STAT 8200 — Design and Analysis of Experiments for Research Workers — Lecture Notes

Basics of Experimental Design

Terminology

Response (Outcome, Dependent) Variable: (y) The variable who's distribution is of interest.

- Could be quantitative (size, weight, etc.) or qualitative (pass/fail, quality rated on 5 point scale).
  - I'll assume the former (easier to analyze).
  - Typically interested in mean of y and how it depends on other variables.
  - E.g., differences in mean response between varieties, drugs.

Explanatory (Predictor, Independent) Variables: (x's) Variables that explain (predict) variablility in the response variable.

- E.g., variety, rainfall, predation, soil type, subject age.

Factor: A set of related treatments or classifications used as an explanatory variable.

- Often qualitative (e.g., variety), but can be quantitative (0, 100, or 200 units fertilizer).

Treatment or Treatment Combination: A particular combination of the levels of all of the treatment factors.

Nuisance Variables: Other variables that influence the response variable but are not of interest.

- E.g., rainfall, level of predation, subject age.
- **Systematic bias** occurs when treatments differ with respect to a nuisance variable. If so, it becomes a **confounding variable** or **confounder**.

Experimental Units: The units or objects that are independently assigned to a specific experimental condition.

- E.g., a plot assigned to receive a particular variety, a subject assigned a particular drug.

Measurement Units: The units or objects on which distinct measurements of the response are made.

- Not necessarily same as exp'tal units. Distinction is *very* important!
- E.g., a plant or fruit within a plot.

Experimental Error: variation among experimental units that have received the same experimental conditions.

- The standard against which differences between treatments are to be judged.
- Treatment differences must be large *relative to* the variability we would expect in the absence of a treatment effect (experimental error) to infer the difference is real (statistical significance).
- If two varieties have mean yields that differ by d units, no way to judge how large d is unless we can estimate the experimental error (requires replication).

Elements of Experimental Design:

(1) **Randomization:** Allocate the experimental units to treatments at random (by chance).

- Makes the treatment groups *probabilistically alike* on *all* nuisance factors, thereby avoiding systematic bias.

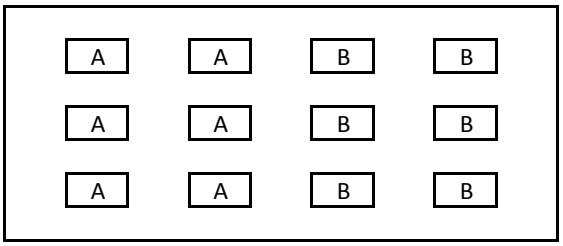
  Example 1: Two varieties (A,B) are to be compared with respect to crop yield. Suppose a crop row consisting of 100 plants is divided into plots of 10 plants. The two varieties are assigned to plots systematically, with variety A in every other plot:

A	B	A	B	A	B	A	B	A	B

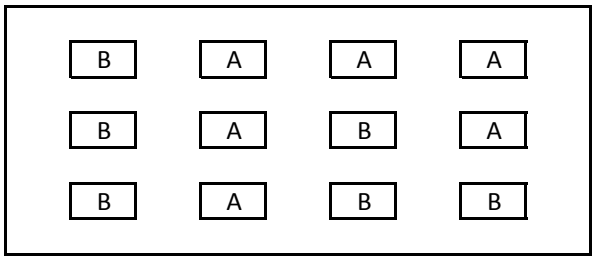
- Suppose there is a fertility gradient along this row. Then even if the varieties are equivalent, this we will observe better yield in variety B.

Randomization:

Example 2: A greenhouse has 12 benches on which plants of two varieties are to be grown. Suppose we assign as follows:



Whereas if we use a **Completely Randomized Design (CRD):**



- Randomization will tend to neutralize all nuisance variables.
- Also induces statistical independence among experimental units.

(2) **Replication:** Repeating the experimental run (one entire set of experimental conditions) using additional similar, independent, experimental units.

- Allows estimation of the experimental error without which treatment differences CANNOT be inferred.
- Increases the precision/power of the experiment.

BEWARE OF PSEUDO-REPLICATION!

Example 3: Suppose we randomize plots in a crop row to two treatments as so:

A	B	B	B	A	B	A	A	A	B

And we measure the size of all 10 plants in each plot.
$\Rightarrow$ we have 50 measurements per treatment.
$\Rightarrow$ we have 5 plots per treatment.

Q: *What's the sample size per treatment that determines our power to statistically distinguish varieties A and B?*

A: 5/treatment not 50. The experimental unit here is the plot, not the plant.

Replication:

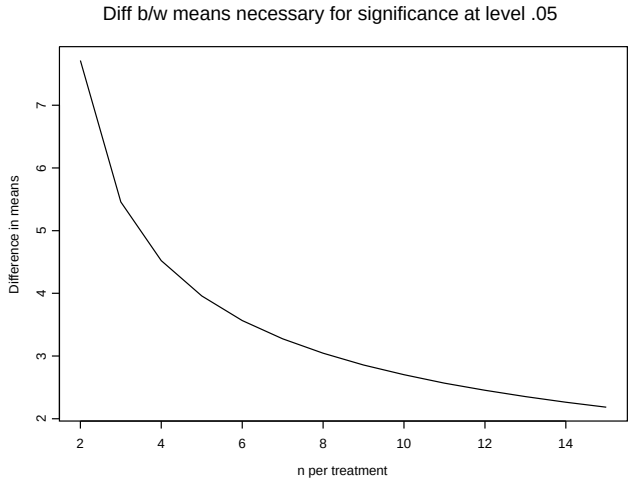
- Taking multiple measurements per experimental unit is called **subsampling** or **pseudo-replication.**
  - It is a useful means to reduce measurement error in characterizing the response at the experimental unit level.
  - If not interested in estimating this measurement error, easiest analysis is to average the subsamples in each experimental unit and analyze these averages as "the data".
  - How to allocate resources between experimental units and measurements units complicated, but generally more bang for adding experimental units over measurements units. Little gains to go beyond 2 or 3 m.u.s/e.u.

Replication:

What determines number of replicates?

- Available resources. Limitations on cost, labor, time, experimental material available, etc.
- Sources of variability in system and their magnitudes.
- Size of the difference to be detected.
- Required significance level ($\alpha = 0.05$?)
- Number of treatments

Effect of number of replicates/treatment on smallest difference in treatment means that can be detected at $\alpha = .05$ in a simple one-way design:



Replication:

- For a given desired probability of detecting a treatment effect of interest and under specific assumptions regarding the factors from the previous page, one can compute the sample size necessary.

- These calculations are most easily done with the help of sample size/power analysis software

  - SAS Analyst,
  - NQuery Advisor,
  - Online calculators like Russ Lenth's: http://www.math.uiowa.edu/∼ rlenth/Power/).

- They also require detailed assumptions which are best made on the basis of past or preliminary data.

  - This is the hard part.

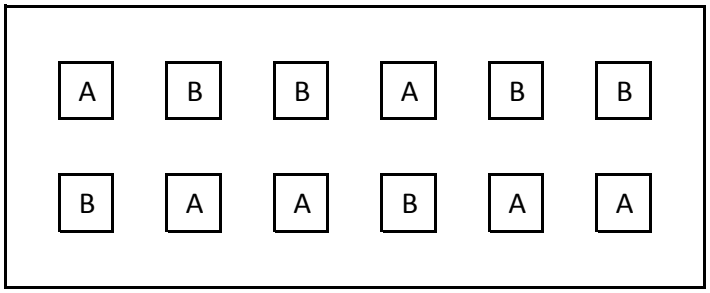
(3) **Blocking:** To detect treatment differences, we'd like the experimental units to be as similar (homogeneous) as possible except for the treatment received.

- We must balance this against practical considerations, and the goal of **generalizability.** For these reasons, the experimental units must be somewhat heterogeneous.
- The idea of blocking is to divide experimental units into homogeneous subgroups (or **blocks**) within which all treatments are observed. Then treatment comparisons can be made between similar units in the same block.
- Reduces experimental error and increases the precision (power, sensitivity) of an experiment.

Blocking:

Example: In our greenhouse example, our completely randomized assignment of varieties A, B happened to assign variety B to all three benches on the west end of the building. Suppose the heater is located on the west end so that there is a temperature gradient from west to east.

This identifiable source of heterogeneity among the experimental units is a natural choice of blocking variable. Solution: rearrange the benches as so and assign both treatments randomly within each column:



- Design above is called a **Randomized Complete Block Design (RCBD).**
- Blocking is often done within a given site. E.g., a large field is blocked to control for potential heterogeneity within the site. This is useful, but only slightly, as it results in only small gains relative to a CRD.
- However, if a known source of variability exists where there is likely to be a gradient in the characteristics of the plots, then blocking within a site is definitely worthwhile.

Blocking:

- In the presence of a gradient, plots should be oriented as follows*:

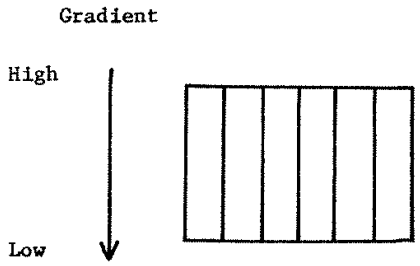


Figure 2.6 Orientation of plots with respect to a gradient.

and, if blocking is done:

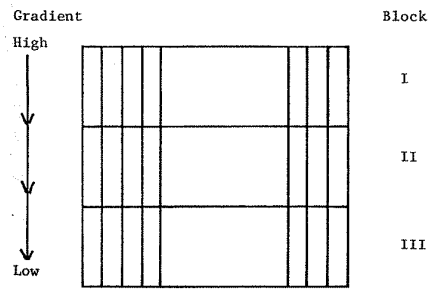


Figure 3.3 Placement of blocks with respect to a gradient.

- Placement of blocks should take into account physical features of the site:

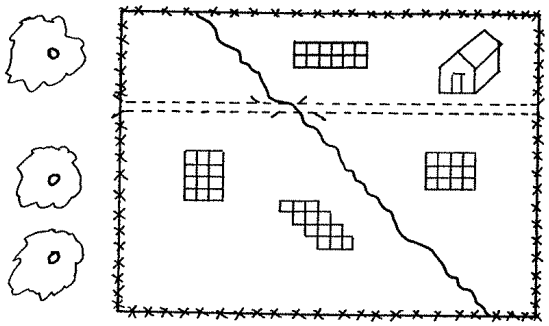


Figure 3.4 Placement of blocks to accommodate physical features of the research site.

* Plots on this page from Petersen (1994).

Blocking:

There are a number of factors that often create heterogeneity in the experimental units that can, and typically should, form the basis of blocks:

- Region.
- Time (season, year, day).
- Facility (e.g., if multiple greenhouses are to be used, or multiple labs to take measurements, if patients recruited from multiple clinics).
- Personnel to conduct the experiment.

Block effects are typically assumed not to interaction with treatment effects.

- E.g., while we allow that a region (block) effect might raise the yield for all varieties, we assume that differences in mean yield between varieties are the same in each region.
- If each treatment occurs just once in each block, this assumption MUST be made in order to analyze the data and the analysis will likely be wrong if this assumption is violated.
- If treatment effects are expected to differ across blocks,
  - use ≥ 2 reps of every treatment within each block, and
  - consider whether the blocking factor is more appropriately considered to be a treatment factor.

(4) **Use of Factorial Treatment Structures.** If more than one treatment factor is to be studied, generally better to study them in combination, rather than separately.

Factorial or **crossed** treatment factors:

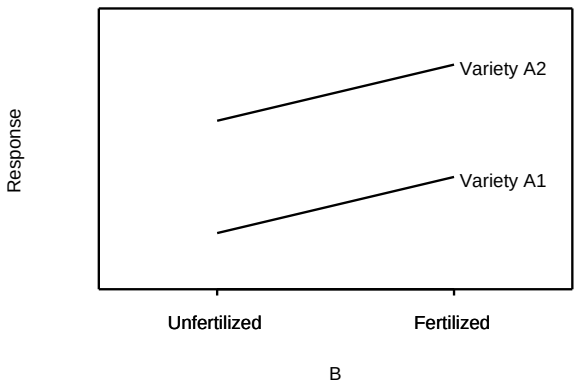
		Fertilized? (B)	
		No (B1)	Yes (B2)
Variety (A)	A1	A1,B1	A1,B2
	A2	A2,B1	A2,B2

- Increases generalizability, efficiency.
- Allows **interactions** to be detected.

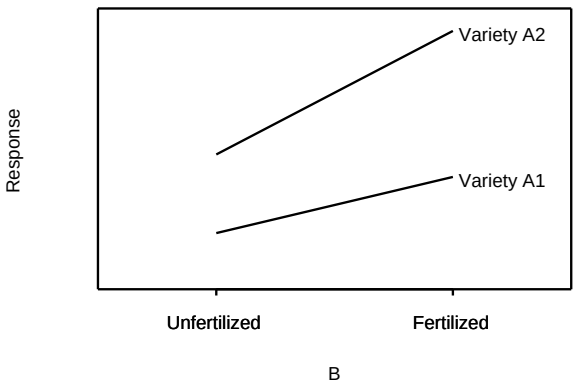
Factorial Treatment Structure and Interactions:

- An interaction occurs when the effect of one treatment factor differs depending on the level at which the other treatment factor(s) are set.

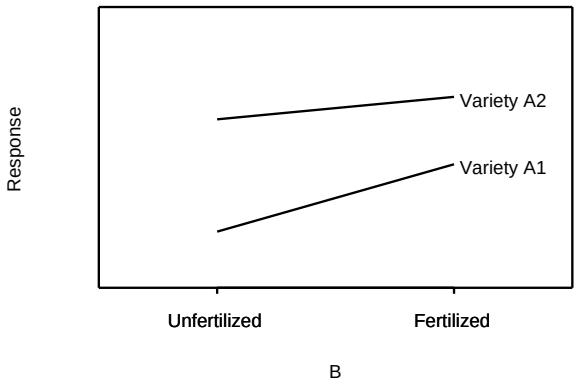
No interaction between A and B



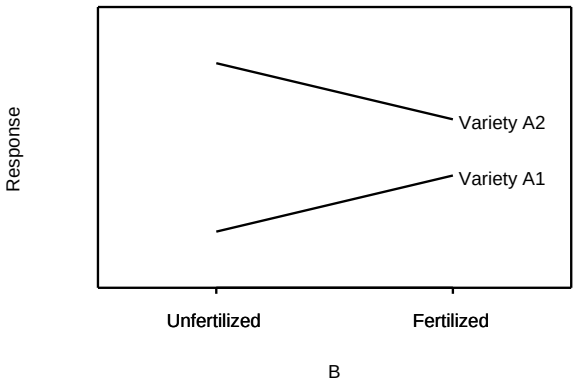
Quantitative synergistic interaction



Quantitative inhibitory interaction



Qualitative interaction between A and B



(5) **Balance:** A balanced experiment is one in which the replication is the same under each level of each experimental factor.

E.g., an experiment with three treatment groups each with 10 experimental units is balanced; an experiment with three treatment groups of sizes 2, 18, and 10 is unbalanced.

Occasionally, the response in a certain treatment or comparisons with a certain treatment are of particular interest. If so, then extra replication in that treatment may be justified. Otherwise, it is desirable to achieve as much balance as possible subject to practical constraints.

- Increases power of experiment.
- Simplifies statistical analysis.

(6) **Limiting Scope/Use of Sequential Experimentation:** Large experiments with many factors and many levels of the factors are

- hard to perform,
- hard to analyze, and
- hard to interpret.

- If the effects of several factors are of interest, best to do several small experiments and build up to an understanding of the entire system.
- Can either use several factors each at a small number (e.g., 2) of levels, or can do sequential experiments each examining a subset of factors (2 or 3).

(7) **Adjustment for Covariates:** Nuisance variables other than the blocking factors that affect the response are often measured and compensated for in the analysis.

- E.g., measure and adjust for rainfall differences in statistical analysis (can't block by rainfall).
- Avoids systematic bias.
- Increases the precision of the experiment.

(8) **Use of a comparison group.**

- (Almost) Always!
- A null or placebo treatment as a "control" appropriate in most (not all) contexts.
- Alternatively or additionally, a standard treatment can be used for comparison.
  - E.g., a variety in long-term use that stays stable over time.

(9) **Use of multiple sizes of experimental units.** E.g., split-plot experiments.

- Adds flexibility/practicality to an experiment.
- Allows investigation of additional factors.

Steps in Experimentation:

- Statement of the objectives
- Identify sources of variation
  - Selection of treatment factors and their levels
  - Selection of experimental units, measurement units
  - Selection of blocking factors, covariates.
- Selection of experimental design
- Consideration of the data to be collected (what variables, how measured)
- (Perhaps) run a pilot experiment
- Outline the statistical analysis
- Choice of number of replications
- Conduct the experiment
- Analyze data and interpret results
- Write up a thorough, readable, stand-alone report summarizing the research

Structures of an Experimental Design:

Treatment Structure: The set of treatments, treatment combinations, or populations under study. (The selection and arrangement of treatment factors.)

Design Structure: The way in which experimental units are grouped together into homogeneous units (blocks).

These structures are combined with a method of randomization to create an experimental design.

Types of Treatment Structures:

(1) One-way

(2) n-way Factorial. Two or more factors combined so that every possible treatment combination occurs (factors are **crossed**).

(3) n-way Fractional Factorial. A specified fraction of the total number of possible combinations of n treatments occur (e.g., Latin Square).

Types of Design Structures:

(1) Completely Randomized Designs. All experimental units are considered as a single homogeneous group (no blocks). Treatments assigned completely at random (with equal probability) to all units.

(2) Randomized Complete Block Designs. Experimental units are grouped into homogeneous blocks within which each treatment occurs c times ($c = 1$, usually).

(3) Incomplete Block Designs. Fewer than the total number of treatments occur in each block.

(4) Latin Square Designs. Blocks are formed in two directions with n experimental units in each combination of the row and column levels of the blocks.

(5) Nested (Hierarchical) Design Structures. The levels of one blocking factor are superficially similar but not the same for different levels of another blocking factor.

> For example, suppose we measure tree height growth in 5 plots on each of 3 stands of trees. The plots are nested within stands (rather than crossed with stands). This means that plot 1 in stand 1 is not the same as plot 1 in stand 2.

Some Examples

Example 1: A one-way layout (generalization of the two sample t-test)
An egg producer is trying to develop a low cholesterol egg. 40 female chicks are randomly divided into four groups, each of which is reared on a different type of feed. Otherwise, the chicks are treated identically. When the hens reach laying age, 10 eggs from each hen are collected and the average cholesterol level per egg is measured.

Treatment Structure:
Design Structure:
Response variable:
Explanatory variable:
Measurement units:
Experimental units:

Example 2: A randomized complete block design (generalization of the design underlying the paired t-test)

Now suppose that the 40 chicks used for the experiment came from 5 different broods, 8 chicks per brood. We are concerned that chicks from different broods might have different characteristics. Here it is appropriate to use brood as a blocking factor. Design: randomly assign the 4 feed types separately for each brood group, such that two chicks from each brood get each type of feed.

Treatment Structure:
Design Structure:
Explanatory variables:

Example 3: Balanced Incomplete Block Design
Now suppose that only 12 chicks are available for the study and they were drawn from four different broods, 3/brood. In this situation all four treatments cannot occur for each brood.

Design:

	Brood			
Feed	1	2	3	4
1	y_{11}	y_{12}	•	y_{14}
2	•	y_{22}	y_{23}	y_{24}
3	y_{31}	y_{32}	y_{33}	•
4	y_{41}	•	y_{43}	y_{44}

Balanced because each pair of treatments occurs together the same number of times.

Treatment Structure:
Design Structure:

Example 4: Latin Square

An investigator is interested in the effects of tire brand (brands A,B,C,D) on tread wear. Suppose that he wants 4 observations per brand.

Latin square design:

	Wheel Position			
Car	1	2	3	4
1	A	B	C	D
2	B	C	D	A
3	C	D	A	B
4	D	A	B	C

In this design 4 replicates are obtained for each treatment without using all possible wheel position $\times$ tire brand combinations. (Economical design)

Treatment Structure:
Design Structure:

Considerations when Designing an Experiment

- Experimental design should give unambiguous answers to questions of interest.
- Experimental design should be "optimal". That is, it should have more power (sensitivity) and estimate quantities of interest more precisely than other designs.
- Experiment should be of a manageable size. Not too big to perform, analyze, or interpret.
- Objectives of the experiment should be clearly defined.
  - What questions are we trying to answer?
  - Which questions are most important?
  - What populations are we interested in generalizing to?
- Appropriate response and explanatory variables must be determined and nuisance variables should be identified.
  - What levels of the treatment factors will be examined?
  - Should there be a control group?
  - Which nuisance variables will be measured?
- The practical constraints of the experiment should be identified.
  - How much time, money, personnel, raw material will it take, and how much do I have?
  - Is it practical to assign and apply the experimental conditions to experimental units as called for by the experimental design.

- Identify blocking factors.
  - In what respects that can be expected to be relevant to the response variable are experimental units dissimilar?
  - Will the experiment be conducted over several days (weeks, seasons, years)? Then block by day (week/season/year).
  - Will the experiment involve several experimenters? (e.g., several different crews to measure trees, lab technicians, etc.). Then block by experimenter.
  - Will the experiment involve units from different locations, or units that dispersed geographically/spatially? Then form blocks by location, or as groups of units that are near one another.
- Statistical analysis of the experiment should be planned in detail to meet the objectives of the experiment.
  - What model will be used?
  - How will nuisance variables be accounted for?
  - What hypotheses will be tested?
  - What "effects" will be estimated?
- Experimental design should be economical, practical, timely.

All of the experiments in this course will be analyzed using special cases of the **general linear model**:

$$y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_k x_{ik} + e_i$$

$\beta_0, \beta_1, \ldots, \beta_k$ are unknown parameters to be estimated from the data and e_i is a random variable (the error term) accounting for all of the variability in the y's not explained in the systematic component of the model.

- $e_1, \ldots, e_n$ are assumed to be independent, each with mean 0, and common variance σ^2.

Regression: x's are continuous variables

ANOVA: x's are indicator variables representing the levels of (usually qualitative) factors

ANCOVA: x's are mixed indicator and continuous variables

Preliminaries (Review, hopefully)

Basic Statistical Concepts

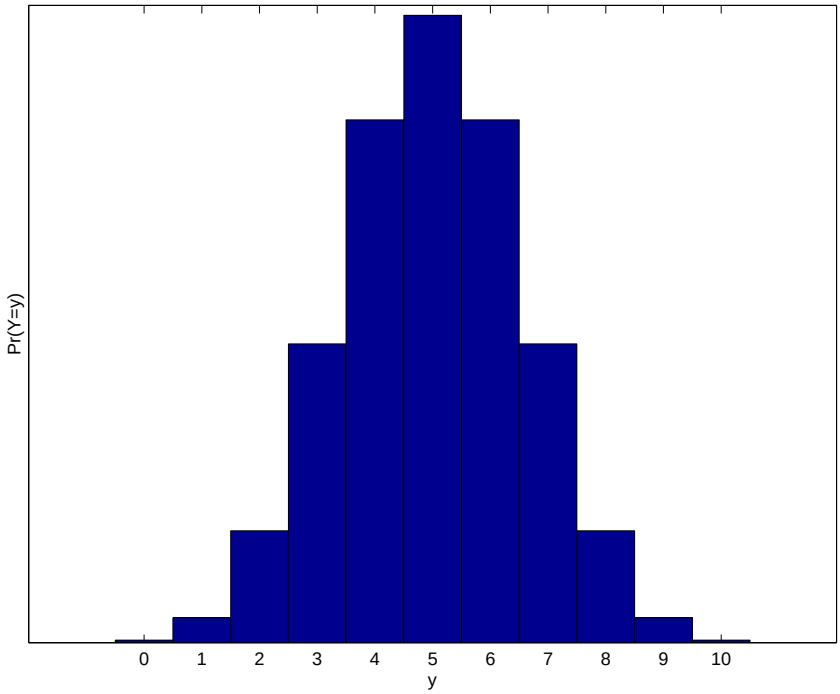
Measures of location (center): **mean, median, mode, mid-range.**

Measures of dispersion (spread): **variance (std. deviation), range, inter-quartile range**.

Suppose we're interested in a random variable Y (e.g., Y =time of operation of a randomly selected PC hard drive until it fails).

This random variable has a **distribution** (sometime its value is small, sometimes its large) which is a **probability distribution** (I can't say for sure when Y will take value y, but I know the probability that $Y = y$, or the probability that Y will take a value between y_1 and y_2).

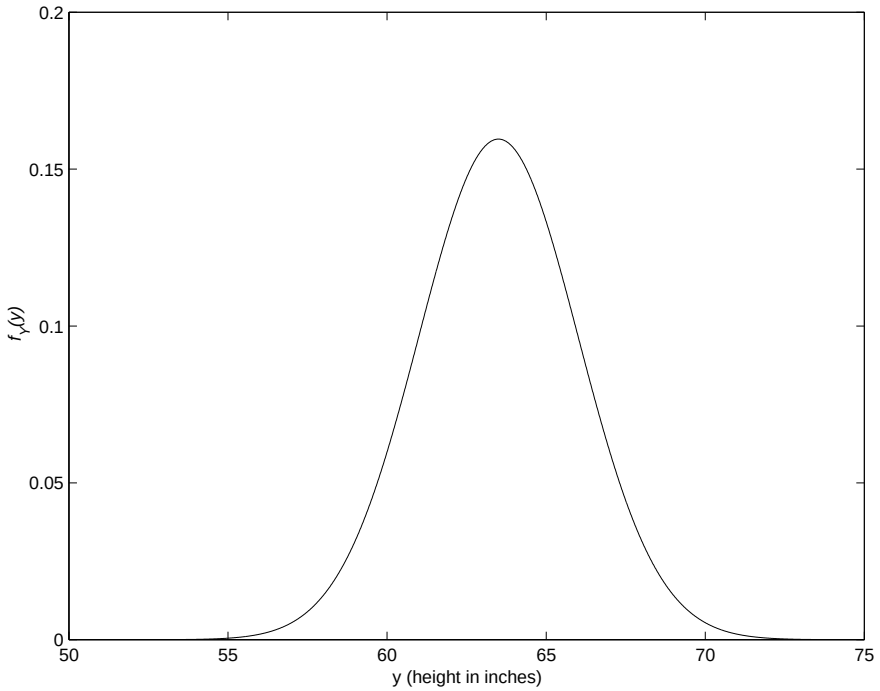
If Y is discrete (e.g., Y =number of heads out of ten fair coin flips) here's an example of probability distribution



- We will only be concerned with continuous random variables or discrete random variables which take on enough possible values so that they can be treated as continuous.

For a continuous random variable, the probability that it takes on any particular value is 0, so plotting $\Pr(Y = y)$ against y doesn't describe a continuous probability distribution. Instead we use the **probability density function** $f_Y(y)$ which gives (roughly) the probability that Y takes a value "close to" y.

E.g., assuming Y =adult female height in the U.S. is normally distributed with mean 63.5 in. and s.d. 2.5 in., here's the distribution of female heights



- Essentially, most statistical problems boil down to questions about what that distribution looks like.

The two most important aspects of what the distribution looks like are

1. where it is located (mean, median, mode, etc.)
2. how spread out it is (variance, std. dev., range, etc.)

The mean, or **expected value** (often written as μ), of a probability distribution is the mean of a random variable with that distribution, taken over the entire population.

The variance (often written as σ^2) of a probability distribution is the variance of a r.v. with that distribution, taken over the entire population.

- The population standard deviation σ is simply the (positive) square root of the population variance: $\sigma = \sqrt{\sigma^2}$.

A few useful facts about expected values and population variances:

Let c represent some constant, and let X, Y denote two random variables with population means μ_X, μ_Y and population variances σ_X^2, σ_Y^2, respectively. Then:

1. $\mathrm{E}(c) = c$.
2. $\mathrm{E}(cY) = c\mathrm{E}(Y) = c\mu_Y$.
3. $\mathrm{E}(X + Y) = \mathrm{E}(X) + \mathrm{E}(Y) = \mu_X + \mu_Y$.
4. $\mathrm{E}(XY) = \mathrm{E}(X)\mathrm{E}(Y)$ if (and only if) X and Y are statistically independent (i.e., knowing something about X doesn't tell you anything about Y).
5. $\mathrm{var}(c) = 0$.
6. $\mathrm{var}(cX) = c^2\mathrm{var}(X) = c^2\sigma_X^2$.
   Notice that this implies $\mathrm{Pop.S.D.}(cX) = c\mathrm{Pop.S.D.}(X) = c\sigma_X$.
7. $\mathrm{var}(X \pm Y) = \mathrm{var}(X) + \mathrm{var}(Y) \pm 2\mathrm{cov}(X, Y)$ where $\mathrm{cov}(X, Y) = \mathrm{E}[(X - \mu_X)(Y - \mu_Y)]$ is the **covariance** between X and Y, and measures the strength of the linear association between them. Note that
   - $\mathrm{cov}(X, Y) = \mathrm{cov}(Y, X)$
   - $\mathrm{cov}(cX, Y) = c\mathrm{cov}(X, Y)$, and
   - if X, Y are independent, then $\mathrm{cov}(X, Y) = 0$
8. This last property implies that $\mathrm{var}(X \pm Y) = \mathrm{var}(X) + \mathrm{var}(Y)$ if X, Y are independent of each other.

To find out about a probability distribution for Y (how Y behaves over the entire population), we typically examine behavior in a randomly selected subset of the population — a sample: $(y_1, \ldots, y_n)$.

We can compute sample quantities corresponding to the population quantities of interest:

Sample mean: $\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$ estimates the pop. mean μ_Y.

Sample variance: $s_Y^2 = \frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y})^2$ estimates σ_Y^2.

Digression on Notation:

Recall the summation notation: $\sum_{i=1}^{k} x_i$ is shorthand for $x_1 + x_2 + \cdots + x_k$.

Often we'll use a double sum when dealing with a variable that has two indexes. E.g.,

Subject	Trt 1	Trt 2	Trt 3
1	y_{11}	y_{21}	y_{31}
2	y_{12}	y_{22}	y_{32}
$\vdots$	$\vdots$	$\vdots$	$\vdots$
n	y_{1n}	y_{2n}	y_{3n}

Here, y_{ij} denotes the response of subject j in group i in an experiment in which each of n subject receives all three treatments.

- This is an example of what's known as a **crossover design.**

If we sum over one index, then we get either a group total or a subject total depending upon which index we choose:

Sum of obs's in group i:

$$\sum_{j=1}^{n} y_{ij} = y_{i1} + \cdots + y_{in} =$$

Sum of obs's obtained on subject j:

$$\sum_{i=1}^{3} y_{ij} = y_{1j} + y_{2j} + y_{3j} =$$

Sum of all observations over groups and subjects (grand total):

$$\begin{aligned}\sum_{i=1}^{3}\sum_{j=1}^{n} y_{ij} &= \sum_{i=1}^{3}(y_{i1} + \cdots + y_{in}) \\ &= (y_{11} + \cdots + y_{1n}) + (y_{21} + \cdots + y_{2n}) + (y_{31} + \cdots + y_{3n}) \\ &=\end{aligned}$$

- A notational convenience we'll often use is to replace an index by a '$\cdot$' to indicate summation over that index.
- If we add a bar, we indicate that we average over the index that has been replaced by '$\cdot$':

$$\bar{y}_{i\cdot} = \frac{1}{n}\sum_{j=1}^{n} y_{ij} = \frac{1}{n} y_{i\cdot} = \text{mean of obs's in group } i$$

$$\bar{y}_{\cdot j} = \frac{1}{3}\sum_{i=1}^{3} y_{ij} = \frac{1}{3} y_{\cdot j} = \text{mean of obs's for subject } j$$

$$\bar{y}_{\cdot\cdot} = \frac{1}{3n}\sum_{i=1}^{3}\sum_{j=1}^{n} y_{ij} = \frac{1}{3n} y_{\cdot\cdot} = \text{grand mean of all obs's}$$

There are a few probability distributions that are particularly important in this course:

- Normal (Gaussian)
- t distribution
- F distribution
- χ^2 (chi-square) distribution

The normal distribution is, by now, well known to you. The other three all arise by considering the distribution of certain functions of normally distributed random variables.

The Normal Distribution

$Y \sim N(\mu, \sigma^2)$ means that Y is a random variable with a normal distribution with parameters $\mu(= \mathrm{E}(Y))$ and $\sigma^2(= \mathrm{var}(Y))$.

That is, the r.v. Y has density

$$f_Y(y) = \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{-(y-\mu)^2}{2\sigma^2}}, \quad -\infty < y < \infty,$$

where $-\infty < \mu < \infty$, and $\sigma^2 > 0$.

If $Y \sim N(\mu, \sigma^2)$ then $Z = \frac{Y-\mu}{\sigma} \sim N(0, 1)$ (standard normal).

The Central Limit Theorem is the main reason that the normal distribution plays such a major role in statistics:

In plain English, the CLT says (roughly speaking) that any random variable that can be computed as a sum or mean of n independent, identically distributed (or iid, for short) random variables (e.g., measurements taken on a simple random sample) will have a distribution that is well approximated by the normal distribution as long as n is large.

How large must n be? Depends on the problem (on the distribution of the Y's).

Importance?

- Many sample quantities of interest (including means, totals, proportions, etc.) are sums of iid random variables, so the CLT tells us that these quantities are approximately normally distributed.
- Furthermore, the iid part is often not strictly necessary, so there are a great many settings in which random variables that are sums or means of other random variables are approximately normally distributed.
- The CLT also suggests that many elementary random variables themselves (i.e., measurements on experimental units or sample members rather than sums or means of such measurements) will be approximately normal because it is often plausible that the uncertainty in a random variable arises as the sum of several (often many) independent random quantities.
    - E.g., reaction time depends upon amount of sleep last night, whether you had coffee this morning, did you happen to blink, etc.

The Chi-square Distribution

If $Y_1, \ldots, Y_n \overset{iid}{\sim} N(0,1)$, then

$$X = Y_1^2 + \cdots + Y_n^2$$

has a (central) chi-square distribution with d.f. $= n$, the number of squared normals that we summed up. To denote this we write $X \sim \chi^2(n)$.

Important Result:

If $Y_1, \ldots, Y_n \sim N(\mu, \sigma^2)$ then

$$\frac{SS}{\sigma^2} = \frac{\sum_{i=1}^n (Y_i - \bar{Y})^2}{\sigma^2} \sim \chi^2(n-1)$$

- Here, SS stands for "sum of squares" or, more precisely, "sum of squared deviations from the mean."

(Student's) t Distribution

If $Z \sim N(0,1)$, $X \sim \chi^2(n)$, and Z and X are independent, then the r.v.

$$t = \frac{Z}{\sqrt{X/n}}$$

has a (central) t distribution with d.f. $= n$ and we write $t \sim t(n)$.

F Distribution

If $X_1 \sim \chi^2(n_1)$, $X_2 \sim \chi^2(n_2)$, and X_1 and X_2 are independent, then the r.v.

$$F = \frac{X_1/n_1}{X_2/n_2}$$

has a (central) F distribution with n_1 numerator d.f., and n_2 denominator d.f. We write $F \sim F(n_1, n_2)$.

Note that the square of a random variable with a $t(n)$ distribution has an $F(1, n)$ distribution:

If

$$t = \frac{Z}{\sqrt{X/n}} \sim t(n),$$

then

$$t^2 = \frac{Z^2/1}{X/n} \sim F(1, n)$$

Testing for equality of 2 pop. means

Suppose that two temporary employees are used for clerical duties and it is desired to hire one of them full time.

To decide which person to hire, the boss decides to determine whether there's a difference in the typing speed between the two workers.

Completely Randomized design:

- 20 letters need to be typed.
- 10 are randomly selected and assigned to Empl 1, the remaining 10 assigned to Empl 2.

Response: Y =words/minute on each of the letters.

Treatment Structure: One-way, because there's a single treatment factor, employee, with two levels (Empl 1, Empl 2).

Design structure: completely randomized

Experimental design: one-way layout

Experimental unit: the letters (10 per treatment).

Assumptions:

Two independent samples. Let y_{ij} = words/minute for Empl i $(i = 1, 2)$ when typing letter j $(j = 1, \ldots, 10)$.

From Empl 1:

$$y_{11}, \ldots, y_{1,10} \overset{iid}{\sim} N(\mu_1, \sigma_1^2)$$

From Empl 2:

$$y_{21}, \ldots, y_{2,10} \overset{iid}{\sim} N(\mu_2, \sigma_2^2)$$

and suppose that the observed result is $\bar{y}_1 = 65$, $\bar{y}_2 = 60$.

Hypothesis of interest:

$$H_0 : \mu_1 = \mu_2 \qquad \text{versus} \qquad H_A : \mu_1 \neq \mu_2$$

or, equivalently,

$$H_0 : \mu_1 - \mu_2 = 0 \qquad \text{versus} \qquad H_A : \mu_1 - \mu_2 \neq 0$$

How we test this depends upon what we know (or what we are willing to assume) about σ_1^2, σ_2^2.

In any case it makes sense to examine $\bar{y}_1 - \bar{y}_2$. If this quantity is "far from zero" or "large" in absolute value then we reject H_0.

- "Large" in absolute value is relative to the natural variability that this quantity would have if the null hypothesis were true. That is, $\bar{y}_1 - \bar{y}_2$ should be far from 0 relative to its standard error.
- **standard error**: the estimated standard deviation of a statistic.
  - E.g., if $\bar{x}$ is the mean of n independent random variables each with variance σ^2 then s.e.$(\bar{x}) = \hat{\sigma}/\sqrt{n}$ where $\hat{\sigma}$ is an estimate of σ (usually the sample standard deviation).

The standard error of $\bar{y}_1 - \bar{y}_2$ is $\sqrt{\hat{\text{var}}(\bar{y}_1 - \bar{y}_2)}$.

- Here, '^' means "estimated".

Since $\bar{y}_1, \bar{y}_2$ are the means of two independent samples, these quantities are independent, so

$$\begin{aligned} \text{var}(\bar{y}_1 - \bar{y}_2) &= \text{var}(\bar{y}_1) + \text{var}(\bar{y}_2) \\ &= \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2} \end{aligned}$$

No matter what we assume about σ_1^2 and σ_2^2, an appropriate test statistic is

$$\frac{\bar{y}_1 - \bar{y}_2}{\text{s.e.}(\bar{y}_1 - \bar{y}_2)} = \frac{\bar{y}_1 - \bar{y}_2}{\sqrt{\hat{\text{var}}(\bar{y}_1 - \bar{y}_2)}}.$$

- However, what the best way is to estimate $\text{var}(\bar{y}_1 - \bar{y}_2)$ depends on our assumptions on σ_1^2 and σ_2^2, leading to distinct test statistics.

Cases:

Case 1: σ_1^2, σ_2^2 both known (may or may not be equal);

For example, we happen to know that the variance of words/minute for Empl 1 is $\sigma_1^2 = 121$, and the variance of words/minute for Empl 2 is $\sigma_2^2 = 81$. Then

$$\text{var}(\bar{y}_1 - \bar{y}_2) = \frac{121}{10} + \frac{81}{10} \qquad \text{(there's nothing to estimate)}$$

$$\Rightarrow \qquad \text{s.e.}(\bar{y}_1 - \bar{y}_2) = \sqrt{\frac{121 + 81}{10}}$$

from which we obtain our test statistic:

$$z = \frac{\bar{y}_1 - \bar{y}_2}{\text{s.e.}(\bar{y}_1 - \bar{y}_2)} = \frac{\bar{y}_1 - \bar{y}_2}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$
$$= \frac{65 - 60}{\sqrt{\frac{121+81}{10}}} = 1.11$$

The $p-$value for this test tells us how much evidence our result provides against the null hypothesis: $p =$ the probability of getting a result at least as extreme as the one obtained.

$$p = \Pr(z \leq -1.11) + \Pr(z \geq 1.11) = 2\Pr(z \geq 1.11) = .267$$

- Comparison of the $p-$value against a pre-selected significance level, α (.05, say), leads to a conclusion. In this case, since $.267 > .05$ we fail to reject H_0 based on a significance level of .05. (There's insufficient evidence to conclude that the mean typing speeds of the two workers differ.)

Case 2: σ_1^2, σ_2^2 unknown, but assumed equal ($\sigma_1^2 = \sigma_2^2 = \sigma^2$, say).

$$\text{var}(\bar{y}_1 - \bar{y}_2) = \frac{\sigma^2}{n_1} + \frac{\sigma^2}{n_2} = \sigma^2 \left(\frac{1}{n_1} + \frac{1}{n_2} \right)$$

Two possible estimators come to mind:

$$s_1^2 = \text{sample variance from } 1^{\text{st}} \text{ sample}$$
$$s_2^2 = \text{sample variance from } 2^{\text{nd}} \text{ sample}$$

Under the assumption that $\sigma_1^2 = \sigma_2^2 = \sigma^2$, both are estimators of the same quantity, σ^2, each based on only a portion of the total number of relevant observations available.

Better idea: combine these two estimators by taking their (weighted) average:

$$\hat{\sigma}^2 = s_P^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}$$

$$\Rightarrow \qquad \text{s.e.}(\bar{y}_1 - \bar{y}_2) = \sqrt{\hat{\sigma}^2 \left(\frac{1}{n_1} + \frac{1}{n_2} \right)} = \sqrt{s_P^2 \left(\frac{1}{n_1} + \frac{1}{n_2} \right)}$$

$$\Rightarrow \qquad \text{test stat.} = t = \frac{\bar{y}_1 - \bar{y}_2}{\sqrt{s_P^2 \left(\frac{1}{n_1} + \frac{1}{n_2} \right)}}$$
$$\sim t(n_1 + n_2 - 2)$$

Suppose we calculate $s_1^2 = 110.2$, $s_2^2 = 89.4$ from the data. Then

$$s_P^2 = \frac{(10 - 1)110.2 + (10 - 1)89.4}{10 + 10 - 2} = 99.8, \quad t = \frac{65 - 60}{\sqrt{99.8 \left(\frac{1}{10} + \frac{1}{10} \right)}} = 1.12$$

$$\Rightarrow \qquad p = \Pr(t_{18} \leq -1.12) + \Pr(t_{18} \geq 1.12) = 2\Pr(t_{18} \geq 1.12) = .277$$

Case 3: σ_1^2, σ_2^2 both unknown but assumed different.

$$\Rightarrow \qquad \text{var}(\bar{y}_1 - \bar{y}_2) = \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}$$

$$\Rightarrow \qquad \text{s.e.}(\bar{y}_1 - \bar{y}_2) = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

$$\Rightarrow \qquad \text{test stat.} = t = \frac{\bar{y}_1 - \bar{y}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} \dot{\sim} t(\nu)$$

where

$$\nu = \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\frac{(s_1^2/n_1)^2}{n_1 - 1} + \frac{(s_2^2/n_2)^2}{n_2 - 1}} \qquad \text{(round down to nearest integer)}$$

- Here, $\dot{\sim}$ means "is approximately distributed as".

For our results,

$$t = \frac{65 - 60}{\sqrt{\frac{110.2}{10} + \frac{89.4}{10}}} = 1.12, \quad \nu = \frac{\left(\frac{110.2}{10} + \frac{89.4}{10}\right)^2}{\frac{(110.2/10)^2}{10-1} + \frac{(89.4/10)^2}{10-1}} = 17.8$$

$$\Rightarrow \qquad p \approx \Pr(t_{17} \leq -1.12) + \Pr(t_{17} \geq 1.12) = 2\Pr(t_{17} \geq 1.12) = .278$$

Problem with completely randomized design:

What if the 10 letters assigned to Empl 1 happen to be more/less difficult to type?

- Randomization provides some protection against this happening, but no guarantee.

Solution: blocking.

Select 10 letters and have both workers type each letter.

Letter (i)	Empl 1	Empl 2	d_i
1	y_{11}	y_{21}	$d_1 = y_{11} - y_{21}$
2	y_{12}	y_{22}	$d_2 = y_{12} - y_{22}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$
10	$y_{1,10}$	$y_{2,10}$	$d_{10} = y_{1,10} - y_{2,10}$

Each letter is a block in which both treatments are observed.

Treatment Structure: one-way
Design Structure: Complete block
Experimental Design: Randomized complete block design (RCBD)

Notice that we no longer have 20 independent observations!

- We can expect y_{11}, y_{21} to be more similar than y_{11}, y_{23} or y_{11}, y_{14} say, because y_{11}, y_{21} are measurements on the same letter.
- $\Rightarrow$ Two-sample $t-$test is no longer an appropriate analysis.

We are really most interested in the average difference in typing speed of the two workers.

Let μ_d =mean difference in typing speed over population of all letters.

Hypothesis of interest:

$$H_0 : \mu_d = 0 \qquad \text{versus} \qquad \mu_d \neq 0$$

Notice that this is a one-sample testing problem:

$$d_1, \ldots, d_{10} \overset{iid}{\sim} N(\mu_d, \sigma_d^2)$$

and we can base our test on $\bar{d} = \frac{1}{10}\sum_{i=1}^{10} d_i$, the sample estimator of μ_d.

2 Cases:

Case 1: σ_d^2 known.

$$\text{var}(\bar{d}) = \frac{\sigma_d^2}{n}$$

$$\Rightarrow \qquad \text{s.e.}(\bar{d}) = \sqrt{\frac{\sigma_d^2}{n}}$$

$$\Rightarrow \qquad \text{test stat.} = z = \frac{\bar{d}}{\sqrt{\sigma_d^2/n}} \sim N(0,1)$$

Case 2: σ_d^2 unknown, so must be estimated.

$$\text{var}(\bar{d}) = \frac{\sigma_d^2}{n}$$

- We estimate σ_d^2 with s_d^2, the sample variance of the d_i's.

$$\Rightarrow \qquad \text{s.e.}(\bar{d}) = \sqrt{\frac{s_d^2}{n}}$$

$$\Rightarrow \qquad \text{test stat.} = t = \frac{\bar{d}}{\sqrt{s_d^2/n}} \sim t(n-1)$$

Suppose we get the following data:

Letter (i)	Empl 1	Empl 2	d_i
1	72	61	11
2	78	69	9
$\vdots$	$\vdots$	$\vdots$	$\vdots$
10	74	61	13
			$\bar{d} = 5$
			$s_d^2 = 76.2$

Then

$$t = \frac{5}{\sqrt{76.2/10}} = 1.81 \quad \Rightarrow \quad p = 2\text{Pr}(t_9 > 1.81) = 0.104$$

The One-way Layout

An Example: Gasoline additives and octane.

Suppose that the effect of a gasoline additive on octane is of interest. An investigator obtains 20 one-liter samples of gasoline and randomly divides these samples into 5 groups of 4 samples each. The groups are assigned to receive 0, 1, 2, 3, or 4 cc/liter of additive and octane measurements are made. The resulting data are as follows:

Treatment	Observations			
A (0cc/l)	91.7	91.2	90.9	90.6
B (1cc/l)	91.7	91.9	90.9	90.9
C (2cc/l)	92.4	91.2	91.6	91.0
D (3cc/l)	91.8	92.2	92.0	91.4
E (4cc/l)	93.1	92.9	92.4	92.4

- This is an example of a one-way (single factor) layout (design).
  - Comparisons among the mean octane levels in the five groups are of interest.
  - Analysis is a generalization of the two sample t-test of equality of means.

In general, we have a single treatment factor with $a \geq 2$ (5 in example) levels (treatments), and n_i ($n_1 = \cdots = n_5 = 4$ in example) replicates for each treatment.

Data:

$$y_{ij}, \quad i = 1, \ldots, a, \quad j = 1, \ldots, n_i.$$

Model:

$$y_{ij} = \mu_i + e_{ij} \qquad \text{(means model)}$$

Three assumptions are commonly made about the e_{ij}s:

(1) e_{ij}, $i = 1, \ldots, a$, $j = 1, \ldots, n_i$, are independent;

(2) e_{ij}, $i = 1, \ldots, a$, $j = 1, \ldots, n_i$, are identically distributed with mean 0 and variance σ^2 (all have same variance);

(3) e_{ij}, $i = 1, \ldots, a$, $j = 1, \ldots, n_i$, are normally distributed.

Alternative **equivalent** form of the model:

$$y_{ij} = \mu + \alpha_i + e_{ij} \quad \text{where } \sum_{i=1}^{a} \alpha_i = 0. \qquad \text{(effects model)}$$

- Here, the restriction $\sum_{i=1}^{a} \alpha_i = 0$ is part of the model. This restriction gives the parameters the following interpretations:

  μ : the grand mean, averaging across all treatment groups

  α_i : the treatment effect (deviation up or down from the grand mean) of the i^{th} treatment

- The relationship between the parameters of the cell means model (the μ_i's) and the parameters of the effects model (μ and the α_i's) is simply:

$$\mu_i = \mu + \alpha_i.$$

Technical points:

- the restriction $\sum_i \alpha_i = 0$ is not strictly necessary in the effects model. Even without it, the effects model is equivalent to the means model. However, without the restriction, the effects model is **overparameterized** and that causes some technical complications in the use of the model. In addition, without the sum-to-zero restriction, the parameters of the effects model don't have the nice interpretations that I've given above.

- It is also possible to use the restriction $\sum_i n_i \alpha_i = 0$ (as in our book), or any one of a large number of other restrictions instead of $\sum_i \alpha_i = 0$. Under the restriction $\sum_i n_i \alpha_i = 0$, μ has a slightly different interpretation: it represents a weighted average (weighted by sample size) of the treatment means rather than a simple average.

Fixed vs. Random Effects:

If the treatments observed are of interest in and of themselves, rather than as representative of some population from which they were drawn, then α_i $i = 1, \ldots, a$, are fixed (non-random), but unknown, parameters $\Rightarrow$ **Fixed Effects Model**.

If the treatments observed can be thought of as a random sample from the population of treatments of interest then it is appropriate to consider the α_i to be random variables $\Rightarrow$ **Random Effects Model**.

- For random effects model additional assumptions are required to completely describe the model (later).

Fixed Effects Model

Notice that in fixed effects models, the model for y consists of two parts:

- what we assume about the mean of y; and
- error, which we can think of as "everything else", or the part of the model that accounts for the fact that, for a given experimental unit, y is typically not exactly equal to its assumed mean.

E.g., in the effects version of our one-way layout model,

$$y_{ij} = \underbrace{\mu + \alpha_i}_{=\mathrm{E}(y_{ij})} + \underbrace{e_{ij}}_{\text{error}}$$

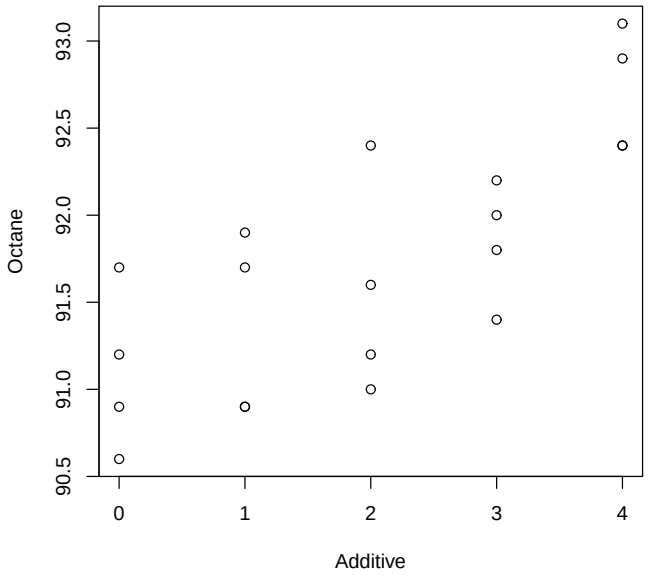
This model is assumed to describe y_{ij} for each i and j. So it is saying that for each response in the ith treatment, its mean is assumed to be $\mu + \alpha_i$ or, in the cell-means version of the model, μ_i.

That is, the model says that

$$\begin{aligned}
&\text{the mean of each } y_{11}, ..., y_{1n_1} \text{ (each obs in treatment 1)} = \mu + \alpha_1 = \mu_1 \\
&\text{the mean of each } y_{21}, ..., y_{2n_2} \text{ (each obs in treatment 2)} = \mu + \alpha_2 = \mu_2 \\
&\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad \vdots \\
&\text{the mean of each } y_{a1}, ..., y_{an_a} \text{ (each obs in treatment } a) = \mu + \alpha_a = \mu_a
\end{aligned}$$

- I.e., for data from a treatments, model says there are a treatment means. Simple!

A picture:



Fitting the model: (Ordinary) Least Squares

Recall the cell means version of our model:

$$y_{ij} = \mu_i + e_{ij}, \quad \text{for each} \quad j = 1, \ldots, n_i, i = 1, \ldots, a.$$

Our model fits well if the error term e_{ij} is small in magnitude for all i and j. Therefore, estimate μ_i, $i = 1, \ldots, a$, with the quantities that make

$$e_{ij}^2 = (y_{ij} - \mu_i)^2$$

small over all i, j.

- Intuitively, it would make just as much sense to try to minimize $|e_{ij}|$ over all i, j. However, it turns out that minimizing e_{ij}^2 is much easier to handle mathematically, and is "better" in a specific statistical sense.

Least Squares: Minimize

$$L = \sum_{i=1}^{a} \sum_{j=1}^{n_i} e_{ij}^2 = \sum_{i=1}^{a} \sum_{j=1}^{n_i} (y_{ij} - \mu_i)^2$$

with respect to $\mu_1, ..., \mu_a$, to obtain estimators $\hat{\mu}_1, ..., \hat{\mu}_a$.

Minimizing the least squares criterion L is a calculus problem. The way to do it is to take derivatives of L, set them equal to zero and solve the resulting equations, which are called the **normal equations**

The $\hat{\mu}_i$s that solve the normal equations are as follows:

$$\hat{\mu}_1 = \bar{y}_{1\cdot}, \quad \hat{\mu}_2 = \bar{y}_{2\cdot}, \quad ..., \quad \hat{\mu}_a = \bar{y}_{a\cdot}$$

or, in general,

$$\hat{\mu}_i = \bar{y}_{i\cdot}.$$

- That is, we estimate the population mean for the ith treatment with the sample mean of the data in our experiment from the ith treatment. Simple!

In the effects version of the model, $y_{ij} = \mu + \alpha_i + e_{ij}$, so the least squares criterion becomes

$$L = \sum_{i=1}^{a}\sum_{j=1}^{n_i} e_{ij}^2 = \sum_{i=1}^{a}\sum_{j=1}^{n_i}\{y_{ij} - (\mu + \alpha_i)\}^2$$

which, along with the restriction† $\sum_i \alpha_i = 0$ leads to estimators

$$\hat{\mu} = \frac{1}{a}(\bar{y}_{1\cdot} + \cdots + \bar{y}_{a\cdot}), \quad \hat{\alpha}_i = \bar{y}_{i\cdot} - \frac{1}{a}(\bar{y}_{1\cdot} + \cdots + \bar{y}_{a\cdot}), \quad i = 1, \ldots, a$$

- In the balanced case, these estimators simplify to become:

$$\hat{\mu} = \bar{y}_{\cdot\cdot}, \quad \hat{\alpha}_i = \bar{y}_{i\cdot} - \bar{y}_{\cdot\cdot}, \quad \hat{\mu}_i = \bar{y}_{i\cdot}. \qquad (*)$$

Note that there is no disagreement between the cell means and effects version of the model. They are completely consistent. The cell means model says that the data from the ith treatment have mean μ_i, and the effects model just breaks up that μ_i into two pieces:

$$\mu_i = \mu + \alpha_i$$

The consistency in the two model versions can be see in that the above relationship holds for the parameter estimators too:

$$\hat{\mu}_i = \bar{y}_{i\cdot} = \hat{\mu} + \hat{\alpha}_i$$

---

† Under the alternative restriction, $\sum_{i=1}^{a} n_i\alpha_i = 0$, we get the estimators given in (*) in both the balanced and unbalanced cases.

Example Gasoline Additives (Continued)

	Treatment	Observations				Total	Mean
i		y_{i1}	y_{i2}	y_{i3}	y_{i4}	$y_{i\cdot}$	$\bar{y}_{i\cdot}$
1	A	91.7	91.2	90.9	90.6	364.4	91.10
2	B	91.7	91.9	90.9	90.9	365.4	91.35
3	C	92.4	91.2	91.6	91.0	366.2	91.55
4	D	91.8	92.2	92.0	91.4	367.4	91.85
5	E	93.1	92.9	92.4	92.4	370.8	92.70

So, the parameter estimates from the cell means model are

$$\hat{\mu}_1 = \bar{y}_{1\cdot} = 91.10, \quad \hat{\mu}_2 = \bar{y}_{2\cdot} = 91.35, \quad ..., \quad \hat{\mu}_5 = \bar{y}_{5\cdot} = 92.70.$$

Or, if we prefer the effects version of the model,

$$y_{\cdot\cdot} = 364.4 + \cdots + 370.8 = 1834.2, \quad \hat{\mu} = \bar{y}_{\cdot\cdot} = 1834.2/20 = 91.71$$

$$\begin{aligned}
\hat{\alpha}_1 &= 91.10 - 91.71 = -0.61 \\
\hat{\alpha}_2 &= 91.35 - 91.71 = -0.36 \\
\hat{\alpha}_3 &= 91.55 - 91.71 = -0.16 \\
\hat{\alpha}_4 &= 91.85 - 91.71 = 0.14 \\
\hat{\alpha}_5 &= 92.70 - 91.71 = 0.99
\end{aligned}$$

■

- Under assumptions (1) and (2) on the e_{ij}s, the method of least squares gives the Best (minimum variance) Linear Unbiased Estimators (BLUE) of the parameters.
  - The point being that we use least-squares to fit the model not just because it makes sense intuitively, but there's also theory that establishes that it is an optimal approach in some well-defined sense.

There's one more parameter of the model: σ^2, the error variance.

How do we estimate σ^2?

The value of the least squares criterion L when evaluated at our fitted model (what we get when we plug in our parameter estimates) is a measure of how well our model fits the data (its the sum of squared differences between the actual and fitted values):

$$L_{\min} = \sum_i \sum_j (y_{ij} - \hat{\mu}_i)^2 = \sum_i \sum_j [y_{ij} - \bar{y}_{i\cdot}]^2$$
$$\equiv SS_E \quad \text{the } \textbf{Sum of Squares due to Error}$$

The **Mean Squares due to Error** is defined to be

$$MS_E = \frac{SS_E}{N - a}$$

and has expected value

$$\mathrm{E}(MS_E) = \frac{\mathrm{E}(SS_E)}{N-a} = \frac{\sigma^2(N-a)}{N-a} = \sigma^2. \qquad (**)$$

- That is, MS_E is an unbiased estimator of σ^2 and therefore, to complete the process of "fitting the model" we take as our estimator of the error variance
$$\hat{\sigma}^2 = MS_E.$$

- The divisor in MS_E (in this case $N-a$) is called the **error degrees of freedom** or d.f.$_E$.

- MS_E is analogous to s_P^2 in the t-test with equal but unknown variances. It is a pooled estimate of σ^2.

- Note that the second equality in (**) follows from the fact that that $SS_E/\sigma^2 \sim \chi^2(N-a)$ where $N = n_1 + \cdots + n_a$. This was a result we noted earlier in the notes when we introduced the chi-square distribution.

Inferences on the μ_is or α_is

The one-way ANOVA is designed to test whether or not all treatment means are equal or, equivalently, whether or not all treatment effects (deviations from the grand mean) are 0. I.e., ANOVA tests the null hypothesis

$$\begin{aligned} & H_0 : \mu_1 = \mu_2 = \cdots = \mu_a \\ \text{or, equivalently,} \quad & H_0 : \alpha_1 = \alpha_2 = \cdots = \alpha_a = 0 \\ \text{versus} \quad & H_1 : \mu_i \neq \mu_{i'} \quad \text{for at least one pair } i \neq i' \end{aligned}$$

Decomposition of SS_T:

The **total (corrected) sum of squares**,

$$SS_T = \sum_i \sum_j (y_{ij} - \bar{y}_{\cdot\cdot})^2$$

is a measure of the total variability in the data. Notice

$$\begin{aligned} SS_T &= \sum_i \sum_j (y_{ij} - \bar{y}_{i\cdot} + \bar{y}_{i\cdot} - \bar{y}_{\cdot\cdot})^2 \\ &= \sum_i \sum_j (\bar{y}_{i\cdot} - \bar{y}_{\cdot\cdot})^2 + \sum_i \sum_j (y_{ij} - \bar{y}_{i\cdot})^2 + \underbrace{2 \sum_i \sum_j (\bar{y}_{i\cdot} - \bar{y}_{\cdot\cdot})(y_{ij} - \bar{y}_{i\cdot})}_{=0} \\ &= \sum_i n_i (\bar{y}_{i\cdot} - \bar{y}_{\cdot\cdot})^2 + \sum_i \sum_j (y_{ij} - \bar{y}_{i\cdot})^2. \end{aligned}$$

That is,

$$SS_T = SS_{Trt} + SS_E.$$

SS_T has $N - 1$ d.f., SS_{Trt} has $a - 1$ d.f., and SS_E has $N - a$ d.f., so we also have a decomposition of the total d.f.:

$$\begin{aligned} & \text{d.f.}_T = \text{d.f.}_{Trt} + \text{d.f.}_E \\ \Rightarrow \quad & N - 1 = (a - 1) + (N - a) \end{aligned}$$

What are degrees of freedom?

Remember when we introduced the chi-square distribution, we said that it was the distribution of a random variable defined as a sum of k independent, squared, (standard) normal random variables. Such a random variable has a chi-square distribution, which has a parameter called the degrees of freedom.

- So, sums of squares formed from normal random variables (e.g., SS_T, SS_{Trt}, SS_E) are chi-square distributed, each with a certain degrees of freedom, which counts the number of independent terms in the sum.
- Roughly speaking, a sum of squares quantifies variability of one kind or another, and the d.f. for a sum of squares counts the number of independent pieces of information that goes into that quantification of variability.
  - E.g., SS_T quantifies the total variability in the data and there are $N-1$ independent pieces of information that go into that quantification.
  - SS_{Trt} quantifies the variability of the treatment means around the grand mean (between-treatment variability). There are $a-1$ independent pieces of information that go into SS_{Trt}.
  - SS_E quantifies within treatment variability or, looked at another way, the variability in the data other than that which is due to differences in the treatment means.

Notice,

$$SS_E = \sum_i \sum_j (y_{ij} - \bar{y}_{i\cdot})^2 = \sum_i (n_i - 1)s_i^2,$$

where s_i^2 is the sample variance within the i^{th} treatment, so

$$\begin{aligned} MS_E &= \frac{SS_E}{\sum_i (n_i - 1)} \\ &= \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2 + \cdots + (n_a - 1)s_a^2}{(n_1 - 1) + \cdots + (n_a - 1)} \\ &= s_P^2 \quad \text{when } a = 2 \end{aligned}$$

a pooled estimator of σ^2.

How does the decomposition of SS_T into $SS_{Trt} + SS_E$ help us test for differences in the treatment means?

We've seen that

$$\mathrm{E}(MS_E) = \mathrm{E}\left(\frac{SS_E}{\mathrm{d.f.}_E}\right) = \sigma^2.$$

I.e., MS_E estimates σ^2.

In addition, when H_0 is true, it can be shown that

$$MS_{Trt} = \frac{SS_{Trt}}{\mathrm{d.f.}_{Trt}}$$

also estimates σ^2. Therefore, if H_0 is true MS_{Trt}/MS_E should be ≈ 1.

However, when H_0 is false, it can be shown that MS_{Trt} estimates something bigger than σ^2. Therefore, to determine whether H_0 is true or not, we can look at how much larger than 1 MS_{Trt}/MS_E is. This ratio of mean squares becomes our test statistic for H_0.

That is, by some tedious calculations using the rules for how expectations work (see p.31) it can be shown that

$$\begin{aligned}\mathrm{E}(MS_{Trt}) &= \sigma^2 + \frac{\sum_i n_i \alpha_i^2}{a-1} \quad \text{in general} \\ &= \sigma^2 \quad \text{if } H_0 \text{ is true.}\end{aligned}$$

Therefore,

$$\frac{MS_{Trt}}{MS_E} = \begin{cases} \dfrac{\text{estimator of something larger than } \sigma^2}{\text{estimator of } \sigma^2} & \text{if } H_0 \text{ is false;} \\ \dfrac{\text{estimator of } \sigma^2}{\text{estimator of } \sigma^2} & \text{if } H_0 \text{ is true.} \end{cases}$$

- If $\frac{MS_{Trt}}{MS_E} >> 1$ then it makes sense to reject H_0.

How much larger than 1 should MS_{Trt}/MS_E be to reject H_0?

- Should be large in comparison with its distribution under H_0.
- Notice that $\frac{MS_{Trt}}{MS_E}$ can be written as

$$\frac{MS_{Trt}}{MS_E} = \frac{SS_{Trt}/\text{d.f.}_{Trt}}{SS_E/\text{d.f.}_E}$$

  or the ratio of two chi-square distributed sums of squares, each divided by its d.f.. It can also be shown that SS_{Trt} and SS_E are independent.

Therefore, under H_0,

$$F = \frac{MS_{Trt}}{MS_E} \sim F(a-1, N-a)$$

and our test statistic becomes an F-test. We reject H_0 for large values of F in comparison to an $F(a-1, N-a)$ distribution.

Result: An α-level test of $H_0 : \alpha_1 = \cdots = \alpha_a = 0$ is: Reject H_0 if

$$F > F_\alpha(a-1, N-a)$$

- Reporting simply the test result (reject/not reject) is not as informative as reporting the $p-$value. The p-value quantifies the strength of the evidence provided by the data against the null hypothesis not just whether that evidence was sufficient to reject.

The test procedure may be summarized in an **ANOVA Table**:

Source of Variation	Sum of Squares	d.f.	Mean Squares	E(MS)	F
Treatments	SS_{Trt}	$a-1$	MS_{Trt}	$\sigma^2 + \frac{\sum n_i \alpha_i^2}{a-1}$	$\frac{MS_{Trt}}{MS_E}$
Error	SS_E	$N-a$	MS_E	σ^2	
Total	SS_T	$N-1$			

A Note on Computations:

We have defined SS_T, SS_{Trt}, and SS_E as sums of squared deviations. Equivalent formulas for the SS_T and SS_{Trt} are as follows:

$$SS_T = \sum_{i=1}^{a}\sum_{j=1}^{n_i} y_{ij}^2 - \frac{y_{\cdot\cdot}^2}{N}$$

$$SS_{Trt} = \sum_{i=1}^{a} \frac{y_{i\cdot}^2}{n_i} - \frac{y_{\cdot\cdot}^2}{N}$$

SS_E is typically computed by subtraction:

$$SS_E = SS_T - SS_{Trt}$$

Gasoline Additive Example (Continued):

- See handout labeled gasadd1.sas.

$$SS_{Trt} = \frac{(364.4)^2 + (365.4)^2 + (366.2)^2 + (367.4)^2 + (370.8)^2}{4} - \frac{(1834.2)^2}{20}$$
$$= 6.108$$

Similarly, $SS_T = 9.478$ and $SS_E = 3.370$

Our test statistic is

$$F = \frac{\text{SS}_{\text{Trt}}/(a-1)}{\text{SS}_{\text{E}}/(N-a)} = \frac{6.108/4}{3.370/15} = 6.80.$$

To obtain the $p-$value we compare with the $F(4, 15)$ distribution.

From Table D.5 we see that the upper 0.05 point of this distribution is $F_{.05}(4, 15) = 3.06$, $F_{0.01}(4, 15) = 4.89$, $F_{.001}(4, 15) = 8.25$.

Since the observed test statistic (6.80) falls between the upper 0.01 and 0.001 points we know that our $p-$value must be between 0.01 and 0.001. From the SAS output (p.3) we obtain the exact $p-$value of 0.0025. This result leads us to reject $H_0 : \mu_1 = \cdots = \mu_a$ and conclude that there are differences among the additives.

We estimate σ^2 using $MS_E = 3.370/15 = 0.2246$.

Confidence Interval for a Treatment Mean:

An estimate of $\text{var}(\hat{\mu}_i) = \text{var}(\bar{y}_{i\cdot})$ is

$$\hat{\text{var}}(\bar{y}_{i\cdot}) = \frac{\hat{\sigma}^2}{n_i} = \frac{0.2246}{4} = 0.05625$$

A $100(1-\alpha)\%$ CI for μ_i is given by

$$\bar{y}_{i\cdot} \pm t_{\alpha/2}(N-a)\sqrt{\hat{\text{var}}(\bar{y}_{i\cdot})}$$

For example, a 95% CI for μ_1 is

$$\bar{y}_{1\cdot} \pm \underbrace{t_{0.05/2}(20-5)}_{=2.131} \sqrt{\hat{\text{var}}(\bar{y}_{1\cdot})}$$
$$= 91.10 \pm 2.131 \underbrace{\sqrt{.05625}}_{=.2370} = (90.59, 91.61)$$

This F test can also be derived and understood as a **test of nested models.**

In general, parameter restrictions in a linear model (e.g., hypotheses that that set certain model parameters to zero) may be tested by comparing SS_E for the unrestricted/full/more general model with SS_E for the restricted/partial/simpler model.

Idea:

- SS_E (full) quantifies lack of fit in full model,
- SS_E (partial) quantifies lack of fit in partial model.
- d.f.$_E$ (partial) − d.f.$_E$ (full) quantifies how much more complex the full model is (how many more parameters have to be estimated).

$\Rightarrow$ $\frac{SS_E \text{ (partial)} - SS_E \text{ (full)}}{\text{d.f.}_E \text{ (partial)} - \text{d.f.}_E \text{ (full)}}$ quantifies how much better the full model fits, relative to how many more parameters it "costs" to fit it.

  - If this value is "large" then the improvement in fit for the full model is worth its extra complexity, where "large" is measured relative to experimental error.

In general, the test statistic to compare a full and partial model is

$$F = \frac{SS_{H_0}/\text{d.f.}_{H_0}}{MS_E \text{ (full model)}}$$

where

$$SS_{H_0} = SS_E \text{ (partial)} - SS_E \text{ (full)}$$

and

$$\text{d.f.}_{H_0} = \text{d.f.}_E \text{ (partial)} - \text{d.f.}_E \text{ (full)}$$

or

$$F = \frac{\{SS_E \text{ (partial)} - SS_E \text{ (full)}\}/\{\text{d.f.}_E \text{ (partial)} - \text{d.f.}_E \text{ (full)}\}}{MS_E \text{ (full model)}}$$

For example, the one-way anova F test compares

full model:

$$y_{ij} = \mu + \alpha_i + e_{ij}$$

partial model:

$$y_{ij} = \mu + e_{ij}$$

In this case, it is easy to show mathematically that

$$F = \frac{SS_{H_0}/\text{d.f.}_{H_0}}{MS_E(\text{full model})} = \frac{MS_{\text{Trt}}(\text{full})}{MS_E(\text{full})}.$$

The equivalence can be demonstrated using our **Gasoline Additives Example.**

- See gasadd1.R where we use the anova() function to obtain a test of nested models. Notice that this gives an identical result to the one given in the anova table for the one-way model (the full model).

Comparisons (Contrasts) among Treatment Means

Once H_0 is rejected we usually want more information: which μ_is differ and by how much? To answer these questions we can make

(1) *a priori* (planned) comparisons; or
(2) Data-based (unplanned, *a posteriori*, or *post hoc*) comparisons. (A.K.A. Data-snooping.)

Ideally, we avoid (2) altogether. The experiment should be thought out well enough so that all hypotheses of interest take the form of planned comparisons. But comparisons of type (2) are sometimes necessary, particularly in preliminary or exploratory studies.

It is important to understand the **multiple comparisons problem** inherent in doing multiple comparisons of either type, but especially of type (2).

- When performing a single statistical hypothesis test, we try to avoid incorrectly rejecting the null hypothesis (a "Type I Error") for that test by setting this probabilitity of such an error to be low.
  - This probability is called the significance level, α, and we typically set it to be $\alpha = .05$ or some other small value.
- This approach controls the probability of a Type I error on that one test.
- However, when we conduct multiple hypothesis tests, the probability of making *at least one* Type I error increases the more tests we perform.
  - The more chances to make a mistake you have the more likely it is that you will make a mistake.
- The problem is exacerbated when doing post hoc tests because post hoc hypotheses are typically chosen by examining the data, doing multiple (many, usually) informal (perhaps even subconscious) comparisons and deciding to do formal hypothesis tests for those comparisons that look "promising".
  - That is, even just a single post hoc hypothesis test really involves many implicit comparisons, which inflates its Type I error probability.

- We'll return to the multiple comparisons problem and statistical methods to help "solve" it later.

Contrasts:

A **contrast** takes the form

$$\psi = \sum_{i=1}^{a} c_i \mu_i \qquad \text{where} \qquad \sum_i c_i = 0$$

and is estimated by

$$C = \sum_i c_i \hat{\mu}_i = \sum_i c_i \bar{y}_{i\cdot}.$$

For example, suppose we have three treatments with population means μ_1, μ_2, μ_3 that we estimate with the corresponding sample means $\bar{y}_{1\cdot}, \bar{y}_{2\cdot}, \bar{y}_{3\cdot}$.

A contrast among the treatment population means is a linear combination of the form

$$\psi = c_1\mu_1 + c_2\mu_2 + c_3\mu_3 \quad \text{such that} \quad c_1 + c_2 + c_3 = 0$$

Simplest example: pairwise contrast $\mu_i - \mu_{i'}$

In our three treatment example we could compare treatments 1 and 3 with the pairwise contrast

$$\psi = 1\mu_1 + 0\mu_2 + (-1)\mu_3 = \mu_1 - \mu_3$$

which we estimate with

$$C = \bar{y}_{1\cdot} - \bar{y}_{3\cdot}$$

Since our responses (the y_{ij}'s) are independent normal r.v.'s $\Rightarrow$ the treatment means (the $\bar{y}_{i\cdot}$'s) are independent normal r.v.'s. And because a linear combination of normal r.v.'s is normal,

$$\sum_i c_i \bar{y}_{i\cdot} \sim N\left(\sum_i c_i \mu_i, \sigma^2 \sum_i \frac{c_i^2}{n_i}\right)$$

We can test $H_0 : \psi = 0$ versus $H_1 : \psi \neq 0$ using a t-test. In a $t-$test we look at a test statistic of the form

$$t = \frac{\text{(param. est.)} - \text{(null value)}}{\text{s.e. of param. est.}}$$

Where the standard error of an estimator is defined to be its estimated standard deviation (the square root of its estimated variance).

So, for a test of $H_0 : \psi = 0$ we use the test statistic

$$t = \frac{C - 0}{\text{s.e.}(C)} = \frac{\sum_i c_i \bar{y}_{i\cdot}}{\sqrt{\text{MS}_\text{E} \sum_i \frac{c_i^2}{n_i}}}.$$

Notice we use MS_E to estimate σ^2.

Under H_0, $t \sim t(\text{d.f.}_E)$ so we compare t to $t(N-a)$ to obtain our $p-$value.

Equivalently,

$$\begin{aligned} F = t^2 &= \frac{(\sum_i c_i \bar{y}_{i\cdot})^2 / \sum_i c_i^2/n_i}{MS_E} \\ &= \frac{SS_C/1}{MS_E} = \frac{MS_C}{MS_E} \\ &\sim F(1, N-a) \end{aligned}$$

so we compare F to a $F(1, N-a)$ distribution and reject H_0 at significance level α if $F > F_\alpha(1, N-a)$.

Example: Scab disease in potatoes (Cochran and Cox, section 4.3).

Scab disease does not thrive in acidic soil. Based on this fact, an experiment was conducted to investigate the effects of applying the soil-acidifying compound sulphur as a preventive for scab disease.

General Objective: Reduce scab disease.

Experimental Objectives:

1. Determine whether or not sulphur reduces scab disease when applied to the soil.

2. Determine the best time of year to apply sulphur.

3. Determine the best dosage level.

Response Variable: "Scab Index" defined as the percentage of scabbed surface area for a randomly selected sample of 100 potatoes.

Treatments: Seven treatments were selected for study. These consisted of a control treatment (0 lbs./acre sulphur) and both a spring and fall application for each of three sulphur dosages (300, 600, and 1200 lbs./acre).

The treatment structure here is really a two-way factorial structure. There are two treatment factors: Amount of sulphur with levels 0, 300, 600 and 1200, and time of application, with levels spring and fall.

If we combine the levels of the two factors we get 8 treatments. However, notice that there's absolutely no difference between applying 0 in the spring, and applying zero in the fall, so there are really only 7 distinct treatments.

- For now, we are going to ignore the underlying two-way treatment structure here and consider the 7 treatments as seven levels of a single treatment factor:

$$T_1 = \text{Control}, \quad T_2 = \text{F3}, \quad T_3 = \text{F6}, \quad T_4 = \text{F12},$$
$$T_5 = \text{S3}, \quad T_6 = \text{S6}, \quad T_7 = \text{S12}.$$

The experiment was conducted in a completely randomized design in which 32 plots were randomly assigned to the 7 treatments so that 8 replicates were obtained in the control treatment and four replicates in each other treatment.

The questions of interest here were

(1) Is there an effect of treating the soil with sulphur?

(2) What time of year should the soil be treated?

(3) What is the effect of dose?

These questions can be answered through the use of planned comparisons. Each question corresponds to a hypothesis that a contrast of the form $c_1\mu_1 + \cdots + c_7\mu_7$ is equal to 0. Appropriate choices for these contrasts are as follows:

(1) $\psi_1 = 6\mu_1 - \mu_2 - \mu_3 - \mu_4 - \mu_5 - \mu_6 - \mu_7$ ($c_1 = 6$, $c_2 = \cdots = c_7 = -1$)
(2) $\psi_2 = \mu_2 + \mu_3 + \mu_4 - \mu_5 - \mu_6 - \mu_7$ ($c_1 = 0$, $c_2 = c_3 = c_4 = 1$, $c_5 = c_6 = c_7 = -1$)
(3) Two contrasts:
    (a) $\psi_3 = 2\mu_2 - \mu_3 - \mu_4 + 2\mu_5 - \mu_6 - \mu_7$ (300 vs. 600 & 1200)
    (b) $\psi_4 = -\mu_3 + \mu_4 - \mu_6 + \mu_7$ (600 vs. 1200)

- These two contrasts can be tested separately, or simultaneously.

We can also determine whether or not there is an interaction between dose and time of application by testing the contrasts formed by multiplying the contrast coefficients in (2) and (3):

(4) Interaction:
    (a) $\psi_5 = 2\mu_2 - \mu_3 - \mu_4 - 2\mu_5 + \mu_6 + \mu_7$
    (b) $\psi_6 = -\mu_3 + \mu_4 + \mu_6 - \mu_7$

- Again, these two contrasts can be tested separately or simultaneously.

Computations:

$$\begin{aligned} C_1 &= 6\bar{y}_{1\cdot} - \bar{y}_{2\cdot} - \bar{y}_{3\cdot} - \bar{y}_{4\cdot} - \bar{y}_{5\cdot} - \bar{y}_{6\cdot} - \bar{y}_{7\cdot} \\ &= 6(22.625) - 9.50 - 15.50 - 5.75 - 16.75 - 18.25 - 14.25 = 55.75 \end{aligned}$$

$$\begin{aligned} \hat{\text{var}}(C_1) &= MS_E \sum_i \frac{c_i^2}{n_i} \\ &= 44.9\big[6^2/8 + (-1)^2/4 + (-1)^2/4 + (-1)^2/4 + (-1)^2/4 \\ &\qquad + (-1)^2/4 + (-1)^2/4\big] \\ &= 44.9(6) = 269.4 \end{aligned}$$

so

$$t = \frac{C_1 - 0}{\text{s.e.}(C_1)} = \frac{55.75}{\sqrt{269.4}} = 3.40.$$

Since $t_{.05/2}(32-7) = 2.060$, we reject $H_0 : \psi_1 = 0$ at $\alpha = .05$, and conclude that there is a difference in the mean scab index between the control and active treatments. (Adding sulphur helps reduce scab disease.)

Alternatively, we could use the equivalent F test:

$$SS_{C_1} = \frac{C_1^2}{\sum_i \frac{c_i^2}{n_i}} = \frac{55.75^2}{6} = 518.010$$

$$F = \frac{MS_{C_1}}{MS_E} = \frac{518.010/1}{44.915} = 11.53$$

- The p-value is the same with either test: $p = .0023$.

The CONTRAST and ESTIMATE statements in PROC GLM

- See handout scab1.sas.

In PROC GLM in SAS, a constant term is always included in all models unless you specify otherwise by using the NOINT option on the MODEL statement. Therefore by default, ANOVA models are parameterized with effects models. E.g., the one way ANOVA model is parameterized as $y_{ij} = \mu + \alpha_i + e_{ij}$ rather than $y_{ij} = \mu_i + e_{ij}$. A contrast is a linear combination of the model parameters $\mu, \alpha_1, \alpha_2, \ldots, \alpha_a$.

The MODEL statement here would be

```
model y=A;
```

where A is the treatment factor with a levels corresponding to the effects $\alpha_1, \ldots, \alpha_a$ (A must be specified as a factor by including A in a CLASS statement).

The syntax of the CONTRAST statements is

```
contrast 'contrast label' intercept c0 A c1 c2 ... ca;
```

Here, c0, c1, ..., ca are the contrast coefficients corresponding to μ, $\alpha_1, \ldots, \alpha_a$, respectively. Here, "intercept" indicates that the next coefficient will be for μ, and "A" indicates that the next coefficients will be for $\alpha_1, \ldots, \alpha_a$.

- Note that, by default, the levels of factor A are ordered alphabetically. This ordering can be changed with the ORDER= option on the PROC GLM statement. ORDER=data orders the levels as they appear in the data set. The ordering of the factor levels is important, because the contrast coefficients are matched to the levels of the factor in the order that SAS is currently using. The factor level ordering used can be seen in the summary of the CLASS variables that appears at the beginning of the PROC GLM output.

- SAS allows you to omit terms from the CONTRAST statement. How it fills in the contrast coefficients for omitted terms is complicated, in general, but for the one-way anova models, omitted coefficients are assumed to equal 0.

  - E.g., the following three CONTRAST statements for the scab disease example are equivalent. Each one tests $\mu_2 - \mu_3 = 0$:

```
CONTRAST 'mu2-mu3 (a)' intercept 0 trt 0 1 -1 0 0 0 0;
CONTRAST 'mu2-mu3 (b)' trt 0 1 -1 0 0 0 0;
CONTRAST 'mu2-mu3 (c)' trt 0 1 -1;
```

Another feature of the CONTRAST statement worth knowing about is that it allows one to test more than one contrast simultaneously.

- Suppose we want to test a dose effect in the scab disease example. We could test $H_0 : \psi_3 = 0$ and $H_0 : \psi_4 = 0$ separately. This would give us tests of two distinct aspects of the dose effect.

- Alternatively, we might want to test the "entire" dose effect at once; that is, we might want to test

$$H_0 : (\psi_3 = 0 \quad \underline{\text{and}} \quad \psi_4 = 0)$$

  (i.e., $H_0 : \psi_3 = \psi_4 = 0$).

- This can be done by specifying both contrasts ψ_3 and ψ_4 on the same CONTRAST statement, separated by a comma:

```
contrast 'dose' trt  0  2 -1 -1  2 -1 -1,
               trt  0  0 -1  1  0 -1  1;
```

  The resulting simultaneous test is a two degree of freedom test, one for each hypothesis being tested in the combined simultaneous hypothesis.

The ESTIMATE statement has a very similar syntax to CONTRAST. It is used to obtain an estimate of a linear combination of the model parameters rather than test a hypothesis concerning a linear combination of the model parameters.

- For example, to estimate $\mu_1 - \frac{1}{6}(\mu_2 + \ldots + \mu_7)$ (the difference between the control treatment mean and the average of the 6 active treatment means) we would say:

```
estimate 'control-trt' trt  1 -.1666667 -.1666667 -.1666667
                              -.1666667 -.1666667 -.1666667;
```

or, equivalently but better because it avoids rounding error in the contrast coefficients,

```
estimate 'control-trt' trt  6 -1 -1 -1 -1 -1 -1/divisor=6;
```

Scab Disease Example — Conclusions:

(1) Some differences among the treatment means exist since $H_0 : \mu_1 = \cdots = \mu_7$ was rejected ($p = .0103$).

(2) Sulphur reduces scab disease since
   (a) we rejected $H_0 : \psi_1 = 0$ ($p = .0023$); and
   (b) control mean is higher than average of the active treatment means ($C_1 = 9.29 > 0$).

(3) Fall is best time for sulphur application since
   (a) we rejected $H_0 : \psi_2 = 0$ ($p = .0332$); and
   (b) fall means are smaller than spring means.

(4) Evidence does not suggest that dose matters since $H_0 : \psi_3 = \psi_4 = 0$ was not rejected. May as well use 300cc/acre.

(5) No evidence of an interaction since $H_0 : \psi_5 = \psi_6 = 0$ was not rejected.

Orthogonal Contrasts

Two contrasts $\psi_1 = \sum_i c_{1i}\mu_i$, $\psi_2 = \sum_i c_{2i}\mu_i$ are **orthogonal** if $\sum_i c_{1i}c_{2i}/n_i = 0$ ($\sum_i c_{1i}c_{2i} = 0$ in balanced case).

- Sample versions of orthogonal contrasts are independent. Consider the balanced case:

$$\begin{aligned}\text{cov}\left(\sum_i c_{1i}\bar{y}_{i\cdot}, \sum_j c_{2j}\bar{y}_{j\cdot}\right) &= \sum_i \sum_j c_{1i}c_{2j}\text{cov}(\bar{y}_{i\cdot}, \bar{y}_{j\cdot}) \\ &= \sum_i c_{1i}c_{2i}\text{var}(\bar{y}_{i\cdot}) \\ &= \frac{\sigma^2}{n}\sum_i c_{1i}c_{2i} = 0\end{aligned}$$

- The interpretation of this independence is that orthogonal contrasts correspond to distinct, non-redundant comparisons among the treatment means. When we test hypotheses on each of several contrasts that are mutually orthogonal, we are asking a set of "non-overlapping" or "non-redundant" questions in some sense.

- At most $a - 1$ mutually orthogonal contrasts can be constructed on a means.

- Orthogonal contrasts can be used to partition the treatment SS into $a - 1$ independent components each with d.f. = 1:

$$SS_{Trt} = SS_{C_1} + SS_{C_2} + \cdots + SS_{C_{a-1}}$$

  for any set $\{C_1, \ldots, C_{a-1}\}$ of mutually orthogonal sample contrasts.

Example: In a balanced experiment comparing mean response from 2 drugs (μ_1, μ_2) and a placebo (μ_3) a natural set of orthogonal contrasts is

$$\begin{aligned}\psi_1 &= \mu_1 - \mu_2 \\ \psi_2 &= \mu_1 + \mu_2 - 2\mu_3\end{aligned}$$

ψ_1 and ψ_2 are orthogonal since

$$\sum_i c_{1i}c_{2i} = (1)(1) + (-1)(1) + (0)(-2) = 0.$$

Scab Disease Example:
The 7 treatments can be represented by a tree diagram:

If we construct contrasts between the branches at the same level, then these contrasts will be orthogonal.

	A	B	C	D	E	F	G
ψ_1 :							
ψ_2 :							
ψ_3 :							
ψ_4 :							

Two contrasts that are not orthogonal are

$$\psi_1 = \mu_1 - \frac{1}{6}(\mu_2 + \mu_3 + \mu_4 + \mu_5 + \mu_6 + \mu_7) \quad \text{and}$$
$$\psi_2 = \mu_1 - \frac{1}{5}(\mu_2 + \mu_3 + \mu_4 + \mu_5 + \mu_6)$$

- It is clear that the sample versions of these contrasts will not be independent and that there is some redundancy in asking the questions (i) does $\psi_1 = 0$? and (ii) does $\psi_2 = 0$? Clearly, if we reject (i), we are more likely to reject (ii) and vice versa.

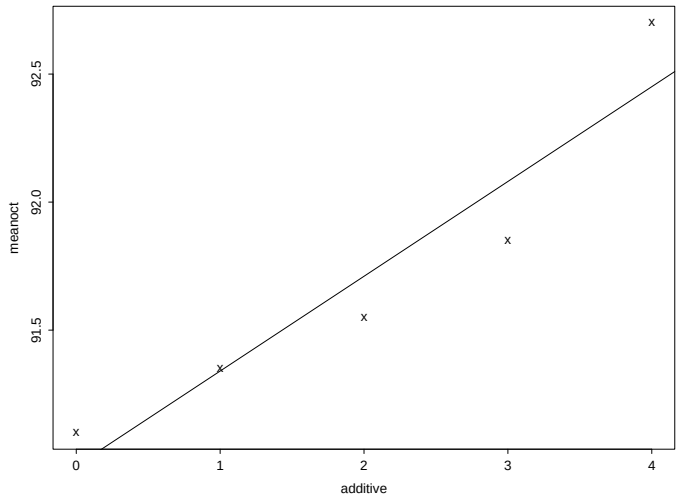
Orthogonal Polynomials — Refer to Gasoline Example

Orthogonal polynomials are a specific type of orthogonal contrast useful when a treatment factor is quantitative. They are especially convenient and easy to use when the treatments are evenly spaced and replication is balanced (e.g., the gasoline example).

Suppose we plot mean octane vs. amount of additive.

Do we have a straight line relationship?

If so, is the slope equal to 0?



If we have a evenly spaced treatments we may test contrasts corresponding to polynomials of degree $1, 2, \ldots, a-1$. That is, we can test the models

$$\begin{aligned}
&\text{degree 1 (linear):} && \mu_i = \beta_0 + \beta_1 x_i \\
&\text{degree 2 (quadratic):} && \mu_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 \\
&\text{degree 3 (cubic):} && \mu_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \beta_3 x_i^3 \\
&\qquad\vdots \\
&\text{degree } a-1: && \mu_i = \beta_0 + \beta_1 x_i + \cdots + \beta_{a-1} x_i^{a-1} \qquad (*)
\end{aligned}$$

Orthogonal contrast coefficients for testing these polynomials may be found in Table D.6 in our text, p.630.

- In Table D.6, the number of treatments a is given as g and contrasts are given for testing a polynomial of order (degree) 1 (linear), order 2 (quadratic), ..., up to degree $a-1$.

Orthogonal polynomial contrasts are tested as are any other contrasts. If ψ_{lin} is the linear contrast estimated by C_{lin} then $H_0 : \psi_{lin} = 0$ is rejected if

$$\frac{MS_{C_{lin}}}{MS_E} > F_\alpha(1, N-a)$$

- Testing the null hypothesis $H_0 : \psi_{lin} = 0$ is equivalent to testing $H_0 : \beta_1 = 0$ in model $(*)$. Therefore, if we reject H_0 we know that a linear model is effective in explaining the variability in the treatment means (the trend in the means). It is still possible that a higher degree polynomial will be more effective in explaining this variability.

- That is, we know that the β_1 term is necessary, but not if it is sufficient; we still might do better by adding one or more of the terms corresponding to $\beta_2, \ldots, \beta_{a-1}$.

- We can test for nonlinearity ($H_0 : \beta_2, \ldots, \beta_{a-1}$ all equal to 0) by examining the **lack of fit** based on the linear model:

  $$SS_{L.O.F.} = SS_{Trt} - SS_{C_{lin}} = SS_{C_{quad}} + SS_{C_{cub}} + \cdots + SS_{C_{(a-1)ic}}$$

  which has degrees of freedom

  $$\text{d.f.}_{L.O.F.} = \text{d.f.}_{Trt} - \text{d.f.}_{C_{lin}} = a - 2$$

An α-level test for lack of fit can be performed by comparing

$$F = \frac{SS_{L.O.F.}/\text{d.f.}_{L.O.F.}}{MS_E} = \frac{MS_{L.O.F.}}{MS_E}$$

to $F(\text{d.f.}_{L.O.F.}, N - a)$. If the resulting $p-$value is less than α, the appropriate model is nonlinear.

The above test of lack of nonlinearity is equivalent to testing

$$H_0 : \psi_{quad} = \psi_{cub} = \cdots = \psi_{(a-1)^{\text{ic}}} = 0$$

That is, the test of nonlinearity can be done with a single simultaneous test that all orthogonal polynomial contrasts except the linear one are 0.

- In SAS, this can be done with the CONTRAST statement by specifying all orthogonal polynomial contrasts other than the linear one on a single CONTRAST statement, with each contrast specification separated by commas. See gasadd1.sas.

Typically, it is of interest to address the question of whether the relationship between the mean of y and the quantitative factor is linear or not. That is usually where the use of orthogonal polynomials ends. But occasionally, if the relationship is not linear, it may be of interest to check whether or not the relationship is quadratic.

That is, if the relationship is nonlinear we may want to proceed and compare

$$\frac{MS_{C_{quad}}}{\text{MS}_\text{E}} \quad \text{to} \quad F(1, N-a).$$

If

$$\frac{MS_{C_{quad}}}{\text{MS}_\text{E}} > F_\alpha(1, N-a)$$

then the β_0, β_1 and β_2 terms belong in the model. To determine whether higher order terms are also necessary, we test for lack of fit based on the second degree (quadratic) model. That is, we would then need to test

$$H_0 : \psi_{cub} = \cdots = \psi_{(a-1)^{\text{ic}}} = 0$$

- We can continue in this manner if necessary to check for cubic, quartic, etc. relationships.
- **Caveat:** But remember, two points determine a line, three points determine a quadratic curve, four points determine a cubic curve, etc. That means when $a = 2$, a straight line relationship is guaranteed to hold just as an artifact of only studying two levels of the factor! That doesn't mean the true relationship is linear. Similarly, the means for a three-level quantitative factor ($a = 3$) are guaranteed to be perfectly fit by a quadratic curve. Etc.
  - So relationships found via orthogonal polynomials should be taken with a grain of salt when there are not many factor levels.
- So, unless a is large (5 or more, say) and there's good reason to hypothesize a quadratic or higher-order relationship, it is best to only test whether or not that relationship is linear.

Gasoline Additive Example:

- Here we do some computations "by hand", but also refer to gasadd1.sas and its output to see how to do these things in SAS.

  From Table D.6, we see that the linear contrast coefficients are (-2,-1,0,1,2). From the output, $SS_{C_{lin}} = 5.476$ on 1 d.f., so $MS_{C_{lin}} = 5.476$ and we conclude that the trend is at least linear since

$$F = \frac{MS_{C_{lin}}}{MS_E} = \frac{5.476}{0.225} = 24.37$$

  exceeds its critical value. We can test

$$\begin{aligned} & H_0 : \text{mean octanes are linear in the amount of additive} \\ \text{versus} \quad & H_1 : \text{mean octanes are nonlinear} \end{aligned}$$

  by computing

$$SS_{L.O.F.} = SS_{Trt} - SS_{C_{lin}} = 6.108 - 5.476 = 0.632$$

  on $\text{d.f.}_{L.O.F.} = 3$. We do not reject H_0 since

$$F = \frac{MS_{L.O.F.}}{MS_E} = \frac{0.632/3}{0.225} = 0.94$$

  is not significant.

  Conclusion: Octane is linear in the amount of additive over the range 0–4 cc/l. We can use SAS Proc Reg to obtain the relationship

$$\hat{y}_{ij} = 90.97 + 0.37x_i$$

Multiple Comparisons Procedures

Recall that when performing a hypothesis test, there are two types of errors that we could make: Type I errors and Type II errors.

Our Conclusion:		The Truth: H_0 is True	The Truth: H_0 is False
	Fail to Reject H_0	Correct	Type II Error
	Reject H_0	Type I Error	Correct

- Recall also that the significance level α of a test is the type I error rate for that test.

The problem: Suppose we have K hypotheses,

$$H_1, H_2, \ldots, H_K,$$

and we choose to test each one at significance level α.

If $H_1, \ldots, H_K$ are all true, then the probability of making at least one type I error (rejecting at least one of the H_i's) is larger (often much larger) than α.

Suppose the K tests statistics appropriate for testing $H_1, \ldots, H_K$ happen to be independent. In this case,

$$\begin{aligned}\Pr[\text{at least one } H_i \text{ rejected}|\text{all } H_i \text{ are true}] &= 1 - \Pr[\text{all } H_i \text{ accepted}|\text{all } H_i \text{ are true} \\ &= 1 - (1-\alpha)^K > \alpha.\end{aligned}$$

So as $K \to \infty$, Pr[at least one Type I error] $\to 1$.

- It is not as easy to compute the type I error rate for the family of hypotheses $H_1, \ldots, H_K$ when their test statistics are not independent, but the problem persists: In general, (unless the tests are perfectly correlated) the probability of at least one type I error in the family of inferences will be greater than the per-inference type I error rate.

Multiple comparison procedures are designed to allow one to conduct several inferences (perform several tests or compute several confidence intervals) without exceeding a pre-specified error rate for the entire "family" of inferences.

Error Rates:

Comparisonwise Error Rate (CWER): the probability of a type I error for any single comparison (or inference).

Familywise Error Rate (FWER): In a collection (family) of inferences (e.g., tests) the FWER is the probability of making at least one type I error among all inferences in the family.

- The FWER is called the **experimentwise error rate** when the family is the collection of all inferences performed when analyzing the data from the experiment.

False Discovery Rate (FDR): By a "discovery", we mean a rejection of the null hypothesis. The FDR is the number of false discoveries (type I errors) divided by the total number of discoveries in the family of tests.

Strong Familywise Error Rate (SFWER): the probability of making one or more false discoveries. Controlling the SFWER is the most stringent criterion for simultaneous testing.

- Difference between FWER and SFWER: FWER is the type I error rate for the family assuming all hypotheses in the family are true; SFWER does not assume all hypotheses are true, it is just the probability of one or more incorrect rejections of a null hypothesis in the family.

Error Rate for Simultaneous Confidence Intervals: If we are computing confidence intervals for K parameters (might be contrasts), this is the probability that at least one interval does not cover its corresponding parameter. This is the most stringent criterion for simultaneous construction of confidence intervals.

Two most commonly controlled error rates are CWER (that is, no adjustment for simultaneity) and the FWER.

Procedures for Controlling Simultaneous Error Rates:

1. Bonferroni
2. Fisher's Least Significant Difference (LSD)
3. Scheffé's
4. Tukey's Honest Significant Difference (Pairwise Comparisons)
5. Others: REGWR, SNK, Duncan's, Dunnet's, etc. are discussed in Ch.4 of our text.

1. **The Bonferroni Method:**

- Very simple, general method, with wide applicability.
- Can be very conservative, though, especially as the number of inferences, K, grows; so most useful for small K.
- Controls the SFWER (and, consequently, also controls FDR and FWER) and produces simultaneous confidence intervals.

Suppose we are interested in testing K hypotheses

$$H_1 : \psi_1 = 0, \quad H_2 : \psi_2 = 0, \quad \ldots, \quad H_K : \psi_K = 0$$

on contrasts

$$\psi_1 = \sum_i c_{1i}\mu_i, \quad \psi_2 = \sum_i c_{2i}\mu_i, \quad \ldots, \quad \psi_K = \sum_i c_{Ki}\mu_i.$$

Let

$$\begin{aligned} R_j &= \text{event of rejecting the hypothesis } H_j \\ T_j &= \text{event that } H_j \text{ is true} \end{aligned} \qquad j = 1, \ldots, K.$$

Then

$$\begin{aligned} \text{SFWER} &= \Pr[\text{reject at least one true } H_j] \\ &= \Pr\left[\bigcup_{j=1}^{K}(R_j \cap T_j)\right] \\ &\leq \sum_{j=1}^{K} \Pr(R_j \cap T_j) \quad \text{(by Bonferroni Inequality)} \\ &\leq \sum_{j=1}^{K} \underbrace{\Pr(R_j|T_j)}_{\text{CWER for } H_j} \end{aligned}$$

Idea of Bonferroni method is just to split the SFWER α equally and allocate α/K to $\Pr(R_1|T_1), \Pr(R_2|T_2), \ldots, \Pr(R_K|T_K)$.

With this strategy, we get

$$\text{SFWER} \leq \sum_{j=1}^{K} \Pr(R_j|T_j) = \sum_{j=1}^{K} \frac{\alpha}{K} = \alpha$$

so that by setting CWER $= \alpha/K$ for each comparison, we ensure that SFWER $\leq \alpha$.

Operationally, this just means that we test each of the $j = 1, \ldots, K$ hypotheses as follows: reject $H_j : \sum_{i=1}^{a} c_{ji}\mu_i = 0$ if

$$t = \frac{|\sum_i c_{ji}\bar{y}_{i\cdot}|}{\sqrt{MS_E \sum_i \frac{c_{ji}^2}{n_i}}} > t_{\alpha/(2K)}(N-a).$$

- That is, we just do the usual t-test but compare t with $t_{\alpha/(2K)}(N-a)$ rather than $t_{\alpha/2}(N-a)$. Equivalently, compare the usual p-value to α/K rather than α.

Simultaneous confidence intervals with a controlled error rate for the contrasts $\psi_1, \ldots, \psi_K$ can be formed based on the Bonferroni method. One can be at least $100(1-\alpha)\%$ confident that $\psi_1, \ldots, \psi_K$ will <u>all</u> be contained in the intervals

$$\sum_i c_{ji}\bar{y}_{i\cdot} \pm t_{\alpha/(2K)}(N-a)\sqrt{MS_E \sum_i \frac{c_{ji}^2}{n_i}}, \quad j = 1, \ldots, K.$$

- Any critical value of the t distribution can be obtained with the tinv function in SAS. Suppose you want the upper $.05/(2K)$ critical value of a t distribution with 9 d.f., and where $K = 3$. This value is called the $1 - .05/(2(3)) = 1 - .008333 = .9917^{\text{th}}$ quantile of the $t(9)$ distribution. It can be obtained with the following short SAS program:

```
data junk;
  tcrit=tinv(.9917,9);
run;
proc print;
run;
```

- There is also an finv function for quantiles of the F distribution. To obtain the upper α^{th} critical value of the $F(a,b)$ distribution (the $1-\alpha^{\text{th}}$ quantile) use finv$(1-\alpha, a, b)$.

2. **Fisher's LSD Method:**

- Controls the FWER but not the FDR or SFWER. Does not produce simultaneous confidence intervals with controlled error rate.

- Fisher's LSD is typically substantially more powerful than Bonferroni, Scheffé.

Step 1. Do an F-test of $H_0 : \mu_1 = \cdots = \mu_a$ at level α. If reject do step 2; otherwise stop.

Step 2. Test contrasts, each with CWER α. That is, do t-test for contrast or F-test for contrast (either one) at level α.

- By making step 2 conditional on rejecting H_0 in step 1, we have controlled the FWER. However, if H_0 does not hold, then the combined type I error rate with this procedure may be $> \alpha$ (i.e., we haven't controlled the SFWER).

- Fisher's LSD is sometimes called the "protected LSD". This method should be distinguished from ordinary LSD, which just does step 2 without first checking the overall ANOVA F test for significance. Ordinary LSD is not a multiple comparison technique at all (it ignores multiplicity).

- The term "LSD" stands for least significant difference. The reason for this terminology is as follows:

  Suppose the contrasts that we are interested in are **pairwise contrasts**.

  - By a pairwise contrast we mean a contrast of the form $\mu_j - \mu_k$ (a comparison between a pair of treatment means).

  Then the LSD procedure is a t-test of $H_0 : \mu_j - \mu_k = 0$, which has rejection rule: reject H_0 if

$$|\bar{y}_{j\cdot} - \bar{y}_{k\cdot}| > t_{\alpha/2}(N-a)\sqrt{MS_E\left(\frac{1}{n_j}+\frac{1}{n_k}\right)}$$

$$= \underbrace{t_{\alpha/2}(N-a)\sqrt{MS_E\left(\frac{2}{n}\right)}}_{\text{the "LSD"}} \quad \text{in the balanced case.}$$

  Notice that in the balanced case where $n_1 = n_2 = \cdots = n_a = n$ the right-hand side does not depend on either j or k. This means that no matter which pair of means we compare, we conclude that they are significantly different from one another if the difference between the corresponding sample means exceeds the LSD.

  - That is, the LSD is the least difference between the sample means that will result in concluding that they are significantly different from one-another.

E.g., in the gasoline additives example, the LSD is

$$LSD = t_{\alpha/2}(N-a)\sqrt{MS_E\left(\frac{2}{n}\right)}$$

$$= 2.131\sqrt{.2247\left(\frac{2}{4}\right)} = .7144,$$

which means that any pair of sample means that differ by more than .7144 will be declared significantly different from one-another by the LSD method.

3. **Scheffé's Method:**

- Appropriate for testing any number of contrasts, including those that were suggested by examining the data (i.e., o.k. for data-snooping)!
- Produces simultaneous C.I.'s with controlled error rate and controls SFWER for tests.
- Can be very conservative (have low power for detecting real differences).

Let $\psi = \sum_{i=1}^{a} c_i \mu_i$ represent a generic population contrast with sample version $C = \sum_{i=1}^{a} c_i \bar{y}_{i\cdot}$. Scheffé has shown that

$$\Pr\left\{ \max_{C} \underbrace{\frac{(C-\psi)^2}{MS_E \sum_i c_i^2/n_i}}_{F \text{ test statistic for contrast}} > (a-1)F_\alpha(a-1, N-a) \right\} = \alpha. \qquad (*)$$

Therefore, for <u>any</u> contrast C we will have

$$\Pr\left\{ \frac{(C-\psi)^2}{MS_E \sum_i c_i^2/n_i} > (a-1)F_\alpha(a-1, N-a) \right\} \leq \alpha.$$

Therefore, if we use $(a-1)F_\alpha(a-1, N-a)$ as our critical value for our F-test of $H_0 : \psi = 0$ rather than the usual critical value of $F_\alpha(1, N-a)$, then we can go data snooping and test as many contrasts as we want, while maintaining SFWER $\leq \alpha$.

Simultaneous confidence intervals based on $(*)$ for any set of contrasts $\psi_1 = \sum_i c_{1i}\mu_i, \psi_2 = \sum_i c_{2i}\mu_i, \ldots$ can be formed. One can be at least $100(1-\alpha)\%$ confident that $\psi_1, \psi_2, \ldots$ will be contained in the intervals

$$\sum_i c_{qi}\bar{y}_{i\cdot} \pm \sqrt{(a-1)F_\alpha(a-1, N-a)MS_E \sum_i \frac{c_{qi}^2}{n_i}}, \quad q = 1, 2, \ldots$$

4. **Tukey's Honest Significant Difference**

- Strictly for pairwise comparisons.
- Produces simultaneous confidence intervals for pairwise differences that control the combined type I error rate. For tests, controls the SFWER.
- Tends to be more powerful than Bonferroni (produces shorter confidence intervals for the pairwise differences between means).

Tukey's method is based on a statistic called the **Studentized range** of the treatment means. This is just the range of the means divided by their standard errors:

$$\text{Studentized range} = \max_j \frac{\bar{y}_{j\cdot}}{\sqrt{MS_E/n}} - \min_k \frac{\bar{y}_{k\cdot}}{\sqrt{MS_E/n}}.$$

- Here, we have assumed that we are in the balanced case where the number of replicates per treatment is n for all treatments.
- The Studentized range is a measure of how far apart the largest and smallest treatment means are.

The probability distribution of the Studentized range under $H_0 : \mu_1 = \cdots = \mu_a$ can be calculated. It is tabulated in Table D.8 of our text and can be computed in statistical software packages such as SAS.

This Studentized range distribution depends on two parameters: a, the number of groups, and d.f._E, the degrees of freedom for error. Let $q_\alpha(a, \text{d.f.}_E)$ denote the upper α^{th} critical value of this distribution.

Tukey's "Honest" significant difference (HSD) is an alternative to using the LSD for pairwise comparisons. In the balanced case, the HSD is

$$\text{HSD} = \frac{q_\alpha(a, \text{d.f.}_E)}{\sqrt{2}} \sqrt{MS_E \left(\frac{1}{n} + \frac{1}{n}\right)}.$$

- In the balanced case, we conclude that a pair of means μ_j, μ_k are significantly different from one another if $|\bar{y}_{j\cdot} - \bar{y}_{k\cdot}| > HSD$.

In the unbalanced case, the HSD is no longer the same for all pairs on means. In this case, the HSD for comparing μ_j, μ_k is

$$\text{HSD}_{jk} = \frac{q_\alpha(a, \text{d.f.}_E)}{\sqrt{2}} \sqrt{MS_E \left(\frac{1}{n_j} + \frac{1}{n_k}\right)},$$

and we conclude that a pair of means μ_j, μ_k are significantly different from one another if $|\bar{y}_{j\cdot} - \bar{y}_{k\cdot}| > HSD_{jk}$.

Tukey's method can also be used to compute simultaneous confidence intervals for $\mu_j - \mu_k$ for all pairs μ_j, μ_k with combined error rate of α. These intervals are given by

$$\bar{y}_{j\cdot} - \bar{y}_{k\cdot} \pm \text{HSD}_{jk} \quad \text{for each pair } j, k.$$

Recommendations:

For simultaneous confidence intervals on multiple contrasts or other linear combinations of the model parameters:

1. If you wish to form confidence intervals on all pairwise differences among the treatment means, use Tukey's HSD method; for simultaneous intervals on other quantities, use Bonferroni intervals.

For hypothesis tests corresponding to planned comparisons:

2. Use Tukey's HSD for all pairwise comparisons; Dunnett's method for all pairwise comparisons with a single reference mean (the control, standard, best or worst treatment); and use Fisher's LSD for all other planned contrasts.

For unplanned comparisons and data snooping:

3. Use Scheffé's method.

- These recommendations are not universally agreed upon, but they are what I'll expect you to follow in this course when deciding which multiple comparison procedure to use for homework and test problems unless I tell you explicitly to use some other approach. Other recommendations are given in our text and elsewhere.

- There are many opinions about multiple comparisons, many of which are equally valid. Although for a given error rate it is often possible to choose an optimal procedure, it is often unclear which error rate should be controlled in a given situation.

- The choice of error rate and α-level is really not a statistical issue, nor even a question that is amenable to objective analysis. Similarly, the choice of the "family" of inferences for which the error rate is to be controlled is also subjective. In some contexts there are guidelines based on custom that can be followed, but ultimately these choices are up to the researcher to decide based upon his/her own tolerance for the risk of making incorrect conclusions.

Power Analysis/Choice of Sample Size

Error Types:

There are two types of errors that can be made in choosing between H_0 and H_A:

I. **Type I Error:** Reject H_0 when H_0 is true.

II. **Type II Error:** Fail to reject H_0 when H_0 is false.

- The probability of I is called α, and it is usually fixed at $\alpha = .05$ or $\alpha = .01$ by the investigator.
- The probability of II is called β and it cannot be fixed (set) by the investigator.
  - Note that the **power** of a hypothesis test is

$$\begin{aligned} \text{power} &= \Pr(\text{reject } H_0 | H_0 \text{ is false}) \\ &= 1 - \beta, \end{aligned}$$

    or the probability of establishing our scientific proposition given that it is true.
- Want power to be high (want to have high probability of detecting the effect we're looking for) and β to be low.
  - However, power depends upon lots of things we can't control and a few we can.

Power/Sample Size Analysis for the two-sample t test:

Suppose we have an active treatment (a new drug say) and a control treatment (placebo, say). Suppose we have a random sample of $2n$ subjects, n randomly assigned to active treatment, and n to control group.

Let μ_1, μ_2 be population means under the two treatments, and let σ_1, σ_2 be the corresponding population standard deviations.

- For simplicity, suppose the distribution of the response in the two populations is normal, with same s.d.: $\sigma_1 = \sigma_2 = \sigma$.

Picture:

- See http://wise.cgu.edu/power_applet/power.asp

Power depends upon:

(A) Difference in the means, $|\mu_1 - \mu_2|$. (Unknown)

  - A big difference is easier to detect than a small one.

(B) σ. (Unknown)

  - If there's lots of variability in the population, it will be hard to detect a difference in the means.

- For power/sample size calculations, (A) and (B) are often combined into a single quantity called **effect size.**

(C) Sample size n. (Can be chosen, so known)

  - Sample size is a measure of the amount of information available about the population. The more information we have, the more powerful our inferences about the population will be.

- Sometimes, sample size is fixed by constraints. Then question is how much power will we have for that sample size.

- More often, we fix power (we want to design a study with a certain level of power - 80%, say), calculate the power for each of several increasing values of n, and then choose the smallest n that gives us the power that has been chosen.

(D) α, the Type I error rate. (Can be chosen, but in practice is usually fixed by convention)

  - High α values make it easier to reject H_0, which will certainly increase power, but at the expense of Type I error.

Typical Power Analysis for Two-sample t Test:

1. Fix $\alpha = .05$ (or .01, or some other value).

2. Assume values for $\mu_1 - \mu_2$ and σ based upon previous research (yours or from literature), educated guesses, and considerations of clinical/practical/scientific (not statistical) significance.

3. Select a desired level of power, 80%, say. This may be dictated by the funding agency, if power analysis is part of a grant proposal, or chosen by the investigator (represents risk of not being able to detect true effect, so how much risk are you comfortable with?).

4. Compute power for each of several values of n. Select smallest n that gives you power $\geq$ the level selected in 3.

Typically, step 2 is the hardest part. One way that people often try to simplify this step is to specify a generic or "canned" *effect size*.

Effect size. In the two-sample t-test, it can be shown that although the power of the test depends on $|\mu_1 - \mu_2|$ and σ, it only depends upon these quantities through their ratio:

$$\frac{|\mu_1 - \mu_2|}{\sigma} \equiv d$$

which is known as the effect size* for the 2-sample t test.

- In other experimental design/statistical hypothesis testing frameworks quantities analogous to d can sometimes (but not always — especially in more complex contexts) be identified and defined as "the effect size".

There are some famous definitions (due to Cohen, 1988) of what constitutes a "small", "medium" or "large" effect size. These effect size descriptions (and their misuse) are especially common in the social sciences.

It is tempting, but not recommended, to avoid consideration of $|\mu_1 - \mu_2|$ and σ separately and instead simply specify d, the effect size.

I especially encourage you to avoid taking a generic "small", "medium", or "large" effect size as your target (e.g., computing the sample size necessary to achieve 80% power to detect a small/medium/large effect size d) without close consideration of the specific research context at hand. See the article by R. Lenth handed out in class for more on this topic.

* Or, more accurately, the standardized effect size for the 2-sample t test.

How is power calculated?

Remember, power is the probability of rejecting H_0 given that it is false. Whether or not H_0 is rejected is determined by a comparison of a test statistic with a critical value.

- Therefore, power computations are based upon the statistical test employed in the analysis!

In the two-sample t test situation, the null and alternative hypotheses are

$$H_0 : \mu_1 - \mu_2 = 0, \quad \text{vs.} \quad H_A : \mu_1 - \mu_2 \neq 0,$$

(assuming a two-tailed alternative) and the rejection rule is: reject H_0 if

$$t = \frac{|\bar{y}_1 - \bar{y}_2|}{s_P\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} > t_{\alpha/2}(n_1 + n_2 - 2),$$

where

$\bar{y}_1, \bar{y}_2$	are the sample means in groups 1 and 2
s_P	is the pooled sample standard deviation
n_1, n_2	are the sample sizes in the two treatment groups

- We can allow $n_1 \neq n_2$, but typically it is best to design **balanced** experiments, so we allocate equal sample sizes $n_1 = n_2 = n$ to the two groups. Thus, the rule becomes: reject H_0 if

$$t = \frac{|\bar{y}_1 - \bar{y}_2|}{s_P\sqrt{\frac{2}{n}}} > t_{\alpha/2}(2n - 2).$$

Now that we have the test statistic, it is clear from the definition of power that

$$\text{power} = \Pr\left\{ \frac{|\bar{y}_1 - \bar{y}_2|}{s_p\sqrt{\frac{2}{n}}} > t_{\alpha/2}(2n-2) \middle| \mu_1 - \mu_2 \neq 0 \right\}. \qquad (*)$$

- Now it is clear why power depends upon $|\mu_1 - \mu_2|$ (affects size of $|\bar{y}_1 - \bar{y}_2|$), σ (affects size of s_P), n and α.
- Actually, note that $|\bar{y}_1 - \bar{y}_2|$ and s_p only figure into the rejection rule through their ratio, which is the sample version of the effect size in this context:

$$\text{effect size} = \frac{|\mu_1 - \mu_2|}{\sigma}.$$

How is the probability in (*) calculated?

Under the classical assumptions of a t test (normal samples, equal variances, independence), statistical distribution theory can be used to show that under H_A, the t statistic in (*) follows a **non-central t distribution.**

- This is a parametric distribution (like the normal distribution) with two parameters: the degrees of freedom ($2n-2$ here), and the non-centrality parameter (a function of the effect size).

Therefore, just like we use tables of the normal distribution to figure out the probability that a normal random variable exceeds some given value, we can figure out the probability in (*) from tables or computer programs that give non-central t probabilities.

Example — An Actual Power Analysis:

Suppose we wish to conduct a study to determine whether the mean blood pressure of 35- to 39-year-old oral contraceptive (OC) users is different from that of comparable non OC users.

Response: Systolic blood pressure.

Design: Two equal sized samples of OC users and non OC users will be taken from among health workers in Boston, MA.

Statistical Analysis: Two-sample t test assuming equal, but unkown variances.

- We assume that systolic BP among OC users is normal with mean μ_1 and variance σ^2 and systolic BP of non-OC users is also normal with mean μ_2 and variance σ^2.
- We are interested in testing

$$H_0 : \mu_1 - \mu_2 = 0 \quad \text{vs.} \quad H_A : \mu_1 - \mu_2 \neq 0$$

  and we do not know $|\mu_1 - \mu_2|$ or σ^2.

Previous Research:

- In a pilot study, 4 OC users and 6 non-OC users in the targeted age range were identified from the researcher's home hospital and tested for systolic BP.
- It was found that the mean and SD among OC users were $\bar{y}_1 = 132.86$ and $s_1 = 15.34$ and among non-OC users the mean and SD were $\bar{y}_2 = 127.44$, $s_2 = 18.23$.

Assumptions for Sample Size Calculation:

Although there was a difference of $\bar{y}_1 - \bar{y}_2 = 132.86 - 127.44 = 5.42$ in mean BP among OC users and non-users in the pilot study, this is not necessarily the value we want to assume in computing $|\mu_1 - \mu_2|$ the unstandardized effect size we wish to be able to detect.

Instead $|\mu_1 - \mu_2|$ should be chosen based upon the answer to questions such as:

i. How much of a difference in mean BP do we expect OC use to be associated with? or
ii. How much of a difference in mean BP would be clinically significant (e.g., correspond to an elevated health risk)?

Suppose that we decide that a 5% increase (or reduction) in BP associated with OC use would be clinically significant. If we assume that μ_2 the population mean systolic BP for non OC users is 127.44 (our sample value), then a 5% increase in mean BP would be $1.05(127.44) - 127.44 = 6.37$ units, so we take $|\mu_1 - \mu_2| = 6.37$.

Because we don't know σ, we'll estimate it by the pooled standard deviation from our pilot study:

$$s_P = \sqrt{\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}} = \sqrt{\frac{(4-1)(15.34)^2 + (6-1)(18.23)^2}{4 + 6 - 2}} = 17.2$$

These values give an effect size of

$$\frac{|\mu_1 - \mu_2|}{\sigma} = 6.37/17.2 = .37$$

- This effect size can be used to determine the non-centrality parameter of the t distribution describing the distribution of the two-sample t-statistic under H_A. Then tables, plots (called operating characteristic curves), or computer functions giving non-central t probabilities can be used to look up the power.

- Alternatively, computer programs can be used that short-cut this process.
  - SAS does sample size and power computations in its "Analyst Application".
  - There are also numerous online sample size/power tools. In particular, we will use Russ Lenth's Java Applets for Power and Sample Size (http://www.stat.uiowa.edu/∼rlenth/Power/index.html).
- SAS Analyst is started from the Solutions menu in SAS. Click on "Solutions", then "Analysis", then "Analyst" to start this application.
- The click on the "Statistics" menu, click on "Sample Size" and click on whichever statistical analysis for which you want to compute power or sample size (in this case, click on "Two-Sample t-test...").
- These steps will lead to a dialog box in which you can specify the necessary assumptions for either a sample size calculation (for a given power value, or range of power values), or a power calculation (for a given sample size or range of sample sizes).
- Some programs require direct input of the effect size or noncentrality parameter. SAS Analyst requires the actual means and SDs. Note however, that if we change the means and SD without changing the effect size, the answer remains the same.
- For the values of $|\mu_1 - \mu_2|$ and σ assumed in this example, it turns out that 116 subjects per group are required to achieve at least 80% power using a two-sample t test and a significance level of $\alpha = .05$.

Sample Size/Power Analysis for the One-way Layout:

The two-sample design analyzed with a t test is among the simplest settings for a power analysis. Often, however, the design and statistical test will be more complex.

- E.g., Suppose we have a treatments to compare rather than just 2. This is a one-way layout (design), for which a one-way analysis of variance is appropriate.

The same principles underlying power analysis in the two-sample situation apply here as well, but the test statistic differs, and the issue of effect size is more complex.

In particular, for a groups, the probability of rejecting

$$H_0 : \mu_1 = \mu_2 = \cdots = \mu_a$$

depends not just on the difference between two means, but the difference between all possible pairs of means — i.e., on the spacing of the means.

In addition, the statistical test is now an F test, not a t test.

In fact, it can be shown that the t test is a special case of the F test corresponding to $a = 2$.

- So, power for a two sample comparison of means is based on the non-central t, and power for more general F tests based on the general linear model (e.g., the one-way anova model) is based on the non-central F distribution.

For the one-way ANOVA, the noncentrality parameter of the F distribution depends upon the spacing of the population means $\mu_1, \ldots, \mu_a$.

- So, power analysis for the one-way anova, can be done by specifying the values of all of the population means $\mu_1, \mu_2, \ldots, \mu_a$ and a value for the common standard deviation σ. Collectively, these specifications determine the effect size in the one-way anova context.

- Alternatively, one can take a conservative approach to power, and calculate the power assuming only a value for the difference

  $$|\mu_{\max} - \mu_{\min}|$$

  together with the assumption that the rest of the means are spread out (spaced) in the least favorable (for rejecting H_0) configuration possible. This places all of the means except $\mu_{\min}$ and $\mu_{\max}$ together, halfway between $\mu_{\min}$ and $\mu_{\max}$:

- Power computed here will be less than or equal to the power under any other possible configuration of the means with the same value of $|\mu_{\max} - \mu_{\min}|$. Therefore, in practice, we can expect the study actually conducted to be more powerful.

Example: Average daily weight gains compared among pigs receiving 4 levels of vitamin B_{12} in their diet.

Suppose that from past data we estimate $\sigma = 0.015$ lbs./day and we hope to be able to detect a difference $\mu_{\max} - \mu_{\min} \geq 0.03$ lbs/day.

Fix $\alpha = 0.05$ and suppose we want Power ≥ 0.90 for a balanced design.

- In SAS Analyst, now choose "Statistics", "Sample Size" and then "One-Way ANOVA...". The resulting dialog box asks for "CSS of means." What is required here is

$$CSS = \sum_{i=1}^{a} (\mu_i - \mu)^2.$$

- Under the least-favorable configuration of the means

$$\begin{aligned} CSS &= \{(\mu_{\max} - \mu_{\min})/2\}^2 + \{(\mu_{\max} - \mu_{\min})/2\}^2 + 0 + \cdots + 0 \\ &= (\mu_{\max} - \mu_{\min})^2/2 = (.03)^2/2 = .00045. \end{aligned}$$

- This leads to a minimum of $n = 9$ per group to achieve 90% power or more.
- Russ Lenth's software asks for "SD[treatment]" instead of CSS. This quantity is just the standard deviation of the assumed population treatment means or $\sqrt{CSS/(a-1)} = \sqrt{.00045/3} = .01225$. Notice that we obtain the same answer, $n = 9$ subjects per groups to achieve at least 90% power for this problem.

Power Analysis for other Types of Problems:

Power analysis for more complex designs/methods of analysis are available, but

1. additional assumptions/predictions about the data-generating mechanism are necessary;

2. the methods can be (much) more difficult to understand and to implement; and in some cases,

3. power analysis methods may not exist at all.

- Power analysis is easiest in the classical, normal-theory linear model for simple designs.

- Often, more complex designs can be simplified and thought of in terms of simpler designs for power/sample size purposes.

  - E.g., for a two-way layout with two factors with 2 and 3 levels, respectively, think of this as a one way layout with $2 \times 3 = 6$ treatments.

- Often, sample size is much more heavily determined by resource constraints and other practical matters than power. In many cases, the right sample size is "as many as you can afford".

  - In these cases, power analysis' role is to tell you whether the study is worth doing at all, or

  - to satisify a bureaucratic requirement (e.g., of a funding agency, or course instructor).

- Finally, we have emphasized power analysis/sample size determination for hypothesis tests. Alternatively, if we are more interested in forming confidence intervals, then sample size can be determined to obtain an interval that has a high probability of not exceeding a given width. Another possibility is to compute sample size or power in the one-way anova that is determined by the hypothesis test for a certain contrast of interest rather than by the overall F test of equal means. See Ch.7 of our text for discussion of these approaches.