ST 501

Experimental Statistics for Biological Sciences I

Lecture Notes

M. Davidian

Department of Statistics

North Carolina State University

©1998 by Marie Davidian

Contents

1 Introduction to Statistics

Complementary Reading: Steel, Torrie, & Dickey (STD), Chapter 1

1.1 Motivation

The purpose of the discussion in this chapter is to stimulate you to start thinking about the important issues upon which statistical methodology is based.

STATISTICS: The development and application of theory and methods to the collection (design), analysis, and interpretation of observed information from planned (or unplanned) experiments.

BIOMETRY: The development and application of statistical methods for biological experiments (often planned).

TYPICAL OBJECTIVES: Some examples

(i) Determine which of 3 fertilizer compounds produces highest yield

(ii) Determine which of 2 drugs is more effective for controlling a certain disease in humans

(iii) Determine whether an activity such as smoking causes a response such as lung cancer

Examples (i) and (ii) represent situations where the scientist has the opportunity to plan (design) the experiment. Such a preplanned investigation is called a **controlled experiment**. The goal is to compare **treatments** (fertilizers, drugs).

In example (iii), the scientist may only **observe** the phenomenon of interest. The **treatment** is smoking, but the experimenter has no control over who smokes. Such an investigation is called an **observational study**.

In this course, we will focus mostly on **controlled experiments**, which leads us to thinking about **design of experiments**.

KEY ISSUE: We would like to make conclusions based on the **data** arising as the result of an experiment. We would moreover like the conclusion to apply **in general**. For example, in (i), we would like to claim that the fertilizers produce different yields in general based on the particular data from a single experiment.

1.2 Some terminology

POPULATION: The entire "universe" of possibilities. For example, in (ii), the population is **all** patients afflicted with the disease.

SAMPLE: A part of the population that we can **observe**. Observation of a sample gives rise to information on the phenomenon of interest, the **data**.

Using this terminology, we may refine our statement of our objective. We would like to make statements about the **population** based on observation of **samples**. For example, in (ii), we obtain 2 **samples** of diseased patients, and subject one to drug 1, the other to drug 2.

In agricultural, medical, and other biological applications, the most common objective is the comparison of two or more **treatments**. We will thus often talk about statistical **inference** in the context of comparing treatments.

PROBLEM: A sample we observe is only one of **many** possible samples we might have seen instead. That is, in (ii), one sample of patients would be expected to be similar to another, but not identical. Plants will differ due to biological variability, which may cause different reactions to fertilizers in example (i).

RESULT: There is **uncertainty** about **inference** we make on a population based on observation of samples.

1.3 Basis for statistical methodology

The premise of statistical inference is that we attempt to control and assess the **uncertainty** of **inferences** we make on the population of interest based on observation of samples.

KEY: Allow explicitly for **variation** in and among samples. Statistics is thus often called the "study of variation."

FIRST STEP: Set up or **design** experiments to **control variability** as much as possible. This is certainly possible in situations such as field trials in agriculture, clinical trials in medicine, reliability studies in engineering, and so on. It is not entirely possible in observational studies, where the samples "determine themselves."

PRINCIPLES OF DESIGN: **Common sense** is the basis for most of the ideas for designing scientific investigations:

- *Acknowledgment of potential sources of variation.* Suppose it is suspected that males may react differently to a certain drug from females. In this situation, it would make sense to assign the drugs to the samples with this in mind instead of with no regard to the gender of participants in the experiment. If this assignment is done correctly, differences in treatments may be assessed despite differences in response due to gender. If gender is ignored, actual differences in treatments could be obscured by differences due to gender.

- *Confounding.* Suppose in example (i), we select all plants to be treated with fertilizer 1 from one nursery, all for fertilizer 2 from another nursery, etc. Or, alternatively, suppose we keep all fertilizer 1 plants in one greenhouse, all fertilizer 2 plants in another greenhouse, etc. These choices may be made for convenience or simplicity, but introduce problems: under such an arrangement, we will not know whether any differences we might observe are due to actual differences in the effects of the drugs or to differences in the origin or handling of plants.

  In such a case, the effects of treatments are said to be **confounded** with the effects of, in this case, nursery or greenhouse.

  Here is another example. Consider a clinical trial to compare a new, experimental treatment to the standard treatment. Suppose a doctor assigns patients with advanced cases of disease to a new experimental drug and assigns patients with mild cases to the standard treatment, thinking that the new drug is promising and should thus be given to the sicker patients. The new drug may perform poorly relative to the standard drug because the conditions under which it was tested were more serious. The effects of the drugs are **confounded** with the seriousness of the disease.

MORAL: To take adequate account of variation and to avoid confounding, we would like the elements of our samples to be as **alike** as possible **except** for the treatments. Assignment of treatments to the samples should be done so that potential sources of variation do not obscure treatment differences. No amount of fancy statistical analysis can help an experiment that was conducted without paying attention to these issues!!!

SECOND STEP: Interpret the **data** by taking appropriate account of sources and magnitude of variation and the design; that is, by **assessing** variability.

PRINCIPLE OF STATISTICAL INFERENCE: Because variation is involved, and samples represent only a part of the population, we may not make statements that are **absolute**. Rather, we must temper our statements by acknowledging the **uncertainty** due to variation and sampling. Statistical methods incorporate this **uncertainty** into statements interpreting the **data**.

The appropriate methods to use are dictated to a large extent by the **design** used to collect the data. For example:

- In example (ii) comparing 2 drugs, if we are concerned about possible gender differences, we might obtain 2 samples of men and 2 samples of women, and treat one of each with drug 1, the other of each with drug 2. With such a design, we should be able to gain insight into differences actually due to the drugs as well as variability due to gender. A method to assess drug differences that takes into account the fact that part of variation observed is attributable to a known feature, gender, would then be appropriate. Intuitively, a method that did not take this into account would be less useful.

MORAL: Design and statistical analysis go hand in hand.

1.4 First notions of experimental design

It is obvious from the above discussion that a successful experiment will be one where we consider the issues of variation, confounding, and so on **prior** to collecting data. That is, ideally, we **design** an experiment before **performing** it. Some fundamental concepts of experimental design are as follows.

RANDOMIZATION: This device is used to ensure that samples are indeed as alike as possible except for the treatments. **Randomization** is a treatment assignment mechanism – rather than assign the treatments **systematically**, assign them so that, once all acknowledged sources of variation are accounted for, it can be assumed that no obscuring or confounding effects remain. Here is an example.

Suppose we wish to compare 2 fertilizers in a certain type of plant (we restrict attention to just 2 fertilizers for simplicity). Suppose we are going to obtain plants and then assign them to receive one fertilizer or the other. How best to do this?

We would like to perform the assignment so that no plant will be more likely to get a potentially "better" treatment than another. We'd also like to be assured that, prior to getting treatment, plants are otherwise basically alike. This way, we may feel confident that differences among the treatments that might show up reflect a "real" phenomenon. A simple way to do this is to obtain a sample of plants from a single nursery, which ensures that they are basically alike, and then determine 2 samples by a **coin flip**:

- Heads = Fertilizer 1, Tails = Fertilizer 2
- Ensures that all plants had an equal chance of getting either fertilizer; that is, all plants are **alike** except for the treatment
- This is a **random** process

Such an assignment mechanism, based on chance (random), is called **randomization**. Randomization is the cornerstone of the design of controlled experiments.

POTENTIAL PROBLEM: Ultimately, we wish to make **general** statements about the efficacy of the treatments. Although using randomization ensures the plants are as alike as possible except for the treatments and that the treatments were fairly assigned, we have only used plants from one nursery. If plants are apt to respond to fertilizers differently because of nursery of origin, this may limit our ability to make general statements.

SCOPE OF INFERENCE: The **scope of inference** for an experimental design is limited to the **population** from which the samples are drawn. In the design above, the population is plants from the single nursery. To avoid limited scope of inference, we might like to instead use plants from more than one nursery. However, to avoid **confounding**, we do not wish to assign the treatments systematically by nursery.

EXTENSION: Include more nurseries but use the same principles of treatment assignment. For example, suppose we identify 3 nurseries. By using plants from all 3, we **broaden** the scope of inference. If we repeated the above process of randomization to assign the treatments for **each** nursery, we would have increased our scope of inference, but at the same time ensured that, once we have accounted for the potential source of variation, **nursery**, plants receiving each treatment are basically alike except for the treatments.

One can imagine that this idea of recognizing potential sources of variation and then assigning treatments randomly to ensure fair comparisons, thus increasing the scope of inference, may be extended to more complicated situations. For example, in our example, another source of variation might be the greenhouses in which the treated plants are kept, and there may be more than 2 treatments of interest.

We will return to these issues later. For now, keep in mind that sound experimental design rests on these key issues:

- Identifying and accounting for potential sources of variation
- Randomly assigning treatments after accounting for sources of variation
- Making sure the scope of the experiment is sufficiently broad.

ASIDE: In an **observational study**, the samples are already determined **by the treatments** themselves. Thus, a natural question is whether there is some **hidden** (**confounding**) factor that causes the responses observed. For example, in (iii), is there some underlying trait that in fact causes **both** smoking **and** lung cancer? This possibility limits the inference we may make from such studies. In particular, we can not infer a **causal** relationship treatment (smoking) and response (lung cancer, yes or no). We may only observe that an **association** appears to exist. Because the investigator does not have **control** over how the treatment is applied, interpretation of the results is not straightforward.

If aspects of an experiment can be **controlled**, that is, the experiment may be **designed** up front (treatment assignment determined in advance), there is an obvious advantage in terms of what conclusions we may draw!

An excellent and much more comprehensive discussion of these issues may be found in *Statistics for Experimenters: An Introduction to Design, Data Analysis, and Model Building*, by G.E.P. Box, W.G. Hunter, and J.S. Hunter. The focus of the book is more on industrial rather than biological experimentation, but the basic principles are the same.

We will return to these issues more formally later in the course, but keep them in mind as we progress.

2 Data

Complementary Reading: STD, Chapter 2

2.1 Introduction

The purpose of this chapter is to introduce concepts and associated terminology regarding the collection, display and summary of **data**.

DATA: The actual observations of the phenomenon of interest based on samples from the population.

EXAMPLES:

(i) Yields in bushels/acre from 2 samples of plots planted with a certain variety of wheat, receiving one of 2 types of fertilizer

(ii) Times to abatement of symptoms for 2 samples of patients treated with 2 different drugs

(iii) The numbers of defective items in each of 20 randomly chosen lots from a manufacturing assembly line

(iv) The classifications of each in a group of 40 patients as having "high," "average," or "low" systolic blood pressure.

ONE GOAL OF STATISTICS: Present and summarize data in a meaningful way.

TERMINOLOGY: A **variable** is a characteristic that changes (i.e. shows variability) from unit to unit (e.g. subjects, plots, etc)

Data are observations of **variables**

2.2 Types of variables

QUALITATIVE VARIABLES: Numerical measurements on the phenomenon of interest are not possible. Rather, the observations are **categorical** (as in (iv) above).

QUANTITATIVE VARIABLES: The observations are in the form of **numerical** values

- *Discrete:* Possible values of the variable differ by **fixed amounts**
- *Continuous:* All values in a given range are possible. We may be limited in recording the exact values by the precision and/or accuracy of the measuring device.

EXAMPLES:

(i) *Variable*: Yield. Quantitative, Continuous. (Yield could take on conceivably **any** fraction of bushel/acre; however, we are limited in observing that fraction by how precisely we can measure fractions)

(ii) *Variable*: Time. Quantitative, Continuous. (Limited only by the precision of our time-recording method)

(iii) *Variable*: Number of defectives. Quantitative, Discrete.

(iv) *Variable*: Blood pressure rating. Qualitative (categorical, with 3 categories: high, average, low)

As we will see later, appropriate statistical methods are dictated in part by the nature of the variable in question.

2.3 Random samples

We have already discussed the notion of **randomization** in the design of an experiment. The underlying goal is to ensure that samples may be assumed to be "representative" of a population of interest. In our nursery example in Chapter 1, then, one population of interest might be that of plant subjected to treatment 1. The **random** assignment to determine which plants receive treatment 1 thus may be thought of as an attempt to obtain a representative **sample** from plants from this population, so that data obtained from the sample will be free from confounding and bias.

In general, the way in which we will view a "representative sample" is one chosen so that any member of the population has an equal chance of being in the sample. In the nursery example, then, it is assumed that the plants ending up in the sample receiving treatment 1 may be thought of as being chosen from the population of all plants were they to receive treatment 1. All plants in the overall population may have equally ended up in this sample. The justification for this assumption is that **randomization** was used to determine the sample.

The idea that samples have been chosen randomly is a foundation of much of the theory underlying the statistical methods that we will study. It is instructive, in order to shape our thinking later when we talk about the formal theory, to think about a **conceptual model** for random sampling.

2.4 Sampling – a model

One way to think about random sampling in a simple way is to think about drawing from a box.

POPULATION: Slips of paper in a box, one slip per individual (plant, patient, plot, etc)

SAMPLE: To obtain a **random sample**, draw slips of paper from the box such that, on each draw, all slips have a equal chance of being chosen; i.e. **completely random** selection.

TWO WAYS TO DRAW:

- *With replacement*: Ideal – on each draw, all population members have the same chance of being in the sample. This is simple to think about, but the drawback is that an individual could conceivably end up in the sample more than once.
- *Without replacement*: This is like real life – the number of slips in the box decreases with each draw, because slips are not replaced. Thus, the chance of being in the sample **increases** with the number of draws (size of the sample).

FACT: If the population is **large** relative to the size of the sample to be chosen, this increase is **negligible**. Thus, we may for practical purposes view drawing **without replacement** (real life) as drawing **with replacement** (ideal). The populations of interest (all plants, all patients, all plots, etc) are usually huge relative to the size of the samples we use for experimentation.

WHY IS THIS IMPORTANT? To simplify the theory, standard statistical methods like those we will study in this course are predicated on the notion that the samples are **completely random**, which would follow if the samples were indeed chosen **with replacement**. This fact allows us to view the samples as effectively having been drawn in this way, i.e. **with replacement** from populations of interest. This is the **model** that we will use when thinking about the properties of data in order to develop statistical methods.

PRACTICAL IMPLEMENTATION: We have already talked about using the flip of a coin to randomize treatments and hence determine random samples. When there are several treatments, the **randomization** is accomplished by a random process that is a little more sophisticated than a simple coin flip, but is still in the same spirit. Random number tables were used in the past to decide how to assign the treatments randomly. More recently, random number generators within high-level computing languages may be used to generate quite complex sampling plans. A statistician can help you in developing a randomization plan for your experiment.

For our purposes throughout the course, then, we will assume that if sampling from the population has been carried out with careful objectivity in procuring members of the population to which to assign treatments, and the assignment has been made using these principles, then this model applies. Henceforth, then, we will use the term **sample** interchangeably with **random sample**, and we will develop methods based on the assumption of random sampling. We will use the term **sample** both to refer to the physical process and the data arising from it.

2.5 Presentation of data

TWO GOALS:

- Summarize and display the sample information
- Use the information to learn about the underlying population

We will focus on **quantitative data** for now.

EXAMPLE: To illustrate the methods, we will consider data on cherry trees reported by Cook & Weisberg (1982, *Residuals and Influence in Regression*, p. 66). The data are the diameter at chest height (4.5 ft above ground level), height, and volume for a sample of 31 black cherry trees in the Allegheny National Forest in Pennsylvania. The data were collected as part of a study on determining an easy way to estimate the volume of a tree (and eventually amount of timber in a specified area of the forest) using its height and diameter.

DATA ON 31 CHERRY TREES

Diameter (in)	*Height (ft)*	*Volume(cu ft)*
8.3	70	10.3
8.6	65	10.3
8.8	63	10.2
10.5	72	16.4
10.7	81	18.8
10.8	83	19.7
11.0	66	15.6
11.0	75	18.2
11.1	80	22.6
11.2	75	19.9
11.3	79	24.2
11.4	76	21.0
11.4	76	21.4
11.7	69	21.3
12.0	75	19.1
12.9	74	22.2
12.9	85	33.8
13.3	86	27.4
13.7	71	25.7
13.8	64	24.9
14.0	78	34.5
14.2	80	31.7
14.5	74	36.3
16.0	72	38.3
16.3	77	42.6
17.3	81	55.4
17.5	82	55.7
17.9	80	58.3
18.0	80	51.5
18.0	80	51.0
20.6	87	77.0

GRAPHICAL DISPLAY: The goal of graphical display of data is to provide a visual impression of the characteristics of the data from a sample. The hope is that the characteristics of the sample are a likely indication of the characteristics of the population from which it was drawn. One characteristic that is well-displayed visually is the way in which data values are **distributed** across the sample. That is, which values in a sample of data appear **most frequently** or, equivalently, are **most likely**? Are there certain ranges of values in which very few observations fell in the sample?

HISTOGRAM: A graph suited to this purpose is the **histogram**. A histogram displays the distribution of the data visually by representing the frequency or likelihood of values in the sample by **area**.

FREQUENCY TABLE: In order to form a histogram, as well as to display quantitatively the information on distribution of values in the sample, a tabular summary known as a **frequency table** is constructed. The idea is to partition the range of data values into smaller ranges, and determine the frequency with which observations in the sample fall into each range. The way in which the frequencies differ across the ranges is evidence of the way in which possible data values are distributed for the entire population.

METHOD: There is no specific "right" method for constructing a frequency table. The choice of how to partition the range of data values is flexible depending on the nature of the data. Usually, the way the partitioning is carried out depends on the size of the sample and where the data values fall. For simplicity, certain conventions are adopted – these should be clearly stated. Our method for constructing a histogram will consist of the following steps and conventions:

(i) To partition the range, **group** the data into **class intervals** of **equal length** (in general, the class intervals may be of different lengths; we will adopt equal lengths for simplicity)

Convention: Class intervals **include** their left endpoint, but **not** their right endpoint (we illustrate below for the cherry tree data).

(ii) Calculate the **frequency** for each interval = # of observations falling into each class interval

(iii) Calculate **relative frequency** = **proportion** of observations falling into each class interval, i.e.

$$\text{relative frequency} = \frac{\text{frequency}}{\text{total \# of observations in sample}}.$$

Step (iii) converts the raw frequencies into dimensionless quantities to be viewed on a **relative** basis – the relative frequency is the proportion of observations taking on values in a particular range.

Note that the sum of the relative frequencies **must** equal 1, so that the total proportion = 1.

EXAMPLE: We now construct a frequency table for the diameter data from the cherry tree study. For convenience, the data are presented above in ascending diameter order. Inspection of the data shows that the diameters range from 8.3 in to 20.6 in. Using the convention that class intervals have **equal width**, we need to pick a width that makes sense across the whole range. A width of 2 in seems reasonable – this gives 7 class intervals in the range of the data.

Here is the frequency table. Note that, using the convention in (i) above, the observation 12.0 in falls in the **3rd** class interval, not the 2nd (12.0 is the left endpoint of the 3rd interval).

FREQUENCY TABLE FOR CHERRY TREE DATA

Class Interval	*Frequency*	*Relative Frequency*	*Height*
8 – 10	3	0.097	0.049
10 – 12	11	0.355	0.177
12 – 14	6	0.193	0.097
14 – 16	3	0.097	0.049
16 – 18	5	0.161	0.081
18 – 20	2	0.065	0.032
20 – 22	1	0.032	0.016
	31	1.000	

CONSTRUCTING A HISTOGRAM: A histogram represents the information in the frequency table graphically. Specifically, it represents the **relative frequencies** by **area**, giving a pictorial display of how the data values are **distributed**. Note that since

$$\text{Sum over all class intervals of relative frequency} = 1,$$

we have

$$\textbf{Total area of histogram} = \mathbf{1}.$$

A picture of the distribution of the data values gives an idea of how **likely** they were to appear in the sample. Hopefully, if the sample is truly "representative," this gives an indication of the way they are likely to occur in the entire population.

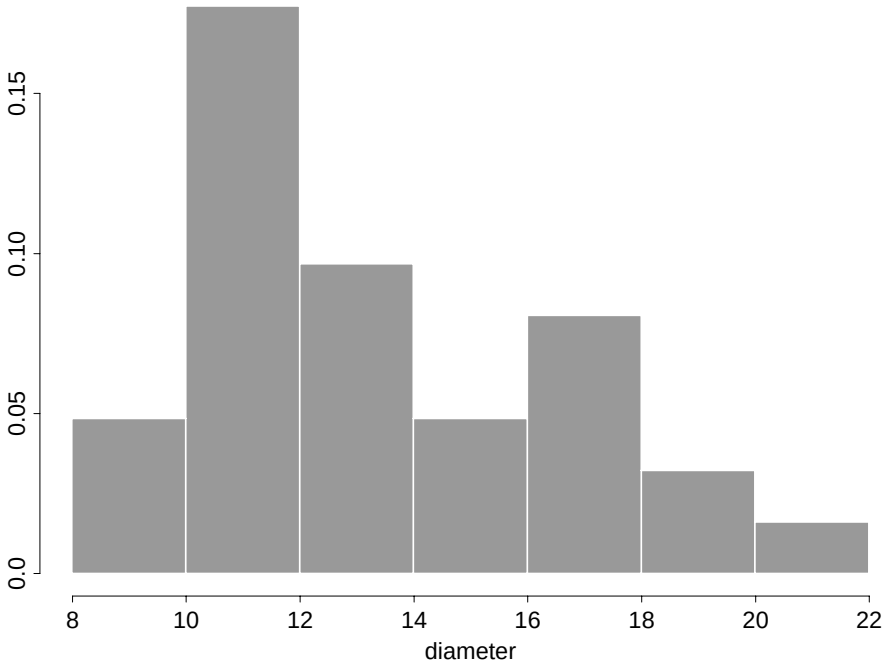
To construct the histogram, one forms for each class interval a **block** with **area** equal to the **relative frequency** for the interval. Recall that

$$\text{area} = \text{width of interval} \times \text{height}$$

so that one constructs the block such that

$$\text{height} = \frac{\text{relative frequency}}{\text{width}}.$$

EXAMPLE: We now construct a histogram for the cherry tree diameter data. The blocks for each class interval have heights as calculated in the frequency table.



Note that the histogram gives the visual impression that a large proportion of the trees have diameters in the range 10 to 14 in. (From the frequency table, we see that this proportion is $0.355 + 0.193 \approx 0.55$, or 55%.) The proportion of trees "tails off" as diameters get larger.

SUMMARY: The **frequency table** and **histogram** are two convenient ways to present the distribution of the data. The frequency table does this numerically, the histogram does it graphically.

The hope is that the form of the frequency table of histogram for the data will reflect the **distribution** of the values in the population. For example, if we could measure the diameter of **every** cherry tree in the forest (the population), and were to construct a histogram of them, we would hope that the histogram based on our sample of 31 trees from this population would bear resemblance to the histogram for the population. We will return to the idea of a "population histogram" in the next chapter.

2.6 Descriptive statistics

A histogram provides an overall visual impression of the character of the data. From it, we get a sense of the "center" of the data, how "spread out" they are, and so on. It is also desirable to summarize these notions **quantitatively**.

IDEA: Summarize data by quantifying the notions of "**center**" and "**spread**" or "**dispersion**." That is, define relevant measures that summarize these notions.

ONE OBJECTIVE: By quantifying center and spread for a sample, we hope to get an idea of these same notions for the **population** from which the sample was drawn. As above, if we could measure the diameter of **every** cherry tree in the forest, and quantify the "center" and "spread" of all of these diameters, we would hope that the "center" and "spread" values for our sample of 31 trees from this population would bear resemblance to the population values.

NOTATION: For the remainder of this discussion (and, in fact, the course), we adopt the following standard notation to describe a sample of data:

$$n = \text{ size of the sample} \quad Y = \text{ the variable of interest}$$

$$Y_1, Y_2, \ldots, Y_n = \text{ observations on the variable for the sample.}$$

MEASURES OF CENTER: There are several different ways to define the notion of "center." Here, we focus on the two most common:

- *Mean or Average*: This is the most common notion of center. For our data, the **sample mean** is defined as

$$\bar{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i = \textbf{sample mean}. \tag{2.1}$$

  That is, the sample mean is the **average** of the sample values. The notation "$\bar{Y}$" is standard; the "bar" indicates the "averaging" operation being performed. We may think of $\bar{Y}$ as the "balancing point" of the histogram for the data.

  Note that in (2.1) we have used **summation** notation; this is described in detail in section 2.10.

The **population mean** is the corresponding quantity for the population. As is commonplace, we will use the Greek symbol μ to denote the population mean:

$$\mu = \text{ mean of all possible values for the variable for the population}$$

One may think of μ as the value that would be obtained if the averaging operation in (2.1) were applied to **all** the values in the population, e.g., all cherry trees.

- *Median*: The **median** is the value such that, when the data are put into **ascending order**, 50% of the observations are above and 50% are below this value. Thus, the median quantifies the notion of center differently from the mean – the assessment of "center" is based on the **likelihood** of the observations (their distribution) rather than by averaging.

  It should be clear that, with this definition, the value chosen as the **median** of a sample need not be unique. If n is **odd**, then the definition may be applied exactly – the median is the "center" observation. For example, if the ordered sample data are

  3 5 6 8 11

  then the median is 6, the value in the "center."

  However, is n is **even**, the "center" value is no longer clear, e.g. suppose now that the data are

  3 5 6 8

  According to the definition, **any** number between 5 and 6 would qualify as the median! To avoid confusion in such situations, a **convention** is adopted. In particular, when n is **even**, the median is defined as the **average** of the two "middle" values. In the example here, then the median would be defined as

  $$(5 + 6)/2 = 5.5.$$

  The notion of median for a population is obvious – the median for the population is the value such that 50% of **all values** in the population are on either side.

- *Mean versus Median*: It should be clear that the mean and median **need not** coincide. For example, for the data set

  3 5 6 8 200

  the **median** = 6 but the **mean** = 222/5 = 44.4! Furthermore, from the looks of this data set, 200 seems to be very **unusual** when compared with the rest of the data; perhaps it is a mistake. This illustrates another feature: the **median** is less likely to be affected by an "aberrant" or "outlying" observation – one that does not fit in with the pattern exhibited by the rest of the data. As a result, the median is preferred over the mean for particular types of data, such as incomes or housing prices. A single "chairman of the board" in a data set of annual income values would produce a **mean** that would give a very inflated impression of the "center" of annual incomes, while the median would give a more realistic picture, as it would not be as affected by the chairman's huge salary!

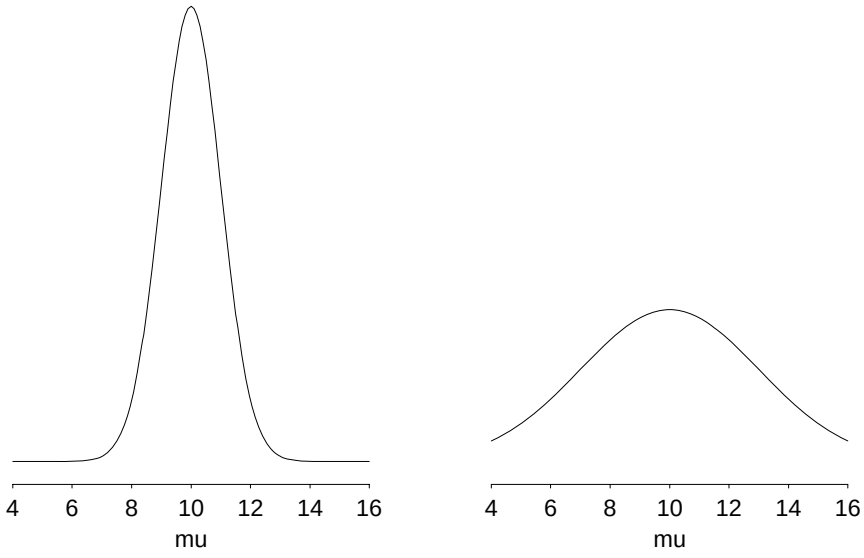
EXAMPLE: For the cherry tree diameter data, we have $n = 31$. The sample mean is

$$\bar{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i = \frac{8.3 + \cdots + 20.6}{31} = \frac{410.7}{31} = 13.25.$$

As n is odd, the median may be determined immediately as the center value 12.9.

MEASURES OF SPREAD: Two data sets (or populations) may have the same **mean**, but may be "spread" about this mean value very **differently**. For illustration, consider the following 2 "smoothed out" histograms.

Both histograms have the same mean, 10. The one on the left has almost all of the likely values concentrate the range 7 to 12; values outside this range are very unlikely. The one on the right, on the other hand, has nonnegligible area over the whole displayed range from 4 to 16. Thus, the "spread" of likely values is much greater.



For a set of data, a measure of **spread** for an individual observation is the **sample deviation**

$$(Y_i - \bar{Y}).$$

Intuition suggests that the **mean** or **average** of these deviations ought to be a good measure of how the observations are spread about the mean for the sample. However, it is straightforward to show that (see section 2.10)

$$\sum_{i=1}^{n}(Y_i - \bar{Y}) = 0$$

for **any** set of data!!

Reason: Some deviations are negative, some are positive, and they cancel each other out.

Remedy: Devise a measure that maintains the **magnitude** of each deviation but **ignores** their **signs**.

- A sensible measure of "spread" is concerned with how "spread out" the observations are; direction is not important.

Variance and standard deviation: The most common idea of ignoring the sign of the deviations is to **square** them.

- *Sample variance*: For a sample of size n, this is defined as

$$\frac{1}{n-1}\sum_{i=1}^{n}(Y_i - \bar{Y})^2. \tag{2.2}$$

  This looks like an average, but with $(n-1)$ instead of n as the divisor – the reason for this will be discussed later in the course. ($(n-1)$ is called the **degrees of freedom**, as we'll see.)

  Note that if the original data are measured in some units, then the sample variance has (units)^2. Thus, sample variance does not measure spread on the **same** scale of measurement as the data.

- *Sample standard deviation*: To get a measure of spread on the **original scale** of measurement (with the **same** units as the data), take the **square root** of the sample variance. This quantity, which thus measures spread in data units, is called the **sample standard deviation**

$$s = \sqrt{s^2} = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(Y_i - \bar{Y})^2}.$$

Computation: The sample variance and standard deviation are generally available on hand-held calculators. It is instructive, however, to examine hand calculation of them, as we will discuss momentarily.

First, it may be shown (see section 2.10) that

$$\sum_{i=1}^{n}(Y_i - \bar{Y})^2 = \sum_{i=1}^{n} Y_i^2 - \frac{1}{n}\left(\sum_{i=1}^{n} Y_i\right)^2. \tag{2.3}$$

This formula is of interest for several reasons:

- Formula (2.3) is preferred when doing hand calculation over (2.2). The reason is to avoid **propagation of error**. In particular, it is common to **round** the value of the sample mean $\bar{Y}$ before calculating the deviations $(Y_i - \bar{Y})$. The error in performing this rounding will carry through the calculation. Because (2.3) involves only sums, such error does not arise.

  The second term on the right hand side of (2.3) may be written as $n\bar{Y}^2$, but this is not recommended for hand calculation, for the same reason.

- Breaking the sum of squared deviations into two pieces highlights a concept that will be important to our thinking later. Write

$$SS = \sum_{i=1}^{n} (Y_i - \bar{Y})^2.$$

  This is the sum of squared deviations, but is often called the **Sum of Squares adjusted for the mean**. The reason may be deduced from (2.3). The two components are called

$$\sum_{i=1}^{n} Y_i^2 = \text{unadjusted sum of squares of the data}$$

$$\frac{1}{n}\left(\sum_{i=1}^{n} Y_i\right)^2 = \text{correction or adjustment for the mean } = n\bar{Y}^2$$

  Thus, we may interpret SS as taking each observation and "centering" (correcting) its magnitude about the sample mean, which itself is calculated from the data. We'll see the importance of this thinking later.

Range: Another simple measure of spread is the **range**

$$\text{range} = \text{difference between highest and lowest values.}$$

This measure gives a quick summary, but is not as informative or useful.

EXAMPLE: For the cherry tree data ($n = 31$), we have

$$\begin{aligned} SS &= \Sigma Y_i^2 - \frac{1}{n}(\Sigma Y_i)^2 \\ &= \left((8.3^2 + \cdots + 20.6^2) - (410.7)^2/31\right) \\ &= 5736.550 - (410.7)^2/31) \\ &= 295.437 \end{aligned}$$

Thus, $s^2 = SS/(n-1) = 295.437/(31-1) = 9.848$, and $s = 3.138$.

The range $= 20.6 - 8.3 = 12.3$.

Note: Here, we have reported all figures to 3 decimal places; however, the **actual calculations** were done by maintaining the figures in the memory of the calculator. When these numbers are used for further calculations in the next section, we use the **actual** not **rounded** values in those calculations. This avoids **error propagation** – if we keep rounding values as we proceed through a series of hand calculations, the effects of rounding each time **add up** and may cause the final result to be markedly different from the true value of the quantity we are calculating! It is thus wise to maintain true values of quantities to be used in subsequent calculations in the memory. Better yet, many statistical calculations are built-in to computer software!

2.7 Statistical inference

We have already defined measures of center and spread for a set of data and mentioned that our goal, besides summarizing and describing data for a sample, is to get an idea of center and spread for the **whole population**.

We already discussed the notion of a **population mean**, denoted by μ. We may define in an analogous fashion **population variance** and **population standard deviation**. These quantities may be thought of as the measures that would be obtained if we could calculate variance and standard deviation based on all the values in the population. Denote the population variance as σ^2 and population standard deviation as σ; this notation is customary.

TERMINOLOGY: A **parameter** is a quantity describing the **population**, e.g.

- μ = population mean
- σ^2 = population variance

Greek letters are commonly used to denote **population** quantities, i.e. **parameters**.

Parameter values are usually **unknown**. Thus, in practice, sample values are used to get an idea of the values of the parameters for the population.

MORE TERMINOLOGY:

- *Estimator*: An **estimator** is a quantity describing the sample that is used as a "guess" for the value of the corresponding population parameter. For example,

$$\bar{Y} \text{ is an estimator for } \mu$$

$$s^2 \text{ is an estimator for } \sigma^2.$$

  - The term **estimator** is usually used to denote the "abstract" quantity
  - The term **estimate** is usually used to denote the actual **numerical** value one might calculate in practice.

- *Statistic*: A quantity derived from the sample observations, e.g. $\bar{Y}$ and s^2 are **statistics**.

INTUITIVELY: The larger the sample, the closer estimates will be to the **true** (but unknown) population parameter values.

- If we are going to use a statistic to estimate a parameter, we would like to have some sense of "how close."
- Just as the data values exhibit variability, so do statistics, because statistics are based on the data!
- Thus, if we wish to use **statistics** to **estimate** population **parameters**, we must consider this variability when we try to determine "how close" an estimate is to the parameter.

- A standard way of quantifying "closeness" of an estimator to the (unknown) parameter is to calculate a measure of **how variable** the estimator is.
- That is, the value of the statistic we end up with, calculated from the **data** we have (with sample size n), is only a **single** value among all the possible values it could have taken on based on all the **other** data sets of size n we may have ended up with!
- Thus, the **variability** of the statistic is with respect to the spread of all possible values it could take on.
- It is customary, then, report not only estimates, but also their variability!

ILLUSTRATION: The **variance** of the **sample mean**. We wish to report the sample mean as an estimate of the population mean μ. We need a measure of how **variable** sample means from samples of size n are across all the possible data sets of size n we might have ended up with.

- Think now of the **population** of **all possible values** of the sample mean $\bar{Y}$ – this population consists of all $\bar{Y}$ values that may be calculated from all possible samples of size n that may be drawn from the population of data values (i.e. the population of Y values).
- This population will **itself** exhibit **spread**, and **itself** may be thought of as having a **mean** and **variance**.
- Statistical theory may be used to show that this mean and variance are as follows

$$\text{Mean of population of } \bar{Y} \text{ values } = \mu \stackrel{\text{def}}{=} \mu_{\bar{Y}} \tag{2.4}$$

$$\text{Variance of population of } \bar{Y} \text{ values } = \frac{\sigma^2}{n} \stackrel{\text{def}}{=} \sigma^2_{\bar{Y}}. \tag{2.5}$$

  The symbols on the far right in each expression are the customary notation for representing these quantities – the subscript "$\bar{Y}$" emphasizes that these are the mean and variance of the population of $\bar{Y}$ values, **not** the population of Y values.
- Equation (2.4) shows that the mean of $\bar{Y}$ values is **the same** as the mean of Y values! This makes intuitive sense – we expect $\bar{Y}$ to be similar to the mean of the population of Y values, μ. (2.4) confirms this, and suggests that $\bar{Y}$ is a sensible estimator for μ!
- Furthermore, the quantity σ^2/n in (2.5) represents how variable $\bar{Y}$ values are. Note that this **depends on the sample size**, n!

IN PRACTICE: The quantity $\sigma^2_{\bar{Y}}$ depends on σ^2, which is **unknown**! Thus, if we want to provide a measure of how variable $\bar{Y}$ values are, the standard approach is to **estimate** $\sigma^2_{\bar{Y}}$ by replacing σ^2, the unknown population (of Y values) variance by its estimate, the sample variance s^2.

- That is, calculate
$$s^2_{\bar{Y}} = \frac{s^2}{n}$$
and use as an estimate of $\sigma^2_{\bar{Y}}$. This notation is standard.

- Of course, this is a **variance**, so does not have the same units of the data. It is thus customary to instead work with the square root of this quantity, which **does** have the same units.

- Define
$$s_{\bar{Y}} = \sqrt{\frac{s^2}{n}} = \frac{s}{\sqrt{n}} \tag{2.6}$$
$s_{\bar{Y}}$ is referred to as the **standard error** (**of the mean**) and is an estimate of the **standard deviation** of all possible $\bar{Y}$ values from samples of size n

- It is important to keep the distinction between s and $s_{\bar{Y}}$ clear:
$$s = \text{ standard deviation of } Y \text{ values}$$
$$s_{\bar{Y}} = \text{ standard deviation of } \bar{Y} \text{ values.}$$

- The **standard error** is an estimate of the **variability** (on the original scale of measurement) associated with using the sample mean $\bar{Y}$ as an estimate of μ.

IMPORTANT: As n increases, $\sigma_{\bar{Y}}$ and $s_{\bar{Y}}$ **decrease** like $\sqrt{n}$. Thus, the **larger** the sample size, the **less variable** (more precise, reliable) the sample mean $\bar{Y}$ will be as an estimator of μ! As intuition suggests, larger sample sizes give "better" estimates!

We will discuss this notion in greater detail later in the course.

COEFFICIENT OF VARIATION: Often, we wish to get an idea of variability on a **relative** basis; that is, we would like to have a **unitless** measure that describes variation in the data (population)

- This is particularly useful if we wish to compare the results of several experiments for which the data are observations on the same variable.
- The problem is that difference **experimental conditions** may lead to different variability. Thus, even though the variable Y may be the same, the variability may be different.
- The **coefficient of variation** is a relative measure that expresses variability as a proportion (percentage) of the mean. For a population with mean μ, standard deviation σ, the definition is

$$CV = \frac{\sigma}{\mu} \text{ as a proportion}$$

$$CV = \frac{\sigma}{\mu} \times 100\% \text{ as a percentage}$$

- **Interpretation**: expresses variability in relation to the average size of the thing being measured.
- As μ and σ are unknown parameters, CV is estimated by replacing them by their estimates from the data, e.g.

$$CV = \frac{s}{\bar{Y}}.$$

- CV is also a useful quantity when attention is focused on a single set of data – it provides an impression of the amount of variation in the data relative to the size of the thing being measured; thus, if CV is large, it is an indication that we will have a difficult time learning about the "signal" (μ) because of the magnitude of the "noise" (σ).

EXAMPLE: For the cherry tree data, we have (taking into account the statement on propagation of error above)

$$s^2_{\bar{Y}} = \frac{9.848}{31} = 0.318$$

$$s_{\bar{Y}} = \sqrt{\frac{9.848}{31}} = 0.564.$$

$$CV = \frac{3.138}{13.25} = 0.237$$

or 23%. A CV value of this magnitude is usually considered pretty "low."

2.8 The linear additive model

It is worth mentioning now a way of thinking about observations that will be a basis for much of our later development. It is instructive to think that data are observations on the **mean** of a population, but, due to **sampling** and imperfect **experimental conditions**, they are not **exactly** like the mean. Rather, they exhibit **variability** that may be thought of as additive "error." That is, all observations would be **exactly alike** (and equal to μ) except for **error** due to the fact we are **sampling** from the population, using imperfect measuring devices, and so on. We think of this error as being added to the mean. Thus, an observation Y_i may be thought of as the sum

$$Y_i = \mu + \epsilon_i,$$

where ϵ_i is an "error" due to sampling, biological variation, etc that makes Y_i **differ** from μ. In this sense

- $\bar{Y}$ is an estimate of μ
- s is an estimate of the variability of "errors"
- Thus, the SS

  $$\sum_{i=1}^{n}(Y_i - \bar{Y})^2$$

  is often referred to as the **error sum of squares**.

2.9 Using SAS to obtain descriptive statistics

The following is a SAS program that creates a temporary (while the program is running) data set containing the cherry tree data, prints out the data, and computes various descriptive statistics for each of the variables (diameter, height, and volume). Following the program is the generated output. The program is heavily documented, containing comments describing how the output was obtained, and thus is fairly self-explanatory.

The program resides in a file. It is always a good idea to name files in a mnemonic fashion. For example, name files containing SAS programming statements with a "`.sas`" extension. A good name for this one might be `cherry.sas`. When a file with a `.sas` extension is run through SAS, two new files are created: a file with a `.log` extension, containing a listing of the program and any error messages generated and a file with a `.lst` extension, containing the program output. Here, for example, these files would be `cherry.log` and `cherry.lst`, respectively.

PROGRAM:

```
*******************************************************;
*                                                     ;
*              ST 511 EXAMPLE 2.1                     ;
*      USING SAS TO COMPUTE DESCRIPTIVE STATISTICS    ;
*                                                     ;
*   THE CHERRY TREE DATA SET FROM COOK & WEISBERG     ;
*   P. 66.  THE DATA ARE THE DIAMETER, HEIGHT, AND    ;
*   VOLUME OF 31 BLACK CHERRY TREES IN THE ALLEGHENY  ;
*   NATIONAL FOREST IN PENNSYLVANIA.  THE DATA WERE   ;
*   ORIGINALLY COLLECTED TO PROVIDE A BASIS FOR       ;
*   DETERMINING AN EASY WAY OF ESTIMATING THE VOLUME  ;
*   OF A TREE (AND EVENTUALLY THE AMOUNT OF TIMBER IN ;
*   A SPECIFIED AREA OF THE FOREST) USING ITS HEIGHT  ;
*   AND DIAMETER.  FOR NOW, WE USE THESE DATA TO      ;
*   ILLUSTRATE THE COMPUTATION OF SAMPLE MEANS,       ;
*   VARIANCES, AND OTHER ATTRIBUTES USING SAS FOR     ;
*   THE VARIABLES D (DIAMETER), H (HEIGHT), AND       ;
*   V (VOLUME).                                       ;
*                                                     ;
*******************************************************;
*;
*******************************************************;
*                                                     ;
*   INVOKE SAS OPTIONS STATEMENT TO FORMAT OUTPUT     ;
*   HERE, WE LIMIT THE WIDTH OF THE OUTPUT TO 80      ;
*   CHARACTERS AND THE LENGTH OF THE PAGE TO 59       ;
*   LINES                                             ;
*                                                     ;
*******************************************************;

OPTIONS LS=80 PS=59;
```

```
*******************************************************;
*                                                     ;
*   CREATE A SAS DATA SET NAMED "TREES".  THE         ;
*   FIRST VARIABLE IS DIAMETER (D), THE SECOND IS     ;
*   HEIGHT (H), AND THE THIRD VOLUME (V).  THERE      ;
*   ARE 31 OBSERVATIONS ON EACH VARIABLE (31 TREES).  ;
*                                                     ;
*******************************************************;
*;
DATA TREES;
  INPUT D H V;
  CARDS;
8.3 70 10.3
8.6 65 10.3
8.8 63 10.2
10.5 72 16.4
10.7 81 18.8
10.8 83 19.7
11.0 66 15.6
11.0 75 18.2
11.1 80 22.6
11.2 75 19.9
11.3 79 24.2
11.4 76 21.0
11.4 76 21.4
11.7 69 21.3
12.0 75 19.1
12.9 74 22.2
12.9 85 33.8
13.3 86 27.4
13.7 71 25.7
13.8 64 24.9
14.0 78 34.5
```

```
14.2 80 31.7
14.5 74 36.3
16.0 72 38.3
16.3 77 42.6
17.3 81 55.4
17.5 82 55.7
17.9 80 58.3
18.0 80 51.5
18.0 80 51.0
20.6 87 77.0
;
*******************************************************;
*                                                     ;
*   PRINT OUT THE DATA SET "TREES" USING THE SAS      ;
*   PROCEDURE "PRINT."  "DATA=TREES" SPECIFIES THE    ;
*   DATA SET TO BE PRINTED OUT.                       ;
*   THE "TITLE" STATEMENT PRINTS A HEADING FOR THE    ;
*   OUTPUT.                                           ;
*                                                     ;
*******************************************************;
*;
PROC PRINT DATA=TREES;
  TITLE 'THE CHERRY TREE DATA SET OF COOK & WEISBERG'; RUN;
*;
*******************************************************;
*                                                     ;
*   USE PROC MEANS TO COMPUTE THE DESCRIPTIVE STAT-   ;
*   ISTICS FOR EACH VARIABLE.  "DATA=TREES" INSTRUCTS ;
*   SAS TO PERFORM THE "MEANS" PROCEDURE ON THE DATA- ;
*   SET "TREES."  THE PROCEDURE AUTOMATICALLY PRINTS  ;
*   CERTAIN STATISTICS -- WHICH ONES DEPENDS ON THE   ;
*   THE VERSION OF THE SAS SYSTEM YOU ARE USING AND   ;
*   THUS WHETHER YOU ARE USING A PC, A UNIX WORK-     ;
*   STATION, OR WHATEVER.  HERE, WE ASK THE PROC      ;
```

```
*   TO PRINT THE SPECIFIC STATISTICS WE WOULD LIKE:   ;
*   THE MEAN, STANDARD DEVIATION, MIN & MAX VALUES,   ;
*   THE STANDARD ERROR FOR THE MEAN, THE SUM OF THE   ;
*   VALUES OF THE VARIABLE, ITS SAMPLE VARIANCE, AND  ;
*   THE COEFFICIENT OF VARIATION EXPRESSED AS A       ;
*   PERCENTAGE.                                       ;
*   THE "VAR" STATEMENT TELLS THE PROC TO PERFORM THE ;
*   COMPUTATIONS FOR EACH OF THE VARIABLES D, H, AND  ;
*   V.  THE 'TITLE' STATEMENT WILL PRINT A HEADING    ;
*   FOR THE OUTPUT.                                   ;
*                                                     ;
*****************************************************;
*;
PROC MEANS DATA=TREES N MEAN STD MIN MAX STDERR SUM VAR CV;
  VAR D H V;
  TITLE 'DESCRIPTIVE STATISTICS FOR THE CHERRY TREE DATA'; RUN;
*;
*****************************************************;
*                                                   ;
*   HERE, WE GET FURTHER STATISTICS FOR THE DATA.    ;
*   WE ASK FOR THE RANGE, THE SUM OF THE             ;
*   SQUARES OF THE OBSERVATIONS (USS), AND THE ERROR  ;
*   SUM OF SQUARES (CSS) FOR EACH VARIABLE.  MORE     ;
*   POSSIBLE STATISTICS ARE LISTED IN THE SAS         ;
*   DOCUMENTATION FOR PROC MEANS.                     ;
*                                                     ;
*****************************************************;
*;
PROC MEANS DATA=TREES RANGE USS CSS;
  VAR D H V;
  TITLE 'MORE STATISTICS FOR THE CHERRY TREE DATA'; RUN;
```

OUTPUT:

%%%

THE CHERRY TREE DATA SET OF COOK & WEISBERG 1

OBS	D	H	V
1	8.3	70	10.3
2	8.6	65	10.3
3	8.8	63	10.2
4	10.5	72	16.4
5	10.7	81	18.8
6	10.8	83	19.7
7	11.0	66	15.6
8	11.0	75	18.2
9	11.1	80	22.6
10	11.2	75	19.9
11	11.3	79	24.2
12	11.4	76	21.0
13	11.4	76	21.4
14	11.7	69	21.3
15	12.0	75	19.1
16	12.9	74	22.2
17	12.9	85	33.8
18	13.3	86	27.4
19	13.7	71	25.7
20	13.8	64	24.9
21	14.0	78	34.5
22	14.2	80	31.7
23	14.5	74	36.3
24	16.0	72	38.3
25	16.3	77	42.6
26	17.3	81	55.4

27	17.5	82	55.7
28	17.9	80	58.3
29	18.0	80	51.5
30	18.0	80	51.0
31	20.6	87	77.0

%%%

DESCRIPTIVE STATISTICS FOR THE CHERRY TREE DATA 2

Variable	N	Mean	Std Dev	Minimum	Maximum
D	31	13.2483871	3.1381386	8.3000000	20.6000000
H	31	76.0000000	6.3718129	63.0000000	87.0000000
V	31	30.1709677	16.4378464	10.2000000	77.0000000

Variable	Std Error	Sum	Variance	CV
D	0.5636263	410.7000000	9.8479140	23.6869484
H	1.1444114	2356.00	40.6000000	8.3839644
V	2.9523244	935.3000000	270.2027957	54.4823308

%%%

MORE STATISTICS FOR THE CHERRY TREE DATA 3

Variable	Range	USS	CSS
D	12.3000000	5736.55	295.4374194
H	24.0000000	180274.00	1218.00
V	66.8000000	36324.99	8106.08

The SAS system has other procedures, or `PROC`s, available for making descriptive summaries of data. Here, we saw the use of one of these, `PROC MEANS`. Others include `PROC UNIVARIATE` and `PROC FREQ`, to name a few. Details may be found in the SAS documentation.

2.10 Summation notation

The Greek letter capital sigma, "Σ," is customarily used to denote the operation of **summing** the quantities to which it is applied. If we think of our data

$$Y_1, Y_2, \ldots, Y_n$$

and functions of them, the symbol

$$\sum_{i=1}^{n}$$

indicates **summation** of the elements that appear after it over the **index** i. This notation is quite convenient for expressing complicated operations on the data concisely. For example,

$$\sum_{i=1}^{n} Y_i = Y_1 + \cdots + Y_n$$

$$\sum_{i=1}^{n} (Y_i - \bar{Y})^2 = (Y_1 - \bar{Y})^2 + \cdots + (Y_n - \bar{Y})^2,$$

where

$$\bar{Y} = \frac{1}{n} \sum_{i=1}^{n} Y_i$$

is the sample mean.

PROPERTIES: Let a be any constant number. Let $f(Y_i)$ and $g(Y_i)$ be any **functions** of Y_i, e.g. $f(Y_i) = Y_i^2$.

(i) $\sum_{i=1}^{n} a = na$; i.e., add a to itself n times

(ii) $\sum_{i=1}^{n} \{a\, f(Y_i)\} = a \sum_{i=1}^{n} f(Y_i)$

(iii) $\sum_{i=1}^{n} \{f(Y_i) + g(Y_i)\} = \sum_{i=1}^{n} f(Y_i) + \sum_{i=1}^{n} g(Y_i)$ and similarly for subtraction.

EXAMPLES: Properties (i)–(iii) may be used to show the following. It is often customary to write Σ for short when the limits of summation and index are clear from the context, and we do so here and in the remainder of the course where it causes no ambiguity.

1. $\Sigma\{3Y_i^3\} = 3\Sigma Y_i^3$

2. $\Sigma\{2(Y_i^2 - 3)^2\} = \Sigma\{2(Y_i^4 - 6Y_i^2 + 9)\} = \Sigma\{2Y_i^4 - 12Y_i^2 + 18\} = 2\Sigma Y_i^4 - 12\Sigma Y_i^2 + 18n.$

3. $\Sigma(Y_i - \bar{Y}) = \Sigma Y_i - n\bar{Y} = \Sigma Y_i - n\left(\frac{1}{n}\Sigma Y_i\right) = 0.$

4. $\Sigma(Y_i - \bar{Y})^2 = \Sigma\{Y_i^2 - 2\bar{Y}Y_i + \bar{Y}^2\} = \Sigma Y_i^2 - 2\bar{Y}\Sigma Y_i + n\bar{Y}^2$. Now

$$2\bar{Y}\Sigma Y_i = 2\bar{Y}n\left(\frac{1}{n}\Sigma Y_i\right) = 2n\bar{Y}^2$$

   so that $\Sigma(Y_i - \bar{Y})^2 = \Sigma Y_i^2 - n\bar{Y}^2$. Furthermore, the second term is

$$n\bar{Y}^2 = n\left(\frac{1}{n}\Sigma Y_i\right)^2 = \frac{1}{n}\left(\Sigma Y_i\right)^2.$$

When you become familiar with this notation, you will probably recognize equalities like those in examples 1–3 without a second thought. Calculations like those in example 4 are a bit harder.

3 Probability Distributions

Complementary Reading: STD, Chapter 3

3.1 Introduction

We have already discussed the notion that statistical methodology is founded on the desire to learn about a **population** based on observation of a **random sample** from the population. The **data** are observations on a **variable** of interest.

We have also discussed the notions of **randomization** and **random samples**. The word "random" implies that **chance** is involved. We are familiar with the idea of **chance** in our everyday lives:

- The **chance** of rain is 40% today.
- I'll **probably** pass ST 511.

The notion of **chance** in the context of statistics arises because of the fact that we deal with **random samples**. The particular sample we end up with is determined by a **chance** mechanism (recall our analogy to drawing from a box). As we have alluded to already, the sample could have turned out another way. Thus, the data we have available to us are as they are **by chance**.

RESULT: Methods to make statements about the population from data are based on taking the random nature of the sample into account; i.e. they involve statements about chance.

PROBABILITY: More formally, the notion of chance is quantified by the notion of **probability**. A discussion of **probability** thus provides a formal framework for describing chance associated with data.

3.2 Probability

In order to talk about chance associated with random samples, it is necessary to talk about probability. It is best to think about things very simply at first, so, as is customary, we do so. This may seem simplistic and even irrelevant; in particular, it is standard to think about very simple situations in which to develop the terminology and properties first, and then extend the ideas behind them to real situations. Thus, we describe the ideas in terms of what is probably the most simple situation where chance is involved, the flip of a coin. The ideas, however, are more generally applicable.

TERMINOLOGY: We illustrate the generic terminology in the context of flipping a coin.

- *(Random) Experiment*: A process for which no outcome may be predicted with **certainty**. For example, if we toss a coin once, the outcome, "heads" (H) or "tails" (T), may not be declared prior to the coin landing. With this definition, we may think of choosing a (random) sample and observing the results as a random experiment – the eventual values of the data we collect may not be predicted with certainty.

- *Sample space*: The **sample space** is the set of all possible (mutually exclusive) outcomes of an experiment. We will use the notation $\mathcal{S}$ in this section to denote the sample space. For example, for the toss of a single coin, the sample space is

  $$\mathcal{S} = \{H, T\}.$$

  By **mutually exclusive**, we mean that the outcomes do not overlap, i.e. they describe **totally distinct** possibilities. For example, the coin comes up either H or T on any toss – it can't do both!

- *Event*: An **event** is a possible result of an experiment. For example, if the experiment consists of two tosses of a coin, the sample space is

  $$\mathcal{S} = \{HH, TH, HT, TT\}.$$

  Each element in $\mathcal{S}$ is a possible result of this experiment. We will use the notation E in this section to denote an event.

  We may think of other events as well, e.g.

  $$E_1 = \{ \text{ see } \textbf{exactly } 1 \ H \text{ in 2 tosses } \} = \{TH, HT\}$$

  $$E_2 = \{ \text{ see } \textbf{at least } 1 \ H \text{ in 2 tosses } \} = \{HH, TH, HT\}.$$

  Thus, events may be combinations of elements in the sample space.

- *Probability function*: P – assigns a number between 0 and 1 to an event. Thus, P quantifies the notion of the **chance** of an event occurring. Properties:
  - For **any** event E, $0 \leq P(E) \leq 1$.
  - If $\mathcal{S}$ is composed of mutually exclusive outcomes denoted by O_i, i.e.

$$\mathcal{S} = \{O_1, O_2, \ldots\},$$

    then

$$P(\mathcal{S}) = \sum_i P(O_i) = 1.$$

    Thus, $\mathcal{S}$ describes **everything** that could possibly happen, since, intuitively, we assign the probability "1" to an event that **must** occur. A probability of "0" (zero) implies that an event **cannot** occur.

We may think of the probability of an event E occurring intuitively as

$$P(E) = \frac{\text{\# of outcomes in } \mathcal{S} \text{ associated with } E}{\text{total \# of possible outcomes in } \mathcal{S}}.$$

For example, in our experiment consisting of two tosses of a coin,

$$P(E_1) = \frac{2}{4} = \frac{1}{2}$$

$$P(E_2) = \frac{3}{4}.$$

3.3 Random variables

The development of statistical methods for analyzing data has its foundations in probability. Because our sample is chosen **at random**, an element of **chance** is introduced, as noted above. Thus, our observations are themselves best viewed as **random**, that is, subject to **chance**.

We thus term the variable of interest a **random variable** to emphasize this principle. **Data** represent observations on the **random variable**. These may be (for **quantitative** random variables) discrete or continuous.

NOTATION: Y = random variable (often abbreviated r.v.). $Y_1, Y_2, \ldots, Y_n$ are observations on Y.

RESULT: Events of interest may be formulated in terms of random variables. In our coin toss experiment, let

$$Y = \#\ H \text{ in 2 coin tosses.}$$

Then we may represent our events as

$$E_1 = \{Y = 1\}, \quad E_2 = \{Y \geq 1\}.$$

Furthermore, the probabilities of the events may also be written in terms of Y, e.g.

$$P(E_1) = P(Y = 1), \quad P(E_2) = P(Y \geq 1).$$

If we did our coin toss experiment n times (each time consists of 2 tosses of the coin), and recorded the value of Y each time, we would have data $Y_1, \ldots, Y_n$, where Y_i is the number of heads seen the ith time we did the experiment.

3.4 Probability distribution of a random variable

To understand this concept, it is easiest to first consider the case of a **discrete** r.v. Y. Thus, Y takes on values that we can think about separately. Our r.v. Y corresponding to the coin tossing experiment is a discrete random variable; it may only take on the values 0, 1, or 2.

PROBABILITY DISTRIBUTION FUNCTION: Let y denote a possible value of Y. The function

$$f(y) = P(Y = y)$$

is the **probability distribution function** for Y. $f(y)$ is the probability we associate with an observation on Y taking the value y.

If we think in terms of **data**, that is, observations on the r.v. Y, then we may think of the **population** of all possible data values. From this perspective, $f(y)$ may be thought of as the **relative frequency** (in the population) with which the value y occurs in the population.

Recall that a **histogram** for a sample summarizes the relative frequencies with which the data takes on values, and these relative frequencies are represented by **area**. This gives a pictorial view of how the values are **distributed**.

If we think of probabilities as the relative frequencies with which Y would take on values, then it seems natural to think of representing probabilities in a similar way.

PROBABILITY HISTOGRAM: A graph representing **probabilities** by area.

To understand probability and the notion of probability histogram, we will consider a particular probability distribution function, the **binomial** distribution. Doing this accomplishes two things:

- It will introduce us to the basic idea of a probability distribution function and histogram
- It will introduce us to a particular function that is useful for describing data that are the result of counting up the number of times out of some total number that a response of interest was observed.
- For example, in an experiment to test the efficacy of insecticide, suppose 30 insects are placed in a chamber, the insecticide is applied, and we count the number (out of 30) dead after 1 hour. The random variable of interest here is Y = number of dead insects out of 30 in 1 hour, taking on values 0, 1, ..., 30. The relative frequencies of these values are described by the binomial distribution.

BINOMIAL DISTRIBUTION: Consider more generally the following experiment:

(i) k **unrelated** trials are performed

(ii) Each trial has two possible outcomes, e.g. for a coin toss, H or T; for the insects, each insect is a trial, with outcomes dead or alive.

(iii) For each trial, the probability of the outcome we are interested in (e.g., dead in the insect example) is equal to some value p, $0 \leq p \leq 1$.

For the k trials, we are interested in the number of trials resulting in the outcome of interest. Let S = "success" denote this outcome. Then the **random variable** of interest is

$$Y = \# \text{ of } S \text{ in } k \text{ trials.}$$

Y may thus take on the values $0, 1, \ldots, k$.

To fix ideas, consider an experiment consisting of k coin tosses, and suppose we are interested in the number of H observed in the k tosses. For more realistic situations, the principles are the same, so we use coin tossing as an easy, "all-purpose" illustration. Then $Y = \#\ H$ in k tosses, and $S = H$. Furthermore, if our coin is "fair" (not weighted to come up H or T more frequently), then

$$p = \frac{1}{2} = P(H).$$

It turns out that the form of the probability distribution function $f(y)$ of Y may be derived mathematically. The expression is

$$f(y) = P(Y = y) = \left(\begin{array}{c} k \\ y \end{array} \right) p^y (1-p)^{k-y}, \quad y = 0, 1, \ldots, k,$$

where

$$\left(\begin{array}{c} k \\ y \end{array} \right) = \frac{k!}{y!(k-y)!}.$$

The notation $x!$ is "x **factorial**" $= x(x-1)(x-2)\cdots(2)(1)$, that is, the product of x with all positive whole numbers smaller than x. By convention, $0! = 1$.

For our purposes, we will not dwell on the form of $f(y)$, except to note its intuitive interpretation. If we do k trials, and y are "successes," then $k - y$ of them are not "successes." There are a number of ways that this can happen in k trials – the expression $\left(\begin{array}{c} k \\ y \end{array} \right)$ turns out to quantify this number of ways.

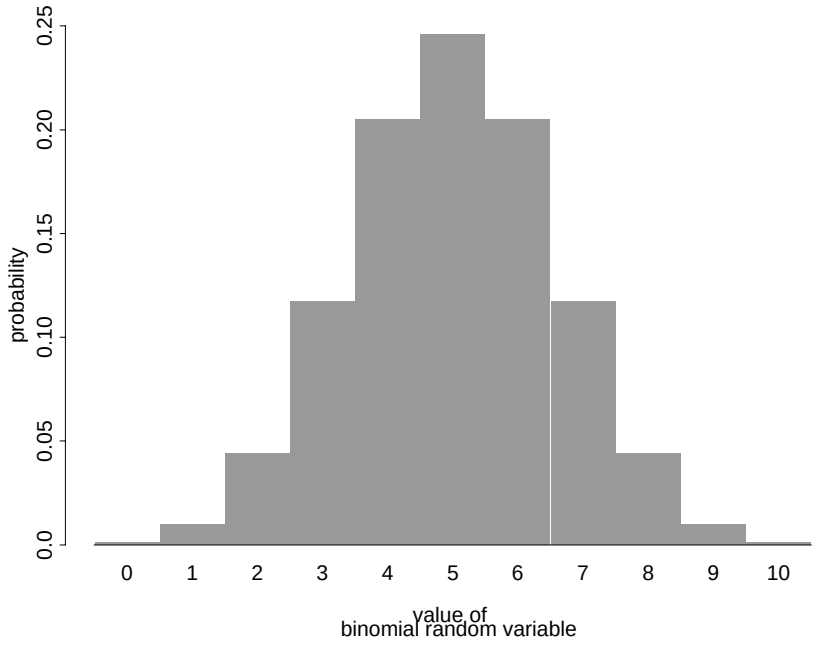
We are not so interested in this particular $f(y)$ here – our main purpose is to establish the idea that probability distribution functions **do exist** and may be calculated.

EXAMPLE: If $p = 1/2$ (fair coin) and $k = 2$, then, applying the formula, the probability of getting (exactly) one head ($y = 1$) is

$$f(1) = P(Y = 1) = \left(\begin{array}{c} 2 \\ 1 \end{array} \right) \left(\frac{1}{2}\right)^1 \left(1 - \frac{1}{2}\right)^1 = \frac{2 \times 1}{(1)(1)} \left(\frac{1}{2}\right) \left(\frac{1}{2}\right) = 2\left(\frac{2}{4}\right) = \frac{1}{2}$$

as we calculated previously by intuition.

Before we look at more calculations, consider the **probability histogram** associated with the r.v. Y with $k = 10$, $p = 1/2$. Suppose, as discussed earlier, we **graph** $f(y)$ by drawing a block of width 1 (like a class interval) and height $f(y)$ (which is like a relative frequency with class interval width = 1) centered at $y = 0, 1, \ldots, 10$. It turns out that this looks like



The area of each block is $P(Y = y)$, and the **total area** = 1, as it should; everything that can possibly occur is represented by the graph. Note that, as intuition suggests, the "most likely" value is 5 (half H, half T), with the remaining values in either direction progressively less likely (getting 10 out of 10 heads would be pretty unusual, but possible). Thus, the probability histogram gives a nice picture of the likelihood of seeing different numbers of H in 10 coin tosses.

INTERPRETATION: Now think of the relationship with **data**. Suppose we were to repeat the experiment n times – that is, **each** experiment consists of 10 tosses. For the ith experiment, Y_i denotes the number of H seen. We thus observe data $Y_1, \ldots, Y_n$. Suppose we plot the histogram for the n observations in our data set.

- Intuitively, if n were a very large number, we would expect that the histogram for the data ought to resemble the probability histogram.
- That is, the relative frequencies with which things happen in our sample of n experiments should resemble those we would expect if we could do experiments forever (thus completely determining the population of all possible numbers of H).

Thus, the probability histogram may be thought of as the analog for the population to a histogram for a sample. As we take larger and larger samples, we will get better and better representation of the population by the sample.

We may thus think of $f(y)$, the probability distribution function, as telling us about the population from which data arise. We would hope that the relative frequencies we see in the sample would be close to those of $f(y)$ for large samples.

MEAN AND VARIANCE: Thinking about $f(y)$ as a **model** for the population suggests that we consider other population quantities. In particular, for data from an experiment like this, what would the **population mean** and μ **variance** σ^2 be like? It turns out that mathematical derivations may be used to show that

$$\mu = kp, \quad \sigma^2 = kp(1-p).$$

We would thus expect, if we did the experiment n times, that the **sample mean** $\bar{Y}$ and **sample variance** s^2 would be "close" to these values (and be **estimators** for them).

For the example, with $k = 10$, $p = 1/2$, we have $\mu = 5$, which is indeed the "balancing point" of the probability histogram. We also have $\sigma^2 = 2.25$.

CALCULATING BINOMIAL PROBABILITIES AND HISTOGRAMS: We illustrate the calculations in a simpler case, where $k = 4$ and $p = 1/2$. In this case,

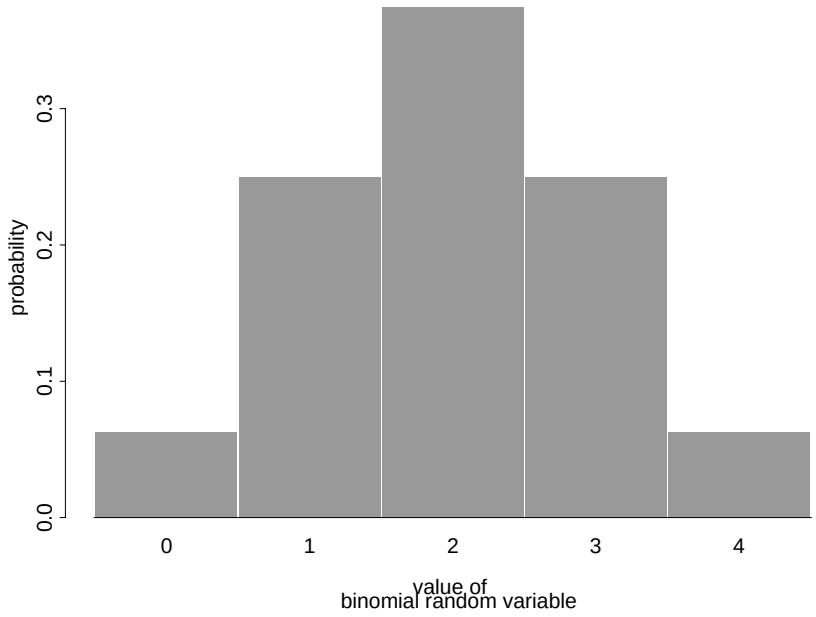
$$f(0) = \left(\begin{array}{c} 4 \\ 0 \end{array} \right) \left(\frac{1}{2}\right)^0 \left(1 - \frac{1}{2}\right)^4 = \frac{4!}{0!\,4!} \left(\frac{1}{2}\right)^4 = (1)0.0625 = 0.0625$$

$$f(1) = \left(\begin{array}{c} 4 \\ 1 \end{array} \right) \left(\frac{1}{2}\right)^1 \left(1 - \frac{1}{2}\right)^3 = \frac{4!}{1!\,3!} \left(\frac{1}{2}\right)^4 = (4)0.0625 = 0.2500.$$

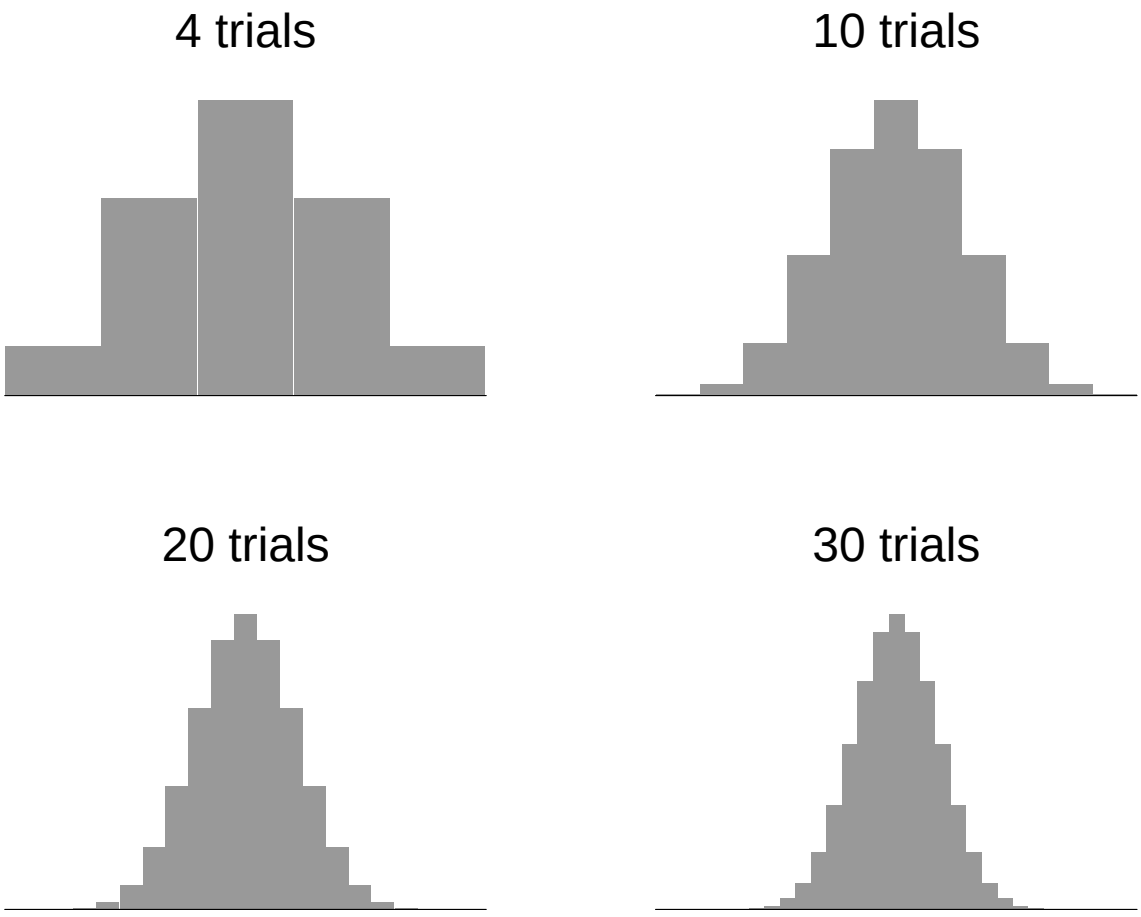
The full set of values is

y	$f(y)$
0	0.0625
1	0.2500
2	0.3750
3	0.2500
4	0.0625
	1.0000

To construct the probability histogram, draw a block of width 1 over each possible value taken on by the binomial r.v. Because width=1, height=$f(y)$.



FINAL NOTE: In the following figure, we plot probability histograms with $p = 1/2$ for a number of different k values. Note that by the time $k = 30$, the **shape** of the histogram is getting **smoothed out** and resembles in fact a **bell-shaped** curve! This is no accident; we will discuss this phenomenon more shortly!



CONTINUOUS RANDOM VARIABLES: Many of the variables of interest in scientific investigations are **continuous**; thus, they can take on any value. For example, suppose we obtain a sample of n pigs and weigh them. Thus, the **random variable** of interest is

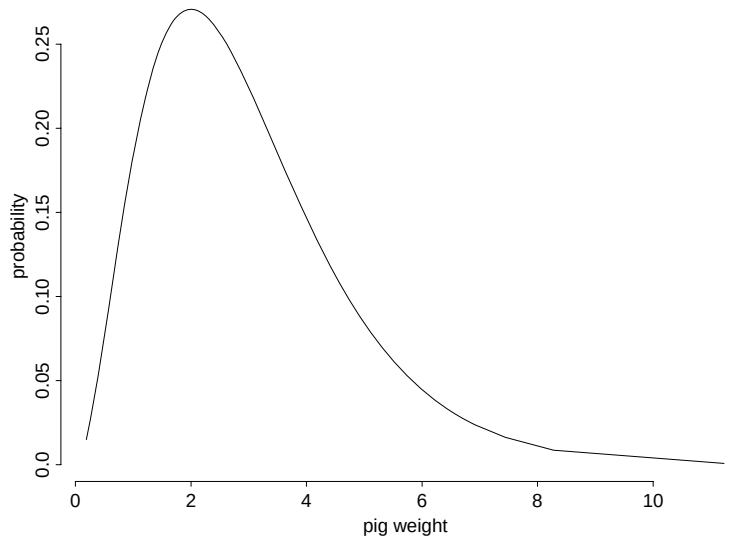
$$Y = \text{ weight of a pig}$$

and the data are $Y_1, \ldots, Y_n$, the observed weights for our n pigs. Y is a random variable because the pigs were drawn at random from the population of all pigs. Furthermore, all pigs do not weigh exactly the same; they exhibit **random variation** due to biological and other factors.

GOAL: Find a function like $f(y)$ for a discrete r.v. that describes the probability of observing a pig weighing y units.

This function, f, would thus serve as a **model** describing the **population** of pig weights – how they are distributed and how they vary.

Because pig weights may be **anything**, however, we would not expect f to be as straightforward as it was for a **discrete** r.v. We would expect that, rather than be "jagged," with blocks corresponding to individual values y, that f would resemble a **smooth curve**, capturing all possible values, something like



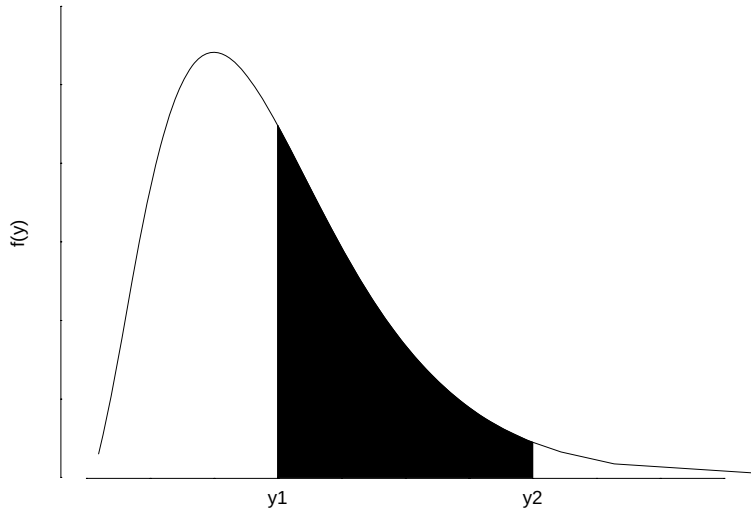
PROBABILITY DENSITY: A function $f(y)$ that describes the distribution of values taken on by a continuous random variable is called a **probability density function**. If we could determine f for a particular r.v. of interest, its graph would have the same interpretation as a **probability histogram** did for a discrete r.v. If we were to take a sample of size n and construct the sample histogram, we would expect it to have the roughly same shape as f – as n increases, we'd expect it to look more and more like f.

Thus, a probability density function describes the **population** of Y values, where Y is a continuous r.v.

TECHNICAL NOTE: For a **continuous** r.v., we do not think about the probability that Y is **exactly equal** to some value y, as we did for a discrete r.v. The reason is that, due to the limitations imposed by the **precision** of measuring devices (e.g. a scale for taking weight measurements, an assay for measuring concentrations of contaminants in test samples, etc), we can never observe Y **exactly**. Thus, we instead speak of the probability that Y falls into an **interval**; although we may not be able to see the exact value of Y, the precision is good enough to allow us to state that the value falls in an interval.

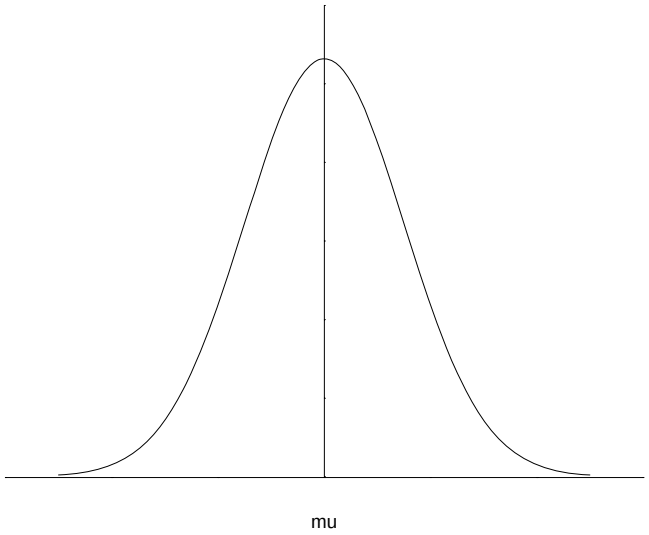
In terms of the graph of $f(y)$, the probability that Y takes on a value in the interval between some y_1 and y_2, where $y_1 < y_2$, is

$$P(y_1 < Y < y_2) = \text{ area of shaded region.}$$



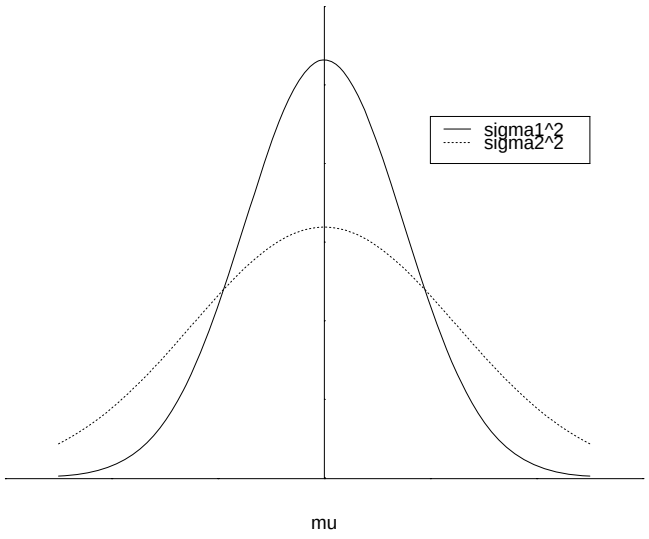
DETERMINING A SUITABLE f: It turns out that for many types of continuous measurement data (so continuous r.v.s Y) , there is a certain function that seems to provide a good description of the probabilities associated with the measurements. The probability distribution associated with this probability density function is called the **normal** (or **Gaussian**) distribution.

In particular, if the population has mean μ and variance σ^2, then the graph of the **normal density function** looks like



Note that the graph is **symmetric**.

The following graph illustrates how variance is a notion of spread. Superimposed on the same axes are two normal density functions with the **same mean** μ but with different variances $\sigma_1^2 < \sigma_2^2$.



RESULT: Often, for this type of data, we postulate that the **population** of all possible measurements is well-described by a **normal** distribution with mean μ and variance σ^2. Eventually, we will see that, because in real life μ and σ^2 are unknown, one objective is to **estimate** them. The procedures for estimating them are based on the normal assumption. For now, we will assume μ and σ^2 are **known** in order to explore the properties of the normal distribution. We'll see how this is used later.

NOTATION: Write $\mathcal{N}(\mu, \sigma^2)$ to denote the normal distribution with mean μ, variance σ^2. Write

$$Y \sim \mathcal{N}(\mu, \sigma^2)$$

to say that the r.v. Y has this distribution.

NORMAL APPROXIMATION TO THE BINOMIAL DISTRIBUTION: Recall that, as the number of trials k grows, the probability histogram for a binomial random variable with $p = 1/2$ has shape suspiciously like the normal density function! The former is jagged, while the normal density is a **smooth** curve; however, as k gets larger, the jagged edges get smoother. You can imagine that if k was very large, the jaggedness would be almost imperceptible.

Mathematically, it may be shown that, as k gets large, for p near $1/2$, the probability distribution function and histogram for the binomial distribution looks more and more like the normal probability density function and its graph!

Thus, if one is dealing with a **discrete** binomial random variable (e.g. the insecticide example), and the number of trials is relatively large, the smooth, continuous normal distribution is often used to approximate the binomial. This allows one to take advantage of the statistical methods we will discuss that are based on the assumption that the normal density is a good description of the population. We thus confine much of our attention in this course to methods for data that may be viewed (at least approximately) as arising from a population that may be described by a normal distribution.

3.5 The standard normal distribution

The probability distribution function f for a normal distribution has a very complicated form. Thus, it is not possible to evaluate easily probabilities for a normal r.v. Y the way we could for a binomial. Luckily, however, these probabilities may be calculated on a computer. Some computer packages have these probabilities "built in;" they are also widely available in tables in the special case when $\mu = 0$ and $\sigma^2 = 1$. For example, see Table A.4 of STD. It turns out that these tables are all that is needed to evaluate probabilities for **any** μ and σ^2, as we'll see.

It is instructive to learn how to use such tables, because the necessary operations give one a better understanding of how probabilities are represented by **area**.

We will learn how to evaluate normal probabilities when μ and σ^2 are **known** first; we will see later how we may use this knowledge to develop statistical methods for estimating them when they are not known.

SITUATION: Suppose $Y \sim \mathcal{N}(\mu, \sigma^2)$.

We wish to calculate probabilities such as

$$P(Y \leq y), \quad P(Y \geq y), \quad P(y_1 \leq Y \leq y_2), \tag{3.1}$$

that is, probabilities for **intervals** associated with values of Y.

TECHNICAL NOTE: When dealing with probabilities for **any** continuous r.v., we do not make the distinction between **strict inequalities** like "$<$" and "$>$" and inequalities like "$\leq$" and "$\geq$". This is for the reason discussed above – the limitations on our ability to see the **exact** values make it impossible to distinguish between Y being **exactly** equal to a value y and Y being equal to a value **extremely close** to y. Thus, the probabilities in (3.1) could equally well be written

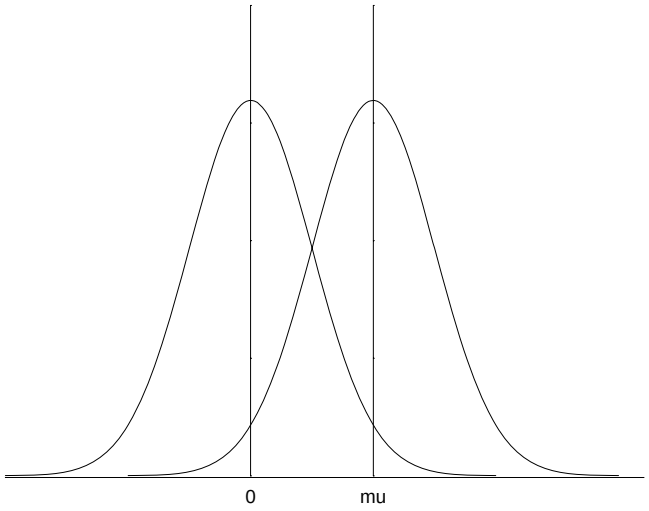
$$P(Y < y), \quad P(Y > y), \quad P(y_1 < Y < y_2). \tag{3.2}$$

The method for determining probabilities of either type (3.1) and (3.2) is thus **identical.**

THE STANDARD NORMAL DISTRIBUTION: Consider the event

$$(y_1 \leq Y \leq y_2).$$

If we **know** μ and were to **subtract** it from **every possible** Y value in the normal population, the effect would be to shift the entire graph downward by μ, like this



Thus, if we think of the r.v. $Y - \mu$, it is clear that this is **also** a normal r.v., but now with **mean** 0 and the same variance σ^2.

If we furthermore **know** σ^2, and thus the standard deviation σ (same units as Y), suppose we divide every possible value of Y by the value σ. This will yield all possible values of the r.v.

$$\frac{Y}{\sigma}.$$

Note that this r.v. is "unitless"; e.g. if Y is measured in grams, so is σ, so the units "cancel." Rather, this r.v. has "units" of **standard deviation**; that is, if it takes value 1, this says that the value of Y is 1 standard deviation to the right of the mean. In particular, the standard deviation of Y/σ is 1.

COMBINING: Define

$$Z = \frac{Y - \mu}{\sigma}.$$

Then Z will have mean zero and standard deviation 1. (It has units of standard deviation of the original r.v. Y). It will also be normally distributed, just like Y, as all we have done is shift the mean and scale by the standard deviation.

Hence, we call Z a **standard normal** r.v., and we write $Z \sim \mathcal{N}(0, 1)$.

Applying this to each component of our event of interest, we see that

$$(y_1 \leq Y \leq y_2) \Leftrightarrow \left(\frac{y_1 - \mu}{\sigma} \leq Z \leq \frac{y_2 - \mu}{\sigma} \right).$$

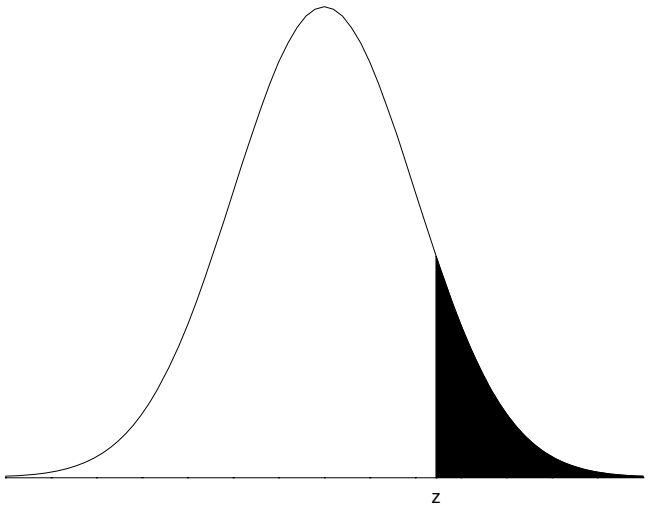
Similarly, we have

$$(Y \geq y) \Leftrightarrow \left(Z \geq \frac{y - \mu}{\sigma} \right).$$

RESULT: If we want to find probabilities about events concerning Y, and we **know** μ and σ, all we need is a table of probabilities for a standard normal r.v. Z.

3.6 Finding probabilities for a standard normal r.v.

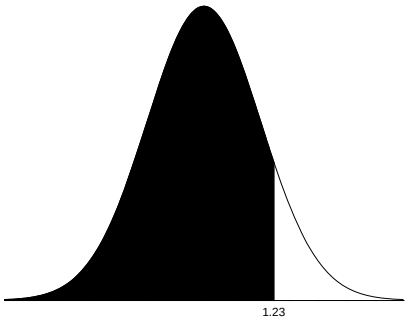
First, then, we learn how to find probabilities for $Z \sim \mathcal{N}(0,1)$. We will illustrate using Table A.4 of STD. The body of this table has values $P(Z \geq z)$ (or, equivalently, $P(Z > z)$) for values of $z \geq 0$. This may be used to find all types of probabilities. That is, the probabilities in the table correspond to areas of the form

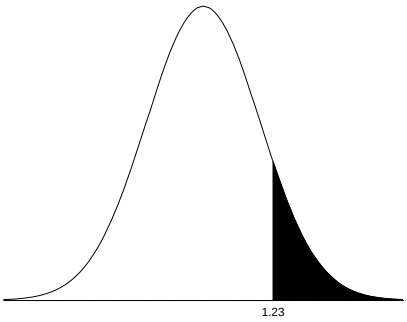


For example, it is immediate from the table that $P(Z \geq 1.23) = 0.1093$.

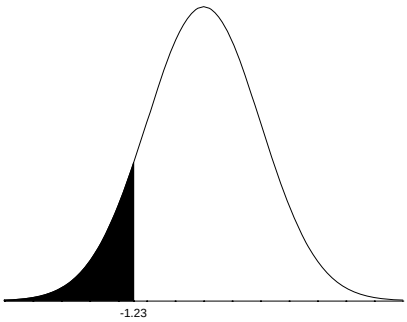
CAUTION: Many statistics texts contain a table for standard normal probabilities. Different tables may present this information differently, however. Our discussion here applies to tables like Table A.4 of STD. The tables usually have some description of how the probabilities are represented; be sure you understand how it works if you use a different table.

EXAMPLES: Recall that the total area is equal to 1; we use this fact to find probabilities for all types of events about Z.

(1) $P(Z \leq 1.23) =$ 

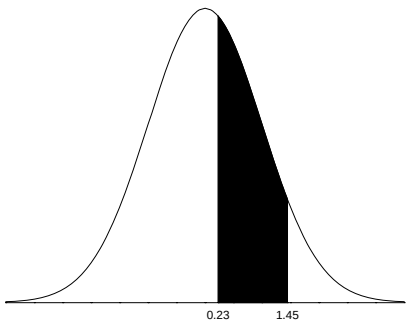
$= 1 -$ 

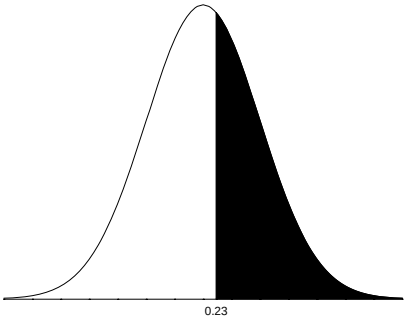
$= 1 - 0.1093 = 0.8907.$

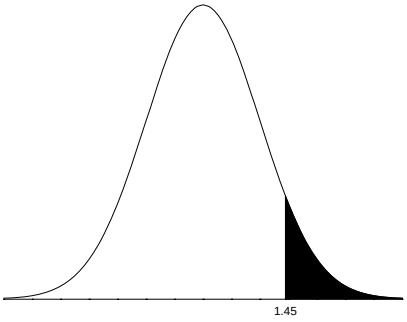
(2) $P(Z \leq -1.23) =$ 

The normal probability density is **symmetric**, thus,

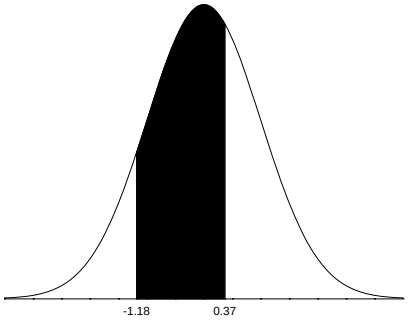
$$P(Z \leq -1.23) = P(Z \geq 1.23) = 0.1093.$$

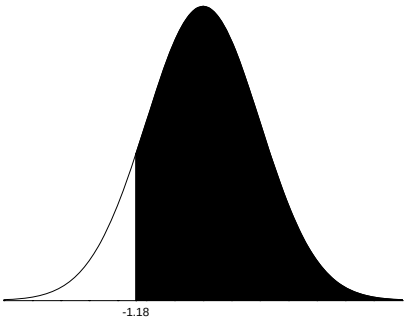
(3) $P(0.23 \leq Z \leq 1.45) =$ 

$=$ 

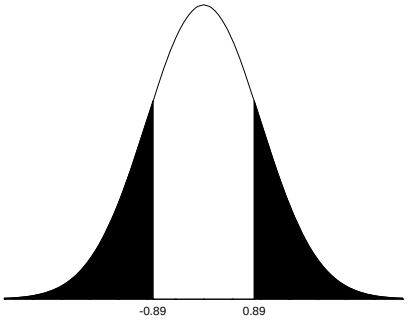
$-$ 

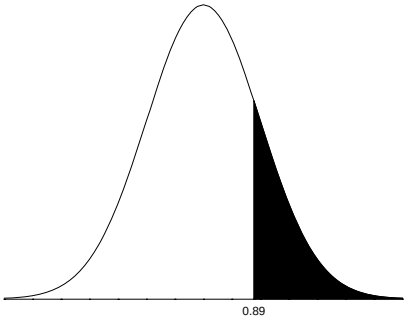
$= 0.4090 - 0.0735 = 0.3355$. By **symmetry**, $P(-1.45 \leq Z \leq -0.23) = 0.3355$ as well.

(4) $P(-1.18 \leq Z \leq 0.37) =$ 

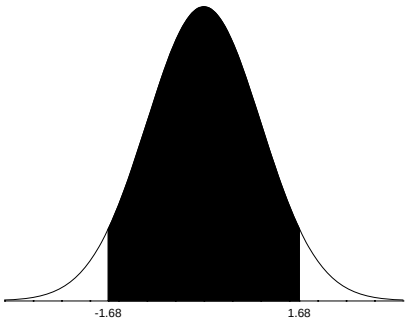
$=$ 

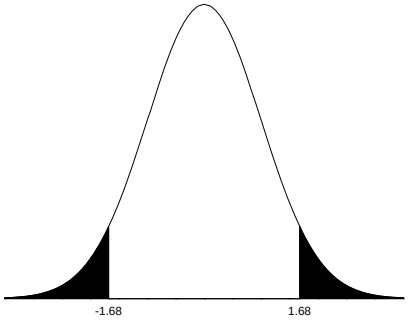
$-0.3557 = (1 - 0.1190) - 0.3557 = 0.5253.$

(5) $P(|Z| \geq 0.89) =$ 

$= 2\times$ 

$= 2(0.1867) = 0.3734.$

(6) $P(|Z| \leq 1.68) =$ 

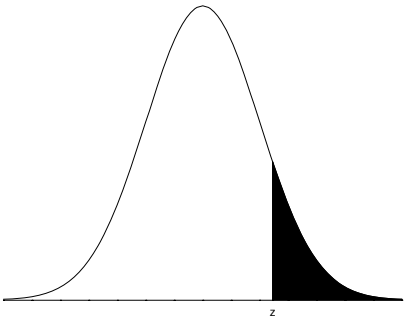
$= 1-$ 

$= 1 - 2(0.0465) = 0.9070.$

Sometimes, we are given the probability and wish to find the value z associated with it:

(7)$P(Z \geq z) = 0.05$ Clearly, this z must be **positive**, as it has very little probability to its right.

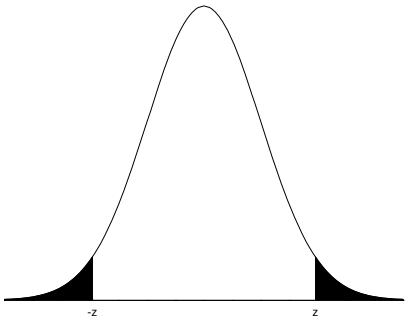
Pictorially, we want to find z so that the shaded region has area 0.05:



Looking through the body of the table, we find $P(Z \geq 1.64) = 0.0495$ and $P(Z \geq 1.65) = 0.0505$. Thus, one could **interpolate** between these two values and obtain $z = 1.645$ (their average).

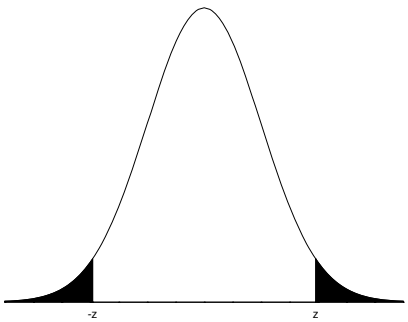
In general, pick the closest value in the body of the table. There is no wrong or right way to interpolate values – usually, picking the closest one suffices. However one does it, one should try to be consistent.

(8) $P(|Z| \geq z) = 0.05$. Again a picture helps:



We want z such that each shaded area equals 0.025, so that the total equals 0.05. Thus, z is positive and satisfies $P(Z \geq z) = 0.025$. From the table, $z = 1.96$.

(9) $P(|Z| \leq z) = 0.975$. Consider



We want z so that the unshaded region has area 0.975 and the shaded region has total area 0.025. Thus, **each** shaded portion should have area 0.0125, and hence z should satisfy $P(Z \geq z) = 0.0125$. From the table, $z = 2.24$.

3.7 Finding probabilities for any normal r.v.

IDEA: Transform probability statements about $Y \sim \mathcal{N}(\mu, \sigma^2)$ in to statements about Z, then use the methods above.

EXAMPLES: Suppose $\mu = 8$, $\sigma^2 = 4$, i.e. $Y \sim \mathcal{N}(8, 4)$. Thus, $\sigma = \sqrt{4} = 2$.

(1) $P(Y \geq 9.5) = P(Z \geq 0.75)$;

$$Z = \frac{Y - 8}{2},$$

so we perform the same operation on 9.5:

$$\frac{9.5 - 8}{2} = 0.75.$$

From the table, the desired probability is 0.2266.

(2) $P(6 \leq Y \leq 10) = P(-1 \leq Z \leq 1) = 1 - 2(0.1587) = 0.6826$, because

$$\frac{6 - 8}{2} = -1, \quad \frac{10 - 8}{2} = 1.$$

(3) $P(|Y| \geq 8.6) = P(Y \leq -8.6) + P(Y \geq 8.6)$. We have

$$\frac{-8.6 - 8}{2} = \frac{-16.6}{2} = -8.3 \text{ and } \frac{8.6 - 8}{2} = \frac{0.6}{2} = 0.3.$$

Thus, this is equivalent to

$$P(Z \leq -8.3) + P(Z \geq 0.3) = 0 + 0.3821 = 0.3821.$$

Note that the first probability does not even appear in the table; for a standard normal, -8.3 is very far out in the left-hand "tail" of the graph, so the probability is virtually 0.

3.8 Statistical inference

For the remainder of this part of the course, we assume that we are interested in a r.v. Y that may be viewed (exactly or approximately) as following a $\mathcal{N}(\mu, \sigma^2)$ distribution. Suppose that we have observed data $Y_1, \ldots, Y_n$.

In real situations, of course, we may be willing to assume that our data arise from some normal distribution, but we do not know that values of μ or σ^2. As we have discussed, one goal is to use $Y_1, \ldots, Y_n$ to estimate μ and σ^2. We use **statistics** like $\bar{Y}$ and s^2 as **estimators** for these unknown parameters.

Because $\bar{Y}$ and s^2 are based on observations on a random variable, they **themselves** are random variables. Thus, we may think about the populations of all possible values they may take on (from all possible samples of size n). It is natural to thus think of the **probability distributions** associated with these populations.

STATISTICAL METHODS: The fundamental principle behind the methods we will discuss is as follows. We base what we are willing to say about μ and σ^2 on how **likely** the values of the statistics $\bar{Y}$ and s^2 we saw from our data would be if some values μ_0 and σ_0^2 were the **true** values of these parameters.

To assess how **likely**, we need to understand the probabilities with which $\bar{Y}$ and s^2 take on values. That is, we need the probability distributions associated with $\bar{Y}$ and s^2! We will discuss this more formally later in the course.

PROBABILITY DISTRIBUTION OF $\bar{Y}$: It turns out that if Y is normal, then the distribution of all possible values of $\bar{Y}$ is **also** normal! Thus, if we want to make statements about "how likely" it is that $\bar{Y}$ would take on certain values, we may calculate these using the normal distribution.

We have

$$\bar{Y} \sim \mathcal{N}(\mu_{\bar{Y}}, \sigma^2_{\bar{Y}}), \quad \mu_{\bar{Y}} = \mu, \quad \sigma_{\bar{Y}} = \frac{\sigma}{\sqrt{n}}.$$

We may use these facts to **transform** events about $\bar{Y}$ into events about a standard normal r.v. Z as before.

EXAMPLE: Suppose $\bar{Y}$ is based on a sample of size $n = 25$ observations on a random variable $Y \sim \mathcal{N}(6, 9)$. Thus $\mu = 6$, $\sigma = 3$. To find

$$P(\bar{Y} \geq 6.9) = P(Z \geq 1.5) = 0.0668,$$

we have $\mu_{\bar{Y}} = 6$, $\sigma_{\bar{Y}} = 3/\sqrt{25} = 3/5 = 0.6$, so that

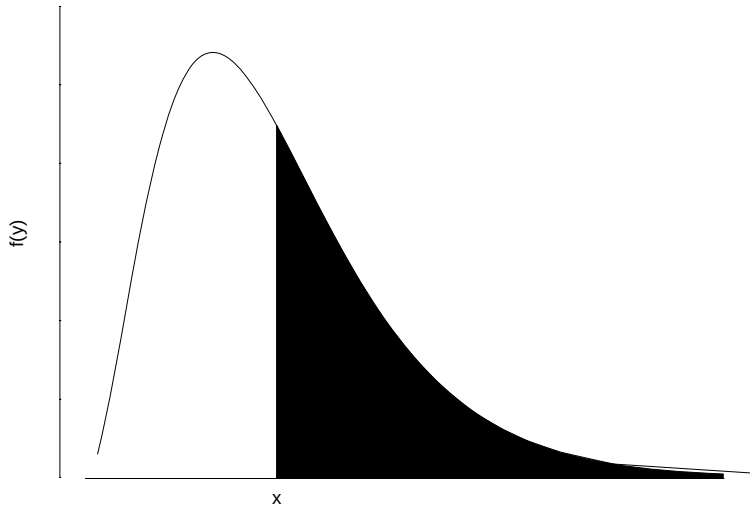
$$Z = \frac{\bar{Y} - 6}{0.6} \text{ and } \frac{6.9 - 6}{0.6} = 1.5.$$

3.9 The χ^2 distribution

PROBABILITY DISTRIBUTION OF s^2: When the data are observations on a r.v. $Y \sim \mathcal{N}(\mu, \sigma^2)$, then it may be shown mathematically that the values taken on by

$$\frac{(n-1)s^2}{\sigma^2}$$

are well-represented by another distribution different from the normal. This quantity is still continuous, of course, so the distribution has a probability density function. This density, when plotted, looks like



The distribution with this density, and thus the distribution that plays a role in describing probabilities associated with s^2 values, is called the **Chi-square distribution with** $(n-1)$ **degrees of freedom**. (We discuss the notion of degrees of freedom in a moment.) This is also often written χ^2 (the Greek letter).

Tables of the probabilities associated with the values of a r.v. with the χ^2 distribution are readily available. Table A.5 of STD is typical of these tables. The body of the table contains values x such that, if χ^2_v denotes a χ^2 random variable with v degrees of freedom, then

$$P(\chi^2_v \geq x) = \text{ probability in the top row}$$

for degrees of freedom v in the left-hand column. (This is different from the normal table, whose body contains probabilities). That is, the values x satisfying a picture like that above are in the body of the table; the area of the shaded region is across the top.

Note that the χ^2 distribution is **not symmetric**. Again, as always, the total area is equal to 1.

EXAMPLES OF USING THE TABLE:

(1) Suppose $v = 13$. Find x so that $P(\chi^2_v \geq x) = 0.050$. From the $v = 13$ row of the table, and the 0.050 column, we have immediately that $x = 22.4$.

(2) Suppose $v = 24$. What is $P(\chi^2_v \geq 29.3)$? In the row for $v =$ degrees of freedom 24, we find

$$P(\chi^2_v \geq 28.2) = 0.250 \text{ and } P(\chi^2_v \geq 33.2) = 0.100.$$

Thus, from the table, the best we can tell is that the probability of interest is somewhere between these two probabilities, i.e.

$$0.100 \leq P(\chi^2_v \geq 29.3) \leq 0.250.$$

Thus, note that the table does not allow us to "zero in" on the probabilities like the normal table did. Often, as we'll see, however, we may only be interested in whether a probability is "small" or "large," so this may suffice. Alternatively, many computer languages and packages have the ability to calculate such probabilities exactly.

(3) Suppose $v = 14$. Find x so that $P(\chi^2_v \leq x) = 0.25$. Because the total area is 1, we have

$$0.25 = P(\chi^2_v \leq x) = 1 - P(\chi^2_v \geq x),$$

so that the same x satisfies $P(\chi^2_v \geq x) = 1 - 0.25 = 0.75$. From the table, we read off the value $x = 10.2$.

(4) Suppose $v = 18$. Find $P(\chi^2_v \leq 8.94)$. By the same reasoning,

$$P(\chi^2_v \leq 8.94) = 1 - P(\chi^2_v \geq 8.94).$$

Now $P(\chi^2_v \geq 8.94)$ is between the probabilities for $x = 8.23$, 0.975, and $x = 9.39$, 0.95, listed in the table. Thus, the probability of interest is between $1 - 0.975 = 0.025$ and $1 - 0.95 = 0.05$, i.e.

$$0.025 \leq P(\chi^2_v \leq 8.94) \leq 0.050.$$

3.10 Student's t distribution

Recall that one of our objectives will be to develop statistical methods to **estimate** μ, the population mean, using the obvious estimator, $\bar{Y}$. We would like to be able to make statements about "how likely" it is that $\bar{Y}$ would take on certain values. We saw above that this involves appealing to the normal distribution, as $\bar{Y} \sim \mathcal{N}(\mu, \sigma^2_{\bar{Y}})$.

A problem with this in real life is, of course, that σ^2, and hence $\sigma^2_{\bar{Y}}$, is **not known**, but must **itself** be **estimated**. Thus, even if we are not interested in σ^2 in its own right, if we are interested in μ, we still need σ^2 to make the inferences we desire!

An obvious approach would be to **replace** $\sigma_{\bar{Y}}$ in our **standard normal** statistic

$$\frac{\bar{Y} - \mu}{\sigma_{\bar{Y}}}$$

by the obvious estimator, $s_{\bar{Y}}$, and consider instead the statistic

$$\frac{\bar{Y} - \mu}{s_{\bar{Y}}}. \tag{3.3}$$

(As we will develop formally later, the value for μ would be a "candidate" value for which we are trying to assess the likelihood of seeing a value of $\bar{Y}$ like the one we saw.)

RESULT: It turns out that when we replace $\sigma_{\bar{Y}}$ by the **estimate** $s_{\bar{Y}}$, the resulting statistic (3.3) no longer follows a **standard normal distribution**. The presence of the **estimate** $s_{\bar{Y}}$ in the denominator serves to **add variation**. Rather, the statistic has a **different** distribution, which is centered at 0 and has the **same**, **symmetric**, **bell shape** as the normal, but whose probabilities in the extreme "tails" are larger than those of the normal distribution.

STUDENT'S t DISTRIBUTION: The probability distribution describing the probabilities associated with the values taken on by the quantity (3.3) is called the **(Student's) t distribution with $(n-1)$ degrees of freedom** for a sample of size n.

The name "Student" sometimes attached to this distribution arose because the statistician who derived the mathematical form of the distribution was too shy to publish his result under his real name, but, rather, published under the pseudonym "Student!"

NOTATION: We will write t_v to denote a random variable with the t distribution with v degrees of freedom.

Tables of the probabilities associated with the values of a r.v. with the t distribution are readily available. Table A.3 of STD is typical of these tables. The body of the table contains values t that the r.v. might take on rather than probabilities, just like the χ^2 table.

EXAMPLES OF USING THE TABLE: These calculations are made using Table A.3 of STD. To avoid confusion, we will do **all** calculations referring to the **bottom row** of Table A.3, which contains probabilities of the form

$$P(t_\nu \geq t) \text{ for } t \geq 0.$$

The numbers in the columns above each probability in this bottom row are the t values satisfying this statement.

If you use a table different from this one, be sure you understand how it works.

(1) Suppose $\nu = 10$. Find the value t such that $P(t_\nu \geq t) = 0.05$. From the bottom row of the table, we find 0.05, and go up this column to the row for $\nu = 10$, which yields $t = 1.812$.

(2) Suppose $\nu = 10$. Find the value of t such that $P(|t_\nu| \geq t) = 0.05$. Just like for the normal distribution, the t distribution is **symmetric**, which we may use to our advantage. Thus, this problem is similar to example (8) on p. 53. Note that the event of interest is satisfied if **either**

$$(t_\nu \leq -t) \text{ or } (t_\nu > t),$$

Because of the **symmetry**, the probabilities of each of these events is **the same**. Thus, just as in the normal example (8), t is such that each probability is 1/2 of the total 0.05, or 0.025. Thus, we are interested in

$$P(t_\nu \geq t) = 0.025,$$

which, from the bottom row, reading up to $\nu = 10$, is 2.228.

(3) Similarly, $P(t_6 \geq t) = 0.05$ gives $t = 1.943$ while $P(|t_6| \geq t) = 0.05$ gives $P(t_6 \geq t) = 0.025$, so that $t = 2.447$.

(4) $P(|t_9| \leq t) = 0.90 = 1 - P(|t_9| \geq t)$. Thus, rearranging, t must satisfy $P(|t_9| \geq t) = 0.10$. As above, this means t must satisfy $P(t_9 \geq t) = 0.05$. From the bottom row, reading up, we find $t = 1.833$.

(5) Similarly, $P(t_9 \leq t) = 0.85 = 1 - P(t_9 \geq t)$, so that $P(t_9 \geq t) = 0.15$, and $t = 1.100$.

In some cases, the values desired won't be in the table. In this case, we report a **range**, just as for the χ^2 distribution.

(6) Find $P(t_4 \geq 3.258)$. From the table, for the row $v = 4$, we find $P(t_4 \geq 2.776) = 0.025$ and $P(t_4 \geq 3.747) = 0.01$. Thus, the best we can say is that the desired probability is between these 2 probabilities, or

$$0.01 \leq P(t_4 \geq 3.258) \leq 0.025.$$

Often, as we'll see, just knowing a range is sufficient for our purposes.

(7) $P(|t_{11}| \leq 1.863) = 1 - P(|t_{11}| \geq 1.863)$. From the row for $v = 11$, we find that

$$P(t_{11} \geq 1.796) = 0.05 \text{ and } P(t_{11} \geq 2.201) = 0.025;$$

these numbers bracket 1.863. By **symmetry**, we thus know that

$$P(|t_{11}| \geq 1.796) = 0.10 \text{ and } P(|t_{11}| \geq 2.201) = 0.05.$$

Thus, $0.05 \leq P(|t_{11}| \geq 1.863) \leq 0.10$. Going back to our original problem, then, using $1 - 0.10 = 0.90$ and $1 - 0.05 = 0.95$,

$$0.90 \leq P(|t_{11}| \leq 1.863) \leq 0.95.$$

(8) $P(-1.4 \leq t_3 \leq 3.1) = P(t_3 \geq -1.4) - P(t_3 \geq 3.1)$. Using **symmetry**, we have

$$P(t_3 \geq -1.4) = 1 - P(t_3 \leq -1.4) = 1 - P(t_3 \geq 1.4).$$

Thus, the desired probability may be rewritten as

$$1 - P(t_3 \geq 1.4) - P(t_3 \geq 3.1).$$

We thus want to find the **largest** and **smallest** values this expression could possibly take on in order to find the entire range of possible probabilities. From the table, using the methods above, we find

$$0.10 \leq P(t_3 \geq 1.4) \leq 0.15 \text{ and } 0.025 \leq P(t_3 \geq 3.1) \leq 0.05.$$

Thus, our expression of interest is **largest** when the two probabilities are both at their smallest, or $1 - 0.10 - 0.025 = 0.875$. Similarly, the expression is **smallest** when the two probabilities are at their largest, or $1 - 0.15 - 0.05 = 0.800$. We conclude

$$0.800 \leq P(-1.4 \leq t_3 \leq 3.1) \leq 0.875.$$

3.11 Degrees of freedom

For both the statistics

$$\frac{(n-1)s^2}{\sigma^2} \text{ and } \frac{\bar{Y}-\mu}{s_{\bar{Y}}},$$

which follow the χ^2 and t distributions, respectively, the notion of **degrees of freedom** has arisen. The probabilities associated with each of these statistics depend on the **sample size** n through the degrees of freedom value $(n-1)$. What is the meaning of this?

Note that both statistics depend on s^2, and recall that

$$s^2 = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \bar{Y})^2.$$

Recall also that it is **always** true that

$$\sum_{i=1}^{n}(Y_i - \bar{Y}) = 0.$$

Thus, if we know that values of $(n-1)$ of the observations in our sample, we may always compute the last value, because the deviations about $\bar{Y}$ of all n of them **must** sum to zero. Thus, s^2 may be thought of as being based on $(n-1)$ "independent" deviations – the final deviation can be gotten from the other $(n-1)$.

The term **degrees of freedom** thus has to do with the fact that there are $(n-1)$ "free" or "independent" quantities upon which the r.v.s above are based. Thus, we would expect their **distributions** to depend on this notion as well.

4 Estimation, Inference, and Sampling Distributions

Complementary Reading: STD, Chapter 4

4.1 Introduction

So far, we have made allusions to the fundamental ideas underlying statistical methods, but we really haven't done this formally. In this chapter, we will begin to discuss some of the formal notions underlying **statistical inference**.

RECALL: The basic ideas

- We would like to learn about a **population(s)** based on observations of a **random sample(s)**.
- The element of **chance** is introduced, hence the name **random variable**, because the values in the sample arise by chance, because of sampling from the population and biological variation.
- The result is that statements we make about the population will involve statements about **probabilities**.

ESTIMATION: A particular way to say something about the population based on a sample is to assign **likely** values based on the sample to the **unknown parameters** describing the population. We have already discussed this notion of **estimation**, e.g.,

$$\bar{Y} \text{ is an } \textbf{estimator} \text{ for } \mu, \quad s^2 \text{ is an } \textbf{estimator} \text{ for } \sigma^2.$$

Now we get a little more precise about **estimation**.

Note that these estimators are not the **only** possibilities. For example, recall that

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(Y_i - \bar{Y})^2;$$

an alternative estimator for σ^2 would be

$$s^2_* = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \bar{Y})^2,$$

where we have replaced the divisor $(n-1)$ by n.

OBVIOUS QUESTION: If we can identify **competing** estimators for population parameters of interest, how can we decide among them?

Recall that estimators such as $\bar{Y}$ and s^2 may be thought of as having their own underlying **populations** (that may be described by probability distributions). That is, for example, we may think of the population of all possible $\bar{Y}$ values corresponding to all of the possible samples of size n we might have ended up with. For this population, we know that

$$\text{Mean of population of } \bar{Y} = \mu_{\bar{Y}} = \mu. \tag{4.1}$$

RESULT: The property (4.1) says that the mean of the probability distribution of $\bar{Y}$ values is **equal** to the **parameter** we try to estimate by $\bar{Y}$. This seems intuitively like a desirable quality.

UNBIASEDNESS: In fact, it is, and has a formal name! An estimator is said to be **unbiased** if the mean of the probability distribution is equal to the population parameter to be estimated by the estimator. Thus, $\bar{Y}$ is an **unbiased estimator** of μ.

Clearly, if we have two competing estimators, then, we would prefer the one that is **unbiased**! Thus, **unbiasedness** may be used as a criterion for choosing among competing estimators.

EXAMPLE: Recall our competing estimators for σ^2 above,

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(Y_i - \bar{Y})^2 \text{ and } s_*^2 = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \bar{Y})^2.$$

- It may be shown mathematically that the estimator s_*^2, which is the **average** of the n squared deviations (division by n), has population with mean equal to

  $$\left(\frac{n-1}{n}\right)\sigma^2$$

  rather than σ! That is, s_*^2 is **not** an unbiased estimator for σ^2! Rather, it is **biased** downward,, because it estimates $(n-1)/n$ times σ^2 rather then σ^2 itself! Note that if n were small, the leading term $(n-1)/n$ could be much smaller than 1, so that s_*^2 could be much smaller than it should be to be a credible estimate of σ^2.

- Thus, note that our usual estimate, s^2, represents an attempt to obtain an **unbiased estimator**! By dividing by $(n-1)$ rather that n, we counteract the bias!

- Consequently, s^2 is the accepted estimator rather than s_*^2, as it is **unbiased**!

MINIMUM VARIANCE: What if we can identify two competing estimators that are **both** unbiased? On what grounds might we prefer one over the other?

Unbiasedness is clearly a desirable property, but we can also think of other desirable properties an estimator might have. For example, as we have discussed previously, we would also like our estimator to be as "close" to the true values as possible – that is, in terms of the **probability distribution** of the estimator, we would like it to have **small variance**! This would mean that the possible values that the estimator could take on (across all possible samples we might have ended up with) exhibit only **small** variation – they don't vary a lot depending on which sample we ended up with!

Thus, if we have two unbiased estimators, choose the one with **smaller** variance.

Ideally, then, we'd like to use an estimator that is unbiased and has **the smallest** variance among all such candidates. Such an estimator is given the name **minimum variance unbiased**.

- It turns out that, for normally distributed data Y, the estimators $\bar{Y}$ (for μ) and s^2 (for σ^2) have this desirable property!

RESULT: We use $\bar{Y}$ and s^2 as estimators for population mean and variance because they have the desirable properties of **unbiasedness** and **minimum variance**.

4.2 Confidence interval for μ

An estimator is a "likely" value. Because of chance, it is of course too much to expect that $\bar{Y}$ and s^2 would be **exactly** equal to μ and σ^2, respectively, for any given data set of size n. Although they may not be "exactly" equal to the value they are estimating, because they are "good" estimators in the above sense, they are likely to be "close."

- Instead of reporting only the single value of the estimator, we report an **intervals** (based on the estimator) and state that it is likely that the true value of the parameter is in the interval.
- "likely" means that **probability** is involved.

Here, we discuss the notion of such an interval, known as a **confidence interval**, for the particular situation where we wish to estimate μ by the sample mean $\bar{Y}$.

Suppose for now that $Y \sim \mathcal{N}(\mu, \sigma^2)$, where μ and σ are **unknown**. We have available a random sample of observations $Y_1, \ldots, Y_n$ and wish to estimate μ. Of course, our estimator is $\bar{Y}$.

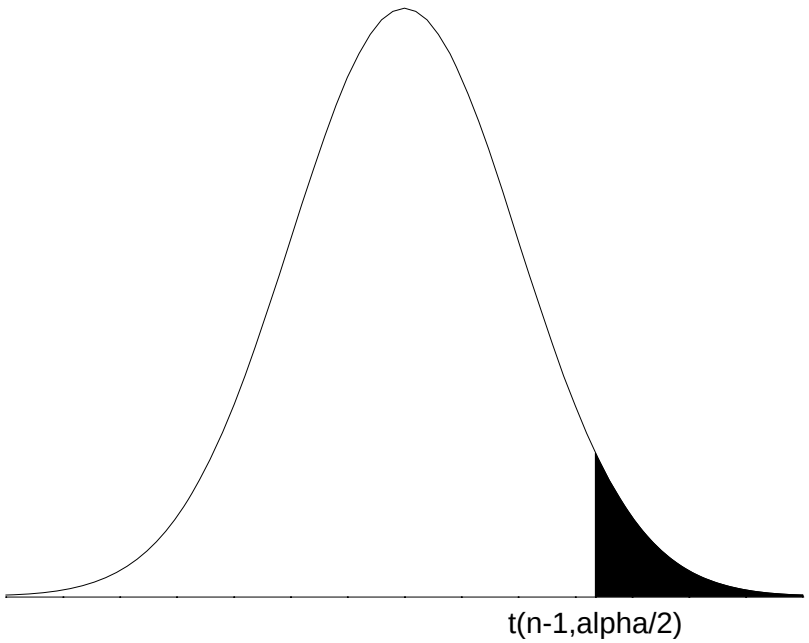
Recall from section 3.10 of the course that, if we wish to make **probability** statements involving $\bar{Y}$, this is complicated by the fact that σ^2 is **unknown**. Thus, even if we are not interested in σ^2 in its own right, we can't ignore it and must estimate it anyway. In particular, recall that we will be interested in the statistic

$$\frac{\bar{Y} - \mu}{s_{\bar{Y}}}$$

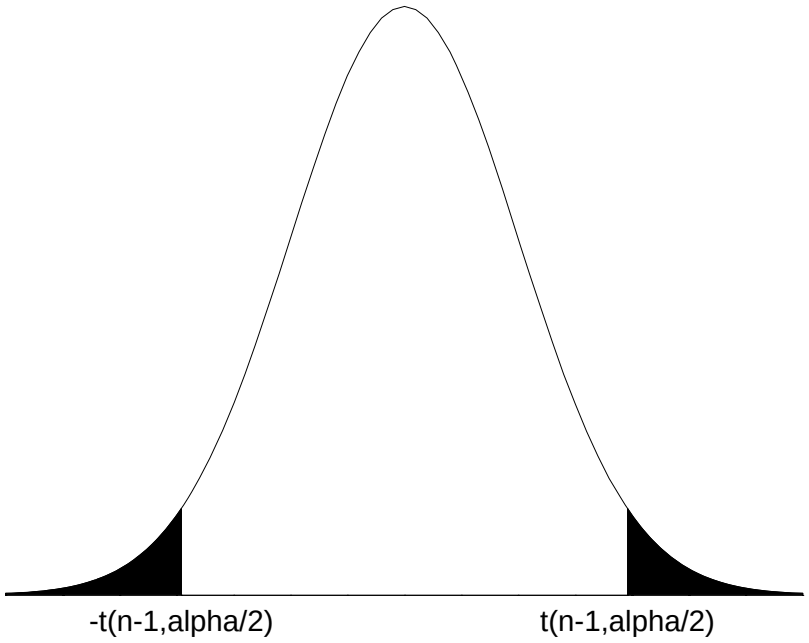
if we wish to make probability statements about $\bar{Y}$ without knowledge of σ^2.

What kind of probability statement do we wish to make? The **chance** or **randomness** that we must contend with arises because $\bar{Y}$ is based on a random sample. The value μ we wish to estimate is a **fixed** (but unknown) quantity. Thus, our probability statements intuitively should have something to do with the uncertainty of trying to get an understanding of the **fixed** value of μ using the **variable** estimator $\bar{Y}$.

With this in mind, consider the following probability statement. Let $t_{n-1,\alpha/2}$ be the point such that the shaded region under the Student's t density with $(n-1)$ degrees of freedom has area $\alpha/2$ (so that the unshaded region has area $(1-\alpha)/2$.



Thus, by symmetry, we know that $t_{n-1,\alpha/2}$ is such that each shaded region has area $\alpha/2$, with $(1-\alpha)$ in the middle:



or

$$P(-t_{n-1,\alpha/2} \leq \frac{\bar{Y}-\mu}{s_{\bar{Y}}} \leq t_{n-1,\alpha/2}) = 1-\alpha. \tag{4.2}$$

If we rewrite (4.2) by algebra (applying the same operation to each inequality), we obtain

$$P(\bar{Y} - t_{n-1,\alpha/2}s_{\bar{Y}} \leq \mu \leq \bar{Y} + t_{n-1,\alpha/2}s_{\bar{Y}}) = (1-\alpha) \tag{4.3}$$

It is important to interpret (4.3) correctly. Even though the μ appears in the middle, this is **not** a probability statement about μ – remember, μ is a **fixed constant**! Rather, the probability in (4.3) has to do with the quantities on either side of the inequalities – these quantities depend on $\bar{Y}$ and $s_{\bar{Y}}$, and thus are subject to chance. Thus, the probability $(1-\alpha)$ refers to the probability that these lower and upper **random** values in the inequalities lie on either side of the **constant** value μ.

DEFINITION: The interval

$$(\bar{Y} - t_{n-1,\alpha/2} s_{\bar{Y}} \ , \ \bar{Y} + t_{n-1,\alpha/2} s_{\bar{Y}})$$

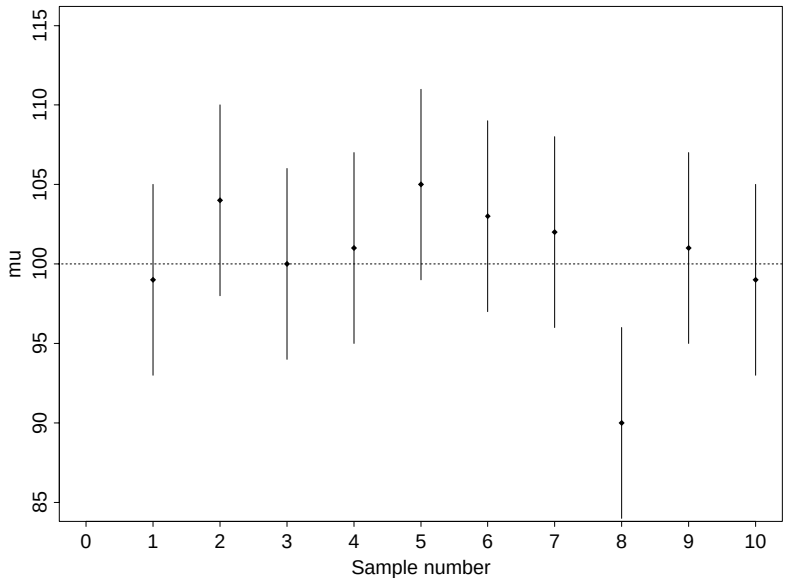
is called a $100(1-\alpha)\%$ **confidence interval** for μ. For example, if $\alpha = 0.05$, then $(1-\alpha) = 0.95$, and the interval would be called a **95% confidence interval**. In general, the value $(1-\alpha)$ is called the **confidence coefficient**.

We will often abbreviate this as "CI."

INTERPRETATION: As noted above, the probability associated with the confidence interval has to do with the endpoints, **not** with probabilities about the value μ! We might be tempted to say "the probability that μ falls in the interval is 0.95," thinking that the endpoints are fixed and μ may or may not be between them. But the endpoints are what varies here – the value of μ is fixed.

- The interval (its endpoints) is based on $\bar{Y}$, the estimate, and $s_{\bar{Y}}$, which has to do with the variability of $\bar{Y}$. Thus, the interval is **random** – **different** samples would end up giving **different** interval (because different samples would lead to different $\bar{Y}$ and $s_{\bar{Y}}$ values).
- μ is **fixed** but **unknown**. The **interval** is thus the only thing that may change, and this depends on the sample.
- Thus, the probability has to do with the likelihood of the **sample** providing endpoints that would be on either side of μ!

This is most easily seen by illustration. μ, the population mean, is unknown, but is a fixed quantity. Our sample is only one of many zillions of possible samples we might have obtained (by chance). Each possible sample would lead a different interval – some would "cover" μ, others would not. For example, if the true value of $\mu = 100$, here are 10 possible intervals from 10 possible samples. The solid "dots" on each interval correspond to the estimate of μ, $\bar{Y}$, from each sample. All of the intervals "cover" μ except that corresponding to "sample 8," which does not. If these are 95% confidence intervals, if we could examine all the possible intervals (here we've only looked at 10), we would find that 95% of these would "cover" μ, and 5% would not.



The confidence coefficient $(1-\alpha)$ is thus the probability that such a sample will produce an interval covering μ. Thus, it measures the "confidence" we have that our sampling procedure and interval construction technique will yield an interval containing μ.

A confidence interval either covers μ or it **doesn't**, so, once it has been constructed, it doesn't make sense to talk about the probability that it covers μ. We instead talk about our confidence in the statement "the interval covers μ." The confidence coefficient quantifies this.

BOTTOM LINE: A confidence interval is thus a statement about the **process** of selecting samples and the methods used. That is, it has to do with the quality of the experiment!

How might we interpret confidence intervals in real situations, then?

- The "width" of the confidence interval must be evaluated on scientific grounds.
- For example, suppose we are investigating a new cancer treatment, and we would be interested by **any** minute improvement in response. If our experiment leads to a 95% confidence interval for mean response that is quite "wide," encompassing a range of values that are both good and bad, then the interpretation is that, for the purposes of being as precise in understanding the treatment as we'd like to be, our experimental procedure is inadequate!
- The confidence interval is thus a statement about the adequacy of our experimental procedure!

EXAMPLES:

(1) The following data are from Box, Hunter, and Hunter (1978, *Statistics for Experimenters*) and represent measurements of dissolved oxygen concentration (mg/L) in 6 test samples:

```
2.62    2.65    2.79    2.83    2.91    3.57
```

Assume that dissolved oxygen concentrations may be thought of as being well-represented by a normal distribution (continuous measurements), $\mathcal{N}(\mu, \sigma^2)$. We would like to obtain a confidence interval for μ, population mean dissolved oxygen concentration for such samples. We have $n = 6$,

$$\bar{Y} = \frac{17.37}{6} = 2.895 \text{ mg/L},$$

$$s^2 = \frac{1}{5}(50.893 - 50.286) = \frac{0.60675}{5} = 0.121.$$

Thus, we obtain $s = 0.348$ (mg/L) and $s_{\bar{Y}} = 0.348/\sqrt{6} = 0.142$ (mg/L).

Thus, our estimate for μ is $\bar{Y} = 2.895$ mg/L and a 95% confidence interval for μ based on this experiment is

$$(2.896 \pm (2.571)(0.142)) = (2.529\,,\, 3.261) \text{ mg/L},$$

where we have used from the table for the t distribution with 5 degrees of freedom $t_{5,0.025} = 2.571$. A 90% confidence interval, using $t_{5,0.05} = 2.015$, would be $(2.609, 3.181)$ mg/L.

(2) The following data are from Finney (1978, *Statistical Method in Biological Assay*, p. 179) and are from an experiment to investigate the influence of different doses of vitamin A on weight gain over a 3 week period. For 5 rats receiving 2.5 units of vitamin A, the following weight increases (mg) were observed:

```
35    49    51    43    27
```

Again, because these are continuous measurement data, the normal assumption seems reasonable; thus, we assume that the population of weight increases for all possible rats receiving 2.5 units of vitamin A may be approximated by a $\mathcal{N}(\mu, \sigma^2)$ probability distribution. We obtain, with $n = 5$,

$$\bar{Y} = \frac{205}{5} = 41 \text{ mg},$$

$$s^2 = \frac{1}{4}(8805 - \frac{(205)^2}{5}) = \frac{400}{4} = 100.$$

Thus, we obtain $s = 10$ (mg) and $s_{\bar{Y}} = 10/\sqrt{5} = 4.472$ (mg).

Thus, our estimate for μ is $\bar{Y} = 41$ mg and a 95% confidence interval for μ based on this experiment is

$$(41 \pm (2.776)(4.472)) = (28.585\,,\ 53.415) \text{ mg},$$

where we have used from the table for the t distribution with 4 degrees of freedom $t_{4,0.025} = 2.776$. A 90% confidence interval, using $t_{4,0.05} = 2.132$, would be $(31.466, 50.534)$ mg/L.

It may seem counterintuitive that a 90% confidence interval would be **narrower** than a 95% interval, as in these examples; however, this is to be expected! For the 90% interval, we may be a little more stringent with the width, as we aren't requiring such a high a level of confidence – if we wish greater confidence of 95%, however, we must widen the interval in order to be "more confident" that it will cover the fixed value μ.

4.3 Formal statistical inference

The confidence interval procedure for the mean of a population we described in the last section is a form of what is known as **statistical inference**. We have used this term previously, but now we are a bit more formal about exactly what we mean.

Because we base what we say about the population on sample evidence, any conclusions we wish to make based on the sample have an element of uncertainty about them. We wish to assess and summarize formally this uncertainty. i.e. assess the quality of our sample evidence. Thus, we temper any inferences we make about the population with a indication of our assessment of the sample evidence. This indication is in terms of probability, as we are expressing how likely things in the sample are.

A confidence interval is just one way we might make such an assessment. The interval is a statement about the quality of the sample evidence regarding the population mean.

We will study formal statistical inference in more detail throughout the rest of the course. In particular, we will discuss ideas related to confidence intervals that address questions like "Is the population mean equal to 10?" In answering such questions, as you might expect, we need to express the "level of confidence" we have in our data for answering such a question.

SAMPLING DISTRIBUTION: Although we are already familiar with this concept, we now give it a formal name. We have seen that if the population may be viewed as **approximately normal**, and we are interested in making inference about the population mean μ based on our sample value, $\bar{Y}$, we use the fact that

$$\frac{\bar{Y} - \mu}{s_{\bar{Y}}} \sim t_{n-1},$$

where by this notation we mean that the statistic on the left hand side follows a t distribution with $(n-1)$ degrees of freedom. In particular, we use this result to make probability statements (confidence intervals) about the sample evidence.

This kind of result, where we appeal to the **probability distribution** of a **statistic**, is central to statistical methodology. Because the probability distribution of the statistic, in this case that of

$$\frac{\bar{Y} - \mu}{s_{\bar{Y}}},$$

is that of all possible values of the statistic across all possible **samples**, such a distribution is referred to as a **sampling distribution**. The Student's t is an example of a sampling distribution; so is the χ^2 distribution. We will in fact encounter other sampling distributions as well.

Formal statistical inference is thus based on comparing **statistics** to appropriate **sampling distributions** to make formal probability statements that characterize the quality of the sample evidence for answering the question at hand.

4.4 Confidence interval for a difference of population means

Rarely in real life is our interest confined to a single population. Rather, we are usually interested in conducting experiments to **compare** populations. For example, an experiment may be set up because we which to gather evidence about the difference among yields (on the average) obtained with several different rates of fertilizer application. More precisely, we would like to make a statement about whether **treatments** give **truly different** responses.

The simplest case is where we wish to compare just 2 such treatments. We will develop this case here first, then discuss extension to more than 2 treatments later in the course.

EXPERIMENTAL PROCEDURE: Take 2 random samples of **experimental units** (plants, subjects, plots, rats, etc). Each unit in the first sample receives treatment 1, each in the other receives treatment 2. We would like to make a statement about the difference in responses to the treatments based on this setup.

EXAMPLE: Suppose we wish to compare the effects of two concentrations of a toxic agent on weight loss in rats. We select a random sample of rats from the population of interest and then randomly assign each rat to receive either concentration 1 or concentration 2. The variable of interest is

$$Y = \text{ weight loss for a rat.}$$

Until the rats receive the treatments, we may assume them all to have arisen from a **common** population for which Y has some mean μ and variance σ^2. Because these are continuous measurement data, it is reasonable to assume further that $Y \sim \mathcal{N}(\mu, \sigma^2)$.

Once the treatments are administered, however, the 2 samples become **different**. One convenient way to view this is to think of **two** populations:

Population 1	population of Y's under treatment 1
Population 2	population of Y's under treatment 2

That is, populations 1 and 2 may be thought of as the original population with all possible rats treated with treatment 1 and 2, respectively. We may thus regard our samples as being randomly selected from these two populations. Because of the nature of the data, it is further reasonable to think about two **random variables**, Y_1 and Y_2, one corresponding to each population, and to think of them as being **normally distributed**:

Population 1	$Y_1 \sim \mathcal{N}(\mu_1, \sigma_1^2)$
Population 2	$Y_2 \sim \mathcal{N}(\mu_2, \sigma_2^2)$

NOTATION: Because we are now thinking of 2 populations, we must adjust our notation accordingly, so we may talk about two different random variables **and** observations on each of them.

Write Y_{ij} to denote the observation from the jth unit receiving the ith treatment; that is, the jth value observed on random variable Y_i.

With this definition, we may thus view our data as follows:

$$\begin{aligned} Y_{11}, Y_{12}, \ldots, Y_{1n_1} \quad & n_1 = \# \text{ units in sample from population 1} \\ Y_{21}, Y_{22}, \ldots, Y_{2n_2} \quad & n_2 = \# \text{ units in sample from population 2} \end{aligned}$$

In this framework, we may now cast our question as follows:

$$\text{Difference in response for the 2 treatments?} \;\Rightarrow\; \text{Is } \mu_1 \text{ different from } \mu_2?$$

More formally, then, we may look at the **difference** $(\mu_1 - \mu_2)$:

$$\begin{aligned} (\mu_1 - \mu_2) = 0 \quad & \text{no difference} \\ (\mu_1 - \mu_2) \neq 0 \quad & \textbf{real} \text{ difference} \end{aligned}$$

OBVIOUS STRATEGY: Base investigation of this **population mean difference** on the data from the 2 samples by **estimating** the difference.

FACTS: It may be shown mathematically that, if both of the random variables (one for each treatment) are normally distributed, then the following facts are true:

- The random variable $(Y_1 - Y_2)$ satisfies

$$(Y_1 - Y_2) \sim \mathcal{N}(\mu_1 - \mu_2, \sigma_D^2),$$

  where $\sigma_D^2 = \sigma_1^2 + \sigma_2^2$. That is, the population of all possible values of the difference between Y_1 and Y_2 is also well-represented by a normal distribution.

- Define

$$\bar{D} = \bar{Y}_1 - \bar{Y}_2,$$

where

$$\bar{Y}_1 = \frac{1}{n_1}\sum_{j=1}^{n_1} Y_{1j}, \quad \bar{Y}_2 = \frac{1}{n_2}\sum_{j=1}^{n_2} Y_{2j}.$$

That is, $\bar{D}$ is the difference in sample means for the two samples. Then, just as for the sample mean from a single sample, the difference in sample means is also normally distributed, i.e.,

$$\bar{D} \sim \mathcal{N}(\mu_1 - \mu_2, \sigma^2_{\bar{D}}), \quad \sigma^2_{\bar{D}} = \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}.$$

Thus, the mean of the population of all possible differences in sample means from all possible samples from the 2 populations is the difference in means for the original populations, by analogy to the single-population case. Thus, the statistic

$$\frac{\bar{D} - (\mu_1 - \mu_2)}{\sigma_{\bar{D}}}.$$

would follow a **standard normal** distribution.

INTUITION: Use $\bar{D}$ as an estimator of $(\mu_1 - \mu_2)$. As before, for a single population, we would like to report an **interval** assessing the quality of the sample evidence; that is, give a confidence interval for $(\mu_1 - \mu_2)$.

In practical situations, σ_1^2 and σ_2^2 will be **unknown**. The obvious strategy would be to replace them by **estimates**. We will consider this first in the simplest case:

- $n_1 = n_2 = n$, i.e. the two samples are of the same size.
- $\sigma_1^2 = \sigma_2^2 = \sigma^2$, i.e. the variances of the 2 populations are the **same**. This essentially says that application of the 2 different treatments affects **only** the mean of the response, not the variability. In many circumstances, this is not unreasonable. We will discuss this further later in the course.

These simplifying assumptions are not necessary to the methods we now consider, but simply make the notation a bit easier, so that the concepts will not be obscured by so many symbols!

Under these conditions,

$$\bar{Y}_1 = \frac{1}{n}\sum_{j=1}^{n} Y_{1j}, \quad \bar{Y}_2 = \frac{1}{n}\sum_{j=1}^{n} Y_{2j}$$

and

$$\sigma^2_{\bar{D}} = 2\left(\frac{\sigma^2}{n}\right).$$

If we wish to replace $\sigma^2_{\bar{D}}$ and hence σ^2 by an estimate, we must first determine a suitable estimate under these conditions. Because the variance is considered to be **the same** in both populations, it makes sense to use the information from **both** samples to come up with such an estimate. That is, "pool" information from the two samples to arrive at a single estimate.

A "pooled" estimate of the common σ^2 is given by

$$s^2 = \frac{(n-1)s_1^2 + (n-1)s_2^2}{2(n-1)} = \frac{s_1^2 + s_2^2}{2}, \tag{4.4}$$

where s_1^2 and s_2^2 are the sample variances from each sample. We use the same notation as for a single sample, s^2, as again we are estimating a single population variance (but now from 2 samples). Note that the "pooled" estimate is just the **average** of the two sample variances, which makes intuitive sense. We have written s^2 as we have in the middle of (4.4) to highlight the general form – when we discuss unequal sample sizes later, it will turn out that the estimator for a common σ^2 will be a **weighted average** of the two sample variances. Here, the weighting is equal because the sample sizes are equal.

Thus, an obvious estimator for $\sigma^2_{\bar{D}}$ is

$$s^2_{\bar{D}} = 2\left(\frac{s^2}{n}\right).$$

Just as in the single-population case, we would thus consider the statistic

$$\frac{\bar{D} - (\mu_1 - \mu_2)}{s_{\bar{D}}}, \quad s_{\bar{D}} = \sqrt{\frac{2}{n}}s. \tag{4.5}$$

It may be shown that this statistic has a Student's t distribution with $2(n-1)$ degrees of freedom.

CONFIDENCE INTERVAL FOR $(\mu_1 - \mu_2)$: By the same reasoning as in the single-population case, then, we may use the last fact to construct a confidence interval for $(\mu_1 - \mu_2)$. In particular, by writing down the same type of probability statement and rearranging, and using the fact that the statistic (4.5) has a $t_{2(n-1)}$ distribution, we have for confidence coefficient $(1-\alpha)$,

$$P\left(-t_{2(n-1),\alpha/2} \le \frac{\bar{D} - (\mu_1 - \mu_2)}{s_{\bar{D}}} \le t_{2(n-1),\alpha/2}\right) = (1-\alpha),$$

$$P(\bar{D} - t_{2(n-1),\alpha/2}s_{\bar{D}} \le \mu_1 - \mu_2 \le \bar{D} + t_{2(n-1),\alpha/2}s_{\bar{D}}) = (1-\alpha).$$

The **confidence interval** for $(\mu_1 - \mu_2)$ is thus

$$(\bar{D} - t_{2(n-1),\alpha/2} s_{\bar{D}}\,,\ \bar{D} + t_{2(n-1),\alpha/2} s_{\bar{D}}).$$

INTERPRETATION: The interpretation is the same as for a single sample and single population. The confidence interval is a statement about the quality of the evidence in the samples from the 2 populations about the **fixed population parameter** $(\mu_1 - \mu_2)$.

EXAMPLE: (Dixon and Massey, 1969, *Introduction to Statistical Analysis.*) The following data concern two types of rations, A and B, being fed to pigs. An experiment was conducted in which 12 randomly selected pigs were fed ration A, and 12 were fed ration B, with the goal of determining whether there is a difference in the weight gains (lbs) for pigs fed the two different rations.

A:	31	34	29	26	32	35	38	34	30	29	32	31
B:	26	24	28	29	30	29	32	26	31	29	32	28

Let population 1 (2) be that of pigs fed ration A (B). Here, then, $n_1 = n_2 = n = 12$.

The usual calculations give $\bar{Y}_1 = 31.75$, $\bar{Y}_2 = 28.6667$. Also, $(n-1) = 11$, and $11s_1^2 = 112.25$, $11s_2^2 = 66.64$, so that

$$s^2 = \frac{112.25 + 66.64}{2 \times 11} = 8.1314.$$

(We did not do the division by 11 to reduce round-off error.) Thus, we have

$$\bar{D} = 3.0833 \text{ and } s_{\bar{D}} = \sqrt{\frac{2(8.1314)}{12}} = 1.1641.$$

From tables of the t distribution, we have $t_{22,0.025} = 2.074$ (for $\alpha = 0.05$). Thus, a 95% confidence interval for $(\mu_1 - \mu_2)$ is

$$(3.0833 - (2.074)(1.1641)\,,\ 3.0833 + (2.074)(1.1641)) = (0.6689\,,\ 5.4978).$$

QUESTION: What do you suppose the interval suggests about the quality of the sample evidence regarding the statement that the difference in mean weight gain between the rations is 0?

We will examine such questions in the next chapter of the course.

5 Inference on Means and Hypothesis Testing

Complementary Reading: STD, Chapter 5

5.1 Introduction

In the last section of the course, we began to discuss formal statistical inference, focusing in particular on inference on a population mean for a single population and on the difference of two means under some simplifying conditions (equal sample sizes and variances). We saw in these situations how to

- **Estimate** a single population mean or difference of means.
- Make a formal, probabilistic statement about the sampling procedure used obtain the data; i.e. how to construct a **confidence interval** for a single population mean or difference of means.

Both estimation and construction of confidence intervals are ways of getting an idea of the (unknown) value of a population parameter of interest (mean or difference of means) and taking into account the uncertainty involved because of sampling and biological variation.

In this chapter, we delve more deeply into the notions of **statistical inference**. As with our first exposure to these ideas, **probabilistic** statements will play an important role.

A RELATED IDEA: We have already discussed confidence intervals as a way of quantifying the quality of the inferences we may make from our samples. Often, we have in mind specific questions about the **values** of unknown population parameters; for example, is the difference of two means greater than zero? A formal framework and terminology has been developed under which to answer such specific questions taking into account the uncertainty involved. This framework is based on the same notions used in the construction of confidence intervals. We will discuss this framework first in the context of a single population, then describe similar results for other situations.

NORMALITY ASSUMPTION: As we have previously stated, all of the methods we will discuss are based on the assumption that the data are **approximately normally distributed**; that is, the **normal distribution** provides a reasonable description of the population(s) of the random variable(s) of interest. It is important to recognize that this is exactly that, **an assumption**. It is often valid, but does not necessarily have to be. **Always** keep this in mind. The procedures we describe may lead to misleading inferences if this assumption is seriously violated.

5.2 Hypothesis tests or tests of significance

PROBLEM: Often, we take observations on a sample with a **specific** question in mind. For example, consider the data on weight gains of rats treated with vitamin A discussed in the last chapter of the course.

Suppose that we know from several years of experience that the average (mean) weight gain of rats of this age and type during a 3 week period when they are **not** treated with vitamin A is 27.8 mg.

QUESTION: If we treat rats of this age and type with 2.5 units of vitamin A, how does this affect 3-week weight gain? That is, if we could administer 2.5 units of vitamin A to the entire population of rats of this age and type, would the (population) mean weight gain **change** from what it would be if we did not?

Of course, we cannot administer vitamin A to all rats, nor are we willing to wait for several years of accumulated experience to comment. The obvious strategy, as we have discussed, is to plan to obtain a **sample** of such rats, treat them with vitamin A, and view the sample as being drawn (randomly) from the (unobservable) population of rats treated with vitamin A. This population has (unknown) mean μ.

We carry out this procedure, and obtain **data** (weight gains for each rat in the sample). Clearly, our question of interest may be regarded as a question about μ. Either

(i) $\mu = 27.8$ mg; that is, vitamin A treatment **does not** affect weight gain, and the mean is what we know from vast past experience, 27.8 mg, despite administration of vitamin A.

(ii) $\mu \neq 27.8$ mg; that is, vitamin A treatment **does** have an affect on weight gain.

Statements (i) and (ii) are called **statistical hypotheses** (about the value of μ in this case).

(i) $\mu = 27.8$ mg states that the phenomenon of interest has **no effect**. We call such a hypothesis the **null hypothesis** and write

$$H_0 : \mu = 27.8.$$

[(ii) $\mu \neq 27.8$ mg states what we suspect or hope is the case; here, that vitamin A **does** have an effect. We call such a hypothesis an **alternative** hypothesis (to H_0) and write

$$H_1 : \mu \neq 27.8.$$

A formal statistical procedure for "deciding" between H_0 and H_1 is called a **hypothesis test** or **test of significance**.

IDEA: We base our "decision" on observation of a **sample** from the population with mean μ, thus, our decision is predicated on the quality of the sampling procedure and the inherent biological variation in the thing we are studying in the population. Thus, as with confidence intervals, **probability** will be involved.

Suppose in truth that $\mu = 27.8$ mg. (i.e. vitamin A has no effect, H_0). For the particular sample we ended up with, recall we observed $\bar{Y} = 41.0$ mg (from a sample size of $n = 5$). The key question is thus:

> How **likely** is it that we would see a sample yielding a sample mean $\bar{Y} = 41.0$ mg if it is indeed true that the population mean $\mu = 27.8$ mg??

- If it is "likely," then 41.0 is not particularly "unusual;" thus, we would not discount H_0 as an explanation for what is going on. (We **do not reject** H_0 as an explanation.)
- If it is "not likely," then 41.0 is "unusual" and "unexpected." This would cause us to think that perhaps H_0 is **not** a good explanation for what is going on. (**Reject** H_0 as an explanation, as it seems unlikely.)

As you might expect, we characterize the notion of "likely" in terms of probability.

FORMAL METHOD : "Pretend" H_0 is true and assess the probability of seeing the $\bar{Y}$ value we saw for our particular sample.

- If this probability is **small**, **reject** H_0.
- If this probability is **not small**, **do not reject** H_0.

How "small" is **small**?

Consider the generic situation where

$$H_0 \mu = \mu_0 \text{ vs. } H_1 \mu \neq \mu_0,$$

where μ_0 is the value of interest ($\mu_0 = 27.8$ mg in the rat example). If we assume ("pretend") H_0 is true, then we assume that $\mu = \mu_0$, and we would like to determine the probability of seeing a sample mean $\bar{Y}$ (our "best guess" for the value of μ) like the one we ended up with.

Recall that if the r.v. of interest Y is **normal**, then we know that, **under our assumption that** $\mu = \mu_0$

$$\frac{\bar{Y} - \mu_0}{s_{\bar{Y}}} \sim t_{n-1}. \tag{5.1}$$

That is, a sample mean calculated from a sample drawn from the population of all Y values, when "centered" and "scaled," behaves like a t r.v.

Intuitively, a "likely" value of $\bar{Y}$ would be one for which $\bar{Y}$ is "close" to μ_0. Equivalently, we would expect the value of the statistic (5.1) to be "small" (close to zero) in magnitude, i.e.

$$\left| \frac{\bar{Y} - \mu_0}{s_{\bar{Y}}} \right| \text{ close to } 0.$$

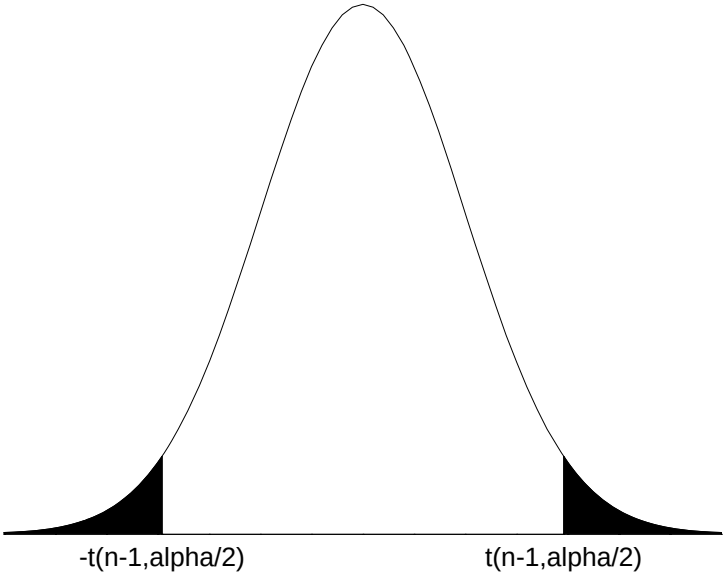
On the other hand, an "unlikely" value of $\bar{Y}$ would be one for which the statistic is "large" in magnitude.

To formalize the notion of "unlikely," suppose we decide that if the probability of seeing the value of $\bar{Y}$ that we saw is **less than** some small value α, say $\alpha = 0.05$, then things are sufficiently unlikely for us to be concerned that our "pretend" assumption $\mu = \mu_0$ may not hold. We would thus feel that the evidence in the sample (the value of $\bar{Y}$ we saw) is strong enough to refute our original assumption that H_0 is true.

Formally, we know that the probabilities corresponding to values of the statistic in (5.1) follow the t distribution with $(n-1)$ degrees of freedom. We thus know that there is a value $t_{n-1,\alpha/2}$ such that

$$P\left(\left| \frac{\bar{Y} - \mu_0}{s_{\bar{Y}}} \right| > t_{n-1,\alpha/2} \right) = \alpha;$$

pictorially speaking, $t_{n-1,\alpha/2}$ is that value such that each shaded region has area $\alpha/2$, for a total of α:



Values of the statistic that are greater in magnitude than $t_{n-1,\alpha/2}$ are thus "unlikely" in the sense that the chance of seeing them is less than α, the "cut-off" probability for "unlikeliness" we have specified.

The value of the statistic (5.1) **we saw** in our sample is thus a realization of a t r.v. Thus, if the value we saw is greater in magnitude than $t_{n-1,\alpha/2}$, we would consider the $\bar{Y}$ we got to be "unlikely" (likely to be seen with probability $\leq \alpha$), and we would **reject** H_0 as the explanation for what is really going on in favor of the other explanation, H_1.

IMPLEMENTATION: Compare the value of the statistic (5.1) to the appropriate value $t_{n-1,\alpha/2}$. In the rat example, we take $\mu_0 = 27.8$ mg (H_0 assumed true). From the calculations in the last chapter of the course, we have $n = 5$, $s_{\bar{Y}} = 4.472$, so that

$$\left| \frac{\bar{Y} - \mu_0}{s_{\bar{Y}}} \right| = \left| \frac{41.0 - 27.8}{4.472} \right| = 2.952.$$

From the table of the t distribution, if we take $\alpha = 0.05$, we have $t_{4,0.025} = 2.776$. Comparing the value of the statistic we saw to this value gives

$$2.952 > 2.776.$$

We thus **reject** H_0 – the evidence in our sample is strong enough to support the contention that mean weight gain is **different from** $\mu_0 = 27.8$, that is, H_1; i.e. vitamin A does have an effect on weight gain.

TERMINOLOGY: The statistic

$$\frac{\bar{Y} - \mu_0}{s_{\bar{Y}}}$$

is called a **test statistic**. A test statistic is a function of the sample information that is used as a basis for "deciding" between H_0 and H_1.

ALTERNATIVELY: Another, equivalent way to think about this procedure is in terms of **probabilities** rather than the value of the test statistic. Our test statistic is a r.v. with a t_{n-1} distribution. Thus, instead of finding the "cut-off" value for our chosen α and comparing the magnitude of the statistic for our sample to this value, we find the probability of seeing a value of the statistic with the same magnitude as that we saw, and compare this probability to α. That is, if t_{n-1} represents a r.v. with the t distribution with $(n-1)$ degrees of freedom, find

$$P\left(|t_{n-1}| > \text{ value of test statistic we saw }\right)$$

and compare this probability to α. If the probability is $< \alpha$, then the value of our test statistic must fall on the horizontal axis in the shaded region in the previous figure.

In our example, from the t table with $n - 1 = 4$, we find

$$0.02 < P(|t_4| > 2.952) < 0.05.$$

Thus, the probability of seeing what we saw is between 0.02 and 0.05 ("small"), and thus $< \alpha = 0.05$. The value of the test statistic we saw, 2.952, is sufficiently "unlikely," and we **reject** H_0.

These two ways of conducting the hypothesis test are **equivalent**.

- In the first, we think about the size of the value of the test statistic for our data. If it is "large," then it is "unlikely." "Large" depends on the probability α we have chosen to define "unlikely."
- In the second, we think directly about the probability of seeing what we saw. If the probability is "small" ("smaller" than α), then the test statistic value we saw was "unlikely."
- A "large" test statistic and a "small" probability are equivalent.
- An advantage of performing the hypothesis test the second way is that we calculate the probability of seeing what we saw – this is useful for thinking about just how "strong" the evidence in the data really is.

TERMINOLOGY: The value α, which is chosen **in advance** to quantify the notion of "likely," is called the **significance level** or **error rate** for the hypothesis test. Formally, because we perform the hypothesis test assuming H_0 is **true**, α is thus the probability of **rejecting** H_0 when it really is **true**.

When will we reject H_0? There are two scenarios:

(i) H_0 really is **not true**, and this caused the "large" value of the test statistic we saw (equivalently, the small probability of seeing a statistic like the one we saw).

(ii) H_0 is in fact **true**, but it turned out that we ended up with an "unusual" sample that caused us to reject H_0 nonetheless. Hopefully, this is **not** the case.

The situation in (ii) is a **mistake** – we end up making an incorrect judgment between H_0 and H_1. Unfortunately, because we are dealing with a chance mechanism (random sampling), it is always possible that we might make such a mistake (because of **uncertainty**).

A mistake like that in (ii) is called a **Type I Error**. The hypothesis testing procedure above ensures that we make a Type I error with probability at most α. This explains why α is often called the "error rate."

TERMINOLOGY: When we reject H_0, we say formally that we "reject H_0 at **level of significance** α." This states clearly what criterion we used to determine "likely;" if we do not state the level of significance, others have no sense of how stringent or lenient we were in our determination. An observed value of the test statistic leading to rejection of H_0 is said to be **(statistically) significant at level** α. Again, stating α is essential. **A common fault in reports of the results of data analyses is failure to state the level of significance**!

ONE-SIDED AND TWO-SIDED TESTS: For the weight gain example, we just considered the particular set of hypotheses

$$H_0 : \mu = 27.8 \text{ mg vs. } H_1 : \mu \neq 27.8 \text{ mg.}$$

Suppose we are fairly hopeful that vitamin A not only **has** an effect of some sort on weight gain, but in fact causes rats to **gain** more weight than they would if they were untreated. We would like to be able to make a statement based on our observed data that addresses this particular issue.

Under these conditions, it might be of more interest to specify a different alternative hypothesis:

$$H_1 : \mu > 27.8.$$

How would we test H_0 against **this** alternative? As we now see, the **principles** underlying the approach are similar to those above, but the procedure must be modified to accommodate the particular **direction** of a departure from H_0 in which we are interested.

INTUITION: If H_0 were really **not true**, and H_1 really is, we would expect to see a value of $\bar{Y}$ "larger" than 27.8. Equivalently, if we calculated

$$\frac{\bar{Y} - \mu_0}{s_{\bar{Y}}}$$

under the assumption $\mu_0 = 27.8$, we would expect to see a large, **positive** value of this statistic. Before, we only cared about the statistic being large in **magnitude**, regardless of direction, but, if our interest is in this new alternative hypothesis, we now care about direction!

Just as before, consider a generic set of hypotheses

$$H_0 : \mu = \mu_0 \text{ vs. } H_1 : \mu > \mu_0.$$

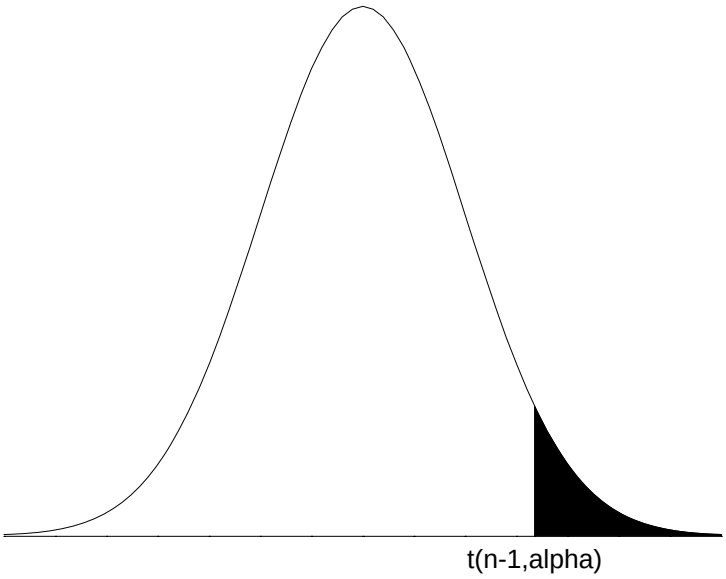
If we again "pretend" that H_0 is true, and that the random variable of interest, Y, is **normal**, then the statistic has a t_{n-1} distribution.

Using the same intuition as before, a "likely" value of $\bar{Y}$ under these conditions would be one where the value of the statistic would be close to zero. On the other hand, a value of $\bar{Y}$ that we would expect if H_0 were not true but instead H_1 were would be large and positive. We thus are interested only in the situation where $\bar{Y}$ is sufficiently far from μ_0 in the **positive direction**.

We know that there is a value $t_{n-1,\alpha}$ such that

$$P\left(\frac{\bar{Y} - \mu_0}{s_{\bar{Y}}} > t_{n-1,\alpha}\right) = \alpha;$$

pictorially speaking, the shaded region has probability α.



Note here that the **entire** probability α is concentrated in the one (positive) "tail," because we are only interested in the concept of "unlikely" as it pertains to large, positive values of the statistic. Values of the statistic that are greater than the value $t_{n-1,\alpha}$ are thus "unlikely" in this sense.

IMPLEMENTATION: Compare the value of the statistic to the appropriate value $t_{n-1,\alpha}$, now being sure to keep account of the **sign** (positive or negative). We have

$$\frac{\bar{Y} - \mu_0}{s_{\bar{Y}}} = \frac{41.0 - 27.8}{4.472} = 2.952.$$

From the table of the t distribution, if we take $\alpha = 0.05$, we have $t_{4,0.05} = 2.132$. Comparing the value of the statistic we saw to this value gives

$$2.952 > 2.132.$$

We thus **reject** H_0 – the evidence in our sample is strong enough to support the contention that mean weight gain is **greater than** $\mu_0 = 27.8$, that is, H_1; i.e. vitamin A increases weight gain.

Alternatively, consider

$$P(t_{n-1} > \text{ value of test statistic we saw}),$$

where t_{n-1} is a r.v. with the t distribution with $(n-1)$ degrees of freedom. From the t table, we find

$$0.01 < P(t_4 > 2.952) < 0.025,$$

so that the probability of seeing a value of $\bar{Y}$ like the one we did (and hence a value of the test statistic) is well below the level of significance $\alpha = 0.05$.

We thus **reject** H_0 if favor of H_1: there is evidence in the sample to support the contention that the true population mean weight gain is greater than 27.8 mg.

TERMINOLOGY: The test of hypotheses of the form

$$H_0 : \mu = \mu_0 \text{ vs. } H_1 : \mu \neq \mu_0$$

is called a **two-sided** hypothesis test – the alternative hypothesis specifies that μ is **different from** μ_0, but may be on "either side" of it. Similarly, a test of hypotheses of the form

$$H_0 : \mu = \mu_0 \text{ vs. } H_1 : \mu > \mu_0$$

or

$$H_0 : \mu = \mu_0 \text{ vs. } H_1 : \mu < \mu_0$$

is called a **one-sided** hypothesis test – the alternative in which we are interested lies to "one side" of the null hypothesis value.

TERMINOLOGY: For either type of test, the value we look up in the t distribution table to which we compare the value of the test statistic is called the **critical value** for the test. In the one-sided test above, the critical value was 2.132; in the two-sided test, it was 2.776. Note that the critical value depends on the chosen level of significance α and the sample size n. The region of the t distribution that leads to rejection of H_0 is called the **critical region**. For example, in the one-sided test above, the critical region was

$$\frac{\bar{Y} - \mu_0}{s_{\bar{Y}}} > 2.132.$$

If we think in terms of **probabilities,**, there is a similar notion. Consider the two-sided test. The probability

$$P\left(\left|\frac{\bar{Y}-\mu_0}{s_{\bar{Y}}}\right| > \text{ what we saw}\right)$$

is called the **p-value**. The **p-value** is compared to $\alpha/2$ in a two-sided test in the alternative method of testing the hypotheses. Reporting a p-value gives more information than just reporting whether or not H_0 was rejected. For example, if $\alpha = 0.05$, and the p-value $= 0.049$, yes, we might reject H_0, but the evidence in the sample might be viewed as "borderline." Reporting the p-value accurately reflects this dilemma. On the other hand, if the p-value $= 0.001$, clearly, we reject H_0; however, the p-value indicates that the chance of seeing what we saw is **very** unlikely (1/1000!).

CHOOSING α: So far, our discussion has assumed that we have *a priori* specified a value α that quantifies our feelings about "likely." How does one decide on an appropriate value for α in real life?

Recall that we mentioned the notion of a particular type of **mistake** we might make, that of a **Type I error**. Because we perform the hypothesis test under the assumption that H_0 is true, this means that the probability we **reject** H_0 when it **is true** at most α.

Choosing α thus has to do with how serious a mistake a Type I error might be in the particular applied situation. This is best illustrated by example:

Suppose the question of interest concerns the efficacy of a costly new drug for the treatment of certain disease in humans, and the new drug has potentially dangerous side effects. Suppose a study is conducted where sufferers of the disease are randomly assigned to receive either the standard treatment or the new drug (this is called a **clinical trial**), and suppose that the r.v. of interest Y is survival time for sufferers of the disease. It is known from years of experience with the standard drug that the mean survival time is some value μ_0. We hope that the new drug is **more** effective in the sense that it **increases** survival, in which case it would be worth its additional expense and the risk of side effects.

We thus consider

$$H_0 : \mu = \mu_0 \text{ vs. } H_1 : \mu > \mu_0,$$

where $\mu =$ mean survival time under treatment with the new drug.

The data are analyzed, and suppose, unbeknownst to us, our sample leads us to commit a Type I error – we end up **rejecting** H_0 when it is **not true** and claim that the new drug is more effective than the standard drug when in reality it is not! Because the new drug is so expensive and carries the possibility of dangerous side effects, this could be a **costly** mistake, as patients would be paying more with the risk of dangerous side effects for no real gain over the standard treatment!

In a situation like this, it is intuitively clear that we would probably like α to be **very small**, so that the chance we end up rejecting H_0 when we really shouldn't is small. In situations where the consequences are not so serious if we make a Type I error, we might choose α to be larger.

ANOTHER KIND OF MISTAKE: You have probably already considered that rejecting H_0 when it really is true (i.e. a Type I error) is not the only kind of mistake we could make.

- The sample data might be "unusual" in such a way that we end up **not rejecting** H_0 when it really **is not** true (so we shuld have rejected!)
- This type of mistake is called a **Type II Error**.

Because a Type II error is also a mistake, we would like the probability of committing such an error, β, say, to also be **small**. In many situations, a Type II error is not as serious a mistake as a Type I error. In our drug example, if we commit a Type II error, we infer that the new drug is **not** effective when it really **is**. Although this, too, is undesirable, as we are discarding a potentially better treatment, we are no worse off than before we conducted the test, whereas, if we commit a Type I error, we will unduly expose patients to unnecessary costs and risks for no gain.

Later, we will discuss another way of thinking about Type II error and the idea of "balancing" the probabilities α and β.

GENERAL PROCEDURE: Here we summarize the steps in conducting a test of hypotheses about a single population mean. The same principles we discuss here will be applicable in **any** testing situation.

1. Determine the question of interest. This is the first and foremost issue – no experiment should be conducted unless the scientific questions are well-formulated. (This has implications for **design**, as we'll discuss later.)

2. Express the question of interest in terms of 2 hypotheses about μ:

   - H_0: **null** hypothesis – the hypothesis of **no effect**. If μ_0 is a specified value, $H_0 : \mu = \mu_0$.
   - H_1: **alternative** hypothesis – the thing we suspect or hope is true. Depending on the question of interest, H_1 will be two-sided, $ha : \mu \neq \mu_0$ or one-sided, $H_1\mu > \mu_0$ or $H_1\mu < \mu_0$.

3. Choose the significance level α, usually a small value like 0.05. The particular situation (severity of making a Type I error) will dictate the value.

4. Conduct the experiment, collect the data, determine critical value, and calculate the test statistic. Perform the hypothesis test, either **rejecting** or **not rejecting** H_0 in favor of H_1.

REMARKS:

- We do not even collect data until the question of interest has been established!

- You will often see the phrase "Accept H_0" used in place of "do not reject H_0." This terminology may be misleading. If we **do** reject H_0, we are saying that the sample evidence is **sufficiently strong** to suggest that H_0 is probably **not true**. On the other hand, if we **do not** reject H_0, we do not because the sample does not contain enough evidence to say it probably not true. This **does not** imply that there is enough evidence to say that it probably **is true**! Tests of hypotheses are set up so that we **assume** H_0 is true and then try to refute it – if we can't, this doesn't mean the assumption **is** true, only that we couldn't reject it. It could well be that the alternative hypothesis H_1 is indeed true, but, because we got an "unusual" sample, we couldn't reject H_0 – this doesn't make H_0 true!

  Another way to think of this: We conduct the test based on the presumption of some value μ_0; in the rat example, $\mu_0 = 27.8$. Suppose that we conducted a hypothesis test and **did not reject** H_0. Now suppose that we changed the value of μ_0; say, in the rat example, to $\mu_0 = 27.7$, and performed a hypothesis test and, again, **did not reject** H_0. Both 27.8 **and** 27.7 can't be true! If we "accepted" H_0 in each case, we'd have a conflict!

- The significance level and critical region are **not** cast in stone. The results of hypothesis tests should not be viewed with **absolute** yes/no interpretation, but rather as **guidelines** for aiding us in interpreting experimental results and deciding what to do next. Often, experimental conditions are so complicated that we can never be entirely assured that the assumptions necessary to validate **exactly** our statistical methods are satisfied. For example, the assumption of normality may be only an approximation, or may in fact be downright unsuitable. It is thus important to keep this firmly in mind. It has become very popular in a number of applied disciplines to do all tests of hypotheses at level 0.05 regardless of the setting, and to strive to find "p-value less than 0.05;" however, if one is realistic, a p-value of, say, 0.049 must be interpreted with these cautionary remarks in mind.

- Many statistical packages such as SAS do not report whether or not a hypothesis was rejected at a given level of significance α, but rather report the exact probability of seeing what we saw or something larger under the assumption that H_0 is true, That is, they report the **p-value**. This way, a test at any level of significance may be performed by comparing the p-value reported in the output to the chosen α.

5.3 Relationship with confidence intervals

Before we discuss more exotic hypotheses and hypothesis tests, we point out something alluded to at the beginning of our discussion. We have already seen that a **confidence** interval for a single population mean is based on the probability statement

$$P\left(-t_{n-1,\alpha/2} \leq \frac{\bar{Y}-\mu}{s_{\bar{Y}}} \leq t_{n-1,\alpha/2}\right) = 1-\alpha. \tag{5.2}$$

As we have seen, a **two-sided** hypothesis test is based on a probability statement of the form

$$P\left(\left|\frac{\bar{Y}-\mu}{s_{\bar{Y}}}\right| > t_{n-1,\alpha/2}\right) = \alpha. \tag{5.3}$$

Comparing (5.2) and (5.3), a little algebra shows that they are **the same** (except for the strict versus not-strict inequalities "$\leq$" and "$>$", which are irrelevant for continuous r.v.s).

Thus, choosing a "small" level of significance α in a hypothesis test is the same as choosing a "large" confidence coefficient $(1-\alpha)$. Furthermore, for the **same** choice α, (5.2) and (5.3) show that the following two statements are equivalent:

(i) Reject $H_0 : \mu = \mu_0$ at level α based on $\bar{Y}$

(ii) μ_0 is not contained in a $100(1-\alpha)\%$ confidence interval for μ based on $\bar{Y}$.

That is, two-sided hypothesis tests about μ and confidence intervals for μ yield the **same** information (recall the question at the end of chapter 4 of the course).

ONE-SIDED CONFIDENCE BOUNDS: What about **one-sided** hypothesis tests? There is a similar notion to a confidence interval in this situation. When we are interested in estimating μ in a situation where we suspect of hope that μ is "large" (or "small"), we would presumably only be interested in the smallest (largest) acceptable value of μ at a given confidence coefficient $(1-\alpha)$. That is, if we hope μ is "large," we might want only a "lower bound" on the value of μ.

Again, because

$$\frac{\bar{Y}-\mu}{s_{\bar{Y}}} \sim t_{n-1},$$

we have

$$P\left(\frac{\bar{Y}-\mu}{s_{\bar{Y}}} < t_{n-1,\alpha}\right) = 1-\alpha \quad \Rightarrow \quad P(\mu > \bar{Y} - t_{n-1,\alpha}s_{\bar{Y}}) = 1-\alpha$$

(by algebra – try it!) Thus, a **lower confidence bound** for μ with confidence coefficient $(1-\alpha)$ is defined as

$$\bar{Y} - t_{n-1,\alpha}s_{\bar{Y}}.$$

By an entirely similar argument, a **upper confidence bound** for μ with confidence coefficient $(1-\alpha)$ is

$$\bar{Y} + t_{n-1,\alpha}s_{\bar{Y}}.$$

EXAMPLE: For the rat weight gain data, because we suspect that vitamin A leads to increased weight gain, that is, $\mu > 27.8$, (μ "large"), we may be interested in reporting a lower confidence bound on the mean weight gain μ using vitamin A along with our estimate $\bar{Y} = 41.0$. With $(1-\alpha) = 0.95$, $t_{4,0.05} = 2.132$, $s_{\bar{Y}} = 4.472$,, we have

$$41.0 - (2.132)(4.472) = 31.467 \text{ mg.}$$

Thus, 31.467 mg is a 95% lower confidence bound for μ.

NOTE: If we were interested in testing the null hypothesis $\mu = 27.8$ against the one-sided alternative at level $\alpha = 0.05$, we could do it by considering the 95% lower confidence bound. Because the null hypothesis value 27.8 is **below** the lower confidence bound, it is not contained in this "one-sided" interval. This is equivalent to rejecting H_0 in favor of $\mu > 27.8$ (which is what we did in the previous section).

5.4 Tests of hypotheses for the mean of a single population

We introduced the basic underpinnings of tests of hypotheses in the context of a single population mean. Here, we summarize the procedure for convenience. The same reasoning underlies the development of tests for other situations of interest; these are described in subsequent sections.

General form: $H_0 : \mu = \mu_0$ vs.

one-sided: $H_1 : \mu > \mu_0$ or $H_1 : \mu < \mu_0$
two-sided: $H_1 : \mu \neq \mu_0$

Test statistic:

$$t = \frac{\bar{Y} - \mu_0}{s_{\bar{Y}}}$$

Procedure: Reject H_0 if

one-sided: $t > t_{n-1,\alpha}$ or $t < -t_{n-1,\alpha}$
two-sided: $|t| > t_{n-1,\alpha/2}$

for level of significance α

NOTE: For a one-sided test with alternative of the form $H_1 : \mu < \mu_0$, note that we compare the raw value of the test statistic to the **negative** of $t_{n-1,\alpha}$. For this test, intuition suggests that we would reject H_0 for large, **negative** values of the statistic; thus, values in the left-hand "tail" of the t-distribution would be required. Because of **symmetry**, this is just the "reverse" of what we do for $H_1 : \mu > \mu_0$.

EXAMPLE: (Zar, 1974, *Biostatistical Analysis*, p. 100). It is thought that the body temperature of intertidal crabs exposed to air is less than the ambient temperature. Body temperatures were obtained from a random sample of 8 such crabs exposed to an ambient temperature of 25.4 degrees Celsius.

```
25.8    24.6    26.1    24.9    25.1    25.3    24.0    24.5
```

Assume that body temperatures are approximately normally distributed. Let μ be the mean body temperature for the population of intertidal crabs exposed to an ambient temperature of 25.4 degrees Celsius. Then we wish to test

$$H_0 : \mu = 25.4 \text{ deg. C vs. } H_1 : \mu < 25.4 \text{ deg. C.}$$

This is a one-sided test. Take $\alpha = 0.05$. We have $n = 8$, $-t_{7,0.05} = -1.895$,

$$\bar{Y} = \frac{200.3}{8} = 25.04, \quad s^2 = 0.4798, \quad s_{\bar{Y}} = 0.245$$

so that

$$\frac{\bar{Y} - \mu_0}{s_{\bar{Y}}} = \frac{25.04 - 25.4}{0.245} = -1.470.$$

The value of the test statistic, -1.470 is **not less than** the critical value -1.895; we **do not reject** H_0 at level of significance 0.05. There is not enough evidence in the sample to suggest that the mean body temperature of intertidal crabs exposed to air at 25.4 degrees Celsius is indeed less that 25.4.

REMARK: A philosophical point: can we say anything about mean body temperature of crabs at an ambient temperature **other than** 25.4 degrees Celsius based on the results of this experiment?

5.5 Testing the difference of two population means

As we have discussed, the usual situation in practice is that in which we would like to **compare** two competing treatments or compare a treatment to a control. Formally, our model for this situation is to think of two populations, one corresponding to the population of interest were every member treated with one treatment, and the other accordingly. We have already discussed confidence intervals for the difference in population means; now, using the same principles as for a single mean, we describe testing hypotheses about the difference.

SCENARIO: 2 populations:

Population 1 $Y_1 \sim \mathcal{N}(\mu_1, \sigma_1^2)$, $Y_{11}, Y_{12}, \ldots, Y_{1n_1} \Rightarrow \bar{Y}_1,\ s_1^2$

Population 2 $Y_2 \sim \mathcal{N}(\mu_2, \sigma_2^2)$, $Y_{21}, Y_{22}, \ldots, Y_{2n_2} \Rightarrow \bar{Y}_2,\ s_2^2$

IMPORTANT: Because the two samples do not involve the same "experimental units" (the things giving rise to the responses), they may be thought of as **independent** (totally unrelated).

General form: $H_0 : \mu_1 - \mu_2 = \delta$ vs.

one-sided: $H_1 : \mu_1 - \mu_2 > \delta$

two-sided: $H_1 : \mu_1 - \mu_2 \neq \delta$

where δ is some value. Most often, $\delta = 0$, so that the null hypothesis corresponds to the hypothesis of **no difference** between the population means.

More generally, we might be interested in other values of δ. For example, if a small difference between the population means is not **scientifically important**, a small difference is effectively equivalent to "no effect." In this case, the null hypothesis is that μ_1 and μ_2 differ by a "scientifically unimportant" amount (δ) against the alternative that they differ by some greater amount.

Note that we have only written the one-sided test one way. A one-sided hypothesis for the difference of two treatment means can always be written with a ">" by identifying the population thought to have the larger mean as population 1. We will adopt this convention.

Test statistic: As in the case of constructing confidence intervals for $\mu_1 - \mu_2$, intuition suggests that we base inference on $\bar{Y}_1 - \bar{Y}_2$. The test statistic is

$$t = \frac{\bar{D} - \delta}{s_{\bar{D}}}.$$

where $\bar{D} = \bar{Y}_1 - \bar{Y}_2$ and $s_{\bar{D}}$ is an estimate of $\sigma_{\bar{D}}$ (more in a moment).

Note that the test statistic is constructed under the assumption that the mean of $\bar{D}$, $\mu_1 - \mu_2$, is equal to δ. This is analogous to the one-sample case – we perform the test under the assumption that H_0 is **true**.

The exact form of the test statistic t and how the test is conducted depend on the situation. Recall that we derived in chapter 4 confidence intervals for the population mean difference in the simple case where (1) $n_1 = n_2 = n$ and (2) $\sigma_1^2 = \sigma_2^2$. Here, we present more general results.

THE CASE OF EQUAL VARIANCES, $\sigma_1^2 = \sigma_2^2 = \sigma^2$: As we have already discussed, it is often reasonable to assume that the 2 populations have **common** variance σ^2. One interpretation is that the phenomenon of interest (say two different treatments) affects only the mean of the response, not its variability ("signal" may change with changing treatment, but "noise" stays the same). (In a moment, we discuss formal hypothesis tests for equality of variances). In this case, we "pool" the data from both samples to estimate the common variance σ^2. The obvious estimator is

$$s^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{(n_1 - 1) + (n_2 - 1)}, \tag{5.4}$$

where s_1^2 and s_2^2 are the sample variances for each sample. Thus, (5.4) is a **weighted average** of the two sample variances, where the "weighting" is in accordance with the sample sizes. We have already discussed such "pooling" when the sample size is the **same**, in which case this reduces to a **simple average**. (5.4) is a generalization to allow differential weighting of the sample variances when the sample sizes are different. Clearly, the sample with the greater sample size should get "more weight."

Recall that in general,

$$\sigma^2_{\bar{D}} = \frac{\sigma^2_1}{n_1} + \frac{\sigma^2_2}{n_2}.$$

When the variances are the same, this reduces to

$$\sigma^2_{\bar{D}} = \sigma^2 \left(\frac{1}{n_1} + \frac{1}{n_2} \right),$$

which of course may be estimated by plugging in the "pooled" estimator for σ^2. We thus arrive at

$$s^2_{\bar{D}} = s^2 \left(\frac{1}{n_1} + \frac{1}{n_2} \right), \quad s_{\bar{D}} = s\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}.$$

It is straightforward to see that these formulas reduce to those we have already specified when the sample sizes are the same.

Note that the total **degrees of freedom** across both samples is

$$(n_1 - 1) + (n_2 - 1) = n_1 + n_2 - 2,$$

so the $t_{n_1+n_2-2}$ distribution is relevant. The degrees of freedom of course reduce to $2(n-1)$ for common sample size n.

Procedure: For level of significance α,

Reject H_0 if

one-sided: $t > t_{n_1+n_2-2,\alpha}$

two-sided: $|t| > t_{n_1+n_2-2,\alpha/2}$

Confidence intervals: Extending our previous results, a $100(1-\alpha)\%$ confidence interval for $\mu_1 - \mu_2$ would be

$$(\bar{D} - t_{n_1+n_2-2,\alpha/2} s_{\bar{D}}, \bar{D} + t_{n_1+n_2-2,\alpha/2} s_{\bar{D}}).$$

By reasoning similar to that for a single sample, a **one-sided lower confidence bound** for $\mu_1 - \mu_2$ would be

$$\bar{D} - t_{n_1+n_2-2,\alpha} s_{\bar{D}}.$$

The relationship between hypothesis tests and confidence intervals is the same as in the single sample case:

- **two-sided** – if δ is not contained in the confidence interval, we have little confidence in the statement that $\mu_1 - \mu_2 = \delta \Leftrightarrow$ reject H_0.
- **one-sided** – if δ is less than the confidence bound, we have little confidence in the statement that $\mu_1 - \mu_2 > \delta \Leftrightarrow$ reject H_0.

EXAMPLE: (Box, Hunter, and Hunter, 1978, *Statistics for Experimenters*, p. 94) A randomized design used in the comparison of standard versus modified fertilizer mixtures for tomato plants.

The following data were obtained in an experiment conducted by a gardener whose objective was to discover whether a change in the fertilizer mixture applied to his tomato plants would result in improved yield. He had 11 plants in a row – 5 were given the standard mixture A and the remaining 6 were given the supposedly improved (modified) mixture B. To ensure fair and random assignment, the As and Bs were randomly allocated to each plant position in the row to give the design shown below. To do this, the gardener took 11 playing cards, 5 red and 6 black, shuffled them, and dealt them out in a row, letting red = A, black = B. He then fertilized his plants accordingly. (We will discuss more about experimental design later in the course – note that this is a method of **randomization**).

Position in row :	1	2	3	4	5	6	7	8	9	10	11
Fertilizer:	A	A	B	B	A	B	B	B	A	A	B
Pounds of tomatoes:	29.9	11.4	26.6	23.7	25.3	28.5	14.2	17.9	16.5	21.1	24.3

Standard A	Modified B
29.9	26.6
11.4	23.7
25.3	28.5
16.5	14.2
21.1	17.9
	24.3

Assume that the yields are approximately normally distributed. Let B = population 1, A = population 2, as we believe the modified fertilizer will have the larger mean. The question of interest is "Does the new mixture B give superior yield?" Thus, $\delta = 0$ and we test

$$H_0 : \mu_1 - \mu_2 = 0 \text{ vs. } H_1 : \mu_1 - \mu_2 > 0.$$

Take $\alpha = 0.05$. We have $n_1 = 6$, $n_2 = 5$ so that $n_1 + n_2 - 2 = 9$. $t_{9,0.05} = 1.833$.

$$\bar{Y}_1 = 22.53 \qquad \bar{Y}_2 = 20.84$$

$$s_1^2 = \frac{147.533}{5} = 29.51 \qquad s_2^2 = \frac{209.992}{4} = 52.50$$

$$s^2 = \frac{5(29.51) + 4(52.50)}{9} = 39.73$$

$$\bar{D} = \bar{Y}_1 - \bar{Y}_2 = 1.69, \quad s_{\bar{D}} = \sqrt{39.73\left(\frac{1}{6} + \frac{1}{5}\right)} = 3.82$$

$$\frac{\bar{D} - \delta}{s_{\bar{D}}} = \frac{1.69}{3.82} = 0.44$$

Because 0.44 is not greater than 1.833, we **do not reject** H_0. There is not enough evidence in the sample to suggest that the new mixture produces superior yield.

(A one-sided lower confidence bound for $\mu_1 - \mu_2$ is $1.69 - (1.833)(3.82) = -5.3121$. Because $\delta = 0$ is greater than the lower confidence bound, -5.3121, we have little confidence in the statement that $\mu_1 - \mu_2 > 0$. The lower bound includes possibilities for the difference that are in the opposite direction from H_0!)

EXAMPLE: The pig data. The goal was to determine if 2 different rations fed to pigs result in different weight gains. We let population 1 (2) be pigs fed ration A (B). Again, we are interested in whether the rations are different, so $\delta = 0$. We test

$$H_0 : \mu_1 - \mu_2 = \delta = 0 \text{ vs. } H_1 : \mu_1 - \mu_2 \neq 0.$$

We will use $\alpha = 0.05$. $n_1 = n_2 = n = 12$ so that $n_1 + n_2 - 2 = 2(n-1) = 22$. From before,

$$\bar{Y}_1 = 31.75, \quad s_1^2 = \frac{112.5}{11} \qquad \bar{Y}_2 = 28.67, \quad s_2^2 = \frac{66.64}{11}$$

$$\bar{D} = 3.0833, \quad s_{\bar{D}} = 1.1641, \quad t_{22,0.025} = 2.074.$$

Thus

$$\frac{\bar{D} - \delta}{s_{\bar{D}}} = \frac{3.0833}{1.1641} = 2.469.$$

We have $2.469 > 2.074$, so we **reject** H_0. The evidence in the samples is sufficiently strong to suggest that there is a difference in weight gains between the two rations.

THE CASE OF UNEQUAL VARIANCES, $\sigma_1^2 \neq \sigma_2^2$: Cases arise commonly in practice where it is unreasonable to assume that the variances of the two populations are the same. (In a moment, we will consider a hypothesis test for equality of variances and how to design experiments to get around this problem.)

Unfortunately, things get a bit more complicated when the variances are unequal. In this case, we may not "pool" information, because we do not believe that the variances are the same. Instead, we use the two sample variances to estimates $\sigma_{\bar{D}}$ in the obvious way:

$$s_{\bar{D}} = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}.$$

Note that because we can't pool the sample variances, we can't pool degrees of freedom, either. It may be shown mathematically that, under these conditions, if we use $s_{\bar{D}}$ calculated in this way in the denominator of our test statistic

$$\frac{\bar{D} - \delta}{s_{\bar{D}}},$$

the statistic no longer has **exactly** a t distribution! In particular, it is not clear what to use for degrees of freedom, as we have estimated two variances separately!

It turns out that an approximation is available that may be used under these circumstances. One calculates the quantity

$$\text{"`effective degrees of freedom"} = \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\left(\frac{s_1^2}{n_1}\right)^2/(n_1 - 1) + \left(\frac{s_2^2}{n_2}\right)^2/(n_2 - 1)}.$$

Clearly, this is **not** an integer.

One then **rounds** the "effective degrees of freedom" to the nearest integer. The approximate "effective degrees of freedom" are then used as if they were exact; one finds the critical value as follows:

one-sided test: $t_{edf,\alpha}$, where *edf* is the **rounded** "effective degrees of freedom."

two-sided test: $t_{edf,\alpha/2}$, with *edf* defined above..

It is important to recognize that this is only an **approximation** – the true distribution of the test statistic is no longer **exactly** t, but we use the t distribution and the *edf* as an approximation to the true distribution. Thus, care must be taken in interpreting the results – one should be aware that "borderline" results may not be trustworthy.

EXAMPLE: (Zar,, 1984, *Biostatistical Analysis*, p. 103) It is thought that the mean clutch size of ducks raised in captivity is smaller than that of ducks breeding in the wild. Suppose it is reasonable to assume (we will examine this in a moment) that variability in clutch size is different for ducks raised in captivity from that for ducks breeding in the wild. Assume that clutch size is approximately normally distributed, so that

$$\begin{array}{lll} \text{Population 1:} & \text{Wild} & Y_1 \sim \mathcal{N}(\mu_1, \sigma_1^2) \\ \text{Population 2:} & \text{Captive} & Y_2 \sim \mathcal{N}(\mu_2, \sigma_2^2) \end{array}$$

We wish to test $H_0 : \mu_1 - \mu_2 = \delta = 0$ vs. $H_1 : \mu_1 - \mu_2 > 0$. We will use $\alpha = 0.10$. Data are collected for two random samples of ducks, one sample from each population.

Captive	Wild
10	9
11	8
12	11
11	12
10	10
11	13
11	11
	10
	12

The usual calculations give

$$n_1 = 9, \quad \bar{Y}_1 = 10.667, \quad s_1^2 = \frac{20.00}{8} = 2.500$$

$$n_2 = 7, \quad \bar{Y}_2 = 10.857, \quad s_2^2 = \frac{2.86}{6} = 0.477$$

Thus, $\bar{D} = 10.667 - 10.857 = -0.190$,

$$s_{\bar{D}} = \sqrt{\frac{2.500}{9} + \frac{0.477}{7}} = 0.5882$$

and $edf \doteq 11.485 \approx 11$. We thus look up $t_{11,0.10} = 1.363$. The test statistic is

$$\frac{\bar{D} - \delta}{s_{\bar{D}}} = \frac{-0.190}{0.5882} = -0.323.$$

This is clearly not greater than the critical value, so we **do not reject** H_0. There is not sufficient evidence in these samples to suggest that clutch size is larger for ducks breeding in the wild.

5.6 Testing equality of variances

It turns out that it is possible to construct hypothesis tests to investigate whether or not 2 **variances** for 2 independent samples drawn from 2 populations are equal.

WARNING: Testing hypotheses about variances is a **harder** problem than testing hypotheses about means. This is because it is easier to get an understanding of the "signal" in a set of data than the "noise" – sample means are **better** estimators of the population means than sample variances are of the population variances using the same sample size. Moreover, for the test we are about to discuss to be valid, the assumption of **normality** is **critical**. Thus, tests for equality of variances should be interpreted with **caution**.

We will defer the motivation behind this type of test until we discuss analysis of variance techniques later in the course. We will see that investigating differences between means may be thought of as investigating differences between variances in a certain sense.

We wish to test

$$H_0 : \sigma_1^2 = \sigma_2^2 \text{ vs. } H_1 : \sigma_1^2 \neq \sigma_2^2.$$

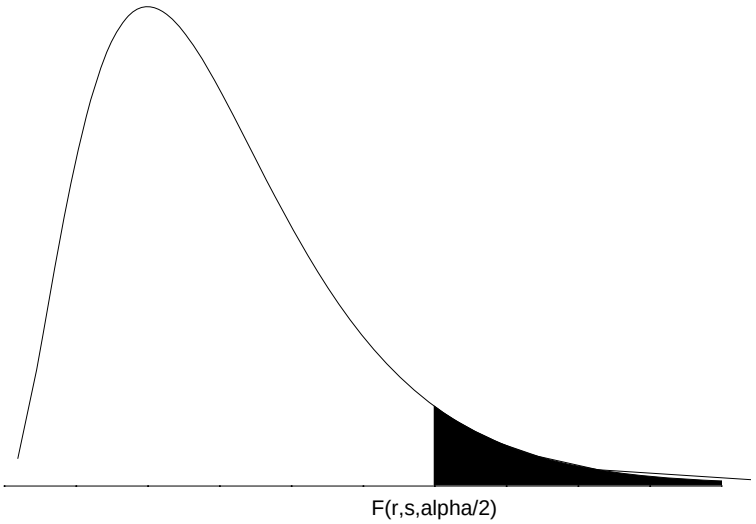
It turns out that the appropriate test statistic is the **ratio** of the two sample variances

$$F = \frac{\text{larger of } (s_1^2, s_2^2)}{\text{smaller of } (s_1^2, s_2^2)}.$$

THE F DISTRIBUTION: The **sampling distribution** of the test statistic F may be derived mathematically, just as for the distributions of our test statistics for means. In general, for **normal data**, a ratio of 2 sample variances from independent samples with sample sizes n_N (numerator) and n_D (denominator) has what is known as the F **distribution with** $(n_N - 1)$ **and** $(n_D - 1)$ **degrees of freedom**. We write a r.v. with this distribution as F_{n_N-1,n_D-1}.

This distribution has shape similar to that of the χ^2 distribution. Tables of the probabilities associated with this distribution are widely available; Table A.6 of STD is an example. For degrees of freedom r (numerator) and s (denominator), the table contains the value taken on by F such that the probability of a larger value is in the second column from the left.

For example, if the probability we were interested in were $\alpha/2$ for some value α, we would look for $\alpha/2$ in the second left column for the denominator degrees of freedom of interest (s). We would then read over to the appropriate numerator column (r) and down. The value $F_{r,s,\alpha/2}$ appearing in the body of the table would then satisfy the following picture, with shaded area $= \alpha/2$.



Procedure: Reject H_0 at level of significance α if $F > F_{r,s,\alpha/2}$.

EXAMPLE: For the duck data, we wish to test whether or not ducks raised in captivity have different variability in clutch size from ducks bred in the wild. Here, we will use $\alpha = 0.10$. From our previous calculations, we have

$$s_1^2 > s_2^2, \text{ so } F = \frac{s_1^2}{s_2^2} = \frac{2.500}{0.477} = 5.241.$$

From the tables of the F distribution, $F_{8,6,0.05} = 4.15$. Because $5.241 > 4.15$, we **reject** H_0 at level of significance $\alpha = 0.10$. There is evidence in these data to suggest that there is a real difference in variability in clutch size in the 2 populations.

Alternatively, from the F distribution table,

$$0.025 < P(F_{8,6} > 5.241) < 0.05;$$

$\alpha/2 = 0.05$, and the **p-value** is less that 0.05. Thus, we **reject** H_0 at level of significance 0.10, with the same interpretation as above.

5.7 Comparing population means using fully paired comparisons

As we will see over and over, the **analysis** of a set of data is dictated by the **design**. For example, in the developments so far on testing for differences in means and variances, it is necessary that the 2 samples be **independent** (i.e. completely unrelated); this is a requirement for the underlying mathematical theory to be valid. Furthermore, although we haven't made much of it, the fact that the samples are independent is a consequence of experimental **design** – the experimental units in each sample **do not overlap** and were assigned treatments randomly (e.g. recall the tomato example). Thus, the methods we have discussed are appropriate for the particular **experimental design**. If we **design** the experiment differently, then different methods will be appropriate.

If it is suspected **in advance** that σ_1^2 and σ_2^2 may not be equal, an alternative strategy is to use a different **experimental design** to conduct the experiment. It turns out that for this design, the appropriate methods of analysis do not depend on whether the variances for the two populations are the same. In addition, the design may make **more efficient** use of experimental resources.

The idea is to make comparisons **within pairs** of experimental units that may tend to be **more alike** than other pairs. The effect is that appropriate methods for analysis are based on considering differences **within** pairs rather than differences between the two samples overall, as in our previous work. The fact that we deal with pairs may serve to eliminate a source of uncertainty, and thus lead to **more precise** comparisons.

The type of design is best illustrated by example.

EXAMPLE: (Dixon and Massey, 1969, *Introduction to Statistical Analysis*, p. 122). A certain stimulus is thought to produce an increase in mean systolic blood pressure in middle-aged men.

One way to design an experiment to investigate this issue would be to randomly select a group of middle-aged men, and then randomly assign each man to either receive the stimulus or not. We would think of 2 populations, those of all middle-aged men with and without the stimulus, and would be interested in testing whether the mean for the stimulated population is greater than that for the unstimulated population. We would have 2 **independent** samples, one from each population, and the methods of the previous sections would be applicable.

In this set-up, variability **among** all men as well as variability **within** the two groups may make it difficult for us to detect differences. In particular, recall that variability in the sample mean difference $\bar{D}$ is characterized by the estimate $s_{\bar{D}}$, which appears in the **denominator** of our test statistic. If $s_{\bar{D}}$ is **large**, the test statistic will be **small**, and it is likely that H_0 will **not** be rejected. Even if there is a **real difference**, the statistical procedure may have a difficult time identifying it because of all the variability.

If we could eliminate the effect of some of the variability inherent in experimental material, we might be able to overcome this problem. In particular, if we **designed** the experiment in a different way, we might be able eliminate the impact of a source of variation, thus ending up with a **more sensitive** statistical test (that will be more likely to detect a **real difference** if one exists).

A better design that is in this spirit is as follows, and seems like a natural approach in a practical sense as well. Rather than assigning men to receive on treatment or the other, obtain a response from **each** man under each treatment! That is, obtain a random sample of middle-aged men and take 2 readings on each man, **with** and **without** the stimulus. This might be carried out using a **before–after** strategy, or, alternatively, the ordering for each man could be different. We will discuss the issues associated with this sort of choice later in the course; for now, it is important to keep in mind that **ordering** could be an issue, depending on the application. We will thus assume for simplicity that we measurements on each man are taken in a before–after fashion and that there is no consequence to order.

In this **alternative design**, because readings of each type are taken on the **same man**, the difference between before and after readings on a given man should be **less variable** than the difference between a before response on one man and an after response on another man! The **man-to-man** variation inherent in the latter difference is **not present** in the difference between readings taken on the **same** man. To summarize,

Design	Type of Difference	Sources of Variation
1	across men	among men, within men
2	within men	within men

In this second design, we still may think of **2 populations**, those of all men with and without the stimulus. What changes in the second design is how we have "sampled" from these populations. The 2 samples are no longer **independent**, because they involve the same men. Thus, different statistical methods are needed.

RESULT: In many situations, a natural form of **pairing** suggests itself. It is wise to set up the experiment this way if possible, because, as we have described, differences **within** experimental units may be **less variable** than those **across** experimental units. In some situations, of course, it may be impractical or impossible to do this.

APPROACH: As you might expect, instead of analyzing the before and after results separately (separate means, sample variances that we might "pool"), we analyze the **paired differences**. The rationale is as follows.

As noted above, we may still think of 2 populations, with associated means μ_1 and μ_2. However, for the purposes of deriving an appropriate analysis method, it is convenient to also think of another population – that of all possible **differences** between before and after measurements. Intuitively, we would expect the mean of this population to be

$$\mu_1 - \mu_2.$$

What we observe under this experimental design are "random drawings" from this population. We may formalize this as follows.

Let Y_1 and Y_2 be the 2 random variables representing the paired observations; e.g. in our example,

$$\begin{aligned} Y_1 &= \text{systolic blood pressure (SBP) after stimulus} \\ Y_2 &= \text{systolic blood pressure (SBP) before stimulus} \end{aligned}$$

(Note that we have designated the population with the suspected **larger** mean as Population 1, in keeping with our convention from before.)

The data are the pairs (Y_{1j}, Y_{2j}), pairs of observations from the jth man, $j = 1, \ldots, n$ (where there are n men in total). Let

$$D_j = Y_{1j} - Y_{2j} = \text{ difference for the } j\text{th pair.}$$

The relevant population is thus the population of the r.v. D, on which we have observations $D_1, \ldots, D_n$! If we think of the random variables Y_1 and Y_2 as having the means μ_1 and μ_2, our hypotheses are

$$H_0 : \mu_1 - \mu_2 = \delta \text{ vs. } H_1 : \mu_1 - \mu_2 > \delta$$

for a one-sided test and

$$H_0 : \mu_1 - \mu_2 = \delta \text{ vs. } H_1 : \mu_1 - \mu_2 \neq \delta$$

for a two-sided test, where, as before, δ is often 0.

These hypotheses may be regarded as a test about the mean of the population of all possible **differences** (i.e. the r.v. D), which has this same mean. To remind ourselves of our perspective, we could think of the mean of the population of differences as

$$\Delta_{diff} = \mu_1 - \mu_2,$$

and express our hypotheses equivalently in terms of Δ_{diff}, e.g.

$$H_0 : \Delta_{diff} = \delta \text{ vs. } H_1 : \Delta_{diff} > \delta.$$

Note that once we begin thinking this way, it is clear what we have – we are interested in testing hypotheses concerning the value of a **single population mean**, Δ_{diff} (that of the hypothetical population of differences)! Thus, the appropriate analysis is that for a single population mean based on a single sample applied to the observed **differences** $D_1, \ldots, D_n$!

We thus compute the "sample mean" and "sample variance" for the sample of differences, and proceed as in section 5.4. Compute

$$\bar{D} = \frac{1}{n}\sum_{j=1}^{n} D_j = \text{ sample mean,}$$

and

$$s_D^2 = \frac{1}{n-1}\left(\sum_{j=1}^{n} D_j^2 - \frac{\left(\sum_{j=1}^{n} D_j\right)^2}{n}\right) = \text{ sample variance.}$$

It turns out that the sample mean of the the differences is **algebraically equivalent** to the difference of the individual sample means, that is

$$\bar{D} = \bar{Y}_1 - \bar{Y}_2,$$

so that the calculation may be done either way. The **standard error** for the sample mean $\bar{D}$ is, by analogy to the single sample case,

$$s_{\bar{D}} = \sqrt{\frac{s_D}{n}} = \frac{s_D}{\sqrt{n}}.$$

Note that we have used the **same** notation, $s_{\bar{D}}$, as we did in the case of two independent samples, but the calculation and interpretation are different here.

We thus have

Test statistic:

$$t = \frac{\bar{D} - \delta}{s_{\bar{D}}}$$

REMARK: Note that the denominator of our test statistic depends on s_D, the sample standard deviation of the **differences**. This shows formally the important point – the relevant variation for comparing differences is that **within pairs** – this sample standard deviation is measuring precisely this quantity, the variability among differences on pairs!

Procedure: For level of significance α,

Reject H_0 if

one-sided: $t > t_{n-1,\alpha}$

two-sided: $|t| > t_{n-1,\alpha/2}$

EXAMPLE: Suppose the following are the data from an experiment conducted on $n = 12$ men, where before and after measurements were obtained on each:

After (Y_1)	Before (Y_2)	Difference $D = (Y_1 - Y_2)$
128	120	8
131	124	7
131	130	1
127	118	9
132	140	-8
125	128	-3
141	140	1
137	135	2
118	126	-8
132	130	2
129	126	3
135	127	8

We wish to test the one-sided hypotheses $H_0 : \mu_1 - \mu_2 = 0$ vs. $H_1 : \mu_1 - \mu_2 > 0$. We will use $\alpha = 0.05$. We obtain

$$\bar{D} = \frac{22}{12} = 1.833, \quad s_D^2 = 33.97, \quad , s_{\bar{D}} = \sqrt{\frac{33.97}{12}} = 1.68.$$

$$t = \frac{1.833}{1.68} = 1.09.$$

Because 1.09 does not exceed $t_{11,0.05} = 1.796$, we **do not reject** H_0. The evidence in these data is not strong enough to suggest that the stimulus raises SBP in middle-aged men. (Alternatively, $0.10 < P(t_{11} > 1.09) < 0.15$, so this probability is not less than 0.05.)

REMARK: The device of pairing observations is a special case of an experimental design principle called **blocking**. This idea is an important foundation of experiments in biology, agriculture, engineering, and other applications, as we will discuss later in the course. The basic idea: **limit** the effect of a potential source of variation so that real differences, if they exist, will be more likely to be detected.

Here is another example. Consider the tomato plant example. Although the gardener used a reasonable randomization scheme, one might want to consider that plants **close together** might tend to be **more alike** in terms of yield because they are more likely to receive the same amount of water, sunlight, and so on. These considerations might be used as a basis for pairing of plants (experimental units): suppose there were 12 plants, 6 to be allocated to each fertilizer. The gardener could have instead randomly assigned fertilizers A and B **within** each pair or **block** of 2 adjacent plants. The arrangement might look like:

(B A) (B A) (A B) (B A) (A B) (B A)

with parentheses indicating a pair or block. Variability among all plants would be eliminated as a source of variation. The only relevant source of variation to be taken into account for the purposes of comparing the fertilizers would be that **within** pairs of plants. We discuss this fundamental idea more in later chapters.

REMARK: In the blood pressure example, the notion of a "pair" of "experimental" units is a bit abstract. In the plant example, each pair consists of two **physically distinct** plants (experimental units), while in the blood pressure example, there is only one man per "pair." However, if we think abstractly, we may think of a "pair" – the man **before** and the man **after** the stimulus. The "experimental units" in the pair are these two abstract "men." We will discuss the notion of an experimental unit more precisely in chapter 6.

5.8 Linear additive model

In chapter 2, we introduced a **model** for explaining an observation on a random variable Y as the mean μ of the population of Y plus an **additive error** representing unexplained variability inherent in the observation that makes it different from μ. In particular, we wrote

$$Y_i = \mu + \epsilon_i.$$

This type of model is a useful representation that may be used as a framework for understanding **sources of variation**. We have just discussed a different **design**, that of the **pairing** of experimental units (observations), as an alternative to the design in which observations are obtained from 2 independent samples. Such a model formalizes the idea that the paired design eliminates the effect of variation across pairs of experimental units on the hypothesis test.

To see this, consider a model for the case where an experiment is conducted with 2 independent samples, one from each of 2 populations of interest. An observation from such an experiment is Y_{ij}, where as before we mean the observation from the jth experimental unit from the ith sample (population). We may think of Y_{ij} in terms of the following linear additive model:

$$Y_{ij} = \mu + \tau_i + \epsilon_{ij}.$$

In this model, μ may be thought of as the "overall mean;" that is, the mean response we would get **before** the treatments were applied. τ_i may be thought of as a "treatment effect;" that is, the **change** in mean that results from applying treatment i. Thus,

$$\mu_i = \mu + \tau_i \text{ for } i = 1, 2.$$

The **error** ϵ_{ij} represents everything else – sampling, biological variation – anything else unexplained that makes Y_{ij} differ from the mean for its population, μ_i. Thus, this includes **all** variation among experimental units, from all possible sources.

For this model, if we let

$$\bar{\epsilon}_i = \frac{1}{n_i}\sum_{j=1}^{n}\epsilon_{ij},$$

the average of the errors for treatment i, then it may be shown algebraically that the sample variance for treatment i,

$$s_i^2 = \frac{1}{n_i - 1}\sum_{j=1}^{n}(Y_{ij} - \bar{Y}_i)^2 = \frac{1}{n_i - 1}\sum_{j=1}^{n}(\epsilon_{ij} - \bar{\epsilon}_i)^2;$$

that is, the sample variance of the Y_{ij} is equal to that of the errors ϵ_{ij} for treatment i! This shows explicitly that the sample variance for a sample, hence the denominator of the usual test statistic, $s_{\bar{D}}$, is measuring the unexplained variability in the data, hence, the term "error sum of squares!" Here, then, it measures **all** variation among with experimental units, without regard to possible different components of this variation.

Now consider an appropriate model for the case where an experiment is conducted according to the **paired** design. If we use this design, we have a legitimate basis for **pairing** experimental units because they may be "alike" in some way. Thus, we may think of **two ways** in which experimental units may vary – **by pairs** and **within pairs**. An observation from such an experiment is again Y_{ij}, where now the subscripts represent the observation for treatment i from the jth **pair**. The model is

$$Y_{ij} = \mu + \tau_i + \rho_j + \epsilon_{ij}.$$

In this model, the interpretations of μ and τ_i are the same as above. The difference is the addition of the term ρ_j. This term may be thought of as the "effect" of being in the jth pair; that is, observations on a particular pair j differ from the mean for the treatment in question by an amount unique to that pair, ρ_j. Note that the two observations Y_{1j} and Y_{2j} on the jth pair **share** this component, as they should (both observations are on the same pair, so should change in the **same** way). The variation among the ρ_j (how they differ) characterizes that **across** pairs of experimental units. Again, the term ϵ_{ij} represents all other unexplained sources of variation. Here, what is left over as unexplained is the variation **within** pairs of experimental units, so it is this variation that the ϵ_{ij} represent.

Comparing the two models, we see now the difference. In the 2-independent-sample model, there is no term ρ_j, because there is no **link** between observations in each sample (they are all independent). The design and its model do not attempt to distinguish different ways in which experimental units may vary. By pairing when appropriate, we are "explaining" more of the variation by something we can identify (the pairing), leaving less "unexplained" (and out of our control) in the "error" term ϵ_{ij}.

Finally, note that for the paired comparison model, if we consider the difference for the jth pair, we get

$$D_j = Y_{1j} - Y_{2j} = (\tau_1 - \tau_2) + (\epsilon_{1j} - \epsilon_{2j}).$$

The effect of ρ_j, which represents variation **across** pairs of experimental units, **disappears**! The only "error" that is left is that in the ϵ_{ij}, which represents variation **within** pairs. This illustrates formally our point in the remark on p. 107 – the relevant variation when analyzing differences from a paired comparison experiment is that **within** pairs. When we compute s_D, the sample standard deviation of the differences, the effect of variation **among** pairs does not play a role and hence is **eliminated** from consideration, leading to **more precise** inferences!

This can be used to advantage in setting up experiments. From this discussion, it is apparent that, if we use a **paired** design, the variability among pairs will not play a role in our ability to detect treatment differences. Thus, we may include in the experiment pairs of **different types**, thereby **widening** the **scope of inference** of the experiment. For example, in the blood pressure study, men of different ethnic groups, ages, and so on could be included. We will discuss this idea more formally later.

As you can see, although the notation is a bit cumbersome, once mastered, it can be quite helpful for understanding why certain experimental designs are preferred over others! We will use these models quite a bit in our later discussion.

5.9 Power, sample size, and detection of differences

We now return to the notion of incorrect inferences in hypothesis tests.

Recall that there are two types of "mistakes" we might make when conducting such a test:

Type I error reject H_0 when it really is true

Type II error do not reject H_0 when it really isn't true

Recall that, because we conduct the test under the assumption that H_0 is true, the probability of rejecting H_0 when it is true is exactly equal to the level of significance α chosen for the test; that is

$$P(\text{Type I error}) = \alpha.$$

We define

$$\beta = P(\text{Type II error}).$$

Because both Type I and II errors are mistakes, we would ideally like both α and β to be **small**.

Because Type I error is often more serious (this, of course, depends on the particular application), the usual approach is to fix the Type I error (i.e. that level of significance α) first. So, far, we have not discussed the implications of Type II error and how it might be taken into account when setting up an experiment.

POWER OF A TEST: If we **do not** commit a Type II error, then we

reject H_0 when H_0 is **not** true, i.e. infer H_1 is true when H_1 **really is** true.

Thus, if we **do not** commit a Type II error, we have made a correct judgment; moreover, we have done precisely what we hoped to do – **detect a departure from** H_0 when in fact such a **departure** (difference) **really does exist**!

We call

$$1 - \beta = P(\text{reject } H_0 \text{ when } H_0 \text{ is } \textbf{not} \text{ true})$$

the **power** of the hypothesis test.

Clearly, **high power** is a desirable property for a test to have:

low probability of Type II error $\Leftrightarrow$ **high power**

$\Leftrightarrow$ high probability to detect a difference if one exists.

INTUITION: **Power** of a test is a function of **how different** the true value of μ or $\mu_1 - \mu_2$ is from the null hypothesis value μ_0 or δ. If the difference between the true value and the null hypothesis value is **small**, we might not be too successful at detecting this. If the difference is **large**, we are apt to be more successful.

We can be more formal about this. To illustrate, we will consider the simple case of testing hypotheses about the value of the mean of a single population, μ. The same principles hold for tests on differences of means and in fact **any** test.

Consider in particular the **one-sided** test

$$H_0 : \mu = \mu_0 \text{ vs. } H_1 : \mu > \mu_0.$$

Recall that the test statistic is

$$\frac{\bar{Y} - \mu_0}{s_{\bar{Y}}}.$$

This test statistic is based on the idea that a **large** observed value of $\bar{Y}$ is evidence that the true value of μ is **greater** than μ_0. We reject H_0 when

$$\frac{\bar{Y} - \mu_0}{s_{\bar{Y}}} > t_{n-1,\alpha}.$$

To simplify our discussion, let us assume that we **know** σ^2, the variance of the population, and hence we **know** $\sigma_{\bar{Y}}$. Under these conditions, we would replace the usual t statistic above by the statistic with $\sigma_{\bar{Y}}$ in place of $s_{\bar{Y}}$. The statistic would of course be

$$\frac{\bar{Y} - \mu_0}{\sigma_{\bar{Y}}} \sim \mathcal{N}(0, 1) \tag{5.5}$$

if H_0 is true. Thus, in this situation, rather than compare the statistic to the t distribution, we would compare it to the **standard normal** distribution.

Let z_α denote the value satisfying

$$P(Z > z_\alpha) = \alpha,$$

for a standard normal r.v. Z. Then we would conduct the test at level of significance α by rejecting H_0 when

$$\frac{\bar{Y} - \mu_0}{\sigma_{\bar{Y}}} > z_\alpha.$$

Now if H_0 is **not true**, then it must be that $\mu \neq \mu_0$ but, instead, $\mu =$ some **other value**, say $\mu_1 > \mu_0$. Under these conditions, in reality, the statement in (5.5) is **not true**. Instead, the statistic that really has a standard normal distribution is

$$\frac{\bar{Y} - \mu_1}{\sigma_{\bar{Y}}} \sim \mathcal{N}(0,1). \tag{5.6}$$

What we would like to do is evaluate **power**; thus, we need probabilities under these conditions (when H_0 is not true).

Thus, consider the **power** of the test when H_0 is not true because $\mu = \mu_1 > \mu_0$. We conduct the test under the assumption (5.5), rejecting when

$$\frac{\bar{Y} - \mu_0}{\sigma_{\bar{Y}}} > z_\alpha.$$

Thus, the power of the test is

$$\begin{aligned} 1-\beta &= P(\text{reject } H_0 \text{ when } \mu = \mu_1) \\ &= P\left(\frac{\bar{Y} - \mu_0}{\sigma_{\bar{Y}}} > z_\alpha\right) \\ &= P\left(\frac{\bar{Y} - \mu_1}{\sigma_{\bar{Y}}} + \frac{\mu_1 - \mu_0}{\sigma_{\bar{Y}}} > z_\alpha\right), \end{aligned}$$

where this last expression is obtained by adding and subtracting the quantity $\mu_1/\sigma_{\bar{Y}}$ to the left hand side of the inequality.

Rearranging, we get

$$1 - \beta = P\left(\frac{\bar{Y} - \mu_1}{\sigma_{\bar{Y}}} > z_\alpha - \frac{\mu_1 - \mu_0}{\sigma_{\bar{Y}}}\right).$$

Now, under what is **really** going on, the quantity on the left hand side of the inequality in this probability statement is a standard normal r.v., as noted in (5.6). Thus, the probability $(1-\beta)$ is the probability that a standard normal r.v. Z **exceeds** the value

$$z_\alpha - \frac{\mu_1 - \mu_0}{\sigma_{\bar{Y}}}$$

$$\text{i.e. } 1 - \beta = P\left(Z > z_\alpha - \frac{\mu_1 - \mu_0}{\sigma_{\bar{Y}}}\right).$$

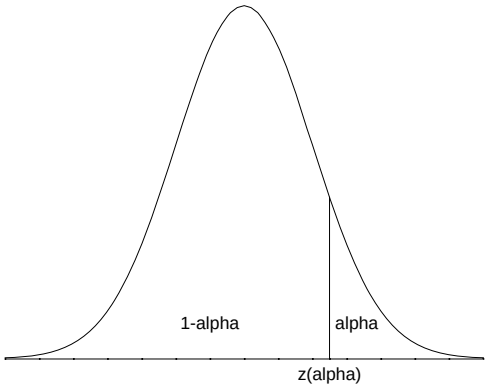
That is, **power** is a function of μ_1, what is **really** going on!

As an illustration, consider the following. Let

$$\Delta = \frac{\mu_1 - \mu_0}{\sigma_{\bar{Y}}}.$$

If H_0 is really true, then $\mu_1 = \mu_0$, and we have

$$P(\text{reject } H_0) = P(Z > z_\alpha) = \alpha,$$



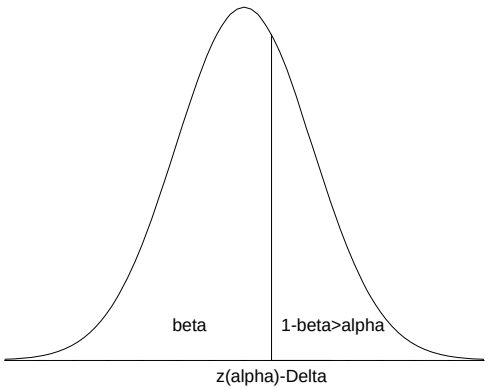
If H_0 is **not** true, then we have just shown that

$$P(\text{reject } H_0, H_0 \text{ not true }, \mu = \mu_1) = 1 - \beta = P(Z = z_\alpha - \Delta).$$

Note that

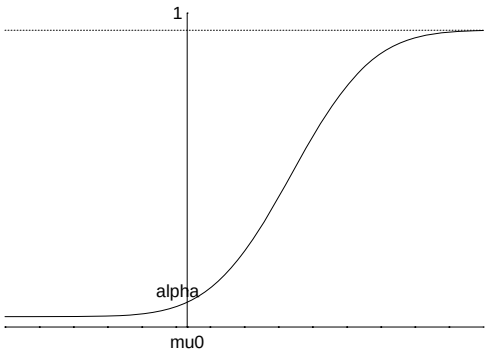
$$\Delta = \frac{\mu_1 - \mu_0}{\sigma_{\bar{Y}}} > 0 \quad \Rightarrow \quad z_\alpha > z_\alpha - \Delta.$$

Thus



Comparing the two pictures, we see that the **farther** μ_1, the true value of μ, is from the hypothesized value μ_0, the **larger** the power $1 - \beta$ will be (and the smaller β will be).

We can in fact **graph** the power $1-\beta$ for the test as a function of possible values μ_1. Remember when $\mu_1 = \mu_0$, so that H_0 is true, the probability we reject H_0 is α. Thus, we would expect the power to equal α under this scenario. The graph looks like



Note that as μ_1 gets greater and greater than μ_0, power **increases**, as we noted above. When the true value of μ is actually **less than** the null value μ_0, the power gets very small – because the test is **not set up** to detect differences in the other direction, it is very bad at doing so!

RESULTS:

- Our ability to detect a difference (departure) from $\mu = \mu_0$ increases depending on how big the difference is (how far μ_1 is from μ_0).
- Thus, if the true value of μ is **much larger** than the hypothesized value μ_0, we have a good chance (how good is measured by power) of detecting this, and a low probability of committing a Type II error.
- If the true value of μ is different from μ_0, but not by very much, we will not have a good chance of detecting this (again measured by power), and we will have a high probability of making a Type II error.

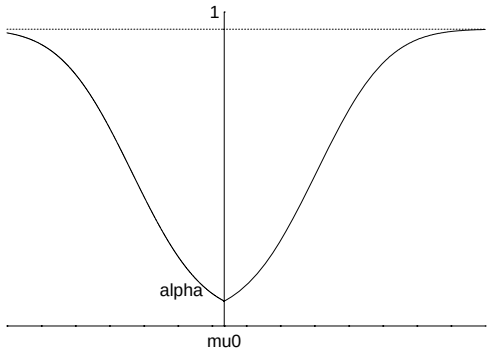
We can make the same sort of argument for a two-sided test

$$H_0 : \mu = \mu_0 \text{ vs. } H_1 : \mu \neq \mu_0.$$

The calculations are harder, but the same principles apply. We get

$$\begin{aligned} 1-\beta &= P\left(\left|\frac{\bar{Y}-\mu_0}{\sigma_{\bar{Y}}}\right| > z_{\alpha/2}\right) \\ &= P\left(\frac{\bar{Y}-\mu_0}{\sigma_{\bar{Y}}} > z_{\alpha/2}\right) + P\left(\frac{\bar{Y}-\mu_0}{\sigma_{\bar{Y}}} < -z_{\alpha/2}\right) \\ &= P\left(\frac{\bar{Y}-\mu_1}{\sigma_{\bar{Y}}} > z_{\alpha/2} - \frac{\mu_1-\mu_0}{\sigma_{\bar{Y}}}\right) + P\left(\frac{\bar{Y}-\mu_1}{\sigma_{\bar{Y}}} < -z_{\alpha/2} - \frac{\mu_1-\mu_0}{\sigma_{\bar{Y}}}\right) \end{aligned}$$

Here, we must concern ourselves with whether $\mu_1 > \mu_0$ or $\mu_1 < \mu_0$. It turns out that a graph of the power $1-\beta$ as a function of μ_1 looks like



Note here that power increases if μ_1 is far from μ_0 in **either direction**. The test is designed to detect departures of either type.

We may pursue a similar argument for any hypothesis test, although the details get harder. If we consider a test about the difference of two means,

$$H_0 : \mu_1 - \mu_2 = \delta \text{ vs. } \mu_1 - \mu_2 > \delta,$$

then δ is the difference under the null hypothesis. If δ_1, say, is the **true** difference, then it should be clear that power will be a function of δ_1; in particular how far δ_1 is from δ.

Even though we conducted these arguments for an "idealized" situation where σ^2 is **known**, the same principles apply (although the details are harder) for the case where it is not and we use the usual t statistics.

THE EFFECT OF SAMPLE SIZE: We have seen that power measures our ability to detect a departure from H_0 if there **really is one**. Is there a way we can increase our ability to detect such departures, particularly when they may be **small**?

Again consider the one-sample case for illustration; the same principles apply in general. Note that the test statistic depends on the **sample size**, n, through the denominator (e.g. $\sigma_{\bar{Y}}$ in the "ideal" case or $s_{\bar{Y}}$ in the "real" case). Recall that as n **increases**, $\sigma_{\bar{Y}}$ (or $s_{\bar{Y}}$) gets **smaller** like $1/\sqrt{n}$.

Thus, if n gets bigger, it follows that

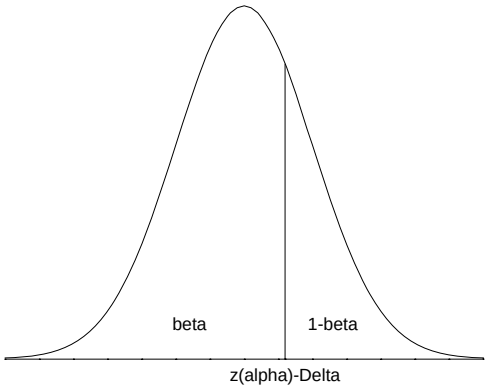
$$\Delta = \frac{\mu_1 - \mu_0}{\sigma_{\bar{Y}}}$$

gets **larger**, because, its denominator gets **smaller**. This implies that

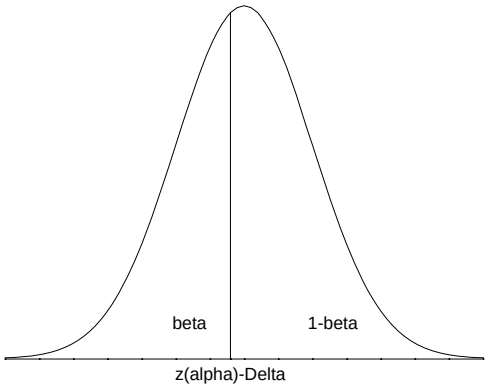
$$z_\alpha - \Delta = z_\alpha - \frac{\mu_1 - \mu_0}{\sigma_{\bar{Y}}}$$

gets **smaller** with **increasing** n! This suggests that, the larger n is, the **greater** the area to the right of $z_\alpha - \Delta$, so the higher the power!

Here are pictures to illustrate. With the same μ_0, μ_1, and α (so the only thing changing is sample size),



larger $n \Downarrow$



RESULT: **Larger** sample size **increases** power, and thus the ability to detect small departures.

- If it is important to detect a small departure, then increase the sample size!

REMARK: How meaningful detection of a **small** departure from H_0 is depends on the application; this may involve biological/ethical/resource considerations. For example, unless a new, more expensive fertilizer produces a substantial increase in mean yield over another, the additional expense may not be warranted. In this situation, only a rather large departure from H_0 might be meaningful. On the other hand, if a new drug for a disease can be shown to produce even a **small** improvement in terrible symptoms or even survival, then, ethically, we would want to know and switch to the new drug. In this case, detection of a small departure is required.

5.10 Balancing α and β and sample size determination

The theoretical results of the previous section show that sample size helps to determine power. We also discussed the idea of a "meaningful" (in a scientific sense) difference to be detected.

This suggests that, if, for a given application, we can state what we believe to be a **scientifically meaningful** difference, we might be able to determine the appropriate sample size to ensure **high power** to detect this difference. This is exactly the idea behind standard procedures for choosing sample size!

In particular, recall that when we spelled out the steps one should go through in setting up and carrying out an experiment, we noted that we should choose the level of significance α based on the severity of committing a Type I error. We now see that, once α has been set, we might also like to choose β, and thus the **power** $1-\beta$ of detecting a meaningful difference at this level of significance.

It follows that, once we have determined

- The level of significance, α
- A scientifically meaningful departure from H_0
- Power, $1-\beta$, with which we would like to be able to detect such a difference

we would like to determine the appropriate sample size to achieve these objectives. For example, we might want $\alpha = 0.05$ and an 80% chance of detecting a particular difference of interest. That is, $1-\beta = 0.80$, or $\beta = 0.20$. Here, the probability of committing a Type II error is 0.20 (less serious than a Type I error, for which the probability has been set at 0.05).

It turns out that the theoretical arguments in the last section form the basis for appropriate formulæ for computing the sample size given α, β, and some notion of a "meaningful difference." We will give these results for the case of a comparison between 2 population means, either using 2 independent samples or using a paired design.

Let z_α be the value such that $P(Z > z_\alpha) = \alpha$ for a standard normal r.v. Z.

Procedure: For a test between 2 population means μ_1 and μ_2, choose the sample size so that

$$\text{one-sided test } n = \frac{(z_\alpha + z_\beta)^2 \zeta_D}{\text{diff}^2}$$
$$\text{two-sided test } n = \frac{(z_{\alpha/2} + z_\beta)^2 \zeta_D}{\text{diff}^2}.$$

Here, diff is the meaningful difference we would like to detect, and, depending on the type of design

2 independent samples	$\zeta_D = 2\sigma^2$
Paired comparison	$\zeta_D = \sigma_D^2$.

Here, σ^2 is the (assumed common) variance for the 2 populations and σ_D^2 is the true variance of the population of differences D_j.

For the 2 independent samples case, the value of n obtained is the number of experimental units needed in **each** sample. For the paired comparison case, n is the total number of experimental units (each will be seen twice).

SLIGHT PROBLEM: We usually do not know σ_D^2 or σ^2. Some practical solutions are as follows:

- Rather than express the "meaningful difference," diff, in terms of the actual units of the response (e.g. diff = 5 mg for the rat experiment), express it in units of the standard deviation of the appropriate response.

  For example, for a test based on 2 independent samples, we might state that we wish to detect a difference the size of one standard deviation of the response. We would thus take diff $= \sigma$, so that the factor

  $$\frac{\zeta_D}{\text{diff}^2} = \frac{2\sigma^2}{\sigma^2} = 2.$$

  For a paired comparison, we might specify the difference to be detected in terms of standard deviation of a response difference D. If we wanted a one standard deviation difference, we would take diff $= \sigma_D$.

- Another approach is to substitute estimates for σ_D^2 or σ^2 from previous studies.

ANOTHER SLIGHT PROBLEM: The tests we actually carry out in practice are based on the t distribution, because we don't know σ^2 or σ_D^2 but rather estimate them from the data. In the procedures above, however, we used the values z_α, z_α, and z_β from the standard normal distribution. There are several things one can do:

- Nothing formal. The sample sizes calculated this way are only rough guidelines, because so much approximation is involved; e.g. having to estimate or "guess at" σ^2 or σ_D^2, assume the data are normal, and so on. Thus, one might regard the calculated sample size as a conservative choice, and use a slightly bigger sample size in real application.

- Along these lines, theoretical calculations may be used to adjust for the fact that the tests are based on the t rather than standard normal distribution. Specifically, one may calculate an appropriate "correction factor" to inflate the sample size slightly.

  (i) Use the appropriate formula to get n

  (ii Multiply n by the **correction factor**

$$\frac{2n+1}{2n-1} \text{ for 2 independent samples}$$

$$\frac{n+2}{n} \text{ for a paired design.}$$

REMARK: Calculation of an appropriate sample size, at least as a rough guideline, should always be undertaken. There is no point in spending resources (time and money) to do an experiment that has very little chance of detecting the scientifically meaningful difference in which one is interested! This should always be carried out **in advance** of performing an experiment – knowing the sample size was too small after the fact is not very helpful!

EXAMPLE: Consider the pig weight gain example first introduced in chapter 4. Recall that there were 2 independent samples of pigs, $n = 12$ in each sample. **In advance**, suppose the investigators stated that they wished to perform a two-sided test to detect a mean weight gain difference between the rations of 1 standard deviation of the response (weight gain in lbs), σ. Thus, diff $= \sigma$, and hence $\zeta_D = 2\sigma^2/\sigma^2 = 2$. Suppose that they also stated that they would like to use a level of significance $\alpha = 0.05$ and have 60% power to detect the difference of interest (so $\beta = 0.40$).

We have $z_{0.025} = 1.96$, $z_{0.40} \approx 0.25$, so that

$$n = 2(z_{0.025} + z_{0.40})^2 = 9.77 \approx 10.$$

The "correction factor" is

$$\frac{2n+1}{2n-1} == \frac{21}{19} = 1.105,$$

which gives the revised sample size $10 \times 1.105 = 11.05 \approx 12$, rounding up to the next largest integer.

REMARK: A very important job for an investigator contemplating an experiment is to determine what he or she would like to detect; that is, what is an important, scientifically relevant difference. Without this subject-matter know-how involved, statistics can't really help!

5.11 Using SAS to perform hypothesis tests about the difference of two population means

Here, we show by means of two additional examples, how to use the SAS procedures `PROC TTEST` and `PROC MEANS` to perform tests about a difference of means. `PROC TTEST` is appropriate for the situation where the data arise from 2 **independent** samples. `PROC MEANS` is used when the data were collected according to a **paired design**.

EXAMPLE 1 – 2 INDEPENDENT SAMPLES: We wish to test

$$H_0 : \mu_1 - \mu_2 = 0 \text{ vs. } H_1 : \mu_1 - \mu_2 \neq 0 \text{ or } > 0.$$

The data we use are from STD, problem 5.5.5. From an area planted with one variety of guayule, 27 plants were selected at random. Of these, 15 were offtypes (O) and 12 were aberrants (A). The question of interest was whether or not there is a difference in mean rubber percentage for the 2 types of plants (a two-sided test). For the 27 plants, rubber percentages were as follows:

```
O:      6.21  5.70  6.04  4.47  5.22  4.45  4.84  5.88  5.82  6.09  5.59  6.06
        5.59  6.74  5.55

A:      4.28  7.71  6.48  7.71  7.37  7.20  7.06  6.40  8.93  5.91  5.51  6.36
```

PROC TTEST computes the t statistic for testing the difference both under the assumption of equal variances ($\sigma_1^2 = \sigma_2^2$) and that of unequal variances, and provides the value of the statistic. It also gives the probability of seeing a t r.v. as large or larger than the one seen (i.e. the p-value). One must **be careful** – PROC TTEST gives p-values for a **two-sided** test only, that is it gives $P(|t| > t$ we saw). To perform a test using the output, we may thus either

- Compare the value of the statistic in the output to the appropriate critical value for our α as if we had calculated it by hand (one- or two-sided)
- For a **two-sided** test, compare the p-value in the output to our α level directly. For a **one-sided** test, compare **one-half** the probability in the output to α. This follows because probability given in the output $= P(|t| > t$ we saw$) = 2P(t > t$ we saw) by symmetry, as long as the t value we saw is positive. If we set up the test so that the population with the higher mean is population 1, this should be okay – under these circumstances, seeing a negative statistic would mean that the null hypothesis would definitely not be rejected, as we're looking for large, positive values.

PROC TTEST also gives the value of F for testing whether the 2 population variances are the same. For this test, recall that we usually compare the observed value of the test statistic to $F_{n_1-1,n_2-1,\alpha/2}$, and reject if F is larger at level α. PROC TTEST gives $2 \times P(F >$ F value we saw), so we compare the probability in the output directly to α, and reject H_0 is the probability is smaller than α. We may thus conduct the test two ways:

- Compare the reported value F to the appropriate critical value from the table as if we had done the calculations by hand.
- Compare the probability given in the output directly to α.

PROC TTEST is **not** appropriate for conducting tests for **paired designs**.

NOTES:

- `DATA` step. There are 2 variables, `TYPE` specifying treatment group and `RUBPER` giving the response. `TYPE` is a **character variable**; it takes on the values "`O`" and "`A`;" the "`$`" after `TYPE` in the `INPUT` statement tells SAS that this is a character variable. The "@@" tells SAS that the data are being entered in one long string rather than line-by-line.

- The output gives a summary of the data, including sample means $(\bar{Y}_1, \bar{Y}_2)$, sample variances (s_1^2, s_2^2), and standard errors for the mean of each sample

$$\left(\sqrt{\frac{s_1^2}{n_1}}, \sqrt{\frac{s_2^2}{n_2}} \right).$$

  The column `T` gives the value of the test statistic computed under the assumptions of unequal and equal variance. The column `DF` gives the appropriate degrees of freedom (*edf* or $n_1 + n_2 - 2$). `Prob>|T|` gives the probability we compare to α in each case for a two-sided test. The bottom line gives the result of the F test for equality of variance. Here, $F = 3.52$. From the F table, $F_{11,14,0.025} \approx 3.10$; because $3.52 > 3.10$, we reject the hypothesis that the variances are the same at level 0.05. Alternatively, the probability given, $2 \times P(F >$ F we saw$) = 0.0297 < 0.05$, so we reject. Given that there is evidence to suggest the variances are not equal, we may prefer to use the unequal variances t test, with $t = 2.9239$ and $edf = 15.9 \approx 16$. From the t table, $t_{16,0.025} = 2.120 < 2.9239$, so we reject H_0. Alternatively, the output gives $P(|t_{16}| > 2.9329 = 0.0100 < 0.05$.

- If we had instead wanted to conduct a one-sided test, we would compare $= 2.9329$ to $t_{16,0.05} = 1.746$ from the t table; alternatively, we would compare the probability $(0.0100)/2 = 0.005$ to $\alpha = 0.05$.

PROGRAM:

```
*******************************************************;
*                                                     ;
*              ST 511   EXAMPLE 5.1                   ;
*     PROBLEM 5.5.5 OF STEEL, TORRIE & DICKEY         ;
*                                                     ;
*    USING PROC TTEST OF SAS TO PERFORM A TEST OF     ;
*    HYPOTHESIS THAT THE MEANS OF 2 INDEPENDENT       ;
*    SAMPLES ARE DIFFERENT.  THE PROCEDURE PERFORMS   ;
*    THE USUAL T-TEST,, BOTH UNDER THE ASSUMPTION OF  ;
*    EQUAL VARIANCES AND UNEQUAL VARIANCES.  THE      ;
```

```
*    PROCEDURE ALSO PERFORMS THE F-TEST FOR THE       ;
*    EQUALITY OF THE VARIANCES.                       ;
*                                                     ;
*    HERE, THE RESPONSES ARE "RUBBER PERCENTAGES,"    ;
*    EITHER FROM "OFFTYPES" OR "ABERRANTS."  THUS,    ;
*    WE DEFINE THE CLASS VARIABLE "TYPE" WITH 2       ;
*    VALUES "O" OR "A" REPRESENTING THE 2 TYPES, AND  ;
*    GIVE THE ASSOCIATED RUBBER PERCENTAGE "RUBPER"   ;
*    FOR EACH IN THE DATA STEP TO CREATE THE DATA SET ;
*    "TREES."                                         ;
*                                                     ;
*****************************************************;

*****************************************************;
*                                                     ;
*   INVOKE SAS OPTIONS STATEMENT TO FORMAT OUTPUT     ;
*   HERE, WE LIMIT THE WIDTH OF THE OUTPUT TO 80      ;
*   CHARACTERS AND THE LENGTH OF THE PAGE TO 59       ;
*   LINES                                             ;
*                                                     ;
*****************************************************;

OPTIONS LS=80 PS=59 NODATE;

*****************************************************;
*                                                     ;
*    THE "$" INDICATES THAT "TYPE" IS A CHARACTER     ;
*    VARIABLE TAKING ON THE CHARACTER VALUES "O" OR   ;
*    "A".  THE "@@" MEANS THAT INSTEAD OF GIVING EACH ;
*    OBSERVATION ON A SEPARATE LINE, WE ARE STRINGING ;
*    THEM ALONG TO SAVE SPACE AS BELOW:               ;
*                                                     ;
*****************************************************;
*;
```

```
DATA TREES;
    INPUT TYPE $ RUBPER @@;
    CARDS;
O 6.21 O 5.70 O 6.04 O 4.47 O 5.22 O 4.45 O 4.84 O 5.88 O 5.82
O 6.09 O 5.59 O 6.06 O 5.59 O 6.74 O 5.55
A 4.28 A 7.71 A 6.48 A 7.71 A 7.37 A 7.20 A 7.06 A 6.40 A 8.93
A 5.91 A 5.51 A 6.36
;
*;
PROC PRINT DATA=TREES;
  TITLE 'RUBBER TREE DATA -- PROBLEM 5.5.5 OF STEEL, TORRIE & DICKEY'; RUN;
*;
PROC TTEST DATA=TREES;
*;
*****************************************************;
*                                                   ;
*   THE "CLASS" STATEMENT TELLS SAS THAT VARIABLE   ;
*   "TYPE" IS NOT NUMERIC BUT INSTEAD DESCRIBES THE ;
*   GROUPS (O OR A HERE) FROM WHICH THE OBSERVATIONS ;
*   ARISE.  THE "VAR" STATEMENT TELLS SAS TO PERFORM ;
*   THE HYPOTHESIS TEST USING THE VARIABLE "RUBPER" ;
*   AS THE OBSERVATIONS.                            ;
*                                                   ;
*****************************************************;
*;
    CLASS TYPE;
    VAR RUBPER;
    TITLE 'HYPOTHESIS TEST FOR THE DIFFERENCE IN MEAN RUBBER';
    TITLE2 'PERCENTAGE -- PROBLEM 5.5.5 OF STEEL, TORRIE & DICKEY'; RUN;
```

OUTPUT:

%%%

RUBBER TREE DATA -- PROBLEM 5.5.5 OF STEEL, TORRIE & DICKEY 1

OBS	TYPE	RUBPER
1	0	6.21
2	0	5.70
3	0	6.04
4	0	4.47
5	0	5.22
6	0	4.45
7	0	4.84
8	0	5.88
9	0	5.82
10	0	6.09
11	0	5.59
12	0	6.06
13	0	5.59
14	0	6.74
15	0	5.55
16	A	4.28
17	A	7.71
18	A	6.48
19	A	7.71
20	A	7.37
21	A	7.20
22	A	7.06
23	A	6.40
24	A	8.93
25	A	5.91
26	A	5.51
27	A	6.36

%%

HYPOTHESIS TEST FOR THE DIFFERENCE IN MEAN RUBBER 2
PERCENTAGE -- PROBLEM 5.5.5 OF STEEL, TORRIE & DICKEY

TTEST PROCEDURE

Variable: RUBPER

TYPE	N	Mean	Std Dev	Std Error	Minimum	Maximum
A	12	6.74333333	1.20490161	0.34782513	4.28000000	8.93000000
O	15	5.61666667	0.64219119	0.16581305	4.45000000	6.74000000

Variances	T	DF	Prob>|T|
Unequal	2.9239	15.9	0.0100
Equal	3.1193	25.0	0.0045

For H0: Variances are equal, F' = 3.52 DF = (11,14) Prob>F' = 0.0297

EXAMPLE 2 – PAIRED COMPARISON:We wish to test

$$H_0 : \mu_1 - \mu_2 = 0 \text{ vs. } H_1 : \mu_1 - \mu_2 \neq 0 \text{ or } > 0$$

where the experiment was conducted according to a paired design.

The data we use are from Dixon and Massey, 1969, p. 126. The data are from an experiment performed to determine whether pollinated hop plants produce higher yield than unpollinated ones. 7 hop plants were used; one half of each plant was pollinated and the other half was not, and yields were reported for each half as follows:

Pollinated:	0.78	0.76	0.43	0.92	0.86	0.59	0.68
Unpollinated:	0.21	0.12	0.32	0.29	0.30	0.20	0.14

The question of interest was to determine if pollination of hop plants results in higher mean yield. We thus wish to perform a one-sided test with pollinated plants identified as population 1.

We have already used `PROC MEANS` to summarize a set of data. You may have noticed that this procedure prints out a "`T`" value and an associated probability. We will now see how to exploit this to perform a paired comparison t test for a difference of means. Recall that in a paired comparison, we think of our data as the differences D_j from each pair and test if the population of differences has mean zero. The t statistic given by `PROC MEANS` for each variable to which it is applied is the appropriate statistic for testing whether the **single** population mean for that variable is 0. Thus, if we construct a data set containing the **differences** as a variable, and apply `PROC MEANS` to that variable, the resulting t test will be exactly the one we want!

The following SAS program does this for the hop plant data. Note the definition of the new variable `DIFF`.

NOTES:

- The `VAR` statement instructs SAS to apply the `PROC` only to the variable `DIFF`. We ask specifically for the sample mean ($\bar{D}$), the sample standard error for the mean ($s_{\bar{D}}$), the t statistic, and the probability of seeing $|t|$ this large. From the output, $t = 6.95$. Thus, for a test at level $\alpha = 0.05$, with $n = 7$, we have $t_{6,0.05} = 1.943$. $6.95 > 1.943$, so we reject H_0. Alternatively, the output gives

  $$P(|t_6| > 6.95) = 0.0004.$$

  Because we want a one-sided test, we need $P(t_6 > 6.95)$. By symmetry, this is $0.0004/2 = 0.0002$. Since $0.00002 < 0.05$, we reject H_0.

- This shows that we must be careful when using this procedure to conduct tests. The output is geared toward **two-sided** tests, so the probability printed is one about $|t|$.. When we perform one-sided tests, because we want a test based on t alone. Don't forget to do as in the example, and take half the printed probability before comparing to α for a one-sided test!

PROGRAM:

```
/*-----------------------------------------------------
|                                                     ;
|                  ST 511   EXAMPLE 5.2               ;
|       A PAIRED COMPARISON HYPOTHESIS TEST USING     ;
|       THE HOP PLANT EXAMPLE OF DIXON & MASSEY,      ;
|       P. 126, PROBLEM 19.                           ;
```

```
|                                                    ;
|       THE DATA ARE FROM AN EXPERIMENT PERFORMED     ;
|       ON 7 HOP PLANTS.  ONE HALF OF EACH PLANT WAS  ;
|       POLLINATED, THE OTHER HALF WAS NOT.           ;
|       THE YIELD OF THE SEED OF EACH HOP PLANT       ;
|       FROM THE POLLINATED AND UNPOLLINATED HALVES   ;
|       WERE TABULATED.  THUS, THE OBSERVATIONS ARE   ;
|       THE YIELDS FROM THE PLANTS -- FOR EACH PLANT  ;
|       WE HAVE A PAIR OF OBSERVATIONS -- POLLINATED  ;
|       AND UNPOLLINATED YIELD.                       ;
|                                                    ;
|       CREATE A SAS DATA SET CONTAINING THE          ;
|       POLLINATED YIELD "POLL" AND THE UNPOL-        ;
|       LINATED YIELD "UNPOLL" FOR EACH OF THE 7      ;
|       PLANTS.  CREATE A NEW VARIABLE, "DIFF"        ;
|       THE DIFFERENCE IN THE YIELDS:                 ;
|                                                    ;
|               DIFF = POLL - UNPOLL                  ;
|                                                    ;
-------------------------------------------------------*/

OPTIONS LS=80 PS=59 NODATE;

DATA HOPS;
    INPUT POLL UNPOLL;
    DIFF=POLL-UNPOLL;
    CARDS;
0.78 0.21
0.76 0.12
0.43 0.32
0.92 0.29
0.86 0.30
0.59 0.20
0.68 0.14
```

```
;
PROC PRINT DATA=HOPS; RUN;
;
/*-----------------------------------------------------;
|                                                      ;
|       THE QUESTION OF INTEREST IS WHETHER OR         ;
|       NOT POLLINATION PRODUCES HIGHER YIELD.         ;
|       NOTE THAT WE HAVE DEFINED THE POPULATIONS      ;
|       SO THAT THE FIRST POPULATION (POLL) IS THE     ;
|       ONE WE SUSPECT HAS THE LARGER MEAN.            ;
|                                                      ;
|       TO CONDUCT THE TEST FOR THIS PAIRED COMPAR-    ;
|       ISON EXPERIMENT, WE USE PROC MEANS APPLIED     ;
|       TO THE VARIABLE "DIFF."  THE OPTIONS           ;
|       REQUESTED IN THE PROC MEANS STATEMENT ARE      ;
|       FOR THE MEAN AND THE STANDARD ERROR OF THE     ;
|       MEAN (MEAN, STDERR) AND FOR THE VALUE OF       ;
|       THE T STATISTIC (T) AND THE PROBABILITY OF     ;
|       SEEING IT: LARGER THAN WHAT WE SAW (PRT).      ;
|       WE COMPARE THIS PROBABILITY TO THE LEVEL       ;
|       OF SIGNIFICANCE AS DESCRIBED IN THE HANDOUT    ;
|                                                      ;
------------------------------------------------------*/
*;
PROC MEANS MEAN STDERR T PRT;
    VAR DIFF;
    TITLE 'PAIRED COMPARISON T TEST';
    TITLE2 'HOP PLANT DATA'; RUN;
```

OUTPUT:

```
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%

                                                                               1
```

OBS	POLL	UNPOLL	DIFF
1	0.78	0.21	0.57
2	0.76	0.12	0.64
3	0.43	0.32	0.11
4	0.92	0.29	0.63
5	0.86	0.30	0.56
6	0.59	0.20	0.39
7	0.68	0.14	0.54

%%%

PAIRED COMPARISON T TEST 2

HOP PLANT DATA

Analysis Variable : DIFF

Mean	Std Error	T	Prob>|T|
0.4914286	0.0707588	6.9451265	0.0004

6 Principles of Experimental Design

Complementary Reading: STD, Chapter 6

6.1 Introduction

In our introduction to the course in chapter 1, we touched briefly on some of the ideas underlying **experimental design**. Now that we have developed a statistical framework for characterizing data, variation, and so on, we are in a position to return to these issues and be a bit more formal.

WHAT IS AN EXPERIMENT? An **experiment** is an investigation set up to provide answers to a question or questions of interest.

- In our context, an experiment is most likely to involve a **comparison of treatments** (e.g. fertilizers, drugs, rations, methods, varieties, etc.).
- The outcome of carrying out the experiment is **information** in the form of **observations** on a **response** (e.g. yield, weight gain, change in temperature, percent decomposition, etc.)
- As we have seen, because of **uncertainty** in the responses due to **sampling** and **biological variation**, we cannot provide **definitive**, absolute answers to the question(s) of interest based on such observations.
- But we **can** make inferences that incorporate and quantify the inherent **uncertainty**.
- Understanding this is the key to developing a good experimental **design**.

SOME IMPORTANT THEMES: The following ideas will arise frequently throughout the rest of the course. We've mentioned some of them already. It is important to keep these in mind as you read the discussion in this chapter.

- Before an experiment may be designed, the questions of interest must be **well-formulated**.
- No formal experimentation should be carried out until this has happened!
- The investigator and the statistician must work **together** to identify the important features and thus an appropriate design.

- There is no one "right" way to design an experiment. But there **are** bad ways, good ways, and better ways. Which way or ways are better depend on the features of the particular situation, as we will discuss.

- How an experiment was designed **does** dictate how it should be analyzed. We have already seen this in the case of comparing two treatment means – whether the experimental units on each treatment came from two **independent** samples or were based on **pairing** determined the type of analysis. Thus, **design and analysis go hand in hand**.

REMARK: It should be clear from the above that experimental design and statistical analysis have as much to do with **philosophy**, **creative thinking**, and **appreciation** of the scientific issues as they do with **mathematics** and **computation**! The entire discipline of **statistics** would not even **exist** without other disciplines where scientific inquiry is involved! We will come to appreciate this more in the ensuing discussion!

6.2 Roles of investigator and statistician

IDEA: The investigator and the statistician will, of necessity, have **different** perspectives on an experiment. If the **work together** from the **initial planning stage**, the result will be an investigation **designed** to provide the information necessary to give insight into important questions of **scientific interest**.

THE INVESTIGATOR:

- *Formulate broad question(s) of interest.* The first step in experimental design is deciding what the problem is! In almost all situations, this is the domain of the **investigator**, who has the scientific expertise. The statistician is generally not in a position to judge this and thus can't tell the investigator which questions to study!

- *Decide which comparisons are relevant.* By this, we mean getting more specific. Once the general question(s) of interest have been determined, the investigator must decide what specific issues are to be the focus of the particular experiment. One experiment can't answer all the questions in the universe!

- *Decide what is a meaningful scientific difference.* The investigator, based on scientific aspects, should have a general idea of what kind of differences among treatments would be important to know about. These not be precise (e.g. a 26.3% difference), but a general "ballpark" should be identified. As we have seen, this is required in order to determine appropriate sample size so that resources will not be expended in vain. It is inappropriate to decide what a meaningful difference is based on the maximum available sample size!!

- *Identify the resources available.* The investigator should have some idea of limitations in terms of money, time, personnel, logistics, etc. that may be imposed. It may turn out that a suitable experiment cannot be conducted with the available resources.

- *Identify peculiarities of the situation.* This is very specific to the situation. The investigator should think of **anything** (even if it may seem irrelevant) that might have an affect on how the experiment may be carried out. For example, if an experiment is to be conducted in a greenhouse, and there is an air conditioner at one end dripping condensation that might affect systematically the outcome of interest for plants placed near it, this will be important for the statistician to know. Or it may be physically impossible or undesirable for two treatments to be applied to adjacent plots in a field experiment.

- *Decide the desired scope of interest.* The investigator should have some sense of the desired applicability of results. For example, if he or she is comparing 2 drugs, and would like to be able to make recommendations on treatment for both men and women, then subjects of both genders will need to be recruited.

THE STATISTICIAN:

- *Identify the relevant population(s).* The ultimate objective for the statistician is to cast the problem in a formal statistical model framework. The first step is to identify, based on the investigator's desired scope of interest and comparisons, the **population(s)** from which **sampling** will be required.

- *Identify an appropriate probability model.* Based on the kind of response to be observed, the statistician must determine how to represent the populations in terms of probability distribution models. For example, if the response to be collected is weight gain under 2 different rations (a **continuous** response), the normal distribution may be an appropriate approximate model for responses on each ration.

- *Cast the questions(s) of interest as statistical hypotheses.* In the context of the probability model, express the scientific questions of interest in terms of **population parameters**. For example, comparing the weight gains for the 2 rations may be cast as a question about the means of the normal distributions for the responses on each ration. In different problems, different probability models and parameters may be appropriate.

- *Design the experiment.* Taking into account the limitations on resources, peculiarities, and meaningful scientific differences identified by the investigator, determine an appropriate procedure for drawing **samples** for the populations of interest. This will involve assigning the treatments to experimental material in such a way that

  - samples are representative of the population
  - no confounding or bias is possible
  - the relevant comparisons may be addressed (statistical hypotheses may be tested)
  - meaningful differences can be detected with high probability (power) if they exist

  This may also involve telling the investigator this **cannot** be accomplished with the available resources!

It doesn't stop here. The investigator and statistician will most likely go through several **iterations** of these activities. For example

- The statistician may come up with a design that the investigator realizes is impossible to carry out because of a peculiarity he/she forgot to mention. Taking this new issue into consideration, the statistician can come up with a new design.

- The statistician may determine that the available resources are not sufficient to detect differences with the desired precision. Based on this consideration, the investigator may decide to scale back the scope of inference or seek additional resources.

RESULT: The investigator and statistician must **work together** from the outset.

6.3 Statistical issues in experimental design

We now discuss some of the notions of design from the statistician's point of view.

TERMINOLOGY: We've already been using some of this terminology; now, we formalize the definitions. In particular, our precise definition of **experimental unit** will be quite important.

- *Treatment*. The procedure whose effects is to be measured and compared with other procedures.
- *Experimental unit*. The unit of experimental material to which one application of the treatment is applied.
- *Experimental design*. From a statistician's point of view, a design is a plan for obtaining and using experimental material in order to allow comparisons of among treatments. More specifically, it is a plan for applying the treatments to experimental units in such a way that experimental units are alike except for the treatments.

REMARK: If we think about the precise definition of **experimental unit** given here, we see that we must be a bit careful. Recall the hop plant example we used to illustrate the use of SAS for testing treatment differences in a **paired design** at the end of chapter 5. Each hop plant as divided in half, and one half was pollinated (treatment 1) and the other half was not (treatment 2). What is the experimental unit here? Because the treatments were applied to "half" plants, the experimental unit is a "half plant."

A harder example is the blood pressure study study in section 5.7. Here, each man was seen **before** and **after** stimulus, and we thought of the "pair" as the abstract "men" before and after. That is, the treatments were applied to each "man" in this "pair" – the "before man" and the "after man." We could almost think of this as "cutting the man in half," as in the hop plant example. The **experimental units** here are the abstract men in a "pair."

In both of these examples, the "pairing" is meaningful in the sense that we'd expect the halves of the same plant to respond more similarly than halves from different plants. Similarly, "halves" from the same man would respond more similarly than "halves" of different men (e.g. the "before" for one man and the "after" for another).

The tomato example at the end of section 5.7 is easy. Here, tomato plants were paired in a meaningful way (adjacent to one another in the row), and the experimental units in a pair were the two plants in that pair.

KEY CONSIDERATIONS: We have already seen from our discussion of comparing two treatments that how well we are able to (i) estimate a difference and (ii) determine whether a real difference may exist in treatment means depends on the **standard error** of the difference. Consider for definiteness the case where we have **two independent samples** of size n with the **same** variance σ^2 in each population. Recall that our test statistic is

$$\frac{\bar{D} - \delta}{s_{\bar{D}}},$$

where

$$s_{\bar{D}} = \sqrt{\frac{2s^2}{n}}. \tag{6.1}$$

Here, s^2 is the estimate of σ^2. $s_{\bar{D}}$ measures how **precisely** we can estimate the difference and determines the size of our test statistic (and thus power).

As we have already discussed, it is clear that the size of $s_{\bar{D}}$ depends on two critical components:

- The sample size, n
- The variance of the response in the population, σ^2.

It should thus be clear that two key aspects that must be considered in designing an experiment are this variance and the number of experimental units.

Although we have identified these issues by thinking of the simple situation of comparing two treatments based on independent samples, these two issues arise more generally.

SAMPLE SIZE: In the two sample case, the sample size n represents the **number of experimental units** seen on each treatment. As we have already discussed, if we increase n, we can decrease the magnitude of $s_{\bar{D}}$, thereby increasing **precision** and **power**.

It is thus obvious that any experimental design must have a **sufficient number** of experimental units on each treatment. This is well under our control and limited only by the resources available.

REPLICATION: Because the number of experimental units per treatment is obviously an important consideration, it is given a special name. The term **replication** refers to the number of experimental units on each treatment. A treatment is said to be **replicated** if it is applied to more than one experimental unit. To a statistician, this term refers precisely to this issue, the number of experimental units on each treatment, under **any** design. Be sure you use it only in this way.

- In the two independent sample design above, the number of replicates on each treatment is n.
- In a **paired** design with n pairs, the number of replicates is also n. The **design**, however, is different.

We may thus state our ideas on sample size more formally.

- The number of replicates per treatment is a key factor in determining precision and power.
- If we have a fixed total number of experimental units available, part of experimental design is thus how to make the best use of them for detecting treatment differences. It should be clear that under these conditions we would be better off with **a few** treatments with lots of replicates per treatment rather than **many** treatments with fewer replicates on each. The same total number of replicates can lead to 2 very different designs, one good and one bad in this case!
- Thus, if limited resources are available, it is better to reduce the number of treatments to be considered or postpone the experiment rather than to proceed with too few replicates.

VARIATION: This aspect is harder. To get an understanding of why, we now think about the variance, σ^2, in our two sample problem more carefully. In such an experiment, each **treatment** is **replicated** n times. In section 5.8, we wrote down a **linear additive model** for a single response observed on an experimental unit receiving treatment i:

$$Y_{ij} = \mu + \tau_i + \epsilon_{ij}.$$

The first component in the model, μ, represents the mean of responses for all experimental units before treatment application and τ_i represents the change in mean for each treatment. Together, these components characterize the treatment mean $\mu_i = \mu + \tau_i$. The treatment mean is **fixed** for each treatment and does not vary. What **does** vary are the observations, because of inherent biological differences in the sampled experimental units to which the treatments were applied.

The final component, ϵ_{ij}, may thus be thought of as characterizing the way in which the response Y_{ij} varies because it arose from observing the particular experimental unit. That is, the ϵ_{ij} characterize the inherent variation in experimental units that makes them yield different responses. Recall that the variance σ^2 describes variation in the population of all possible responses. From this perspective, we may thus think of σ^2 as characterizing the variation in ϵ_{ij} values in the population of experimental units; that is, how experimental units vary!

We thus see that **precision** of estimation and **power** for testing differences depends the **inherent variation** in the experimental material (units).

- If this variation is large, our ability to provide good estimates and detect differences will be limited. Inspection of (6.1) shows that one way we might try to overcome this variability when we are studying a response that is inherently **highly variable** across experimental units is to use an **enormous** number of replicates (enormous number of experimental units per treatment), big enough to make the ratio

  $$\frac{\sigma^2}{n}$$

  small. If we are limited by available resources, however, we may be unable to get a sample size big enough!

- Unlike replication, we cannot **completely** control variability. There will **always** be some variability attributable to experimental units present.

This seems pretty hopeless! However, by being **clever**, paying careful attention to the nature of the experimental material, we may be able to reduce the **effects** of variability on our ability to make precise inferences **by design**. To discuss this, we need some more terminology.

EXPERIMENTAL ERROR: If we wish to compare treatments, the **experimental unit**, having received one application of the treatment, is the relevant unit of experimental material to consider when assessing the available information. We define **experimental error** to be a measure of the variation among experimental units that hopefully measures mainly **inherent** variation among them.

- In the above discussion, then, σ^2 quantifies what we are regarding as experimental error in the two sample case. In the model, ϵ_{ij} represents how experimental error arises.

IDEA: To "be clever," consider the following:

- If we wish to detect differences among treatments, then we hope that most of the variability in results is due to the **systematic** effect of treatments (e.g. τ_i in the model above) rather than **random** variation in the experimental units to which the treatments were applied, e.g., **experimental error**.
- Clearly then, if we could reduce the magnitude of experimental error somehow, we'd be in a better position to detect differences.
- As we just discussed, we can't get rid of variation entirely! But we **can** think of how it arises!
- For example, if we are trying to compare two drugs, the experimental units would be **subjects** to whom we administer the drugs. Subjects may vary in their responses to the drugs because they are just inherently different, but they may also differ in their responses for **systematic** reasons, such as gender, age, etc. Thus, part of the variation in experimental units may be due to **systematic** causes we can identify!
- If we could attribute some of the variation in experimental units to systematic sources, we could reduce the our assessment of inherent variation among them!
- That is, reduce our assessment of **experimental error**!

IMPLEMENTATION: To take advantage of this idea, the obvious strategy is to set up the experiment so that the **systematic** variation among experimental units may be **separated** from the **inherent** variation.

- If we **group** experimental units according to systematic features they share, such as gender, we can hopefully explain part of the variation by that we observe **across groups**.
- The remaining variation will be that arising **within groups** of experimental units (that we can't explain and thus view as inherent). This variation would comprise **experimental error**.
- Because the units in the groups are in the groups because they share a common feature, they are apt to be "alike," thus, the hope is that experimental error will be small.
- We have already seen an example this idea – the **paired design**. Recall that the linear additive model we thought of was

$$Y_{ij} = \mu + \tau_i + \rho_j + \epsilon_{ij}.$$

"Pair" was identified as a systematic feature that might explain some of the variability, and is represented in the model by the systematic component ρ_j. Once we account for ρ_j, ϵ_{ij} represents what we attribute to **experimental error**. This assessment of experimental error must be smaller than if we hadn't identified the systematic feature "pair."

RESULT: Experimental design is founded on the principle of reducing what we regard as experimental error by meaningful grouping of experimental units. Although we can't eliminate inherent variability completely, we can try to be careful about what we consider to be inherent variability!

RANDOMIZATION: We have mentioned the idea of **randomization** many times so far. In terms of experimental design, randomization involves the assignment of treatments to experimental units, based on the chosen design, by some chance mechanism. When meaningful grouping is involved, this may still be done, as we will discuss later in the course. Again, the purpose is to ensure that no treatment is somehow favored or handicapped, so that the replicates receiving each treatment are representative of the population except for the treatments. As we discussed in chapter 1, systematic, rather than random, assignment, may lead to **confounding**.

As we have also discussed, randomization ensures that observations represent **random samples** from populations of interest. This ensures validity of statistical methods.

SUMMARY: A good experimental design attempts to

- Ensure sufficient **replication** of treatments
- Reduce the effects of **experimental error** by meaningful grouping of experimental units.

As we will see in subsequent chapters, this second aspect may be accomplished in a number of ways.

6.4 Experimental unit vs. sampling unit

In the preceding discussion, we have seen how important the notions of **experimental unit** and **replication** are to experimental design. Correctly identifying the relevant experimental unit is one of the most important aspects of design.

However, in many experiments, confusion may arise. In our work so far, we have considered only cases where a **single** observation of the response is made on each experimental unit; however, it is common practice to take **more than one observation** on an experimental unit.

SAMPLING UNIT: The **fraction** of the experimental unit upon which a single observation is made.

To understand this important distinction and its implications for design, consider the following examples:

	Treatment	*Experimental Unit*	*Sampling Unit*	*Observation*
(i)	Food rations	20 swine in a pen	a single pig	weight gain
(ii)	Insecticides	50 plants on a plot	a single plant	# insects
(iii)	Drugs	a single patient	a single patient	survival time
(iv)	Variety	3 row plot of plants	a single row	yield

In example (iii), the experimental unit and sampling unit are **the same**. In the others:

(i) It is common to confine animals to the same pen or cage. Thus, it is simpler to feed the entire pen the same ration rather than to feed them individually. Moreover, this also ensures that no mistaken mixing of rations might occur among animals. However, it is logical to observe a separate weight gain for each animal. Because the **whole pen** receives the same ration, it constitutes an experimental unit. The observations, weight gains, are measured on each pig **within** an experimental unit, thus, they constitute sampling units.

(ii) Similarly, it is easier to spray a whole plot with insecticide rather than individual plants, but logical to count insects on each plant.

(iv) It is logistically simpler to plant large areas with the same variety; here, a 3-row plot. The rows may be sufficiently separated that it is possible to measure yield on each row separately, so this is often the case.

PRINCIPLE: It is the **experimental unit** that is the relevant unit of experimental material to consider when assessing available information, **not** the sampling unit.

ILLUSTRATION: Here is an example. Suppose the following experiment is conducted, with the goal of comparing two treatments in mice. Two mice are obtained. One mouse is given treatment A, the other, treatment B. **500** observations are taken on each mouse.

- *What is the experimental unit*? Mouse.
- *What is the sampling unit*? A single measurement on a mouse.
- *How many replicates per treatment*? One!

In this example, we may know quite a bit about each mouse, with 500 observations on each! But we know very little about how the treatments compare **in the population** of such mice! We've only seen **one** mouse on each treatment; thus, we do not know if observed differences we might see are due to a real difference in the treatments or just the fact that these 2 mice are quite different (and may in fact have responded quite differently to the **same** treatment)! In particular, the **number of replicates** (sample size) is $n = 1$, which is pretty useless for getting good precision or power!

This is an extreme, perhaps contrived example, but it illustrates a general point. It is a **common mistake** for investigators to use **too few** replicates per treatment but take many sampling units on each! This is likely due to failure to realize the difference, and confusion about what is meant by **replication**.

As we have discussed, the **number of replicates** (and not the number of sampling units per replicate), is the important feature determining precision and power. Thus, it is better to have a **large** number of experimental units and a **small** number of sampling units on each than the other way around!

6.5 Experimental procedure

A final word of caution is in order. So far, when we have discussed **variation**, we have mentioned systematic and inherent sources. We have seen that good experimental design attempts to exploit our understanding of these components of variation.

However, there is **another** source of variation about which statistics and design can do **nothing**. If the execution of an experiment is faulty or sloppy, this may introduce variation into the results. For example

- *Inaccuracy* may result from untidy record keeping or mistakes in entering data onto collection forms or a computer.
- *Bias* may result if a measuring device such as a scale is "off" in a systematic way.

These phenomena contaminate results in a way we cannot deal with by statistical methods, which are based on **chance**. **No** type of statistical method, no matter how sophisticated, can compensate for variation induced by **faulty technique**!

7 One Way Classification and Analysis of Variance

Complementary Reading: STD, Chapter 7

7.1 Introduction

As we discussed in chapter 6, the purpose of an experiment is often to investigate differences among treatments. In particular, in our statistical model framework, we would like to compare the (population) **means** of the responses to each treatment. We have already discussed designs (two independent samples, pairing) for comparing two treatment means. In this chapter, we begin our study of more complicated problems and designs by considering the comparison of **more than** two treatment means.

Recall that we argued in chapter 6 that, in order to detect differences if they really exist, we must try to control the effects of **experimental error**, so that any variation we observe can be attributed mainly to the effects of the treatments rather than to differences among the experimental units to which the treatments are applied. We discussed the idea that **designs** involving meaningful **grouping** of experimental units are the key to reducing the effects of experimental error, by identifying components of variation among experimental units that may be due to something besides inherent biological variation among them. The paired design for comparing two treatments is an example of such a design.

Before we can talk about grouping in the more complicated scenario involving more than two treatments, it makes sense to talk about the simplest setting in which we compare several treatment means. This is basically an extension of the "two independent samples" design to more than two treatments.

ONE WAY CLASSIFICATION: Consider an experiment to compare several treatment means set up as follows. We obtain (randomly, of course) experimental units for the experiment, and randomly assign them to treatments so that each experimental unit is observed under one of the treatments. In this situation, the samples corresponding to the treatment groups are **independent** (the experimental units in each treatment sample are **unrelated**). We do not attempt to **group** experimental units according to some factor (e.g. gender, etc.).

In this experiment, then, the only way in which experimental units may be "classified" is with respect to which treatment they received. Other than the treatments, they are viewed as basically alike. Hence, such an arrangement is often called a **one way classification**.

USEFULNESS: When experimental units are thought to be basically alike, and are thus expected to exhibit a small amount of variation from unit-to-unit, grouping them really would not add much precision to an experiment.

EXAMPLE: In a laboratory experiment for which the experimental material may be some chemical mixture to be divided into beakers (the experimental units) to which treatments will be applied, and experimental conditions are the same for all beakers, we would not expect much variation among the beakers before the treatments were applied. In this situation, grouping beakers would be pointless; there is not even an identifiable basis for doing so. We would thus expect that, once we apply the treatments, any variation in responses across beakers will mainly be due to the treatments, as beakers are pretty much alike otherwise.

COMPLETE RANDOMIZATION: If there is no basis for grouping, and thus treatments are to be simply assigned to experimental units without regard to any other factors, then, as noted above, this should be accomplished according to some chance (random) mechanism. All experimental units should have an equal chance of receiving any of the treatments. When randomization is carried out in this way, it is called **complete randomization**. This is to distinguish the scheme for treatment allocation from more complicated methods involving **grouping**, which we will talk about later.

ADVANTAGES:

- Simplicity of implementation.
- Simplicity of analysis.
- The size of the experiment is limited only by the availability of experimental units. No special considerations for different types of experimental units are required.

DISADVANTAGES:

- **Experimental error**, our assessment of variation believed to be inherent among experimental units (not systematic), includes **all** (both inherent and potential systematic) sources. If it turns out unexpectedly that some of the variation among experimental units is indeed due to a systematic component, it will not be possible to "separate it out" of experimental error, and comparisons will lack precision. In such a situation, a more complicated design involving grouping should have been used up front.
- Thus, we run the risk of low precision and power if something unexpected arises.

7.2 Analysis of variance

CURIOSITY: We wish to determine if differences exist among **means** for responses to treatments; however, the general procedure for inferring whether such differences exist is called analysis of **variance**!

ANALYSIS OF VARIANCE: This is the name given to a given to a general class of procedures that are based roughly on the following idea. We have already spoken loosely of attributing **variation** to treatments as being "equivalent" to determining if a difference exists in the underlying population treatment **means**. It turns out that it may be shown that there is actually a more formal basis to this loose way of speaking, and it is this basis that gives the procedure its name.

It is easiest to understand this in the context of the **one way classification**; however, the basic premise is applicable to more complicated designs that we will discuss later.

NOTATION: To facilitate our further development, we will change slightly our notation for denoting a **sample mean**. As we will see shortly, we will need to deal with several different types of means for the data, and this notation makes it a bit easier to keep straight which mean is which.

Let t denote the **number of treatments**. Let

$$Y_{ij} = \text{ response on the } j\text{th experimental unit on treatment } i$$

Here, $i = 1, \ldots, t$.

(We consider first the case where only **one** observation is taken on each experimental unit, so that the experimental unit = the sampling unit.)

We will consider for simplicity the case where the same number of experimental units, that is, **replicates** are assigned to each treatment. To highlight the term **replication**, we let

$$r = \text{ number of experimental units, or replicates, per treatment.}$$

Thus, r replaces our previous notation, n,

We will denote the **sample mean for treatment** i by

$$\bar{Y}_{i\cdot} = \frac{1}{r}\sum_{j=1}^{r} Y_{ij}.$$

The only difference between this notation and our previous notation is the use of the "$\cdot$" in the subscript. This usage is fairly standard, and reminds us that the mean was taken by summing over the subscript in the second position, j (that is, summing over the replicates for treatment i). It will become clear shortly why this is done.

Also define

$$\bar{Y}_{\cdot\cdot} = \frac{1}{rt}\sum_{i=1}^{t}\sum_{j=1}^{r} Y_{ij}.$$

Note that the total number of observations in the experiment is $r \times t = rt$; r replicates on each of t treatments. Thus, $\bar{Y}_{\cdot\cdot}$ represents the sample mean of **all the data**, across all replicates **and** treatments. The double dots make it clear that the summing has been performed over both subscripts.

SET-UP: Consider first the case of $t = 2$ treatments with two independent samples. Suppose that the population variance is the **same** for each treatment and equal to σ^2.

Recall that our test statistic for the hypotheses

$$H_0 : \mu_1 - \mu_2 = 0 \text{ vs. } H_1 : \mu_1 - \mu_2 \neq 0$$

was (in our new notation),

$$\frac{\bar{Y}_{1\cdot} - \bar{Y}_{2\cdot}}{s_{\bar{D}}},$$

where now $\bar{D} = \bar{Y}_{1\cdot} - \bar{Y}_{2\cdot}$. Here, we have taken $\delta = 0$ and considered the two-sided alternative hypothesis, as we are interested in just a **difference**.

For $t > 2$ treatments, there is no obvious generalization of this set-up!

When $t = 2$, it is obvious that the difference of the 2 sample means is the thing to inspect, because we are interested in $\mu_1 - \mu_2$. However, now, we have t population means, say

$$\mu_1, \mu_2, \ldots, \mu_t.$$

Thus, the hypotheses aren't so simple anymore – we can't express the notion of treatments "differing" by a simple difference! In particular, the null hypothesis is now

$$H_0: \text{ the } \mu_i \text{ are all equal}$$

and the alternative is

$$H_1: \text{ the } \mu_i \text{ are not all equal}$$

IDEA: The idea to generalize the $t = 2$ case is to think instead about estimating **variances**, as follows. This may seem totally irrelevant, but when you see the end result, you'll see why!

Assume that the data are **normally distributed** with the **same variance** for all t treatment populations, that is

$$Y_{ij} \sim \mathcal{N}(\mu_i, \sigma^2).$$

How would we estimate σ^2? The obvious approach is to generalize what we did for 2 treatments and "pool" the sample variances across all t treatments. If we write s_i^2 to denote the sample variance for the data on treatment i, then

$$s_i^2 = \frac{1}{r-1}\sum_{j=1}^{r}(Y_{ij} - \bar{Y}_{i\cdot})^2.$$

The estimate would be the **average** of all t sample variances (because r is the same for all samples), so the "pooled" estimate would be

$$\begin{aligned}&\frac{(r-1)s_1^2 + (r-1)s_2^2 + \cdots + (r-1)s_t^2}{t(r-1)} \\ &= \frac{\sum_{j=1}^{r}(Y_{1j} - \bar{Y}_{1\cdot})^2 + \sum_{j=1}^{r}(Y_{2j} - \bar{Y}_{2\cdot})^2 + \cdots \sum_{j=1}^{r}(Y_{tj} - \bar{Y}_{t\cdot})^2}{t(r-1)}. \end{aligned} \tag{7.1}$$

As in the case of 2 treatments, this estimate makes sense **regardless** of whether H_0 is true. It is based on deviations from each mean separately, through the sample variances, so it doesn't matter whether the true means are different or the same – it is still a sensible estimate.

Now recall that, if a sample arises from a normal population, then the sample mean is **also** normally distributed. Thus, this should hold for **each** of our t samples, which, in our new notation, may be written

$$\bar{Y}_{i\cdot} \sim \mathcal{N}\left(\mu_i, \frac{\sigma^2}{r}\right) \tag{7.2}$$

(that is, $\sigma^2_{\bar{Y}_{i\cdot}} = \sigma^2/r$).

Now consider the null hypothesis; under H_0, all the treatment means are the same and thus equal the same value, μ, say. That is,

$$\text{Under } H_0, \quad \mu_i = \mu, \quad i = 1, \ldots, t.$$

Under this condition, (7.2) becomes

$$\bar{Y}_{i\cdot} \sim \mathcal{N}\left(\mu, \frac{\sigma^2}{r}\right)$$

for all $i = 1, \ldots, t$. Thus, if H_0 **really were true**, we could view the sample means

$$\bar{Y}_{1\cdot}, \bar{Y}_{2\cdot}, \ldots, \bar{Y}_{t\cdot}$$

as being just a **random sample** from a normal population with mean μ and variance σ^2/r.

Consider under these conditions how we might estimate the variance of this population, σ^2/r. The obvious estimate would be the **sample variance** of our "random sample" from this population, the t sample means. Recall that sample variance is just the sum of **squared deviations** from the sample mean, divided by sample size -1. Here, our sample size is t, and the sample mean is

$$\begin{aligned} \frac{1}{t}\sum_{i=1}^{t} \bar{Y}_{i\cdot} &= \frac{1}{t}\sum_{i=1}^{t}\left(\frac{1}{r}\sum_{j=1}^{r} Y_{ij}\right) \\ &= \frac{1}{rt}\sum_{i=1}^{t}\sum_{j=1}^{r} Y_{ij} = \bar{Y}_{\cdot\cdot} \end{aligned}$$

That is, the mean of the sample means is just the sample mean of all the data (this is not always true, but is in this case because r is the same sample size for all treatments.)

Thus, the sample variance we would use as an estimator for σ^2/r is

$$\frac{1}{t-1}\sum_{i=1}^{t}(\bar{Y}_{i\cdot} - \bar{Y}_{\cdot\cdot})^2.$$

This suggests another estimator for σ^2 **itself**, namely, r times this, or

$$r \times \frac{1}{t-1} \sum_{i=1}^{t} (\bar{Y}_{i\cdot} - \bar{Y}_{\cdot\cdot})^2. \tag{7.3}$$

.

REMARK: Note that we derived the estimator for σ^2 given in (7.3) under the assumption that the treatment means were all the **same**. If they really **were not** the same, then, intuitively, this estimate of σ^2 would tend to be **too big**, because the deviations about the sample mean $\bar{Y}_{\cdot\cdot}$ of the $\bar{Y}_{i\cdot}$ will include **two** components

1. A component attributable to **random variation** among the $\bar{Y}_{i\cdot}$s
2. A component attributable to the **systematic difference** among the means μ_i.

The first component will be present even if the means are the **same**; the second component will only be present when they **differ**.

RESULT: We now have derived **two** estimators for σ^2 :

- The first, the "pooled" estimate given in (7.1), **will not** be affected by whether or not the means are the different. This estimate reflects how **individual observations** differ from their means, regardless of the values of those means; thus, it reflects **only** variation attributable to how experimental units differ among themselves.
- The second, derived assuming the means are the same, and given in (7.3), **will** be affected by whether the means are are different. This estimate reflects not only how **individual observations**, through their sample means, differ, but **also** how the means might differ. That is, this estimate reflects **both** variation attributable to how experimental units differ among themselves **and** attributable to differences caused by the **treatments** (different means)!

IMPLICATION: Recall that we derived the second estimator for σ^2 under the assumption that H_0 is **true**. Thus, if H_0 **really were** true, we would expect **both** estimators for σ^2 to be about the same size, since in this case both would reflect only variation attributable to experimental units. If, on the other hand, H_0 really is **not true**, we would expect the second estimator to be **larger**.

With this in mind, consider the ratio

$$F = \frac{\text{estimator for } \sigma^2 \text{ based on sample means (7.3)}}{\text{estimator for } \sigma^2 \text{ based on individual deviations (7.1)}}$$

We now see that if

- H_0 is true, ratio will be **small**
- H_0 is not true, ratio will be **large**!

THE F RATIO: The result is that we may base inference on **treatment means** (whether they differ) on this ratio of **estimators for variance**! We may use this ratio as a **test statistic** for testing H_0 vs. H_1.

Recall from chapter 5 that a ratio of two sample variances for 2 independent populations has a F **distribution**. Recall also that our approach to hypothesis testing is to assume H_0 is true, look at the value of a test statistic, and evaluate how "likely" it is if H_0 is true. If H_0 true in our situation here, then

- The numerator is a sample variance of "data" $\bar{Y}_{i\cdot}$, $i = 1, \ldots, t$, from a $\mathcal{N}(\mu, \sigma^2/r)$ population
- The denominator is a ("pooled") sample variance of the data Y_{ij}.

It turns out, as we will see shortly, that if H_0 is true we may further view these 2 sample variances as **independent**, even though the are based on the **same observations**.

It thus follows that we have the ratio of 2 "independent" sample variances if H_0 is true, so that

$$F \sim F_{(t-1),t(r-1)}$$

The **degrees of freedom** $(t-1)$ (numerator) and $t(r-1)$ (denominator) correspond to the degrees of freedom for each sample variance. That is, the statistic F has the $F_{(t-1),t(r-1)}$ distribution as its **sampling distribution**.

As we will demonstrate formally in a moment, this fact may be used to conduct hypothesis tests about differences among the treatment means.

INTERESTING FACT: It turns out that, in the case of $t = 2$ treatments, the ratio F reduces to

$$F = \frac{(\bar{Y}_{1\cdot} - \bar{Y}_{2\cdot})^2}{s^2_{\bar{D}}} = t^2,$$

the **square** of the usual t statistic. Here, F will have a $F_{1,2(r-1)}$ distribution.

It is furthermore true that, when the **numerator** degrees of freedom for a F distribution are equal to 1, and the denominator degrees of freedom = some value ν, say, then the **square root** of the F random variable has a t_ν distribution. (You can convince yourself by comparing the square roots of the values in the F table for numerator df = 1 to the appropriate values in the t table.)

Thus, when $t = 2$, comparing the ratio F to the F distribution is the **same** as comparing the usual t statistic to the t distribution. That is, it's the **same test procedure**.

This shows that our argument to generalize the comparison for $t = 2$ treatments to $t > 2$ treatments based on the idea of estimating the variance σ^2 of the observations in 2 different ways is valid – it reduces to what we would expect in the case $t = 2$.

ANALYSIS OF VARIANCE: The reason for the use of this term should now be obvious! In fact, this same reasoning may be extended to more complicated situations beyond the simple one way classification setting, as we'll see later in the course.

7.3 Linear additive model

It is again convenient to write down a model for an observation, to highlight the possible sources of variation. For the general one way classification with t treatments, we may classify an individual observation as being on the jth experimental unit in the ith treatment group as

$$Y_{ij} = \mu + \tau_i + \epsilon_{ij}, \quad i = 1, \ldots, t, \quad j = 1, \ldots, r_i,$$

where

- t = number of treatments
- r_i = number of replicates on treatment i. In general, this may be different for different treatments, so we add the subscript i. (What we called n_i in the case $t = 2$ is now r_i.)
- $\mu_i = \mu + \tau_i$ is the mean of the **population** describing responses on experimental units receiving the ith treatment.
- μ may be thought of as the "overall" mean with no treatments
- τ_i is the change in mean (deviations from μ) associated with treatment i

We have written the model generally here to allow for **unequal replication**. We will see shortly that the idea of an F ratio may be generalized to this case.

This model is just an extension of that we used in the case of 2 treatments, and, as in that case, shows that we may think of observations varying about an overall mean because of the **systematic** effect of treatments and the **random** variation in experimental units.

7.4 Fixed vs. random effects

Recall that τ_i represents the "effect" (deviation) associated with getting treatment i. Depending on the situation, our further interpretation of τ_i, and in fact of the treatments themselves, may differ. Consider the following examples.

EXAMPLE 1: Suppose $t = 3$ and that each treatment is a different fertilizer mixture for which mean yields are to be compared. Here, we are interesting in comparing 3 specific treatments. If we repeated the experiment again, these 3 fertilizers would always constitute the treatments of interest.

EXAMPLE 2: Suppose a factory operates a large number of machines to produce a product and wishes to determine whether the mean yield of these machines differs. It is impractical for the company to keep track of yield for **all** of the many machines it operates, so a **random sample** of 5 such machines is selected, and observations on yield are made on these 5 machines. The hope is that the results for the 5 machines involved in the experiment may be **generalized** to gain insight into the behavior of **all** of the machines.

In Example 1, there is a **particular** set of treatments of interest. If we started the experiment next week instead of this week, we would still be interested in this same particular set – it would not vary across other possible experiments we might do.

In Example 2, the treatments are the 5 machines chosen from all machines at the company, chosen by random selection. If we started the experiment next week instead of this week here, we might end up with a **different** set of 5 machines with which to do the experiment. In fact, whatever 5 machines we end up with, these **particular** machines are not the specific machines of interest. Rather, interest focuses on the **population** of **all** machines operated by the company. The question of interest now is not about the **particular** treatments involved in the experiment, but the **population** of **all** such treatments!

This distinction is quite an important one. In a situation like Example 1, the particular treatments in the experiment are the only ones of interest, so there is no uncertainty involved. In a situation like Example 2, there **is** additional uncertainty involved, because the treatments are no longer fixed!

We thus make the following distinction in our model:

- In a case like Example 1, the τ_i are best regarded as **fixed** quantities, as they describe a particular set of conditions. In this situation, the τ_i are referred to as **fixed effects**.
- In a case like Example 2, the τ_i are best regarded as **random variables**! Here, the particular treatments in the experiment may be thought of as being drawn from a population of all such treatments, so there is chance involved. We hence think of the τ_i as random variables with some mean and **variance** σ_τ^2. This variance characterizes the variability in the population of all possible treatments, in our example, the variability across all machines owned by the company. If machines are quite different in terms of yield, σ_τ^2 will be **large**. If yields are consistent across machines, σ_τ^2 will be **small**. In this situation, the τ_i are referred to as **random effects**.

You might expect that these 2 situations would lead to **different** considerations for testing. In the **random treatment effects** case, there is **additional uncertainty** involved, because the treatments we use aren't the only ones of interest. It turns out that in the particular simple case of assessing treatment differences for the one way classification, the methods we will discuss are valid for either case. **However**, in more complicated designs, this is **not necessarily** the case. We will discuss this issue more later.

7.5 Model restriction

PROBLEM: We have seen that $\bar{Y}_{i\cdot}$ is an estimator for μ_i for a sample from population i. $\bar{Y}_{i\cdot}$ is our "best" indication of the mean response for population i. But if we think about our model, $\mu_i = \mu + \tau_i$, which breaks μ_i into 2 components, we do not know **how much** of what we see, $\bar{Y}_{i\cdot}$, is due to the original population of experimental units **before** treatments were applied (μ) and how much is due to the effect of the treatment (τ_i).

In particular, the linear additive model we write down to describe the situation actually contains elements we can never hope to get a sense of from the data at hand! More precisely, the best we can do is estimate the individual means μ_i using $\bar{Y}_{i\cdot}$; we cannot hope to estimate the individual treatment effects τ_i without additional knowledge or assumptions.

TERMINOLOGY: Mathematically speaking, a model that contains components that cannot be estimated is said to be **overparameterized**. We have more **parameters** then we can estimate from the available information. Thus, although the linear additive model is a nice device for focusing our thinking about the data, it is **overparameterized** from a mathematical point of view.

ONE APPROACH: It may seem that this is an "artificial" problem – why write down a model for which one can't estimate all its components? The reason is, as above, to give a nice framework for **thinking** about the data – for example, the model allows us to think of **fixed** or **random** treatment effects, depending on the type of experiment.

To reconcile our desire to have a helpful model for thinking and the mathematics, the usual approach is to impose some sort of **assumption**. A standard way to think about things is to suppose that the overall mean μ can be thought of as the **mean** or **average** of the individual treatment means μ_i, that is

$$\mu = \frac{1}{t}\sum_{i=1}^{t}\mu_i,$$

just as

$$\bar{Y}_{\cdot\cdot} = \frac{1}{t}\sum_{i=1}^{t}\bar{Y}_{i\cdot}.$$

This implies that

$$\mu = \frac{1}{t}\sum_{i=1}^{t}(\mu+\tau_i) = \mu + \frac{1}{t}\sum_{i=1}^{t}\tau_i,$$

so that it must be that

$$\sum_{i=1}^{t}\tau_i = 0.$$

The condition $\sum_{i=1}^{t}\tau_i = 0$ goes along with the interpretation of the τ_i as "deviations" from an overall mean. The treatments "affect" the response in different "directions;" some of the τ_i must be negative and others positive for them all to sum to zero.

The **restriction** $\sum_{i=1}^{t}\tau_i = 0$ is thus one you will see often in work on analysis of variance. Basically, it has no effect on our objective, investigating differences among treatment means. All the restriction does is impose a particular interpretation on our linear additive model.

You will often see the null and alternative hypotheses written in terms of τ_i instead of μ_i. Note that, under this interpretation, if all treatment means were the **same** (H_0), then the τ_i must **all** be zero.

This interpretation is valid in the case where the τ_i are **fixed effects**. When they are **random effects**, the interpretation is similar. We think of the τ_i themselves as having population mean 0, analogous to them averaging to zero above. This is the analog of the restriction in the case of random effects. If there are no differences across treatments, then they do not **vary**, that is, $\sigma^2_\tau = 0$. We defer further discussion of this issue for now.

7.6 Assumptions for analysis of variance

Before we turn to using the above framework and ideas to develop formal methods, we restate for completeness the assumptions underlying our approach.

- The observations, and hence the errors, are **normally distributed**
- The observations have the **same variance** σ^2.
- All observations, both across and within samples, are **unrelated** (independent).

The assumptions provide the basis for concluding that the sampling distribution of the statistic F upon which we will base our inferences is really the F distribution.

IMPORTANT: The assumptions above are **not necessarily** true for any given situation. In fact, they are probably never **exactly true**. For many data sets, they may be a reasonable **approximation**, in which case the methods we will discuss will be fairly reliable. In other cases, they may be seriously violated; here, the resulting inferences may be **misleading**!

If the underlying distribution of the data really **is not** normal and/or the variances across treatment groups are **not** the same, then the rationale we used to develop the statistic is lost.

If the data really are **not** normal, hypothesis tests may be **flawed** in the sense that the **true level of significance** is **greater** than the chosen level α, and we may claim there is a difference when there really **is not** a difference! We may think we are seeing a difference in means, when actually we are seeing lack of normality.

For some data, it may be possible to get around these issues somewhat. So far, we have written down our model in an **additive** form. However, there are physical situations where a more plausible model is one that has error enter in a **multiplicative** way:

$$Y_{ij} = \mu^* \tau_i^* \epsilon_{ij}^*.$$

Such a model is often appropriate for growth data, or many situations where the variability in response tends to **get larger** as the response gets larger.

If we take **logarithms**, we may write this as

$$\log Y_{ij} = \mu + \tau_i + \epsilon_{ij}, \quad \mu = \log \mu^*, \quad \tau_i = \log \tau_i^*, \quad \epsilon_{ij} = \log \epsilon_{ij}^*, \quad .$$

Thus, the **logarithms** of the observations satisfy a linear, additive model.

This is the rationale behind the common practice of **transforming** the data. It is often the case that many types of biological data seem to be close to normally distributed with constant variance on the **logarithm** scale, but not at all on their **original scale**. The data are thus analyzed on this scale instead.

We will discuss data transformation more later in the course. Other transformations besides the log may be more appropriate in some circumstances.

IMPORTANT: It is beyond the scope of this course to discuss methods for **diagnosing** violations of the assumptions and for determining appropriate transformations. The best approach would be to seek the advice of a **statistician**!

For the remainder of our discussion of analysis of variance in this and subsequent chapters, we will assume that the above assumptions are reasonable either on the original or transformed scale. Keep in mind at all times that these are **assumptions**, and must be verified before the methods may be considered valid.

7.7 ANOVA for one way classification with equal replication

ACRONYM: The popular acronym for **AN**analysis **Of VA**riance is **ANOVA**.

We begin with the simplest case, where $r_i = r$ for all i. We will also assume that the τ_i have **fixed effects**.

Recall our argument to derive the form of the F ratio statistic

$$F = \frac{\text{estimator for } \sigma^2 \text{ based on sample means (7.3)}}{\text{estimator for } \sigma^2 \text{ based on individual deviations (7.1)}}.$$

In particular, the components may be written

- **Numerator**:
$$\frac{r \times \sum_{i=1}^{t}(\bar{Y}_{i\cdot} - \bar{Y}_{\cdot\cdot})^2}{t-1} = \frac{\text{Treatment SS}}{\text{degrees of freedom for treatments}}$$
- **Denominator**:
$$\frac{\sum_{i=1}^{t}\sum_{j=1}^{r}(Y_{ij} - \bar{Y}_{i\cdot})^2}{t(r-1)} = \frac{\text{Error SS}}{\text{degrees of freedom for error}}$$

Here, we define the quantities **Treatment SS** and **Error SS** and their **degrees of freedom** as given above, where, as before, SS = Sum of Squares. These names make intuitive sense. The Treatment SS is part of the estimator for σ^2 that includes a component due to variation in **treatment means**. The use of the term Error SS is as before – the estimator for σ^2 in the denominator only assesses apparent variation across experimental units.

"CORRECTION" TERM: It is convenient to define

$$C = \frac{\left(\sum_{i=1}^{t}\sum_{j=1}^{r} Y_{ij}\right)^2}{rt}.$$

ALGEBRAIC FACTS: It is convenient for getting insight (and for hand calculation) to express the SSs differently. It is possible to show that

- Treatment SS $= r\sum_{i=1}^{t}(\bar{Y}_{i\cdot} - \bar{Y}_{\cdot\cdot})^2$
$$= \frac{\sum_{i=1}^{t}\left(\sum_{j=1}^{r} Y_{ij}\right)^2}{r} - C.$$
- Error SS $= \sum_{i=1}^{t}\sum_{j=1}^{r}(Y_{ij} - \bar{Y}_{i\cdot})^2$
$$= \sum_{i=1}^{t}\sum_{j=1}^{r} Y_{ij}^2 - \frac{\sum_{i=1}^{t}\left(\sum_{j=1}^{r} Y_{ij}\right)^2}{r}.$$

Consider that the **overall , "total" variation** in all the data, if we do not consider that different treatments were applied, would obviously be well-represented by the **sample variance** for **all the data**, lumping them all together without regard to treatment. There are rt total observations; thus, this sample variance would be

$$\frac{\sum_{i=1}^{t}\sum_{j=1}^{r}(Y_{ij}-\bar{Y}_{\cdot\cdot})^2}{rt-1};$$

each deviation is of course taken about the overall mean of all rt observations.

ALGEBRAIC FACT: It may be shown that the numerator of the overall sample variance may be written

$$\sum_{i=1}^{t}\sum_{j=1}^{r}(Y_{ij}-\bar{Y}_{\cdot\cdot})^2=\sum_{i=1}^{t}\sum_{j=1}^{r}Y_{ij}^2-C.$$

This quantity is called the **Total SS**. Because it is the numerator of the overall sample variance, it may be thought of as measuring how observations vary about the overall mean, **without** regard to treatments. That is, it measures the **total variation**.

We are now in a position to gain insight. From the algebraic facts above, note that (check!)

$$\text{Treatment SS} \;+\; \text{Error SS} \;=\; \text{Total SS}. \tag{7.4}$$

(7.4) illustrates a fundamental point – the Total SS, which characterizes **overall variation** in the data without regard to the treatments, may be **partitioned** into two **independent** components:

- Treatment SS, measuring how much of the overall variation is in fact due to the treatments (in that the treatment means **differ**)
- Error SS, measuring the remaining variation, which we attribute to inherent variation among experimental units.

F STATISTIC: If we now define

$$MS_T = \text{Treatment MS} = \frac{\text{Treatment SS}}{\text{degrees of freedom for treatments}} = \frac{\text{Treatment SS}}{t-1}$$

$$MS_E = \text{Error MS} = \frac{\text{Error SS}}{\text{degrees of freedom for error}} = \frac{\text{Error SS}}{t(r-1)}$$

then we may write our F statistic as

$$F=\frac{\text{Treatment MS}}{\text{Error MS}}.$$

We now get some insight into why F has an $F_{t-1,t(r-1)}$ distribution. The components in the numerator and denominator are "independent" in the sense that the partition the Total SS into two "orthogonal" components. (A formal mathematical argument is possible.)

We summarize this information in a table:

One Way ANOVA table – Equal Replication

Source of variation	DF	SS Definition	SS Computation	MS	F
Among Treatments	$t-1$	$r\sum_{i=1}^{t}(\bar{Y}_{i\cdot}-\bar{Y}_{\cdot\cdot})^2$	$\frac{\sum_{i=1}^{t}\left(\sum_{j=1}^{r}Y_{ij}\right)^2}{r}-C$	MS_T	F
Error (Within Treatments)	$t(r-1)$	$\sum_{i=1}^{t}\sum_{j=1}^{r}(Y_{ij}-\bar{Y}_{i\cdot})^2$	by subtraction	MS_E	
Total	$rt-1$	$\sum_{i=1}^{t}\sum_{j=1}^{r}(Y_{ij}-\bar{Y}_{\cdot\cdot})^2$	$\sum_{i=1}^{t}\sum_{j=1}^{r}Y_{ij}^2-C$		

STATISTICAL HYPOTHESES: The question of interest in this setting is to determine if the means of the t treatment populations are different. We may write this formally as

$$H_{0,T}: \mu_1=\mu_2=\cdots=\mu_t \text{ vs. } H_{1,T}: \text{ The } \mu_i \text{ are not all equal.}$$

This may also be written in terms of the τ_i under the restriction $\sum_{i=1}^{t}\tau_i=0$ as

$$H_{0,T}: \tau_1=\tau_2=\cdots=\tau_t=0 \text{ vs. } H_{1,T}: \text{ The } \tau_i \text{ are not all equal.}$$

It is important to note that the alternative hypothesis **does not specify** the **way** in which the treatment means (or deviations) differ. The best we can say based on our statistic is that they differ **somehow**. The numerator of the statistic can be large because the means differ in a huge variety of different configurations. Some of the means may be different while the others are the same, all might differ, and so on.

We add the subscript "T" to remind ourselves that this particular test is with regard to treatment means – later, when we consider more exotic designs, there will be other tests we may wish to perform as well.

TEST PROCEDURE: Reject $H_{0,T}$ in favor of $H_{1,T}$ at level of significance α if

$$F > F_{(t-1),t(r-1),\alpha}.$$

NOTE: This is analogous to a two-sided test when $t = 2$ – we do not state in $H_{1,T}$ the **order** in which the means differ, only that they do. The range of possibilities of how they differ is just more complicated when $t > 2$.

We use α instead of $\alpha/2$ here because we have no choice as to which MS appears in the numerator and which appears in the denominator. (Compare to the test of equality of variance in the case of 2 treatments in chapter 5.)

EXAMPLE: (Sokal and Rohlf, 1981, *Biometry*, p. 219–221.) The following data record the length of pea sections, in ocular units ($\times$0.114 mm), grown in tissue culture with auxin present. The purpose of the experiment was to test the effects of the addition of various sugars on growth as measured by length. Pea plants were randomly assigned to one of 5 treatment groups: Control (no sugar added), 2% glucose added, 2% fructose added, 1% glucose + 1% fructose added, and 2% sucrose added. 10 observations were obtained for each group of plants.

Here, then, the **individual plants** to which the treatments were applied are the **experimental units**. Because only one observation was taken on each, they are also the **sampling units**.

We assume that the measurements are approximately normally distributed, which seems reasonable for such continuous measurement data, with the same variance.

We have $t = 5$, $r = 10$.

	Control	2% glucose	2% fructose	1% glucose + 1% fructose	2% sucrose
	75	57	58	58	62
	67	58	61	59	66
	70	60	56	58	65
	75	59	58	61	63
	65	62	57	57	64
	71	60	56	56	62
	67	60	61	58	65
	67	57	60	57	65
	76	59	57	57	62
	68	61	58	59	67
$\sum_{j=1}^{r} Y_{ij}$	701	593	582	580	641
$\bar{Y}_{i\cdot}$	70.1	59.3	58.2	58.0	64.1

We calculate:

$$C = \frac{\left(\sum_{i=1}^{t}\sum_{j=1}^{r} Y_{ij}\right)^2}{rt} = \frac{(701 + \cdots + 641)^2}{5 \times 10} = \frac{(3097)^2}{50} = 191,828.18$$

$$\frac{\sum_{i=1}^{t}\left(\sum_{j=1}^{r} Y_{ij}\right)^2}{r} = \frac{(701^2 + \cdots + 641^2)}{10} = \frac{1,929,055)}{10} = 192,905.50.$$

$$\sum_{i=1}^{t}\sum_{j=1}^{r} Y_{ij}^2 = 75^2 + 67^2 + \cdots + 67^2 = 193,151.00.$$

Thus,

$$\text{Treatment SS } = 192,905.50 - 191,828.18 = 1077.32$$

$$\text{Total SS } = 193,151.00 - 191,828.18 = 1322.82$$

$$\text{Error SS } = \text{ Total SS } - \text{ Treatment SS } = 1322.82 - 1077.32 = 245.50.$$

We also have $t - 1 = 4$, $t(r-1) = 5(9) = 45$, so that

$$MS_T = \frac{1077.32}{4} = 269.33, \quad MS_E = \frac{245.50}{45} = 5.46, \quad F = \frac{269.33}{5.46} = 49.33.$$

We summarize the computations in an analysis of variance table:

Analysis of Variance – Pea Section Data

Source of variation	DF	SS	MS	F
Among Treatments	4	1077.32	269.33	49.33
Error (Within Treatments)	45	245.50	5.46	
Total	49	1322.82		

To perform the hypothesis test for differences among the treatment means, we compare F to the appropriate value from the F table. For level of significance $\alpha = 0.05$, we have

$$2.53 < F_{4,45,0.05} < 2.61,$$

so that $49.33 > F_{4,45,0.05}$. We thus **Reject** $H_{0,T}$. There is evidence in these data to suggest that the mean lengths of pea sections are different depending upon which (if any) sugar was added.

NOTE: We **can not** tell from these results **which** means are larger or smaller than which other means – only that there **is** a difference.

In section 7.13, we show how to use SAS to conduct this type of analysis.

7.8 ANOVA for one way classification with unequal replication

We now generalize the ideas of ANOVA to the case where the r_i are **not** all equal. This may be the case by design or because of mishaps during the experiment that result in lost or unusable data. When $r_i \equiv r$ for all i, the procedure we now discuss reduces to that we have previously described for the case of equal replication. Here, again, our discussion assumes that the τ_i have **fixed effects**.

When the r_i are not all equal, the the total number of observations in the data set is

$$N = \sum_{i=1}^{t} r_i.$$

Thus, for purposes of algebra, we redefine the correction factor as

$$C = \frac{\left(\sum_{i=1}^{t}\sum_{j=1}^{r_i} Y_{ij}\right)^2}{\sum_{i=1}^{t} r_i} = \frac{\left(\sum_{i=1}^{t}\sum_{j=1}^{r} Y_{ij}\right)^2}{N}.$$

The **overall sample mean** of all observations is now

$$\bar{Y}_{\cdot\cdot} = \frac{\sum_{i=1}^{t}\sum_{j=1}^{r_i} Y_{ij}}{\sum_{i=1}^{t} r_i} = \frac{\sum_{i=1}^{t}\sum_{j=1}^{r_i} Y_{ij}}{N}$$

and the sample mean for the ith treatment is now

$$\bar{Y}_{i\cdot} = \frac{1}{r_i}\sum_{j=1}^{r_i} Y_{ij}.$$

It may be shown that the **Total SS**, which is still defined as

$$\sum_{i=1}^{t}\sum_{j=1}^{r}(Y_{ij} - \bar{Y}_{\cdot\cdot})^2,$$

with $\bar{Y}_{\cdot\cdot}$ as above, is equivalent to

$$\text{Total SS } = \sum_{i=1}^{t}\sum_{j=1}^{r} Y_{ij}^2 - C,$$

where the correction factor C is now defined as above.

Recall in our argument for the equal replication case, the ratio of two different estimators for σ^2 formed the basis for our test statistic, the F ratio, and the sums of squares associated with numerator and denominator, the Treatment and Error SS, respectively, summed to equal the Total SS. It is possible to derive similar quantities in the unequal replication case. We had

$$F = \frac{\text{estimator for } \sigma^2 \text{ based on sample means}}{\text{estimator for } \sigma^2 \text{ based on individual deviations}}.$$

Error SS: Our estimator for σ^2 with unequal replication (for the denominator) based on individual deviations would again be based on "pooling" sample variances across treatments. Now

$$s_i^2 = \frac{1}{r_i - 1} \sum_{j=1}^{r_i} (Y_{ij} - \bar{Y}_{i\cdot})^2.$$

The estimate for σ^2 would be the **weighted average** of all t sample variance (in an obvious extension of what we did for two samples), that is,

$$\begin{aligned} &\frac{(r_1 - 1)s_1^2 + (r_2 - 1)s_2^2 + \cdots + (r_t - 1)s_t^2}{\sum_{i=1}^t r_i - t} \\ &= \frac{\sum_{j=1}^{r_1} (Y_{1j} - \bar{Y}_{1\cdot})^2 + \sum_{j=1}^{r_i} (Y_{2j} - \bar{Y}_{2\cdot})^2 + \cdots \sum_{j=1}^{r_i} (Y_{tj} - \bar{Y}_{t\cdot})^2}{\sum_{i=1}^t r_i - t}. \end{aligned}$$

We now divide by the total number of degrees of freedom,

$$(r_1 - 1) + \cdots (r_t - 1) = \sum_{i=1}^t r_i - t = N - t.$$

(Compare to the $t = 2$ case.)

As in the equal replication case, this numerator makes sense **regardless** of whether H_0 is true. Note that the estimate may be written

$$\sum_{i=1}^t \sum_{j=1}^{r_i} (Y_{ij} - \bar{Y}_{i\cdot})^2 \stackrel{def}{=} \text{ Error SS},$$

and we may thus write the denominator of our test statistic as

$$\frac{\text{Error SS}}{\text{degrees of freedom for error}} = \text{ Error MS}.$$

It is possible to show algebraically that the Error SS may be written as

$$\sum_{i=1}^t \sum_{j=1}^{r_i} (Y_{ij} - \bar{Y}_{i\cdot})^2 = \sum_{i=1}^t \sum_{j=1}^{r_i} Y_{ij}^2 - \sum_{i=1}^t \left(\frac{\left(\sum_{j=1}^{r_i} Y_{ij} \right)^2}{r_i} \right).$$

Treatment SS: If we want an estimator of σ^2 by considering deviations among treatments, as before, we need to consider the treatment sample means $\bar{Y}_{i\cdot}$. Note now that, if all the means really were the same and equal to μ, we would have

$$\bar{Y}_{i\cdot} \sim \mathcal{N}\left(\mu, \frac{\sigma^2}{r_i}\right), \tag{7.5}$$

We would still like our estimator for σ^2 based on the sample means to be based on the squared deviations $(\bar{Y}_{i\cdot} - \bar{Y}_{\cdot\cdot})^2$ as before; but, note from (7.5) that the sample means no longer have the same variance. However, (7.5) suggests that we might consider the squared deviations $r_i(\bar{Y}_{i\cdot} - \bar{Y}_{\cdot\cdot})^2$; the multiplication by r_i puts all squared deviations accounts for the different sample sizes, so that the sum of these "adjusted" squared deviations, divided by appropriate degrees of freedom, will again estimate σ^2. We thus consider

$$\sum_{i=1}^{t} r_i(\bar{Y}_{i\cdot} - \bar{Y}_{\cdot\cdot})^2 \stackrel{def}{=} \text{ Treatment SS.}$$

Note that is $r_i = r$ for all treatments $i = 1, \ldots, t$, then this reduces to the Treatment SS for equal replication.

This expression is as before based on $(t-1)$ **independent quantities**, so the degrees of freedom are $(t-1)$. We thus have our estimator for σ^2 based on sample means

$$\frac{\sum_{i=1}^{t} r_i(\bar{Y}_{i\cdot} - \bar{Y}_{\cdot\cdot})^2}{t-1} = \frac{\text{Treatment SS}}{\text{degrees of freedom for treatments}}.$$

Algebraically, it may be shown that

$$\text{Treatment SS} = \sum_{i=1}^{t}\left(\frac{\left(\sum_{j=1}^{r_i} Y_{ij}\right)^2}{r_i}\right) - C.$$

From the algebraic representations of Total, Treatment, and Error SS above, it is easy to see that we again have

$$\text{Total SS} \;=\; \text{Treatment SS} \;+\; \text{Error SS}.$$

Thus, the interpretation is the same as in the equal replication case: we have a ratio of two "independent" sample variances. This ratio will be **small** if there are really no differences among the treatment means $\mu_1, \ldots, \mu_t$, and **large** if they are. Formally, define now

$$MS_T = \frac{\text{Treatment SS}}{t-1} = \text{ Treatment MS}, \quad MS_E = \frac{\text{Error SS}}{\sum_{i=1}^{t} r_i - t} = \text{ Error MS}, \quad F = \frac{MS_T}{MS_E}.$$

The statistic F will have an $F_{t-1,N-t}$ distribution, where, as above, we have written $N = \sum_{i=1}^{t} r_i$ for brevity.

We summarize this information in a table:

One Way ANOVA table – Unequal Replication

Source of variation	DF	SS Definition	SS Computation	MS	F
Among Treatments	$t-1$	$\sum_{i=1}^{t} r_i(\bar{Y}_{i\cdot} - \bar{Y}_{\cdot\cdot})^2$	$\sum_{i=1}^{t} \left(\frac{\left(\sum_{j=1}^{r_i} Y_{ij}\right)^2}{r_i} \right) - C$	MS_T	F
Error (Within Treatments)	$\sum_{i=1}^{t} r_i - t$ $(= N - t)$	$\sum_{i=1}^{t}\sum_{j=1}^{r}(Y_{ij} - \bar{Y}_{i\cdot})^2$	by subtraction	MS_E	
Total	$N-1$	$\sum_{i=1}^{t}\sum_{j=1}^{r}(Y_{ij} - \bar{Y}_{\cdot\cdot})^2$	$\sum_{i=1}^{t}\sum_{j=1}^{r} Y_{ij}^2 - C$		

STATISTICAL HYPOTHESES: The question of interest is the same, is to determine if the means of the t treatment populations are different. We write

$$H_{0,T}: \mu_1 = \mu_2 = \cdots = \mu_t \text{ vs. } H_{1,T}: \text{ The } \mu_i \text{ are not all equal,}$$

or, equivalently under the model restriction $\sum_{i=1}^{t} \tau_i = 0$ as

$$H_{0,T}: \tau_1 = \tau_2 = \cdots = \tau_t = 0 \text{ vs. } H_{1,T}: \text{ The } \tau_i \text{ are not all equal.}$$

Again, remember that the the alternative hypothesis **does not specify** the **way** in which the treatment means (or deviations) differ.

TEST PROCEDURE: Reject $H_{0,T}$ in favor of $H_{1,T}$ at level of significance α if

$$F > F_{t-1,N-t,\alpha}.$$

EXAMPLE: (Zar, 1974, *Biostatistical Analysis*, p. 134.) Each of 19 randomly selected pigs is assigned at random to one of 4 diet regimes. The data are the body weights of the pigs, in pounds, after having been raised on the diets.

Here, the **individual pigs** are the **experimental units** as well as the **sampling units** (one measurement per pig). We assume that the measurements are approximately normally distributed, which seems reasonable for these data, with approximately constant variance σ^2.

We have $t = 4$, $N = \sum_{i=1}^{t} r_i = 19$.

	Diet 1	Diet 2	Diet 3	Diet 4
	133.8	151.2	225.8	193.4
	125.3	149.0	224.6	185.3
	143.1	162.7	220.4	182.8
	128.9	145.8	212.3	188.5
	135.7	153.5		198.6
r_i	5	5	4	5
$\sum_{j=1}^{r_i} Y_{ij}$	666.8	762.2	883.1	948.6
$\bar{Y}_{i\cdot}$	133.36	152.44	220.78	189.72

We calculate:

$$C = \frac{\left(\sum_{i=1}^{t}\sum_{j=1}^{r_i} Y_{ij}\right)^2}{N} = \frac{(666.8 + \cdots + 948.6)^2}{19} = \frac{10,632,164.49}{19} = 559,587.60$$

$$\sum_{i=1}^{t}\left(\frac{\left(\sum_{j=1}^{r_i} Y_{ij}\right)^2}{r_i}\right) = \frac{666.8^2}{5} + \cdots + \frac{948.6^2}{5} = 580,049.01.$$

$$\sum_{i=1}^{t}\sum_{j=1}^{r_i} Y_{ij}^2 = 133.8^2 + 125.3^2 + \cdots + 198.6^2 = 580,671.41.$$

Thus,

$$\text{Treatment SS} = 580,049.01 - 559,587.60 = 20461.41$$

$$\text{Total SS} = 580,671.41 - 559,587.60 = 21083.81$$

$$\text{Error SS} = \text{Total SS} - \text{Treatment SS} = 21083.81 - 20461.41 = 622.40.$$

We also have $t - 1 = 3$, $N - t = 19 - 4 = 15$, so that

$$MS_T = \frac{20461.41}{3} = 6820.47, \quad MS_E = \frac{622.40}{15} = 41.49, \quad F = \frac{6820.47}{41.49} = 164.39.$$

We summarize the computations in an analysis of variance table:

Analysis of Variance – Pig Weight Data

Source of variation	DF	SS	MS	F
Among Treatments	3	20461.41	6820.47	164.39
Error (Within Treatments)	15	622.40	41.49	
Total	18	21083.81		

To perform the hypothesis test for differences among the treatment means, we compare F to the appropriate value from the F table. For level of significance $\alpha = 0.05$, we have

$$F_{3,15,0.05} = 3.29,$$

so that $164.39 >> F_{3,15,0.05}$. We thus **Reject** $H_{0,T}$. There is strong evidence in these data to suggest that the mean weights are different under the different diets.

In section 7.13, we show how to use SAS to conduct this type of analysis.

7.9 A closer look at the F ratio

By thinking about the **additive linear model**, we may gain more insight into **why** the F ratio is a reasonable basis for assessing differences in treatment means when the treatment effects are best regarded as **fixed**. We may also see how the F ratio provides an appropriate test when the treatment effects are best regarded as **random effects**.

To simplify our discussion, we will look at the case of **equal replication** here. In the case of a one way classification with unequal replication, the basic idea is the same, but the arithmetic expressions and other considerations a more complicated. STD describe the results in their section 7.5.

RECALL: F is the ratio of two different estimates for variation among responses on experimental units, σ^2,

$$F = \frac{MS_T}{MS_E},$$

where MS_T is based on the differences among treatment sample means and MS_E on deviations "within" treatments.

We discussed the following notions. For our model

$$Y_{ij} = \mu + \tau_i + \epsilon_{ij},$$

- **Fixed treatment effects** – the τ_i are **fixed values** corresponding to the fixed amounts by which the mean for treatment i differs from the overall mean. The treatments in the experiment are the **only** ones of interest. This situation is sometimes called "Model I."
- **Random treatment effects** – the τ_i are **random variables** themselves. The particular τ_i in the experiment correspond to the deviations for the treatments that ended up in the experiment by **random sampling** from the **population** of all possible treatments. Thus, the τ_i have mean 0 and variance σ_τ^2 representing the population of all possible treatment deviations. This is sometimes called "Model II."

In the **fixed effects** case, under our linear additive model, $\mu_i = \mu + \tau_i$, and we test $H_{0,T} : \mu_1 = \cdots = \mu_t$, which, under the **restriction** $\sum_{i=1}^t \tau_i = 0$ becomes $H_{0,T} : \tau_1 = \cdots = \tau_t = 0$. We talk about the **random effects** case momentarily.

If we were to repeat an experiment of either type endlessly, then we may think of the **populations** of all possible values of the Treatment MS and Error MS, MS_T and MS_E, respectively, that might arise. These populations will have **means** of their own, for which MS_T and MS_E from a particular experiment are **estimates**. It is possible to derive the forms of these **means** for each case (fixed or random effects). These means are called **expected mean squares**, and thus describe what each MS is estimating. These are summarized below:

Source of variation	Degrees of freedom	Expected Mean Square	
		Fixed effects	Random effects
Treatments MS_T	$t-1$	$\sigma^2 + \frac{r\sum_{i=1}^t \tau_i^2}{t-1}$	$\sigma^2 + r\sigma_\tau^2$
Error MS_E	$t(r-1)$	σ^2	σ^2

FIXED EFFECTS: From this table, in the case where the τ_i are best regarded as having fixed effects, when we use MS_T as an estimate of σ^2 when forming the F ratio, what we are **really** estimating is the quantity

$$\sigma^2 + \frac{r\sum_{i=1}^{t}\tau_i^2}{t-1}.$$

The added term expresses the possible **extra variation** due to differences among treatment means (treatment deviations, τ_i).

$$\text{This extra variation } \frac{r\sum_{i=1}^{t}\tau_i^2}{t-1} \text{ is in terms of the fixed treatment effects.}$$

If $H_{0,T}$ is true, all the $\tau_i = 0$, and thus this term is itself zero. Thus, when $H_{0,T}$ is **true**, both MS_T and MS_E are reasonable estimators for σ^2, and the F ratio will be small, reflecting the fact that the extra term is 0. When $H_{0,T}$ is **not true**, the extra term will be nonzero, and the F ratio will be large.

RANDOM EFFECTS: From this table, in the case where the τ_i are best regarded as having random effects, when we use MS_T as an estimate of σ^2 when forming the F ratio, what we are **really** estimating is the quantity

$$\sigma^2 + r\sigma_\tau^2.$$

We quantify the **variation** in the population of all possible treatments by σ_τ^2, the **variance** of the random variables τ_i. Thus, we see that the the extra term expresses the possible **extra variation** due to the possibility of **variation** in the population of all treatment deviations. That is,

$$\text{This extra variation } r\sigma_\tau^2 \text{ is in terms of the variance of the random treatment effects.}$$

If there were **no variation** in this population (so that all possible treatments are **the same**), we would express this as $\sigma_\tau^2 = 0$, and we see that the extra variation term would be zero.

From this discussion, we see that when we are in the case of random treatment effects, the **hypotheses** we are really testing may be expressed as

$$H_{0,T} : \sigma_\tau^2 = 0 \text{ vs. } H_{1,T} : \sigma_\tau^2 > 0.$$

Under $H_{0,T}$, there is **no variation** in the population of all possible treatment effects τ_i, and all τ_i must have the same value. Furthermore, both MS_T and MS_E under $H_{0,T}$ are reasonable estimators for σ^2, and the F ratio will be small. When $H_{0,T}$ is **not true**, the extra term will be nonzero, and the F ratio will be large.

These observations formalize how we have been interpreting the F ratio in either case all along!

ESTIMATING σ^2: Also from the table, we see that, regardless of whether $H_{0,T}$ is true, MS_E is always a valid estimator for σ^2 – the mean of the population of all possible MS_E values is σ^2.

Often, as a "rough measure" of the inherent variability in an experiment (that not due to systematic effects of treatments, only to the variation due to experimental units), a **coefficient of variation** is computed. This CV is an estimate of the quantity

$$\frac{\sigma^2}{\mu}.$$

Under our interpretation of the model, μ is the average treatment mean, so gives a rough idea of the magnitude of "signal" in the data. Thus, this quantity represents the magnitude of inherent variation in responses on experimental units relative to the size of the thing being measured (recall the discussion in chapter 2); that is, the "noise" relative to the "signal." This quantity gives a sense of the "quality" of the information – if it is **large**, the inherent variability in experimental units is large, and inference may be difficult. If it is **small**, we have a low "noise-to-signal" ratio, so we should hope to be able to figure out what is going on!

The point is that the CV represents the magnitude of what we attributing to **inherent variation** in experimental material; it has nothing to do with the treatments, so measures something we have to "live with", unless we may identify **other systematic** sources of variation (that may have formed the basis for grouping).

The obvious estimator is

$$\frac{\sqrt{MS_E}}{\bar{Y}_{\cdot\cdot}}.$$

EXAMPLE: For the pea section data, we had $MS_E = 5.46$, so that $\sqrt{MS_E} = 2.34$, and $\bar{Y}_{\cdot\cdot} = 61.94$. Thus, the estimated CV is roughly

$$\frac{2.34}{61.94} = 0.038;$$

The estimate of CV is thus 3.8%, which is pretty small; this says that the size of the "signal" relative to the inherent "noise" is pretty low!

7.10 Subsampling and linear additive model

We have discussed previously the difference between an **experimental unit** and a **sampling unit**. The experimental unit, being the element of experimental material that received an application of the treatment, is the entity of interest for assessing **experimental error**. In many experiments, we may have several sampling units on each experimental unit.

EXAMPLE: Recall our example of administering rations (treatments) to pigs. Suppose we keep the swine in a pen of 20 animals, and feed them the treatment by introducing a trough of ration into the pen.

- **Experimental unit** – pen. The treatment is "applied" to the entire pen
- **Sampling unit** – individual animal. The weight gain for each pit is recorded at the end of the experiment.

IMPORTANCE: The importance of correctly identifying the experimental unit is as follows. In order to assess accurately the effects of the treatments, we must be assured that the treatments **did indeed** get applied as we intend, so that we may realistically attribute differences observed to the treatments. If, in the pig example, we treated individual pigs as the experimental unit, we see that this assurance would not be fulfilled – Once the ration trough is introduced into the pen, we have **no control** over how it is "applied" to individual pigs. Larger animals might "squeeze out" smaller, weaker ones from getting to the trough, so that the treatment would not be "applied" equally to all pigs in the pen. Thus, different animals in the pen might exhibit different weight gains simply because they **did not** receive the same "application" of the treatment! Treating the pigs as experimental units would incorrectly assume that they all **did** receive the same "application" of treatment!

MORAL: A way to think about the experimental unit is as the element of experimental material over which we have **control** over the application of the treatment.

LINEAR ADDITIVE MODEL: So far, we have discussed the analysis of variance procedure within the context of **one observation per experimental unit**; that is the experimental and sampling units were **the same**. We now consider how the procedure might be extended to the case of **more than one** sampling unit per experimental unit. This is referred to as **subsampling**.

To facilitate our discussion, it is convenient to think of a **linear additive model** for an observation under these conditions.

When we have subsampling, we may classify an **individual observation** (on a **sampling unit** now) as the kth subsample on the jth experimental unit receiving treatment i. We thus index each observation by three subscripts. We may think of the following model:

$$Y_{ijk} = \mu + \tau_i + \epsilon_{ij} + \delta_{ijk}, \quad i = 1, \ldots, t, \quad j = 1, \ldots, r_i, \quad k = 1, \ldots, s_{ij}.$$

Here,

- μ and τ_i are as before.
- t = number of treatments as before, r_i = number of replicates (experimental units) on treatment i, and s_{ij} = the number of sampling units (subsamples) on the jth experimental unit receiving treatment i.
- ϵ_{ij} = "error" associated with the jth experimental unit receiving treatment i. ϵ_{ij} quantifies how this experimental unit varies in the population of all experimental units.
- δ_{ijk} = additional "error" associated with the particular sampling unit (the kth), which we might refer to as **sampling error**.

To understand the distinction between the two "error" components, consider the following example. Suppose an experiment is conducted to compare the effect of 3 different doses of a toxic agent (treatments) on the birth weights of rats. For each dose, several pregnant female rats are given the particular dose. Thus, the female rats are the experimental unit, as they receive an application of the treatment. For each mother, suppose that a birth weight is recorded for each rat pup. Thus, the rat pups are the sampling units. For a given mother rat, all of her pups are of course not **identical**; rather, they exhibit variability among themselves. Furthermore, mother rats vary inherently across themselves. In the model, ϵ_{ij} characterizes the mother-to-mother variation. The δ_{ijk} characterizes the **additional** variation that might be present because all rat pups on a given mother are not exactly alike.

EXPERIMENTAL ERROR: How we think about **experimental error** under these conditions becomes a little more complex. Recall that we think of experimental error as measuring the inherent variation in responses on experimental units; that is, the variation in the data that we attribute to things other than (the systematic effects of) the treatments. If we think about this variation in the current context, it is clear that there are now **two** sources of inherent variation that may make responses on experimental units differ: variation due to differences among experimental units and among sampling units within them.

Thus, if we wish to assess how much of the **overall variation** in the data is due to systematic effects of the treatments, we must weigh this against the variation in the data due to inherent, unexplained sources. Both variation among experimental units (e.g. mother rats) **and** among sampling units (e.g. rat pups within mother rats) contribute to this latter variation.

The result is that our assessment of **experimental error** must measure the variation **both** among and within experimental units – in our linear additive model, then, it must measure the **total** variability associated with the error terms ϵ_{ij} **and** δ_{ijk}. We will see how this is done shortly.

TERMINOLOGY: A model such as this is called a **nested** model. The data may be classified according to a **hierarchical** structure: experimental units within treatment groups and then sampling units within experimental units. The units at the "inner" level are entirely contained within a unit at the "outer" level of this hierarchy, hence the term "nested." In the rat example, rat pups are **nested** within mother rats (which in turn are **nested** within treatment dose groups).

7.11 ANOVA for one way classification with equal replication and subsampling

Although the model and the notion of experimental error seem much more complicated in this case than when there was only one sampling unit per experimental unit, the principles for developing a sensible test of whether treatment means differ are the same. In particular, we again construct a F ratio of the following form:

$$F = \frac{\text{Estimate of all variation, including that due to treatment mean differences}}{\text{Estimate of all variation except that due to treatment mean differences}}.$$

Such a ratio will be an indication of how much of the variation is attributable to differences in treatment means.

To construct such a ratio formally, we must think of appropriate MSs for the numerator and denominator.

EQUAL REPLICATION CASE: We will consider the case of equal replication and equal numbers of sampling units per experimental unit, that is,

$$r_i = r, \qquad s_{ij} = s.$$

That is, on each treatment there are r experimental units, each with s sampling units.

We will discuss the implications of unequal numbers of sampling units at the end of this section.

NOTATION: Define

$$\bar{Y}_{i\cdot\cdot} = \frac{1}{rs}\sum_{j=1}^{r}\sum_{k=1}^{s} Y_{ijk} = \text{ mean for treatment } i$$

$$\bar{Y}_{ij\cdot} = \frac{1}{s}\sum_{k=1}^{s} Y_{ijk} = \text{ mean on sampling units for } j\text{th replicate on treatment } i$$

$$\bar{Y}_{\cdots} = \frac{1}{trs}\sum_{i=1}^{t}\sum_{j=1}^{r}\sum_{k=1}^{s} Y_{ijk} = \textbf{ overall} \text{ sample mean } = \frac{1}{t}\sum_{i=1}^{t}\bar{Y}_{i\cdots}$$

For algebraic convenience, denote the **correction factor** as

$$C = \frac{\left(\sum_{i=1}^{t}\sum_{j=1}^{r}\sum_{k=1}^{s} Y_{ijk}\right)^2}{trs}.$$

As in the simpler case of one sampling unit per experimental unit, a measure of the **overall**, "**total**" variation in the data, that does not take into account that treatments were applied or the fact that there are both experimental units and sampling units within them, would again be the **sample variance** for **all the responses**, lumping them together without regard to treatments or type of unit. There are trs total observations; thus, this sample variance is

$$\frac{\sum_{i=1}^{t}\sum_{j=1}^{r}\sum_{k=1}^{s}(Y_{ijk} - \bar{Y}_{\cdots})^2}{trs - 1}.$$

It may be shown that the numerator may be expressed as

$$\sum_{i=1}^{t}\sum_{j=1}^{r}\sum_{k=1}^{s}(Y_{ijk} - \bar{Y}_{\cdots})^2 = \sum_{i=1}^{t}\sum_{j=1}^{r}\sum_{k=1}^{s} Y_{ijk}^2 - C.$$

By analogy to the simpler case, we call this quantity the **Total SS**, measuring as before **total variation** in the data.

By analogy to the case where we had one sampling unit per experimental unit, the **numerator** of a suitable F ratio for testing treatment differences should be an estimator of variance including a component of variation among the treatment means. Intuitively, this would be based on the deviations

$$(\bar{Y}_{i\cdot\cdot} - \bar{Y}_{\cdot\cdot\cdot}),$$

the deviations of the treatment sample means from the overall average. In particular, define

$$\text{Treatment SS} = rs\sum_{i=1}^{t}(\bar{Y}_{i\cdot\cdot} - \bar{Y}_{\cdot\cdot\cdot})^2 = \frac{\sum_{i=1}^{t}\left(\sum_{j=1}^{r}\sum_{k=1}^{s}Y_{ijk}\right)^2}{rs} - C.$$

The appropriate variance estimator would be the sample variance of the $\bar{Y}_{i\cdot\cdot}$'s, and would thus be based on $(t-1)$ degrees of freedom. The appropriate estimator is thus

$$\text{Treatment MS} = \frac{\text{Treatment SS}}{t-1}.$$

Again by analogy to the simpler case, for the **denominator**, we need an estimator of variance reflecting **experimental error**. From our discussion above, this includes **all** variation **except** that due to the treatments. Intuitively, this should be based on the quantities

$$(\bar{Y}_{ij\cdot} - \bar{Y}_{i\cdot\cdot}).$$

The $\bar{Y}_{ij\cdot}$'s represent the information on individual experimental units (averaged over sampling units); thus, this characterizes how individual experimental units vary about their treatment means. In particular, define

$$\text{Experimental Error SS} = \sum_{i=1}^{t}\sum_{j=1}^{r}s(\bar{Y}_{ij\cdot} - \bar{Y}_{i\cdot\cdot})^2 = \frac{\sum_{i=1}^{t}\sum_{j=1}^{r}\left(\sum_{k=1}^{s}Y_{ijk}\right)^2}{s} - \frac{\sum_{i=1}^{t}\left(\sum_{j=1}^{r}\sum_{k=1}^{s}Y_{ijk}\right)^2}{rs}$$

algebraically.

The appropriate variance estimator would be Experimental Error SS divided by its degrees of freedom. Note that there are tr experimental unit averages $\bar{Y}_{ij\cdot}$ and t treatment sample means $\bar{Y}_{i\cdot\cdot\cdot}$. This suggests that there are $tr - t = t(r-1)$ independent quantities involved in this SS. To verify this, consider the following argument.

Consider the quantities

$$(\bar{Y}_{ij\cdot} - \bar{Y}_{\cdots}).$$

These measure the deviation of individual experimental units about the overall average, thus, **do not** take into account the effects possible differences among treatment means in assessing this variation. Define accordingly

$$\text{Among Experimental Units SS } = \sum_{i=1}^{t}\sum_{j=1}^{r} s(\bar{Y}_{ij\cdot} - \bar{Y}_{\cdots})^2 = \frac{\sum_{i=1}^{t}\sum_{j=1}^{r}\left(\sum_{j=1}^{r}\sum_{k=1}^{s} Y_{ijk}\right)^2}{s} - C$$

algebraically. This SS clearly is based on $rt - 1$ independent quantities, because (try it)

$$\sum_{i=1}^{t}\sum_{j=1}^{r}(\bar{Y}_{ij\cdot} - \bar{Y}_{\cdots}) = 0,$$

so we can always reconstruct the final experimental unit mean if we know the rest. Thus, the "Among Experimental Unit SS" has $rt - 1$ degrees of freedom. It is important not to confuse this SS with the Experimental Error SS above – this one is not useful, except for helping understand degrees of freedom (coming up), as it measures variation in experimental units **including** that due to treatments.

Note from the algebraic representations of these SSs that we have

$$\text{Among Experimental Units SS } = \text{ Treatment SS } + \text{ Experimental Error SS.}$$

We thus see that this SS may be broken down into two "independent" components. The total degrees of freedom are $rt - 1$, as above, and Treatment SS has $t - 1$ degrees of freedom. Thus, we are left with

$$(rt - 1) - (t - 1) = t(r - 1),$$

as expected. Thus, define

$$\text{Experimental Error MS } = \frac{\text{Experimental Error SS}}{t(r-1)}.$$

We now have our two "variance estimators" to form an F ratio suitable for testing treatment differences. However, before we finish, we note that we may characterize yet another part of the overall variation in the data. Recall that we have two sources of variation – that among experimental units and that among sampling units. Experimental error takes into account both. We may in fact identify that due **only** to the tendency for sampling units within an experimental unit to vary. Intuitively, an estimate for this type of variation would be based on the quantities

$$(Y_{ijk} - \bar{Y}_{ij\cdot}),$$

which measures how responses on individual sampling units vary about the average for the experimental unit from which they came. A SS which measures this variation would thus be

$$\sum_{i=1}^{t}\sum_{j=1}^{r}\sum_{k=1}^{s}(Y_{ijk} - \bar{Y}_{ij\cdot})^2 = \sum_{i=1}^{t}\sum_{j=1}^{r}\sum_{k=1}^{s} Y_{ijk}^2 - \frac{\sum_{i=1}^{t}\sum_{j=1}^{r}\left(\sum_{k=1}^{s} Y_{ijk}\right)^2}{s}$$

algebraically. It is natural to refer to this SS as **Sampling Error (Within Experimental Units) SS**.

From the algebraic representations above, observe that

$$\text{Total SS} \;=\; \text{Treatment SS} \;+\; \text{Experimental Error SS} \;+\; \text{Sampling Error SS}.$$

That is, the **total variation** in the data may be partitioned into three components – that taking into account variation among experimental units and treatment means, that due to total variation in experimental units except for the treatments, and that due to variation on sampling units within experimental units. Inspecting the degrees of freedom, we see that

$$trs - 1 = (t-1) + t(r-1) + tr(s-1),$$

so that Sampling Error SS should have $tr(s-1)$ degrees of freedom.

We may now collect this altogether and contemplate testing. Define

$$MS_T = \frac{\text{Treatment SS}}{t-1}$$

and

$$MS_E = \frac{\text{Experimental Error SS}}{t(r-1)}.$$

The ratio

$$F_T = \frac{MS_T}{MS_E}$$

will be large if there really are differences in treatment means, as the numerator will be large relative to the denominator.

We may also define

$$MS_S = \frac{\text{Sampling Error SS}}{tr(s-1)}.$$

Recall that the Experimental Error SS, and hence MS_E, characterizes variation among experimental units due to **all sources except** the treatments applied, including both that due to inherent differences in experimental units **and** differences in sampling units within them. Sampling Error SS, and hence MS_S, represents only the latter source. Thus, intuitively, we would expect the ratio

$$F_S = \frac{MS_E}{MS_S}$$

to be large if there were **additional** variation inherent in experimental units beyond that due just to sampling within them. Thus, using F_S, we also have the possibility of investigating the nature of variation among experimental units.

It may be shown, because of the "independence" of the various SSs, that

$$F_T \sim F_{t-1,t(r-1)} \text{ and } F_S \sim F_{t(r-1),tr(s-1)}.$$

We may summarize all of this in an analysis of variance table; we describe the computations below.

One Way ANOVA table – Subsampling with Equal Replication

Source of variation	DF	SS Definition	MS	F
Among Treatments	$t-1$	$rs\sum_{i=1}^{t}(\bar{Y}_{i\cdot\cdot}-\bar{Y}_{\cdots})^2$	MS_T	F_T
Experimental Error	$t(r-1)$	$s\sum_{i=1}^{t}\sum_{j=1}^{r}(\bar{Y}_{ij\cdot}-\bar{Y}_{i\cdot\cdot})^2$	MS_E	F_S
Sampling Error	$tr(s-1)$	$\sum_{i=1}^{t}\sum_{j=1}^{r}\sum_{k=1}^{s}(Y_{ijk}-\bar{Y}_{ij\cdot})^2$	MS_S	
Total	$trs-1$	$\sum_{i=1}^{t}\sum_{j=1}^{r}\sum_{k=1}^{s}(Y_{ijk}-\bar{Y}_{\cdots})^2$		

COMPUTATION: It is easiest to use the following procedure:

1. Calculate the correction factor

$$C=\frac{\left(\sum_{i=1}^{t}\sum_{j=1}^{r}\sum_{k=1}^{s}Y_{ijk}\right)^2}{trs}$$

and the Total SS

$$\sum_{i=1}^{t}\sum_{j=1}^{r}\sum_{k=1}^{s}Y_{ijk}^2-C.$$

2. Calculate the Treatment SS

$$\frac{\sum_{i=1}^{t}\left(\sum_{j=1}^{r}\sum_{k=1}^{s}Y_{ijk}\right)^2}{rs}-C.$$

3. Calculate the Among Experimental Units SS

$$\frac{\sum_{i=1}^{t}\sum_{j=1}^{r}\left(\sum_{k=1}^{s}Y_{ijk}\right)^2}{s}-C$$

4. Find the Experimental Error SS by subtraction:

$$\text{Experimental Error SS} = \text{Among Experimental Units SS} - \text{Treatment SS}.$$

5. Find the Sampling Error SS by subtraction:

$$\text{Sampling Error SS} = \text{Total SS} - \text{Among Experimental Units SS}.$$

Thus, although Among Experimental Units SS is not an interesting quantity for our tests, it is useful for computation.

STATISTICAL HYPOTHESES: The major question of interest is to determine if the means of the t treatment populations differ:

$$H_{0,T}: \mu_1 = \cdots = \mu_t \text{ vs. } H_{1,T}: \text{ The } \mu_i \text{ are not all equal}$$

(for the fixed effects case).

As noted above, another hypothesis we may wish to test, and for which we have information available because of the subsampling, is to determine whether experimental error is due mainly just to variation among individual sampling units or if in fact environmental or other differences among experimental units are greater than those within experimental units.

To state this hypothesis formally, and to gain insight into the nature of both hypothesis tests and the suitability of the F ratios for testing them, we construct a table of **expected mean squares** under our linear additive model, just as we did in the simpler case of one sampling unit per experimental unit. In our model, we have **two** random errors:

- ϵ_{ij}, representing errors due to experimental units. Call the **variance** of this error, representing the variance in the population of experimental units, σ^2_ϵ.
- δ_{ijk}, representing errors due to sampling units. Call the variance of this error, representing the variance in the population of such errors, σ^2.

CONVENTION: It is convention in most texts on analysis of variance to use the symbol "σ^2" to denote the variance associated with the "smallest" unit of measurement; that is, the unit on which **individual observations** arise. Here, this is the sampling unit. In the case of one sampling unit per experimental unit, we used σ^2 to denote the variance in the population of experimental units (the "smallest" unit of measurement in that setting); here, however, the usage is different. We probably should have called this variance σ^2_ϵ as above, to be consistent, but this is typically not how it is done, unfortunately. So, keep in mind that here σ^2_ϵ is the variance of interest with regard to experimental units.

Source of variation	Degrees of freedom	Expected mean square
Treatments MS_T	$t-1$	$\sigma^2 + s\sigma^2_\epsilon + \left(\frac{tr\sum_{i=1}^t \tau_i^2}{t-1} \text{ or } tr\sigma^2_\tau\right)$ (depending on whether τ_i are fixed or random effects)
Experimental Error MS_E	$t(r-1)$	$\sigma^2 + s\sigma^2_\epsilon$
Sampling Error MS_S	$tr(s-1)$	σ^2

Several things may be gleaned from the table:

- Whether the treatment effects are **fixed** or **random**, the statistic F_T is appropriate. Note that the MS_E estimates

$$\sigma^2 + s\sigma^2_\epsilon,$$

which takes into account variation both among and within experimental units. MS_T estimates the same quantity **plus** a term representing the **extra variation** due to treatments

$$\frac{tr\sum_{i=1}^t \tau_i^2}{t-1} \text{ or } tr\sigma^2_\tau.$$

Thus, F_T is indeed expected to be **large** if there really are differences among treatments.

- In the case of **random** effects, the hypotheses about treatments should be written

$$H_{0,T}: \sigma^2_\tau = 0 \text{ vs. } H_{1,T}: \sigma^2_\tau > 0.$$

- The hypotheses tested by F_S are thus

$$H_{0,S}: \sigma^2_\epsilon = 0 \text{ vs. } H_{1,S}: \sigma^2_\epsilon > 0.$$

TEST PROCEDURE: For testing $H_{0,T}$ vs. $H_{1,T}$ at level of significance α, reject $H_{0,T}$ if

$$F_T > F_{t-1,t(r-1),\alpha}.$$

For testing $H_{0,S}$ vs. $H_{1,S}$ at level α, reject $H_{0,S}$ if

$$F_T > F_{t(r-1),tr(s-1),\alpha}.$$

EXAMPLE: (Zar, 1974, *Biostatistical Analysis*, p. 144.) Three different drugs for the treatment of high cholesterol are produced by three different manufacturers, each of which produces its drug at one of 2 different plants, as shown below. The measurements given are cholesterol concentrations (mg/100 ml of plasma) for human females treated with the drugs. Assume that a reasonable model for the distribution of such measurements is a normal distribution. A question of interest is whether cholesterol levels for female subjects differ among the three drugs (thus, manufacturers).

Drug (Manufacturer) 1		Drug (Manufacturer) 2		Drug (Manufacturer) 3	
Plants		Plants		Plants	
P_{11}	P_{12}	P_{21}	P_{22}	P_{31}	P_{32}
102	103	108	109	104	105
104	104	110	108	106	107

The 2 plants are different for each of the different drugs (manufacturers). Furthermore, for each plant, 2 different females are used. Thus, the only classification scheme for these data that makes sense is a **nested** one, with

Treatments	Drugs (Manufacturers)	$t = 3$
Experimental Units	Plants	$r = 2$
Sampling Units	Female subjects	$s = 2$

Calculations:

$$C = \frac{\left(\sum_{i=1}^{t}\sum_{j=1}^{r}\sum_{k=1}^{s} Y_{ijk}\right)^2}{trs} = \frac{1270^2}{12} = 143,408.33$$

$$\text{Total SS } = \sum_{i=1}^{t}\sum_{j=1}^{r}\sum_{k=1}^{s} Y_{ijk}^2 - C = 134,480.00 - 134,408.33 = 71.67$$

$$\text{Treatment SS } = \frac{\sum_{i=1}^{t}\left(\sum_{j=1}^{r}\sum_{k=1}^{s} Y_{ijk}\right)^2}{rs} - C = 134,369.50 - 134,408.33 = 61.17$$

$$\text{Among Experimental Units SS } = \frac{\sum_{i=1}^{t}\sum_{j=1}^{r}\left(\sum_{k=1}^{s} Y_{ijk}\right)^2}{s} - C = 134,471.00 - 134,408.33 = 62.67$$

Thus,

$$\text{Experimental Error SS } = 62.67 - 61.17 = 1.50$$

$$\text{Sampling Error SS } = 71.67 - 62.67 = 9.00.$$

Source of variation	DF	SS	MS	F
Among Treatments	2	61.17	30.58	61.16
Experimental Error	3	1.50	0.50	0.33
Sampling Error	6	9.00	1.50	
Total	11	71.67		

We use level of significance $\alpha = 0.05$.

For testing $H_{0,T}$ vs. $H_{1,T}$, we have $F_{2,3,0.05} = 9.55$ and $F_T = 61.16 > 9.55$. **Reject** $H_{0,T}$; there is strong evidence in these data to suggest that there is a difference in mean cholesterol level for the 3 drugs.

For testing $H_{0,S}$ vs. $H_{1,S}$, we have $F_{3,9,0.05} = 4.76$ and $F_S = 0.33$. **Do not reject** $H_{0,S}$; there is not enough evidence to suggest that there is a difference in mean cholesterol level due to a difference among plants within manufacturers for the 3 treatments.

UNEQUAL REPLICATION AND NUMBERS OF SUBSAMPLES: When the numbers of replicates and/or subsamples are not the same, the same general procedure we used above may be used to construct an analysis of variance table, partitioning Total SS into components. However, the **imbalance** leads to some difficulties. Degrees of freedom become more difficult to calculate. Recall in our table of expected mean squares above, each expected MS was equal to the one below it **plus** a term representing **extra variation**. This nice property no long holds exactly when the numbers of subsamples are not the same. Intuitively, if we have different numbers of subsamples on each experimental unit, the **quality** of information on each experimental unit is **different**. We would thus expect trying to sort out the different sources of variation to be much harder.

The result is that **exact** tests of hypotheses may no longer be carried out. Rather, **approximate tests** must be conducted. This is analogous to the situation where we had 2 populations with **unequal variances** in chapter 5 – we had to resort to an **approximate** t test. The basic problem is the same – we no longer have the same quality of information (measured by variance) on all experimental units under study.

See STD, section 7.8, for a description of the procedure. Just keep in mind that **care** must be taken under these circumstances.

7.12 Variance components

We have seen in our linear additive models for the one way classification, both with and without subsampling, that we construct MSs to estimate the variability due to different sources, e.g. treatments, experimental units, sampling units.

Recall from our tables of expected means squares that this involves **variances** of various error terms, e.g. in the case of **random treatment effects** and subsampling with equal replication and numbers of subsamples, MS_T estimates

$$\sigma^2 + s\sigma^2_\epsilon + tr\sigma^2_\tau.$$

Quantities such as σ^2, σ^2_ϵ, and σ^2_τ are called **variance components**. Here, the variability we might observe across treatments is associated with 3 sources: individual sampling units (σ^2), experimental units (σ^2_ϵ), and treatments (σ^2_τ).

In an experiment involving subsamples, we may in the planning stages wonder how best to allocate the available resources. We might

- have many experimental units with few sampling units on each
- have less experimental units with more sampling units on each

Generally, as we have discussed, we would like to have enough experimental units to get a good idea of the population; however, we may fine-tune how we do this if we understand the relative sizes of experimental and sampling error, as well as considerations such as whether subsamples are expensive or difficult to obtain or whether it is impractical to have too many experimental units.

Previous experimental data may be used to plan future experiments and give the investigator information on these questions. We may **estimate** the relative magnitudes of the variance components σ^2 and σ^2_ϵ to help guide our planning.

- MS_E estimates $\sigma^2 + s\sigma^2_\epsilon$.
- MS_S estimates σ^2; call this $\hat{\sigma}^2$.
- Thus, by algebra, we may estimate σ^2_ϵ by

$$\hat{\sigma}^2_\epsilon = \frac{MS_E - MS_S}{s}.$$

Given such estimates of σ^2 and σ^2_ϵ, we may now consider future experiments. Say we wish to conduct a future experiment with t treatments and are considering different numbers of replicates, r^*, say, and different numbers of sampling units on each, s^*. We may compute MS_E and the associated degrees of freedom $tr^*(s^* - 1)$ for each different scenario, and compare. At what point does increasing r^* and decreasing s^* and vice versa not really add anything? This information, along with associated costs of doing the experiment different ways, can be used to decide how best to use resources in the future. See STD, section 7.9, for more.

7.13 Using SAS to perform analysis of variance for one way classification

Here, we give three examples of using the SAS procedure `PROC GLM`, which stands for "General Linear Model," to construct the analysis of variance for one way classification experiments. As the name suggests, thinking about a linear additive model for the data helps with use of the procedure.

`PROC GLM` requires that the data be in a temporary data set with variables denoting the treatments and, if subsampling is present, the experimental units. Such variables are referred to as `CLASS` (`CLASSIFICATION`) variables, as they tell SAS how observations are to be classified. For example, in the pea section data, with treatments 5 different sugar mixtures, we might use a `CLASS` variable `SUGAR`.

`PROC GLM` also requires a `MODEL` statement. The `MODEL` statement tells the procedure how to construct the analysis. For example, consider the one way classification model with one sampling unit per experimental unit:

$$Y_{ij} = \mu + \tau_i + \epsilon_{ij}.$$

Suppose treatments are classified according to a `CLASS` variable `A`. To tell SAS that the analysis of variance corresponding to this model is desired, we would use the `PROC GLM` statement

```
MODEL = A;
```

Note that the `MODEL` statement does not include a symbol for μ; this is understood. It also does not include a symbol for the "smallest" unit of measurement error term, which is ϵ_{ij} in this case. In the model with subsampling (equal numbers of replicates, sampling units),

$$Y_{ijk} = \mu + \tau_i + \epsilon_{ij} + \delta_{ijk},$$

this is a bit harder. See the third example for an illustration.

These ideas are best illustrated by examples.

EXAMPLE 1: One way classification with equal numbers of replicates per treatment – the pea section data The `MEANS` statement used with `PROC GLM` here is **optional**; it requests that the sample means corresponding to the `CLASS` variables listed in the statement be printed.

PROGRAM:

```
*********************************************;
*                                           ;
*          EXAMPLE 7.1     ST 511           ;
*                                           ;
*      USING PROC GLM IN THE CASE OF        ;
*    EQUAL REPLICATION WITH ONE OBSERVATION ;
*          PER EXPERIMENTAL UNIT            ;
*     THE PEA SECTION DATA OF SOKAL & ROHLF ;
*                                           ;
*********************************************;

*****************************************************;
*                                                   ;
*   INVOKE SAS OPTIONS STATEMENT TO FORMAT OUTPUT    ;
*   HERE, WE LIMIT THE WIDTH OF THE OUTPUT TO 80     ;
*   CHARACTERS AND THE LENGTH OF THE PAGE TO 59      ;
*   LINES                                           ;
*                                                   ;
*****************************************************;

OPTIONS LS=80 PS=59 NODATE;

DATA PEAS;
  INPUT SUGAR $ LENGTH @@;
  CARDS;
CNTL  75  CNTL  67  CNTL  70  CNTL  75  CNTL  65  CNTL  71
CNTL  67  CNTL  67  CNTL  76  CNTL  68
GLU2  57  GLU2  58  GLU2  60  GLU2  59  GLU2  62  GLU2  60
GLU2  60  GLU2  57  GLU2  59  GLU2  61
FRU2  58  FRU2  61  FRU2  56  FRU2  58  FRU2  57  FRU2  56
FRU2  61  FRU2  60  FRU2  57  FRU2  58
GLU1FRU1  58  GLU1FRU1  59  GLU1FRU1  58  GLU1FRU1  61
```

```
GLU1FRU1  57  GLU1FRU1  56  GLU1FRU1  58  GLU1FRU1  57
GLU1FRU1  57  GLU1FRU1  59
SUC2  62  SUC2  66  SUC2  65  SUC2  63  SUC2  64  SUC2  62
SUC2  65  SUC2  65  SUC2  62  SUC2  67
;
PROC PRINT;
  TITLE 'THE PEA DATA OF SOKAL & ROHLF';
  TITLE2 'ONE WAY CLASSIFICATION BY TREATMENT';
  TITLE3 'TREATMENT = SUGAR (5 LEVELS)'; RUN;
*;
PROC GLM;
  CLASS SUGAR;
  MODEL LENGTH = SUGAR;
  MEANS SUGAR; RUN;
```

OUTPUT:

```
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%

                    THE PEA DATA OF SOKAL & ROHLF                         1
                  ONE WAY CLASSIFICATION BY TREATMENT
                     TREATMENT = SUGAR (5 LEVELS)

                      OBS    SUGAR      LENGTH

                        1    CNTL         75
                        2    CNTL         67
                        3    CNTL         70
                        4    CNTL         75
                        5    CNTL         65
                        6    CNTL         71
                        7    CNTL         67
                        8    CNTL         67
                        9    CNTL         76
                       10    CNTL         68
                       11    GLU2         57
                       12    GLU2         58
                       13    GLU2         60
                       14    GLU2         59
                       15    GLU2         62
                       16    GLU2         60
                       17    GLU2         60
                       18    GLU2         57
                       19    GLU2         59
                       20    GLU2         61
                       21    FRU2         58
                       22    FRU2         61
                       23    FRU2         56
                       24    FRU2         58
```

25	FRU2	57
26	FRU2	56
27	FRU2	61
28	FRU2	60
29	FRU2	57
30	FRU2	58
31	GLU1FRU1	58
32	GLU1FRU1	59
33	GLU1FRU1	58
34	GLU1FRU1	61
35	GLU1FRU1	57
36	GLU1FRU1	56
37	GLU1FRU1	58
38	GLU1FRU1	57
39	GLU1FRU1	57
40	GLU1FRU1	59
41	SUC2	62
42	SUC2	66
43	SUC2	65
44	SUC2	63
45	SUC2	64
46	SUC2	62
47	SUC2	65
48	SUC2	65
49	SUC2	62
50	SUC2	67

%%%

THE PEA DATA OF SOKAL & ROHLF 2

ONE WAY CLASSIFICATION BY TREATMENT

TREATMENT = SUGAR (5 LEVELS)

General Linear Models Procedure

Class Level Information

Class	Levels	Values
SUGAR	5	CNTL FRU2 GLU1FRU1 GLU2 SUC2

Number of observations in data set = 50

%%%

THE PEA DATA OF SOKAL & ROHLF 3

ONE WAY CLASSIFICATION BY TREATMENT

TREATMENT = SUGAR (5 LEVELS)

General Linear Models Procedure

Dependent Variable: LENGTH

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	4	1077.3200000	269.3300000	49.37	0.0001
Error	45	245.5000000	5.4555556		
Corrected Total	49	1322.8200000			

R-Square	C.V.	Root MSE	LENGTH Mean
0.814412	3.770928	2.3357131	61.940000

```
Source                  DF      Type I SS     Mean Square   F Value     Pr > F

SUGAR                    4   1077.3200000     269.3300000     49.37     0.0001

Source                  DF    Type III SS     Mean Square   F Value     Pr > F

SUGAR                    4   1077.3200000     269.3300000     49.37     0.0001



%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%

                         THE PEA DATA OF SOKAL & ROHLF                        4
                      ONE WAY CLASSIFICATION BY TREATMENT
                          TREATMENT = SUGAR (5 LEVELS)

                        General Linear Models Procedure

                 Level of        ------------LENGTH-----------
                 SUGAR       N        Mean                SD

                 CNTL       10     70.1000000       3.98469293
                 FRU2       10     58.2000000       1.87379591
                 GLU1FRU1   10     58.0000000       1.41421356
                 GLU2       10     59.3000000       1.63639169
                 SUC2       10     64.1000000       1.79195734
```

EXAMPLE 2: One way classification with unequal numbers of replicates – the pig diet data.

PROGRAM:

```
***********************************************;
*                                             ;
*       EXAMPLE 7.2        ST 511             ;
```

```
*                                                ;
*     USING PROC GLM IN THE CASE OF UNEQUAL      ;
*    REPLICATION IN A ONE-WAY CLASSIFICATION     ;
*    ILLUSTRATED BY THE PIG DIET DATA OF ZAR     ;
*                                                ;
************************************************;

*****************************************************;
*                                                   ;
*   INVOKE SAS OPTIONS STATEMENT TO FORMAT OUTPUT    ;
*   HERE, WE LIMIT THE WIDTH OF THE OUTPUT TO 80     ;
*   CHARACTERS AND THE LENGTH OF THE PAGE TO 59      ;
*   LINES                                           ;
*                                                   ;
*****************************************************;

OPTIONS LS=80 PS=59 NODATE;

DATA PIGS;
  INPUT DIET $ WEIGHT @@;
  CARDS;
DIET1  133.8  DIET1  125.3  DIET1  143.1  DIET1  128.9  DIET1  135.7
DIET2  151.2  DIET2  149.0  DIET2  162.7  DIET2  145.8  DIET2  153.5
DIET3  225.8  DIET3  224.6  DIET3  220.4  DIET3  212.3
DIET4  193.4  DIET4  185.3  DIET4  182.8  DIET4  188.5  DIET4  198.6
;
PROC PRINT;
  TITLE 'THE PIG DIET DATA OF ZAR';
  TITLE2 'ILLUSTRATION OF USING PROC GLM FOR A ONE-WAY ANOVA';
  TITLE3 'WITH UNEQUAL REPLICATION'; RUN;
*;
PROC GLM;
  CLASS DIET;
  MODEL WEIGHT = DIET;
```

```
  MEANS DIET; RUN;
```

OUTPUT:

```
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%

                            THE PIG DIET DATA OF ZAR                            1
              ILLUSTRATION OF USING PROC GLM FOR A ONE-WAY ANOVA
                            WITH UNEQUAL REPLICATION

                             OBS    DIET     WEIGHT

                               1    DIET1     133.8
                               2    DIET1     125.3
                               3    DIET1     143.1
                               4    DIET1     128.9
                               5    DIET1     135.7
                               6    DIET2     151.2
                               7    DIET2     149.0
                               8    DIET2     162.7
                               9    DIET2     145.8
                              10    DIET2     153.5
                              11    DIET3     225.8
                              12    DIET3     224.6
                              13    DIET3     220.4
                              14    DIET3     212.3
                              15    DIET4     193.4
                              16    DIET4     185.3
                              17    DIET4     182.8
                              18    DIET4     188.5
                              19    DIET4     198.6

%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
```

THE PIG DIET DATA OF ZAR 2

ILLUSTRATION OF USING PROC GLM FOR A ONE-WAY ANOVA

WITH UNEQUAL REPLICATION

General Linear Models Procedure

Class Level Information

Class	Levels	Values
DIET	4	DIET1 DIET2 DIET3 DIET4

Number of observations in data set = 19

%%%

THE PIG DIET DATA OF ZAR 3

ILLUSTRATION OF USING PROC GLM FOR A ONE-WAY ANOVA

WITH UNEQUAL REPLICATION

General Linear Models Procedure

Dependent Variable: WEIGHT

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	20461.405763	6820.468588	164.38	0.0001
Error	15	622.399500	41.493300		
Corrected Total	18	21083.805263			

R-Square	C.V.	Root MSE	WEIGHT Mean
0.970480	3.753460	6.4415293	171.61579

Source	DF	Type I SS	Mean Square	F Value	Pr > F
DIET	3	20461.405763	6820.468588	164.38	0.0001

Source	DF	Type III SS	Mean Square	F Value	Pr > F
DIET	3	20461.405763	6820.468588	164.38	0.0001

%%%

THE PIG DIET DATA OF ZAR 4

ILLUSTRATION OF USING PROC GLM FOR A ONE-WAY ANOVA

WITH UNEQUAL REPLICATION

General Linear Models Procedure

Level of DIET	N	WEIGHT Mean	WEIGHT SD
DIET1	5	133.360000	6.80793654
DIET2	5	152.440000	6.40023437
DIET3	4	220.775000	6.10593973
DIET4	5	189.720000	6.35035432

EXAMPLE 3: One way classification with equal replication and subsampling (equal numbers of sampling

units) – the cholesterol data. Here, we must define **two** `CLASS` variables:

`MANUFACT`	Treatments
`PLANT`	Experimental units

To tell SAS that we wish to fit the model

$$Y_{ijk} = \mu + \tau_i + \epsilon_{ij} + \delta_{ijk},$$

we must tell it that plants are **nested** within manufacturers. In the `PROC GLM` statements, the specification

```
PLANT(MANUFACT)
```

tells SAS that the `PLANT` factor is only meaningful **within** a particular `MANUFACT` level. If we identify a particular plant **within** a particular manufacturer, we have identified a particular experimental unit – this specification is the SAS way of doing this.

Note the use of the `RANDOM` statement. Here, `PLANT` identifies something **random** – the ϵ_{ij} terms in the model. With the `RANDOM` statement, SAS will calculate a table of **expected mean squares** treating `PLANT` as such a random factor. Compare this to the one we calculated earlier. The table is useful for determining how the appropriate F ratios are to be constructed.

Note how important this is here. By default, SAS **always** computes **all** F ratios by using the MS corresponding to the "smallest" unit of measurement, which it always calls `ERROR`, in the denominator. Here, this corresponds to **sampling units**. For testing treatment differences, then, the F ratio computed by SAS is not appropriate! (This of course may be seen from the table of expected mean squares given by the `RANDOM` statement.) We thus must **specifically request** that SAS compute the appropriate test. This is done in the `TEST` statement – here, we request that the F ratio with the Treatment SS in the denominator (`H=MANUFACT`) and Experimental Error SS in the denominator (`E=PLANT(MANUFACT)` be constructed.

We also ask for means on the treatments ($\bar{Y}_{i\cdot\cdot}$) and on individual experimental units ($\bar{Y}_{ij\cdot}$) be calculated in the `MEANS` statement.

In the output, you will notice that the basic analysis of variance table contains only two lines: `MODEL` and `ERROR`. Below, the `MODEL` line is broken down into components. In fact, in this case, these correspond to

`MODEL` = Among Experimental Units

`ERROR` = Sampling Error

and the Among Experimental Units SS is partitioned into Treatment SS and Experimental Error SS in the lower table.

The "F value" given in the upper table is **irrelevant**. In fact, that in the lower table for `MANUFACT` is **incorrect** – we asked for the correct one using the `RANDOM` statement!

MORAL: SAS is a valuable computational tool, but it is **not** a black box. It does not know how your experiment was conducted! You must tell it via `MODEL` statements. Furthermore, the **user** is responsible for ensuring that the correct test statistics are computed. SAS computes F ratios by default, but they may be totally inappropriate for the matter at hand. The user must be able to understand and interpret the output correctly!

PROGRAM:

```
*******************************************;
*                                         ;
*      EXAMPLE 7.3        ST 511          ;
*                                         ;
*      USING PROC GLM ON A ONE-WAY        ;
*   CLASSIFICATION WITH SUBSAMPLING (EQUAL ;
```

```
*   REPLICATION AND SUBSAMPLE NUMBERS)        ;
*                                             ;
*   ILLUSTRATION WITH THE CHOLESTEROL DATA    ;
*                   OF ZAR                    ;
*                                             ;
*********************************************;

OPTIONS LS=80 PS=59 NODATE;


*;
DATA CHOLEST;
  INPUT MANUFACT $ PLANT $ CONC;
  CARDS;
M1 P1 102
M1 P1 104
M1 P2 103
M1 P2 104
M2 P1 108
M2 P1 110
M2 P2 109
M2 P2 108
M3 P1 104
M3 P1 106
M3 P2 105
M3 P2 107
;
PROC PRINT;
  TITLE 'THE CHOLESTEROL DATA OF ZAR';
  TITLE2 'ANALYSIS OF VARIANCE USING PROC GLM';
  TITLE3  'IN THE CASE OF SUBSAMPLING -- EQUAL';
  TITLE4 'NUMBERS OF REPLICATIONS AND SUBSAMPLES'; RUN;
*;
PROC GLM;
  CLASS MANUFACT PLANT;
```

```
MODEL CONC = MANUFACT PLANT(MANUFACT);
RANDOM PLANT(MANUFACT);
TEST H=MANUFACT E=PLANT(MANUFACT);
MEANS MANUFACT PLANT(MANUFACT); RUN;
```

OUTPUT:

```
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%

                      THE CHOLESTEROL DATA OF ZAR                             1
                   ANALYSIS OF VARIANCE USING PROC GLM
                   IN THE CASE OF SUBSAMPLING -- EQUAL
                  NUMBERS OF REPLICATIONS AND SUBSAMPLES

                     OBS    MANUFACT    PLANT    CONC

                       1       M1         P1      102
                       2       M1         P1      104
                       3       M1         P2      103
                       4       M1         P2      104
                       5       M2         P1      108
                       6       M2         P1      110
                       7       M2         P2      109
                       8       M2         P2      108
                       9       M3         P1      104
                      10       M3         P1      106
                      11       M3         P2      105
                      12       M3         P2      107

%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%

                      THE CHOLESTEROL DATA OF ZAR                             2
                   ANALYSIS OF VARIANCE USING PROC GLM
                   IN THE CASE OF SUBSAMPLING -- EQUAL
```

NUMBERS OF REPLICATIONS AND SUBSAMPLES

General Linear Models Procedure

Class Level Information

Class	Levels	Values
MANUFACT	3	M1 M2 M3
PLANT	2	P1 P2

Number of observations in data set = 12

%%%

THE CHOLESTEROL DATA OF ZAR 3

ANALYSIS OF VARIANCE USING PROC GLM

IN THE CASE OF SUBSAMPLING -- EQUAL

NUMBERS OF REPLICATIONS AND SUBSAMPLES

General Linear Models Procedure

Dependent Variable: CONC

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	5	62.66666667	12.53333333	8.36	0.0112
Error	6	9.00000000	1.50000000		
Corrected Total	11	71.66666667			

R-Square	C.V.	Root MSE	CONC Mean
0.874419	1.157239	1.2247449	105.83333

Source	DF	Type I SS	Mean Square	F Value	Pr > F
MANUFACT	2	61.16666667	30.58333333	20.39	0.0021
PLANT(MANUFACT)	3	1.50000000	0.50000000	0.33	0.8022

Source	DF	Type III SS	Mean Square	F Value	Pr > F
MANUFACT	2	61.16666667	30.58333333	20.39	0.0021
PLANT(MANUFACT)	3	1.50000000	0.50000000	0.33	0.8022

%%%

THE CHOLESTEROL DATA OF ZAR 4

ANALYSIS OF VARIANCE USING PROC GLM

IN THE CASE OF SUBSAMPLING -- EQUAL

NUMBERS OF REPLICATIONS AND SUBSAMPLES

General Linear Models Procedure

Source	Type III Expected Mean Square
MANUFACT	Var(Error) + 2 Var(PLANT(MANUFACT)) + Q(MANUFACT)
PLANT(MANUFACT)	Var(Error) + 2 Var(PLANT(MANUFACT))

%%%

THE CHOLESTEROL DATA OF ZAR 5

ANALYSIS OF VARIANCE USING PROC GLM

IN THE CASE OF SUBSAMPLING -- EQUAL

NUMBERS OF REPLICATIONS AND SUBSAMPLES

General Linear Models Procedure

Level of MANUFACT	N	CONC Mean	CONC SD
M1	4	103.250000	0.95742711
M2	4	108.750000	0.95742711
M3	4	105.500000	1.29099445

Level of PLANT	Level of MANUFACT	N	CONC Mean	CONC SD
P1	M1	2	103.000000	1.41421356
P2	M1	2	103.500000	0.70710678
P1	M2	2	109.000000	1.41421356
P2	M2	2	108.500000	0.70710678
P1	M3	2	105.000000	1.41421356
P2	M3	2	106.000000	1.41421356

%%%

THE CHOLESTEROL DATA OF ZAR 6

ANALYSIS OF VARIANCE USING PROC GLM

IN THE CASE OF SUBSAMPLING -- EQUAL

NUMBERS OF REPLICATIONS AND SUBSAMPLES

General Linear Models Procedure

Dependent Variable: CONC

Tests of Hypotheses using the Type III MS for PLANT(MANUFACT) as an error term

Source	DF	Type III SS	Mean Square	F Value	Pr > F
MANUFACT	2	61.16666667	30.58333333	61.17	0.0037

8 Multiple Comparisons

Complementary Reading: STD, Chapter 8

8.1 Introduction

As we mentioned in chapter 7, when we test the usual hypotheses regarding differences among more than 2 treatment means, if we reject the null hypothesis, the best we can say is that there **is** a difference among the treatment means **somewhere**. Based on this analysis, we **cannot** say how these differences occur. For example, it may be that all the means are the same except one. Alternatively, all means may differ from all others. However, on the basis of this test, we cannot tell.

The concept of **multiple comparisons** is related to trying to glean more information from the data on the nature of the differences among means. For reasons that will become clear shortly, this is a difficult issue, and even statisticians do not always agree. As you will see, the issue is one of "**philosophy**" to some extent.

Understanding the principles and the problem underlying the idea of multiple comparisons is thus much more important than being familiar with the many formal statistical procedures!

In this chapter, we will discuss the issues, and then consider only a few procedures. STD contains descriptions of many more. The discussion of multiple comparisons is really only meaningful when the number of treatments $t \geq 3$; if $t = 2$, then there are only two treatment means and thus only one possible comparison of interest, whether the two means differ. Thus, throughout the chapter, we assume that there are at least 3 treatments under consideration.

8.2 Principles – "planned" versus "families" of comparisons

RECALL: When we perform a F test (a test based on an F ratio) in the analysis of variance framework for the difference among t treatment means, the only inference we make as a result of the test is that the means **considered as a group** differ somehow if we reject the null hypothesis:

$$H_{0,T}: \mu_1 = \cdots = \mu_t \text{ vs. } H_{1,T}: \text{ The } \mu_i \text{ are not all equal.}$$

We **do not** and **cannot** make inference as to **how** they differ. Consider the following scenarios:

- If $H_{0,T}$ is **not rejected**, it could be that there is a **real** difference between, say, 2 of the t treatments, but it is "getting lost" by being considered with all the other possible comparisons among treatment means.
- If $H_{0,T}$ is **rejected**, we are naturally interested in the specific nature of the differences among the t means. Are they **all** different from one another? Are only some of them different, the rest all the same? One is naturally tempted to look at the sample means for each treatment – are there differences that are "suggested" by these means?

To consider these issues, we must recall the framework in which we test hypotheses. Recall that the **level of significance** for any hypothesis test

$$\alpha = P(\text{ reject } H_0 \text{ when } H_0 \text{ is true}) = P(\text{ Type I error }).$$

Thus, the level α specifies the probability of making a "mistake" and saying that there is evidence to suggest the alternative hypothesis is true when it really isn't. The probability of making such an error is controlled to be no more than α by the way we do the test.

IMPORTANT: The level α, and thus the probability of making such a mistake, applies **only** to the particular H_0 under consideration. Thus, in the analysis of variance situation above, where we test $H_{0,T}$ vs. $H_{1,T}$, α applies **only** to the comparison among **all** the means together – either they differ **somehow** or they do not. The probability that we end up saying the differ **somehow** when they don't is thus no more than α, that is all. α does not pertain to, say, any further consideration of the means, say, comparing them 2 at a time. To see this, consider some examples.

EXAMPLE – $t = 3$ TREATMENTS: Suppose there are $t = 3$ treatments under consideration. We are certainly interested in the question "do the means differ?" but we may also be interested in **how**. Specifically, does, say, $\mu_1 = \mu_2$, but $\mu_3 \neq \mu_1$ or μ_2? Here, we are interested in what we can say about 3 **separate** comparisons. We'd like to be able to combine the 3 comparisons somehow to make a statement about **how** the means differ.

Suppose, to address this, we decide to perform 3 **separate** t tests for the 3 possible differences of means, **each** with level of significance α. Then, we have

$$P(\text{ Type I error comparing } \mu_1 \text{ vs. } \mu_2) = \alpha$$

$$P(\text{ Type I error comparing } \mu_1 \text{ vs. } \mu_3) = \alpha$$

$$P(\text{ Type I error comparing } \mu_2 \text{ vs. } \mu_3) = \alpha.$$

That is, in **each test**, we have probability of α of inferring that the 2 means under consideration differ when they really **do not**.

Because we are performing more than one hypothesis test, the probability we make this mistake in **at least one** of the tests is no longer α! This is simply because we are performing more than one test! For example, it may be shown mathematically that, if $\alpha = 0.05$,

$$P(\text{ Make at least one Type I error in the 3 tests }) \approx 0.14!$$

Suppose we perform the 3 separate tests, and we reject the null hypothesis in the first two tests, but not in the third (μ_2 vs. μ_3), with each test at level $\alpha = 0.05$. We then proclaim at the end that "There is sufficient evidence in these data to say that μ_1 differs from μ_2 and μ_1 differs from μ_3. We do not have sufficient evidence to say that μ_2 and μ_3 differ from each other." Next to this statement, we say "at level of significance $\alpha = 0.05$." What is wrong with doing this?

From the calculation above, the chance that we said something wrong in this statement is **not** 0.05! Rather, it is 0.14! The chance we have made a mistake in claiming a difference in at least one of the tests is 0.14 – almost **3 times** the chance we are claiming!

In fact, for larger values of t (more treatments), if we make a separate test for each possible **pairwise** comparison among the t means, each at level α, things only get worse. For example, if $\alpha = 0.05$ and $t = 10$, the chance we make at least one Type I error, and say in our overall statement that there is evidence for a difference in a pair when there isn't, is almost 0.90!!! That is, if we try to combine the results of all these tests together into a statement that tries to sort out all the differences, it is **very possible** (almost certain) that we will claim a difference that really doesn't exist somewhere in our statement!

RESULT: If we wish to "sort out" differences among all the treatment means, we cannot just compare them all separately without having a much higher chance of concluding something wrong!

ANOTHER PROBLEM: Recognizing the problems associated with "sorting out" differences above, suppose we decide to look at the sample treatment means $\bar{Y}_{i\cdot}$ and compare **only those** that appear to possibly be different. Won't this get around this problem, as we're not likely to do as many separate tests?

NO! To see this, suppose we inspect the sample means, and decide to conduct a t test at level $\alpha = 0.05$ for a difference in the two treatment means observed to have the **highest** and **lowest** sample means $\bar{Y}_{i\cdot}$ among all those in the experiment.

Recall that the data are just **random samples** from the treatment populations of interest. Thus, the sample means $\bar{Y}_{i\cdot}$ could have ended up the way they did because:

- There **really is** a difference in the population means **OR**
- We got "unusual" samples, and there really is **no difference**!

Because of the chance mechanism involved, either of these explanations is possible.

Returning to our largest and smallest sample means, then, the large difference we have observed could be due just to chance – we got some "unusual" samples, even though the means **really don't differ**! Because we have **already seen** in our samples a large difference, however, it turns out that, even if this is the case, we will still be **more likely** to reject the null hypothesis of no difference in the 2 means! That is, although we claim that we only have a 5% chance of rejecting the null hypothesis when it's true, the chance we actually do this is higher!

Here, with $\alpha = 0.05$, it turns out that the **true** probability that we reject the null hypothesis that the means with smallest and largest observed $\bar{Y}_{i\cdot}$ values are the same when it is true is

- actually 0.13 if $t = 3$!
- actually 0.60 if $t = 10$!

RESULT: If we test something on the basis of what we observed in our samples, our chance of making a Type I error is much greater than we think.

DILEMMA: The above discussion shows that there are clearly problems with trying to get a handle on how treatment means differ. In fact, this brings to light some philosophical issues.

"PLANNED COMPARISONS": This is best illustrated by example. Suppose that some university researchers are planning to conduct an experiment involving 4 different treatments:

- The standard treatment, which is produced by a certain company and is in widespread use.
- A new commercial treatment manufactured by a rival company, which hopes to show it is better than the standard, so that they can market the new treatment for enormous profits, taking away the market share of their competitor.
- Two experimental treatments developed by the university researchers. These are being developed by two new procedures, one designed by the university researchers, the other by rival researchers at another, more prestigious university. The university researchers hope to show that their treatment is better and publish their results, both for enormous academic glory and to humiliate their rivals in print.

The administration of the company is trying to decide whether they should begin planning marketing strategy for their treatment; they are thus uninterested for this purpose in the two university treatments. Because the university researchers are setting up an experiment, the company finds it less costly to simply pay the researchers to include their treatment and the standard in the study rather than for the company to do a separate experiment of their own.

On the other hand, the university researchers are mainly interested in the experimental treatments. In fact, their main question is which of them shows more promise.

In this situation, the main comparison of interest as far as the company is concerned is that between their new treatment and the standard. Thus, although the actual experiment involves 4 treatments, they really only care about the pairwise comparison of the two; that is, the difference

$$\mu_{\text{standard}} - \mu_{\text{new (company)}}. \tag{8.1}$$

Similarly, for the university researchers, their main question involves the difference in the pair

$$\mu_{\text{experimental (us)}} - \mu_{\text{experimental (them)}}.$$

In this situation, the comparison of interest depends on the interested party. Of course, as usual, each party would like to control the probability of making a Type I error in their particular comparison. For example, the company may be satisfied with level of significance $\alpha = 0.05$, as usual, and probably would have used this same significance level had they conducted the experiment with just the two treatments on their own.

In this example, **prior** to conduct of the experiment, specific comparisons of interest among various treatment means have been identified, **regardless** of the overall outcome of the experiment. Because these comparisons were identified **in advance**, we do not run into the problem we did in comparing the treatments with the largest and smallest sample means after the experiment, because the tests will be performed **regardless**. Nor do we run into the problem of sorting out differences among all means. As far as the company is concerned, their question about (8.1) is a separate test, for which they want the probability of concluding a difference when there isn't one to be at most 0.05. A similar statement could be made about university researchers.

RESULT: Of course, this made-up scenario is a bit idealized, but it illustrates the main point. If specific questions involving particular treatment means are of **independent interest**, are identified as such **in advance** of seeing experimental results, and would be investigated regardless of outcome, then it may be legitimate to perform the associated tests at level of significance α, without concern about the outcome of other tests.

In this case, the level of significance α applies **only** to the test in question, and may be proclaimed **only** in talking about that single test!

TERMINOLOGY: If a specific comparison is made in this way from data from a larger experiment, we are controlling the **comparisonwise** (Type I) error rate at α. For each comparison of this type that is made, the probability of a Type I error is α. The results of several comparisons **may not** be combined in to a single statement with claimed Type I error rate α.

ADVANTAGE: As we will discuss shortly, because the full experiment contains information about **experimental error in addition** to that from only the two treatments of interest for either of these comparisons, it is possible to **gain precision** for testing such **planned comparisons**.

"FAMILIES OF COMPARISONS": Recall the pea section data of chapter 7. There were 4 sugar treatments and a control (no sugar). Suppose a main question of interest is to sort out how the sugar treatments differ from the control; that is, the investigators would like to be able to make a statement as to which sugar treatments lead to different mean pea section lengths from the control and which don't. Here, then, even though the investigators have performed the experiment and conducted the F test, which led to them to infer that there **are** differences, they would like to investigate the **reasons** for these differences, with specific interest in how the sugars each compare to the control as a possible source of the differences.

The investigators would like to make a **single statement** about the differences. This statement involves **4 pairwise comparisons** – each sugar treatment against the control. From our previous discussion, it is clear that testing **each** pairwise comparison against the control at level of significance α and then making such a statement would involve a chance of declaring differences that really don't exist **greater** than that indicated by α.

SOLUTION: For such a situation, then, the investigators would like to avoid this difficulty. They are interested in ensuring that the **family** of 4 comparisons they wish to consider (all sugars against the control) has **overall** probability of making at least one Type I error to be no more than α, i.e.

$$P(\text{ at least one Type I error among all 4 tests in the family }) \leq \alpha.$$

TERMINOLOGY: When a question of interest about treatment means involves a **family** of several comparisons, we would like to control the **familywise** error rate at α, so that the overall probability of making at least one mistake (declaring a difference that doesn't exist) is controlled at α.

It turns out that a number of statistical methods have been developed that ensure that the level of significance for a **family** of comparisons is no more than a specified α. These are often called **multiple comparison** procedures.

REMARK: The above discussion is admittedly simplified, and there is much debate among statisticians about different perspectives on the problem of making comparisons among subsets of means or all means from an experiment with several treatments. However, the idea that there **is** an issue is in itself important. Thus, as mentioned earlier, the most important thing you can take away from this chapter may well be an appreciation of the philosophical debate!

8.3 The least significant difference

First, we consider the situation where we have **planned in advance** of the experiment to make certain comparisons among treatment means. Each comparison is of interest in its own right, and thus is to be viewed as separate. Statements regarding each such comparison will **not** be combined.

Thus, here, we wish to control the **comparisonwise** error rate at α.

IDEA: Despite the fact that each comparison is to be viewed **separately**, we can still take advantage of **all** information in the experiment on experimental error.

To fix ideas, suppose we have an experiment involving t treatments, and we are interested in comparing two treatments a and b, with means μ_a and μ_b. That is, we wish to test

$$H_0 : \mu_a = \mu_b \text{ vs. } H_1 : \mu_a \neq \mu_b.$$

We wish to control the comparisonwise error rate at α.

If we only had sample information for treatments a and b alone, we would base our test on the usual t statistic

$$\frac{|\bar{Y}_{a\cdot} - \bar{Y}_{b\cdot}|}{s_{\bar{Y}_{a\cdot} - \bar{Y}_{b\cdot}}};$$

the numerator is the usual (absolute) difference in means, and the denominator represents the usual standard error for the difference. Suppose there are r_a and r_b replicates, respectively, in each sample. Then

$$s_{\bar{Y}_{a\cdot} - \bar{Y}_{b\cdot}} = s\sqrt{\frac{1}{r_a} + \frac{1}{r_b}},$$

where s is the usual pooled estimate of variance σ^2.

Note that, because we have an experiment involving all t treatments, **all** observations from all t treatments contribute to our knowledge of σ^2 (experimental error) through MS_E! Thus, the idea is to **take advantage** of this by replacing the denominator of the t test statistic by something involving an estimate of σ from all t treatments.

TEST STATISTIC: As our test statistic for H_0 vs. H_1, use instead

$$\frac{|\bar{Y}_{a\cdot} - \bar{Y}_{b\cdot}|}{s_{\bar{Y}_{a\cdot}-\bar{Y}_{b\cdot}}}, \quad s_{\bar{Y}_{a\cdot}-\bar{Y}_{b\cdot}} = s\sqrt{\frac{1}{r_a}+\frac{1}{r_b}}, \quad s = \sqrt{MS_E}.$$

That is, instead of basing the estimate of σ^2 on only the two treatments in question, use the estimate from all t treatments.

RESULT: This will lead to a **more precise** test: The estimate of the standard deviation of the treatment mean difference from all t treatments will be a more precise estimate, because it is based on **more information**! Specifically, the standard error based on only the two treatments will have only $r_a + r_b - 2$ degrees of freedom, while that based on all t will have

$$\sum_{j=1}^{r} r_i - t = N - t$$

degrees of freedom!

TEST PROCEDURE: Perform the test using all information by rejecting H_0 in favor of H_1 if

$$\frac{|\bar{Y}_{a\cdot} - \bar{Y}_{b\cdot}|}{s_{\bar{Y}_{a\cdot}-\bar{Y}_{b\cdot}}} > t_{N-t,\alpha/2}.$$

A quick look at the t table will reveal the advantage of this test. The degrees of freedom, $N - t$ for estimating σ^2 (experimental error), are **greater** than those from only 2 treatments. The corresponding **critical value** is thus **smaller**, giving more chance for rejection of H_0 if it's really true!!

Note that the test procedure may be rewritten as follows: Reject H_0 if

$$|\bar{Y}_{a\cdot} - \bar{Y}_{b\cdot}| > s_{\bar{Y}_{a\cdot}-\bar{Y}_{b\cdot}}\ t_{N-t,\alpha/2}, \quad s = \sqrt{MS_E}.$$

TERMINOLOGY: If we decide to compare 2 treatment means from a larger experiment involving t treatments, the value

$$s_{\bar{Y}_{a\cdot}-\bar{Y}_{b\cdot}}\ t_{N-t,\alpha/2} = s\sqrt{\frac{1}{r_a}+\frac{1}{r_b}}\ t_{N-t,\alpha/2}, \quad s=\sqrt{MS_E}$$

is called the **least significant difference** (LSD) for the test of H_0 vs. H_1 above, based on the entire experiment. From above, we reject H_0 at level α if

$$|\bar{Y}_{a\cdot}-\bar{Y}_{b\cdot}| > LSD.$$

EQUAL REPLICATION: Suppose that $r_i = r$ for all t treatments.. Then, for **any** such comparison between two means, the value of LSD is **the same**:

$$s_{\bar{Y}_{a\cdot}-\bar{Y}_{b\cdot}} = s\sqrt{\frac{2}{r}}, \quad s=\sqrt{MS_E}, \quad LSD = s\sqrt{\frac{2}{r}}\ t_{rt-t,\alpha/2}.$$

Thus, in experiments with equal replication, all such pairwise comparisons of interest require only a single calculation.

WARNING: It is critical to remember that the LSD procedure is only valid if the paired comparisons are **genuinely** of **independent interest**.

EXAMPLE: For the pea section data of chapter 7, we had equal replication with $r = 10$, $t = 5$. Let μ_1 denote the mean for the control treatment and $\mu_2, \ldots, \mu_5$ denote those for the sugar treatments.

Suppose that it is was of interest, in advance of the experiment, one investigator was interested in the particular question of whether 2% glucose (Treatment 2) differs from the control:

$$H_{0,2}: \mu_1 = \mu_2 \text{ vs. } H_{1,2}: \mu_1 \neq \mu_2.$$

Another investigator was interested in the specific question of 2% fructose (Treatment 3) vs. the control:

$$H_{0,3}: \mu_1 = \mu_3 \text{ vs. } H_{1,3}: \mu_1 \neq \mu_3.$$

We have

$$\bar{Y}_{1\cdot} = 70.1, \quad \bar{Y}_{2\cdot} = 59.3, \quad \bar{Y}_{3\cdot} = 58.2, \quad , s = \sqrt{MS_E} = 2.3357.$$

$rt - t = t(r-1) = 45$, $t_{45,0.025} \approx 2.01$. Thus

$$LSD = (2.3357)\sqrt{2/10}\ (2.01) \approx 2.10.$$

Because there is equal replication, we may use this value for both comparisons:

$$|\bar{Y}_{2\cdot} - \bar{Y}_{1\cdot}| = 10.8 > 2.10$$

$$|\bar{Y}_{3\cdot} - \bar{Y}_{1\cdot}| = 11.9 > 2.10.$$

RESULT: $H_{0,2}$ is rejected at level of significance 0.05; there is sufficient evidence to suggest that the glucose treatment yields mean pea section lengths different from the control. $H_{0,3}$ is rejected at level of significance 0.05; there is sufficient evidence to suggest that the fructose treatment yields mean pea section lengths different from the control.

INTERPRETATION: Each of these is a **separate** comparison to be viewed on its own. We **do not** combine the information from these two **separate** tests in to a single statement about how both glucose and fructose differ from the control at level $\alpha = 0.05$.

- *Valid statement*: There is evidence to suggest that the mean pea section length for plants treated with 2% glucose is different from that of untreated plants at level 0.05.
- *Invalid statement*: There is evidence to suggest that the mean pea section lengths for plants treated with 2% glucose or 2% fructose are different at level 0.05.

This is a subtle point, but an important one. The significance level for each test applies **only** to that test. For the **invalid** statement above, the **true** significance level would be

$$P(\text{ make at least one Type I error in the 2 tests }) > 0.05.$$

WARNING: Statistical packages such as SAS will generally compute the LSD test for **all possible** pairwise comparisons (see section 8.6). It is important to remember that each comparison is to be viewed **separately**!

8.4 Contrasts

Before we discuss the notion of multiple comparisons (for families of comparisons), we consider the case where we may be interested in comparisons that **can not** be expressed as a difference in a **pair** of means.

EXAMPLE: Suppose that for the pea section experiment, a particular question of interest was to compare sugar treatments containing fructose to those that do not:

μ_3, μ_4 do contain fructose

μ_2, μ_5 do not contain fructose

It is suspected that treatments 3 and 4 are similar in terms of resulting pea section length, and that treatments 2 and 5 are similar, and that lengths from 3 and 4, **on the average**, are **different** from lengths from 2 and 5, **on the average**.

Thus, the question of interest is to compare the **average** mean pea section length for treatments 3 and 4 to the **average** mean pea section length for treatments 2 and 5.

To express this formally in terms of the treatment means, we are thus interested in the comparison between

$$\frac{\mu_3 + \mu_4}{2} \text{ and } \frac{\mu_2 + \mu_5}{2};$$

that is, the **average** of μ_3 and μ_4 vs. the **average** of μ_2 and μ_5.

We may express this formally as a set of hypotheses:

$$H_0 : \frac{\mu_3 + \mu_4}{2} = \frac{\mu_2 + \mu_5}{2} \text{ vs. } H_1 : \frac{\mu_3 + \mu_4}{2} \neq \frac{\mu_2 + \mu_5}{2}.$$

These may be rewritten by algebra as:

$$H_0 : \mu_3 + \mu_4 - \mu_2 - \mu_5 = 0 \text{ vs. } H_1 : \mu_3 + \mu_4 - \mu_2 - \mu_5 \neq 0. \tag{8.2}$$

Similarly, suppose a particular question of interest was whether sugar treatments differ on average in terms of mean pea section length from the control. We thus would like to compare the **average** mean length for treatments 2–5 to the mean for treatment 1. We may express this formally as a set of hypotheses:

$$\begin{aligned} H_0 &: \frac{\mu_2 + \mu_3 + \mu_4 + \mu_5}{4} = \mu_1, \text{ or } 4\mu_1 - \mu_2 - \mu_3 - \mu_4 - \mu_5 = 0 \\ &\text{vs.} \\ H_1 &: \frac{\mu_2 + \mu_3 + \mu_4 + \mu_5}{4} \neq \mu_1, \text{ or } 4\mu_1 - \mu_2 - \mu_3 - \mu_4 - \mu_5 \neq 0 \end{aligned} \tag{8.3}$$

TERMINOLOGY: A linear function of treatment means of the form

$$\sum_{i=1}^{t} c_i \mu_i$$

such that the constants c_i sum to zero, i.e.

$$\sum_{i=1}^{t} c_i = 0$$

is called a **contrast**.

Both of the functions in (8.2) and (8.3) are **contrasts**:

$\mu_3 + \mu_4 - \mu_2 - \mu_5$	$c_1 = 0,\ c_2 = -1,\ c_3 = 1,\ c_4 = 1,\ c_5 = -1$	$\sum_{i=1}^{5} c_i = 0$
$4\mu_1 - \mu_2 - \mu_3 - \mu_4 - \mu_5$	$c_1 = 4,\ c_2 = -1,\ c_3 = -1,\ c_4 = -1,\ c_5 = -1$	$\sum_{i=1}^{5} c_i = 0$

INTERPRETATION: Note that, in each case, if the means really were **all** equal, then, because the **coefficients** c_i sum to zero, the contrast itself will be equal to zero. If they are different, it will not. Thus, the null hypothesis says that there are no differences. The alternative says that the particular combination of means is different from zero, which reflects real differences among functions of the means.

Note that, of course, a **pairwise comparison** is a **contrast**, e.g.

$$\mu_2 - \mu_1$$

is a contrast with $c_1 = -1$, $c_2 = 1$, $c_3 = c_4 = c_5 = 0$.

ESTIMATION: As intuition suggests, the best (in fact, **unbiased**) estimator of any contrast $\sum_{i=1}^{t} c_i \mu_i$ is

$$Q = \sum_{i=1}^{t} c_i \bar{Y}_{i\cdot}.$$

It turns out that, for the population of all such Q's for a particular set of c_i's, the variance of a contrast is

$$\sigma_Q^2 = \sigma^2 \sum_{i=1}^{t} \frac{c_i^2}{r_i},$$

which may be estimated by replacing s^2 by the pooled estimate, MS_E. Thus, a **standard error** estimate for the contrast is

$$s_Q = s\sqrt{\sum_{i=1}^{t} \frac{c_i^2}{r_i}}, \quad s = \sqrt{MS_E}.$$

It may be easily verified (try it) that this expression reduces to our usual expression when the contrast is a pairwise comparison.

TEST PROCEDURE: For testing hypotheses of the general form

$$H_0 : \sum_{i=1}^{t} c_i \mu_i = 0 \text{ vs. } H_1 : \sum_{i=1}^{t} c_i \mu_i \neq 0$$

at level of significance α when the comparison is **planned**, we use the same procedure as for the LSD test. We reject H_0 in favor of H_1 if

$$|Q| = |\sum_{i=1}^{t} c_i \bar{Y}_{i\cdot}| > s_Q \; t_{N-t,\alpha/2}.$$

EXAMPLE: For the pea section data, we illustrate with the first set of hypotheses regarding whether or not fructose is present. We have

$$|Q| = |\sum_{i=1}^{5} c_i \bar{Y}_{i\cdot}| = |58.2 + 58.0 - 59.3 - 64.1| = 7.2.$$

$$s_Q = 2.3357\sqrt{\frac{0}{10} + \frac{1^2}{10} + \frac{1^2}{10} + \frac{-1^2}{10} + \frac{-1^2}{10}} = 1.0446$$

$$s_Q \; t_{45,0.025} = (1.0446)(2.01) = 2.969.$$

Thus, we have $7.2 > 2.969$, and we reject H_0 at level of significance $\alpha = 0.05$. There is evidence to suggest that there is a difference, on the average, in mean pea section length depending on whether or not fructose is present.

REMINDER: The result of this test is not to be compared with any other. The level of significance α pertains only to the question at hand.

8.5 Families of comparisons

We now consider the problem of **combining statements**. Suppose we wish to make a single statement about a **family** of contrasts (e.g. several pairwise comparisons) at some specified level of significance α.

EXAMPLE: Suppose, in the pea section example, we wish to make a statement about how **all** sugar treatments compare against the control. This involves the 4 contrasts

$$\mu_2 - \mu_1, \quad \mu_3 - \mu_1, \quad \mu_4 - \mu_1, \quad \mu_5 - \mu_1.$$

To ensure that we control our chance of declaring differences between any of the sugar treatments and control when they are really not present at α, we require

$$P(\text{ at least 1 Type I error in the 4 comparisons}) \leq \alpha.$$

There are a number of methods for making statements about **families of comparisons** that control the overall **familywise** level of significance at a value α. We discuss three of these; STD discuss several more. The basic premise behind each is the same.

BONFERRONI METHOD: This method is based on modifying individual t tests. Suppose we specify c contrasts C_1, C_2,...,C_c and wish to test the hypotheses

$$H_{0,k} : C_k = 0 \text{ vs. } H_{1,k} : C_k \neq 0, \quad k = 1, \ldots, c$$

while controlling the overall **familywise** Type I error rate for all c tests as a group to be $\leq \alpha$.

It may be shown mathematically that, if one makes c tests, each at level of significance α/c, then

$$P(\text{ at least 1 Type I error in the } c \text{ tests }) \leq \alpha.$$

Thus, in the Bonferroni procedure, we use for each test the t value corresponding to the appropriate degrees of freedom and α/c and conduct each test as usual! That is, for each contrast $C_k = \sum_{i=1}^{t} c_i \mu_i$, reject $H_{0,k}$ if

$$|Q_k| > s_Q \; t_{N-t,\alpha/(2c)}, \quad Q_k = \sum_{i=1}^{t} c_i \bar{Y}_{i\cdot}.$$

EXAMPLE: For the pea section data, suppose we wished to make a statement about how all sugar treatments compare against the control. This involves the 4 pairwise contrasts

$$\mu_2 - \mu_1, \quad \mu_3 - \mu_1, \quad \mu_4 - \mu_1, \quad \mu_5 - \mu_1.$$

For each contrast, we have $s_Q = (2.3357)\sqrt{2/10} = 1.0446$. For an **overall**, **familywise** level of $\alpha = 0.05$, we use $\alpha/c = 0.05/4 = 0.00625$, so we obtain $t_{45,0.00625} \approx 2.9$ (by rough interpolation; the SAS output has the exact value). Thus, for each pairwise contrast, we compare the absolute sample mean difference to

$$s_Q \; t_{45,0.00625} = (1.0446)(2.9) \approx 3.029.$$

Comparing each sample mean difference in the family to this value:

$$|\bar{Y}_{2\cdot} - \bar{Y}_{1\cdot}| = 10.8 > 3.029$$

$$|\bar{Y}_{3\cdot} - \bar{Y}_{1\cdot}| = 11.9 > 3.029$$

$$|\bar{Y}_{4\cdot} - \bar{Y}_{1\cdot}| = 12.1 > 3.029$$

$$|\bar{Y}_{5\cdot} - \bar{Y}_{1\cdot}| = 6.0 > 3.029.$$

We may state "At level of significance $\alpha = 0.05$, there is sufficient evidence that each sugar treatment mean differs from the control." The chance that we have declared at least one wrong difference is controlled at $\alpha = 0.05$.

The Bonferroni method is valid with any type of contrast, and may be used with unequal replication.

SCHEFFÉ'S METHOD: The idea is to ensure that, for **any** family of contrasts, the family level is α by requiring that the probability of a Type I error is $\leq \alpha$ for the family of **all possible contrasts**. That is, control things so that

$$P(\text{ at least 1 Type I error among tests of } \textbf{all possible contrasts} \text{ }) \leq \alpha.$$

Because the level is controlled for **any** family of contrasts, it will be controlled for the one of interest.

The procedure is to compute the quantity

$$S = \sqrt{(t-1)\, F_{t-1,N-t,\alpha}}.$$

For all contrasts of interest, C_k, reject the corresponding null hypothesis $H_{0,k}$ if

$$|Q_k| > s_Q S.$$

EXAMPLE: For the pea section data and the family of comparisons of each sugar treatment vs. control, we have

$$S = \sqrt{(5-1)\,F_{4,45,\alpha}} \approx \sqrt{4(2.57)} = 3.206.$$

(2.57 was interpolated from the F table.) We thus have, for each comparison, $s_Q\,S = (1.0446)(3.206) = 3.349$. Comparing each sample mean difference in the family to this value:

$$|\bar{Y}_{2\cdot} - \bar{Y}_{1\cdot}| = 10.8 > 3.349$$

$$|\bar{Y}_{3\cdot} - \bar{Y}_{1\cdot}| = 11.9 > 3.349$$

$$|\bar{Y}_{4\cdot} - \bar{Y}_{1\cdot}| = 12.1 > 3.349$$

$$|\bar{Y}_{5\cdot} - \bar{Y}_{1\cdot}| = 6.0 > 3.349.$$

We reach the same conclusion as by the Bonferroni method, with the same careful statement of the results. The Scheffé procedure may be used with equal replication and any kinds of contrasts. Note that, because it controls the Type I error probability for the family of **all** possible contrasts, the critical value is larger than for the Bonferroni method, which only considers 4. This need not always be the case, however, because the two methods have a different mathematical basis.

TUKEY'S METHOD: For this method, we **must** have **equal** replication, and the contrasts of interest **must** all be **pairwise comparisons**. Because only pairwise contrasts are of interest, the method takes advantage of this fact. It ensures that the familywise error rate for **all possible** pairwise comparisons is controlled at α; if this is true, then any subset of pairwise comparisons will also have this property. It turns out that there is a special distribution, given in Table A.8 of STD, for example, that is applicable in this case.

The procedure is to compute

$$T = \frac{1}{\sqrt{2}}\,q_\alpha(t, N-t),$$

where $q_\alpha(t, N-t)$ is the value found from this table. For example, $q_{0.05}(5, 45) \approx 4.04$ (the value closest in the table). If C_k is the kth pairwise comparison, we would reject the corresponding null hypothesis $H_{0,k}$, if

$$|Q_k| > s_Q\;T.$$

EXAMPLE: Here, as above, $s_Q = 1.0446$, and we have $T = (1/\sqrt{2})(4.04) = 2.857$. Thus, $s_Q T = (1.0446)(2.857) = 2.984$. For our family of all sugar treatments versus control, we have

$$|\bar{Y}_{2\cdot} - \bar{Y}_{1\cdot}| = 10.8 > 2.984$$

$$|\bar{Y}_{3\cdot} - \bar{Y}_{1\cdot}| = 11.9 > 2.984$$

$$|\bar{Y}_{4\cdot} - \bar{Y}_{1\cdot}| = 12.1 > 2.984$$

$$|\bar{Y}_{5\cdot} - \bar{Y}_{1\cdot}| = 6.0 > 2.984.$$

(This differs slightly from the SAS results in section 8.6, as we've had to interpolate from the table.)

We may make the same statement as with the other two methods.

COMPARISON: In the pea section example, all three methods considered here yielded qualitatively the same conclusions. This need not always be the case, however. Here is how the three methods we have discussed compare:

- The Tukey method may **not** be used with unequal replication, while the other two may.
- If **all** pairwise comparisons are of interest, the Tukey method is preferred over the Bonferroni and Scheffé methods in cases of equal replication, because it is then **always** true that

  $$T < S, \quad T < t_{N-t,\alpha/(2c)}.$$

  We will thus be more likely to reject the relevant hypotheses with the Tukey method (i.e. we will be **less conservative**).
- Because the Scheffé method considers all possible contrasts, not just ones involving pairs, if the number of contrasts is **small**, the Bonferroni method may be preferred. In this case, it is likely that

  $$S > t_{N-t,\alpha/(2c)}.$$

It is **legitimate** to conduct the hypothesis tests using all of these methods (where applicable) and then choose the one that rejects most often. This is valid because the "cut-off" values S, T, and $t_{N-t,\alpha/(2c)}$ **do not depend** on the data. Thus, we are not choosing on the basis of observed data (which of course would be **illegitimate**, as discussed earlier).

A PROBLEM WITH MULTIPLE COMPARISONS: Regardless of **which** method one uses, there is always a problem when conducting multiple comparisons. Because the number of comparisons being made may be **large** (e.g. all possible pairwise comparisons for $t = 10$ treatments!), and we wish to control the overall probability of at least one Type I error to be **small**, we are likely to have **low power** (that is, likely to have a difficult time detecting real differences among the comparisons in our family). This is for the simple reason that, in order to ensure we achieve overall level α, we must use critical values for each comparison **larger** than we would if the comparisons were each made separately at level α (inspect the pea section results for an example). Thus, although the goal is worthy, that of trying to sort out differences, we may be quite unsuccessful at achieving it!

This problem has tempted some investigators to try to figure out ways around the issue; for example, claiming that certain comparisons were of interest in advance when they really weren't, so as to salvage an experiment with no "significant" results. This is, of course, **inappropriate**!!! The **only** way to ensure enough power to test all questions of interest is to **design** the experiment with a large enough sample size!!!

8.6 Using SAS to perform multiple comparisons

EXAMPLE 1: The pea section data. Refer to chapter 7 for a complete description. The program here is a modified version of that in section 7.13, using `PROC GLM` to obtain the analysis of variance. Here, we specify **options** in the `MEANS` statement to request the LSD, Bonferroni, Scheffé, and Tukey procedures. Note that the method used to show whether pairwise comparisons were significant is to place the same letter (A, B, etc) next to the mean.

WARNING: Be sure you understand that the `LSD` results are on a **comparisonwise** basis, while the others are on a **familywise** basis (see the output).

PROGRAM:

```
*******************************************;
*                                         ;
*        ST 511      EXAMPLE 8.1          ;
*                                         ;
*     USING SAS TO DO MULTIPLE COMPARISON  ;
*     TESTS -- EQUAL REPLICATION.          ;
```

```
*     HERE, WE USE PROC GLM WITH THE         ;
*     OPTIONS BON, SCHEFFE, AND TUKEY        ;
*     TO PERFORM TESTS FOR ALL PAIRWISE      ;
*     COMPARISONS ON THE PEA SECTION DATA    ;
*                                            ;
*     ALSO USE LSD FOR PLANNED PAIRWISE      ;
*     COMPARISONS                            ;
*                                            ;
*********************************************;

OPTIONS LS=80 PS=59 NODATE;

DATA PEAS;
  INPUT SUGAR $ LENGTH @@;
  CARDS;
CNTL  75  CNTL  67  CNTL  70  CNTL  75  CNTL  65  CNTL  71
CNTL  67  CNTL  67  CNTL  76  CNTL  68
GLU2  57  GLU2  58  GLU2  60  GLU2  59  GLU2  62  GLU2  60
GLU2  60  GLU2  57  GLU2  59  GLU2  61
FRU2  58  FRU2  61  FRU2  56  FRU2  58  FRU2  57  FRU2  56
FRU2  61  FRU2  60  FRU2  57  FRU2  58
GLU1FRU1  58  GLU1FRU1  59  GLU1FRU1  58  GLU1FRU1  61
GLU1FRU1  57  GLU1FRU1  56  GLU1FRU1  58  GLU1FRU1  57
GLU1FRU1  57  GLU1FRU1  59
SUC2  62  SUC2  66  SUC2  65  SUC2  63  SUC2  64  SUC2  62
SUC2  65  SUC2  65  SUC2  62  SUC2  67
;
PROC PRINT;
  TITLE 'THE PEA DATA OF SOKAL & ROHLF';
  TITLE2 'ONE WAY CLASSIFICATION BY TREATMENT';
  TITLE3 'TREATMENT = SUGAR (5 LEVELS)'; RUN;
*;
PROC GLM;
  CLASS SUGAR;
```

```
MODEL LENGTH = SUGAR;

MEANS SUGAR / LSD BON SCHEFFE TUKEY; RUN;
```

OUTPUT:

```
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%

                    THE PEA DATA OF SOKAL & ROHLF                           1
                 ONE WAY CLASSIFICATION BY TREATMENT
                     TREATMENT = SUGAR (5 LEVELS)

                      OBS    SUGAR       LENGTH

                        1    CNTL          75
                        2    CNTL          67
                        3    CNTL          70
                        4    CNTL          75
                        5    CNTL          65
                        6    CNTL          71
                        7    CNTL          67
                        8    CNTL          67
                        9    CNTL          76
                       10    CNTL          68
                       11    GLU2          57
                       12    GLU2          58
                       13    GLU2          60
                       14    GLU2          59
                       15    GLU2          62
                       16    GLU2          60
                       17    GLU2          60
                       18    GLU2          57
                       19    GLU2          59
                       20    GLU2          61
                       21    FRU2          58
                       22    FRU2          61
```

23	FRU2	56
24	FRU2	58
25	FRU2	57
26	FRU2	56
27	FRU2	61
28	FRU2	60
29	FRU2	57
30	FRU2	58
31	GLU1FRU1	58
32	GLU1FRU1	59
33	GLU1FRU1	58
34	GLU1FRU1	61
35	GLU1FRU1	57
36	GLU1FRU1	56
37	GLU1FRU1	58
38	GLU1FRU1	57
39	GLU1FRU1	57
40	GLU1FRU1	59
41	SUC2	62
42	SUC2	66
43	SUC2	65
44	SUC2	63
45	SUC2	64
46	SUC2	62
47	SUC2	65
48	SUC2	65
49	SUC2	62
50	SUC2	67

%%%

THE PEA DATA OF SOKAL & ROHLF 2

ONE WAY CLASSIFICATION BY TREATMENT

TREATMENT = SUGAR (5 LEVELS)

General Linear Models Procedure

Class Level Information

Class	Levels	Values
SUGAR	5	CNTL FRU2 GLU1FRU1 GLU2 SUC2

Number of observations in data set = 50

%%%

THE PEA DATA OF SOKAL & ROHLF 3

ONE WAY CLASSIFICATION BY TREATMENT

TREATMENT = SUGAR (5 LEVELS)

General Linear Models Procedure

Dependent Variable: LENGTH

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	4	1077.3200000	269.3300000	49.37	0.0001
Error	45	245.5000000	5.4555556		
Corrected Total	49	1322.8200000			

R-Square	C.V.	Root MSE	LENGTH Mean
0.814412	3.770928	2.3357131	61.940000

Source	DF	Type I SS	Mean Square	F Value	Pr > F
SUGAR	4	1077.3200000	269.3300000	49.37	0.0001

Source	DF	Type III SS	Mean Square	F Value	Pr > F
SUGAR	4	1077.3200000	269.3300000	49.37	0.0001

%%%

THE PEA DATA OF SOKAL & ROHLF 4

ONE WAY CLASSIFICATION BY TREATMENT

TREATMENT = SUGAR (5 LEVELS)

General Linear Models Procedure

T tests (LSD) for variable: LENGTH

NOTE: This test controls the type I comparisonwise error rate not the experimentwise error rate.

Alpha= 0.05 df= 45 MSE= 5.455556

Critical Value of T= 2.01

Least Significant Difference= 2.1039

Means with the same letter are not significantly different.

T Grouping	Mean	N	SUGAR

A	70.100	10	CNTL
B	64.100	10	SUC2
C	59.300	10	GLU2
C			
C	58.200	10	FRU2
C			
C	58.000	10	GLU1FRU1

%%%

THE PEA DATA OF SOKAL & ROHLF 5

ONE WAY CLASSIFICATION BY TREATMENT

TREATMENT = SUGAR (5 LEVELS)

General Linear Models Procedure

Tukey's Studentized Range (HSD) Test for variable: LENGTH

NOTE: This test controls the type I experimentwise error rate, but generally has a higher type II error rate than REGWQ.

Alpha= 0.05 df= 45 MSE= 5.455556

Critical Value of Studentized Range= 4.018

Minimum Significant Difference= 2.9681

Means with the same letter are not significantly different.

Tukey Grouping	Mean	N	SUGAR

A	70.100	10	CNTL
B	64.100	10	SUC2
C	59.300	10	GLU2
C			
C	58.200	10	FRU2
C			
C	58.000	10	GLU1FRU1

%%%

THE PEA DATA OF SOKAL & ROHLF 6

ONE WAY CLASSIFICATION BY TREATMENT

TREATMENT = SUGAR (5 LEVELS)

General Linear Models Procedure

Bonferroni (Dunn) T tests for variable: LENGTH

NOTE: This test controls the type I experimentwise error rate, but generally has a higher type II error rate than REGWQ.

Alpha= 0.05 df= 45 MSE= 5.455556

Critical Value of T= 2.95

Minimum Significant Difference= 3.0836

Means with the same letter are not significantly different.

Bon Grouping	Mean	N	SUGAR

```
                    A            70.100     10  CNTL

                    B            64.100     10  SUC2

                    C            59.300     10  GLU2
                    C
                    C            58.200     10  FRU2
                    C
                    C            58.000     10  GLU1FRU1
```

%%%

THE PEA DATA OF SOKAL & ROHLF 7

ONE WAY CLASSIFICATION BY TREATMENT

TREATMENT = SUGAR (5 LEVELS)

General Linear Models Procedure

Scheffe's test for variable: LENGTH

NOTE: This test controls the type I experimentwise error rate but generally has a higher type II error rate than REGWF for all pairwise comparisons

Alpha= 0.05 df= 45 MSE= 5.455556

Critical Value of F= 2.57874

Minimum Significant Difference= 3.3548

Means with the same letter are not significantly different.

Scheffe Grouping	Mean	N	SUGAR

```
A            70.100     10  CNTL

B            64.100     10  SUC2

C            59.300     10  GLU2
C
C            58.200     10  FRU2
C
C            58.000     10  GLU1FRU1
```

EXAMPLE 2: The pig diet data – unequal replication. Here, as in the first example, we modify the program from section 7.13 by adding the options `LSD`, `SCHEFFE`, and `BON` in the `MEANS` statement. Note that we may not use the Tukey method here, as we do not have equal replication. Note also that the way of indicating significant results is different from that in the case of equal replication. Here, the actual comparisons are printed out, and "`***`" denotes either **comparisonwise** significance (LSD) or **familywise** significance (Bonferroni, Scheffé).

The output also gives a **confidence interval** for the difference in each case. Can you determine how these were constructed? (Remember the relationship between hypothesis tests and confidence intervals).

PROGRAM:

```
*********************************************;
*                                           ;
*          ST 511    EXAMPLE 8.2            ;
*                                           ;
*    USING SAS TO DO MULTIPLE COMPARISONS    ;
*    ON ALL PAIRWISE COMPARSIONS OF MEANS    ;
*                                           ;
*     USING PROC GLM IN THE CASE OF UNEQUAL  ;
```

```
*    REPLICATION IN A ONE-WAY CLASSIFICATION   ;
*    ILLUSTRATED BY THE PIG DIET DATA OF ZAR   ;
*                                              ;
*    ALSO USE LSD OPTIONS FOR PLANNED PAIR-    ;
*    WISE COMPARISONS                          ;
*                                              ;
**********************************************;

OPTIONS LS=80 PS= 59 NODATE;

DATA PIGS;
  INPUT DIET $ WEIGHT @@;
  CARDS;
    DIET1  133.8  DIET1  125.3  DIET1  143.1  DIET1  128.9  DIET1  135.7
    DIET2  151.2  DIET2  149.0  DIET2  162.7  DIET2  145.8  DIET2  153.5
    DIET3  225.8  DIET3  224.6  DIET3  220.4  DIET3  212.3
    DIET4  193.4  DIET4  185.3  DIET4  182.8  DIET4  188.5  DIET4  198.6
;
PROC PRINT;
  TITLE 'THE PIG DIET DATA OF ZAR';
  TITLE2 'ILLUSTRATION OF USING PROC GLM FOR A ONE-WAY ANOVA';
  TITLE3 'WITH UNEQUAL REPLICATION';
  TITLE4 'TO DO MULTIPLE COMPARISION TESTS ON ALL';
  TITLE5 'PAIRWISE COMPARSIONS OF MEANS'; RUN;

************************************************************;
*                                                          ;
*  THE SLASH (/) FOLLOWED BY THE WORDS "LSD", "BON"        ;
*  AND "SCHEFFE" TELLS SAS TO PERFORM THE LSD, BONFERRONI   ;
*  AND SCHEFFE PROCEDURES ON THE DIET MEANS                 ;
*                                                          ;
************************************************************;

PROC GLM;
```

```
CLASS DIET;
MODEL WEIGHT = DIET;
MEANS DIET / LSD BON SCHEFFE; RUN;
```

OUTPUT:

```
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%

                    THE PIG DIET DATA OF ZAR                          1
             ILLUSTRATION OF USING PROC GLM FOR A ONE-WAY ANOVA
                    WITH UNEQUAL REPLICATION
                TO DO MULTIPLE COMPARISION TESTS ON ALL
                   PAIRWISE COMPARSIONS OF MEANS

                     OBS    DIET     WEIGHT

                       1    DIET1     133.8
                       2    DIET1     125.3
                       3    DIET1     143.1
                       4    DIET1     128.9
                       5    DIET1     135.7
                       6    DIET2     151.2
                       7    DIET2     149.0
                       8    DIET2     162.7
                       9    DIET2     145.8
                      10    DIET2     153.5
                      11    DIET3     225.8
                      12    DIET3     224.6
                      13    DIET3     220.4
                      14    DIET3     212.3
                      15    DIET4     193.4
                      16    DIET4     185.3
                      17    DIET4     182.8
```

```
                    18    DIET4     188.5
                    19    DIET4     198.6

%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%

                          THE PIG DIET DATA OF ZAR                          2
                ILLUSTRATION OF USING PROC GLM FOR A ONE-WAY ANOVA
                          WITH UNEQUAL REPLICATION
                     TO DO MULTIPLE COMPARISION TESTS ON ALL
                          PAIRWISE COMPARSIONS OF MEANS

                              General Linear Models Procedure
                                  Class Level Information

                Class     Levels    Values

                DIET           4    DIET1 DIET2 DIET3 DIET4

                         Number of observations in data set = 19

%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%

                          THE PIG DIET DATA OF ZAR                          3
                ILLUSTRATION OF USING PROC GLM FOR A ONE-WAY ANOVA
                          WITH UNEQUAL REPLICATION
                     TO DO MULTIPLE COMPARISION TESTS ON ALL
                          PAIRWISE COMPARSIONS OF MEANS

                              General Linear Models Procedure

Dependent Variable: WEIGHT
```

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	20461.405763	6820.468588	164.38	0.0001
Error	15	622.399500	41.493300		
Corrected Total	18	21083.805263			

R-Square	C.V.	Root MSE	WEIGHT Mean
0.970480	3.753460	6.4415293	171.61579

Source	DF	Type I SS	Mean Square	F Value	Pr > F
DIET	3	20461.405763	6820.468588	164.38	0.0001

Source	DF	Type III SS	Mean Square	F Value	Pr > F
DIET	3	20461.405763	6820.468588	164.38	0.0001

%%

THE PIG DIET DATA OF ZAR 4

ILLUSTRATION OF USING PROC GLM FOR A ONE-WAY ANOVA

WITH UNEQUAL REPLICATION

TO DO MULTIPLE COMPARISION TESTS ON ALL

PAIRWISE COMPARSIONS OF MEANS

General Linear Models Procedure

T tests (LSD) for variable: WEIGHT

NOTE: This test controls the type I comparisonwise error rate not the experimentwise error rate.

Alpha= 0.05 Confidence= 0.95 df= 15 MSE= 41.4933

Critical Value of T= 2.13145

Comparisons significant at the 0.05 level are indicated by '***'.

DIET Comparison	Lower Confidence Limit	Difference Between Means	Upper Confidence Limit	
DIET3 - DIET4	21.845	31.055	40.265	***
DIET3 - DIET2	59.125	68.335	77.545	***
DIET3 - DIET1	78.205	87.415	96.625	***
DIET4 - DIET3	-40.265	-31.055	-21.845	***
DIET4 - DIET2	28.597	37.280	45.963	***
DIET4 - DIET1	47.677	56.360	65.043	***
DIET2 - DIET3	-77.545	-68.335	-59.125	***
DIET2 - DIET4	-45.963	-37.280	-28.597	***
DIET2 - DIET1	10.397	19.080	27.763	***
DIET1 - DIET3	-96.625	-87.415	-78.205	***
DIET1 - DIET4	-65.043	-56.360	-47.677	***
DIET1 - DIET2	-27.763	-19.080	-10.397	***

%%%

THE PIG DIET DATA OF ZAR 5

ILLUSTRATION OF USING PROC GLM FOR A ONE-WAY ANOVA

WITH UNEQUAL REPLICATION

TO DO MULTIPLE COMPARISION TESTS ON ALL

PAIRWISE COMPARSIONS OF MEANS

General Linear Models Procedure

Bonferroni (Dunn) T tests for variable: WEIGHT

NOTE: This test controls the type I experimentwise error rate but generally has a higher type II error rate than Tukey's for all pairwise comparisons.

Alpha= 0.05 Confidence= 0.95 df= 15 MSE= 41.4933

Critical Value of T= 3.03628

Comparisons significant at the 0.05 level are indicated by '***'.

DIET Comparison	Simultaneous Lower Confidence Limit	Difference Between Means	Simultaneous Upper Confidence Limit	
DIET3 - DIET4	17.935	31.055	44.175	***
DIET3 - DIET2	55.215	68.335	81.455	***
DIET3 - DIET1	74.295	87.415	100.535	***
DIET4 - DIET3	-44.175	-31.055	-17.935	***
DIET4 - DIET2	24.910	37.280	49.650	***
DIET4 - DIET1	43.990	56.360	68.730	***
DIET2 - DIET3	-81.455	-68.335	-55.215	***

DIET2 - DIET4	-49.650	-37.280	-24.910	***
DIET2 - DIET1	6.710	19.080	31.450	***
DIET1 - DIET3	-100.535	-87.415	-74.295	***
DIET1 - DIET4	-68.730	-56.360	-43.990	***
DIET1 - DIET2	-31.450	-19.080	-6.710	***

%%%

THE PIG DIET DATA OF ZAR 6

ILLUSTRATION OF USING PROC GLM FOR A ONE-WAY ANOVA

WITH UNEQUAL REPLICATION

TO DO MULTIPLE COMPARISION TESTS ON ALL

PAIRWISE COMPARSIONS OF MEANS

General Linear Models Procedure

Scheffe's test for variable: WEIGHT

NOTE: This test controls the type I experimentwise error rate but generally has a higher type II error rate than Tukey's for all pairwise comparisons.

Alpha= 0.05 Confidence= 0.95 df= 15 MSE= 41.4933

Critical Value of F= 3.28738

Comparisons significant at the 0.05 level are indicated by '***'.

DIET Comparison	Simultaneous Lower Confidence Limit	Difference Between Means	Simultaneous Upper Confidence Limit	

DIET3 - DIET4	17.485	31.055	44.625	***
DIET3 - DIET2	54.765	68.335	81.905	***
DIET3 - DIET1	73.845	87.415	100.985	***
DIET4 - DIET3	-44.625	-31.055	-17.485	***
DIET4 - DIET2	24.486	37.280	50.074	***
DIET4 - DIET1	43.566	56.360	69.154	***
DIET2 - DIET3	-81.905	-68.335	-54.765	***
DIET2 - DIET4	-50.074	-37.280	-24.486	***
DIET2 - DIET1	6.286	19.080	31.874	***
DIET1 - DIET3	-100.985	-87.415	-73.845	***
DIET1 - DIET4	-69.154	-56.360	-43.566	***
DIET1 - DIET2	-31.874	-19.080	-6.286	***

9 Multi-Way Classification and Analysis of Variance

Complementary Reading: STD, Chapter 9

9.1 Introduction

Recall from chapter 7 that when no sources of variation other than the treatments are anticipated, grouping observations will probably not add very much precision, i.e. reduce our assessment of experimental error. If the experimental units are expected to be fairly uniform, then, a **completely random design** will probably be sufficient.

In many situations, however, **other** sources of variation are anticipated.

EXAMPLES:

- In an agricultural field experiment, adjacent plots in a field will tend to be "more alike" than those far apart.
- Observations made with a particular measuring device or by a particular individual may be more alike than those made by different devices or individuals.
- Plants kept in different greenhouses may be more alike than those from different greenhouses.
- Patients treated at the same hospital may be more alike than those treated at different hospitals.

In such cases, it is clear that there is a potential source of **systematic** variation we may identify in advance. This suggests that we may wish to **group** experimental units in a meaningful way on this basis.

RESULT: When experimental units are considered in meaningful groups, they may be thought of as being **classified** not only according to **treatment** but also according to

- Position in field
- Device or observer
- Greenhouse
- Hospital.

OBJECTIVE: In an experiment, we seek to investigate differences among treatments – by accounting for differences due to effects of phenomena such as those above, a possible source of variation will (hopefully) be **excluded** from our assessment of experimental error. The result will be increased ability to detect **treatment differences** if they exist. **Designs** involving meaningful **grouping** of experimental units are the key to reducing the effects of experimental error, by identifying components of variation among experimental units that may be due to something besides inherent biological variation among them. The paired design for comparing two treatments is an example of such a design.

MULTI-WAY CLASSIFICATION: If experimental units may be classified not only according to treatment but to other meaningful factors, things obviously become more complicated. In this chapter, we discuss designs involving more than one way of classifying experimental units. In particular, we consider:

- Two-way classification, where experimental units may be classified by treatment and another meaningful grouping factor
- A form of three-way classification (one of which, of course, is treatment) called a **Latin square**.

9.2 Randomized complete block design

When experimental units may be **meaningfully grouped**, e.g. by area of field, device, greenhouse, hospital, and so on, clearly, a **completely randomized** design will be suboptimal. In this situation, an alternative strategy for assigning treatments to experimental units, which takes advantage of the grouping, may be used.

RANDOMIZED COMPLETE BLOCK DESIGN:

- The **groups** are called **blocks**.
- **Each** treatment appears the **same** number of times in **each** block; hence, the term **complete block design**.
- The simplest case is that where each treatment appears **exactly once** in each block. Here, because

  $$\text{number of } \textbf{replicates} \;=\; \text{number of experimental units for each treatment},$$

  we have

  $$\text{Number of replicates} \;=\; \text{number of blocks} \;= r.$$

- **Blocks** are often called **replicates** for this reason.
- To set up such a design, **randomization** is used in the following way:
  - Assign experimental units to blocks on the basis of the meaningful grouping factor (greenhouse, device, etc)
  - Now **randomly assign** the treatments to experimental units **within** each block.

  Hence, the term **randomized complete block design**: each block is **complete**, and **randomization** occurs within each block.

RATIONALE: Experimental units within blocks are alike as possible, so observed differences among them should be mainly attributable to the treatments. To ensure this interpretation holds, in the conduct of the experiment, all experimental units within a block should be treated as uniformly as possible:

- In a field, all plots should be harvested at the same time of day.
- All measurements using a single device should be made by the same individual if different people use it in a different way.
- All plants in a greenhouse should be watered at the same time of day or by the same amount.

ADVANTAGES:

- **Greater precision** is possible than with a completely random design with one way classification.
- Increased **scope of inference** is possible because more experimental conditions may be included.

DISADVANTAGES:

- If there is a large number of treatments, a **large** number of experimental units per block will be required. Thus, large variation among experimental units within blocks might still arise, with the result that no precision is gained, but experimental procedure is more complicated. In this case, other designs may be more appropriate.

In this chapter, we consider first randomized complete block designs. Treatment is of course one of the classifications, and is the one of interest in any multi-way classification. The other classification(s) are most often not of explicit interest in their own right, but are chosen to increase precision for detecting differences among treatments, because experimental error is reduced. The idea is to account for acknowledged, potential sources of variation **up front**, in the **design** of an experiment, so that if treatment differences do exist, this will be elucidated in the most efficient manner possible.

It turns out that the same ideas of constructing F ratios that will be large if treatment differences exist, that is, **analysis of variance**, may be extended to this setting.

9.3 Linear additive model for two-way classification

We assume here that **one observation** is taken on each experimental unit (i.e. sampling unit = experimental unit). Assume that a **randomized complete block design** is used with exactly one experimental unit per treatment per block.

For the two-way classification with t treatments, we may classify an individual observation as being from the jth block on the ith treatment:

$$Y_{ij} = \underbrace{\mu + \tau_i}_{\mu_i} + \beta_j + \epsilon_{ij} = \underbrace{\mu_i + \beta_j}_{\mu_{ij}} + \epsilon_{ij} = \mu_{ij} + \epsilon_{ij},$$

$$i = 1, \ldots, t; \qquad j = 1, \ldots, r,$$

where

- t = number of treatments
- r = number of replicates on treatment i; that is, the **number of blocks**
- μ = overall mean (as before)
- τ_i = effect of the ith treatment (as before)
- $\mu_i = \mu + \tau_i$ mean of the population for the ith treatment
- β_j = effect of the jth block
- $\mu_{ij} = \mu + \tau_i + \beta_j = \mu_i + \beta_j$ mean of the population for the ith treatment in the jth block
- ϵ_{ij} = "error" describing all other sources of variation (e.g. inherent variation among experimental units not attributable to treatments or blocks).

In this model, the effect of the jth block, β_j, is a deviation from the overall mean μ attributable to being an experimental unit in that block. It is a systematic deviation, the same for all experimental units in the block, thus formally characterizing the fact the experimental units in the same block are "alike."

In fact, this model is just an extension of that for the **paired** design with **two** treatments considered in chapter 5. In that model, we had a term ρ_j, the effect of the jth **pair**. Here, it should be clear that this model just extends the idea behind **meaningful pairing** of observations to groups larger than two (and more than two treatments). Thus, the **paired** design is just a special case of a **randomized complete block design** in the case of two treatments.

FIXED AND RANDOM EFFECTS: As in the one way classification, the treatment **and** block effects, τ_i and β_j, may be regarded as **fixed** or **random**. We have already discussed the notion of regarding **treatments** as having fixed or random effects. We may also apply the same reasoning to **blocks**.

Note that there are a number of possibilities:

- **Both** τ_i and β_j are best regarded as having **fixed effects**. In this case, both describe a particular set of conditions that will not vary from experiment to experiment. For example:

Treatments	3 specific drugs
Blocks	2 breeds of dog

- **Both** the τ_i and β_j are best regarded as having **random effects**. That is, τ_i and β_j are thought of as random variables drawn from the populations of all possible treatments and blocks, respectively, with variances σ^2_τ and σ^2_β. For example:

Treatments	3 machines (chosen at random from all machines at a company)
Blocks	4 machine operators (chosen at random from all operators employed at the company)

- We may also have the situation of a "mixed" model, containing both fixed and random effects. Most often,

  Treatment effects τ_i are **fixed**
  Block effects β_j are **random**

  For example:

Treatments	3 fertilizers
Blocks	2 greenhouses

  That is, if we wish our inferences to apply to all possible greenhouses, we regard the greenhouses as a **random** sample from the population of all possible greenhouses.

  If, on the other hand, the scope of inference for our experiment is only the effects of fertilizers for a particular company that maintains 2 greenhouses, we could regard them as **fixed** block effects.

COMPLICATION: It turns out that, unlike in the case of the one way classification with one sampling unit per experimental unit, the distinction between fixed and random effects becomes very important in higher way classifications. In particular, although the computation of quantities in an analysis of variance table may be the same, the **interpretation** of these quantities, i.e. what the MSs involved estimate, will **depend** on what's fixed and what's random. We will point this out in the course of our discussion.

MODEL RESTRICTION: Just as in the one way classification, the linear additive model above is **overparameterized**. If we think about the mean for the ith treatment in the jth block, μ_{ij}, it should be clear that we do not know how much of what we see is attributable to each of the components μ, τ_i, and β_j.

ONE APPROACH: Again, the model is mainly a device for helping us to think about the sources of variation in the data. Again, then, to reconcile our desire to have a helpful model for thinking and the mathematics, we impose **restrictions** that then govern how we think about the model.

The usual approach is to impose the restrictions

$$\sum_{i=1}^{t} \tau_i = 0 \text{ and } \sum_{j=1}^{r} \beta_j = 0.$$

One way to think about this is to suppose that the **overall** mean μ may be thought of as the **average** of the μ_{ij}, that is

$$\mu = \frac{1}{rt}\sum_{i=1}^{t}\sum_{j=1}^{r}\mu_{ij} = \frac{1}{rt}\sum_{i=1}^{t}\sum_{j=1}^{r}(\mu + \tau_i + \beta_j) = \mu + \frac{1}{t}\sum_{i=1}^{t}\tau_i + \frac{1}{r}\sum_{j=1}^{r}\beta_j.$$

We thus see that, in order for μ to have this interpretation, we must have $\sum_{i=1}^{t}\tau_i = 0$ and $\sum_{j=1}^{r}\beta_j = 0$, which are the restrictions above.

As before, these restrictions go along with the interpretation of τ_i and β_j as "deviations" from an overall mean. The treatments "affect" the response in different "directions;" some of the τ_i must be negative and others positive for them all to sum to zero. Now, so do the blocks!

It is thus common to see null and alternative hypotheses written in terms of τ_i and β_j in this context, as we will exhibit shortly.

This interpretation is valid in the case where the τ_i and β_j are **fixed effects**. When they are **random effects**, the interpretation is similar, as in the one way case.

9.4 Analysis of variance for two-way classification – randomized complete block design with no subsampling

ASSUMPTIONS: Before we turn to the analysis, we reiterate the assumptions that underlie the validity of the methods:

- The observations, and hence the errors, are **normally distributed**
- The observations have the **same variance**.

These assumptions are necessary for F ratios we construct to have the F sampling distribution. Recall our discussion on the validity of assumptions in chapter 7 – the same issues apply here (and, indeed, **always**).

IDEA: As for the one way classification, we **partition** the Total SS, which measures **all variation** in the data from all sources, into "independent" components describing variation attributable different sources. These components are the numerators of estimators for "variance." We develop the idea by thinking of **both** the treatment and block effects τ_i and β_j as being **fixed**.

NOTATION: Define

$$\bar{Y}_{\cdot\cdot} = \frac{1}{rt}\sum_{i=1}^{t}\sum_{j=1}^{r} Y_{ij} = \text{ overall sample mean}$$

$$\bar{Y}_{i\cdot} = \frac{1}{r}\sum_{j=1}^{r} Y_{ij} = \text{ sample mean for treatment } i \text{ (over all blocks)}$$

$$\bar{Y}_{\cdot j} = \frac{1}{t}\sum_{i=1}^{t} Y_{ij} = \text{ sample mean for block } j \text{ (over all treatments)}$$

$$C = \frac{\left(\sum_{i=1}^{t}\sum_{j=1}^{r} Y_{ij}\right)^2}{rt} = \text{ correction factor}$$

The Total SS is, as usual, the numerator of the **overall sample variance** for all the data, without regard to treatments **or**, now, blocks:

$$\text{Total SS } = \sum_{i=1}^{t}\sum_{j=1}^{r}(Y_{ij} - \bar{Y}_{\cdot\cdot})^2 = \sum_{i=1}^{t}\sum_{j=1}^{r} Y_{ij}^2 - C.$$

As before, this is based on $rt - 1$ independent quantities (degrees of freedom).

As before, the numerator of a F ratio for testing treatment mean differences ought to involve a variance estimate containing a component due to variation among the treatment means. This quantity will be identical to that in chapter 7, by the same rationale! We thus have

$$\text{Treatment SS } = r\sum_{i=1}^{t}(\bar{Y}_{i\cdot} - \bar{Y}_{\cdot\cdot})^2 = \frac{\sum_{i=1}^{t}\left(\sum_{j=1}^{r} Y_{ij}\right)^2}{r} - C$$

This will have $t-1$ degrees of freedom, again by the same argument. This assessment of variation effectively **ignores** the blocks, as it is based on **averaging** over them.

By an entirely **similar** argument with **blocks** in place of treatments, we may arrive at an analogous quantity:

$$\text{Block SS } = t\sum_{j=1}^{r}(\bar{Y}_{\cdot j} - \bar{Y}_{\cdot\cdot})^2 = \frac{\sum_{j=1}^{r}\left(\sum_{i=1}^{t} Y_{ij}\right)^2}{t} - C$$

This will have $r-1$ degrees of freedom. This assessment of variation effectively **ignores** the treatments, as it is based on **averaging** over them.

ERROR SS: Again, we need a SS that represents the variation we are attributing to **experimental error**. If we are to have the same situation as in the one way classification case, where all SSs are additive and sum to the Total SS, we must have

$$\text{Block SS } + \text{ Treatment SS } + \text{ Error SS } = \text{ Total SS.}$$

Solving this equation for Error SS and doing the algebra, we arrive at

$$\text{Error SS } = \sum_{i=1}^{t}\sum_{j=1}^{r}(Y_{ij} - \bar{Y}_{i\cdot} - \bar{Y}_{\cdot j} + \bar{Y}_{\cdot\cdot})^2. \tag{9.1}$$

That is, if it were indeed true that we could **partition** Total SS into these components, then (9.1) would have to be the quantity characterizing experimental error. Inspection of this quantity seems pretty hopeless for attaching such an interpretation!

INTERPRETATION: In fact, it is **not** hopeless. Recall our **model restrictions**

$$\sum_{i=1}^{t} \tau_i = 0 \text{ and } \sum_{j=1}^{r} \beta_j = 0$$

and what they imply, namely, that the τ_i and β_j may be regarded as **deviations** from an overall mean μ. Consider then **estimation** of these quantities under the restrictions and this interpretation: the obvious estimators are

$$\hat{\mu} = \bar{Y}_{..}, \quad \hat{\tau}_i = \bar{Y}_{i.} - \bar{Y}_{..}, \quad \hat{\beta}_j = \bar{Y}_{.j} - \bar{Y}_{..}$$

The "hats" denote estimation of these quantities. Thus, if we wanted to estimate $\mu_{ij} = \mu + \tau_i + \beta_j$, the estimator would be

$$\hat{\mu}_{ij} = \bar{Y}_{..} + (\bar{Y}_{i.} - \bar{Y}_{..}) + (\bar{Y}_{.j} - \bar{Y}_{..}).$$

Because in our linear additive model

$$\epsilon_{ij} = Y_{ij} - \mu_{ij}$$

characterizes whatever variation we are attributing to experimental error, we would hope that an appropriate Error SS would be based on an **estimate** of ϵ_{ij}. Such an estimate is

$$\begin{aligned} Y_{ij} - \hat{\mu}_{ij} &= Y_{ij} - \{\bar{Y}_{..} + (\bar{Y}_{i.} - \bar{Y}_{..}) + (\bar{Y}_{.j} - \bar{Y}_{..})\} \\ &= Y_{ij} - \bar{Y}_{i.} - \bar{Y}_{.j} + \bar{Y}_{..} \end{aligned}$$

by algebra. Note that this quantity is precisely that in (9.1); thus, our candidate quantity for Error SS does indeed make sense – it is the sum of squared deviations of the individual observations, Y_{ij} from their "sample mean," an estimate of all "left over" variation, $\hat{\epsilon}_{ij}$.

The degrees of freedom associated with Error SS is of course the number of independent quantities on which it depends. Considering our partition

$$\text{Block SS} \ + \ \text{Treatment SS} \ + \ \text{Error SS} \ = \ \text{Total SS},$$

we have degrees of freedom

$$(r-1) + (t-1) + (t-1)(r-1) = rt - 1;$$

that is, for the degrees of freedom to add up, Error SS must have $(t-1)(r-1)$ degrees of freedom. In fact, we may justify this intuitively. We arrived at the form of Error SS under the restrictions that the τ_i and β_j sum to zero. Thus, these restrictions involve $t-1$ and $r-1$ independent quantities, respectively. Thus, there are $(t-1)(r-1)$ combinations of independent quantities.

We summarize this information in an ANOVA table as follows:

Two Way ANOVA table – Randomized Complete Block Design

Source of variation	DF	SS Definition	SS Computation	MS	F
Among Blocks	$r-1$	$t\sum_{j=1}^{r}(\bar{Y}_{\cdot j}-\bar{Y}_{\cdot\cdot})^2$	$\frac{\sum_{j=1}^{r}\left(\sum_{i=1}^{t}Y_{ij}\right)^2}{t}-C$	MS_B	$F_B=\frac{MS_B}{MS_E}$
Among Treatments	$t-1$	$r\sum_{i=1}^{t}(\bar{Y}_{i\cdot}-\bar{Y}_{\cdot\cdot})^2$	$\frac{\sum_{i=1}^{t}\left(\sum_{j=1}^{r}Y_{ij}\right)^2}{r}-C$	MS_T	$F_T=\frac{MS_T}{MS_E}$
Error	$(t-1)(r-1)$	$\sum_{i=1}^{t}\sum_{j=1}^{r}(Y_{ij}-\bar{Y}_{i\cdot}-\bar{Y}_{\cdot j}+\bar{Y}_{\cdot\cdot})^2$	by subtraction	MS_E	
Total	$rt-1$	$\sum_{i=1}^{t}\sum_{j=1}^{r}(Y_{ij}-\bar{Y}_{\cdot\cdot})^2$	$\sum_{i=1}^{t}\sum_{j=1}^{r}Y_{ij}^2-C$		

Here,

$$MS_T=\frac{\text{Treatment SS}}{t-1},\quad MS_B=\frac{\text{Block SS}}{r-1},\quad MS_E=\frac{\text{Error SS}}{(t-1)(r-1)}.$$

STATISTICAL HYPOTHESES: The primary question of interest in this setting is to determine if the means of the t treatment populations are different. We may write this formally as

$$H_{0,T}:\mu_1=\mu_2=\cdots=\mu_t \text{ vs. } H_{1,T}: \text{ The } \mu_i \text{ are not all equal.}$$

This may also be written in terms of the τ_i under the restriction $\sum_{i=1}^{t}\tau_i=0$ as

$$H_{0,T}:\tau_1=\tau_2=\cdots=\tau_t=0 \text{ vs. } H_{1,T}: \text{ The } \tau_i \text{ are not all equal.}$$

TEST PROCEDURE: Reject $H_{0,T}$ in favor of $H_{1,T}$ at level of significance α if

$$F_T > F_{t-1,(t-1)(r-1),\alpha}.$$

A secondary question of interest might be whether there is a systematic effect of blocks. We may write a set of hypotheses for this under our model restrictions as

$$H_{0,B}: \beta_1 = \cdots \beta_r = 0 \text{ vs. } H_{1,B}: \text{ The } \beta_j \text{ are not all equal.}$$

TEST PROCEDURE: Reject $H_{0,B}$ in favor of $H_{1,B}$ at level of significance α if

$$F_B > F_{r-1,(t-1)(r-1),\alpha}.$$

NOTE: In most experiments, whether or not there are block differences is not really a main concern, because, by considering blocks up front we have acknowledged them as a possible nontrivial source of variation. If we test whether block effects are different and reject $H_{0,B}$, then, by blocking, we have probably increased the precision of our experiment, our original objective.

EXPECTED MEAN SQUARES: We developed the above in the case where both τ_i and β_j have **fixed** effects. It is instructive to examine the **expected mean squares** under our linear additive model under this condition as well as the cases where (i) **both** τ_i and β_j are **random** and (ii) the case where τ_i is **fixed** but β_j are **random** (the "mixed" case). This allows insight into the suitability of the test statistics given above.

Source of	Expected Mean Square		
variation	Both fixed	Both random	Mixed
MS_B	$\sigma^2_\epsilon + t\frac{\sum_{j=1}^r \beta_j^2}{r-1}$	$\sigma^2_\epsilon + t\sigma^2_\beta$	$\sigma^2_\epsilon + t\sigma^2_\beta$
MS_T	$\sigma^2_\epsilon + r\frac{\sum_{i=1}^t \tau_i^2}{t-1}$	$\sigma^2_\epsilon + r\sigma^2_\tau$	$\sigma^2_\epsilon + r\frac{\sum_{i=1}^t \tau_i^2}{t-1}$
MS_E	σ^2_ϵ	σ^2_ϵ	σ^2_ϵ

Here, σ^2_ϵ is the variance associated with the ϵ_{ij}, that is, corresponding to what we are attributing to experimental error in this situation.

From this table, we have

- *Both τ_i and β_j fixed*: F_T and F_B are both appropriate. The MSs in the numerators of these statistics, MS_T and MS_B estimate σ^2_ϵ (estimated by MS_E) **plus** an additional term that is equal to zero under $H_{0,T}$ and $H_{0,B}$, respectively.

- *Both τ_i and β_j random*: Note here that MS_T and MS_B estimate σ^2_ϵ **plus** an extra term involving the variances σ^2_τ and σ^2_β, respectively, characterizing variability in the populations of all possible treatments and blocks. Thus, under these conditions, F_T and F_B are appropriate for testing

$$H_{0,T}: \sigma^2_\tau = 0 \text{ vs. } H_{1,T}: \sigma^2_\tau > 0,$$

$$H_{0,B}: \sigma^2_\beta = 0 \text{ vs. } H_{1,B}: \sigma^2_\beta > 0.$$

  These hypotheses are the obvious ones of interest when τ_i and β_j are **random**.

- *"Mixed" model*: For τ_i fixed and β_j random, the same observations as above apply. F_T and F_B are appropriate for testing

$$H_{0,T}: \tau_1 = \tau_2 = \cdots = \tau_t = 0 \text{ vs. } H_{1,T}: \text{ The } \tau_i \text{ are not all equal and}$$

$$H_{0,B}: \sigma^2_\beta = 0 \text{ vs. } H_{1,B}: \sigma^2_\beta > 0,$$

  respectively.

EXAMPLE: (Balaam, 1972, *Fundamentals of Biometry*, p. 170. The following data are yields in bushels/acre from an agricultural experiment set out in a randomized complete block design. The experiment was designed to investigate the differences in yield for 7 hybrid varieties of wheat, labeled A–G here. A field was divided into 5 blocks, each containing 7 plots. In each plot, the 7 plots were assigned at random to be planted with the 7 varieties, one plot for each variety. A yield was recorded for each plot.

Here, then, the **plots** are the **experimental units**, and, because only one observation was taken on each, they are also the **sampling units**. The randomization of treatments to experimental units was carried out **within** each block.

We assume that the measurements are approximately normally distributed, which seems reasonable for such continuous measurement data, with the same variance.

We have $t = 7$, $r = 5$.

	Variety							
Block	A	B	C	D	E	F	G	$\sum_{i=1}^{t} Y_{ij}$
I	10	9	11	15	10	12	11	78
II	11	10	12	12	10	11	12	78
III	12	13	10	14	15	13	13	90
IV	14	15	13	17	14	16	15	104
V	13	14	16	19	17	15	18	112
$\sum_{j=1}^{r} Y_{ij}$	60	61	62	77	66	67	69	$\sum_{i=1}^{t}\sum_{j=1}^{r} Y_{ij} = 462$

Calculations:

$$C = \frac{\left(\sum_{i=1}^{t}\sum_{j=1}^{r} Y_{ij}\right)^2}{rt} = \frac{(462)^2}{(7)(5)} = 6098.4$$

$$\frac{\sum_{i=1}^{t}\left(\sum_{j=1}^{r} Y_{ij}\right)^2}{r} = \frac{(60^2 + \cdots + 69^2)}{5} = 6140.0$$

$$\frac{\sum_{j=1}^{r}\left(\sum_{i=1}^{t} Y_{ij}\right)^2}{t} = \frac{(78^2 + \cdots + 112^2)}{7} = 6232.6$$

$$\sum_{i=1}^{t}\sum_{j=1}^{r} Y_{ij}^2 = 10^2 + \cdots + 18^2 = 6314.0$$

Thus,

Treatment SS $= 6140.0 - 6098.4 = 41.6$

Block SS $= 6232.6 - 6098.4 = 134.2$

Total SS $= 6314.0 - 6098.4 = 215.6$

Error SS = Total SS − Treatment SS − Block SS $= 215.6 - 134.2 - 41.6 = 39.8$.

Analysis of Variance – Variety Data

Source of variation	DF	SS	MS	F
Blocks	4	134.2	33.54	20.21
Treatments	6	41.6	6.93	4.18
Error	24	39.8	1.66	
Total	34	215.6		

To perform the hypothesis test for differences among the treatment means, we compare F to the appropriate value from the F table. For level of significance $\alpha = 0.05$, we have $F_{6,24,0.05} = 2.51$, and $4.18 > 2.51$. We thus **Reject** $H_{0,T}$. There is evidence in these data to suggest that there are differences in mean yields among the varieties.

To test the hypothesis on block differences, we find $F_{4,24,0.05} = 2.78$. We have $20.21 > 2.87$, thus we **Reject** $H_{0,B}$. There is strong evidence in these data to suggest differences in mean yield across blocks.

In section 9.9, we show how to use SAS to conduct this type of analysis.

USEFULNESS OF BLOCKING: Note from these results that the **blocking** served to explain much of the overall variation. To appreciate this further, suppose that we had **not** blocked the experiment, but instead had just conducted the experiment according to a **completely random design**. Suppose that we ended up with the same data as in the experiment above.

Under these conditions, variety is the only classification factor for the plots, and we would construct the following analysis of variance using the methods of chapter 7:

Analysis of Variance – Variety Data

Source of variation	DF	SS	MS	F
Treatments	6	41.6	6.93	1.12
Error	28	174.0	6.21	
Total	34	215.6		

The test for differences in mean yield for the varieties (treatments) would be to compare $F_T = 1.12$ to $F_{6,28,0.05} = 2.45$. Note that we would thus **not reject** the null hypothesis of no treatment differences at level $\alpha = 0.05$!

MORAL: In the one way classification experiment and analysis, there is no accounting for the variation in the data that is actually attributable to a **systematic** source, position in the field (the factor used to block the experiment above)! The one way analysis has no choice but to attribute this variation to **experimental error**; that is, it regards this variation as just part of the **inherent** variation among experimental units that we can not explain. The result is that the Error SS in the one way analysis contains both variation due to position in field (which is actually **systematic** variation) and inherent variation.

Here, note that

$$134.2 + 39.8 = 174.0 \text{ and } 4 + 24 = 28.$$

That is, the Error SS for the one way classification analysis, which actually may be regarded as **ignoring** the blocks (because what we really did was "pretend" they didn't exist), is equal to the sum of the Block SS and Error SS for the two way classification analysis. Thus, note that the blocking serves to **partition** what is regarded as inherent variation in an analysis ignoring blocks into two "independent" components, one due to what is actually **systematic** (explainable) variation and one due to what, after accounting for blocks, we attribute to inherent (unexplainable) variation.

Thus, in the one way analysis, MS_E is **too big**, and we could not reject $H_{0,T}$. By blocking the experiment, and explicitly acknowledging position in the field as a potential source of variation, MS_E was sufficiently reduced so that we **could** identify variety differences.

Two things to take away from this exercise:

- Blocking may be an effective means of explaining variation (increasing precision) so that differences among treatments that may really exist are more likely to be detected.
- The data from an experiment set up according to a particular design should be analyzed according to the appropriate procedure for that design! The above shows that if we set up the experiment according to a randomized complete block design, but then analyzed it as if it had been set up according to a completely randomized design, **erroneous inference** results, in this case, failure to identify real differences in treatments! **The design of an experiment dictates the analysis**!!!

9.5 ANOVA for two-way classification – randomized complete block design with subsampling

In the previous sections, we assumed that exactly one observation was taken on each of the rt experimental units in a randomized complete block design. We now consider the case where each of the rt experimental units has **more than one** sampling unit, where, as in the one way classification case, the sampling units are chosen randomly from within the experimental units.

We consider only the case of an **equal number** s of sampling units per experimental unit. Just as in the one way classification, when numbers of sampling units are **unequal**, the quality of information on different experimental units **differs**, and the analysis is more complicated (see chapter 7).

LINEAR ADDITIVE MODEL: As before, it is useful to write down a linear additive model. Here, we may identify an observation (now on a **sampling unit** within an experimental unit) as being the kth observation from the jth block on the ith treatment. The subscripts for i and j identify the **experimental unit** (one per particular treatment/block combination); then k identifies the sampling unit on this experimental unit. The model is:

$$Y_{ijk} = \underbrace{\mu + \tau_i}_{\mu_i} + \beta_j + \epsilon_{ij} + \delta_{ijk} = \underbrace{\mu_i + \beta_j}_{\mu_{ij}} + \epsilon_{ij} + \delta_{ijk} = \mu_{ij} + \epsilon_{ij} + \delta_{ijk},$$

$$i = 1, \ldots, t; \qquad j = 1, \ldots, r; \qquad k = 1, \ldots, s,$$

where μ, μ_i, μ_{ij}, τ_i, β_j are defined as before, t is still the number of treatments, r the number of blocks, so that there are still rt experimental units involved. We now have

- s = number of sampling units on each experimental unit.
- ϵ_{ij} now represents the "error" associated with the particular experimental unit (the one getting treatment i in block j), with variance σ^2_ϵ (the variance of the population of experimental units)
- δ_{ijk} = "error" associated with the response on the kth sampling unit on the experimental unit getting treatment i in block j. The δ_{ijk} thus represent the fact that all sampling units on a particular experimental unit are not exactly alike. Assume they vary with variance σ^2.

The motivation for the form of the analysis of variance table and the composition of the sums of squares is the same as for the one way classification, as described in chapter 7. The only difference is the presence of the component β_j representing the systematic effect due to the jth block. As you may expect (especially considering the discussion at the end of the last section), the addition of this component means that, algebraically, what previously was Experimental Error SS for the one way classification is further decomposed in this setting into two "independent" pieces: a Block SS and what is now Experimental Error SS. The measure of systematic variation due to the blocking factor is extracted from our measure of variation among experimental units. Recall from our discussion in chapter 7 that **experimental error** measures the **two** sources of inherent variation in experimental units that may make responses on experimental units differ: variation due to differences among experimental units and among sampling units within them. The Sampling Error SS measures solely inherent variation in sampling units.

Again, this is a **nested** model; see the discussion in chapter 7.

We thus omit a detailed description, and simply present the relevant notation and ANOVA table below. the treatments.

NOTATION: Define

$$\bar{Y}_{i\cdot\cdot} = \frac{1}{rs}\sum_{j=1}^{r}\sum_{k=1}^{s} Y_{ijk} = \text{ mean for treatment } i \text{ over all blocks, sampling units}$$

$$\bar{Y}_{\cdot j\cdot} = \frac{1}{ts}\sum_{j=1}^{r}\sum_{k=1}^{s} Y_{ijk} = \text{ mean for block } j \text{ over all treatments, sampling units}$$

$$\bar{Y}_{ij\cdot} = \frac{1}{s}\sum_{k=1}^{s} Y_{ijk} = \text{ mean on sampling units for } j\text{th replicate on treatment } i$$

$$\bar{Y}_{\cdots} = \frac{1}{trs}\sum_{i=1}^{t}\sum_{j=1}^{r}\sum_{k=1}^{s} Y_{ijk} = \textbf{ overall } \text{sample mean .}$$

For algebraic convenience, denote the **correction factor** as

$$C = \frac{\left(\sum_{i=1}^{t}\sum_{j=1}^{r}\sum_{k=1}^{s} Y_{ijk}\right)^2}{trs}.$$

We give the ANOVA table with definitions of the various SSs; formulas for calculation follow.

Two-Way ANOVA table – RCBD with Subsampling

Source of variation	DF	SS Definition	MS	F
Among Blocks	$r-1$	$ts\sum_{j=1}^{r}(\bar{Y}_{.j.}-\bar{Y}_{...})^2$	MS_B	F_B
Among Treatments	$t-1$	$rs\sum_{i=1}^{t}(\bar{Y}_{i..}-\bar{Y}_{...})^2$	MS_T	F_T
Experimental Error	$(t-1)(r-1)$	$s\sum_{i=1}^{t}\sum_{j=1}^{r}(\bar{Y}_{ij.}-\bar{Y}_{...})^2$ $-$ Block SS $-$ Treatment SS	MS_E	F_S
Sampling Error	$tr(s-1)$	$\sum_{i=1}^{t}\sum_{j=1}^{r}\sum_{k=1}^{s}(Y_{ijk}-\bar{Y}_{ij.})^2$	MS_S	
Total	$trs-1$	$\sum_{i=1}^{t}\sum_{j=1}^{r}\sum_{k=1}^{s}(Y_{ijk}-\bar{Y}_{...})^2$		

CALCULATIONS:

- Correction factor
$$C=\frac{\left(\sum_{i=1}^{t}\sum_{j=1}^{r}\sum_{k=1}^{s}Y_{ijk}\right)^2}{trs}$$
- Total SS
$$\sum_{i=1}^{t}\sum_{j=1}^{r}\sum_{k=1}^{s}Y_{ijk}^2-C$$
- Block SS
$$\frac{\sum_{j=1}^{r}\left(\sum_{i=1}^{t}\sum_{k=1}^{s}Y_{ijk}\right)^2}{ts}-C$$
- Treatment SS
$$\frac{\sum_{i=1}^{t}\left(\sum_{j=1}^{r}\sum_{k=1}^{s}Y_{ijk}\right)^2}{rs}-C$$
- Experimental Error
$$\frac{\sum_{i=1}^{t}\sum_{j=1}^{r}\left(\sum_{k=1}^{s}Y_{ijk}\right)^2}{s}-C-\text{ Block SS }-\text{ Treatment SS}$$
- Find the Sampling Error SS by subtraction:

$$\text{SamplingError SS }=\text{Total SS }-\text{ Block SS }-\text{ Treatment SS }-\text{ Experimental Error SS.}$$

As usual,

$$MS_T = \frac{\text{Treatment SS}}{t-1}, \quad MS_B = \frac{\text{Block SS}}{r-1}, \quad MS_E = \frac{\text{Error SS}}{(t-1)(r-1)}.$$

Just as in the one way classification, **Experimental Error** is the appropriate error for the denominator of the F ratios. We will see verification of this momentarily from a table of expected mean squares.

STATISTICAL HYPOTHESES: The major question of interest is to determine if the means of the t treatment populations differ:

$$H_{0,T}: \mu_1 = \cdots = \mu_t \text{ vs. } H_{1,T}: \text{ The } \mu_i \text{ are not all equal}$$

or equivalently under the restriction $\sum_{i=1}^{t} \tau_i = 0$,

$$H_{0,T}: \tau_1 = \cdots \tau_t = 0 \text{ vs. } H_{1,T}: \text{ The } \tau_i \text{ are not all equal}$$

(for the fixed effects case).

TEST PROCEDURE: For testing $H_{0,T}$ vs. $H_{1,T}$ at level of significance α, reject $H_{0,T}$ if

$$F_T > F_{t-1,(t-1)(r-1),\alpha}.$$

For testing whether there is a systematic effect due to blocks, the formal hypotheses are (under the restriction $\sum_{j=1}^{r} \beta_j = 0$)

$$H_{0,B}: \beta_1 = \cdots \beta_r = 0 \text{ vs. } H_{1,B}: \text{ The } \beta_j \text{ are not all equal.}$$

TEST PROCEDURE: Reject $H_{0,B}$ in favor of $H_{1,B}$ at level of significance α if

$$F_B > F_{r-1,(t-1)(r-1),\alpha}.$$

As in the one way case, we may also wish to test whether experimental error is due mainly just to variation among individual sampling units or whether there are also differences among experimental units themselves. Using the same rationale as in chapter 7, and recalling the definitions of σ^2_ϵ and σ^2, the formal hypotheses are

$$H_{0,S}: \sigma^2_\epsilon = 0 \text{ vs. } \sigma^2_\epsilon > 0.$$

TEST PROCEDURE: For testing $H_{0,S}$ vs. $H_{1,S}$ at level α, reject $H_{0,S}$ if

$$F_T > F_{(t-1)(r-1),tr(s-1),\alpha}.$$

That these tests are appropriate in not only the fixed effects case but in other cases as well is evident from the table of expected mean squares:

Source of	Expected Mean Square		
variation	Both fixed	Both random	Mixed
MS_B	$\sigma^2 + s\sigma^2_\epsilon + st\frac{\sum_{j=1}^r \beta_j^2}{r-1}$	$\sigma^2 + s\sigma^2_\epsilon + st\sigma^2_\beta$	$\sigma^2 + s\sigma^2_\epsilon + st\sigma^2_\beta$
MS_T	$\sigma^2 + s\sigma^2_\epsilon + sr\frac{\sum_{i=1}^t \tau_i^2}{t-1}$	$\sigma^2 + s\sigma^2_\epsilon + sr\sigma^2_\tau$	$\sigma^2 + s\sigma^2_\epsilon + sr\frac{\sum_{i=1}^t \tau_i^2}{t-1}$
MS_E	$\sigma^2 + s\sigma^2_\epsilon$	$\sigma^2 + s\sigma^2_\epsilon$	$\sigma^2 + s\sigma^2_\epsilon$

Note that in each case, MS_E estimates $\sigma^2 + s\sigma^2_\epsilon$, which takes into account variation both among and within experimental units (through the variances σ^2_ϵ and σ^2, respectively). In each case, both MS_T and MS_B estimate this quantity **plus** a term representing the **extra variation** due to treatments or blocks (through either τ_i, β_j or σ^2_τ, σ^2_β, depending on the model). For the fixed effects case, these terms will be zero if $H_{0,T}$ or $H_{0,B}$ is true. Similarly, when effects of either type are random, it is clear that the hypotheses being tested are thus

$$H_{0,T}: \sigma^2_\tau = 0 \text{ vs. } H_{1,T}: \sigma^2_\tau > 0.$$

$$H_{0,B}: \sigma^2_\beta = 0 \text{ vs. } H_{1,B}: \sigma^2_\beta > 0.$$

Thus, in all cases, the tests based on F_T and F_B above are valid.

It is also clear, as in the one way classification case, that the hypotheses tested by F_S are

$$H_{0,S}: \sigma^2_\epsilon = 0 \text{ vs. } H_{1,S}: \sigma^2_\epsilon > 0.$$

EXAMPLE: The operators of a nursery with 2 greenhouses would like to investigate differences among 3 fertilizers they might use on plants they are growing for commercial sale. They are interested in plant heights (mm) after 6 weeks. To set up the experiment, they randomly select 12 similar seedlings and randomly allocate them to 6 trays, 2 per tray. The trays are then randomly allocated to be placed in the 2 greenhouses, 3 trays per greenhouse. For each greenhouse, the 3 fertilizers are assigned to the trays by a random mechanism, so that each of the 3 trays receives a different fertilizer. At the end of 6 weeks, the heights of each seedling (two in teach tray) are measured.

In this setup, the trays are thus the **experimental units** to which the treatments (fertilizers) are applied. The treatments were randomly assigned to receive treatments, one tray per treatment, within each greenhouse (block); thus, this is a randomized complete block design. The 2 seedlings in each tray are the **sampling units**. To summarize:

Treatments	Fertilizers	$t = 3$
Blocks	Greenhouses	$r = 2$
Experimental Units	Trays	$rt = 6$
Sampling Units	Seedlings	$s = 2$

Fertilizer	Greenhouse		$\sum_{j=1}^{r}\sum_{k=1}^{s} Y_{ijk}$
	I	II	
1	47	46	
	43	40	176
2	62	67	
	68	71	268
3	41	42	
	39	46	168
$\sum_{i=1}^{t}\sum_{k=1}^{s} Y_{ijk}$	300	312	$\sum_{i=1}^{t}\sum_{j=1}^{r}\sum_{k=1}^{s} Y_{ijk} = 612$

Calculations:

$$C = \frac{\left(\sum_{i=1}^{t}\sum_{j=1}^{r} Y_{ij}\right)^2}{rt} = \frac{612^2}{12} = 31212$$

$$\text{Total SS } = \sum_{i=1}^{t}\sum_{j=1}^{r}\sum_{k=1}^{s} Y_{ijk}^2 - C = 32854 - 31212 = 1642$$

$$\text{Block SS } = \frac{\sum_{j=1}^{r}\left(\sum_{i=1}^{t}\sum_{k=1}^{s} Y_{ijk}\right)^2}{ts} - C = 31224 - 31212 = 12$$

$$\text{Treatment SS } = \frac{\sum_{i=1}^{t}\left(\sum_{j=1}^{r}\sum_{k=1}^{s} Y_{ijk}\right)^2}{rs} - C32756 - 31212 = 1544$$

$$\begin{aligned}\text{Experimental Error} &= \frac{\sum_{i=1}^{t}\sum_{j=1}^{r}\left(\sum_{k=1}^{s} Y_{ijk}\right)^2}{s} - C - \text{ Block SS } - \text{ Treatment SS} \\ &= (90^2 + 86^2 + \cdots + 88^2)/2 - 31212 - 12 - 1544 \\ &= 32792 - 31212 - 12 - 1544 = 24.\end{aligned}$$

$$\text{Sampling Error } = 1642 - 12 - 1544 - 24 = 62.$$

Source of variation	DF	SS	MS	F
Blocks	1	12	12	1.00
Treatments	2	1544	772	64.33
Experimental Error	2	24	12	1.16
Sampling Error	6	62	10.3	
Total	11	1642		

We use level of significance $\alpha = 0.05$.

For testing $H_{0,T}$ vs. $H_{1,T}$, we have $F_{2,2,0.05} = 19.00$ and $F_T = 64.33 > 19.00$. **Reject** $H_{0,T}$; there is strong evidence in these data to suggest that there is a difference in mean 6-week height for the 3 fertilizers.

Was blocking beneficial? We gain insight by testing $H_{0,B}$ vs. $H_{1,B}$, we have $F_{1,2,0.05} = 18.51$ and $F_B = 1.00$. Thus, we **do not reject** $H_{0,B}$; there is not enough evidence in these data to suggest that there are differences in the heights attained after 6 weeks due to greenhouse.

For testing whether the trays vary in addition to the variation among seedlings within them, we test $H_{0,S}$ vs. $H_{1,S}$. $F_{2,6,0.05} = 5.14$ and $F_S = 1.16$. **Do not reject** $H_{0,S}$; there is not enough evidence to suggest that trays vary in terms of 6-week heights above and beyond just the variation in seedlings within them.

9.6 ANOVA for two-way classification – randomized complete block design with more than one experimental unit per treatment per block

So far, we have been concerned with the situation where each treatment is seen **only once** within each block. We may take only a single observation on each experimental unit or sample them more than once, but the fact remains that each treatment is observed on **more than one** experimental unit in each block.

There is an intuitively apparent drawback to this type of design. Because we only see each treatment **once** in each block, we do not have sufficient information to determine whether part of what we observe for a particular treatment/block combination is actually due to something **systematic**. This is best discussed in the context of an example.

MOTIVATING EXAMPLE: STD discuss an example in which a farmer would like to determine the effect on yield of 3 cultivars (the treatments), one of which is known to be drought-resistant. The farmer is interested in results as they pertain to yields he himself might realize; thus he is interested in how the cultivars differ in the area of his farm where he might plant them. He thus divides this area into 3 blocks in order to conduct an experiment, basing the blocks on what he knows about the wetness in various parts of this area. (The blocks may be regarded as **fixed** here, as they are in the only area of interest to the farmer.) One of the blocks is on a hillside and is quite dry.

It is natural to expect that the drought-resistant cultivar might give higher yields in the dry block relative to the other cultivars in the dry block as opposed to how they might compare in the other blocks, because the drought-resistant cultivar is specifically designed to give superior yields under dry conditions.

Such an effect is not **random**, but is a **systematic** effect attributable to the particular cultivar/block combination. If the farmer conducted the experiment so that each block was divided into 3 plots, each plot was randomized to a different cultivar, and each plot was harvested to give a single yield value, he would have only one experimental unit (plot) per treatment/block combination. From this information, he would not be able to determine whether a high yield for the drought-resistant cultivar in the dry block **really was** a result of this suspected systematic effect or just a **chance result** for the particular plot! He might suspect it, but he will never be sure!

This issue may be seen easily if we consider an linear additive model for this experiment. As usual, let Y_{ij} be the yield for cultivar i in block j, with τ_i denoting the effect of cultivar i, β_j the effect of the jth block, and ϵ_{ij} denoting the "error" associated with the plot receiving combination i, j ("inherent variation" in the experimental unit). If we also believe there is a **systematic effect** associated with seeing **particular** cultivar–block combinations, then the model should contain a **further component** representing systematic deviations in mean response due to particular combinations. Let

$$(\tau\beta)_{ij} = \text{ the additional effect of using cultivar } i \text{ in block } j.$$

Then a linear additive model that describes Y_{ij} would be

$$Y_{ij} = \mu + \tau_i + \beta_j + \underbrace{(\tau\beta)_{ij} + \epsilon_{ij}} \ . \tag{9.2}$$

Inspection of this model illustrates the problem. The highlighted terms $(\tau\beta)_{ij}$ and ϵ_{ij} both pertain to what happens when treatment i is used in block j. The part represented by $(\tau\beta)_{ij}$ is due solely to systematic effects of the treatment/block combination, while the part represented by ϵ_{ij} has to do with the (random) nature of the plot to which the combination is "applied." Together, they represent the total variation in the response specific to the combination i, j. Without further information, however, we have no way of knowing how much of the variation is due to the systematic component and how much is due to the (random) plot component!

TERMINOLOGY: When the difference in mean response among treatments is **different** depending on block, an **interaction** between treatments and blocks is said to exist.

In the example, we expect the difference among cultivars to be **different** in the dry block than it might be in other blocks. Thus, we suspect a cultivar-block (representing wetness) interaction.

To understand this further, consider the model (9.2), and suppose that there are 2 treatments ($i = 1, 2$) and 2 blocks ($j = 1, 2$) for simplicity. The part of the model that describes the mean response for treatment i in block j is

$$\mu + \tau_i + \beta_j + (\tau\beta)_{ij}, \quad i, j, = 1, 2.$$

Consider the mean difference between treatments (treatment 1 − treatment 2) in each block:

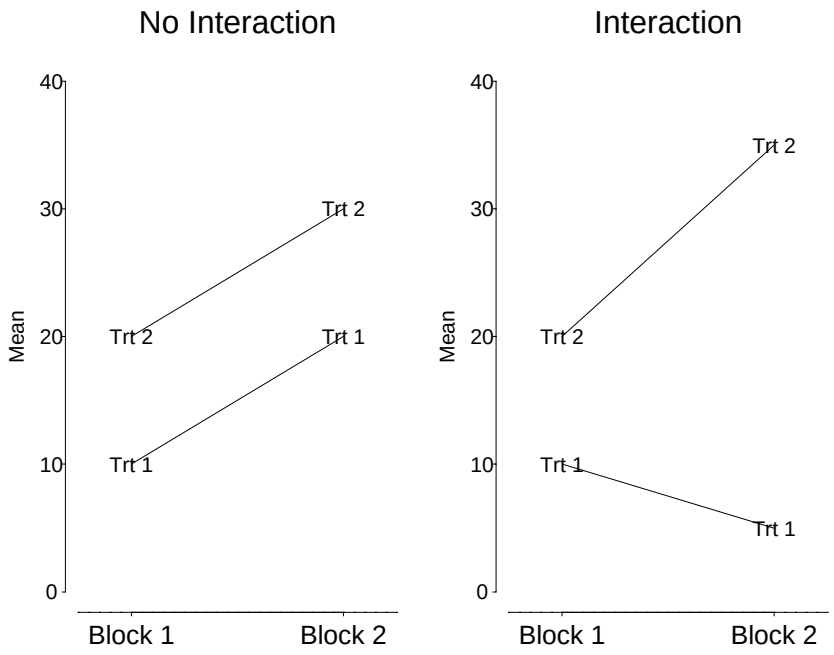
$$\begin{aligned}
\text{Block 1} \quad & \{\mu + \tau_1 + \beta_1 + (\tau\beta)_{11}\} - \{\mu + \tau_2 + \beta_1 + (\tau\beta)_{21}\} = (\tau_1 - \tau_2) - \{(\tau\beta)_{11} - (\tau\beta)_{21}\} \\
\text{Block 2} \quad & \{\mu + \tau_1 + \beta_2 + (\tau\beta)_{12}\} - \{\mu + \tau_2 + \beta_2 + (\tau\beta)_{21}\} = (\tau_1 - \tau_2) - \{(\tau\beta)_{12} - (\tau\beta)_{22}\}
\end{aligned}$$

Note that the difference **depends** on which block is under consideration! If $(\tau\beta)_{ij} = 0$ for all i, j, then we suspect there is **no** systematic effect due to particular block/treatment combinations. Note that in this case the difference in treatment mean response reduces to

$$(\tau_1 - \tau_2) \text{ inn } \textbf{each} \text{ block!}$$

That is, if there is no such systematic effect present, then the difference in treatment mean is **the same** regardless of block. If there **is** such an effect, it is not, and the component $(\tau\beta)_{ij}$ characterizes how the differences differ!

This is illustrated in the following picture. In the picture, the means for each of the two treatments are plotted for each block. For each treatment, the means are then joined by a line across blocks, so that each line indicates the change in treatment mean from Block 1 to Block 2 for each treatment. With **no interaction**, the change in mean going from Block 1 to Block 2 is **the same** for each treatment (that is, $= \tau_1 - \tau_2$). **With Interaction**, the change in mean going from Block 1 to Block 2 is **different** for each treatment, reflecting the different mean differences given above due to the $(\tau\beta)_{ij}$!



RESULT: If we suspect that differences in treatment means might be different depending on the conditions represented by the blocks, then an experiment that has one experimental unit allocated to each treatment/block combination will not provide the necessary information to determine whether this is the case. Any differences that are actually due to this systematic effect will be attributed to **inherent variation**, because we have no way of determining otherwise.

SOLUTION: If we were to observe each treatment/block combination on **more than one** experimental unit, we would be in a position to sort out what is systematic and what isn't. This would entail setting up the experiment as follows. For t treatments, form r blocks, each consisting of st experimental units. Randomly assign the t treatments to the st experimental units in each block, so that s experimental units are allocated to each treatment. Suppose we take a single observation on each experimental unit.

To understand how this will help, consider the linear additive model for such an experiment.

LINEAR ADDITIVE MODEL: We now formally write down a linear additive model for this situation. Here, we may identify an observation as being on the kth experimental unit in the jth block receiving the ith treatment. **All three subscripts** are required to identify a particular experimental unit, on which there is is **single** observation taken, Y_{ijk}. Note that, although we again use k as a subscript as in the previous section, it is indexing something different: here, k is indexing the **multiple** (> 1) experimental units receiving treatment i in block j.

(If furthermore **subsamples** were taken on each experimental unit, we would need yet **another** subscript to identify the individual sampling units. We do not consider this extension here, but confine our attention to the case where the experimental unit and sampling unit are the same.)

The model is:

$$Y_{ijk} = \mu + \tau_i + \beta_j + (\tau\beta)_{ij} + \epsilon_{ijk} \tag{9.3}$$

$$i = 1, \ldots, t; \qquad j = 1, \ldots, r; \qquad k = 1, \ldots, s,$$

where μ, τ_i, β_j are defined as before, t is still the number of treatments, and r is the number of blocks. There are now rt treatment/block combinations, and k indexes the experimental units on each combination. Thus, s is the number of experimental units seen at each combination, and there are rts total experimental units (and observations) in all.

We also have

- $(\tau\beta)_{ij}$ represents the **interaction effect** of the ith treatment and jth block. This is a systematic effect that describes how the mean might change because this particular combination was used.
- ϵ_{ijk} represents the "error" associated with the kth experimental unit getting treatment i in block j with variance σ^2_ϵ (the variance of the population of experimental units).

Inspection of the model (9.3) shows why we may now expect to get a handle on the systematic interaction effect. Because we have s experimental units all subjected to the i, j combination, their **sample mean** $\bar{Y}_{ij\cdot}$ should contain information on $(\tau\beta)_{ij}$. The variation about this mean would tell us about the inherent variation in experimental units.

INTERPRETATION: Compare this model to that for subsampling in the last section. In this model, the components $(\tau\beta)_{ij}$ and ϵ_{ijk} replace ϵ_{ij} and δ_{ijk} in the subsampling model; other than this change in symbols, the models look very similar. **However**, the interpretation is quite different! It is important that you feel comfortable with the fact that, although the models have a similar form, they represent experiments that are very different.

Algebraically, it turns out that the **same computations** apply in constructing an analysis of variance relevant to model (9.3) as did in the subsampling model. Again, however, the **interpretation** of the components of this analysis is very different.

NATURE OF $(\tau\beta)_{ij}$: In the above discussion, the systematic **interaction** effect is regarded as a **fixed** effect; e.g. in the cultivar example, the both blocks and treatments were the only ones of interest and hence **fixed effects** themselves. Any additional effect of these in combination would hence **also** be **fixed**.

The development we now undertake applies **only** to the case where **all effects** τ_i, β_j, and $(\tau\beta)_{ij}$ are regarded as **fixed**. As we will discuss shortly, unlike the two previous two-way classification models we have considered in previous sections, the distinction between **fixed** and **random** effects in this model becomes important, as it **changes** the analysis.

MODEL RESTRICTION: Just as in previous models, **restrictions** must be imposed in order for the model to have the usual interpretation. The restrictions here are

$$\sum_{i=1}^{t} \tau_i = 0, \quad \sum_{j=1}^{r} \beta_j = 0, \quad \sum_{i=1}^{t} (\tau\beta)_{ij} = 0 \text{ for each } j, \quad \sum_{j=1}^{r} (\tau\beta)_{ij} = 0 \text{ for each } i.$$

These restrictions allow μ to maintain its interpretation as the "overall mean."

Rather than go through all the arguments, we simply write down the analysis of variance table and the calculations.

NOTATION: Define

$$\bar{Y}_{i\cdot\cdot} = \frac{1}{rs}\sum_{j=1}^{r}\sum_{k=1}^{s} Y_{ijk} = \text{ mean for treatment } i \text{ over all blocks, experimental units}$$

$$\bar{Y}_{\cdot j\cdot} = \frac{1}{ts}\sum_{j=1}^{r}\sum_{k=1}^{s} Y_{ijk} = \text{ mean for block } j \text{ over all treatments, experimental units}$$

$$\bar{Y}_{ij\cdot} = \frac{1}{s}\sum_{k=1}^{s} Y_{ijk} = \text{ mean on experimental units on treatment } i \text{ in block } j$$

$$\bar{Y}_{\cdots} = \frac{1}{trs}\sum_{i=1}^{t}\sum_{j=1}^{r}\sum_{k=1}^{s} Y_{ijk} = \textbf{ overall} \text{ sample mean .}$$

For algebraic convenience, denote the **correction factor** as

$$C = \frac{\left(\sum_{i=1}^{t}\sum_{j=1}^{r}\sum_{k=1}^{s} Y_{ijk}\right)^2}{trs}.$$

Two-Way ANOVA table – RCBD with More than One Experimental Unit per Treatment/Block

Source of variation	DF	SS Definition	MS	F
Among Blocks	$r-1$	$ts\sum_{j=1}^{r}(\bar{Y}_{\cdot j\cdot}-\bar{Y}_{\cdots})^2$	MS_B	F_B
Among Treatments	$t-1$	$rs\sum_{i=1}^{t}(\bar{Y}_{i\cdot\cdot}-\bar{Y}_{\cdots})^2$	MS_T	F_T
Interaction	$(t-1)(r-1)$	$s\sum_{i=1}^{t}\sum_{j=1}^{r}(\bar{Y}_{ij\cdot}-\bar{Y}_{\cdots})^2$ − Block SS − Treatment SS	MS_I	F_I
Experimental Error	$tr(s-1)$	$\sum_{i=1}^{t}\sum_{j=1}^{r}\sum_{k=1}^{s}(Y_{ijk}-\bar{Y}_{ij\cdot})^2$	MS_E	
Total	$trs-1$	$\sum_{i=1}^{t}\sum_{j=1}^{r}\sum_{k=1}^{s}(Y_{ijk}-\bar{Y}_{\cdots})^2$		

CALCULATIONS:

- Correction factor
$$C = \frac{\left(\sum_{i=1}^{t}\sum_{j=1}^{r}\sum_{k=1}^{s} Y_{ijk}\right)^2}{trs}$$
- Total SS
$$\sum_{i=1}^{t}\sum_{j=1}^{r}\sum_{k=1}^{s} Y_{ijk}^2 - C$$
- Block SS
$$\frac{\sum_{j=1}^{r}\left(\sum_{i=1}^{t}\sum_{k=1}^{s} Y_{ijk}\right)^2}{ts} - C$$
- Treatment SS
$$\frac{\sum_{i=1}^{t}\left(\sum_{j=1}^{r}\sum_{k=1}^{s} Y_{ijk}\right)^2}{rs} - C$$
- Interaction SS
$$\frac{\sum_{i=1}^{t}\sum_{j=1}^{r}\left(\sum_{k=1}^{s} Y_{ijk}\right)^2}{s} - C - \text{ Block SS } - \text{ Treatment SS}$$
- Find the Experimental Error SS by subtraction:

$$\text{Experimental Error SS } = \text{Total SS } - \text{ Block SS } - \text{ Treatment SS } - \text{ Interaction SS.}$$

As usual,

$$MS_T = \frac{\text{Treatment SS}}{t-1}, \quad MS_B = \frac{\text{Block SS}}{r-1}, \quad MS_I = \frac{\text{Interaction SS}}{(t-1)(r-1)}, \quad MS_E = \frac{\text{Experimental Error SS}}{rt(s-1)}.$$

For the case of **fixed effects only**, **Experimental Error** is the appropriate error for the denominator of the F ratios. We will discuss the basis for this, and why things are different when there are random effects, shortly.

STATISTICAL HYPOTHESES: The major interest focuses on the treatments. Note that if we suspect an **interaction**, then we believe that differences among the treatments may themselves be different under the different conditions represented by the blocks. We thus may be interested in a test on this issue **first**; if the differences among treatments are very different depending on which block we are in, this may have implications for how we interpret the results of the experiment. Thus, it is often common practice to test for interaction first and inspect a graph of the treatment means like the one we made for the simplest case of $r = 2$, $t = 2$.

The formal hypotheses for interaction, under our model restrictions, are as follows:

$H_{0,I}$: There are no interaction effects, i.e. all $(\tau\beta)_{ij} = 0$ vs. $H_{1,I}$: At least one $(\tau\beta)_{ij}$ is not 0.

TEST PROCEDURE: To conduct the test at level of significance α, reject $H_{0,I}$ if

$$F_I > F_{(t-1)(r-1),rt(s-1),\alpha}.$$

TREATMENT MAIN EFFECTS: Recall the restrictions we imposed on the $(\tau\beta)_{ij}$ in our linear additive model. They must sum to zero both across blocks and across treatments. This allows a certain interpretation of the test based on F_T. The hypotheses being tested are

$$H_{0,T} : \tau_1 = \cdots \tau_t = 0 \text{ vs. } H_{1,T} : \text{ The } \tau_i \text{ are not all equal.}$$

From the model,

$$Y_{ijk} = \mu + \tau_i + \beta_j + (\tau\beta)_{ij} + \epsilon_{ijk},$$

recall that the mean for treatment i in block j is

$$\mu_{ij} = \mu + \tau_i + \beta_j + (\tau\beta)_{ij},$$

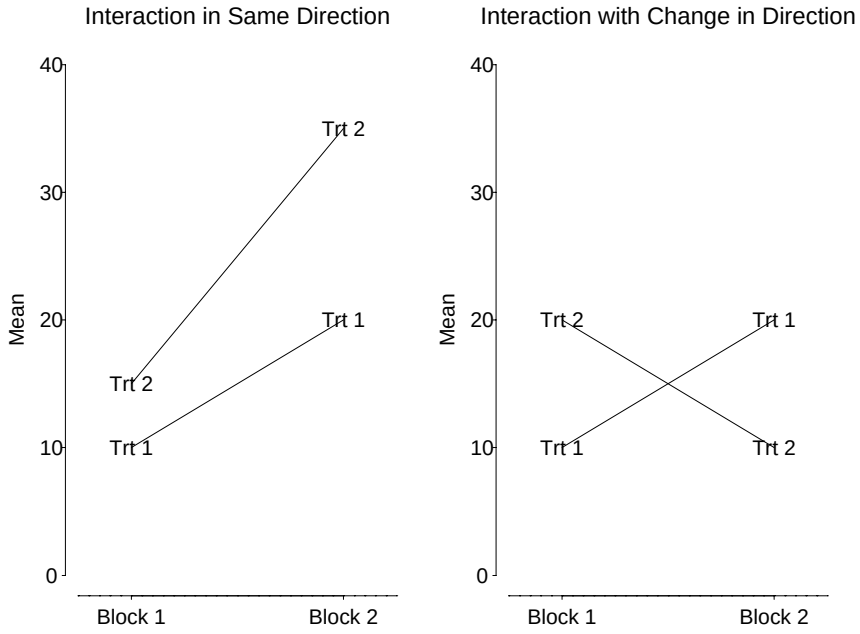
say. This mean pertains to the particular effect of treatment i under the conditions in block j. If we think of the **average** of such means for treatment i across the conditions in **all** blocks, this mean would be

$$\frac{1}{r}\sum_{j=1}^{r}\mu_{ij} = \mu + \tau_i + \frac{1}{r}\sum_{j=1}^{r}\beta_j + \frac{1}{r}\sum_{j=1}^{r}(\tau\beta)_{ij} = \mu + \tau_i,$$

because of the restrictions $\sum_{j=1}^{r}\beta_j = 0$ and $\sum_{j=1}^{r}(\tau\beta)_{ij} = 0$ for any treatment i.

We now see what we are testing when we test $H_{0,T}$ vs. $H_{1,T}$. If all τ_i are 0, this says that the means for all treatments, **averaged across blocks**, are the same. This is often referred to as testing **main effects** of the treatments. If indeed there is really **no interaction**, then differences in treatment means are the same in all blocks, and we don't need to think about averaging across blocks. In this case, the interpretation of the test is as before.

Whether this test is meaningful depends on the situation. To see this, consider an example. Suppose that $r = 2$, $t = 2$, and consider the two situations depicted in the picture.



In the left hand situation, the differences between the 2 treatments is indeed different in each block, but Trt 2 always gives a higher mean responses. The average mean across blocks for Trt 1 will be **lower** than that for Trt 2. In this situation, then, the test of main effects might be useful – for this type of behavior, knowing whether one treatment is really better than the other on average is useful, because they differ in the **same direction**.

In the right hand situation, the difference between the 2 treatments is such that it **changes direction**. Here, Trt 2 is better than Trt 1 under the conditions of Block 1, but exactly the opposite is true for Block 2! Note, however, that the **average** of the means for each treatment is **the same**! A test for main effects would find no difference. But this is really **uninteresting** – in this situation, this average isn't useful, because it represents something that can never really happen under real conditions (like those in Blocks 1 and 2).

TEST PROCEDURE: For testing $H_{0,T}$ vs. $H_{1,T}$ at level of significance α, reject $H_{0,T}$ if

$$F_T > F_{t-1,rt(s-1),\alpha}.$$

BLOCK MAIN EFFECTS: The situation for blocks is entirely analogous to that for treatments; in the above argument, average over treatments instead of blocks. The same discussion of whether this is interesting, if indeed one is interested in blocks at all, depends on the situation. Usually, as in the case of no interaction, testing blocks may be uninteresting. If one does, the formal hypotheses are

$$H_{0,B}: \beta_1 = \cdots \beta_r = 0 \text{ vs. } H_{1,B}: \text{ The } \beta_j \text{ are not all equal.}$$

TEST PROCEDURE: Reject $H_{0,B}$ in favor of $H_{1,B}$ at level of significance α if

$$F_B > F_{r-1,rt(s-1),\alpha}.$$

EXPECTED MEAN SQUARES: The tests above are specific to the case where **all effects are fixed**. We now verify that they are indeed valid, and look at what happens under other conditions.

When the **block** effects are **random**, we think of the blocks as arising from a population of all possible blocks. Analogous to how we restrict the β_j to sum to zero when they are fixed, we think of the **random** β_j as coming from a population with mean 0 and some variance σ^2_β. Similarly, for **random** treatment effects, we think of the τ_i as coming from a population with mean 0 and variance σ^2_τ.

Regardless of whether the τ_i are fixed or random, if the β_j are, then how should we think about **interaction**? Recall that in the case where all components are fixed, we thought of the $(\tau\beta)_{ij}$ as fixed deviations unique to particular treatment/block combinations that have the effect of making the differences in treatment means different for different blocks. When blocks are **random**, we may think of each block in the population of all possible blocks as having associated with it a $(\tau\beta)_{ij}$ value – if treatment i were applied to a block chosen from this population, the associated deviation would be this value. From this perspective, it seems sensible to think of the $(\tau\beta)_{ij}$ as being **random** as well. One can envision a population of $(\tau\beta)_{ij}$ values for each treatment i containing all the possible deviations that could arise for each possible block! If we think of the $(\tau\beta)_{ij}$ in this way, then, analogous to our model restrictions in the fixed case, we think of them as having mean 0 and some variance $\sigma^2_{\tau\beta}$, say.

With this interpretation, here are the **expected mean squares** under various conditions (the "mixed" situation means treatments are fixed, blocks are random).

Source of variation	Expected Mean Square: Both fixed	Both random	Mixed
MS_B	$\sigma^2_\epsilon + st\frac{\sum_{j=1}^r \beta_j^2}{r-1}$	$\sigma^2_\epsilon + s\sigma^2_{\tau\beta} + st\sigma^2_\beta$	$\sigma^2_\epsilon + st\sigma^2_\beta$
MS_T	$\sigma^2_\epsilon + sr\frac{\sum_{i=1}^t \tau_i^2}{t-1}$	$\sigma^2_\epsilon + s\sigma^2_{\tau\beta} + sr\sigma^2_\tau$	$\sigma^2_\epsilon + s\sigma^2_{\tau\beta} + sr\frac{\sum_{i=1}^t \tau_i^2}{t-1}$
MS_I	$\sigma^2_\epsilon + s\frac{\sum_{i=1}^t \sum_{j=1}^r (\tau\beta)_{ij}^2}{(t-1)(r-1)}$	$\sigma^2_\epsilon + s\sigma^2_{\tau\beta}$	$\sigma^2_\epsilon + s\sigma^2_{\tau\beta}$
MS_E	σ^2_ϵ	σ^2_ϵ	σ^2_ϵ

- The table shows that the tests described above when all components have **fixed** effects are indeed valid.

- Note, however, that when **blocks** are random, in either the **both random** or **mixed** cases, the EMS for MS_T contains an additional term involving $\sigma^2_{\tau\beta}$. Thus, in the mixed case, for example, the numerator and denominator of the usual $F_T = MS_T/MS_E$ ratio estimates

$$\frac{\sigma^2_\epsilon + s\sigma^2_{\tau\beta} + sr\frac{\sum_{i=1}^t \tau_i^2}{t-1}}{\sigma^2_\epsilon}.$$

  That is, the F ratio **no longer** has the property it will be small when the $\tau_i = 0$, because the numerator still involves the variance σ^2_β pertaining to blocks!

- An F ratio that **does** have the need property is

$$F_T^* = \frac{MS_T}{MS_I}.$$

- Note further that the F_B ratio is suitable when treatment effects are **fixed** but, analogous to that for treatments, needs to be replaced by $F_B^* = MS_B/MS_I$ when both block and treatment effects are random.

RESULT: Whether effects are fixed or random will have implications for how one tests hypotheses of interest when interaction is present. It is thus critical to be realistic about how you wish inferences to apply.

EXAMPLE: (Zar, *Biostatistical Analysis*, p. 164) The following experiment was set up to determine the effect of a certain hormone treatment on plasma calcium levels (in mg/100 ml) for a certain type of bird. It was thought that sex (gender) of the birds might also play a role. (Random) samples of 10 male birds and 10 female birds were obtained. For each sample, 5 birds were randomly assigned to receive the hormone treatment, and the remaining 5 did not. Thus, we may view this as a RCBD – the **blocks** are sexes ($r = 2$), which are obviously **fixed**. Within each block, **experimental units** (the birds) were randomized to receive the treatments (hormone yes or no, $t = 2$), with $s = 5$ birds on each treatment/block combination. A single plasma calcium level was recorded for each bird. The treatments are obviously **fixed** as well.

It was suspected that the difference between having the hormone treatment or not might be different depending on whether a bird were male or female. That is, a hormone/sex **interaction** is suspected.

REMARK: If we had obtained only 2 male and 2 female birds, and randomly allocated the treatments (yes or no) to the males and females, we would not be able to investigate interaction!

	Males (Block 1)	Females (Block 2)	$\sum_{j=1}^{r}\sum_{k=1}^{s} Y_{ijk}$
Hormone Treatment	32.0	39.1	
(yes)	23.8	26.2	
	28.8	21.3	301.5
	25.0	35.8	
	29.3	40.2	
No Hormone Treatment	14.5	16.5	
(no)	11.0	18.4	
	10.8	12.7	135.0
	14.3	14.0	
	10.0	12.8	
$\sum_{i=1}^{t}\sum_{k=1}^{s} Y_{ijk}$	199.5	237.0	$\sum_{i=1}^{t}\sum_{j=1}^{r}\sum_{k=1}^{s} Y_{ijk} = 436.5$

The algebra is exactly the same as for the case of subsampling, so is omitted. Of course, remember that the **interpretation** is quite different!

Source of variation	DF	SS	MS	F
Blocks	1	70.3125	70.3125	3.07
Treatments	1	1386.1125	1386.1125	60.53
Interaction	1	4.9005	4.9005	0.21
Experimental Error	16	366.3720	22.8982	
Total	19	1827.6975		

We use level of significance $\alpha = 0.05$. For all tests, the appropriate F value is $F_{1,16,0.05} = 4.49$.

We first test for interaction. We have $F_I = 0.21$, which does not exceed 4.49. Thus, we **do not reject** $H_{0,I}$. There is not enough evidence in these data to suggest a sex-hormone interaction. Note that the sample means for each combination are

Hormone/Sex	
yes, M	27.78
no, M	12.12
yes, F	32.52
no, F	14.88

From these means, we see that the difference between treatments (yes−no) is in the **same direction** for both males and females. We also see that the differences for males (15.66) and for females (17.64) are fairly comparable in magnitude. The test for interaction is failing to reject because it finds the difference to be similar (15.66 vs. 17.64) in each block.

Even though we did not reject the null hypothesis of interaction, it could be that we did not because we did not have sufficient precision. Thus, it still makes sense to look at the **sample evidence**. Because things go in the same direction for males and females, the test of $H_{0,T}$ vs. $H_{1,T}$ for main effects is interesting. We have $F_T = 60.50 >> 4.49$. **Reject** $H_{0,T}$: There is strong evidence to suggest that mean plasma calcium levels differ depending on whether or not a bird is treated, averaged across both sexes.

Was blocking beneficial? For $H_{0,B}$ vs. $H_{1,B}$, $F_B = 3.07$, so we **do not reject** the null hypothesis. There does not seem to be enough evidence to suggest that, averaged across the treatments, mean response is different for males and females.

This example is implemented using SAS in section 9.9.

9.7 Three-way classification – the Latin square

MOTIVATION: There are often situations where it may be necessary to account for **two** sources of variation by blocking. If the number of treatments and levels of each blocking factor is large, the size of the experiment may become unwieldy or resources may be limited. Thus, in agricultural field experiments (and other situations, as we'll see shortly in an example), a particular setup is often used that allows differences among treatments to be assessed with less resources.

In field experiments, the physical layout is that of a **square** with rows of plots. For 4 treatments A, B, C, D, the layout would be

A D C B
B C A D
D A B C
C B D A

This type of setup would be useful when, for example, variability due to soil differences, etc., arises in two directions. Each plot would constitute a **single experimental unit**.

A similar layout might be used in other settings. For example, a greenhouse experiment might be laid out on a long bench. For 4 treatments A–D, pots might be laid out in groups of 4 with space in between:

A D C B — B C A D — D A B C — C B D A

We may think of this abstractly as a "square" as well. Each group is a row, and the positions within each row are the columns. Such an arrangement might be useful if we are concerned about two factors: position in the greenhouse (rows) and being on an "end" or between two other plants.

This particular kind of setup, with **two** blocking variables ("rows" and "columns"), in which the numbers of rows, columns, and treatments are the **same**, is known as a **Latin square**.

FEATURES:

- The design is laid out, either physically or more abstractly, such that each treatment appears **exactly once** in each row and each column.
- Thus, **all possible** combinations of treatments and blocking factors are **not seen**. A single experimental unit is placed in each position.
- This means that far fewer resources need be used relative to an experiment where all possible combinations are included. **However**, intuitively, this means that we have **no hope** of assessing possible **interaction** effects!
- Thus, a Latin square design is only appropriate if we believe interaction effects will be negligible or irrelevant.

NOTATION: Because the numbers of treatments, rows, and columns are the same, the number of **replicates** on each treatment is equal to the number of treatments, row, and columns. We will denote this as r.

For given value of r, there may be **several** ways to construct a Latin square. This is actually a mathematical exercise. Extensive listings of ways to construct Latin squares for different values of r are often given in texts on experimental design.

- **Randomization** consists of choosing one of the possible designs for given r at random. Then, randomly assign the letters A, B, C, D, etc to the treatments of interest.

LINEAR ADDITIVE MODEL: To write down a model, we need to be a bit careful with the notation. The key is that, although we have 3 classifications (row, column, treatment), we **do not** have $r \times r \times r = r^3$ observations; rather, we only have $r \times r = r^2$. Thus, our previous practice of letting subscripts index each classification does not extend straightforwardly to this situation. We thus use the following notation:

$$Y_{ij(t)} = \text{ observation for the treament appearing in row } i, \text{ column } j.$$

The notation emphasizes that, for **each** i, j combination, only one of the r treatments is valid.

The linear model is written

$$Y_{ij(t)} = \mu + \beta_i + \kappa_j + \tau_{(t)} + \epsilon_{ij},$$

where

- $i, j = 1, \ldots, r$
- μ is an "overall" mean
- β_i and κ_j represent the effects of the ith row and jth column
- $\tau_{(t)}$ represents the effect of the treatment appearing at position i, j
- ϵ_{ij} = "error" associated with the experimental unit appearing at position i, j.

MODEL RESTRICTION: As usual, this model is overparameterized. Thus, we impose the usual restrictions, which imply that the effects β_i, κ_j, and $\tau_{(t)}$ may be interpreted as deviations from the overall mean, μ, due to row, column,treatment:

$$\sum_{i=1}^{r} \beta_i = 0, \quad \sum_{j=1}^{r} \kappa_j = 0, \quad \sum_{t=1}^{r} \tau_{(t)} = 0.$$

NOTATION: To set up the analysis of variance, we drop the subscript (t) on Y for convenience, and write

$$\bar{Y}_{\cdot\cdot} = \frac{1}{r^2} \sum_{i=1}^{r} \sum_{j=1}^{r} Y_{ij} = \text{ overall sample mean}$$

$$\bar{Y}_{i\cdot} = \frac{1}{r} \sum_{j=1}^{r} Y_{ij} = \text{ sample mean for row } i$$

$$\bar{Y}_{\cdot j} = \frac{1}{r} \sum_{i=1}^{r} Y_{ij} = \text{ sample mean for column } j$$

$$\bar{Y}_t = \text{ average of } Y \text{ values on treatment } t$$

We write the correction factor as

$$C = \frac{\left(\sum_{i=1}^{r} \sum_{j=1}^{r} Y_{ij}\right)^2}{r^2}.$$

As usual, sums of squares for rows, columns, and treatments will be based on deviations measuring how the means for each row, column, and treatment differ from the overall mean:

$$(\bar{Y}_{i\cdot} - \bar{Y}_{\cdot\cdot}), \quad (\bar{Y}_{\cdot j} - \bar{Y}_{\cdot\cdot}), \quad (\bar{Y}_{t} - \bar{Y}_{\cdot\cdot}).$$

The Total SS will be as usual

$$\sum_{i=1}^{r}\sum_{j=1}^{r}(Y_{ij} - \bar{Y}_{\cdot\cdot})^2.$$

It turns out that, in order that we have the usual **partition** of Total SS into "independent" components for Rows, Columns, and Treatments, the quantity representing Experimental Error SS would have to be based on

$$(Y_{ij} - \bar{Y}_{i\cdot} - \bar{Y}_{\cdot j} - \bar{Y}_{t} + 2\bar{Y}_{\cdot\cdot}). \tag{9.4}$$

This appears to have no intuitive rationale. However, under our model restrictions, as before, the obvious estimators of β_i, κ_j, and $\tau_{(t)}$ are the **deviations** of the sample means from the overall sample mean:

$$\hat{\mu} = \bar{Y}_{\cdot\cdot}, \quad \hat{\beta}_i = \bar{Y}_{i\cdot} - \bar{Y}_{\cdot\cdot}, \quad \hat{\kappa}_j = \bar{Y}_{\cdot j} - \bar{Y}_{\cdot\cdot}, \quad \hat{\tau}_{(t)} = \bar{Y}_{t} - \bar{Y}_{\cdot\cdot}.$$

If we plug in these estimates into

$$\epsilon_{ij} = Y_{ij} - (\mu + \beta_i + \kappa_j + \tau_{(t)}),$$

algebra shows that we obtain (9.4). Thus, Experimental Error SS in the following ANOVA table does indeed measure what we attribute to inherent variation among experimental units.

We do not consider subsampling; our model above supposes exactly one observation is taken on the experimental unit at position i, j.

Three-Way ANOVA table – Latin Square

Source of variation	DF	SS Definition	MS	F
Columns	$r-1$	$r\sum_{j=1}^{r}(\bar{Y}_{\cdot j}-\bar{Y}_{\cdot\cdot})^2$	MS_C	$F_C=\frac{MS_C}{MS_E}$
Rows	$r-1$	$r\sum_{i=1}^{r}(\bar{Y}_{i\cdot}-\bar{Y}_{\cdot\cdot})^2$	MS_R	$F_R=\frac{MS_R}{MS_E}$
Treatments	$r-1$	$r\sum_{t=1}^{r}(\bar{Y}_{t}-\bar{Y}_{\cdot\cdot})^2$	MS_T	$F_T=\frac{MS_T}{MS_E}$
Experimental Error	$(r-1)(r-2)$	$\sum_{t=1}^{r}\sum_{i=1}^{r}\sum_{j=1}^{r}(Y_{ij}-\bar{Y}_{i\cdot}-\bar{Y}_{\cdot j}-\bar{Y}_{t}+2\bar{Y}_{\cdot\cdot})^2$	MS_E	
Total	r^2-1	$\sum_{i=1}^{r}\sum_{j=1}^{r}(Y_{ij}-\bar{Y}_{\cdot\cdot})^2$		

CALCULATIONS:

- Total SS
$$\sum_{i=1}^{r}\sum_{j=1}^{r}Y_{ij}^2-C$$
- Column SS
$$\frac{\sum_{j=1}^{r}\left(\sum_{i=1}^{r}Y_{ij}\right)^2}{r}-C$$
- Row SS
$$\frac{\sum_{i=1}^{r}\left(\sum_{j=1}^{r}Y_{ij}\right)^2}{r}-C$$
- Treatment SS
$$\frac{\sum_{t=1}^{r}\left(Y_{t\cdot}\right)^2}{r}-C$$
, where $Y_{t\cdot}$ is the sum of the observations on treatment t.
- Find the Experimental Error SS by subtraction:

$$\text{Experimental Error SS} = \text{Total SS} - \text{Column SS} - \text{Treatment SS} - \text{Row SS}.$$

The degrees of freedom for Experimental Error may be verified by considering the corresponding partition of those for Total SS.

Regardless of whether factors are fixed or random, the F ratios given in the table are valid for testing the usual sets of hypotheses.

STATISTICAL HYPOTHESES AND TESTS: The main focus is of course on the treatments. Under our model restrictions, the hypotheses are

$$H_{0,T}: \tau(1) = \cdots = \tau_{(r)} = 0 \text{ vs. } H_{1,T}: \text{ Not all } \tau_{(t)} \text{ equal.}$$

We reject $H_{0,T}$ at level of significance α if

$$F_T > F_{r-1,(r-1)(r-2),\alpha}.$$

Depending on the context, we may be interested in the factors corresponding to rows and columns. The hypotheses and tests are

$$H_{0,C}: \kappa_1 = \cdots = \kappa_r = 0 \text{ vs. } H_{1,C}: \text{ Not all } \kappa_j \text{ equal.}$$

$$H_{0,R}: \beta_1 = \cdots = \beta_r = 0 \text{ vs. } H_{1,C}: \text{ Not all } \beta_i \text{ equal.}$$

We reject the null hypotheses if

$$F_C > F_{r-1,(r-1)(r-2),\alpha} \text{ and } F_R > F_{r-1,(r-1)(r-2),\alpha},$$

respectively.

The statement of these hypotheses is appropriate for **fixed effects**; for random effects, we would restate them in terms of variance components, but the tests would be the same.

EXAMPLE: Here is an example of an "abstract" Latin square. (Box, Hunter, and Hunter, *Statistics for Experimenters*, p. 245) A study was conducted to investigate differences in reduction of oxides of nitrogen in automobile emissions obtained from 4 different gasoline additives. 4 cars and 4 drivers were used in the study. Even though the cars were identical models, it was thought that there might be slight systematic differences in their performance. Similarly, it was thought that differences in driving style might also affect performance. A Latin square design was used to eliminate the differences due to car and driver effects from the assessment of experimental error. It was not thought that there would be nonnegligible differences in the way the additives yield different reductions for different cars or drivers; that is, no **interaction** is suspected. Thus, a Latin square is appropriate, and allows additive differences to be evaluated without every driver driving every car with every additive.

The treatments are the 4 gasoline additives. We let Rows = Drivers, Columns = Cars. The data are coded measures of the reductions of oxides of nitrogen.

		Car				
		1	2	3	4	Row Total
Driver	I	A	B	D	C	92
		21	26	20	25	92
	II	D	C	A	B	
		23	26	20	27	96
	III	B	D	C	A	
		15	13	16	16	60
	IV	C	A	B	D	
		17	15	20	20	72
Column Total		76	80	76	88	320

Treatment totals (the Y_t.s): A: $21+20+16+15 = 72$, B: $26+27+15+20 = 88$, C: $25+26+16+17 = 84$, C: $20 + 23 + 13 + 20 = 76$.

Source of variation	DF	SS	MS	F
Columns (Car)	3	24	8.00	3.0
Rows (Driver)	3	216	72.00	27.0
Treatment	3	40	13.33	5.0
Experimental Error	6	16	2.67	
Total	15	296		

We use level of significance $\alpha = 0.05$. For all tests, the appropriate F value is $F_{3,6,0.05} = 4.76$. We thus **reject** $H_{0,T}$ – there is evidence in these data to suggest that there are differences in mean reduction of oxides of nitrogen among the additives.

Note that we also reject the null hypothesis for rows, which corresponds to drivers. There is evidence to suggest that different driving habits increase the variability in reduction of oxides. By using this design, it was possible to separate out the effect of drivers from experimental error while keeping the size of the experiment relatively small.

This example is implemented using SAS in section 9.9.

9.8 More on violation of assumptions

Throughout the course, we have made the point that the statistical methods we are studying are based on certain assumptions. Analysis of variance methods rely on the assumptions of **normality** and **constant variance**, with additive error.

We have already discussed the notion that this may be a reasonable assumption for many forms of continuous measurement data. We have also discussed that often a **logarithmic** transformation is useful for achieving approximate normality and constant variance for many types of continuous data as well.

However, there are many situations where our data are in the form of **counts** or **proportions**, which are not continuous across a large range of values. For example, we may count the numbers of birds of a certain species observed in a specific area or the proportions of insects surviving insecticide treatment.

In these situations, our interest still lies in assessing differences among treatments; however, the assumptions of **normality** and **constant variance** are certainly violated. It is well known for both count and proportion data that variance is **not constant** but rather depends on the size of the mean! Furthermore, histograms for such data are often highly **asymmetric**,

The methods we have discussed may still be used in these situations **provided that** a suitable transformation is used. That is, although the distribution of Y may not be normal with constant variance for all treatments and blocks, but it may be possible to **transform** the data and analyze them on the **transformed** scale, where these assumptions are more realistic.

Selection of an appropriate transformation of the data, h, say, is often based on the type of data. The values $h(Y_{ij})$ are treated as the data and analyzed in the usual way. Some common transformations are

- *Square root*: $h(Y) = \sqrt{Y}$. This is often appropriate for **count** data with small values.
- *Logarithmic*: $h(Y) = \log Y$. We have already discussed the use of this transformation for data where errors tend to have a multiplicative effect, such as growth data. Sometimes, the log transformation is useful for count data over a large range.
- *Arc sine*: $h(Y) = \arcsin(\sqrt{Y})$ or $\sin^{-1}\sqrt{Y}$. This transformation is appropriate when the data are in the form of percentages or proportions.

A deeper discussion of violation of assumptions is worth an entire course in itself. The basic message of this discussion is that it is often the case that the standard assumptions on which analysis of variance is predicated are violated. Data transformation is one way of handling violations of the assumptions. Your best bet is to consult a **statistician**!

9.9 Using SAS to perform analysis of variance for multi-way classification

Here, we give four examples of using the SAS procedure `PROC GLM`, "General Linear Model," to construct the analysis of variance for multi-way classification experiments. The considerations are the same as for the one way classification case. Thinking about the **linear additive model** for the particular problem is helpful in setting up the appropriate `MODEL` statement in `PROC GLM`. This is best illustrated in the context of our examples.

EXAMPLE 1: Randomized complete block design with one experimental unit per treatment/block combination and no subsampling – the Variety Data.

The linear additive model here is

$$Y_{ij} = \mu + \tau_i + \beta_j + \epsilon_{ij}.$$

The two classification (`CLASS`) variables are `BLOCK` and `VARIETY`, and the response is `YIELD`. The linear additive model implies the following model statement

```
MODEL YIELD = BLOCK VARIETY.
```

As in the one way case, the overall mean μ is understood.

PROGRAM:

```
/*****************************************************
*                                                    *
*      ST 511                  EXAMPLE 9.1           *
*                                                    *
*   USING PROC GLM TO PERFORM TWO-WAY ANOVA WITH     *
*   ONE EXPERIMENTAL UNIT PER TREATMENT/BLOCK        *
*   COMBINATION                                      *
*                                                    *
*****************************************************/

OPTIONS LS=80 PS=59 NODATE;

DATA BUSHELS;
   INPUT BLOCK $ VARIETY $ YIELD @@;
   CARDS;
1 A 10 1 B 9 1 C 11 1 D 15 1 E 10 1 F 12 1 G 11
2 A 11 2 B 10 2 C 12 2 D 12 2 E 10 2 F 11 2 G 12
3 A 12 3 B 13 3 C 10 3 D 14 3 E 15 3 F 13 3 G 13
4 A 14 4 B 15 4 C 13 4 D 17 4 E 14 4 F 16 4 G 15
5 A 13 5 B 14 5 C 16 5 D 19 5 E 17 5 F 15 5 G 18
;
PROC PRINT;
   TITLE 'AGRICULTURAL EXPERIMENT DATA (BUSHELS/ACRE)'; RUN;
*;
PROC GLM;
   CLASS BLOCK VARIETY;
   MODEL YIELD = BLOCK VARIETY;
   TITLE2 'TWO-WAY ANOVA -- RANDOMIZED COMPLETE BLOCK DESIGN'; RUN;
```

OUTPUT:

%%%

AGRICULTURAL EXPERIMENT DATA (BUSHELS/ACRE) 1

OBS	BLOCK	VARIETY	YIELD
1	1	A	10
2	1	B	9
3	1	C	11
4	1	D	15
5	1	E	10
6	1	F	12
7	1	G	11
8	2	A	11
9	2	B	10
10	2	C	12
11	2	D	12
12	2	E	10
13	2	F	11
14	2	G	12
15	3	A	12
16	3	B	13
17	3	C	10
18	3	D	14
19	3	E	15
20	3	F	13
21	3	G	13
22	4	A	14
23	4	B	15
24	4	C	13
25	4	D	17
26	4	E	14
27	4	F	16
28	4	G	15
29	5	A	13

30	5	B	14
31	5	C	16
32	5	D	19
33	5	E	17
34	5	F	15
35	5	G	18

%%%

AGRICULTURAL EXPERIMENT DATA (BUSHELS/ACRE) 2

TWO-WAY ANOVA -- RANDOMIZED COMPLETE BLOCK DESIGN

General Linear Models Procedure

Class Level Information

Class	Levels	Values
BLOCK	5	1 2 3 4 5
VARIETY	7	A B C D E F G

Number of observations in data set = 35

%%%

AGRICULTURAL EXPERIMENT DATA (BUSHELS/ACRE) 3

TWO-WAY ANOVA -- RANDOMIZED COMPLETE BLOCK DESIGN

General Linear Models Procedure

Dependent Variable: YIELD

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	10	175.77142857	17.57714286	10.59	0.0001
Error	24	39.82857143	1.65952381		
Corrected Total	34	215.60000000			

R-Square	C.V.	Root MSE	YIELD Mean
0.815266	9.759281	1.2882251	13.200000

Source	DF	Type I SS	Mean Square	F Value	Pr > F
BLOCK	4	134.17142857	33.54285714	20.21	0.0001
VARIETY	6	41.60000000	6.93333333	4.18	0.0052

Source	DF	Type III SS	Mean Square	F Value	Pr > F
BLOCK	4	134.17142857	33.54285714	20.21	0.0001
VARIETY	6	41.60000000	6.93333333	4.18	0.0052

EXAMPLE 2: Randomized complete block design with more than one experimental unit per treatment/block combination – the Hormone-Sex Data.

The linear additive model here is

$$Y_{ij} = \mu + \tau_i + \beta_j + (\tau\beta)_{ij} + \epsilon_{ij}.$$

The two classification (`CLASS`) variables are `SEX` (the blocks) and `HORMONE` (the treatments), and the response is `PLASMA`. Here, we have the **interaction** term $(\tau\beta)_{ij}$. To specify an interaction term in the `MODEL` statement of `PROC GLM`, the syntax is, in this example, `SEX*HORMONE`; the asterisk "*" tells SAS to construct the interaction term. The linear additive model thus implies the following model statement

```
MODEL PLASMA = SEX HORMONE SEX*HORMONE.
```

As in the one way case, the overall mean μ is understood.

The interpretation of the output given here is relevant to the case where **both** treatments and blocks are regarded as **fixed**, which is clearly the case for the hormone data.

We use the SAS procedure PLOT to produce a crude "interaction plot" of the sample means for each hormone/sex combination. This displays graphically the means given earlier, and shows visually that there does not appear to be an interaction.

PROGRAM:

```
/************************************************
*                                               *
*          ST 511            EXAMPLE 9.2        *
*                                               *
*    USING PROC GLM TO PERFORM TWO-WAY ANOVA     *
*    WITH MORE THAN ONE EXPERIMENTAL UNIT PER    *
*    TREATMENT/BLOCK COMBINATION AND INTERACTION *
*                                               *
************************************************/

OPTIONS LS=80 PS=59 NODATE;

DATA HORMONES;
  INPUT SEX $ HORMONE $ PLASMA @@;
  CARDS;
M Y 32.0 M Y 23.8 M Y 28.8 M Y 25.0 M Y 29.3
M N 14.5 M N 11.0 M N 10.8 M N 14.3 M N 10.0
F Y 39.1 F Y 26.2 F Y 21.3 F Y 35.8 F Y 40.2
F N 16.5 F N 18.4 F N 12.7 F N 14.0 F N 12.8
*;
PROC PRINT;
   TITLE ' HORMONE TREATMENT EXAMPLE'; RUN;
*;
PROC GLM;
   CLASS SEX HORMONE;
```

```
   MODEL PLASMA = SEX HORMONE SEX*HORMONE;
   TITLE2 'TWO-WAY ANOVA WITH INTERACTION'; RUN;

/************************************************
*                                               *
*   USE PROC MEANS TO OBTAIN THE SAMPLE MEANS   *
*   FOR EACH SEX/HORMONE COMBINATION AND MAKE   *
*   AN INTERACTION PLOT USING PROC PLOT.  WE    *
*   USE THE TREATMENT SYMBOL FOR PLOTTING       *
*                                               *
************************************************/

PROC SORT; BY SEX HORMONE; RUN;

PROC MEANS MEAN NOPRINT;
  BY SEX HORMONE;
  VAR PLASMA;
  OUTPUT OUT=HMEANS MEAN=SHMEAN; RUN;

PROC PLOT; PLOT SHMEAN*SEX=HORMONE; RUN;
```

OUTPUT:

%%%

HORMONE TREATMENT EXAMPLE 1

OBS	SEX	HORMONE	PLASMA
1	M	Y	32.0
2	M	Y	23.8
3	M	Y	28.8
4	M	Y	25.0
5	M	Y	29.3
6	M	N	14.5
7	M	N	11.0
8	M	N	10.8
9	M	N	14.3
10	M	N	10.0
11	F	Y	39.1
12	F	Y	26.2
13	F	Y	21.3
14	F	Y	35.8
15	F	Y	40.2
16	F	N	16.5
17	F	N	18.4
18	F	N	12.7
19	F	N	14.0
20	F	N	12.8

%%%

HORMONE TREATMENT EXAMPLE 2

TWO-WAY ANOVA WITH INTERACTION

General Linear Models Procedure

Class Level Information

Class	Levels	Values
SEX	2	F M
HORMONE	2	N Y

Number of observations in data set = 20

%%%

HORMONE TREATMENT EXAMPLE 3

TWO-WAY ANOVA WITH INTERACTION

General Linear Models Procedure

Dependent Variable: PLASMA

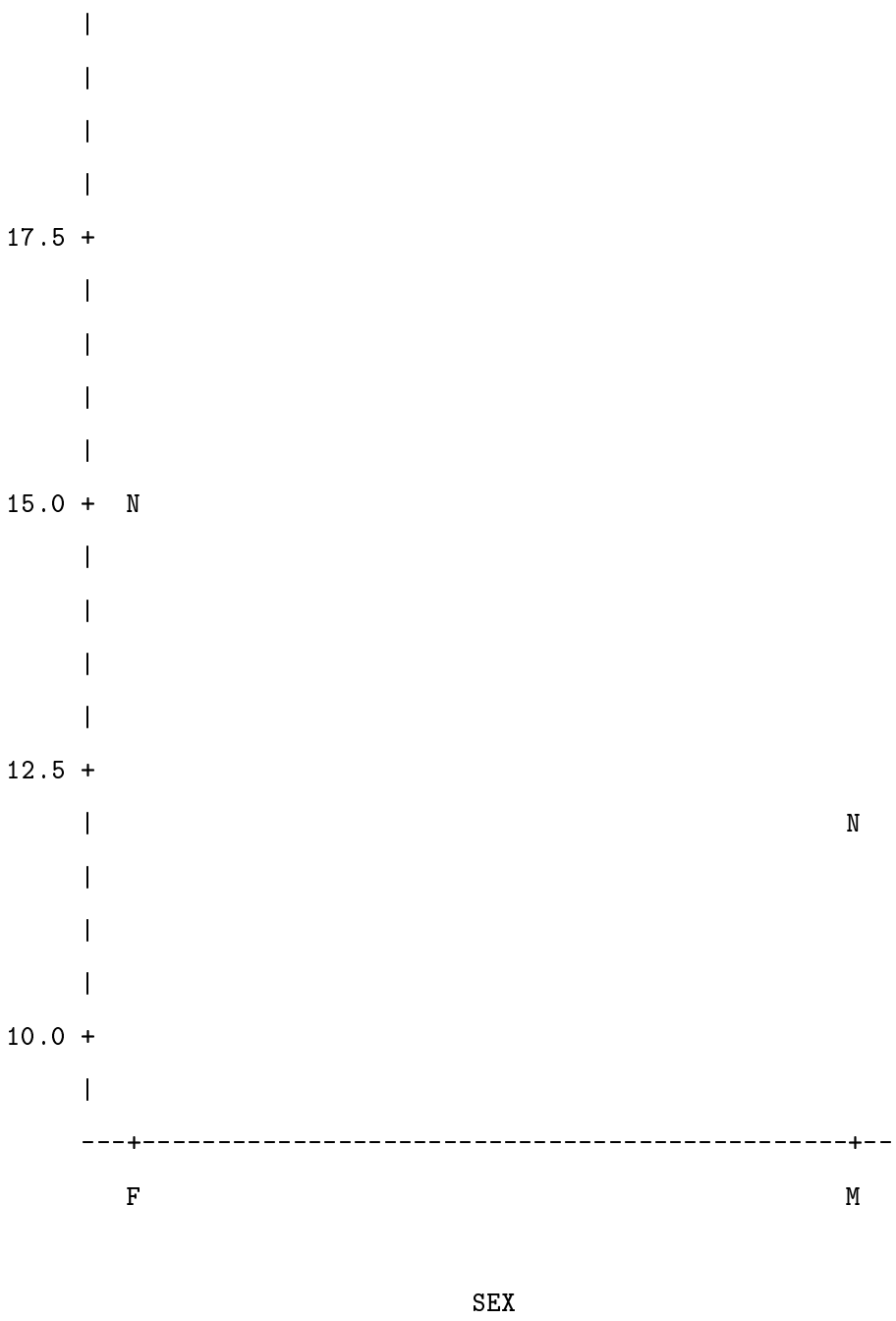
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	1461.3255000	487.1085000	21.27	0.0001
Error	16	366.3720000	22.8982500		
Corrected Total	19	1827.6975000			

R-Square	C.V.	Root MSE	PLASMA Mean
0.799545	21.92537	4.7852116	21.825000

Source	DF	Type I SS	Mean Square	F Value	Pr > F
SEX	1	70.3125000	70.3125000	3.07	0.0989
HORMONE	1	1386.1125000	1386.1125000	60.53	0.0001
SEX*HORMONE	1	4.9005000	4.9005000	0.21	0.6499

Source	DF	Type III SS	Mean Square	F Value	Pr > F
SEX	1	70.3125000	70.3125000	3.07	0.0989
HORMONE	1	1386.1125000	1386.1125000	60.53	0.0001
SEX*HORMONE	1	4.9005000	4.9005000	0.21	0.6499

```
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%


                            HORMONE TREATMENT EXAMPLE                          4

                         TWO-WAY ANOVA WITH INTERACTION


               Plot of SHMEAN*SEX.  Symbol is value of HORMONE.


         SHMEAN |
           32.5 +  Y
                |
                |
                |
                |
           30.0 +
                |
                |
                |
                |                                                  Y
           27.5 +
                |
                |
                |
                |
           25.0 +
                |
                |
                |
                |
           22.5 +
                |
                |
                |
                |
           20.0 +
```



EXAMPLE 3: Randomized complete block design with one experimental unit per treatment/block combination and subsampling – the Nursery Data.

The linear additive model here is

$$Y_{ijk} = \mu + \tau_i + \beta_j + \epsilon_{ij} + \delta_{ijk}.$$

Here, the classification (`CLASS`) variables are `HOUSE` (the blocks) and `FERT` (the treatments), and the

response is HEIGHT. Although there is no interaction term in this model, because the calculations of the SSs are algebraically the same as for the model with interaction and more than one experimental unit per treatment/block, the syntax we use to get SAS to compute the analysis of variance table is the same – in the MODEL statement, we include a term "HOUSE*FERT" in this case. We thus use

```
MODEL HEIGHT = HOUSE FERT HOUSE*FERT.
```

As in the one way case, the overall mean μ is understood. The **difference** is in how we construct the F ratios. Recall that SAS automatically uses the MS for the "smallest" unit of measurement for the denominator of **all** F ratios. This was okay for the previous example, but not here. Here, the **experimental error** SS is being calculated by the same algebraic computation used to obtain the HOUSE*FERT line of the ANOVA table; thus, the MS corresponding to this line is the correct denominator for the F statistics for testing treatments and blocks. We thus use a TEST statement to request that these test statistics be computed. The F ratios produced by default, appearing in the upper tables in the output, are **inappropriate** and should be ignored.

PROGRAM:

```
/************************************************
*                                               *
*          ST 511             EXAMPLE 9.3       *
*                                               *
*    USING PROC GLM TO PERFORM TWO-WAY ANOVA     *
*    WITH ONE EXPERIMENTAL UNIT PER              *
*    TREATMENT/BLOCK COMBINATION AND             *
*    SUBSAMPLING                                 *
*                                               *
************************************************/

OPTIONS LS=80 PS=59 NODATE;

DATA NURSERY;
  INPUT FERT HOUSE $ HEIGHT @@;
  CARDS;
  1 I 47 1 I 43 1 II 46 1 II 40
  2 I 62 2 I 68 2 II 67 2 II 71
```

```
  3 I 41 3 I 39 3 II 42 3 II 46
*;
PROC PRINT;
   TITLE ' NURSERY EXAMPLE'; RUN;
*;
PROC GLM;
   CLASS FERT HOUSE;
   MODEL HEIGHT = HOUSE FERT HOUSE*FERT;
   TEST H=FERT HOUSE E=FERT*HOUSE;
   TITLE2 'TWO-WAY ANOVA WITH SUBSAMPLING'; RUN;
```

OUTPUT:

```
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%

                                NURSERY EXAMPLE                          1

                          OBS    FERT    HOUSE    HEIGHT

                            1       1       I         47
                            2       1       I         43
                            3       1       II        46
                            4       1       II        40
                            5       2       I         62
                            6       2       I         68
                            7       2       II        67
                            8       2       II        71
                            9       3       I         41
                           10       3       I         39
                           11       3       II        42
                           12       3       II        46

%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%

                                NURSERY EXAMPLE                          2
```

TWO-WAY ANOVA WITH SUBSAMPLING

General Linear Models Procedure

Class Level Information

Class	Levels	Values
FERT	3	1 2 3
HOUSE	2	I II

Number of observations in data set = 12

%%%

NURSERY EXAMPLE 3

TWO-WAY ANOVA WITH SUBSAMPLING

General Linear Models Procedure

Dependent Variable: HEIGHT

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	5	1580.0000000	316.0000000	30.58	0.0003
Error	6	62.0000000	10.3333333		
Corrected Total	11	1642.0000000			

R-Square	C.V.	Root MSE	HEIGHT Mean

0.962241	6.303040	3.2145503		51.000000

Source	DF	Type I SS	Mean Square	F Value	Pr > F
HOUSE	1	12.0000000	12.0000000	1.16	0.3226
FERT	2	1544.0000000	772.0000000	74.71	0.0001
FERT*HOUSE	2	24.0000000	12.0000000	1.16	0.3747

Source	DF	Type III SS	Mean Square	F Value	Pr > F
HOUSE	1	12.0000000	12.0000000	1.16	0.3226
FERT	2	1544.0000000	772.0000000	74.71	0.0001
FERT*HOUSE	2	24.0000000	12.0000000	1.16	0.3747

Tests of Hypotheses using the Type III MS for FERT*HOUSE as an error term

Source	DF	Type III SS	Mean Square	F Value	Pr > F
FERT	2	1544.0000000	772.0000000	64.33	0.0153
HOUSE	1	12.0000000	12.0000000	1.00	0.4226

EXAMPLE 4: Latin square – the Gasoline Additives Data.

The linear additive model here is

$$y_{ij(t)} = \mu + \tau_i + \kappa_j + \tau_{(t)} + \epsilon_{ij}.$$

Here, we have three classification (`CLASS`) variables `DRIVER`, `CAR`, and `ADD` (for additives). The response is `REDUCE` (for reduction in oxides of nitrogen). To get SAS to construct the analysis of variance for this model is easy. The SSs corresponding to each term are algebraically those that SAS computes when it sees a `CLASS` variable standing alone in a `MODEL` statement. It thus computes the 3 required SSs for these factors from the `MODEL` statement

```
MODEL REDUCE = DRIVER CAR ADD;
```

As in the one way case, the overall mean μ is understood. The error SS is automatically computed as what is "left over," that is, corresponding to ϵ_{ij} (the "smallest" unit of measurement) in the linear model.

PROGRAM:

```
/************************************************
*                                               *
*          ST 511             EXAMPLE 9.4       *
*                                               *
*    USING PROC GLM TO PERFORM THREE-WAY ANOVA   *
*    FOR A BASIC LATIN SQUARE                    *
*                                               *
************************************************/

OPTIONS LS=80 PS=59 NODATE;

DATA OXIDES;
  INPUT CAR DRIVER $ ADD $ REDUCE @@;
  CARDS;
1 I A 21 2 I B 26 3 I D 20 4 I C 25
1 II D 23 2 II C 26 3 II A 20 4 II B 27
1 III B 15 2 III D 13 3 III C 16 4 III A 16
1 IV C 17 2 IV A 15 3 IV B 20 4 IV D 20
```

```
;

PROC PRINT;
   TITLE ' OXIDE REDUCTION EXAMPLE'; RUN;

PROC GLM;
   CLASS DRIVER CAR ADD;
   MODEL REDUCE = DRIVER CAR ADD;
   TITLE2 'ANOVA FOR BASIC LATIN SQUARE'; RUN;
```

OUTPUT:

```
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%

                          OXIDE REDUCTION EXAMPLE                          1

                    OBS    CAR    DRIVER    ADD    REDUCE

                      1     1      I         A       21
                      2     2      I         B       26
                      3     3      I         D       20
                      4     4      I         C       25
                      5     1      II        D       23
                      6     2      II        C       26
                      7     3      II        A       20
                      8     4      II        B       27
                      9     1      III       B       15
                     10     2      III       D       13
                     11     3      III       C       16
                     12     4      III       A       16
                     13     1      IV        C       17
                     14     2      IV        A       15
                     15     3      IV        B       20
                     16     4      IV        D       20
```

%%%

OXIDE REDUCTION EXAMPLE 2

ANOVA FOR BASIC LATIN SQUARE

General Linear Models Procedure

Class Level Information

Class	Levels	Values
DRIVER	4	I II III IV
CAR	4	1 2 3 4
ADD	4	A B C D

Number of observations in data set = 16

%%%

OXIDE REDUCTION EXAMPLE 3

ANOVA FOR BASIC LATIN SQUARE

General Linear Models Procedure

Dependent Variable: REDUCE

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	9	280.00000000	31.11111111	11.67	0.0037

Error	6	16.00000000	2.66666667	
Corrected Total	15	296.00000000		

R-Square	C.V.	Root MSE	REDUCE Mean
0.945946	8.164966	1.6329932	20.000000

Source	DF	Type I SS	Mean Square	F Value	Pr > F
DRIVER	3	216.00000000	72.00000000	27.00	0.0007
CAR	3	24.00000000	8.00000000	3.00	0.1170
ADD	3	40.00000000	13.33333333	5.00	0.0452

Source	DF	Type III SS	Mean Square	F Value	Pr > F
DRIVER	3	216.00000000	72.00000000	27.00	0.0007
CAR	3	24.00000000	8.00000000	3.00	0.1170
ADD	3	40.00000000	13.33333333	5.00	0.0452

10 Simple Linear Regression and Correlation

Complementary Reading: STD, Chapters 10, 11

10.1 Introduction

So far, we have focused our attention on problems where the main issue is identifying differences among treatment means. In this setting, we based our inferences upon observations on a random variable Y under the various experimental conditions (treatments, blocks, etc).

Another problem that arises in the biological and physical sciences, economics, industrial applications, and biomedical settings is that of investigating the **relationship** between two (or more) variables. Depending on the nature of the variables and the observations on them (more on this in a moment), the methods of **regression analysis** or **correlation analysis** are appropriate.

In reality, our development of the methods for identifying differences among treatment means, those of **analysis of variance**, are in fact very similar to **regression analysis** methods, as will become apparent in our discussion. Both sets of methods are predicated on representing the data by a **linear, additive model**, where the model includes components representing both **systematic** and **random** sources of variation. The common features will become evident over the course of our discussion.

In this chapter, we will restrict our study to the **simplest** case in which we have **two** variables for which the relationship between them is reasonably assumed to be a **straight line**. (In the case of **correlation analysis**, we will have to be a bit more precise about what we mean by this, as you will see.) Note, however, that the areas of regression analysis and correlation analysis are much broader than indicated by our introduction here.

TERMINOLOGY: It is important to clarify the usage of the term **linear** in statistics. **Linear**, as we have been using it

throughout the course so far, refers to how a component of an equation describing a relationship enters that relationship. For example, in the one way classification model, recall that we represented an observation as

$$Y_{ij} = \mu + \tau_i + \epsilon_{ij}.$$

This equation is said to be **linear** because the components μ, τ_i, and ϵ_{ij} enter directly. Contrast this with an example of an equation **nonlinear** in μ and τ_i:

$$Y_{ij} = \exp(\mu + \tau_i) + \epsilon_{ij}.$$

This equation still has an **additive** error, but the **parameters** of interest μ and τ_i enter in a **nonlinear** fashion, through the exponential function. The term **linear** is thus used in statistical applications in this broad way, to indicate that parameters enter in a straightforward rather than complicated way.

The term **linear regression** has a similar interpretation, as we will see. The term **simple linear regression** refers to the particular case where the relationship is a **straight line**. The use of the term **linear** thus refers to how parameters come into the model for the data in a **general** sense, and does not necessarily mean that the relationship need be a **straight line**, except when prefaced by the term **simple**. We'll see examples shortly.

SCENARIO: We are interested in the relationship between two variables, which we will call X and Y. We **observe** pairs of X and Y values on each of a **sample** of **experimental units**, and we wish to use them to say something about the relationship. How we view the relationship is dictated by the situation:

"EXPERIMENTAL" DATA: Here, observations on X and Y are **planned** as the result of an experiment, lab procedure, etc. For example

- X = dose of a drug, Y = response such as change in blood pressure for a human subject
- X = concentration of toxic substance, Y = number of mutant offspring observed for a pregnant rat

In these examples, we are **in control** of the values of X (e.g. we choose the doses or concentrations) and we observe the resulting Y.

"OBSERVATIONAL" DATA: Here, we **observe** both X and Y values, neither of which is under our control. For example,

- X = weight, Y = height of a human subject
- X = average heights of plants in a plot, Y = yield

In the experimental data situations, there is a distinction between what we call X and what we call Y, because the former is under dictated by the investigator. It is standard to use these symbols in this way. In the observational data examples, there is not necessarily such a distinction. We will have more to say about this momentarily. In any event, if we use the symbols in this way, then what we call Y is always understood to be something we observe, while X may or may not be.

RELATIONSHIPS BETWEEN TWO VARIABLES: In some situations, scientific theory may suggest that 2 variables are **functionally related**, e.g.

$$Y = g(X).$$

Here, g is some function. The form of g may follow from some particular theory. Even if there is no suitable theory, we may still suspect some kind of **systematic** relationship between X and Y, and may be able to identify a function g that provides a reasonable empirical description.

OBJECTIVE: Based on a sample of observations on X and Y, formally describe and assess the relationship between them.

PRACTICAL PROBLEM: In most situations, the values we **observe** for Y (and sometimes X, certainly in the case of observational data) are **not exact**. In particular, due to biological variation among experimental units and the sampling of them, imprecision and/or inaccuracy of measuring devices, and so on, we may only observe values of Y (and also possibly X) with some **error**. Thus, based on a sample of (X, Y) pairs, our ability to see the relationship **exactly** is obscured by this error.

RANDOM VS. FIXED X: Given these issues, it is natural to think of Y (and perhaps X) as **random variables**. How we do this is dictated by the situation, as above:

- *Experimental data*: Here, X (dose, concentration) is **fixed** at predetermined levels by the experimenter. Thus, X is best viewed as a **fixed quantity** (like **treatment** in our previous situations). Y, on the other hand, which is subject to biological and sampling variation and error in measurement, is a **random variable**. Clearly, the values for Y we do get to see will be related to the fixed values of X.

- *Observational data*: Consider Y = height, X = weight. In this case, neither weight nor height is a fixed quantity; both are subject to **variation**. Thus, **both** X and Y must be viewed as **random variables**. Clearly, the values taken on by these two random variables are related or associated somehow.

STATISTICAL MODELS: These considerations dictate how we think of a formal **statistical model** for the situation:

- *Experimental data*: Here, a natural way to think about Y is by representing it as

$$Y = g(X) + \epsilon. \tag{10.1}$$

  Here, then, we believe the function g describes the relationship, but values of Y we observe are not exactly equal to $g(X)$ because of the errors mentioned above. The **additive** "error" ϵ characterizes this, just as in our previous models. In this situation, the following terminology is often used:

$$\begin{aligned} Y &= \textbf{response or dependent variable} \\ X &= \textbf{concomitant or independent variable, covariate} \end{aligned}$$

  These terms seem natural; in this case, we wish to characterize how the response changes with, or **depends** on, the value of X, which **we** control **independently**. The distinction between the interpretations of X and Y and the suitability of this terminology is clear.

- *Observational data*: In this situation, there is really not much distinction between X and Y, as **both** are seen with error. Here, the terms **independent** and **dependent** variable may be misleading. For example, if we have observed pairs of X = weight, Y = height, it is not necessarily if we should be interested in a relationship

$$Y = g(X) \text{ or } X = h(Y),$$

say. Even if we have in our mind that we want to think of the relationship in a particular way, say $Y = g(X)$, it is clear that the above model (10.1) is not really appropriate, as it does not take into account "error" affecting X.

We will begin our discussion of these problems in the next section by considering **regression** models that are appropriate when we have **experimental data**; that is, when it is legitimate to think of a model like (10.1). We will then discuss a **probability model** that is better suited to describing observational data. Methods for each type of model, **regression analysis** and **correlation analysis**, will then be the focus.

10.2 Simple linear regression model

STRAIGHT LINE MODEL: Consider the particular situation of experimental data, where it is legitimate to regard X as fixed. It is often reasonable to suppose that the relationship between Y and X, which we have called in general g, is in fact a **straight line**. We may write this as

$$Y = \beta_0 + \beta_1 X + \epsilon \tag{10.2}$$

for some values β_0 and β_1. Here, then, g is a straight line with

Intercept	β_0	The value taken on at $X = 0$
Slope	β_1	Expresses the **rate of change** in Y, i.e. β_1 = change in Y brought about by a change of one unit in X.

ISSUE: The problem is that we do not know β_0 or β_1. To get information on their values, the typical experimental setup is to choose values X_i, $i = 1, \ldots, n$, and observe the resulting responses $Y_1, \ldots, Y_n$, so that the **data** consist of the pairs (X_i, Y_i), $i = 1, \ldots, n$. The data are then used to **estimate** β_0 and β_1, i.e. to **fit** the model to the data, in order to

- quantify the relationship between Y and X
- use the relationship to **predict** a new response Y_0 we might observe at a given value X_0 (perhaps one not included in the experiment)
- use the relationship to **calibrate** – given a new Y_0 value we might see, for which the corresponding value X_0 is **unknown**, estimate the value X_0.

The model (10.2) is referred to as a **simple linear regression** model. The term **regression** refers to the postulated relationship.

- The **regression** relationship in this case is the straight line $\beta_0 + \beta_1 X$
- The **parameters** β_0 and β_1 characterizing this relationship are called **regression coefficients** or **regression parameters**.

If we think of our **data**, (X_i, Y_i), we may thus think of a model for Y_i as follows:

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i = \mu_i + \epsilon_i, \quad \mu_i = \beta_0 + \beta_1 X_i. \tag{10.3}$$

This looks very much like our linear additive model in the analysis of variance, but with only one observation on each "treatment." That is,

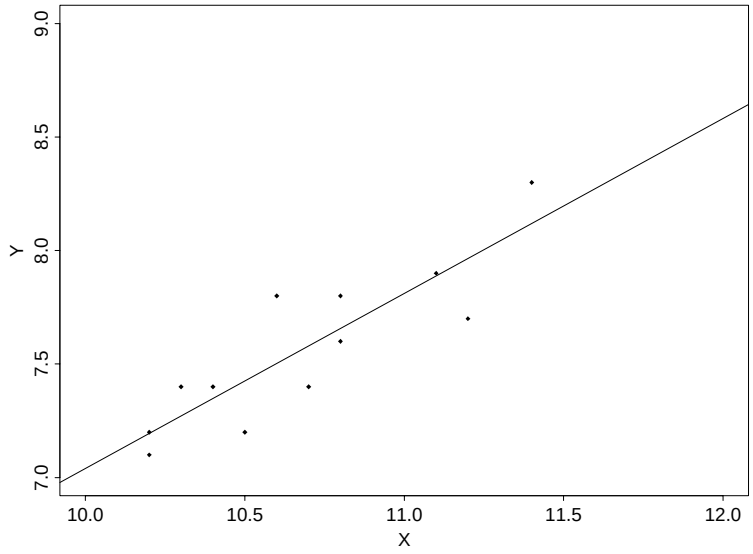
$$\mu_i = \beta_0 + \beta_1 X_i$$

is the **mean** of observations we would see at the particular setting X_i.

In the one way classification situation, we would certainly not be able to estimate each mean μ_i with a single observation; however, here, because we also have the variable X, and are willing to represent μ_i as a particular function of X (here, the straight line), we will be able to take advantage of this functional relation to estimate μ_i. Hence, we only need the single subscript i.

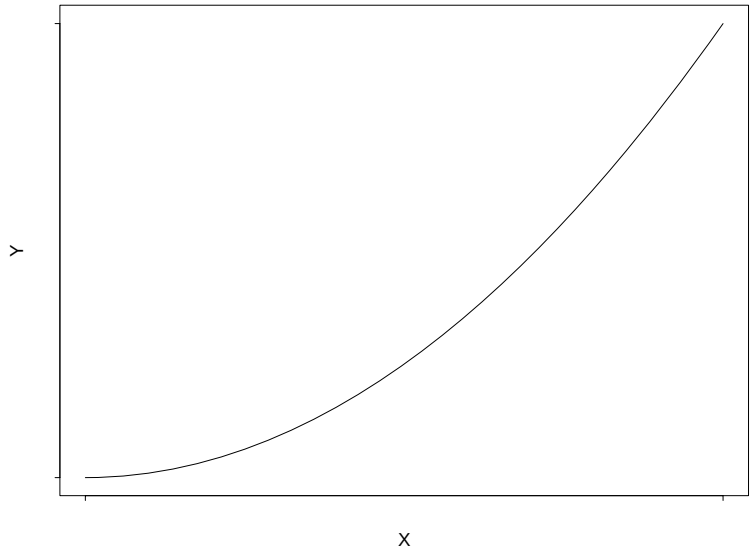
The "errors" ϵ_i characterize all the sources of inherent variation that cause Y_i to not exactly equal its mean, $\mu_i = \beta_0 + \beta_1 X_i$. We may think of this as **experimental error** – all unexplained inherent variation due to the experimental unit.

The following picture summarizes the situation. At any X_i value, the **mean** of responses Y_i we might observe is $\mu_i = \beta_0 + \beta_1 X_i$. The means are on a **straight line** over all X values. Because of "error", at any X_i value, the Y_i **vary** about the mean μ_i, so do not lie on the line, but are scattered about it.



OBJECTIVE: For the simple linear regression model, **fit** the line to the data to serve as our "best" characterization of the relationship based on the available data. More precisely, **estimate** the parameters β_0 and β_1 that characterize the **mean** at any X value.

REMARK: We may now comment more precisely on the meaning of the term **linear** regression. In practice, the regression **need not** be a **straight line**, nor need there be a **single** independent variable X. For example, the underlying relationship between Y and X (that is, the **mean**), may be more adequately represented by a **curve** like



For such a situation, a better model might be a **quadratic** function of X, e.g.

$$\beta_0 + \beta_1 X + \beta_2 X^2 \tag{10.4}$$

Or, if Y is some measure of **growth** of a plant and X is time, we would eventually expect the relationship to **level off** when X gets large, as plants can not continue to get large without bound! A popular model for this is the **logistic growth function**

$$\frac{\beta_1}{1 + \beta_2 e^{\beta_3 X}} \tag{10.5}$$

The curve for this looks much like that above, but would begin to "flatten out" if we extended the picture for large values of X.

In the quadratic model (10.4), note that, although the function is no longer a straight line, it is still a straightforward function of the **regression parameters** characterizing the curve, $\beta_0, \beta_1, \beta_2$. In particular, β_0, β_1, and β_2 enter in a **linear** fashion in the sense we discussed in the previous section.

Contrast this with the logistic growth model (10.5). Here, the **regression parameters** characterizing the curve **do not** enter the model in a straightforward fashion. In particular, the parameters β_2 and β_3

appear in a quite complicated way, in the denominator. This function is thus **not linear** as a function of β_1, β_2, and β_3; rather, it is better described as **nonlinear**.

It turns out that **linear** functions are much easier to work with then **nonlinear** functions. Although we will work strictly with the **simple linear regression** model, be aware that the methods we discuss extend easily to more complex **linear** models like (10.4). They do not extend as easily to **nonlinear** models.

10.3 The bivariate normal distribution

Consider the situation of **observational** data. Because **both** X and Y are subject to error, **both** are random variables that are somehow related.

RECALL: A **probability distribution** provides a formal description of the population of possible values that might be taken on by a random variable. So far, we have only discussed this notion in the context of a **single** random variable.

It is possible to **extend** the idea of a probability distribution to **two** random variables. Such a distribution is called, for obvious reasons, a **bivariate** probability distribution. This distribution describes not only the populations of possible values that might be taken on by the two random variables, but also how those values are taken on **together**.

Consider our X and Y. Formally, we would think of a probability distribution function

$$f(x, y)$$

that describes the populations of X and Y and how they are related; that is, how X and Y values vary **together**.

BIVARIATE NORMAL DISTRIBUTION: Recall that the **normal** distribution is often a reasonable description of a population of continuous measurements. When both X and Y are continuous measurements, a reasonable assumption is that they are **both** normally distributed. However, we also expect them to **vary together**.

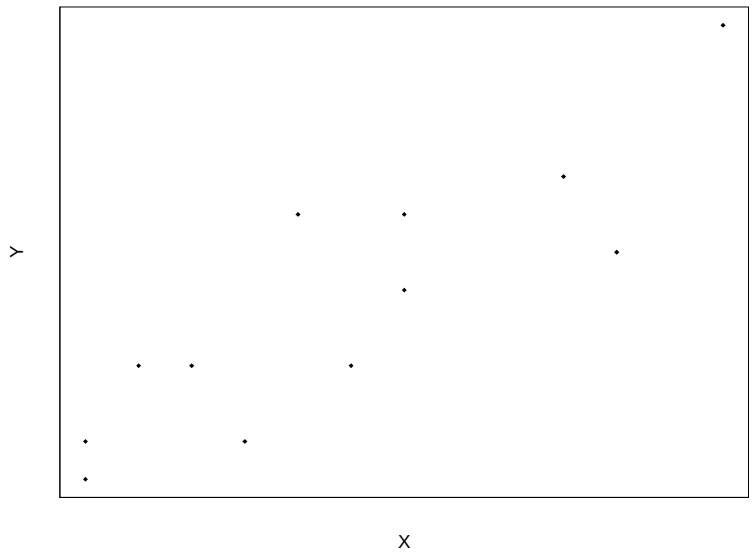
The **bivariate normal distribution** is a probability distribution with a probability density function $f(x, y)$ for both X and Y such that

- The 2 random variables X and Y each have normal distributions with means μ_X and μ_Y, and variances σ_X^2 and σ_Y^2, respectively.
- The relationship between X and Y is characterized by a quantity ρ_{XY} such that $-1 \leq \rho_{XY} \leq 1$.
- ρ_{XY} is referred to as the **correlation coefficient** between the two random variables X and Y and measures the **linear association** between values taken on by X and values taken on by Y.

  $\rho_{XY} = 1$ all possible values of X and Y lie on a straight line with **positive slope**
  $\rho_{XY} = -1$ all possible values of X and Y lie on a straight line with **negative slope**
  $\rho_{XY} = 0$ there is no relationship between X and Y.

The value $\rho_{XY} = 0$ basically says that X and Y values do not **vary** together – for any X value we might observe, any Y value is possible. If $\rho_{XY} \neq 0$, it implies that there is some restriction about how the X and Y values happen together.

The values $\rho_{XY} = -1$ and 1 represent **extremes** of how X and Y values happen together, what may be considered **perfect** association – any pair (X_i, Y_i) we might see **must** lie exactly on a line. This is unlikely to be observed in practice. What we are more likely to observe are pairs that exhibit behavior like this:



For such a situation, there is clearly a tendency for the values to vary together in a **positive** way, but not to the extreme of lying all on a straight line. A positive value of ρ_{XY} **between** 0 and 1 would characterize this situation. Similarly, a negative value intermediate between -1 and 0 would indicate a situation like that above, but with tendency in the negative direction.

We will discuss interpretation of correlation more later in this chapter. For now, note that, just as we thought of simple linear regression model for experimental data as an appropriate representation of the data, here, we think of the bivariate normal distribution. This model is an appropriate framework, as it keeps X and Y on equal footing.

OBJECTIVE: Given our observed data pairs (X_i, Y_i), we would like to **quantify** the degree of association. To do this, we **estimate** ρ_{XY}.

10.4 Comparison of regression and correlation models

We have identified two appropriate **statistical models** for thinking about the problem of assessing association between two variables X and Y. These may be thought of as

- *Fixed* X: Postulate a model for the **mean** of the **random variable** Y as a function of the fixed quantity X (in particular, we focused on a straight line in X). Estimate the parameters in the model to characterize the relationship.
- *Random* X: Characterize the (linear) relationship between X and Y by the **correlation** between them (in a bivariate normal probability model) and estimate the **correlation** parameter.

It turns out that the arithmetic operations for regression analysis under the first scenario and correlation analysis under the second are **the same**! That is, to **fit** the regression model by estimating the intercept and slope parameters and to **estimate** the correlation coefficient, we use the **same** operations on our data!

The important issue is in the **interpretation** of the results.

SUBTLETY: In settings where X is best regarded as a random variable, many investigators still want to fit regression models treating X as fixed. This is because, although correlation describes the "degree of association" between X and Y, it doesn't characterize the relationship in a way suitable for some purposes.

For example, an investigator may desire to **predict** the yield of a plot based on **observing** the average height of plants in the plot. The correlation coefficient does not allow this. He thus would rather fit a **regression model**, even though X is **random**.

Is this legitimate? If we are careful about the interpretation, it may be.

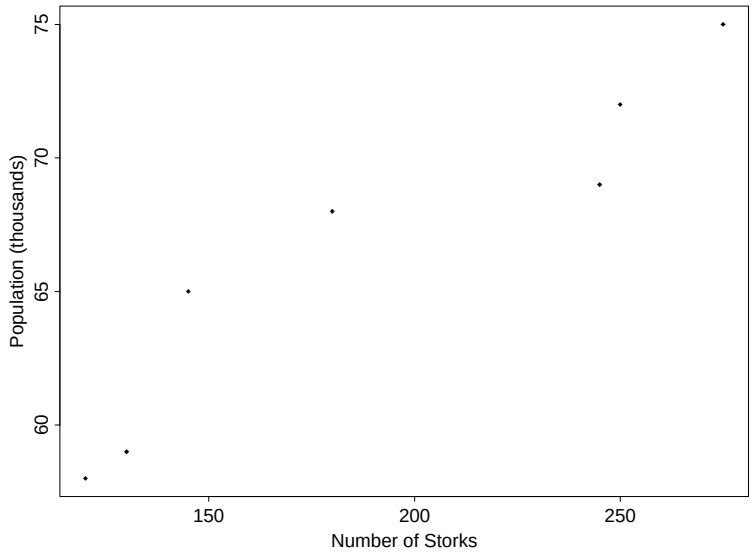
If X and Y are really both observed random variables, and we fit a regression to characterize the relationship, technically, any subsequent analyses based on this are regarded as **conditional** on the values of X involved. This means that we essentially regard X as "fixed," even though it isn't. However, this may be okay for the prediction problem above. **Conditional** on having seen a particular average height, he wants to get a "best guess" for yield. He is **not** saying that he could **control** heights and thereby influence yields, only that, **given** he sees a certain height, he might be able to say something about the associated yield.

This subtlety is an important one. Inappropriate use of statistical techniques may lead one to erroneous or irrelevant inferences. Is it best to **consult a statistician** for help in identifying both a suitable model framework and the conditions under which regression analysis may be used with observational data.

10.5 "Causation vs. Correlation"

Investigators are often tempted to infer a **causal** relationship between X and Y when they fit a regression model or perform a correlation analysis. A significant **association** (of the straight line type or any type) between X and Y in either situation **does not necessarily** imply a **causal** relationship. This is best illustrated by what may seem to be a nonsensical example.

EXAMPLE: (Box, Hunter, and Hunter, *Statistics for Experimenters*, p. 8) The following plot show the population of Oldenberg, Germany at the the end of each of the 7 years 1930–1939 (Y) plotted against the number of storks observed that year in Oldenberg (X). It was common in the past in the US and other places for parents to tell their young children that storks bring babies rather than having to explain where babies **really** come from!



From the plot, there appears to be a strong association between X and Y. Few people, however, would infer that the increase in X (storks) **caused** the observed increase in Y! This may indeed seem nonsensical and contrived, but, nonetheless, many investigators are guilty of making such interpretations in other contexts!

Often, when two variables X and Y appear to have a strong association like this, it may be because **both** X and Y are in fact each associated with a **third** variable, say W. In the example, both Y and X increased with W = time, so it is understandable that X and Y appear to increase together. Here, then, the linear association between X and Y is basically a nonsensical thing to be concerned with – shooting storks will not decrease the birth rate of humans!

MORAL: This phenomenon is the basis of the remark "Correlation does not necessarily imply causation." An investigator should be aware of the temptation to infer causation in setting up a study, and be on the lookout for "lurking" variables like W above that are actually the driving force behind observed results. Box, Hunter, and Hunter give another example on p. 493–494.

10.6 Fitting a simple linear regression model – the method of least squares

Now that we have discussed some of the **conceptual** issues involved in studying the relationship between two variables, we are ready to describe practical implementation. We do this first for fitting a simple linear regression model.

Throughout this discussion, assume that it is legitimate to regard the X's as fixed.

For observations (X_i, Y_i), $i = 1, \ldots, n$, we postulate the simple linear regression model

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i, \quad i = 1, \ldots, n.$$

We wish to fit this model by **estimating** the intercept and slope **parameters** β_0 and β_1.

ASSUMPTIONS: For the purposes of making inference about the true values of intercept and slope, making predictions, and so on, we make the following assumptions. These are often reasonable, and we will discuss violations of them later in this chapter.

1. The observations $Y_1, \ldots, Y_n$ are **independent** in the sense that they are not related in any way. For example, they are derived from different animals, subjects, etc. They might also be measurements on the **same** subject, but taken far apart enough in time to where the value at one time is totally unrelated to that at another.

2. The observations $Y_1, \ldots, Y_n$ have the **same variance**, σ^2. Each Y_i is observed at a possibly different X_i value, and is thought to have mean $\mu = \beta_0 + \beta_1 X_i$. At each X_i value, we may thus think of the possible values for Y_i and how they might vary. This assumption says that, regardless of which X_i we consider, this variation in possible Y_i values is the same.

3. The observations Y_i are each normally distributed with mean $\mu_i = \beta_0 + \beta_1 X_i$, $i = 1 \ldots, n$, and variance σ^2 (the same, as in 2 above). That is, for each X_i value, we think of all the possible values taken on by Y as being well-represented by a normal distribution.

THE METHOD OF LEAST SQUARES: There is no one way to estimate β_0 and β_1. The most widely accepted method is that of **least squares**. This idea is intuitively appealing. It also turns out to be, mathematically speaking, the appropriate way to estimate these parameters under the assumption of normality 3 above.

For each Y_i, note that

$$Y_i - (\beta_0 + \beta_1 X_i) = \epsilon_i,$$

that is, the **deviation** $Y_i - (\beta_0 + \beta_1 X_i)$ is a measure of the vertical distance of the observation Y_i from the line $\beta_0 + \beta_1 X_i$ that is due to the inherent variation (represented by ϵ_i). This deviation may be negative or positive.

A natural way to measure the **overall deviation** of the observed data Y_i from their means, the regression line $\beta_0 + \beta_1 X_i$, due to this error is

$$\sum_{i=1}^{n} \{Y_i - (\beta_0 + \beta_1 X_i)\}^2.$$

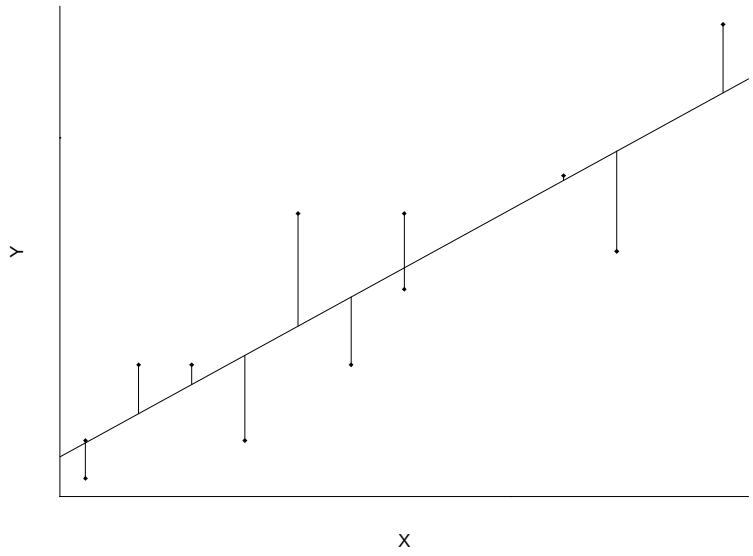
This has the same appeal as a sample variance – we ignore the signs of the deviations but account for their magnitude.

The method of **least squares** derives its name from thinking about this measure. In particular, we want to find the estimates of β_0 and β_1 that are the "most plausible" to have generated the data. Thus, a natural way to think about this is to choose as estimates the values $\hat{\beta}_0$ and $\hat{\beta}_1$ that make this measure of overall variation as small as possible (that is, which **minimize** it). This way, we are attributing as much of the overall variation in the data as possible to the assumed straight line relationship.

Formally, then $\hat{\beta}_0$ and $\hat{\beta}_1$ **minimize**

$$\sum_{i=1}^{n} (Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i)^2.$$

This is illustrated in the following picture. The line fitted by **least squares** is the one that makes the sum of the squares of all the vertical discrepancies as small as possible.



To find the form of the estimators $\hat{\beta}_0$ and $\hat{\beta}_1$, calculus may be used to solve this minimization problem. Define

$$S_{XY} = \sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y}) = \sum_{i=1}^{n} X_i Y_i - \frac{\left(\sum_{i=1}^{n} X_i\right)\left(\sum_{i=1}^{n} Y_i\right)}{n}$$

$$S_{XX} = \sum_{i=1}^{n}(X_i - \bar{X})^2 = \sum_{i=1}^{n} X_i^2 - \frac{\left(\sum_{i=1}^{n} X_i\right)^2}{n}$$

$$S_{YY} = \sum_{i=1}^{n}(Y_i - \bar{Y})^2 = \sum_{i=1}^{n} Y_i^2 - \frac{\left(\sum_{i=1}^{n} Y_i\right)^2}{n},$$

where $\bar{X}$ and $\bar{Y}$ are the sample means of the X_i and Y_i values, respectively. Then the calculus arguments show that the values $\hat{\beta}_0$ and $\hat{\beta}_1$ minimizing the sum of squared deviations above satisfy

$$\hat{\beta}_1 = \frac{S_{XY}}{S_{XX}}, \quad \hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X}.$$

The second formula in each case is preferred for hand calculation.

Thus, the **fitted** straight line is given by

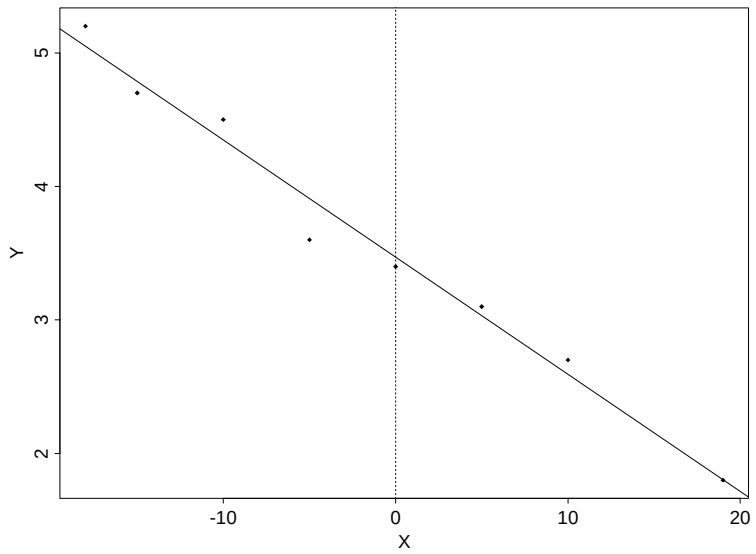
$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i.$$

The "hat" on the Y_i emphasizes the fact that these values are our "best guesses" for the means at each X_i value and that the **actual values** $Y_1, \ldots, Y_n$ we observed may not fall on the line. The $\hat{Y}_i$ are often called the **predicted values**; they are the estimated values of the **means** at the X_i.

EXAMPLE: (Zar, *Biostatistical Analysis*, p. 225) The following data are rates of oxygen consumption of birds (Y) measured at different temperatures (X). Here, the temperatures were set by the investigator, and the Y was measured, so the assumption of fixed X is justified.

X (degrees Celsius)	Y (ml/g/hr)
-18	5.2
-15	4.7
-10	4.5
-5	3.6
0	3.4
5	3.1
10	2.7
19	1.8

Here is a plot of the data. It is always advisable to plot the data before analysis, to ensure that the model assumptions seem valid.



CALCULATIONS: We have $n = 8$.

$$\sum_{i=1}^{n} Y_i = 29, \quad \bar{Y} = 3.625, \quad \sum_{i=1}^{n} Y_i^2 = 114.04$$

$$\sum_{i=1}^{n} X_i = -14, \quad \bar{X} = -1.75, \quad \sum_{i=1}^{n} X_i^2 = 1160,$$

$$\sum_{i=1}^{n} X_i Y_i = -150.4.$$

$$S_{XY} = -150.4 - \frac{(29)(-14)}{8} = -99.65$$

$$S_{XX} = 1160 - \frac{(-14)^2}{8} = 1135.5$$

$$S_{YY} = 114.0 - \frac{29^2}{8} = 8.915.$$

Thus, we obtain

$$\hat{\beta}_1 = \frac{-99.65}{1135.5} = -0.0878, \quad \hat{\beta}_0 = 3.625 - (-0.0878)(-1.75) = 3.4714.$$

The fitted line

$$\hat{Y}_i = 3.4714 + -0.0878 X_i$$

is superimposed on the plot. The vertical dashed line indicates $X = 0$ – the intercept estimate 3.4714 is the value at $X = 0$.

10.7 Assessing the fitted regression

Recall for a single sample, we use $\bar{Y}$ as our estimate of the mean and use the standard error $s_{\bar{Y}}$ as our estimate of precision of $\bar{Y}$ as an estimator of the mean. Here, we wish to do the same thing. How precisely have we estimated the intercept and slope parameters, and, for that matter, the line overall? Specifically, we would like to quantify

- The precision of the estimate of the line
- The variability in the estimates $\hat{\beta}_0$ and $\hat{\beta}_1$.

Consider the identity

$$Y_i - \bar{Y} = (\hat{Y}_i - \bar{Y}) + (Y_i - \hat{Y}_i).$$

Algebra and the fact that $\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i = \bar{Y} + \hat{\beta}_1(X_i - \bar{X})$ may be used to show that

$$\sum_{i=1}^{n}(Y_i - \bar{Y})^2 = \sum_{i=1}^{n}(\hat{Y}_i - \bar{Y})^2 + \sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2. \tag{10.6}$$

The quantity on the left hand side of this expression is one you should recognize – it is the **Total SS** for the set of data. For **any** set of data, we may always compute the Total SS as the sum of squared deviations of the observations from the (overall) mean, and it serves a measure of the overall variation in the data.

Thus, (10.6) represents a **partition** of our assessment of overall variation in the data, Total SS, into two independent components.

- $(\hat{Y}_i - \bar{Y})$ is the deviation of the **predicted value** of the ith observation from the overall mean. $\bar{Y}$ would be the estimate of mean response at all X values we would use if we did not believe X played a role in the values of Y. Thus, this deviation measures the difference between going to the trouble to have a **separate** mean for each X value versus just using a single, **common** mean as the model! We would thus expect the sum of squared deviations

  $$\sum_{i=1}^{n}(\hat{Y}_i - \bar{Y})^2$$

  to be **large** if using separate means via the regression model is much better than using a single mean. Using a single mean effectively **ignores** the X_i, so we may think of this as as measuring the variation in the observations that may be explained by the regression line $\beta_0 + \beta_1 X_i$.

- $(Y_i - \hat{Y}_i)$ is the deviation of the predicted value for the ith observation (our "best guess" for its mean) and the **observation itself** (that we observed). Hence, the sum of squared deviations

  $$\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2$$

  measures any additional variation of the observations about the regression line; that is, the inherent variation in the data at each X_i value that causes observations not to lie on the line.

Thus, the overall variation in the data, as measured by Total SS, may be broken down into two components that each characterize parts of the variation:

- **Regression SS** $= \sum_{i=1}^{n}(\hat{Y}_i - \bar{Y})^2$, which measures that portion of the variability that may be explained by the regression relationship (so is actually attributable to a **systematic** source, the assumed straight line relationship between Y and X).
- **Error SS** (also called **Residual SS**) $= \sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2$, which measures the inherent variability in the observations (e.g., **Experimental error**).

RESULT: We hope that the (straight line) regression relationship explains a good part of the variability in the data. A large value of **Regression SS** would in some sense indicate this.

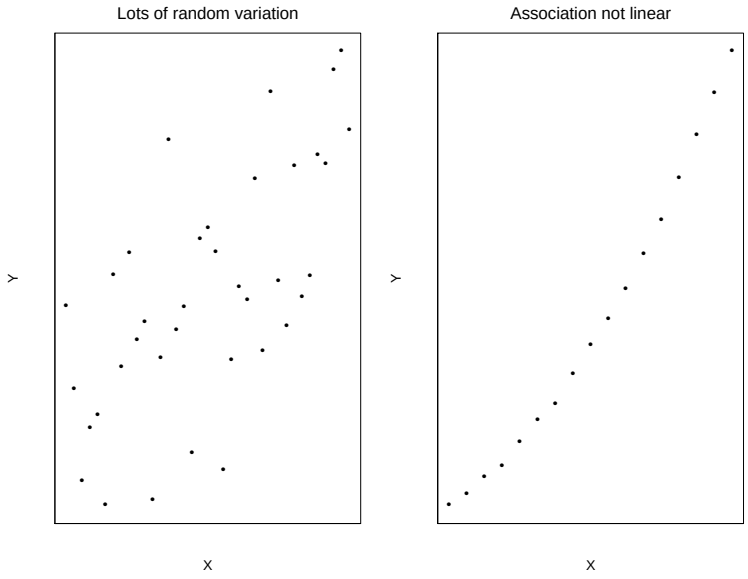
COEFFICIENT OF DETERMINATION: One measure of this is the ratio

$$R^2 = \frac{\text{Regression SS}}{\text{Total SS}}.$$

R^2 is called the **coefficient of determination** or the **multiple correlation coefficient**. (This second name arises from the fact that it turns out to be algebraically the value we would use to "estimate" the correlation between the Y_i and $\hat{Y}_i$ values, and is not to be confused with correlation as we have discussed it previously.)

Intuitively, R^2 is a measure of the "proportion of total variation in the data **explained** by the assumed straight line relationship with X." Note that we must have $0 \leq R^2 \leq 1$, because both components are nonnegative and the numerator can be no larger than the denominator. Thus, an R^2 value close to 1 is often taken as evidence that the regression model does "a good job" at describing the variability in the data, better than if we just assumed a common mean (and ignored X_i).

IMPORTANT: It is critical to understand what R^2 does and does not measure. R^2 is computed under the assumption that the simple linear regression model is correct; i.e. that it is a good description of the underlying relationship between Y and X. Thus, it assesses, **if the relationship between X and Y really is a straight line**, how much of the variation in the data may actually be attributed to that relationship rather than just to inherent variation. If R^2 is small it may be that there is a lot of random inherent variation in the data, so that, although the straight line is a reasonable model, it can only explain so much of the observed overall variation. This is the case in the plot on the left hand side.



Alternatively, R^2 may be close to 1, but the straight line model may not be the most appropriate model. This is the case in the right hand plot. R^2 may be quite "high" here, but in a sense that it irrelevant, because it assumes that the straight line model is correct when a better model actually exists.

ANALYSIS OF VARIANCE: The partition of Total SS above has the same interpretation as in the situations we have already discussed. Thus, it is common practice to summarize the results in an **analysis of variance** table. Note that Total SS has $n-1$ degrees of freedom, as **always**. It may be shown that

$$\text{Regression SS } = \sum_{i=1}^{n}(\hat{Y}_i - \bar{Y})^2 = \hat{\beta}_1 \sum_{i=1}^{n}(X_i - \bar{X})^2.$$

$\hat{\beta}_1$ is a single function of the Y_i, thus it is a **single** independent quantity. Thus, we see that Regression SS has 1 degree of freedom. By subtraction, Error SS has $n-2$ degrees of freedom.

Analysis of Variance – Simple Linear Regression

Source	DF	SS	MS	F
Regression	1	$\sum_{i=1}^n (\hat{Y}_i - \bar{Y})^2$	$MS_R = \frac{SS}{1}$	$F_R = \frac{MS_R}{MS_E}$
Error	$n-2$	$\sum_{i=1}^n (Y_i - \hat{Y}_i)^2$	$MS_E = \frac{SS}{n-2}$	
Total	$n-1$	$\sum_{i=1}^n (Y_i - \bar{Y})^2$		

CALCULATIONS:

$$\text{Regression SS } = \frac{S_{XY}^2}{S_{XX}}.$$

Total SS ($= S_{YY}$) is calculated in the usual way. Thus, Error (Residual) SS may be found by subtraction.

It turns out that the **expected mean squares**, that is, the values estimated by MS_R and MS_E are

MS_R	$\sigma^2 + \beta_1^2 \sum_{i=1}^n (X_i - \bar{X})^2$
MS_E	σ^2

Thus, is $\beta_1 = 0$, the two MSs we observe should be about the same, and we would expect F_R to be small. However, note that $\beta_1 = 0$ implies that the **true** regression line is

$$Y_i = \beta_0 + \epsilon_i;$$

that is, there is no association with X (slope $= 0$) and thus all Y_i have the **same mean** β_0, **regardless** of the value of X_i! There is a straight line relationship, but it has **slope** 0, which effectively means no relationship!

If $\beta_1 \neq 0$, then we would expect the ratio F_R to be **large**. It is possible to show mathematically that a test of the hypotheses

$$H_0 : \beta_1 = 0 \text{ vs. } H_1 : \beta_1 \neq 0$$

may be carried out by comparing F_R to the appropriate value from a $F_{1,n-2}$ distribution. That is, the statistic F may be shown to have this distribution if H_0 is true. Thus, the procedure would be

$$\text{Reject } H_0 \text{ at level of significance } \alpha \text{ if } F_R > F_{1,n-2,\alpha}.$$

The interpretation of the test is as follows. **Under the assumption that a straight line relationship exists**, we are testing whether or not the **slope** of this relationship is in fact zero. A zero slope means that there is no systematic change in mean along with change in X; that is, no association. It is important to recognize that if the true relationship is **not a straight line**, then this may be a meaningless test. For example, if the true relationship is something like a sine wave, that curves up and down over the range of the X's, then the concept of a "slope" is meaningless.

EXAMPLE: For the oxygen consumption data, assume that the data are approximately normally distributed with constant variance. We have

$$\text{Total SS } = S_{YY} = 8.915, \quad S_{XY} = -99.65, \quad S_{XX} = 1135.5$$

from before, with $n = 8$. Thus,

$$\text{Regression SS } = \frac{(-99.65)^2}{1135.5} = 8.745$$

$$\text{Error (Residual) SS } = 8.915 - 8.745 = 0.170.$$

Analysis of Variance Table for the Oxygen Data

Source	DF	SS	MS	F
Regression	1	8.745	8.745	308.927
Error (Residual)	6	0.170	0.028	
Total	7	8.915		

We have $F_{1,6,0.05} = 5.99$. $F_R = 308.927 >> 5.99$, thus, we **reject** H_0 at level of significance $\alpha = 0.05$. There is strong evidence in these data to suggest that, under the assumption that the simple linear regression model is appropriate, the slope is **not** zero, so that an association appears to exist. Note that

$$R^2 = \frac{8.745}{8.915} = 0.981;$$

thus, as the straight line assumption appears to be consistent with the visual evidence in the plot of the data, it is reasonable to conclude that the straight line relationship explains a very high proportion of the variation in the data (the fact that Y_i values are different is mostly due to the relationship with X).

ESTIMATE OF σ^2: If we desire an estimate of the variance σ^2 associated with inherent variation in the Y_i values (due to variation among experimental units, sampling, and measurement error), from the expected mean squares above, the obvious estimate is the Error (Residual) MS. That is, denoting the estimate by s^2,

$$s^2 = MS_E.$$

10.8 Confidence intervals for regression parameters and means

Because β_0, β_1, and in fact the entire regression line are population parameters that we have estimated, we wish to attach some measure of precision to our estimates of them.

STANDARD ERRORS AND CONFIDENCE INTERVALS FOR REGRESSION PARAMETERS: It turns out that it may be shown, under our assumptions, that, if the relationship really is a straight line, the standard deviations of the populations of all possible $\hat{\beta}_1$ and $\hat{\beta}_0$ values are

$$SD(\hat{\beta}_1) = \frac{\sigma}{\sqrt{S_{XX}}}, \quad SD(\hat{\beta}_0) = \frac{\sigma\sqrt{\sum_{i=1}^n X_i^2}}{\sqrt{nS_{XX}}},$$

respectively. Because σ is not known, we estimate these standard deviations by replacing σ by the estimate s. We thus obtain the estimated standard deviations

$$EST\ SD(\hat{\beta}_1) = \frac{s}{\sqrt{S_{XX}}}, \quad EST\ SD(\hat{\beta}_0) = \frac{s\sqrt{\sum_{i=1}^n X_i^2}}{\sqrt{nS_{XX}}},$$

respectively. These are often referred to, analogous to a single sample mean, as the **standard errors** of $\hat{\beta}_1$ and $\hat{\beta}_0$.

It may also be shown under our assumptions that

$$\frac{\hat{\beta}_1 - \beta_1}{EST\ SD(\hat{\beta}_1)} \sim t_{n-2} \text{ and } \frac{\hat{\beta}_0 - \beta_0}{EST\ SD(\hat{\beta}_0)} \sim t_{n-2}. \tag{10.7}$$

These results are similar in spirit to those for a single mean and difference of means in chapter 5; the t distribution is relevant rather than the normal because we have replaced σ by an estimate (with $n-2$ degrees of freedom).

Because we are estimating the **true** parameters β_1 and β_0 by these estimates, it is common practice to provide a **confidence interval** for the true values β_1 and β_0, just as we did for a sample mean or difference of means. The derivation is in the same spirit as those given in chapters 4 and 5, and the interpretation is the same. By "inverting" probability statements about the quantities in (10.7) in the same fashion, we arrive at the following $100(1-\alpha)\%$ confidence intervals:

Interval for β_1:

$$\{\hat{\beta}_1 - t_{n-2,\alpha/2} EST\ SD(\hat{\beta}_1)\,,\ \hat{\beta}_1 + t_{n-2,\alpha/2} EST\ SD(\hat{\beta}_1)\}$$

Interval for β_0:

$$\{\hat{\beta}_0 - t_{n-2,\alpha/2} EST\ SD(\hat{\beta}_0)\,,\ \hat{\beta}_0 + t_{n-2,\alpha/2} EST\ SD(\hat{\beta}_0)\}$$

The interpretation is as follows: Suppose that zillions of experiments were conducted using the **same** (fixed) X_i values as those in the observed experiment. Suppose that for each of these, we fitted the regression line by the above procedures and calculated $100(1-\alpha)\%$ confidence intervals for β_1 (and β_0). Then for $100(1-\alpha)\%$ of these, the **true** value of β_1 (β_0) would fall between the endpoints. The endpoints are a function of the **data**; thus, whether or not β_1 (β_0) falls within the endpoints is a function of the **experimental procedure**. Thus, just as in our earlier cases, the confidence interval is a statement about the quality of the experimental procedure for learning about the value of β_1 (β_0).

EXAMPLE: For the oxygen consumption data, $n = 8$, $t_{6,0.025} = 2.447$, $\hat{\beta}_0 = 3.4714$, $\hat{\beta}_1 = -0.0878$, $s^2 = 0.028$, so that

$$EST\ SD(\hat{\beta}_0) = \frac{\sqrt{0.028}\sqrt{1160}}{\sqrt{8(1135.5)}} = 0.06012$$

$$EST\ SD(\hat{\beta}_1) = \frac{\sqrt{0.028}}{\sqrt{1135.5}} = 0.0050$$

Thus, if we take $\alpha = 0.05$, a 95% confidence interval for

$$\beta_0\text{: } 3.4714 \pm (2.447)(0.601) = (3.3216, 3.6185)$$

$$\beta_1\text{: } -0.0878 \pm (2.447)(0.0050) = (-0.0999, -0.0755)$$

As above, the interpretation of these intervals is as a statement about the quality of the experimental procedure. For example, β_1 has units of the rate of change of oxygen consumption per unit change in temperature, ml/g/hr per degree Celsius. The interval may be considered to be pretty **narrow** (≈ 0.025); the smallest observed response is **72 times** this value!

STANDARD ERROR AND CONFIDENCE INTERVAL FOR THE MEAN: Our interest in the values of β_0 and β_1 is usually because we are interested in the characteristics of Y at particular X values. Recall that $\mu_i = \beta_0 + \beta_1 X_i$ is the **mean** of the Y_i values at the value of X X_i. Thus, just as we are interested in estimating the mean of a single sample to give us an idea of the "center" of the distribution of possible values, we may be interested in estimating the mean of Y_i values at a particular value of X.

For example, an experiment may have been conducted to characterize the relationship between concentration of a suspected toxicant (X) and a response like number of mutated cells (Y). Based on the results, the investigators may wish to estimate the numbers of mutations that might be seen **on average** at other concentrations not considered in the experiment. That is, they are interested in the "typical" (mean) number of mutations at particular X values.

Consider this problem in general. Suppose we have fitted a regression line to data, and we wish to estimate the **mean** response at a new value X_0. That is, we wish to estimate

$$\mu_0 = \beta_0 + \beta_1 X_0.$$

The obvious estimator for μ_0 is the value of this expression with the estimates of the regression parameters plugged in, that is,

$$\hat{\mu}_0 = \hat{\beta}_0 + \hat{\beta}_1 X_0.$$

In the example, $\hat{\mu}_0$ is thus our estimate of the average number of mutations at some concentration X_0. Note that, of course, the estimate of the mean will depend on the value of X_0.

Because μ_0 is an **estimate** based on our sample, we again would like to attach to it some estimate of precision. It may be shown mathematically that the variance of the distribution of all possible $\hat{\mu}_0$ values (based on the results of all possible experiments giving rise to all possible values of $\hat{\beta}_0$ and $\hat{\beta}_1$) is

$$\sigma^2 \left(\frac{1}{n} + \frac{(X_0 - \bar{X})^2}{S_{XX}} \right).$$

We may thus estimate the standard deviation of $\hat{\mu}_0$ by

$$EST\ SD(\hat{\mu}_0) = s \sqrt{\frac{1}{n} + \frac{(X_0 - \bar{X})^2}{S_{XX}}}.$$

It may be shown that, under our assumptions,

$$\frac{\hat{\mu}_0 - \mu_0}{EST\ SD(\hat{\mu}_0)} \sim t_{n-2}.$$

Thus, using our standard argument to construct a $100(1-\alpha)\%$ confidence interval for a population parameter based on such a result, we have that a $100(1-\alpha)\%$ confidence interval for μ_0, the **true mean** of all possible Y values at the fixed value X_0, is

$$\{\hat{\mu}_0 - t_{n-2,\alpha/2} EST\ SD(\hat{\mu}_0)\,,\ \hat{\mu}_0 + t_{n-2,\alpha/2} EST\ SD(\hat{\mu}_0)\}.$$

The interpretation is the same as above.

LENGTH OF CONFIDENCE INTERVAL: The confidence interval for μ_0 will of course be different depending on the value of X_0. In fact, the expression for $EST\ SD(\hat{\mu}_0)$ above will be **smallest** if we choose $X_0 = \bar{X}$ and will get larger the farther X_0 is from $\bar{X}$ in **either direction**. This implies that the precision with which we expect to estimate the mean value of Y **decreases** the farther X_0 is from the "middle" of the original data. This makes intuitive sense – we would expect to have the most "confidence" in our fitted line as an estimate of the true line in the "center" of the observed data. The result is that the confidence intervals for μ_0 will be wider the farther X_0 is from $\bar{X}$.

IMPLICATION: If the fitted line will be used to estimate means for values of X besides those used in the experiment, it is important to use a range of X's which contains the future values of interest, X_0, preferably more toward the "center."

EXTRAPOLATION: It is sometimes desired to estimate the mean based on the fit of the straight line for values of X_0 **outside** the range of X's used in the original experiment. This is called **extrapolation**. In order for this to be valid, we must believe that the straight line relationship holds for X's outside the range where we have observed data! In some situations, this may be reasonable; in others, we may have no basis for making such a claim without data to support it. It is thus very important that the investigator have an **honest** sense of the relevance of the straight line model for values outside those used in the experiment if inferences such as estimating the mean for such X_0 values are to be reliable. In the event that such an assumption is deemed to be relevant, note from the above discussion that the quality of the estimates of the μ_0 for X_0 outside the range is likely to be poor.

Note, of course, that we may in fact be interested in the mean at values of X that **were** included in the experiment. The procedure above is of course valid in this case.

EXAMPLE: In the oxygen consumption example, suppose we desire a 95% confidence interval for the mean of possible Y values at $X_0 = 2.5$ degrees. Note that this value is close to the "center" of X's in the experiment.

We have $\hat{\beta}_0 = 3.4714$, $\hat{\beta}_1 = -0.0878$, $s^2 = 0.028$, $S_{XX} = 1135.5$, and $\bar{X} = -1.75$. Thus,

$$\hat{\mu}_0 = 3.4714 + (-0.0878)(2.5) = 3.2519 \text{ml/g/hr}$$

$$EST\ SD(\hat{\mu}_0) = \sqrt{0.028}\sqrt{\frac{1}{8} + \frac{\{2.5 - (-1.75)\}^2}{1135.5}} = 0.0632.$$

Thus, the confidence interval for μ_0, the mean at $X_0 = 2.5$, is $3.2519 \pm (2.447)(0.0632)$, or

$$(3.0973, 3.4065) \text{ ml/g/hr}.$$

10.9 Prediction and calibration

PREDICTION: Sometimes, depending on the context, we may be interested not in the **mean** of possible Y values at a particular X_0 value, but in fact the **actual value** of Y we might **observe** at X_0. This distinction is important. The estimate of the mean at X_0 provides just a general sense about values of Y we might see there – just the "center" of the distribution. This may not be adequate for some applications. For example, consider a stockbroker who would like to learn about the value of a stock based on observed previous information. The stockbroker does not want to know about what might happen "on the average" at some future time X_0; she is concerned with the **actual value** of the stock at that time, so that she may make sound judgments for her clients. The stockbroker would like to **predict** or **forecast** the **actual value**, Y_0 say, of the stock that might be **observed** at X_0.

In this kind of situation, we are interested not in the **population parameter** μ_0, but rather the **actual value** that might be taken on by a **random variable**, Y_0, say. In the context of our model, we are thus interested in the "future" observation

$$Y_0 = \beta_0 + \beta_1 X_0 + \epsilon_0,$$

where ϵ_0 is the "error" associated with Y_0 that makes it differ from the mean at X_0, μ_0. It is important to recognize that Y_0 is not a **parameter** but a **random variable**; thus, we do not wish to **estimate** a fixed quantity, but instead learn about the value of a **random** quantity.

Now, our "best guess" for the value Y_0 is **still** our estimate of the "central" value at X_0, the mean μ_0. We will write

$$\hat{Y}_0 = \hat{\beta}_0 + \hat{\beta}_1 X_0$$

to denote this "best guess." Note that this is identical to our estimate for the mean, $\hat{\mu}_0$; however, we use a different symbol in this context to remind ourselves that we are interested in Y_0, not μ_0. We call $\hat{Y}_0$ a **prediction** or **forecast** rather than an "estimate" to make the distinction clear.

Of course, just as we do in estimation of fixed parameters, we would still like to have some idea of how well we can predict/forecast. To get an idea, we would like to characterize the **uncertainty** that we have about $\hat{Y}_0$ as a guess for Y_0, but, because it is not a parameter, it is not clear what to do. Our usual notion of a standard error and confidence interval does not seem to apply.

We **can** write down an assessment of the likely size of the error we might make in using $\hat{Y}_0$ to characterize Y_0. Intuitively, there will be two sources of error:

- Part of the error in $\hat{Y}_0$ comes from the fact that we don't know β_0 and β_1 but must estimated them from the observed data.
- **Additional error** arises from the fact that what we are really doing is trying to "hit a moving target!" That is, Y_0 itself is a **random variable**, so itself is **variable**! Thus, additional uncertainty is introduced because we are trying to characterize a quantity that **itself** is uncertain!

The assessment of uncertainty thus should be composed of **two** components. An appropriate measure of uncertainty is the standard deviation of $\hat{Y}_0 - Y_0$ – that is, the variability in the deviations between $\hat{Y}_0$ and the thing we are trying to "hit," Y_0. This variance turns out to be

$$\sigma^2 + \sigma^2 \left(\frac{1}{n} + \frac{(X_0 - \bar{X})^2}{S_{XX}} \right).$$

The extra σ^2 added on is accounting for the additional variation above (i.e. the fact that Y_0 itself is variable). We estimate the associated standard deviation of this variance by substituting s^2 for σ^2:

$$EST\ ERR(\hat{Y}_0) = s\sqrt{1 + \left(\frac{1}{n} + \frac{(X_0 - \bar{X})^2}{S_{XX}} \right)}.$$

We call this "$EST\ ERR$" to remind ourselves that we this is an estimate of the "error" between $\hat{Y}_0$ and Y_0, each of which is random.

The usual procedure is to use this estimated uncertainty to construct what might be called an "uncertainty interval" for Y_0 based on our observed data. Such an interval is usually called a **prediction interval**. A $100(1-\alpha)\%$ interval is given by

$$\{\hat{Y}_0 - t_{n-2,\alpha/2} EST\, ERR(\hat{Y}_0)\,,\ \hat{Y}_0 + t_{n-2,\alpha/2} EST\, ERR(\hat{Y}_0)\}.$$

Note that this interval is **wider** that the confidence interval for the mean μ_0. This is because we are trying to forecast the value of a random variable rather than estimate just a single population parameter. Understandably, we cannot do the former as well as the latter, because Y_0 varies as well.

EXAMPLE: For the oxygen consumption data, we have $\hat{Y}_0 = 3.2159$.

$$\hat{\mu}_0 = 3.4714 + (-0.0878)(2.5) = 3.2519 \text{ml/g/hr}$$

$$EST\, SD(\hat{\mu}_0) = \sqrt{0.028}\sqrt{1 + \frac{1}{8} + \frac{\{2.5 - (-1.75)\}^2}{1135.5}} = 0.1787.$$

Thus, the confidence interval for μ_0, the mean at $X_0 = 2.5$, is $3.2519 \pm (2.447)(0.1787)$, or

$$(2.8147, 3.6891).$$

This interval is an estimate of the range of Y_0 values between which 95% of the area under the probability histogram falls.

CALIBRATION: Suppose we have fitted the regression line and now, for a value Y_0 of Y we have observed, we would like to **estimate** the unknown corresponding value of X, say X_0.

As an example, consider a situation where interest focuses on two different methods of calculating the age of a tree. One way is by counting tree rings. This is considered to be very accurate, but requires sacrificing the tree. Another way is by a carbon-dating process. Suppose that data are obtained for n trees on X = age by the counting method, Y = age by carbon dating. Here, technically, both of these might be considered random variables; however, the goal of the investigators was to determine the relationship of the **more variable**, less reliable carbon data method to the accurate counting method. That is, **given** a tree is of age X (which could be determined exactly by the counting method), what would the associated carbon method value look like? Thus, for their purposes, regression analysis is appropriate. From the observed pairs (X_i, Y_i), a straight line model seems reasonable, and is fitted to the data. Now suppose that the carbon data method is applied to a tree not in the study yielding an age Y_0. What can we say about the true age of the tree, X_0, (that is, its age by the very accurate counting method) without sacrificing the tree?

The idea is to use the fitted line to estimate X_0. Note that X_0 is a **fixed value**, thus, it is perfectly legitimate to want to estimate it. The obvious choice for an estimator, $\hat{X}_0$, say, is found by "inverting" the fitted regression line:

$$\hat{X}_0 = \frac{Y_0 - \beta_0}{\beta_1}.$$

Because $\hat{X}_0$ is an **estimate** of a **parameter**, based on information from our original sample, we ought to also report **standard error** and/or **confidence interval**.

It turns out that deriving such quantities is a much harder mathematical exercise than for estimating a mean or for prediction, and a description of this is beyond the scope of our discussion here. This is because the estimated regression parameters appear both in the numerator and denominator of the estimate, which leads to mathematical difficulties. Be aware that this may indeed be done; if you have occasion to make calibration inference, you should consult a statistician for help with attaching estimates of precision to your calibrated values (the estimates $\hat{X}_0$.)

EXAMPLE: For the oxygen consumption data, suppose that a bird not in the original sample is measured to have an oxygen consumption rate of 4.4 ml/g/hr. What is our estimate for the temperature at which this reading was taken?

$$\hat{X}_0 = \frac{4.2 - 3.4714}{-0.0878} = -8.2984 \text{ deg. C.}$$

10.10 Violation of assumptions

Earlier in this chapter, we stated assumptions under which the methods for simple linear regression we have discussed yield valid inferences. Just as we discussed in chapters 7 and 9, it is often the case in practice that one or more of the assumptions is violated. As in those situations, there are several ways in which the assumptions may be violated.

- *Nonconstant variance*: Recall in the regression situation that the **mean response** changes with X. Thus, in a given experiment, the responses may arise from distributions with means across a **large range**. The usual assumption is that the **variance** of these distributions is the **same**. However, it is often the case the the variability in responses changes, most commonly in an **increasing** fashion, with changing X and mean values. This is thus likely to be of concern in problems where the response means cover a large range.

  We have already discussed the idea of **data transformation** as a way of handling this violation of the usual assumptions. In the regression context, this may be done in a number of ways. One way is to invoke an appropriate transformation, and then postulate a regression model on the transformed scale. Sometimes, in fact, it may be that, although the data do **not** appear to follow a straight line relationship with X on the **original scale**, they may on some **transformed scale**.

  Another approach is to proceed with a modified method known as **weighted least squares**; this is described by STD in their chapter 10. This method, however, requires that the variances are **known**, which is rarely the case in practice.

  A number of **diagnostic procedures** have been developed for helping to determine if nonconstant variance is an issue and how to handle it. Other approaches to transforming data are also available. Discussion of these and of weighted least squares is beyond the scope of this course. The best strategy is to consult a statistician to help with both diagnosis and identification of the best methods for a particular problem.

- *Nonnormality*: Also, as we have discussed previously, the normal distribution may not provide a realistic model for some types of data, such as those in the form of **counts** or **proportions**. **Transformations** like those discussed in chapter 9 may be used in the regression context as well. In addition, there are other approaches to investigating the association between such responses Y and a covariate X that we have not discussed here. Again, a statistician can help you determine the best approach.

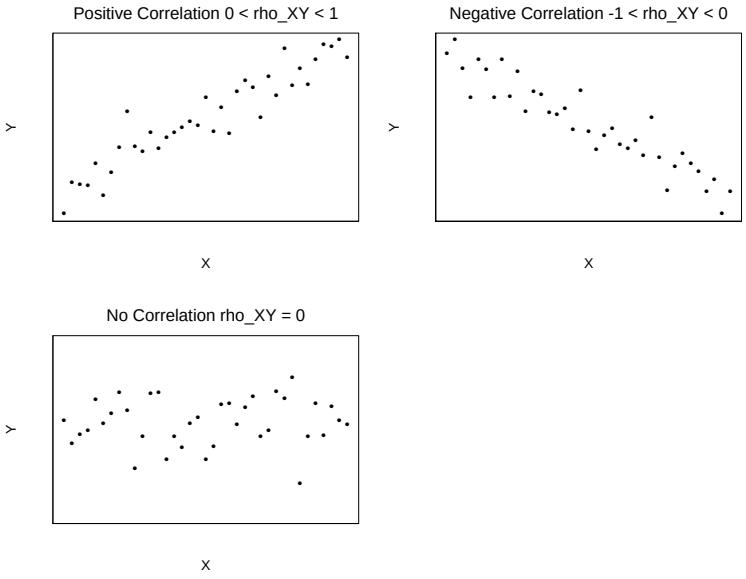
- *Outliers*: Another phenomenon that can make the normal approximation unreasonable is the problem of **outliers**; i.e. data points that do not seem to fit well with the pattern of the rest of the data. In the context of straight line regression, an outlier might be an observation that falls far off the apparent approximate straight line trajectory followed by the remaining observations. Practitioners may often "toss out" such anomalous points, which may or may not be a good idea, depending on the problem. If it is clear that an "outlier" is the result of a mishap or a gross recording error, than this may be acceptable. On the other hand, if no such basis may be identified, the outlier may in fact be a genuine response; in this case, it contains information about the process under study, and may be reflecting a legitimate phenomenon. In this case, "throwing out" an outlier may lead to misleading conclusions, because a legitimate feature is being ignored. Again, there are sophisticated **diagnostic** procedures for identifying outliers and deciding how to handle them. A statistician can help you with these issues.

10.11 Correlation analysis

Throughout this discussion, we regard **both** Y and X as random variables such that the **bivariate normal distribution** provides an appropriate, approximate model for their **joint distribution**.

CORRELATION: Recall that the **correlation coefficient** ρ_{XY} is a measure of the degree of (linear) association between two random variables. ρ_{XY} satisfies $-1 \leq \rho_{XY} \leq 1$, with $\rho_{XY} = 1$ denoting a "perfect" positive association, $\rho_{XY} = -1$ denoting a "perfect" negative association, and $\rho_{XY} = 0$ denoting "no association." This was described earlier in this chapter.

In practice, "perfect" association is rarely observed. We are more likely to observe intermediate associations or no association, as illustrated in the following picture:



INTERPRETATION: It is very important to understand what correlation **does not** measure. Investigators sometimes confuse the value of the correlation coefficient and the **slope** of an apparent underlying straight line relationship. These do not have anything to do with each other:

- The correlation coefficient may be virtually equal to 1, implying an almost perfect association. But the slope may be very **small** at the same time. Although there is indeed an almost perfect association, the **rate of change** of Y values with X values may be very **slow**.
- The correlation coefficient may be very small, but the apparent "slope" of the relationship could be very steep. In this situation, it may be that, although the rate of change of Y values with X values is fast, there is large inherent variation in the data.

ESTIMATION: For a particular set of data, of course, ρ_{XY} is **unknown**. We may estimate ρ_{XY} from a set of n pairs of observations (X_i, Y_i), $i = 1, \ldots, n$, by the **sample correlation coefficient**

$$r_{XY} = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^{n}(X_i - \bar{X})^2 \sum_{i=1}^{n}(Y_i - \bar{Y})^2}}.$$

In our shorthand notation,

$$r_{XY} = \frac{S_{XY}}{\sqrt{S_{XX} S_{YY}}}.$$

For hand calculation, one should use the preferred forms of S_{XY}, S_{XX}, and S_{YY}.

Note that the **same** calculations are involved as were need for regression analysis! In fact, recall in regression analysis that

$$\sqrt{\text{Regression SS}} = \frac{S_{XY}}{\sqrt{S_{XX}}}.$$

The quantity r^2_{XY} is thus often called the **coefficient of determination** (like R^2) in this setting, where correlation analysis is appropriate. However, it is important to recognize that the **interpretation** is **different**. Here, we are not acknowledging a straight line relationship; rather, we are just modeling the data in terms of a bivariate normal distribution with correlation ρ_{XY}. Thus, the former interpretation for the quantity r^2_{XY} has no meaning here.

Likewise, the idea of **correlation** really only has meaning when both variables Y and X are **random variables**.

CONFIDENCE INTERVAL: Because r_{XY} is an **estimator** of the population parameter ρ_{XY}, it would be desirable to report, along with the estimate itself, a **confidence interval** for ρ_{XY}.

There is no one way to carry out these analyses. One common approach is an **approximation** known as **Fisher's Z transformation**. The method is based on the mathematical result that the quantity

$$Z' = 0.5 \log\left(\frac{1 + r_{XY}}{1 - r_{XY}}\right)$$

has an **approximate** normal distribution with mean and variance

$$0.5 \log\left(\frac{1 + \rho_{XY}}{1 - \rho_{XY}}\right) \text{ and } \frac{1}{n-3},$$

respectively, when n is **large**.

This result may be used to construct an approximate $100(1-\alpha)\%$ confidence interval for the **mean**

$$0.5 \log\left(\frac{1 + \rho_{XY}}{1 - \rho_{XY}}\right),$$

where log is natural logarithm. This confidence interval is

$$\left(Z' - z_{\alpha/2}\sqrt{\frac{1}{n-3}}\,,\; Z' + z_{\alpha/2}\sqrt{\frac{1}{n-3}}\right), \tag{10.8}$$

where, as before, $z_{\alpha/2}$ is the value such that, for a standard normal r.v. Z, $z_{\alpha/2}$ satisfies $P(Z > z_{\alpha/2})$. This confidence interval may then be **transformed** to obtain a confidence interval for ρ_{XY} itself as follows.

Let Z'_L and Z'_U be the lower and upper endpoints of the interval (10.8). We illustrate the approach for Z'_L. To obtain the lower endpoint for an approximate confidence interval for ρ_{XY} itself, calculate

$$\frac{\exp(2Z'_L) - 1}{\exp(2Z'_L) + 1}.$$

This value is the lower endpoint of the ρ_{XY} interval; to obtain the upper endpoint, apply the same formula to Z'_U.

HYPOTHESIS TEST: We may also be interested in **testing hypotheses** about the value of ρ_{XY}. The usual hypotheses tested are analogous in spirit to what is done in straight line regression – the null hypothesis is the hypothesis of "no association." Here, then, we test

$$H_0 : \rho_{XY} = 0 \text{ vs. } H_1 : \rho_{XY} \neq 0.$$

It is important to recognize what is being tested here. The alternative states simply that ρ_{XY} is **different from** zero. The **true** value of the correlation coefficient could be **quite small** and H_1 would be true. Thus, if the null hypothesis is rejected, this is not necessarily an indication that there is a "strong" association, just there is evidence that there is **some** association.

Of course, as always, if we **do not** reject H_0, this does not mean that we **do** have enough evidence to infer that there is **not** an association! This is particularly critical here, as we discuss in a moment.

The procedure for testing H_0 vs. H_1 is intuitively reasonable: we **reject** H_0 if the confidence interval **does not** contain 0.

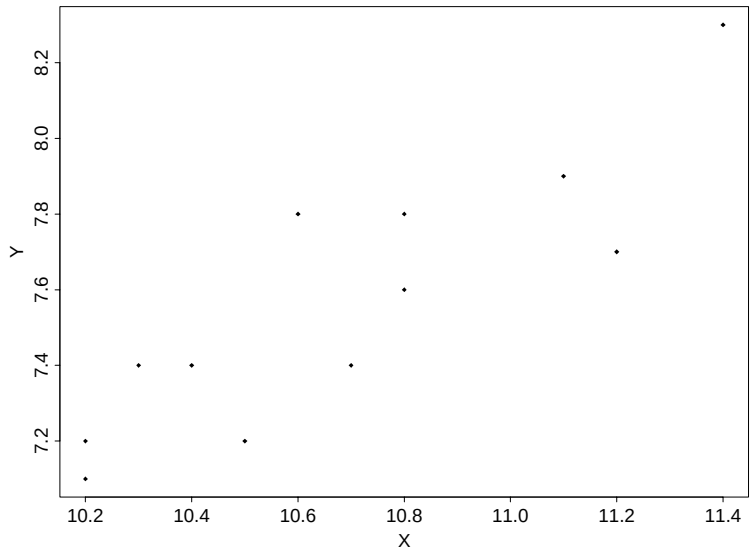
It is possible to modify this procedure to test whether ρ_{XY} is equal to some other value besides zero; STD describe this on p. 295. However, be aware that most statistical packages provide by default the test for $H_0 : \rho_{XY} = 0$ only.

WARNING: This procedure is **only approximate**, even under our bivariate normal assumption. It is an example of the type of approximation that is often made in difficult problems, that of approximating the behavior of a statistic under the condition that the sample size, n, is **large**. If n is small, the procedure is likely to be unreliable. Moreover, it is worth noting that, intuitively, trying to understand the underlying association between 2 **random variables** is likely to be very difficult with a small number of pairs of observations. Thus, testing aside, one should be very wary of over-interpretation of the estimate of ρ_{XY} when n is small – one "outlying" or "unusual" observation could be enough to affect the computed value substantially! It thus may be very difficult to detect when ρ_{XY} is different from 0 with a small sample size!

EXAMPLE: The following data are measurements on wing length (X) and tail length (Y) for a sample of $n = 12$ birds:

Wing length (X, cm)	Tail length (Y, cm)
10.4	7.4
10.8	7.6
11.1	7.9
10.2	7.2
10.3	7.4
10.2	7.1
10.7	7.4
10.5	7.2
10.8	7.8
11.2	7.7
10.6	7.8
11.4	8.3

A plot of the data reveals a possible positive association.



We obtain

$$\sum_{i=1}^{n} X_i = 128.2, \quad \sum_{i=1}^{n} X_i^2 = 1371.32, \quad S_{XX} = 1.717,$$

$$\sum_{i=1}^{n} Y_i = 90.8, \quad \sum_{i=1}^{n} Y_i^2 = 688.40, \quad S_{YY} = 1.347,$$

$$\sum_{i=1}^{n} X_i Y_i = 971.37, \quad S_{XY} = 1.323.$$

Thus, our estimate of ρ_{XY} is

$$r_{XY} = \frac{1.323}{\sqrt{(1.717)(1.347)}} = 0.8704.$$

The estimate is fairly close to 1. We calculate an approximate 95% confidence interval:

$$Z' = 0.5 \log\left(\frac{1 + 0.8704}{1 - 0.8704}\right) = 1.335.$$

We have $z_{0.025} = 1.96$, $\sqrt{1/(n-3)} = 1/3$, so that the interval is

$$\{1.335 - (1.96)(1/3)\,,\ 1.335 - (1.96)(1/3)\} = (0.681, 1.988).$$

We transform this to an interval for ρ_{XY} itself:

$$\frac{\exp\{2(0.681)\} - 1}{\exp\{2(0.681)\} + 1} = 0.592, \quad \frac{\exp\{2(1.988)\} - 1}{\exp\{2(1.988)\} + 1} = 0.963.$$

Thus, the interval is $(0.592, 0.963)$. The interval does not contain 0; thus, if we were test H_0 vs. H_1 above, we would clearly **reject** H_0 and infer that there is evidence to suggest that there is some association between wing and tail lengths (under the assumption that our bivariate normal distribution model is correct). The confidence interval does support the contention that the association is reasonably high; from the plot, there is a definite positive trend, obscured by some moderate variation.

10.12 Using SAS to perform simple linear regression and correlation analyses

EXAMPLE 1: Simple linear regression.

To perform simple linear regression (or indeed, more complicated regression) analyses in SAS, we may use **either** `PROC GLM` or `PROC REG`. `PROC GLM` is, as the name "General Linear Model" implies, a very general procedure, while `PROC REG` is intended specifically for regression analysis. Thus, we use `PROC REG`. The procedure also has a number of **diagnostic** features and the ability to calculate confidence intervals for the mean and prediction intervals. We also illustrate how `PROC PLOT` may be used to produce a plot of the data with the fitted regression line superimposed.

The syntax for specifying the model to be fitted is similar to that in `PROC REG`. The systematic part of the (linear additive) model here is

$$\beta_0 + \beta_1 X,$$

and is represented by the statement

```
MODEL Y = X.
```

Thus, in `PROC REG` syntax, the **intercept** is always understood, just like the overall mean in `PROC GLM`. The `X` in the `MODEL` statement tells SAS to include a term with a slope coefficient for X.

The "`/`" after this statement followed by `P R CLM` asks SAS to compute for each X_i value in the data set the predicted value ($\hat{Y}_i$), the **residual** $Y_i - \hat{Y}_i$, and the 95% confidence limits (endpoints of the confidence interval) for the mean μ_i at each X_i value (that is, taking in turn $X_0 = X_i$ for each i). Prediction limits may be obtained by specifying `CLI`; other possibilities are given in the SAS documentation for `PROC REG`.

The output gives an analysis of variance table, R^2 (`R-SQUARE`), and the parameter estimates $\hat{\beta}_0$ and $\hat{\beta}_1$ and their estimated standard errors. We may use these to construct confidence intervals for β_0 and β_1.

The final statement `OUTPUT OUT=OXYGEN2 P=PRED` instructs SAS to create a new data set named `OXYGEN2` containing everything in the original data set `OXYGEN` **plus** a **new** variable named `PRED`, containing the predicted values $\hat{Y}_i$ at each X_i. In `PROC PLOT`, we operate on this new data set. The statement `PLOT Y*X='*'` instructs SAS to create a plot with "`*`" as the plotting symbol. The second part, `PRED*X='P' / OVERLAY`, asks SAS to also plot the predicted values against X and **overlay** this plot on top of that of the data. If we were to connect the "`P`"s appearing on the plot, we would get a graph of the fitted straight line.

PROGRAM:

```
*********************************************;
*                                           ;
*          ST 511  USING PROC REG TO        ;
*         FIT A SIMPLE LINEAR REGRESSION    ;
*                                           ;
*    APPLIED TO THE OXYGEN CONSUMPTION DATA  :
*                                           ;
*********************************************;

OPTIONS LS=80 PS=59 NODATE;

*********************************************;
*                                           ;
*    CREATE A SAS DATA SET "OXYGEN"         ;
*    CONTAINING TWO VARIABLES X AND Y       ;
*                                           ;
*********************************************;

DATA OXYGEN;  INPUT X Y;
  CARDS;
-18  5.2
-15  4.7
-10  4.5
 -5  3.6
  0  3.4
  5  3.1
 10  2.7
 19  1.8
;

PROC PRINT; TITLE 'OXYGEN CONSUMPTION DATA'; RUN;
```

```
*******************************************;
*                                         ;
*     CALL PROC REG TO FIT THE REGRESSION     ;
*     LINE. THE MODEL STATEMENT TELLS SAS     ;
*     TO FIT THE SIMPLE LINEAR REGRESSION     ;
*     MODEL                               ;
*             Y = B0 + B1 X + E           ;
*                                         ;
*     THE LETTERS P, R, CLM AFTER THE "/"     ;
*     TELL SAS TO COMPUTE THE PREDICTED       ;
*     VALUES FROM THE FITTED LINE, THE        ;
*     RESIDUALS (Y - PREDICTED VALUE) AND     ;
*     FOR EACH X VALUE, A 95% CONFIDENCE      ;
*     INTERVAL FOR THE MEAN PREDICTED VALUE   ;
*                                         ;
*     THE "OUTPUT OUT=OXYGEN2" STATEMENT      ;
*     CREATES A NEW DATA SET CONTAINING THE   ;
*     OLD DATA SET PLUS THE VALUES OF THE     ;
*     PREDICTED VALUES.  WE DO THIS SO WE MAY ;
*     PLOT THE FITTED REGERSSION LINE ON TOP  ;
*     OF THE DATA BELOW, USING PROC PLOT.     ;
*                                         ;
*******************************************;

PROC REG;
  MODEL Y = X / P R CLM;
  OUTPUT OUT = OXYGEN2 P = PRED; RUN;

*******************************************;
*                                         ;
*     WE NOW USE PROC PLOT TO PLOT THE DATA   ;
*     AND THE FITTED LINE SUPERIMPOSED UPON   ;
*     IT.  WE SPECIFY THE DATA SET OXYGEN2,   ;
*     SINCE IT CONTAINS THE DATA THE THE      ;
```

```
*     PREDICTED VALUES "PRED".                ;
*                                             ;
*********************************************;

PROC PLOT DATA=OXYGEN2;
  PLOT Y*X = '*' PRED*X = 'P' / OVERLAY;
  TITLE2 'PLOT OF THE DATA AND THE FITTED REGRESSION';
  TITLE3 ' Y = OXYGEN CONSUMPTION, X = TEMPERATURE';
  TITLE4 'THE DATA POINTS - *, THE FITTED LINE = P'; RUN;
```

OUTPUT:

%%%

OXYGEN CONSUMPTION DATA 1

OBS	X	Y
1	-18	5.2
2	-15	4.7
3	-10	4.5
4	-5	3.6
5	0	3.4
6	5	3.1
7	10	2.7
8	19	1.8

%%%

OXYGEN CONSUMPTION DATA 2

Model: MODEL1

Dependent Variable: Y

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Prob>F
Model	1	8.74515	8.74515	308.933	0.0001
Error	6	0.16985	0.02831		
C Total	7	8.91500			

Root MSE	0.16825	R-square	0.9809

Dep Mean	3.62500	Adj R-sq	0.9778
C.V.	4.64135		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	T for H0: Parameter=0	Prob > |T|
INTERCEP	1	3.471422	0.06012323	57.738	0.0001
X	1	-0.087759	0.00499296	-17.576	0.0001

%%%

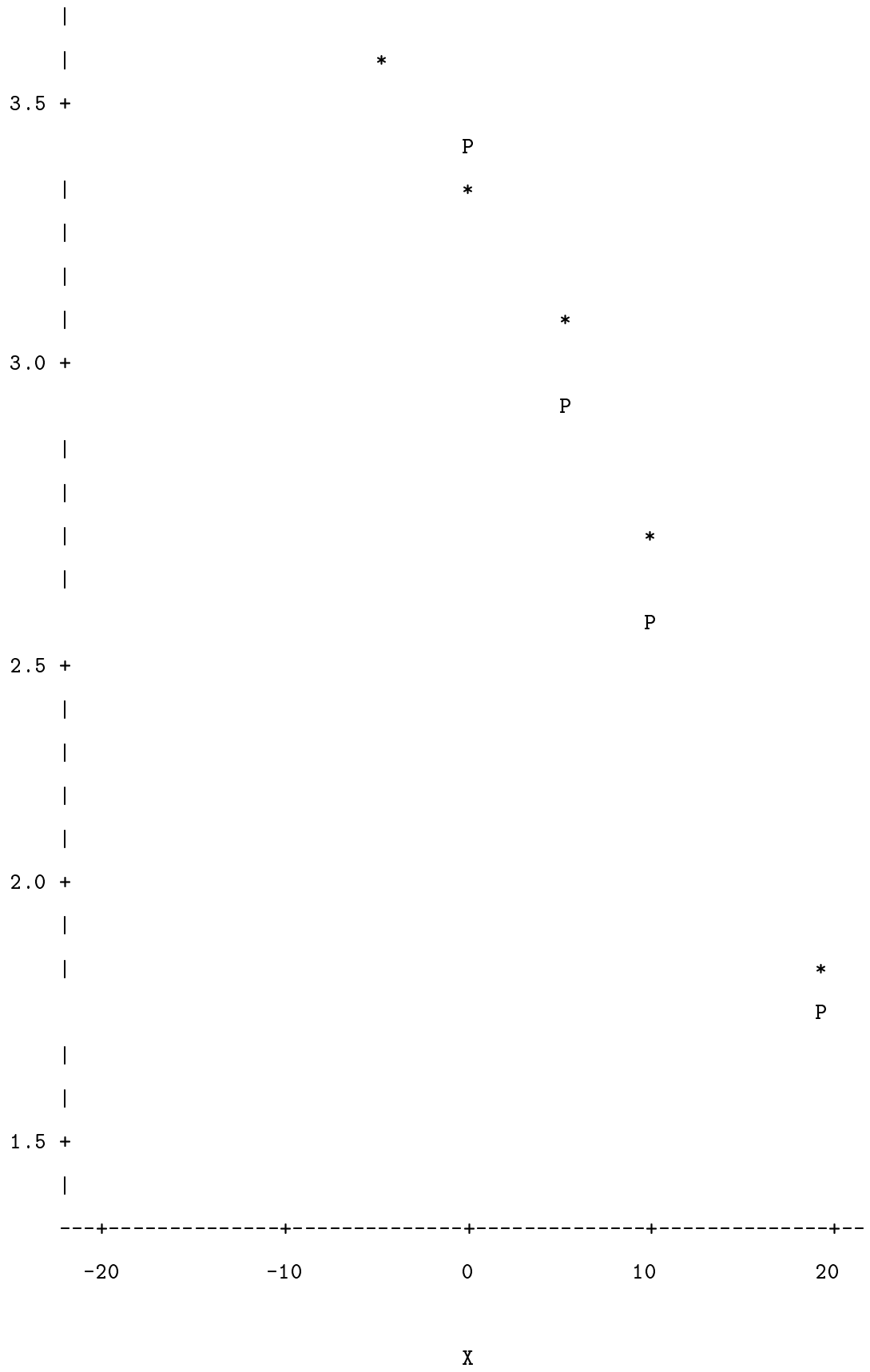
OXYGEN CONSUMPTION DATA 3

Obs	Dep Var Y	Predict Value	Std Err Predict	Lower95% Mean	Upper95% Mean	Residual	Std Err Residual
1	5.2000	5.0511	0.101	4.8049	5.2973	0.1489	0.135
2	4.7000	4.7878	0.089	4.5701	5.0055	-0.0878	0.143
3	4.5000	4.3490	0.072	4.1720	4.5261	0.1510	0.152
4	3.6000	3.9102	0.062	3.7593	4.0611	-0.3102	0.157
5	3.4000	3.4714	0.060	3.3243	3.6185	-0.0714	0.157
6	3.1000	3.0326	0.068	2.8653	3.1999	0.0674	0.154
7	2.7000	2.5938	0.084	2.3894	2.7983	0.1062	0.146
8	1.8000	1.8040	0.119	1.5117	2.0963	-0.00401	0.118

Obs	Student Residual	-2-1-0 1 2	Cook's D
1	1.104	|** |	0.339
2	-0.615	*| |	0.073
3	0.994	|* |	0.112

```
     4    -1.982  |   ***|      |     0.305
     5    -0.455  |      |      |     0.015
     6     0.438  |      |      |     0.019
     7     0.727  |      |*     |     0.086
     8    -0.034  |      |      |     0.001

Sum of Residuals                      0
Sum of Squared Residuals         0.1698
Predicted Resid SS (Press)       0.2645
```

```
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%

                          OXYGEN CONSUMPTION DATA                            4
                  PLOT OF THE DATA AND THE FITTED REGRESSION
                    Y = OXYGEN CONSUMPTION, X = TEMPERATURE
                   THE DATA POINTS - *, THE FITTED LINE = P

                     Plot of Y*X.     Symbol used is '*'.
                     Plot of PRED*X.  Symbol used is 'P'.

       Y |
         |
     5.5 +
         |
         |
         |     *
         |
               P
     5.0 +
         |
         |
                    P
         |          *
         |
     4.5 +                 *
         |
         |
                           P
         |
         |
     4.0 +
         |
                                   P
         |
```


3.5
3.0
2.5
2.0
1.5
-20
-10
0
10
20
X

EXAMPLE 2: Using SAS to compute sample correlation coefficients.

The SAS procedure `PROC CORR` is one of several procedures that computes sample correlation coefficients. Here, we use it to obtain the estimate of the correlation. The statement `VAR WING TAIL` tells SAS to compute the sample correlation coefficients among all pairs of variables in the list – here, there are only two, so only one correlation is possible. The procedure may thus be used to estimate correlations among all possible combinations of more than two variables.

The output contains some simple statistics for each variable in the list. "`PEARSON CORRELATION COEFFICIENTS`" are the sample correlations r_{XY}. Note that the results are presented in the form of a **matrix**. The first number in each "cell" of the matrix is the estimated correlation between the two relevant variables. The second number below it is the **p-value** for the test of $H_0 : \rho_{XY} = 0$ vs. $H_1 : \rho_{XY} \neq 0$. So, for a test of H_0 from the output, we compare this p-value to our chosen level of significance α.

PROGRAM:

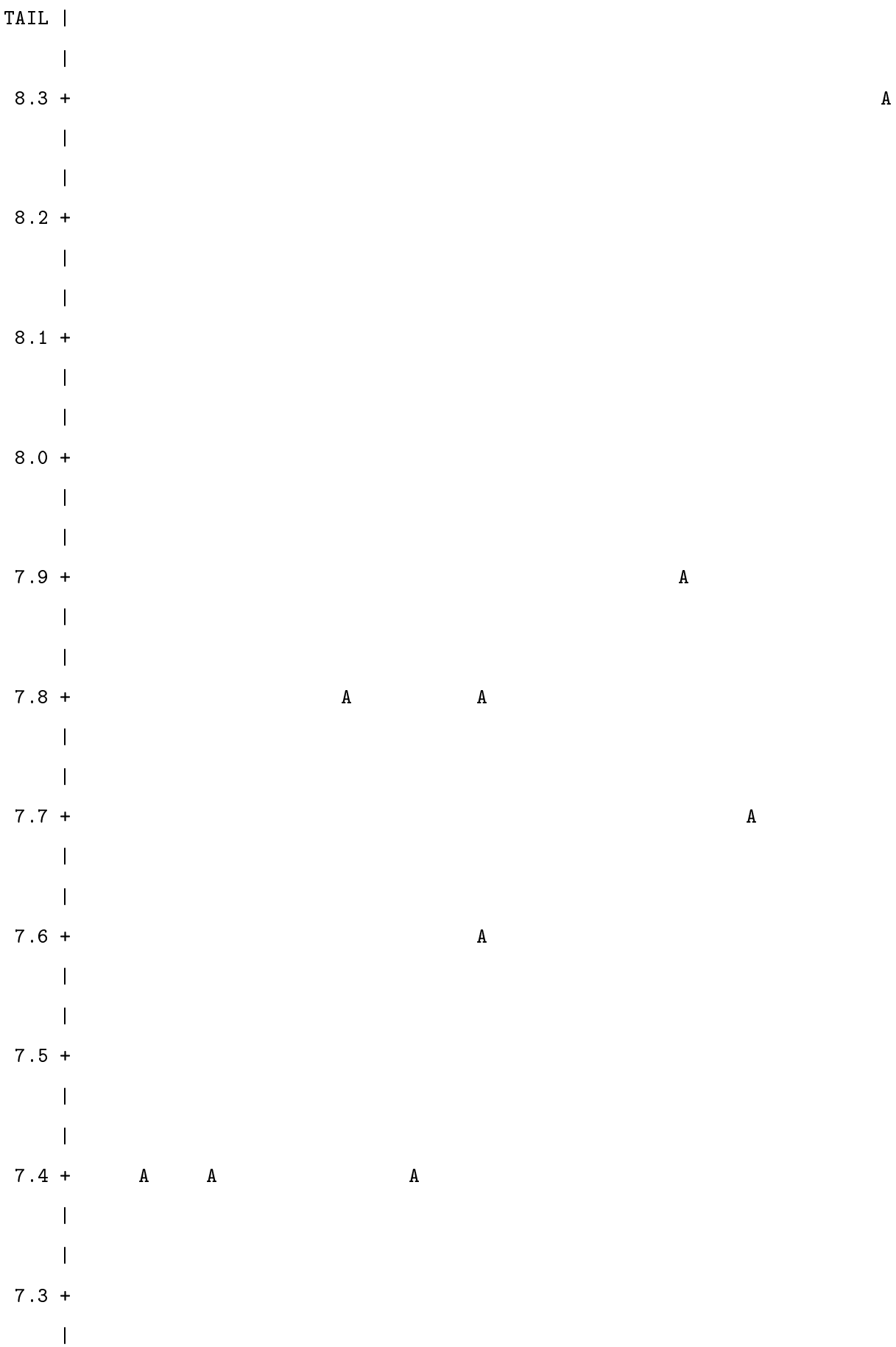
```
*********************************************;
*                                           ;
*      ST 511  USING PROC CORR TO COMPUTE   ;
*      THE SAMPLE CORRELATION COEFFICIENT   ;
*                                           ;
*********************************************;

OPTIONS LS=80 PS=59 NODATE;

*********************************************;
*                                           ;
*      APPLIED TO THE BIRD DATA             :
*                                           ;
*********************************************;

DATA BIRDS;  INPUT WING TAIL;
  CARDS;
10.4  7.4
10.8  7.6
```

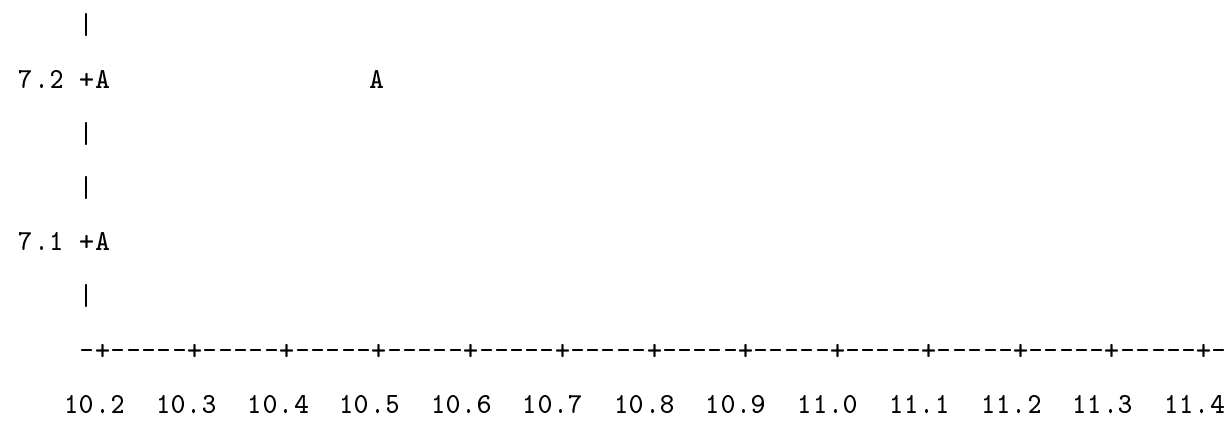
```
11.1  7.9
10.2  7.2
10.3  7.4
10.2  7.1
10.7  7.4
10.5  7.2
10.8  7.8
11.2  7.7
10.6  7.8
11.4  8.3
;

PROC PRINT;  TITLE 'WING AND TAIL LENGTHS FOR 12 BIRDS'; RUN;

*********************************************;
*                                           ;
*      PRODUCE A PLOT OF THE DATA WITH      ;
*      PLOT                                 ;
*                                           ;
*********************************************;

PROC PLOT;  PLOT TAIL*WING;
  TITLE2 'PLOT OF TAIL VERSUS WING LENGTHS'; RUN;

*********************************************;
*                                           ;
*      CALL PROC CORR TO COMPUTE THE        ;
*      CORRELATION BETWEEN WING AND TAIL    ;
*      LENGTH.  THIS PRODUCES THE VALUE OF  ;
*      THE CORRELATION AND THE PROBABILITY  ;
*      OF SEEING WHAT WE SAW OR SOMETHING   ;
*      MORE EXTREME UNDER HO: CORR = 0      ;
*      COMPARE THIS PROBABILITY TO ALPHA    ;
*      TO TEST AGAINST H1: CORR NOT = 0     ;
```

```
*                                              ;
***********************************************;


PROC CORR;  VAR WING TAIL;
  TITLE2 'CORRELATION ANALYSIS - BIRD DATA'; RUN;
```

OUTPUT:

```
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%

                       WING AND TAIL LENGTHS FOR 12 BIRDS                      1

                              OBS    WING    TAIL

                                1    10.4     7.4
                                2    10.8     7.6
                                3    11.1     7.9
                                4    10.2     7.2
                                5    10.3     7.4
                                6    10.2     7.1
                                7    10.7     7.4
                                8    10.5     7.2
                                9    10.8     7.8
                               10    11.2     7.7
                               11    10.6     7.8
                               12    11.4     8.3


%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%

                       WING AND TAIL LENGTHS FOR 12 BIRDS                      2
                        PLOT OF TAIL VERSUS WING LENGTHS

             Plot of TAIL*WING.  Legend: A = 1 obs, B = 2 obs, etc.
```


TAIL
8.3
8.2
8.1
8.0
7.9
7.8
7.7
7.6
7.5
7.4
7.3



%%%

WING AND TAIL LENGTHS FOR 12 BIRDS 3

CORRELATION ANALYSIS - BIRD DATA

Correlation Analysis

2 'VAR' Variables: WING TAIL

Simple Statistics

Variable	N	Mean	Std Dev	Sum	Minimum	Maximum
WING	12	10.68333	0.39505	128.20000	10.20000	11.40000
TAIL	12	7.56667	0.34989	90.80000	7.10000	8.30000

Pearson Correlation Coefficients / Prob > |R| under Ho: Rho=0 / N = 12

	WING	TAIL

```
WING        1.00000        0.87035
             0.0            0.0002

TAIL        0.87035        1.00000
             0.0002         0.0
```