

Introduction to Design and Analysis of Experiments

Stats 101B

What is this course about?

- How to collect data
- Specifically, how to collect data in a controlled experiment
- Good data collection leads to the ability to detect 'actual' differences between groups
- Can lead to cause-and-effect type conclusions

Statistical Paradigm

- Observed data = signal + random noise
- Mathematically:

$$y_i = \text{model} + \epsilon_i$$

where $\epsilon_i \sim N(0, 3)$

- If noise too large, signal is obscured
- We can deal with noise through
 - wise choice of statistical method
 - careful data collection procedures

Practical Applications

- Chicken and egg problem
 - To analyze data, you need to know how it was collected
 - To understand how to collect data, you need to know how to analyze it
- With the rise of 'Big Data' where do statisticians/data scientists see data collection in a controlled environment?



Employee Stories



Louise
Mgmt. Service
▶ WATCH VIDEO

Research/
Contracts/Grants

▶ FIND OUT MORE

Weather & Surf Report

Santa Monica	65°
H: 70° L: 48°	
5 DAY	
TUESDAY NIGHT	
73° 48°	
WEDNESDAY NIGHT	
76° 50°	
THURSDAY NIGHT	
74° 51°	
FRIDAY NIGHT	
71° 52°	

Click To Apply**Statistician****Job Duties**

Select appropriate test formulas and perform standard computations such as multiple and partial correlation coefficients, t-tests, Chi-square, nonparametric tests, and analysis of variance working from written or verbal instructions; prepare control cards to adopt particular computer programs for selected research problems, which may include such factors as n-way analysis of variance, analysis for unbalanced design, discriminant analysis and step-wise regression; perform complicated computations independently on research problems such as regression analysis, analysis of covariance where the design is not complex and the mathematical model is relatively simple; participate in meetings with departmental representatives for the purpose of modifying codes, designing and changing forms, and making reports during selected long-term projects; process pre-coded data and the preparation of related reports; facilitate in report preparation and statistical review of existing reports; and communicate with outside database sources for secure transmission of properly formatted data. Position is for approximately 10 hours/week. This is a 25% part-time position that will last until 7-30-14.

Job Qualifications

Prior experience working with statistics and statistical models, or related professional activities. Excellent organizational skills. Ability to summarize and clearly present statistical, financial and programmatic information. Ability to effectively identify trends and potential problems, recommend solutions, and execute appropriate actions. Excellent computer skills and working knowledge of Microsoft Office Programs required. Experience writing and preparing reports and grant proposal. Excellent verbal and written communication skills. Working knowledge of both medical and gambling terminology required. Knowledge of UCLA policies and procedures preferred. Bachelor's degree or equivalent work experience required. Master's degree preferred.

Click To Apply

[Back to Search Result](#) ▶ [Bookmark Job](#) ▶ [Search Again](#) ▶

Summary Information

Job Title: Statistician
UCLA Title: Statistician
Job Num.: H65288
Work Hours: 10 hrs./week flexible



Careers

Search by Keyword

SEARCH JOBS[Email similar jobs to me](#)

Share this Job



Data Scientist Job

Apply now »**Date:** Dec 14, 2013**Location:** New York City, NY, US

You may be a great fit for our team if you are excited about working on high impact, real world problems using huge (sometimes slightly messy) data sets, including billions of transactions, to unlock valuable insights and power new products.

Responsibilities:

- Frame and conduct complex analyses and experiments using tremendously large (e.g. 10^6 to 10^{10} records), complex (not always well-structured, highly variable) data sets
- Answer product questions by using appropriate statistical techniques on available data
- Analyze and interpret the results of product experiments; ! ! !
- Design and implement scripts, programs, databases, and other software components
- Draw conclusions and effectively communicate findings with both technical and non-technical team members
- Review relevant academic and industry research to identify useful algorithms, techniques, libraries, etc.
- Share your knowledge by communicating your findings through sensible presentation and accessible language and mentor staff on relevant procedures and techniques
- Work closely with a product engineering team to identify and answer important product questions;
- Develop best practices for instrumentation and experimentation and communicate those to product engineering teams.

Requirements:

- Bachelors Degree in Math, Statistics, Computer Science, or other quantitative discipline
- 3+ years experience in R, Python, Java, or other languages appropriate for large scale analysis of numerical and textual data
- 2+ years experience with data mining, machine learning, statistical modeling tools and underlying algorithms
- 2+ years experience with relational databases and SQL
- 2+ years experience working with extremely large data sets

CAREER SEARCH

Nike does more than outfit the world's best athletes. We are a place to explore potential, obliterate boundaries, and push out the edges of what can be.

[SEARCH JOBS](#)

[Nike Jobs](#) > [Product Development Jobs](#) > [United States Product Development Jobs](#)

JOB DESCRIPTION

[APPLY NOW](#)

WEAR TEST ANALYST I

Job Location: Portland, OR

As our Wear Test Analyst I for Global Apparel, you will utilize sensory testing protocols to measure athlete's perceptions, ultimately enhancing the final product within Development and Innovation calendar timelines. You will work with the Wear Test Manager to prioritize tests based on innovation and potential impact to business. You will help manage, execute and monitor the product testing seasonal plan from innovation research through commercialization. You will formalize and present test findings while ensuring they are archived for future knowledge. You will ensure product is tested to Nike performance standards, and be responsible to recruit and maintain a pool of qualified testers. In addition, you will ensure constant, effective communication with regions, business units, and promo and innovation teams in order to align goals, timelines, protocols, and potential outcomes of testing.



[CAREERS AT NIKE, INC.](#)

[CAREER SEARCH](#)

[CAREER AREAS](#)

[BENEFITS](#)

[LOCATIONS](#)

[INTERNSHIPS](#)

[TEMPORARY OPPORTUNITIES - U.S.](#)

Qualifications

Requirements for the position include:

- Bachelor's degree, preferably in a science related field
- Minimum 1 year relevant work experience
- In lieu of a Bachelor's degree, a minimum of 4 years' work experience
- Strong written and verbal communication skills
- Demonstrated computer skills, including familiarity with word processing, spreadsheets, databases, and statistics programs
- Experience statistically analyzing data (both subjective and objective) and making recommendations
- Strong information seeking skills and drive for results
- Ability to multitask and manage a varied workload
- Proven influencing skills
- Ability to communicate with and draw out information from athletes at all levels
- Knowledge of statistics, statistical process control and design of experiments is a plus
- Passion for and knowledge of sports is a plus
- Sensory evaluation experience is preferred

NIKE, Inc. is committed to employing a diverse workforce. Qualified applicants will receive consideration without regard to race, color, religion, sex, national origin, age, sexual orientation, gender identity, gender expression, veteran status, or disability.

COMMON SEARCHES

retail apparel Design Converse Merchandising Hurley

Digital

SIGN UP FOR JOB ALERTS

ENTER EMAIL 

Get a jump on your career at Nike Inc. Find out about the hottest opportunities to join our team.

Job ID: 00110409



Store

Mac

iPod

iPhone

iPad

iTunes

Support



Jobs at Apple

Corporate

Retail

Students

My Profile

Search

127 job(s) found

Sign In

My Favorites ▾

Clear all filters

Show filters ▾

Save this mix ▾

Showing 1-20 of 127

Previous

Next

Job Title
Statistical Engineering Analyst – Wireless Design
iOS Data Scientist
Senior Analyst, Advanced Analytics
Senior Analyst, Advanced Analytics
Senior Analyst, Advanced Analytics
Statistician/Algorithm Analyst
iOS Battery Life Tools/Analytics Manager
Senior Product Engineer
SoC Product Engineer
Algorithms R&D Intern
Maps Data Sciences Engineer
Location Analytics Software Engineer
iOS Location Analytics Software Engineer
Kernel and Performance Tools Enaineer

Statistician/Algorithm Analyst

Job Number: 26639090

Posted: Dec. 2, 2013

Santa Clara Valley, California, United St...

Weekly Hours: 40.00

Job Summary

We are a cross functional group of statisticians and subject matter experts tackling a multitude of challenges across Apple Maps and iTunes. We focus on the algorithmic evaluation of everything from search and recommendations to routing and geocoding. Our group works directly with the engineering teams responsible for ranking, recommendation and other alorithms and tackles every type of problem in statistics, including design of experiments, AB testing, machine learning, importance sampling, hierarchical bayesian models for spelling and synonyms, etc.

Key Qualifications

- Solid experience with data analysis in R, S+, or Matlab
- Experience with MySQL or other DB
- Demonstrated ability to work independently
- Experience with at least one scripting language (preferably Python)
- Ability to draw conclusions from data and recommend actions
- Excellent written and verbal presentation skills



All Jobs

My Applications

Starred



★ Statistician/Engineering Analyst

Mountain View, CA, USA

Technical Infrastructure · Full-time

Know someone who would be interested?

APPLY NOW**Find connections**

Sign in to see your connections at Google

As a Statistician/Engineering Analyst, you will evaluate and improve Google's products, such as Ads and Search. You will collaborate with a multi-disciplinary team of engineers and analysts on a wide range of problems. This position will bring analytical rigor and statistical methods to the challenges of measuring Search and Ads quality, improving consumer products, and understanding the behavior of end-users, advertisers, and publishers.

Please note that this position is the same as Data Scientist.

Responsibilities

- Research, develop, and apply methods for measuring and analyzing Google products.
- Develop new algorithms and methods for optimizing product performance, and adoption.
- Research new ways for modeling and predicting end-user behavior.
- Lead investigations into multiple streams of product data.
- Design experiments to answer targeted questions. Conduct exploratory data analysis in high dimensions.



Minimum qualifications

- BA/BS degree in Statistics or other quantitative disciplines such as Engineering, Applied Mathematics, etc or equivalent practical experience.
- Experience with large data sets using statistical software (R, S-Plus, Matlab, or similar) and large databases (SQL).

Preferred qualifications

- MS or PhD in Statistics or other quantitative disciplines such as Engineering, Applied Mathematics, etc.
- Broad work experience with large data sets. Strong experimental design and analysis skills.
- Considerable practical experience in quantitative analysis. Demonstrated leadership and self-direction.
- Specific positions can benefit from



Get more out of your job search

- See customized job recommendations
- Connect with people at Google
- Get email updates when new jobs open
- Star jobs you like to quickly find them later

SIGN INDon't have a Google account? [Sign up](#)

VIEWERS OF THIS ALSO VIEWED

Corporate Engineering Data Analyst
Mountain View, CA, USA

Quantitative Analyst
Boulder, CO, USA

Quantitative Analyst, YouTube
San Bruno, CA, USA

VIDEOS TO WATCH





Join Our Talent Network

Receive updates and job notifications



Search by

Profession | Location

Right Now @ Microsoft

Life @ Microsoft

Interviewing @ Microsoft

: DATA SCIENTIST, YAMMER JOB

Keyword Search



Data Scientist, Yammer Job

Apply Now

Date: Dec 11, 2013

Location: Redmond, WA, US

Job Category: Software Engineering: Program Management

Location: Redmond, WA, US

Job ID: 849099-121540

Division: Applications and Services Engineering

Social media has dramatically changed the way we share and connect with friends and family, and it will have an even more profound impact on the way companies operate. Yammer and Microsoft provide a secure, private social network for your company. Yammer empowers employees to be more productive and successful by enabling them to collaborate easily, make smarter decisions faster, and self-organize into teams to take on any business challenge. It is a way of working that naturally drives business alignment and agility, reduces cycle times, engages employees and improves relationships with customers and partners.

Day-to-day work varies greatly, but here's a small sample of things you might do:

- Help internal customers understand relevant data, and how it should impact their decisions.
- Work with Product and Engineering teams to define criteria and measure success of new features.
- Help determine priorities by estimating the potential impact of projects.
- Triage problems with our product using user engagement data.

Reasons this job is awesome:

- Autonomy - You'll actually own your projects. This means working directly with your customers from beginning to end on all of your projects.
- Great team - Yammer analysts are exceptional thinkers and a lot of fun. When you love everyone you work with, every day is great.
- Impact - From project scope to priority to implementation, you have an impact on what and how things get made. You can point to features and say "I helped make that happen."
- Free Hot Sauce - We will give you a free bottle of hot sauce when you join. !

We're looking for someone with:

- Amazing logical problem-solving abilities.
- A background in analysis of large datasets. This typically means a degree in a quantitative field, or relevant work experience. Team members have backgrounds as varied as economics, mobile gaming, physics, psychology, and statistics.
- No fear of working in a space with ambiguous answers.
- Good communication abilities - the ability to turn numbers into words and words into decisions.

Bonus points for:

- Knowledge of Excel, SQL, R (or similar stats packages)
- Experience with A/B Testing



How will we be introduced to experimental design?

- Basic review of Stats 10
 - Confidence interval and hypothesis testing for means
 - Independent and paired t-test
- ANOVA
- Examining important features of experiments
- Different (efficient) experimental designs



- You each have access to The Island
- <http://island.maths.uq.edu.au/index.php>
- First HW due Monday

RStudio

- A user-friendly interface for R statistical software package
- Created by a loosely-organized community of volunteer programmers.
- Download at <http://topaz.rstudio.org/>



Let's Review Stat 10

- Parameter vs Statistic
- Distribution of
 - Population
 - Sample
- Sampling Distribution
- Confidence Intervals
- Hypothesis Testing

Parameter vs Statistic

JANUARY 7, 2014



In New Year, Half Are Looking Forward to Midterm Elections

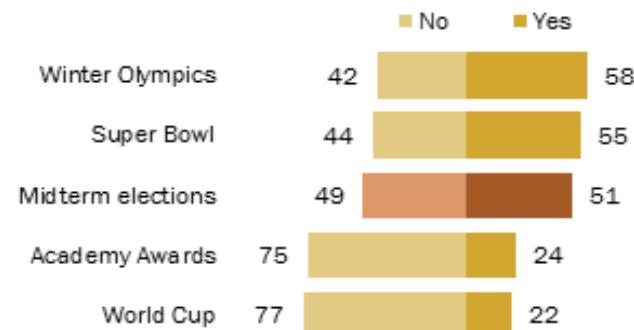
More Republicans than Democrats Are Anticipating Midterms

As 2014 begins and the midterm election campaigns heat up, about half of the public (51%) is especially looking forward to November's congressional elections while 49% are not looking forward to them.

The new national survey by the Pew Research Center, conducted Jan. 2-5 among 1,005 adults, finds that 58% are looking forward to next month's Winter Olympics in Sochi, Russia, while 42% are not. A majority (55%) also is looking forward to next month's Super Bowl.

Looking Ahead to Major Events of 2014

% saying they are "especially looking forward to" each



Survey conducted Jan. 2-5, 2014.

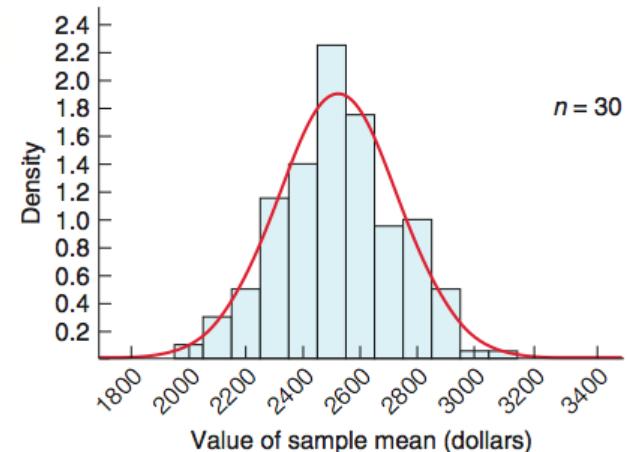
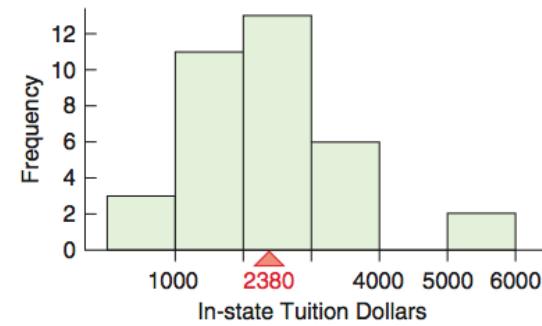
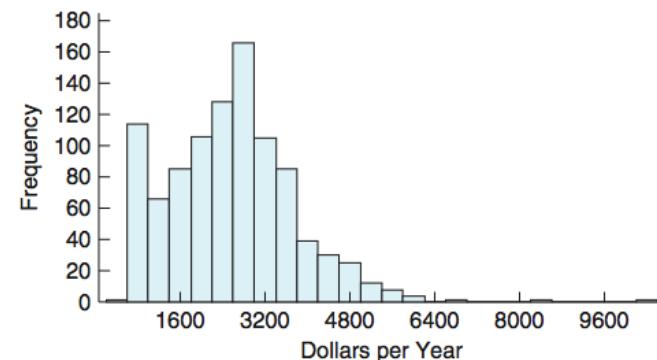
PEW RESEARCH CENTER

Distributions

Population

| Sample

Sampling



Confidence Intervals

- Make use of sampling distribution to construct confidence intervals
- For example we are 95% confident that the true population mean is in the interval

$$\bar{x} \pm 1.96 \frac{\sigma}{\sqrt{n}}$$

- What does 95% confident mean?
- Where does 1.96 come from?

Confidence Interval Example

A random sample of 35 two- year colleges in 2008 had a mean tuition of \$2380 with a standard deviation of \$1160. Construct a 95% CI. Interpret.

$$2380 \pm 2.032 \frac{1160}{\sqrt{35}}$$

$$2380 \pm 398.43$$

$$(1981.57, 2778.43)$$

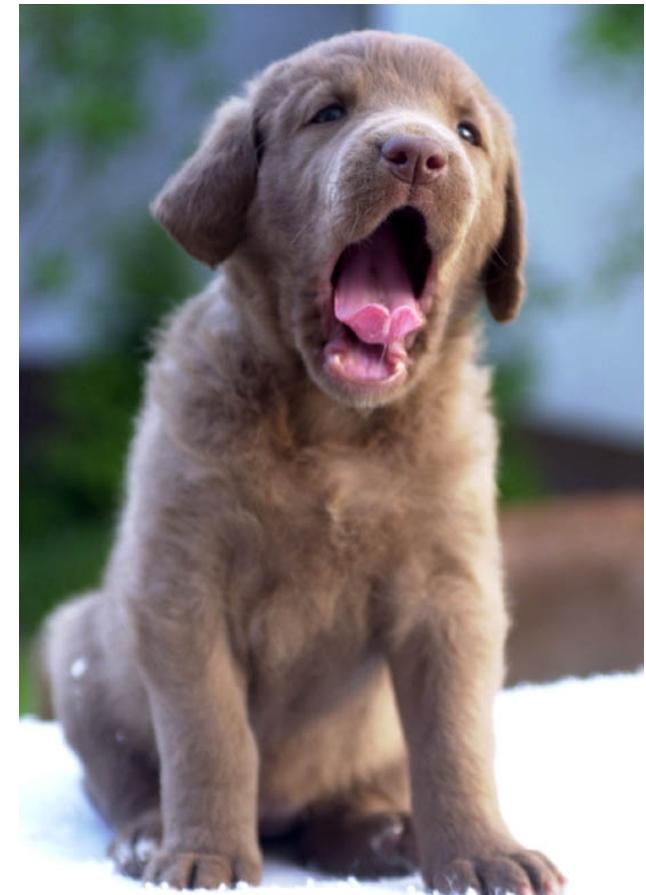
Confidence Interval Interpretation

We are 95% confident that the mean tuition could be as low as \$1982 or as high as \$2778

Mythbusters: Is Yawning Contagious?

<http://www.youtube.com/watch?v=7XYpdYjiLTU>

	Stimulus	Control
Yawn	10	4
No Yawn	24	12



Chance Model

- Null hypothesis is that the response varies only as a result of random chance
- We can simulate this: take 50 cards. On 14 write 'yawn' and on the rest 'No yawn'
- Get 34 people to represent the treatment group and 16 to be the control group
- Shuffle the cards and pass them out. What are the results?

Original Data to Simulation

	Stimulus	Control
Yawn	10	4
No Yawn	24	12

Difference in proportions
between groups is
originally 29% - 25% = 4%

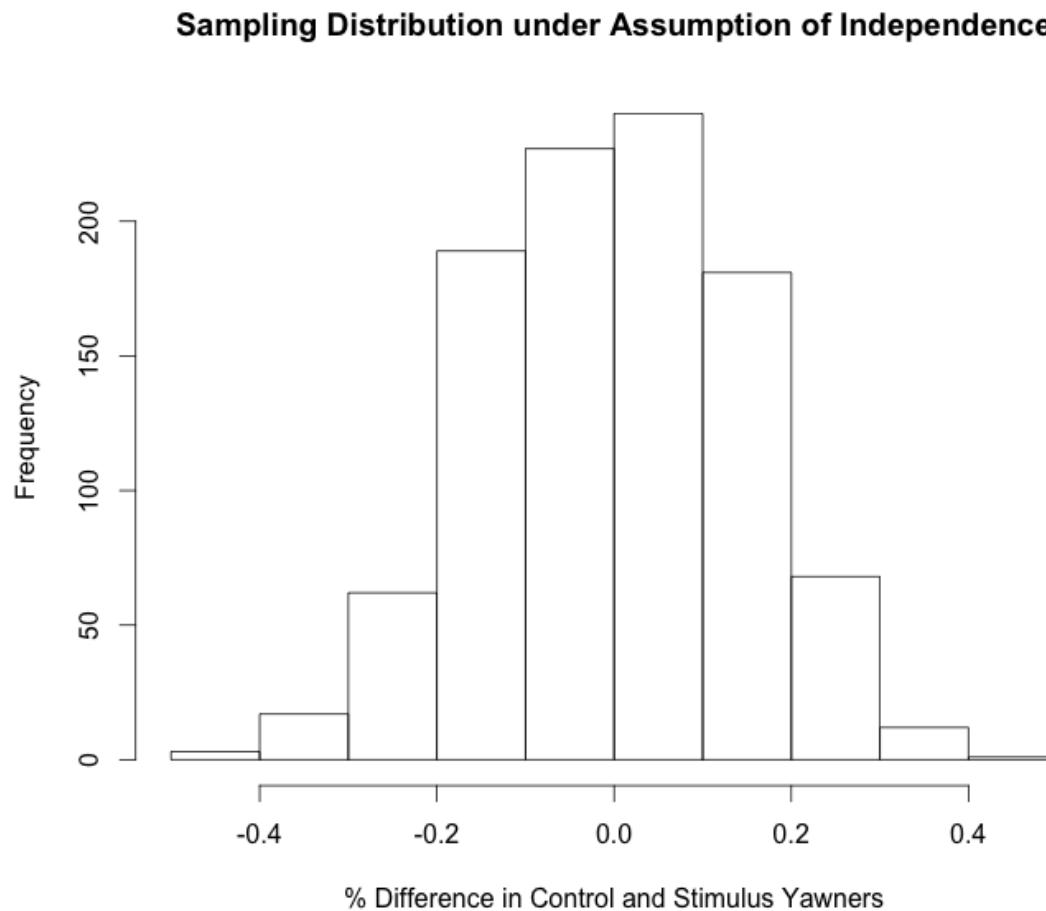
	Stimulus	Control
Yawn	9	5
No Yawn	25	11

Difference in proportions
between groups in a single
simulation 31% - 26% = 5%

Fairly close, although a slightly higher result by chance. What happens if we simulate this 1000 times?

Chance Model Distribution

Is this myth really “confirmed”?



Basic Structure of HT

- State Hypotheses.
 - Null: Starting point, status quo
 - Alternative: Alternative claim
- Compute test statistic
- Determine if your observed value is 'rare'
 - If the theoretical sampling distribution of the test statistic is known, use this to determine if the test statistic is rare
- Make conclusion to reject or fail to reject null hypothesis

Parametric vs Permutation

- Parametric
 - Simple to calculate
 - Must meet more conditions
- Permutation
 - No assumption of distribution
 - Can compute test statistic for different measures
 - Computationally more difficult and time consuming

T-Test

- Does mother's smoking affect birthweight of her baby?
- Does dieting work?
- On the Island:
Does taking adrenalin improve basic arithmetic skills?

Independent T-test

The test statistic for comparing means is

$$\bar{x}_1 - \bar{x}_2 \pm t_{df}^* \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

with degrees of freedom

$$df = \min(n_1 - 1, n_2 - 1)$$

Independent T-test Example

On the Island, I have randomly selected 10 inhabitants to take shot of adrenalin and then take a basic arithmetic test and another 10 inhabitants to simply take the basic arithmetic test. Is there a significant difference in basic arithmetic for those who had a shot of adrenalin and those who did not?

	Mean	Standard deviation
Adrenalin	38.750	1.488
Control	37.125	2.532

R Code: Independent T-test

`t.test(control, exp)`

```
data: control and exp
t = -1.565, df = 11.32, p-value = 0.1451
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
-3.9024816  0.6524816
sample estimates:
mean of x mean of y
 37.125    38.750
```

Paired T-test

- Paired samples
 - before and after comparisons
 - objects are related somehow (twins, siblings, spouses)
 - Experimenters have deliberately matched subjects to have similar characteristics
- Knowing the value of a subject in one group tells you something about the value in the other

R Code: Paired T-test

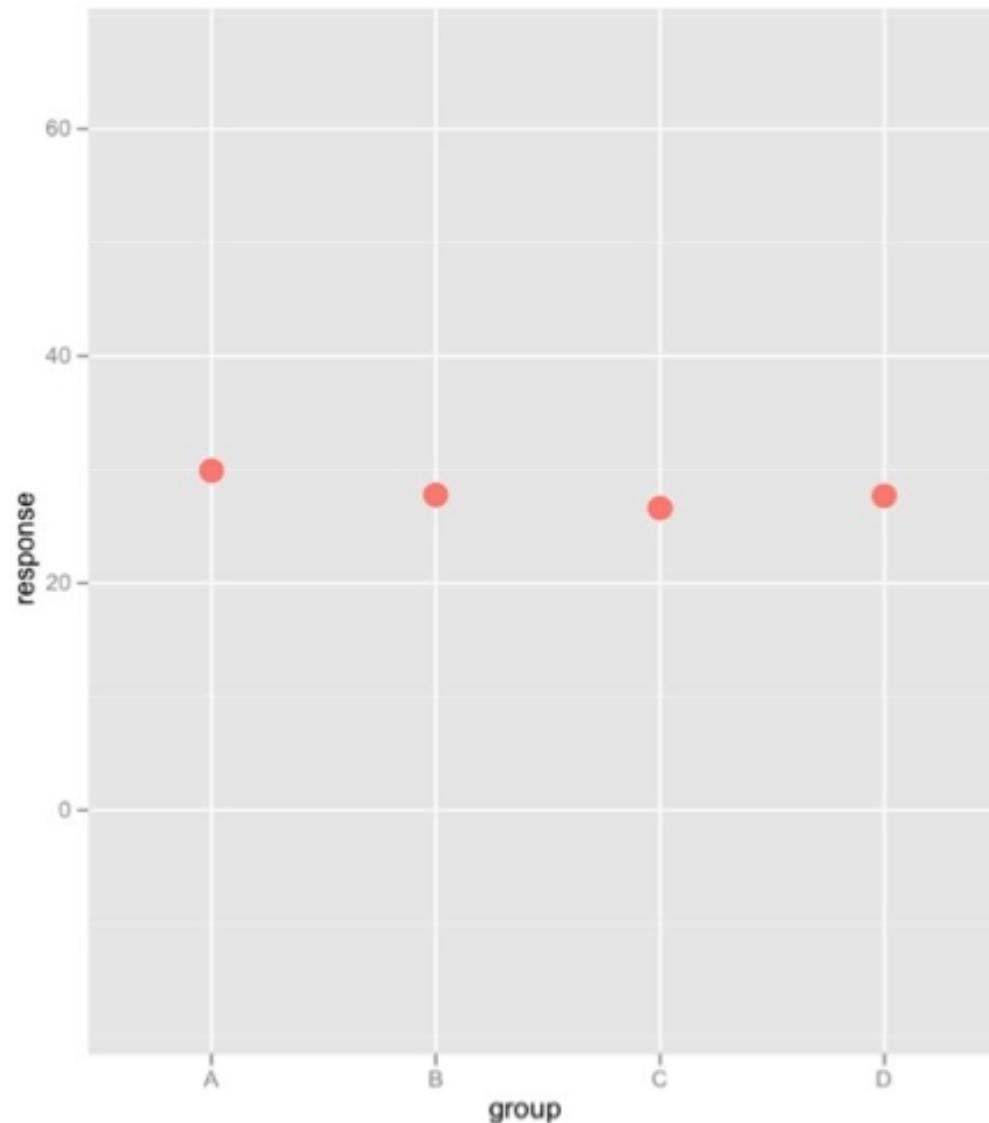
```
t.test(before, after, paired=TRUE)
```

```
data: before and after
t = -0.1982, df = 9, p-value = 0.8473
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
-1.241088 1.041088
sample estimates:
mean of the differences
-0.1
```

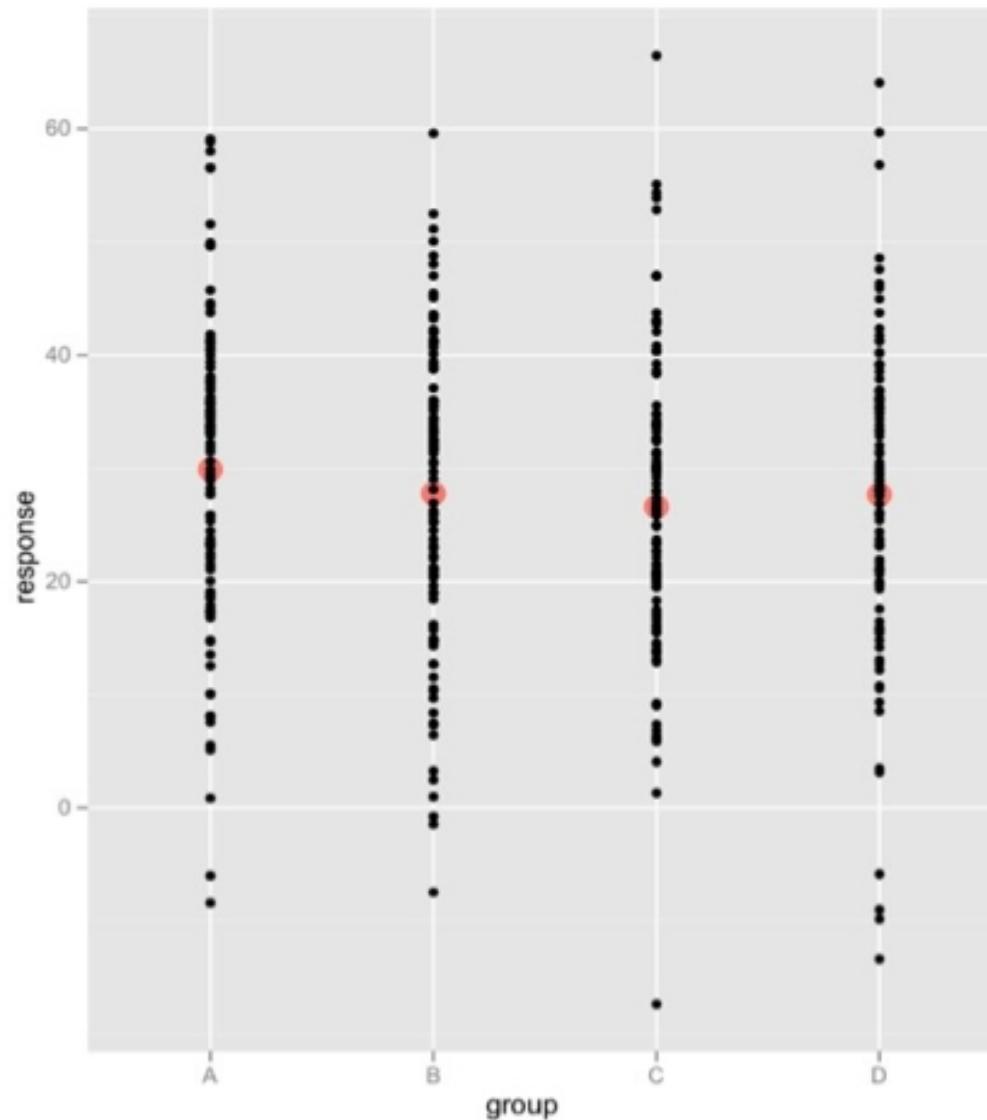
ANalysis Of VAriance (ANOVA)

- Invented by R.A. Fisher in 1920s
- Approach for comparing means of multiple groups
- Observed variance is partitioned into components attributable to different sources of variation
- Heavily used in analysis of experimental data

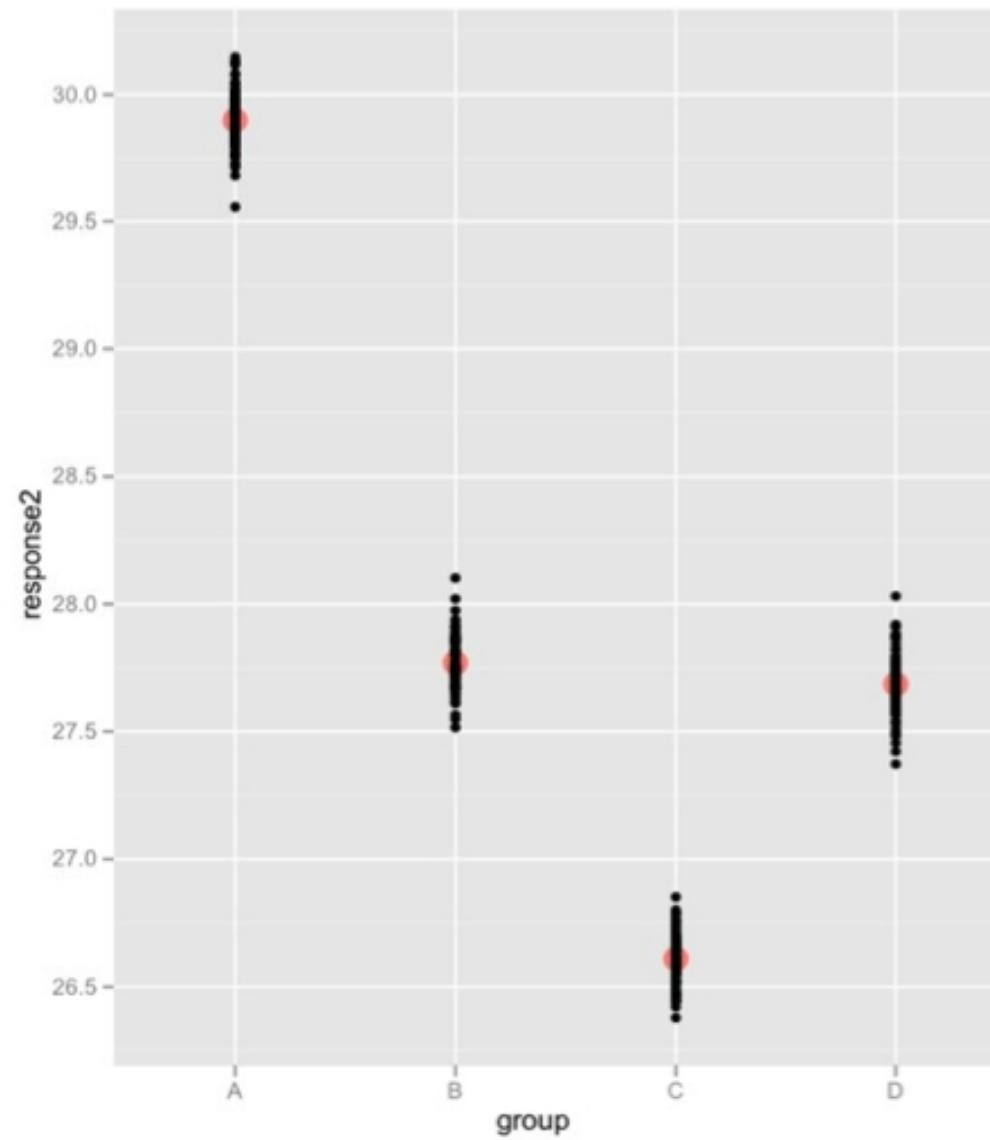
4 Group Means



How do these groups compare?



How about these groups?



So...

- Although we are comparing means and it is the focus in our hypotheses, it is really the variation that is behind our conclusion

One-Way ANOVA

- Generalization of two-sample t-test
- Compare more than two means of different groups
- Compares a numerical response across a categorical variable with several categories
 - Ex: Randomly assign subjects to one of 3 levels of caffeine consumption. Measure blood pressure over a week. Is the mean blood pressure different in the 3 groups?
- Multiple two-sample t-tests results in increased chance of type I error

One-way 'Means Model'

$$y_{ij} = \mu_i + \epsilon_{ij} \quad \epsilon_{ij} \sim N(0, \sigma)$$

$$i = 1 \dots a, j = 1 \dots n$$

i – counts the number of treatment groups

j – counts individual replications

called the 'means model' b/c it focuses on means.

μ_i for example, represents mean blood pressure for Red Bull drinkers

Hypotheses One-way 'Means Model'

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

$$H_a : \mu_i \neq \mu_j$$

for at least one pair of (i,j)

R Code: One-way 'Means Model'

- Simply implemented through `lm`

```
m1.lm=lm(Times~Diet,data=leafhopper)
summary(m1.lm)
```

One-way 'Effects Model'

Define $\mu_i = \mu + \tau_i$

so now

$$y_{ij} = \mu + \tau_i + \epsilon_{ij}$$

- μ represents the grand mean (benchmark): the overall blood pressure of all people in the caffeine study
- τ_i represents the effect of being not on just any of the caffeine levels but on the Red Bull level

Hypotheses One-way 'Effects Model'

$$H_0 : \tau_1 = 0 \quad \text{for all } i$$

$$H_a : \tau_i \neq 0 \quad \text{for at least one } i$$

R Code: One-way 'Effects Model'

Implement through the **aov** command

```
#The aov command uses 'effect' coding  
m1.aov=aov(Times~Diet,data=leafhopper)  
model.tables(m1.aov)
```

ANOVA conclusions

- Only tells us if any of the group means are different from the rest
- Does not tell us specifically which means are different
- Be careful when reporting effect size. Differs depending on which model you have specified

Paper Airplane Experiment

- Next week we will all collect some data concerning paper airplanes
- Handout for the airplane model online
- Create groups of 4
- Collect materials over the weekend
 - Paper: 8.5 x 11 regular, construction
 - Paper clips (at least 4)
 - Stopwatch of somekind (phone will work)

Chapter 4 Slides

The Content of an Experiment

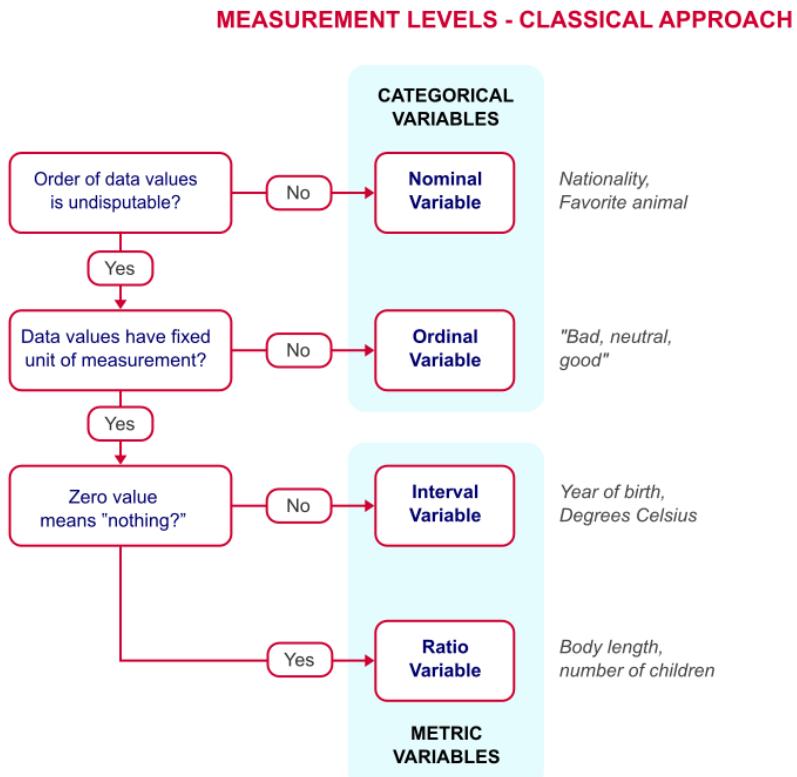
Response Variable

- The measurement you use to judge the effect of the conditions.
- We will be considering one response at a time in this course

Types of Response Variables (levels of measurements)

- Nominal – name of a category. Groups have no meaningful order
- Ordinal – categories that are ordered but have no meaningful order
- Interval – a number and the distance between numbers have meaning, but there is no natural zero
- Ratio – numbers that both have meaningful distance and natural lowest value

Type of Response Variable



Analysis of Variance is:

- Impossible for Nominal Data
- Sometimes appropriate for Ordinal data, but at risk of losing information.
- Generally appropriate for interval and ratio data

Types of Response Examples:

The first level of measurement is **nominal level of measurement**. In this level of measurement, the numbers in the variable are used only to classify the data. In this level of measurement, words, letters, and alpha-numeric symbols can be used. Suppose there are data about people belonging to three different gender categories. In this case, the person belonging to the female gender could be classified as F, the person belonging to the male gender could be classified as M, and transgendered classified as T. This type of assigning classification is nominal level of measurement.

The second level of measurement is the **ordinal level of measurement**. This level of measurement depicts some ordered relationship among the variable's observations. Suppose a student scores the highest grade of 100 in the class. In this case, he would be assigned the first rank. Then, another classmate scores the second highest grade of an 92; she would be assigned the second rank. A third student scores a 81 and he would be assigned the third rank, and so on. The ordinal level of measurement indicates an ordering of the measurements.

The third level of measurement is the **interval level of measurement**. The interval level of measurement not only classifies and orders the measurements, but it also specifies that the distances between each interval on the scale are equivalent along the scale from low interval to high interval. For example, an interval level of measurement could be the measurement of anxiety in a student between the score of 10 and 11, this interval is the same as that of a student who scores between 40 and 41. A popular example of this level of measurement is temperature in centigrade, where, for example, the distance between 94°C and 96°C is the same as the distance between 100°C and 102°C.

The fourth level of measurement is the **ratio level of measurement**. In this level of measurement, the observations, in addition to having equal intervals, can have a value of zero as well. The zero in the scale makes this type of measurement unlike the other types of measurement, although the properties are similar to that of the interval level of measurement. In the ratio level of measurement, the divisions between the points on the scale have an equivalent distance between them.

The researcher should note that among these levels of measurement, the nominal level is simply used to classify data, whereas the levels of measurement described by the interval level and the ratio level are much more exact.

Principles in Designing Experiments

- 1) Replication 2) Randomization 3) Blocking

Reliability and Validity: How good is your Response?

Reliability and validity both refer to the soundness of the connections between the goal of the experiment and your choice of response, that is, between the kinds of conclusions you hope to draw and the kind of evidence you plan to gather.

Reliability: is concerned with Repeatability

Validity: is Concerned with Relevance

Replication

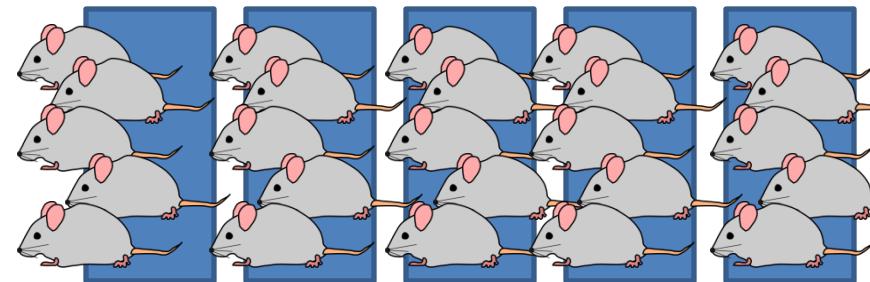
- One observation doesn't work
- Replication refers to use of several independent observations
- Independent being there is no connection between
- measurement of one observation and another
 - simply put knowledge of one observation tells us nothing about the another
- Helps to attain a more *reliable estimate of the effect* of each treatment

Replication Examples

- Does coffee improve memory? Ask an Islander to take a memory test. Then do the same thing after drinking a cup of coffee. Are these independent?
- If we have a system where we are in control of a pressure knob. We plan on running two sequential trials setting the pressure to 30 psi. Between each trial we must set the pressure to an intermediate setting and then reset the pressure to 30. Why?

Replication Examples

- Say we have a lab with ten cages of five laboratory mice each. We apply two treatments to each of the ten cages.
- How many replications do we have?



Caffeine Memory Boost Evidenced In Study At Johns Hopkins University

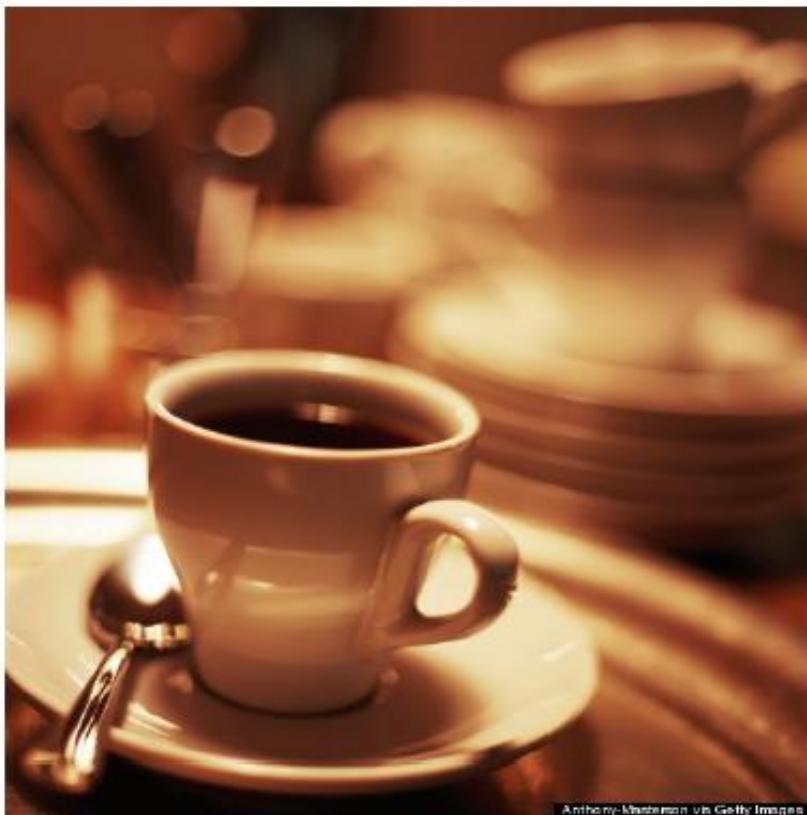
Huffington Post UKPA | Posted: 12/01/2014 18:42 GMT | Updated: 12/01/2014 18:42 GMT



Students take heed...

A double espresso shot after swotting for an exam might help to jog those elusive memories, new research suggests.

Scientists have found the first clear evidence of caffeine's memory-boosting effect, and shown that it lasts at least 24 hours.



Anthony Masterson via Getty Images

Just one double espresso could be enough to notice an effect

Volunteers took part in a double-blind trial in which they were either given a 200 milligram caffeine pill or dummy placebo tablet five minutes after studying a series of images.

Tests a day later proved that the memory of those who took caffeine had been enhanced at a deep level.

The amount of caffeine used was roughly equivalent to a double shot of strong espresso coffee.

US lead researcher Dr Michael Yassa, assistant professor of psychological and brain sciences at Johns Hopkins University in Baltimore, said: "We've always known that caffeine has cognitive-enhancing effects, but its particular effects on strengthening memories and making them resistant to forgetting has never been examined in detail in humans."

"We report for the first time a specific effect of caffeine on reducing forgetting over 24 hours."

More than 100 participants took part in the study, none of whom were regular users of caffeinated products.

Before being given the caffeine pill or placebo, they were asked to identify a series of pictured objects as either outdoor or indoor items.

The next day, both groups were tested on their ability to recognise the images they had been shown earlier. Some of the images were the same as the ones they had seen, some were new, and some similar but not identical.

Although all the volunteers correctly identified "new" and "old" pictures, those who had taken the caffeine pill were better able to spot "similar" images.

Participants not dosed with caffeine were more likely to be fooled into thinking the similar pictures were the ones viewed the previous day.

Recognising the difference between two similar but not identical items reflects a deep level of memory retention, said the team writing in the journal *Nature Neuroscience*.

Volunteers took part in a **double-blind trial** in which they were either given a 200 milligram caffeine pill or dummy placebo tablet five minutes after studying a series of images.

Tests a day later proved that the memory of those who took caffeine had been enhanced at a deep level.

The amount of caffeine used was roughly equivalent to a double shot of strong espresso coffee.

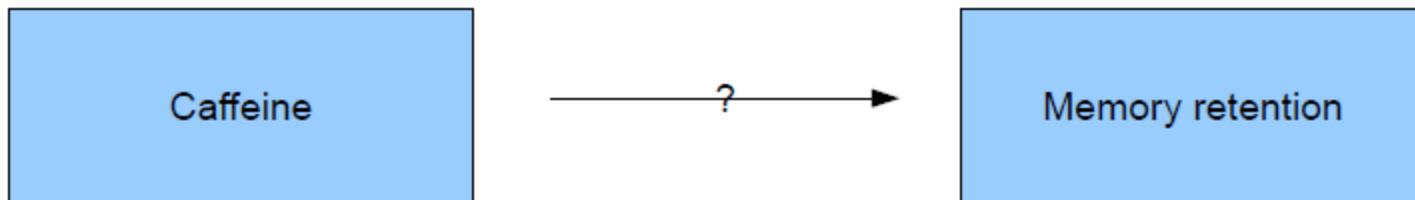
US lead researcher Dr Michael Yassa, assistant professor of psychological and brain sciences at Johns Hopkins University in Baltimore, said: "We've always known that caffeine has cognitive-enhancing effects, but its particular effects on strengthening memories and making them resistant to forgetting has never been examined in detail in humans."

Caffeine Study

- Suppose we change the study. Volunteers were told to drink as much coffee, Red Bull, Mountain Dew, 5-hr Energy as they wanted, or not. They were shown a number of pictures and then they were tested on the pictures from the previous day. We find that caffeine drinkers had higher memory retention than non-caffeine drinkers.
- Can we conclude it was because of the caffeine?

Cause-and-Effect

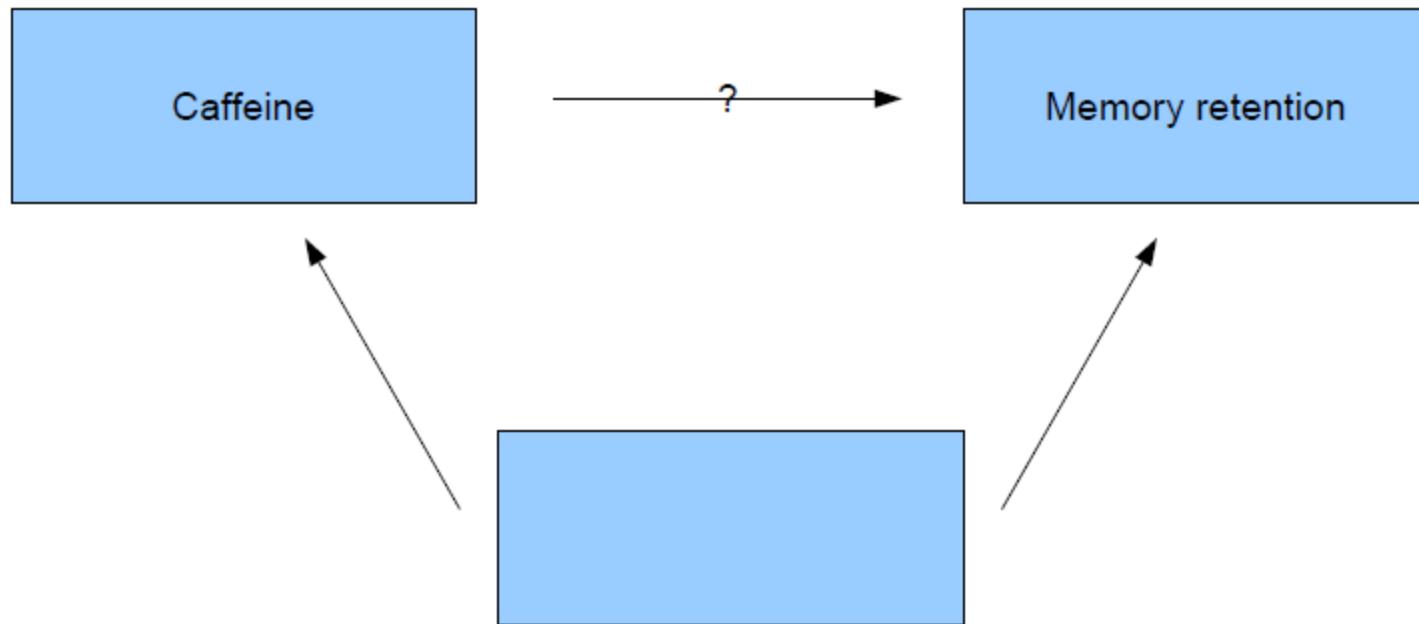
- Does caffeine cause a change in memory retention?



Confounding Factors

- A confounding factor confounds our ability to make cause-and-effect conclusions.
- Serves as an alternative explanation for the association we observed (caffeine consumption associated with higher memory retention)
- To do this, it has to explain both the variability in the treatment and the association with the response

Name a Potential Confounder



Randomized Assignment

- Subjects or objects are randomly assigned to treatment groups or the order of the treatment is randomized.
- Selection Bias occurs in observational studies when the process of selecting groups to be compared confounds the effects of interest with other effects.
- Two influences on the response are confounded if the design makes it impossible to isolate the effects of one from the effects of the other.
- Why does this solve the problem of confounding factors?

Randomization

- In our example, if the sample sizes are fairly large, if they are shuffled randomly, the distribution of nuisance factors should be about the same for each level of the treatment factor
- It ensures that any bias of the experimenter is avoided and that any unknown differences between the units are unlikely to consistently favor one particular treatment

Blinding

- When we know what treatment was assigned, it's difficult not to let that knowledge influence our assessment of the outcome
- Two main classes of individuals who can affect the outcome of the experiment
- Those who influence the results (subjects)
- Those who evaluate results (statisticians)
- If the individuals in either group do not know which group they are in the study is blind

Placebos

- Often simply applying any treatment can induce an effect
- A 'fake' treatment that looks just like the treatment being tested is called a placebo
- The placebo effect occurs when taking a sham treatment results in a change in the response variable
- <http://cheezburger.com/16159489>

Blocking

- Suppose that the treatment is truly associated with the response. i.e. caffeine really does cause higher memory retention
- Suppose we suspect people of similar ages might see similar effects from caffeine
- We can block these people by putting those with similar age groups consumptions into the same group. Ex: Young Adults, Middle-Aged, Senior Citizens
- Within each group, we randomly assign some to use caffeine and some to no caffeine

Blocking

- Randomize within blocks
- This helps us control for the *nuisance factor* of age
- Age might be a confounder. Why not just leave it to the randomization to take care of?

Blocking and Nuisance Factors

- Blocking a nuisance factor can lead to an increase in power – our ability to detect real effects
- We have 'eliminated' one source of variation in age, and so get more precise measurements

Blocking

- A block consists of (relatively) homogeneous (i.e. similar) observations/objects.
- Randomization is performed within each block.

Types of Design Factors

- **Design** factors are those that affect the response and are of interest in the study
- **Held-constant** factors are those that may affect the response, but we're not interested in studying and we have no control over their value. So these can be set on one value and held constant
- **Allowed-to-vary** factors are those which might cause some variation in response, but we're comfortable with allowing this to be 'solved' by the randomization

Nuisance Factors

- Controllable – use blocking
- Uncontrollable – measure the values and perform 'statistical controls'
- Noise – the variation caused will be 'modeled' in the noise term, potentially diminishing our precision.

Cause and effect diagrams

- The Cause & Effect (CE) diagram, also sometimes called the ‘fishbone’ diagram, is a tool for discovering all the possible causes for a particular effect. The effect being examined is normally some troublesome aspect of product or service quality, such as ‘a machined part not to specification’, ‘delivery times varying too widely’, ‘excessive number of bugs in software under development’, and so on, but the effect may also relate to internal processes such as ‘high rate of team failures’.
- The major purpose of the CE Diagram is to act as a first step in problem solving by generating a comprehensive list of possible causes. It can lead to immediate identification of major causes and point to the potential remedial actions or, failing this, it may indicate the best potential areas for further exploration and analysis. At a minimum, preparing a CE Diagram will lead to greater understanding of the problem.
- The CE Diagram was invented by Professor Kaoru Ishikawa of Tokyo University, a highly regarded Japanese expert in quality management. He first used it in 1943 to help explain to a group of engineers at Kawasaki Steel Works how a complex set of factors could be related to help understand a problem. CE Diagrams have since become a standard tool of analysis in Japan and in the West in conjunction with other analytical and problem-solving tools and techniques.
- CE Diagrams are also often called Ishikawa Diagrams, after their inventor, or Fishbone Diagrams because the diagram itself can look like the skeleton of a fish.

Cause and Effect Diagram

- **Use it when you start investigating a problem**
- Construct a CE Diagram whenever you need to investigate the causes or contributing factors for an effect (be it a quality characteristic or other outcome) which is of concern to you. This will most likely be after you have conducted a general investigation of problems for a particular function, product, or service, and ranked them using a Pareto Chart. The effect ranked highest provides the starting point for a CE Diagram.
- For example, you may just have completed an investigation of all the reasons recorded for goods being returned by customers and found that the highest incidence relates to incorrect goods being sent. A CE Diagram can be constructed to explore the possible causes for this.
- Developing a CE Diagram in a team meeting is a very effective technique for,
- concentrating team members' attention on a specific problem
- pooling, and reflecting back, team thinking
- constructing a picture of the problem at hand without resorting to the tight discipline of a flowchart

How to draw CE diagram

This is a three step process.

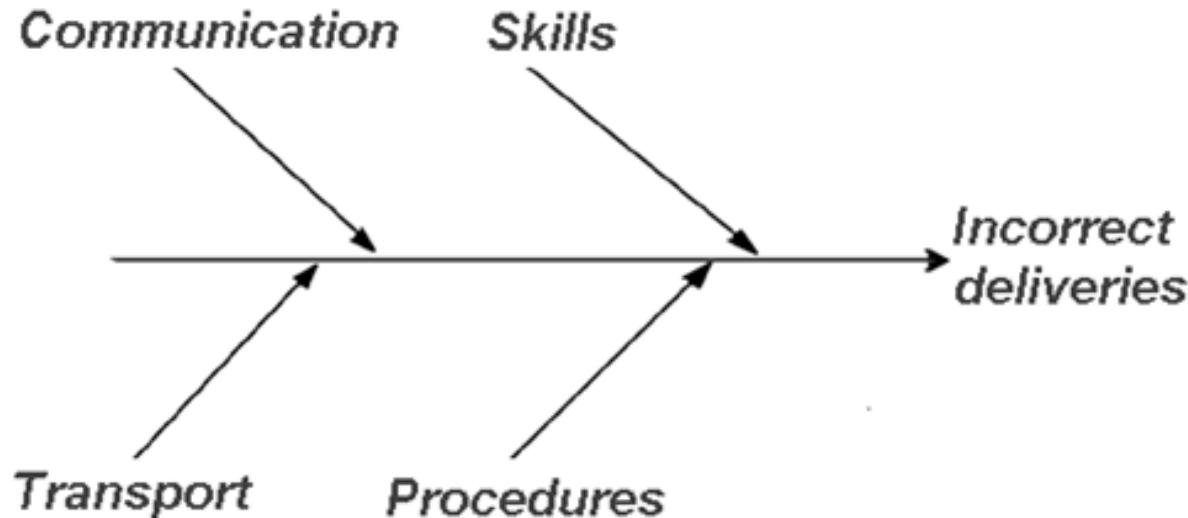
Step 1

Write down the effect to be investigated and draw the ‘backbone’ arrow to it. In the example shown below the effect is ‘Incorrect deliveries’.



Step 2

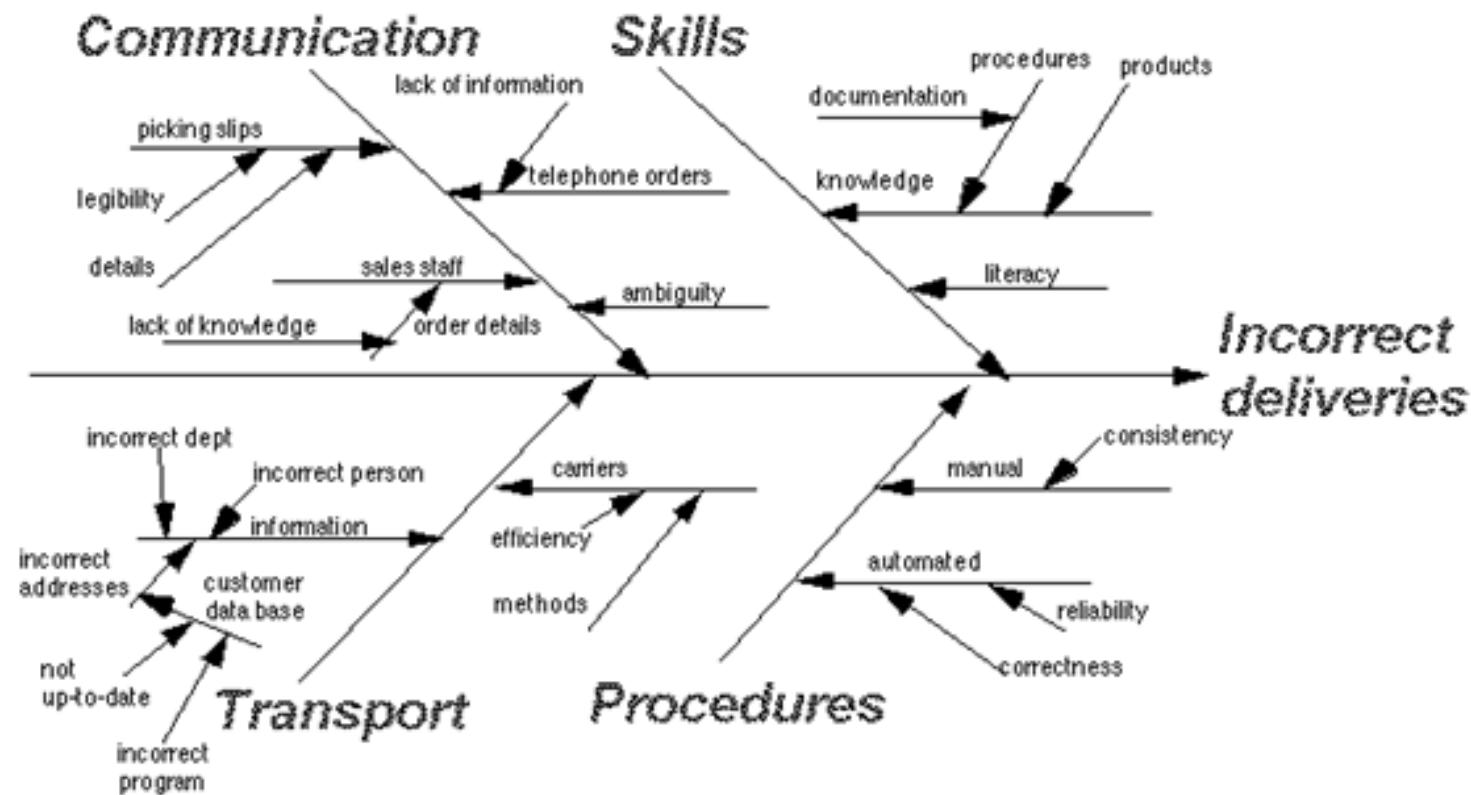
Identify all the broad areas of enquiry in which the causes of the effect being investigated may lie. For incorrect deliveries the diagram may then become:



For manufacturing processes, the broad areas of enquiry which are most often used are Materials (raw materials), Equipment (machines and tools), Workers (methods of work), and Inspection (measuring method).

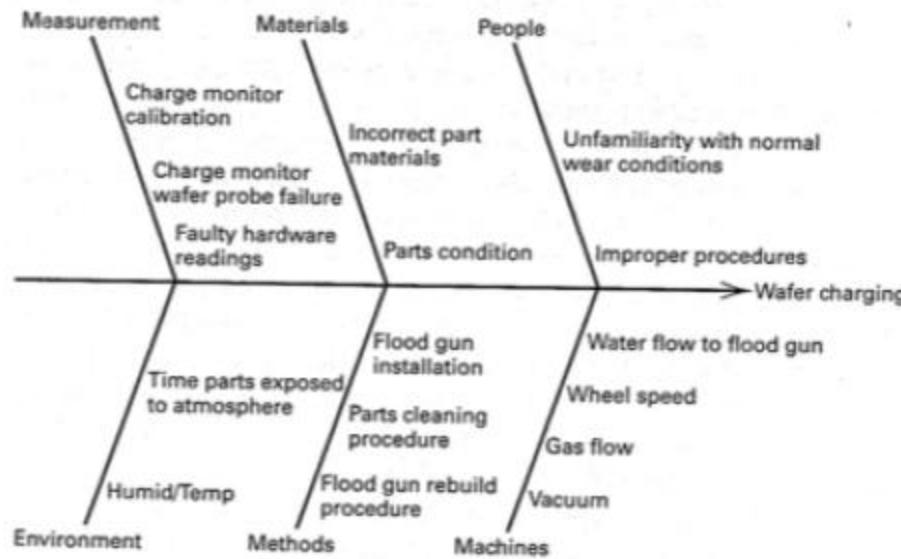
Step 3

This step requires the greatest amount of work and imagination because it requires you (or you and your team) to write in all the detailed possible causes in each of the broad areas of enquiry. Each cause identified should be fully explored for further more specific causes which, in turn, contribute to them.



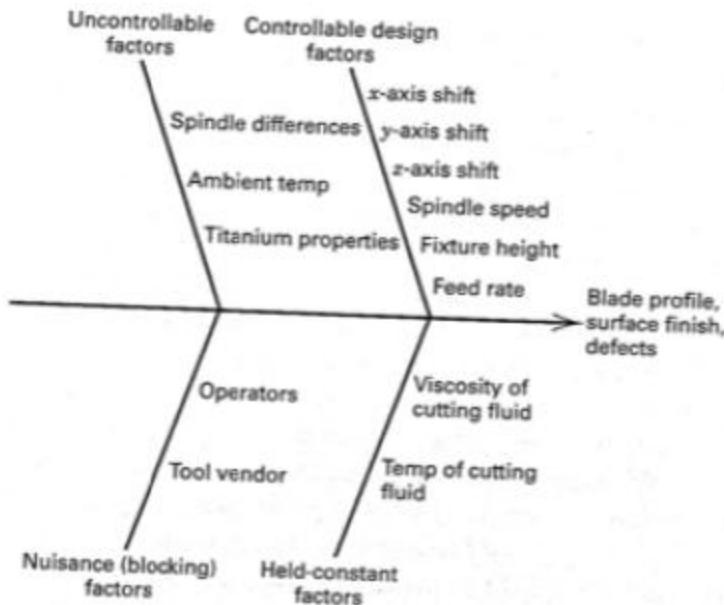
You continue this process of branching off into more and more directions until every possible cause has been identified. The final result will represent a sort of a 'mind dump' of all the factors relating to the effect being explored and the relationships between them.

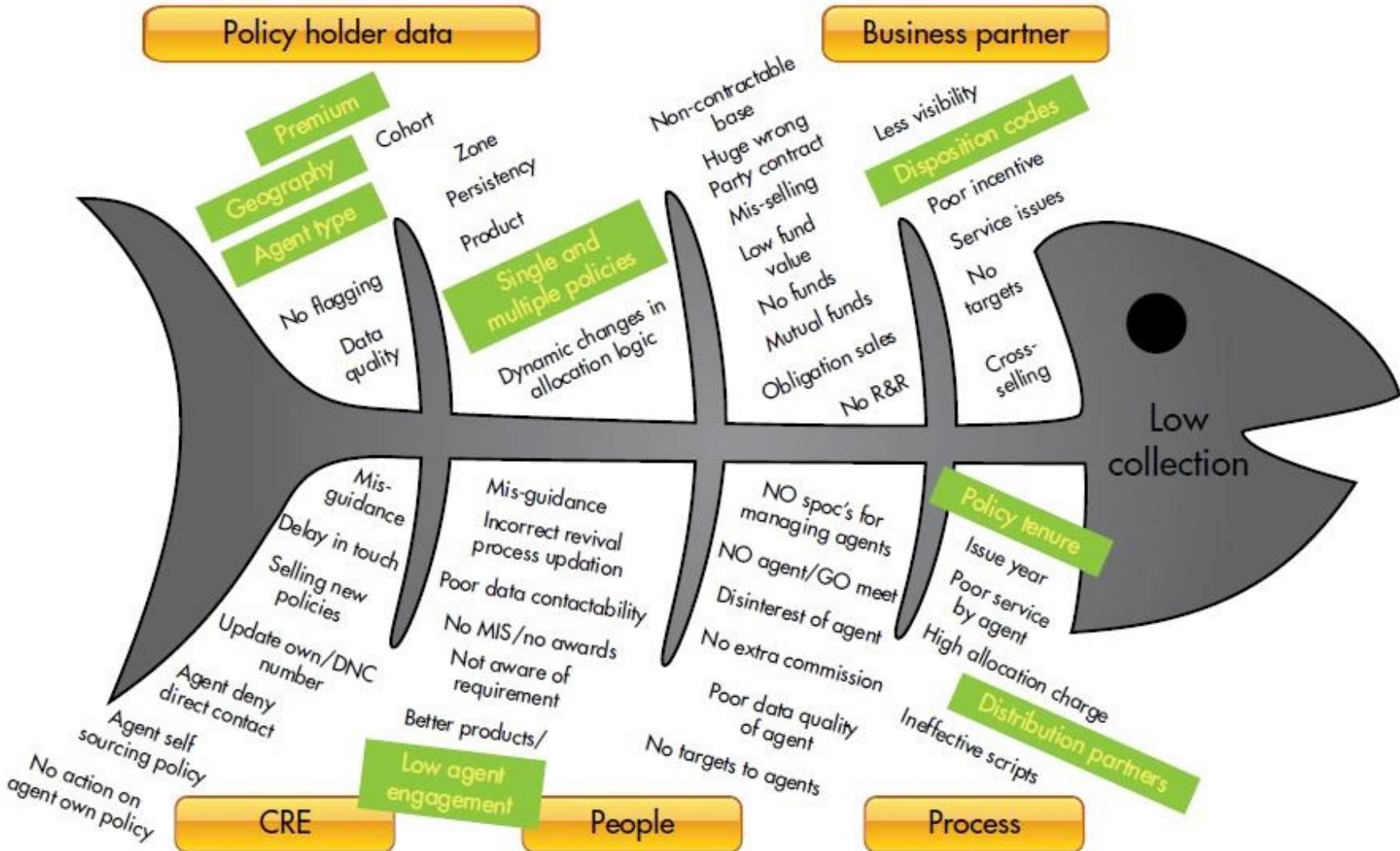
Cause-and-Effect Diagram

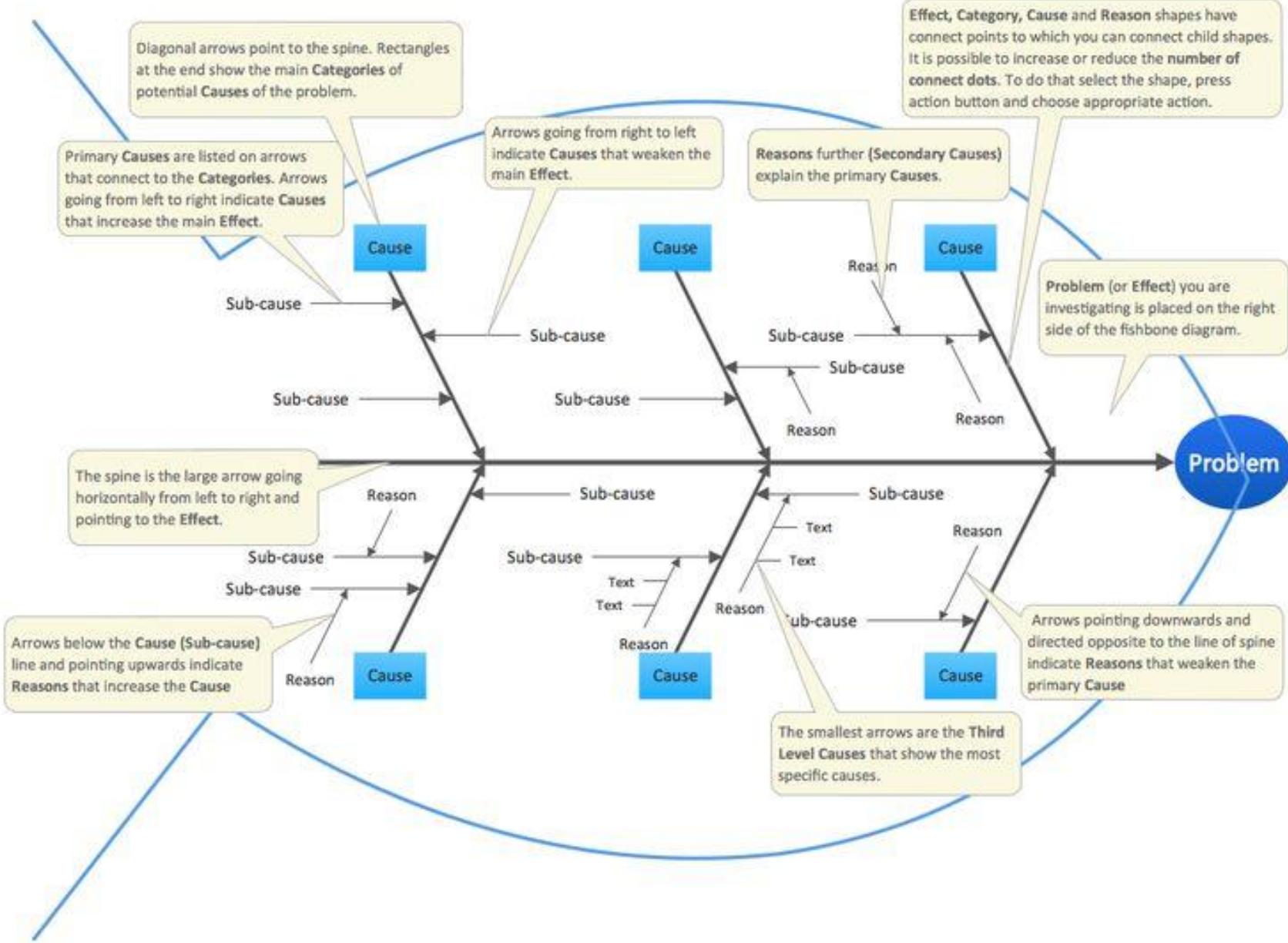


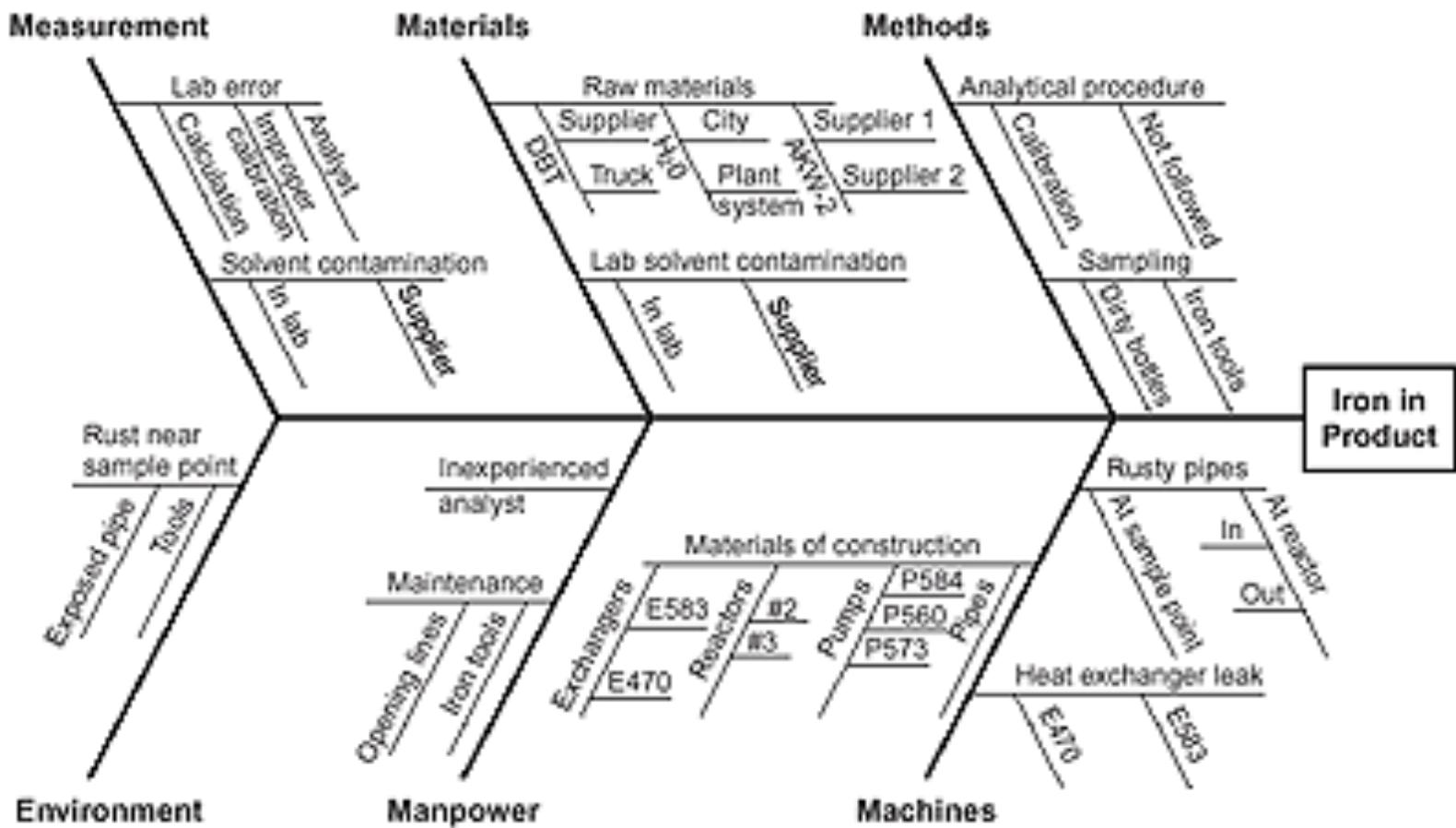
■ FIGURE 1.10 A cause-and-effect diagram for the etching process experiment

Cause-and-Effect Diagram









Paper Airplane Lab Experiment



Questions:



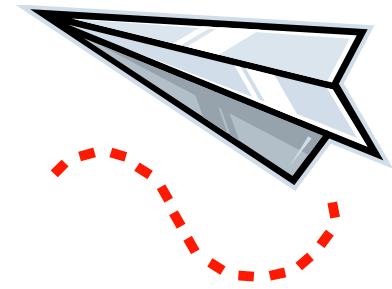
- Have you flown a paper airplane before? (Hopefully not in this class)
- Do you always use the same type of paper?
- Do you always use the same design?
- Do you want it to fly straight or do tricks?

Introduction:



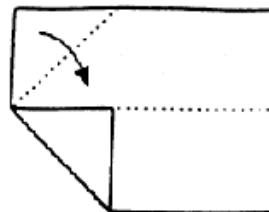
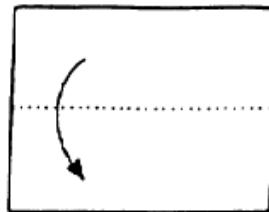
- We are going to design an experiment to test paper airplane flight distance.
- We want the planes to fly as far as they can.
- We need to think about how we are going to design and perform the experiment.
- What things do we need to think about? (Think about the steps of the Scientific Method)

Problem:



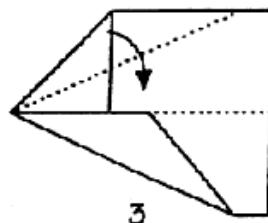
- What question are we trying to answer?
 - We want to design an experiment to test how the Shape of the plane will affect the flight distance of the paper airplane.
 - How does adding paper clips to a paper airplane affect its flight? (if Possible)

How to make a paper plane

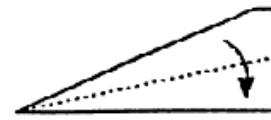


1

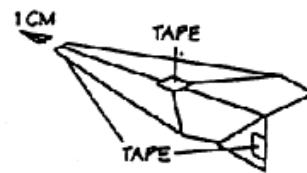
2



3

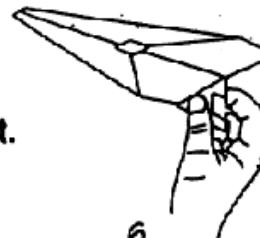


4



5

The Dart.



6

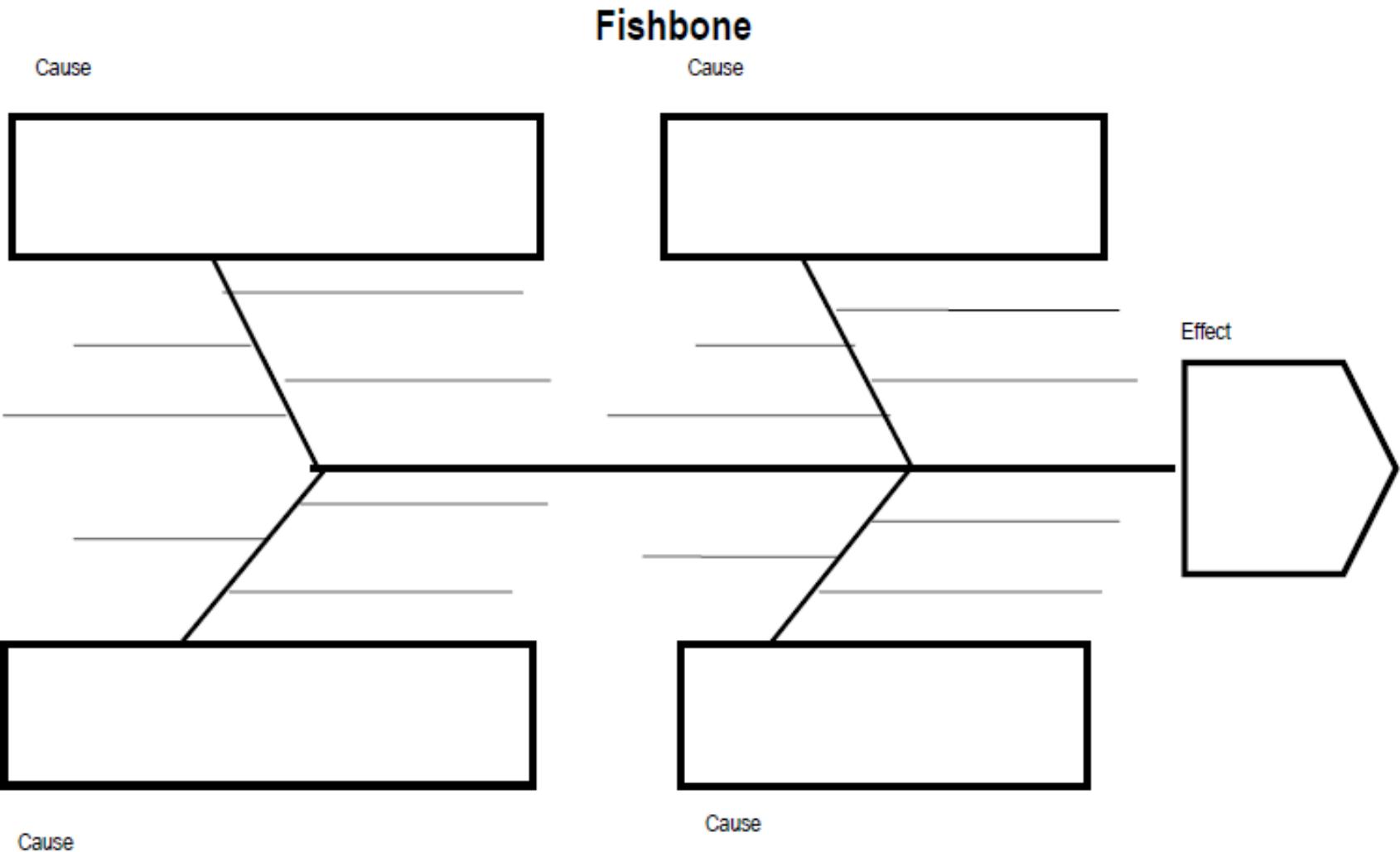
Paper Airplanes

- What can we measure that will tell us something about airplane performance?
- What factors contribute to good or bad performance?
- Work in groups of 4. Decide what you will measure and how you will test it.

Create a Fish-Bone Diagram (Airplane Experiment)



Fishbone Template



how this calibration will be maintained during the experiment. The gauge or measurement system capability (or measurement error) is also an important factor. If gauge capability is inadequate, only relatively large factor effects will be detected by the experiment or perhaps additional replication will be required. In some situations where gauge capability is poor, the experimenter may decide to measure each experimental unit several times and use the average of the repeated measurements as the observed response. It is usually critically important to identify issues related to defining the responses of interest and how they are to be measured *before* conducting the experiment. Sometimes designed experiments are employed to study and improve the performance of measurement systems. For an example, see Chapter 13.

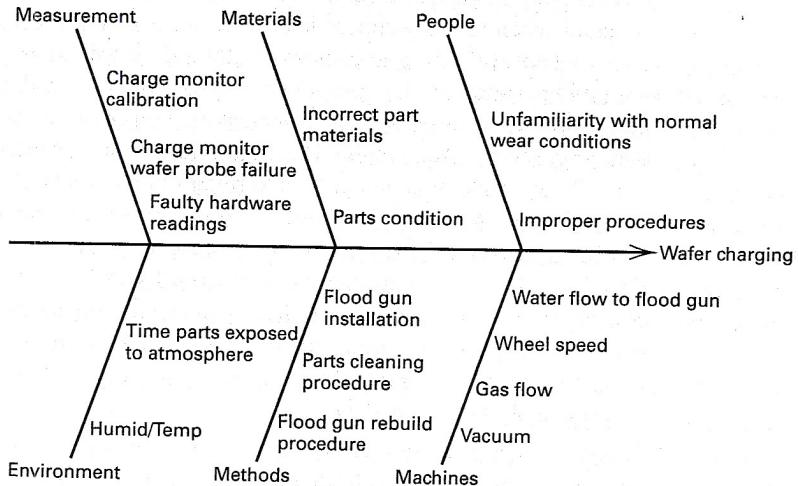
3. Choice of factors, levels, and range. (As noted in Table 1.1, steps 2 and 3 are often done simultaneously or in the reverse order.) When considering the factors that may influence the performance of a process or system, the experimenter usually discovers that these factors can be classified as either **potential design factors** or nuisance factors. The potential design factors are those factors that the experimenter may wish to vary in the experiment. Often we find that there are a lot of potential design factors, and some further classification of them is helpful. Some useful classifications are **design factors**, **held-constant factors**, and **allowed-to-vary** factors. The design factors are the factors actually selected for study in the experiment. Held-constant factors are variables that may exert some effect on the response, but for purposes of the present experiment these factors are not of interest, so they will be held at a specific level. For example, in an etching experiment in the semiconductor industry, there may be an effect that is unique to the specific plasma etch tool used in the experiment. However, this factor would be very difficult to vary in an experiment, so the experimenter may decide to perform all experimental runs on one particular (ideally “typical”) etcher. Thus, this factor has been held constant. As an example of allowed-to-vary factors, the experimental units or the “materials” to which the design factors are applied are usually nonhomogeneous, yet we often ignore this unit-to-unit variability and rely on randomization to balance out any material or experimental unit effect. We often assume that the effects of held-constant factors and allowed-to-vary factors are relatively small.

Nuisance factors, on the other hand, may have large effects that must be accounted for, yet we may not be interested in them in the context of the present experiment. Nuisance factors are often classified as **controllable**, **uncontrollable**, or **noise factors**. A controllable nuisance factor is one whose levels may be set by the experimenter. For example, the experimenter can select different batches of raw material or different days of the week when conducting the experiment. The blocking principle, discussed in the previous section, is often useful in dealing with controllable nuisance factors. If a nuisance factor is uncontrollable in the experiment, but it can be measured, an analysis procedure called the **analysis of covariance** can often be used to compensate for its effect. For example, the relative humidity in the process environment may affect process performance, and if the humidity cannot be controlled, it probably can be measured and treated as a covariate. When a factor that varies naturally and uncontrollably in the process can be controlled for purposes of an experiment, we often call it a noise factor. In such situations, our objective is usually to find the settings of the controllable design factors that minimize the variability transmitted from the noise factors. This is sometimes called a process robustness study or a robust design problem. Blocking, analysis of covariance, and process robustness studies are discussed later in the text.

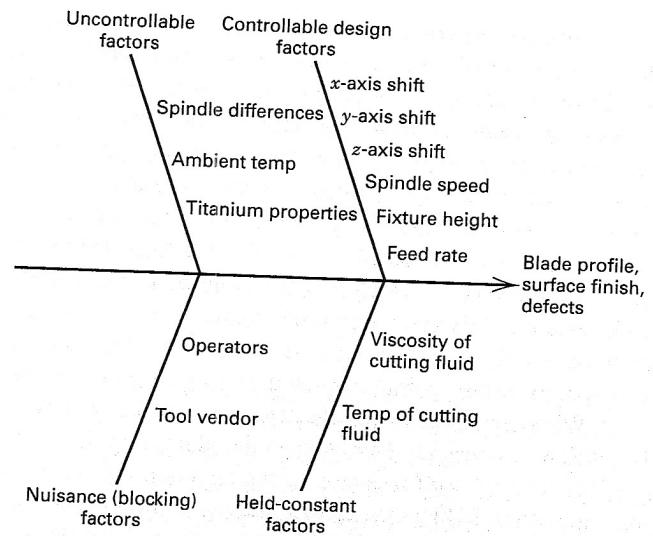
Once the experimenter has selected the design factors, he or she must choose the ranges over which these factors will be varied and the specific levels at which runs will be made. Thought must also be given to how these factors are to be controlled at the desired values and how they are to be measured. For instance, in the flow solder experiment, the engineer has defined 12 variables that may affect the occurrence of solder defects. The experimenter will also have to decide on a region of interest for each variable (that is, the range over which each factor will be varied) and on how many levels of each variable to use. **Process knowledge** is required to do this. This process knowledge is usually a combination of practical experience and theoretical understanding. It is important to investigate all factors that may be of importance and to be not overly influenced by past experience, particularly when we are in the early stages of experimentation or when the process is not very mature.

When the objective of the experiment is **factor screening** or **process characterization**, it is usually best to keep the number of factor levels low. Generally, two levels work very well in factor screening studies. Choosing the region of interest is also important. In factor screening, the region of interest should be relatively large—that is, the range over which the factors are varied should be broad. As we learn more about which variables are important and which levels produce the best results, the region of interest in subsequent experiments will usually become narrower.

The **cause-and-effect diagram** can be a useful technique for organizing some of the information generated in pre-experimental planning. Figure 1.10 is the cause-and-effect diagram constructed while planning an experiment to resolve problems with wafer charging (a charge accumulation on the wafers) encountered in an etching tool used in semiconductor manufacturing. The cause-and-effect diagram is also known as a **fishbone diagram** because the “effect” of interest or the response variable is drawn along the spine of the diagram and the potential causes or design factors are organized in a series of ribs. The cause-and-effect diagram uses the traditional causes of measurement, materials, people, environment, methods, and machines to organize the information and potential design factors. Notice that some of the individual causes will probably lead directly to a design factor that



■ FIGURE 1.10 A cause-and-effect diagram for the etching process experiment



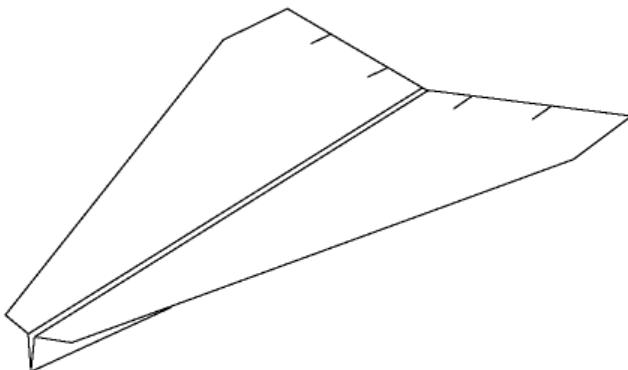
■ FIGURE 1.11 A cause-and-effect diagram for the CNC machine experiment

will be included in the experiment (such as wheel speed, gas flow, and vacuum), while others represent potential areas that will need further study to turn them into design factors (such as operators following improper procedures), and still others will probably lead to either factors that will be held constant during the experiment or blocked (such as temperature and relative humidity). Figure 1.11 is a cause-and-effect diagram for an experiment to study the effect of several factors on the turbine blades produced on a computer-numerical-controlled (CNC) machine. This experiment has three response variables: blade profile, blade surface finish, and surface finish defects in the finished blade. The causes are organized into groups of controllable factors from which the design factors for the experiment may be selected, uncontrollable factors whose effects will probably be balanced out by randomization, nuisance factors that may be blocked, and factors that may be held constant when the experiment is conducted. It is not unusual for experimenters to construct several different cause-and-effect diagrams to assist and guide them during preexperimental planning. For more information on the CNC machine experiment and further discussion of graphical methods that are useful in preexperimental planning, see the supplemental text material for this chapter.

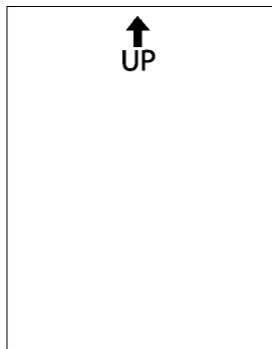
We reiterate how crucial it is to bring out all points of view and process information in steps 1 through 3. We refer to this as **pre-experimental planning**. Coleman and Montgomery (1993) provide worksheets that can be useful in pre-experimental planning. Also see the **supplemental text material** for more details and an example of using these worksheets. It is unlikely that one person has all the knowledge required to do this adequately in many situations. Therefore, we strongly argue for a team effort in planning the experiment. Most of your success will hinge on how well the pre-experimental planning is done.

4. **Choice of experimental design.** If the above pre-experimental planning activities are done correctly, this step is relatively easy. Choice of design involves consideration of sample size (number of replicates), selection of a suitable run order for the experimental trials, and determination of whether or not blocking or other randomization restrictions are involved. This book discusses some of the more important types of

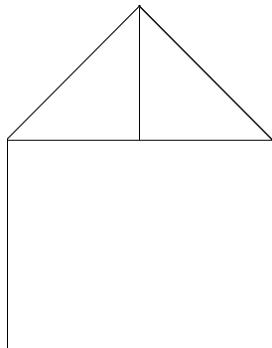
Arrow



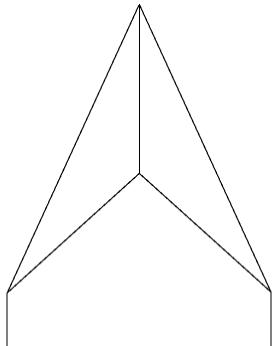
This plane is easy to fold and flies straight and smooth. Add a small amount of up elevator for long level flights.



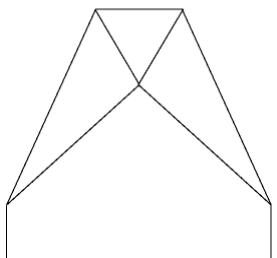
Orient the template with the “UP” arrow at the top of the page. Then, flip the paper over onto its backside, so that you cannot see any of the fold lines.



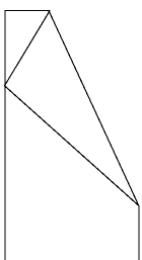
Pull the top right corner down toward you until fold line 1 is visible and crease along the dotted line. Repeat with the top left corner.



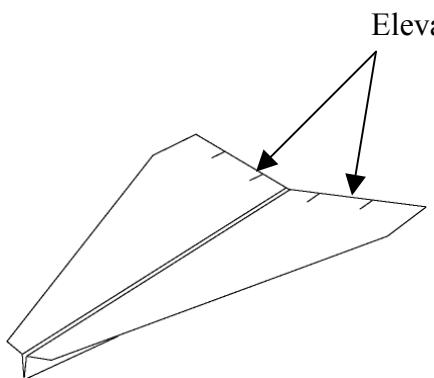
Fold the right side over again and crease along fold line 2.
Repeat with the left side.



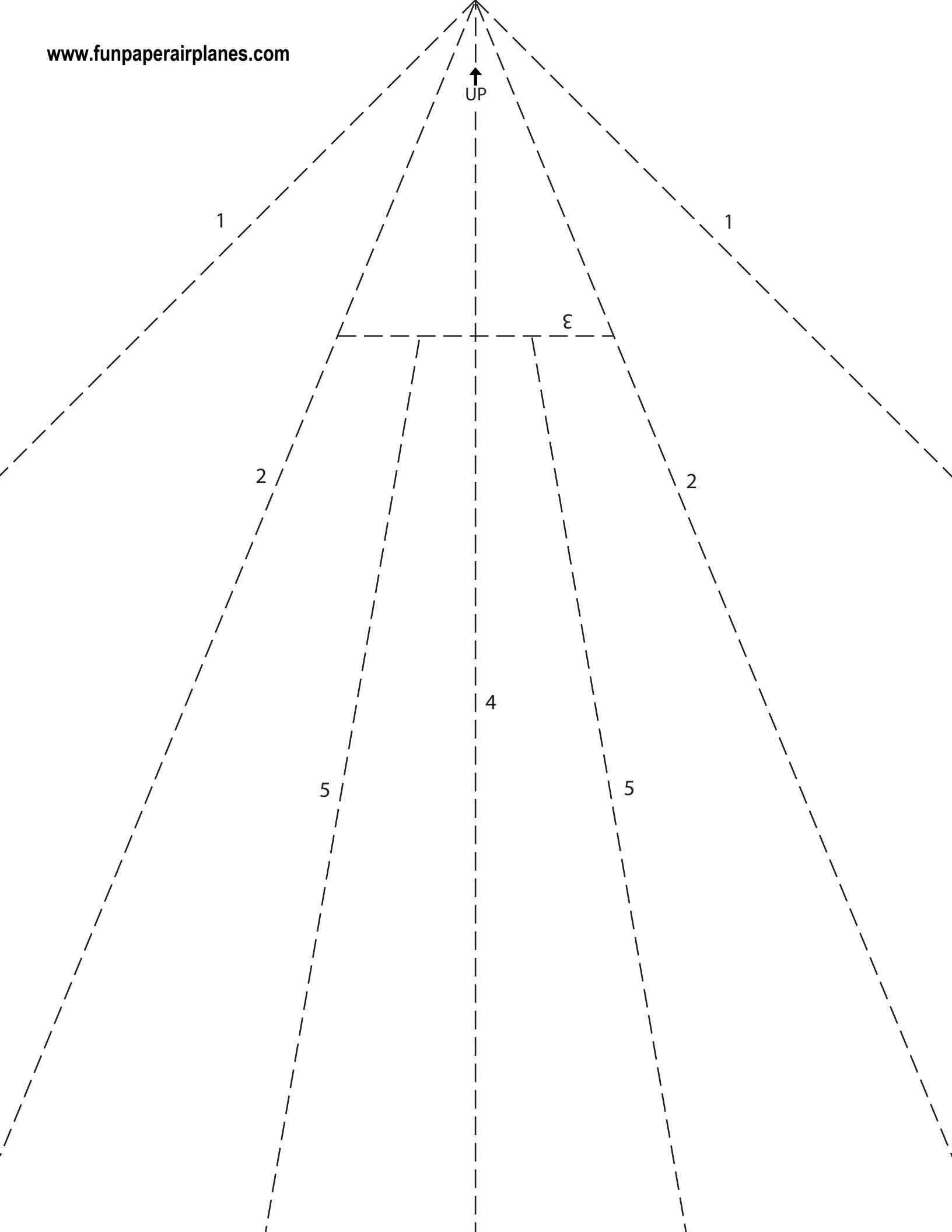
Fold the tip down toward you and crease along fold line 3.



Now, flip the paper over. Then, fold the left side over onto the right side and crease along fold line 4 so that the outside edges of the wings line up.



Fold the wings down along fold lines 5. Partially open the folds you just created so that the wings stick out straight. Cut two slits, one inch apart, along the back edge of each wing for elevator adjustments. Add wing dihedral by tilting the wings up slightly away from the fuselage. The wings will have a slight "V" shape when viewed from the front. Read the Introduction for more information about dihedral. Now you are ready to fly!



Drug	95% CI
carbamazepine	(95, 101)
lamotrigine	(98, 104)
phenytoin	(94, 104)
valproate	(88, 97)

▲ **TABLE 9.1** 95% confidence intervals for the average IQ of children whose mothers took various epilepsy drugs during their pregnancies.

Why did these four intervals lead the researchers to recommend that pregnant women not use valproate as a “first choice” drug for epilepsy? The researchers wrote that “Although the confidence intervals for carbamazepine and phenytoin overlap with the confidence interval for valproate, the confidence intervals for the differences between carbamazepine and valproate and between phenytoin and valproate do not include zero.” What does this tell us?

In this chapter we discuss how confidence intervals can be used to estimate characteristics of a population—in this case, the population of all children of women with epilepsy who took one of these drugs during pregnancy. Confidence intervals can also be used to judge between hypotheses about the means and about differences between means. The population of pregnant women with epilepsy is large, and yet if conditions are right, we can make decisions and reach an understanding about the entire population on the basis of a small sample. At the end of this chapter, we will return to this study and see if we can better understand its conclusions.

SECTION 9.1

Sample Means of Random Samples

As you learned in Chapter 7, we estimate population parameters by collecting a random sample from that population. We use the collected data to calculate a statistic, and this statistic is used to estimate the parameter. Whether we are using the statistic \hat{p} to estimate the parameter p or are using \bar{x} to estimate μ , if we want to know how close our estimate is to the truth, we need to know how far away that statistic is, typically, from the parameter.

Just as we did in Chapter 7 with \hat{p} , we now examine three characteristics of the behavior of the sample mean: its accuracy, its precision, and its probability distribution. By understanding these characteristics, we’ll be able to measure how well our estimate performs and thus make better decisions.

As a reminder, Table 9.2 shows the relation between some commonly used statistics and the parameters they estimate. (This table originally appeared as Table 7.2.)

► **TABLE 9.2**

Statistic (based on data)		Parameter (typically unknown)	
Sample mean	\bar{x}	Population mean	μ (mu)
Sample standard deviation	s	Population standard deviation	σ (sigma)
Sample variance	s^2	Population variance	σ^2
Sample proportion	\hat{p} (p -hat)	Population proportion	p

► Details

Mu-sings

The mean of a population, represented by the Greek character μ , is pronounced “mu” as in *music*.

This chapter uses much of the vocabulary introduced in Chapters 7 and 8, but we’ll remind you of important terms as we proceed. To help you visualize how a sample mean based on randomly sampled data behaves, we’ll make use of the by-now-familiar technique of simulation. Our simulation is slightly artificial, because to do a simulation, we need to know the population. However, after using a simulation to understand how the sample mean behaves in this artificial situation, we will discuss what we do in the real world when we do not know very much about the population.



Accuracy and Precision of a Sample Mean

The reason why the sample mean is a useful estimator for the population mean is that the sample mean is accurate and, with a sufficiently large sample size, very precise. The accuracy of an estimator, you'll recall, is measured by the **bias**, and the precision is measured by the standard error. You will see in this simulation that

1. The sample mean is unbiased when estimating the population mean—that is, on average, the sample mean is the same as the population mean.
2. The **precision** of the sample mean depends on the variability in the population, but the more observations we collect, the more precise the sample mean becomes.

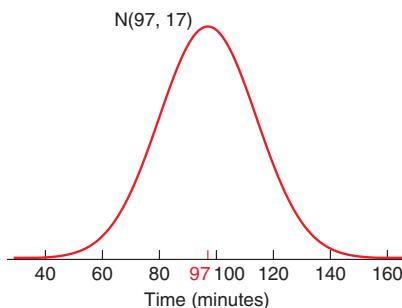
For our simulation, we'll use the population that consists of the finishing times of everyone who ran the Cherry Blossom Ten Mile Run in 2009 (Kaplan 2009). (The finishing time is the amount of time it took to finish the race.) This race is held every spring in Washington, D.C. As in most such races, data are carefully collected on every participant. Rather than showing you the histogram of the finishing times for all 8600 runners, we're going to take advantage of the fact that the distribution closely follows a Normal model with a mean of 97 minutes and a standard deviation of 17 minutes: $N(97, 17)$. Figure 9.1 shows the distribution of this population.

The population parameters are, in symbols,

$$\mu = 97 \text{ minutes}$$

$$\sigma = 17 \text{ minutes}$$

For our simulation, we will randomly sample 30 runners and calculate the average of their finishing times. We'll then repeat this many, many times. (The exact number of repetitions we performed isn't important for this discussion.) We are interested in two questions: (1) What is the typical value of the sample mean? If it is 97 minutes—the value of the population mean—then the sample mean is unbiased. (2) Typically, how far away is a sample mean from 97? In other words, how much spread is in the distribution of sample means? This spread helps us measure the precision of the sample mean as an estimator of the population mean.



Looking Back

Bias and Precision

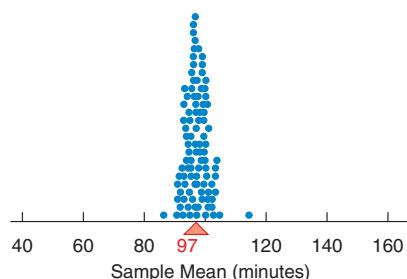
Bias is the mean distance between the sample statistic and the parameter it is estimating. Precision is measured by the standard error, which tells us how far the statistic typically deviates from its center. Review Figure 7.2 to help visualize these two properties.

For example, our very first sample of 30 runners had a mean finishing time of 95.1 minutes. We plotted this sample mean, as well as the means of the many other samples we took, in Figure 9.2 on the next page. The plot is on the same scale as Figure 9.1, the picture of the population distribution, so that you can see how much narrower this distribution of sample means is.

From this dotplot of sample means, we learn that the typical value of the sample means is the same as the population mean of 97 minutes. And we see that the sample mean is a relatively precise estimator: All of the sample means are within about 10 minutes (either above or below) the true mean value of 97 minutes.

◀ FIGURE 9.1 Finishing times in the Cherry Blossom Ten Mile Run follow a Normal model with a mean of 97 minutes and a standard deviation of 17 minutes.

► **FIGURE 9.2** Each dot represents a sample mean based on 30 runners who were randomly selected from the population whose distribution is shown in Figure 9.1. Note that the spread of this distribution is much smaller than the spread of the population, but the center looks to be at about the same place: 97 minutes.



Looking Back

Sampling Distribution

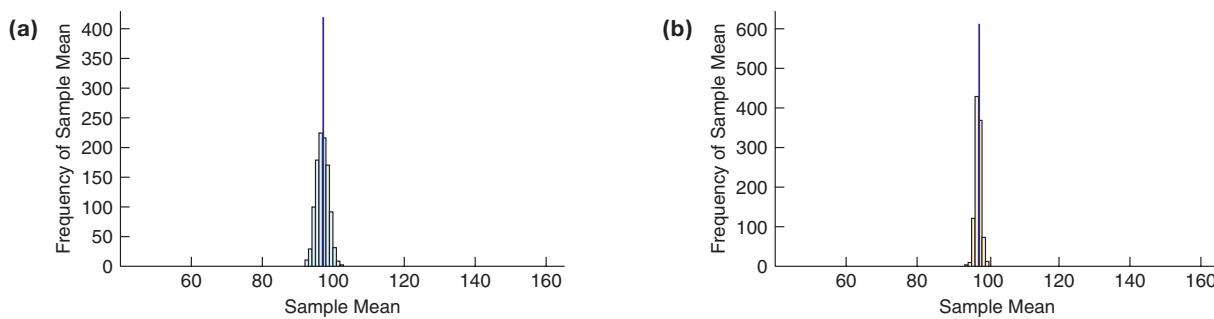
In Chapter 7 we introduced the sampling distribution of sample proportions. The sampling distribution of sample means is the same concept: It is a distribution that gives us probabilities for sample means drawn from a population. The sampling distribution is the distribution of all possible sample means.

Figure 9.2 is a very approximate picture of the **sampling distribution** of the sample mean for samples of size 30. Recall that a sampling distribution is a probability distribution of a statistic; in this case, the statistic is the sample mean. You can think of the sampling distribution as the distribution of *all* possible sample means that would result from drawing repeated random samples of a certain size from the population.

When the mean of the sampling distribution is the same value as the population mean, we say that the statistic is an **unbiased estimator**. This appears to be the case here, because both the mean of the distribution of sample means in Figure 9.2 and the population mean are about 97 minutes.

The standard deviation of the sampling distribution is what we call the **standard error**. The standard error measures the precision of an estimator by telling us how much the statistic varies from sample to sample. For the sample mean, the standard error is smaller than the population standard deviation. We can see this because the spread for the sampling distribution is smaller than the spread of the population distribution. Soon you'll see how to calculate the standard error.

What happens to the center and spread of the sampling distribution if we increase the sample size? Let's start all over with the simulation. But this time, we take a random sample of 100 runners and calculate the mean. We then repeat this many hundreds of times. Figure 9.3 shows the results for this simulation and also for two new



► **FIGURE 9.3** (a) Histogram of a large number of sample means. Each sample mean is based on a sample of 100 randomly selected runners. (b) Sample means based on samples of 500 runners. (c) Sample means based on samples of 1000 runners. Each time, the sampling distribution gets narrower, reflecting a smaller standard error.



simulations where each sample mean is based on 500 runners and then on 1000 runners. The scale of the x -axis is the same as in Figure 9.1. Note that the spread of the distributions becomes quite small—so small, in fact, that we can't get a good look at the shape of the distributions.

What Have We Demonstrated with These Simulations?

Because the sampling distributions are always centered at the population mean, we have demonstrated that the sample mean is an unbiased estimator of the population mean. We saw this for only one type of population distribution: the Normal distribution. But in fact, this is the case for any population distribution.

We have demonstrated that the standard deviation of the sampling distribution, which is called the **standard error** of the sample mean, gets smaller with larger sample size. This is true for any population distribution.

We can be more precise. If the symbol μ represents the mean of the population and if σ represents the standard deviation of the population, then

1. The mean of the sampling distribution is also μ (which tells us that the sample mean is unbiased when estimating the population mean).
2. The standard error is $\frac{\sigma}{\sqrt{n}}$ (which tells us that the standard error depends on the population distribution and is smaller for larger samples).



Looking Back

Sample Proportions from Random Samples

Compare the properties of the sample mean to those of the sample proportion, \hat{p} , given in Chapter 7. The sample proportion is also an unbiased estimator (for estimating the population proportion, p). It has standard error

$$\sqrt{\frac{p(1-p)}{n}}.$$

KEY POINT

For all populations, the sample mean is unbiased when estimating the population mean. The standard error of the sample mean is $\frac{\sigma}{\sqrt{n}}$, so the sample mean is more precise for larger sample sizes.

EXAMPLE 1 iTunes Library Statistics

A student's iTunes library of mp3s has a very large number of songs. The mean length of the songs is 243 seconds, and the standard deviation is 93 seconds. The distribution of song lengths is right-skewed. Using his mp3 player, this student will create a playlist that consists of 25 randomly selected songs.

QUESTIONS

- a. Is the mean value of 243 minutes an example of a parameter or a statistic? Explain.
- b. What should the student expect the average song length to be for his playlist?
- c. What is the standard error for the mean song length of 25 randomly selected songs?

SOLUTIONS

- a. The mean of 243 is an example of a parameter, because it is the mean of the population that consists of all of the songs in the student's library.
- b. The sample mean length can vary, but is typically the same as the population mean: 243 seconds.
- c. The standard error is $\frac{\sigma}{\sqrt{n}} = \frac{93}{\sqrt{25}} = \frac{93}{5} = 18.6$ seconds.

TRY THIS! Exercise 9.9



SECTION 9.2

The Central Limit Theorem for Sample Means

In the last simulation, all of the approximate sampling distributions (Figures 9.2 and 9.3) were Normal. This probably doesn't surprise you, because the population distribution was Normal.

 Looking Back
CLT for Proportions

In Chapter 7, you saw that the Central Limit Theorem applies to sample proportions. Here you'll see that it also applies to sample means.

What might surprise you is that the sampling distribution of the mean is always Normal (or at least approximately Normal), regardless of the shape of the population distribution. (If the sample size is small, however, the approximation can be pretty lousy.) This is the conclusion of the Central Limit Theorem, an important mathematical theorem that tells us that as long as the sample size is large, we can use the Normal distribution to perform statistical inference, regardless of the population the data are sampled from.

The **Central Limit Theorem (CLT)** assures us that no matter what the shape of the population distribution, if a sample is selected such that the following conditions are met, then the distribution of sample means follows an approximately Normal distribution. The mean of this distribution is the same as the population mean. The standard deviation (also called the standard error) of this distribution is the population standard deviation divided by the square root of the sample size. As a rule of thumb, sample sizes of 25 or more may be considered "large."

Condition 1: Random Sample and Independence. Each observation is collected randomly from the population, and observations are independent of each other. The sample can be collected either with or without replacement.

Condition 2: Normal. Either the population distribution is Normal or the sample size is large.

Condition 3: Big Population. If the sample is collected without replacement, then the population must be at least 10 times larger than the sample size.

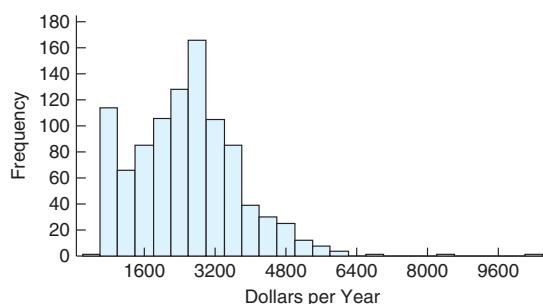
 KEY POINT

The sampling distribution of \bar{x} is approximately $N\left(\mu, \frac{\sigma}{\sqrt{n}}\right)$, where μ is the mean of the population and σ is the standard deviation of the population. The larger the sample size, n , the better the approximation. If the population is Normal to begin with, then the sampling distribution is exactly a Normal distribution.

Visualizing Distributions of Sample Means

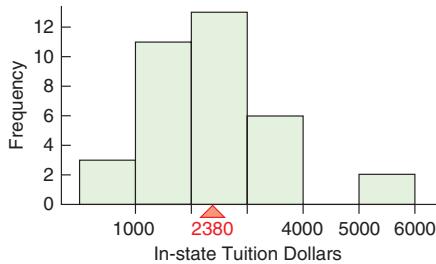
The histogram in Figure 9.4 shows the distribution of in-state tuition and fees for all two-year colleges in the United States for the 2008–2009 academic year (*The Chronicle of Higher Education, Facts & Figures* 2008). Note that the distribution has an unusual feature: It has a mode near 0. This is because all two-year colleges in California charged \$647 for California residents. The distribution is right-skewed.

► **FIGURE 9.4** Distribution of annual tuitions and fees at all two-year colleges in the United States for the 2008–2009 academic year.



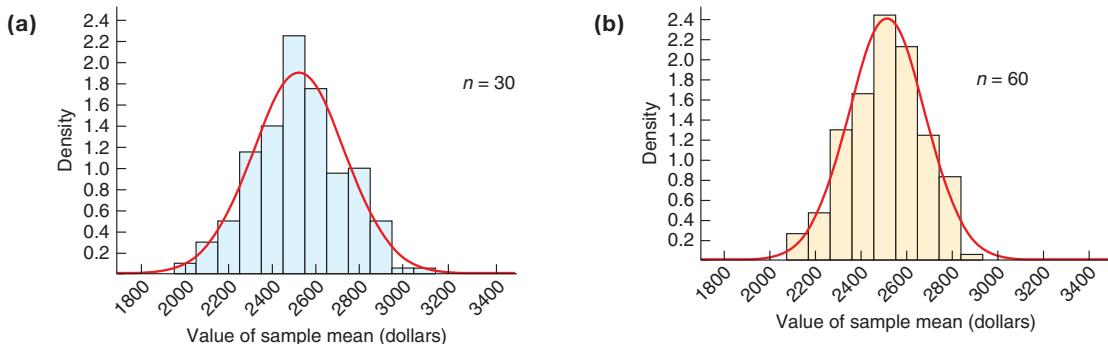
This histogram represents the distribution of a population, because it includes *all* two-year colleges. The mean of this population—the “typical” tuition of all two-year colleges—is \$2535.

Using this distribution, we now show the results of a simulation that should be starting to feel familiar. First, we take a random sample of 30 colleges. The distribution of this sample is shown in Figure 9.5. We find the mean tuition of the 30 colleges in the sample and record this figure; for example, the sample mean for the sample shown in Figure 9.5 is about \$2380.



◀ FIGURE 9.5 Distribution of a sample of 30 colleges taken from the population of all colleges. The mean of this sample, \$2380, is indicated.

We repeat this activity (that is, we sample another 30 colleges from the population of all colleges and record the mean tuition of the sample) 200 times. When we are finished, we have 200 sample mean tuitions, each sample mean based on a sample of 30 colleges. Figure 9.6a shows this distribution. Figure 9.6b shows the distribution of averages when, instead of sampling 30 colleges, we double the number and sample 60 colleges. What differences do you see among the population distribution (Figures 9.4), the distribution of one of the samples (Figure 9.5), the sampling distribution when the sample size is 30 (Figure 9.6a), and the sampling distribution when the sample size is 60 (Figure 9.6b)?



▲ FIGURE 9.6 (a) Distribution of sample means, where each sample mean is based on a sample size of $n = 30$ college tuitions and is drawn from the population shown in Figure 9.4. This is (approximately) the sampling distribution of \bar{x} when $n = 30$. A Normal curve is superimposed. (b) The approximate sampling distribution of \bar{x} when $n = 60$.

Both of the sampling distributions in Figures 9.6a and 9.6b show us the values and relative frequencies for \bar{x} , but they are based on different sample sizes. We see that even though the *population* distribution has an unusual shape (Figure 9.4), the sampling distributions for \bar{x} are fairly symmetric and unimodal. Although the Normal curve that is superimposed doesn’t match the histogram very closely when $n = 30$, the match is pretty good for $n = 60$.

This is exactly what the CLT predicts. When the sample size is large enough, we can use the Normal distribution to find approximate probabilities for the values we might see for \bar{x} when we take a random sample from the population.

The more observations in your sample, the better an approximation the Normal distribution provides. Generally, the CLT provides a useful approximation of the true



Looking Back

Distribution of a Sample vs. Sampling Distribution
Recall that these are two different concepts. The *distribution of a sample*, from Chapter 3, is the distribution of one single sample of data (Figure 9.5). The *sampling distribution*, on the other hand, is the probability distribution of an estimator or statistic such as the sample mean (Figures 9.6a and 9.6b).

probabilities if the sample size is 25 or more. But this is just a rule of thumb. Be aware that you might need larger sample sizes in some situations. Unlike in Chapter 7, where we worked with sample proportions, we can't provide a hard-and-fast rule for sample size. For nearly all examples in this book, though not always in real life, 25 is large enough.

Applying the Central Limit Theorem

The Central Limit Theorem helps us find probabilities for sample means when those means are based on a random sample from a population. Example 2 demonstrates how we can answer probability questions about the sample mean even if we can't answer probability questions about individual outcomes.

EXAMPLE 2 Pulse Rates Are Not Normal

According to one very large study done in the United States, the mean resting pulse rate of adult women is about 74 beats per minute (bpm), and the standard deviation of this population is 13 bpm (NHANES 2003–2004). The distribution of resting pulse rates is known to be skewed right.

QUESTIONS

- Suppose we take a random sample of 36 women from this population. What is the approximate probability that the average pulse rate of this sample will be below 71 or above 77 bpm? (In other words, what is the probability that it will be more than 3 bpm away from the population mean of 74 bpm?)
- Can you find the probability that a single adult woman will have a resting pulse rate more than 3 bpm away from the mean value of 74?

SOLUTION

- It doesn't matter that the population distribution is not Normal. Because the sample size of 36 women is relatively large, the distribution of sample means will be approximately (though not exactly) Normal.

The mean of this Normal distribution will be the same as the population mean: $\mu = 74$ bpm. The standard deviation of this distribution is the standard error:

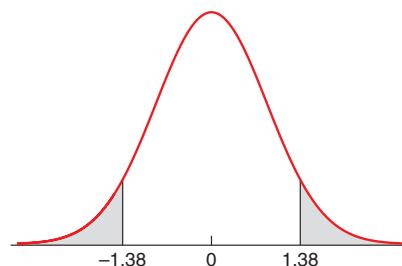
$$SE = \frac{\sigma}{\sqrt{n}} = \frac{13}{\sqrt{36}} = \frac{13}{6} = 2.167$$

To use the Normal table to find probabilities requires that the values of 71 bpm and 77 bpm be converted to standard units:

$$z = \frac{\bar{x} - \mu}{SE} = \frac{71 - 74}{2.167} = \frac{-3}{2.167} = -1.38$$

Figure 9.7 shows that the area that corresponds to the probability that the sample mean pulse rate will be more than 1.38 standard errors away from the population mean pulse rate. This probability is calculated to be about 17%.

► FIGURE 9.7 Area of the Normal curve outside of z-scores of -1.38 and 1.38 .



CONCLUSIONS

- The approximate probability that the average pulse of 36 adult women will be more than 3 bpm away from 74 bpm is about 17%.
- We cannot find the probability for a single woman because we do not know the probability distribution. We know only that it is “right-skewed,” which is not enough information to find actual probabilities.

TRY THIS! Exercise 9.11


Caution
CLT Not Universal

The CLT does not apply to all statistics you run across. It does not apply to the sample median, for example. No matter how large the sample size, you cannot use the Normal distribution to find a probability for the median value. It also does not apply to the sample standard deviation.

Many Distributions

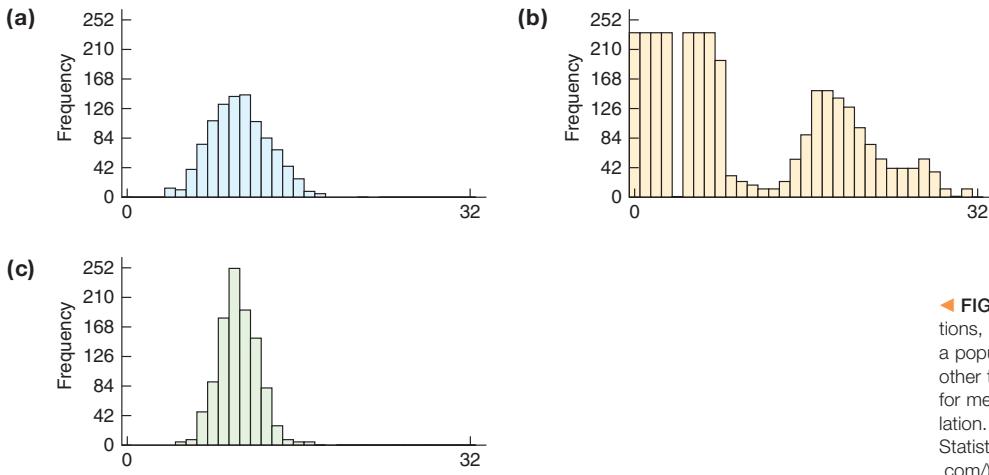
It's natural at this point to feel that you have seen a confusingly large number of types of distributions, but it's important that you keep them straight. The *population distribution* is the distribution of values from the population. Figure 9.4 (two-year college tuitions) is an example of a population distribution because it shows the distribution of *all* two-year colleges. Figure 9.1 (runners' times for all competitors in a race) is another example of a population distribution. For some populations, we don't know precisely what this distribution is. Sometimes we assume (or know) that it is Normal, sometimes we know it is skewed in one direction or the other, and sometimes we know almost nothing.

We then take a random sample of n observations. We can make a histogram of these data. This histogram gives us a picture of the *distribution of the sample*. If the sample size is large, and if the sample is random, then the sample will be representative of the population, and the distribution of the sample will look similar to (but not the same as!) the population distribution. Figure 9.5 is an example of the distribution of a sample of size $n = 30$ taken from the population of two-year college tuitions.

The *sampling distribution* is more abstract. If we take a random sample of data and find the sample mean (the center of the distribution of the sample), and then repeat this many, many times, we will get an idea what the sampling distribution looks like. Figures 9.6a and 9.6b are examples of approximate sampling distributions for the sample mean, based on samples from the two-year college tuition data. Note that these do not share the same shape as the population or the sample; they are both approximately Normal.

EXAMPLE 3 Identify the Distribution

Figure 9.8 shows three distributions. One distribution is a population. The other two distributions are (approximate) sampling distributions. One sampling distribution is based on sample means of size 10, and the other is based on sample means of size 25.



◀ FIGURE 9.8 Three distributions, all on the same scale. One is a population distribution, and the other two are sampling distributions for means sampled from the population. (Source: Rice Virtual Lab in Statistics, <http://onlinestatbook.com/>)

QUESTION Which graph (a, b, or c) is the population distribution? Which shows the sampling distribution for the mean with $n = 10$? Which with $n = 25$?

SOLUTION The Central Limit Theorem tells us that sampling distributions for means are approximately Normal. This implies that Figure 9.8b is not a sampling distribution, so it must be the population distribution from which the samples were taken. We know that the sample mean is more precise for larger samples, and because Figure 9.8a has the larger standard error (is wider), it must be the graph associated with $n = 10$. This means that Figure 9.8c is the sampling distribution of means with $n = 25$.

TRY THIS! Exercise 9.13



SNAPSHOT THE SAMPLE MEAN (\bar{x})

- WHAT IS IT?** ▶ The arithmetic average of a sample of data.
- WHAT DOES IT DO?** ▶ Estimates the mean value of a population, μ . The mean is used as a measure of what is “typical” for a population.
- HOW DOES IT DO IT?** ▶ If the sample was a random sample, then the sample mean is unbiased, and we can make the precision of the estimator as good as we want by taking a large enough sample size.
- HOW IS IT USED?** ▶ If the sample size is large enough (or the population is Normal), we can use the Normal distribution to find the probability that the sample mean will take on a value in any given range. This lets us know how wrong our estimate could be.

The *t*-Distribution

The hypothesis tests and confidence intervals that we will use for estimating and testing the mean are based on a statistic called the ***t*-statistic**:

$$t = \frac{\bar{x} - \mu}{\left(\frac{s}{\sqrt{n}}\right)}$$

The *t*-statistic is very similar to a *z*-score for the sample mean. In the numerator, we subtract the population mean from the sample mean. Then we divide not by the standard error but, rather, by an *estimate* of the standard error.

It would be nice if we could divide by the true standard error. But in real life, we almost never know the value of σ , the population standard deviation. So instead, we replace it with an estimate: the sample standard deviation, s . This gives us an estimate of the standard error:

$$SE_{EST} = \frac{s}{\sqrt{n}}$$

Compare the *t*-statistic to the *z*-statistic, and you will see that we simply replaced σ in the *z*-statistic with s .

$$z = \frac{\bar{x} - \mu}{\left(\frac{\sigma}{\sqrt{n}}\right)}$$

The *t*-statistic does *not* follow the Normal distribution. One reason for this is that the denominator changes with every sample. For this reason, the *t*-statistic is



Looking Back

Sample Standard Deviation

In Chapter 3 we gave the formula for the sample standard deviation:

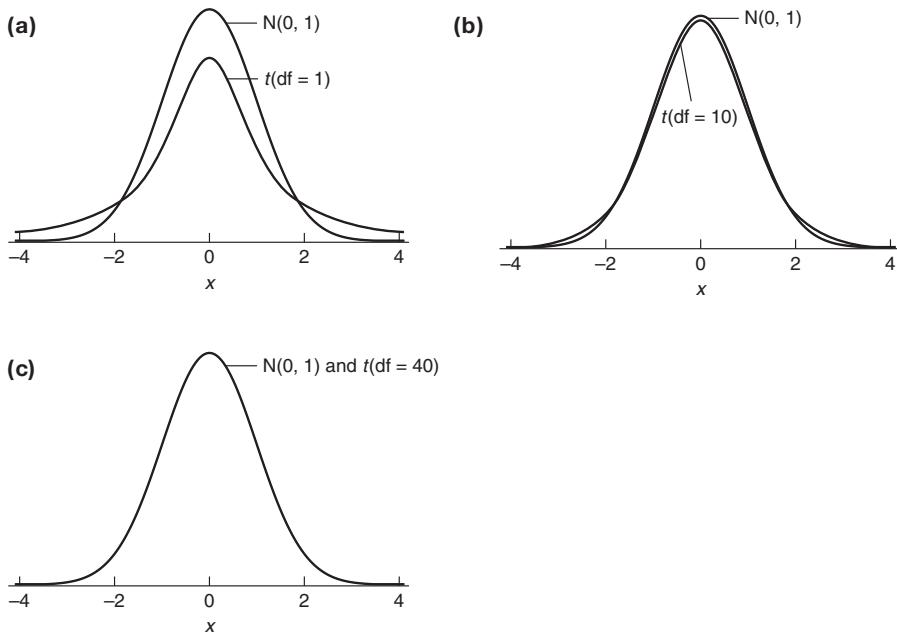
$$s = \sqrt{\frac{\sum(x - \bar{x})^2}{n - 1}}$$

more variable than the z -statistic (whose denominator is always the same.) Instead, the t -statistic follows a distribution called—surprise!—the **t -distribution**. This was Gosset's great discovery at the Guinness brewery. When small sample sizes were used to make inferences about the mean, even if the population was Normal, the Normal distribution just didn't fit the results that well. Gosset discovered a new distribution, which he called the t -distribution, that turned out to be a better model than the Normal for the sampling distribution of \bar{x} when σ is not known.

The t -distribution shares many characteristics with the $N(0, 1)$ distribution. Both are symmetric, are unimodal, and might be described as “bell-shaped.” However, the t -distribution has thicker tails. This means that in a t -distribution, it is more likely that we will see extreme values (values far from 0) than it is in a standard Normal distribution.

The t -distribution's shape depends on only one parameter, called the **degrees of freedom (df)**. The number of degrees of freedom is (usually) an integer: 1, 2, 3, and so on. If df is small, then the t -distribution has very thick tails. As the degrees of freedom get larger, the tails get thinner. Ultimately, when df is infinitely large, the t -distribution is exactly the same as the $N(0, 1)$ distribution.

Figure 9.9 shows t -distributions with 1, 10, and 40 degrees of freedom. In each case, the t -distribution is shown with a $N(0, 1)$ curve so that you can compare them. The t -distribution is the one whose tails are “higher” at the extremes. Note that by the time the degrees of freedom reaches 40 (Figure 9.9c), the t -distribution and the $N(0, 1)$ distribution are too close to tell apart (on this scale).



Details

Degrees of Freedom

Degrees of freedom are related to the sample size: Generally, the larger the sample size, the larger the degrees of freedom. When estimating a single mean, as we are doing here, the number of degrees of freedom is equal to the sample size minus one.

$$df = n - 1$$

◀ FIGURE 9.9 (a) A t -distribution with 1 degree of freedom, along with a $N(0, 1)$ distribution. The t -distribution has much thicker tails. (b) The degrees of freedom are now equal to 10, and the tails are only slightly thicker in the t -distribution. (c) The degrees of freedom are now 40, and the two distributions are visually indistinguishable.

SECTION 9.3

Answering Questions about the Mean of a Population

Do you commute to work? How long does it take you to get there? Is this amount of time typical for others in your state? Which state has the greatest commuting times? This information is important not just to those of us who must fight traffic every day, but also to business leaders and politicians who make decisions about quality of living

and the cost of doing business. The U.S. Census performs periodic surveys that determine, among other things, commuting times around the country. In 2007 (the last time the survey was done), the state of New York had the greatest mean commuting time, which was 31.5 minutes. Vermont was lowest at 21.2 minutes.

These means are estimates of the mean commuting time for all residents in these states who work outside of their homes. We can learn the true mean commuting time only by asking all residents, which is clearly too time-consuming to do very often. Instead, the U.S. Census takes a random sample of U.S. residents to estimate these values.

In this section we present two techniques for answering questions about the population mean. Confidence intervals are used for estimating values. Hypothesis tests are used for deciding whether the value is one thing or another. These are the same methods that were introduced in Chapter 7 (confidence intervals) and Chapter 8 (hypothesis tests) for population proportions, but here you'll see how they are modified to work with means.

Estimation with Confidence Intervals

Confidence intervals are a technique for communicating an estimate of the mean along with a measure of our uncertainty in that estimate. The job of a confidence interval is to provide us with a range of values that, according to the data, are highly plausible values for the unknown population mean. For instance, the range of values for the mean commuting time for all Vermont commuters is 21.1 to 21.3 minutes.

Not all confidence intervals do an equally good job; the "job performance" of a confidence interval is therefore measured with something called the **confidence level**. The higher this level, the better the confidence interval performs. The confidence interval for mean Vermont commuting times is 90%, which means we can be extremely confident that this interval contains the true mean.

Sometimes, you will be in a situation in which you will know only the sample mean and sample standard deviation. In these situations, you can use a calculator to find the confidence interval. However, if you have access to the actual data, you are much better off using statistical software to do all the calculations for you. We will show you how to respond to both situations.

No matter which situation you are in, you will need to judge whether a confidence interval is appropriate for the situation, and you will need to interpret the confidence interval. Therefore, we will discuss these essential skills before demonstrating the calculations.

When Are Confidence Intervals Useful?

A confidence interval is a useful answer to the following questions: "What's the typical value for a variable in this large group of objects or people? And how far away from the truth might this estimate of the typical value be?" You should provide a confidence interval whenever you are estimating the value of a population parameter on the basis of a random sample from that population. For example, judging on the basis of a random sample of 30 adults, what's the typical body temperature of all healthy adults? On the basis of a survey of a random sample of Vermont residents, what's the typical commuting time for all Vermont residents? A confidence interval is useful for answering questions such as these because it communicates the uncertainty in our estimate and provides a range of plausible values.

A confidence interval is not appropriate if there is no uncertainty in your estimate. This would be the case if your "sample" were actually the entire population. For example, it is not necessary to find a confidence interval for the mean score on your class's statistics exam. The population is your class, and all of the scores are known. Thus the population mean is known, and there is no need to estimate it.

Checking Conditions

In order to measure the correct confidence level, the following conditions must hold:

Condition 1: *Random Sample and Independence*. The data must be collected randomly, and each observation must be independent of the others.

Condition 2: Normal. Either the population must be Normally distributed or the sample size must be fairly large (at least 25).

If these conditions do not hold, then we cannot measure the job performance of the interval; the confidence level may be incorrect. This means that we may advertise a 95% confidence level when, in fact, the true performance is much worse than this.

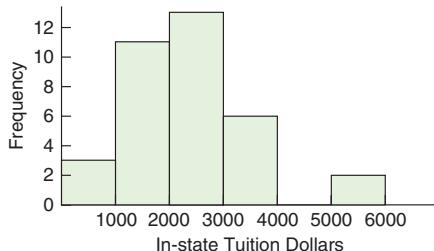
To check the first condition, you must know how the data were collected. This is not always possible, so rather than checking these conditions, you must simply assume that they hold. If they do not, your interval will not be valid.

The requirement for independence means that measurement of one object in the sample does not affect any other. Essentially, if we know the value of any one observation, this knowledge should tell us nothing about the values of other observations. This condition might be violated if, say, we randomly sampled several schools and gave all of the students math exams. The individual math scores would not be independent, because we would expect that students within the same school might have similar scores.

The second condition is due to the Central Limit Theorem. If the population distribution is Normal (or very close to it), then we have nothing to worry about. But if it is non-Normal, then we need a large enough sample size so that the sampling distribution of sample means is approximately Normal. For many applications, a sample size of 25 is large enough, but for extremely skewed distributions, you might need an even larger sample size.

EXAMPLE 4 Is the Cost of College Rising?

Many cities and states are finding it more difficult to offer low-cost college educations. Did the mean cost of attending two-year colleges increase in the United States from the 2007–2008 academic year to 2009–2010? In 2007–2008, the mean cost of *all* two-year colleges was \$2429. A random sample of 35 two-year colleges in the United States found that the average tuition charged in 2008–2009 was \$2380, with a standard deviation of \$1160. Figure 9.10 shows the distribution of this sample. On the basis of this, a 90% confidence interval for the mean cost of attending two-year colleges in 2008–2009 is \$2048 to \$2711.



◀ FIGURE 9.10 Distribution of in-state tuition for a random sample of 35 two-year colleges during the academic year 2008–2009.

QUESTIONS

- Describe the population. Is the number \$2380 an example of a parameter or a statistic?
- Verify that the conditions for a valid confidence interval are met.

SOLUTIONS

- The population consists of all two-year college tuitions (for in-state residents) in the academic year 2008–2009. (There are roughly 1000 two-year colleges in the United States.) The number \$2380 is the mean of a sample of only 35 colleges. Because it is the mean of a sample, it is a statistic.
- The first condition is that the data represent a random sample of independent observations. We are told the sample was collected randomly, so we assume this is true.

Independence also holds, because knowledge about any one school's tuition tells us nothing about other schools in the sample. The second condition requires that the population be roughly Normally distributed or the sample size be equal to or larger than 25. We do not know the distribution of the population, but because the sample size is large enough (bigger than 25), this condition is satisfied.

**TRY THIS!** Exercise 9.15

Interpreting Confidence Intervals

To understand confidence intervals, you must know how to interpret a confidence interval and how to interpret a confidence level.

A confidence *interval* can be interpreted as a range of plausible values for the population parameter. In other words, in the case of population means, we can be confident that if we were to someday learn the true value of the population mean, it would be within the range of values given by our confidence interval. For example, the U.S. Census estimates that the mean commuting time for Vermont residents is 21.1 minutes to 21.3 minutes, with a 90% confidence level. We interpret this to mean that we can be fairly confident that the true mean commute for *all* Vermont residents is between 21.1 and 21.3 minutes. Yes, we could be wrong. The mean might be less than 21.1 minutes, or it might be more than 21.3 minutes. However, we would be rather surprised to find this was the case; we are highly confident that the mean is within this interval.

KEY POINT

A confidence interval can be interpreted as a range of plausible values for the population parameter.

EXAMPLE 5 Evidence for Changing College Costs

Based on a random sample of 35 two-year colleges, a 90% confidence interval for the mean in-state tuition at two-year colleges for the 2009–2010 academic year is \$2048 to \$2711. In the academic year 2007–2008, the mean of all two-year colleges' in-state tuitions was \$2429.

QUESTION Does the confidence interval provide evidence that the mean tuition has increased?

SOLUTION No, it does not. Although we cannot know the population mean of all tuitions in 2009/2010, we are highly confident that it is in the range of \$2048 to \$2711. This range includes the value \$2429, so there is not enough evidence to conclude that the mean tuition has changed.

TRY THIS! Exercise 9.17

Measuring Performance with the Confidence Level

The confidence *level*, which in the case of the interval for the mean Vermont commuting time was 90%, tells us about the method used to find the interval. A value for the level of 90% tells us that the U.S. Census used a method that works in 90% of all samples. In other words, if we were to take many same-sized samples of Vermont residents, and for each sample calculate a 90% confidence interval, then 90% of those intervals would contain the population mean.

The confidence level does *not* tell us whether the interval (21.1 to 21.3) contains or does not contain the population mean. The "90%" just tells us that the method that produced this interval is a pretty good method.

Suppose you decided to purchase an mp3 player online. You have your choice of several manufacturers, and they are rated in terms of their performance level. One manufacturer has a 90% performance level, which means that 90% of the players they produce are good ones, and 10% are defective. Some other manufacturers have lower levels: 80%, 60%, and worse. From whom do you buy? You choose to buy from the manufacturer with the 90% level, because you can be very confident that the player they send you will be good. Of course, once the player arrives at your home, the confidence level isn't too useful. Your player either works or does not work; there's no 90% about it.

Confidence levels work the same way. We prefer confidence intervals that have 90% or higher confidence levels, because then we know that the process that produced these levels is a good process, and therefore, we are confident in any decisions or conclusions we reach. But the level doesn't tell us whether this one particular interval sitting in front of us is good or bad. In fact, we shall never know that, unless we someday gain access to the entire population.

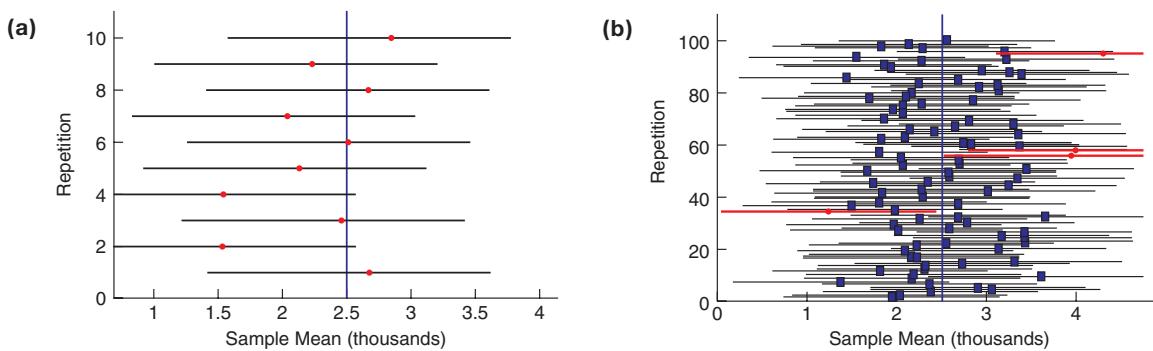
KEY POINT

The confidence level is a measure of how well the method used to produce the confidence interval performs. We can interpret the confidence level to mean that if we were to take many random samples of the same size from the same population, and for each random sample calculate a confidence interval, then the confidence level is the proportion of intervals that "work"—the proportion that contain the population parameter.

Figure 9.11 illustrates this interpretation of confidence levels. From the population of 107 movies that were playing during the week of April 3, 2009 (<http://www.the-numbers.com/>), we took a random sample (with replacement) of 30 movies and calculated the mean revenue per theater in this sample. Because the samples were random, each sample produced a different sample mean. For each sample we also calculated a 95% confidence interval. We repeated this process 100 times, and each time we made a plot of the confidence interval. Figure 9.11a shows the results from the first 10 samples of 30 theaters. All 10 samples produced "good" intervals—intervals that contain the true population mean of \$2429. Figure 9.11b shows what happened after we collected 100 95% confidence intervals. With a 95% confidence interval, we'd expect about 95% of the intervals to be good and 5% to be bad. And in fact, four intervals (shown in red) were bad.

Caution
Confidence Levels Are Not Probabilities

A confidence level, such as 90%, is not a probability. Saying we are 90% confident the mean is between 21.1 minutes and 21.3 minutes does not mean that there is a 90% chance that the mean is between these two values. It either is, or isn't. There's no probability about it.



▲ FIGURE 9.11 (a) Ten different 95% confidence intervals, each based on a separate random sample of 30 movie theaters. The population mean of \$2500 is shown with a vertical bar. All ten intervals are good because they include this population mean. (b) One hundred confidence intervals, each based on a random sample of 30 theaters. Because we are using a 95% confidence level, we expect about 95% of the intervals to be good. In fact, 96 of the 100 turned out to be good, this time. The red intervals are "bad" intervals that do not contain the population mean.

EXAMPLE 6 iPad Batteries

A consumer group wishes to test Apple's claim that the iPad has a 10-hour battery life. With a random sample of 5 iPads, and running them under identical conditions, the group finds a 95% confidence interval for the mean battery life of an iPad to be 9.5 hours to 12.5 hours. One of the following statements is a correct interpretation of the confidence level. The other is a correct interpretation of the confidence interval.

- (i) We are very confident that the mean battery life of all iPads is between 9.5 and 12.5 hours.
- (ii) In about 95% of all samples of 5 iPads, the resulting confidence interval will contain the mean battery life of all iPads.

QUESTION Which of these statements is a valid interpretation of a confidence interval? Which of these statements is a valid interpretation of a confidence level?

SOLUTION Statement (i) interprets a confidence *interval* (9.5, 12.5). Statement (ii) tells us the meaning of the 95% confidence *level*.

TRY THIS! Exercise 9.21

Calculating the Confidence Interval

Confidence intervals for means have the same basic structure as they did for proportions:

$$\text{Estimator} \pm \text{margin of error}$$

As in Chapter 7, the margin of error has the structure

$$\text{Margin of error} = (\text{multiplier}) \times SE$$

The standard error (*SE*) is $SE = \frac{\sigma}{\sqrt{n}}$. Because we usually do not know the standard deviation of the population and hence the *SE*, we replace the *SE* with its estimate. This leads to a formula similar in structure, but slightly different in details, from the one you learned for proportions.

Formula 9.1: One-Sample *t*-Interval

$$\bar{x} \pm m$$

$$\text{where } m = t^*SE_{\text{EST}} \text{ and } SE_{\text{EST}} = \frac{s}{\sqrt{n}}$$

The multiplier t^* is a constant that is used to fine-tune the margin of error so that it has the level of confidence we want. This multiplier is found using a *t*-distribution with $n - 1$ degrees of freedom. SE_{EST} is the estimated standard error.

To compute a confidence interval for the mean, you first need to choose the level of confidence. After that, you need either the original data or these four pieces of information:

1. The sample average, \bar{x} , which you calculate from the data.
2. The sample standard deviation, s , which you calculate from the data.
3. The sample size, n , which you know from looking at the data.
4. The multiplier, t^* , which you look up in a table (or use technology) and which is determined by your confidence level and the sample size n . The value of t^* tells us how wide the margin of error is, in terms of standard errors. For example, if t^* is 2, then our margin of error is two standard errors wide.

The first three steps are pretty straightforward, so let's spend a minute on finding t^* , the multiplier for the margin of error.

The multiplier is based on a t -distribution with $n - 1$ degrees of freedom. The correct values can be found in Table 4 in the Appendix, or you can use technology. Table 4 is organized such that each row represents possible values of t^* for each degree of freedom. The columns contain the values of t^* for a given confidence level. For example, for a 95% confidence level and a sample size of $n = 30$, we use $t^* = 2.045$. We find this in the table by looking in the row with $df = n - 1 = 30 - 1 = 29$ and using the column for a 95% confidence level. Refer to Table 9.3, which is from the table in the Appendix.

EXAMPLE 7 Finding the Multiplier t^*

Suppose we collect a sample of 30 iPads and wish to calculate a 90% confidence interval for the mean battery life.

QUESTION Using Table 9.3, which is from the table in Appendix A, find t^* for a 90% confidence interval when $n = 30$.

DF	Confidence Level			
	90%	95%	98%	99%
28	1.701	2.048	2.467	2.763
29	1.699	2.045	2.462	2.756
30	1.697	2.042	2.457	2.750
34	1.691	2.032	2.441	2.728

◀ TABLE 9.3 Critical Values of t .

SOLUTION We find the number of degrees of freedom from the sample size,

$$df = n - 1 = 30 - 1 = 29$$

And so we find, from Table 9.3, $t^* = 1.699$ (shown underlined)

TRY THIS! Exercise 9.23



It is best to use technology to find the multiplier, because most tables stop at 35 or 40 degrees of freedom. For a 95% confidence level, if you do not have access to technology and the sample size is bigger than 40, it is usually safe to use $t^* = 1.96$ —the same multiplier that we used for confidence intervals for sample proportions (for 95% confidence). The precise value, if we used a computer, is 2.02, but this is only 0.06 unit away from 1.96, so the result is probably not going to be affected in a big way.

The wider the confidence interval, the more confident we will be that it covers the true parameter value. We can always increase our level of confidence by making the margin of error bigger. We do this by choosing larger values for t^* .

Looking Back

Why Not 100%?

In Chapter 7, you learned that one reason why a 95% confidence level is popular is that increasing the confidence level only a small amount beyond 95% requires a much larger margin of error.

EXAMPLE 8 College Tuition Costs

A random sample of 35 two-year colleges in 2008–2009 had a mean tuition (for in-state students) of \$2380, with a standard deviation of \$1160.

QUESTION Find a 90% confidence interval and a 95% confidence interval for the mean in-state tuition of all two-year colleges in 2008–2009. Interpret the intervals. Assume the necessary conditions hold.

SOLUTION We are given the desired confidence level, the standard deviation, and the sample mean, \bar{x} , so the next step is to calculate the estimated standard error.

$$\bar{x} = \$2380$$

$$SE_{EST} = \frac{1160}{\sqrt{35}} = 196.0758$$

$$\bar{x} \pm m = \bar{x} \pm t^*SE_{EST}$$

We find the appropriate values of t^* (from Table 9.3 above):

$$t^* \text{ (for 90\%)} = 1.691$$

$$t^* \text{ (for 95\%)} = 2.032$$

For the 90% confidence interval,

$$\bar{x} \pm t^*SE_{EST} \text{ becomes } 2380 \pm (1.691 \times 196.0758) = 2380 \pm 331.5642$$

$$\text{Lower limit: } 2380 - 331.5642 = 2048.44$$

$$\text{Upper limit: } 2380 + 331.5642 = 2711.56$$

A 90% confidence interval for the mean tuition of all two-year colleges in the 2008–2009 academic year is (\$2048, \$2712). That is, we are 90% confident that the mean tuition (the typical tuition) of all two-year colleges in 2008–2009 was between \$2048 and \$2712.

For the 95% confidence interval,

$$\bar{x} \pm t^*SE_{EST} \text{ becomes } 2380 \pm (2.032 \times 196.0758) = 2380 \pm 398.4260$$

$$\text{Lower Limit: } 2380 - 398.4260 = 1981.570$$

$$\text{Upper Limit: } 2380 + 398.4260 = 2778.4260$$

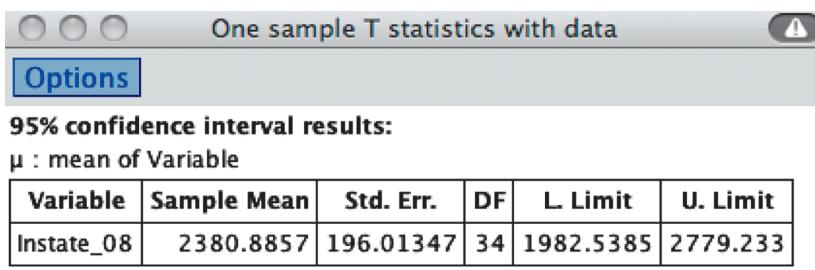
CONCLUSION The 90% confidence interval is (\$2048, \$2712). The 95% confidence interval is (\$1982, \$2778), which is wider. We are 90% confident that the mean tuition of all two-year colleges is between \$2048 and \$2712. We are 95% confident that the mean tuition could be as low as \$1982 or as high as \$2778.

TRY THIS! Exercise 9.25

Tech

If you have access to the original data (and not just to the summary statistics, as we were given in Example 8), then it is always best to use a computer to find the confidence interval for you. Figure 9.12 shows StatCrunch output that shows us what we would see if we had access to the full data for Example 8 and asked the software

► **FIGURE 9.12** StatCrunch output showing the 95% confidence interval for the mean in-state tuition at two-year colleges in the academic year 2008–2009.



to find a 95% confidence interval for the mean tuition of all two-year colleges. The output shows us the estimated mean (\$2380.8857), the standard error (\$196.01347), the degrees of freedom (34), the lower limit of the confidence interval (\$1982.5385), and the upper limit (\$2779.233). Note that these are not exactly the values we calculated in Example 8. The StatCrunch values are more accurate because there is less rounding.

Reporting and Reading Confidence Intervals

There are two ways of reporting confidence intervals. Professional statisticians tend to report (lower boundary, upper boundary). This is what we've done so far in this chapter. Thus we reported the 95% confidence interval for the mean of two-year college tuitions in 2008–2009 as (\$1982, \$2778).

In the press, however, and in some scholarly publications, you'll also see confidence intervals reported as

$$\text{Estimate} \pm \text{margin of error}$$

For the two-year college tuitions, we calculated the margin of error to be \$398.426 for 95% confidence. Thus we could also report the confidence interval as

$$\$2380 \pm \$398$$

This form is useful because it shows our estimate for the mean (\$2380) as well as our uncertainty (the mean could plausibly be \$398 lower or \$398 more).

You're welcome to choose whichever you think best, although you should be familiar with both forms.

Testing a Mean

In Chapter 8 we laid out the foundations of hypothesis testing. Here, you'll see that the same four steps can be used to test hypotheses about means of populations. These four steps are

Step 1: Hypothesize.

State your hypotheses about the population parameter.

Step 2: Prepare.

Get ready to test: Choose and state a significance level. Choose a test statistic appropriate for the hypotheses. State and check conditions required for the computations, and state any assumptions that must be made.

Step 3: Compute to compare.

Compute the observed value of the test statistic in order to compare the null hypothesis value to our observed value. Find the p-value to measure your level of surprise.

Step 4: Interpret.

Do you reject or not reject the null hypothesis? What does this mean?

As an example of testing a mean, consider this “study” one of the authors did. McDonald’s advertises that its ice cream cones have a mean weight of 3.2 ounces ($\mu = 3.2$). A human server starts and stops the machine that dispenses the ice cream, so we might expect some variation in the amount. Some cones might weight slightly more, some cones slightly less. If we weighed all of the McDonald’s ice cream cones at a particular store, would the mean be 3.2 ounces, as the company claims?

One of the authors collected a sample of five ice-cream cones and weighed them on a postage scale. The weights were (in ounces)

4.2, 3.6, 3.9, 3.4, and 3.3

We summarize these data as

$$\bar{x} = 3.68 \text{ ounces}, s = 0.3701 \text{ ounce}$$

Do these observations support the claim that the mean weight is 3.2 ounces? Or is the mean a different value? We'll apply the four steps of the hypothesis test to make a decision.

Step 1: Hypothesize

Hypotheses come in pairs and are always statements about the population. In this case, the population consists of all ice cream cones that have been, will be, or could be dispensed from a particular McDonald's. In this chapter, our hypotheses are about the mean values of populations.

The null hypothesis is the status-quo position, which is the claim that McDonald's makes. An individual cone might weigh a little more than 3.2 ounces, or a little less, but after looking at a great many cones, we would find that McDonald's is right and the mean weight is 3.2 ounces.

We state the null hypothesis as

$$H_0: \mu = 3.2$$

Recall that the null hypothesis always contains an equals sign.

The alternative hypothesis, on the other hand, says that the mean weight is different from 3.2 ounces:

$$H_a: \mu \neq 3.2$$

This is an example of a two-tailed hypothesis. We will reject the null hypothesis if the average of our sample cones is very big (suggesting that the population mean is bigger than 3.2) or very small (suggesting the population mean is less than 3.2). It is also possible to have one-tailed hypotheses, as you will see later in this chapter.



Hypotheses are always statements about population parameters. For the test you are about to learn, this parameter is always μ , the mean of the population.

Details

What Value for α ?

For most situations, using a significance level of 0.05 is a good choice and is recommended by many scientific journals. Values of 0.01 and 0.10 are also commonly used.

Step 2: Prepare

The first step is to set the significance level α (alpha), as we discussed in Chapter 8. The significance level is a performance measure that helps us evaluate the quality of our test procedure. It is the probability of making the mistake of rejecting the null hypothesis when, in fact, the null hypothesis is true. In this case, this is the probability that we will say McDonald's cones do not weigh an average of 3.2 ounces when, in fact, they really do.

The test statistic, called the one-sample t -test, is very similar in structure to the test for one proportion and is based—not surprisingly, given the name of the test—on the t -statistic introduced in Section 9.2. The idea is simple: Compare the observed value of the sample mean, \bar{x} , to the value claimed by the null hypothesis, μ_0 .

Formula 9.2: Test Statistic for the One-Sample t -Test

$$t = \frac{\bar{x} - \mu_0}{SE_{EST}}, \quad \text{where} \quad SE_{EST} = \frac{s}{\sqrt{n}}$$

If conditions hold, the test statistic follows a t -distribution with $df = n - 1$

This test statistic works because it compares the value of the parameter that the null hypothesis says is true, μ_0 , to the estimate of that value that we actually observed in our data. If the estimate is close to the null hypothesis value, then the t -statistic is close to 0. But if the estimate is far from the null hypothesis value, then the t -statistic is far from 0. The farther t is from 0, the worse things look for the null hypothesis.

Anyone can make a decision, but only a statistician can measure the probability that the decision is right or wrong. To do this, we need to know the sampling distribution of our test statistic.

Looking Back

The z-Test

Compare this to the z-test statistic for one proportion in Chapter 8, which has a very similar structure:

$$z = \frac{\hat{p} - p_0}{SE}$$



The sampling distribution will follow the *t*-distribution under these conditions:

Condition 1: *Random Sample and Independence*. The data must be a random sample from a population, and observations must be independent of one another.

Condition 2: *Normal*. The population distribution must be Normal or the sample size must be large. For most situations, 25 is large enough.

Now let's apply this to our ice cream problem. The population for testing the mean ice cream cone weight is somewhat abstract, because a constant stream of ice cream cones is being produced by McDonald's. However, it seems logical that if some cones weigh slightly more than 3.2 ounces and some weigh slightly less, then this distribution of weights should be symmetric and not too different from a Normal distribution. Because our population distribution is Normal, the fact that we have a small sample size, $n = 5$, is not a problem here.

The ice cream cone weights are independent of each other because we were careful, when weighing, to recalibrate the scale, and each cone was obtained on a different day. The cones were not, strictly speaking, randomly sampled, although because the cones were collected on different days and at different times, we will assume that they behave as though they come from a random sample. (But if we're wrong, our conclusions could be *very* wrong!)

Step 3: Compute to compare

The conditions of our data tell us that our test statistic should follow a *t*-distribution with $n - 1$ degrees of freedom. Therefore, we proceed to do the calculations necessary to compare our observed sample mean to the hypothesized value of the population mean, and to measure our surprise.

To find the observed value of our *t*-statistic, we need to find the sample mean and the standard deviation of our sample. These values are given above, but you can easily calculate them from the data.

$$SE_{EST} = \frac{0.3701}{\sqrt{5}} = 0.1655$$

$$t = \frac{3.68 - 3.2}{0.1655} = \frac{0.48}{0.1655} = 2.90$$

The observed sample mean was 2.90—almost 3 standard errors above the value expected by the null hypothesis.

KEY POINT

The *t*-statistic measures how far away (how many standard errors) our observed mean, \bar{x} , lies from the hypothesized value, μ_0 . Values far from 0 tend to discredit the null hypothesis.

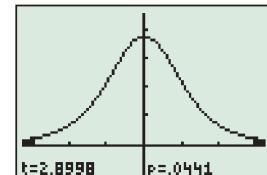
How unusual is such a value, according to the null hypothesis? The *p*-value tells us exactly that—the probability that we would get a *t*-statistic as extreme as or more extreme than what we observed, if in fact the mean is 3.2 ounces.

Because our alternative hypothesis says we should be on the lookout for *t*-statistic values that are much bigger or smaller than 0, we must find the probability in both tails of the *t*-distribution. The *p*-value is shown in the small shaded tails of Figure 9.13. Our sample size is $n = 5$, so our degrees of freedom are $n - 1 = 5 - 1 = 4$.

The *p*-value of 0.044 tells us that if the typical cone really weighs 3.2 ounces, our observations are somewhat unusual. We should be surprised.

Step 4: Interpret

The last step is to compare the *p*-value to the significance level and decide whether to reject the null hypothesis. If we follow a rule that says we will reject whenever the



▲ FIGURE 9.13 The tail areas above 2.90 and below -2.90 are shown as the small shaded areas on both sides. The *p*-value is 0.0441, the probability that if $\mu = 3.2$, a test statistic will be bigger than 2.90 or smaller than -2.90. The distribution shown is a *t*-distribution with 4 degrees of freedom.

Looking Back

p-Values

In Chapter 8 you learned that the p-value is the probability that when the null hypothesis is true, we will get a test statistic as extreme as or more extreme than what we actually saw. (What is meant by “extreme” depends on the alternative hypothesis.) The p-value measures our surprise at the outcome.

p-value is less than or equal to the significance level, then we know that the probability of mistakenly rejecting the null hypothesis will be the value of α .

Our p-value (0.044) is less than the significance level we chose (0.05), so we should reject the null hypothesis and conclude that at this particular McDonald’s, cones do *not* weigh, on average, 3.2 ounces.

This result makes some sense from a public relations standpoint. If the mean were 3.2 ounces, about half of the customers would be getting cones that weighed too little. By setting the mean weight a little higher than what is advertised, McDonald’s can give everyone more than they thought they were getting.

One- and Two-Tailed Alternative Hypotheses

The alternative hypothesis in the ice cream cone test was two-tailed. As you learned in Chapter 8, alternative hypotheses can also be one-tailed. The exact form of the alternative hypothesis depends on the research question. In turn, the form of the alternative hypothesis tells us how to find the p-value.

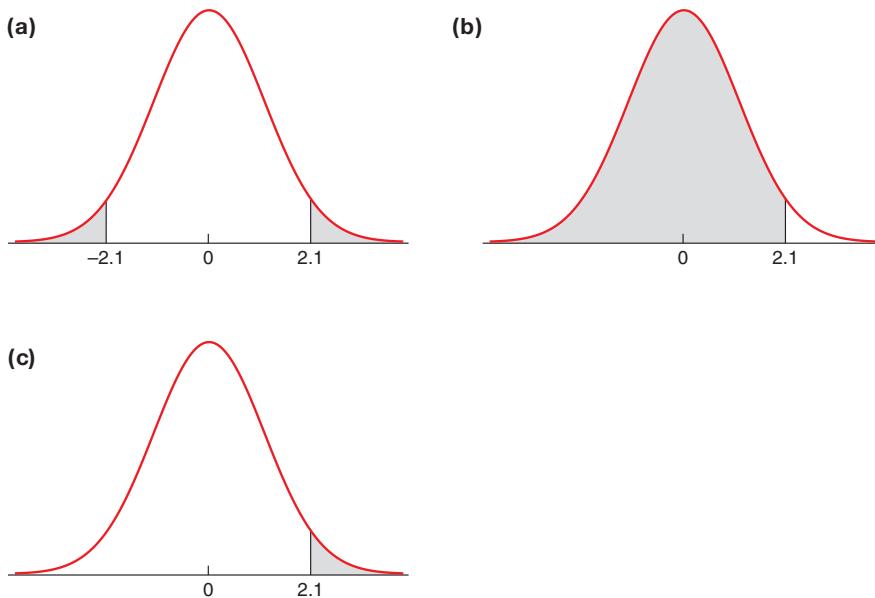
You will always use one of the following three pairs of hypotheses for the one-sample t -test:

Two-Tailed	One-Tailed (Left)	One-Tailed (Right)
$H_0: \mu = \mu_0$ $H_a: \mu \neq \mu_0$	$H_0: \mu = \mu_0$ $H_a: \mu < \mu_0$	$H_0: \mu = \mu_0$ $H_a: \mu > \mu_0$

You choose the pair of hypotheses on the basis of your research question. For the ice cream cone example, we asked if the mean weight is *different* from the value advertised, and so used a two-tailed alternative hypothesis. Had we instead wanted to know whether the mean weight was less than 3.2 ounces, we would have used a one-tailed (left) hypothesis.

Your choice of alternative hypothesis determines how you calculate the p-value. Figure 9.14 shows how to find the p-value for each alternative hypothesis, all using the same t -statistic value of $t = 2.1$ and the same sample size of $n = 30$.

► **FIGURE 9.14** The distributions are t -distributions with $n - 1 = 29$ degrees of freedom. The shaded region in each graph represents a p-value when $t = 2.1$ for (a) a two-tailed alternative hypothesis, (b) a one-tailed (left) hypothesis and (c) a one-tailed (right) hypothesis.



Looking Back

What Does “as extreme as or more extreme than” Mean?

See Chapter 8 for a detailed discussion of how the p-value depends on the alternative hypothesis.

Note that the p-value is always an “extreme” probability; it’s always the probability of the tails (even if the tail is pretty big, as it is in Figure 9.14b).

EXAMPLE 9 Dieting

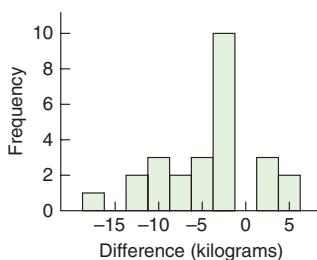
Americans who want to lose weight have many different diets among which to choose. In one study (Dansinger et al. 2005), researchers compared results from four different diets. In this example, though, we look at only a small part of these data to examine whether one of the more popular diets, the Weight Watchers diet, is effective. The researchers examined 40 subjects who were randomly assigned to this diet. Researchers recorded the change in weight after 12 months. The distribution of amount of weight lost in this sample is shown in Figure 9.15. Only 26 of the 40 subjects stayed with the diet for that long, so we have data on only these 26 people.

QUESTION Test the hypothesis that people on the Weight Watchers diet tend to lose weight. Summary statistics are given below. (A negative weight change means the person lost weight.)

$$\bar{x} = -4.6 \text{ kg}, s = 5.4 \text{ kg}$$

(4.6 kilograms is about 10 pounds.)

SOLUTION From Figure 9.15, we see that although a small number of people actually gained weight, the typical experience was a loss in weight. After a year, the average change in weight of the 26 people who stayed on the diet was -4.6 kilograms (about 10 pounds), with a standard deviation of 5.4 kilograms.



◀ FIGURE 9.15 Change in weight for subjects after one year on a low-calorie diet (the Weight Watchers diet). Note that most values are negative, representing people who lost weight.

Our population of interest is the group of all overweight people who might go on the Weight Watchers diet and stick to the diet for one year. Could the mean weight change of this population be negative? If so, then, on average, we could say that people do lose weight on this diet.

Step 1: Hypothesize

Let μ represent the mean weight change of the population.

$$H_0: \mu = 0$$

$$H_a: \mu < 0$$

The null hypothesis says that the mean is 0, because the neutral position here is that no change occurs, on average. This is the same as saying that the diet is ineffective, or not different from no diet at all.

The alternative is that the mean change is negative. This differs from the ice cream cone example, where we only wanted to know whether or not the mean weight was 3.2 ounces. Here we care about the direction of the weight change: Did it go down?

Step 2: Prepare

We will test using a 5% significance level.

We need to check the conditions to see whether the *t*-statistic will follow the *t*-distribution (at least approximately).

Condition 1: Random Sample and Independence. The subjects in this study were not selected randomly from the population of all dieters. But they were selected randomly from a larger group of dieters, because one-fourth of the subjects in this study were randomly assigned to Weight Watchers and the rest to other diets. We will assume that the researchers took care so that observations are independent.

Condition 2: Normal. The distribution of the sample does not look Normal, which makes us suspect that the population distribution is not Normal. But because the sample size is larger than 25, this condition is satisfied.

Step 3: Compute to compare

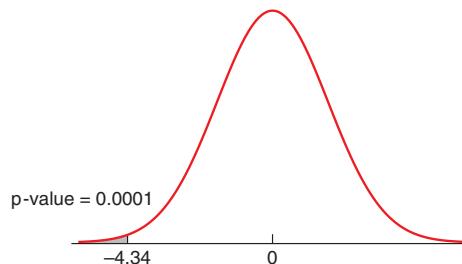
$$SE = \frac{s}{\sqrt{n}} = \frac{5.4}{\sqrt{26}} = 1.0590$$

$$t = \frac{\bar{x} - \mu}{SE} = \frac{-4.6}{1.0590} = -4.34$$

This tells us that our test statistic is 4.34 standard errors below what the null hypothesis expected.

Our p-value here is the area below the observed value, because the alternative cares only whether we get values that are smaller than expected. (Remember, this is a one-tailed hypothesis.) We find the p-value, using technology, to be 0.0001 (Figure 9.16). We use a *t*-distribution with $n - 1 = 25$ degrees of freedom.

► **FIGURE 9.16** The p-value (shaded) for the diet data. The area has been enlarged so that it can be seen.

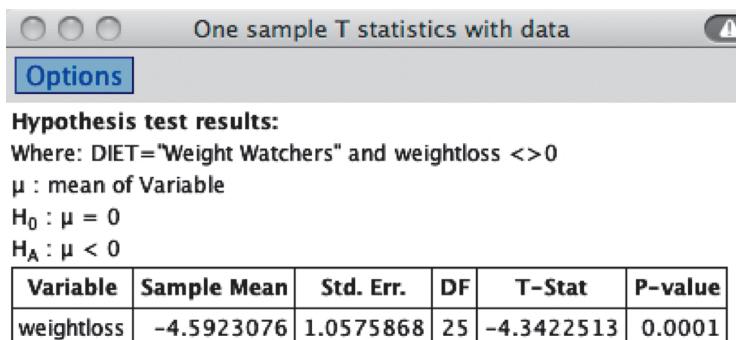
**Step 4: Interpret**

The p-value is much smaller than 0.05, so we conclude that the mean weight change is in fact negative, meaning that people do tend to lose weight after one year on this diet.

TRY THIS! Exercise 9.29

**Tech**

If you have access to the original data, then you should use statistical software to perform the *t*-test. Figure 9.17 shows the results from using StatCrunch to carry out the hypothesis test. We had to tell the computer only the null and alternative hypotheses, and it did the rest. Note that, because of rounding, the values are not precisely the same as those we calculated “by hand” in the example.



◀ FIGURE 9.17 Output from StatCrunch, showing the test that the mean weight loss under Weight Watchers was a negative value (Example 8). The small p-value leads us to conclude that, on average, the dieters really lost weight.



SNAPSHOT ONE SAMPLE t -TEST

WHAT IS IT? ► $t = \frac{\bar{x} - \mu_0}{SE_{EST}}$, where $SE_{EST} = \frac{s}{\sqrt{n}}$

WHAT DOES IT DO? ► It tests hypotheses about a mean of a single population.

HOW DOES IT DO IT? ► If the estimated mean differs from the hypothesized value, then the test statistic will be far from 0. Thus, values of the t -statistic that are unexpectedly far from 0 (in one direction or the other) discredit the null hypothesis.

HOW IS IT USED? ► When proposing values about a single population mean. The observed value of the test statistic is compared to a t -distribution with $n - 1$ degrees of freedom to calculate the p-value. If the p-value is small, you should be surprised by the outcome and reject the null hypothesis.

SECTION 9.4

Comparing Two Population Means

Does your ability to smell depend on whether you are sitting up or lying down? Do gas prices in Denver differ from those in San Francisco? These questions can be answered, in part, by comparing the means of two populations. Although we could construct separate confidence intervals to estimate each mean (for example, the mean smelling ability of people who are lying down compared to the mean smelling ability of people who are sitting), we can construct more precise estimates by focusing on the difference between the two means.

When comparing two populations, it is important to pay attention to whether the data sampled from the populations are two **independent samples** or are, in fact, one sample of related pairs (paired samples). With **paired (dependent) samples**, if you know the value that a subject has in one group, then you know something about the other group, too. In such a case, you have somewhat less information than you might

Caution**Paired (Dependent) vs.
Independent Samples**

One indication that you have paired samples is that each observation in one group is coupled with *one particular observation* in the other group. In this case, the two groups will have the same sample size (assuming no observations are missing).

have if the samples were independent. We begin with some examples to help you see which is which.

Usually, dependence occurs when the objects in your sample are all measured twice (as is common in “before and after” comparisons), or when the objects are related somehow (for example, if you are comparing twins, siblings, or spouses), or when the experimenters have deliberately matched subjects in the groups to have similar characteristics.

EXAMPLE 10 Independent or Dependent Samples?

Here are four descriptions of research studies.

- a. Subjects were tested for their sense of smell twice: once when lying down, once while sitting up. Researchers want to know whether the mean ability to detect smells differs depending on whether one is sitting up or lying down.
- b. Men and women each had their sense of smell measured. Researchers want to know whether, typically, men and women differ in their ability to sense smells.
- c. Researchers randomly assigned overweight people to one of two diets: Weight Watchers and Atkins. Researchers want to know whether the mean weight loss on Weight Watchers was different from that on Atkins.
- d. The numbers of years of education for husbands and wives are compared to see whether the means are different.

QUESTION For each study, state whether it involves two independent samples or paired (that is, dependent) samples.

SOLUTIONS

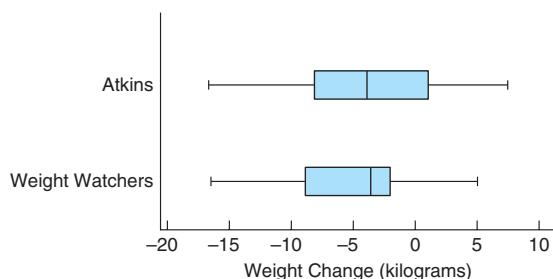
- a. This study has two populations: One population consists of people lying down, the other of people sitting up. However, the samples for the population actually consist of just one group of people. Each person has her or his sense of smell measured twice: once sitting up and once lying down. It seems reasonable to expect that a person who has a very good (or very bad) sense of smell while sitting up might also have a better (or worse) than average sense of smell while lying down. Thus, these samples are *paired* (or *dependent*); knowledge about a value in one sample could give us some information about the other value (because the same people are in both samples).
- b. The two populations consist of men in one and women in the other. As long as the people are not related, knowledge about a measurement of a man could not tell us anything about any of the women. These are *independent* samples.
- c. The two populations are people on the Weight Watchers diet and people on the Atkins diet. We are told that the two samples consist of different people; subjects are randomly assigned to one diet but not to the other. These are *independent* samples.
- d. The populations are matched. Each husband is coupled with one particular wife, so the samples are *paired* (or *dependent*).

TRY THIS! Exercise 9.45

As you shall see, we analyze paired data differently from data that come from two independent samples. Paired data are turned into “difference” scores: We simply subtract one value in each pair from the other. We now have just a single variable, and we can analyze it using the one-sample techniques discussed in Section 9.3. The weight loss data in Example 9 was an example of this. Patients were measured before and after the diet, and these two paired values were subtracted to produce a weight loss measure.

Estimating the Mean Difference with Confidence Intervals (Independent Samples)

The Weight Watchers diet is a very traditional, low-calorie diet. The Atkins diet, on the other hand, limits the amount of carbohydrates. Which diet is more effective? Researchers compared these two diets (as well as two others) by randomly assigning overweight subjects to the two diet groups. The boxplots in Figure 9.18 show summary statistics for the samples' weight losses after one year. The mean weight loss of the Weight Watchers dieters was 4.6 kilograms (about 10 pounds), and the mean loss of the Atkins dieters was 3.9 kilograms (Dansinger et al. 2005).



◀ FIGURE 9.18 Weight change (kilograms) for people randomly assigned to the Atkins diet or the Weight Watchers diet. The medians are close, but because of the skew in the distributions, the sample means are slightly less close.

To guarantee a particular confidence level—for example, 95%—requires that certain conditions hold:

Condition 1: *Random Samples and Independence*. Both samples are randomly taken from their populations, and each observation is independent of any other.

Condition 2: *Independent Samples*. The two samples are independent of each other (not paired).

Condition 3: *Normal*. The populations are approximately Normal, or the sample size in each sample is at least 25. (In special cases, you might need even larger sample sizes.)

If these conditions hold, we can use the following procedure to find an interval with a 95% confidence level.

The formula for a confidence interval comparing two means, when the data are from independent samples, is the same structure as before:

$$(\text{Estimate}) \pm \text{margin of error}$$

which is

$$(\text{Estimate of difference}) \pm t^*(SE_{\text{estimate of difference}})$$

We estimate the difference with

$$(\text{Mean of first sample}) - (\text{mean of second sample})$$

The standard error of this estimator depends on the sample sizes of both samples and also on the standard deviations of both samples:

$$SE_{\text{EST}} = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

We can put these together into a confidence interval:

Formula 9.3: Two-Sample t -Interval

$$(\bar{x}_1 - \bar{x}_2) + t^* \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

The multiplier t^* is based on an approximate t -distribution. If a computer is not available, you can conservatively calculate the degrees of freedom for the t^* multiplier as the smaller of $n_1 - 1$ and $n_2 - 1$, but a computer provides a more accurate value.

Choosing the value of t^* (the critical value of t) by hand to get your desired level of confidence is tricky. For reasons requiring some pretty advanced mathematics to explain, the sampling distribution is not a t -distribution, but only approximately a t -distribution. To make matters worse, to get the approximation to be good requires using a rather complex formula to find the degrees of freedom. If you must do these calculations by hand, we recommend taking a “fast and easy” (but also safe and conservative) approach instead. For t^* , use a t -distribution with degrees of freedom equal to the smaller of $n_1 - 1$ and $n_2 - 1$. That is, use the smaller of the two samples, and subtract 1. For a 95% confidence level, if both samples contain 40 or more observations, you can use 1.96 for the multiplier.

EXAMPLE 11 Comparing Men’s and Women’s Senses of Smell

Researchers studying people’s sense of smell devised a measure of smelling ability (Lundström et al. 2006). If you score high on this scale, you can detect smells better than others. Although it was not a goal of the study, we can use the data collected by these researchers to determine whether women and men differ in their ability to detect smells. The fact that the researchers felt it was important to record the gender of the study participants suggests that there may be some reason to think this sense might vary by gender.

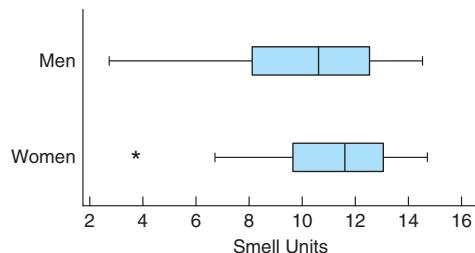
For this example, we compare men and women whose sense of smell was measured while they were lying down. (The subjects’ sense of smell was also measured when they were sitting up.) The summary statistics are

Men: $\bar{x} = 10.0694$, $s = 3.3583$, $n = 18$

Women: $\bar{x} = 11.1250$, $s = 2.7295$, $n = 18$

Boxplots are shown in Figure 9.19.

► **FIGURE 9.19** Distribution of smelling ability for men and women. There is slight skew, and there is one potential outlier (indicated by the asterisk).



QUESTION It looks as if women tend to have the more sensitive sense of smell. But could this difference be due to chance? Find a 95% confidence interval for the mean difference in smelling ability between men and women. Interpret the interval. Assume that the participants in this sample are random samples from the population of all adult men and women.

SOLUTION These data consist of two independent samples: 18 men and 18 women. You might expect that the ability to smell would be Normally distributed across the population, with some people a little better than average and some people a little worse. The boxplots show some skew (and one potential outlier), but with a sample of

size 18, it can be hard to tell whether a distribution is Normal on the basis of the boxplot (or histogram). Thus, even though we are not certain that the Normal condition is fulfilled, we will proceed by assuming that it is. Our assumption is based on some theoretical beliefs about how biological traits such as sense of smell are distributed.

Because the sample sizes of both groups are the same (18), our number of degrees of freedom for t^* is conservatively estimated as the smaller of $18 - 1$ and $18 - 1$, which equals 17. For an approximate 95% confidence interval, we use Table 4 in Appendix A to find $t^* = 2.110$.

Let's call the group of men group 1. (It doesn't matter which we choose for group 1 and which for group 2.)

Estimate of difference: $10.0694 - 11.1250 = -1.0556$

$$m = t^* \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} = t^* \sqrt{\frac{3.3583^2}{18} + \frac{2.7295^2}{18}} = t^* 1.0200$$

$$m = 2.110 \times 1.0200 = 2.1522$$

Therefore, a 95% confidence interval is

$$-1.0556 \pm 2.1522, \text{ or about } (-3.2, 1.1)$$

Because the interval contains zero, we cannot rule out the possibility that the mean difference in the population is 0. This suggests that men and women may not differ in their ability to smell.

TRY THIS! Exercise 9.51



With access to the full data set, and not just to the summary statistics that were provided in Example 11, we can use statistical software to get more accurate calculations. Figure 9.20 shows StatCrunch output for the 95% confidence interval for the difference of the mean smelling ability of men and that of women. The confidence interval is $(-3.1, 1.0)$, which is slightly different (and narrower) than what we found "by hand." The computer-produced interval is more accurate.

Tech

Two sample T statistics with data

Options

95% confidence interval results:

μ_1 : mean of Lying Down where Sex="man"
 μ_2 : mean of Lying Down where Sex="woman"
 $\mu_1 - \mu_2$: mean difference
 (with pooled variances)

Difference	Sample Mean	Std. Err.	DF	L. Limit	U. Limit
$\mu_1 - \mu_2$	-1.0555556	1.0200313	34	-3.1285086	1.0173975

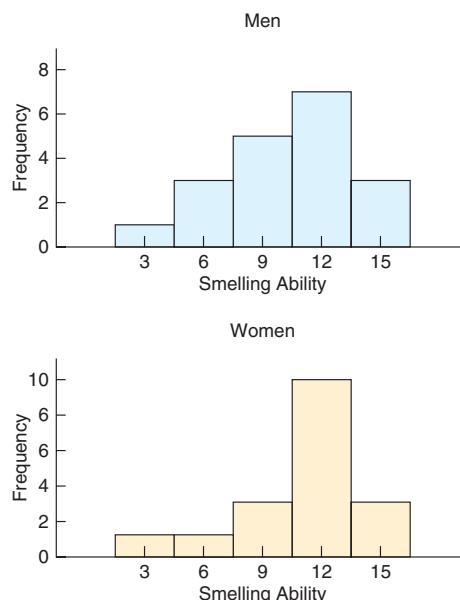
◀ FIGURE 9.20 StatCrunch output for smelling ability confidence interval.

Testing Hypotheses about Mean Differences

Hypothesis tests to compare two means from independent samples follow the same structure we discussed in Chapter 8, although now we have more parameters to compare. We show this structure by revisiting the study to investigate whether men and women differ in their ability to detect smells. In Example 11 you found a confidence interval for the difference in the mean smelling ability of men and women. Here we approach the same data with a hypothesis test.

In Example 11 we used boxplots to investigate the shape of the distribution of smelling ability. Here, we examine histograms (Figure 9.21), which show a more detailed picture of the distributions. Both distributions have roughly the same amount of spread, and the histograms show only a little left skew.

► FIGURE 9.21 Distributions of smelling ability for a sample of 18 men and 18 women.



We call men “population 1” and women “population 2.” Then the symbol μ_1 represents the mean ability to smell for *all* men (while lying down), and μ_2 represents the mean ability to smell for *all* women (while lying down.)

Step 1: Hypothesize

$H_0: \mu_1 = \mu_2$ (men and women have the same means for the sense of smell)

$H_a: \mu_1 \neq \mu_2$ (men and women have different means for the sense of smell)

Step 2: Prepare

The conditions for testing two means are not very different those for from testing one mean and are identical to those for finding confidence intervals of the difference of two means.

Condition 1: *Random Samples and Independent Observations.* Observations are taken at random from two populations, producing two samples. Observations within a sample are independent of one another, which means that knowledge of one value tells us nothing about other observed values in that sample.

Condition 2: *Independent Samples.* The samples are independent of each other. Knowledge about a value in one sample does not tell us anything about any value in the other sample.

Condition 3: *Normal.* Both populations are approximately Normal, or both sample sizes are 25 or more. (In extreme situations, larger sample sizes may be required.)

We might expect the population distributions to be Normally distributed, because measures of biological traits, such as sense of smell, are often Normally distributed.



The sample distributions show a little bit of left skew, but even Normal populations sometimes produce skewed samples when the sample sizes are small, as they are here. The skew is not so great that we would doubt that the populations are Normally distributed, so we assume that condition 2 holds. We must assume that condition 1 (random samples and independence) holds, because we don't know whether the people were randomly sampled. (In fact, they probably were not, because this is rather difficult to do for such studies.) We will assume it is true, understanding that if these people are not representative of the population, we might have substantial bias in our results. It is also safe to assume that condition 3, independent samples, holds, because these are two distinct groups of people. (If researchers had sampled married couples, for example, then this assumption would have been violated.)

Another step in our preparation is to choose a significance level. It is common to use $\alpha = 0.05$, and we will do so for this example.

Step 3: Compute to compare

The test statistic used to test this hypothesis is based on the difference between the sample means. Basically, the test statistic measures how far away the observed difference in sample means is from the hypothesized difference in population means. Yes, you guessed it: The distance is measured in terms of standard errors.

$$t = \frac{\text{(difference in sample means} - \text{what null hypothesis says the difference is)}}{SE_{EST}}$$

Using the test statistic is made easier by the fact that the null hypothesis almost always says that the difference is 0.

Formula 9.4: Two-Sample t -Test

$$t = \frac{\bar{x}_1 - \bar{x}_2 - 0}{SE_{EST}}, \quad \text{where} \quad SE_{EST} = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

If conditions hold, the test statistic follows an approximate t -distribution, where the degrees of freedom are conservatively estimated to be the smaller of $n_1 - 1$ and $n_2 - 1$.

To compare the sense of smell between men and women:

$$\text{Difference in sample means} = \bar{x}_1 - \bar{x}_2 = 10.0694 - 11.125 = -1.0556$$

$$SE_{EST} = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} = \sqrt{\frac{3.3583^2}{18} + \frac{2.7295^2}{18}} = 1.0200$$

$$t = \frac{-1.0556}{1.0200} = -1.03$$

This statistic tells us that the observed difference, -1.0556 , is about one standard error below what the null hypothesis told us to expect.

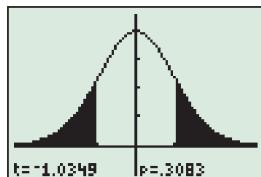
We now measure how surprising this is if the null hypothesis is true. To do this, we need to know the sampling distribution of the test statistic t , because we measure surprise by finding the probability that if the null hypothesis is true, we would see a value as extreme as or more extreme than the value we observed. In other words, we need to find the p-value.

If the conditions listed in the Prepare step hold, then t follows, approximately, a t -distribution with minimum $(n_1 - 1, n_2 - 1)$ degrees of freedom. This approximation can be made even better by adjusting the degrees of freedom, but this adjustment is, for most cases, too complex for a "by hand" calculation. For this reason, we recommend using technology for two-sample hypothesis tests, because you will get more accurate p-values.

Details

Null Hypotheses for Two Means

Mathematically, we can easily adjust our test statistic if the null hypothesis claims that the difference in means is some value other than 0. But in almost all scientific, business, and legal settings, the null hypothesis value will be 0.



▲ FIGURE 9.22 The shaded area represents the p-value for this test: the probability of getting a *t*-statistic more than 1.03 standard errors away from 0 when the null hypothesis is true.

Both sample sizes are 18, so $n_1 = 18$ and $n_2 = 18$. Thus we use $18 - 1 = 17$ for the degrees of freedom.

Our alternative hypothesis is two-tailed and says that the true difference might be much bigger than 0 or much smaller than 0. We therefore find the area under both tails of the *t*-distribution. Figure 9.22 shows this probability as the shaded area under the appropriate *t*-distribution.

The p-value (found with technology) is 0.3083.

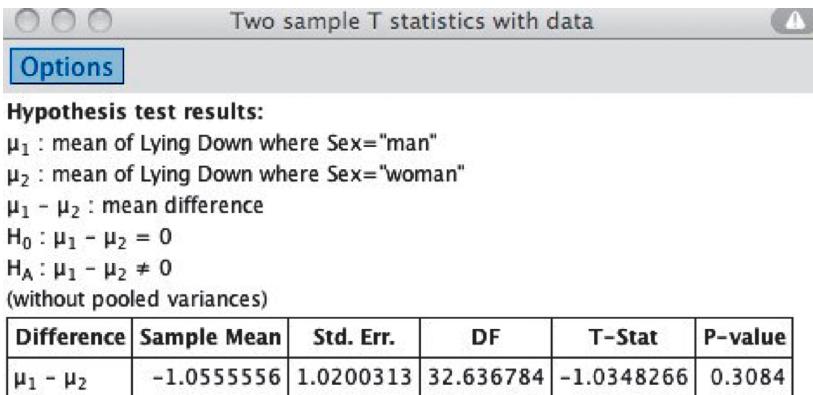
Step 4: Interpret

Again, we compare the p-value to the significance level, α . If the p-value is less than or equal to α , we reject the null hypothesis. In this example, the p-value, 0.308, is bigger than 0.05. If men and women have the same mean smelling ability, then test statistics as extreme as ours occur fairly often. Our test statistic was not a surprise to the null hypothesis—it is pretty much what the null told us we would get. Therefore, we do not reject the null hypothesis.

The previous analysis was done using only the summary statistics provided. If you have the raw data, then you should use computer software to do the analysis. You will get more accurate values and save yourself lots of time. Figure 9.23 shows StatCrunch output for testing whether the mean sense of smell is different for men than for women.

► FIGURE 9.23 StatCrunch output in a test of whether, on average, men's ability to smell is different from women's.

Tech



! Caution

Don't Accept!
Remember from Chapter 8 that we do not "accept" the null hypothesis. It is possible that the sample size is too small (the test has low power) to detect the real difference that exists. Instead, we say that there is not enough evidence for us to reject the null.

! Caution

Don't Pool
When using software to do a two-sample *t*-test, make sure it does the unpooled version. You might have to tell the software explicitly. The unpooled version is more accurate in more situations than the pooled version.

Into the Pool

Some software packages, and some textbooks too, provide for another version of this *t*-test called the "pooled two-sample *t*-test." We have presented the unpooled version (you can see this in the StatCrunch output above the table, where it says "without pooled variances"). The unpooled version is preferred over the other version because the pooled version works only in special circumstances (when the population standard deviations are equal). The unpooled version works reasonably well in all situations, as long as the listed conditions hold.

Test of Two Means: Dependent Samples

With **paired samples**, we turn two samples into one. We do this by finding the difference in each pair.

Recall the study to evaluate smelling ability. Earlier, you saw there were no differences in mean smelling ability between men and women. Are there differences, however, that depend on position? Researchers carried out this study to determine whether people differ in their ability to smell depending on whether they are sitting up or lying down.

**SNAPSHOT TWO SAMPLE *t*-TEST (FROM INDEPENDENT SAMPLES)**

WHAT IS IT? ► A procedure for deciding whether two means, estimated from independent samples, are different. The test statistic used is

$$t = \frac{\bar{x}_1 - \bar{x}_2 - 0}{SE_{EST}}, \text{ where } SE_{EST} = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

WHAT DOES IT DO? ► Provides us with a decision on whether to reject the null hypothesis that the two means are the same and lets us do so knowing the probability that we are making a mistake.

HOW DOES IT DO IT? ► Compares the observed difference in sample means to 0, the value we expect if the population means are equal.

HOW IS IT USED? ► The observed value of the test statistic can be compared to a *t*-distribution.

To test this idea, they measured each subject's sense of smell twice: once while sitting and once while lying down. This is a test of two means, because we are interested in two populations:

Population 1: All people lying down; μ_1 represents the mean ability to smell while lying down.

Population 2: All people sitting up; μ_2 represents the mean ability to smell while sitting up.

However, even though we have two populations, we do not have two *independent* samples. Rather, we have one sample of people who were measured twice. Thus we can change the problem slightly so that, instead of measuring the ability to smell in each position, we measure the *difference* in ability when a person goes from sitting up to lying down.

The first few lines of the data are shown in Table 9.4a.

Subject Number	Sex	Sitting	Lying
1	woman	13.5	13.25
2	woman	13.5	13
3	woman	12.75	11.5
4	man	12.5	12.5

◀ TABLE 9.4a Smelling ability for the first four people sitting and lying.

We create a new variable, call it *difference*, and define it to be the difference between smelling ability sitting up and smelling ability lying down. We show this new variable in Table 9.4b.

Subject Number	Sex	Sitting	Lying	Difference
1	woman	13.5	13.25	0.25
2	woman	13.5	13	0.50
3	woman	12.75	11.5	1.25
4	man	12.5	12.5	0

◀ TABLE 9.4b Difference between smelling ability while sitting up and smelling ability while lying down.

Our hypotheses are now about just one mean: the mean of *difference*.

$$H_0: \mu_{\text{difference}} = 0 \quad (\text{or } \mu_{\text{sitting}} = \mu_{\text{lying}})$$

$$H_a: \mu_{\text{difference}} \neq 0 \quad (\text{or } \mu_{\text{sitting}} \neq \mu_{\text{lying}})$$

Our test statistic is the same as for the one-sample *t*-test:

$$t = \frac{\bar{x}_{\text{difference}} - 0}{SE_{\text{difference}}} \quad \text{where} \quad SE_{\text{difference}} = \frac{s_{\text{difference}}}{\sqrt{n}}$$

We find \bar{x} by averaging the change variable: $\bar{x} = 0.8681$.

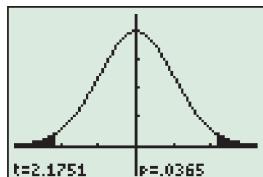
We find $s_{\text{difference}}$ by finding the standard deviation of the difference variable: $s = 2.3946$.

There were 36 participants altogether, so

$$SE_{\text{EST}} = \frac{2.3946}{\sqrt{36}} = 0.3991$$

and then

$$t = \frac{0.8681}{0.3991} = 2.18$$



▲ FIGURE 9.24 A *t*-distribution with $n - 1 = 35$ degrees of freedom. The shaded area represents the p-value for the smell study (sitting vs. lying) and illustrates that if there is no difference in our ability to smell, then our outcome was very unusual and surprising.

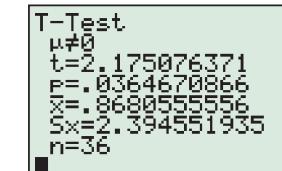
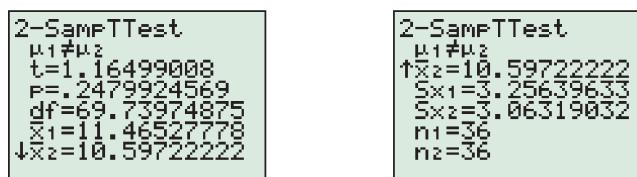
To find the p-value, we use a *t*-distribution (assuming the conditions for a one-sample *t*-test hold) with $n - 1$ degrees of freedom, where n is the number of data pairs. Figure 9.24 shows a *t*-distribution with 35 degrees of freedom. The shaded areas represent the (two-sided) p-value of 0.0365.

Because the p-value is less than 0.05, we reject the null hypothesis. There is evidence that our sense of smell is affected by the position of our body.

Paired *t*-Test vs. Two-Sample *t*-Test

If you have paired data and (incorrectly) do the two-sample *t*-test, you will generally get a p-value that is too big. Figures 9.25 and 9.26 compare the results of doing a two-sample *t*-test on paired data (Figure 9.25) and doing a paired *t*-test on the same data (Figure 9.26.) Note that the test statistic is much larger when you (correctly) use the paired *t*-test to test the paired data; as a result, the p-value is much smaller.

► FIGURE 9.25 TI-83/84 output for two-sample *t*-test.



▲ FIGURE 9.26 TI-83/84 output for paired *t*-test.

The tests produce different values because when we convert the paired data to differences, the resulting differences have a smaller standard deviation than does either sample by itself. This smaller standard deviation leads to a smaller standard error. So even though the numerators of both *t*-statistics (the paired and the two-sample) are the same, the paired *t*-statistic is larger because its denominator is smaller. In Chapter 12 you'll learn more about why studies designed with paired data can be powerful.



SNAPSHOT PAIRED *t*-TEST (DEPENDENT SAMPLES)

WHAT IS IT? ► A procedure for deciding whether two dependent (paired) samples have different means. Each pair is converted to a difference. The test statistic is the same as for the one-sample *t*-test, except that the null-hypothesis value is 0:

$$t = \frac{\bar{x}_{\text{difference}} - 0}{SE_{\text{difference}}}$$

WHAT DOES IT DO? ► Lets us make decisions about whether the means are different, while knowing the probability that we are making a mistake.

HOW DOES IT DO IT? ► The test statistic compares the observed average difference, $\bar{x}_{\text{difference}}$, with the average difference we would expect if the means were the same: 0. Large values discredit the null hypothesis.

HOW IS IT USED? ► If the required conditions hold, the value of the observed test statistic can be compared to a *t*-distribution with $n - 1$ degrees of freedom.

SECTION 9.5

Overview of Analyzing Means

We hope you've been noticing a lot of repetition. The hypothesis test for two means is very similar to the test for one mean, and the hypothesis test for paired data is really a special case of the one-sample *t*-test. Also, the hypothesis tests use almost the same calculations as the confidence intervals, and they impose the same conditions, arranged slightly differently.

All the test statistics (for one proportion, for one mean, for two means, and for two proportions) have this structure:

$$\text{Test statistic} = \frac{(\text{estimated value}) - (\text{null hypothesis value})}{SE}$$

All the confidence intervals have this form:

$$\text{Estimated value} \pm (\text{multiplier}) SE_{\text{EST}}$$

Not all confidence intervals used in statistics have this structure, but most that you will encounter do.

The method for computing a p-value is the same for all tests, although different distributions are used for different situations. The important point is to pay attention to the alternative hypothesis, which tells you whether you are finding a two-tailed or a one-tailed (and *which tail*) p-value.

Confidence Intervals and Hypothesis Tests

In the preceding examples, we reached exactly the same conclusion about men and women and their sense of smell, whether we used a confidence interval or a hypothesis test. This is no coincidence. In fact, it has to be that way. If you have a *two-tailed* alternative hypothesis, then you actually have two choices for how to do the test. Both choices always reach the same conclusion.

Chapter 5:

Randomization and the Basic Factorial Design (BF Design)

Designing Experiments

Every experiment involves a sequence of activities:

1. **Conjecture** – the original hypothesis that motivates the experiment.
2. **Experiment** – the test performed to investigate the conjecture.
3. **Analysis** – the statistical analysis of the data from the experiment.
4. **Conclusion** – what has been learned about the original conjecture from the experiment. Often the experiment will lead to a revised conjecture, and a new experiment, and so forth.

Choosing a Design Structure

Two principles for assigning treatments

- Random assignment
- Blocking

Four design structures based on these principles:

- BF: Basic Factorial
- CB: Complete Block
- LS: Latin Square
- SP/RM: Split Plot/Repeated Measures

Random Assignment

- Randomly assign materials to treatment groups
- This puts all nuisance variation into the 'noise' and lets us fit the ANOVA model
- We assume a 'balanced' design

Diet Effectiveness

- Researchers assigned 116 subjects to four diets:
 - Atkins
 - Ornish
 - Weight Watchers
 - Zone
- How?

Shuffle Some Cards

- Write everyone's name on a card
- Shuffle the cards together many times
- Deal them into four piles

Random Assignment Using R

```
# (No seed here)
>names=1:116

>groups=rep(c("Atkins","Ornish","WW","Zone"),29)

>shuffled=sample(names,116,replace=FALSE)

>mydata=data.frame(Group=groups,Names=shuffled)

>groups[1:20]
[1] "Atkins" "Ornish" "WW" "Zone" "Atkins"
"Ornish" "WW" "Zone" "Atkins" [10] "Ornish" "WW"
"Zone" "Atkins" "Ornish" "WW" "Zone" "Atkins"
"Ornish" [19] "WW" "Zone"
>mydata[1:20,1]
[1] Atkins Ornish WW Zone Atkins Ornish WW Zone
Atkins Ornish WW [12] Zone Atkins Ornish WW Zone
Atkins Ornish WW Zone Levels: Atkins Ornish WW
Zone
```

Pseudo-Random

- Computers use 'pseudo' random numbers determined by an algorithm based on a 'seed' value
- If you know the seed you will always get the same sequence of random numbers
- Set the seed prior to doing a randomization routine
 - Make sure you always get the same results
 - Allows others to go back and reproduce your findings

```
# (with seed)
>set.seed(1234)

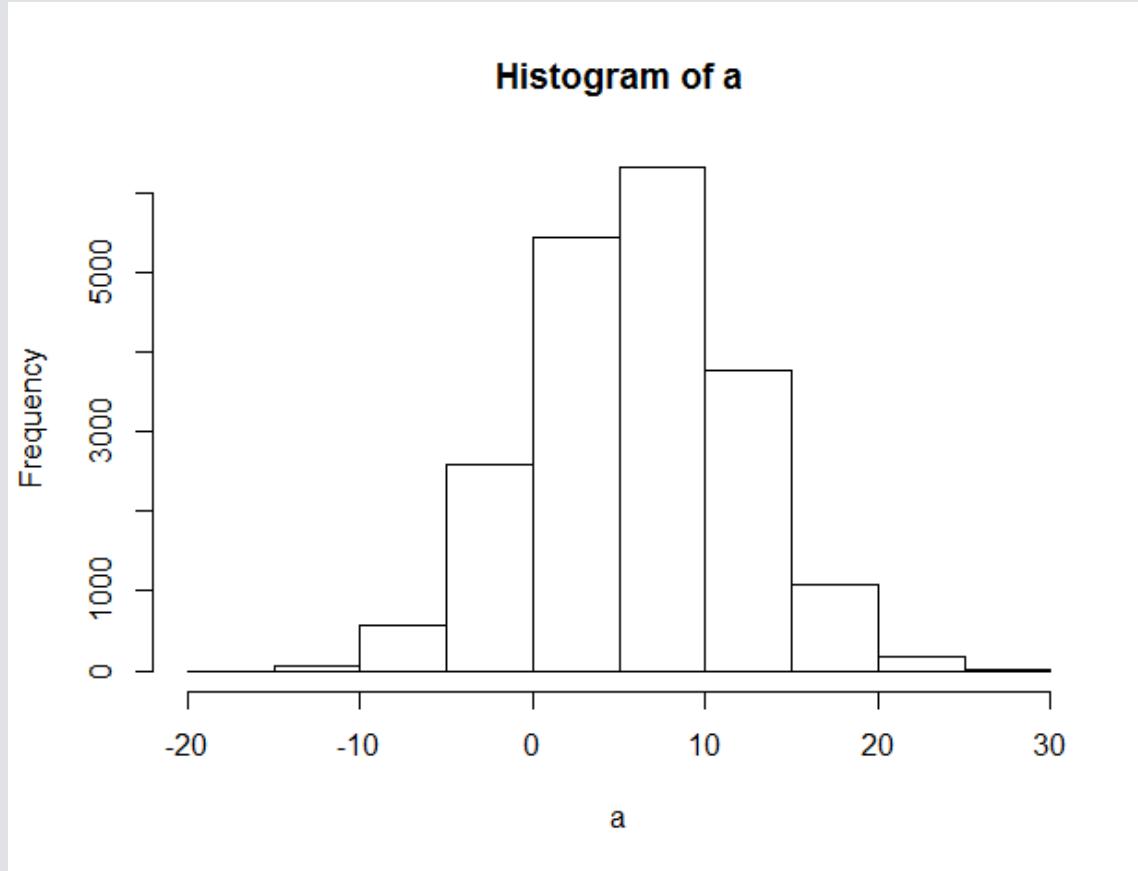
>names=1:116

>groups=rep(c("Atkins","Ornish","WW","Zone"),29)

>shuffled=sample(names,116,replace=FALSE)
>mydata=data.frame(Group=groups,Names=shuffled)
>groups[1:20]
[1] "Atkins" "Ornish" "WW" "Zone" "Atkins" "Ornish"
"WW" "Zone" "Atkins" [10] "Ornish" "WW" "Zone"
"Atkins" "Ornish" "WW" "Zone" "Atkins" "Ornish" [19]
"WW" "Zone"
>mydata[1:20,1] [1] Atkins Ornish WW Zone Atkins
Ornish WW Zone Atkins Ornish WW [12] Zone Atkins
Ornish WW Zone Atkins Ornish WW Zone Levels: Atkins
Ornish WW Zone
```

```
>#using set.seed() before a random process  
  
>set.seed(1234)  
  
>#drawing 4 random digits from a normal distribution  
with mean 3 and sd 2  
  
>rnorm(4,3,2)  
[1] 0.5858685 3.5548585 5.1688824 -1.6913954  
  
>#drawing 4 random digits from a normal distribution  
with mean 6 and sd 6  
  
>rnorm(4,6,6) [1] 8.574748 9.036335 2.551560 2.720209
```

```
> a<-rnorm(20000,6,6)  
>hist(a)
```



```
>summary(a)  
Min. 1st Qu. Median Mean 3rd Qu. Max.  
-18.760 1.908 6.019 5.995 10.040 28.360  
sd(a)  
[1] 5.981181
```

Randomized Basic Factorial

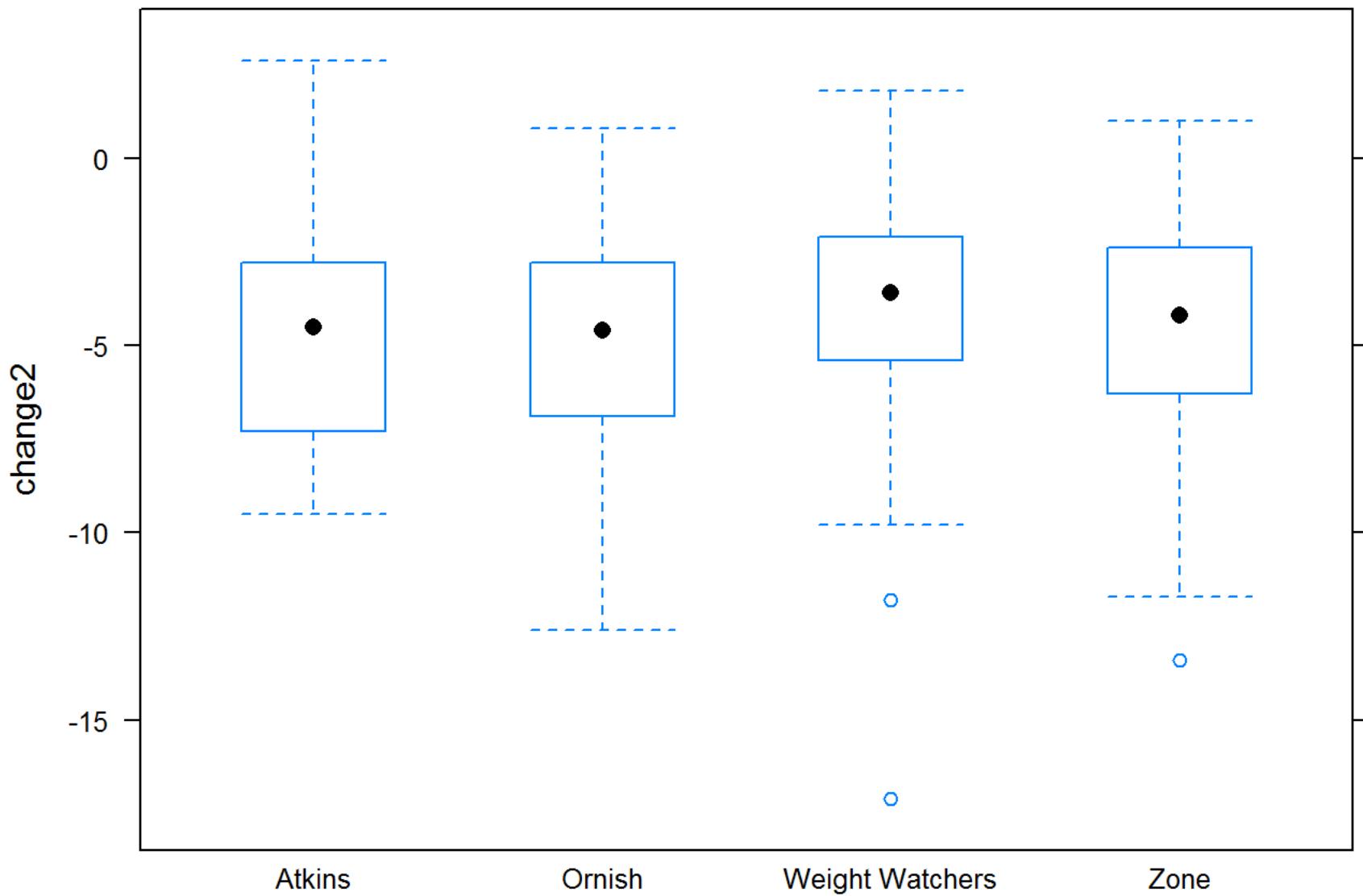
- Use a random method to assign each experimental unit to a treatment
 - Minimizes Bias
 - Ensures that the error follows a chance model and so can be estimated
- Shorthand: RBF

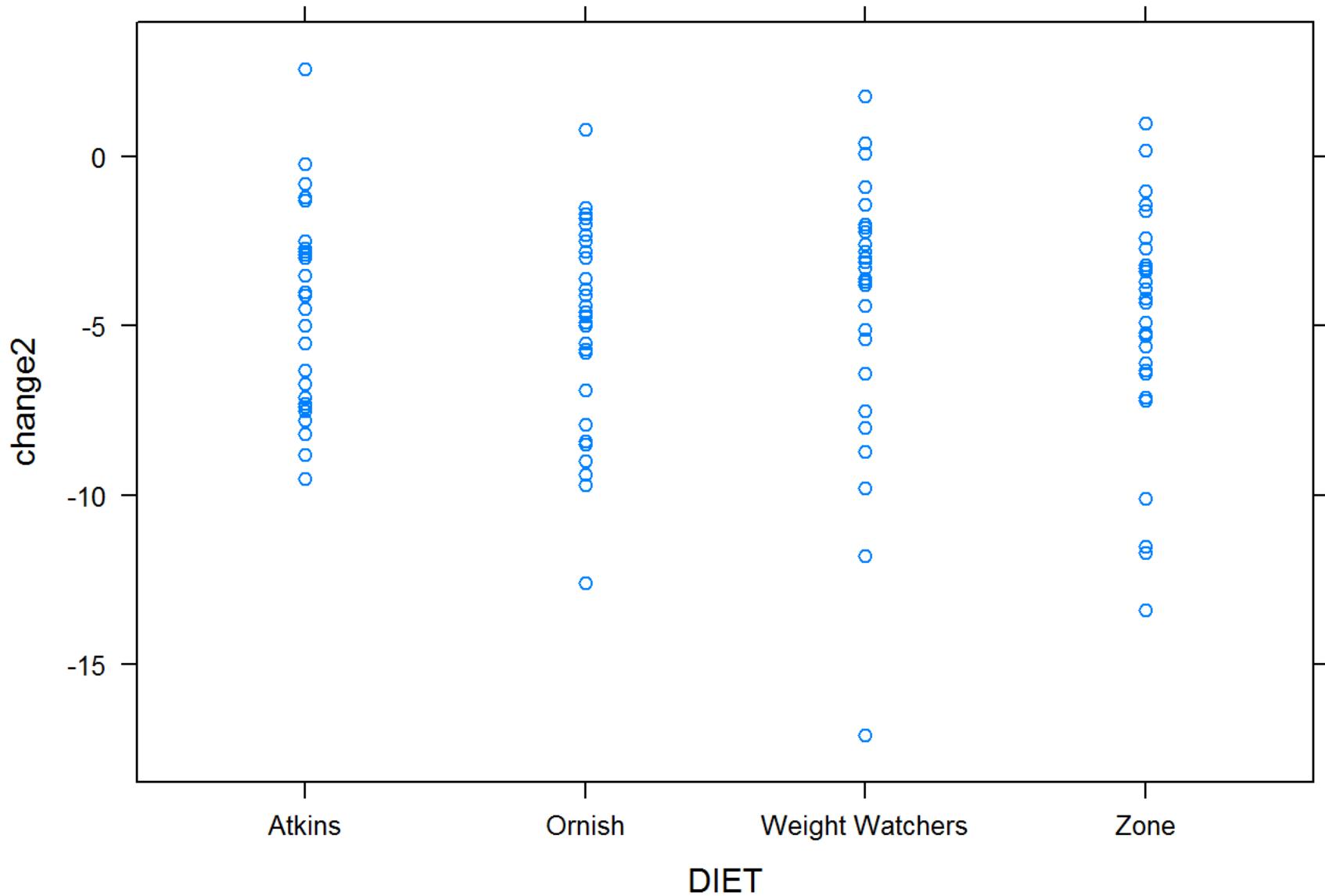
Which diet is most effective?

- Atkins, Ornish, Weight Watchers, Zone
- 116 overweight subjects available
 - Randomly assign subjects to each of the four diets
- Measure weight-loss after 2 months

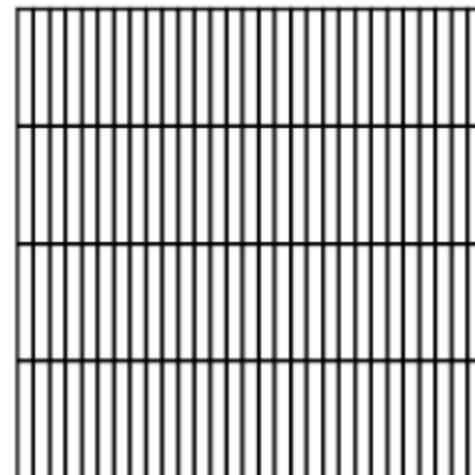
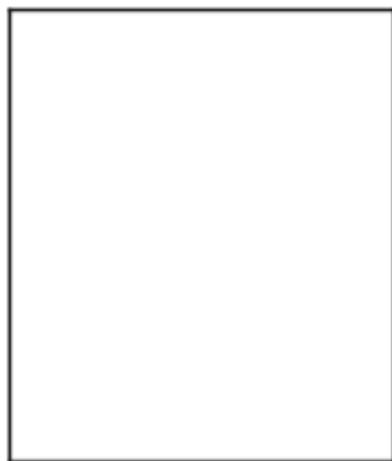
```
# import Diet Data
# diet_balanced <- read.csv("diet_balanced_design.csv")
diet_balanced <- read.csv("~/STAT 101B/Nathan/Week
3/diet_balanced_design.csv")
head(diet_balanced)
## SUBJECT DIET DIETW DIETX DIETY DIETZ AGE SEX WEIGHT_0
BMI_0 ## 1 1 Atkins No No no yes 43 Female 92.3 36.50963 ##
2 2 Atkins No No no yes 23 Male 109.5 37.88927 ## 3 3 Atkins
No No no yes 42 Male 86.5 28.90173 ## 4 4 Atkins No No no
yes 55 Male 118.0 31.67870 ## 5 5 Atkins No No no yes 66
Female 80.2 30.94016 ## 6 6 Atkins No No no yes 37 Female
109.2 40.60083 ## WAIST_0 DROPOUT2 WEIGHT_2 BMI_2 WAIST_2
DROPOUT6 WEIGHT_6 BMI_6 ## 1 123 no 89.8 35.52075 118.0 no
92.0 36.39097 ## 2 119 no 104.0 35.98616 112.0 no 96.2
33.28720 ## 3 103 no 79.2 26.46263 94.5 no 80.4 26.86358 ##
4 110 no 115.0 30.87331 108.5 no 117.4 31.51762 ## 5 103 no
77.5 29.89854 100.5 no 78.0 30.09143 ## 6 113 no 102.5
38.10976 108.0 no 107.3 39.89441
```

Boxplot





Diet Diagrams



One-way 'Effects Model'

$$y_{ij} = \mu + \tau_j + \epsilon_{ij}$$

$$\epsilon_{ij} \sim N(0, \sigma)$$

$$\sum \tau_j = 0$$

y_{ij} is the weight loss of subject i in diet group j

One-way 'Effects Model'

$$y_{ij} = \mu + \tau_j + \epsilon_{ij}$$

- This model says that all of the subjects experience the same 'benchmark' weightloss, μ
- Subjects in diet j share a weightloss, τ_j
- Each subject has his/her own unique weightloss, ϵ_{ij}

Estimating Model Terms

- Estimating effects
 - Inside/Outside Factors
 - Using the computer

Inside/Outside Factors

- One factor is **inside** another if each group of the first factor fits completely inside some group of the second factor

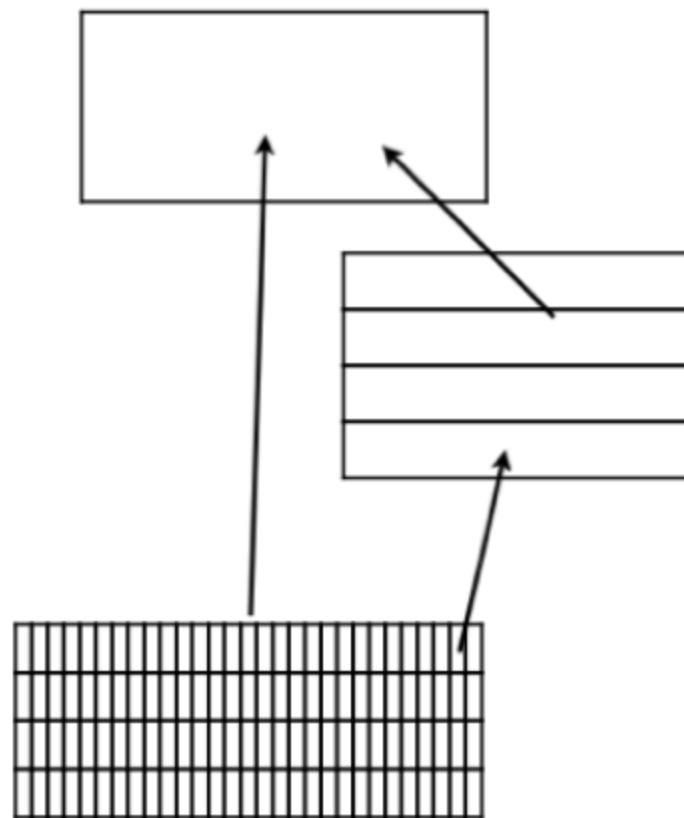
Inside or Outside Benchmark?

- Each of the four diet groups are included in the large benchmark group.
- So diet is **inside** benchmark. Benchmark is **outside** diet.

Residuals?

- There are 29×4 individuals. Each **inside** one of the diet groups
- Every group is **inside** the benchmark group
- So the residual error factor is **inside** the diet and benchmark factors

Inside/Outside diagram



Decomposition

- Break each individual observation into its component parts

Observation = benchmark + structural condition effects + residual error

Estimated Effects

Estimated effect for a factor = Average for the factor – sum of estimated effects for all outside factors

The 'sum of estimated effects of outside factors' is called the **partial fit**

Benchmark

Estimated effect for a factor = Average for the factor – sum of estimated effects for all outside factors

Benchmark does not have any outside factors, so...

benchmark = mean of all observations – 0

benchmark = -4.72

Diet Effects

- One outside factor (benchmark)
 - Atkins = Mean of those in Atkins group – benchmark
 - Ornish = Mean of those in Ornish group – benchmark
 - WW = Mean of those in WW group – benchmark
 - Zone = Mean of those in Zone group – benchmark
- Note: every individual observation within a group has the same estimated effect

```
#Overall Mean
benchmark <- mean(change2)
benchmark

## [1] -4.715517

#Diet Effects
diet_effects <- by(change2,DIET,function(x) mean(x))-benchmark

#OR
atkins <- mean(change2[DIET=='Atkins']) - benchmark
ornish <- mean(change2[DIET=='Ornish']) - benchmark
ww <- mean(change2[DIET=='Weight Watchers']) - benchmark
zone <- mean(change2[DIET=='Zone']) - benchmark

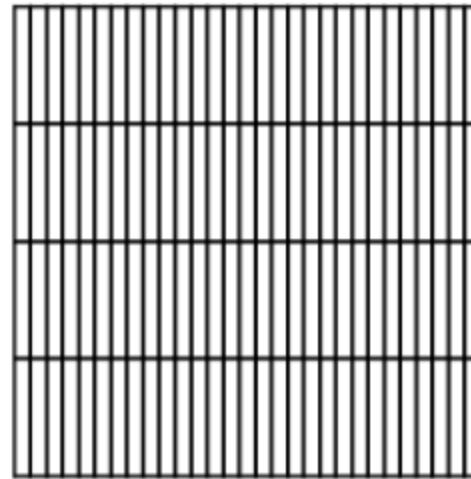
table(atkins,ornish,ww,zone)

## , , ww = 0.325862068965517, zone = -0.0396551724137932
##
##                               ornish
## atkins                  -0.28448275862069
## -0.00172413793103399           1
```

Diet Decomposition

-4.72

-0.00172
-0.284
0.326
-0.0397



Chance Errors

- Two outside factors: benchmark and diet
- Each observation is in its own chance error 'group', the average is just the actual observation
 - ex. observation is the Atkins group has:
 $\text{est. effect} = \text{observation} - \text{benchmark} - \text{Atkins effect}$
 - Note: each of these are unique

```
#Residuals
atkins_R <- change2[DIET=='Atkins'] - benchmark - atkins
ornish_R <- change2[DIET=='Ornish'] - benchmark - ornish
ww_R <- change2[DIET=='Weight Watchers'] - benchmark - ww
zone_R <- change2[DIET=='Zone'] - benchmark - zone

#sum of squares for diet treatment; 3 degrees of freedom
SS_diet <- 29*(atkins^2+ornish^2+ww^2+zone^2)
MS_diet <- SS_diet/3

#sum of squares for residuals; 112 degrees of freedom
SS_residuals <- sum(atkins_R^2+ornish_R^2+ww_R^2+zone_R^2)
MS_residuals <- SS_residuals/112

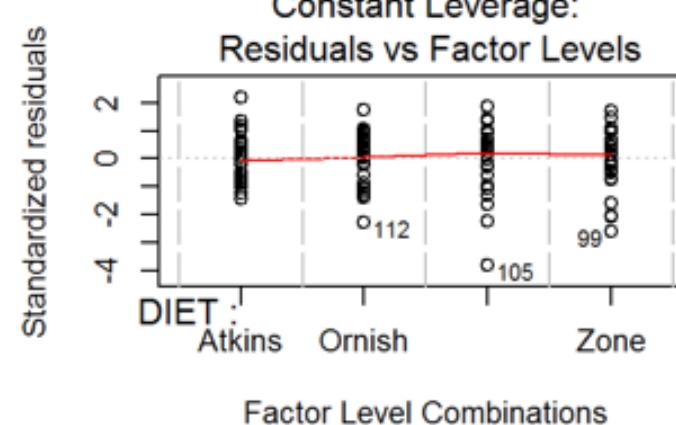
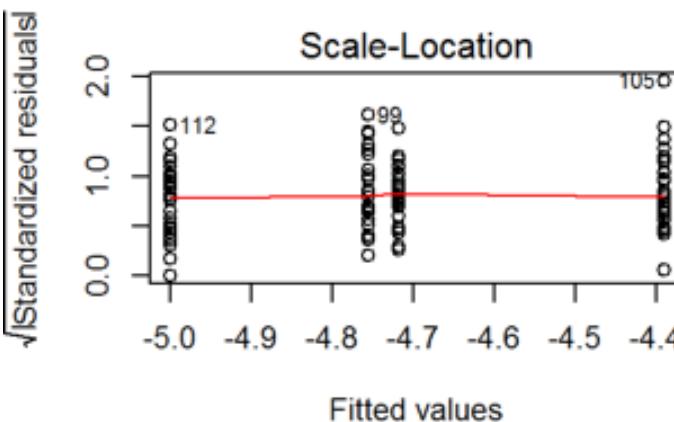
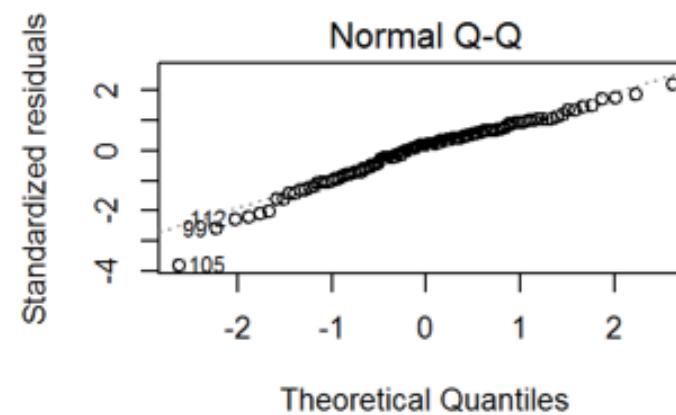
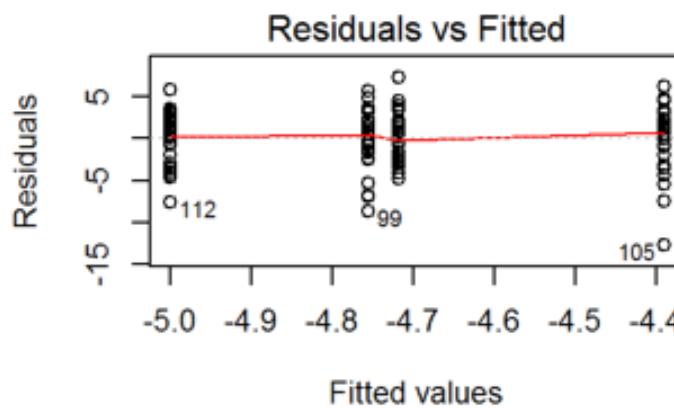
#f-statistic and p-value
F_stat <- MS_diet/MS_residuals
pf(F_stat, 3, 112, lower.tail=FALSE)

## [1] 0.9236083
```

```
#using aov
m1 <- aov(change2~DIET,data=diet_balanced)

par(mfrow=c(2,2))

plot(m1)
```



```
model.tables(m1)
```

```
## Tables of effects
##
## DIET
## DIET
##          Atkins          Ornish Weight Watchers        Zone
##          -0.0017         -0.2845         0.3259       -0.0397
```

```
summary(m1)
```

```
##           Df Sum Sq Mean Sq F value Pr(>F)
## DIET        3    5.5   1.824   0.159  0.924
## Residuals  112 1284.0  11.464
```

```
#checking conditions
```

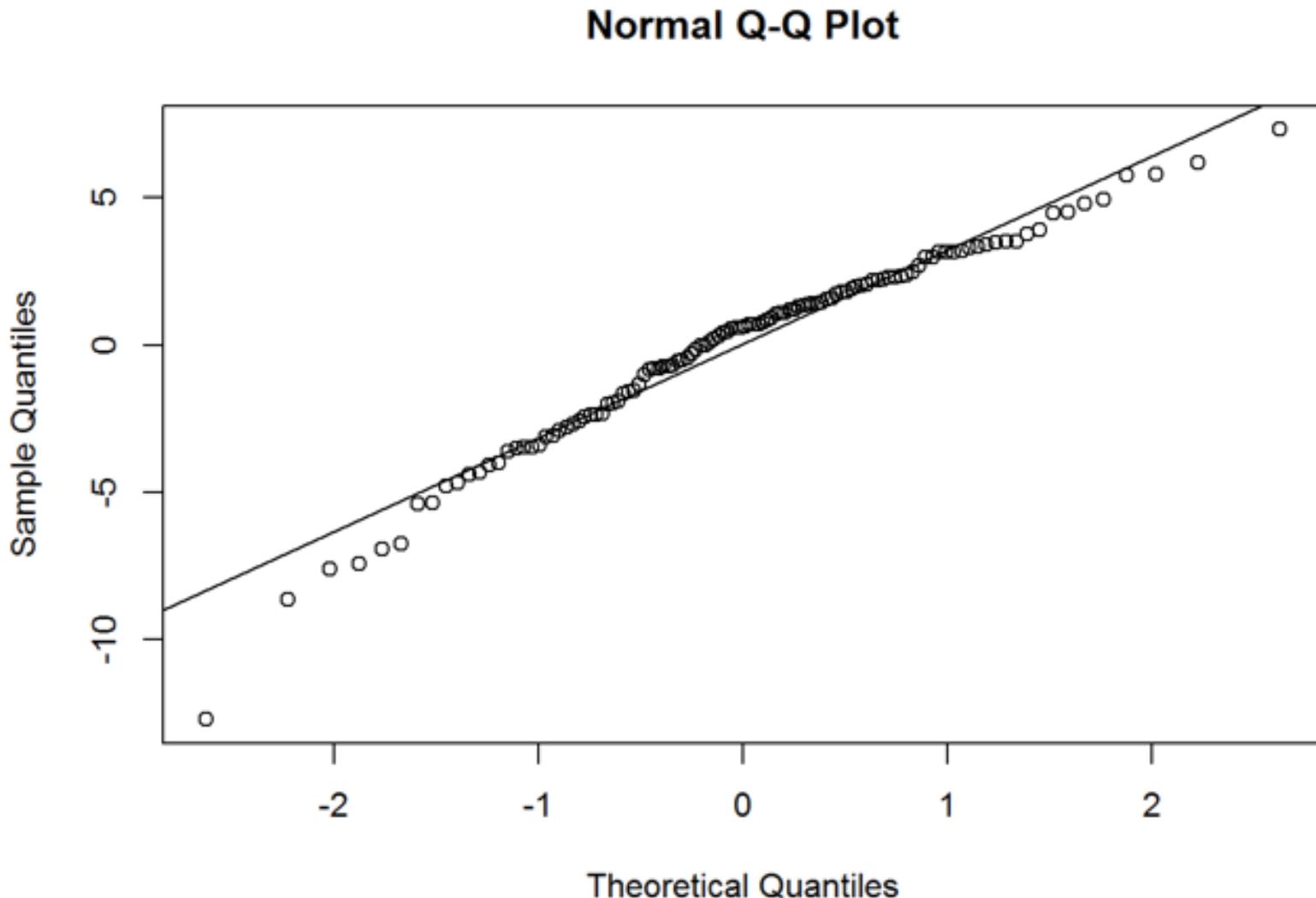
```
par(mfrow=c(1,1))
```

```
qqnorm(m1$residuals)
```

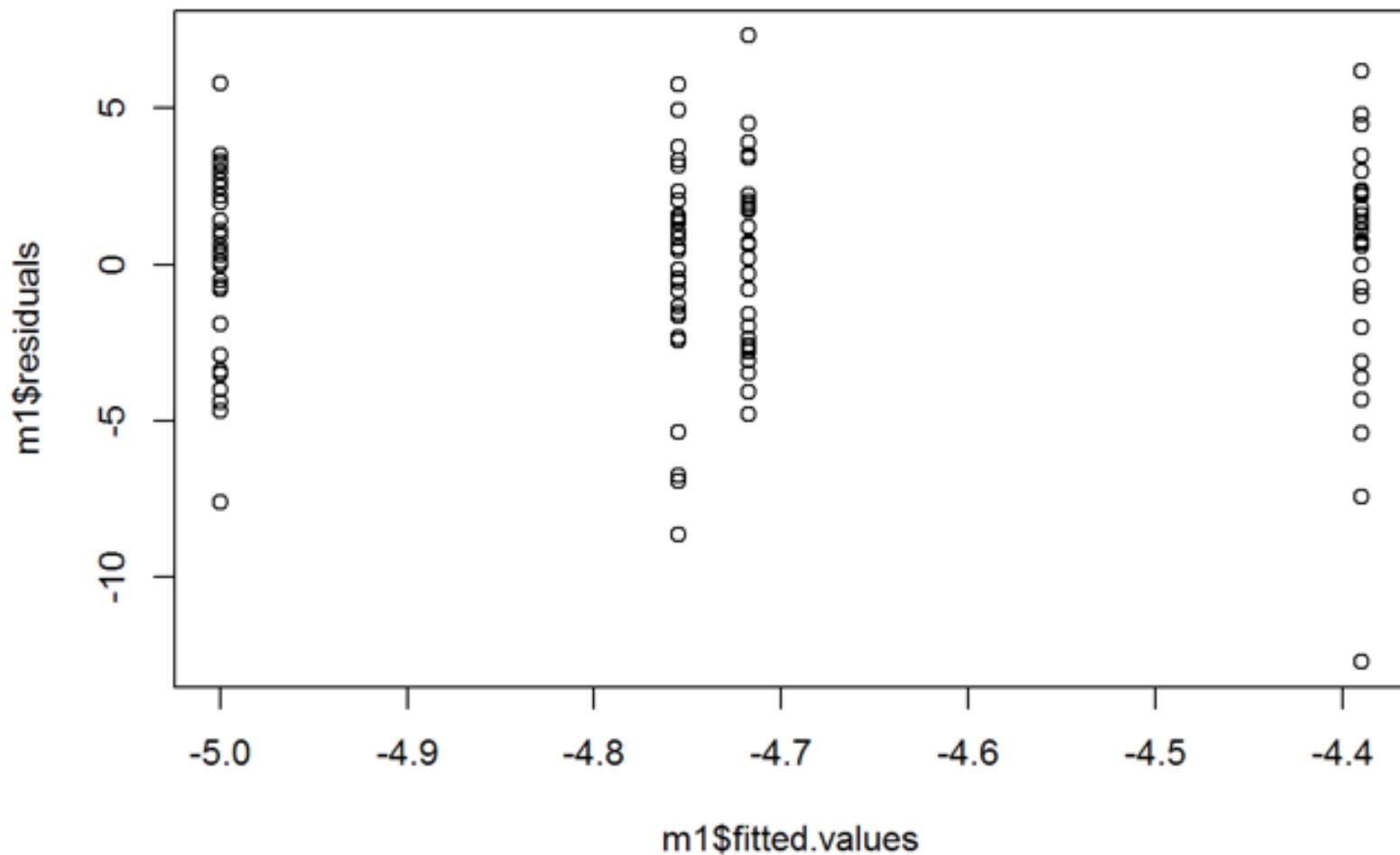
```
qqline(m1$residuals)
```

```
qqnorm(m1$residuals)
```

```
qqline(m1$residuals)
```



```
plot(m1$residuals ~ m1$fitted.values)
```



Hypotheses One-way 'Effects Model'

$$H_0 : \tau_i = 0 \quad \text{for all } i$$

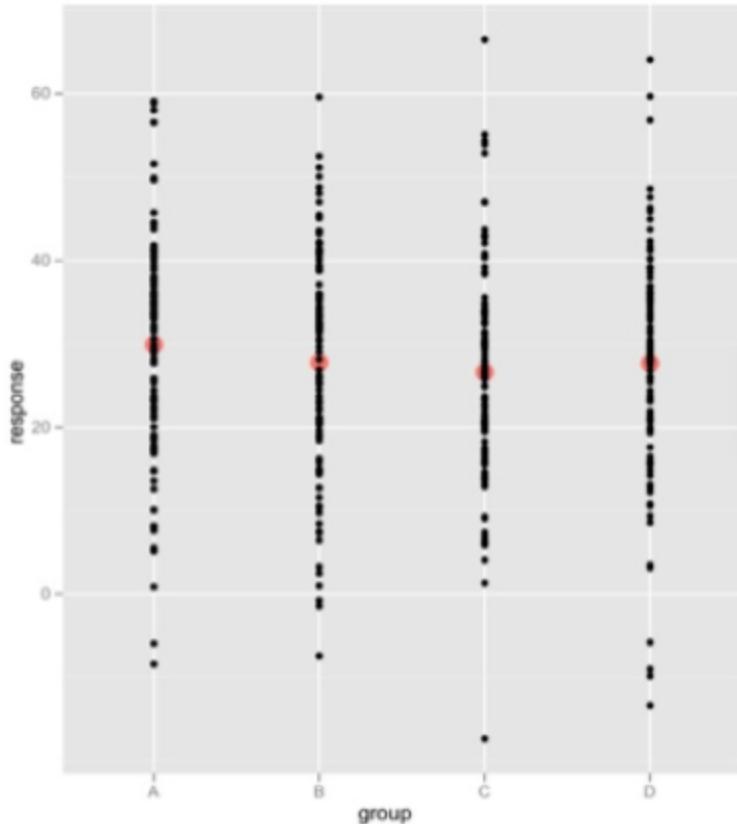
$$H_a : \tau_i \neq 0 \quad \text{for at least one } i$$

- Null: none of the treatments will cause a change
- Alternative: At least one of the treatments will cause a change

How is this done?

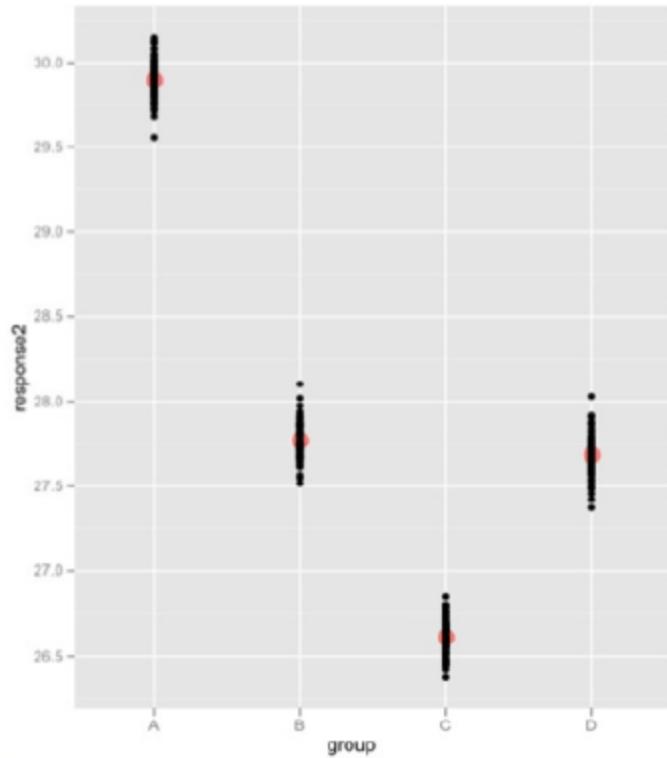
- Compare variability between the groups with the variability in chance errors
- If variability between groups is large relative to chance errors, then this is evidence that observed differences are not due to chance but are 'real'

F-Ratio



Small F-ratio: the variation within groups is bigger than the variation between groups

F-Ratio



Big F-ratio: the variation within groups is smaller than the variation between groups

Measuring Variability

- Measured by the 'sum of squares'
- The total variability in the data is:

$$SS_T = \sum_{i=1}^a \sum_{j=1}^n (y_{ij} - \bar{y}_{..})^2$$

- Can be decomposed into components for each factor
- $SST = SS$ due to treatments + SS due to 'error'

Between Group Variability

- Measured by how much the group means vary about the 'grand' mean

$$SS_{treat} = \sum_{i=1}^a \sum_{j=1}^n (\bar{y}_{i\cdot} - \bar{y}_{..})^2$$

- Sum of Squares due to treatment

Average of the response variable

Within Group Variability

- Measures the variability within each group, added together

$$SS_E = \sum_{i=1}^a \sum_{j=1}^n (\bar{y}_{ij} - \bar{y}_{i\cdot})^2$$

- Sum of squares due to Error

Average of the response variable for subject i and treatment j which is the same as y_{ij} since we have one measurement per subject

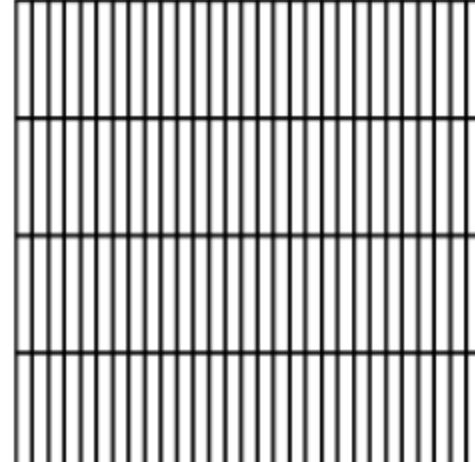
- If the means really differ, then the variability between the group means (SS Treatment) should be greater than the variability within groups (SS Error)
- One difficulty, the more terms we add, the greater the variability becomes.
 - If we add treatment groups, we add variability.
 - If we add subjects we add variability
- So...we take the mean SS by dividing by the 'degrees of freedom'

Degrees of Freedom

- Measures the number of independent observations used to estimate an effect
- Can think of df as the number of bits of information
- The book calls them 'the number of free numbers' in a table

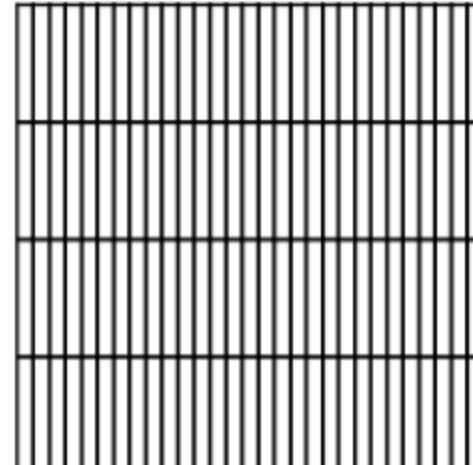
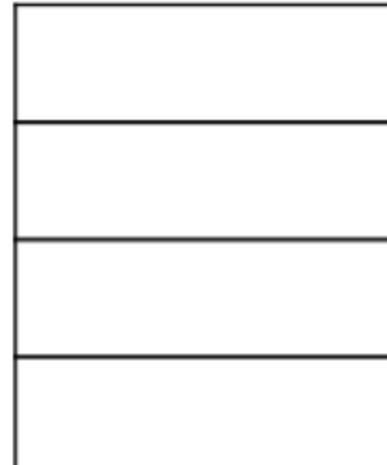
df: benchmark

- df for a factor = number of levels for the factor – sum of df for all outside factors
- There is only one level and no outside factors so, $df=1 - 0 = 1$



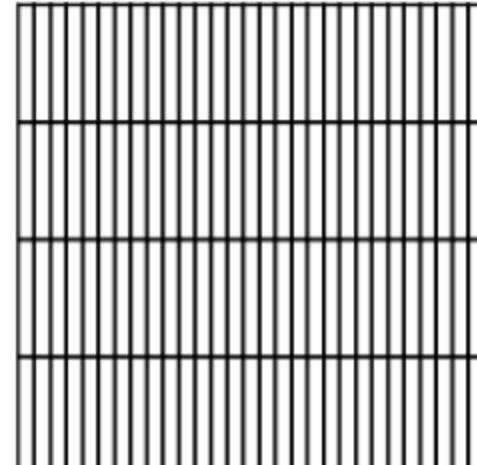
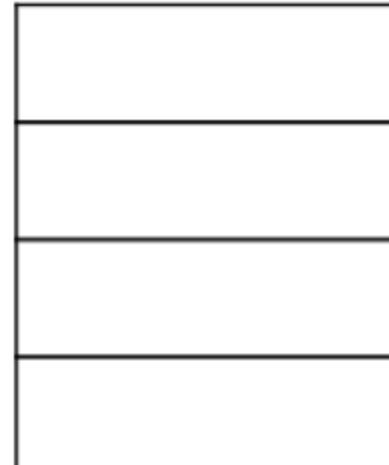
df: diet

- There are four levels and one outside factor using 1 df so, $df=4 - 1 = 3$



df: residuals

- There are 116 'levels' and 2 outside factors whose df total to 4 so, $df = 116 - 4 = 112$



Why is this the case?

- The diet factor has 4 unique estimated effects. A condition was that these sum to one. If you know 3 then you can compute the fourth.
- Hence we have 3 'free' values.

-0.00172
-0.284
0.326
-0.0397

Mean Sum of Squares

- $MS = SS/df$

- $SS(\text{diets}) =$

$$29 \times (\text{atkins}^2 + \text{ornish}^2 + \text{ww}^2 + \text{zone}^2)$$

$$= 5.47$$

- So

$$MS(\text{diets}) = 5.47 / 3 = 1.82$$

Since we have balanced design with 29 subjects in each treatment.

Mean Sum of Squares

- $MS = SS/df$
- $MS(\text{residuals}) =$
 $(\text{atkins.resid}^2 + \text{ornish.resid}^2 + \text{ww.resid}^2 + \text{zone.resid}^2) / 112$
 $= 11.5$

Sampling Distribution of F

- F-statistic = $MSTreat/MSResid$
- F-statistic = $1.82 / 11.5 = 0.158$
- Under assumptions of our model, the F-Statistic follows an F distribution with two different df parameters: (df numerator, df denominator)
- We fail to reject the null when F is small
- P-value = $P(F > F_{\text{obs}}) = P(F > 0.158)$
$$1 - \text{pf}(.158, 3, 112) = 0.924$$

In R...

```
summary(aov(change2~DIET,data=balanced.diet))
      Df Sum Sq Mean Sq F value Pr(>F)
DIET       3      5    1.82    0.16   0.92
Residuals 112 1284    11.46
```

- Conclusion, there is no evidence that diet has an effect on weight loss after two months of dieting

The Completely Randomized Single-Factor Experiment

An Example

A manufacturer of paper used for making grocery bags is interested in improving the tensile strength of the product. Product engineering thinks that tensile strength is a function of the hardwood concentration in the pulp and that the range of hardwood concentrations of practical interest is between 5 and 20%. A team of engineers responsible for the study decides to investigate four levels of hardwood concentration: 5%, 10%, 15%, and 20%. They decide to make up six test specimens at each concentration level, using a pilot plant. All 24 specimens are tested on a laboratory tensile tester, in random order. The data from this experiment are shown in Table 13-1.

The Completely Randomized Single-Factor Experiment

An Example

Table 13-1 Tensile Strength of Paper (psi)

Hardwood Concentration (%)	Observations						Totals	Averages
	1	2	3	4	5	6		
5	7	8	15	11	9	10	60	10.00
10	12	17	13	18	19	15	94	15.67
15	14	18	19	17	16	18	102	17.00
20	19	25	22	23	18	20	127	21.17
							383	15.96

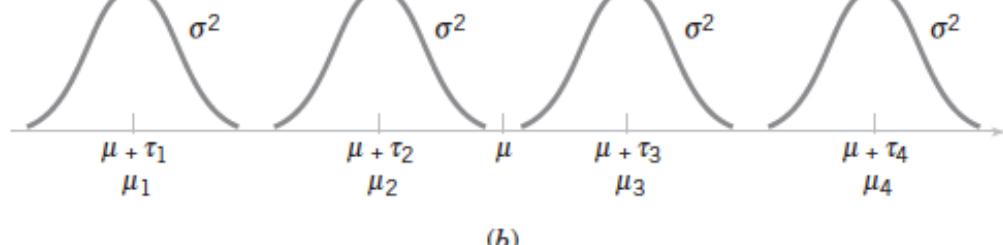
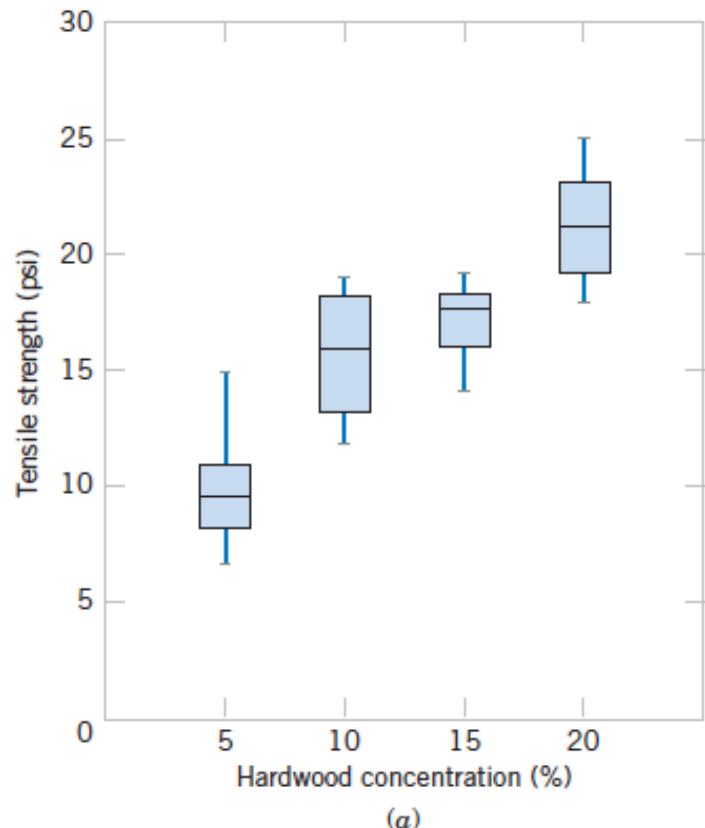
The Completely Randomized Single-Factor Experiment

An Example

- The levels of the factor are sometimes called **treatments**.
- Each treatment has six observations or **replicates**.
- The runs are run in **random** order.

The Completely Randomized Single-Factor Experiment

An Example



(b)

(a) Box plots of hardwood concentration data. (b) Display of the model in Equation 13-1 for the completely randomized single-factor experiment

The Completely Randomized Single-Factor Experiment

The Analysis of Variance

Suppose there are a different levels of a single factor that we wish to compare. The levels are sometimes called **treatments**.

Table 13-2 Typical Data for a Single-Factor Experiment

Treatment	Observations				Totals	Averages
1	y_{11}	y_{12}	\dots	y_{1n}	$y_{1\cdot}$	$\bar{y}_{1\cdot}$
2	y_{21}	y_{22}	\dots	y_{2n}	$y_{2\cdot}$	$\bar{y}_{2\cdot}$
\vdots	\vdots	\vdots	$\vdots \vdots \vdots$	\vdots	\vdots	\vdots
a	y_{a1}	y_{a2}	\dots	y_{an}	$y_{a\cdot}$	$\bar{y}_{a\cdot}$
					$y_{\cdot\cdot}$	$\bar{y}_{\cdot\cdot}$

The Completely Randomized Single-Factor Experiment

The Analysis of Variance

We may describe the observations in Table 13-2 by the linear statistical model:

$$Y_{ij} = \mu + \tau_i + \epsilon_{ij} \begin{cases} i = 1, 2, \dots, a \\ j = 1, 2, \dots, n \end{cases} \quad (13-1)$$

The model could be written as

$$Y_{ij} = \mu_i + \epsilon_{ij} \begin{cases} i = 1, 2, \dots, a \\ j = 1, 2, \dots, n \end{cases}$$

The Completely Randomized Single-Factor Experiment

The Analysis of Variance

Fixed-effects Model

The treatment effects are usually defined as deviations from the overall mean so that:

$$\sum_{i=1}^a \tau_i = 0$$

Also,

$$y_{i\cdot} = \sum_{j=1}^n y_{ij} \quad \bar{y}_{i\cdot} = y_{i\cdot}/n \quad i = 1, 2, \dots, a$$

$$y_{..} = \sum_{i=1}^a \sum_{j=1}^n y_{ij} \quad \bar{y}_{..} = y_{..}/N$$

The Completely Randomized Single-Factor Experiment

The Analysis of Variance

We wish to test the hypotheses:

$$H_0: \tau_1 = \tau_2 = \cdots = \tau_a = 0$$

$$H_1: \tau_i \neq 0 \quad \text{for at least one } i$$

The analysis of variance partitions the total variability into two parts.

The Analysis of Variance

Definition

The sum of squares identity is

$$\sum_{i=1}^a \sum_{j=1}^n (y_{ij} - \bar{y}_{..})^2 = n \sum_{i=1}^a (\bar{y}_{i\cdot} - \bar{y}_{..})^2 + \sum_{i=1}^a \sum_{j=1}^n (y_{ij} - \bar{y}_{i\cdot})^2 \quad (13-5)$$

or symbolically

$$SS_T = SS_{\text{Treatments}} + SS_E \quad (13-6)$$

The Analysis of Variance

The expected value of the treatment sum of squares is

$$E(SS_{\text{Treatments}}) = (a - 1)\sigma^2 + n \sum_{i=1}^a \tau_i^2$$

and the expected value of the error sum of squares is

$$E(SS_E) = a(n - 1)\sigma^2$$

The ratio $MS_{\text{Treatments}} = SS_{\text{Treatments}}/(a - 1)$ is called the **mean square for treatments**.

The Analysis of Variance

The appropriate test statistic is

$$F_0 = \frac{SS_{\text{Treatments}}/(a - 1)}{SS_E/[a(n - 1)]} = \frac{MS_{\text{Treatments}}}{MS_E} \quad (13-7)$$

We would reject H_0 if $f_0 > f_{\alpha, a-1, a(n-1)}$

The Analysis of Variance

Definition

The sums of squares computing formulas for the ANOVA with equal sample sizes in each treatment are

$$SS_T = \sum_{i=1}^a \sum_{j=1}^n y_{ij}^2 - \frac{y_{..}^2}{N} \quad (13-8)$$

and

$$SS_{\text{Treatments}} = \sum_{i=1}^a \frac{y_i^2}{n} - \frac{y_{..}^2}{N} \quad (13-9)$$

The error sum of squares is obtained by subtraction as

$$SS_E = SS_T - SS_{\text{Treatments}} \quad (13-10)$$

The Completely Randomized Single-Factor Experiment

The Analysis of Variance

Analysis of Variance Table

Table 13-3 The Analysis of Variance for a Single-Factor Experiment, Fixed-Effects Model

Source of Variation	Sum of Squares	Degrees of Freedom	Mean Square	F_0
Treatments	$SS_{\text{Treatments}}$	$a - 1$	$MS_{\text{Treatments}}$	$\frac{MS_{\text{Treatments}}}{MS_E}$
Error	SS_E	$a(n - 1)$	MS_E	
Total	SS_T	$an - 1$		

Example

EXAMPLE 13-1 Tensile Strength ANOVA

Consider the paper tensile strength experiment described in Section 13-2.1. This experiment is a CRD. We can use the analysis of variance to test the hypothesis that different hardwood concentrations do not affect the mean tensile strength of the paper.

The hypotheses are

$$H_0: \tau_1 = \tau_2 = \tau_3 = \tau_4 = 0$$

$$H_1: \tau_i \neq 0 \text{ for at least one } i$$

The Completely Randomized Single-Factor Experiment

Example

We will use $\alpha = 0.01$. The sums of squares for the analysis of variance are computed from Equations 13-8, 13-9, and 13-10 as follows:

$$\begin{aligned} SS_T &= \sum_{i=1}^4 \sum_{j=1}^6 y_{ij}^2 - \frac{y_{..}^2}{N} \\ &= (7)^2 + (8)^2 + \dots + (20)^2 - \frac{(383)^2}{24} = 512.96 \end{aligned}$$

$$\begin{aligned} SS_{\text{Treatments}} &= \sum_{i=1}^4 \frac{y_{i..}^2}{n} - \frac{y_{..}^2}{N} \\ &= \frac{(60)^2 + (94)^2 + (102)^2 + (127)^2}{6} - \frac{(383)^2}{24} \\ &= 382.79 \end{aligned}$$

$$\begin{aligned} SS_E &= SS_T - SS_{\text{Treatments}} \\ &= 512.96 - 382.79 = 130.17 \end{aligned}$$

The Completely Randomized Single-Factor Experiment

Example

The ANOVA is summarized in Table 13-4. Since $f_{0.01,3,20} = 4.94$, we reject H_0 and conclude that hardwood concentration in the pulp significantly affects the mean strength of the paper. We can also find a P -value for this test statistic as follows:

$$P = P(F_{3,20} > 19.60) \approx 3.59 \times 10^{-6}$$

Since $P \approx 3.59 \times 10^{-6}$ is considerably smaller than $\alpha = 0.01$, we have strong evidence to conclude that H_0 is not true.

Table 13-4 ANOVA for the Tensile Strength Data

Source of Variation	Sum of Squares	Degrees of Freedom	Mean Square	f_0	P -value
Hardwood concentration	382.79	3	127.60	19.60	3.59 E-6
Error	130.17	20	6.51		
Total	512.96	23			

Table 13-5 Minitab Analysis of Variance Output for Example 13-1

One-Way ANOVA: Strength versus CONC

Analysis of Variance for Strength

Source	DF	SS	MS	F	P
Conc	3	382.79	127.60	19.61	0.000
Error	20	130.17	6.51		
Total	23	512.96			
Individual 95% CIs For Mean Based on Pooled StDev					
Level	N	Mean	StDev	— + — + — + — + —	
5	6	10.000	2.828	(—*—)	
10	6	15.667	2.805	(—*—)	
15	6	17.000	1.789	(—*—)	
20	6	21.167	2.639	(—*—)	
— + — + — + — + —					
Pooled StDev = 2.551				10.0 15.0 20.0 25.0	

Fisher's pairwise comparisons

Family error rate = 0.192

Individual error rate = 0.0500

Critical value = 2.086

Intervals for (column level mean) – (row level mean)

	5	10	15
10	-8.739		
	-2.594		
15	-10.072	-4.406	
	-3.928	1.739	
20	-14.239	-8.572	-7.239
	-8.094	-2.428	-1.094

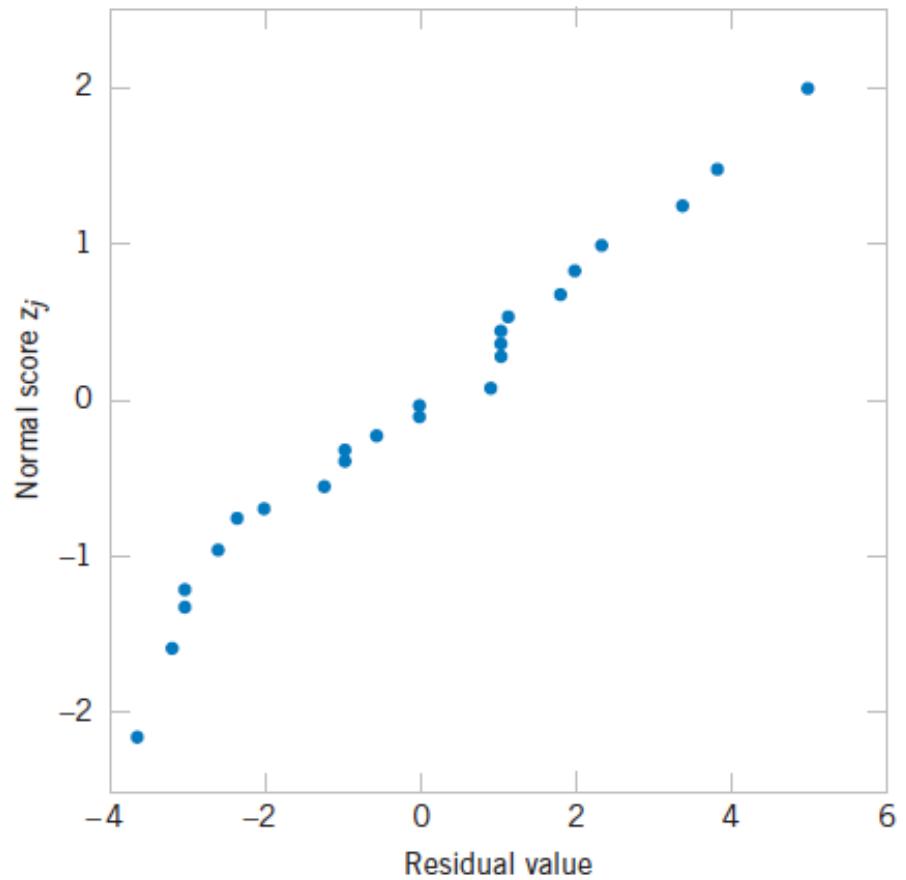
Residual Analysis and Model Checking

Table 13-6 Residuals for the Tensile Strength Experiment

Hardwood Concentration (%)	Residuals					
5	-3.00	-2.00	5.00	1.00	-1.00	0.00
10	-3.67	1.33	-2.67	2.33	3.33	-0.67
15	-3.00	1.00	2.00	0.00	-1.00	1.00
20	-2.17	3.83	0.83	1.83	-3.17	-1.17

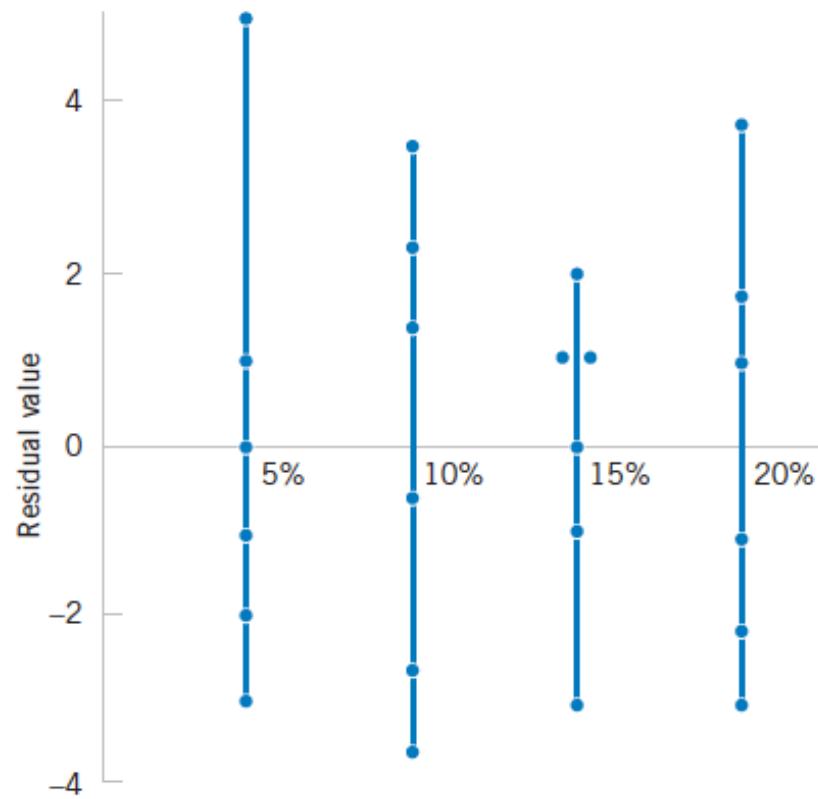
Residual Analysis and Model Checking

Normal probability plot of residuals from the hardwood concentration experiment.



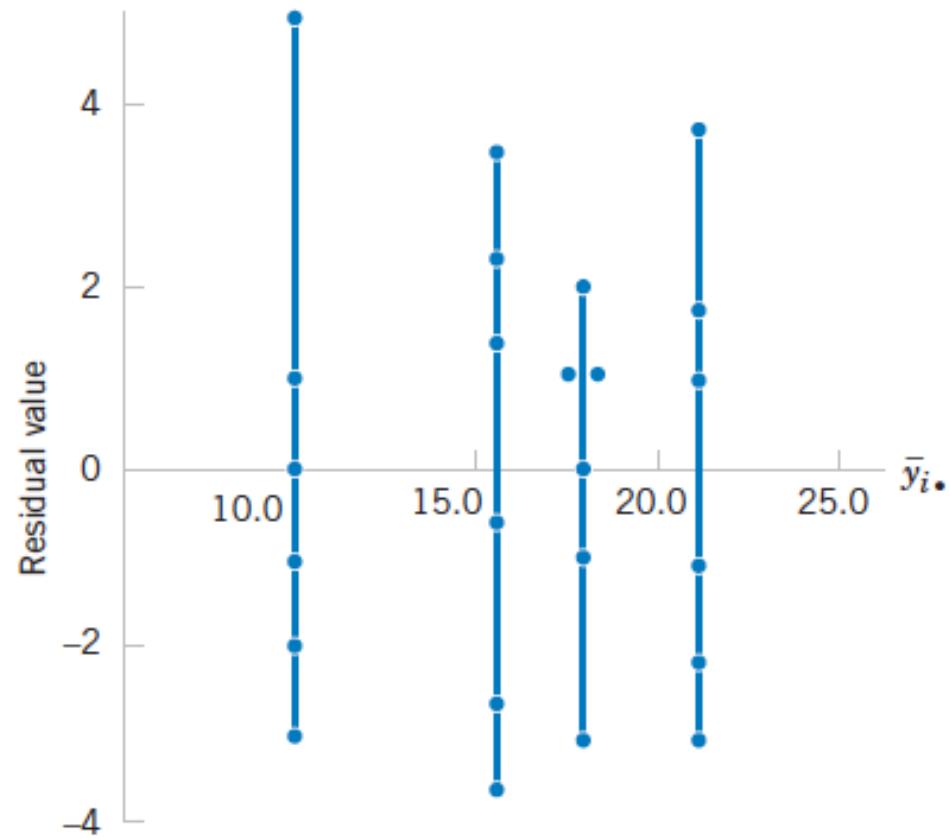
Residual Analysis and Model Checking

Plot of residuals versus factor levels
(hardwood concentration).



Residual Analysis and Model Checking

Plot of residuals versus \bar{y}_i



Required Conditions

- Objects/subjects randomly assigned to treatment levels.
- Independent observations
- Errors are normally distributed in the population
- Amount of variation is same for each treatment level

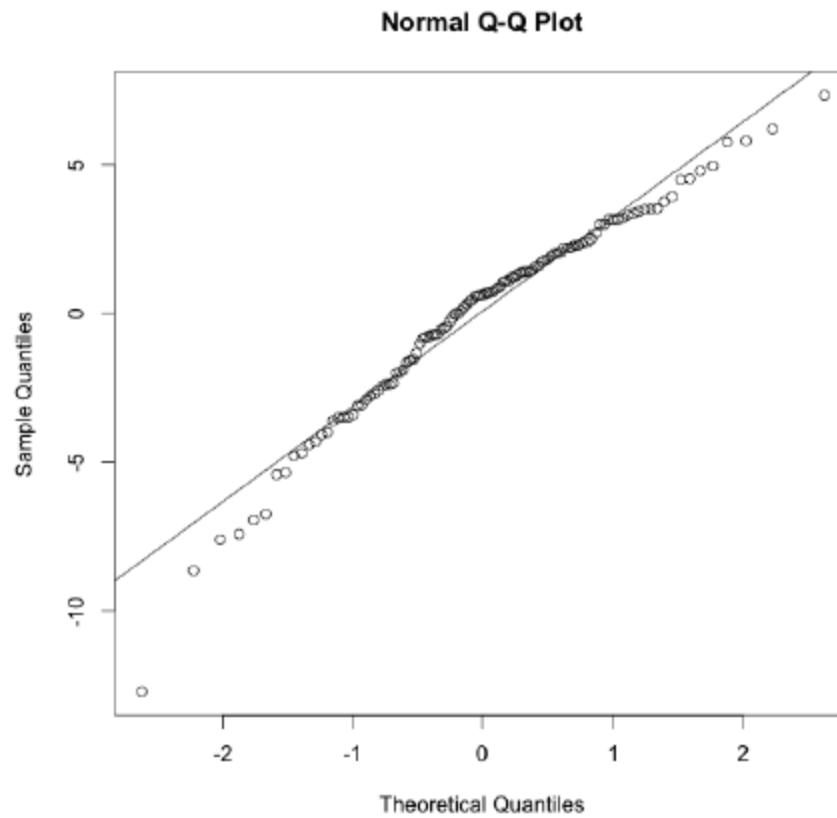
Checking Conditions

- Some of these conditions can be checked by examining the residuals.
- `aov()` will calculate these automatically
 - `fitted.values` contains the predicted values.
 - `residuals` contains the residuals

Diagnostic: Normality Conditions

- `qqnorm()`
- `qqline()`
- This plots the residuals against their expected values if their distribution were truly normal
- Thus if they are truly from a Normal population this plot will be roughly a straight line

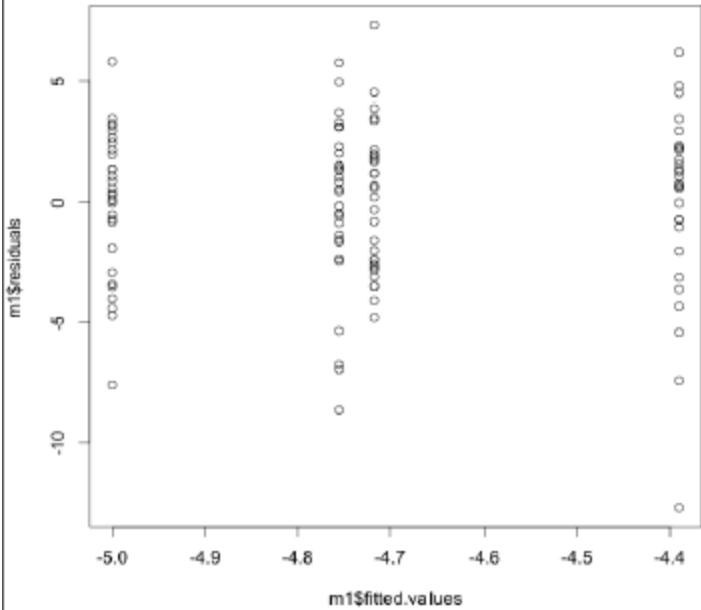
qq plots



Model Specification

- If the shape of the model correctly describes the signal, then the residuals should have no structure

$$\epsilon_{ij} \sim N(0, \sigma)$$



- In each group the residuals should be centered at about 0
- In each group, standard deviation should be about the same

Non-constant Variance

- The major problem occurs when the variance/SD increases steadily with the fitted (decreasing also a problem)
- For balanced designs (equal sample sizes in each treatment level) non-constant variance not a big problem

Other Conditions

- If you know the order in which the data were recorded, plotting the residuals against the order can tell you if there is time-dependence within observation
- This is a serious violation of the independence condition and can invalidate the results of the study
- Ideally, then, your plot will show no trend or structure

You also need to check for leverage points and outliers

Observational version of BF design

Sampling instead of assigning

Example 5.2:

Intravenous Fluids:

Conditions: Three Drug Companies

Material: Six different samples of intravenous fluid from each drug company

Response: Number of contaminated particles of a certain size

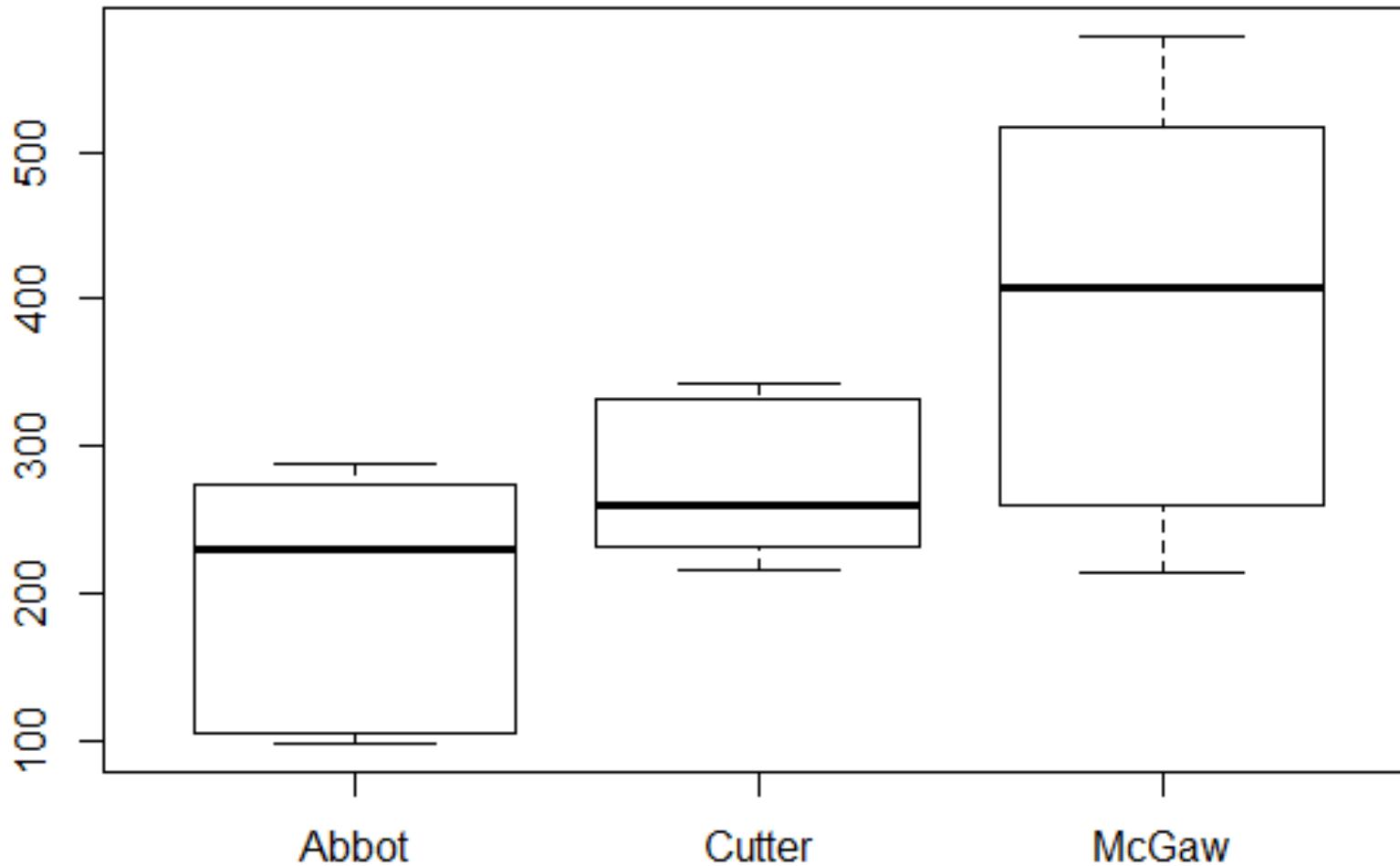
Chapter 5 Section 2 Example:

```
> ch5sec2
```

	DrugCompany	Particles
1	Cutter	255
2	Cutter	265
3	Cutter	343
4	Cutter	332
5	Cutter	232
6	Cutter	217
7	Abbot	106
8	Abbot	289
9	Abbot	99
10	Abbot	275
11	Abbot	221
12	Abbot	240
13	McGaw	578
14	McGaw	516
15	McGaw	214
16	McGaw	413
17	McGaw	401
18	McGaw	260

```
> |
```

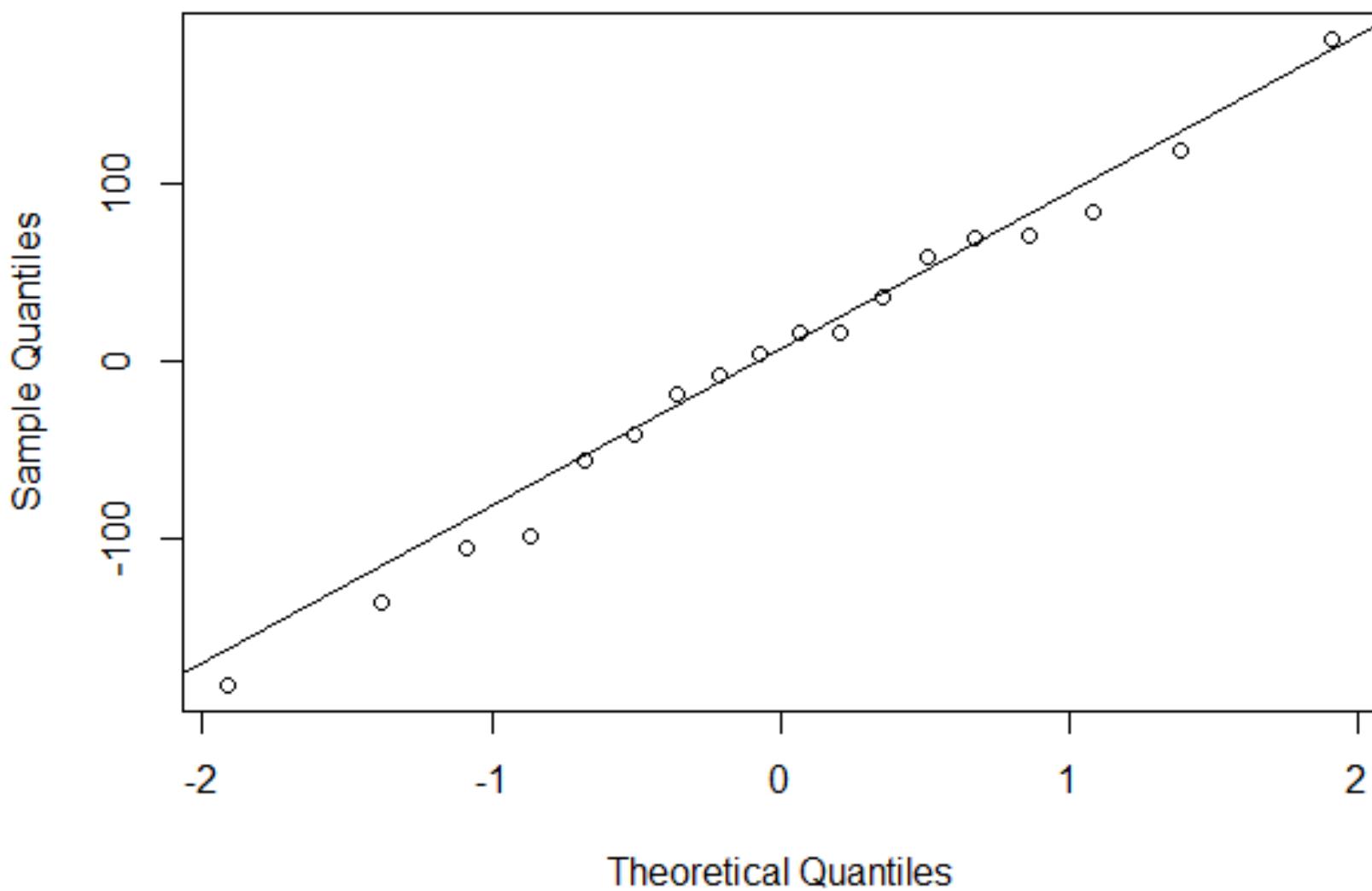
Boxplot of Drug Companies

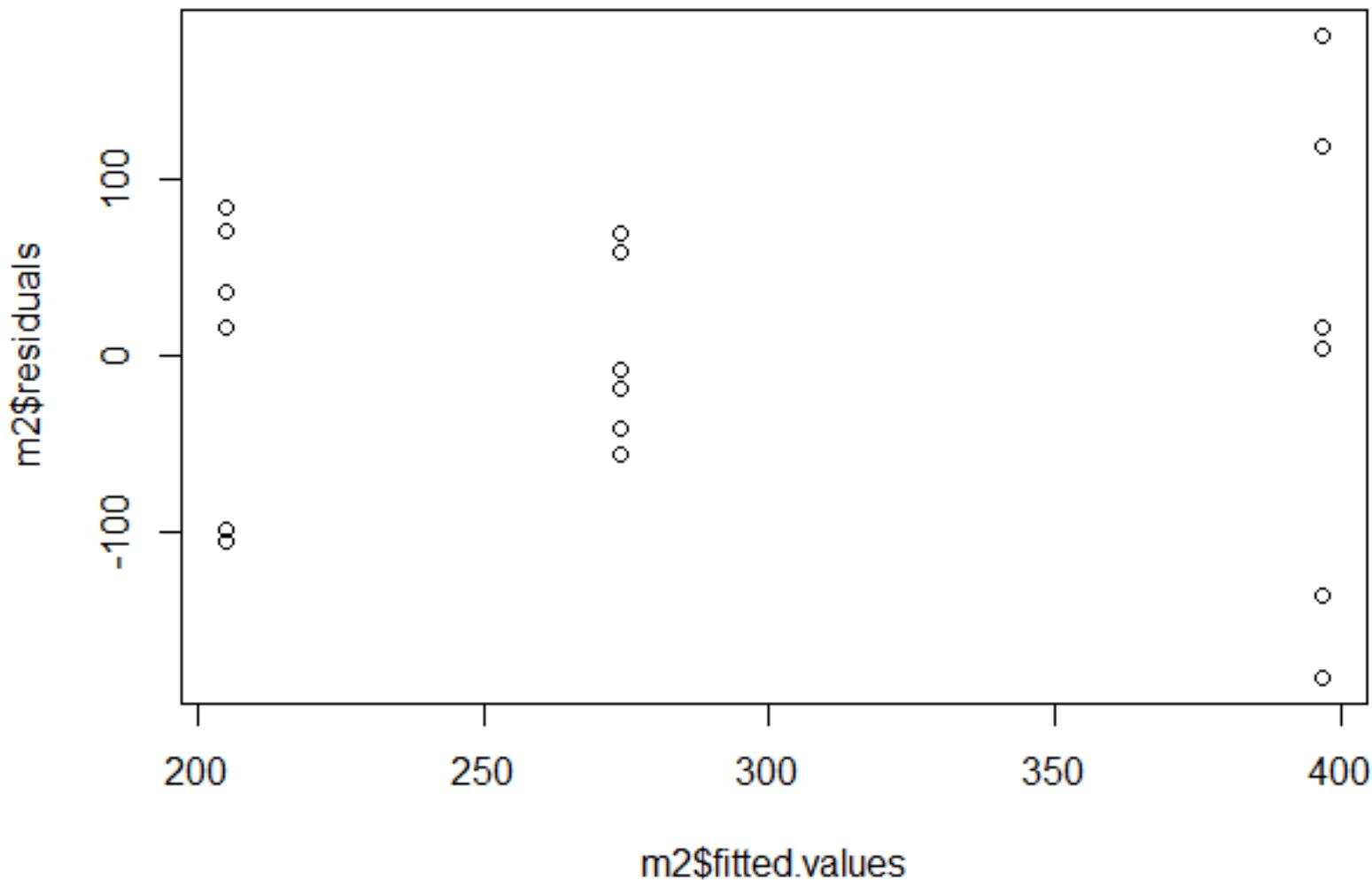


One-Way ANOVA:

```
> ch5sec2 <- read.csv("~/STAT 101B/Nathan/Week 3/intravenous_ex52.csv")
> attach(ch5sec2)
The following objects are masked from ch5sec2 (position 3):
      DrugCompany, Particles
> m2 <- aov(Particles~DrugCompany,data=ch5sec2)
> summary(m2)
      Df Sum Sq Mean Sq F value Pr(>F)
DrugCompany  2 113508   56754   5.771 0.0138 *
Residuals   15 147506    9834
---
Signif. codes:  0 '****' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
> summary(Particles)
  Min. 1st Qu. Median     Mean 3rd Qu.     Max.
  99.0   223.8  262.5   292.0  340.2   578.0
> |
```

Normal Q-Q Plot





Review of R

- TA office Hour
- 3PM to 5PM on Monday at 8141 math science bldg

Basic operations

- ## is a sign of comment

2 + 3 ## summation

2 * 3 ## multiplication

2 - 3

2 / 7

Sqrt(7) ## square root

7 ^ 2 ## exponential

- The order of computation

7 ^ (1 / 2)

7 ^ 1 / 2 ## exponents without parentheses

Exponents operation has **higher** priority

Basic operation

- Assignment

`x = 2` (modern approach)

`x`

`x <- 2` (traditional method)

`x`

`<-` can be used in function definition.

- Logical

`x == 3` ## return TRUE if `x = 3`, otherwise return FALSE

`(x == 3) * 3.4` ## logical value can also be used for computation

Basic functions

- Print function

```
print(x) ## R is case sensitive  
Print(x)
```

- Help function

```
help(exp) ## show help  
? ## equivalent to help function  
example(exp) ## show examples of function
```

Vector and matrix

- Vector

```
z = c(1,4,7) ## creates a vector and assign to z
```

```
print(z)
```

```
z = 1:10 ## create a vector from 1 to 10
```

```
print(z)
```

- Matrix

```
matrix(1:10, nrow = 2, ncol = 5) ## create a 2x5 matrix, the values are from 1 to  
10
```

```
## notice that the matrix is filled by rows.
```

```
print(x[1,1])
```

Vector and matrix

- Combination function

`c()` ## combination of numbers or strings.

“strings” ## generate a string

`c("fee", "fie", "foe")` ## combination of strings.

- Combination function (matrix)

`cbind(c(1,2,3),c(4,5,6))` ## combine by column

`rbind(c(1,2,3),c(4,5,6))` ## combine by row

- Reshape a matrix

`x = matrix(1:10, nrow = 2, ncol = 5)`

`x = matrix(x, nrow = 5, ncol = 2)`

String

- "abcde" defines a string
- `cat(stringA, stringB) ## print the concatenating of two strings`
`cat("The zero occurs at", 2*pi, "radians \n The end")`
- `paste(stringA, stringB) ## concatenate of two strings`
`b = paste("abcde" , "fghj")`
- `substr(stringA, start, end) ## get a substring of stringA`
`c = substr ("abcde" , 2, 4)`

Statements

- If statement

```
if(x == 3) y = 4  
print(y)
```

- For loops

```
cnt = 0  
for(i in 10:20)  
{  
  cat("Iteration ", cnt, ":\n" )  
  print(i)  
  cnt = cnt+1  
  print(cnt)  
}
```

Statements

- While loops

```
while(logical statement)
{
  operations
  ...
}
i = 10; cnt = 0
while(i <=20)
{
  cat("Iteration ", cnt, ":\n" )
  print(i)
  cnt = cnt+1
  i = i+1;
  print(cnt)
}
```

Math and Statistical Function

- `exp(2) ## e function, e^2`
- `mean() ## compute arithmetic mean`
`mean(1:5) ## mean of a vector`
`mean(x) ## mean of a matrix`
`colMeans(x) ## compute mean of a matrix for each column`
`rowMeans(x) ## compute mean of a matrix for each row`
- `sqrt() ## compute square root`
`sqrt(x) ## square root for each element`
- `median() ## compute arithmetic median`
`median(1:4) ## if even, mean of the two middle numbers`
`median(1:5) ## if odd, just the middle number.`
- `var() and sd() ## variance and standard deviation of a vector or matrix`
`var(1:5); sd(1:5) ## the dimension is n-1, 4`
`var(x); sd(x)`

Function

- Define a function

```
FunctionName <- function(argument)
{
  operations
}
```

```
myMean <- function(x)
{
  sum = 0;
  for(ele in x)
    sum = sum + ele;
  mymean = sum / length(x)
  return(mymean)
}
myMean(x)
mean(x)
```

ANOVA

- One way ANOVA (Completely Randomized Design)

```
fit <- aov(y ~ A, data = mydataframe)
```

- Two way ANOVA

```
fit <- aov(y ~ A+B + A:B)
```

Chapter 5 Diet Example

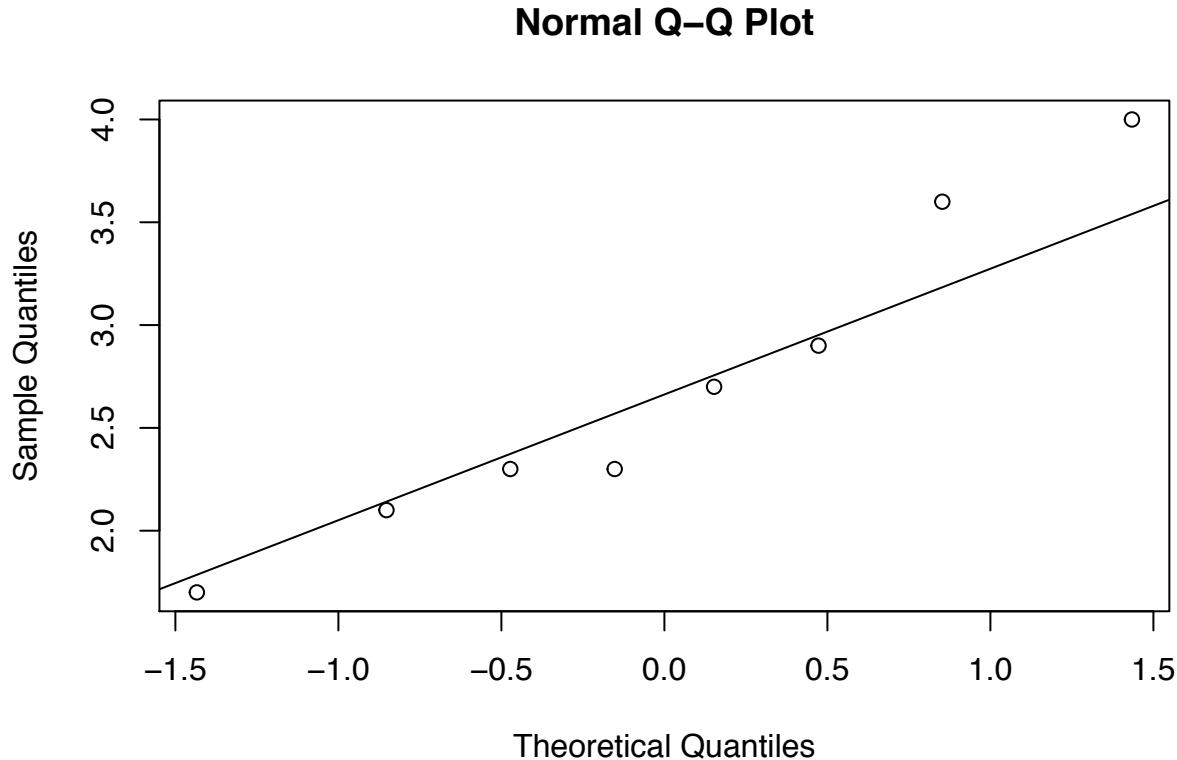
Akram Almohalwas

January 14, 2016

This is an R Markdown document. Markdown is a simple formatting syntax for authoring HTML, PDF, and MS Word documents. For more details on using R Markdown see <http://rmarkdown.rstudio.com>.

When you click the **Knit** button a document will be generated that includes both content as well as the output of any embedded R code chunks within the document. You can embed an R code chunk like this:

```
Diet = c(2.3, 1.7, 4, 3.6, 2.9, 2.7, 2.1, 2.3)
qqnorm(Diet)
qqline(Diet)
```



```
summary(Diet)
```

```
##      Min. 1st Qu. Median     Mean 3rd Qu.    Max.
##    1.700   2.250   2.500   2.700   3.075   4.000
```

```
var(Diet)
```

```
## [1] 0.6028571
```

```

# To get the SS total like we used to do for one-way Anova
ss_t<-var(Diet)*(length(Diet)-1)
ss_t

## [1] 4.22

# In oder to get the Anova summary Example 5.16 page 176 in your textbook
# We do the following:
# First get the sum of all scores squared and that is denoted by Total
# on page 176
sum(Diet^2)

## [1] 62.54

# Since the over all average for data is 2.7, we create a benchmark with 2.7
# for all cells
grand<-c(2.7,2.7,2.7,2.7,2.7,2.7,2.7,2.7)
# Sum of the squares of grand gives Grand Average
sum(grand^2)

## [1] 58.32

# If we subtract grand average from the average per group we get the diet effect
effect<-c(-0.7,-0.7,1.1,1.1,0.1,0.1,-0.5,-0.5)
# sum of the squares of effect gives SS(Diet) in a regular One-Way Anova
var(effect)*7

## [1] 3.92

sum(effect^2)

## [1] 3.92

# In order to get the Residual we subtract the grand average as well as the per
# group average from the observed value
res<-Diet-grand-effect
# sum all the squares of the residuals gives SSE
sum(res^2)

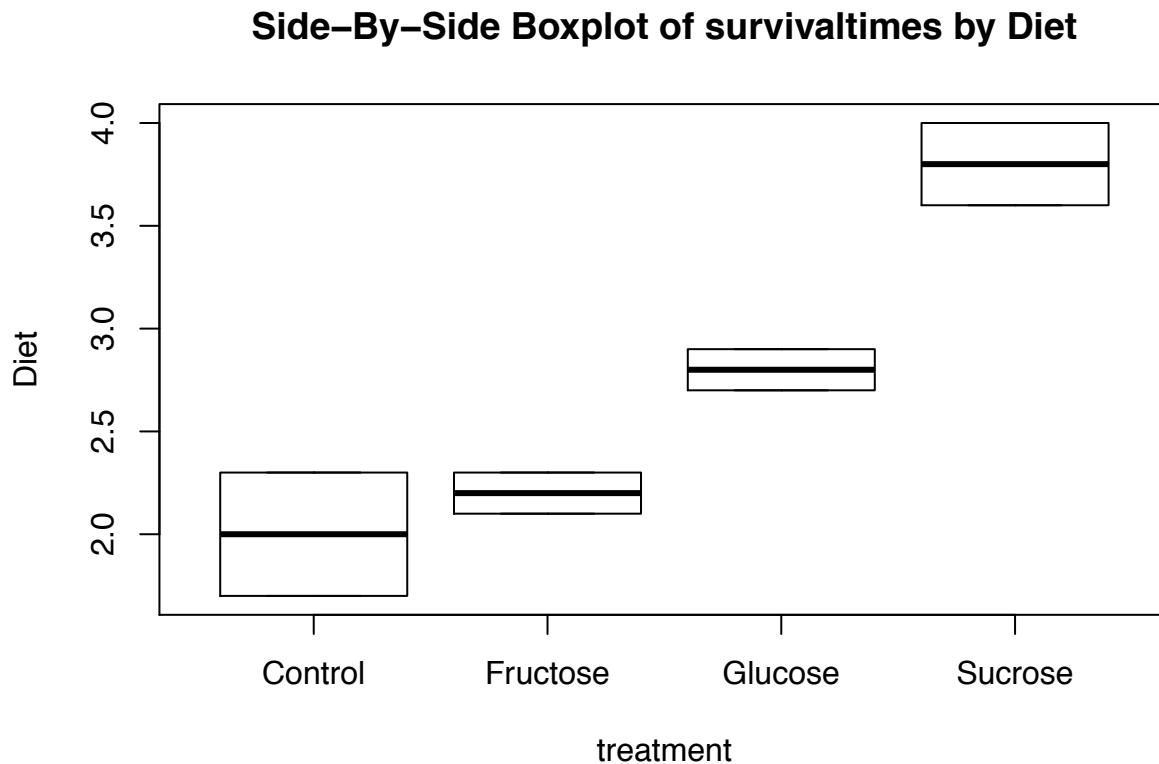
## [1] 0.3

treatment= c(rep("Control",2), rep("Sucrose",2), rep("Glucose",2),rep("Fructose",2))
data1 = data.frame(Diet,treatment)
Diet

## [1] 2.3 1.7 4.0 3.6 2.9 2.7 2.1 2.3

```

```
# Here we create boxplots for the data side-by-side
plot(Diet ~ treatment, data=data1, main="Side-By-Side Boxplot of survivaltimes by Diet")
```



```
mall<-mean(Diet)
mall
```

```
## [1] 2.7
```

```
ma<-mean(Diet[treatment=="Control"])
ma
```

```
## [1] 2
```

```
mb<-mean(Diet[treatment=="Sucrose"])
mb
```

```
## [1] 3.8
```

```
mc<-mean(Diet[treatment=="Glucose"])
mc
```

```
## [1] 2.8
```

```

md<-mean(Diet[treatment=="Fructose"])
md

## [1] 2.2

means<-c(mall,ma(mb,mc,md)
tmeans<-c(ma,mb,mc,md)
mean(tmeans)

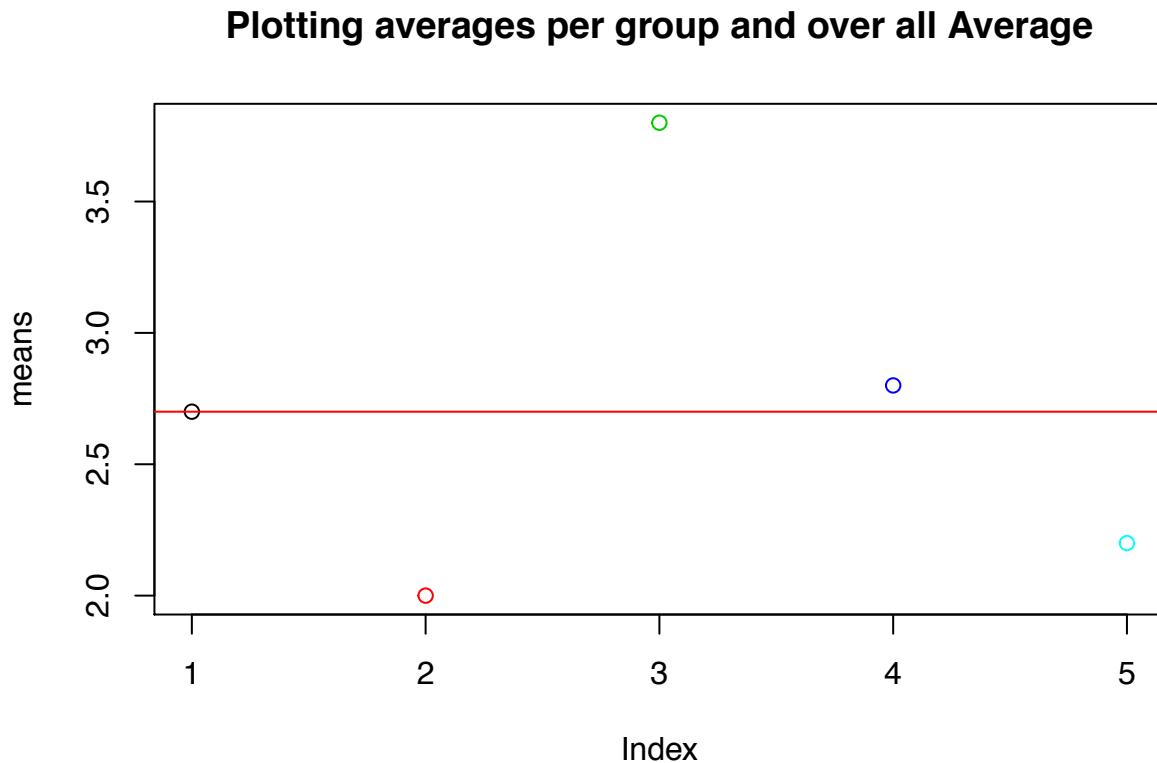
## [1] 2.7

var(tmeans)

## [1] 0.6533333

plot(means,col=1:8,main="Plotting averages per group and over all Average")
abline(h=mall,col="red")

```



```

results <- aov(Diet ~ treatment, data=data1)
summary(results)

```

```

##          Df Sum Sq Mean Sq F value    Pr(>F)

```

```

## treatment      3    3.92    1.307   17.42 0.00925 **
## Residuals     4    0.30    0.075
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

results$residuals

##    1    2    3    4    5    6    7    8
##  0.3 -0.3  0.2 -0.2  0.1 -0.1 -0.1  0.1

sum(Diet^2)

## [1] 62.54

reg<-lm(Diet~treatment)
reg

## 
## Call:
## lm(formula = Diet ~ treatment)
## 
## Coefficients:
##             (Intercept)  treatmentFructose  treatmentGlucose
##                   2.0                  0.2                  0.8
## treatmentSucrose
##                   1.8

summary(reg)

## 
## Call:
## lm(formula = Diet ~ treatment)
## 
## Residuals:
##    1    2    3    4    5    6    7    8
##  0.3 -0.3  0.2 -0.2  0.1 -0.1 -0.1  0.1
## 
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)    
## (Intercept) 2.0000    0.1936 10.328 0.000496 ***
## treatmentFructose 0.2000    0.2739  0.730 0.505681  
## treatmentGlucose 0.8000    0.2739  2.921 0.043192 *  
## treatmentSucrose 1.8000    0.2739  6.573 0.002773 ** 
## ---      
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 0.2739 on 4 degrees of freedom
## Multiple R-squared:  0.9289, Adjusted R-squared:  0.8756 
## F-statistic: 17.42 on 3 and 4 DF,  p-value: 0.009248

```

```
# All pair-wise comparisons between groups
pairwise.t.test(Diet, treatment, p.adjust="bonferroni")
```

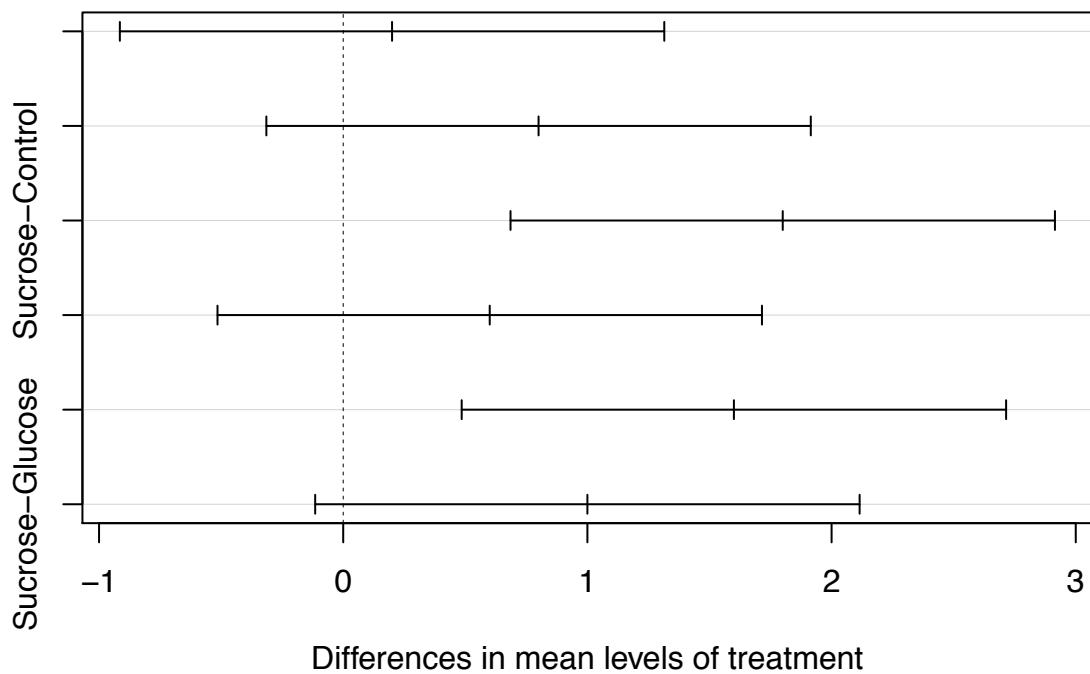
```
##
##  Pairwise comparisons using t tests with pooled SD
##
## data: Diet and treatment
##
##          Control Fructose Glucose
## Fructose 1.000   -      -
## Glucose   0.259   0.562   -
## Sucrose   0.017   0.026   0.130
##
## P value adjustment method: bonferroni
```

```
TukeyHSD(results, conf.level = 0.95)
```

```
## Tukey multiple comparisons of means
## 95% family-wise confidence level
##
## Fit: aov(formula = Diet ~ treatment, data = data1)
##
## $treatment
##          diff      lwr      upr      p adj
## Fructose-Control 0.2 -0.9148496 1.31485 0.8805808
## Glucose-Control  0.8 -0.3148496 1.91485 0.1337642
## Sucrose-Control  1.8  0.6851504 2.91485 0.0095276
## Glucose-Fructose 0.6 -0.5148496 1.71485 0.2677272
## Sucrose-Fructose 1.6  0.4851504 2.71485 0.0145912
## Sucrose-Glucose  1.0 -0.1148496 2.11485 0.0703156
```

```
plot(TukeyHSD(results, "treatment"))
```

95% family-wise confidence level



```
# Sucrose-Control  1.8  0.6851504  2.91485  0.0095276  
# Sucrose-Fructose 1.6  0.4851504  2.71485  0.0145912
```

Chapter 5:

Randomization and the Basic Factorial Design (BF Design)

Designing Experiments

Every experiment involves a sequence of activities:

1. **Conjecture** – the original hypothesis that motivates the experiment.
2. **Experiment** – the test performed to investigate the conjecture.
3. **Analysis** – the statistical analysis of the data from the experiment.
4. **Conclusion** – what has been learned about the original conjecture from the experiment. Often the experiment will lead to a revised conjecture, and a new experiment, and so forth.

Choosing a Design Structure

Two principles for assigning treatments

- Random assignment
- Blocking

Four design structures based on these principles:

- BF: Basic Factorial
- CB: Complete Block
- LS: Latin Square
- SP/RM: Split Plot/Repeated Measures

Random Assignment

- Randomly assign materials to treatment groups
- This puts all nuisance variation into the 'noise' and lets us fit the ANOVA model
- We assume a 'balanced' design

Diet Effectiveness

- Researchers assigned 116 subjects to four diets:
 - Atkins
 - Ornish
 - Weight Watchers
 - Zone
- How?

Shuffle Some Cards

- Write everyone's name on a card
- Shuffle the cards together many times
- Deal them into four piles

Random Assignment Using R

```
# (No seed here)
>names=1:116

>groups=rep(c("Atkins","Ornish","WW","Zone"),29)

>shuffled=sample(names,116,replace=FALSE)

>mydata=data.frame(Group=groups,Names=shuffled)

>groups[1:20]
[1] "Atkins" "Ornish" "WW" "Zone" "Atkins"
"Ornish" "WW" "Zone" "Atkins" [10] "Ornish" "WW"
"Zone" "Atkins" "Ornish" "WW" "Zone" "Atkins"
"Ornish" [19] "WW" "Zone"
>mydata[1:20,1]
[1] Atkins Ornish WW Zone Atkins Ornish WW Zone
Atkins Ornish WW [12] Zone Atkins Ornish WW Zone
Atkins Ornish WW Zone Levels: Atkins Ornish WW
Zone
```

Pseudo-Random

- Computers use 'pseudo' random numbers determined by an algorithm based on a 'seed' value
- If you know the seed you will always get the same sequence of random numbers
- Set the seed prior to doing a randomization routine
 - Make sure you always get the same results
 - Allows others to go back and reproduce your findings

```
# (with seed)
>set.seed(1234)

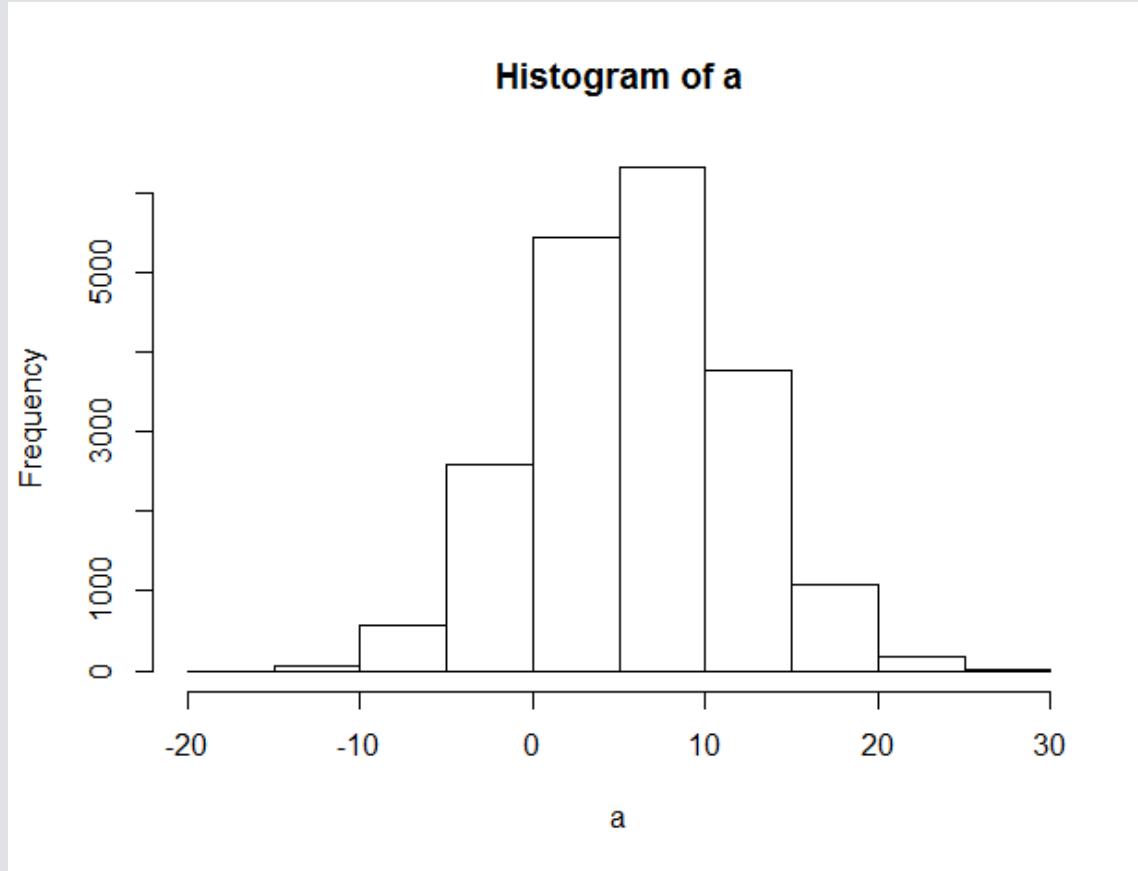
>names=1:116

>groups=rep(c("Atkins","Ornish","WW","Zone"),29)

>shuffled=sample(names,116,replace=FALSE)
>mydata=data.frame(Group=groups,Names=shuffled)
>groups[1:20]
[1] "Atkins" "Ornish" "WW" "Zone" "Atkins" "Ornish"
"WW" "Zone" "Atkins" [10] "Ornish" "WW" "Zone"
"Atkins" "Ornish" "WW" "Zone" "Atkins" "Ornish" [19]
"WW" "Zone"
>mydata[1:20,1] [1] Atkins Ornish WW Zone Atkins
Ornish WW Zone Atkins Ornish WW [12] Zone Atkins
Ornish WW Zone Atkins Ornish WW Zone Levels: Atkins
Ornish WW Zone
```

```
>#using set.seed() before a random process  
  
>set.seed(1234)  
  
>#drawing 4 random digits from a normal distribution  
with mean 3 and sd 2  
  
>rnorm(4,3,2)  
[1] 0.5858685 3.5548585 5.1688824 -1.6913954  
  
>#drawing 4 random digits from a normal distribution  
with mean 6 and sd 6  
  
>rnorm(4,6,6) [1] 8.574748 9.036335 2.551560 2.720209
```

```
> a<-rnorm(20000,6,6)  
>hist(a)
```



```
>summary(a)  
Min. 1st Qu. Median Mean 3rd Qu. Max.  
-18.760 1.908 6.019 5.995 10.040 28.360  
sd(a)  
[1] 5.981181
```

Randomized Basic Factorial

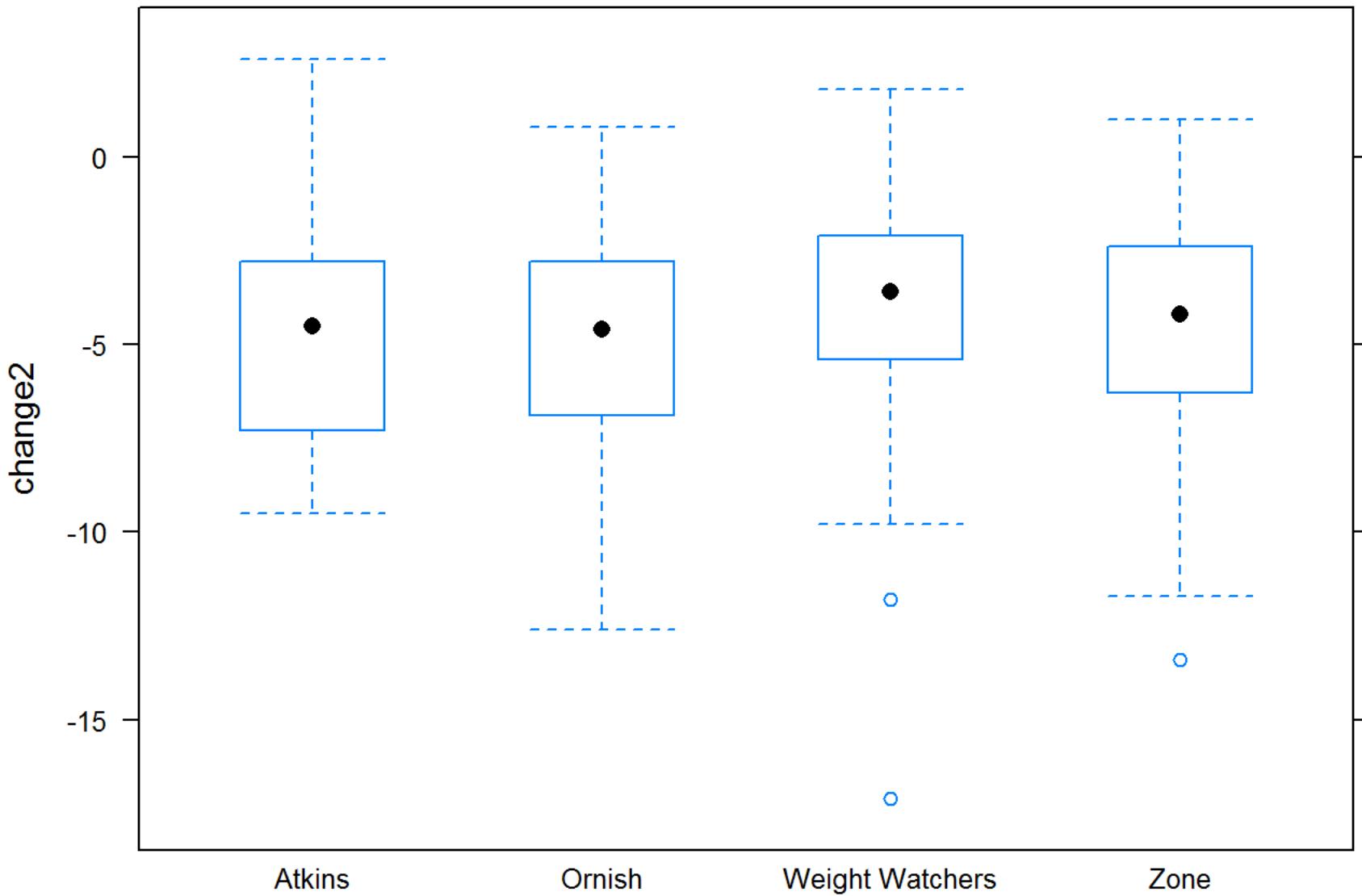
- Use a random method to assign each experimental unit to a treatment
 - Minimizes Bias
 - Ensures that the error follows a chance model and so can be estimated
- Shorthand: RBF

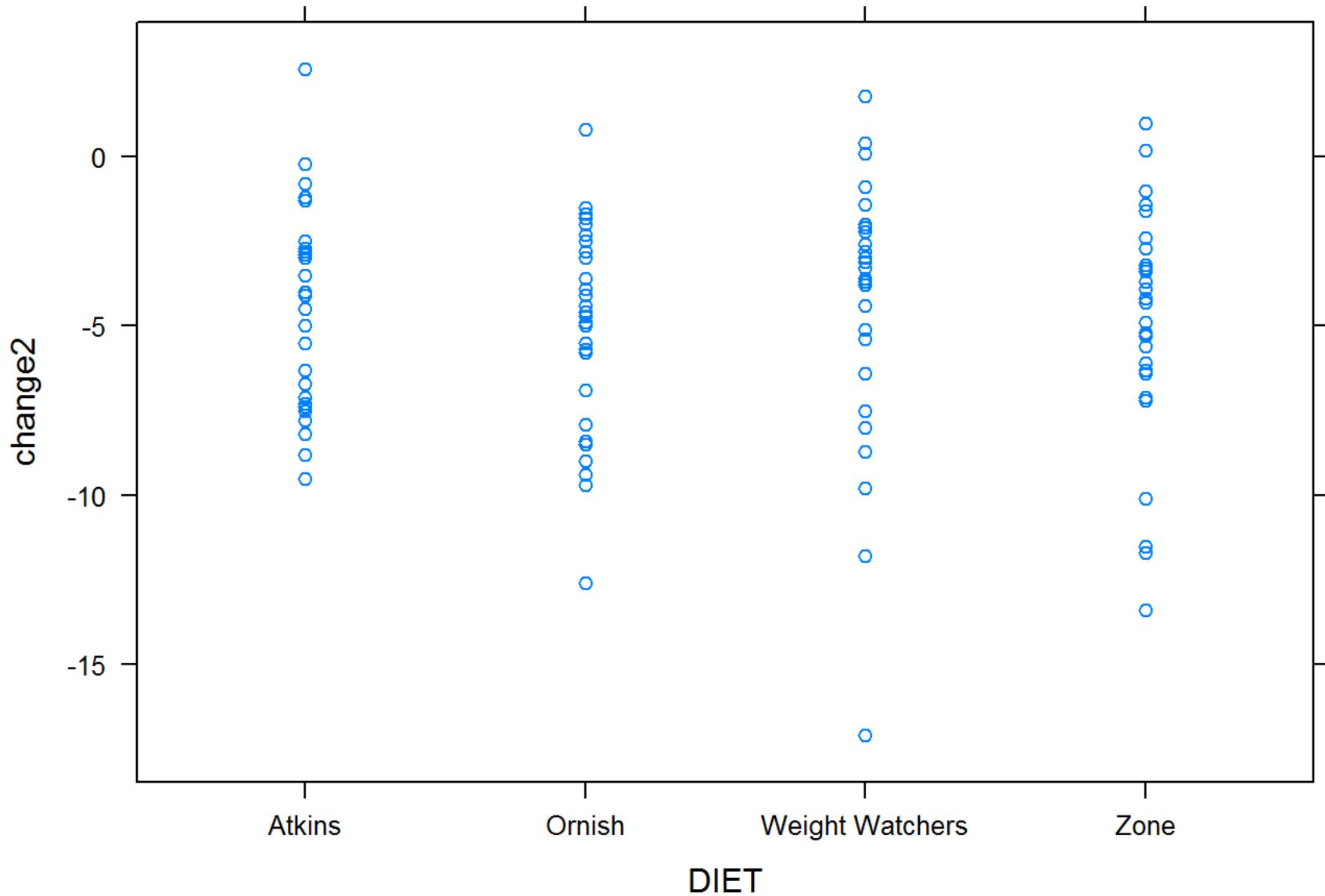
Which diet is most effective?

- Atkins, Ornish, Weight Watchers, Zone
- 116 overweight subjects available
 - Randomly assign subjects to each of the four diets
- Measure weight-loss after 2 months

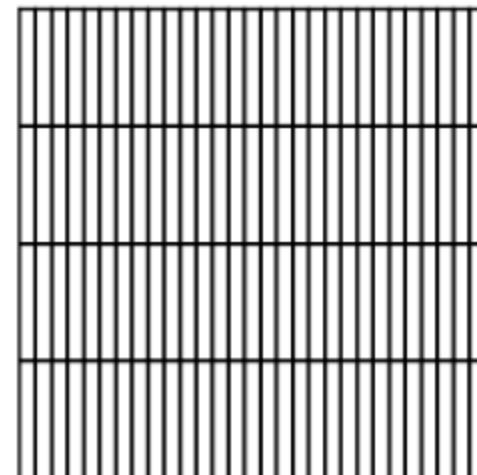
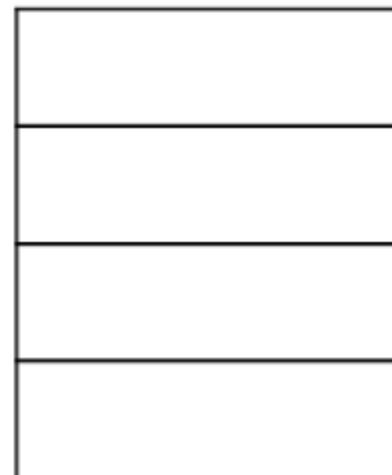
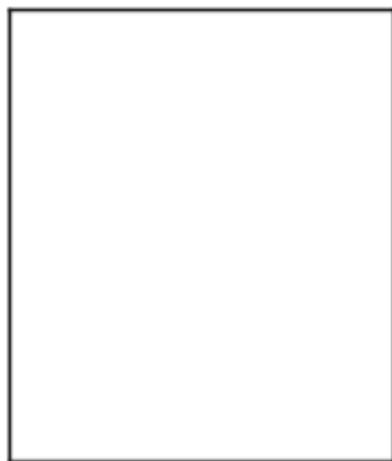
```
# import Diet Data
# diet_balanced <- read.csv("diet_balanced_design.csv")
diet_balanced <- read.csv("~/STAT 101B/Nathan/Week
3/diet_balanced_design.csv")
head(diet_balanced)
## SUBJECT DIET DIETW DIETX DIETY DIETZ AGE SEX WEIGHT_0
BMI_0 ## 1 1 Atkins No No no yes 43 Female 92.3 36.50963 ##
2 2 Atkins No No no yes 23 Male 109.5 37.88927 ## 3 3 Atkins
No No no yes 42 Male 86.5 28.90173 ## 4 4 Atkins No No no
yes 55 Male 118.0 31.67870 ## 5 5 Atkins No No no yes 66
Female 80.2 30.94016 ## 6 6 Atkins No No no yes 37 Female
109.2 40.60083 ## WAIST_0 DROPOUT2 WEIGHT_2 BMI_2 WAIST_2
DROPOUT6 WEIGHT_6 BMI_6 ## 1 123 no 89.8 35.52075 118.0 no
92.0 36.39097 ## 2 119 no 104.0 35.98616 112.0 no 96.2
33.28720 ## 3 103 no 79.2 26.46263 94.5 no 80.4 26.86358 ##
4 110 no 115.0 30.87331 108.5 no 117.4 31.51762 ## 5 103 no
77.5 29.89854 100.5 no 78.0 30.09143 ## 6 113 no 102.5
38.10976 108.0 no 107.3 39.89441
```

Boxplot





Diet Diagrams



One-way 'Effects Model'

$$y_{ij} = \mu + \tau_j + \epsilon_{ij}$$

$$\epsilon_{ij} \sim N(0, \sigma)$$

$$\sum \tau_j = 0$$

y_{ij} is the weight loss of subject i in diet group j

One-way 'Effects Model'

$$y_{ij} = \mu + \tau_j + \epsilon_{ij}$$

- This model says that all of the subjects experience the same 'benchmark' weightloss, μ
- Subjects in diet j share a weightloss, τ_j
- Each subject has his/her own unique weightloss, ϵ_{ij}

Estimating Model Terms

- Estimating effects
 - Inside/Outside Factors
 - Using the computer

Inside/Outside Factors

- One factor is **inside** another if each group of the first factor fits completely inside some group of the second factor

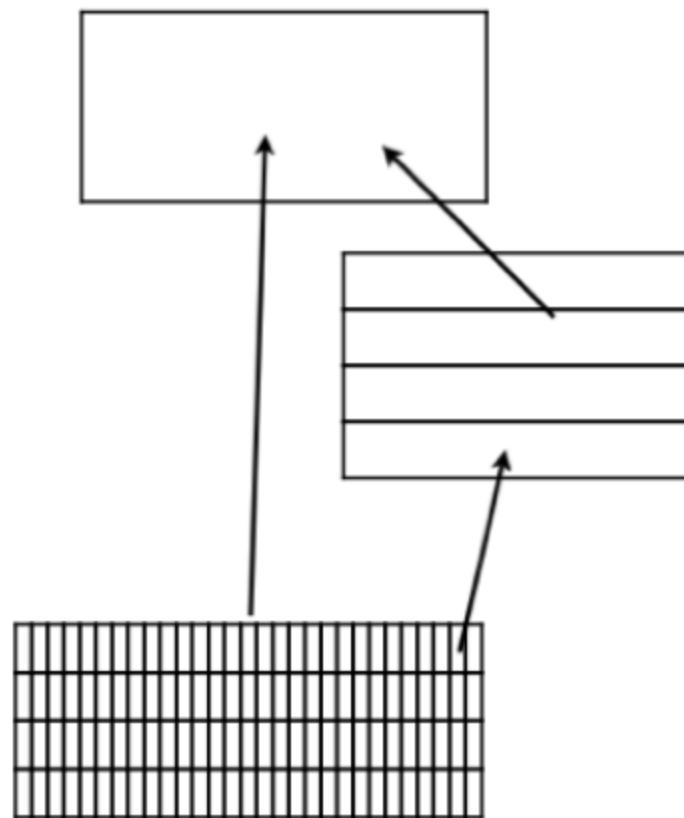
Inside or Outside Benchmark?

- Each of the four diet groups are included in the large benchmark group.
- So diet is **inside** benchmark. Benchmark is **outside** diet.

Residuals?

- There are 29×4 individuals. Each **inside** one of the diet groups
- Every group is **inside** the benchmark group
- So the residual error factor is **inside** the diet and benchmark factors

Inside/Outside diagram



Decomposition

- Break each individual observation into its component parts

Observation = benchmark + structural condition effects + residual error

Estimated Effects

Estimated effect for a factor = Average for the factor – sum of estimated effects for all outside factors

The 'sum of estimated effects of outside factors' is called the **partial fit**

Benchmark

Estimated effect for a factor = Average for the factor – sum of estimated effects for all outside factors

Benchmark does not have any outside factors, so...

benchmark = mean of all observations – 0

benchmark = -4.72

Diet Effects

- One outside factor (benchmark)
 - Atkins = Mean of those in Atkins group – benchmark
 - Ornish = Mean of those in Ornish group – benchmark
 - WW = Mean of those in WW group – benchmark
 - Zone = Mean of those in Zone group – benchmark
- Note: every individual observation within a group has the same estimated effect

```
#Overall Mean
benchmark <- mean(change2)
benchmark

## [1] -4.715517

#Diet Effects
diet_effects <- by(change2,DIET,function(x) mean(x))-benchmark

#OR
atkins <- mean(change2[DIET=='Atkins']) - benchmark
ornish <- mean(change2[DIET=='Ornish']) - benchmark
ww <- mean(change2[DIET=='Weight Watchers']) - benchmark
zone <- mean(change2[DIET=='Zone']) - benchmark

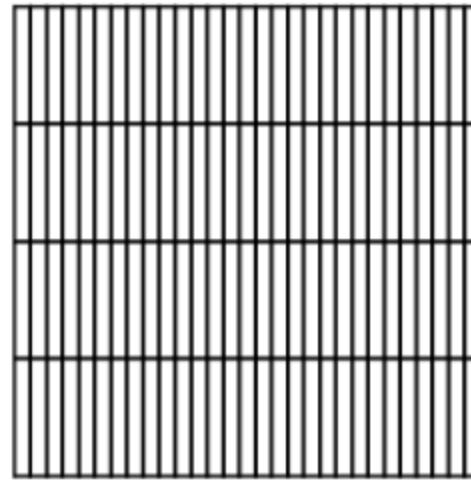
table(atkins,ornish,ww,zone)

## , , ww = 0.325862068965517, zone = -0.0396551724137932
##
##                               ornish
## atkins                  -0.28448275862069
## -0.00172413793103399          1
```

Diet Decomposition

-4.72

-0.00172
-0.284
0.326
-0.0397



Chance Errors

- Two outside factors: benchmark and diet
- Each observation is in its own chance error 'group', the average is just the actual observation
 - ex. observation is the Atkins group has:
 $\text{est. effect} = \text{observation} - \text{benchmark} - \text{Atkins effect}$
 - Note: each of these are unique

```
#Residuals
atkins_R <- change2[DIET=='Atkins'] - benchmark - atkins
ornish_R <- change2[DIET=='Ornish'] - benchmark - ornish
ww_R <- change2[DIET=='Weight Watchers'] - benchmark - ww
zone_R <- change2[DIET=='Zone'] - benchmark - zone

#sum of squares for diet treatment; 3 degrees of freedom
SS_diet <- 29*(atkins^2+ornish^2+ww^2+zone^2)
MS_diet <- SS_diet/3

#sum of squares for residuals; 112 degrees of freedom
SS_residuals <- sum(atkins_R^2+ornish_R^2+ww_R^2+zone_R^2)
MS_residuals <- SS_residuals/112

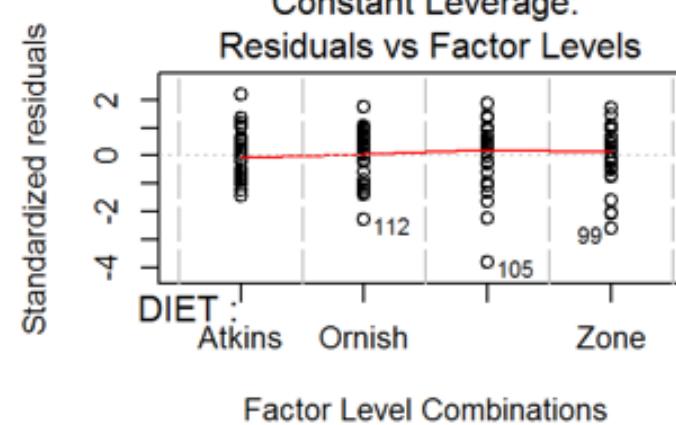
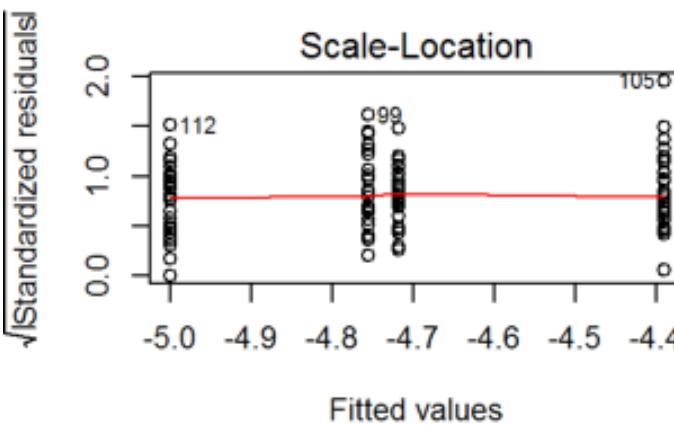
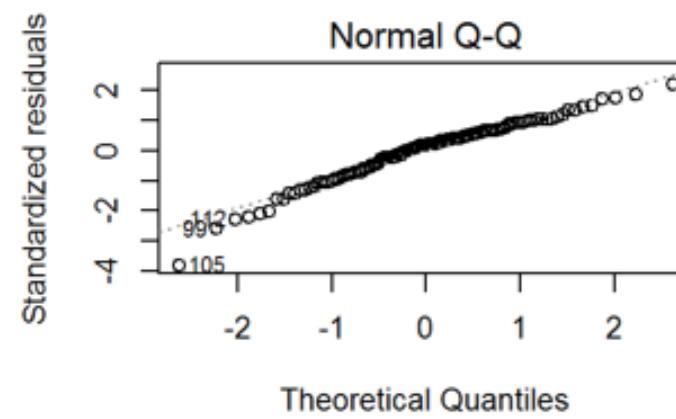
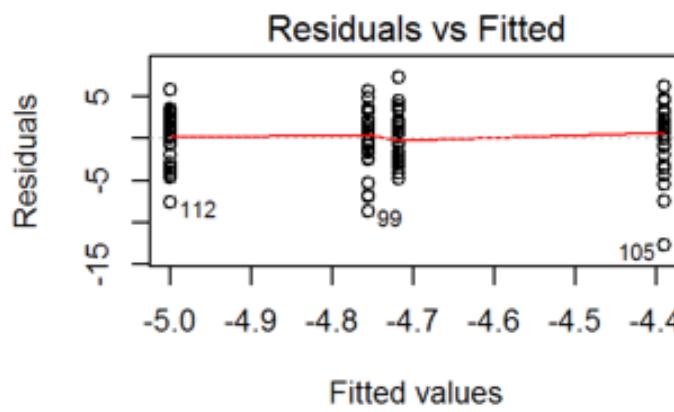
#f-statistic and p-value
F_stat <- MS_diet/MS_residuals
pf(F_stat, 3, 112, lower.tail=FALSE)

## [1] 0.9236083
```

```
#using aov
m1 <- aov(change2~DIET,data=diet_balanced)

par(mfrow=c(2,2))

plot(m1)
```



```
model.tables(m1)
```

```
## Tables of effects
##
## DIET
## DIET
##          Atkins          Ornish Weight Watchers        Zone
##          -0.0017         -0.2845         0.3259        -0.0397
```

```
summary(m1)
```

```
##           Df Sum Sq Mean Sq F value Pr(>F)
## DIET       3   5.5   1.824   0.159  0.924
## Residuals 112 1284.0  11.464
```

```
#checking conditions
```

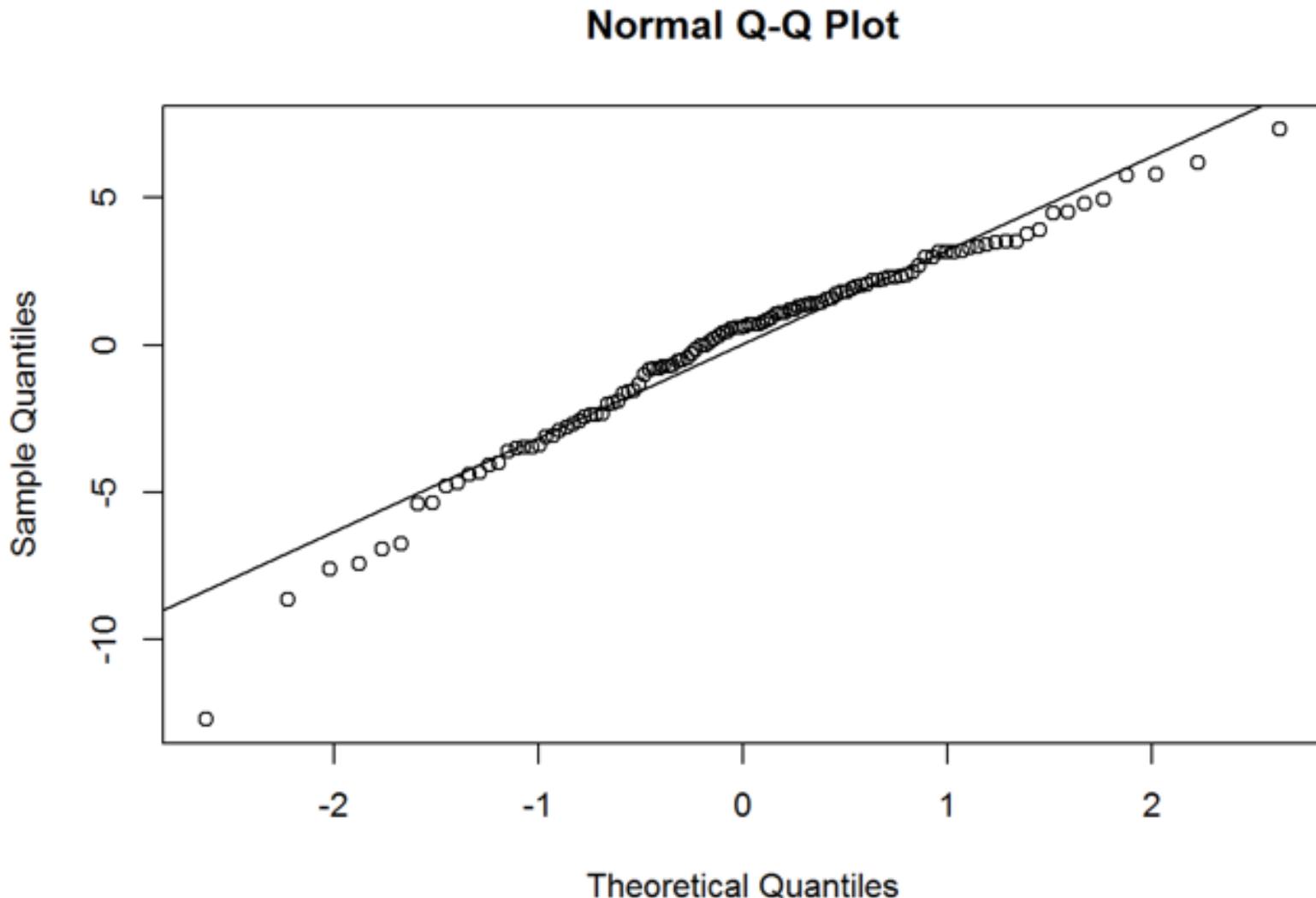
```
par(mfrow=c(1,1))
```

```
qqnorm(m1$residuals)
```

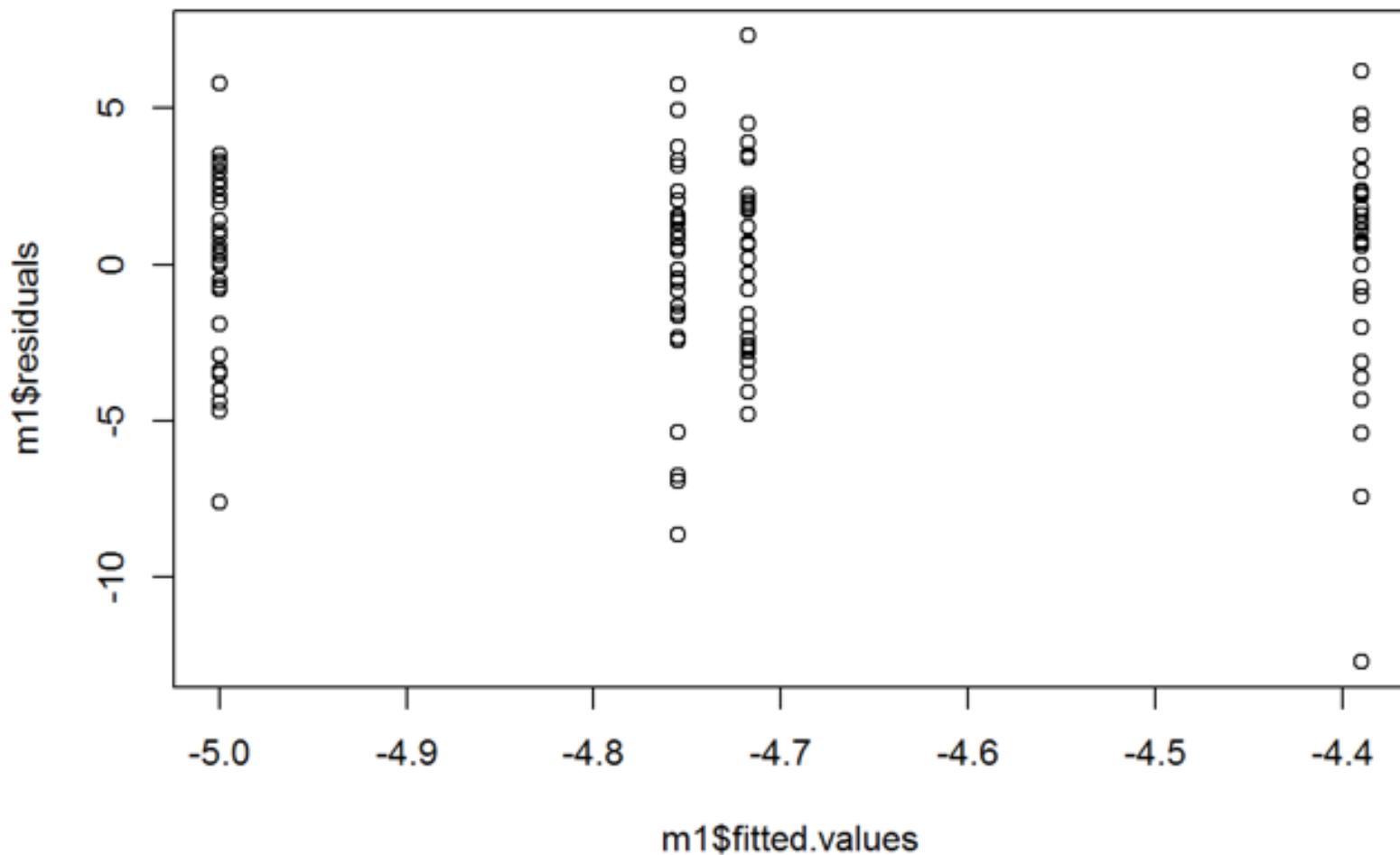
```
qqline(m1$residuals)
```

```
qqnorm(m1$residuals)
```

```
qqline(m1$residuals)
```



```
plot(m1$residuals ~ m1$fitted.values)
```



Hypotheses One-way 'Effects Model'

$$H_0 : \tau_i = 0 \quad \text{for all } i$$

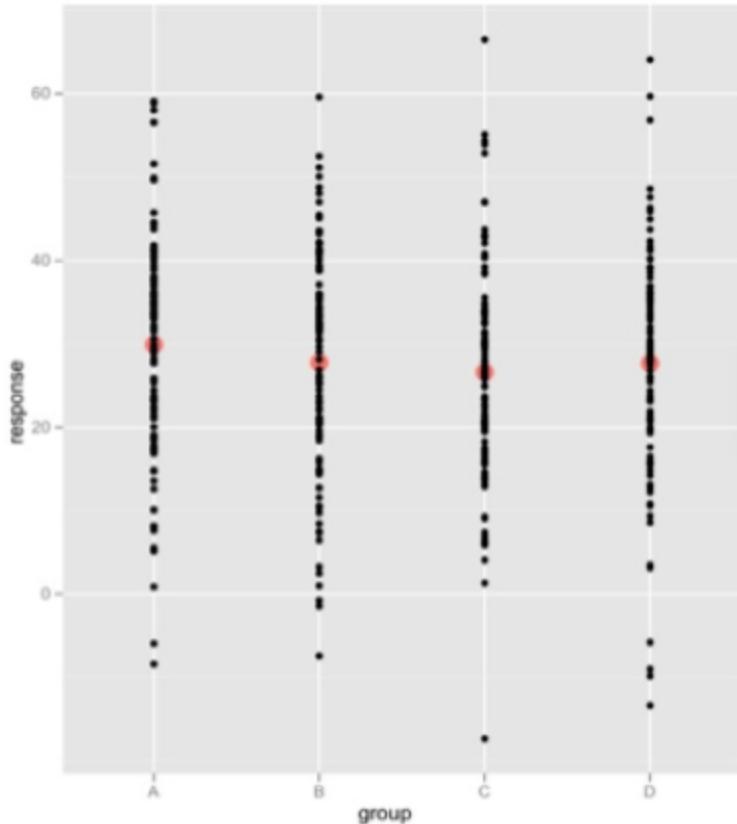
$$H_a : \tau_i \neq 0 \quad \text{for at least one } i$$

- Null: none of the treatments will cause a change
- Alternative: At least one of the treatments will cause a change

How is this done?

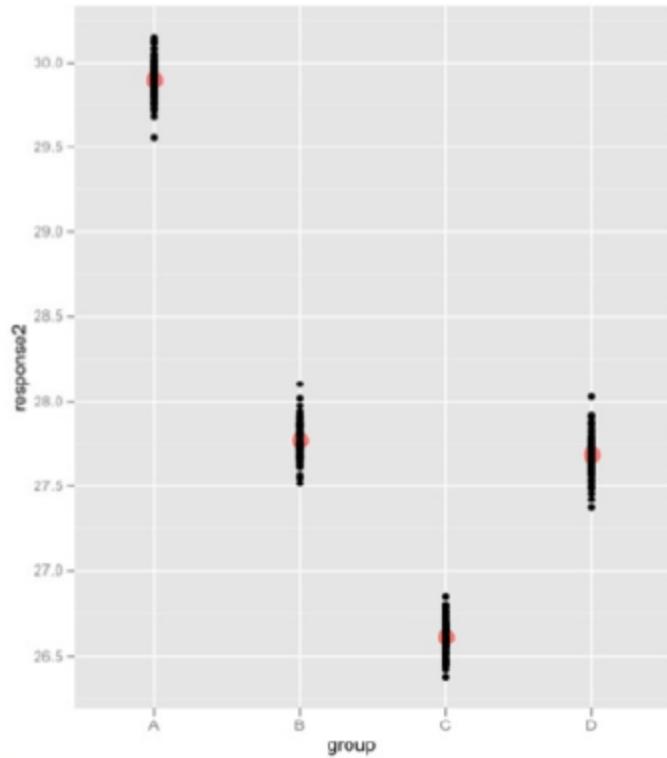
- Compare variability between the groups with the variability in chance errors
- If variability between groups is large relative to chance errors, then this is evidence that observed differences are not due to chance but are 'real'

F-Ratio



Small F-ratio: the variation within groups is bigger than the variation between groups

F-Ratio



Big F-ratio: the variation within groups is smaller than the variation between groups

Measuring Variability

- Measured by the 'sum of squares'
- The total variability in the data is:

$$SS_T = \sum_{i=1}^a \sum_{j=1}^n (y_{ij} - \bar{y}_{..})^2$$

- Can be decomposed into components for each factor
- $SST = SS$ due to treatments + SS due to 'error'

Between Group Variability

- Measured by how much the group means vary about the 'grand' mean

$$SS_{treat} = \sum_{i=1}^a \sum_{j=1}^n (\bar{y}_{i\cdot} - \bar{y}_{..})^2$$

- Sum of Squares due to treatment

Average of the response variable

Within Group Variability

- Measures the variability within each group, added together

$$SS_E = \sum_{i=1}^a \sum_{j=1}^n (\bar{y}_{ij} - \bar{y}_{i\cdot})^2$$

- Sum of squares due to Error

Average of the response variable for subject i and treatment j which is the same as y_{ij} since we have one measurement per subject

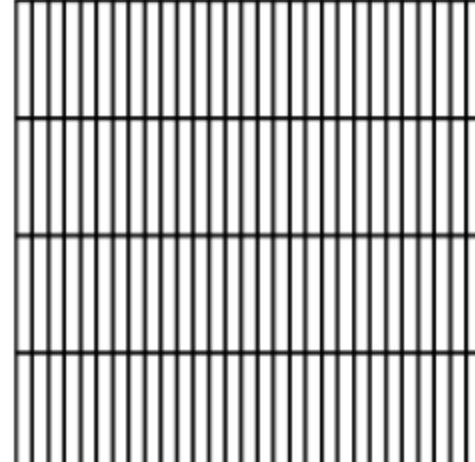
- If the means really differ, then the variability between the group means (SS Treatment) should be greater than the variability within groups (SS Error)
- One difficulty, the more terms we add, the greater the variability becomes.
 - If we add treatment groups, we add variability.
 - If we add subjects we add variability
- So...we take the mean SS by dividing by the 'degrees of freedom'

Degrees of Freedom

- Measures the number of independent observations used to estimate an effect
- Can think of df as the number of bits of information
- The book calls them 'the number of free numbers' in a table

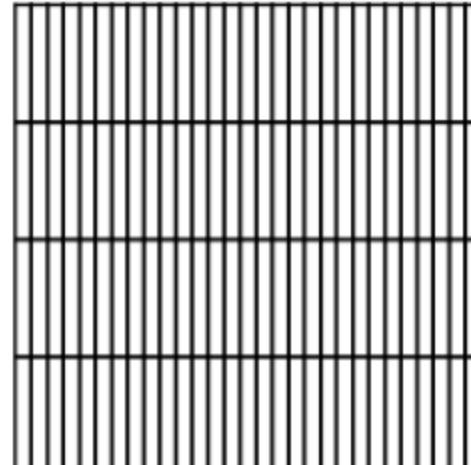
df: benchmark

- df for a factor = number of levels for the factor – sum of df for all outside factors
- There is only one level and no outside factors so, $df=1 - 0 = 1$



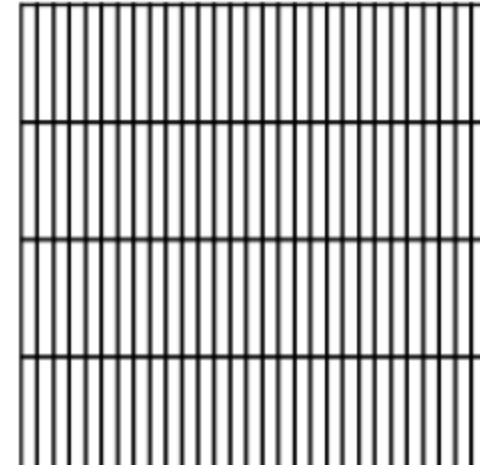
df: diet

- There are four levels and one outside factor using 1 df so, $df=4 - 1 = 3$



df: residuals

- There are 116 'levels' and 2 outside factors whose df total to 4 so, $df = 116 - 4 = 112$



Why is this the case?

- The diet factor has 4 unique estimated effects. A condition was that these sum to one. If you know 3 then you can compute the fourth.
- Hence we have 3 'free' values.

-0.00172
-0.284
0.326
-0.0397

Mean Sum of Squares

- $MS = SS/df$

- $SS(\text{diets}) =$

$$29 \times (\text{atkins}^2 + \text{ornish}^2 + \text{ww}^2 + \text{zone}^2)$$

$$= 5.47$$

- So

$$MS(\text{diets}) = 5.47 / 3 = 1.82$$

Since we have balanced design with 29 subjects in each treatment.

Mean Sum of Squares

- $MS = SS/df$
- $MS(\text{residuals}) =$
 $(\text{atkins.resid}^2 + \text{ornish.resid}^2 + \text{ww.resid}^2 + \text{zone.resid}^2) / 112$
 $= 11.5$

Sampling Distribution of F

- F-statistic = $MSTreat/MSResid$
- F-statistic = $1.82 / 11.5 = 0.158$
- Under assumptions of our model, the F-Statistic follows an F distribution with two different df parameters: (df numerator, df denominator)
- We fail to reject the null when F is small
- P-value = $P(F > F_{\text{obs}}) = P(F > 0.158)$
$$1 - \text{pf}(.158, 3, 112) = 0.924$$

In R...

```
summary(aov(change2~DIET,data=balanced.diet))
      Df Sum Sq Mean Sq F value Pr(>F)
DIET       3      5    1.82    0.16   0.92
Residuals 112  1284   11.46
```

- Conclusion, there is no evidence that diet has an effect on weight loss after two months of dieting

Multiple Comparison Methods

Comparing Multiple Means

Comparing Treatment Means

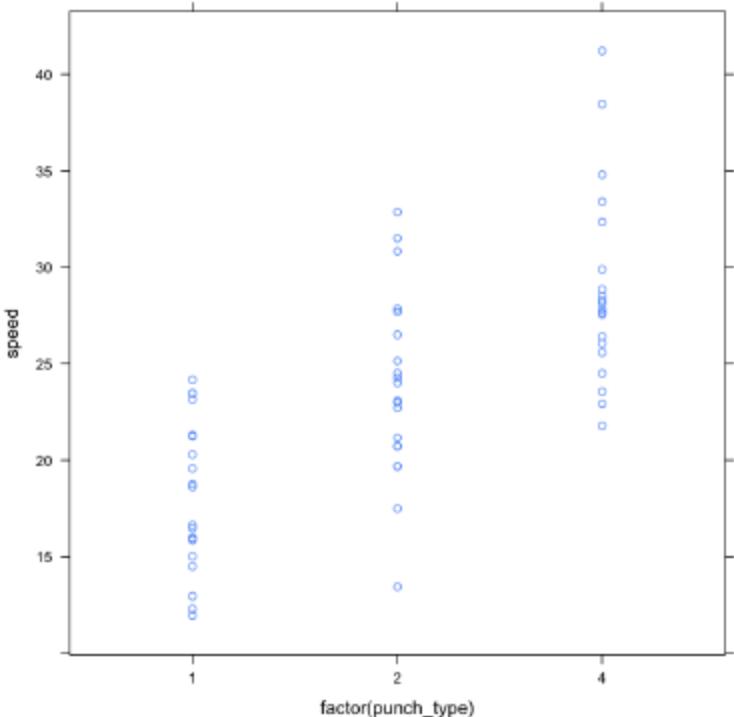
- Suppose ANOVA F-test indicates a difference in group effects. Which groups are different?
- Need to estimate and compare the groups
- How can we find confidence intervals?
- The problem of multiple comparisons

HBO Boxing Data

- Do different punch types have different speeds?
- RBF[1] design has 60 boxers randomly assigned to throw 1 of three punches types (cross, hook, uppercut)



HBO Boxing Data



```
#Initial look at the data
library(lattice)
xyplot(speed~factor(punch_type),data=punches)
by(punches$speed,factor(punches$punch_type),function(x) mean(x))

#ANOVA on boxing data
m1=aov(speed~factor(punch_type),data=punches)
summary(m1)
model.tables(m1)

#check conditions with residuals
qqnorm(m1$residuals)
```

```
#Need to read in punches.Rdata file first
```

```
#Initial look at the data
```

```
library(lattice)
```

```
xyplot(speed~factor(punch_type),data=punches)
```

```
by(punches$speed,factor(punches$punch_type),function(x) mean(x))
```

```
#ANOVA on boxing data
```

```
m1=aov(speed~factor(punch_type),data=punches)
```

```
summary(m1)
```

```
model.tables(m1)
```

```
#Check conditions with residuals
```

```
qqnorm(m1$residuals)
```

```
qqline(m1$residuals)
```

```
plot(m1$residuals ~ m1$fitted.values)
```

```
#####
#multiple comparisons; no corrections
t.test(punches$speed[punches$punch_type=='1'],punches$speed[punches$punch_type=='2']
)
t.test(punches$speed[punches$punch_type=='1'],punches$speed[punches$punch_type=='4']
)
t.test(punches$speed[punches$punch_type=='2'],punches$speed[punches$punch_type=='4']
)

#multiple comparisons; Bonferroni Correction
conflevel=1-.05/3
t.test(punches$speed[punches$punch_type=='1'],punches$speed[punches$punch_type=='2']
,conf.level=conflevel)
t.test(punches$speed[punches$punch_type=='1'],punches$speed[punches$punch_type=='4']
,conf.level=conflevel)
t.test(punches$speed[punches$punch_type=='2'],punches$speed[punches$punch_type=='4']
,conf.level=conflevel)

#Tukey HSD
TukeyHSD(m1)
plot(TukeyHSD(m1))
```

HBO Boxing Data

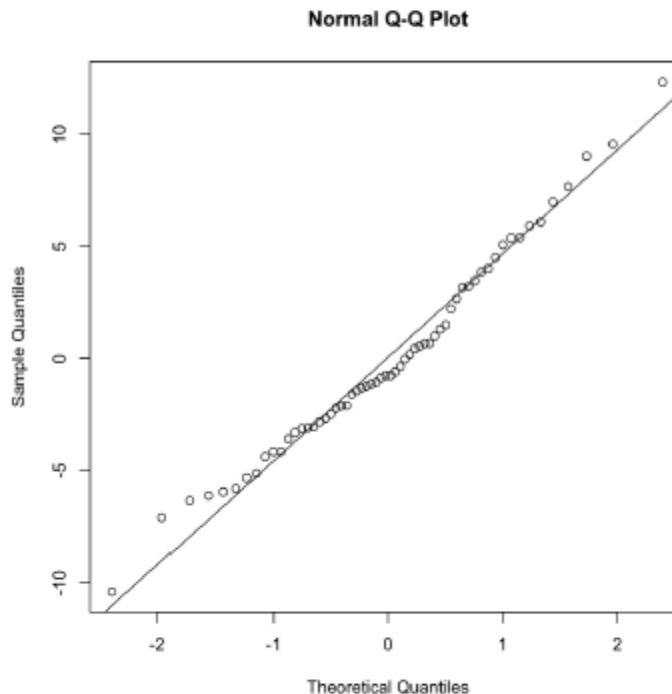
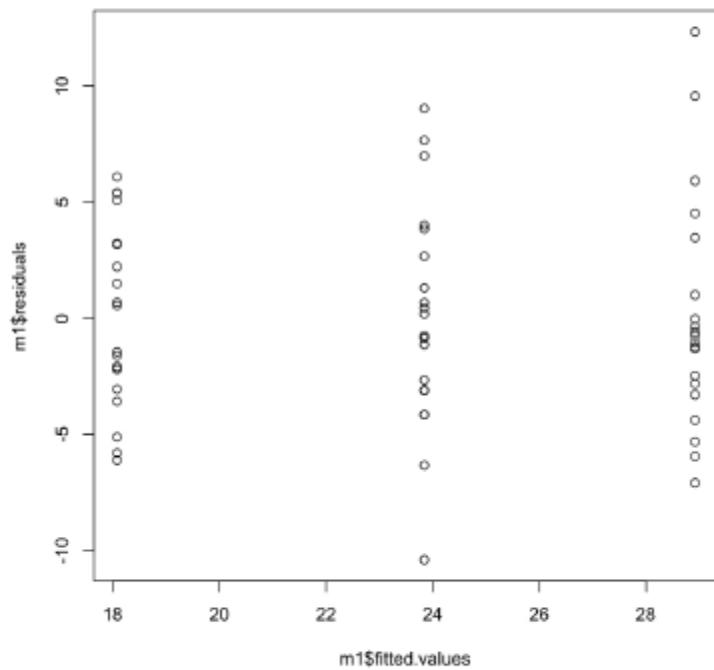
- Conclude that at least one punch type is different from the others but which?

```
> summary(m1)
      DF Sum Sq Mean Sq F value    Pr(>F)
factor(punch_type)  2   1169   584.6   27.67 4.01e-09 ***
Residuals         57   1204    21.1
---
Signif. codes:  0 '****' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
> model.tables(m1)
Tables of effects

  factor(punch_type)
  factor(punch_type)
    1      2      4
-5.523  0.241  5.282
```

HBO Boxing Data

- Should check the conditions.



Multiple Comparisons: Approach I

- Find confidence intervals for all possible differences

$$\mu_1 - \mu_2$$

$$\mu_1 - \mu_3$$

$$\mu_2 - \mu_3$$

- A 95% CI for the difference of two means:

$$(\bar{y}_1 - \bar{y}_2) \pm t_{n_1+n_2-2} S_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$

Multiple Comparisons: Approach I

$$\mu_1 - \mu_2 \quad (-8.57, -2.96)$$

$$\mu_1 - \mu_3 \quad (-13.68, -7.93)$$

$$\mu_2 - \mu_3 \quad (-8.19, -1.89)$$

- Conclusion: All three punches types have different mean punch speeds with uppercuts being fastest and crosses being slowest.

Remembering Hypothesis Testing

- Possible decisions in a hypothesis test are:

		DECISION
NULL Hypothesis	ACCEPT H_0	REJECT H_0
TRUE	<i>Correct Decision</i>	<i>Type I Error = α</i>
FALSE	<i>Type II Error = β</i>	<i>Correct Decision</i>

- Type I error: Reject the null hypothesis when in fact it is true
- Type II error: Fail to reject the null hypothesis when in fact it is false

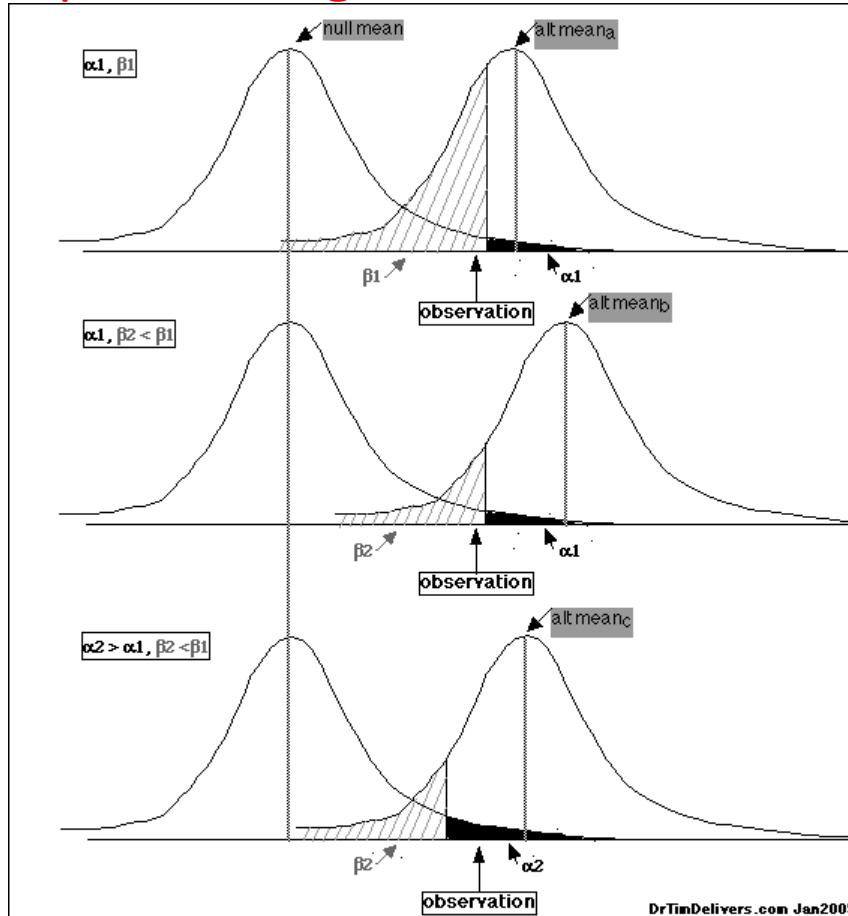
The probability of making a Type I error, α , is chosen by the researcher *before* the sample data is collected.

The **level of significance**, α , is the probability of making a Type I error.

“In Other Words”

As the probability of a Type I error increases, the probability of a Type II error decreases, and vice-versa.

They are NOT complementing of each other



DrTimDelivers.com Jan2005

FIG A Illustration choosing alpha, beta, alternative mean, variances(2) & shapes(2)
6 of the 7 are independent (choosing any 6 determines 7th), 3 if normal curves (fits 2x2)
($1-\alpha$) probability observation belongs to null distribution (α_1, α_2)
($1-\beta$) probability observation belongs to alternative distribution (β_1, β_2)
with error levels α_n for not null and β_n for not alternative
[shown for equal variance and same shape for each distribution]

The complement of β , $1 - \beta$, is the probability of rejecting the null hypothesis when the alternative hypothesis is true. The value of $1 - \beta$ is referred to as the **power of the test**. The higher the power of the test, the more likely the test will reject the null when the alternative hypothesis is true.

The power of a test is the probability that it correctly rejects a false null hypothesis.

Experiment-Wise Error

- Also, the probability that at least one interval will fail increases as the number of groups increases, almost to a certainty

α_f : the experiment-wise error

Number of Treatments	Number of Comparisons	α	α_f
2	1	0.050	0.050
3	3	0.050	0.143
5	10	0.050	0.401
10	45	0.050	0.901
15	105	0.050	0.995

- This seems like too much error in reporting our results. Some 'control' for this is needed.

Multiple Comparisons Methods

Least significance difference (LSD)

Fisher's protected LSD

Bonferroni method

Duncan's multiple range test

Scheffe method

Dunn-Sidak bound

Newman-Keuls's multiple range test

Student-Newman-Keuls method

Tukey's honestly significant (or significantly) difference (HSD)

Studentized range statistic

Ryan-Einot-Gabriel-Welsch method

Dunnett's method

etc...

1 Why is multiple testing a problem?

Say you have a set of hypotheses that you wish to test simultaneously. The first idea that might come to mind is to test each hypothesis separately, using some level of significance α . At first blush, this doesn't seem like a bad idea. However, consider a case where you have 20 hypotheses to test, and a significance level of 0.05. What's the probability of observing at least one significant result just due to chance?

$$\begin{aligned}\mathbb{P}(\text{at least one significant result}) &= 1 - \mathbb{P}(\text{no significant results}) \\ &= 1 - (1 - 0.05)^{20} \\ &\approx 0.64\end{aligned}$$

So, with 20 tests being considered, we have a 64% chance of observing at least one significant result, even if all of the tests are actually not significant. In genomics and other biology-related fields, it's not unusual for the number of simultaneous tests to be quite a bit larger than 20... and the probability of getting a significant result simply due to chance keeps going up.

Methods for dealing with multiple testing frequently call for adjusting α in some way, so that the probability of observing at least one significant result due to chance remains below your desired significance level.

Bonferroni Correction

- Simplest, most conservative method
- Want the overall confidence level for all the tests to remain at 95%
- If you are doing k comparisons (in our example $k = 3$) instead of using k separate intervals at level

$$1 - \alpha$$

Use level

$$1 - \frac{\alpha}{k}$$

What Is the Bonferroni Correction?

Matthew A. Napierala, MD

The Bonferroni correction is an adjustment made to P values when several dependent or independent statistical tests are being performed simultaneously on a single data set. To perform a Bonferroni correction, divide the critical P value (α) by the number of comparisons being made. For example, if 10 hypotheses are being tested, the new critical P value would be $\alpha/10$. The statistical power of the study is then calculated based on this modified P value.

The Bonferroni correction is used to reduce the chances of obtaining false-positive results (type I errors) when multiple pair wise tests are performed on a single set of data. Put simply, the probability of identifying at least one significant result due to chance increases as more hypotheses are tested.

For example, a researcher is testing 20 hypotheses simultaneously, with a critical P value of 0.05. In this case, the following would be true:

- $P(\text{at least one significant result}) = 1 - P(\text{no significant results})$
- $P(\text{at least one significant result}) = 1 - (1-0.05)^{20}$
- $P(\text{at least one significant result}) = 0.64$

Thus, performing 20 tests on a data set yields a 64 percent chance of identifying at least one significant result, even if all of the tests are actually not significant. Therefore, while a given α may be appropriate for each individual comparison, it may not be appropriate for the set of all comparisons.

This fact is potentially problematic because in contemporary orthopaedic research studies, numerous simultaneous tests are routinely performed. Thus, to avoid a large number of spurious positives, α must be lowered to account for the number of comparisons being performed.

The Bonferroni correction is based on the idea that if an experimenter is testing n dependent or independent hypotheses on a set of data, the probability of type I error is offset by testing each hypothesis at a statistical significance level $1/n$ times what it would be if only one hypothesis were tested.

An example of the use of the Bonferroni correction in the orthopaedic literature can be found in a recent article in *The Journal of Bone & Joint Surgery-American*, "The Relationship Between Time to Surgical Débridement and Incidence of Infection After Open High-Energy Lower Extremity Trauma." The purpose of the study was to evaluate the relationship between the timing of the initial treatment of open fractures and the development of subsequent infection.

The study consisted of 315 patients identified as a subgroup of the Lower Extremity Assessment Project (LEAP). For their analysis, the investigators created two outcome measures ("All Infections" and "Major Infections") and used four time-to-treatment variables. In their manuscript, the authors stated the following: "Because two outcome measures were tested against four hypothesized predictors, a Bonferroni-adjusted significance level of 0.00625 was calculated to account for the increased possibility of type-I error." In other words, because 8 hypotheses were being tested on one set of data, the chance of obtaining a false-positive result was 34 percent. Accordingly, the authors used the Bonferroni correction to adjust the P value for each hypothesis to 0.00625 to neutralize this risk.

However, although the Bonferroni correction controls for false positives, it can become very conservative as the number of tests increases. This, in turn, increases the risk of generating false negatives (type II errors).

In sum, the risk of making erroneous false-positive conclusions is increased when testing multiple hypotheses on a single set of data. This fact is often underappreciated by investigators and consumers of orthopaedic literature. The Bonferroni correction is a simple statistical method for mitigating this risk, and its appropriate use can ensure the integrity of studies in which a large number of significance tests are used. Other tests that also control for false positives, without the risk of increasing false negatives, are the Tukey and Dunnett's tests.

Bonferroni Correction

$$\mu_1 - \mu_2 \quad (-8.57, -2.96) \qquad \qquad (-9.23, -2.29)$$

$$\mu_1 - \mu_3 \quad (-13.68, -7.93) \qquad \qquad (-14.36, -7.25)$$

$$\mu_2 - \mu_3 \quad (-8.19, -1.89) \qquad \qquad (-8.93, -1.15)$$

- Confidence level for the 3 intervals on the right is 98.33%. The intervals become wider.
- In extreme cases we can have a significant treatment effect (by F-test) but after this correction no treatment means are different

Tukey HSD

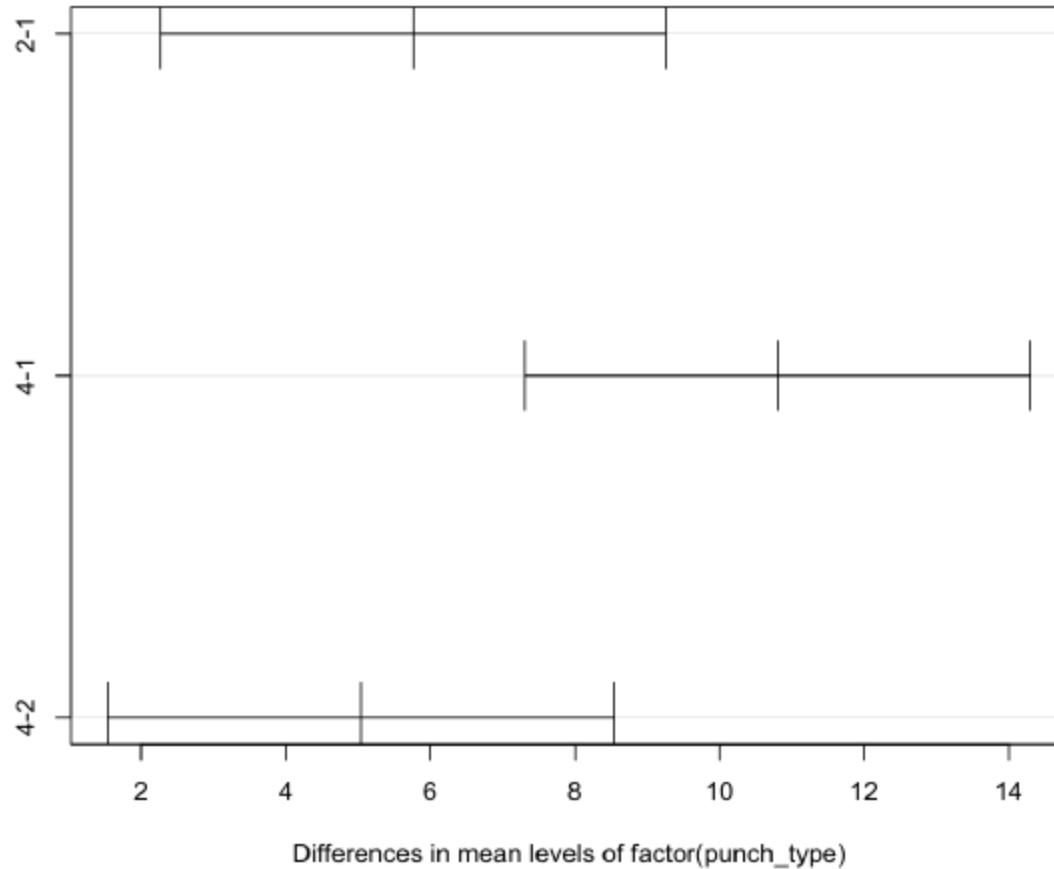
```
> TukeyHSD(m1)
  Tukey multiple comparisons of means
    95% family-wise confidence level

Fit: aov(formula = speed ~ factor(punch_type), data = punches)

$`factor(punch_type)`
      diff      lwr      upr      p adj
2-1  5.763355  2.265334  9.261375 0.0005994
4-1 10.804850  7.306829 14.302870 0.0000000
4-2  5.041495  1.543475  8.539516 0.0028449
```

Tukey HSD

95% family-wise confidence level



Tukey HSD

formula:

$$\frac{M_1 - M_2}{\sqrt{MS_w \left(\frac{1}{n} \right)}} \quad \begin{aligned} M &= \text{treatment/group} \\ &\text{mean} \\ n &= \text{number per} \\ &\text{treatment/group} \end{aligned}$$

Steps

1. Calculate an analysis of variance (e.g., One-way between-subjects ANOVA).
2. Select two means and note the relevant variables (Means, Mean Square Within, and number per condition/group)
3. Calculate Tukey's test for each mean comparison
4. Check to see if Tukey's score is statistically significant with Tukey's probability/critical value table taking into account appropriate df_{within} and number of treatments.

Problem: Susan Sound predicts that students will learn most effectively with a constant background sound, as opposed to an unpredictable sound or no sound at all. She randomly divides twenty-four students into three groups of eight. All students study a passage of text for 30 minutes. Those in group 1 study with background sound at a constant volume in the background. Those in group 2 study with noise that changes volume periodically. Those in group 3 study with no sound at all. After studying, all students take a 10 point multiple choice test over the material. She begins by conducting a [One-way, between-subjects Analysis of Variance](#). She finds a significant F score. The relevant variables from her ANOVA table are:

$$MS_{\text{within}} = 4.18; M_1 = 6; M_2 = 4; M_3 = 3; df_{\text{within}} = 21; n = 8$$

$$M_{1 \text{ vs } M_2} = \frac{6 - 4}{\sqrt{4.18 \left(\frac{1}{8} \right)}} = \frac{2}{\sqrt{4.18 * .125}} = \frac{2}{.72} = 2.767$$

$$M_{1 \text{ vs } M_3} = \frac{6 - 3}{.72} = 4.15^*$$

$$M_{2 \text{ vs } M_3} = \frac{4 - 3}{.72} = 1.38$$

*****(according to the [Tukey's sig/probability table](#), taking into account ($df_{\text{within}} = 21$ and treatments = 3), the mean comparison between means 1 and 3 is statistically significant, but not the other comparisons).

Interpretation: Susan's hypothesis was only partially supported in that those who studied with a constant noise did perform significantly better on the test than those who studied without any noise, but they did not perform significantly better than those who studied with random noise.

The Completely Randomized Single-Factor Experiment

An Example

A manufacturer of paper used for making grocery bags is interested in improving the tensile strength of the product. Product engineering thinks that tensile strength is a function of the hardwood concentration in the pulp and that the range of hardwood concentrations of practical interest is between 5 and 20%. A team of engineers responsible for the study decides to investigate four levels of hardwood concentration: 5%, 10%, 15%, and 20%. They decide to make up six test specimens at each concentration level, using a pilot plant. All 24 specimens are tested on a laboratory tensile tester, in random order. The data from this experiment are shown in Table 13-1.

The Completely Randomized Single-Factor Experiment

An Example

Table 13-1 Tensile Strength of Paper (psi)

Hardwood Concentration (%)	Observations						Totals	Averages
	1	2	3	4	5	6		
5	7	8	15	11	9	10	60	10.00
10	12	17	13	18	19	15	94	15.67
15	14	18	19	17	16	18	102	17.00
20	19	25	22	23	18	20	127	21.17
							383	15.96

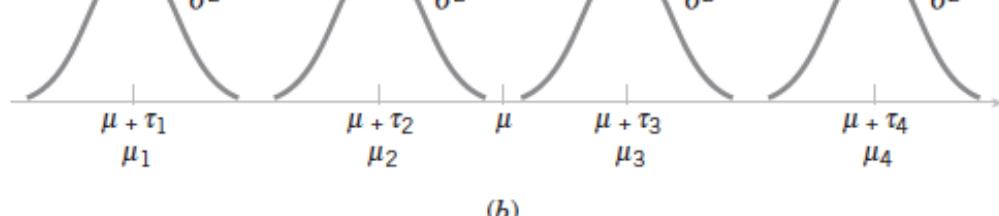
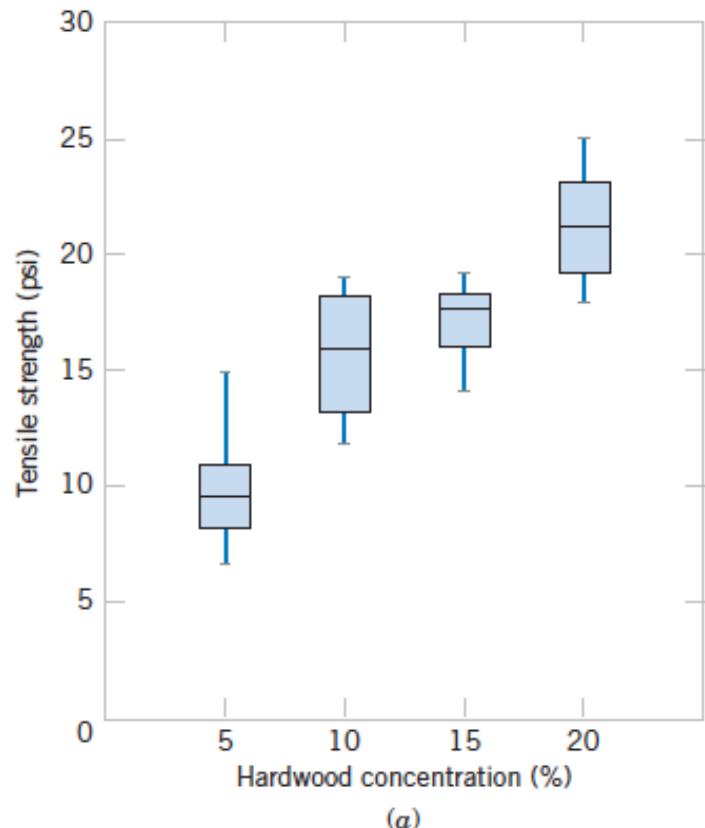
The Completely Randomized Single-Factor Experiment

An Example

- The levels of the factor are sometimes called **treatments**.
- Each treatment has six observations or **replicates**.
- The runs are run in **random** order.

The Completely Randomized Single-Factor Experiment

An Example



(b)

(a) Box plots of hardwood concentration data. (b) Display of the model in Equation 13-1 for the completely randomized single-factor experiment

The Completely Randomized Single-Factor Experiment

The Analysis of Variance

Suppose there are a different levels of a single factor that we wish to compare. The levels are sometimes called **treatments**.

Table 13-2 Typical Data for a Single-Factor Experiment

Treatment	Observations				Totals	Averages
1	y_{11}	y_{12}	\dots	y_{1n}	$y_{1\cdot}$	$\bar{y}_{1\cdot}$
2	y_{21}	y_{22}	\dots	y_{2n}	$y_{2\cdot}$	$\bar{y}_{2\cdot}$
\vdots	\vdots	\vdots	$\vdots \vdots \vdots$	\vdots	\vdots	\vdots
a	y_{a1}	y_{a2}	\dots	y_{an}	$y_{a\cdot}$	$\bar{y}_{a\cdot}$
					$y_{..}$	$\bar{y}_{..}$

The Completely Randomized Single-Factor Experiment

The Analysis of Variance

We may describe the observations in Table 13-2 by the linear statistical model:

$$Y_{ij} = \mu + \tau_i + \epsilon_{ij} \begin{cases} i = 1, 2, \dots, a \\ j = 1, 2, \dots, n \end{cases} \quad (13-1)$$

The model could be written as

$$Y_{ij} = \mu_i + \epsilon_{ij} \begin{cases} i = 1, 2, \dots, a \\ j = 1, 2, \dots, n \end{cases}$$

The Completely Randomized Single-Factor Experiment

The Analysis of Variance

Fixed-effects Model

The treatment effects are usually defined as deviations from the overall mean so that:

$$\sum_{i=1}^a \tau_i = 0$$

Also,

$$y_{i\cdot} = \sum_{j=1}^n y_{ij} \quad \bar{y}_{i\cdot} = y_{i\cdot}/n \quad i = 1, 2, \dots, a$$

$$y_{..} = \sum_{i=1}^a \sum_{j=1}^n y_{ij} \quad \bar{y}_{..} = y_{..}/N$$

The Completely Randomized Single-Factor Experiment

The Analysis of Variance

We wish to test the hypotheses:

$$H_0: \tau_1 = \tau_2 = \cdots = \tau_a = 0$$

$$H_1: \tau_i \neq 0 \quad \text{for at least one } i$$

The analysis of variance partitions the total variability into two parts.

The Analysis of Variance

Definition

The sum of squares identity is

$$\sum_{i=1}^a \sum_{j=1}^n (y_{ij} - \bar{y}_{..})^2 = n \sum_{i=1}^a (\bar{y}_{i\cdot} - \bar{y}_{..})^2 + \sum_{i=1}^a \sum_{j=1}^n (y_{ij} - \bar{y}_{i\cdot})^2 \quad (13-5)$$

or symbolically

$$SS_T = SS_{\text{Treatments}} + SS_E \quad (13-6)$$

The Analysis of Variance

The expected value of the treatment sum of squares is

$$E(SS_{\text{Treatments}}) = (a - 1)\sigma^2 + n \sum_{i=1}^a \tau_i^2$$

and the expected value of the error sum of squares is

$$E(SS_E) = a(n - 1)\sigma^2$$

The ratio $MS_{\text{Treatments}} = SS_{\text{Treatments}}/(a - 1)$ is called the **mean square for treatments**.

The Analysis of Variance

The appropriate test statistic is

$$F_0 = \frac{SS_{\text{Treatments}}/(a - 1)}{SS_E/[a(n - 1)]} = \frac{MS_{\text{Treatments}}}{MS_E} \quad (13-7)$$

We would reject H_0 if $f_0 > f_{\alpha, a-1, a(n-1)}$

The Analysis of Variance

Definition

The sums of squares computing formulas for the ANOVA with equal sample sizes in each treatment are

$$SS_T = \sum_{i=1}^a \sum_{j=1}^n y_{ij}^2 - \frac{y_{..}^2}{N} \quad (13-8)$$

and

$$SS_{\text{Treatments}} = \sum_{i=1}^a \frac{y_i^2}{n} - \frac{y_{..}^2}{N} \quad (13-9)$$

The error sum of squares is obtained by subtraction as

$$SS_E = SS_T - SS_{\text{Treatments}} \quad (13-10)$$

The Completely Randomized Single-Factor Experiment

The Analysis of Variance

Analysis of Variance Table

Table 13-3 The Analysis of Variance for a Single-Factor Experiment, Fixed-Effects Model

Source of Variation	Sum of Squares	Degrees of Freedom	Mean Square	F_0
Treatments	$SS_{\text{Treatments}}$	$a - 1$	$MS_{\text{Treatments}}$	$\frac{MS_{\text{Treatments}}}{MS_E}$
Error	SS_E	$a(n - 1)$	MS_E	
Total	SS_T	$an - 1$		

Example

EXAMPLE 13-1 Tensile Strength ANOVA

Consider the paper tensile strength experiment described in Section 13-2.1. This experiment is a CRD. We can use the analysis of variance to test the hypothesis that different hardwood concentrations do not affect the mean tensile strength of the paper.

The hypotheses are

$$H_0: \tau_1 = \tau_2 = \tau_3 = \tau_4 = 0$$

$$H_1: \tau_i \neq 0 \text{ for at least one } i$$

The Completely Randomized Single-Factor Experiment

Example

We will use $\alpha = 0.01$. The sums of squares for the analysis of variance are computed from Equations 13-8, 13-9, and 13-10 as follows:

$$\begin{aligned} SS_T &= \sum_{i=1}^4 \sum_{j=1}^6 y_{ij}^2 - \frac{y_{..}^2}{N} \\ &= (7)^2 + (8)^2 + \cdots + (20)^2 - \frac{(383)^2}{24} = 512.96 \end{aligned}$$

$$\begin{aligned} SS_{\text{Treatments}} &= \sum_{i=1}^4 \frac{y_{i..}^2}{n} - \frac{y_{..}^2}{N} \\ &= \frac{(60)^2 + (94)^2 + (102)^2 + (127)^2}{6} - \frac{(383)^2}{24} \\ &= 382.79 \end{aligned}$$

$$\begin{aligned} SS_E &= SS_T - SS_{\text{Treatments}} \\ &= 512.96 - 382.79 = 130.17 \end{aligned}$$

The Completely Randomized Single-Factor Experiment

Example

The ANOVA is summarized in Table 13-4. Since $f_{0.01,3,20} = 4.94$, we reject H_0 and conclude that hardwood concentration in the pulp significantly affects the mean strength of the paper. We can also find a P -value for this test statistic as follows:

$$P = P(F_{3,20} > 19.60) \approx 3.59 \times 10^{-6}$$

Since $P \approx 3.59 \times 10^{-6}$ is considerably smaller than $\alpha = 0.01$, we have strong evidence to conclude that H_0 is not true.

Table 13-4 ANOVA for the Tensile Strength Data

Source of Variation	Sum of Squares	Degrees of Freedom	Mean Square	f_0	P -value
Hardwood concentration	382.79	3	127.60	19.60	3.59 E-6
Error	130.17	20	6.51		
Total	512.96	23			

Table 13-5 Minitab Analysis of Variance Output for Example 13-1

One-Way ANOVA: Strength versus CONC

Analysis of Variance for Strength

Source	DF	SS	MS	F	P
Conc	3	382.79	127.60	19.61	0.000
Error	20	130.17	6.51		
Total	23	512.96			
Individual 95% CIs For Mean Based on Pooled StDev					
Level	N	Mean	StDev	— + — + — + — + —	
5	6	10.000	2.828	(—*—)	
10	6	15.667	2.805	(—*—)	
15	6	17.000	1.789	(—*—)	
20	6	21.167	2.639	(—*—)	
— + — + — + — + —					
Pooled StDev = 2.551				10.0 15.0 20.0 25.0	

Fisher's pairwise comparisons

Family error rate = 0.192

Individual error rate = 0.0500

Critical value = 2.086

Intervals for (column level mean) – (row level mean)

	5	10	15
10	-8.739		
	-2.594		
15	-10.072	-4.406	
	-3.928	1.739	
20	-14.239	-8.572	-7.239
	-8.094	-2.428	-1.094

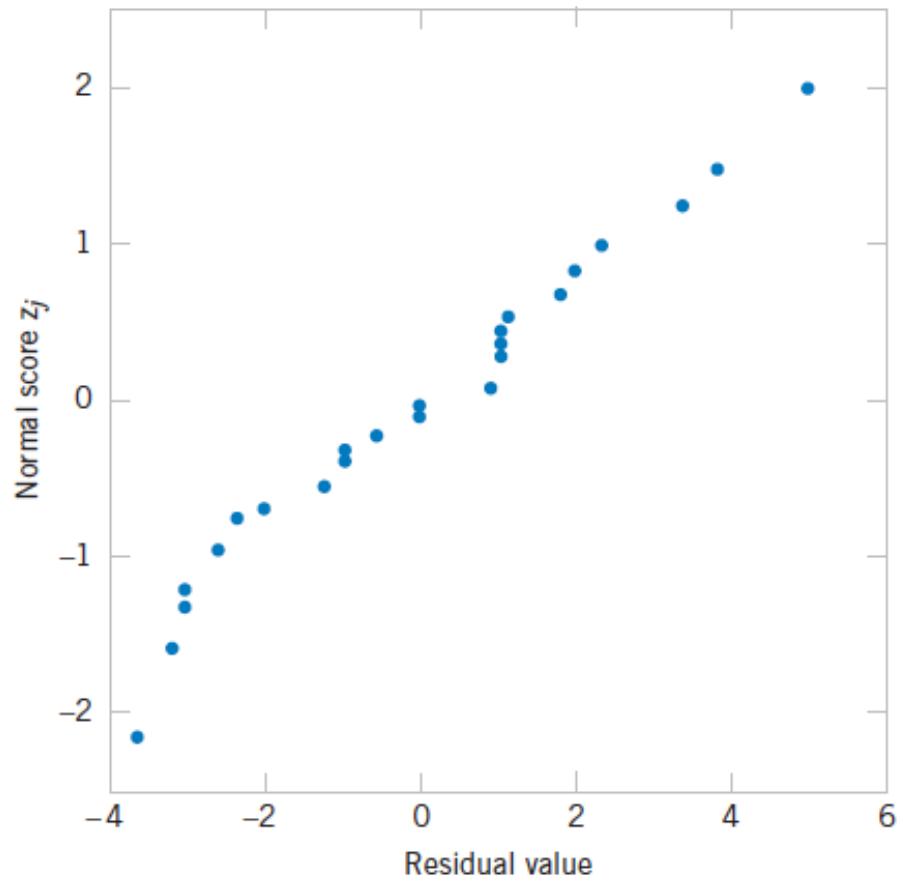
Residual Analysis and Model Checking

Table 13-6 Residuals for the Tensile Strength Experiment

Hardwood Concentration (%)	Residuals					
5	-3.00	-2.00	5.00	1.00	-1.00	0.00
10	-3.67	1.33	-2.67	2.33	3.33	-0.67
15	-3.00	1.00	2.00	0.00	-1.00	1.00
20	-2.17	3.83	0.83	1.83	-3.17	-1.17

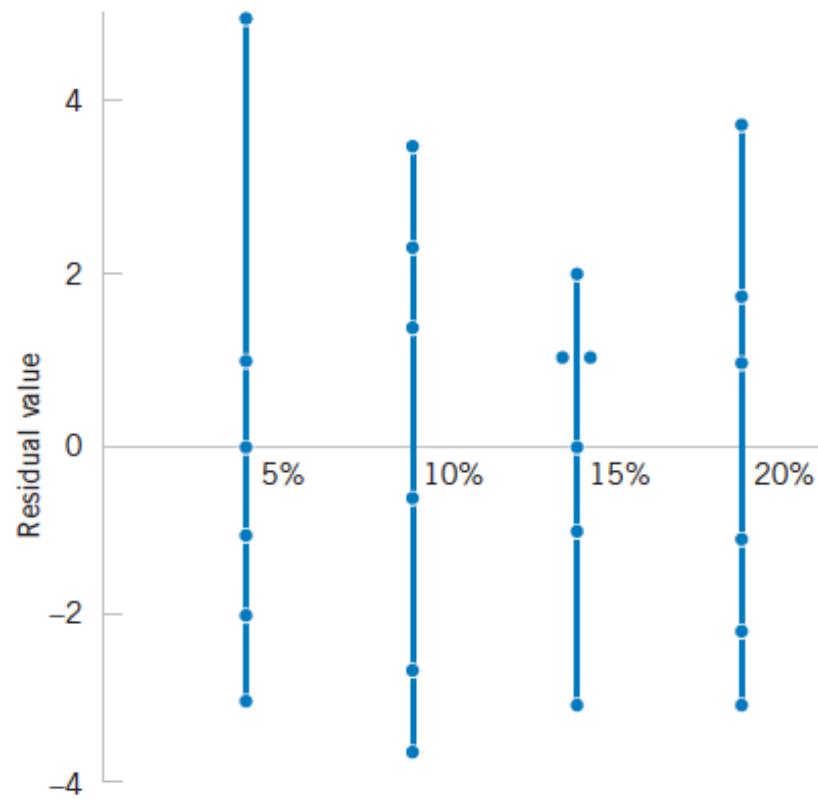
Residual Analysis and Model Checking

Normal probability plot of residuals from the hardwood concentration experiment.



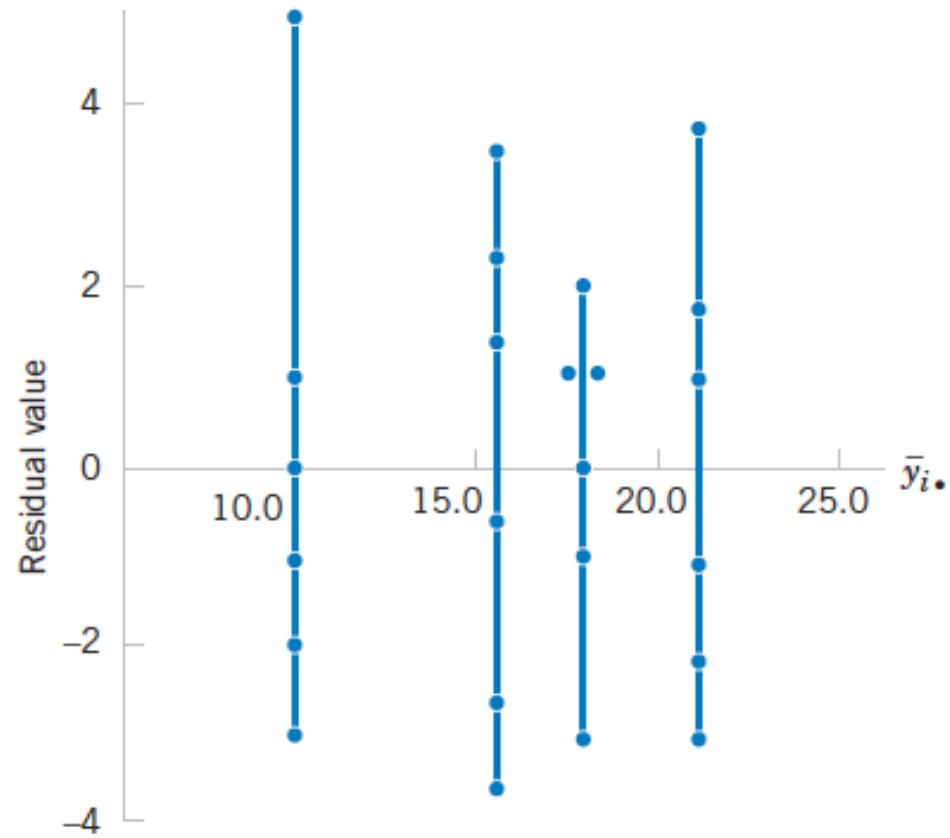
Residual Analysis and Model Checking

Plot of residuals versus factor levels
(hardwood concentration).



Residual Analysis and Model Checking

Plot of residuals versus \bar{y}_i



Required Conditions

- Objects/subjects randomly assigned to treatment levels.
- Independent observations
- Errors are normally distributed in the population
- Amount of variation is same for each treatment level

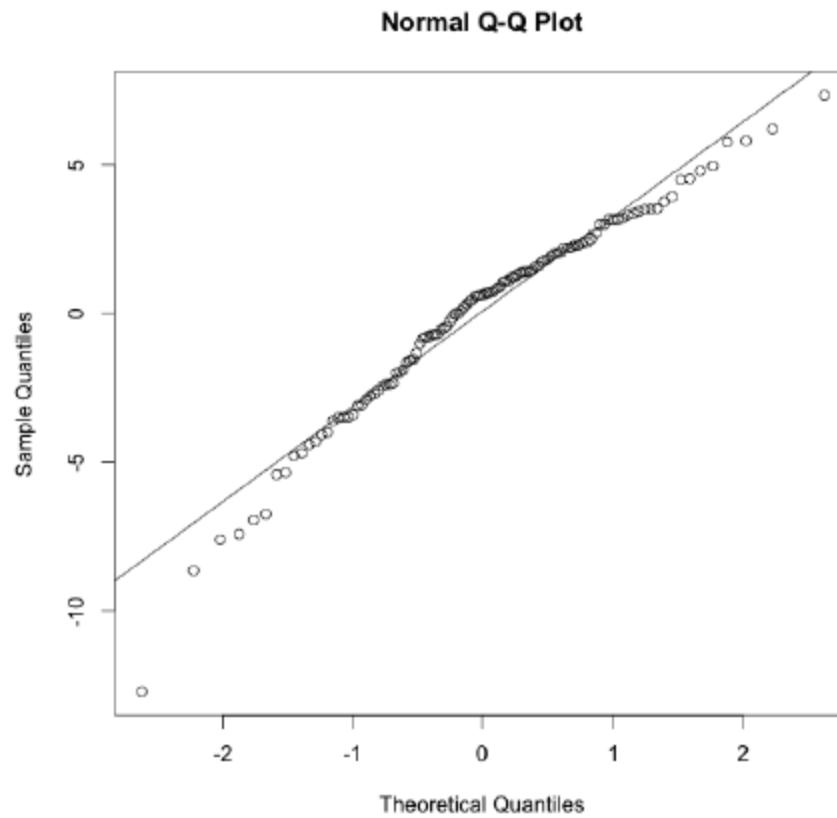
Checking Conditions

- Some of these conditions can be checked by examining the residuals.
- `aov()` will calculate these automatically
 - `fitted.values` contains the predicted values.
 - `residuals` contains the residuals

Diagnostic: Normality Conditions

- `qqnorm()`
- `qqline()`
- This plots the residuals against their expected values if their distribution were truly normal
- Thus if they are truly from a Normal population this plot will be roughly a straight line

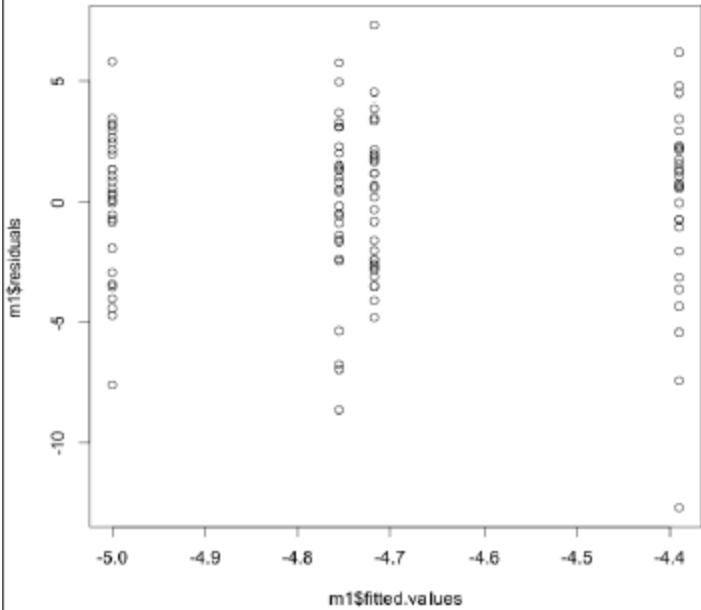
qq plots



Model Specification

- If the shape of the model correctly describes the signal, then the residuals should have no structure

$$\epsilon_{ij} \sim N(0, \sigma)$$



- In each group the residuals should be centered at about 0
- In each group, standard deviation should be about the same

Non-constant Variance

- The major problem occurs when the variance/SD increases steadily with the fitted (decreasing also a problem)
- For balanced designs (equal sample sizes in each treatment level) non-constant variance not a big problem

Other Conditions

- If you know the order in which the data were recorded, plotting the residuals against the order can tell you if there is time-dependence within observation
- This is a serious violation of the independence condition and can invalidate the results of the study
- Ideally, then, your plot will show no trend or structure

You also need to check for leverage points and outliers

Observational version of BF design

Sampling instead of assigning

Example 5.2:

Intravenous Fluids:

Conditions: Three Drug Companies

Material: Six different samples of intravenous fluid from each drug company

Response: Number of contaminated particles of a certain size

Calculation Example

Leafhopper Data

Chapter 5 Diet Example and Particles Count Example

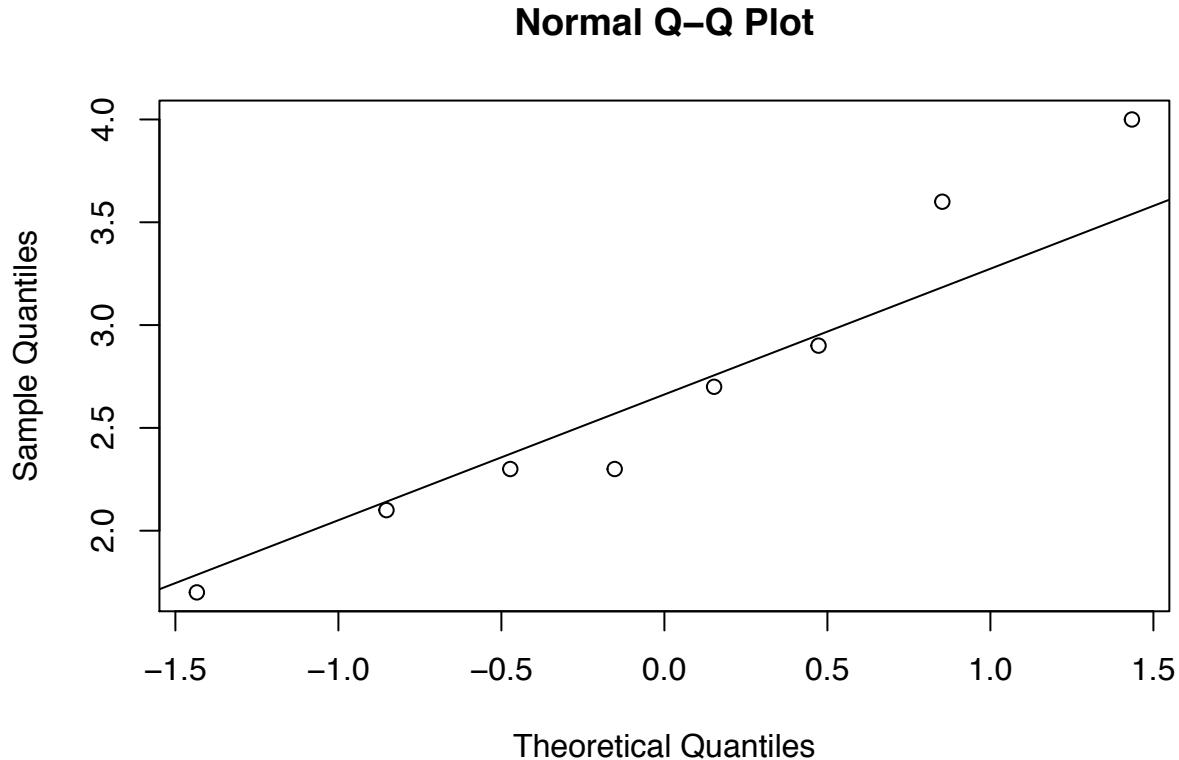
Akram Almohalwas

January 14, 2016

This is an R Markdown document. Markdown is a simple formatting syntax for authoring HTML, PDF, and MS Word documents. For more details on using R Markdown see <http://rmarkdown.rstudio.com>.

When you click the **Knit** button a document will be generated that includes both content as well as the output of any embedded R code chunks within the document. You can embed an R code chunk like this:

```
Diet = c(2.3, 1.7, 4, 3.6, 2.9, 2.7, 2.1, 2.3)
qqnorm(Diet)
qqline(Diet)
```



```
summary(Diet)
```

```
##      Min. 1st Qu. Median      Mean 3rd Qu.      Max.
##    1.700   2.250   2.500   2.700   3.075   4.000
```

```
var(Diet)
```

```
## [1] 0.6028571
```

```

# To get the SS total like we used to do for one-way Anova
ss_t<-var(Diet)*(length(Diet)-1)
ss_t

## [1] 4.22

# In oder to get the Anova summary Example 5.16 page 176 in your textbook
# We do the following:
# First get the sum of all scores squared and that is denoted by Total
# on page 176
sum(Diet^2)

## [1] 62.54

# Since the over all average for data is 2.7, we create a benchmark with 2.7
# for all cells
grand<-c(2.7,2.7,2.7,2.7,2.7,2.7,2.7,2.7)
# Sum of the squares of grand gives Grand Average
sum(grand^2)

## [1] 58.32

# If we subtract grand average from the average per group we get the diet effect
effect<-c(-0.7,-0.7,1.1,1.1,0.1,0.1,-0.5,-0.5)
# sum of the squares of effect gives SS(Diet) in a regular One-Way Anova
var(effect)*7

## [1] 3.92

sum(effect^2)

## [1] 3.92

# In order to get the Residual we subtract the grand average as well as the per
# group average from the observed value
res<-Diet-grand-effect
# sum all the squares of the residuals gives SSE
sum(res^2)

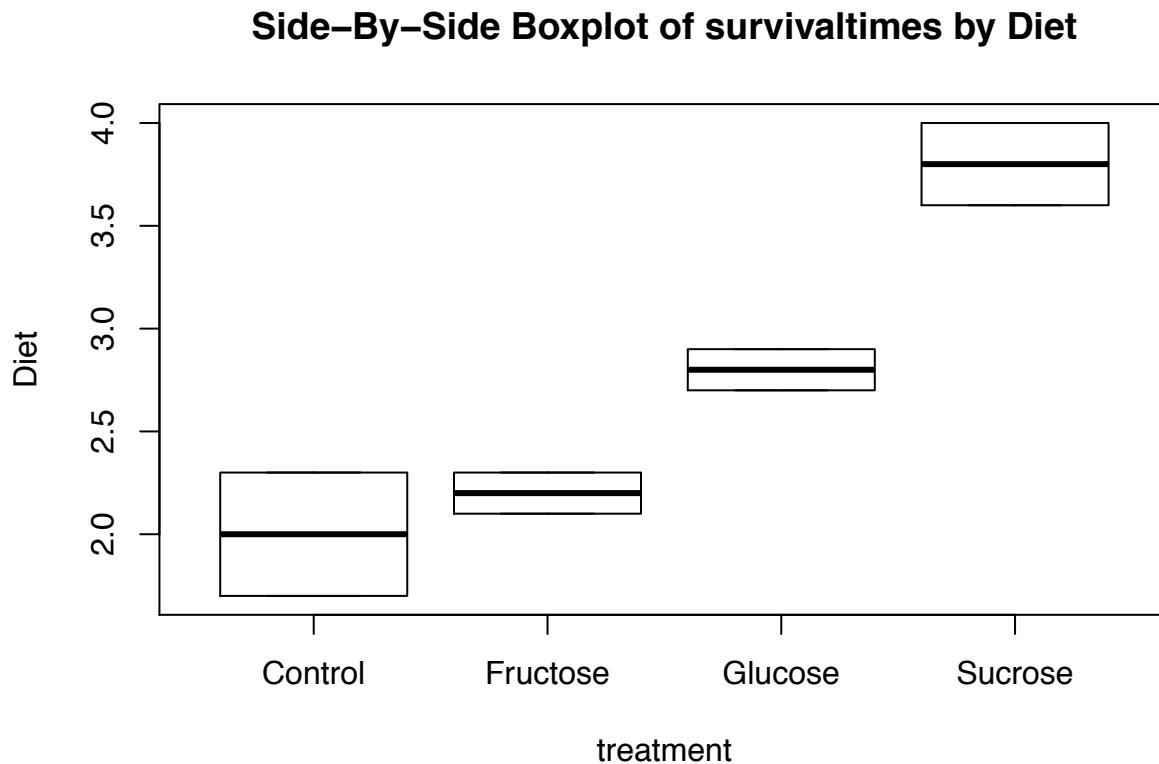
## [1] 0.3

treatment= c(rep("Control",2), rep("Sucrose",2), rep("Glucose",2),rep("Fructose",2))
data1 = data.frame(Diet,treatment)
Diet

## [1] 2.3 1.7 4.0 3.6 2.9 2.7 2.1 2.3

```

```
# Here we create boxplots for the data side-by-side
plot(Diet ~ treatment, data=data1, main="Side-By-Side Boxplot of survivaltimes by Diet")
```



```
mall<-mean(Diet)
mall
```

```
## [1] 2.7
```

```
ma<-mean(Diet[treatment=="Control"])
ma
```

```
## [1] 2
```

```
mb<-mean(Diet[treatment=="Sucrose"])
mb
```

```
## [1] 3.8
```

```
mc<-mean(Diet[treatment=="Glucose"])
mc
```

```
## [1] 2.8
```

```

md<-mean(Diet[treatment=="Fructose"])
md

## [1] 2.2

means<-c(mall,ma(mb,mc,md)
tmeans<-c(ma,mb,mc,md)
mean(tmeans)

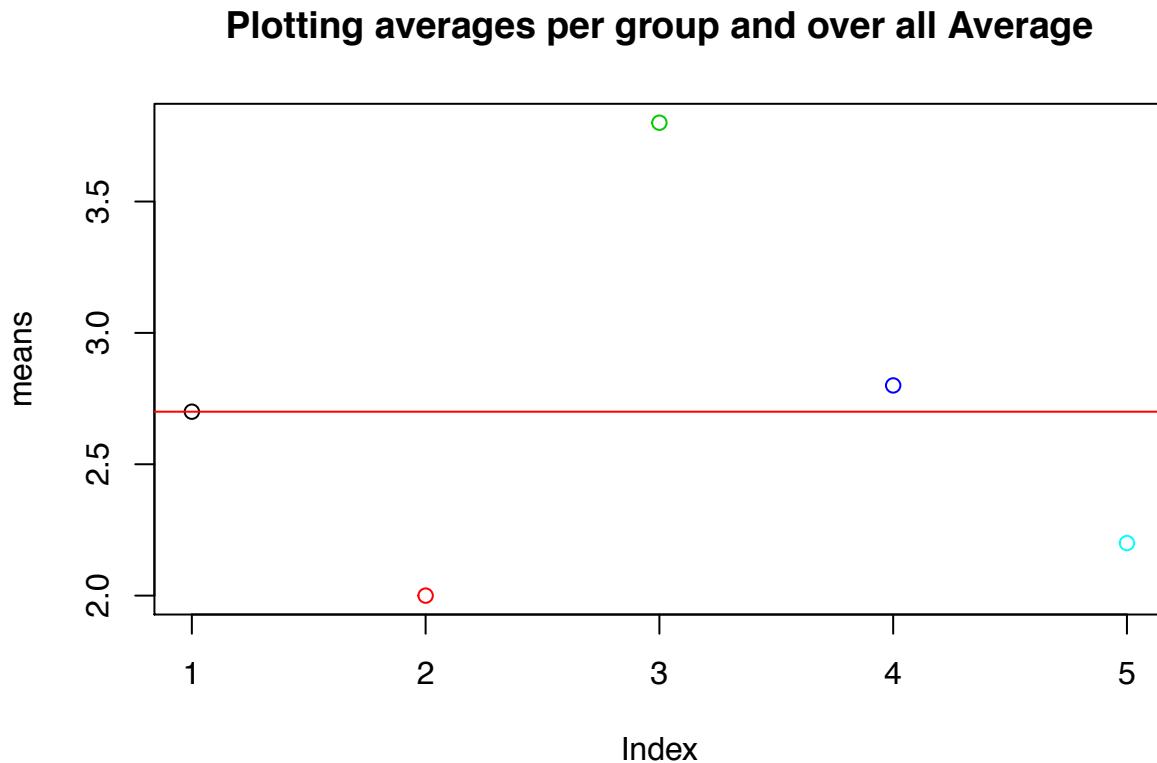
## [1] 2.7

var(tmeans)

## [1] 0.6533333

plot(means,col=1:8,main="Plotting averages per group and over all Average")
abline(h=mall,col="red")

```



```

results <- aov(Diet ~ treatment, data=data1)
summary(results)

```

```

##          Df Sum Sq Mean Sq F value    Pr(>F)

```

```

## treatment      3    3.92    1.307   17.42 0.00925 **
## Residuals     4    0.30    0.075
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

results$residuals

##    1    2    3    4    5    6    7    8
##  0.3 -0.3  0.2 -0.2  0.1 -0.1 -0.1  0.1

sum(Diet^2)

## [1] 62.54

reg<-lm(Diet~treatment)
reg

## 
## Call:
## lm(formula = Diet ~ treatment)
## 
## Coefficients:
##             (Intercept)  treatmentFructose  treatmentGlucose
##                   2.0                  0.2                  0.8
## treatmentSucrose
##                   1.8

summary(reg)

## 
## Call:
## lm(formula = Diet ~ treatment)
## 
## Residuals:
##    1    2    3    4    5    6    7    8
##  0.3 -0.3  0.2 -0.2  0.1 -0.1 -0.1  0.1
## 
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)    
## (Intercept) 2.0000    0.1936 10.328 0.000496 ***
## treatmentFructose 0.2000    0.2739  0.730 0.505681  
## treatmentGlucose 0.8000    0.2739  2.921 0.043192 *  
## treatmentSucrose 1.8000    0.2739  6.573 0.002773 ** 
## ---      
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 0.2739 on 4 degrees of freedom
## Multiple R-squared:  0.9289, Adjusted R-squared:  0.8756 
## F-statistic: 17.42 on 3 and 4 DF,  p-value: 0.009248

```

```
# All pair-wise comparisons between groups
pairwise.t.test(Diet, treatment, p.adjust="bonferroni")
```

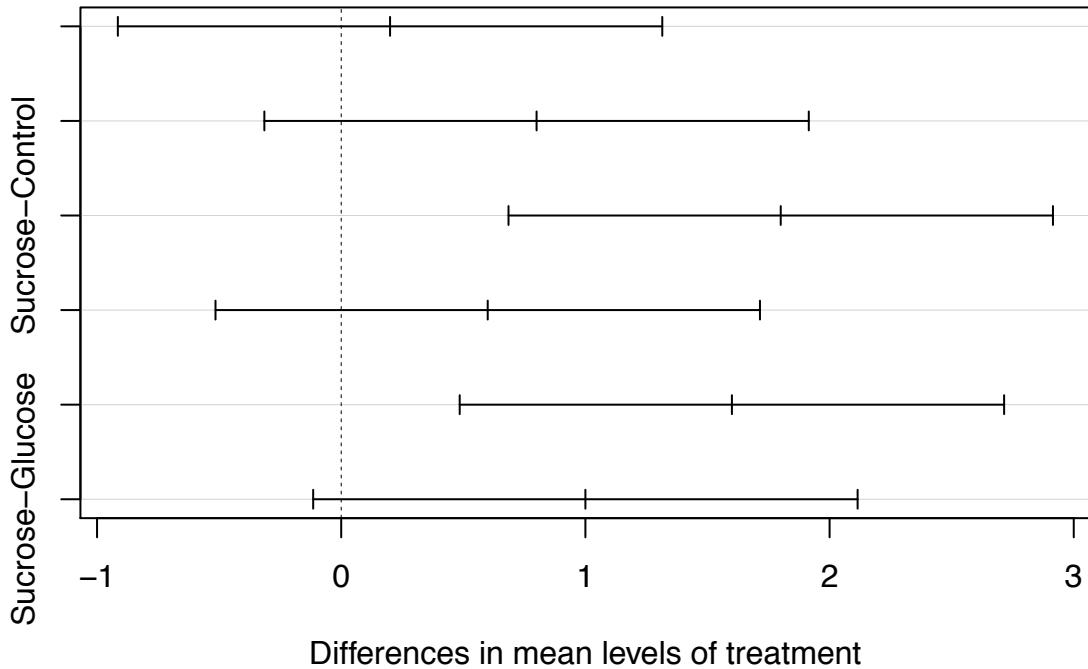
```
##
##  Pairwise comparisons using t tests with pooled SD
##
## data: Diet and treatment
##
##          Control Fructose Glucose
## Fructose 1.000   -      -
## Glucose   0.259   0.562   -
## Sucrose   0.017   0.026   0.130
##
## P value adjustment method: bonferroni
```

```
TukeyHSD(results, conf.level = 0.95)
```

```
## Tukey multiple comparisons of means
## 95% family-wise confidence level
##
## Fit: aov(formula = Diet ~ treatment, data = data1)
##
## $treatment
##          diff      lwr      upr      p adj
## Fructose-Control 0.2 -0.9148496 1.31485 0.8805808
## Glucose-Control  0.8 -0.3148496 1.91485 0.1337642
## Sucrose-Control  1.8  0.6851504 2.91485 0.0095276
## Glucose-Fructose 0.6 -0.5148496 1.71485 0.2677272
## Sucrose-Fructose 1.6  0.4851504 2.71485 0.0145912
## Sucrose-Glucose  1.0 -0.1148496 2.11485 0.0703156
```

```
plot(TukeyHSD(results, "treatment"))
```

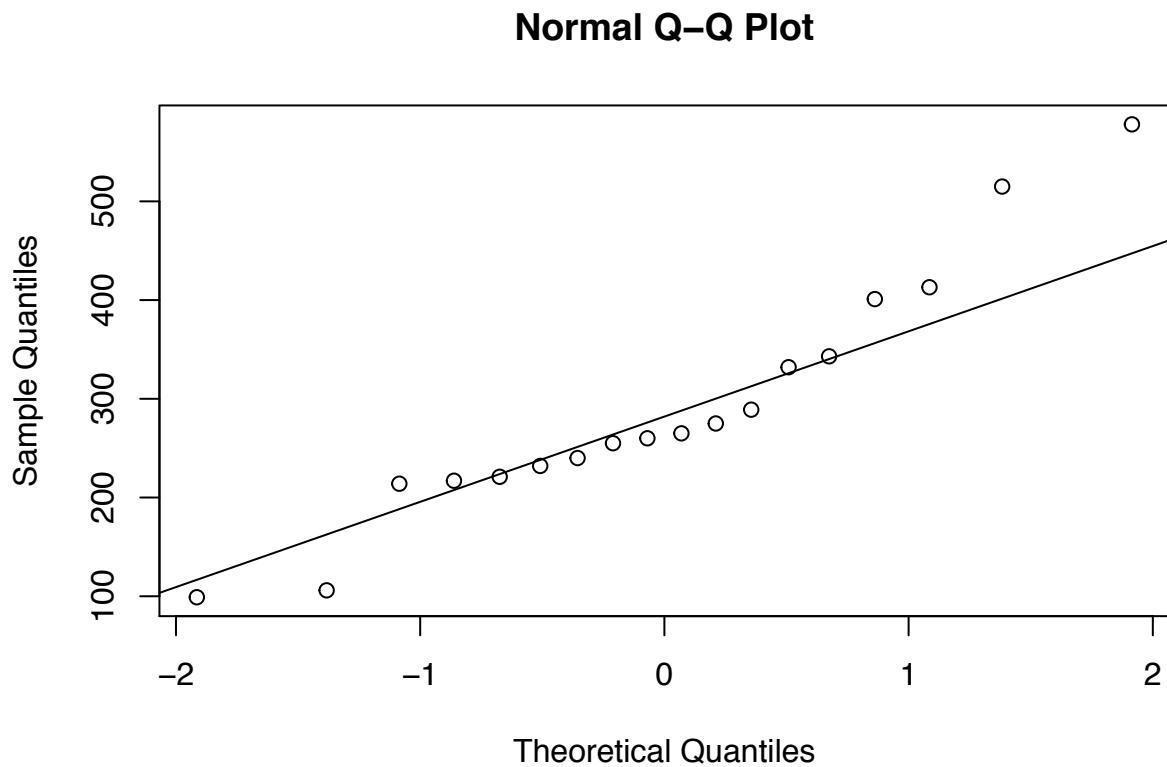
95% family-wise confidence level



```
# Sucrose-Control  1.8  0.6851504 2.91485 0.0095276  
# Sucrose-Fructose 1.6  0.4851504 2.71485 0.0145912
```

Example 2: Intravenous Fluids (on Page 153)

```
particles = c(255, 265, 343, 332, 232, 217, 106, 289, 99, 275, 221, 240, 578, 515, 214, 413, 401, 260)  
qqnorm(particles)  
qqline(particles)
```



```
summary(particles)
```

```
##      Min. 1st Qu. Median   Mean 3rd Qu.   Max.
## 99.0  223.8  262.5 291.9 340.2 578.0
```

`var(particles)`

```
## [1] 15327.47
```

```
# To get the SS total like we used to do for one-way Anova  
sum(particles^2)
```

```
## [1] 1794735
```

```
# Since the over all average for data is 291.9, we create a benchmark with 291.9  
# for all cells
```

```
grand1<-c(291.9,291.9,291.9,291.9,291.9,291.9,291.9)
```

Sum of the squares of grand gives Grand Average

sum(grand1^2)

```
## [1] 1533701
```

```

# Getting the Manufacturer Effect:
sum((particles-grand1)^2)

## [1] 260567

# If we subtract grand average from the average per group we get the manufacturer effect
# In order to get the Residual we subtract the grand average as well as the per
# group average from the observed value
# sum all the squares of the residuals gives SSE
manufacturer= c(rep("Cutter",6), rep("Abbot",6), rep("McGaw",6))
data2 = data.frame(particles,manufacturer)
data2

##      particles manufacturer
## 1          255        Cutter
## 2          265        Cutter
## 3          343        Cutter
## 4          332        Cutter
## 5          232        Cutter
## 6          217        Cutter
## 7          106        Abbot
## 8          289        Abbot
## 9           99        Abbot
## 10         275        Abbot
## 11         221        Abbot
## 12         240        Abbot
## 13         578       McGaw
## 14         515       McGaw
## 15         214       McGaw
## 16         413       McGaw
## 17         401       McGaw
## 18         260       McGaw

# Here we create boxplots for the data side-by-side
plot(particles ~ manufacturer, data=data2, main="Side-By-Side Boxplot of Particles count by Manufacture")

```