Chapter 1

ST 511 - Introduction

Readings: Chapter 1 (for 1.3 just read the 2 that interest you the most)

________________ - the science of designing studies or experiments, collecting data and modeling/analyzing data for the purpose of decisions making and scientific discovery when the available information is both limited and variable.

Why learn statistics?

- We live in a society that collects volumes upon volumes of data.
- Are people looking at the data?
- Are they interpreting the data properly?
- How do we turn raw data into information?
  - to make new policy
  - to make better product
  - to increase yield

Statistics is often called the 'science of learning from data.'

ex: Gas mileage
Suppose we fill 20 of the same model of car with a full tank of gas. Each car will have a different miles per gallon.
Why?

Factors that affect gas mileage:

To summarize the information from the 20 cars we might look at the **average** gas mileage of the 20 cars.

Questions to answer:

- How do we obtain an overall average miles per gallon for this model of car? (Not just for these 20.)
- overall average when driving in city?
- overall average when driving on a highway?
- overall average with low tire pressure?
- overall average when in a city with low tire pressure?

Statistics provides a framework for solving this type of problem!

Method of statistics often follows a 4 step process

- Step 1: Identify the research objective
- Step 2: Collect the information needed to answer the questions
- Step 3: Organize and summarize the information.
- Step 4: Draw conclusions from the information.

(Repeat as necessary to answer research objective.)

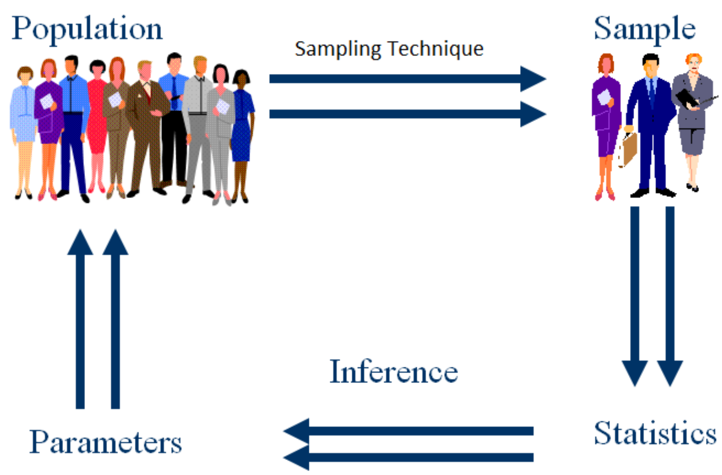
Big ideas in stats:

- ____________________ - all the values, items, or individuals of interest

- ____________________ - a (usually) unknown summary value about the population

- ____________________ - a subset of the population we observe data on

- ____________________ - a summary value calculated from the sample observations



Gas Example:
What is the population, sample, parameter of interest, and statistic (most likely to be used)?

Common Notation in statistics:

Name	Parameter	Statistic	Quantity Measured
Mean	μ	$\bar{Y}$ or $\bar{y}$ or $\bar{X}$ or $\bar{x}$	Center or Location
Proportion	p or π	$\hat{P}$ or $\hat{p}$ or $\hat{\pi}$	Location or Frequency
Standard Deviation (SD)	σ	S or s	Variability or spread
Variance (Var)	σ^2	S^2 or s^2	Variability or spread

Note: $\bar{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i$ and $S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(Y_i - \bar{Y})^2$ where n is the sample size (or number of observed values in the sample).

Many, many, many, more to come!

Question of interest will lead you to which parameter you have interest in. This will also most likely lead you to which type of data you will collect.

Scales (Types) of Data:

- ________________________________ - A variable that is described by attributes or labels
  Subscales:
  Nominal - categories have no ordering (Male, Female) (zip codes)
  Ordinal - can order categories (Lickert scale data) (college football rankings)

- ____________________ - A variable that is described by numerical measurements where arithmetic can be performed
  Subscales:
  Discrete - finite or countable finite number of values (# of flowers on a plant, 0, 1, 2, ...)
  Continuous - any value in an interval is possible (Temperature, $(-459.67 \deg F, \infty)$
  (Some lump these together and call them interval.)

How we summarize and analyze the data will depend on which type of data we have.

ex: SAT (get to know each other a little!)

- 50 total students (16 males and 34 females) where matched on socio-economic background (all had similar income).
- A study was done to examine the effect of preparation atmosphere on SAT scores.
- Two types of atmospheres were investigated (strict vs easy going).
- Students were divided into two groups of 25 (12 males and 13 females in strict class and 4 males 21 females in the easy going class).
- After a 9 week tutoring session the SAT was taken (although 1 in the strict group did not take the exam and 5 in the easy going group did not take the exam).

With a partner or two (introduce yourselves):

1. Determine the research question.
2. Define the population and sample.
3. Define possible parameter(s) of interest.
4. Define possible statistics that might be calculated.
5. Why might the students have been matched on socio-economic background?
6. What issues might you see with the design of this study?
7. What other variables might you collect?

Chapter 2

ST 511 - Sampling and Experiments

Readings: Chapter 2 (for 2.2/2.3 read if interested)

This class is about analyzing data. As scientists, most of the time this data will come from a designed experiment, but the methods used for analysis are also useful for observational studies. However, the conclusions drawn will differ! Let's define what me mean by experimental and observational study.

Observational Study researchers does not interfere or intervene in the process of collecting data.

- Ex: measuring political beliefs in using a poll, measuring yield of a crop based on rainfall

_________________________ researchers manipulate the conditions in which the experiment is done.

- Ex: assigning different fertilizers and irrigation method and measuring crop yield, assigning temperatures of water to tanks containing a fish and observing weight gain

Big difference in conclusions drawn!

- Cannot usually infer causation from observational experiments, but you can from a well-designed experiment.
- Experiments are not always feasible or ethical. i.e. cannot assign people to smoke a pack a day or have expectant mothers drink a certain amount of alcohol.

To describe the methods for creating a well-designed experiment, we first need some definitions.

- **Response Variable** - Variable of interest that characterizes performance or behavior.
- **Explanatory Variables** - Variables that determine the study conditions (can be quantitative or categorical).
- **Factor** - Explanatory variable of interest.
- **Level** - The specified value of a factor (or explanatory variable).

- **Confounding Variable** - Explanatory variable (not of interest) that may mask (or enhance) the effect of a factor.
- **Covariate** - Quantitative confounding variable.
- **Treatment** - A specific experimental condition, either the level of a factor (if only 1 factor) or the combinations of levels from a number of factors.
- **Experimental units (EUs)** - Units on which the treatments are assigned.
- **Measurement (Observational) units** - Units on which values are observed (often the same as EUs, but not always).
- **Replicate** - Name given to EUs that receive the same treatment.
- **Control Treatment** - Benchmark treatment sometimes necessary for comparison (to avoid the *placebo effect*).
- **Experimental Error** - Used to describe the variation in response among EUs that are assigned the same treatment.

Example: An experiment was run to determine the effects of adding phosphorous ($0, 147, 294, 441$ kg/m^2) and nitrogen ($0, 45, 90, 135$ kg/m^2) to the soil of a certain type of grass (a Miscanthus species). The growth of the plant was of interest and at the end of the growing period the plant was dried and the weight recorded with the final measurement being recorded in megagram per hectare (0.1 kg/m^2). Four plots of grass were used in total. Within each plot, each combination of phosphorous and nitrogen was observed. The plots were arranged in a large field in a 4x1 rectangle (north to south). There is a possibility of a water gradient as a stream runs to the north of the field. A partial data table is given here:

Plot	P	N	Dry yield
1	0	135	1.95
1	0	45	3.51
1	0	90	2.87
1	0	0	2.88
1	294	45	2.37
1	294	0	3.5
1	294	135	3.55
1	294	90	4.4
...	...	...	...

Let‘s identify (if possible) the response, explanatory variable(s), factor(s), level(s), confounding variable(s), treatment(s), number of replicates, experimental units, and observational units.

(See example 2.4/2.7 and the resulting discussion for more practice)

Notice that many of the response values are different. What is causing them to be different?

Sources of Variation in the responses of an experiment:

1. **Treatment effect** - we hope there is an effect due to the variables we are setting
2. **Identified confounding variables** - We record some variables that are not of interest, but we think may have an effect on the response.
3. **Unidentified sources (these make up the Experimental Error or error variation) -**
   (a) Inherent variability in experimental units - Experimental units are different!
   Ex: No two people, paper towels, concrete blocks, or even lab rats are exactly the same.
   Consequence: Experimental units respond differently to the same treatment
   (b) Measurement error - Multiple measurements of a same experimental unit typically contain error. If the same experimental unit is measured more than once, will the value be the same?
   Ex: Blood Pressure, Break a water sample in two, measure each for bacteria
   (c) Variations in applying/creating treatments
   The treatment is not clearly defined, leaving room for interpretation.
   Ex: Two researchers mix concrete, one stirs for 10 minutes and one for 20 minutes, will they come out exactly the same? Temperature is of interest but two ovens don't heat exactly the same, etc.
   (d) Effects from any other extraneous (or lurking) variables - Extraneous variables are those variables that are not part of the treatment, but may influence the response.
   Ex: For the oven example, the experiment is done over the course of several days. There may be slight differences due to humidity changes.

Let's identify these in the grass growth example.

(See example 2.5/2.6 and the resulting discussion for more practice)

No matter how hard we try, some experimental error will remain. What we can do is use good experimental design techniques to ensure our study is valid.

DOE is about creating the optimal experiment to determine the effects of different treatments. Different types of experimental designs are then analyzed differently.

Pillars of Experimental Design

1. ____________________ - means that the treatments are randomly allocated to the EUs.
   (a) Every EU has a chance to get a different treatment, so helps protect the results of the analysis against a systematic influence of lurking variables.
   (b) Allows the observed responses to be regarded as a random sample.

   Note: Different randomization schemes lead to different statistical analyses.

   ________________________________ - for t treatments, replicated n_t times each, use a random number generator to assign the treatments to the EUs.

   Most basic randomization design - assumes all EUs are exchangeable.

2. ____________________ - Repetition of an experiment using a large group of subjects to reduce chance variation in the results
   (a) Allows us to generalize the results to the population and increases reliability of conclusions.
   (b) Allows an estimate of variability (an estimate of experimental error) not due to the treatment effect.

   Note: Replication does not mean that we measure the same EUs multiple times, this is called repeated measures. Observations from repeated measures experiments cannot usually be considered independent.

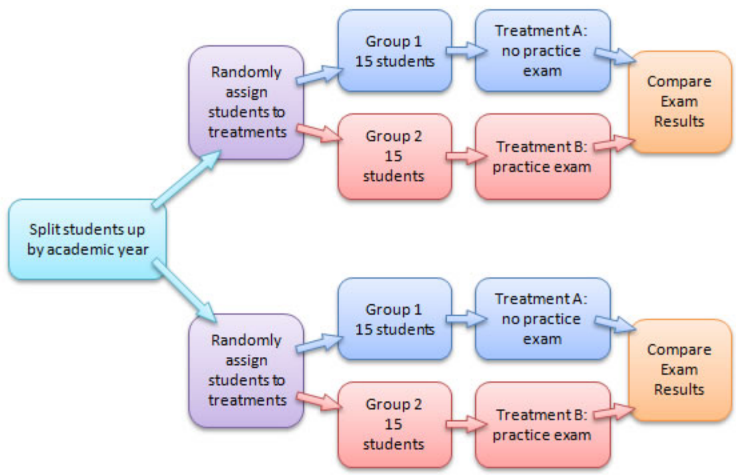
3. **Methods for accounting for/reducing experimental error**

   (a) Controlling Variables - holding certain variables constant across the EUs
   Decreases generalizability, but reduces experimental error.

   We're not interested in the effects of these variables on the response. These variables affect the response in exactly the same manner, so that we don't see the effects on the conclusions. We don't get information on what happens at levels other than the fixed one.

   (b) Blocking - Divide subjects with similar characteristics into 'blocks', and then within each block, randomly assign subjects to treatment groups.

   Blocks - Groups of EUs sharing a common level of a confounding variable.



   Similar to controlling, but allows for increased generalizability. EUs within a block are very similar (decreases experimental error there as all the EUs in a block are affected similarly by the confounding variable). By having enough blocks to cover the range of the population you can still generalize.)

There are also methods for dealing with some explained experimental error during the analysis stage - Namely ANCOVA.

These ideas are very important. Unless you are well versed in statistical methods and ideas you should consult a statistician before investing time and money in an experiment.
'A poorly designed study can never be saved, but a poorly analyzed one has the possibility of being reanalyzed.'

Chapter 3

ST 511 - Descriptive Statistics

Readings: Chapter 3 (you can skip the guidelines for constructing class intervals, stem-and-leaf plots, grouped mean/median)

Recall: Process of a study involves

1. Identify the research objective
2. Collect the information needed to answer the questions
3. Organize and summarize the information.
4. Draw conclusions from the information.

We will now talk about step 3!

So you have data... now what??

	A	B	C	D	E	F	G	H	I	J	K	L
1	Block	Treatment	Welltype	Depth	Month	ON	NH4	NO3	OP	Cl	TOC	ID
521	3	3	Middle	1	10	2.98	1.61	0.68	0.09	3.94	16.56	52.00
522	3	3	Middle	1.5	1	1.07	0.14	0.46	0.06	6.16	21.87	53.00
523	3	3	Middle	1.5	2	1.55	0.02	0.02	0.06	6.27	23.74	53.00
524	3	3	Middle	1.5	3	0.87	0.02	1.88	0.03	4.89	9.83	53.00
525	3	3	Middle	1.5	4	0.19	0.00	0.93	0.01	3.13	5.39	53.00
526	3	3	Middle	1.5	5	0.13	0.02	1.06	0.02	3.15	5.41	53.00
527	3	3	Middle	1.5	6	0.98	0.00	0.92	0.03	2.98	3.47	53.00
528	3	3	Middle	1.5	7	0.35	0.01	0.51	0.03	2.61	1.97	53.00
529	3	3	Middle	1.5	8	0.17	0.02	0.02	0.03	3.48	1.79	53.00
530	3	3	Middle	1.5	9	0.44	0.01	0.00	0.02	5.01	3.74	53.00
531	3	3	Middle	1.5	10	0.00	0.04	0.09	0.04	4.35	3.32	53.00
532	3	3	Middle	2	1	1.07	0.03	0.03	0.06	9.23	20.10	54.00
533	3	3	Middle	2	2	0.99	0.02	0.00	0.06	9.02	15.38	54.00
534	3	3	Middle	2	3	0.39	0.02	0.10	0.05	7.72	8.38	54.00
535	3	3	Middle	2	4	0.00	0.00	0.12	0.02	5.02	3.58	54.00
536	3	3	Middle	2	5	0.10	0.01	0.15	0.03	4.12	5.73	54.00
537	3	3	Middle	2	6	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]

Whether we are describing an observed population or using sampled data to draw an inference from the sample to the population, an insightful description of the data is an important step in drawing conclusions from it.

Good descriptive statistics enable us to make sense of the data by reducing a large set of measurements to a few summary measures that provide a good, rough picture of the original measurements.

Summary measure used for a variable depends on its ________________________.

Our goal will be to describe the variable's ________________________

i.e. the

Two major characteristics of the variable's distribution that we often describe are ________________________

and ________________________

We will mostly deal with quantitative variables and our focus will be on their summary measures. However, we will briefly talk about graphs and statistics for categorical variables.

Categorical Variables

Numerical measure used for categorical variable:

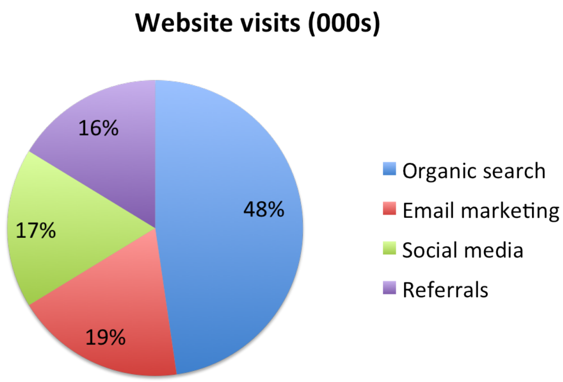
For this simple study, we can find the sample proportion for each categorical variable:

Panel	Type of Wood	Paint thickness in millimeters	Type of water repellent	Weathering time in months
1	Oak	8.5	Solvent-based	6
2	Pine	10.9	Solvent-based	4
3	Oak	9.6	Water-based	8
4	Poplar	8.0	Solvent-based	12
5	Pine	8.3	Water-based	3
6	Poplar	7.9	Water-based	15
7	Poplar	9.8	Water-based	15

The main plots used are ______________________ and ______________________.

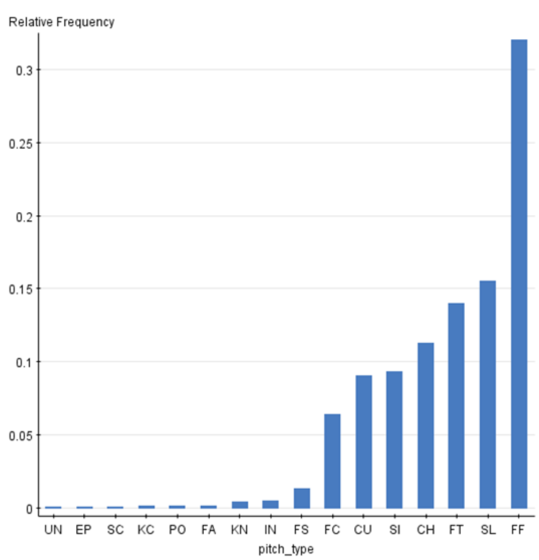
______________________ - use percents to display data.

- Order does not matter (although should order from highest percentage to lowest)



______________________ - Categories along the x-axis. Count, percent, or relative frequency (sample proportion) along the y-axis.

- Order does not matter (although should order from highest percentage to lowest).
- A gap should exist between the bars.



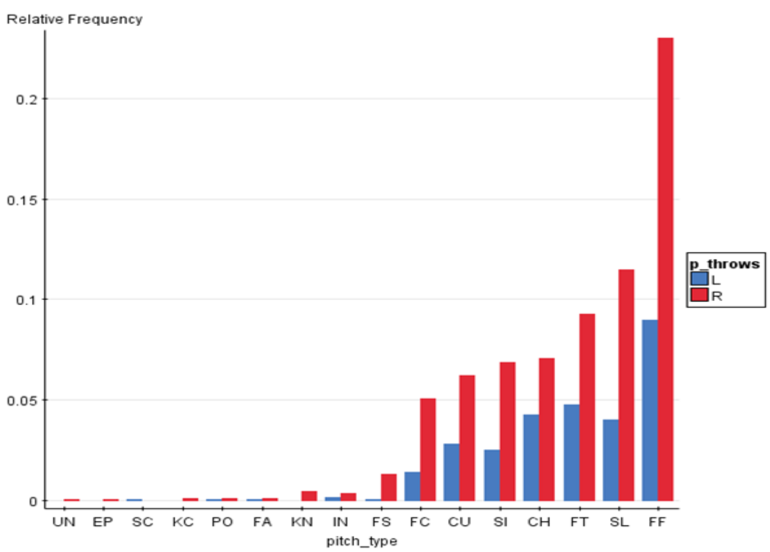
We might also have multiple categorical variables of interest. In which case, a good display of the data is

a ___________________________________.

Let's create one for the paint example from earlier.

Book example on page 103 is nice.

Many other methods exist such as comparative bar charts:



Quantitative Variables

We will again consider the paint example:

Panel	Type of Wood	Paint thickness in millimeters	Type of water repellent	Weathering time in months
1	Oak	8.5	Solvent-based	6
2	Pine	10.9	Solvent-based	4
3	Oak	9.6	Water-based	8
4	Poplar	8.0	Solvent-based	12
5	Pine	8.3	Water-based	3
6	Poplar	7.9	Water-based	15
7	Poplar	9.8	Water-based	15

Numerical measures of location:

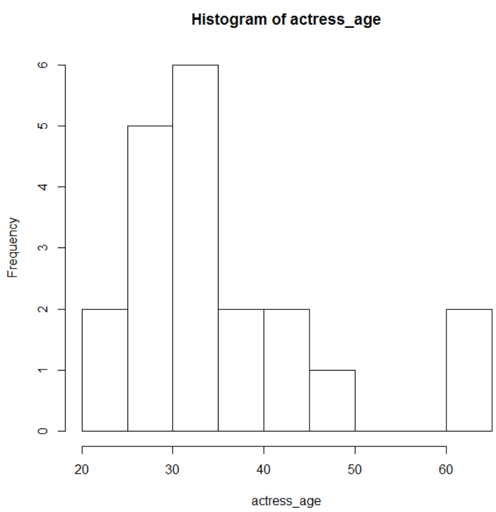
Numerical Measures of Spread

The main plots used are ________________________ and ________________________.

A ________________________ is obtained by splitting the range of the data into equal-sized bins. Then for each bin, we count the number of points that fall into each bin and that is the height of our bar (or use relative frequency - i.e. proportion in category).

- Typically, an observation equal to a boundary value is put in the higher interval.
- Bars should touch!
- Too many classes will spread the data out, thereby not revealing the pattern. Too few classes will lump the data.
- **** This is the most important graphical technique for displaying the distribution of a quantitative variable!

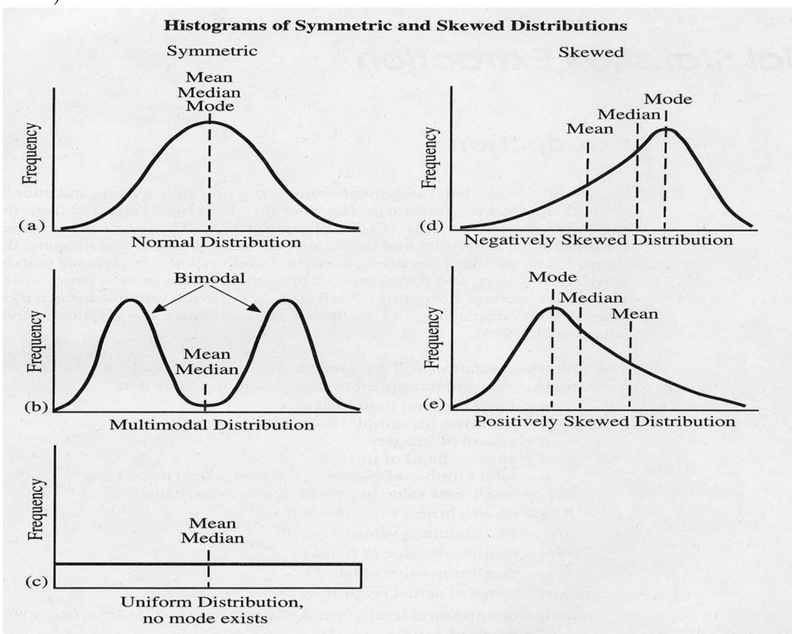
Ex. Ages of the winners of the best actress Academy Award in the recent 20 years (1994-2013) are: 36, 45, 49, 39, 34, 26, 25, 33, 35, 35, 28, 30, 29, 61, 32, 33, 45,29, 62 and 22



What are we looking for in a histogram?

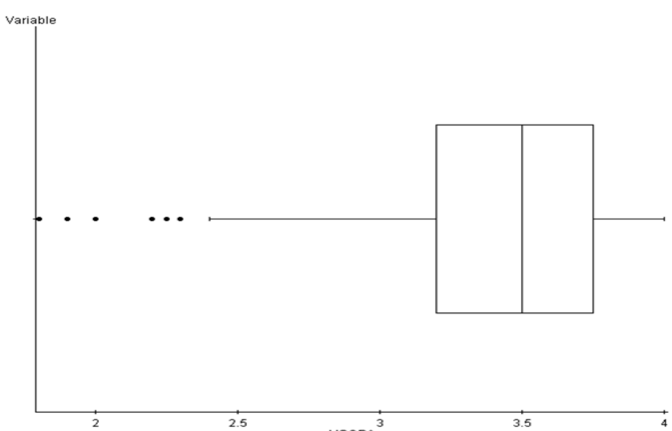
-
-
-

Relationship between mean and median for a histogram (note pictures use smooth curves, but same ideas hold):



A ________________________ displays the five number summary of the data.

Five number summary includes:



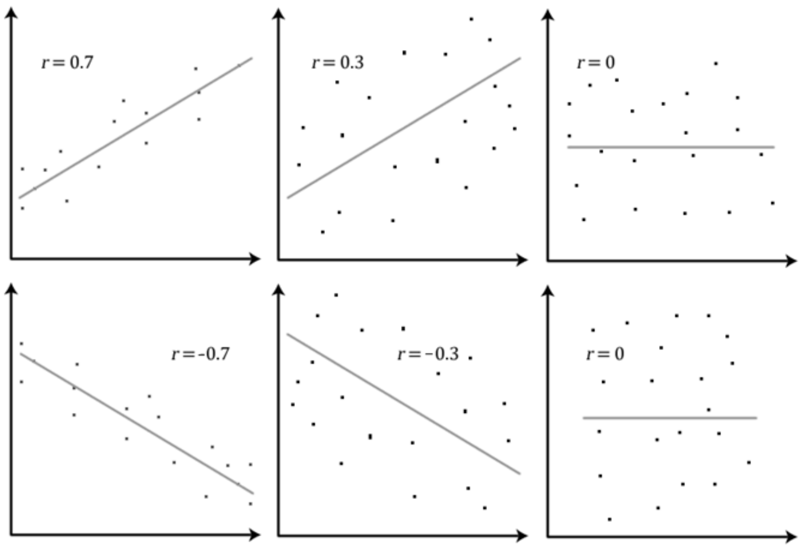
- Measure of center from a boxplot -

- Measures of spread from a boxplot -

- Can tell skewness but not modality!

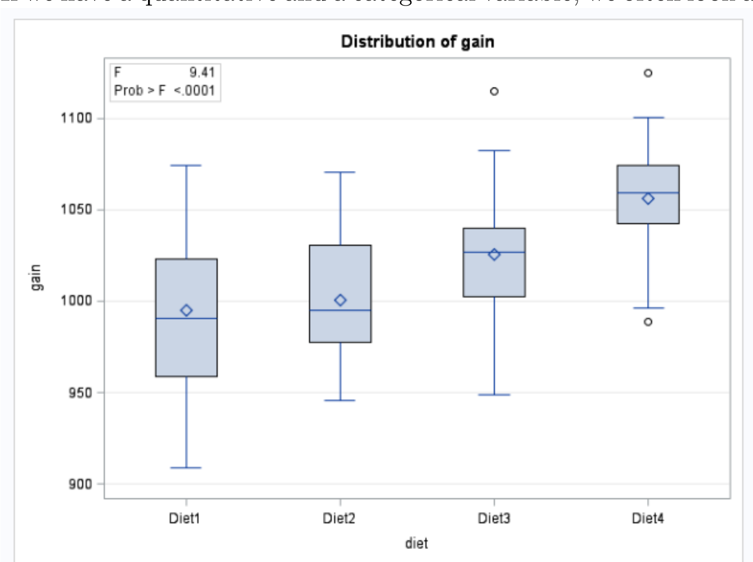
If we have two quantitative variables of interest, we often look at ______________________________

and ______________________________ to inspect the 'linear association' between the variables (call them x and y).



We will look at these more later in the course.

If we have a quantitative and a categorical variable, we often look at ______________________________.



Review -
Numeric Summaries of Location: Mean/median/trimmed mean (quantitative), proportion (qualitative)
Numeric Summaries of Spread: Variance, SD, IQR, CV, Quartiles (quantitative)
Graphical Summaries for categorical: Bar Chart, Pie Graph
Graphical Summaries for quantitative: Histogram, Boxplot

Chapter 4

ST 511 - Probability and Distributions

Readings: Chapter 4 - 4.1-4.4, 4.6-4.8 (ok to skip Poisson Distribution), 4.9-4.10, 4.12, 4.14

Recall: Our goal is to conduct inference. In order to do this, we need a firm understanding of probability.

Interpretation of Probability

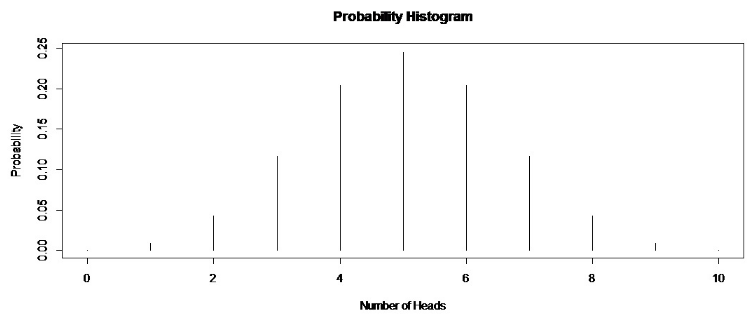
- ________________________________ of an outcome in repeated experiment = # of times outcome observed/# of times experiment was repeated
  - Ex: the chance of rolling snake eyes ($(1, 1)$) on two fair dice
  - Ex: the chance of getting a head on a flipped coin

Probability and Inference

- Ex: We want to see if a coin is fair. If we formulate the research question in terms of parameters, we want to test the hypothesis:

  Suppose the coin is tossed $n = 10$ times and yields $y = 10$ heads.

  - If hypothesis is true, how likely is the observed event?
  - With 10/10 heads, reasonable to conclude coin not fair. What about 9/10 heads? 7/10 heads?
  - To make a decision, need to know ____________________________



The above plot gives the probability of observing a given number of heads from a fair coin in 10 tosses.

Sets and Sample Spaces

A probability model is a mathematical representation of a random phenomenon. Defined by its

-
-
-

Definitions:

- _______________: A collection of **elements**, $a_1, a_2, \ldots$
- ____________________________: the set of all outcomes under consideration
- _______________: Each possible distinct result of a random process (experiment)
- A is a _________________ of B if every element of A belongs to B ($A \subset B$)
- _______________: A collection of outcomes (a subset of $\mathbf{S}$)

Sample Space examples:
Define the sample space **S** for each situation below.
Of the parts manufactured today, randomly select and measure the thickness of **a single part**.

It is known that the thickness must be between 10 and 11 mm.

It is known that the thickness has only three values (low, medium or high).

Experiment asks, does the part thickness meet specifications?

Now two parts are randomly selected and measured.

Do the 2 parts conform to specifications?

Number of conforming parts from the two is measured.

Now, parts are randomly selected until a non-conforming part is found.

More Set definitions:

- ________________ $(A \cup B)$: the set of all points in A or B (including both)
- ________________ $(A \cap B)$: the set of all points in both A and B
- ________________ of A ($\bar{A}$ or A^c): contains elements in S but not in A
- A and B are ________________ or ________________ if $A \cap B = \emptyset$.

Gender of Children - Discrete Example: - A family has two children of different ages. Consider the possible genders of these children. Let a pair FM denote the element in which the younger child is female and the older is male.

1. Sample Space: $S = \{ \qquad\qquad\qquad\qquad \}$
2. Let A be the event in which both children are males, B the event in which there is at least one male, and C the event containing no males.
   List the elements of
   - $A = \{ \qquad\qquad\qquad \}$ $\qquad A \cup C = \{ \qquad\qquad\qquad \}$
   - $B = \{ \qquad\qquad\qquad \}$ $\qquad \bar{A} = \{ \qquad\qquad\qquad \}$
   - $C = \{ \qquad\qquad\qquad \}$ $\qquad A \cup B = \{ \qquad\qquad\qquad \}$
   - $A \cap C = \{ \qquad\qquad\qquad \}$ $\qquad B \cup \bar{A} = \{ \qquad\qquad\qquad \}$

Relating Set Theory to Probability

- An __________________ is any process that can be repeated (theoretically) and has a well-defined set of possible outcomes (sample space)
- An event corresponds to ____
- The Probability of the event is the likelihood or chance that a particular outcome or event from a random experiment will occur. We write

$$P(A) = \text{Probability the event A occurs}$$

- Probabilities are numbers between ___________________________________
- May be written as proportion (0.15), percent (15%), or a fraction (3/20).
- P(Event)=1 implies ____
- P(Event)=0 implies ____

Simplified axioms of probability

- The probability of an event, $P(A)$, a function, must satisfy:

  ____

  ____

  If A and B are disjoint (mutually exclusive) then

  ____

Probability example (Recall Gender of Children ex) - The sample space for this experiment was

$$\mathbf{S} = \{MM, MF, FM, FF\}$$

1. What would be reasonable probabilities for each outcome in S?

2. For A, B, and C, defined earlier, find $P(A)$, $P(B)$, $P(C)$, $P(A \cup C)$, and $P(S)$

Conditional Probability, Independence, and Other Probability Rules

Often we will have knowledge of one event's occurrence. How does that change the probability of another event?

For example, the probability of getting a 1 in the toss of a six-sided die is __________.

If we know that an odd number has fallen, then the probability of occurrence of a 1 is __________.

The __________________________________ of an event A given that an event B has occurred is equal to

provided $P(B) > 0$.

Independent Events
Two events and are said to be __________________________ if and only if **any** one of the following 3 conditions hold:

$$\begin{aligned} P(A|B) &= P(A) \\ P(B|A) &= P(B) \\ P(A \cap B) &= P(A)P(B). \end{aligned}$$

Otherwise, the events are said to be ____________________________________

Independence Examples

- Consider again the "Gender of Children" example. Let A be the event that the younger child is female, and B be the event that the older child is male. Are A and B dependent?

- (Credit Card Example) - The proportion of NCSU students with a VISA card is 0.48, the proportion with a MasterCard is 0.64, the proportion with both is 0.35.

  1. Calculate the conditional probability that a randomly sampled student has a VISA given he/she has a MasterCard.

  2. Are the events 'having a VISA' and 'having a MasterCard' independent?

Laws of Probability
Sometimes probabilities of events can be obtained by using multiplicative and additive rules.

____**:**

$$\boxed{P(A \cap B) = P(B|A)P(A) = P(A|B)P(B)}$$

Notice that if A and B are ________________________, then

$$\boxed{P(A \cap B) = P(A)P(B)}$$

Using the Multiplicative Law
(Urn Example) - An urn contains 10 marbles, 4 are red (R) and 6 are black (B).

1. If 2 are randomly chosen from the urn, what is the probability that both are black?

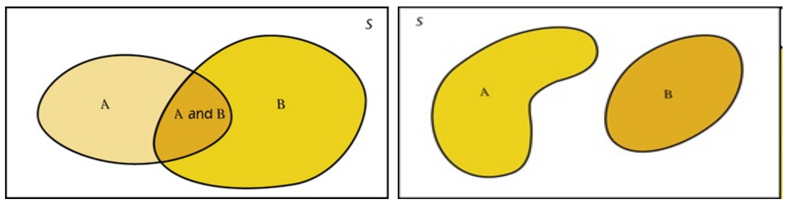
2. If 1 is randomly chosen, then replaced, and then another randomly chosen (making the selections independent events), what is the probability of selecting a red then a black?

3. Flip a fair coin 3 times, find the probability of observing 3 heads (HHH) assuming independent flips.

____:

$$\boxed{P(A \cup B) = P(A) + P(B) - P(A \cap B)}$$

Recall 3rd axiom: if A and B are disjoint, then $P(A \cup B) = P(A) + P(B)$.



Using the Additive Law

- (Credit Card Example) - The proportion of NCSU students with a VISA card is 0.48, the proportion with a MasterCard is 0.64, the proportion with both is 0.35.
  Find the probability that a randomly sampled student has a VISA or MasterCard (or both).

- (Axiom Example) - Can A and B be mutually exclusive if $P(A) = 0.4$ and $P(B) = 0.7$? What if $P(B) = 0.3$?

______________________________:
A special case of additive law is obtained by taking $B = A^c$, then

$$\boxed{P(A) + P(A^c) = 1 \text{ implies } P(A) = 1 - P(A^c)}$$

Ex: In 17th century De M'er'e asked Pascal, which is more likely:
A: rolling at least one six in four throws of a single dice, or
B: rolling at least one double six in 24 throws of a pair of dice?
Find $P(A)$ and $P(B)$.

(See example 4.1 on page 149.)

Random Variables

Motivating Ex.:
Assume that all mens basketball teams playing this season are equally strong. We are interested in

$$Y = \# \text{ of points scored by NC State in a game.}$$

- Before each game, we know the population of possible values.
- Each value occurs with some probability.
- However, we do not know what will be the number of points scored by NC State during the next game.

The outcome is random, hence the # of points scored in a game is a **random variable**.

- A **Random Variable (RV)** is a real-valued function
  - Domain (values it takes in) = ______________________________
  - Range (values it outputs) = ______________________________

  An RV assigns a real number to each outcome in a sample space.

Two Types of RVs we'll discuss

- ______________________________: takes on finite or countably infinite # of values
- ______________________________: takes on a subset of intervals of real numbers

Why do we need to distinguish between these two types of RVs?

Basic Definition and Probability Distributions

Discrete Random Variables - An Example

- **Discrete random variable** assumes only a finite or countably infinite # of values
- Ex: Flip a coin 3 times - Let $Y = \#$ of heads from the 3 tosses
  - Range of Y?
  - Called ________________________________ of the RV
- Each outcome has a _________________________

  To describe the distribution, we need to describe the probability for each outcome in the support!

- Function $P(y) = P(Y = y)$ is called the ____
- Can be represented as a table:

Possible Values of Y	y_1	y_2	...	y_n
Probability for each value	$P(y_1) = P(Y = y_1)$	$P(y_2) = P(Y = y_2)$	...	$P(y_n) = P(Y = y_n)$

Let's find the probability distribution for $Y = \#$ of heads from 3 tosses using a table:

Some other examples of discrete random variables:

- Y = # of textbooks purchased in a semester. Support:
- X = # of plants that bloom from a group of 20 plants. Support:
- Y = # of flips of a coin before first head. Support:

Probability distribution for a discrete random variable must follow the following rules:

- For every y in the support of the RV Y,
- The sum of the probabilities over the entire support must be 1.

Let's check for the coin example.

Rules of probability still apply. For any two distinct values in the support, call them y_1 and y_2

Let's compute $P(Y \geq 2)$ for the coin example.

Summary Characteristics of RVs

Just as in the numerical summaries section we will want to summarize characteristics of the distribution. What are the two major characteristics?

To find the ____ of a **discrete RV**

Let's find the mean of Y from the coin example.

To find the __________________________________ of a **discrete RV**

Let's find the variance of Y from the coin example.

Let X (Any capital can be used to denote a RV) denote the # of male children if a family has 2 children (assume a the probability of a male child is 0.4 and that the children are independent).

- Determine the support of X

- Find $P(x)$, the probability distribution of X using a table

- Show that $P(x)$ meets the two conditions to be a probability distribution for a discrete RV.

- Find $P(X = 0 \textit{ or } X = 2)$

- Find the average number of male children.

- Find the variance of the number of male children.

Binomial Distribution

Recognizing a Distribution

- Note: # of Heads example and the # of male children example are similar.

- Similar experiments with similarly defined RV's yield the same ________________________________

- This particular distribution so common, it is called the ________________________________

- Knowing and being able to recognize common distributions will save us from having to derive things over and over!

When does a RV follow the Binomial Distr.?

Consider the following experiments:

- a coin is flipped, the outcome is either a head or a tail.
- a baby is born, the baby is either born in March or is not.

In each of these examples, an event has two ___________________________________.

For convenience, one of the outcomes can be labeled __________________________ and the other

outcome ________________________________

Bernoulli Trials

- An experiment with only two possible mutually exclusive outcomes (such as S or F) is called a Bernoulli Trial
  - Bernoulli trials are the basis of three 'families' of distributions:
    - ________________________________ distribution
    - ________________________________ distribution
    - ________________________________ distribution

- For a trial denote the probability of success as

- Then the probability of failure is

We have a **Binomial Experiment** if:

1. Full experiment consists of a sequence of ________________________________

2. ____ on each trial (Bernoulli Trials)

3. Probability of success $P(S) = \pi$ is ________________________________, where $0 \leq \pi \leq 1$

Define the RV $Y =$ ____

Then Y is said to follow a binomial distribution.

We write ____ for convenience.

For $Y = \#$ of heads in three tosses

For $X = \#$ of male children from the two

(see example 4.5 and example 4.6 on pages 159/160 for practice picking out binomial experiments)

General form of the Probability Distribution for a Binomial RV

Ex: Suppose we have a Binomial Experiment with $n = 3$ trials and $P(S) = \pi$ where π is an unknown parameter

- Let Y be the $\#$ of successes – ____

Outcome	$P(\text{Outcome})$	y	Reps	$P(Y = y)$
SSS	$\pi\pi\pi = \pi^3$	3	1	π^3
SSF SFS FSS				
SFF FSF FFS				
FFF				

For general n, the event $Y = y$ occurs when there are exactly

- Consider one such outcome w/1^{st} y trials successful last $n - y$ failures:

$$SSS \cdots SFFF \cdots F$$

- Probability of this outcome?

- How many different sequences with exactly y successes in n trials?

The Probability distribution for $Y \sim Bin(n, \pi)$ is:

$$y = 0, 1, 2, \cdots, n, \ 0 \leq \pi \leq 1$$

Binomial Distribution Example

Suppose 60% of NCSU students favor closed-book exams. A random sample (outcomes independent) of 5 NCSU students is drawn.

1. Define Success/Failure, n, π, and a RV Y that follows the Binomial distribution

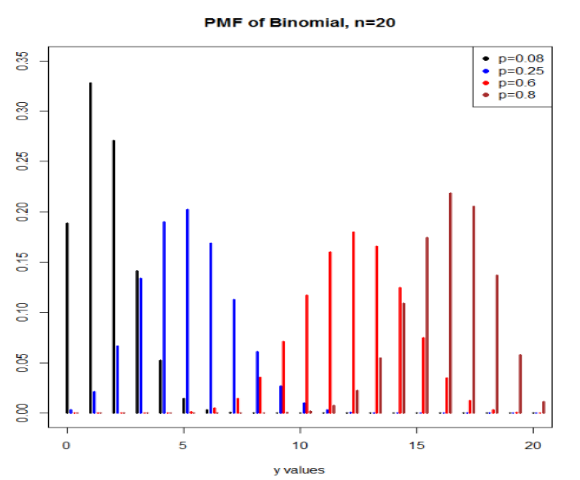
2. Calculate $P(exactly\ 1\ in\ favor)$

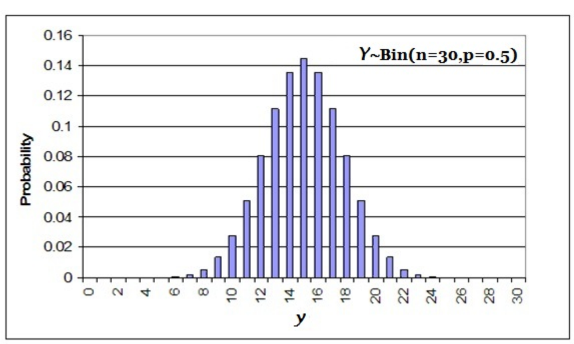
3. Calculate $P(less\ than\ 2\ in\ favor)$

4. Calculate $P(4\ or\ more\ in\ favor)$

(see examples 4.7 and 4.8 on page 162 for more practice with the binomial pmf)

We will want to have general formulas for the mean and variance of a binomial. Consider the following plots:





Binomial Expected Value - If $Y \sim Bin(n, \pi)$, then

Binomial Variance - If $Y \sim Bin(n, \pi)$, then

Binomial Standard Deviation =

Multiple Choice Test Example
Consider a multiple choice test with 20 questions, each with five possible answers (a,b,c,d,e), only one of which is correct. Let Y= # of questions guessed correctly.

1. Let's verify Y follows a binomial, calculate $E(Y)$, and calculate $Var(Y)$.

2. If scores of 50% and higher are passing, find the formula (i.e. don't simplify) for the probability of passing by guessing.

Many other common discrete distributions exist.

Connection with making inference

Hypothesis Testing Idea:

You love Pepsico and their products. They are having a promotion where their bottle caps are either winners (a free Pepsico product) or losers. Your friend claims you will hardly ever win, in fact he thinks only 1 in 20 bottles is a winner. You think the chance of winning is much higher than that.

To prove him wrong you grab 50 randomly selected Pepsico bottles and find that 12 of your caps are winners. How can we show your friend you are most likely correct?

Continuous Random Variables

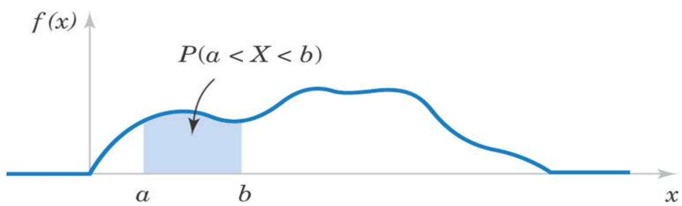
A ______________________________ has an interval or collection of intervals as its support
Ex:

- Y=maximum daily temperature (interval $[-40°F, 130°F]$).
- Y=lifetime (in years) of electronic equipment $0 < Y < \infty$
- Y=weight loss (or gain) after a 6 month period $-\infty < Y < \infty$.

For Discrete RVs we had the probability distribution, $P(y) = P(Y = y)$.

For Continuous RVs we can't assign probability to every y in the support. We now call the probability distribution by

- Probability a randomly chosen value will lie between any 2 given values is represented in terms of the area between the two values under the probability distribution.



A function $f(y)$ is a probability distribution if and only if

1. $f(y)$ is a ______________________________, i.e. $f(y) \geq 0$ for all y
2. The area under $f(y)$ is 1, i.e.

Similar to the discrete case where $P(y) \geq 0$ and $\sum_y P(y) = 1$

We can find probabilities using integrals:

Example: Probabilities from a continuous probability distribution
Let Y be a random variable with probability distribution:

$$f(y) = \begin{cases} (1/2)y & 0 < y < 2 \\ 0 & \text{else} \end{cases}$$

1. Graph the probability distribution.

2. Find $P(1 \le Y \le 2)$ and $P(1 \le Y < 10)$.

Expectations

Definition of Expectation
For a RV Y with probability distribution $f(y)$, the **expected value** of Y or mean of Y is defined as

$\mu_Y = E(Y)$ is then a ____________________________________ of all possible values of Y, with weighting function $f(y)$.

In general, the expectation of a function of Y, $g(Y)$, can be evaluated as

Variance of a Continuous RV
Definition of Variance is still the same as the discrete case:

Example: Let Y be a random variable with probability distribution:

$$f(y) = \begin{cases} 3y^2 & 0 < y < 1 \\ 0 & \text{otherwise} \end{cases}$$

Perhaps Y models the proportion of gas in a tank at a randomly selected time. Calculate μ_Y, σ_Y^2, and σ_Y.

(More practice will be provided in the problem session problems, of course these will be posted online if you can't make it!)

Named Distributions
How do we use continuous RVs?

- As before, for a particular experiment we assume a distribution and find characteristics of interest (probabilities, means, variances, etc)
- As with discrete RVs, many scenarios lead to similar distributions (such as the binomial for discrete RVs)
- The most important continuous distribution is the normal distribution. We will also discuss the t-distribution, χ^2 distribution and F-distribution later in the course.

Normal Distribution

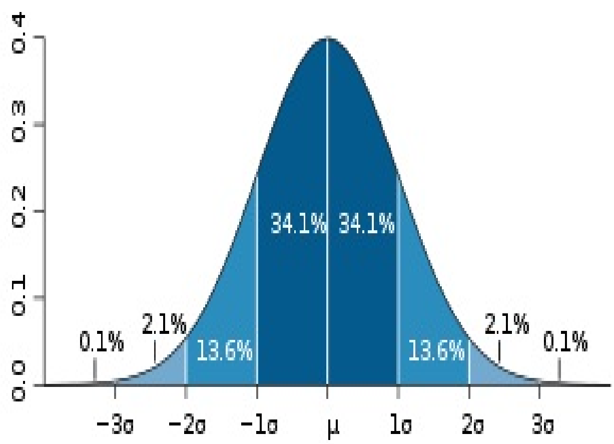
The Normal Distribution $Y \sim N(\mu, \sigma^2)$

A RV Y has a **normal distribution** with mean μ and variance σ^2 if the probability distribution of Y is

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

The constants ________________________ are the 'parameters' of the distribution.

We write $Y \sim N(\mu, \sigma^2)$.
Most Famous Bell-Shaped Curve



Properties of the $N(\mu, \sigma^2)$ **RV**

- ____

- ____

- ____

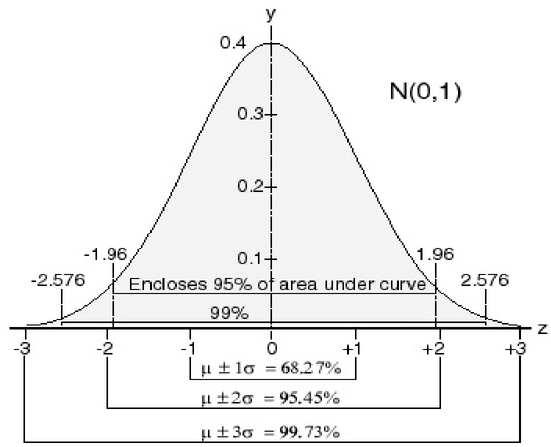
Expectation and Variance of $N(\mu, \sigma^2)$**.**:

$$E(Y) = \_\_\_\_\_\_\_\_\_\_, \quad Var(Y) = \_\_\_\_\_\_\_\_\_\_\_.$$

Note: these are the parameters of the distribution - (i.e. the distribution is ________________________ by its mean and variance)

$Z \sim N(0,1)$ (i.e. a normal distribution with $\mu = 0$ and $\sigma^2 = 1$) is said to follow a

The standard normal is centered at zero and its probabilities are concentrated between (-3,+3).



Standardization of Normal Random Variables Theorem

If $Y \sim N(\mu, \sigma^2)$, then ______________________________ follows the std normal distribution:

Suppose that $Y \sim N(\mu, \sigma^2)$. By standardizing Y, we have

Likewise, if $Z \sim N(0,1)$ _____________________________________

CDF of $Z \sim N(0,1)$: $\Phi(z)$
The CDF (cumulative distribution function) of the standard normal random variable Z is by definition:

No closed form, has to be calculated through ___________________________________ .

Tables or calculators often used.

We will use SAS to find probabilities (for homework, on test you will just need to find the answer in terms of a standard normal distribution). See SAS file on web.

Bottle Example
A company that manufactures and bottles apple juice uses a machine that automatically fills 16-ounce bottles. There is some variation, however, in the amounts of liquid dispensed into the bottles. The amount dispensed has been observed to be approximately normally distributed with mean 16 ounces and standard deviation 1 ounce.
What is the probability a randomly selected bottle will:

- have more than 17.5 ounces?

- have between 15.2 and 16 ounces?

- have less than 15 ounces?

(See example 4.15, 4.16, 4.17 in the book on page 174 for more practice.)

Percentiles of the Normal Distribution
The $(100p)$**th percentile** of Y (also called the **pth quantile** of Y) is the value y that solves $P(Y \leq y) = p$.

Suppose $Z \sim N(0,1)$. Find (see SAS file)

1. the 97.5th percentile of Z: $z =$

2. the 2.5th percentile of Z: $z =$

Suppose $Y \sim N(100,9)$. Then find

1. the 97.5th percentile of Y: $y =$

2. the 2.5th percentile of Y: $y =$

(See examples 4.18 and 4.19 on page 177/178 for more practice.)

SAT/ACT Example
The mathematics portion of the SAT and ACT exams produce scores that are approximately normally distributed. The SAT scores have averaged 480 with a S.D. of 100. The ACT scores average 18 with a S.D. of 6.

1. An engineering school sets 550 as the minimum SAT math score for students. What percentage of students will score below 550 typically?

2. What score should the engineering school set as a comparable standard on the ACT?

Sampling Distributions

Sampling Distribution Example:
Recall: A company that manufactures and bottles apple juice uses a machine that automatically fills 16-ounce bottles. There is some variation, however, in the amounts of liquid dispensed into the bottles. The amount dispensed has been observed to be approximately normally distributed with mean 16 ounces and standard deviation 1 ounce.
We defined the random variable

Now, suppose we think our machine may have broken and the amount dispensed might not actually be 16 any more.

Our goal is now to estimate the true mean amount of liquid in each bottle. We take a random sample of 10 bottles and find the amount each has. What statistic might we use to help answer our question?

Say for that sample the mean amount dispensed was 15.6 ounces. If we then take a second sample of 10 bottles and find the sample mean, will we get the same value? Why/Why not?

Similar to our random variable above, **the idea of finding the sample mean for a sample is a random variable itself**!!

We call the distribution of a statistic, such as the sample mean, the ________________________________ of the statistic.

Let's investigate the **sampling distribution** of the mean - http://www.stat.tamu.edu/ west/ph/sampledist.html

When do we know the sampling distribution of $\bar{Y}$'s form?

Distribution of $\bar{Y}$ from a Normal Population

If the 'parent population' is normal with mean = μ and variance = σ^2, i.e.

and a 'random sample' of size n is taken

then the distribution of $\bar{Y}$ will be normal with mean = μ and variance = σ^2/n

Example: Suppose the yearly rainfall totals for a city in northern California follow a Normal distribution, with a mean of 18 inches and a standard deviation of 6 inches. We take a random sample of 5 years worth of data.

1. Do we know the distribution of the parent population? If so, what is it?

2. Do we know the distribution of the sample mean for n=5? If so, what is it?

3. What is the probability of observing **a rainfall** greater than 12 inches?

4. What is the probability of observing **a sample mean rainfall** (n=5) greater than 12 inches?

When do we know the sampling distribution of $\bar{Y}$'s form?

Distribution of $\bar{Y}$ from a 'Large' sample

Central Limit Theorem (CLT): If the parent population has mean = μ and variance = σ^2 and a 'large' random sample (usually if n $\geq$ 30) is taken then we can use the approximation:

Example: Suppose that we are interested in the mean of hours studied in a week for a certain population of college students. We take a random sample of size $n = 64$ students from that population. Suppose we know from past data that the mean hours studied is 10 and the standard deviation of hours studied is 4.

1. Do we know the distribution of the parent population? If so, what is it?

2. Do we know the distribution of the sample mean for n=64? If so, what is it?

3. What is the probability of observing a student that studies less than 8 hours?

4. What is the probability of observing a sample mean (n=64) greater than 12?

For more practice with probabilities about a sample mean see example 4.24 on page 189 and problem session problems.

Things to note:

The distribution of $\bar{Y}$ from a random sample is centered at the mean of the parent population.

Means are ________________________________ than individual observations.
Also, means from larger samples vary less than mean from smaller samples.

The standard deviation of a statistic is also called the _______________________________________

Every statistic has a sampling distribution. Most do not have a normal distribution, but often for a large sample a normal distribution can be a reasonable approximation. Let's check out the applet again!

Normal approximation to the binomial and to $\underset{\sim}{\hat{\pi}}$ or $\underset{\sim}{\hat{p}}$

Suppose $Y \sim Bin(n, \pi)$. Then $Y = X_1 + X_2 + ... + X_n$ where

$$X_i = \begin{cases} 1 & \text{if trial i is a success} \\ 0 & \text{otherwise} \end{cases}$$

Note: each $X_i \sim Bin(1, \pi)$.

How can we use the CLT to approximate the distribution of $\underset{\sim}{\hat{\pi}} = \underset{\sim}{\hat{p}} = \frac{X_1+X_2+...+X_n}{n}$?

Similarly, we can then approximate the distribution of Y by

Example: Technology underlying hip replacement has changed as these operations become more popular. Still, for many patients, the increased durability has been counterbalanced by an increased incidence of squeaking. Suppose that the probability of a hip squeaking is 0.4. A random sample of 25 people will be taken.
Let Y = # of subjects whose hips developed squeaking.

1. Define the exact and approximate distribution we could use for Y.

2. Calculate $P(Y \leq 10)$ using the normal approximation and using the binomial and compare.

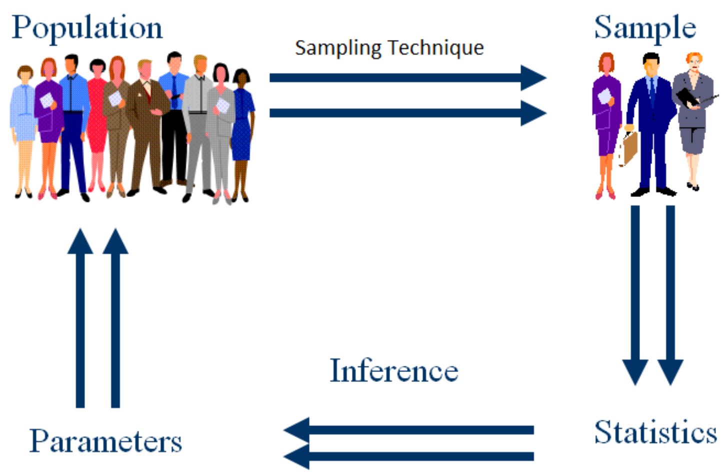
Where this is really useful is when n is very large! For another example see example 4.25 pg 192.

Chapter 5

ST 511 - Inferences About Population Central Values

Readings: Chapter 5 (for 5.8-5.9 read if interested)

Recall our overall idea:



Inference - refers to making mathematical claims about a parameter using sample data.

Two main methods for inference:

1. ______________________________ - Range of values we think contains the parameter.

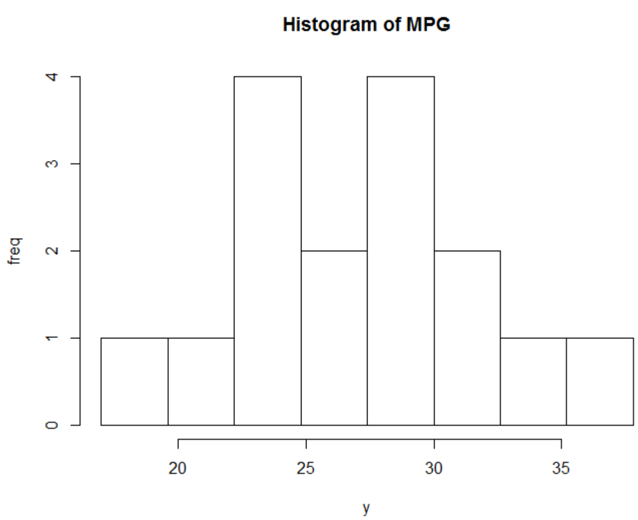
2. ______________________________ - Test of whether a specific parameter value is plausible.

Putting it all together: Engineers at Ford are attempting to improve the overall gas mileage of next year's model of one of their cars. From extensive testing on the previous year's model they know the gas mileage can be well modeled by a normal distribution with true mean gas mileage of 26.9 mpg and true standard deviation of 2.3 mpg.
To investigate if this year's model has improved the average gas mileage, data is collected on 16 automobiles. The average gas mileage of the sample is 28.25 mpg.

- Population -

- Parameter of Interest -

- Random Variables used to answer questions of interest -



- Sample -

- Statistic(s) -

- Inference using a 'Hypothesis Test' -

- Inference using a ‘Confidence Interval’ -

5.2 - Confidence Intervals

Confidence Intervals are better than point estimates, such as $\bar{y}$, as they ________________________________

Intervals that are created have a ____ = the probability **the procedure used** creates an interval that contains the parameter.

$\alpha =$ ____ - usually 0.01, 0.05, or 0.1.

An **observed interval**, such as the one in the example above, is interpreted in the following manner:

We say...

For a confidence level of $(1 - \alpha)$, the general interpretation of an observed CI is...

What we mean by 'confident' for our example above is...

For a general $(1 - \alpha)100\%$ CI, confidence is interpreted as...

Note that the interval is random and the parameter is fixed!

`http://bcs.whfreeman.com/ips4e/cat_010/applets/confidenceinterval.html`

Confidence Interval for μ_Y, the population mean

For a **random sample** of size n, where the population standard deviation, σ_Y, is known, a $(1-\alpha)100\%$ CI for μ_Y is given by

The interval is valid if **either**

1.

2.

Common z values used:

Ex: The length of music videos is of interest to advertisers. Assume we know the standard deviation of the length of music videos is 18 seconds (a dubious assumption we will learn to deal with later). In a random sample of 44 music videos, the average length was found to be 186 seconds. Find a 90% confidence interval for the mean length of all music videos. What does confidence mean here?

(Book examples 5.1 pg 228, 5.2 pg 229 for more practice.)

Factors that affect the width of confidence intervals:

1. Natural Variability -

2. Level of Confidence desired -

3. Sample size n

5.3 - Required sample size to have an interval of a given width:

Ex: Suppose that birth weights for boys are normally distributed. We want to estimate the population mean μ, the overall average birthweight for boys. Assuming that the population standard deviation, σ, is known to be 12 oz (based on past data), what minimum sample size is required if the width of a 90% CI is to be at most 6 ounces?

(Book examples 5.3/5.4 on page 231/232 for more practice.)

5.4/5.6 - Hypothesis Testing for μ_Y.

- Hypothesis testing and confidence interval estimation are related methods and are often both used to analyze the same situation.
- CI is a numerical answer to the question, 'What is the population value?'
- Hypothesis test is used to answer questions about particular values for a parameter.

Goal of hypothesis testing:

Step 1/2 - Determine Null and Alternative Hypotheses

____________________ - a statement or claim regarding parameter(s)

____ - Statement about parameter 'no effect' or 'no difference,' the default belief or status quo.

____ - Statement we hope to prove or give evidence for.

One-sided vs Two-Sided Hypotheses:

For the following examples, define the parameter of interest and determine the null and alternative hypotheses:

1. A certain type of light bulb is advertised as having an average lifetime of 750 hours. A potential customer likes the price and wants to purchase a large amount of them if it can be shown the average lifetime is higher than advertised. A random sample of 20 bulbs was selected and the lifetime of each bulb was determined. The mean was 766.4 hours. It is known that the lifetime of the light bulbs is normally distributed and the true standard deviation is 30.5 hours.

2. The average age of a person on facebook last year was 19.34 years. The standard deviation of the age of facebook users is 8.2 years. Suppose an advertising agency is interested in seeing if the average age is different this year. He randomly selects 150 profiles and finds that the sample mean is 18.99 years old.

There are two possible conclusions from a hypothesis test:

-
  - Observed value is 'significantly different' than hypothesized value (i.e. observed value is unlikely to have occurred simply due to random chance if the hypothesized value were true)
-

We never 'accept' H_0 as the book states. We believe the null hypothesis until we see significant evidence to the contrary. Thus, we either reject the null or we fail to reject (not enough evidence to the contrary). This does not imply that H_0 is true, just that we can't say it isn't true.

- The data collected will not 'prove H_0', but it may lead us to believe that it is pretty unlikely that H_0 is true.

Two types of errors we could make:

- ____________________________ - Reject H_0 when H_0 is 'true'
  - Probability of Type I error = P(Type I Error) =
- ____________________________ - Fail to reject H_0 when H_A is 'true'
  - Probability of Type II error = P(Type II Error) =

We consider the Type I error to be the most serious one. This is why we 'control' the type I error rate by setting it **prior to the experiment**.

Decision	H_0 Is True	H_0 Is False
Fail to reject H_0	no error	type II error
Reject H_0	type I error	no error

Idea follows US justice system. Consider the following example:

A person is on trial for a crime.

- The null hypothesis is H_0: Innocent
- The alternative is H_A: Guilty
  - A type I error would be
  - A type II error would be

For most crimes, a type I error is worse than a type II. This is the same in an experiment, usually making a type I error is worse, so we set *the type I error rate* at α.

For the light bulb example earlier, what would a type I error be in words? a type II error?

Step 3 - Check Assumptions and Find Test Statistic

To make inference, we will need a 'test statistic' (such as $\bar{Y}$ or $Z = \frac{\bar{Y}-\mu_Y}{\sigma_Y/\sqrt{n}}$) that we know the *sampling distribution* of.

Recall: If we are interested in a true mean, we can estimate it using the sample mean (a RV). We know the distribution of the sample mean is

$$\bar{Y} \sim N(\mu_Y, \sigma_Y^2/n)$$

if

We assume the null hypothesis is true and see if we find evidence to the contrary.

Assuming the null hypothesis is true, we can use the test statistic

For the two examples previously given,

- determine which assumptions are met,
- calculate the **observed** value of the test statistic,
- assuming the null is true, draw the sampling distribution and place the observed value of the test statistic on the distribution.

Step 4 - Find Rejection Region (RR) and/or find P-value - Make Decision

________________________________ is determined by our chosen α and the distribution of our test statistic under the null hypothesis.

RR - values of the test statistic for which the null hypothesis will be rejected.

For a '$>$' alternative and an $\alpha = 0.05$ let's find our RR

For a '$<$' alternative and an $\alpha = 0.01$ let's find our RR

For a '$\neq$' alternative and an $\alpha = 0.05$ let's find our RR

For the previous two examples, let's write down our RR and make our decision using $\alpha = 0.05$.

___________________________ - probability of getting a test statistic as extreme or more extreme than the observed value, assuming the null hypothesis is true.

- Measure of how likely it is to get this type of sample (or worse) if H_0 is true.
- A better measure than RR for evaluating the evidence for or against the null.

For a '$>$' alternative the p-value is

For a '$<$' alternative the p-value is

For a '$\neq$' alternative the p-value is

Decision determined by

- If p-value $\leq \alpha$, ____

- If p-value $> \alpha$, ____

For the previous two examples, let's draw the sampling distribution of the test statistic, shade the 'extreme' region, find the p-value, and make our decision using $\alpha = 0.05$.

Step 5 - Draw Conclusions (in the context of the problem)

Interpreting the result means that we say in a formal way what our conclusion means for this problem:

- Fail to reject H_0, we say... At the $\alpha 100\%$ significance level, there is not enough evidence to support the alternative hypothesis that (context).
- Reject H_0, we say... At the $\alpha 100\%$ significance level, there is enough evidence to reject the null hypothesis that (context) in favor of the alternative that (context).

For the previous two examples, let's interpret our results at the 0.05 significance level.

Overview of HT

1. Set up Alternative
2. Set up Null
3. Check Assumptions and Calculate Observed Test Stat
4. Find RR and/or P-value
5. Draw Conclusions in the context of the problem

Example: The average employee tenure (number of years workers have been with their current employer) in 2010 was 4.4 years with a standard deviation of 0.9 years. Tenure is believed to be higher in this year than it was in 2010. A sample of 90 employees produced a mean tenure period of 4.7 years.

1. Assuming the spread remained constant, conduct a 0.01 level (that means use $\alpha = 0.01$) test to determine if average tenure is greater than it was in 2010. (You need to do all 5 steps.)

2. Is the sample mean 4.7 'significantly different' from 4.4? Explain what we mean by significantly different in the context of the problem.

Relationship between CIs and Two-sided Tests

- If the null value μ_0 is contained in a $100(1-\alpha)\%$ CI for μ, then we fail to reject H_0 at level α.
- If the null value μ_0 is NOT contained in a $100(1-\alpha)\%$ CI for μ, then we reject H_0 at level α.

Let's calculate a confidence interval for μ in the employee tenure example.

5.5 - Power and Choosing a Sample Size

______________________ = 1 - P(Type II Error) = 1-β = 1 - P(failing to reject H_0 when H_A is true) = P(reject H_0 when H_A is true)

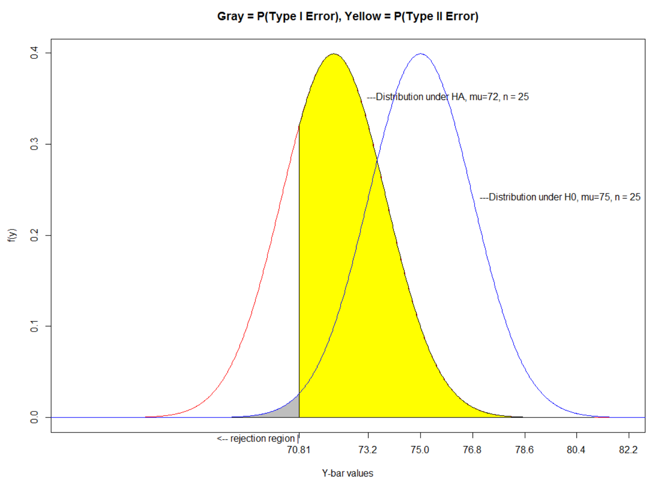
Ideally, we have small type I AND type II error rates (probabilities).
We 'control' the α = P(type I error) = type I error rate by setting this prior to the experiment.
The main way to deal with the type II error rate (or equivalently power) is by increasing the sample size.

Ex: The drying time of a certain type of paint (in minutes) under specified test conditions is known to follow a $N(75, 9^2)$. Chemists have designed a new additive to decrease average drying time. Let Y denote the drying time of this new additive. Lets assume the $Y \sim N(\mu, 9^2)$. We want to determine if there is strong evidence to suggest an improvement in average drying time. Suppose a random sample of 25 drying times is taken and the sample mean is 70.8.

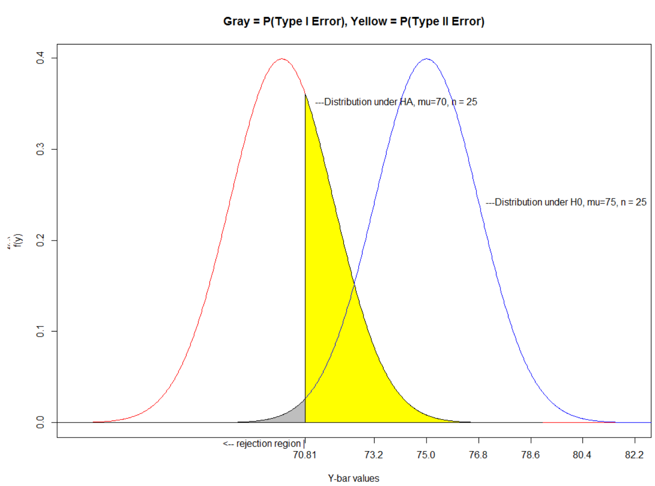
1. Conduct a hypothesis test using rejecting regions and $\alpha = 0.01$.

2. A HT is conducted assuming H_0 is true. Let the random sample of 25 drying times be denoted as $Y_1, Y_2, ..., Y_{25}$. What is the distribution of $\bar{Y}$ generally?

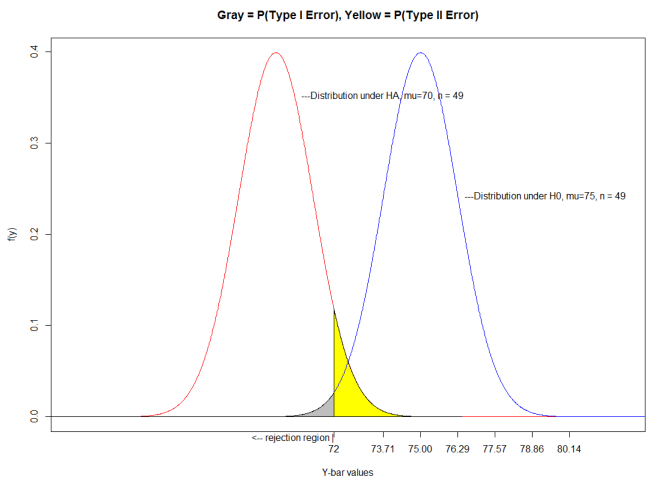
3. Assume now that in actuality the average drying time is truly 72 minutes. Find the probability of a type II error. (We denote this as $\beta(72)$.)



4. Assume now that in actuality the average drying time is 70 minutes. Find the probability of a type II error, $\beta(70)$.



5. Continuing with the assumption that $\mu = 70$, suppose now that a random sample of size n=49 is conducted. Find $\beta(70)$.



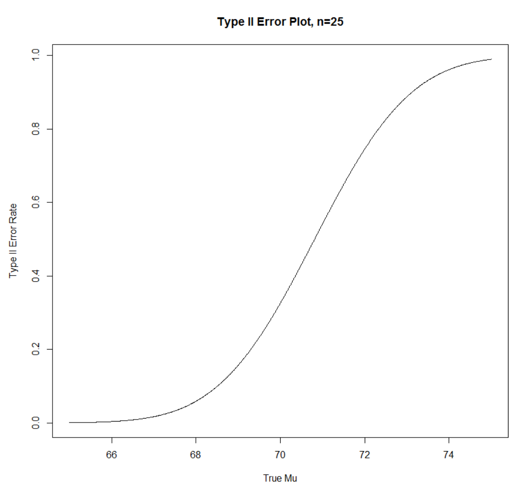
Generally, the type II error rate for a one-tailed test is given by

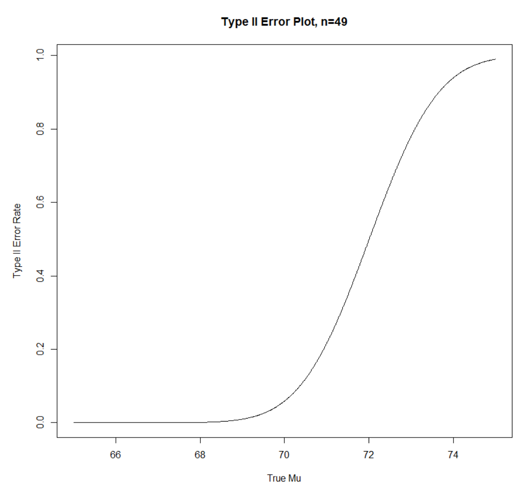
$$\beta(\mu_A) = P\left(Z \leq z_\alpha - \frac{|\mu_0 - \mu_A|}{\sigma/\sqrt{n}}\right)$$

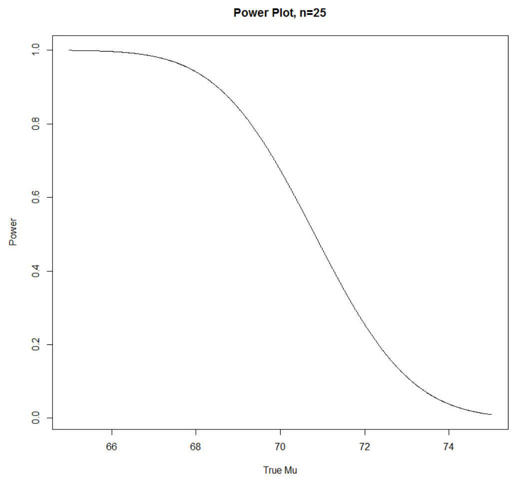
for a two-tailed test it is given by

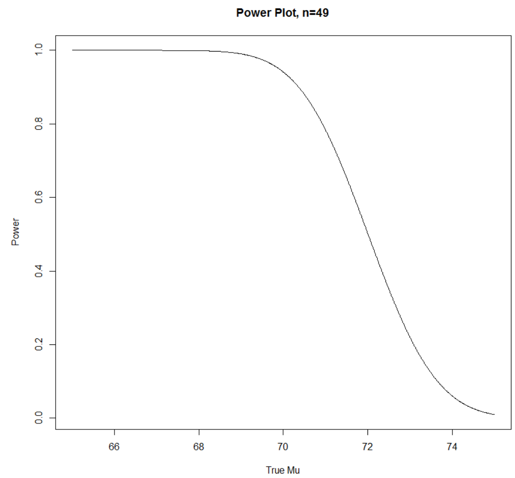
$$\beta(\mu_A) \approx P\left(Z \leq z_{\alpha/2} - \frac{|\mu_0 - \mu_A|}{\sigma/\sqrt{n}}\right)$$

Prior to an experiment, we would assume a value for σ_Y and plot the power (or type II error rate, β) as a function of μ_A.







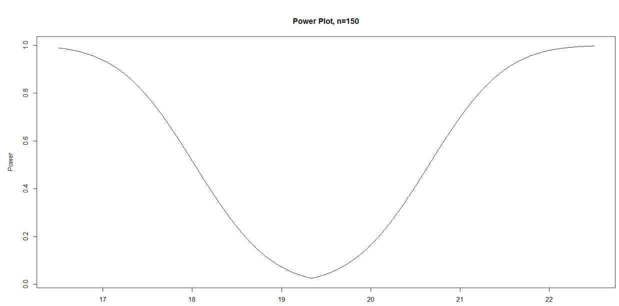


Recall Example: The average age of a person on facebook last year was 19.34 years. The standard deviation of the age of facebook users is 8.2 years. Suppose an advertising agency is interested in seeing if the average age is different this year. He randomly selects 150 profiles and finds that the sample mean is 18.99 years old.

Using $\alpha = 0.05$, our RR is $\{z_{obs} : |z_{obs}| > 1.96\}$.

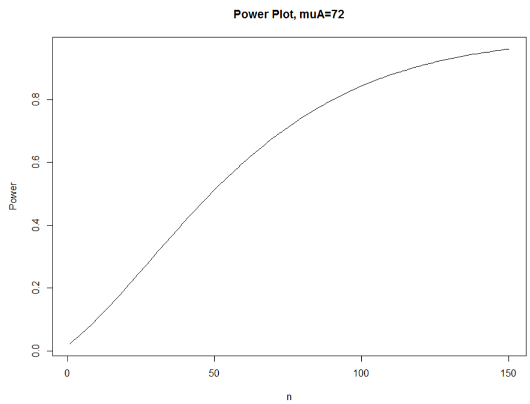
What is the power if $\mu_A = 18$ is the truth? if $\mu_A = 21$ is the truth?

Looking at the power curve below, what is the power when $\mu = 19.34$ is the truth? Why does this make sense?

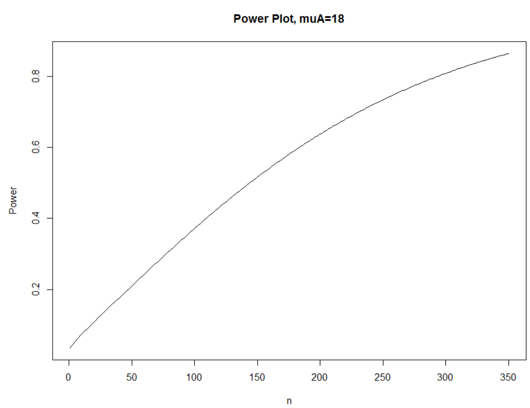


Alternatively, we could fix the value of μ_A and let n vary to determine a given sample size for obtaining a certain power.

Drying time example with $\mu_A = 72$. Plot of power for varying n



Facebook age example with $\mu_A = 18$. Plot of power for varying n



Inference about μ_Y when σ_Y is unknown (Section 5.7)

To use the previous Hypothesis Test or Confidence Interval for μ we need to know the true value of σ_Y^2. In real life this is highly unlikely!

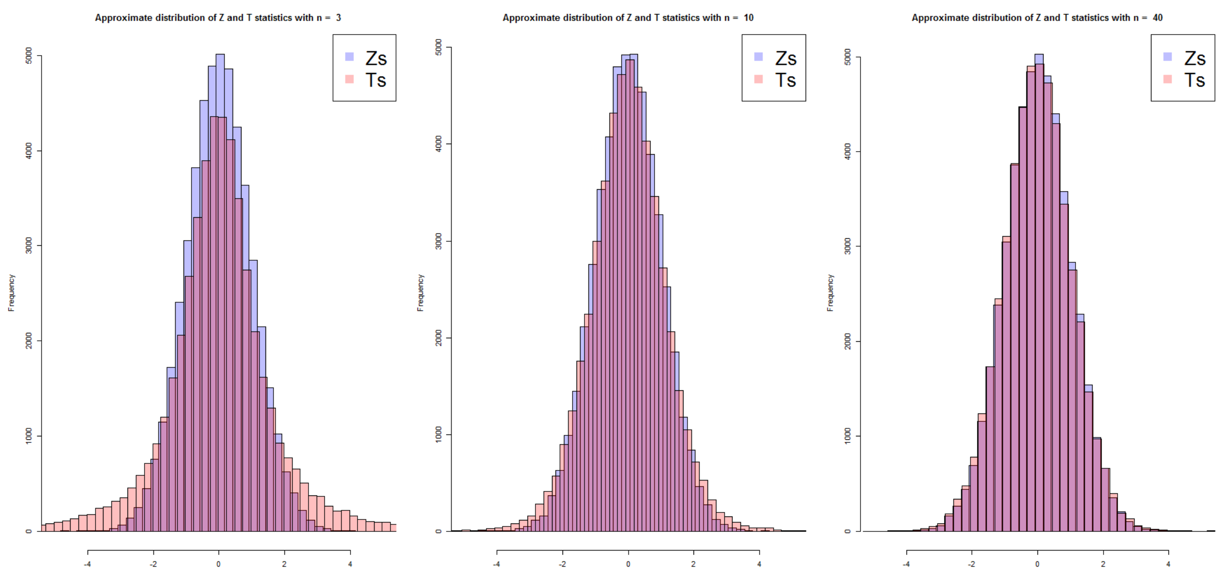
Recall: ________________________________ is the standard deviation of a statistic. For $\bar{Y}$

$$SE(\bar{Y}) = \sigma/\sqrt{n}$$

A good estimator of σ is the sample standard deviation $S = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(Y_i - \bar{Y})^2}$, if we plug this in for σ we get the

Our inference for μ was based on the test statistic

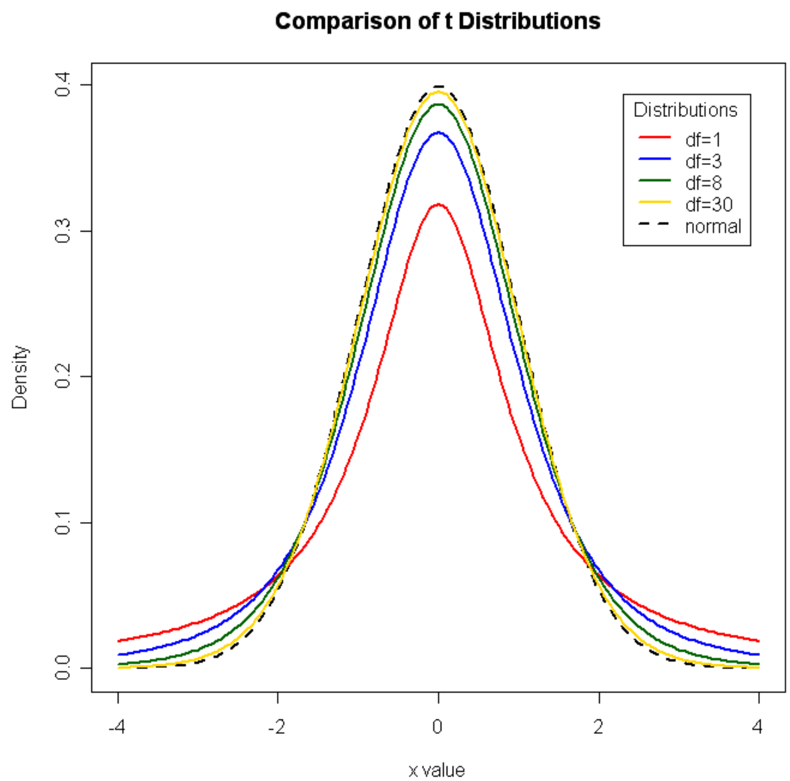
If we plug in our estimator of σ, does



For data from a normal distribution

T-distribution

t-distribution is a bell-shaped distribution centered at 0



Degrees of freedom:

For a random sample of size n where the parent population is *reasonably symmetric and mound shaped*, a $(1-\alpha)100\%$ CI for μ is given by

- $t_{\alpha/2}$ is a multiplier from the t-dist. with df = n-1 (for every df, $t_{\alpha/2}$ will be different).

- Called a 'one-sample t' interval for μ (our previous interval is called a 'one-sample z' interval for μ).

In SAS we can get t_α or $t_{\alpha/2}$ by the following code:

tinv(p,df) returns the pth quantile, 0<p<1, of the t distribution with df degrees of freedom.
To get $t_{\alpha/2}$ we want to find the $1-\alpha/2$ quantile (or use symmetry and take the negative of the $\alpha/2$ quantile). Thus, we can get $t_{\alpha/2}$ with the code:

```
data multiplier;
talpha1=tinv(1-alpha,df);
*or;
talpha2=-tinv(alpha,df);
run;
```

A psychologist claims that the mean age at which children start walking is 12.5 months. Carol wants to check if this claim is true. She took a random sample of 18 children and found that the mean age at which these children started walking was 12.9 months with a standard deviation of 0.80 month. Using a 99% Confidence Level, find a CI for $\mu =$ the true mean age at which children start walking. Be sure to interpret the interval, state the assumptions that are met **or required** for this interval to be valid. Can you conclude that the mean age which all children start walking is different from 12.5 months? (Helpful T_{df} values - $P(T_{18} > 3.55) = 0.01$ $P(T_{18} > 3.88) = 0.005$ $P(T_{17} > 3.57) = 0.01$ $P(T_{17} > 3.90) = 0.005$

For another example of a t interval for μ see example 5.17 on pg 256.

Similarly, we can create a Hypothesis Test for μ_Y when σ_Y is unknown.

HT for μ_Y when σ_Y is unknown

Step 1/2: Setting up hypotheses - No change

Step 3: Check Assumptions/Find Test Statistic

Assumptions: A random sample of size n where the parent population is ______________________________

____

then $T = \frac{\bar{Y}-\mu_0}{S/\sqrt{n}} \sim t_{n-1}$ can be used as a test statistic.

Observed value of test stat -

Step 4: Find RR and/or P-value - Make Decision

Step 5: Interpretation - No change.

We can get probabilities about the t distribution with a particular df using SAS via the code below:

```
data probs;
tprob1=probt(0.9,5); *P(T with 5 df < 0.9);
tprob2=1-probt(2.1,10); *P(T with 10 df > 2.1);
run;
```

Ex: A biology class is asked to find if the average wingspan of monarch butterflies is different from 91 mm. The class caught and measured the wingspans of 13 monarch butterflies. The average length (in millimeters) was found to be 93.5 with a standard deviation of 3.44mm. Conduct a hypothesis test with level 0.01 using p-values. (Use all 5 steps and be sure to state assumptions you must make for this procedure to be valid. Also, how would you investigate that assumption based off your data set?) $P(T_{13} > 2.620) = 0.0106$ $P(T_{12} > 2.620) = 0.0112$ $P(T_{13} > -2.620) = 0.9894$ $P(T_{12} > -2.620) = 0.9888$ $P(T_{90} > 93.5) \approx 0$

Note: If n is 'large' (> 30) then in practice you often use z critical value (multiplier) and use the standard normal to create the RR and to find the p-value.

Ex: Suppose that we are interested in testing whether or not the average NCSU student spent more than \$200 this semester on textbooks. We randomly sample 50 students and ask them how much they spent this semester on textbooks. Suppose that the sample average was \$204.5 and the sample standard deviation was \$20.12. Carry out a hypothesis test using the RR approach with $\alpha = 0.02$. Useful values: $P(T_{49} > 1.68) = 0.05 \quad P(T_{49} > 2.40) = 0.01 \quad P(T_{49} > 2.01) = 0.025 \quad P(T_{49} > 2.11) = 0.02$

Steps 1/2: $\mu =$ (true) average amount spent on textbooks this semester for NCSU students

$$H_0 : \mu = 200 \quad or \quad H_0 : \mu \leq 200$$
$$H_A : \mu > 200$$

Step 3: Since RS and n is large we can use

$$T = \frac{\bar{Y} - \mu_0}{S/\sqrt{n}} \sim t_{n-1}$$

as our test statistic.

$$t_{obs} = \frac{\bar{y} - \mu_0}{s/\sqrt{n}} = \frac{204.5 - 200}{20.12/\sqrt{50}} = 1.582$$

Step 4: RR determined by the alternative hypothesis and our test statistic. Since $T \sim t_{n-1}$ and we have a '$>$' alternative our RR is

$$RR = \{t_{obs} : t_{obs} > 2.11\}$$

Since $1.582 < 2.11$ we fail to reject H_0.

Step 5: At the 2% significance level, there is not enough evidence to support the alternative that the true average amount spend on textbooks this semester by NCSU students is greater than \$200.

Note: If we had used a standard normal instead of the t_{49} our RR would have been

$$RR = \{t_{obs} : t_{obs} > 2.05\}$$

This is pretty close, and so often for very large n we simply use the standard normal distribution instead of the t.

For another HT using the t, see example 5.15 on pg 253.

Chapter 6

ST 511 - Inferences Comparing Two Population Central Values

Readings: Chapter 6 (for 6.3-6.5 read if interested)

Our problems so far have dealt with inference for the mean of only **1 population** of interest. In real life this will not usually be the case. We will start with looking at inference regarding the means of **2 populations** and then in later chapters look at what to do with an arbitrary number of populations.

Motivating Example:
Jocko's garage seems to be giving out really high estimates for insurance claims. To investigate insurance fraud, insurance adjusters take 10 damaged cars and take each one to both Jocko's and a repair shop they trust, Jami's repair shop. Then then get the estimates from the repair shop (in the end, 2 for each car). Data are provided below:

Obs	Jocko	Jami
1	450	255
2	699	720
3	670	499
4	800	760
5	401	225
6	1000	700
7	535	300
8	680	350
9	1100	1000
10	850	770

Here we have two populations: all estimates from Jocko's and all estimates from Jami's repair shop.

Therefore, we have 2 random variables:
Y_i= estimate for the i^{th} randomly selected car at Jocko's
X_i= estimate for the i^{th} randomly selected car at Jami's

We now have two sample sizes:
n_1 (or n_Y) = number sampled at Jocko's
n_2 (or n_X)= number sampled at Jami's.
(Here they are equal, but generally for a two sample problem, they need not be.)

We now have two sample mean **random variables**:
$\bar{Y}$= mean estimate for a randomly selected sample of 10 cars at Jocko's
$\bar{X}$= mean estimate for a randomly selected sample of 10 cars at Jami's

We also have 2 sets of summary statistics (1 for each sample):

The UNIVARIATE Procedure
Variable: Jocko

Moments			
N	10	Sum Weights	10
Mean	718.5	Sum Observations	7185
Std Deviation	225.955871	Variance	51056.0556
Skewness	0.27601197	Kurtosis	-0.6296669
Uncorrected SS	5621927	Corrected SS	459504.5
Coeff Variation	31.4482771	Std Error Mean	71.4535202

Basic Statistical Measures			
Location		Variability	
Mean	718.5000	Std Deviation	225.95587
Median	689.5000	Variance	51056
Mode	.	Range	699.00000
		Interquartile Range	315.00000

The UNIVARIATE Procedure
Variable: Jami

Moments			
N	10	Sum Weights	10
Mean	557.9	Sum Observations	5579
Std Deviation	267.400428	Variance	71502.9889
Skewness	0.14751428	Kurtosis	-1.3512691
Uncorrected SS	3756051	Corrected SS	643526.9
Coeff Variation	47.9298132	Std Error Mean	84.55944

Basic Statistical Measures			
Location		Variability	
Mean	557.9000	Std Deviation	267.40043
Median	599.5000	Variance	71503
Mode	.	Range	775.00000
		Interquartile Range	460.00000

Two parameters of interest:
μ_Y (or μ_1) = (true) mean of all estimates at Jocko's
μ_X (or μ_2) = (true) mean of all estimates at Jami's repair shop.

Goal: Investigate $\mu_D = \mu_{diff} = \mu_1 - \mu_2 = \mu_Y - \mu_X$

What are possible methods of inference for $\mu_{diff} = \mu_1 - \mu_2$?

Distribution	Two Samples are Independent	Two Samples are 'Paired'
$\bar{Y} - \bar{X} \sim Normal$	6.2 - Two-sample t-test	6.4 Paired-t-test
$\bar{Y} - \bar{X} \sim Not\ Normal$	6.3 - Wilcoxon Rank Sum Test	6.5 - Wilcoxon Signed Rank Test

6.4 - Inference for Paired Data (Matched Pairs t or Paired t)

What is paired data?
Each 'unit' receives two treatments. The units could be:

1. A single subject (each subject gets both treatments)
2. Two subjects that have been **matched** together (one receives treatment A and the other receives treatment B)

Ex: Auto example - We have paired data because

How to make inference here? Hypothesis test = paired t-test:
Parameter:

Null hypothesis:

Alternative Hypothesis:

Test Statistic:

RR/p-value:

Conclusions same as for all HT. Note that this test is **equivalent to the one-sample t-test on the differences between the paired data.**

Similarly we can create a confidence interval using the test statistic above:

Note: We do not need to know each variable's sample mean and standard deviation, **only the mean and standard deviation of the differences!**.

Both the HT and the CI can be done very easily in SAS:

```
data autodata;
input Jocko Jami;
datalines;
450 255
699 720
670 499
800 760
401 225
1000 700
535 300
680 350
1100 1000
850 770
;

proc ttest data=autodata;
     paired Jocko*Jami;
run;

/* About this code:
The PAIRED VAR1*VAR2 statement requests the paired t-test.
SAS calculates the differences as VAR1-VAR2.
*/
```

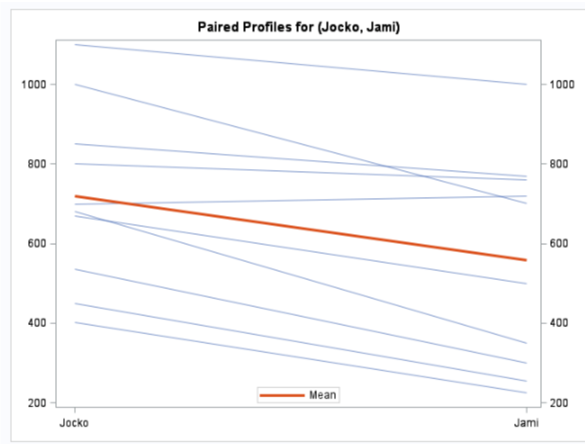
The TTEST Procedure

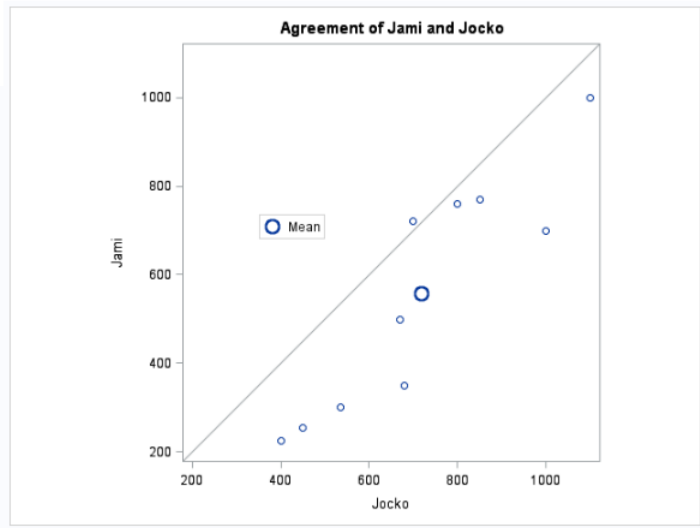
Difference: Jocko - Jami

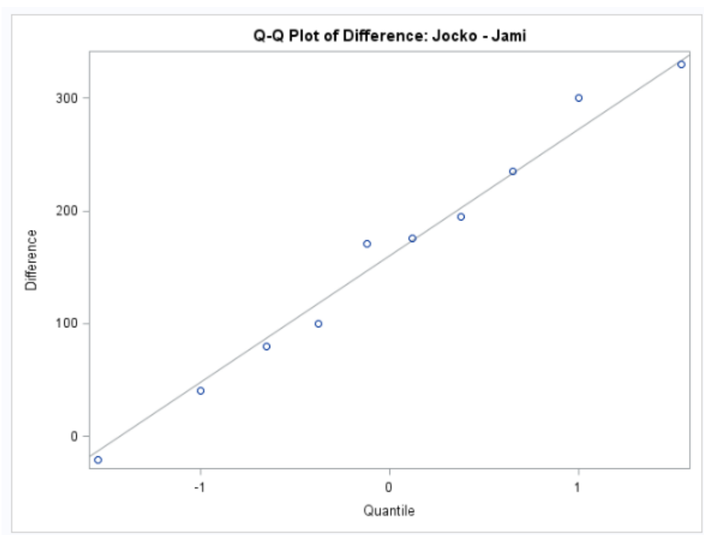
N	Mean	Std Dev	Std Err	Minimum	Maximum
10	160.6	112.0	35.4157	-21.0000	330.0

Mean	95% CL Mean		Std Dev	95% CL Std Dev	
160.6	80.4841	240.7	112.0	77.0336	204.5

DF	t Value	Pr > |t|
9	4.53	0.0014


Paired Profiles for (Jocko, Jami)
1000
800
600
400
200
Mean
Jocko
Jami


Agreement of Jami and Jocko
Mean
Jami
Jocko
200
400
600
800
1000


Q-Q Plot of Difference: Jocko - Jami
Difference
300
200
100
0
-1
0
1
Quantile

One of the two scenarios below has paired data where looking at paired differences makes sense and on scenario has a case where that does not make any sense (even though paired differences for each are given). Identify which of the two scenarios below has paired data - for the example with paired data find a 95% confidence interval (state assumptions needed on the data, how you would inspect the assumption, and interpret the interval): Some values - $P(T_9 > 1.83) = 0.05$ $P(T_9 > 2.26) = 0.025$ $P(T_{22} > 1.72) = 0.05$ $P(T_{22} > 2.07) = 0.025$

1. A nutrition expert is examining a weight loss program to evaluate its effectiveness (i.e., if participants lose weight on the program). Ten subjects are randomly selected for the investigation. Each subjects initial weight is recorded, they follow the program for 6 weeks, and they are again weighed. Is the program effective?
   The data are given below:

Subject	Initial Weight	Final Weight
1	180	165
2	142	138
3	126	128
4	138	136
5	175	170
6	205	197
7	116	115
8	142	128
9	157	144
10	136	130

The UNIVARIATE Procedure
Variable: FminusI

Moments			
N	10	Sum Weights	10
Mean	-6.6	Sum Observations	-66
Std Deviation	5.8156876	Variance	33.8222222
Skewness	-0.2343677	Kurtosis	-1.1697528
Uncorrected SS	740	Corrected SS	304.4
Coeff Variation	-88.116479	Std Error Mean	1.8390819

2. A manufacturer of cat food wants to assure that the packages being produced at the Tennessee plant have the same average weight as the packages being produced at the Wisconsin plant. Samples of 23 packages each were collected from Tennessee plant and Wisconsin plant respectively. The package weights (in ounces) are given below:

Sample	Tennessee	Wisconsin
1	4.67	4.74
2	4.65	4.65
3	4.68	4.60
4	4.59	4.62
⋮	⋮	⋮
23	4.66	4.62

The UNIVARIATE Procedure
Variable: Tenn_Wisc

Moments			
N	23	Sum Weights	23
Mean	-0.0008696	Sum Observations	-0.02
Std Deviation	0.08564796	Variance	0.00733557
Skewness	-0.327649	Kurtosis	-1.2283775
Uncorrected SS	0.1614	Corrected SS	0.16138261
Coeff Variation	-9849.5154	Std Error Mean	0.01785883

Output from SAS to conduct the paired t-test on the weight example.

```
*Conduct paired t-test;
proc ttest data=weight;
paired Final*Initial;
run;
```

The TTEST Procedure

Difference: final - initial

N	Mean	Std Dev	Std Err	Minimum	Maximum
10	-6.6000	5.8157	1.8391	-15.0000	2.0000

Mean	95% CL Mean		Std Dev	95% CL Std Dev	
-6.6000	-10.7603	-2.4397	5.8157	4.0002	10.6172

DF	t Value	Pr > |t|
9	-3.59	0.0059

6.2 - Inference for Two Independent Samples (Two-Sample t)

For the second example on the previous page, we did not have paired data, but rather two samples, one from the Tennessee population and one from the Wisconsin population.

The UNIVARIATE Procedure
Variable: Tenn

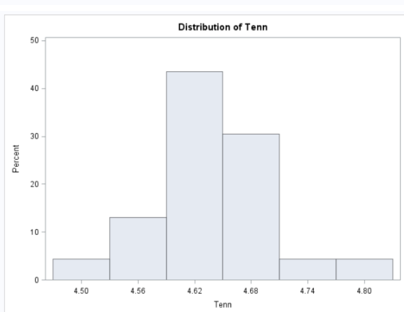
Moments			
N	23	Sum Weights	23
Mean	4.63304348	Sum Observations	106.56
Std Deviation	0.06483789	Variance	0.00420395
Skewness	0.61292717	Kurtosis	1.39541642
Uncorrected SS	493.7896	Corrected SS	0.09248696
Coeff Variation	1.39946657	Std Error Mean	0.01351964

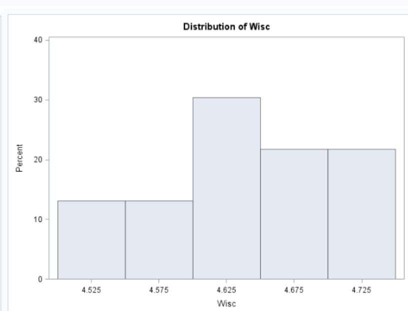
Basic Statistical Measures			
Location		Variability	
Mean	4.633043	Std Deviation	0.06484
Median	4.630000	Variance	0.00420
Mode	4.590000	Range	0.30000
		Interquartile Range	0.08000

The UNIVARIATE Procedure
Variable: Wisc

Moments			
N	23	Sum Weights	23
Mean	4.63391304	Sum Observations	106.58
Std Deviation	0.06058346	Variance	0.00367036
Skewness	-0.142007	Kurtosis	-0.7243252
Uncorrected SS	493.9632	Corrected SS	0.08074783
Coeff Variation	1.30739313	Std Error Mean	0.01263252

Basic Statistical Measures			
Location		Variability	
Mean	4.633913	Std Deviation	0.06058
Median	4.630000	Variance	0.00367
Mode	4.620000	Range	0.22000
		Interquartile Range	0.09000





Define:
Y_i = the weight for the i^{th} randomly selected package from the Tennessee plant
X_i = the weight for the i^{th} randomly selected package from the Wisconsin plant

μ_1 = the mean weights for Tennessee plants
μ_2 = the mean weights for Wisconsin plants

Question of interest (Claim):

What could we do to make inference here?

An 'unbiased' estimate of μ_d is

What is the variance of this quantity?

Let's define the ____________________ between two random variables. $Cov(X, Y)$ is a measure the how the random variables ____

Mathematically:

$Cov(X, Y) = E(XY) - E(X)E(Y)$ - Similar to $Var(X) = E(X^2) - (E(X))^2 = E(XX) - E(X)E(X)$

Generally, for the random variable $aX + bY$ we have

$$Var(aX + bY) = a^2 Var(X) + b^2 Var(Y) + 2ab Cov(X, Y)$$

Since covariance is a measure of how the RV's vary together. If X is independent of Y that means

This implies that if X is independent of Y then $Cov(X, Y) = 0$.

Now back to our quantity $\bar{D}$, what is the variance of this quantity?

Knowing the mean and variance of this quantity is useful, but to use it for inference we must know the

Theorem: If $Y_i \sim^{iid} N(\mu_1, \sigma^2)$ and $X_i \sim^{iid} N(\mu_2, \sigma^2)$ (both parent populations are normal, same variance) where all Y are **independent** of all X (independent samples) then

We can estimate the common variance by

Thus, the **test statistic** we can use for our HT and CI are ex: Back to the catfood example. Let us assume that $Y_i \sim^{iid} N(\mu_1, \sigma^2)$ and $X_i \sim^{iid} N(\mu_2, \sigma^2)$ where $Y's$ and $X's$ are independent (that is, our parent populations are independent normals with equal variance assumed). Let's conduct a hypothesis test at the 0.01 level to determine if the mean weights differ. Would a 99% CI for μ_{diff} contain 0? Why/why not?

Analysis of cat food data using SAS

```
proc ttest data=catfood2;
*Specify that location is categorical;
class location;
*variable that we want to test on;
var weight;
run;
```

The TTEST Procedure

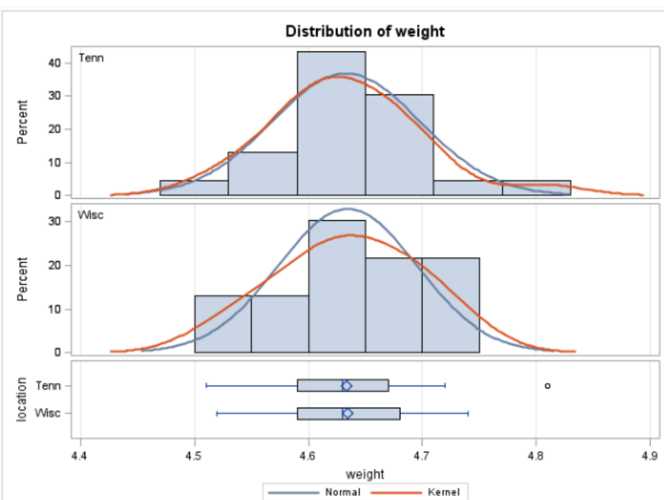
Variable: weight

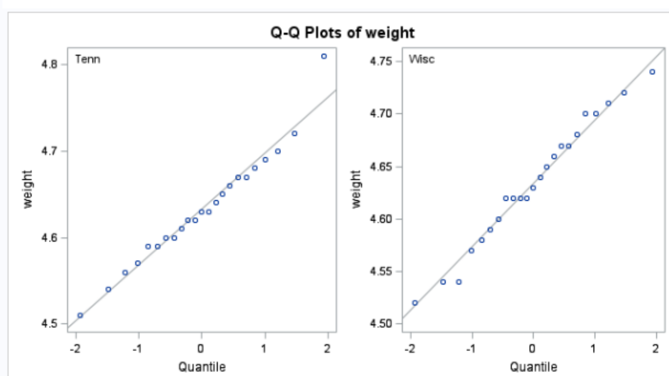
location	N	Mean	Std Dev	Std Err	Minimum	Maximum
Tenn	23	4.6330	0.0648	0.0135	4.5100	4.8100
Wisc	23	4.6343	0.0604	0.0126	4.5200	4.7400
Diff (1-2)		-0.00130	0.0627	0.0185		

location	Method	Mean	95% CL Mean		Std Dev	95% CL Std Dev	
Tenn		4.6330	4.6050	4.6611	0.0648	0.0501	0.0918
Wisc		4.6343	4.6082	4.6605	0.0604	0.0467	0.0855
Diff (1-2)	Pooled	-0.00130	-0.0386	0.0359	0.0627	0.0519	0.0792
Diff (1-2)	Satterthwaite	-0.00130	-0.0386	0.0359			

Method	Variances	DF	t Value	Pr > |t|
Pooled	Equal	44	-0.07	0.9441
Satterthwaite	Unequal	43.785	-0.07	0.9441

Equality of Variances				
Method	Num DF	Den DF	F Value	Pr > F
Folded F	22	22	1.15	0.7447





The equal variance assumption seemed reasonable above. What can we do when it is **not** reasonable?

Theorem: If $Y_i \sim^{iid} N(\mu_1, \sigma_1^2)$ and $X_i \sim^{iid} N(\mu_2, \sigma_2^2)$ (both parent populations are normal, different variance) where all Y are **independent** of all X (independent samples) then

Therefore, $\bar{D} = \bar{Y} - \bar{X}$ still is a good statistic to base our inference on.

Suppose we estimate our standard error using the sample variances:

We can create the test statistic

Issue: What are the degrees of freedom for our test statistic??

Satterthwaite's approximation to degrees of freedom
To approximate the df associated with a t statistic based on a standard error of the form

$$SE = \sqrt{c_1 S_1^2 + c_2 S_2^2 + \cdots + c_k S_k^2}$$

(a linear combination of sample variances), use the **Satterthwaite approximation:**

$$\widehat{df} = \frac{(c_1 S_1^2 + c_2 S_2^2 + \cdots + c_k S_k^2)^2}{(c_1 S_1^2)^2/df_1 + (c_2 S_2^2)^2/df_2 + \cdots + (c_k S_k^2)^2/df_k}$$

Always round down!

Example: Consider an experiment involving the comparison of the mean heart rate following 30 minutes of aerobic exercise among females aged 20 to 24 years (Y variable, group 1) as compared to females aged 30-34 years (X variable, group 2). For this experiment, heart rates are recorded on each participant following 30 minutes of intense aerobic exercise. The sample data and some statistics (not all will be needed) are given below:
$n_1 = 15$, $\bar{y}$=150.22, $s_1^2 = 160$
$n_2 = 10$, $\bar{x} = 141.10$, $s_2^2 = 100$

$$\widehat{SE}\left(\bar{Y}-\bar{X}\right) = \sqrt{\frac{(n_1-1)s_1^2+(n_2-1)s_2^2}{n_1+n_2-2}\left(\frac{1}{n_1}+\frac{1}{n_2}\right)} = \sqrt{\frac{(15-1)160+(10-1)100}{15+10-2}(1/15+1/10)} = 4.768$$

$$\widehat{SE}\left(\bar{Y}-\bar{X}\right) = \sqrt{\frac{s_1^2}{n_1}+\frac{s_2^2}{n_2}} = \sqrt{\frac{160}{15}+\frac{100}{10}} = 4.55$$

$$\widehat{df} = \frac{\left(\frac{s_1^2}{n_1}+\frac{s_2^2}{n_2}\right)^2}{\left(\frac{s_1^2}{n_1}\right)^2/(n_1-1)+\left(\frac{s_2^2}{n_2}\right)^2/(n_2-1)} = \frac{\left(\frac{160}{15}+\frac{100}{10}\right)^2}{\left(\frac{160}{15}\right)^2/(15-1)+\left(\frac{100}{10}\right)^2/(10-1)} = 22.20$$

$$P(T_{23} > 2.50) = 0.01 \quad P(T_{23} > 2.81) = 0.005 \quad P(T_{22} > 2.51) = 0.01 \quad P(T_{22} > 2.82) = 0.005$$

Conduct a hypothesis test at the $\alpha = 0.01$ level assuming the variances of the two population are not equal. Be sure to show all steps (use RR, state the assumptions that must be made and how you would check that assumption). Also, create a 99% confidence interval for the difference in means.

Analysis of heart rate data using SAS

```
proc ttest data=heartrate;
*denote group as a categorical variable;
class group;
var rate;
run;
```

The TTEST Procedure

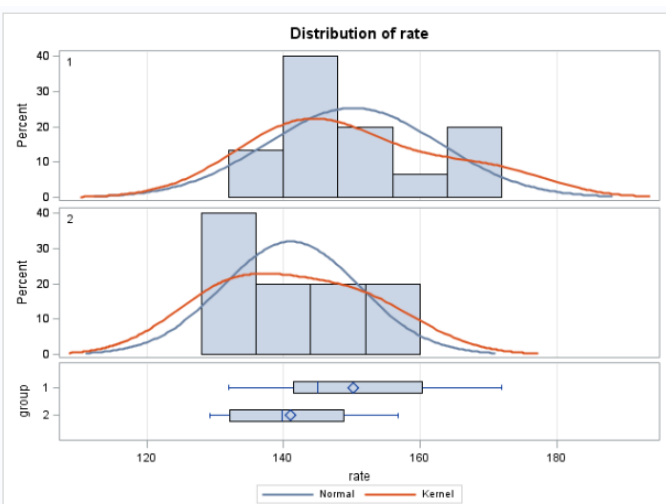
Variable: rate

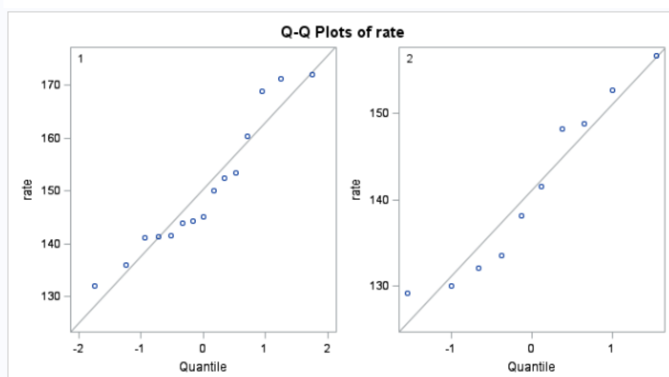
group	N	Mean	Std Dev	Std Err	Minimum	Maximum
1	15	150.2	12.6500	3.2662	132.1	171.9
2	10	141.1	10.0004	3.1624	129.2	156.7
Diff (1-2)		9.1190	11.6849	4.7704		

group	Method	Mean	95% CL Mean		Std Dev	95% CL Std Dev	
1		150.2	143.2	157.2	12.6500	9.2614	19.9503
2		141.1	133.9	148.3	10.0004	6.8786	18.2568
Diff (1-2)	Pooled	9.1190	-0.7492	18.9872	11.6849	9.0817	16.3912
Diff (1-2)	Satterthwaite	9.1190	-0.3045	18.5425			

Method	Variances	DF	t Value	Pr > |t|
Pooled	Equal	23	1.91	0.0685
Satterthwaite	Unequal	22.202	2.01	0.0572

Equality of Variances				
Method	Num DF	Den DF	F Value	Pr > F
Folded F	14	9	1.60	0.4833





Recap of possible inferences for the difference of means based on the normal distribution:

Paired Data: Assume **differences** are a RS and normally distributed
100(1-α)% CI for μ_d is

$$\overline{D} \pm t_{\alpha/2,n-1} S_D/\sqrt{n} = \bar{Y} - \bar{X} \pm t_{\alpha/2,n-1} S_{\bar{Y}-\bar{X}}/\sqrt{n}$$

HT: for $H_0 : \mu_d = \Delta_0 \;\; vs \;\; H_a : \mu_d > \Delta_0 \;\; or \;\; \mu_d < \Delta_0 \;\; or \;\; \mu_d \neq \Delta_0$

$$\text{Test Statistic: } T = \frac{\bar{Y} - \bar{X} - \Delta_0}{S_d/\sqrt{n}}$$

$$RR : \{t_{obs} : t_{obs} > t_{\alpha,n-1}\} \quad or \quad \{t_{obs} : t_{obs} < -t_{\alpha,n-1}\} \quad or \quad \left\{t_{obs} : |t_{obs}| > t_{\alpha/2,n-1}\right\}$$
$$P - value : P(T_{n-1} > t_{obs}) \quad or \quad P(T_{n-1} < t_{obs}) \quad or \quad 2 * P(T_{n-1} > |t_{obs}|)$$

Independent Samples: Assume populations are independent RS's with each population having a normal distribution
Equal Variance (Pooled Variance):
100(1-α)% CI for μ_d is

$$\bar{Y} - \bar{X} \pm t_{\alpha/2,n_1+n_2-2} \sqrt{\frac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{n_1+n_2-2} \left(\frac{1}{n_1} + \frac{1}{n_2}\right)}$$

HT: for $H_0 : \mu_d = \Delta_0 \;\; vs \;\; H_a : \mu_d > \Delta_0 \;\; or \;\; \mu_d < \Delta_0 \;\; or \;\; \mu_d \neq \Delta_0$

$$\text{Test Statistic: } T = \frac{\bar{Y} - \bar{X} - \Delta_0}{\sqrt{\frac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{n_1+n_2-2} \left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

$$RR : \{t_{obs} : t_{obs} > t_{\alpha,n_1+n_2-2}\} \quad or \quad \{t_{obs} : t_{obs} < -t_{\alpha,n_1+n_2-2}\} \quad or \quad \left\{t_{obs} : |t_{obs}| > t_{\alpha/2,n_1+n_2-2}\right\}$$
$$P - value : P(T_{n_1+n_2-2} > t_{obs}) \quad or \quad P(T_{n_1+n_2-2} < t_{obs}) \quad or \quad 2 * P(T_{n_1+n_2-2} > |t_{obs}|)$$

Unequal Variance:
100(1-α)% CI for μ_d is

$$\bar{Y} - \bar{X} \pm t_{\alpha/2,\widehat{df}} \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

HT: for $H_0 : \mu_d = \Delta_0 \;\; vs \;\; H_a : \mu_d > \Delta_0 \;\; or \;\; \mu_d < \Delta_0 \;\; or \;\; \mu_d \neq \Delta_0$

$$\text{Test Statistic: } T = \frac{\bar{Y} - \bar{X} - \Delta_0}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

$$RR : \left\{t_{obs} : t_{obs} > t_{\alpha,\widehat{df}}\right\} \quad or \quad \left\{t_{obs} : t_{obs} < -t_{\alpha,\widehat{df}}\right\} \quad or \quad \left\{t_{obs} : |t_{obs}| > t_{\alpha/2,\widehat{df}}\right\}$$
$$P - value : P(T_{\widehat{df}} > t_{obs}) \quad or \quad P(T_{\widehat{df}} < t_{obs}) \quad or \quad 2 * P(T_{\widehat{df}} > |t_{obs}|)$$

$$\widehat{df} = \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\left(\frac{s_1^2}{n_1}\right)^2/(n_1-1) + \left(\frac{s_2^2}{n_2}\right)^2/(n_2-1)}$$

Chapter 7

ST 511 - Inferences About Variances

Readings: Chapter 7 (for 7.4 read if interested)

We saw in the 2-sample t-test we may have interest in testing if two population variances are equal (i.e. $\sigma_1^2 = \sigma_2^2$).

To investigate this, we first start by looking at inference for a single population variance.

Inference for σ^2

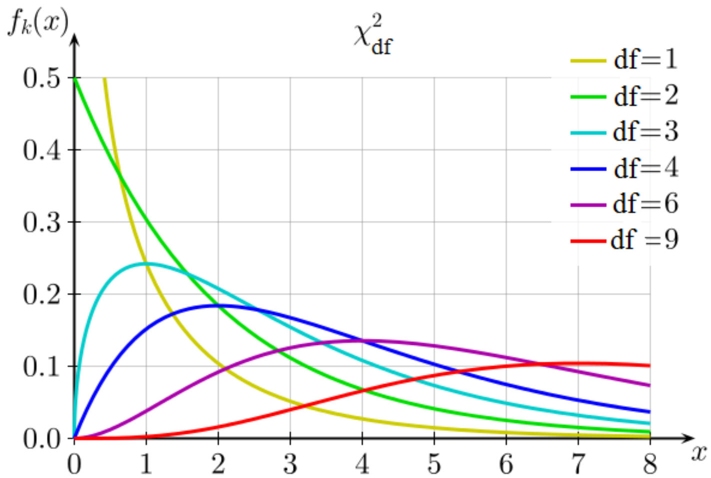
To make inference for σ^2 we need a corresponding statistic...

This is 'unbiased' for σ^2,

To create a CI or HT, we need to know the ____

Theorem: If $Y_i \sim^{iid} N(\mu, \sigma^2)$ (i.e. a RS from a normal parent population) then

Note: Large n will not relax this assumption! We must have the assumed normality here!



Mean = df, Variance = 2(df)

How can we make a $(1-\alpha)100\%$ CI for σ^2?

To get the χ^2_L or χ^2_U values in SAS we can do the following:

```
*Syntax     PROBCHI(x,df)        P(Chi^2_df <x) = returned value
The The PROBCHI function returns the probability that an observation from a chi-square distribution, with
degrees of freedom df is less than or equal to x. This function accepts a noninteger degrees of freedom
parameter df if needed);

*Syntax     QUANTILE(dist, probability, parm-1,...,parm-k)       P(dist<= value returned) = probability
The QUANTILE function computes the probability from various continuous and discrete distributions
'probability' is a numeric constant, variable, or expression that specifies the value of a random variable.
parm-1,...,parm-k are optional shape, location, or scale parameters appropriate for the specific distribution.;


*Find some probabilities and quantile values from a chi-square;
data chisq;
prob1 = probchi(2,2); *Probability chi-sq 2 is less than its mean --- P(Chi^2_2<2);
prob2 = probchi(12.8,4); *P(Chi^2_4<12.8);

quant1 = quantile('chisq',0.95,11); *0.95 quantile from a chi^2_11;
quant2 = quantile('chisq',0.99,15); *0.99 quantile from a chi^2_15;
run;

proc print data=chisq;
title 'Chi-Square values';
run;
```

Chi-Square values

Obs	prob1	prob2	quant1	quant2
1	0.63212	0.98770	19.6751	30.5779

Example: A dairy processing company claims that the variance of the amount of fat in the whole milk processed by the company is 0.25 g^2. You collect a sample of 41 milk containers and find a sample variance of 0.27 g^2. Find a 90% CI for σ^2 = true variance of the amount of fat in the company's whole milk. What do you think of the company's claim? Useful values: $P(\chi^2_{40} > 55.758) = 0.05, P(\chi^2_{40} > 26.509) = 0.95$

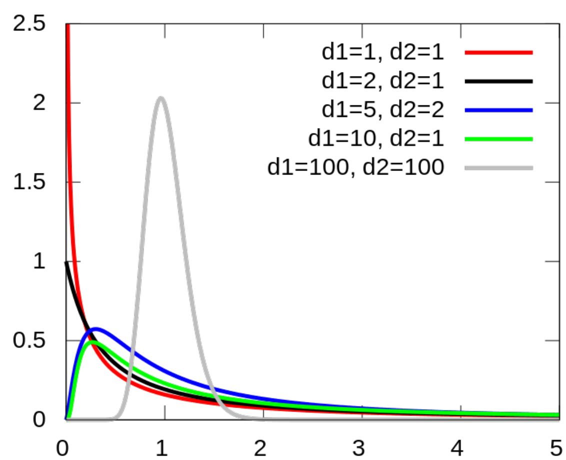
A hypothesis test for $\sigma^2 = \sigma_0^2$ could be done using the test statistic $\chi^2 = \frac{(n-1)s^2}{\sigma_0^2}$. We won't cover this in class.

Both the CI and the HT rely heavily on the normality assumption. If normality does not hold then the interval and test will not be valid!!! In fact, they perform very poorly (they are not robust to this assumption being violated).

Inference for two variances, σ_1^2 and σ_2^2

Now we are ready to compare two variances (as is needed in the two-sample t-test).

Theorem: If $Y_i \sim^{iid} N(\mu_1, \sigma_1^2)$ $(i = 1, ..., n_1)$ and $X_i \sim^{iid} N(\mu_2, \sigma_2^2)$ $(i = 1, ..., n_2)$ where the Y's and X's are independent then



Notice, when comparing variances we are looking at the ratio $\frac{\sigma_1^2}{\sigma_2^2}$ rather than $\sigma_1^2 - \sigma_2^2$. This is because we know the distribution of the statistic above which involves ratios rather than differences. What value is of interest for this ratio?

How can we make a $(1-\alpha)100\%$ CI for σ^2?

To get the F_L or F_U values in SAS we can do the following:

```
*Syntax    PROBF(x,ndf,ddf)    P(F_df1,df2<x) = returned value
The PROBF function returns the probability that an observation from an F distribution,
with numerator degrees of freedom ndf (our df1) and denominator degrees of freedom ddf (our df2)
is less than or equal to x.

*Find some probabilities and quantile values from an F distribution;
data f;
prob1 = probf(10,4,2); *Probability F_4,2 is less than 10 --- P(F_4,2<10);
prob2 = 1-probf(22,3,18); *P(F_3,18>22);

quant1 = quantile('f',0.95,4,2); *0.95 quantile from an F_4,2;
quant2 = quantile('f',0.99,3,18); *0.99 quantile from an F_3,18;
run;

proc print data=f;
title 'F values';
run;
```

F values

Obs	prob1	prob2	quant1	quant2
1	0.90703	.000003016	19.2468	5.09189

Example: A company is comparing methods for producing pipes and wants to choose the method with the least variability. It has taken a sample of the lengths of the pipes using both methods and found the following data and summaries. Find a 99% CI for the ratio of the variances. Values: $P(F_{11,14} > 4.508) = 0.005, P(F_{11,14} > 0.196) = 0.995, P(F_{14,11} > 5.103) = 0.005, P(F_{14,11} > 0.222) = 0.995$

The UNIVARIATE Procedure
Variable: Width

Method=A

Moments			
N	12	Sum Weights	12
Mean	4.06666667	Sum Observations	48.8
Std Deviation	0.88557463	Variance	0.78424242
Skewness	-0.0526034	Kurtosis	-1.6531257
Uncorrected SS	207.08	Corrected SS	8.62666667
Coeff Variation	21.7764253	Std Error Mean	0.25564338

Basic Statistical Measures			
Location		Variability	
Mean	4.066667	Std Deviation	0.88557
Median	4.050000	Variance	0.78424
Mode	3.100000	Range	2.50000
		Interquartile Range	1.70000

The UNIVARIATE Procedure
Variable: Width

Method=B

Moments			
N	15	Sum Weights	15
Mean	4.38	Sum Observations	65.7
Std Deviation	0.65159146	Variance	0.42457143
Skewness	0.15901561	Kurtosis	-1.1801064
Uncorrected SS	293.71	Corrected SS	5.944
Coeff Variation	14.8765173	Std Error Mean	0.16824019

Basic Statistical Measures			
Location		Variability	
Mean	4.380000	Std Deviation	0.65159
Median	4.400000	Variance	0.42457
Mode	3.500000	Range	2.00000
		Interquartile Range	1.20000

The hypothesis test for the ratio of the variances is summarized below:

Null	Alternative	Test Stat	RR
$H_0 : \sigma_1^2 \leq \sigma_2^2$ $H_0 : \sigma_1^2/\sigma_2^2 \leq 1$	$H_A : \sigma_1^2 > \sigma_2^2$ $H_A : \sigma_1^2/\sigma_2^2 > 1$	S_1^2/S_2^2	$\{F_{obs} : F_{obs} \geq F_{\alpha,df1,df2}\}$
$H_0 : \sigma_1^2 = \sigma_2^2$ $H_0 : \sigma_1^2/\sigma_2^2 = 1$	$H_A : \sigma_1^2 \neq \sigma_2^2$ $H_A : \sigma_1^2/\sigma_2^2 \neq 1$	S_1^2/S_2^2	$\{F_{obs} : F_{obs} \geq F_{\alpha/2,df1,df2} \;\; or \;\; F_{obs} \leq F_{1-\alpha/2,df1,df2}\}$

Example: Recall heartrate example from chapter 6. Conduct an HT for equality of variance at the 0.05 level. $P(F_{14,10} > 3.798) = 0.025, P(F_{14,10} > 0.316)0.975$

The TTEST Procedure

Variable: rate

group	N	Mean	Std Dev	Std Err	Minimum	Maximum
1	15	150.2	12.6500	3.2662	132.1	171.9
2	10	141.1	10.0004	3.1624	129.2	156.7
Diff (1-2)		9.1190	11.6849	4.7704		

group	Method	Mean	95% CL Mean		Std Dev	95% CL Std Dev	
1		150.2	143.2	157.2	12.6500	9.2614	19.9503
2		141.1	133.9	148.3	10.0004	6.8786	18.2568
Diff (1-2)	Pooled	9.1190	-0.7492	18.9872	11.6849	9.0817	16.3912
Diff (1-2)	Satterthwaite	9.1190	-0.3045	18.5425			

Method	Variances	DF	t Value	Pr > |t|
Pooled	Equal	23	1.91	0.0685
Satterthwaite	Unequal	22.202	2.01	0.0572

Equality of Variances				
Method	Num DF	Den DF	F Value	Pr > F
Folded F	14	9	1.60	0.4833

setcounterchapter7

Chapter 8

ST 511 - Analysis of Variance for Comparing Means

Readings: Chapter 8 (read 8.1-8.4)

We now turn our focus back to comparing means from different populations. In chapter 6, we saw how to compare the (true) means from two (roughly) normal populations (both with equal and unequal variance).

We may however have interest in comparing the (true) means from more than two populations, say t populations. This type of question often comes up when conducting a completely randomized experiment (CRD).

In a CRD we have N total units (the book uses n_T to represent this number). We have t treatments - recall a treatment is a specific experimental condition which, in an experiment, we assign to the experimental units. This treatment may come from the levels of a single factor or the combinations of levels from several factors.

Let n_i denote the number of replicates for treatment i (i=1,...,t) (i.e. the # of units assigned to that treatment). In a balanced CRD design, we have $n_i = n$ for every treatment i. Thus, the total number of units can be given by $N = nt$ (or in the book notation $n_T = nt$).

The CRD design randomly assigns the treatments to the units (treating every unit as interchangeable).

Therefore, we can consider having t different populations that we now want to compare. For instance, we may want to know if the means are equal for each population (or the standard deviations, medians, etc.)

Example: (some description taken from Goosen, 2014)
Consider having 24 pieces of cheese. Color of the cheese is important in terms of consumer satisfaction. We have interest in how the color differs for 4 different types of corn syrup (26, 42, 55, and 62) (4 treatments). A CRD design is decided upon and we randomly assign each corn syrup type to 6 pieces of cheese (6 replicates for each treatment).

As a response, we measure the color using a 3 part CIE L*a*b* Color System.

- 'L' reflects the lightness of a sample, from black (L = 0) to white (L = 100) and runs from top to bottom.
- 'a' defines the shades from red (positive values) to green (negative values).
- 'b' defines the shades from yellow (positive values) to blue (negative values).

All three of these could be treated as responses (and analyzed together), but for our purposes we will only look at the 'L' response variable.

Again, we will focus on the means of the population. How might we make inference here?

Define

- μ_1 = mean 'L' score for **all** pieces of cheese that with corn syrup 26.
- μ_2 = mean 'L' score for **all** pieces of cheese that with corn syrup 42.
- μ_3 = mean 'L' score for **all** pieces of cheese that with corn syrup 55.
- μ_4 = mean 'L' score for **all** pieces of cheese that with corn syrup 62.

We want to test the hypotheses

$$H_0 : \mu_1 = \mu_2 = \mu_3 = \mu_4 \quad vs \quad H_A : \text{at least one mean differs}$$

For two independent samples, we said if the samples came from normal populations with equal variance we could use

$$T = \frac{\bar{Y} - \bar{X}}{\sqrt{S_p^2(1/n_1 + 1/n_2)}} \quad \text{where} \quad S_p^2 = \frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2}$$

as a test statistic to make our inference. Now we have more than 2 populations, so this exact set-up won't work, but we can do something else.

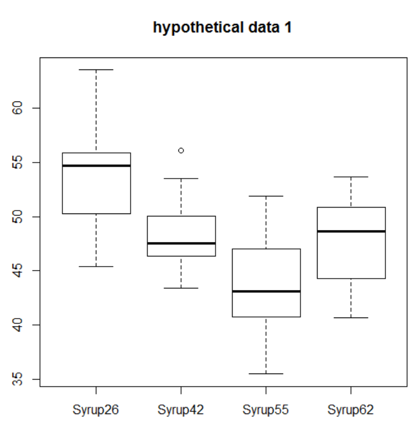
ANOVA for analyzing a CRD

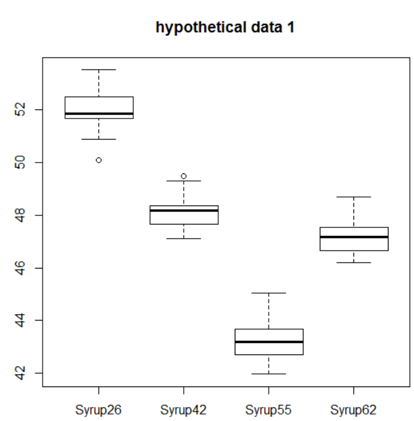
Data and labeling:

Corn Syrup	Replicate #	'L' measurement	Label
26	1	51.89	y_{11}
26	2	51.52	y_{12}
26	3	52.69	y_{13}
26	4	52.06	y_{14}
26	5	51.63	y_{15}
26	6	52.73	y_{16}
42	1	47.21	y_{21}
42	2	48.57	y_{22}
42	3	47.57	y_{23}
42	4	46.85	y_{24}
42	5	48.64	y_{25}
42	6	47.49	y_{26}
55	1	41.43	y_{31}
55	2	42.31	y_{32}
55	3	42.31	y_{33}
55	4	41.49	y_{34}
55	5	42.12	y_{35}
55	6	42.65	y_{36}
62	1	45.99	y_{41}
62	2	46.66	y_{42}
62	3	47.35	y_{43}
62	4	45.83	y_{44}
62	5	46.77	y_{45}
62	6	47.88	y_{46}

We now need two subscripts to represent which observation we are talking about. The first subscript (i) represents the treatment group, where as the second subscript (j) represent the replicate number.

Consider the following two hypothetical set of boxplots for this data. Which would give evidence that the (true) means differ?





ANOVA = Analysis of Variance

In this case, compare variation 'within' groups to variation 'between' groups.

Assumptions:

$$Y_{1j} \sim^{iid} N(\mu_1, \sigma^2)$$
$$Y_{2j} \sim^{iid} N(\mu_2, \sigma^2)$$
$$Y_{3j} \sim^{iid} N(\mu_3, \sigma^2)$$
$$Y_{4j} \sim^{iid} N(\mu_4, \sigma^2)$$

and each sample is independent of one another.

That is, the populations are independent random samples from normally distributed parent populations with equal variances. Rather than write this all out we will just say

$$Y_{ij} \sim^{iid} N(\mu_i, \sigma^2)$$

Within group variation

In two samples, to estimate the common variance σ^2 we used S_p^2. Here we use the same exact idea:

$$MS(E) = MS(W) = S_w^2$$

$$= \frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2 + ... + (n_t - 1)S_t^2}{n_1 + n_2 + ... + n_t - t}$$

For a balanced design we have

$$MS(E) = MS(W) = S_w^2 = \frac{(n-1)(S_1^2 + ... + S_t^2)}{N - t}$$

$$= \frac{(n-1)(S_1^2 + ... + S_t^2)}{nt - t} = \frac{S_1^2 + ... + S_t^2}{t}$$

(i.e. just the simple average of the variances).

With the double subscript our formula for S_i^2 is given by

$$S_i^2 = \frac{\sum_{j=1}^{n}(Y_{ij} - \bar{Y}_{i\bullet})^2}{n-1}$$

where the group mean $\bar{Y}_{i\bullet}$ is given by

$$\bar{Y}_{i\bullet} = \frac{\sum_{j=1}^{n} Y_{ij}}{n}$$

Thus, for a balanced design $MS(E)$ is written

$$MS(E) = \frac{\sum_{i=1}^{t}\sum_{j=1}^{n}(Y_{ij} - \bar{Y}_{i\bullet})^2}{t(n-1)}$$

syrup=26

Analysis Variable : l				
N	Mean	Std Dev	Minimum	Maximum
6	52.0866667	0.5190247	51.5200000	52.7300000

syrup=42

Analysis Variable : l				
N	Mean	Std Dev	Minimum	Maximum
6	47.7216667	0.7295592	46.8500000	48.6400000

syrup=55

Analysis Variable : l				
N	Mean	Std Dev	Minimum	Maximum
6	42.0516667	0.4895066	41.4300000	42.6500000

syrup=62

Analysis Variable : l				
N	Mean	Std Dev	Minimum	Maximum
6	46.7466667	0.7834964	45.8300000	47.8800000

Figure 8.1: summary from proc means. - proc means data=cheese; by syrup; var L; run;

Between group variation

Variation between groups is judged by the variation between the group means:

$$MS(T) = MS(B) = S_b^2 = \frac{\sum_{i=1}^{t}\sum_{j=1}^{n}(\bar{Y}_{i\bullet} - \bar{Y}_{\bullet\bullet})^2}{t-1}$$

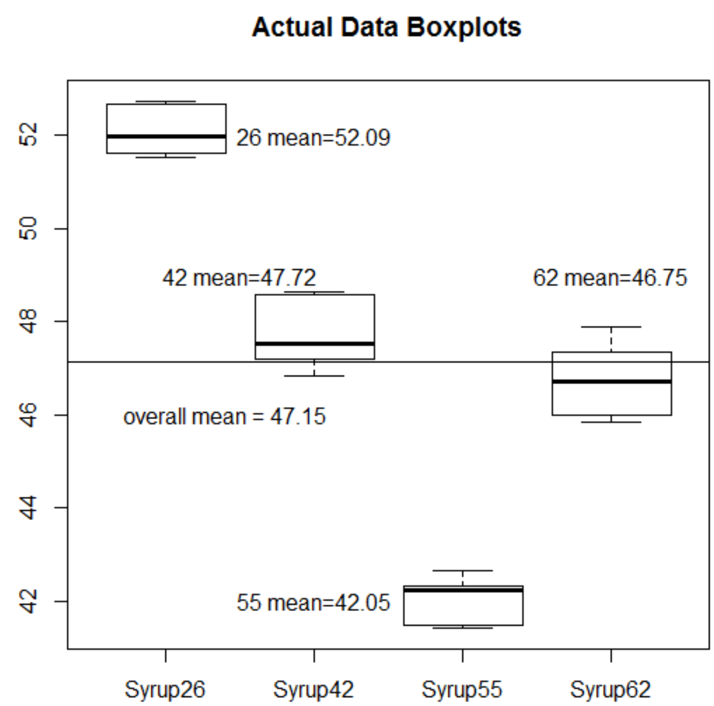
where $\bar{Y}_{\bullet\bullet}$ is the overall mean

$$\bar{Y}_{\bullet\bullet} = \frac{\sum_{i=1}^{t}\sum_{j=1}^{n} Y_{ij}}{nt}$$

The MEANS Procedure

Analysis Variable : l				
N	Mean	Std Dev	Minimum	Maximum
24	47.1516667	3.6913208	41.4300000	52.7300000

Figure 8.2: summary from proc means proc means data=cheese; var L; run;



For our hypotheses:

$$H_0 : \mu_1 = \mu_2 = \mu_3 = \mu_4 \quad vs \quad H_A : \text{at least one mean differs}$$

Test statistic is:

$$F = \frac{MS(T)}{MS(E)} = \frac{S_b^2}{S_w^2} \sim F_{t-1,N-t}$$

We reject H_0 for large values of F, values greater than $F_{t-l,N-t,\alpha}$.

We get a p-value by $P(F_{t-1,N-t} > F_{obs})$.

To get this analysis in SAS we can run the code:

```
proc anova data=cheese;
      class syrup;
      model L = syrup;
      means syrup/tukey;
run;
```

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	305.1189000	101.7063000	245.80	<.0001
Error	20	8.2756333	0.4137817		
Corrected Total	23	313.3945333			

R-Square	Coeff Var	Root MSE	l Mean
0.973594	1.364233	0.643259	47.15167

Source	DF	Anova SS	Mean Square	F Value	Pr > F
syrup	3	305.1189000	101.7063000	245.80	<.0001

Alpha	0.05
Error Degrees of Freedom	20
Error Mean Square	0.413782
Critical Value of Studentized Range	3.95825
Minimum Significant Difference	1.0395

Means with the same letter are not significantly different.			
Tukey Grouping	Mean	N	syrup
A	52.0867	6	26
B	47.7217	6	42
B			
B	46.7467	6	62
C	42.0517	6	55

How do we view the ANOVA 'model'?

We made the assumption that $Y_{ij} \sim^{iid} N(\mu_i, \sigma^2)$. Instead of viewing it in this form, we look at it as

$$Y_{ij} = \mu_i + E_{ij}$$

where $E_{ij} \sim^{iid} N(0, \sigma^2)$. So for $i = 1$ we have

$$Y_{1j} = \mu_1 + E_{1j}$$

which is adding the constant μ_1 to the $N(0, \sigma^2)$ distribution, giving $Y_{1j} \sim^{iid} N(\mu_1, \sigma^2)$.

Usually we then change this to a different parameterization -

$$Y_{ij} = \mu + \tau_i + E_{ij}$$

where μ is the overall (grand) mean, τ_i is the effect for the i^{th} treatment, and $E_{ij} \sim^{iid} N(0, \sigma^2)$.

So for $i = 1$ we have

$$Y_{1j} = \mu + \tau_1 + E_{1j}$$

which is adding the constant $\mu + \tau_1$ to the $N(0, \sigma^2)$ distribution, giving $Y_{1j} \sim^{iid} N(\mu + \tau_1, \sigma^2)$.

Relationship between parameterizations (Note: we assume $\sum_{i=1}^{t} \tau_i = 0$ for *identifiability* purposes):

Population	Mean 1^{st} way	Mean 2^{nd} way	Variance
1	μ_1	$\mu + \tau_1$	σ^2
2	μ_2	$\mu + \tau_2$	σ^2
3	μ_3	$\mu + \tau_3$	σ^2
4	μ_4	$\mu + \tau_4$	σ^2

Our hypothesis now becomes

$$H_0 : \tau_1 = \tau_2 = ... = \tau_t = 0 \quad vs \quad H_A : \text{at least one differs}$$

We then analyze the model using an analysis of variance table (ANOVA table). **Table for balanced one-way ANOVA:**

Source	DF	SS	MS	F
Treatments	$t-1$	$SS(T)$	$MS(T) = \frac{SS(T)}{(t-1)}$	$F = \frac{MS(T)}{MS(E)}$
Error	$t(n-1)$	$SS(E)$	$MS(E) = \frac{SS(E)}{(N-t)}$	
Total	$nt-1$	$SS(TOT)$		

where

$$\begin{aligned} SS(T) &= \sum_{i=1}^{t}\sum_{j=1}^{n}(\bar{y}_{i\bullet} - \bar{y}_{\bullet\bullet})^2 = n\sum_{i=1}^{t}(\bar{y}_{i\bullet} - \bar{y}_{\bullet\bullet})^2 \\ SS(E) &= \sum_{i=1}^{t}\sum_{j=1}^{n}(y_{ij} - \bar{y}_{i\bullet})^2 \\ SS(Tot) &= \sum_{i=1}^{t}\sum_{j=1}^{n}(y_{ij} - \bar{y}_{\bullet\bullet})^2 \end{aligned}$$

Note: SS(T) is also called SS(Between) and SS(E) is also called SS(Within).

We can now see the overall idea of ANOVA.

Consider $SS(Tot)$ (sum of squares total or total sum of squares), which, if we divide by $nt - 1$ we get the sample variance of the y's (over all the treatments). This is a measure of the variation in the response.

We take this $SS(Tot)$ and 'partition' it into a piece due to the different 'sources' in our model. Here the sources are the treatments and error (about the treatments). Thus

$$SS(Tot) = SS(T) + SS(E)$$

Similarly, the degrees of freedom add up.

$$df_{Tot} = df_T + df_E \quad or \quad nt - 1 = t(n-1) + (t-1)$$

The sum of squares represent variability from each source. When we divide by the degrees of freedom, this standardizes that measure of variation (and we call this a mean square).

Our test then becomes the ratio of the $MS(T)$ to the $MS(E)$.

$$F = \frac{MS(T)}{MS(E)}$$

Example:
The following example studies the effect of bacteria on the nitrogen content of red clover plants. The treatment factor is bacteria strain, and it has six levels. Red clover plants are inoculated with the treatments, and nitrogen content is later measured in milligrams. The data are derived from an experiment by Erdman (1946) and are analyzed in Chapters 7 and 8 of Steel and Torrie (1980). Conduct a test to determine if the means are equal at the 0.05 level. Be sure to show all 5 steps (use p-values).

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	5	847.046667	169.409333	14.37	<.0001
Error	24	282.928000	11.788667		
Corrected Total	29	1129.974667			

R-Square	Coeff Var	Root MSE	Nitrogen Mean
0.749616	17.26515	3.433463	19.88667

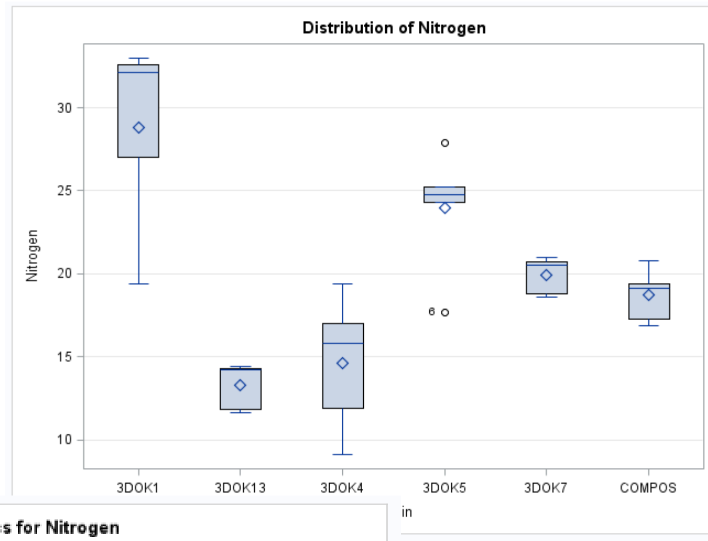
Source	DF	Anova SS	Mean Square	F Value	Pr > F
Strain	5	847.0466667	169.4093333	14.37	<.0001

How do we check our assumptions?

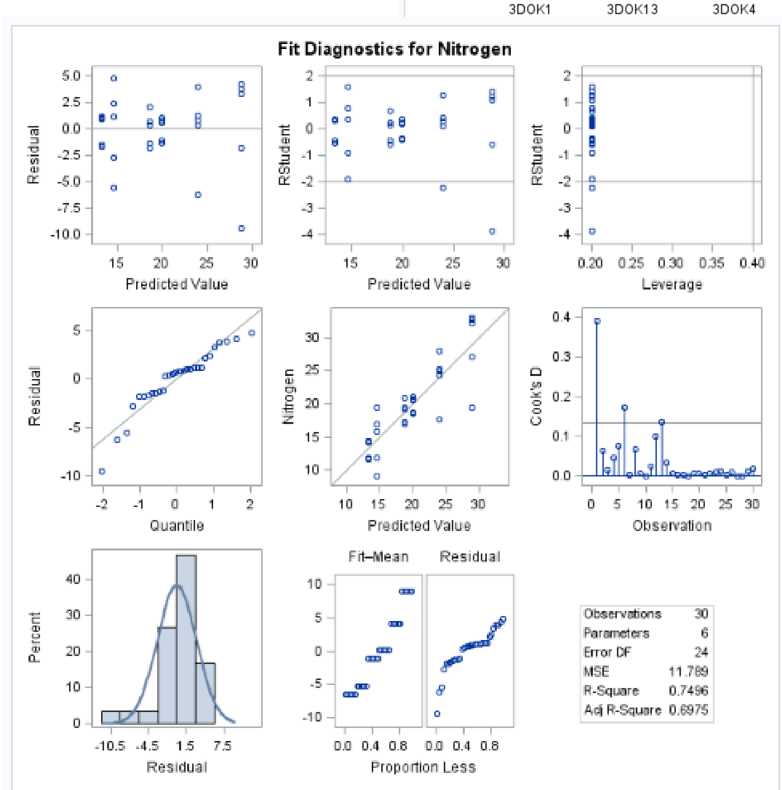
To investigate the constant variance assumption, we can look at side-by-side box plots or residual vs predicted plots. A residual is the observed value minus the predicted value. For the ij^{th} observation the residual is

$$r_{ij} = obs - pred = y_{ij} - \bar{y}_{i\bullet}$$

If the constant variance assumption is violated, one thing we might do is try a transformation of the data. This will be looked at in ST 512.



To check the normality, we look for a straight line in the qq-plot (or quantile vs residual plot) and also we hope to see a roughly normal looking histogram in the bottom left panel. (Note: the transformation idea above may also solve some nonnormality issues.)



If we reject H_0 and conclude that the treatment means differ, the next logical question to ask is which treatment means are the ones that differ.

To answer this question we usually look at all pairwise comparisons of treatment means. That is, if we reject H_0, we would look at

$$\begin{array}{cccc} \mu_1 - \mu_2 & \mu_1 - \mu_3 & \cdots & \mu_1 - \mu_t \\ & \mu_2 - \mu_3 & \cdots & \mu_{t-1} - \mu_t \end{array}$$

to see which differ.

Let's focus on $\mu_1 - \mu_2$. We get an estimator this quantity with the corresponding sample means

$$\hat{\mu}_1 - \hat{\mu}_2 = \bar{Y}_{1\bullet} - \bar{Y}_{2\bullet}$$

The standard error of this quantity can be found by taking the square root of the variance (recall we assume our samples are independent so covariance is 0)

$$\begin{aligned} Var(\bar{Y}_{1\bullet} - \bar{Y}_{2\bullet}) &= (1^2)Var(\bar{Y}_{1\bullet}) + (-1)^2\bar{Y}_{2\bullet} + 2(1)(-1)Cov(\bar{Y}_{1\bullet}, \bar{Y}_{2\bullet}) \\ &= Var(Y_{1j})/n_1 + Var(Y_{2j})/n_2 = \sigma^2/n_1 + \sigma^2/n_2 \end{aligned}$$

For a balanced design we have

$$Var(\bar{Y}_{1\bullet} - \bar{Y}_{2\bullet}) = \sigma^2(1/n + 1/n) = 2\sigma^2/n$$

yielding a standard error of

$$SE(\bar{Y}_{1\bullet} - \bar{Y}_{2\bullet}) = \sqrt{2\sigma^2/n}$$

By the normality assumption on the data we then have a case similar to the two-sample t test with pooled variance!

$$\bar{Y}_{1\bullet} - \bar{Y}_{2\bullet} \sim N(\mu_1 - \mu_2, \sigma^2(1/n_1 + 1/n_2)) = N(\mu_1 - \mu_2, 2\sigma^2/n)$$

We estimate σ^2 by the common pooled estimate (over all the samples, not just these two)

$$MS(E) = S_w^2 = \frac{\sum_{i=1}^{t}\sum_{j=1}^{n}(Y_{ij} - \bar{Y}_{i\bullet})^2}{N - t}$$

Thus, we can for a t-test for testing $H_0 : \mu_1 = \mu_2 \quad vs \quad H_A : \mu_1 \neq \mu_2$ using

$$T = \frac{\bar{Y}_{1\bullet} - \bar{Y}_{2\bullet}}{\sqrt{MS(E)(1/n_1 + 1/n_2)}} = \frac{\bar{Y}_{1\bullet} - \bar{Y}_{2\bullet}}{\sqrt{2MS(E)/n}} \sim t_{N-t}$$

and we can form a $(1-\alpha)100\%$ CI for $\mu_1 - \mu_2$ using

$$\bar{Y}_{1\bullet} - \bar{Y}_{2\bullet} \pm t_{\alpha/2,N-t}\sqrt{MS(E)(1/n_1 + 1/n_2)} \quad or \quad \bar{Y}_{1\bullet} - \bar{Y}_{2\bullet} \pm t_{\alpha/2,N-t}\sqrt{2MS(E)/n}$$

and check if 0 is in the interval.

An issue arises! If we do this for all of the possible pairwise comparisons, we control the type I error rate for each test/interval, but by doing so many tests, the overall probability of making **at least one type I error** may be much more than α. This is known as the multiple comparison correction issue. This will be taken up in ST 512.

```
proc glm data=Clover plots=all;
 class strain;
 model nitrogen = strain;
 lsmeans strain/pdiff cl;
 run;
```

Strain	Nitrogen LSMEAN	95% Confidence Limits	
3DOK1	28.820000	25.650902	31.989098
3DOK13	13.260000	10.090902	16.429098
3DOK4	14.640000	11.470902	17.809098
3DOK5	23.980000	20.810902	27.149098
3DOK7	19.920000	16.750902	23.089098
COMPOS	18.700000	15.530902	21.869098

Least Squares Means for Effect Strain

i	j	Difference Between Means	95% Confidence Limits for LSMean(i)-LSMean(j)	
1	2	15.560000	11.078218	20.041782
1	3	14.180000	9.698218	18.661782
1	4	4.840000	0.358218	9.321782
1	5	8.900000	4.418218	13.381782
1	6	10.120000	5.638218	14.601782
2	3	-1.380000	-5.861782	3.101782
2	4	-10.720000	-15.201782	-6.238218
2	5	-6.660000	-11.141782	-2.178218
2	6	-5.440000	-9.921782	-0.958218
3	4	-9.340000	-13.821782	-4.858218
3	5	-5.280000	-9.761782	-0.798218
3	6	-4.060000	-8.541782	0.421782
4	5	4.060000	-0.421782	8.541782
4	6	5.280000	0.798218	9.761782
5	6	1.220000	-3.261782	5.701782

Least Squares Means for effect Strain
Pr > |t| for H0: LSMean(i)=LSMean(j)
Dependent Variable: Nitrogen

i/j	1	2	3	4	5	6
1		<.0001	<.0001	0.0354	0.0004	<.0001
2	<.0001		0.5311	<.0001	0.0053	0.0194
3	<.0001	0.5311		0.0002	0.0229	0.0738
4	0.0354	<.0001	0.0002		0.0738	0.0229
5	0.0004	0.0053	0.0229	0.0738		0.5794
6	<.0001	0.0194	0.0738	0.0229	0.5794	

Multiple comparison correction needed!