $\sigma$ or $s$.
In general, we should use the t-table rather than the z-table when we have $s$ rather than $\sigma$, but once we are up over 30 observations, the differences are small and the z-table more precise.
Two-Tailed Tests

- What if we want to test a statement such as $\mu = k$ (the population mean is equal to some constant) rather than $\mu > k$ or $\mu < k$. Alternatively, how do we do two tailed tests?
  - 1. When do we use this? When we want to know if something has changed or is different, but we don’t know the direction of that difference. It addresses a threshold issue.
  - Example: Are protected classes treated differently than non-protected classes? This is the actual question asked by Title VII, not whether protected classes are treated worse.

Example: Over a long time, output per employee in a plant has been 30 units per day per employee with variance of 25 units. A number of changes have occurred recently:
- We implemented a system of statistical process control. This increases productivity by reducing the number of inspectors we need, but increases production workers job duties, slowing them slightly.
- We also implemented a JIT inventory and logistics system. This saves costs by reducing inventory and carrying costs, but has led to production slowdowns when parts were not available.
- New machinery we have installed over the last year should have increased the rate of production.
Two-Tailed Tests

- We suspect that production has been affected by these changes, but we don’t know how. We would like to know if there has been a change, we don’t know enough to know whether it has been an improvement or a decline in productivity.
  - $H_A$: productivity per employee has been affected and is no longer 30: $\mu \neq 30$
  - $H_0$: productivity per employee remains as it has been: $\mu = 30$
- We take a sample for 50 days and get a sample mean of 31.5. Is it reasonable to believe that we are still averaging 30 units of output per employee per day?

Two-Tailed Tests: A Graphical Approach

With variance of 25, the standard deviation of samples size 50 is 0.7. As sample size is greater than 30, we expect the sample means to be normally distributed.
Now, before “drawing our sample”, let’s consider a rule for rejecting the hypothesis that nothing has really changed. Let’s only reject our null of no change if the probability of observing a given sample is less than 5% (but it can up in the upper or lower tail).

The 5% critical point for a two tailed test is 1.96. With sample standard deviation of 0.7, our cutpoints for our 5% test are:

- $30 \pm 1.96 \times 0.7$
- $30 \pm 1.37$
- $28.6$ to $31.4$
We take a sample for 50 days and get a sample mean of 31.5. Is it reasonable to believe that we are still averaging 30 units of output per employee per day?
Two-Tailed Tests: A Graphical Approach

Result: Reject in a 10% and 5% test, but not in a 1% test.

• 10% critical point 1.645
• 5% critical point 1.96
• 1% critical point 2.58

\[ P(x = 315 \text{ if } \mu = 30) \]

\[ P(z \geq \frac{x - \mu}{\sigma / \sqrt{n}}) \]

\[ P(z \geq \frac{31.5 - 30}{\frac{25}{\sqrt{50}}} \]

\[ = P(z \geq \frac{1.5}{7.1}) \]

\[ = P(z \geq 2.12) \]

\[ = P(z \geq 2.12) + P(z \leq -2.12) \]

\[ 2 \times P(z \geq 2.12) = 2 \times 1.7 = 3.4\% \]

Two-Tailed Tests: Example

Example (to be reviewed outside of class):

- We are hearing stories that things are different in our Toledo office and decide to check in on this. We don’t know if things are good or bad in Toledo, just that everyone says they are different.

- As one metric of difference, we decide to calculate employee retention in the office. Employees typically remain with our company for 25 months. As we don’t know the direction of the Toledo effect, we need to use a two tailed test.
Two-Tailed Tests: Example

- Our null is that the Toledo office isn’t really any different, it only seems that way
  - \( H_0: \mu_{\text{Toldeo}} = 25 \)
- Our alternative is that Toledo is different.
  - \( H_A: \mu_{\text{Toldeo}} \neq 25 \)
- We use readily available data and find that, for the 26 employees in the Toledo office, mean time with our firm is 22 months and standard deviation for the sample is 8 months. Can we reject the null? Remember, as \( n = 26 \) we have to use a t-table.

\[
P(\bar{x} = 22 \mid \mu = 25) = \\
P(t(26 - 1) \geq |\frac{\bar{x} - \mu}{\sqrt{\frac{\sigma^2}{n}}}|) = \\
P(t(26 - 1) \geq |\frac{22 - 25}{\sqrt{8^2/26}}|) = \\
P(t(26 - 1) \geq |\frac{3}{1.87}|) = P(t(26 - 1) \geq 1.62)
\]

Critical cutoffs for a \( t(25 - 1) \) are
- 10% \( \pm 1.70 \)
- 5% \( \pm 2.06 \)
- 1% \( \pm 2.78 \)

reject at 10% but not at 5%
Example: A colleague of yours who works in another department sits down with you at lunch one day and says, “I notice that your shop has a hire to interview ratio of 6 to 1, mine has a ratio of 5 to 1. I guess that I really know what I am doing. What are you ordering for lunch?”

You go ahead and order, but you are reflecting on the fact that, given Detroit is only 50 miles away, this person’s life is worth no more than $200. However, before moving ahead on this thought, you decide to test whether your colleague is actually performing better than you, or whether this is better explained as a random occurrence (you plan is, if indeed the colleague is doing better, you will learn from her before she is terminated with extreme prejudice, otherwise, she can go tomorrow).

Checking your firms records, you find that:

1. You hired 10 employees in the last year, with an average of six candidates per hire. The standard deviation per hire was 2.
2. Your colleague also hired 10 employees last year, with an average of five candidates per hire and with a standard deviation of 1.8.
Tests for Differences in Means

Now set up the hypothesis:

1. The null is that there is no difference between your performance and that of your colleague.
2. The alternative is that your colleague did indeed outperform you, had a lower ratio of interviews to new hires.

Tests for Differences in Means

Formalizing this: Call your group “mine” and her group “hers”

- \( H_A \): Your colleague has a lower population ratio of interviews to hires:
  - \( \mu_{\text{hers}} < \mu_{\text{mine}} \)
- \( H_O \): There is no difference in the ratio of interviews to hires in the population of their department and my department
  - \( \mu_{\text{hers}} \geq \mu_{\text{mine}} \)

Note that our alternative and null might be written as:

- \( H_A \): \( \mu_{\text{hers}} - \mu_{\text{mine}} < 0 \)
- \( H_O \): \( \mu_{\text{hers}} - \mu_{\text{mine}} \geq 0 \)
Tests for Differences in Means

Now set up the probability problem.

Our typical problem has a single mean and looks like:

\[
p\left( \frac{X - \mu}{\sigma / \sqrt{n}} \leq \text{some numbers} \right) = ?
\]

But now we have multiple population and sample means, so it now looks like:

\[
p\left( \frac{(\bar{X}_{\text{hers}} - \bar{X}_{\text{mine}}) - (\mu_1 - \mu_2)}{\sigma_{\text{hers-mine}}} \leq \frac{(5 - 6) - 0}{??} \right)
\]

Tests for Differences in Means

- So our problem is calculating the standard deviation for the difference “hers-mine”
- If two samples, A and B, are independent, then the standard deviation of the difference in their means is:

\[
\sigma_{A-B} = \sqrt{\frac{\sigma_A^2}{n_A} + \frac{\sigma_B^2}{n_B}}
\]
Tests for Differences in Means

The standard deviation in our problem:

\[
\sigma_A = 2 \\
\sigma_A = 1.8 \\
n_A = n_B = 10 \\
\sqrt{\frac{2^2}{10} + \frac{1.8^2}{10}} \\
\sqrt{\frac{4}{10} + \frac{3.24}{10}} \\
= 0.85088
\]

Tests for Differences in Means

\[
p\left(\bar{x}_{hers} - \bar{x}_{min} < 5 - 6\right)
\]

\[
p\left(\frac{\bar{x}_{hers} - \bar{x}_{min}}{\sigma_{hers-min}} < \frac{5 - 6}{0.85088}\right)
\]

\[
p(t(?)) < -1.175
\]
Tests for Differences in Means

Now we have one last problem, determining our degrees of freedom. The formula for degrees of freedom for a difference in means is:

- Degrees of Freedom = n_{group1} + n_{group2} - 2

So, in this case, our degrees of freedom are 10 + 10 - 2 = 18 and the one sided t-statistic for 18 degrees of freedom is -1.33. As -1.175 is greater than -1.33, we cannot reject the null hypotheses in a 10% test. There is nothing to learn from your colleagues except the cost of rudeness.

Tests for Differences in Means

The key to doing tests for differences in means:

1. Setting up the hypothesis in terms of two population means
2. Translating this into a statement about sample means (setting up the probability)
3. Calculating the joint standard deviation.
Tests for Differences in Means

Example: The United States Bureau of Labor Statistics recently announced that the unemployment rate declined from 5.5% to 5.4%. There were 50,000 households in each sample and the standard deviation of each sample was .161276. Test to determine whether it's reasonable to believe that the unemployment rate has dropped.

First, our alternative is that the unemployment rate this month is lower than the unemployment rate last month:

\[ H_0: \mu_{\text{current}} - \mu_{\text{last}} \leq 0 \]
\[ H_a: \mu_{\text{current}} - \mu_{\text{last}} > 0 \]

So our probability problem is:

\[ P( \bar{x}_{\text{current}} - \bar{x}_{\text{last}} < 0.054 - 0.055 \mid \mu_{\text{current}} - \mu_{\text{last}} > 0) \]
Tests for Differences in Means

And our problem looks like:

\[ p(\bar{x}_{August} - \bar{x}_{July} < 0.054 - 0.055) \]

\[ p(z < \frac{0.054 - 0.055}{\sigma_{A-J}}) \]

\[ \sigma_{A-J} = \sqrt{\frac{.161276^2}{50000} + \frac{.161276^2}{50000}} \]

\[ \sigma_{A-J} = .00102 \]

\[ p(z < \frac{.054-.055}{.00102}) = p(z < -1.000) \]

Note that in this case, we have 50,000 + 50,000 - 2 for our degrees of freedom, so we use a z table rather than the t-table. As the critical value for a 10% one tailed test is 1.28, we cannot reject the null of no change in the population.