Chapter 7: Inference for Distributions

Read chapter 7, then read the notes and try the WEBCT assignment questions. If you need more practice, try the practice questions with answers available on the web.

Chapter 7 Introduction

In the previous chapter, we introduced tests of significance and confidence intervals for the mean when the population variance is known. You should have a good sense of what these techniques do and an understanding of how they operate. In practice, we usually do not know the population variance. In this chapter, the procedures of inference are developed for just such a situation. Researchers are also interested in comparing variances of populations. The F test is introduced as a way of comparing spread or variance of two populations.

General Background

Suppose X_1, X_2, ..., X_n is a sample from a distribution with mean μ and variance σ^2. We want to make inferences about μ and σ.

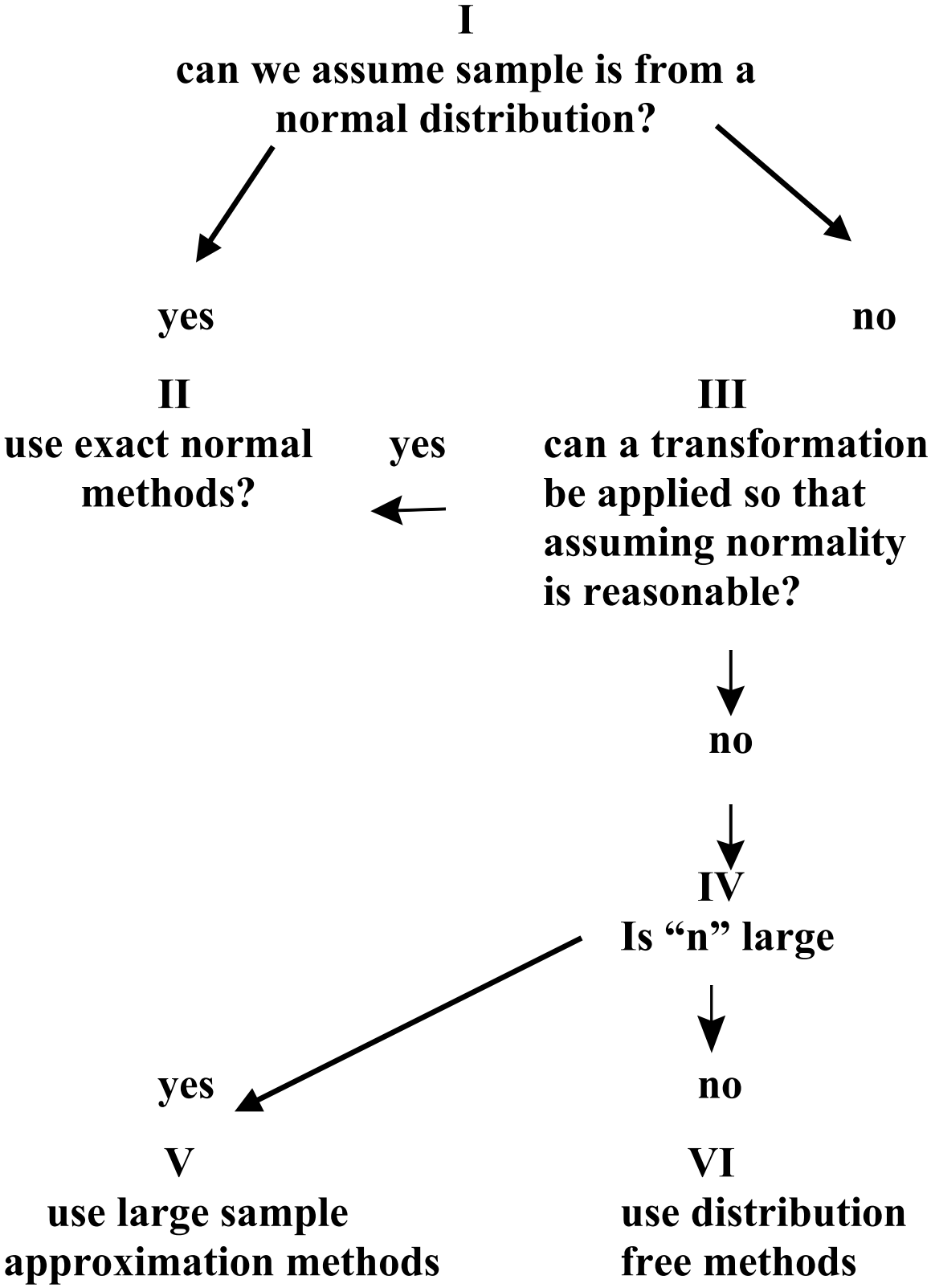
There are many different inference methods. Which do we use?

There is a general approach to inference which we describe here for inferences about μ. The flow chart on the next page helps us as researchers decide on the best technique for the problem at hand. We then briefly discuss each method.

I. Can we assume X_1, X_2,..., X_n is a sample from a N(μ,σ) distribution for some μ, σ?

First do a probability plot of X_1, X_2,...,X_n as first discussed in Chapter 1.

If the plot looks reasonable for a sample of size n from the normal then proceed to II. How do we decide if the plot is reasonable?


I
can we assume sample is from a normal distribution?
yes
no
II
use exact normal methods?
yes
III
can a transformation be applied so that assuming normality is reasonable?
no
IV
Is "n" large
yes
no
V
use large sample approximation methods
VI
use distribution free methods

Simulate a few samples of size n from N (0, 1) and do the normal probability plots. Then compare deviations from a straight line for simulated data with actual data.

View the Video – Inference for One Mean

II Exact methods for the normal distribution

Recall what the Central Limit Theorem (CLT) says when "n" is large. The sampling distribution of the mean $\bar{x}$ normal with the same mean as the population and reduced variance. In symbols we have:

$$\bar{x} \sim N(\mu, \sigma/\sqrt{n})$$

If we standardize we obtain

$$\frac{\bar{x} - \mu}{S_x / \sqrt{n}} \sim N(1, 0)$$

FACT: When $X_1, X_2,...,X_n$ is a sample from a $N(\mu,\sigma)$ distribution we have:

$$\bar{x} \sim N(\mu, \sigma/\sqrt{n})$$

If the variance is unknown, we substitute the sample standard deviation "s" and obtain a t-statistic, which is distributed as a "t" or student distribution.

$$t = \frac{\bar{x} - \mu}{S_x / \sqrt{n}} \sim \text{Student } (n-1) \text{ degrees of freedom}$$

The term "degrees of freedom" (df) is used to adjust for the sample size n. The t distribution changes are based on the sample size -1 or degrees of freedom. Smaller samples have more probability in the tails.

The graph below is of the student (m) or t(m) distribution. Note the greater probability in the tails.

student (m) distribution

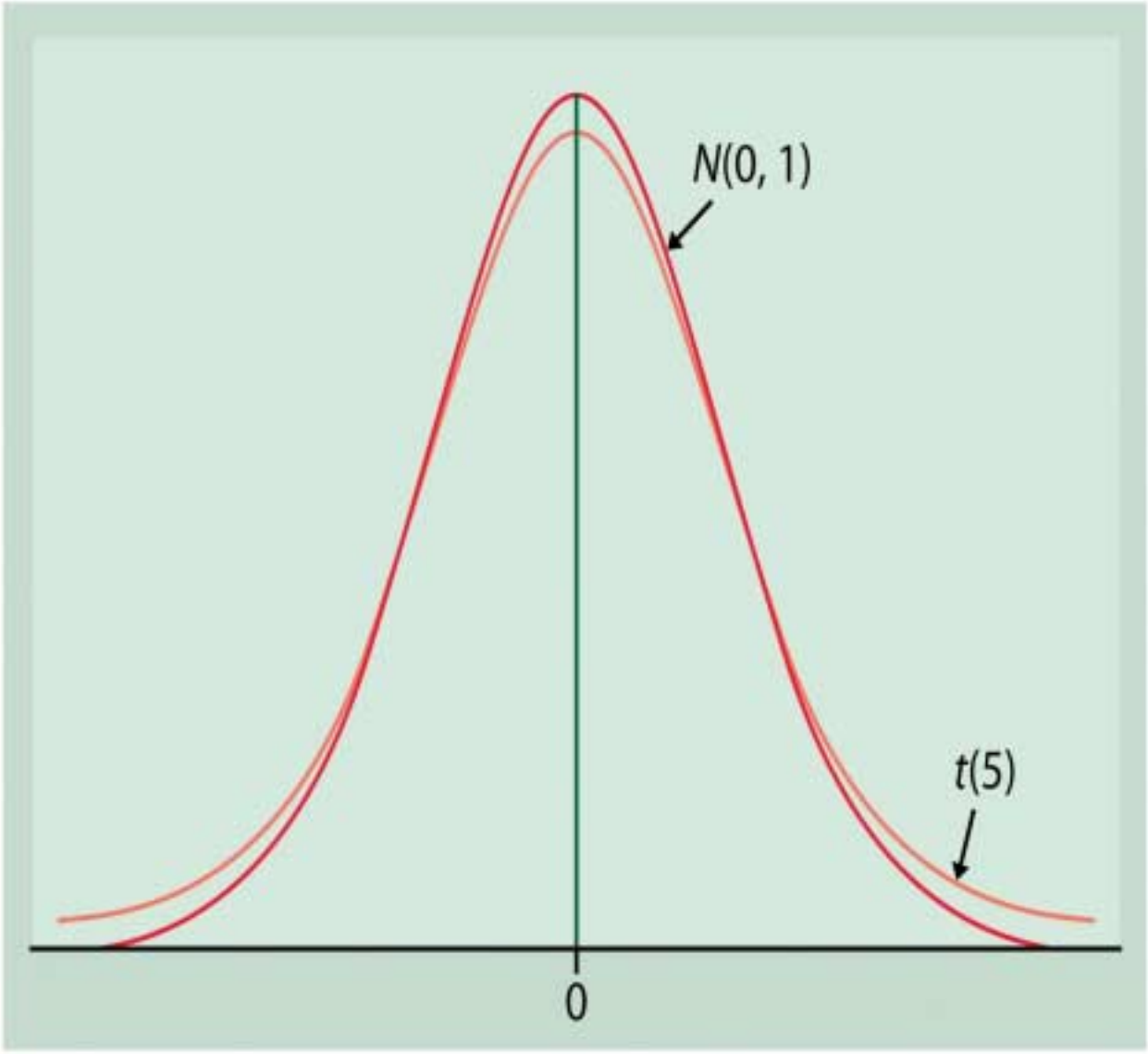


Figure 7.1 Density curves for the standard normal and t(5) distributions. Both are symmetric with centre 0. The t distributions have more probability in the tails than does the standard normal distributions.

Properties of Student (m) or t(m) Distribution

1) The distribution is symmetric about 0. For example, if x ~ Student (m) then

$$P(X>x) = P(X < -x)$$

2) When m $\geq$ 30 the Student distribution approximates the normal. As such, Student (m) = N (0, 1).

3) When m is small the Student (m) has long tails. For example, for the student (1) distribution:

$$P(x < -3) = .1024$$

$$P(x < 3) = .8976$$

Now we can calculate the probability between -3 and +3 for the student (1) distribution. THEREFORE,

> (-3, 3) contains .8976 – .1024 = .7952 of the probability for the Student (1) but (-3,3) contains .9974 of the probability for the N (0, 1)

Now, if $X_1, X_2, ..., X_n$ is a sample from $N(\mu, \sigma)$ then a $1-\alpha$ confidence interval for the mean (as discussed in Chapter 6) is given by:

$$1 - \alpha =$$

$$P\left(-t_{\alpha/2}(n-1) < \frac{\bar{x} - \mu}{S_x/\sqrt{n}} < t_{\alpha/2}(n-1)\right)$$

$$\bar{x} \pm S_x/\sqrt{n}\ t_{\alpha/2}(n-1)$$

where $\alpha/2$ is the probability in the tails of the t distribution. The confidence interval is given below:

$$\bar{x} \pm S_x/\sqrt{n}\ t_{\alpha/2}(n-1)$$

where $\alpha/2$ is the probability in the tails of the t-distribution and the above formula is an exact $(1-\alpha)$ CI for μ.

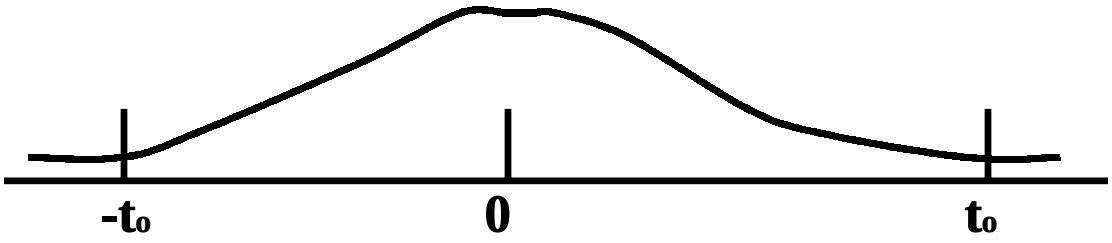
Now to test: $\mathbf{H_o: \mu = \mu_o}$ vs $\mathbf{H_a: \mu \neq \mu_o}$

with an observed value T_o for T we compute the p-value which is equal to

$$= \mathbf{P(T > T_o \text{ or } T < T_o)}$$

$$= \mathbf{2\,(P\,(T > T_o))}$$ because of symmetry of the distribution

where $\tilde{t}$ Student (n-1). The graph below gives the t distribution with value t_o.



The shaded area gives the probability that is calculated above.

Example

In a study examining DDT poisoning a researcher measures the "absolute refactory period" (ARF) or time for a nerve to recover after stimulus. For the sample data(n=4) we have:

$$\bar{x} = 1.75 \text{ msec}$$
$$s = .1291$$
$$s/\sqrt{n} = .0645 \text{ msec}$$

The ARF for unpoisoned animals is 1.3 msec. DDT poisoning should slow nerve recovery. Test this.

$$H_o: \mu = 1.3$$
$$H_a: \mu > 1.3$$

Substitute into the t statistic formula.

$$t = \frac{1.75 - 1.3}{.0645} = 6.98$$

Our t statistic is 6.98 with df=3, Using these numbers and the t table we obtain p=.005. We

conclude there is strong evidence that the ARF has increased.

A 90% confidence interval (CI) with df = 3 gives a t-value with, $t_{\alpha/2} = 2.353$. We substitute into the CI formula and obtain (1.6, 1.9) msec. as our CI.

SPSS EXAMPLE ONE

Click above to see how to conduct a single sample t-test.

III Transformations

Suppose we do a normal probability plot and conclude the assumption of normality is unwarranted for $X_1, X_2,..., X_n$

Many times a simple transformation can 'fix' things and make the data more normally distributed. For example, consider the **logarithmic transformation** for skewed or long tailed data.
If the X_i are always > 0 then replace X by Y=ln X where ln is the logarithmic function. Often Y_1,

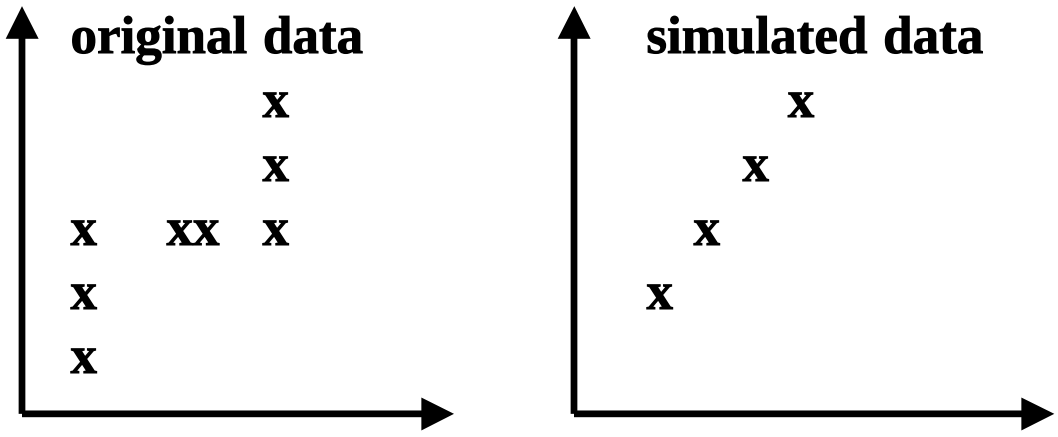
$Y_2,...,Y_n$ will look like a sample from a normal when $X_1, X_2,..., X_n$ doesn't.
Examine a plot of the data to see if the data has a long tail or large skew. If there is you can try the logarithmic transformation.

Example: Survival Times of Guinea Pigs

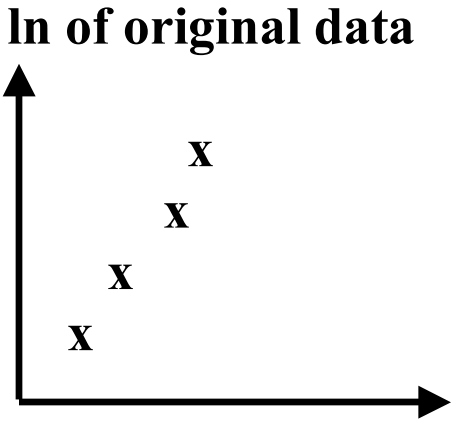
Pigs are injected with bacteria and survival times are recorded.

n = 72

The normal probability plots are given below. The original data is skewed. The simulated data from the normal shows the line pattern typical of normal data.



The ln transformed data is normally distributed as can be seen in the normal probability plot below.



IV Is n large?

If transformation does not achieve normality we ask if n is large enough so that we can use Z.

$$Z = \frac{x - \bar{\mu}}{S_x / \sqrt{n}} \sim N(0, 1)$$

How do we decide if n is large enough?

Answer: Do simulations of the sampling distribution of Z from non-normal distributions.

V Large Sample Methods

If we decide n is "large" we can use approximation methods based on the Central Limit Theorm (CLT).

VI Distribution-Free Methods or Non-Parametric Methods

These methods carry out exact tests and construct exact confidence intervals under very weak assumptions.

NOTE: If we feel confident that a certain assumption holds, then we should use tests and CI's, which depend on those assumptions.

WHY?

1) Tests will be more powerful. In other words, they will detect deviations from H_o with greater probability.

2) Confidence intervals will be more accurate; i.e. shorter.

Now we will see how to compare two means using the t distribution.

Comparing Two Means

(1) Paired Comparisons

Suppose we have a population and quantitative variables Y_1, Y_2 defined on π (π_1, π_2,...)

We want to compare the means of these two variables:

$$\mu_{y1} , \mu_{y2}$$

We want to test the following hypothesis concerning equality of means:

$$\textbf{TEST} \quad H_o: \mu_{y1} = \mu_{y2} \ \text{ vs } \ H_a: \mu_{y1} \neq \mu_{y2}$$

or, another way of saying this is that the difference in means is:

$$H_o: \mu_{y1} - \mu_{y2} = 0 \ \text{ vs } \ H_a: \mu_{y1} - \mu_{y2} \neq 0$$

Hence we want to make inferences about the difference in the two means.

$$\mu_{y1} - \mu_{y2}$$

First generate a sample and estimate using

$$\overline{Y_1} - \overline{Y_2} \quad \textit{on the sample}$$

i.e.,

$Y_{11}(\pi_1), Y_{12}(\pi_2) \ldots (\pi_n)$

$Y_{21}(\pi_1), Y_{22}(\pi_2) \ldots (\pi_n)$

We now have a **paired comparison** (the same person measured twice).

Writing the difference as $d = Y_1 - Y_2$ we can calculate the mean of the differences.

$$\bar{d} = \frac{1}{n} \sum d$$

the variance of the difference is given by

$$s^2_d = \frac{1}{n - 1} \sum (d - \bar{d})^2$$

Now by the CLT we have:

$$\frac{\bar{d} - \mu(y_1 - y_2)}{S_d / \sqrt{n}} \; N \sim (0,1)$$

for large n. With a corresponding confidence interval given below:

$$\bar{d} \pm s_d / \sqrt{n} \; Z_{\alpha/2}$$

If n is not large we use exact methods.

Exact normal methods

We first check to see if the assumption "$d_1, d_2, ..., d_n$ is a sample from the normal" makes sense via a normal probability plot.

If the data is normally distributed we use the t distribution

$$T = \frac{\bar{d} - \mu(y_1 - y_2)}{S_d / \sqrt{n}} \sim Student\ (n-1)$$

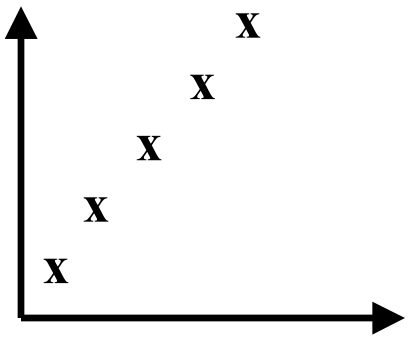
for tests and confidence intervals.

E.g. Design of Controls

A researcher examines the time to move indicator a fixed distance. The indicator has either a right or left thread. Each person works with both threads and time is recorded in seconds to complete the task. The difference in time for right and left thread is recorded in the last column.

	right thread(Y_1)	left thread(Y_2)	Y_1-Y_2=D
Subject 1	**113**	**137**	**-24**
Subject 2	**105**	**105**	**0**
⋮			
Subject 25	**88**	**123**	**-35**

Normal probability plot below for D indicates data is normal.



looks fine against simulated data from N(0,1), n=25

We assume normality for the differences. The test of hypothesis is given below:

H_o: $\mu = 0$ (no difference in turning right or left thread)

H_a: $\mu < 0$ (right - left thread less than zero, right thread easier to turn)

The mean and standard error of the differences is:

$$SD = 22.94/\sqrt{25} = 4.59$$

$$\overline{X} = -13.32 \text{ sec.}$$

The t-statistic is equal to:

$$t = -13.32 / 4.59 = -2.90 \text{ with } df = 24 \ (25-1=24)$$

The p-value is p = .005. There is strong evidence against H_o. **There is overwhelming evidence that the right thread is easier to turn than left.**

SPSS EXAMPLE TWO

Click above to see how to perform the paired difference t-test.

View the Video – Comparing Two Means

(2) Two Independent Samples

Previously we had one population measured twice and called this the paired t test. Now we have 2 populations π_1, π_2 and quantitative variable Y_1, defined on π_1, Y_2 on π_2.

We want to compare the means of each population:

$\mu_{y1} - \mu_{y2}$ using the sample means $\overline{Y}_1 - \overline{Y}_2$

The standard error of estimate for the difference in means is:

$$\sqrt{\frac{s^2_{y1}}{n_1} + \frac{s^2_{y2}}{n_2}}$$

By CLT we have:

$$Z = \frac{\overline{Y}_1 - \overline{Y}_2 - (\mu_{y1} - \mu_{y2})}{\sqrt{\frac{s^2_{y1}}{n_1} + \frac{s^2_{y2}}{n_2}}} \sim N(0,1)$$

for large n_1, n_2 we can derive the necessary tests and CI's.

If the normal probability plot indicates that $Y_1,...,Y_n$ is a sample from a normal distribution and $Y_2...Y_m$ is a sample from the normal distribution. Then use the exact method below that has a t-distribution.

$$T = \frac{\overline{Y}_1 - \overline{Y}_2 - (\mu_{y1} - \mu_{y2})}{\sqrt{\frac{s^2_{y1}}{n_1} + \frac{s^2_{y2}}{n_2}}} \sim \textit{Student (df)}$$

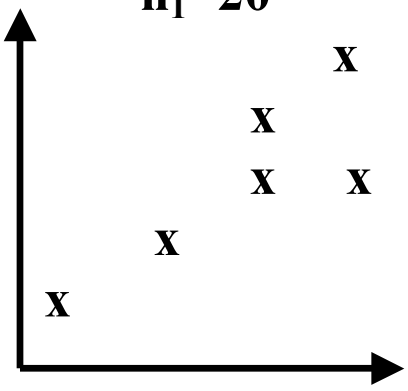
See the textbook for tests and CI's.

Example: **Effect of type of feed on weight gain**

A researcher examines the effect of a new feed on the weight gain of animals. There are two groups of animals, a control group, and a treatment group. She wants to test if the treatment weight is greater than the control. The normal probability plots are given below for both groups.

Control

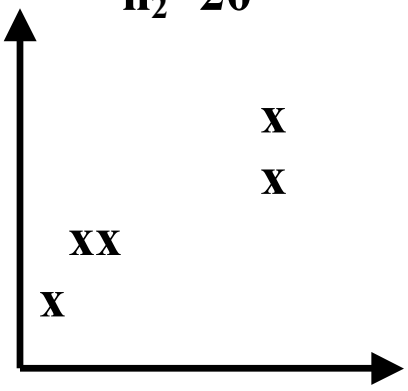
$n_1=20$



normal prob. plot

Treatment

$n_2=20$



normal prob. plot

The plots indicate that it is appropriate to assume normality. **REMEMBER:** In contrast to the paired comparison test where one person is measured twice in the two-sample test there are two distinct groups and each person is measured once.

We have the mean and standard error for each group:

$$\bar{Y}_1 = 366.3 \qquad s_{y1}/\sqrt{20} = 11.0$$

$$\bar{Y}_2 = 403.3 \qquad s_{y2}/\sqrt{20} = 9.6$$

The test of hypothesis is given below. Note the alternative is directional.

$$H_o: \quad \mu_{y1}-\mu_{y2} = 0 \text{ vs}$$
$$H_a: \quad \mu_{y1}-\mu_{y2} < 0 \text{ (directional)}$$

$$t = -2.47$$

P-value = .018 (with *df* = 19)

The df is calculated from the smaller sample size of the two groups since we do not assume equal variances. We conclude that there is strong evidence against H_o that the Control and Treatment weight gain is equal. The treatment increases weight gain.

A 0.95 CI for $\mu_2 - \mu_1$ is (5.58, 67.72) with df = 19.

SPSS EXAMPLE THREE

Click above to see how to compare two groups using the independent samples t test.

Inference for σ^2

We have seen how to make inferences concerning the mean or location of a distribution. Now we will see how to make inferences concerning the spread or variance of a distribution. Let $X_1, X_2, \ldots, X_n$ be a sample from a distribution. We estimate σ by:

$$s_x = \sqrt{\frac{1}{(n-1)} \sum (x - \bar{x})^2}$$

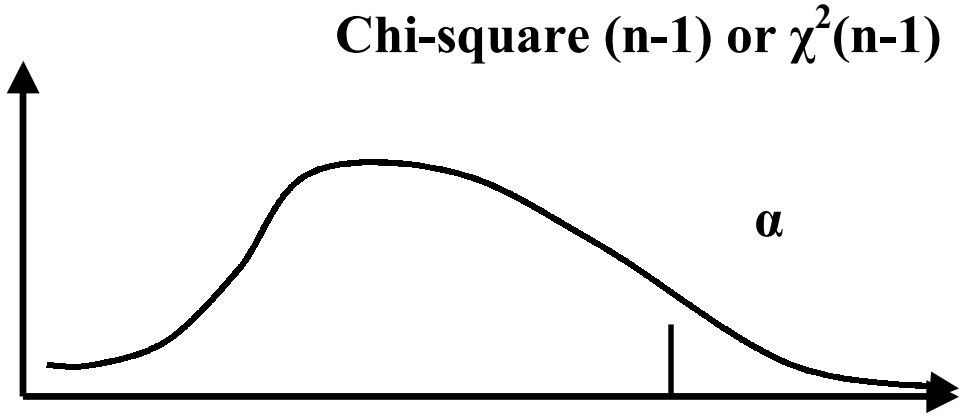
the sample standard deviation.

FACT: If $X_1, X_2,..., X_n$ is a sample from a $N(\mu,\sigma)$ distribution then the following quantity is distributed as a Chi-square(n–1) distribution.

$$\frac{(n-1)s^2_x}{\sigma_2} \sim \text{Chi-square (n-1)}$$

degrees of freedom

The Chi-square distribution is a non-symmetric distribution.



$$\chi^2_{\alpha}(n-1)$$

The value of Chi-square with 10 degrees of freedom and $\alpha = .05$ is given in the χ^2 distribution table as:

$$\chi^2_{.05}(10) = 18.31$$

Example: We want to see if the variance on SSHA test is greater than 20. We sample n=21 people and calculate the following variance , $s^2 = 40$. We want to test:

$$H_0: \sigma^2 = 20$$
$$H_a: \sigma^2 > 20$$

Substitute into the formula on the previous page and we obtain:

$$\chi^2 = \frac{(n-1)s^2}{\sigma_2} = \frac{(21-1)\,40}{20} = 40$$

We look up χ^2 with (21-1) = 20 degrees of freedom in the χ^2 table and obtain the p-value = .005.

We have strong evidence against H_o, and conclude that the variance is larger than 20 on the SSHA test.

TESTING $\sigma_1^2 = \sigma_2^2$

The above Chi-square test is used for one sample, however, many times researchers wish to compare the spread or variance in two samples.

Suppose we have two samples:

$$\mathbf{Y_1 \ldots Y_n \text{ from } N(\mu_1,\sigma_1)}$$

$$\mathbf{Y_2 \ldots Y_m \text{from } N(\mu_2,\sigma_2)}$$

estimate σ^2_1 by

$$\frac{1}{n-1} \sum (Y - \overline{Y}_1)^2 = s^2_1$$

the sample two variance

$$\sigma^2_2 \text{ by } \frac{1}{m-1}\sum (Y - \overline{Y}_2)^2 = s^2_2$$

In order to test the following hypothesis concerning the equality of variances:

$$H_0: \sigma^2_1 = \sigma^2_2$$

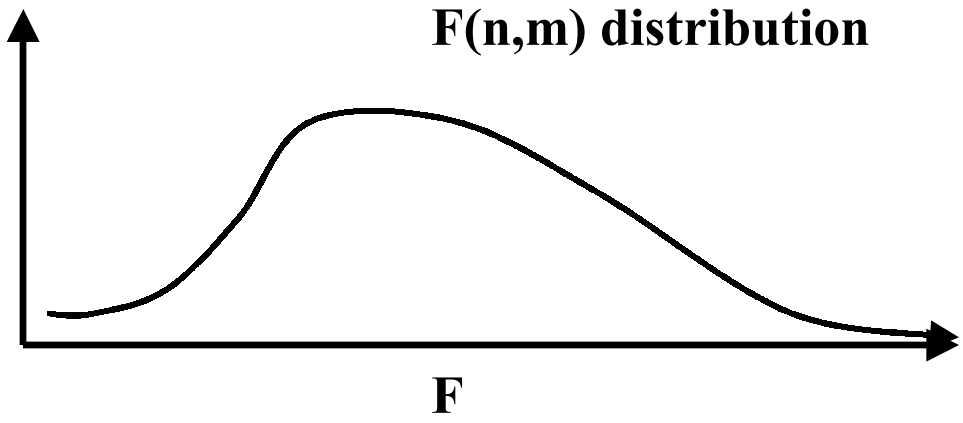
$$H_a: \sigma^2_1 \neq \sigma^2_2$$

We calculate the test statistic F, defined as the ratio of two variances

$$F = \frac{s^2_1}{s^2_2}$$

now F = 1 if H_0 is true

Fact: Under H_o our test statistic will follow an F distribution with n and m degrees of freedom



When calculating the F statistic put the larger variance over smaller variance, then the F ratio $s_1^2/s_2^2 \geq 1$ and consequently the p-value is doubled. i.e. p-value = 2 P ($F > s_1^2/s_2^2$). See the text and the example below.

Two Groups, a Treatment and a Control Group

The researcher is using the Degree Reading Power (DRP) test to compare two groups. Suppose the researcher wants to see if the variance is the same in both groups. The data is given below:

Control	**s = 17.15**	**n=23**
Treatment	**s = 11.01**	**n=21**

We want to test if the variances for both groups are equal.

Test $H_0: \sigma^2_c = \sigma^2_T$ **(variances equal)**

$H_a: \sigma^2_c \neq \sigma^2_T$ **(variances different)**

Put the larger variance over the smaller and calculate the F statistic:

$$F = \frac{s^2_{1\ (larger)}}{s^2_{2\ (smaller)}} = \frac{(17.15)^2}{(11.01)^2} =$$

$$\frac{294.12}{121.22} = 2.42$$

The F-statistic is equal to F = 2.42.

The F distribution under H_o has 23-1=22 degrees of freedom for the larger variance, and 21-1=20 degrees of freedom for the smaller variance.

We write this as F(22,20).

We use table F distribution table, column 25, row 20 and obtain an approximate value of 2.40, which corresponds to a p-value =.025. We double the p-value, 2(.025) = .05 (since the alternative is two-sided). We have some evidence against H_o. We conclude that the variances are different.

Pooling

Consider the problem of comparing means in 2 independent samples, Y_1, Y_2 that was discussed previously. Suppose you know or have decided that the two groups have:

$$\sigma^2_{y1} = \sigma^2_{y2} = \sigma^2$$

They have common variance. We speak of **pooling**, the samples for one estimate of variance.

$$\mathbf{S^2 = \frac{[\sum(Y - \overline{Y}_1)^2 + \sum(Y - \overline{Y}_2)^2]}{n_1 + n_2 - 2}}$$

For inferences about $\mu_1 - \mu_2$ we use the t-statistic.

$$T = \frac{\overline{Y}_1 - \overline{Y}_2 - (\mu_1 - \mu_2)}{S\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \sim \textit{Student (df)}$$

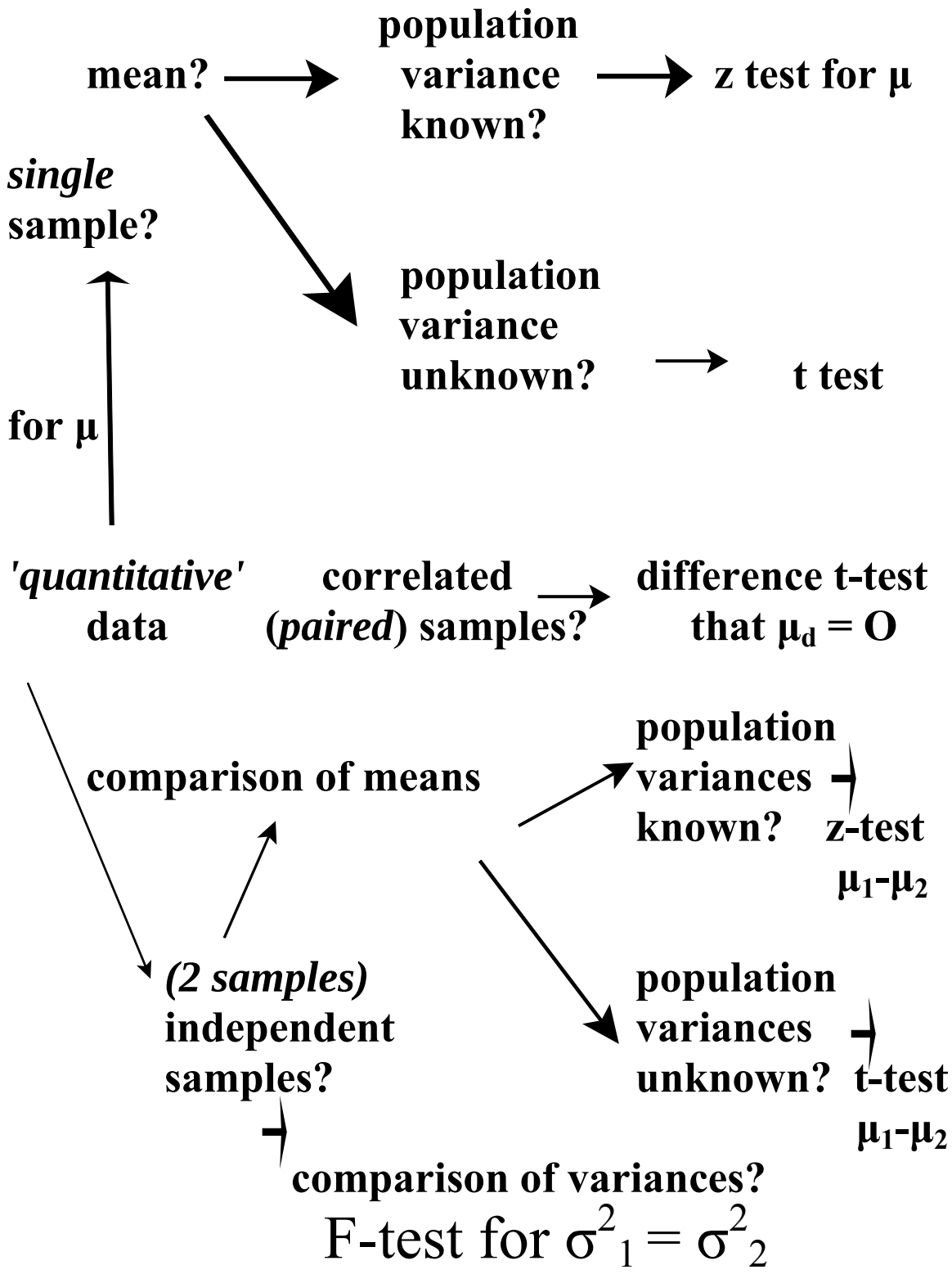
which is distributed as a t distribution under the null hypothesis.

$$T \sim \text{Student } (n_1 + n_2 - 2)$$

This is not that good an idea since the researcher assumes that the variances are equal. This pooling without testing the assumption of equality of variances is commonly done.

On the next page is a flow chart to help you decide what test should be used in a given situation. Most students find problem identification the key, once the problem is identified then the proper test statistic falls into place. There are 3 starting branches….single sample…..correlated(paired) samples …or two samples…..once this is decided determine if the question concerns the mean or variance….and follow the branch to the end point which is the proper test.

A Decision Tree for Basic Tests of Significance



The tree above begins in the middle of the page with quantitative data. There are three choices or branches given, correlated, single or 2 independent samples. The problems in this chapter fall into one of the above classes. Once the class has been determined a careful reading of the problem will determine if inferences are to be made about the mean or variance. The corresponding test statistic is given in the tree. Students find the tree useful since it organizes the decision process in a logical manner.

Chapter 7 Summary

Significance tests and confidence intervals are introduced for the mean of a population. The **one-sample t statistic** has a **t-distribution.** Every **t distribution** has a **degrees of freedom (*df*).** The **paired difference t-test** is introduced as an example of the single sample t test.

The **two sample t-test** is seen as a procedure for comparing two independent groups. The **pooled** and **non-pooled** degrees of freedom are calculated for the above t procedure.

The **F statistic** is a ratio of variances. It is used to compare two population variances to see if they are equal.

Remember: Tips for Success

1) Read the text.

2) Read the notes.

3) Try the assignment.

4) If needed, try the exercise questions.

5) Try the simulations and view the videos if you need more help with a concept.

6) Try the self tests for practice on each chapter of the text at www.whfreeman.com/ips

7) Steady Work = Success