Our third test covers the material from 3/30 to 5/2 on the syllabus—significance tests, correlation and regression coefficients. The *comprehensive* final exam will cover the material on tests 1-3 **and** the new material covered in the final lectures.

A basic overview of the material

The major idea of this material is a test of hypotheses or significance tests. You should consult sections 10-2 and 3 in the book to make sure you are familiar with the vocabulary terms. Additionally, we defined the correlation of two data sets.

The basic idea is that we want to understand something about an **unknown** population parameter. For us, this is either μ, the population mean or $\mu_1 - \mu_2$, the difference of population means. We use the sample mean, $\bar{X}$ or the difference of sample means, $\bar{X}_1 - \bar{X}_2$ to make the inference.

How this is done involves a test of two competing hypotheses: the null hypotheses, H_0 and the alternative hypotheses, H_A. To help decide which hypothesis is preferable a **test statistic** is used whose distribution under the null hypothesis is known.

An observed value of the test statistic can be used two ways to see if the results are statistically significant. The p-value can be computed, as is done on the computer. This is compared to the significance level, α. Alternatively, in class we used the significance level α to find the *critical values* which then give the *acceptance* and *rejection regions*.

Based on our analysis we can then decide to accept or reject the null hypothesis. There are again two ways of phrasing this:

- If the observed value is in acceptance region we *accept*. This is equivalent to the p-value being larger than α or the differences *not* being *statistically significant*.
- If the observed value is in the rejection region we *reject* the null hypothesis. This is the same as the p-value being smaller than α. We can say the differences *are statistically significant*.

The basic skills involve these steps:

1. Specifying the two hypotheses
2. Selecting a test statistic with known sampling distribution
3. Computing the critical values or p-values
4. Making a comparison of the observed value of the test statistic (from a sample).

We learned a few different tests:

Test for mean when variance is known or n is large This test has hypotheses

$$H_0 : \mu = a, \quad H_A : \mu < a \text{ or } \mu \neq a \text{ or } \mu > a$$

where a is some number. The alternative is one of 3 possibilities: less, two-sided or greater.

For these assumptions the test statistic

$$Z = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}} \text{ (when } \sigma \text{ is known) or } \frac{\bar{x} - \mu}{s/\sqrt{n}}$$

has a standard normal distribution. (We still Table B, the t-distribution table.)

Test for mean when n is small This test has the same hypotheses as the previous, only when n is small the distribution of the test statistic is different. We use

$$T = \frac{\bar{x} - \mu}{s/\sqrt{n}}$$

which will have the t-distribution with $n - 1$ degrees of freedom under the null hypothesis.

Two sample test of difference of means When testing the difference of means the hypotheses are

$$H_0 : \mu_1 = \mu_2, \quad H_A : \mu_1 <, \neq, > \mu_2$$

(one of the three signs). The test statistic used varies depending on the assumptions:

The two sample sizes are large We can use

$$T = \frac{\bar{x} - \bar{y}}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}},$$

which will have a normal distribution. (That is, use table B with ∞ for the degrees of freedom.)

The samples are small, and we assume the population variances are equal

$$T = \frac{\bar{x} - \bar{y}}{s_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}},$$

where s_p is the pooled standard deviation and is defined by

$$s_p^2 = \frac{(n_1 - 1)s_1^2 + (n_2 2 - 1)s_2^2}{n_1 + n_2 - 2}.$$

Under H_0 and an assumption that the unknown variances are equal, the distribution of T is the t-distribution with $n_1 + n_2 - 2$ degrees of freedom.

The correlation was defined as a way to measure how strong a linear relationship is when two variables are in a linear relationship. The correlation is defined by

$$r = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum(x_i - \bar{x})^2 \sum(y_i - \bar{y})^2}}.$$

Some basic facts: $-1 \leq r \leq 1$; when r^2 is close to 1, the data is close to a straight line, when r^2 is close to 0 the data is scattered; r is negative if the linear relationship runs down hill, positive if it runs uphill.
The regression line, $\hat{y} = b_0 + b_1 x$, may be thought of as a summary of a linear relationship between two variables. We learned how to compute the coefficients using formulas similar to the correlation coefficient. Some key facts:

- $b_1 > 0$ when $r > 0$, and vice versa.
- The regression line goes through the point $(\bar{x}, \bar{y})$
- The predicted y value for a given x value is given by the corresponding $\hat{y}$.
- The formulas are:

$$b_1 = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sum(x_i - \bar{x})^2}, \quad b_0 = \bar{y} - b_1\bar{x}.$$

Some sample problems
The problems should seem similar after awhile as the procedure is the same: formulate the hypotheses, find the observed value, compute the p-value.
For each of these questions, write down your two hypotheses, your test statistic and how you found the p-value. The first two are done for you,

1. A test is done to see which textbook best helps student learning. Two sections of MTH 123321 were taught with different textbooks. A common final was given, and the scores are presented here:

```
               n   xbar    s
------------------------------
text book A:   18  72     12
text book B:   25  76     10
```

Is there a statistically significant difference at the $\alpha = 0.10$ level?
Answer:

The two populations are presumably the test scores of the hypothetical population of students taking MTH 123321 were they to use one of the two text books. We assume these populations are normally distributed with means μ_1 and μ_2 with unknown variances, that are not assumed to be equal. Our hypotheses are

$$H_0 : \mu_1 = \mu_2, \qquad H_A : \mu_1 \neq \mu_2$$

We use a two-sided alternative here as we are interested in any differences.

The test statistic would be the T without the pooled standard deviation. Its distribution would be the t distribution with 17 degrees of freedom. (We use the smaller of $18-1$ and $25-1$ here.) The critical values correspond to 17 degrees of freedom and 0.05 area in the right tail. That is -1.740 and 1.740.

Our observed value is

```
> (72 - 76)/sqrt(12^2/18 + 10^2/25)

[1] -1.154701
```

This is in the acceptance region, so we accept the null hypothesis of equal means. That is, there is no indication the two text books make a difference.

2. Last month the average time to park on campus was 8 minutes. This month (as more students have dropped out) it seems to take less time. Suppose, the sample average for 10 trips is 7 minutes with a sample standard deviation of 2 minutes. Does this indicate that the average time is less or is the difference explainable by sampling variation? Use $\alpha = 0.05$ for a significance level.

3. A test to determine if echinacea is beneficial in treating the common cold was setup as follows. If a child reported cold symptoms then they were randomly assigned to be given a treatment of echinacea or a placebo treatment. The time to recover was measured and is summarized in the table below

```
group       n   sample mean  sample sd
-------------------------------------
echinacea   200     5.3        2.5
placebo     207     5.4        2.5
```

Is this evidence that the echinacea group had a quicker recovery? Use $\alpha = 0.05$ for a significance level.

4. Is the line at the DMV improved? Historically, it took 65 minutes to do something at the DMV. After changes were made a sample of 21 people found their time was only 59 minutes. Is this evidence that the mean time has decreased? Use $\alpha = 0.05$ for a significance level.

5. A machine sometimes needs recalibration. It should have a mean of 16 and variance of 4, but sometimes the mean gets out of calibration. In a sample of 10 runs the sample mean was 15.8. At a $\alpha = 0.05$ significance level, is this evidence that the population mean is different from 16?

6. Which of these tests do we not know how to handle:

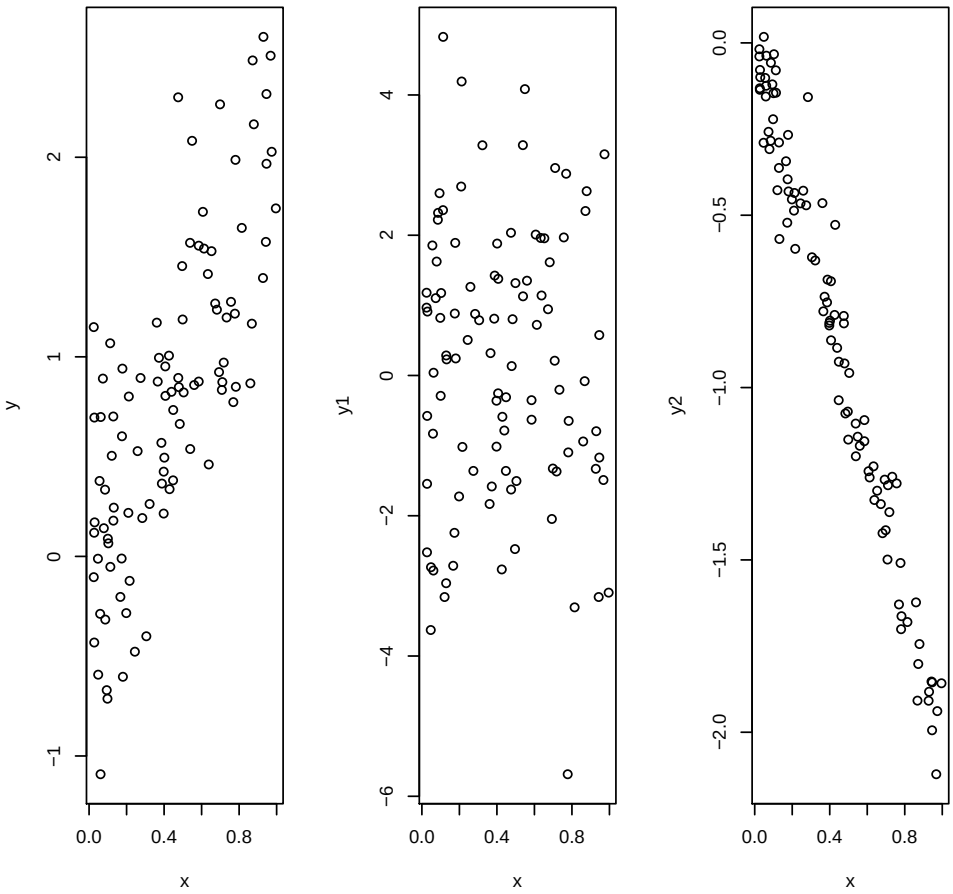
- We want to test
$$H_0 : \mu = 5, \quad H_A : \mu \neq 5$$
We assume the population is not normal, and n, our sample size, is large.
- We want to test
$$H_0 : \mu = 5, \quad H_A : \mu \neq 5$$
We assume the population is normal, but n, our sample size, is small
- We want to test
$$H_0 : \mu = 5, \quad H_A : \mu \neq 5$$
We assume the population is not normal, and n, our sample size, is small.

7. For this data set:

```
x | 1 1 2 2 3 3
y | 1 2 2 3 3 1
```

   (a) Plot the data first, and guess the correlation coefficient r
   (b) Compare with the actual value you find by doing the math
   (c) Sketch a least squares regression line, guess the slope.
   (d) Now find the least squares regression line and compare to your guess.

8. For each of these scatterplot, estimate the value of r, sketch the least-squares regression line.


y
x
y1
x
y2
x