Continuing Statistical Inference III

Comparing Means, Paired Differences

and

Performing Tests in JMP

Recap: Statistical Inference

We've been working with two tools using statistical inference so far, both of which involve a single unknown mean:

Tool 1: Confidence Intervals

Identifying a range of values likely to contain a true mean by creating a confidence interval around a sample mean.

and

Tool 2: Hypothesis Tests

We use a probability connecting the true mean (μ) and the sample mean ($\bar{X}$) to determine if an observed value of the sample mean ($\bar{x}$) is likely to have occured for a specific assumed value of the true mean.

Tool 1: Confidence Intervals

Step 1

We find a probability distribution which connects the true mean (μ) and the sample mean ($\bar{X}$)

- If n is large and we know the true variance is σ^2 then the central limit theorem will get us

$$Z = \frac{\bar{X} - \mu}{\sqrt{\sigma^2/n}} \sim N(0, 1)$$

- If n is large and we do not now the true variance, but can calculate the sample variance using

$$S^2 = \sum_{i=1}^{n} (X_i - \bar{X})^2$$

then we can connect $\bar{X}$ and μ using

$$Z = \frac{\bar{X} - \mu}{\sqrt{S^2/n}} \sim N(0, 1)$$

Tool (1): Confidence Intervals

Step 1

- If n is large and we do not now the true variance, but can calculate the sample variance using

$$S^2 = \sum_{i=1}^{n} (X_i - \bar{X})^2$$

then we can connect $\bar{X}$ and μ using

$$T = \frac{\bar{X} - \mu}{\sqrt{S^2/n}} \sim t_{n-1}$$

where t_{n-1} is a t-distribution with $n - 1$ degrees of freedom.

Recap

Inference

Conf. Ints

Tool 1: Confidence Intervals

Step 2

With the distribution in hand we find determine the value from the distribution which describes our desired confidence interval:

- Large n, known σ^2: Find the value $z_{0.975}$ so that

$$P\left(\frac{\bar{X}-\mu}{\sqrt{\sigma^2/n}} \leq z_{0.975}\right) = P(Z \leq z_{0.975}) = .975$$

  - Exclude extremes: 95% conf that μ is between $\bar{x} \pm z_{0.975}\sqrt{\frac{\sigma^2}{n}}$
  - Exclude large: 97.5% conf that $\mu \leq \bar{x} + z_{0.975}\sqrt{\frac{\sigma^2}{n}}$
  - Exclude small: 97.5% conf that $\mu \geq \bar{x} - z_{0.975}\sqrt{\frac{\sigma^2}{n}}$

Recap

Inference

Conf. Ints

Tool 1: Confidence Intervals

Step 2 (continued)

- Large n, unknown σ^2: Find the value $z_{0.975}$ so that

$$P\left(\frac{\bar{X}-\mu}{\sqrt{S^2/n}} \leq z_{0.975}\right) = P(Z \leq z_{0.975}) = .975$$

  - Exclude extremes: 95% conf that μ is between $\bar{x} \pm z_{0.975}\sqrt{\frac{s^2}{n}}$
  - Exclude large: 97.5% conf that $\mu \leq \bar{x} + z_{0.975}\sqrt{\frac{s^2}{n}}$
  - Exclude small: 97.5% conf that $\mu \geq \bar{x} - z_{0.975}\sqrt{\frac{s^2}{n}}$

Tool 1: Confidence Intervals

Step 2 (continued)

- Small n, unknown σ^2: Find the value $t_{n-1,0.975}$ so that

$$P\left(\frac{\bar{X}-\mu}{\sqrt{S^2/n}} \leq t_{n-1,0.975}\right) = P(T \leq t_{n-1,0.975}) = .975$$

  - Exclude extremes: 95% conf that μ is between $\bar{x} \pm t_{n-1,0.975}\sqrt{\frac{s^2}{n}}$
  - Exclude large: 97.5% conf that $\mu \leq \bar{x} + t_{n-1,0.975}\sqrt{\frac{s^2}{n}}$
  - Exclude small: 97.5% conf that $\mu \geq \bar{x} - t_{n-1,0.975}\sqrt{\frac{s^2}{n}}$

Tool 2: Hypothesis Tests

Step 1

Determine the assumption to be tested and write your hypothesis statement.

Also, determine a level of significance for your test. This is the probability that you make a Type I error (rejecting a true Null Hypothesis).

Step 2

Determine a test statistic that connects the data you can observe and the values set in the null hypothesis.

Determine the value of the test statistic in this specific case using the assumptions of the null hypothesis and the values calculated from your data.

Tool 2: Hypothesis Tests

Step 3

Find the p-value: the probability that you would observe a value *as or more extreme* than your test statistic under the null hypothesis (i.e., using the null hypothesis values).

Step 4

State your conclusion: if the p-value is below the level of significance, we reject the null hypothesis. Otherwise, we fail to reject the null hypothesis.

Notice that both the confidence interval and the hypothesis test involve finding a test statistic. It's just a matter of what other pieces around the probability equation we do or don't know.

Recap

Inference

Conf. Ints

Hypothesis Tests

Tool 2: Hypothesis Tests

Suppose that we are interested in a determining if a machine is still producing gears with an average center bore less than of $0.23\mu m$. We collect 9 such gears and measure the bores:

```
0.24 0.21 0.31 0.18 0.29 0.20 0.32 0.23 0.10
```

The mean is $\bar{x} = 0.2311\mu m$ the sample variance is $s^2 = 0.03889$.

Tool 2: Hypothesis Tests

Performing a test at the 0.05 significance level:

Hypothesis Statement
$H_0 \; : \mu \leq 0.23$
$H_1 \; : \mu > 0.23$

Test Statistic

$$T = \frac{\bar{X} - \mu}{\sqrt{\frac{S^2}{n}}} \sim t_{n-1} \rightarrow T^* = \frac{0.2311 - 0.23}{\sqrt{\frac{0.03889}{9}}} = 0.0167$$

p-value
The probability that an observation from a t_8 would be more extreme than 0.0167 is > 0.10 (Table B.4).

Conclusion
Since the p-value is greater than 0.05 we fail to reject the null hypothesis.
There is not enough statistical evidence to suggest that the mean bore width is not less than 0.23.

Tool 2: Hypothesis Tests

The p-value came from Table B.4 in the text book.

- Looking at the column, we see the degrees of freedom on the left most column.
- The first entry for 8 degrees of freedom is 1.397, which is in the $Q(.9)$ column.
- This means that the probability of a random variable being larger than 1.397 is 0.10 (90% below, 10% above).
- Since there is an area of 0.10 under the curve between 1.397 and ∞, we know there will be *more* than 0.10 under the curve between 0.167 and ∞.
- So even though we can't get the p-value exactly, we can use the table to say that we know it's at least 0.10.

Comparing Means

Comparing Means

Suppose we have two true means, μ_1 and μ_2 that we would like to compare. For instance, X_1 has mean μ_1 and variance σ_1^2 and X_2 has mean μ_2 and variance σ_2^2

- We can gather a sample of size n_1 from Population 1 to get a mean $\bar{x}_1$. At this point we can do hypothesis tests and confidence intervals for μ_1 .
- We can gather a sample of size n_2 from Population 1 to get a mean $\bar{x}_2$. At this point we can do hypothesis tests and confidence intervals for μ_2 .

But how do we make inferences on both at the same time?

Comparing Means

These two facts are very useful:

> **FACT I**
> The sum or difference of two normal random variables also follows a normal distribution.
>
> **FACT II**
> The sum or difference of two t random variables is NOT a t distribution.

And take the example of the difference between to sample means: $\bar{D} = \bar{X}_1 - \bar{X}_2$

$$E(\bar{D}) = E(\bar{X}_1) - E(\bar{X}_2) = \mu_1 - \mu_2$$

and

$$Var(\bar{D}) = Var(\bar{X}_1) + Var(\bar{X}_2) = \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}$$

Comparing Means

<u>Case I: Both large sample sizes, both variances known</u>

Since $\bar{X}_1 \sim N(\mu_1, \frac{\sigma_1^2}{n_1})$ and $\bar{X}_2 \sim N(\mu_2, \frac{\sigma_2^2}{n_2})$ then

$$\bar{D} \sim N\left(\mu_1 - \mu_2, \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}\right) \rightarrow \frac{\bar{D} - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \sim N(0, 1)$$

and since we can collect data and get the value $\bar{d} = \bar{x}_1 - \bar{x}_2$, then we can use

$$Z = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \sim N(0, 1)$$

to make statements about confidence intervals or hypothesis tests.

Comparing Means

Comparing Means

Case II: Both large sample sizes, variances unknown

This is the same as result as Case I, only we replace the variances with the sample variances, S^2.

Comparing Means

Example

Suppose that a researcher wishes to compare the average lengths of two populations of squirrels. A sample of 50 squirrels were taken from each population. For the first population, the mean length was 17.9 cm and the sample variances of the lenghts was 2.8 cm^2. In the second population, the mean length in the sample was 19.6 cm and the sample variance was 1.2 cm.

Construct a 95% confidence interval for the true difference in the populations means.

Comparing Means

Example (continued)

This works like any other Z statistic:

$$(\bar{x}_1 - \bar{x}_2) \pm z_{0.975}\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

which gives

$$(17.9 - 19.6) \pm 1.96\sqrt{\frac{2.8}{50} + \frac{1.2}{50}} = -1.7 \pm 0.554$$

In other words, we are 95% confident that the mean length of the first population is somewhere between 2.254 and 1.146 cm less than the second population.