MA 2B RECITATION 02/23/12

1. t-Tests

We're still looking at different ways to estimate some value by sampling. Suppose we want to estimate the mean. If we know the standard deviation, we could use the test statistic

$$\frac{\overline{X}_i - \mu_0}{\sigma/\sqrt{n}}$$

which follows a normal distribution.

However, if we do not know the standard deviation, we want to use the obvious replacement: the sample standard deviation

$$s = \hat{\sigma} = \sqrt{\frac{\sum_i (X_i - \overline{X})^2}{n-1}}$$

However, because of the variability of the sample standard deviation, our statistic is not normal. It has what we call a **t-distribution with $v = n - 1$ degrees of freedom**. It resembles the normal distribution, but depends on n.

There are two situations where you want to use a t-test:

(a) When the population variance is unknown.
(b) When the sample size is small.

Then we can use the t-test for a hypothesis test or to find confidence intervals.

We will talk about three different versions of t-test, each with their own application. In particular, the assumptions involved in each t-Test are important, so make sure to pay attention to them.

Remark 1.1. The t-test was originated by W.S. Gosset, a statistician working for the Guinness brewing company (the same one that makes the Book of World Records). Since Guinness employees were not allowed to publish under their own names, Gosset published under the pseudonym "Student." This is why the terms here are sometimes referred to as "Student's t-test" or "Student's t-distribution."

1.1. Basic t-Test.

We want to use the statistic

$$T = \frac{\overline{X} - \mu_0}{S/\sqrt{n}}.$$

Under the null hypothesis that $\mu = \mu_0$, T will have a t distribution with $n - 1$ degrees of freedom.

Date: February 23, 2011.

1.2. **Two Population t-Tests.** Here, we have two populations that are *normal* with the *same* standard deviation and *different* means. To test hypotheses involving the *difference* of these means, we can use the statistic

$$T = \frac{\overline{X}_i - \overline{Y}_i - (\mu_X - \mu_Y)}{s\sqrt{\frac{1}{n} + \frac{1}{m}}}$$

where

$$s = \sqrt{\frac{\sum_i (X_i - \overline{X}_i)^2 + \sum_i (\overline{Y}_i - Y_i)^2}{m + n - 2}}$$

where there are n observed X_i's and m observed Y_i's. The statistic T has a t-distribution with $n + m - 2$ degrees of freedom, so we can use essentially the same method as above.

Example 1. The owner of a bakery believes that the smell of fresh baking will encourage customers to purchase goods from his bakery. To investigate this belief, he records the daily sales for 10 days when all the bakery's windows are open, and the daily sales for 10 days when all the windows are closed. The following sales are recorded:

$$\text{open} : 202.0, 204.5, 207.0, 215.5, 190.8, 215.6, 208.8, 187.8, 204.1, 185.7$$
$$\text{closed} : 193.5, 192.2, 199.4, 177.6, 205.4, 200.6, 181.8, 169.2, 172.2, 192.8$$

Assuming that these data are random samples from normal populations with the same variance, is this belief valid?

Solution. $H_0 : \mu_1 = \mu_2$, where 1 corresponds to open windows and 2 to closed windows.

$H_1 : \mu_1 > \mu_2$ (one-tailed)

Say significance level is $\alpha = 0.05$, degrees of freedom $v = 10 + 10 - 2 = 18$. Critical region is $t > 1.734$.

Under H_0, we calculate

$$\overline{X}_i = 202.18, \hat{\sigma}_X^2 = 115.7284; \quad \overline{Y}_i = 188.47, \hat{\sigma}_Y = 156.6534.$$

Therefore,

$$\begin{aligned} s^2 &= \frac{9 \cdot 115.7284 + 9 \cdot 156.6534}{10 + 10 - 2} \\ &= \frac{115.7284 + 156.6534}{2} \quad \text{mean when } n = m \end{aligned}$$

Therefore, $s = 11.67$. Hence,

$$t = \frac{202.18 - 188.47}{11.67\sqrt{\frac{1}{10} + \frac{1}{10}}} = 2.63.$$

This value does lie in the critical region so H_0 is rejected. Thus, there is evidence, at the 5% level of significance, that suggests that fresh baking will encourage customers to purchase goods from the bakery. □

1.3. **Paired t-Tests.** We again have a set of data pairs, and we want to evaluate the *average differences*. We don't need to assume that the values in the pairs are normal, but only that the *differences* are normal. Here, we use the statistic

$$T = \frac{(\overline{X}_i - \overline{Y}_i)}{\frac{1}{\sqrt{n}}\sqrt{\frac{\sum_i((X_i-Y_i)-(\overline{X}_i-\overline{Y}_i))^2}{n-1}}},$$

which has a t distribution with $n-1$ degrees of freedom. This is just a one-population t test applied to the differences.

Example 2. Suppose that a professor wants to see whether some new computer software helps in teaching statistics. Students are randomly assigned into two groups, a control group only receiving standard lessons, and an experimental group that also uses the software. They are selected in pairs of roughly equal mathematical ability, with one from each pair assigned randomly to the control group and the other to the experimental group. At the end, they are given a test to measure their understanding, their results are shown as follows, with the first number in the pair being the control, and the second number being the experimental:

$$(72,75),(82,79),(93,84),(65,71),(76,82),(89,91),(81,85),(58,68),(95,90),(91,92).$$

Assuming percentage marks are normally distributed, does the computer package help students on learning the statistics for the exam?

Proof. $H_0 : \mu_d = 0$, where d represents the "Difference = Experimental - Control."
$H_1 : \mu_d > 0$ (one-tailed)
Significance, say, $\alpha = 0.05$.
Degrees of freedom $v = 10 - 1 = 9$.
Critical region is $t > 1.833$. Under H_0, the test statistic is

$$T = \frac{\overline{d}}{\hat{\sigma}_d/\sqrt{n}}$$

The 10 differences are

$$3, -3, -9, 6, 6, 2, 4, 10, -5, 1.$$

Hence, $\sum d = 15$ and $\sum d^2 = 317$. Therefore,

$$\overline{d} = 1.5, \hat{\sigma_d} = 5.72.$$

Therefore,

$$T = \frac{1.5}{5.72/\sqrt{10}} = 0.83$$

The value does not lie in the critical region so H_0 is not rejected. Thus, there is no evidence, at 5% significance, that the computer package helps the students on learning statistics for the exam. □

2. p-Values

Suppose that we're not interested accepting or rejecting a hypothesis for a certain significance level, but whether we'd reject for *any* given significance level. In other words,

For instance, if you The way we deal with this in statistics is by using the p-value.

The **p-value** of a statistic is the probability of observing something at least as extreme under your null hypothesis. It's easiest to see this in an example.

The main point is that we reject the null for any significance level α such that $\alpha \geq p$.

Example 3. Suppose that a test has been given to all high schoolers in the United States. The mean score is 70 with standard deviation 10. The testmakers suspect that students from California have a higher mean score on the test than elsewhere, because the mean score from a random sample of 64 Californians is 73. Is this strong evidence that the overall mean for Californian students is higher?

Solution. The null hypothesis H_0 says that there is no difference between the mean score for Californians and the mean for the entire population, so $\mu = 70$. The alternative hypothesis claims that the mean for Californians is higher than the population mean, so $\mu > 70$. This is a one-sided test.

Since we know the standard deviation, we should use

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

as our statistic. We have a sample mean of 73 and population standard deviation as 10, so

$$z = \frac{73 - 70}{10/\sqrt{64}} = \frac{3}{1.25} = 2.4.$$

Since this is a one-sided test, the p-value is equal to the probability of observing a value greater than 2.4 in the standard normal distribution, for

$$P(Z > 2.4) = 1 - P(Z \leq 2.4) = 1 - 0.9918 = 0.0082.$$

The p-value is less than 0.01, indicating that it is highly unlikely that these results would be observed under the null hypothesis. The school board can thus confidently reject H_0 given this result, although they can't conclude any additional information about the mean of the distribution. □

3. Confidence Intervals

Given some statistic, a confidence interval is an interval chosen so that the true value of a parameter lies within the interval with a given probability.

Much like finding critical values, we start by specifying the desired level of precision and plugging in what you know, then tweak the bounds until you get your desired result.

Example 4. The adjusted point totals of 8 statistics majors taking Ma 2b at the midterm were taken:

$$181, 178, 186, 184, 189, 176, 180, 182.$$

Find a 95% confidence interval for the mean point total of the remaining statistics majors int he class.

Solution. Since the sample is small and we do not know the variance, we should use the t-test. We first find the mean and standard deviation of the sample:

$$\overline{x}_i = 182, s_{n-1} = 4.24, n = 8.$$

Find the t-value from the table for $v = 7$ and $p = 0.975$, we obtain

$$t_{7,0.975} = 2.365.$$

By putting our data into the formula, we obtain

$$\overline{x}_i \pm t \cdot \frac{s_{n-1}}{\sqrt{n}} = 182 \pm 2.365 \cdot \frac{4.24}{\sqrt{8}} = 182 \pm 3.55.$$

Therefore, the confidence interval is from 178.45 to 185.55. $\square$