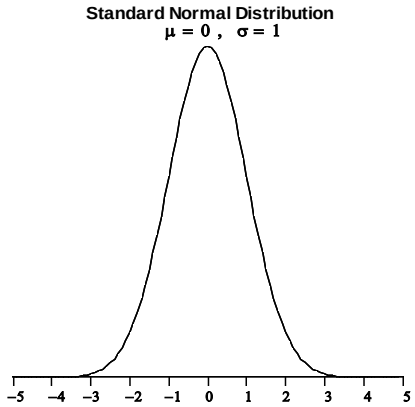
Data Science Enablement Workshop Series

Part 1: Statistical Decision Making

Bradley Novic, Ph.D.
PhaseTwo Analytics, LLC
Data … Knowledge … Intelligence
phasetwoanalytics@gmail.com



Probability Distributions

$$z = \frac{y - \mu}{\sigma}$$

ANOVA

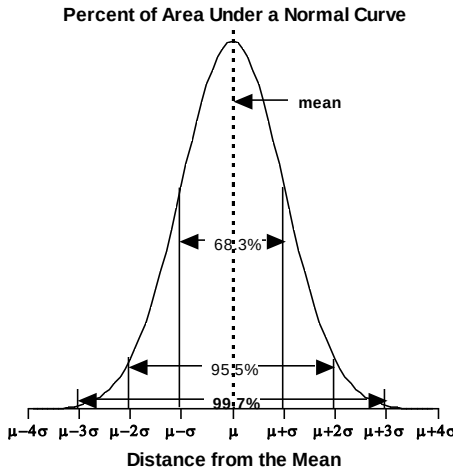
Hypothesis Tests

Statistics For Decision Makers

Power

Confidence

T-tests



PhaseTwo Analytics

Statistics Enables Data Driven Decision Making Under Uncertainty

- Does the distribution of the data from my process follow a Normal distribution?
- Is my process or product capable of meeting a specification?
- Has the mean of my process changed?
- Are the variances of population A & B the same?
- Given my sampling plan, what is the probability that I will reject the Hypothesis that an effect is 0 given that it really is at some effect level δ ?
- Does predictor X have a statistically significant linear or quadratic effect on response Y?
- If I fit a significant 1^{st} order linear model between X and Y, what are limits within which I am 95% confident the next observed X will fall?

Workshop Agenda

- Variability
- Statistical Distributions
  - Discrete vs Continuous
  - Visualizing distributions
  - Common/useful distributions
  - Functional forms
  - Probability calculations
  - The Central Limit Theorem
- Hypothesis Tests of means & variance
  - One sample, Two sample, Paired
  - Confidence Intervals & relation to tests
  - Alpha & Beta errors
  - Power and sample size
  - Testing regression coefficients
- ANOVA (ANalysis Of VAriance)

Preliminaries

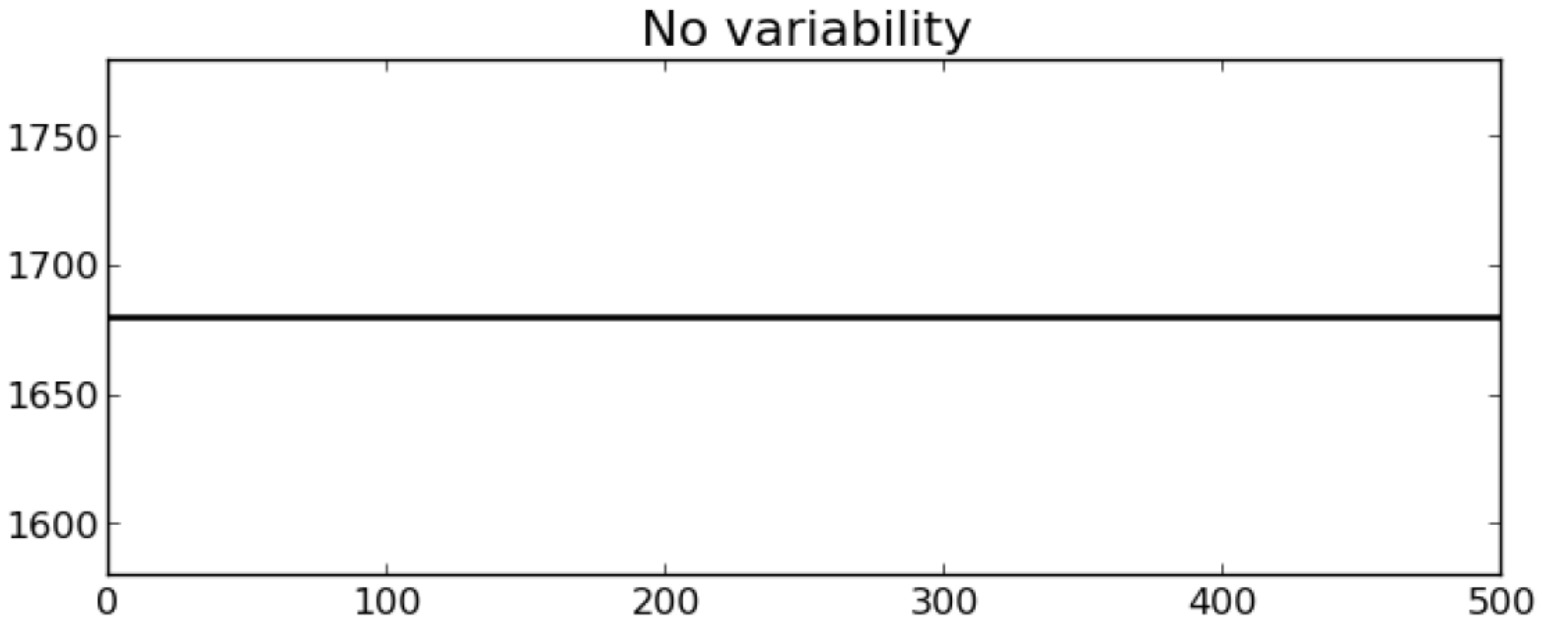
- Check on R & Python requirements
- Check on JMP requirements

- We'll start the class with a review of some basic concepts related to variability, distributions, probability, statistics & visualization … but no calculus (well .. a little ☺)

Data types

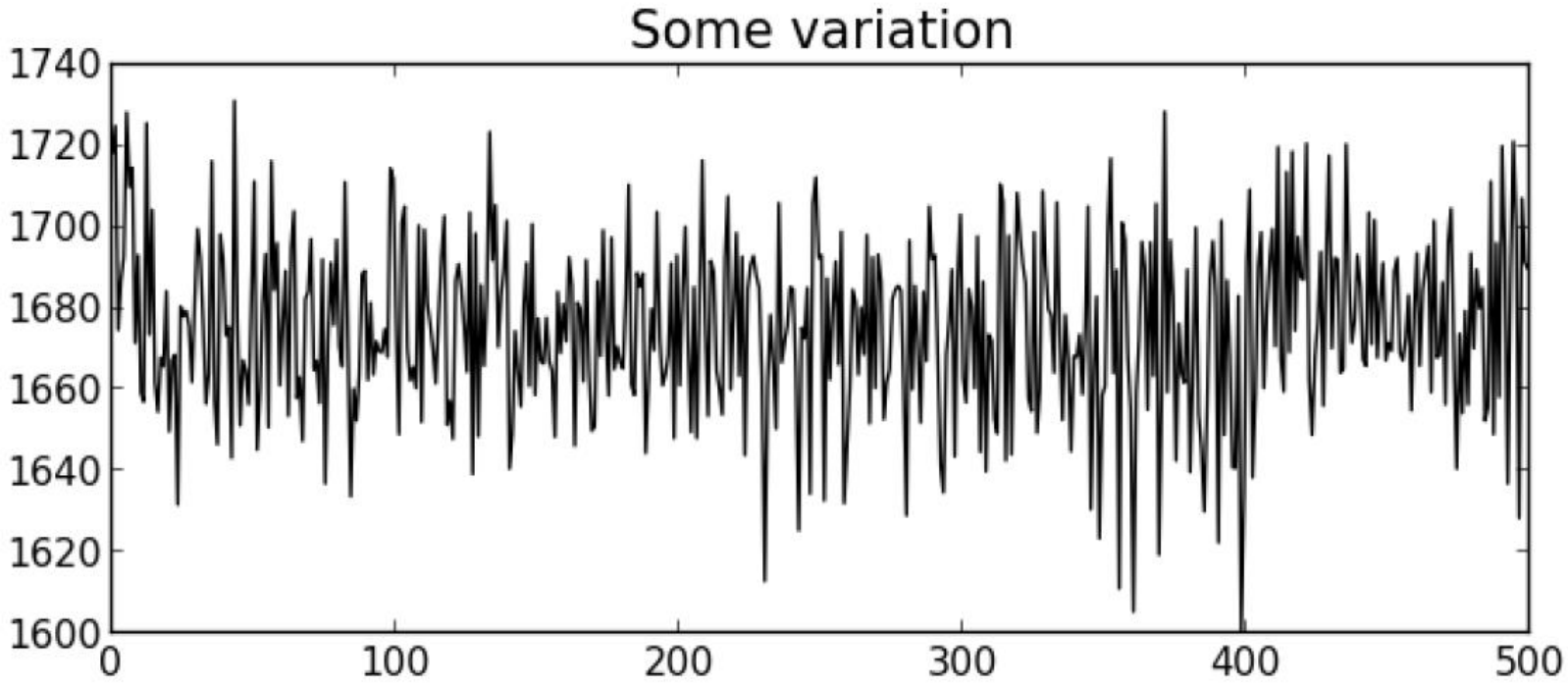
- **Numeric/Continuous:** variables which assume numerical values to an infinite number of significant digits
- **Numeric/Discrete:** variables which assume specific numerical values
- **Categorical/Ordinal:** variable values are categorical, but there is an inherent ranking to the categories; ie. sizes are often small, medium, large, x-large
- **Categorical/Nominal:** variables values are purely categorical; e.g. gender (M/F), Color, Country,

Variability … Observations Vary!!!



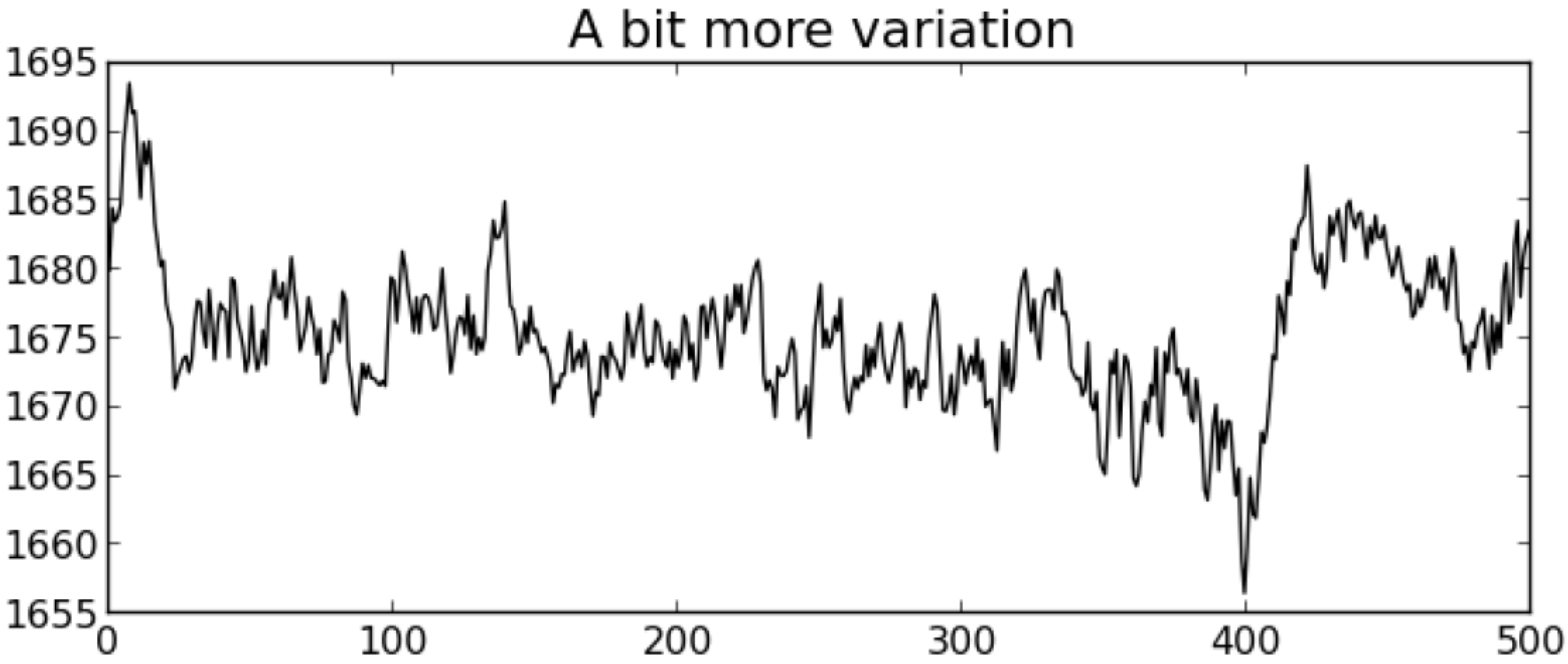
(and this course would be unnecessary)

We have plenty of variability in our recorded data



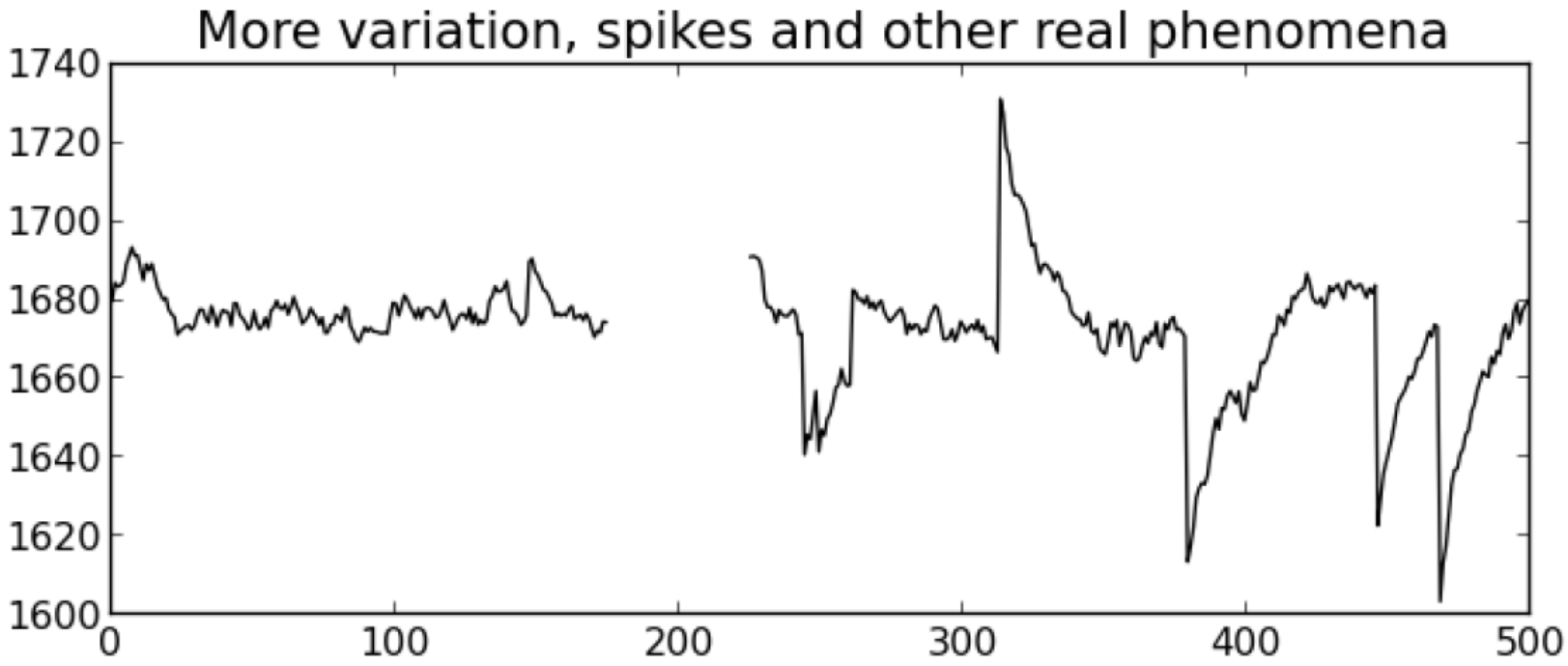
Variability is created from many sources, sometimes intentionally!

- Feedback control: introduces variability
- Operating staff: introduce variability into a process
- Sensor drift, spikes, noise, recalibration shifts, errors in our sample analysis



Variability due to unintentional circumstances

- Raw material properties are not constant
- Production disturbances:
  - external conditions change (ambient temperature, humidity)
  - equipment breaks down, wears out, maintenance shut downs



- AND … measurement system error

RANDOM VARIATION

… unknown causes

… unpredictable

Random Variables and Distributions

- **Random variable :**
  - **a variable whose values are outcomes of a random phenomenon.**
  - **A random variable has a probability distribution, which specifies the probability that its value falls in any given interval.**

Distributions
… quantifying variation

- Wikipedia: a **probability distribution** is a mathematical function that provides the probabilities of occurrence of different possible outcomes of a random variable
  - **Discrete Random Variables**: its distribution can be described by a probability mass function , $p(x_i)$, which assigns a probability to each outcome x_i
  - **Continuous Random Variable**: its distribution can be described by a probability density function or PDF, **f(x),** which assigns probabilities to intervals. Pr($a < X < b$) is the area under the pdf between a & b
  - Outcomes of a random variable produce a data distribution

Probabilities sum(integrate) to 1

- For a discrete rv $\sum_i p(x_i) = 1$

- For a continuous rv $\int f(x)dx = 1$
  (area under the pdf)

Characterizing Distributions

Mean & Variance

μ … **Mean is the center of the distribution**

- For a rv X, $\mu = E[X]$ … the "expected value" of X
- $= \sum_i x_i p(Xi)$ for a discrete distribution with pf P(x)
- $= \int xf(x)$ for a continuous dist with pdf f(x)

σ^2 … **Variance is the 'average' squared distance of any value in the distribution from its mean**

- For a rv X, $\sigma^2 = E[(X- \mu)^2]$
- $= \sum_i (Xi - E[X])2p(Xi)$ for a discrete distribution with pf P(x)
- $= \int (x - E[x])^2 f(x) dx$ for a continuous dist with pdf f(x)

Useful facts about means and variances

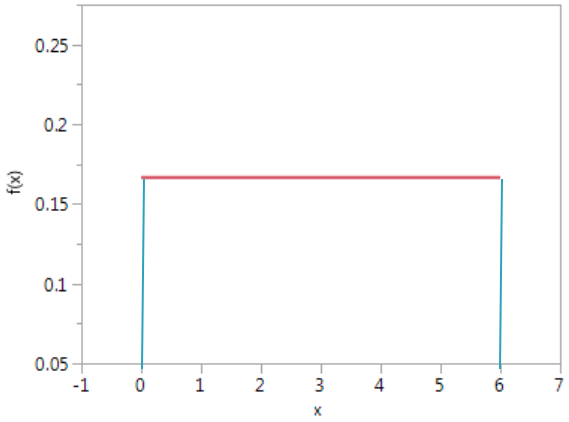
- If X_i is a rv i= 1,n (iid)
  - $E[a_1X_1 + a_2 X_2 + \ldots + a_nX_n] = (a_1+a_2+ \ldots + a_n) E[X_i]$
    - This means $E[\bar{X}] = E[X]$
  - $V[aX] = a^2V[X]$
  - $V[X_1 + X_2] = V[X_1] + V[X_2]$ (assuming the Xs are independent)
    - Note that from the above, $V[\bar{X}] = V[X]/n$

Discrete Distribution Example

- Uniform distribution over integers 1to 5
- The probability function is
  - $P(x=i) = \frac{1}{5}$ for i= 1,5
  - $E[X] = 1(\frac{1}{5}) + 2(\frac{1}{5}) + 3(\frac{1}{5}) + 4(\frac{1}{5}) + 5(\frac{1}{5}) = 3$
  - $V[X] = (1-3)^2(\frac{1}{5}) + (2-3)^2(\frac{1}{5}) + (3-3)^2(\frac{1}{5}) + (4-3)^2(\frac{1}{5}) + (5-3)^2(\frac{1}{5}) = (4+1+0+1+4)/5 = 2$

Continuous Distribution Example

- Uniform continuous distribution from 0 to 6
- f(X) = 1/6 x=0,6
- f(X) = 0 elsewhere



- $E[X] = \int_0^6 xf(x)dx$
- $= \int_0^6 \left(\frac{1}{6}\right) xdx$
- $= \left(\frac{1}{6}\right) (x^2/2) \Big|_0^6$
- $= 3 - 0 = 3$

- $V[X] = \int_0^6 \left(\frac{1}{6}\right) (x-3)^2 dx$
- $= \left(\frac{1}{6}\right) (x-3)^3/3) \Big|_0^6 = 3$

PhaseTwo Analytics

Measures of Center

- Mean (μ or $\bar{X}$)
  - the "balance point" of the distribution; the geometric center
- Median
  - the value at which 50% of the distribution is higher and 50% of the distribution is lower
- Mode
  - the most frequently occurring value of the distribution

Calculating the Sample Mean

The sample mean is calculated as the arithmetic average of the data, by:

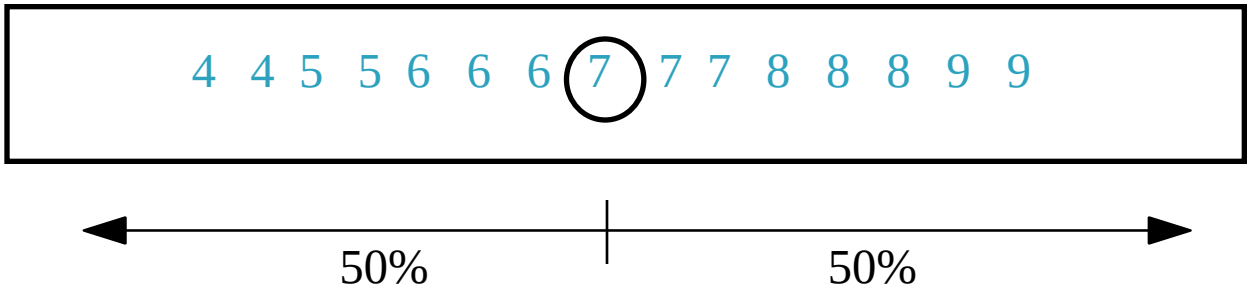
$$\bar{y} = \frac{\sum_{i=1}^{n} y_i}{n} = \frac{y_1 + y_2 + \ldots + y_n}{n}$$

where n is the number of samples

Sample Median

Median . . .

Middle point of the distribution of *ordered* data points



Calculating the Median . . .

For an *odd* number of samples (n):
the value of ordered point number

$$\frac{n+1}{2}$$

For an *even* number of samples (n):
the average value of ordered points

$$\frac{n}{2} \quad \text{and} \quad \frac{n+1}{2}$$

PhaseTwo Analytics

Measures of Dispersion (Spread)

- Standard Deviation (σ or s):
  - the average distance from any point to the mean
- Variance:
  - the standard deviation squared (used for comparing the dispersion between two or more samples)

Sample Standard Deviation and Variance

For a sample of data, the Sample Variance is:

$$s^2 = \frac{\sum_{i=0}^{n} (Xi - \bar{x})^2}{(n-1)} = \text{Sum of Squares/df or SS/df}$$

Sums of squares divided by degrees of freedom are sometimes Referred to by the term "Mean Square" or MS

The sample standard deviation (denoted as s) is calculated as:

$$s = \sqrt{\frac{\sum_{i=0}^{n} (Xi - \bar{x})^2}{(n-1)}}$$, also referred to as the "Root Mean Square"

Populations and Samples

- Population
  - every outcome of the random variable
- Sample
  - a subset of observations taken from the random variable

Parameters Versus Statistics

Parameter

Numerical characteristic of a *population;* denoted with *Greek* letters

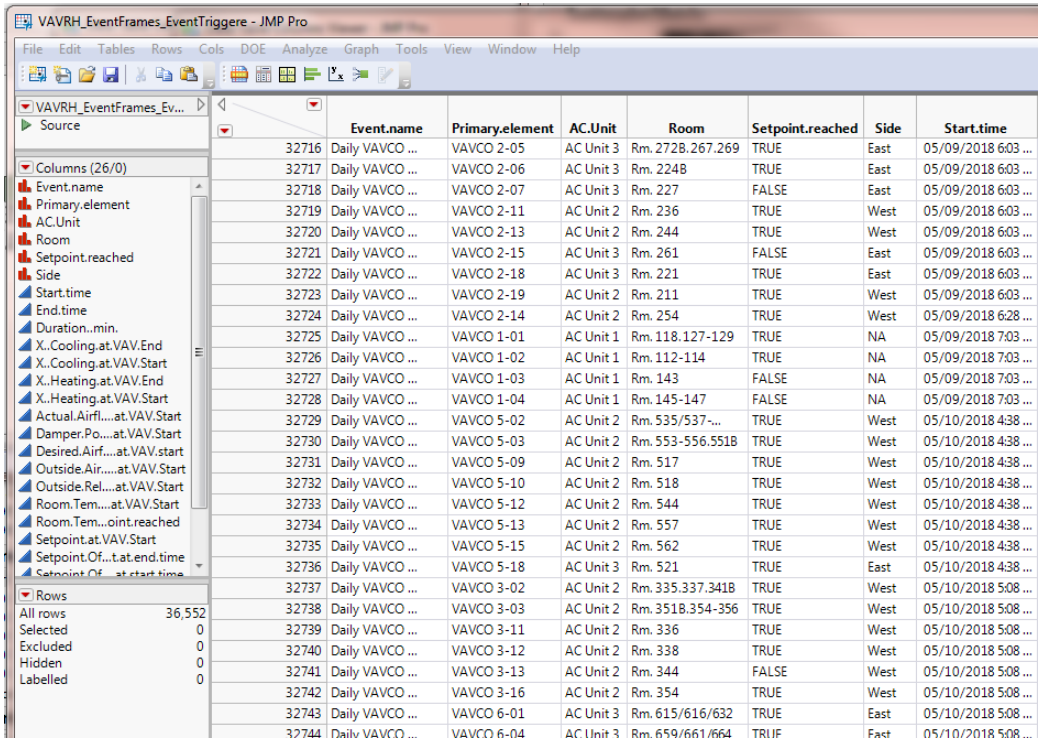
Statistic

Numerical characteristic computed from a population *sample;* denoted with *Latin* letters

Data Characteristic	Population	Sample
data point	X	x_i
mean	μ	$\bar{x}$
standard deviation	σ	s

Exercise 0
familiarization with a dataset

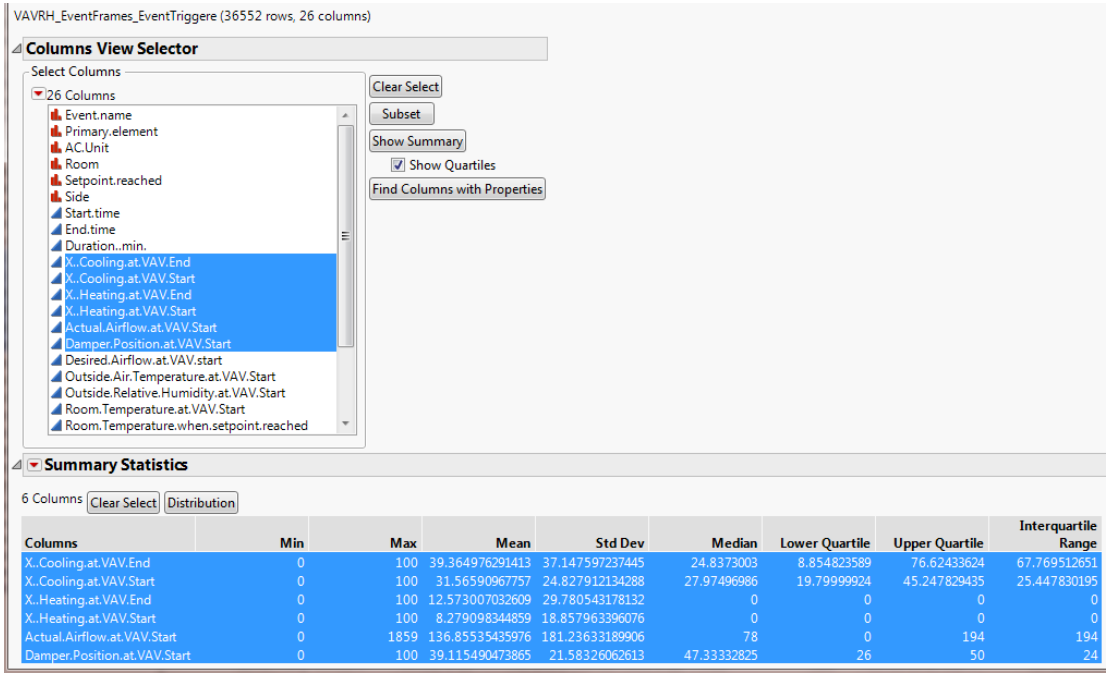
- For the dataset VAVRH_EventFrames_EventTriggere
  - How many rows and columns does it have? …
    - Examine the data table. Row & Column info is listed on the left sidebar
  - What are the names of the columns and their data types? …
    - Examine the data table. Data types are visually displayed (continuous, categorical, nominal)



	Event.name	Primary.element	AC.Unit	Room	Setpoint.reached	Side	Start.time
32716	Daily VAVCO ...	VAVCO 2-05	AC Unit 3	Rm. 272B.267.269	TRUE	East	05/09/2018 6:03 ...
32717	Daily VAVCO ...	VAVCO 2-06	AC Unit 3	Rm. 224B	TRUE	East	05/09/2018 6:03 ...
32718	Daily VAVCO ...	VAVCO 2-07	AC Unit 3	Rm. 227	FALSE	East	05/09/2018 6:03 ...
32719	Daily VAVCO ...	VAVCO 2-11	AC Unit 2	Rm. 236	TRUE	West	05/09/2018 6:03 ...
32720	Daily VAVCO ...	VAVCO 2-13	AC Unit 2	Rm. 244	TRUE	West	05/09/2018 6:03 ...
32721	Daily VAVCO ...	VAVCO 2-15	AC Unit 3	Rm. 261	FALSE	East	05/09/2018 6:03 ...
32722	Daily VAVCO ...	VAVCO 2-18	AC Unit 3	Rm. 221	TRUE	East	05/09/2018 6:03 ...
32723	Daily VAVCO ...	VAVCO 2-19	AC Unit 2	Rm. 211	TRUE	West	05/09/2018 6:03 ...
32724	Daily VAVCO ...	VAVCO 2-14	AC Unit 2	Rm. 254	TRUE	West	05/09/2018 6:28 ...
32725	Daily VAVCO ...	VAVCO 1-01	AC Unit 1	Rm. 118.127-129	TRUE	NA	05/09/2018 7:03 ...
32726	Daily VAVCO ...	VAVCO 1-02	AC Unit 1	Rm. 112-114	TRUE	NA	05/09/2018 7:03 ...
32727	Daily VAVCO ...	VAVCO 1-03	AC Unit 1	Rm. 143	FALSE	NA	05/09/2018 7:03 ...
32728	Daily VAVCO ...	VAVCO 1-04	AC Unit 1	Rm. 145-147	FALSE	NA	05/09/2018 7:03 ...
32729	Daily VAVCO ...	VAVCO 5-02	AC Unit 2	Rm. 535/537-...	TRUE	West	05/10/2018 4:38 ...
32730	Daily VAVCO ...	VAVCO 5-03	AC Unit 2	Rm. 553-556.551B	TRUE	West	05/10/2018 4:38 ...
32731	Daily VAVCO ...	VAVCO 5-09	AC Unit 2	Rm. 517	TRUE	West	05/10/2018 4:38 ...
32732	Daily VAVCO ...	VAVCO 5-10	AC Unit 2	Rm. 518	TRUE	West	05/10/2018 4:38 ...
32733	Daily VAVCO ...	VAVCO 5-12	AC Unit 2	Rm. 544	TRUE	West	05/10/2018 4:38 ...
32734	Daily VAVCO ...	VAVCO 5-13	AC Unit 2	Rm. 557	TRUE	West	05/10/2018 4:38 ...
32735	Daily VAVCO ...	VAVCO 5-15	AC Unit 2	Rm. 562	TRUE	West	05/10/2018 4:38 ...
32736	Daily VAVCO ...	VAVCO 5-18	AC Unit 3	Rm. 521	TRUE	East	05/10/2018 4:38 ...
32737	Daily VAVCO ...	VAVCO 3-02	AC Unit 2	Rm. 335.337.341B	TRUE	West	05/10/2018 5:08 ...
32738	Daily VAVCO ...	VAVCO 3-03	AC Unit 2	Rm. 351B.354-356	TRUE	West	05/10/2018 5:08 ...
32739	Daily VAVCO ...	VAVCO 3-11	AC Unit 2	Rm. 336	TRUE	West	05/10/2018 5:08 ...
32740	Daily VAVCO ...	VAVCO 3-12	AC Unit 2	Rm. 338	TRUE	West	05/10/2018 5:08 ...
32741	Daily VAVCO ...	VAVCO 3-13	AC Unit 2	Rm. 344	FALSE	West	05/10/2018 5:08 ...
32742	Daily VAVCO ...	VAVCO 3-16	AC Unit 2	Rm. 354	TRUE	West	05/10/2018 5:08 ...
32743	Daily VAVCO ...	VAVCO 6-01	AC Unit 3	Rm. 615/616/632	TRUE	East	05/10/2018 5:08 ...
32744	Daily VAVCO ...	VAVCO 6-04	AC Unit 3	Rm. 659/661/664	TRUE	East	05/10/2018 5:08 ...

Exercise 0
familiarization with a dataset

- For the dataset VAVRH_EventFrames_EventTriggere
  - What does some of the data look like?
    - In JMP you work directly with the data table in view and menu above it
  - What are some of the characteristics of the data?
    - Via “Columns Viewer” Menu->Cols->Columns Viewer
    - then highlight columns & click “show quantiles” & click the “Show summary” box

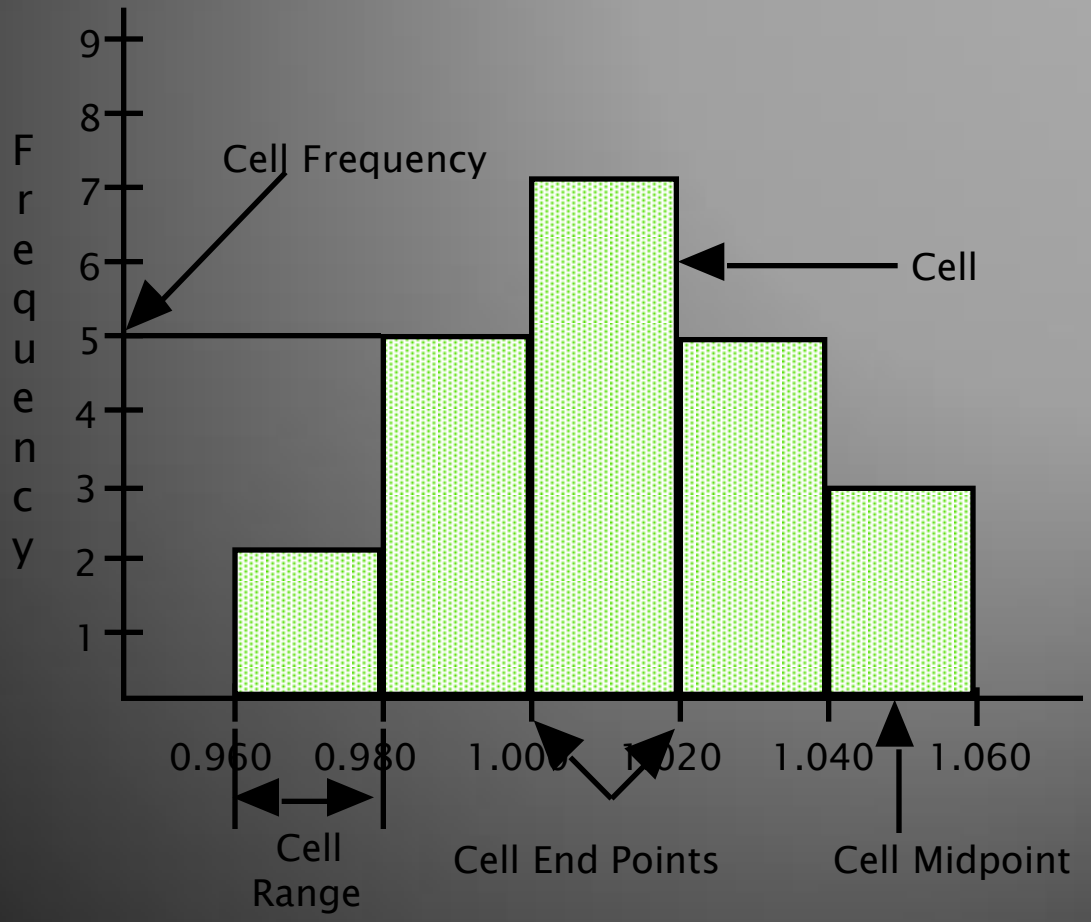


Columns	Min	Max	Mean	Std Dev	Median	Lower Quartile	Upper Quartile	Interquartile Range
X..Cooling.at.VAV.End	0	100	39.364976291413	37.147597237445	24.8373003	8.854823589	76.62433624	67.769512651
X..Cooling.at.VAV.Start	0	100	31.56590967757	24.827912134288	27.97496986	19.79999924	45.247829435	25.447830195
X..Heating.at.VAV.End	0	100	12.573007032609	29.780543178132	0	0	0	0
X..Heating.at.VAV.Start	0	100	8.279098344859	18.857963396076	0	0	0	0
Actual.Airflow.at.VAV.Start	0	1859	136.85535435976	181.23633189906	78	0	194	194
Damper.Position.at.VAV.Start	0	100	39.115490473865	21.58326062613	47.33332825	26	50	24

Visualizing Distributions

- Histograms
- Box (and whisker) plot

Histogram



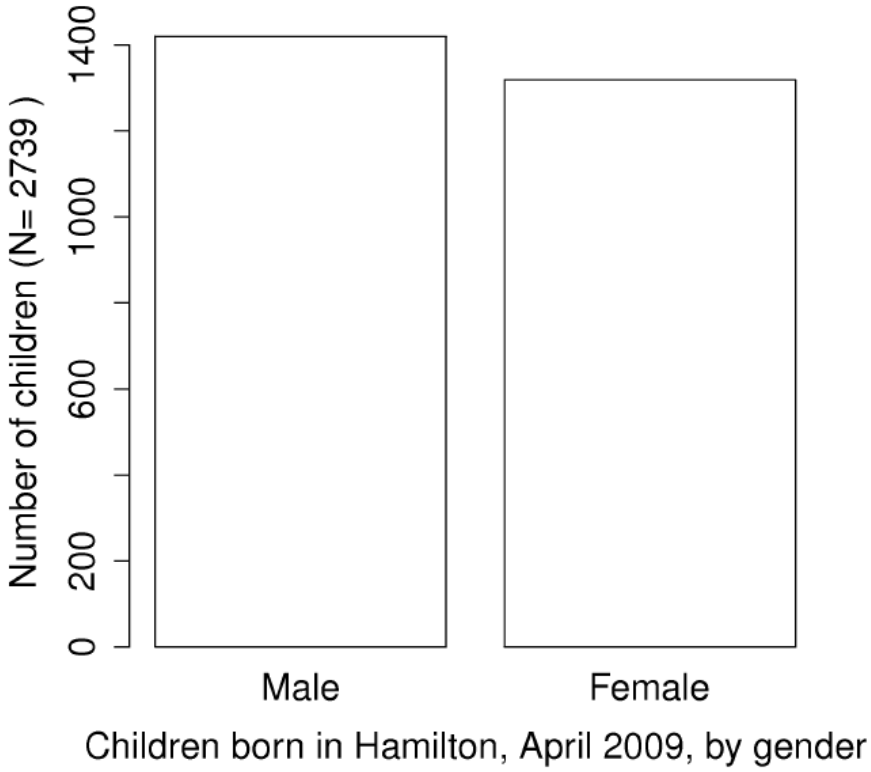
PhaseTwo Analytics

Histogram

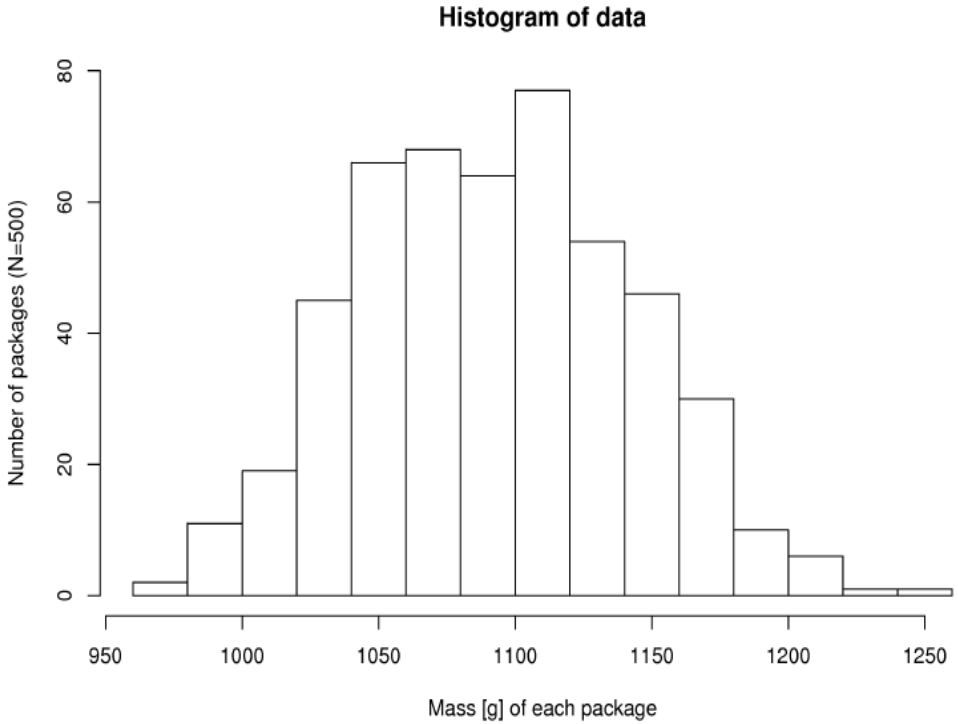
- Graphical display of a data distribution
- Data values are displayed on the x-axis (horizonal axis)
- The range of data values are broken into equally spaced intervals, called cells
- The number or proportion of data points within each interval is displayed as height on the y-axis (vertical axis)

Histograms summarize variation in a measured variable

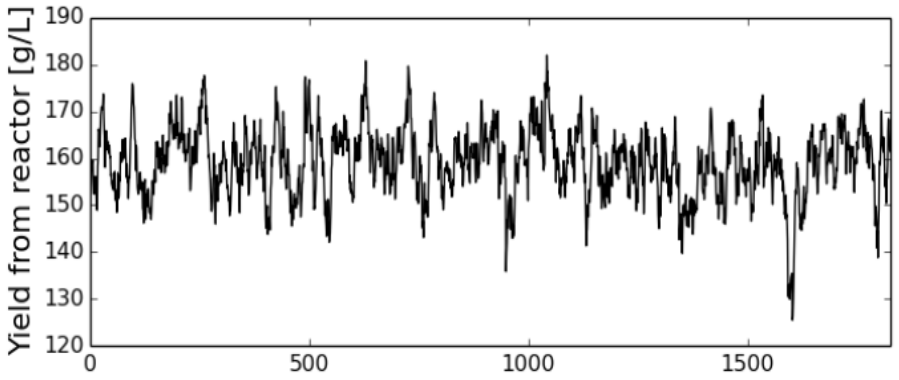
Shows *number* of samples that occur in a **category**: called a **frequency distribution**

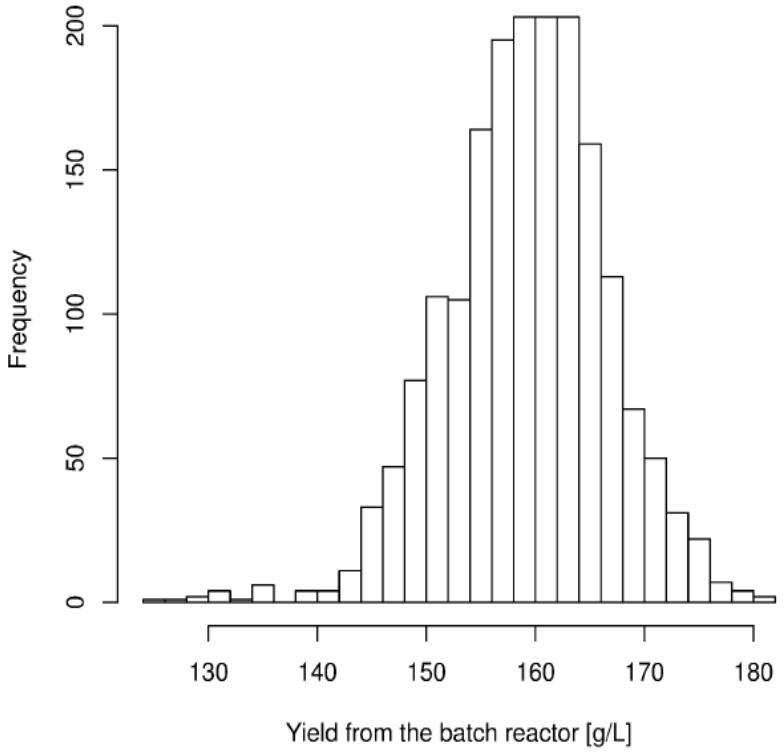


Histograms for a **continuous** variable use category bins (usually of equal size)

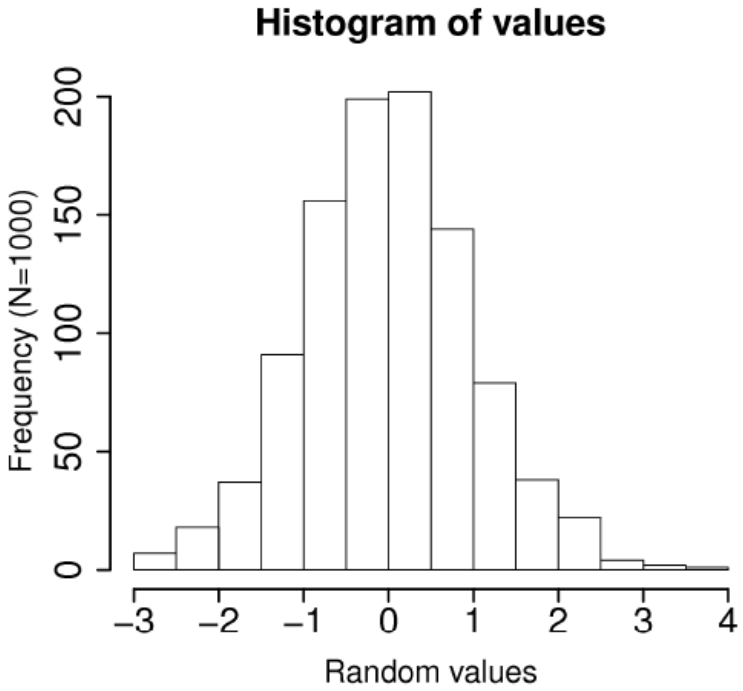


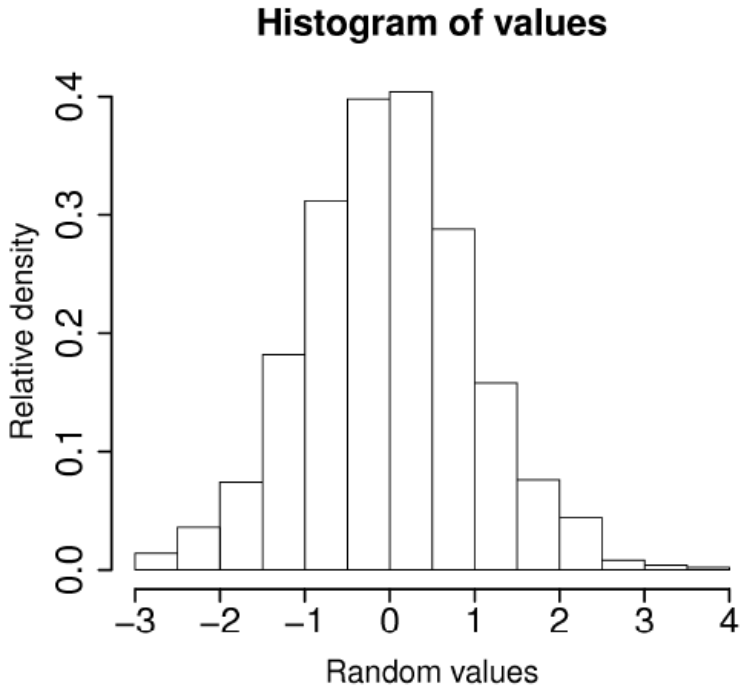
Data from a reactor: production yield, measured in g/L



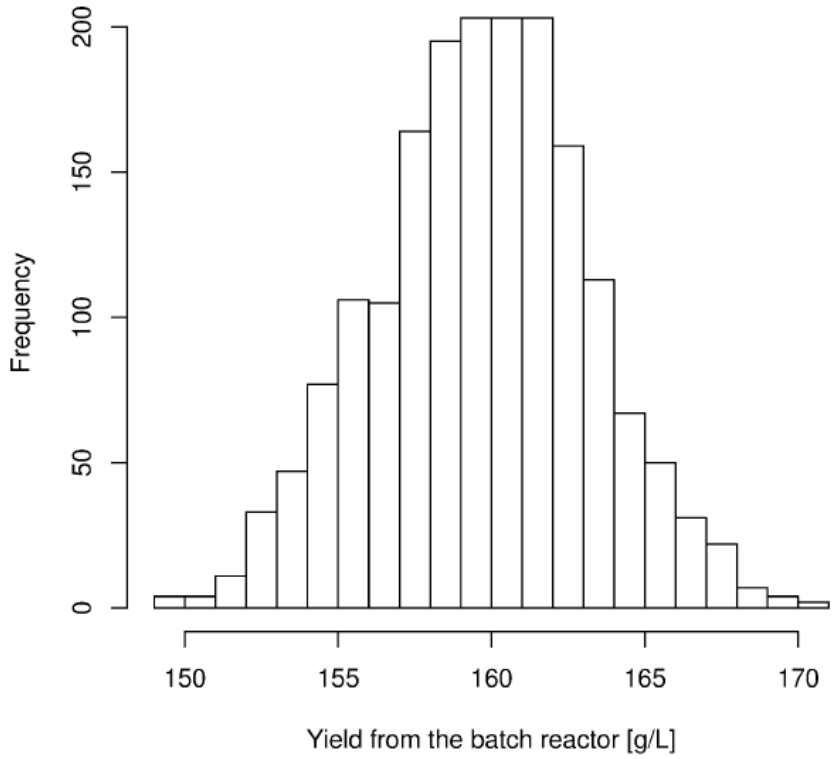


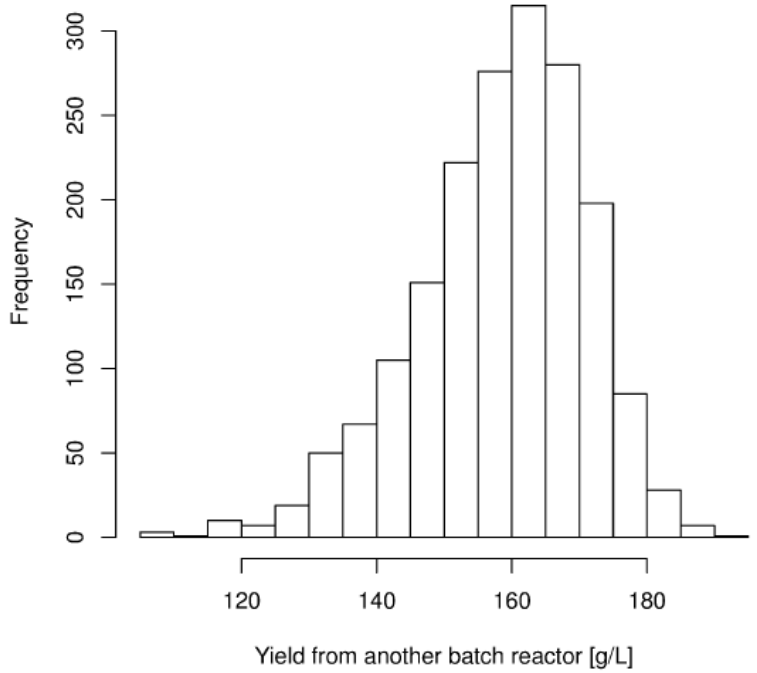
Frequency distribution: you can also use a relative frequency



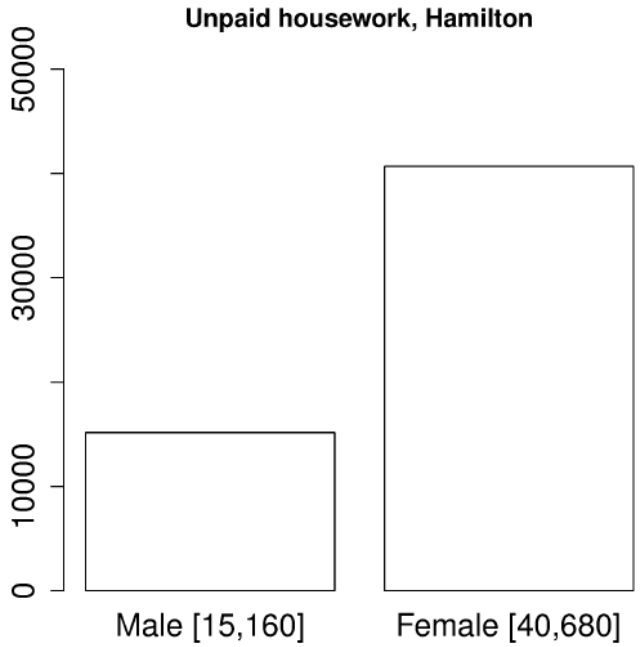


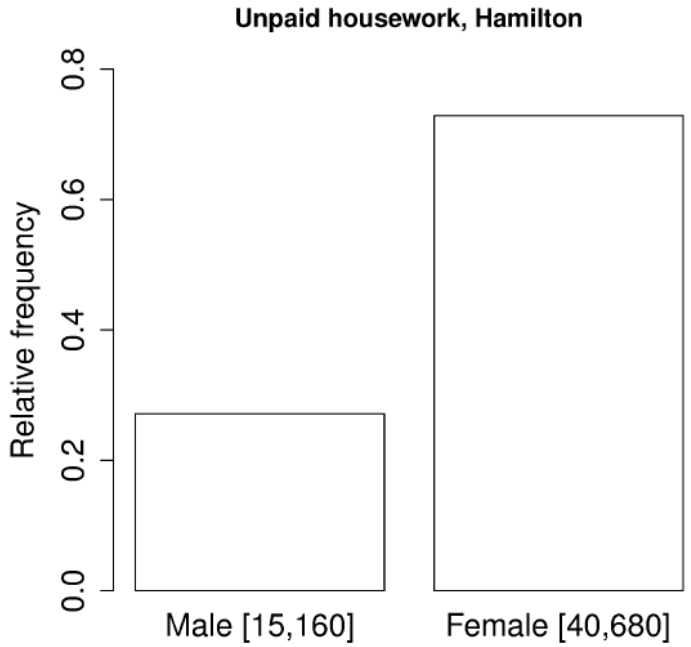
Data from a reactor: production yield, measured in g/L



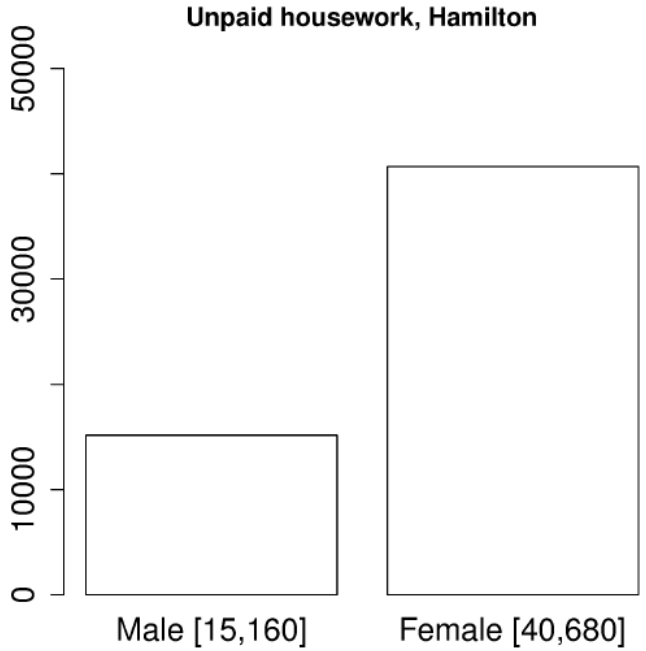


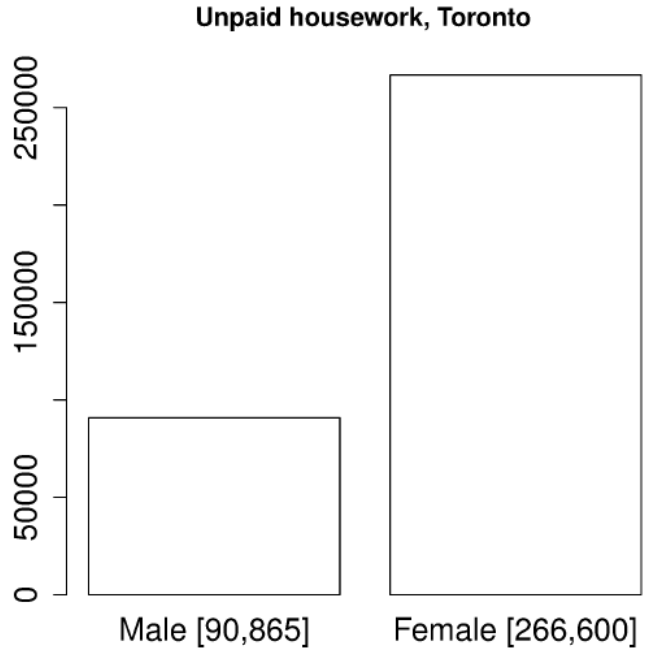
Unpaid housework in two cities: fair comparison are harder with frequency data alone



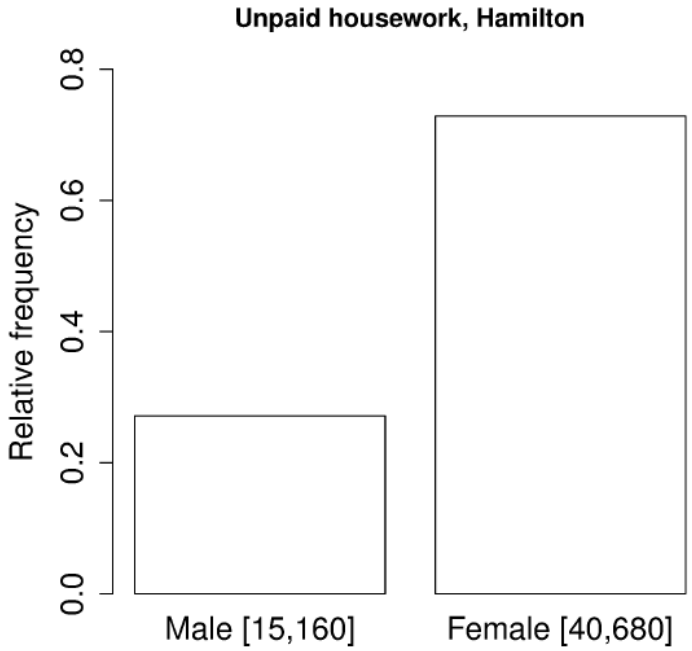


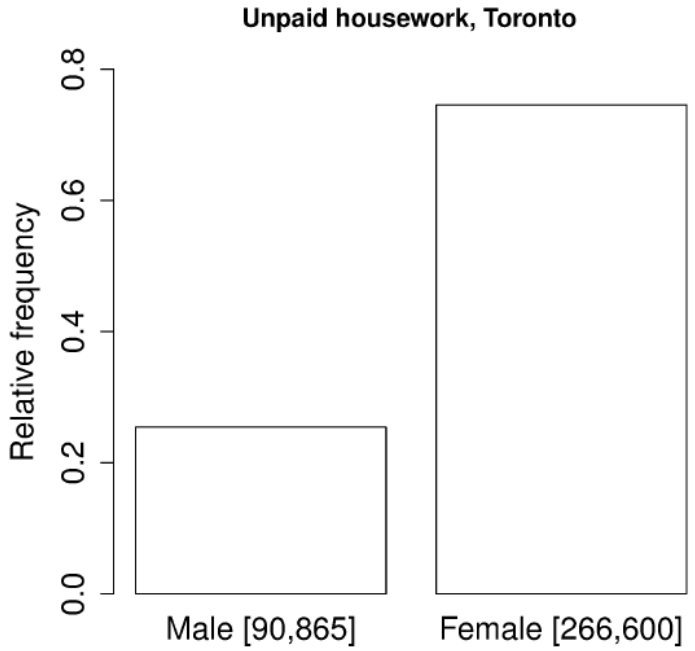
Unpaid housework in two cities: fair comparison are harder with frequency data alone





Unpaid housework in two cities: easier comparison with *relative* frequency data

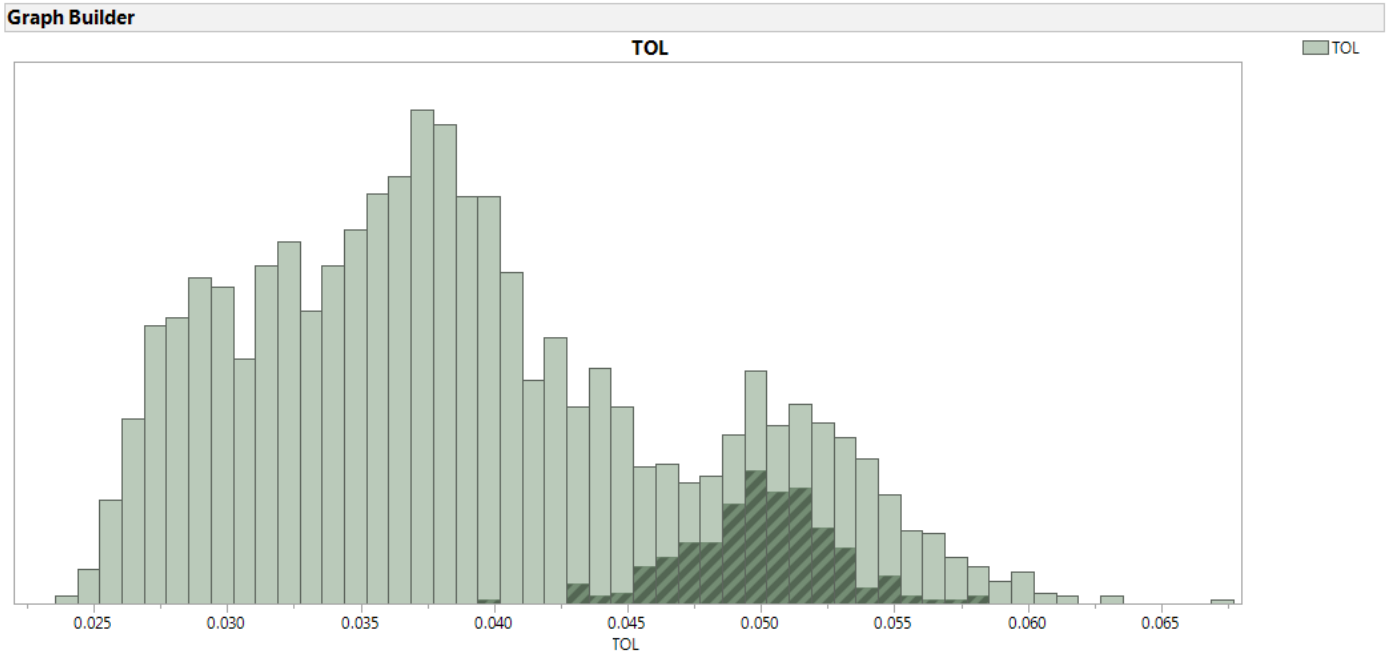




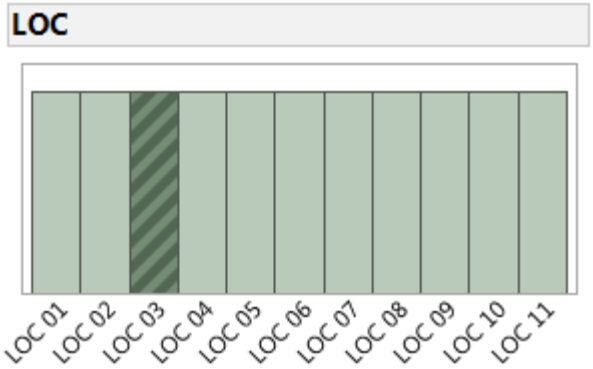
Histograms Using Software

- Data set: Tolerance Stack
- Part tolerance data has been collected from a manufacturing process. The tolerances are measured at 11 locations on the part (LOC 01 – LOC 11).
- Create a histogram to visualize the distribution of tolerances over the entire part
- Using histograms for each location, assess the location effect of the tolerance measurements.
- Assess the ability to hold an upper tolerance of .045

Histogram of Tolerance Measurements For Entire Part

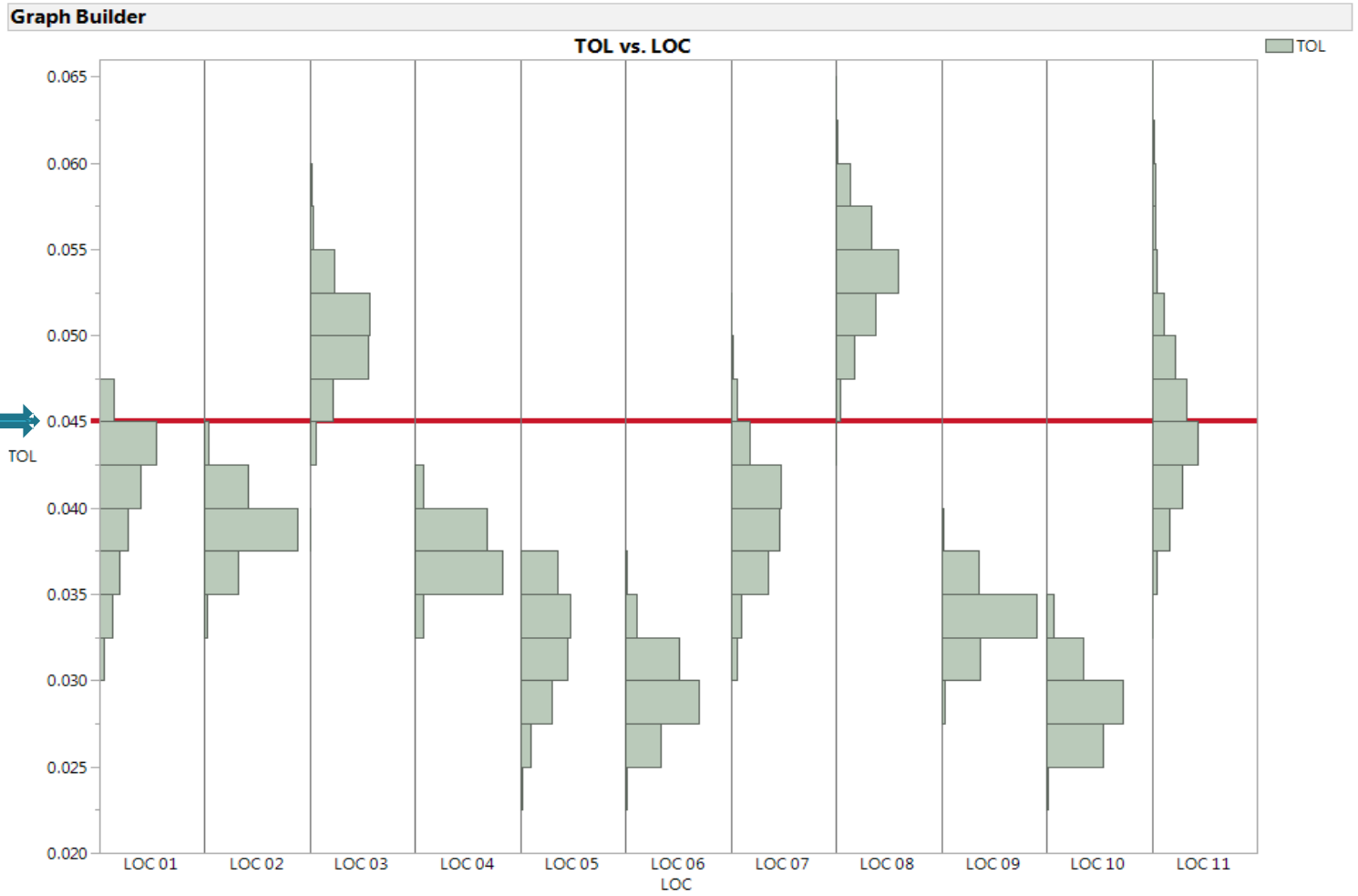


LOCATION EFFECTS



PhaseTwo Analytics

Histogram Stack Over Locations



Spec Limit

Problem Locations become Apparent

PhaseTwo Analytics

BOXPLOTS … but first

Quantiles of a Distribution

pth Quantile is the value which divides the range of data values so that p percent (fraction) of the values are less than that data value

- ***Quartiles*** -- Quantiles which divide the distribution into four equal parts (25th, 50th, 75th, and 100th quantiles)
- ***Median*** -- The 2nd quartile (50th quantile)

FYI…Formula for Calculating Quantiles

Define:

Q(p) = the quantile for the p^{th} percentage

$y_{(i)}$ = the i^{th} largest data point

Calculate $(n+1)p = i + g$,

where n = number of data points

i = integer portion of the product

g = fractional portion of the product

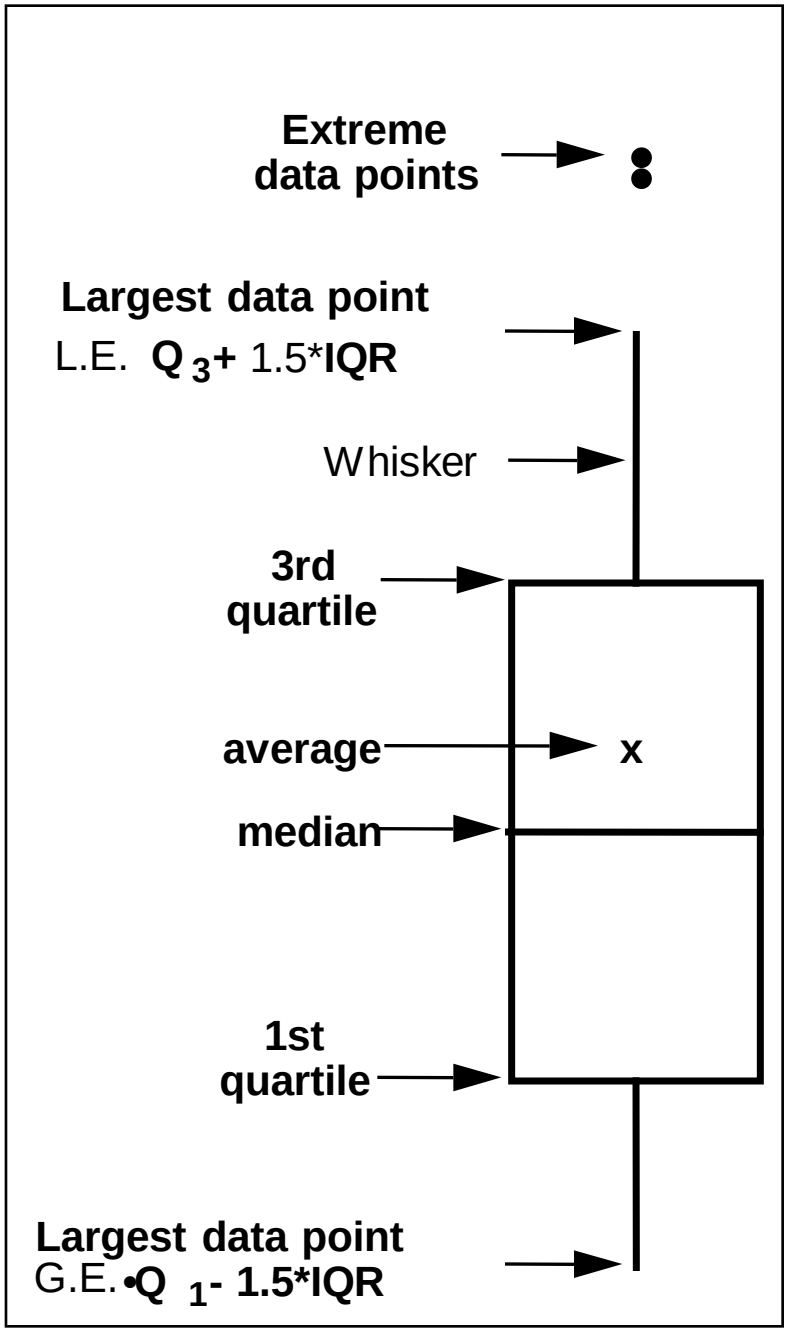
$Q(p) = (1-g)y_{(i)} + gy_{(i+1)}$, for $\frac{1}{n+1} \le p \le \frac{n}{n+1}$

Boxplots … or Box and Whisker Plots

An alternative to a histogram for displaying the distribution of data using lines, dots, and rectangles to indicate the data quartiles

IQR = (Q3 - Q1)

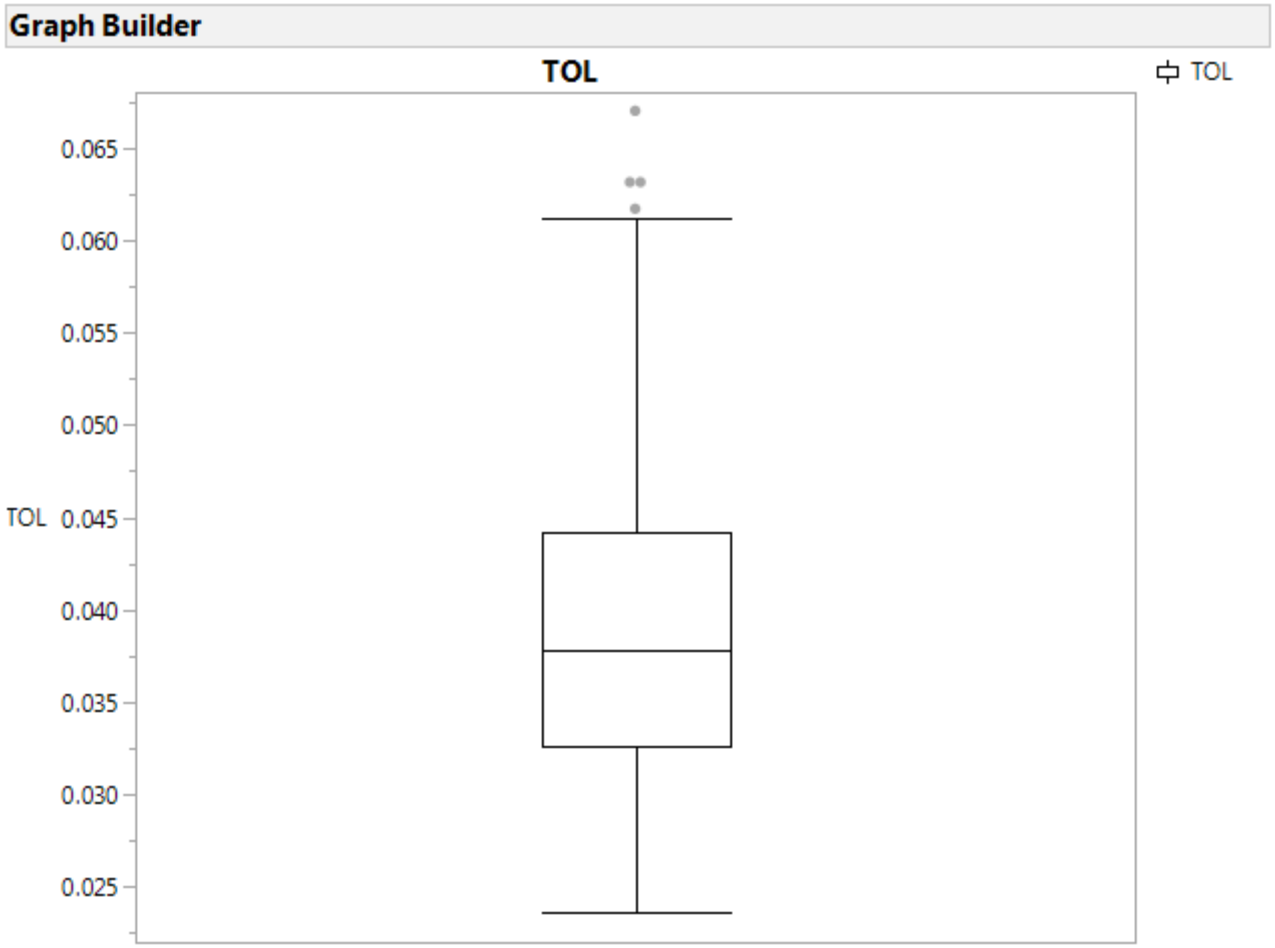
- The box has lines that extend from each end, sometimes called *whiskers*.
- The whiskers extend from the ends of the box to the outermost data point that falls within the distances computed as follows:
- 1st quartile – 1.5*(interquartile range)
- 3rd quartile + 1.5*(interquartile range)
- If the data points do not reach the computed ranges, then the whiskers are determined by the upper and lower data point values (not including outliers).



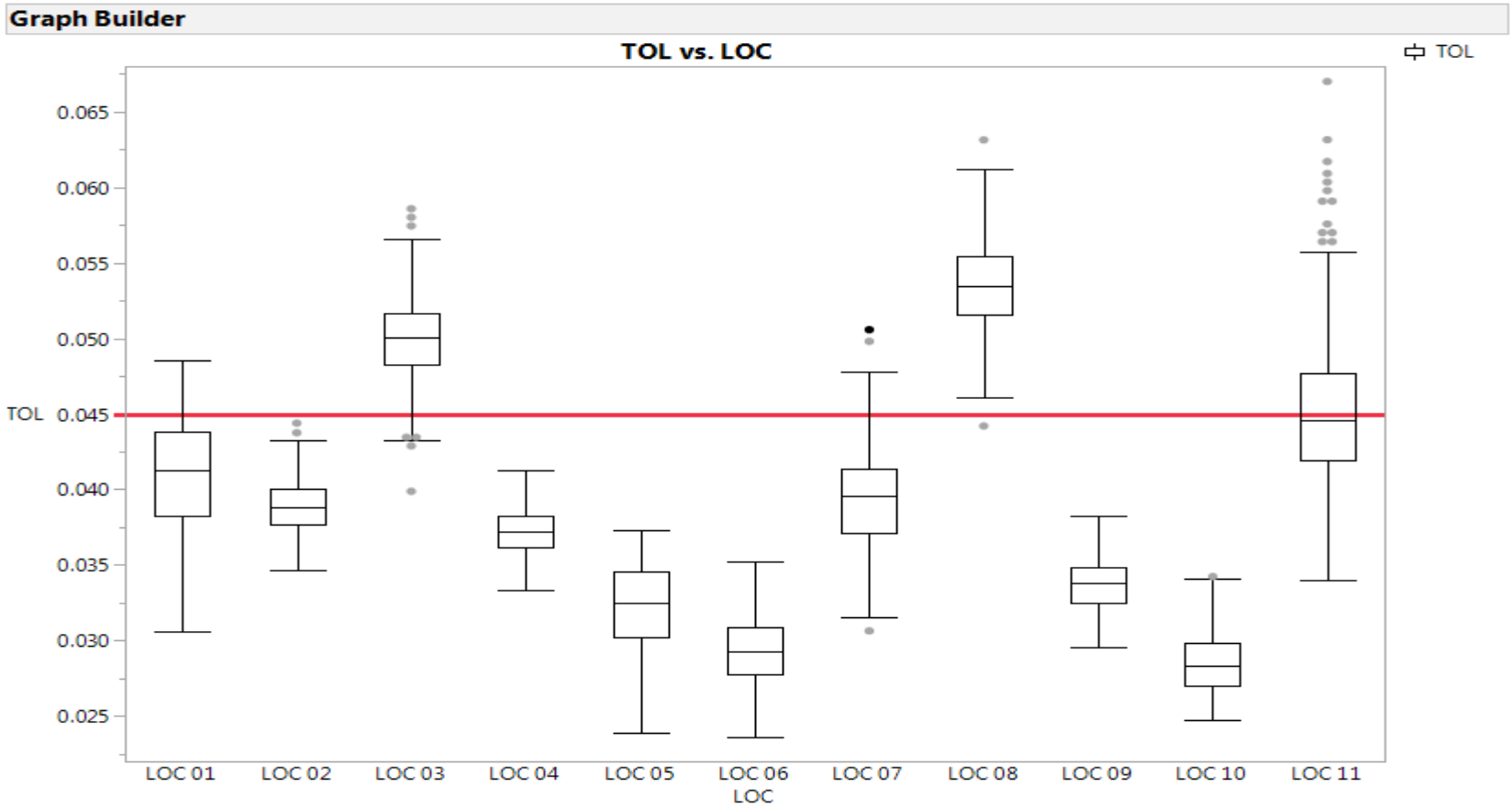
Boxplots Using Software

- Data set: Tolerance Stack
- Create a Boxplot to visualize the overall distribution of tolerances over the entire part
- Using boxplots for each location, assess the location effect of the tolerance measurements.

Boxplots: Entire Part



Boxplot: by Location

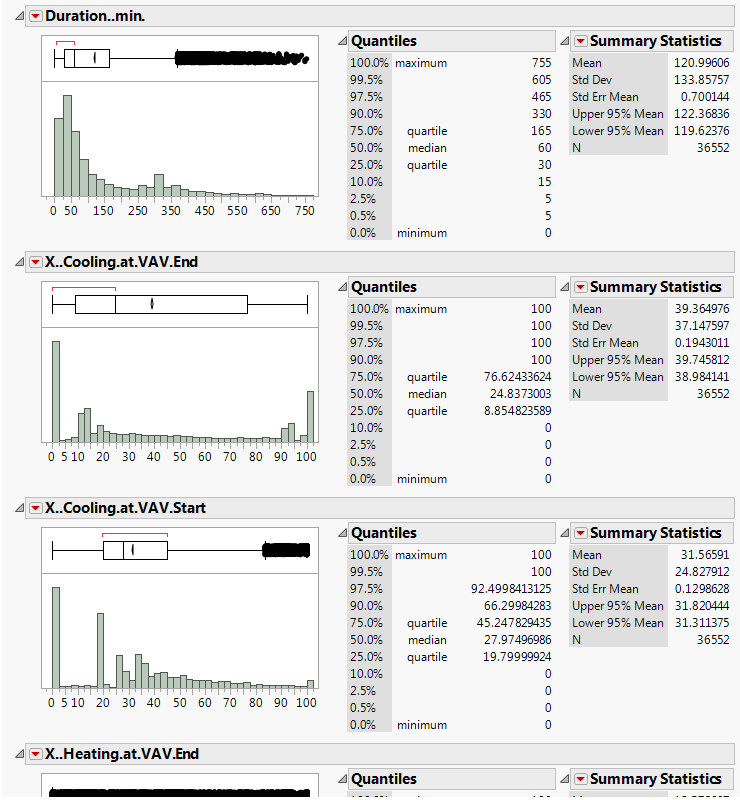


PhaseTwo Analytics

Exercise 0 revisited in JMP

familiarization with a dataset

- For the dataset VAVRH_EventFrames_EventTriggere
  - What does some of the data look like?
    - In JMP you work directly with the data table in view and menu above it
  - What are some of the characteristics of the data?
    - Via quick JMP visualizations



PhaseTwo Analytics

Other Quick Visualization Tools

- Bivariate plots (Y vs X)

- Scatterplot matrices

Commonly Used Distributions

- Discrete
  - Bernoulli
  - Binomial
  - Uniform
- Continuous
  - Normal
  - Student's t
  - Chi Squared
  - F Distribution
  - Uniform

Bernoulli (Binary) Distribution

- a Bernoulli rv is a discrete probability distribution which takes on the value X=1 with prob p and X=0 with prob 1-p
  - Pr(X=1) = P
  - Pr(X=0) = 1-p

- E[X] = 1*p + 0*(1-p) = p
- V[X] = p(1-p) …. why?

- Used to represent binary outcomes
  - Heads/tails, Pass/Fail, Success/Failure, etc

- Suppose Pr(Pass) = .7 and Pr(fail)=.3
  - Then Pr (Pass,Pass,Pass) = (.7) (.7) (.7) = .343 … assuming independence
  - Pr(Pass,Fail,Pass,fail,Pass,Fail) = (.7)(.3)(.7)(.3)(.7)(.3)

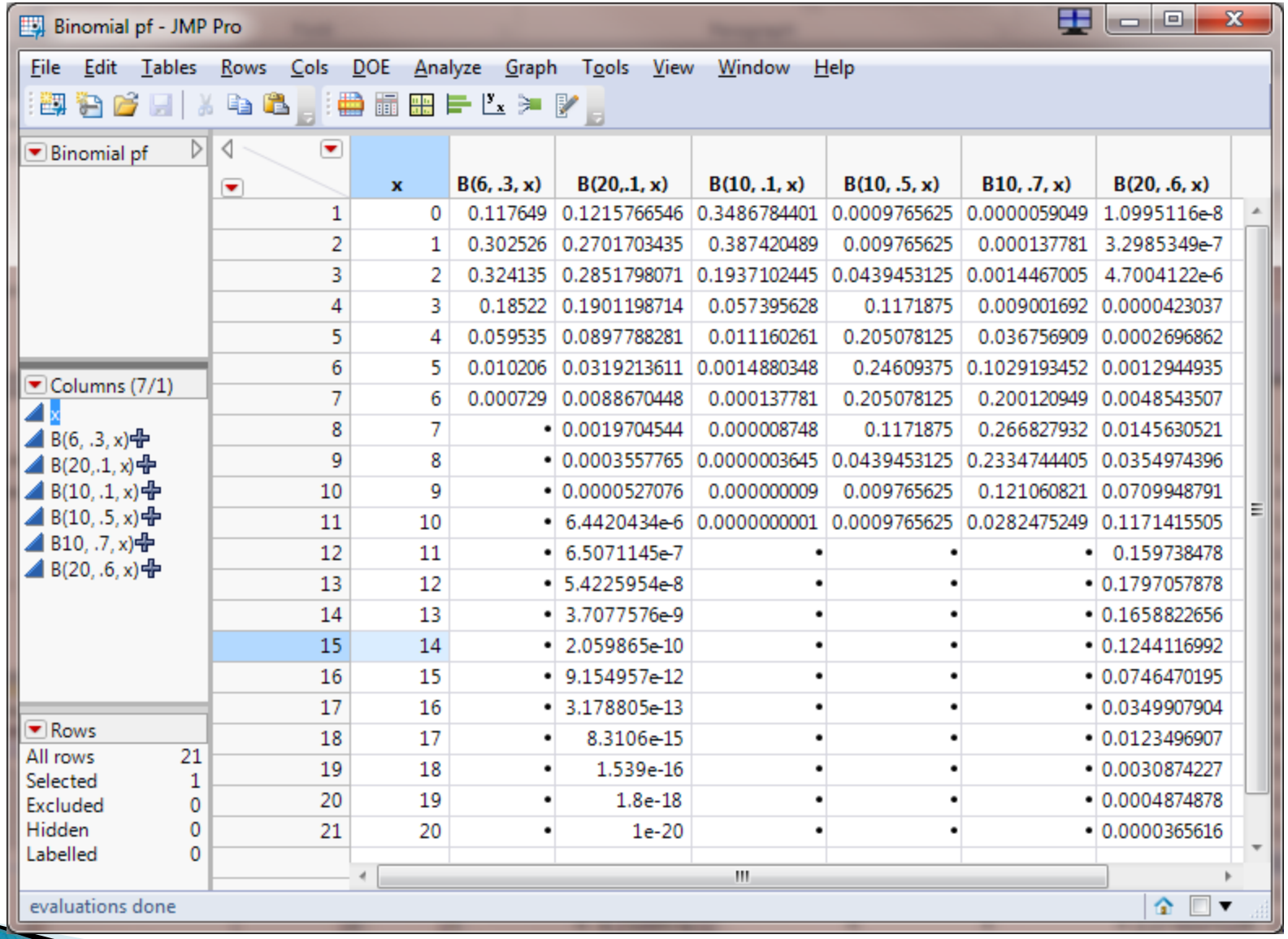
Binomial Distribution

- Let X1, X2, X3, …, Xn be a random sample (iid) of n observations from Bernoulli rv
- Then X=$\sum$ Xi (the number of times X=1)
  - Can be used, e.g., to represent the number of occurrences of a binary event … scrap, failure, heads, etc
- follows a Binomial Distribution, and
- $$\Pr(X = k) = \binom{n}{k} p^k (1-p)^{n-k}$$ is the probability function

$$E[X] = E[X_1 + \cdots + X_n] = E[X_1] + \cdots + E[X_n] = \underbrace{p + \cdots + p}_{n \text{ times}} = np$$

$$\operatorname{Var}(X) = \operatorname{Var}(X_1 + \cdots + X_n) = \operatorname{Var}(X_1) + \cdots + \operatorname{Var}(X_n) = n \operatorname{Var}(X_1) = np(1-p)$$

Binomial prob function calculations – B(n,p,x)



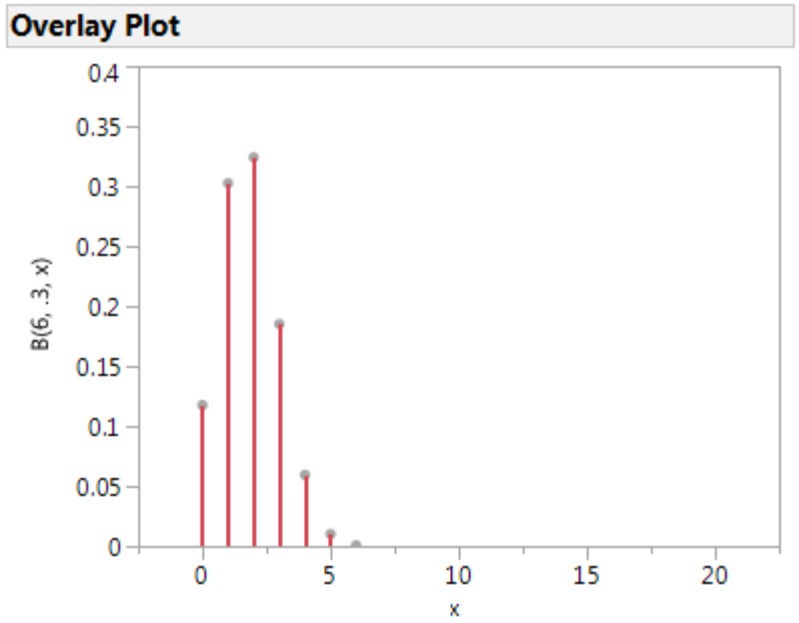
	x	B(6, .3, x)	B(20,.1, x)	B(10, .1, x)	B(10, .5, x)	B10, .7, x)	B(20, .6, x)
1	0	0.117649	0.1215766546	0.3486784401	0.0009765625	0.0000059049	1.0995116e-8
2	1	0.302526	0.2701703435	0.387420489	0.009765625	0.000137781	3.2985349e-7
3	2	0.324135	0.2851798071	0.1937102445	0.0439453125	0.0014467005	4.7004122e-6
4	3	0.18522	0.1901198714	0.057395628	0.1171875	0.009001692	0.0000423037
5	4	0.059535	0.0897788281	0.011160261	0.205078125	0.036756909	0.0002696862
6	5	0.010206	0.0319213611	0.0014880348	0.24609375	0.1029193452	0.0012944935
7	6	0.000729	0.0088670448	0.000137781	0.205078125	0.200120949	0.0048543507
8	7	•	0.0019704544	0.000008748	0.1171875	0.266827932	0.0145630521
9	8	•	0.0003557765	0.0000003645	0.0439453125	0.2334744405	0.0354974396
10	9	•	0.0000527076	0.000000009	0.009765625	0.121060821	0.0709948791
11	10	•	6.4420434e-6	0.0000000001	0.0009765625	0.0282475249	0.1171415505
12	11	•	6.5071145e-7	•	•	•	0.159738478
13	12	•	5.4225954e-8	•	•	•	0.1797057878
14	13	•	3.7077576e-9	•	•	•	0.1658822656
15	14	•	2.059865e-10	•	•	•	0.1244116992
16	15	•	9.154957e-12	•	•	•	0.0746470195
17	16	•	3.178805e-13	•	•	•	0.0349907904
18	17	•	8.3106e-15	•	•	•	0.0123496907
19	18	•	1.539e-16	•	•	•	0.0030874227
20	19	•	1.8e-18	•	•	•	0.0004874878
21	20	•	1e-20	•	•	•	0.0000365616

PhaseTwo Analytics

Binomial PF Plots & Prob Calculations

B(6,.3,x) x=0,6

x	B(6, .3, x)
0	0.117649
1	0.302526
2	0.324135
3	0.18522
4	0.059535
5	0.010206
6	0.000729



P(X<2) = P(X=0) + P(X=1) = .11765 + .30253 = .42018

P(X>=2) = P(X=2) + P(X=3) + P(X=4) + P(X=5) + P(X=6) = .32414+.18522+.05954+.01021+.00073 =

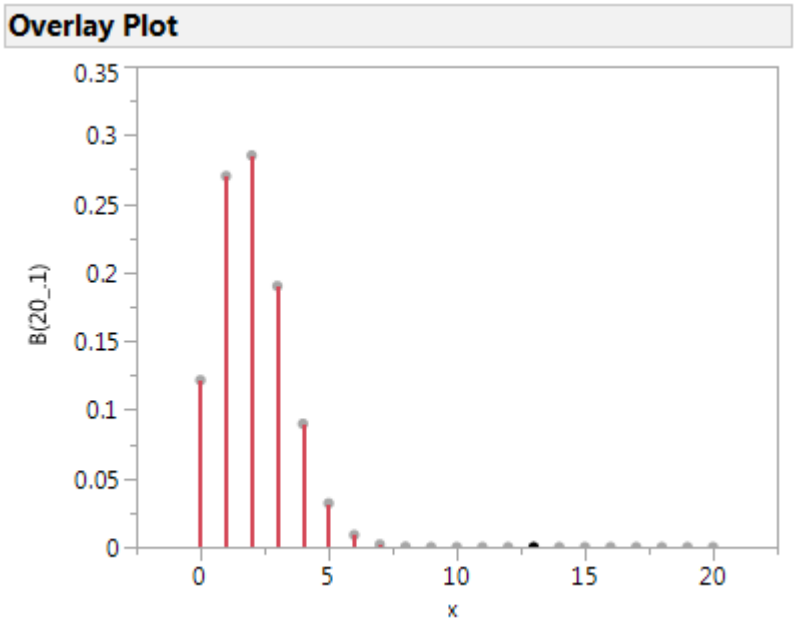
BUT, MORE EASILY = 1- P(X<2) = 1 - .42018 = .57982

WHY???

Binomial PF Plots & Prob Calculations

B(20,.1,x) x=0,20

x	B(20,.1, x)
0	0.1215766546
1	0.2701703435
2	0.2851798071
3	0.1901198714
4	0.0897788281
5	0.0319213611
6	0.0088670448
7	0.0019704544
8	0.0003557765
9	0.0000527076
10	6.4420434e-6
11	6.5071145e-7
12	5.4225954e-8
13	3.7077576e-9
14	2.059865e-10
15	9.154957e-12
16	3.178805e-13
17	8.3106e-15
18	1.539e-16
19	1.8e-18
20	1e-20



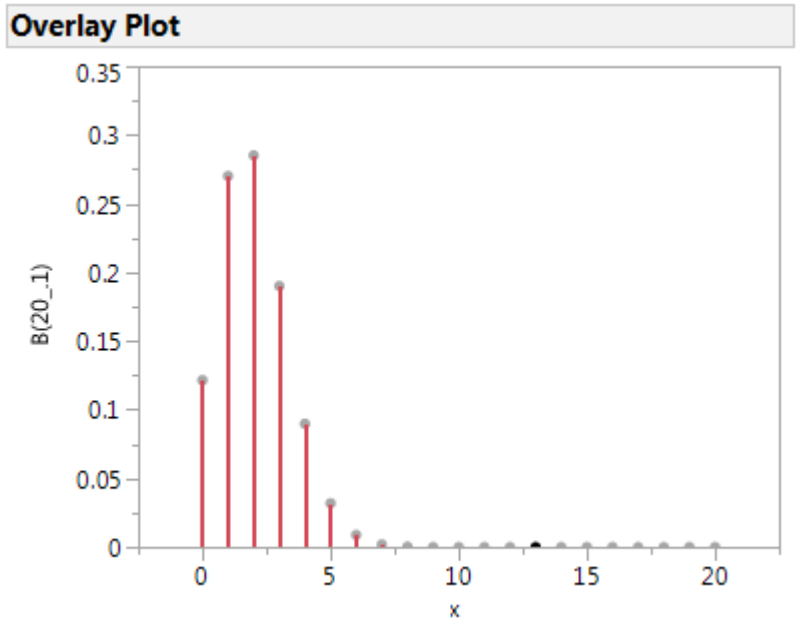
Calculate the P(X>=1)

Give me **two** ways of doing the calculation

Binomial PF Plots & Prob Calculations

B(20,.1,x) x=0,20

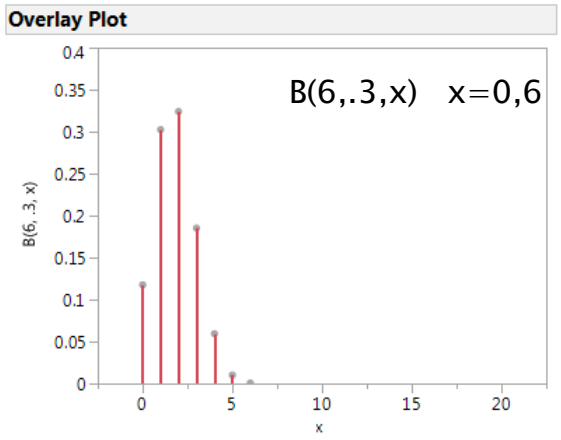
x	B(20,.1, x)
0	0.1215766546
1	0.2701703435
2	0.2851798071
3	0.1901198714
4	0.0897788281
5	0.0319213611
6	0.0088670448
7	0.0019704544
8	0.0003557765
9	0.0000527076
10	6.4420434e-6
11	6.5071145e-7
12	5.4225954e-8
13	3.7077576e-9
14	2.059865e-10
15	9.154957e-12
16	3.178805e-13
17	8.3106e-15
18	1.539e-16
19	1.8e-18
20	1e-20

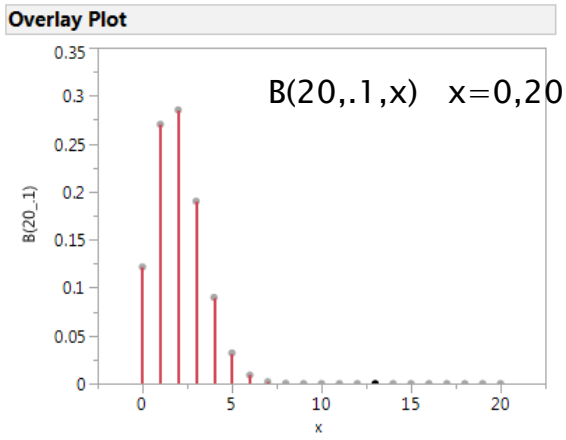


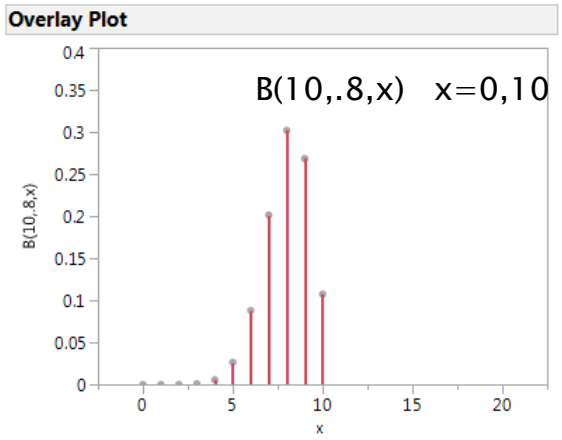
P(X>=1) = 1 – P(X=0)
= 1– .12158 = .87842

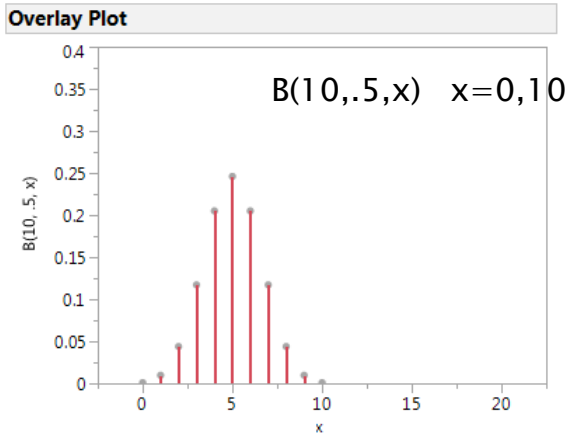
PhaseTwo Analytics

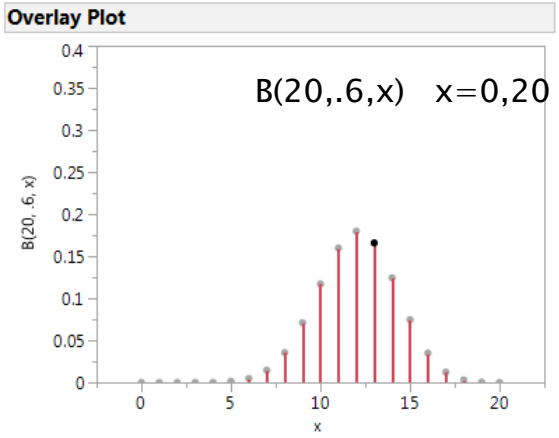
Examine Binomial PF Plots





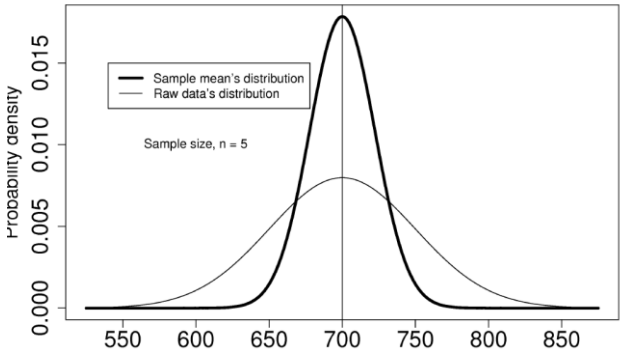




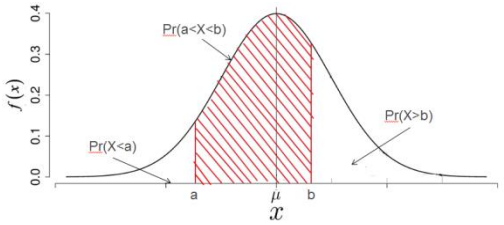


Let's make some shape observations

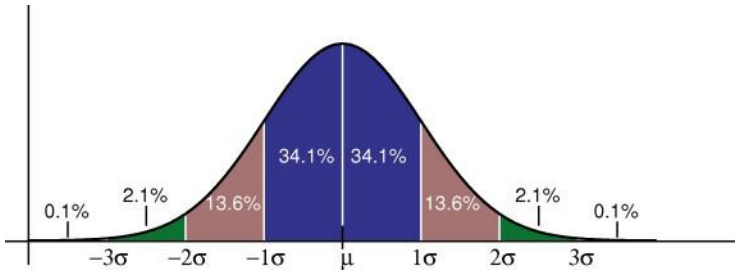
The Normal Distribution

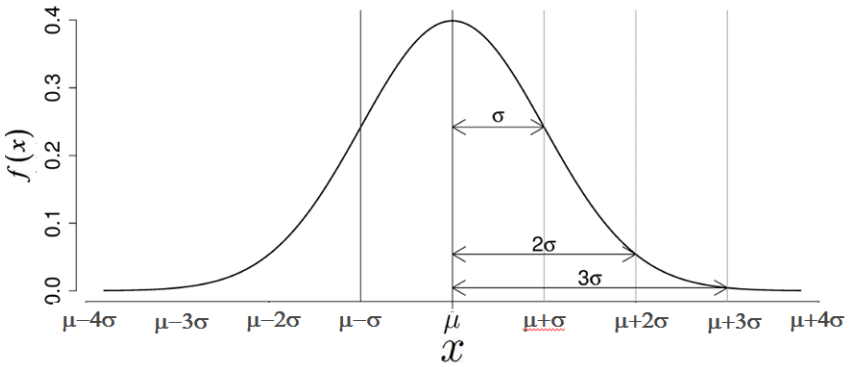


$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$



$$\Pr(a<X<b) = \int_a^b f(X)\,dx$$





"Bell-shaped curve"

PhaseTwo Analytics

The Normal Distribution

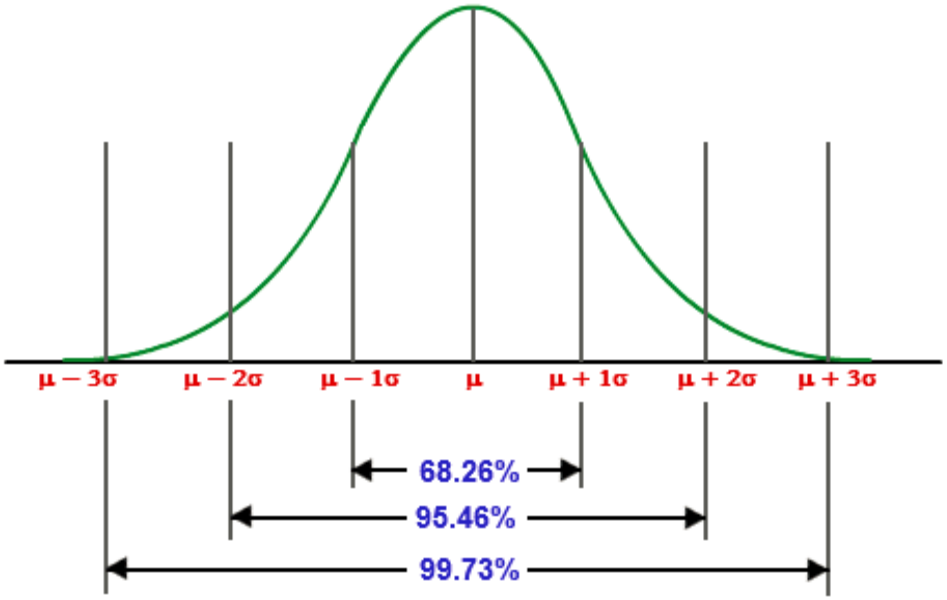
- Representative of the distribution of many naturally occurring phenomena that we measure
  - Population heights, weights & other physical dimensions
  - Repeated instrument measurements
  - Measurements of many phenomena in the physical & life sciences
- The Central Limit Theorem for **averages** of random variables *greatly* extends the usage of the normal distribution … (we will discuss this later)
- Fully characterized by its mean μ and variance σ^2
- Follows the pdf:

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$

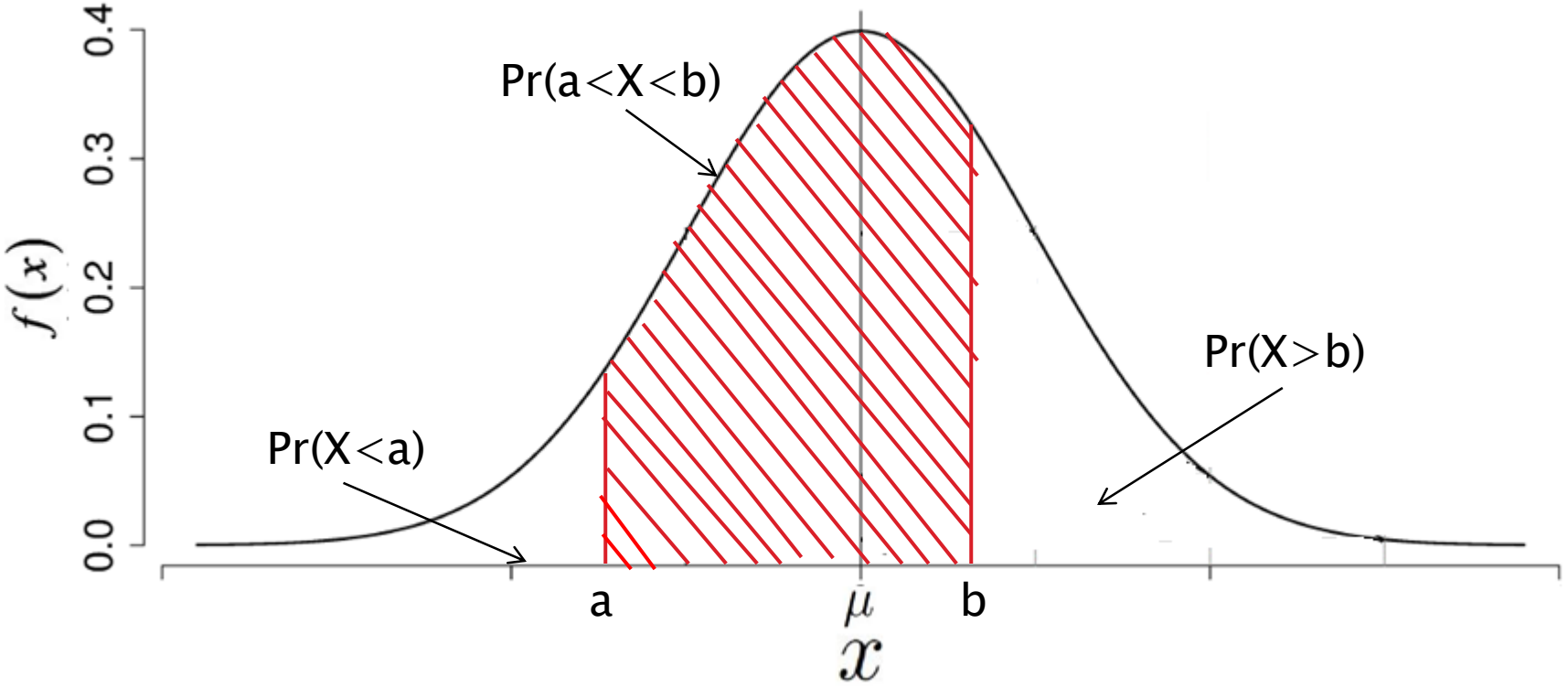
- **Is symmetric about μ**

Normal Distribution Properties

- Recall: Pr(a < X < b) is the area under the pdf between a & b
- Here are important facts related to Normal dist probabilities:

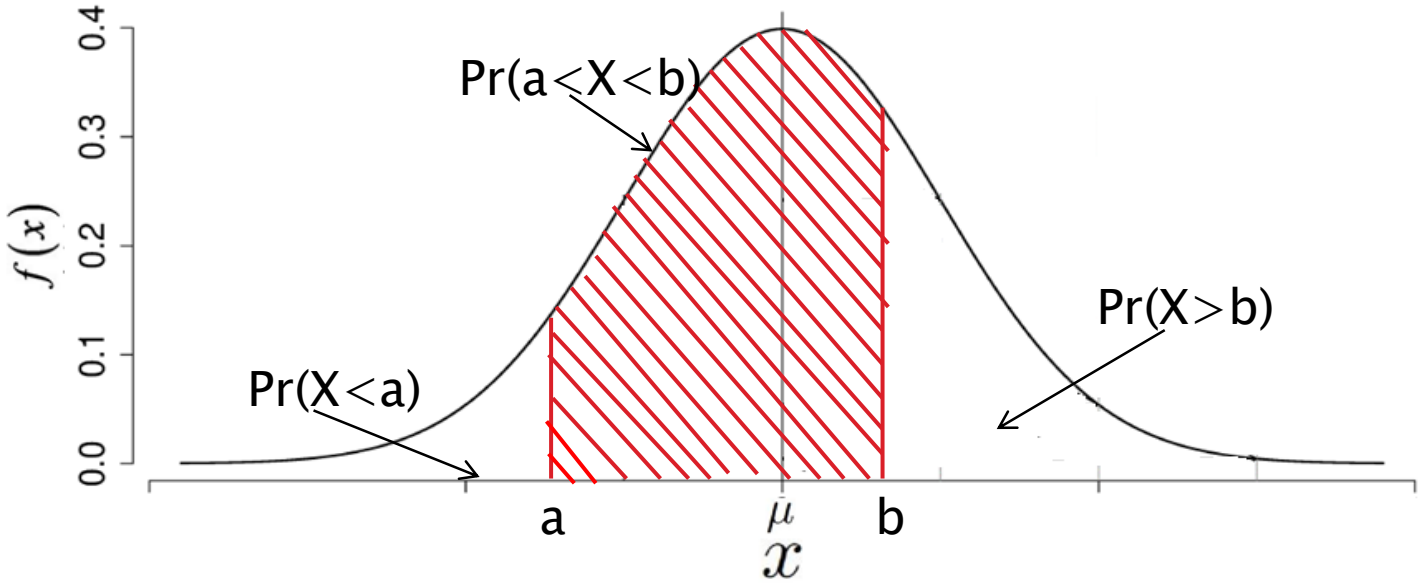


- Pr($\mu-1\sigma < X < \mu+1\sigma$)= .6826
- Pr($\mu-2\sigma < X < \mu+2\sigma$)= .9546
- Pr($\mu-3\sigma < X < \mu+3\sigma$)= .9973
- Pr($X > \mu+2\sigma$) = .023 .. How do we get this???



$$\Pr(a<X<b) = \int_a^b f(X)\,dx \qquad \int_a^b \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$

AND … from the laws of probability distributions



$Pr(X<a) + Pr(a<X<b) + Pr(X>b) = 1$

$Pr(X>a) = 1 - Pr(X<a)$

$Pr(X<b) = 1 - Pr(X>b)$

Etc, etc, etc

Calculating Normal Probabilities the Z distribution

- Since there are infinite combinations of μ and σ, tabulating probabilities of the normal distribution is an impossible task
- Tabulating probabilities still has value even though probability calculations are easy to do
- But note the following:
  - For any normally distributed rv X, by centering and scaling:

    $Z = \frac{X-\mu}{\sigma}$ follows a normal distribution with mean $= 0$ and sd $= 1$

  - Any probability statement for X can be converted to a statement in Z
  - e.g. $\Pr(a<X<b) = \Pr(\frac{a-\mu}{\sigma} < \frac{X-\mu}{\sigma} < \frac{b-\mu}{\sigma}) = \Pr(\frac{a-\mu}{\sigma} < Z < \frac{b-\mu}{\sigma})$
  - The Normal distribution with mean 0 and sd 1 is referred to as the **Standard Normal Distribution or Z Distribution** and is extensively tabulated

Statistical tables (tabular form)

z	q = cumulative area under the normal distribution
-3.00	0.001350
-2.75	0.002980
-2.50	0.006210
-2.25	0.01222
-2.00	0.02275
-1.80	0.03593
-1.50	0.06681
-1.00	0.1587
-0.50	0.3085
0.00	**0.5**
0.50	0.6915
1.00	0.8413
1.50	0.9332
1.80	0.9641
2.00	0.9773
2.25	0.9878
2.50	0.9938
2.75	0.9970
3.00	0.9987

q = cumulative area under the normal distribution	z
0.001	-3.090
0.0025	-2.807
0.005	-2.576
0.01	-2.326
0.025	-1.960
0.05	-1.645
0.075	-1.440
0.1	-1.282
0.3	-0.5244
0.5	**0.0**
0.7	0.5244
0.9	1.282
0.925	1.440
0.95	1.645
0.975	1.960
0.99	2.326
0.995	2.576
0.9975	2.807
0.999	3.090

Find:

Pr (Z < 1.645) =

Pr (Z > 1.645) =

Pr (Z < 1.96) =

Pr (Z < −1.645) =

Pr (|Z| < 1.645) =

Pr (|Z| > 1.96) =

Example

1. Assume $x =$ biological activity of a drug, $x \sim \mathcal{N}(26.9, 9.3)$

   What is the probability of $x \leq 30.0$?

Example

1. Assume $x =$ biological activity of a drug, $x \sim \mathcal{N}(26.9, 9.3)$

   What is the probability of $x \leq 30.0$?

- First create z variable: $z \approx \dfrac{30 - 27}{\sqrt{9}} = \dfrac{3}{3} = 1$
- Find the area under the standard normal distribution: $z \leq 1$
- So the area $\approx 84\%$

The Central Limit Theorem

Regardless of the distribution of a random variable, X, with mean μ and standard deviation σ :

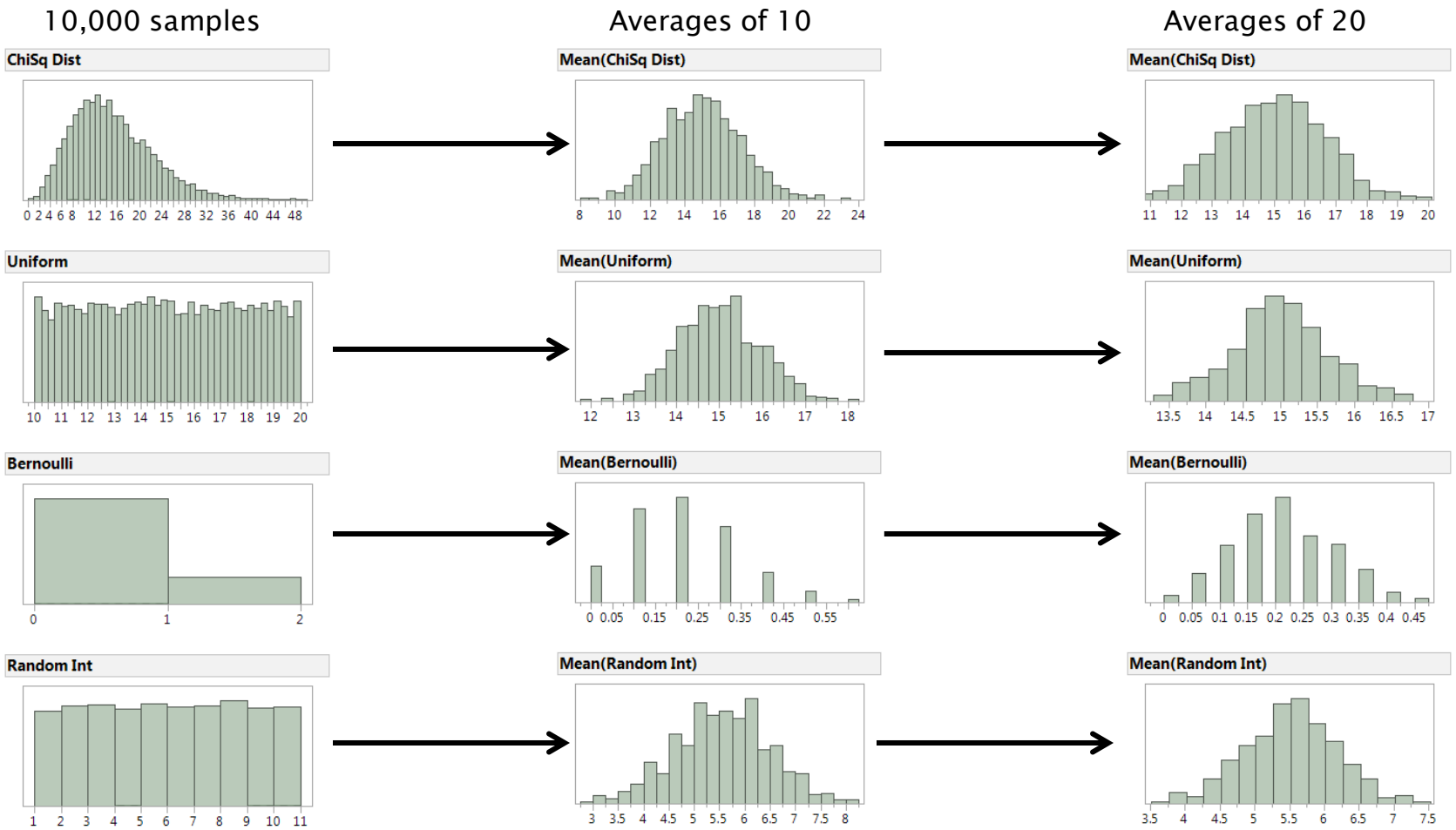
The distribution of sample means, $\bar{X}$, calculated from n *independent samples* from X will have mean μ, standard deviation $\boldsymbol{\sigma}/\sqrt{n}$, and the distribution will be *increasingly* ***normally distributed*** as n gets larger

Key Points

1. Applies to any distribution (iid)
2. Averages tend to a Normal Distribution
3. Variances decrease by a factor of 1/n

The Central Limit Theorem In Action

10,000 samples from non-normal distributions



PhaseTwo Analytics

Z distribution revisited

- Recall:
  - *For a normally distributed rv* ***X*** with mean μ and standard deviation σ :

$$Z = \frac{X-\mu}{\sigma}$$ follows a normal distribution with mean = 0 and sd = 1 (the Standard Normal Dist)

- *In light of the Central Limit Theorem* we can now say:
  - For an average , $\bar{X}$ from a sample of n independent observations from any rv **X** with mean μ and standard deviation σ :

$$Z = \frac{\bar{X}-\mu}{\sigma/\sqrt{n}}$$ follows a normal distribution (approx) with mean = 0 and sd = 1 (standard Normal Dist)

For future reference, note the following

$$Z = \frac{\text{statistic} - \text{mean(statistic)}}{\text{s.d.(statistic)}}$$

Where the “Statistic” follows a Normal Dist

HYPOTHESIS TESTS OF THE MEAN μ

- Questions like these get asked:
  - Has the mean tensile strength of parts coming out of my process changed?
  - Has the model-based approach to the NOC troubleshooting process resulted in a reduction in the mean Alarm Duration?
  - Is the voter turnout to Tuesday's election significantly greater than the historical mean?
- Based on the question, Data may be collected (a trial) in order to answer the question

HYPOTHESIS TESTS OF THE MEAN … has μ Increased???

- Suppose a rv X has mean μ_0 and **known variance σ^2**
- You take a random sample $x_1, x_2, x_3, \ldots, x_n$ from X to test the hypothesis
  - $H_0: \mu \leq \mu_0$ (the null hypothesis)

  versus
  - $H_a: \mu > \mu_0$ (the alternative hypothesis)

- **What's the sensible approach** to testing this hypothesis based on the data collected even if you don't know statistics?

HYPOTHESIS TESTS OF THE MEAN … has μ Increased???

- Common sense approach:
  - Compute $\bar{X}$ and if $\bar{X}$ is 'significantly' larger than μ_0, then we say that there is evidence to reject the hypothesis that $\mu \leq \mu_0$

- In other words, there is some outcome of $\bar{X}$, let's call it x_c, such that if $\bar{X} \geq x_c$, we reject the hypothesis that $\mu \leq \mu_0$ … *because under that hypothesis,* $\bar{X}$ *exceeding* x_c *would be highly unlikely!!*
- i.e. Under H_0 the probability that $\bar{X} \geq x_c$ is too small

HYPOTHESIS TESTS OF THE MEAN
… has μ Increased ???

- So, **choose x_c** so that **$Pr(\bar{X} > x_c)=\alpha$** , where alpha is small (typically .05)
- Note: **there is an equivalent value z_c of the Z distribution**, such that

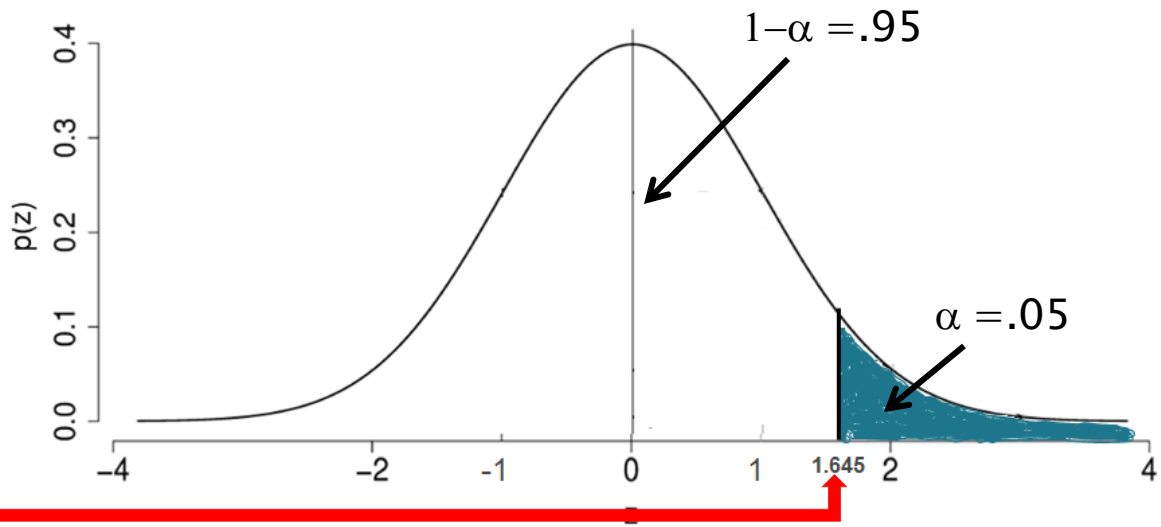
  $Pr(\bar{X} > x_c) = \alpha$ will be equivalent to $Pr(z>z_c) = \alpha$, because

- $Pr(\bar{X} > x_c) = Pr\left(\frac{\bar{X}-\mu_0}{\sigma/\sqrt{n}} > \frac{x_c-\mu_0}{\sigma/\sqrt{n}}\right) = Pr(z>z_c)$

  where $z_c = \frac{x_c-\mu_0}{\sigma/\sqrt{n}}$
- So … we determine if $\bar{X}$ is a rare occurrence by determining if its Z equivalent is rare
- This is the criterion used to reject $H_0: \mu \leq \mu_0$
- $Pr(z>z_c)$ is called the significance level of the test (α)
- $(1-\alpha)$ is called the confidence level

Critical points of the Z distribution… z_α for common choices of α

q = cumulative area under the normal distribution	z
0.001	-3.090
0.0025	-2.807
0.005	-2.576
0.01	-2.326
0.025	-1.960
0.05	-1.645
0.075	-1.440
0.1	-1.282
0.3	-0.5244
0.5	**0.0**
0.7	0.5244
0.9	1.282
0.925	1.440
0.95	1.645
0.975	1.960
0.99	2.326
0.995	2.576
0.9975	2.807
0.999	3.090



Pr(Z> 1.645) = .05

Pr(Z< −1.645) = .05 (symmetry)

Pr(Z> 1.96) = .025

Pr(|Z|> 1.96)= Pr(Z> 1.96)+ Pr(Z< −1.96) = .05

HYPOTHESIS TEST OF THE MEAN μ

$H_0: \mu \leq \mu_0$ vs $H_a: \mu > \mu_0$ (known σ^2)

- **Procedure**:
  - Collect data: $x_1, x_2, x_3, \ldots, x_n$
  - Calculate $\bar{X}$
  - Calculate $z = \frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}}$
  - Select α (confidence level)
  - If $z > z_\alpha$, reject the null hypothesis that $\mu \leq \mu_0$ ... OR
  - Calculate p-value = Pr ($Z > z$) . If p<= .05 then reject!

Example … manually w table:

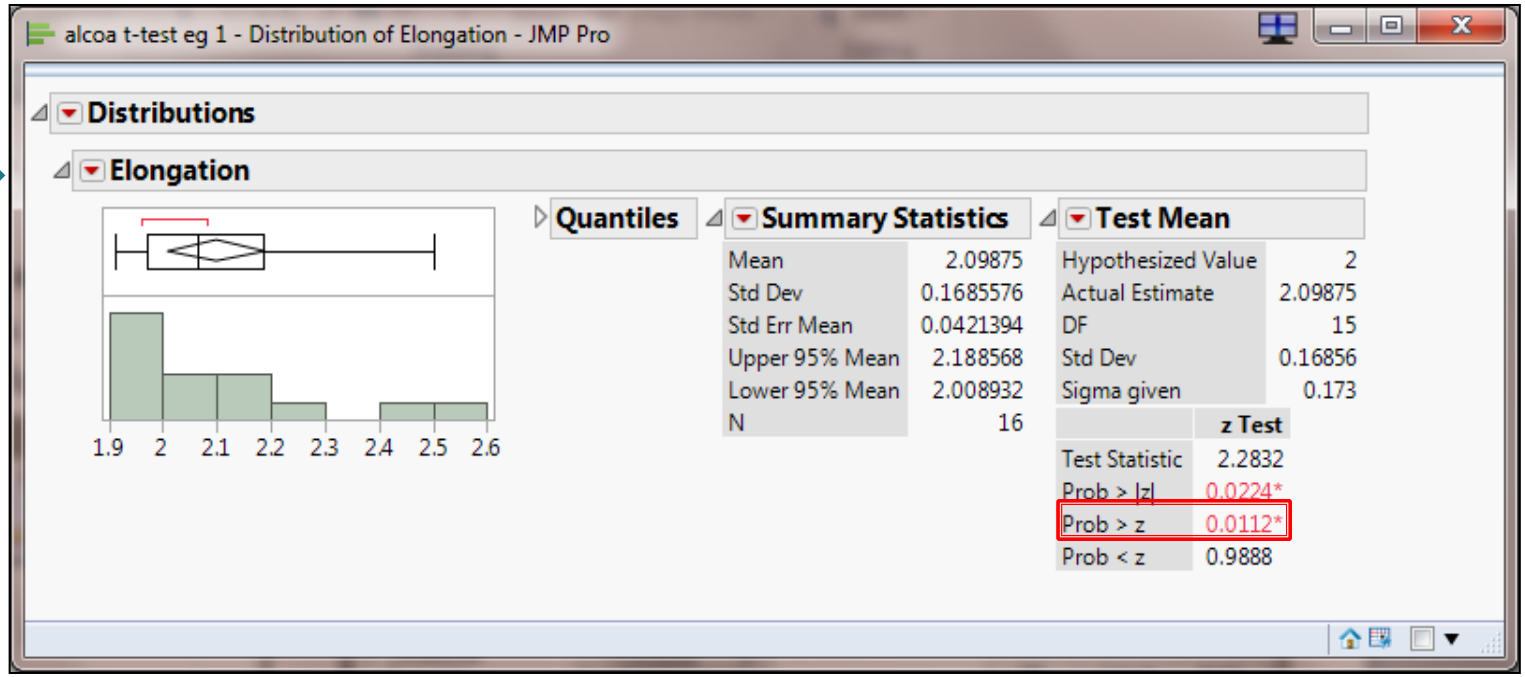
- % elongation of metal samples is measured to check to see if the mean elongation has increased from its historical 2% level. σ^2 is assumed to be .03
- Conduct a hypothesis test of $H_0: \mu \leq \mu_0$ versus $H_a: \mu > \mu_0$
- From the data we calculate:
  - $\bar{X} = 2.099$
  - $\sigma^2 = .03$ so $\sigma = .173$
  - n= 16
  - If $\alpha = .05$ then $z_\alpha = 1.645$
  - So, $z = \frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}} = \frac{2.099 - 2}{.173/4} = 2.28$
  - Since $Z > z_\alpha$, we reject $H_0: \mu \leq \mu_0$

	Elongation
1	2.42
2	2.09
3	2.17
4	1.91
5	2.5
6	1.99
7	2.19
8	2.08
9	1.99
10	2.11
11	1.97
12	2.05
13	2.22
14	1.96
15	1.97
16	1.96

Example continued … using software

Use JMP's "Distribution platform" to conduct "Test of Mean"

	Elongation
1	2.42
2	2.09
3	2.17
4	1.91
5	2.5
6	1.99
7	2.19
8	2.08
9	1.99
10	2.11
11	1.97
12	2.05
13	2.22
14	1.96
15	1.97
16	1.96



PhaseTwo Analytics

Example: Using R & Python

HYPOTHESIS TEST OF THE MEAN μ
$H_0: \mu \geq \mu_0$ vs $H_a: \mu < \mu_0$ (σ^2 known)

- Because of symmetry, the procedure is just the mirror image of the previous case with a focus now on the left side of the distribution

- Procedure:
  - Collect data: $x_1, x_2, x_3, \ldots, x_n$
  - Calculate $\bar{X}$
  - Calculate $z = \frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}}$
  - Select α (confidence level)
  - If $z < -z_\alpha$, reject the null hypothesis that $\mu \geq \mu_0$

HYPOTHESIS TEST OF THE MEAN μ
$H_0: \mu = \mu_0$ vs $H_a: \mu \neq \mu_0$ (σ^2 known)
Two-sided Tests

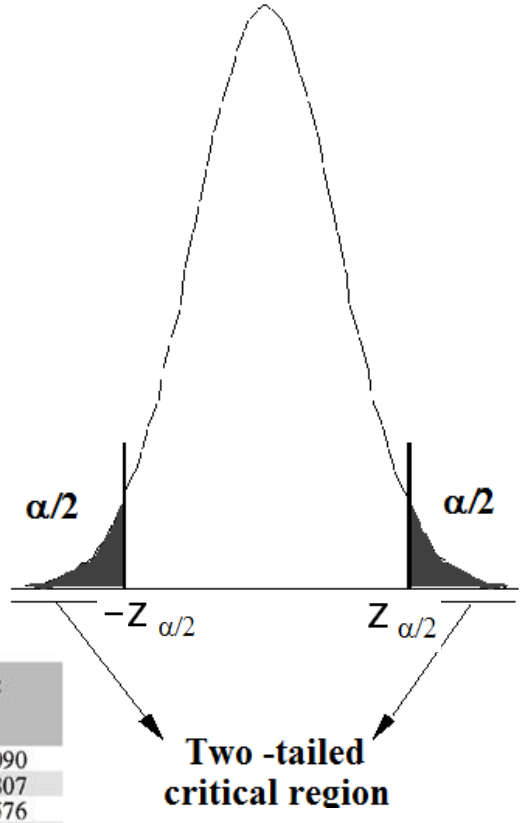
- If you think the mean has changed but aren't sure if it's a change in one direction or the other then a two-sided test is needed.
- Evidence to reject H_0 would occur if $\bar{X}$ is either much larger or much smaller
- The test procedure development is very similar to before
- We reject if the Z statistic is either to large or too small where the probability of seeing values $\leq \alpha$ determines the rejection region and symmetry is invoked

TWO-SIDED HYPOTHESIS TEST OF THE MEAN μ

$H_0: \mu = \mu_0$ vs $H_a: \mu \neq \mu_0$ (σ^2 known)

- Procedure:
  - Collect data: $x_1, x_2, x_3, \ldots, x_n$
  - Calculate $\bar{X}$
  - Calculate $z = \frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}}$
  - Select α (confidence level)
  - If $|z| > z_{\alpha/2}$, reject H_0
  - Note from the Z tables:
    - $P(|Z| > 1.96) = .05$

    so $Z < -1.96$ and $Z > 1.96$ define the critical region for $\alpha = .05$



q = cumulative area under the normal distribution	z
0.001	-3.090
0.0025	-2.807
0.005	-2.576
0.01	-2.326
0.025	-1.960
0.05	-1.645
0.075	-1.440
0.1	-1.282
0.3	-0.5244
0.5	**0.0**
0.7	0.5244
0.9	1.282
0.925	1.440
0.95	1.645
0.975	1.960
0.99	2.326
0.995	2.576
0.9975	2.807
0.999	3.090

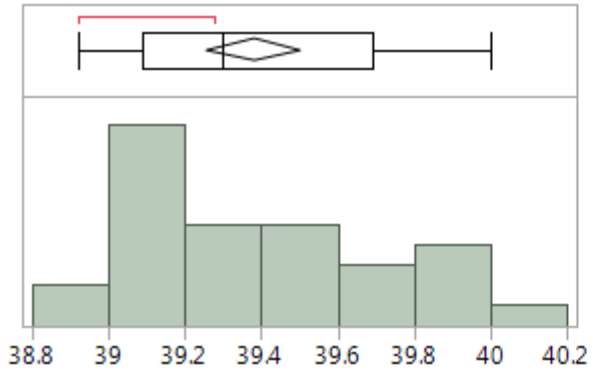
Two-sided tests … Coating Thickness example

- Coating thickness of tablets in a drug mfg process must meet a spec of $\mu = 40$. The Standard deviation is known to be $\sigma = .3$ (historically)
- Data are gathered to determine if the mean thickness meets the spec
- Test the hypothesis at a 95% confidence level $(\alpha = .05)$
  - From the data: $\bar{X} = 39.38$ and n= 30
  - the critical region is $|Z| > z_{.025}$ where $z_{.025} = 1.96$
  - Reject if Z< -1.96 or Z> 1.96
  - $z = \frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}} = \frac{39.38 - 40}{.3/\sqrt{30}} = -11.32$
  - **Since z < -1.96 we reject the hypothesis that $\mu = 40$**

Tablet	Coating thickness
1	39.04
2	39.05
3	39.09
4	38.92
5	39.04
6	39.13
7	39.32
8	39.23
9	39.28
10	39.2
11	39.12
12	39.09
13	39.02
14	39.87
15	39.45
16	39.68
17	39.49
18	39.47
19	39.53
20	40
21	39.91
22	39.96
23	39.81
24	39.78
25	39.72
26	39.59
27	39.38
28	39.16
29	39.14
30	38.93

Two-sided tests … Coating Thickness example Using JMP

Coating thickness



Summary Statistics	
Mean	39.38
Std Dev	0.3342826
Std Err Mean	0.0610314
Upper 95% Mean	39.504823
Lower 95% Mean	39.255177
N	30

Test Mean	
Hypothesized Value	40
Actual Estimate	39.38
DF	29
Std Dev	0.33428
Sigma given	0.3

	z Test
Test Statistic	-11.320
Prob > |z|	<.0001*
Prob > z	1.0000
Prob < z	<.0001*

HYPOTHESIS TESTS OF THE MEAN

σ^2 *unknown*

- Generally we do not know σ^2
- The test under this condition is an extension of the case when σ^2-is known

HYPOTHESIS TEST OF THE MEAN: σ^2 *unknown*

$H_0: \mu \leq \mu_0$ vs $H_a: \mu > \mu_0$

- Previously we tested H_0 vs H_a using the Z statistic,

  $z = \frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}}$, **under the assumption that σ^2 is known**

- A sensible approach would be to use "s" (the sample standard deviation) in place of σ

- Problem? S will add significant variability to Z and it would follow a different, but well known, distribution

- That distribution is referred to as the "t" distribution: $t = \frac{\bar{X} - \mu_0}{s/\sqrt{n}}$

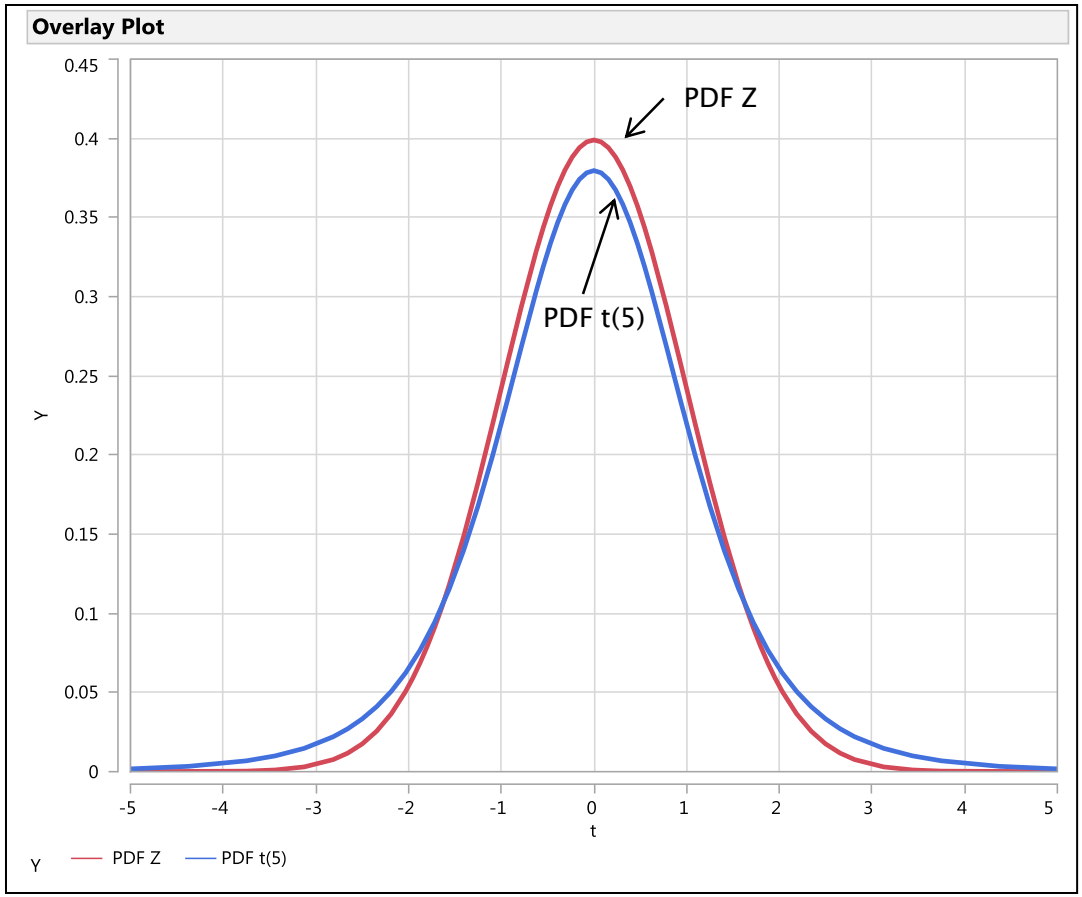
- **The t distribution has one parameter $\upsilon = n-1$**, which refers to the degrees of freedom (df) of s^2.

- s^2 has n−1 df because it loses a degree of freedom by estimating μ using $\bar{X}$

- Recall the formula:

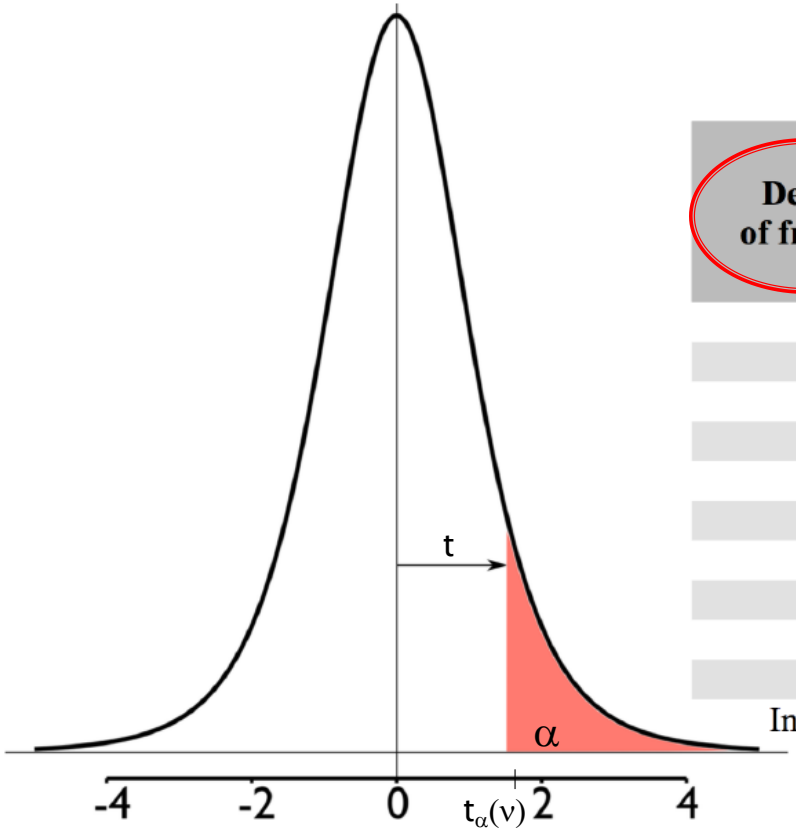
$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

t dist vs Z dist



Noticeable difference in tail probabilities (areas)
The t distribution is "fatter"

Use tables of the t distribution for tail areas



t distribution

Degrees of freedom	t value when area under the tail is						
	0.4	0.25	0.1	0.05	0.025	0.01	0.005
1	0.325	1.000	3.08	6.31	12.7	31.8	63.7
2	0.289	0.816	1.89	2.92	4.30	6.97	9.92
3	0.277	0.765	1.64	2.35	3.18	4.54	5.84
4	0.271	0.741	1.53	2.13	2.78	3.75	4.60
5	0.267	0.727	1.48	2.02	2.57	3.37	4.03
10	0.260	0.700	1.37	1.81	2.23	2.76	3.17
15	0.258	0.691	1.34	1.75	2.13	2.60	2.95
20	0.257	0.687	1.33	1.72	2.09	2.53	2.85
30	0.256	0.683	1.31	1.70	2.04	2.46	2.75
60	0.254	0.679	1.30	1.67	2.00	2.39	2.66
Infinite	0.253	0.674	1.28	1.64	1.96	2.33	2.58

$Pr(t(5) > 2.02) = .05$

$Pr(t(10) > 1.81) = .05$

$Pr(t(10) < -1.81) = .05$ (symmetry)

$Pr(t(10) > 2.23) = .025$

$Pr(|t(10)| > 2.23) = .025 + .025 = .05$

HYPOTHESIS TEST OF THE MEAN μ

$H_0: \mu \leq \mu_0$ vs $H_a: \mu > \mu_0$ (σ^2 unknown)

- Procedure:
  - Collect data: $x_1, x_2, x_3, \ldots, x_n$
  - Calculate $\bar{X}$
  - Calculate s
  - Calculate $t = \frac{\bar{X} - \mu_0}{s/\sqrt{n}}$
  - Select α (confidence level) and determine the threshold value, $t_\alpha(\nu)$
  - If $t > t_\alpha(\nu)$, reject the null hypothesis that $\mu \leq \mu_0$
  - or … calculate the p-value $p = P(\mathbf{t} > t)$

    reject if $p < \alpha$

HYPOTHESIS TEST OF THE MEAN μ
$H_0: \mu \geq \mu_0$ vs $H_a: \mu < \mu_0$ (σ^2 unknown)

- Procedure:
  - Collect data: $x_1, x_2, x_3, \ldots, x_n$
  - Calculate $\bar{X}$
  - Calculate s
  - Calculate $t = \frac{\bar{X} - \mu_0}{s/\sqrt{n}}$
  - Select α (confidence level) and determine the threshold value, $-t_\alpha(\nu)$
  - If $t < -t_\alpha(\nu)$, reject the null hypothesis that $\mu \geq \mu_0$
  - OR … if the calculated p-value $\leq \alpha$, then reject H_0

HYPOTHESIS TEST OF THE MEAN μ
$H_0: \mu = \mu_0$ vs $H_a: \mu \neq \mu_0$ (σ^2 unknown)

- Procedure:
  - Collect data: $x_1, x_2, x_3, \ldots, x_n$
  - Calculate $\bar{X}$ and s
  - Calculate $t = \frac{\bar{X} - \mu_0}{s/\sqrt{n}}$
  - Select α (confidence level) and determine the threshold value, $t_{\alpha/2}(\nu)$ where $\nu = n-1$
  - If $|t| > t_{\alpha/2}(\nu)$, reject the null hypothesis that $\mu = \mu_0$
  - OR … if the calculated p-value $\leq \alpha$, then reject H_0
  - e.g. If n=30, $\alpha = .05$
    $t_{.025}(29) \cong 2.04$

Degrees of freedom	t value when area under the tail is						
	0.4	0.25	0.1	0.05	0.025	0.01	0.005
1	0.325	1.000	3.08	6.31	12.7	31.8	63.7
2	0.289	0.816	1.89	2.92	4.30	6.97	9.92
3	0.277	0.765	1.64	2.35	3.18	4.54	5.84
4	0.271	0.741	1.53	2.13	2.78	3.75	4.60
5	0.267	0.727	1.48	2.02	2.57	3.37	4.03
10	0.260	0.700	1.37	1.81	2.23	2.76	3.17
15	0.258	0.691	1.34	1.75	2.13	2.60	2.95
20	0.257	0.687	1.33	1.72	2.09	2.53	2.85
30	0.256	0.683	1.31	1.70	2.04	2.46	2.75
60	0.254	0.679	1.30	1.67	2.00	2.39	2.66
Infinite	0.253	0.674	1.28	1.64	1.96	2.33	2.58

Coating Thickness example revisited …σ^2 unknown ***solve manually***

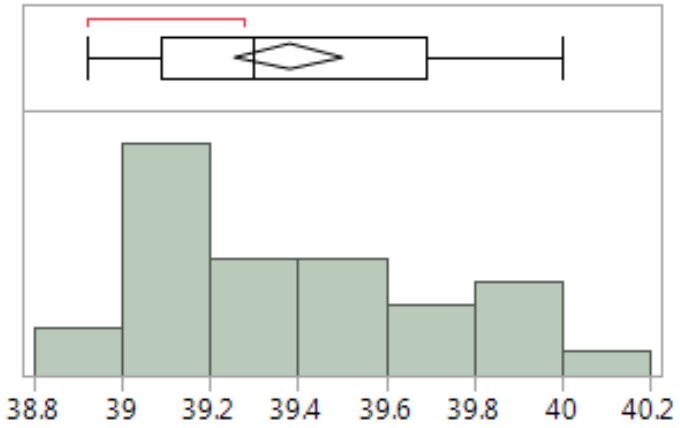
- Coating thickness of tablets in a drug mfg process must meet a spec of $\mu = 40$.
- σ^2 **is assumed to be unknown**
- We are testing, $H_0: \mu = 40$ vs $H_a: \mu \neq 40$
- Test the hypothesis at a 95% confidence level $(\alpha = .05)$

- From the data: $\bar{X} = 39.38$, s= .3343 and n= 30

- $t = \frac{\bar{X} - \mu_0}{s/\sqrt{n}} = \frac{39.38 - 40}{.3343/\sqrt{30}} = -10.158$
  - Since $|t| > 2.04$ **we reject $H_0: \mu = 40$**
  - (if $\alpha = .05$ so $t_{.025}(29) \cong 2.04$)

Using Software

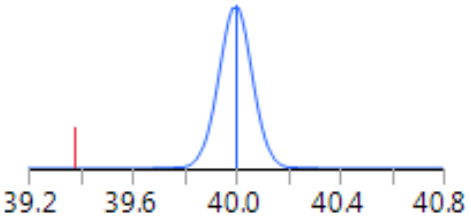
Coating thickness



Summary Statistics	
Mean	39.38
Std Dev	0.3342826
Std Err Mean	0.0610314
Upper 95% Mean	39.504823
Lower 95% Mean	39.255177
N	30

Test Mean	
Hypothesized Value	40
Actual Estimate	39.38
DF	29
Std Dev	0.33428

	t Test
Test Statistic	-10.159
Prob > |t|	<.0001*
Prob > t	1.0000
Prob < t	<.0001*



Confidence Intervals For The Mean

And Their Relationship to Hypothesis Tests

Confidence Intervals For The Mean

- Testing vs Estimation
  - In Testing we ask "Is it μ_0 or not?"
    - The mean being μ_0 or not is specifically important
    - A product specification whose violation may have legal consequences
- Estimation provides a "best guess" plus a range of other statistically plausible guesses

Confidence Intervals for the mean μ
σ^2 known

- A confidence interval is a random interval for a parameter like μ
- A 95% interval will capture μ with probability(confidence) .95

- It is derived as follows:
  - If x_i are iid samples from a rv X with mean μ and variance σ^2
  - Then … $z = \frac{\bar{X}-\mu_0}{\sigma/\sqrt{n}}$ follows a Standard Normal Distribution
  - So, Pr $(-z_{\alpha/2} < \frac{\bar{X}-\mu}{\sigma/\sqrt{n}} < z_{\alpha/2}) = 1-\alpha$ BUT, by rearranging terms …
  - Pr $(\bar{X} - z_{\alpha/2}(\sigma/\sqrt{n}) < \mu < \bar{X} + z_{\alpha/2}(\sigma/\sqrt{n}) = 1-\alpha$

The interval $(\bar{X} - z_{\alpha/2}(\sigma/\sqrt{n}) \ , \ \bar{X} + z_{\alpha/2}(\sigma/\sqrt{n})$ is called a $(1-\alpha)$*100% confidence interval for μ

Connection to Hypothesis Tests

Since values of μ that satisfy

$$-z_{\alpha/2} < \frac{\bar{x}-\mu}{\sigma/\sqrt{n}} < z_{\alpha/2}$$ *(i.e. are in the confidence interval)*

lead to acceptance of H_0, the confidence interval contains all μ that lead to acceptance of H_0

Coating Thickness example

Find the 95% confidence interval for μ, $\sigma = .3$ (known)

- $\bar{X}$ =39.38 , n= 30 , $z_{.025} = 1.96$
- $\bar{X} - z_{.025}(\sigma/\sqrt{n}) = 39.38 - 1.96(.3/5.477) = \mathbf{39.26}$
- $\bar{X} + z_{.025}(\sigma/\sqrt{n}) = 39.38 + 1.96(.3/5.477) = \mathbf{39.49}$
- 95% confidence interval is : (39.27 , 39.49)

Confidence Intervals for the mean μ

*** σ^2 unknown ***

- The derivation is identical to what was done when σ^2 is known except with unknown σ^2 we use the t distribution
  - We've previously shown:
  - $\Pr\left(-t_{\alpha/2}(n-1) < \frac{\bar{X}-\mu}{s/\sqrt{n}} < t_{\alpha/2}(n-1)\right) = 1-\alpha$ … and by rearranging terms
  - or … $\Pr\left(\bar{X} - t_{\alpha/2}(n-1)\,(s/\sqrt{n}) < \mu < \bar{X} + t_{\alpha/2}(n-1)\,(s/\sqrt{n})\right) = 1-\alpha$

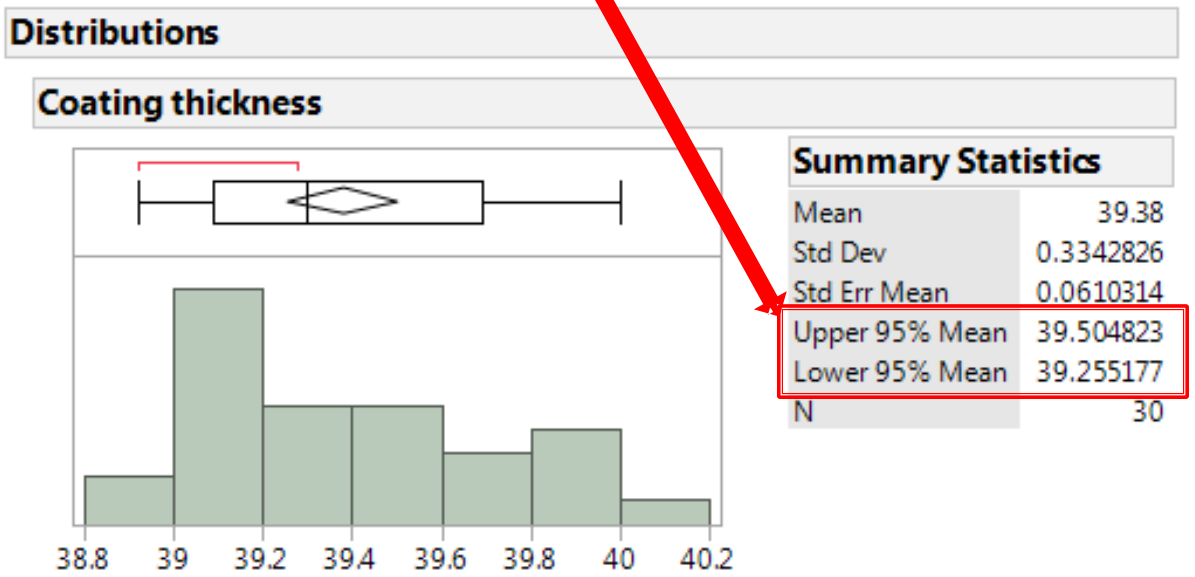
- The interval $\left(\bar{X} - t_{\alpha/2}(n-1)\,(s/\sqrt{n}),\ \bar{X} + t_{\alpha/2}(n-1)(s/\sqrt{n})\right)$ is a $(1-\alpha)*100\%$ confidence interval for μ

Coating Thickness example
Construct a 95% confidence interval for μ, σ^2 unknown

- $\bar{X}$ =39.38 , n= 30 , s= .33428 $t_{.025}(29) = 2.04$
- $\bar{X} - t_{.025}(29)(s/\sqrt{n}) = 39.38 - 2.04(.3343/5.477) = 39.255$
- $\bar{X} + t_{.025}(29)(s/\sqrt{n}) = 39.38 + 2.04(.3343/5.477) = 39.5045$
- 95% confidence interval is : (39.255 , 39.5045)
- Next, lets use software

Coating Thickness example Construct a 95% confidence interval for μ, σ^2 unknown

- $\bar{X}$ =39.38 , n= 30 , s= .33428 $t_{.025}(29) = 2.04$
- $\bar{X} - t_{.025}(29)\,(s/\sqrt{n}) = 39.38 - 2.04(.3343/5.477) = 39.255$
- $\bar{X} + t_{.025}(29)\,(s/\sqrt{n}) = 39.38 + 2.04(.3343/5.477) = 39.5045$
- 95% confidence interval is : (39.255 , 39.5045)



Class Lab #1: Visualizing Distributions, Means Test and Confidence Intervals.

- The datasets referenced below contain room temperature measurements from 2 VAV units (VAVCO 3-06 and VAVCO 4-06) that control temperature in two zones of the SLTC building in San Leandro. The two zones are in identical floor locations but on different floors. A question arose as to whether the mean zones temperatures were comparable which led to a study of the data.

Class Lab #1:

a. Data: VAV_3-06_4-06.xls VAV_3-06_4-06_Stacked.xls

b. Create a histogram and boxplot of variable RmTemp in file VAV_3-06_4-06_Stacked.xls

c. Visually compare the room temperatures(RmTemp) of VAV 3-06 and VAV 4-06 using Plots, Histograms or Boxplots.

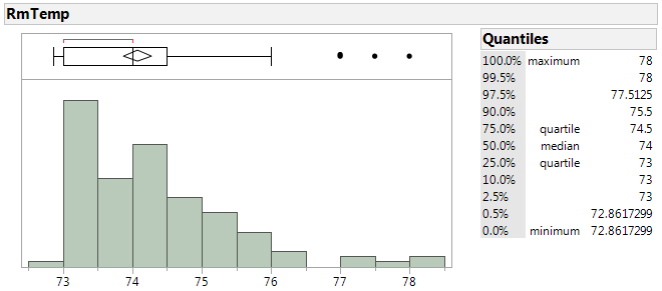
d. Does the data for the VAV 3-06 unit support a hypothesis that the mean room starting temperature is73.5 degrees?

e. Construct a 95% confidence interval for the mean room start temperature for VAV 3-06.

f. Based on the confidence interval, would you accept or reject the hypothesis that the mean temperature is 74 degrees?

Lab #1 Solutions

b.



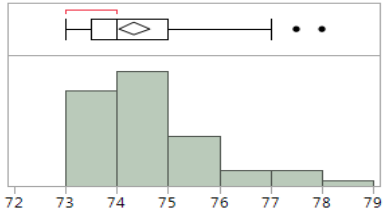
Quantiles		
100.0%	maximum	78
99.5%		78
97.5%		77.5125
90.0%		75.5
75.0%	quartile	74.5
50.0%	median	74
25.0%	quartile	73
10.0%		73
2.5%		73
0.5%		72.8617299
0.0%	minimum	72.8617299

Using the file VAV_3-06_4-06_Stacked
Menu: Analyze -> Distribution
Drag the RmTemp column to the Y,columns box & click OK

c.

Distributions VAV Unit= VAVCO 3-06

RmTemp

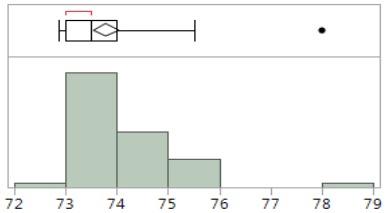


Quantiles		
100.0%	maximum	78
99.5%		78
97.5%		77.75
90.0%		76
75.0%	quartile	75
50.0%	median	74
25.0%	quartile	73.5
10.0%		73
2.5%		73
0.5%		73
0.0%	minimum	73

Summary Statistics	
Mean	74.338983
Std Dev	1.1687035
Std Err Mean	0.1521522
Upper 95% Mean	74.643549
Lower 95% Mean	74.034417
N	59

Distributions VAV Unit= VAVCO 4-06

RmTemp

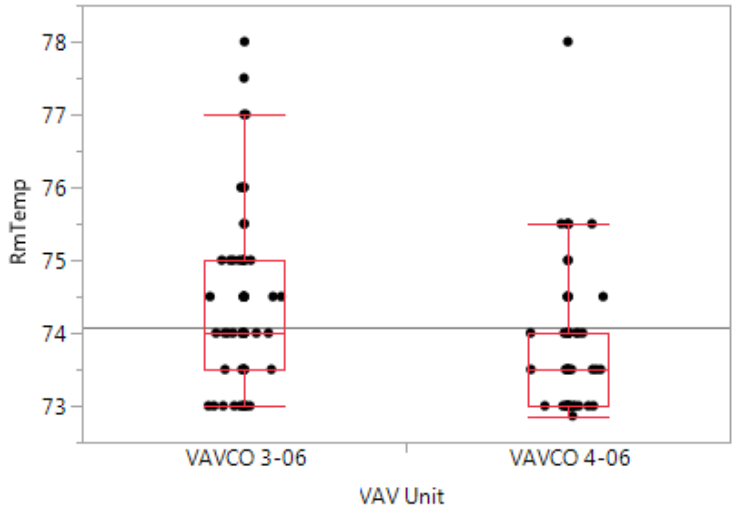


Quantiles		
100.0%	maximum	78
99.5%		78
97.5%		76.75
90.0%		75.5
75.0%	quartile	74
50.0%	median	73.5
25.0%	quartile	73
10.0%		73
2.5%		72.9308649
0.5%		72.8617299
0.0%	minimum	72.8617299

Summary Statistics	
Mean	73.794267
Std Dev	0.9808372
Std Err Mean	0.1276941
Upper 95% Mean	74.049874
Lower 95% Mean	73.538659
N	59

Comparison using histograms
Menu: Analyze -> Distribution
Drag the RmTemp column to the Y,columns box &
Drag the VAVunit column to the "By" box -> click OK

Oneway Analysis of RmTemp By VAV Unit

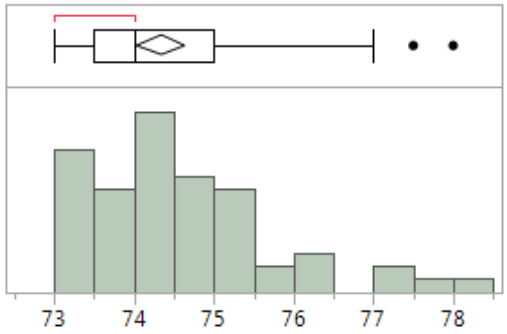


Comparison using "Fit Y by X"
Menu: Analyze -> Fit Y by X
Drag the RmTemp column to the Y,response box &
Drag the VAVunit column to the "X,Factor" box -> click OK
Click the red triangle -> Disply options -> Boxplots to add an overlay of boxplots to the data

Lab #1 Solutions

Using the file VAV_3-06_4-06 :
Menu: Analyze -> Distribution
Drag the Rm.Temp.VAVCO 3-06 column to the Y,columns box & click OK
Click the red triangle on the Rm.Temp.VAVCO 3-06 bar and select "Test Mean"
Enter 73.5 for the specified mean & click OK
The p-value for the two-sided test is given (Prob > |t|)
The confidence interval for the mean is also given

Rm.Temp.VAVCO 3-06



Quantiles		
100.0%	maximum	78
99.5%		78
97.5%		77.75
90.0%		76
75.0%	quartile	75
50.0%	median	74
25.0%	quartile	73.5
10.0%		73
2.5%		73
0.5%		73
0.0%	minimum	73

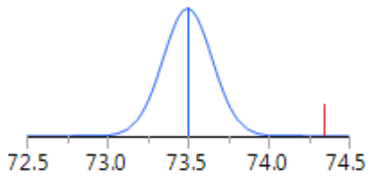
Summary Statistics	
Mean	74.338983
Std Dev	1.1687035
Std Err Mean	0.1521522
Upper 95% Mean	74.643549
Lower 95% Mean	74.034417
N	59

e.

Test Mean	
Hypothesized Value	73.5
Actual Estimate	74.339
DF	58
Std Dev	1.1687

t Test	
Test Statistic	5.5141
Prob > |t|	<.0001*
Prob > t	<.0001*
Prob < t	1.0000

d. reject

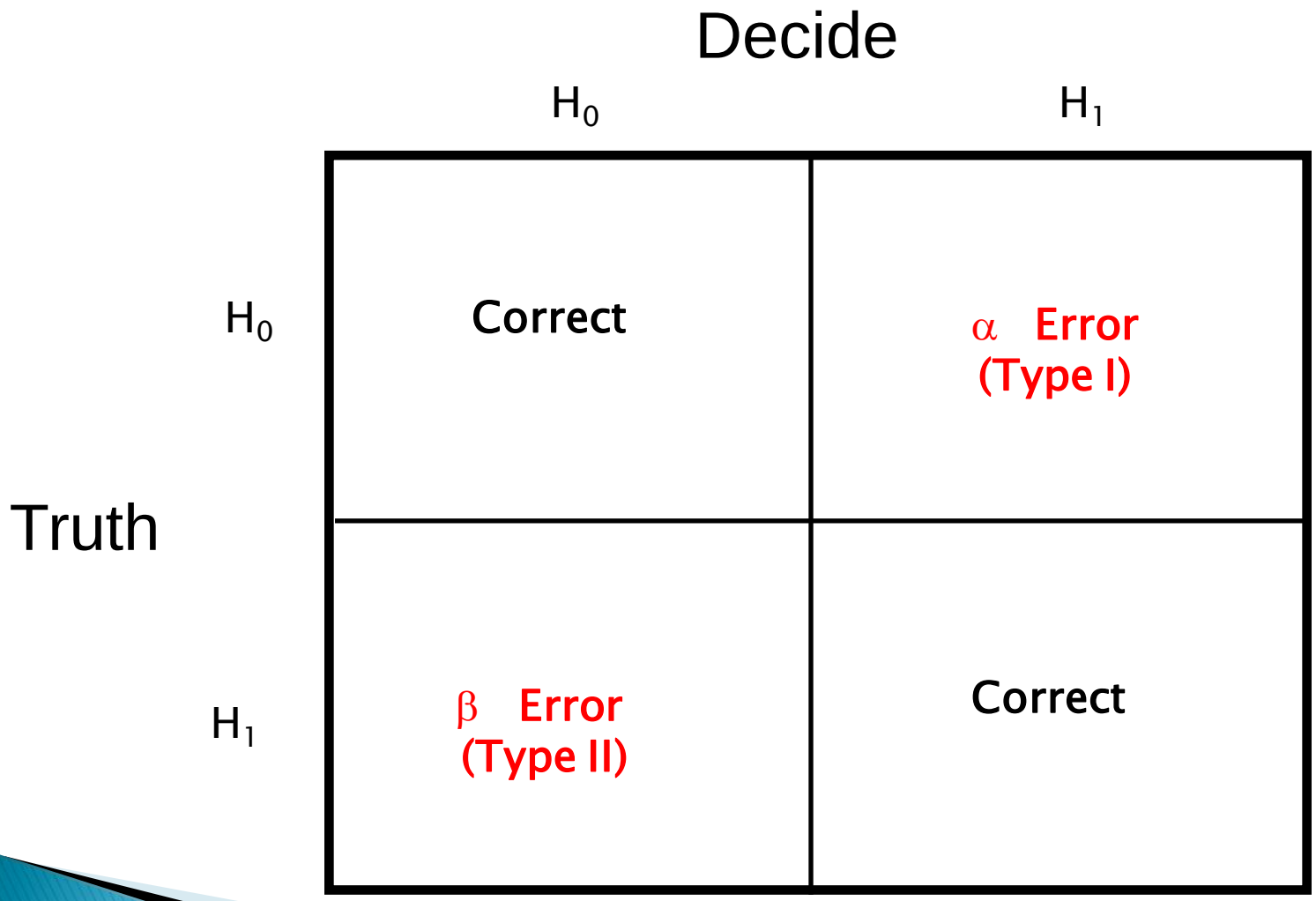


f. Since the confidence interval does not contain 74, the hypothesis that the mean temp is 74 is rejected

HYPOTHESIS TESTING RISKS

Risk is associated with error

Hypothesis Testing Errors



Hypothesis Testing Risks
Alpha and Beta Errors

Alpha α error = P(Reject H_0 | H_0 true)
(Type I error)

Beta β error = P(Accept H_0 | H_1 true)
(Type II error)

Alpha α error = P(Reject H_0 | H_0 true) (Type I error)

For the case where:

$H_0: \mu \leq \mu_0$ vs $H_1: \mu > \mu_0$

We reject H_0 when $Z > z_{\alpha}$ ... where $P(Z > z_{\alpha} \mid \mu = \mu_0) = \alpha$

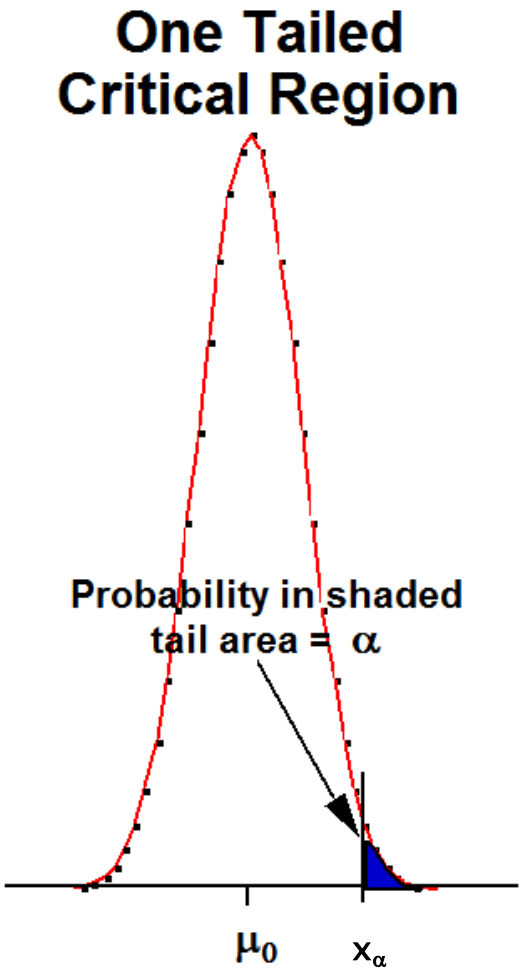
but recall that this is equivalent to rejecting when

$\bar{X} > x_{\alpha}$ where $x_{\alpha} = \mu_0 + z_{\alpha}\,\sigma/\sqrt{n}$

So, $P(\bar{X} > x_{\alpha} \mid \mu = \mu_0) = \alpha$

P(Reject H_0 | H_0 true) = Type I error = α

TYPE I ERROR IS FIXED BY OUR CHOICE

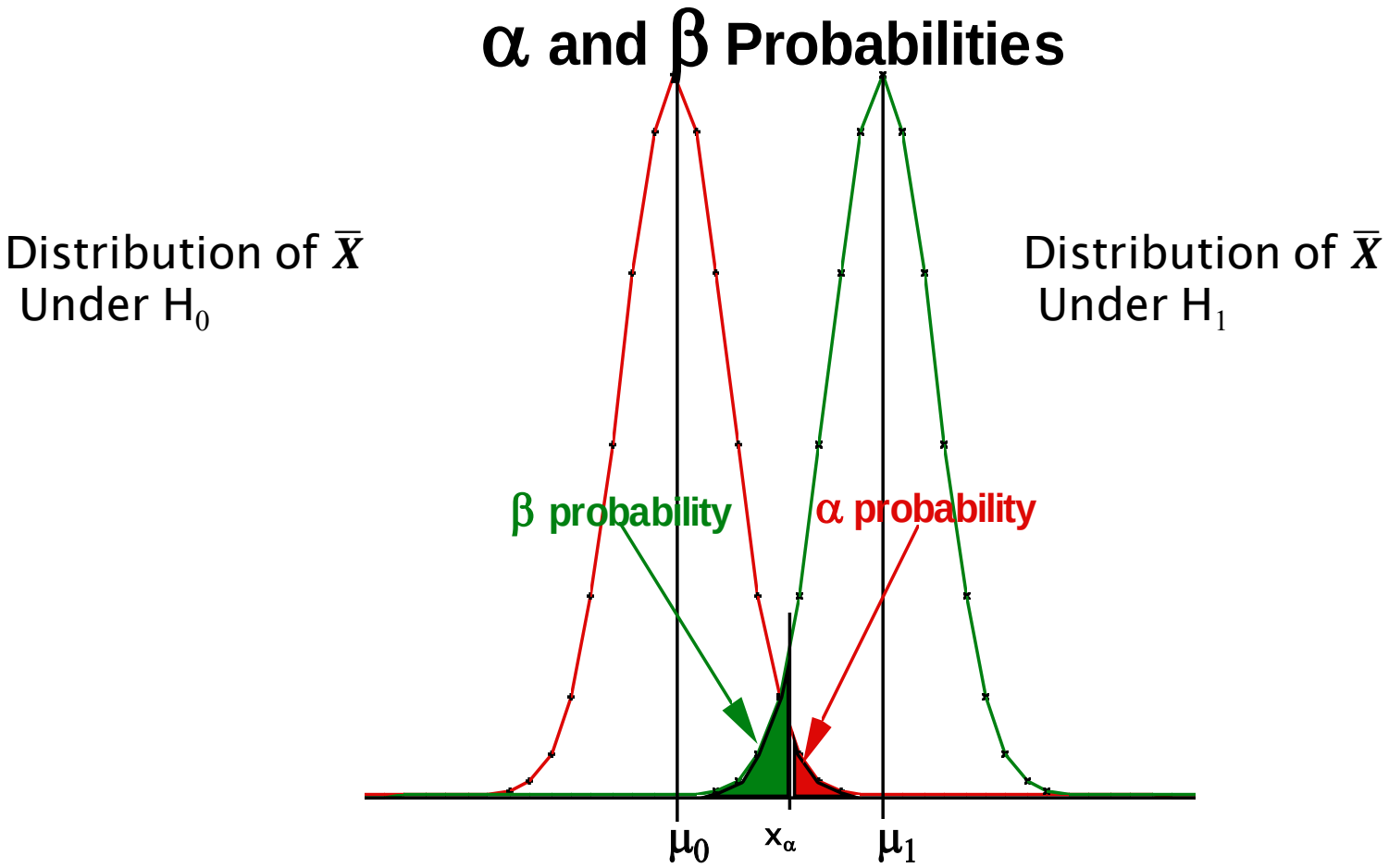


What about the Type II error β ???

β = P(Accept H_0 | $\mu = \mu_1$)
TYPE II ERROR

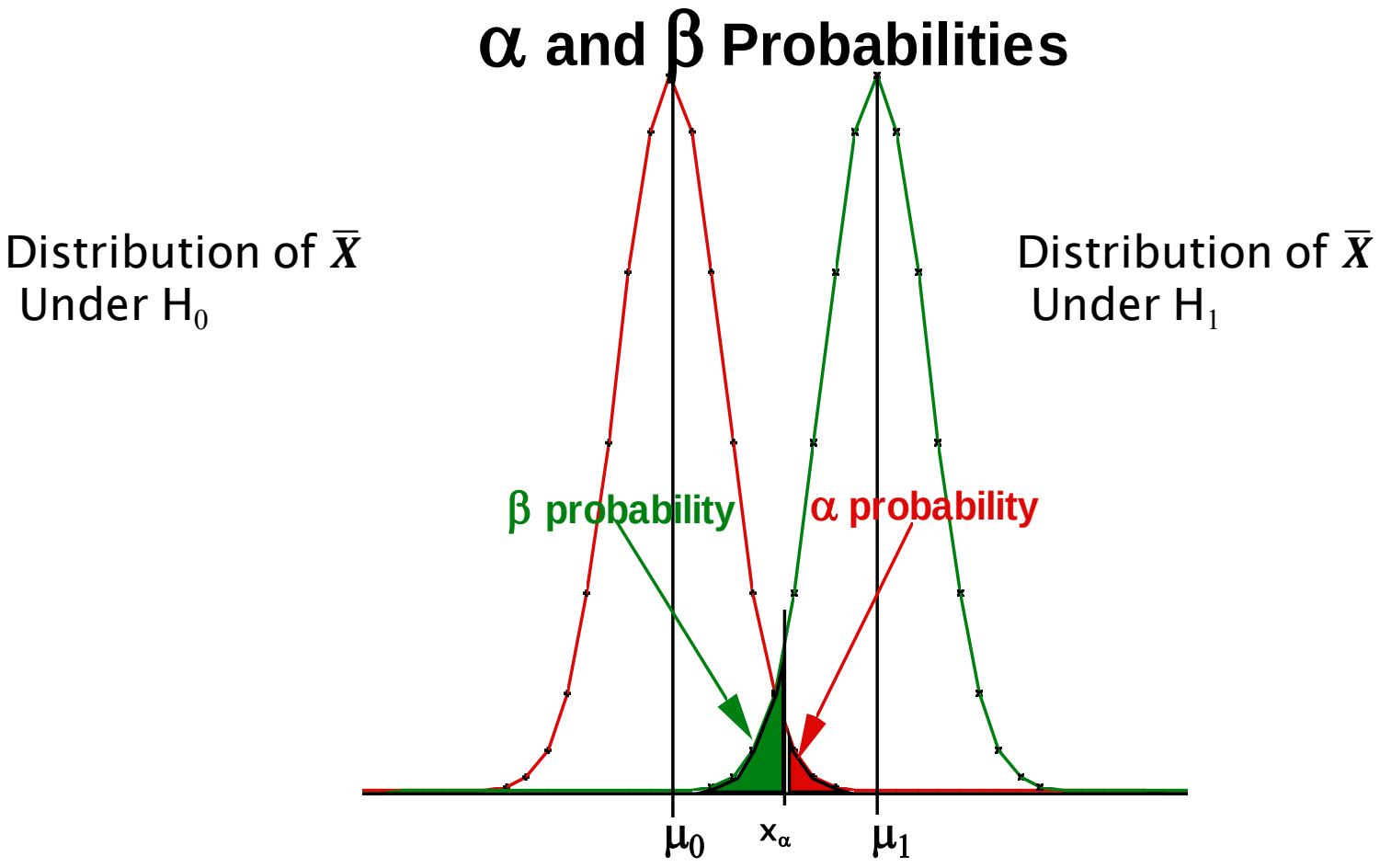
- Type II error depends on a specific choice of the alternative hypothesis $\mu = \mu_1$
- The investigator should think about alternative Hypothetical means that are of importance!
- If it is important to determine that $\mu = \mu_1$ then is β an acceptable level of risk ???
- If not … what can you do???

$\beta = P(\text{Accept } H_0 \mid \mu = \mu_1)$
TYPE II ERROR



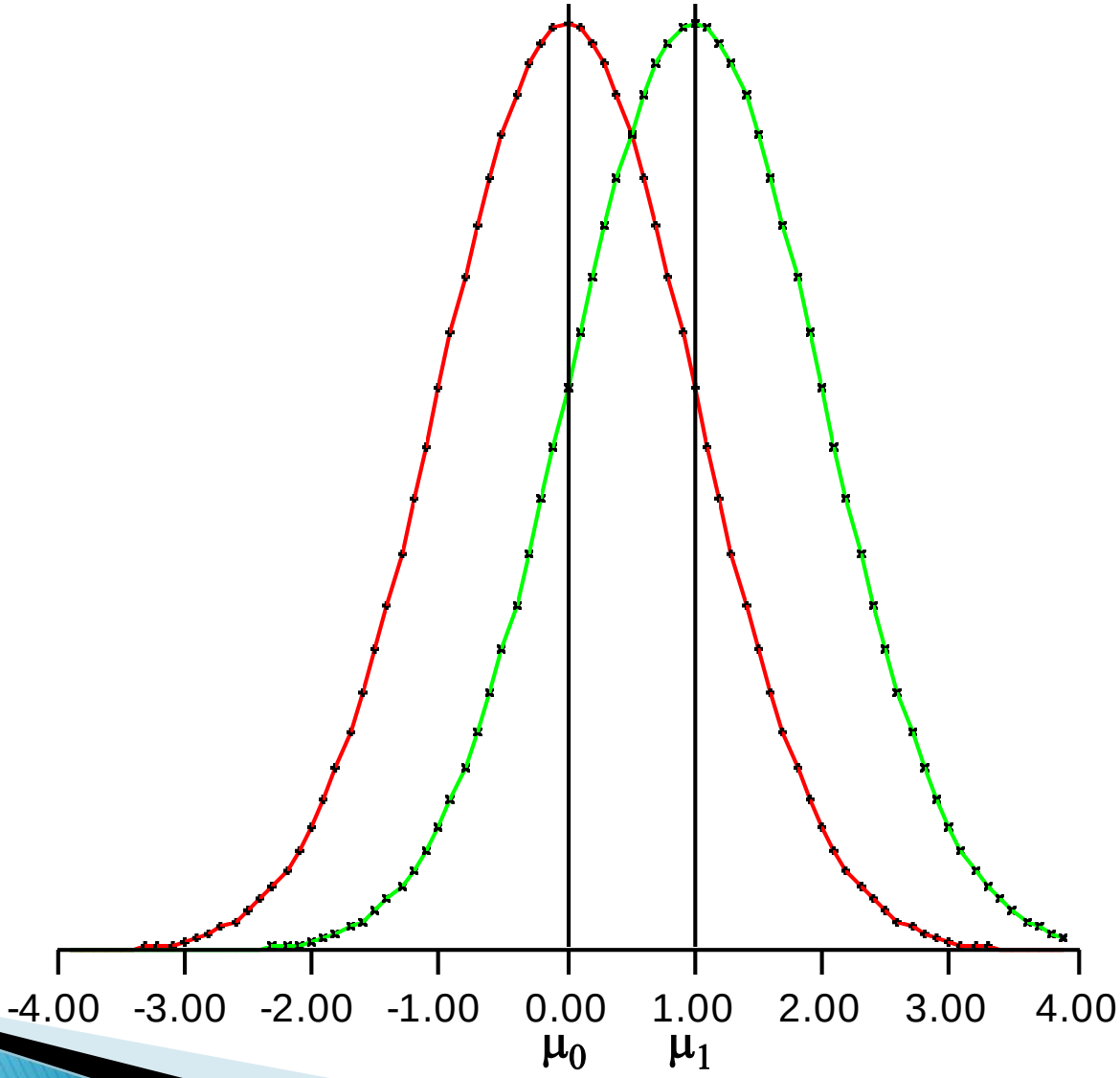
Notice that the overlap in the distributions impacts β

For fixed α, what influences the overlap in the curves and, as a result, the size of β ???



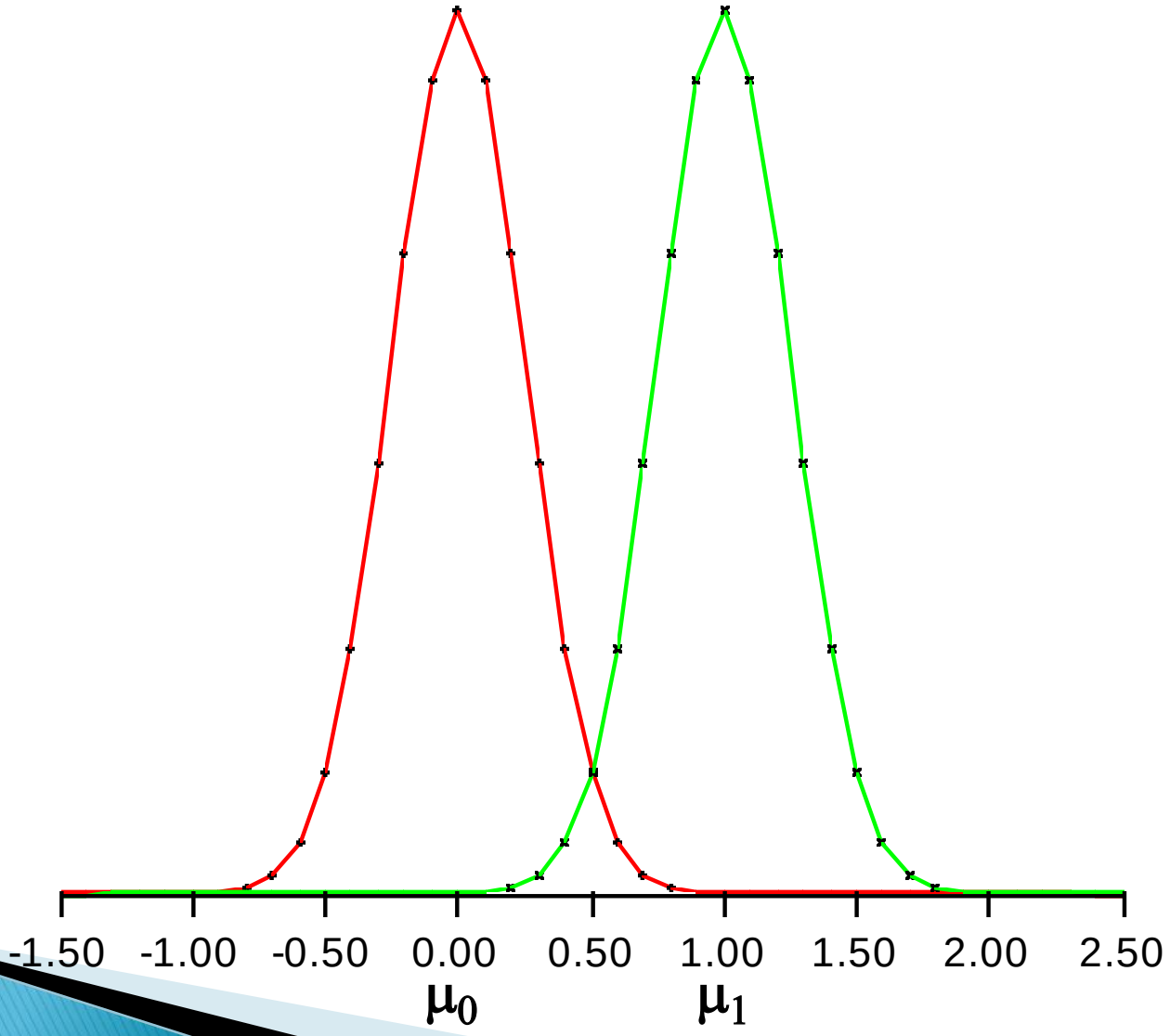
Two Normal Distributions, With $\sigma = 1$

$\mu_0 = 0$ $\mu_1 = 1$



PhaseTwo Analytics

Two Normal Distributions, With $\sigma = .25$



Recall: The Central Limit Theorem

Regardless of the distribution of a random variable, X, with mean μ and standard deviation σ :

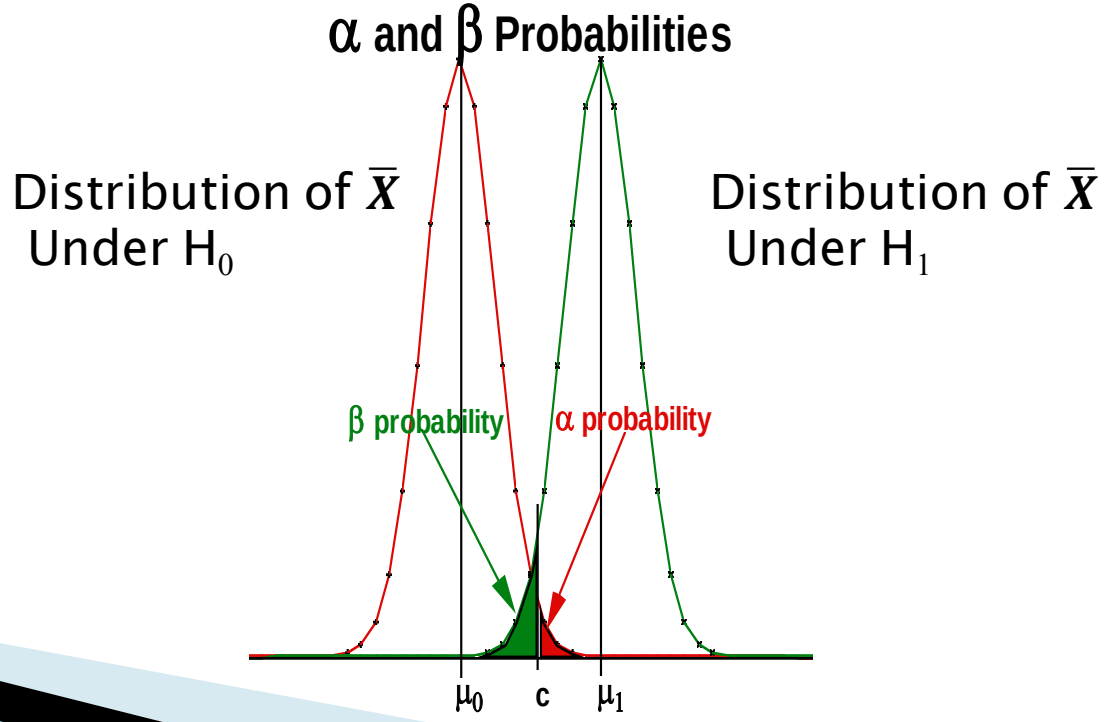
The distribution of sample means, $\bar{X}$, calculated from n independent samples from X will have mean μ, standard deviation $\sigma/\sqrt{n}$, and the distribution will be *increasingly normally distributed* as n gets larger

- ✓ **So: To separate distributions, increase sample size n**
- ✓ **Taking a sample n=16 decreases the s.d. of $\bar{X}$ by 4 !**

Power = 1 – β
Pr (Reject H_0 | H_1 is true)

What affects Power/ β

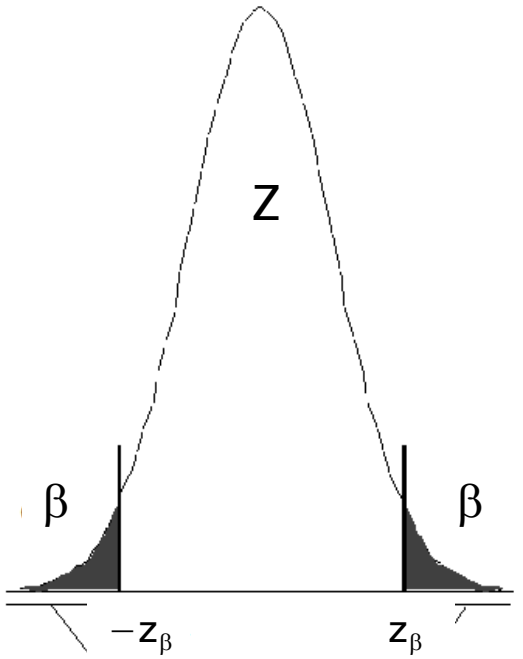
- Choice of α
- Mean Separation $\mu_1 - \mu_0$
- σ
- Sample size n



PhaseTwo Analytics

Determining the sample size needed to achieve a desired β

- $\beta = P(\text{Accept } H_0 \mid \mu = \mu_1)$
- $= P\left(\frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}} < z_\alpha \mid \mu = \mu_1 \right)$
- $= P\left(\frac{\bar{X} - \mu_1}{\sigma/\sqrt{n}} < \frac{(\mu_0 - \mu_1) + z_\alpha \sigma/\sqrt{n}}{\sigma/\sqrt{n}} \mid \mu = \mu_1 \right)$
- $= P\left(Z < \frac{(\mu_0 - \mu_1)}{\sigma/\sqrt{n}} + z_\alpha \right) = \beta$ **if**
- $\frac{(\mu_0 - \mu_1)}{\sigma/\sqrt{n}} + z_\alpha = -z_\beta$ solving for n
- $$n = \left[\frac{(z_\alpha + z_\beta)\sigma}{(\mu_1 - \mu_0)} \right]^2$$



PhaseTwo Analytics

Example

Suppose:

$\mu_1 - \mu_0 = 1 \quad \sigma = 1 \quad \alpha=5\% \ (z_\alpha=1.645) \quad \beta=2.5\% \quad (z_\beta=1.96)$

$$n = \left\lceil \frac{\sigma(z_\alpha + z_\beta)}{\mu_1 - \mu_0} \right\rceil^2 = \left\lceil \frac{1(1.96 + 1.645)}{1} \right\rceil^2 = 12.996$$

So, for the values of μ_0, μ_1 and σ above, a sample of $n=13$ will guarantee that not only will we reject H_0 $\alpha=5\%$ of the time when it's true, but if the mean is really μ_1 (or farther from μ_0), we will fail to reject H_0 only 2.5% of the time yielding a Power of 97.5%

Determining n When σ is Unknown **but** an estimate, s, based on ν df exists

$$n = \left\lceil \frac{s\left(t_{\beta} + t_{\alpha}\right)}{\mu_1 - \mu_0} \right\rceil^2$$

Note: z values are not affected by n … BUT

t values change as n changes

So … determining n for a t-test is iterative

Example

Assume the same conditions as for the previous z example:

$\mu_1 - \mu_0 = 1$ $\sigma = 1$ α=5% (z_α=1.645) β=2.5% (z_β=1.96)

Assume s= 1 is an estimate of σ based on ν = 12 df

- In a 1st iteration use t_α and t_β with ν = 12 df

When $\nu = 12$ $t_\beta = 2.179$ $\beta = 2.5\%$

$t_\alpha = 1.782$ $\alpha = 5\%$

$\mu_1 = 1$ $\mu_0 = 0$ $s = 1$

$$n = \left\lceil \frac{s(t_\beta + t_\alpha)}{\mu_1 - \mu_0} \right\rceil^2 = \left[\frac{1(2.179 + 1.782)}{1} \right]^2 = 15.689 = 16$$

Example continued

- Now, assume that the sample size for the 2nd iteration is 16

so $\nu = 15$ $t_\beta = 2.131$ $(\beta = 2.5\%)$

$t_\alpha = 1.753$ $(\alpha = 5\%)$

$$n = \left\lceil \frac{s(t_\beta + t_\alpha)}{\mu_1 - \mu_0} \right\rceil^2 = \left\lceil \frac{1(2.131+1.753)}{1} \right\rceil^2 = 15.085 = 16$$

Always round up!

n has stopped changing
therefore n=16 samples are needed

Comparing n required when σ is Known to when σ is Unknown

With *known* σ need 13 samples

With *unknown* σ need 16 samples

3 samples are the cost of not knowing σ

Procedure for Determining n When σ is Unknown

1. Start with the equivalent z values for α and β and derive a value for n using an estimate for σ

2. Recalculate n based on the degrees of freedom (n) solved for in the previous step and using the t-values for α and β

3. Repeat step 2 with the most recent value of n

4. Stop the iteration when you get the same value for n twice in a row

Testing "known" σ^2

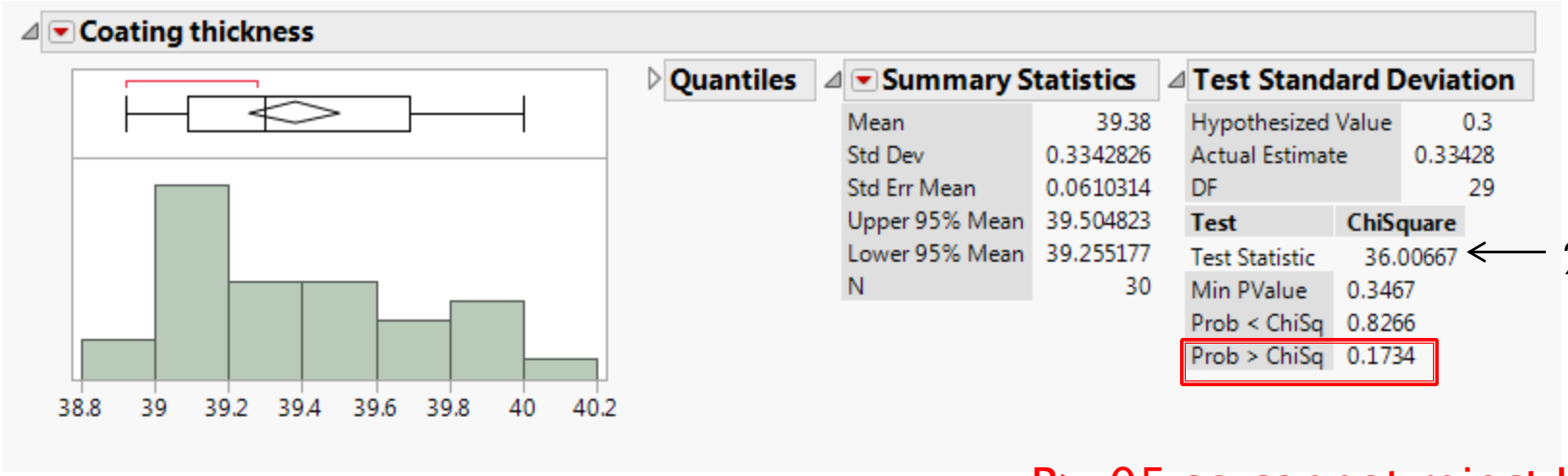
- Suppose the "known" value, σ_0^2, of σ^2 is questioned
  - A change is made to a process design that could affect the "historical" value of σ^2
- PROCEDURE:
  - Confirm the hypothesis that $\sigma^2 = \sigma_0^2$ using the χ^2 based hypothesis test for variances described below
  - If H_0 is not rejected, continue with the hypothesis test of the mean
  - If H_0 is rejected, follow the procedure for σ^2 unknown

χ^2 hypothesis test for variances

- Hypothesis: H_0: $\sigma=\sigma_0$ vs H_a: $\sigma > \sigma_0$
- Form the test statistic: $\chi^2_\nu = \frac{s^2\nu}{\sigma^2}$
- Most software will easily calculate a p-value … $P(\chi_{n-1}^2 > \chi^2)$
- Assumes Xs are independent and Normally distributed!

Coating Thickness example revisited …

- In the coating thickness example we assumed the Standard deviation was $\sigma = .3$
- From the data n= 30 and s= .33428 (10% larger than assumed)
- Let's test the assumption that $\sigma^2 = .3$



$$\chi^2_\nu = \frac{s^2\nu}{\sigma^2}$$

P>.05 so cannot reject H_0

Comments on testing "known" σ^2

- It is a bit unusual to be unsure about a mean μ but claim to know σ^2
- The "unknown σ^2" approach to testing μ is the path usually taken

COMPARING TWO POPULATIONS

Hypothesis Testing: Comparing means of 2 Populations

- Comparing means of populations is commonplace in data science
  - Is the mean tensile strength of parts coming from Process 1 different than the mean from Process 2?
  - Is the mean NOC Alarm Duration in North America longer than that of the APAC region?
  - Is the mean temp at VAV start higher on the 3rd floor than the fifth floor?

Typical Hypotheses

- $H_0: \mu_1 \leq \mu_2$ vs $H_1: \mu_1 > \mu_2$
- $H_0: \mu_1 \geq \mu_2$ vs $H_1: \mu_1 < \mu_2$
- $H_0: \mu_1 = \mu_2$ vs $H_1: \mu_1 \neq \mu_2$

- BUT, these hypotheses are equivalent to:

- $H_0: \mu_1 - \mu_2 \leq 0$ vs $H_1: \mu_1 - \mu_2 > 0$
- $H_0: \mu_1 - \mu_2 \geq 0$ vs $H_1: \mu_1 - \mu_2 < 0$
- $H_0: \mu_1 - \mu_2 = 0$ vs $H_1: \mu_1 - \mu_2 \neq 0$

NOTE: the form of these hypotheses is the same as the single sample mean hypotheses!!!

REMEMBER

$$Z = \frac{\text{statistic} - \text{mean(statistic)}}{\text{s.d.(statistic)}}$$

Where the “Statistic” follows a Normal Dist

TWO SAMPLE TESTING FRAMEWORK

- Suppose n_1 samples are taken from population 1 and n_2 from population 2
  - $X_{11}, x_{12}, x_{13}, \ldots, x_{1n_1}$ iid mean μ_1 and variance σ_1^2
  - $X_{21}, x_{22}, x_{23}, \ldots, x_{2n_2}$ iid mean μ_2 and variance σ_2^2
  - $\overline{X_1}$ will follow a normal distribution with mean μ_1, variance $\frac{\sigma_1^2}{n_1}$
  - $\overline{X_2}$ will follow a normal distribution with mean μ_2, variance $\frac{\sigma_2^2}{n_2}$
  - $\overline{X_1} - \overline{X_2}$ will follow a normal distribution with mean $\mu_1 - \mu_2$, and variance $\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}$
  - AND, $Z = \frac{(\overline{X_1} - \overline{X_2}) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$ follows the familiar standard normal (Z) distribution and **forms the basis of hypothesis tests of** $\mu_1 - \mu_2$

Testing $H_0: \mu_1 - \mu_2 = 0$ vs $H_1: \mu_1 - \mu_2 \neq 0$

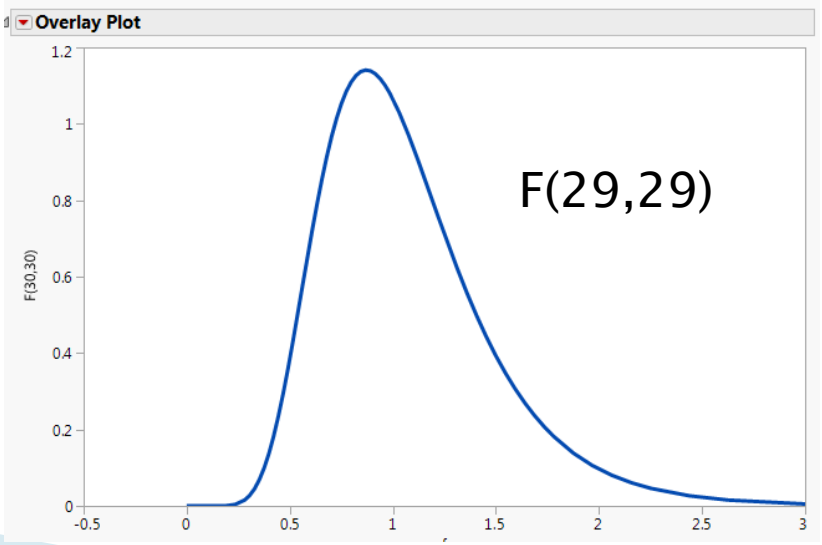
- Case 1 : $\sigma_1^2 = \sigma_2^2$ then test H_0 using
  - $Z = \dfrac{(\overline{X_1} - \overline{X_2})}{\sigma\sqrt{\dfrac{1}{n_1} + \dfrac{1}{n_2}}}$ reject if $|z| > z_{\alpha/2}$
- Case 2 : $\sigma_1^2 \neq \sigma_2^2$ then test H_0 using
  - $Z = \dfrac{(\overline{X_1} - \overline{X_2})}{\sqrt{\dfrac{\sigma_1^2}{n_1} + \dfrac{\sigma_2^2}{n_2}}}$ reject if $|z| > z_{\alpha/2}$

BUT … A REALITY CHECK

- It is rare to know the true σ^2 for each population
- So, will skip exercises for this scenario and go to the realistic situation where σ^2 is unknown
- BUT NOTE … When σ^2 is unknown, a test is needed to 1st determine if, although unknown, $\sigma_1^2 = \sigma_2^2$

How to test for equality of variances

- There are a number of tests
- F-test commonly used
- Under the hypothesis that $\sigma_1^2 = \sigma_2^2$
  - $F = S_1^2 / S_2^2$
  - follows an F distribution with (n_1-1), (n_2-1) df
  - **Assumes that the observations are iid and follow a Normal dist**
  - Software will calculate a p-value for the calculated F. Reject the hypothesis that if $p \leq .05$



Testing framework for unknown variances

- Suppose n_1 samples are taken from population 1 and n_2 from population 2
  - $X_{11}, x_{12}, x_{13}, \ldots, x_{1n_1}$ iid mean μ_1 and **unknown** variance σ_1^2
  - $X_{21}, x_{22}, x_{23}, \ldots, x_{2n_2}$ iid mean μ_2 and **unknown** variance σ_2^2
  - Then, $\overline{X_1}$ will follow a normal distribution with mean μ_1, variance $\frac{\sigma_1^2}{n_1}$
  - $\overline{X_2}$ will follow a normal distribution with mean μ_2, variance $\frac{\sigma_2^2}{n_2}$
  - $\overline{X_1} - \overline{X_2}$ will follow a normal distribution with mean $\mu_1 - \mu_2$, and variance $\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}$
  - AND, $t = \frac{(\overline{X_1} - \overline{X_2}) - (\mu_1 - \mu_2)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$ follows a **t** distribution and forms the basis for testing

Case 1: $H_0: \mu_1 - \mu_2 = 0$ vs $H_1: \mu_1 - \mu_2 \neq 0$ and $\sigma_1^2 = \sigma_2^2$ (unknown)

- Test the hypothesis that $\sigma_1^2 = \sigma_2^2$. If $\sigma_1^2 = \sigma_2^2$
- Test using $t = \dfrac{(\overline{X_1} - \overline{X_2})}{s_p\sqrt{\dfrac{1}{n_1}+\dfrac{1}{n_2}}}$
- Where $s_p = \sqrt{\dfrac{(n_1-1)s_1^2+(n_2-1)s_2^2}{n_1+n_2-2}}$ is the "pooled" sd
- t has n_1+n_2-2 d.f.
- We reject H_0 when $|t| > t_{\alpha/2}(\nu)$ where $\nu = n_1+n_2-2$

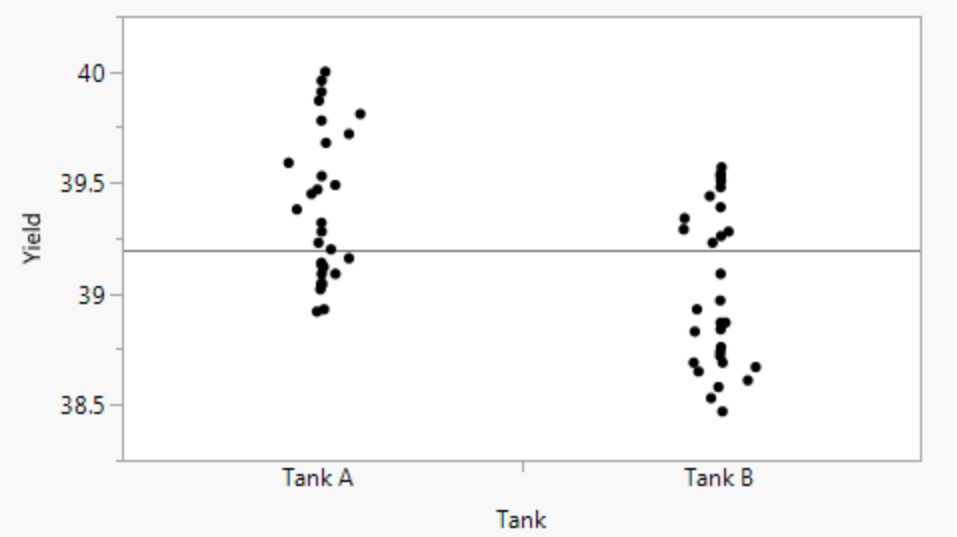
Example: Tank Yield Data

- Batch yield data have been gathered from two reactor tanks in a chemical production process.
- We would like to compare the means but let's first test for equality of variances. Most software provide the means to quickly conduct the F-test shown previously.
- We will conduct a two-tailed test since going into the experiment we don't have a specific alternative hypothesis

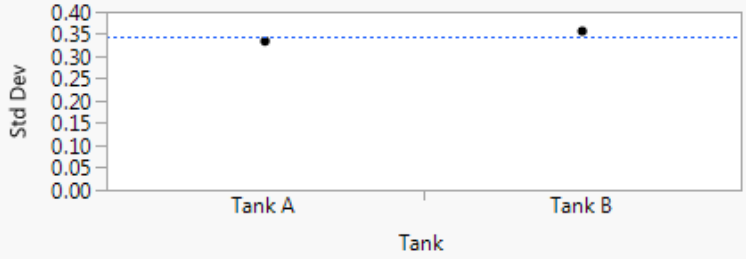
Example: Tank Yield Data

TEST FOR VARIANCE EQUALITY

Menu: Analyze -> Fit Y by X
Drag the Yield column to the Y,response box &
Drag the Tank column to the "X,Factor" box -> click OK
Click the red triangle -> "Unequal Variances"



Tests that the Variances are Equal



Level	Count	Std Dev	MeanAbsDif to Mean	MeanAbsDif to Median
Tank A	30	0.3342826	0.2880000	0.2840000
Tank B	30	0.3565803	0.3193111	0.3110000

Test	F Ratio	DFNum	DFDen	p-Value
O'Brien[.5]	0.3319	1	58	0.5668
Brown-Forsythe	0.2913	1	58	0.5915
Levene	0.6177	1	58	0.4351
Bartlett	0.1188	1	.	0.7304
F Test 2-sided	1.1379	29	29	0.7304

Cannot reject H_0: $\sigma_1^2 = \sigma_2^2$

Example: Tank yield data

- Next: Test the H_0: $\mu_A - \mu_B = 0$ vs H_1: $\mu_1 - \mu_2 \neq 0$
- For the case: $\sigma_1^2 = \sigma_2^2$
- Manually,
- $X_A = 39.38$, $X_B = 39.012$, $s_A^2 = .1117449$, $s_A^2 = .1271495$, $n_A = 30$, $n_B = 30$
- $s_p^2 = (s_A^2 + s_A^2)/2 = .1194472$ so $s_p = .345611$

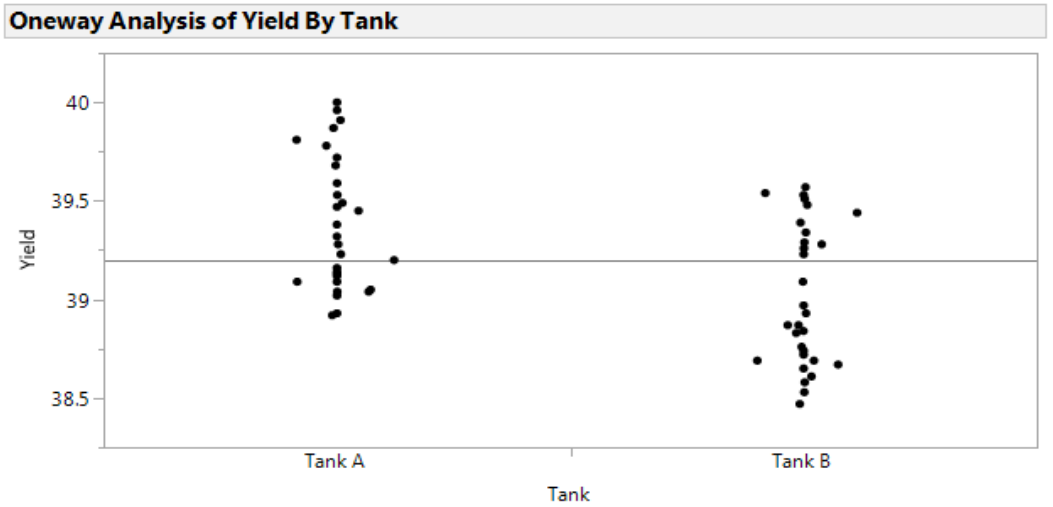
- So, $$t = = \frac{(\overline{X_1} - \overline{X_2})}{s_p\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}} = \frac{(39.38 - 39.012)}{.345611\sqrt{\frac{1}{30}+\frac{1}{30}}} = 4.124$$

- For $= .05$, $t_{\alpha/2}(58) = 2$ since $4.124 > 2$ we reject H_0 !!!

Now conduct the test using JMP, R, Python

- Using JMP

Menu: Analyze -> Fit Y by X
Drag the Yield column to the Y,response box &
Drag the Tank column to the "X,Factor" box -> click OK
Click the red triangle -> "Means/ANOVA/Pooled t"

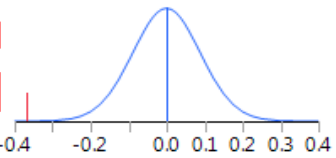


Oneway Anova

Pooled t Test

Tank B-Tank A
Assuming equal variances

Difference	-0.36767	t Ratio	-4.12014
Std Err Dif	0.08924	DF	58
Upper CL Dif	-0.18904	Prob > |t|	0.0001*
Lower CL Dif	-0.54629	Prob > t	0.9999
Confidence	0.95	Prob < t	<.0001*



Case 2: $H_0: \mu_1 - \mu_2 = 0$ vs $H_1: \mu_1 - \mu_2 \neq 0$ and $\sigma_1^2 \neq \sigma_2^2$ (unknown)

- Recall that whether the variances are equal or not, the statistic
- $t = \dfrac{(\overline{X_1} - \overline{X_2})}{\sqrt{\dfrac{s_1^2}{n_1} + \dfrac{s_2^2}{n_2}}}$ follows a t distribution.

- If $\sigma_1^2 \neq \sigma_2^2$, the df for the t distribution is

$$\nu = \frac{\left(\dfrac{s_1^2}{n_1} + \dfrac{s_2^2}{n_2}\right)^2}{\left(\dfrac{s_1^2}{n_1}\right)^2\left(\dfrac{1}{n_1 - 1}\right) + \left(\dfrac{s_2^2}{n_2}\right)^2\left(\dfrac{1}{n_2 - 1}\right)}$$

Example: Comparing Means of Weight Data $\sigma_1^2 \neq \sigma_2^2$ (unknown)

20 daily weight measurements are taken from metal production at two plants. Are the mean daily weights from the plants different?

North Plant Data

48.06	49.25	30.84	51.14
24.42	25.52	44.54	38.92
34.24	47.21	28.69	28.64
31.60	34.24	44.65	22.83
46.84	31.56	45.90	33.65

$\overline{X_N} = 37.137$ $s_N = 9.246$

South Plant Data

43.76	78.91	31.61	45.06
53.23	35.22	58.53	22.37
64.04	65.50	56.25	56.12
30.88	40.33	60.33	62.95
74.83	32.10	34.88	64.36

$\overline{X_S} = 50.564$ $s_S = 16.084$

Weight Data Variance Test

First test whether the two standard deviations are the same using the F test (manually)

H_0: $\sigma_S = \sigma_N$ Let $\alpha = 5\%$

$\overline{X_S} = 50.564$ $s_S = 16.084$

$\overline{X_N} = 37.137$ $s_N = 9.246$

$F = S_s^2 / S_N^2 = \frac{(16.084)^2}{(9.246)^2} = 3.026$ with df (19,19)

Can be shown(later) that Prob of getting a result as or more extreme = .0201, so

P-value = .0201 so we reject H_0: $\sigma_S = \sigma_N$

and conclude *the two processes have different standard deviations*

Weight Data: Testing the means

- Now test that the means are equal using the t test for unequal variances
- H_0: $\mu_N - \mu_S = 0$ and $\alpha = .05$
- $$t = \frac{(\overline{X_1} - \overline{X_2})}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} = \frac{(37.137 - 50.564)}{\sqrt{\frac{9.246^2}{20} + \frac{16.084^2}{20}}} = -3.237$$

$$\nu = \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\left(\frac{s_1^2}{n_1}\right)^2\left(\frac{1}{n_1 - 1}\right) + \left(\frac{s_2^2}{n_2}\right)^2\left(\frac{1}{n_2 - 1}\right)} = \frac{\left(\frac{9.246^2}{20} + \frac{16.084^2}{20}\right)^2}{\left(\frac{9.246^2}{20}\right)^2\left(\frac{1}{19}\right) + \left(\frac{16.084^2}{20}\right)^2\left(\frac{1}{19}\right)} = 30.321$$

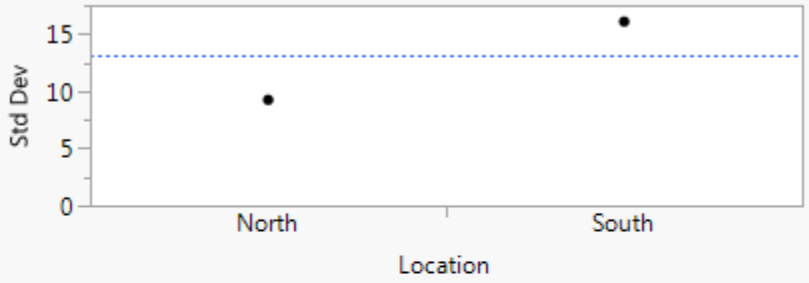
- For $= .05$, $t_{\alpha/2}(30) = 2.042$ since $3.237 > 2.042$ we reject H_0 !!!

Now do using JMP, R, Python

Menu: Analyze -> Fit Y by X
Drag the Weight column to the Y,response box &
Drag the Location column to the "X,Factor" box -> click OK
Click the red triangle -> "Unequal Variances"

Now, Click the red triangle -> t-test

Tests that the Variances are Equal

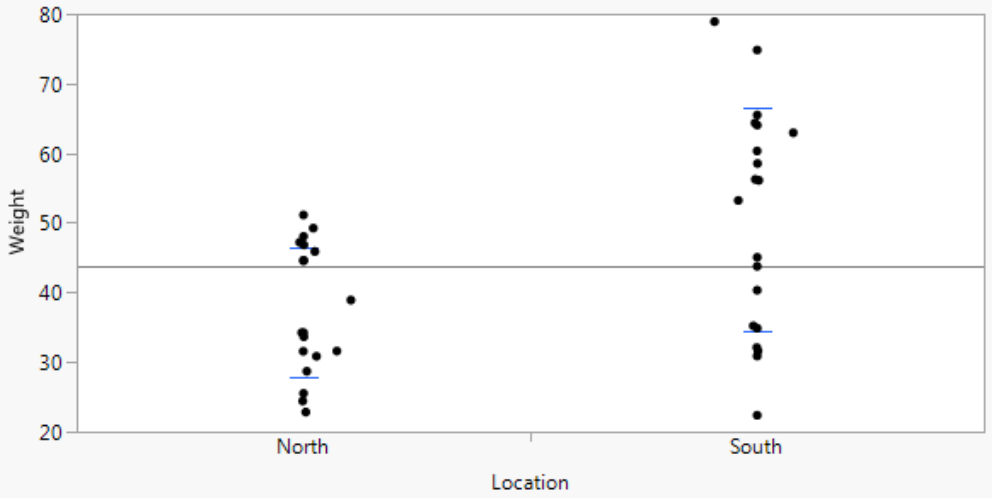


Level	Count	Std Dev	MeanAbsDif to Mean	MeanAbsDif to Median
North	20	9.24585	8.22770	7.93800
South	20	16.08397	13.88570	13.61900

Test	F Ratio	DFNum	DFDen	p-Value
O'Brien[.5]	8.3023	1	38	0.0065*
Brown-Forsythe	5.9085	1	38	0.0199*
Levene	9.1516	1	38	0.0044*
Bartlett	5.4065	1	.	0.0201*
F Test 2-sided	3.0262	19	19	0.0200*

The equal variance hypothesis is rejected

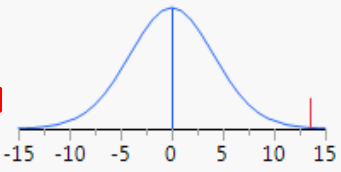
Oneway Analysis of Weight By Location



t Test

South-North
Assuming unequal variances

Difference	13.4260	t Ratio	3.23645
Std Err Dif	4.1484	DF	30.32092
Upper CL Dif	21.8943	Prob > |t|	0.0029*
Lower CL Dif	4.9577	Prob > t	0.0015*
Confidence	0.95	Prob < t	0.9985



The equal means hypothesis is rejected

Confidence Intervals for $\mu_1 - \mu_2$

Confidence intervals for $\mu_1 - \mu_2$

- Case 1: $\sigma_1^2 = \sigma_2^2$ (unknown)
- Since $\dfrac{(\overline{X_1} - \overline{X_2}) - (\mu_1 - \mu_2)}{s_p\sqrt{\dfrac{1}{n_1}+\dfrac{1}{n_2}}}$ follows a **t** distribution, it lies between $\pm t_{\alpha/2}(\nu)$ with probability $1-\alpha$. Rearranging the probability statement yields:
- Pr $\left((\overline{X_1} - \overline{X_2}) - t_{\alpha/2}(\nu)\left(s_p\sqrt{\dfrac{1}{n_1}+\dfrac{1}{n_2}}\right) < (\mu 1 - \mu 2) < (\overline{X_1} - \overline{X_2}) + t_{\alpha/2}(\nu)\left(s_p\sqrt{\dfrac{1}{n_1}+\dfrac{1}{n_2}}\right) \right) = 1 - \alpha$
- where $\nu = n_1 + n_2 - 2$
- So, a $(1-\alpha)$*100% confidence interval for $(\mu 1 - \mu 2)$ is:

$$(\overline{X_1} - \overline{X_2}) - t_{\alpha/2}(\nu)\left(s_p\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}\right) < (\mu 1 - \mu 2) < (\overline{X_1} - \overline{X_2}) + t_{\alpha/2}(\nu)\left(s_p\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}\right)$$

Confidence intervals for $\mu_1 - \mu_2$

- **Case 2:** $\sigma_1^2 \neq \sigma_2^2$
- Since $\dfrac{(\overline{X_1} - \overline{X_2}) - (\mu_1 - \mu_2)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$ follows a **t** distribution, it lies between $\pm t_{\alpha/2}(\nu)$ with probability $1 - \alpha$. Rearranging the probability statement yields:
- $\Pr\left((\overline{X_1} - \overline{X_2}) - t_{\alpha/2}(\nu)\left(\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}\right) < (\mu 1 - \mu 2) < (\overline{X_1} - \overline{X_2}) + t_{\alpha/2}(\nu)\left(\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}\right) \right) = 1 - \alpha$
- So, a $(1 - \alpha)$*100% confidence interval for $(\mu 1 - \mu 2)$ is:

$$(\overline{X_1} - \overline{X_2}) - t_{\alpha/2}(\nu)\left(\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}\right) < (\mu 1 - \mu 2) < (\overline{X_1} - \overline{X_2}) + t_{\alpha/2}(\nu)\left(\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}\right)$$

And recall,

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

Example: Tank yield data revisited

- For the Tank Yield data, construct a 95% confidence interval for $\mu_A - \mu_B$
- Use the confidence interval to test the hypothesis that $\mu_A - \mu_B = 0$
- From the formula: a $(1-\alpha)$*100% confidence interval for $(\mu 1 - \mu 2)$ is:

$$(\overline{X_A} - \overline{X_B}) - t_{\alpha/2}(\nu)\left(s_p\sqrt{\frac{1}{n_A}+\frac{1}{n_B}}\right) < (\mu_A - \mu_B) < (\overline{X_A} - \overline{X_B}) + t_{\alpha/2}(\nu)\left(s_p\sqrt{\frac{1}{n_A}+\frac{1}{n_B}}\right)$$

$X_A = 39.38$, $X_B = 39.012$, $n_A = 30$, $n_B = 30$, $s_p = .345611$, $t_{.025}(58) = 2$

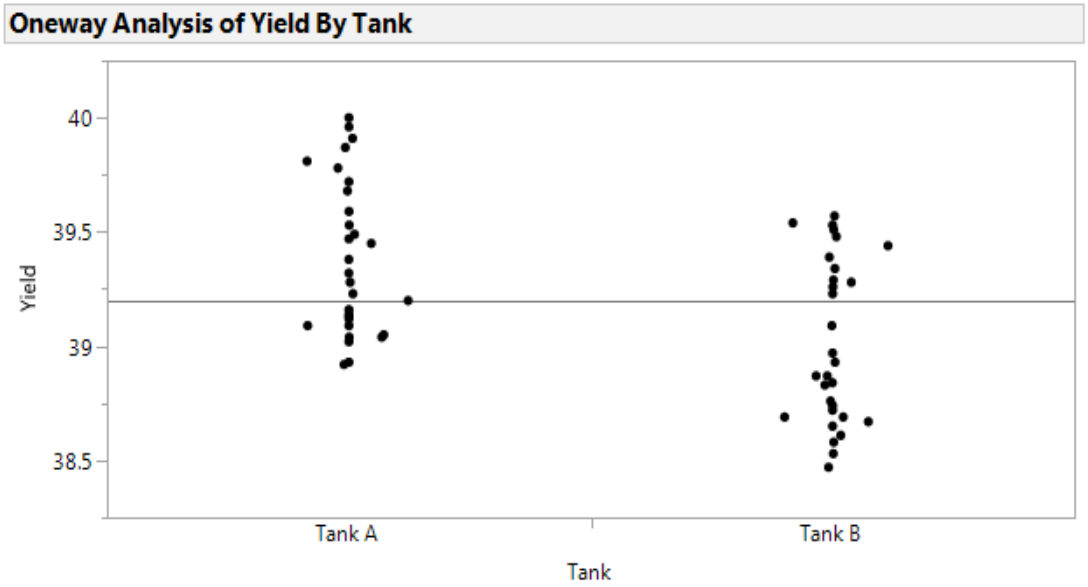
$(.3677 - 2(.345611)(.258199)\ ,\ .3677 + 2(.345611)(.258199)) =$

Or, (.189227, .54617)

Since the confidence interval does not contain 0, the null hypothesis that $\mu_A - \mu_B = 0$ is rejected

Example: Tank yield data revisited
JMP software solution

Menu: Analyze -> Fit Y by X
Drag the Yield column to the Y,response box &
Drag the Tank column to the "X,Factor" box -> click OK
Click the red triangle -> "Means/ANOVA/Pooled t"

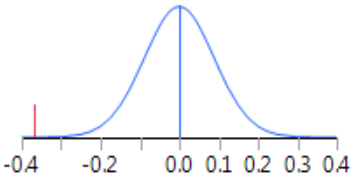


Oneway Anova

Pooled t Test

Tank B-Tank A
Assuming equal variances

Difference	-0.36767	t Ratio	-4.12014
Std Err Dif	0.08924	DF	58
Upper CL Dif	-0.18904	Prob > |t|	0.0001*
Lower CL Dif	-0.54629	Prob > t	0.9999
Confidence	0.95	Prob < t	<.0001*



Example: Construct a confidence interval for $\mu_N - \mu_S$ using the plant metal weight data set and either JMP, R or Python

20 daily weight measurements are taken from metal production at two plants. Are the plants performing differently?

North Plant Data

48.06	49.25	30.84	51.14
24.42	25.52	44.54	38.92
34.24	47.21	28.69	28.64
31.60	34.24	44.65	22.83
46.84	31.56	45.90	33.65

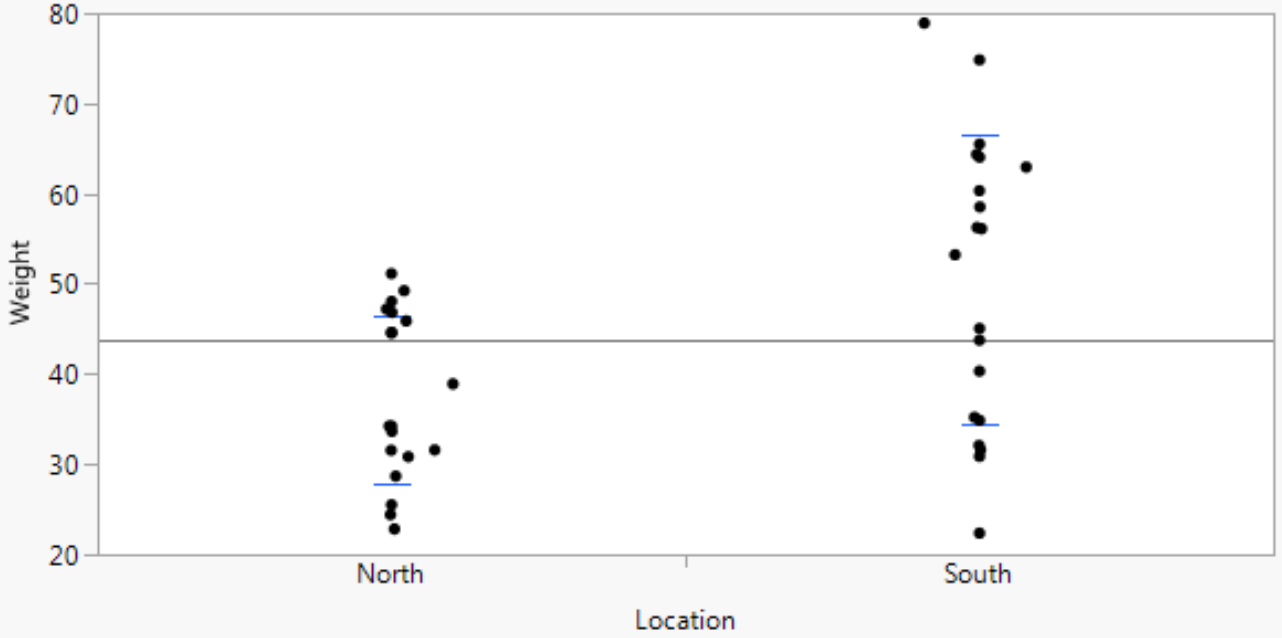
$\overline{X_N} = 37.137$ $s_N = 9.246$

South Plant Data

43.76	78.91	31.61	45.06
53.23	35.22	58.53	22.37
64.04	65.50	56.25	56.12
30.88	40.33	60.33	62.95
74.83	32.10	34.88	64.36

$\overline{X_S} = 50.564$ $s_S = 16.084$

Oneway Analysis of Weight By Location

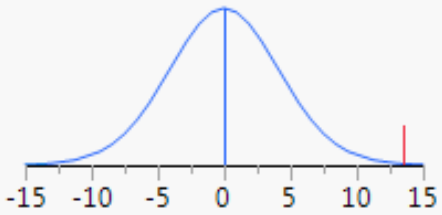


t Test

South-North

Assuming unequal variances

Difference	13.4260	t Ratio	3.23645
Std Err Dif	4.1484	DF	30.32092
Upper CL Dif	21.8943	Prob > |t|	0.0029*
Lower CL Dif	4.9577	Prob > t	0.0015*
Confidence	0.95	Prob < t	0.9985



Commentary on testing for equality of variances as a precursor for testing for equality of means

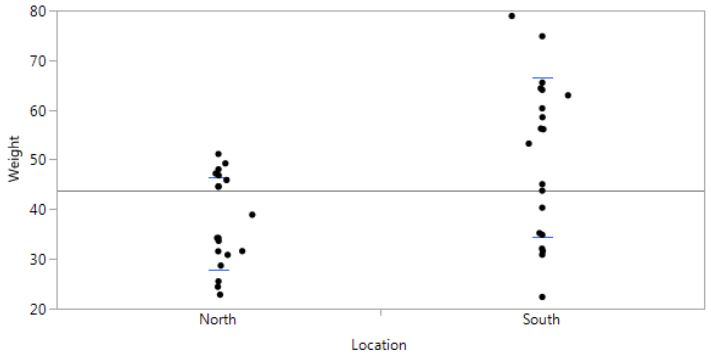
- **Research on robustness done by Prof. George Box in the 1950's concluded the following**
  - The t test is not only robust to the assumption of normality but also to slight differences in variance
  - Many data sets don't follow a normal distribution
  - The test for equality of variances is highly sensitive to the normality assumption to the point that it can double for a Normality test
  - Welch, 1937, showed that when group sizes are equal or near equal, testing means is affected surprisingly little by variance inequality
  - Box recommends proceeding with the t test and not to be concerned about equality of variances unless the variances are obviously very different
  - When groups of observations are of unequal size and variance differences appear large, Welch's proposed test on means will be robust to variance inequality and to non-normality

"To make the preliminary test on variances is rather like putting to sea in a rowing boat to find out whether conditions are sufficiently calm for an ocean liner to leave port"

– George E. P. Box
Biometrika (1953)

Revisiting the Weight data problem, $\sigma_1^2 \neq \sigma_2^2$

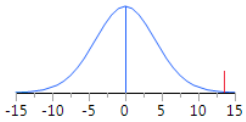
Oneway Analysis of Weight By Location



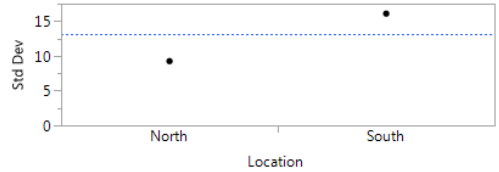
t Test

South-North
Assuming unequal variances

Difference	13.4260	t Ratio	3.23645
Std Err Dif	4.1484	DF	30.32092
Upper CL Dif	21.8943	Prob > |t|	0.0029*
Lower CL Dif	4.9577	Prob > t	0.0015*
Confidence	0.95	Prob < t	0.9985



Tests that the Variances are Equal



Level	Count	Std Dev	MeanAbsDif to Mean	MeanAbsDif to Median
North	20	9.24585	8.22770	7.93800
South	20	16.08397	13.88570	13.61900

Test	F Ratio	DFNum	DFDen	p-Value
O'Brien[.5]	8.3023	1	38	0.0065*
Brown-Forsythe	5.9085	1	38	0.0199*
Levene	9.1516	1	38	0.0044*
Bartlett	5.4065	1	.	0.0201*
F Test 2-sided	3.0262	19	19	0.0200*

Welch's Test

Welch Anova testing Means Equal, allowing Std Devs Not Equal

F Ratio	DFNum	DFDen	Prob > F
10.4746	1	30.321	0.0029*

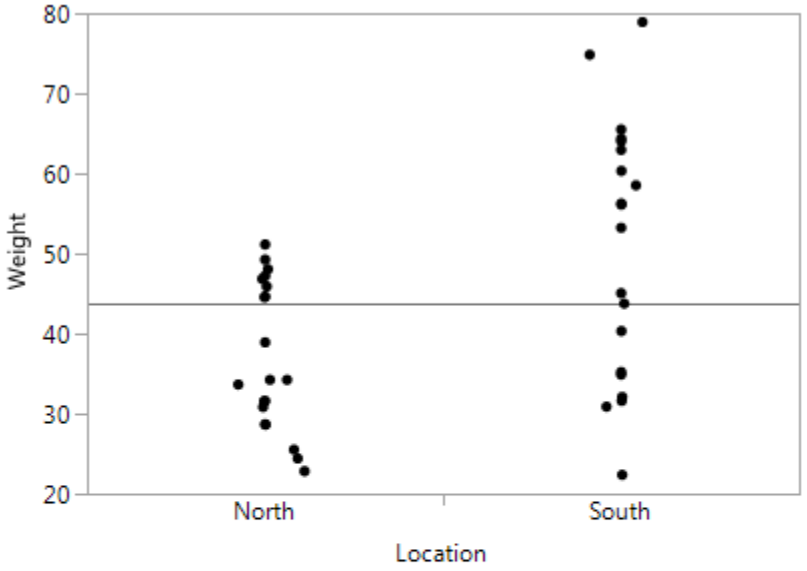
t Test
3.2365

PhaseTwo Analytics

Revisiting the Weight data problem, $\sigma_1^2 = \sigma_2^2$ (The variances aren't greatly different. Box would suggest that you not worry about the potential variance difference)

Test assuming equal variance

Oneway Analysis of Weight By Location



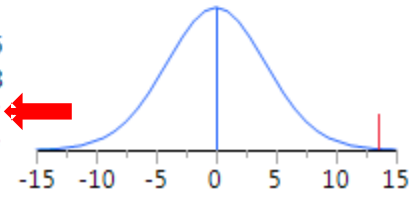
Oneway Anova

Pooled t Test

South-North

Assuming equal variances

Difference	13.4260	t Ratio	3.23645
Std Err Dif	4.1484	DF	38
Upper CL Dif	21.8239	Prob > |t|	0.0025*
Lower CL Dif	5.0281	Prob > t	0.0013*
Confidence	0.95	Prob < t	0.9987



PhaseTwo Analytics

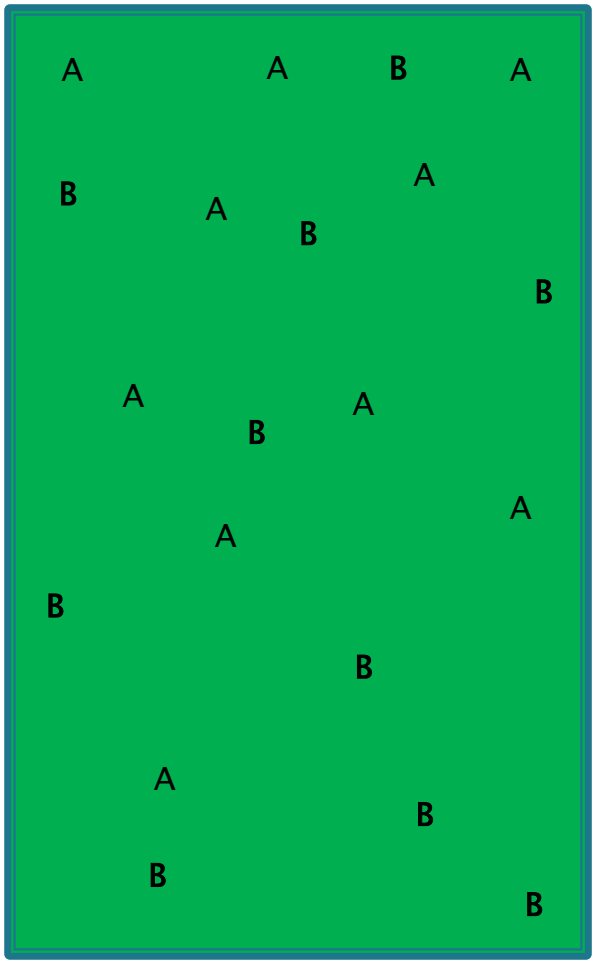
Paired Comparisons: Agricultural example

- We wish to compare the mean yield of corn of Fertilizer A vs Fertilizer B by applying the fertilizers in n randomly selected areas of a farm. The plan is to gather the resultant yield data and test the hypothesis that the means are equal or not (one sided test might be appropriate)

- Problem: there are natural gradients in soil nutrients that will add variability to each measurement. That added variability could blur the comparison between A & B.

- Solution: Conduct each experiment in pairs so that each pair is in near enough proximity to that there is little difference in the soil. For each pair of measurements (X_{Ai}, X_{Bi}), each difference $di = X_{Bi} - X_{Ai}$ is a one sample estimate of the mean difference $\mu_B - \mu_A$ and the difference 'blocks out' the soil nutrient differences across the farm.

Unpaired



Paired

Paired Comparisons

- Let μ be the overall mean yield
- Let τ be the effect of fertilizer
- Let λ be the effect of the location
- X_{Ai} is the Yield at location i using fertilizer A

- Mean of $X_{Ai} = \mu + \tau_A + \lambda_i = \mu_A + \lambda_i$
- Mean of $X_{Bi} = \mu + \tau_B + \lambda_i = \mu_B + \lambda_i$

- By differencing …
- The mean of $d_i = (X_{Bi} - X_{Ai}) = \mu_B - \mu_A$
- We'll say $\mu_d = \mu_B - \mu_A$
- The test of hypothesis about μ_d becomes a simple t test based on the average, $\bar{d}$, of the differences d_i
- $t = \dfrac{\bar{d} - \mu_d}{s_d/\sqrt{n}}$ where s_d is the sd estimate of the differences, is used to test hypotheses about μ_d

Example (Tomato Yield)

- The following table summarizes data from a paired comparison study of Fertilizers A & B applied to plots of tomato plants. The resultant yields are the number of tomatoes picked from each plot

	Location	Yield A	Yield B	d
1	1	132	140	8
2	2	82	88	6
3	3	109	112	3
4	4	143	142	-1
5	5	107	118	11
6	6	66	64	-2
7	7	95	98	3
8	8	108	113	5
9	9	88	93	5
10	10	133	136	3

We'd like to test the hypothesis that the Mean difference in yields is 0

The following plots yield visually revealing information

Fig 1 indicates that the distributions overlap quite a bit if pairing is not considered
Fig 2 tells quite a different story
Fig 3 indicates the majority of the differences are positive

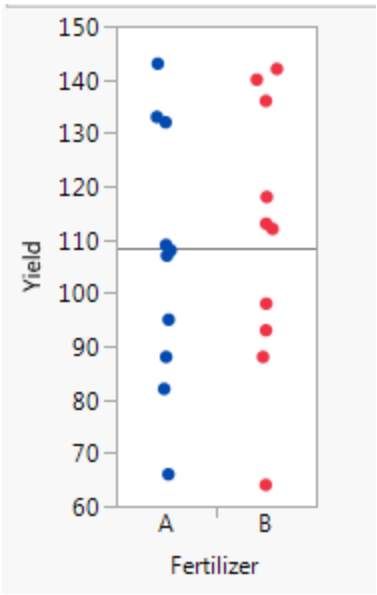


Fig 1

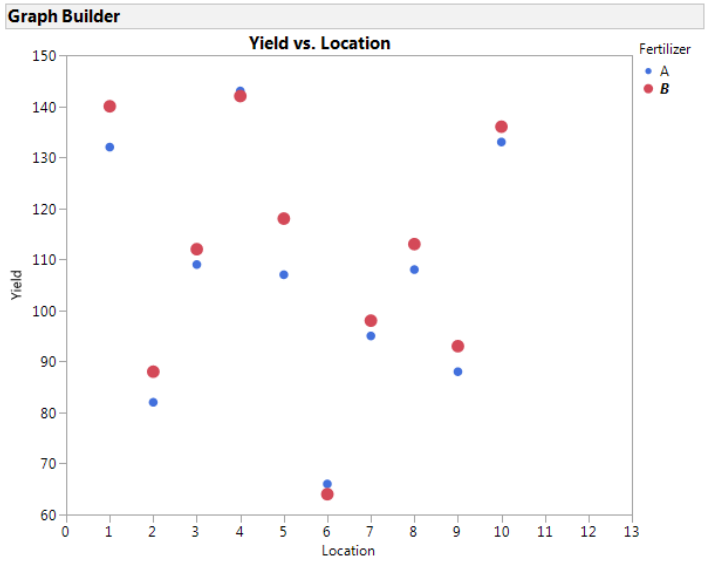


Fig 2

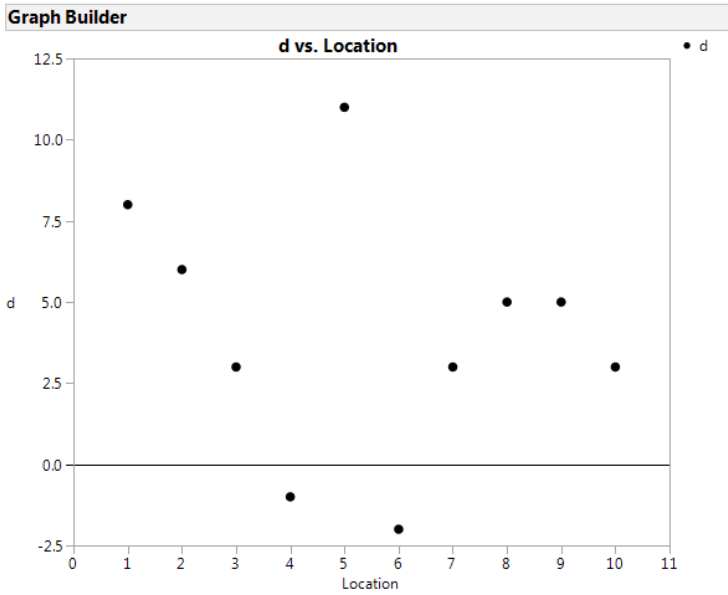


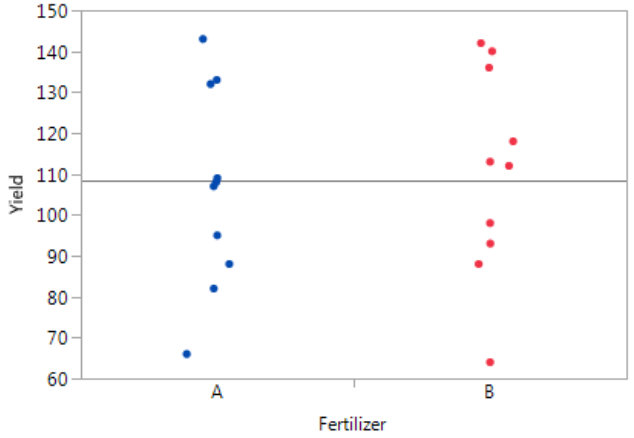
Fig 3

PhaseTwo Analytics

A difference in conclusions

UNPAIRED

Oneway Analysis of Yield By Fertilizer



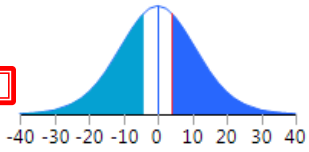
Oneway Anova

Pooled t Test

B-A

Assuming equal variances

Difference	4.100	t Ratio	0.368911
Std Err Dif	11.114	DF	18
Upper CL Dif	27.449	Prob > |t|	0.7165
Lower CL Dif	-19.249	Prob > t	0.3582
Confidence	0.95	Prob < t	0.6418

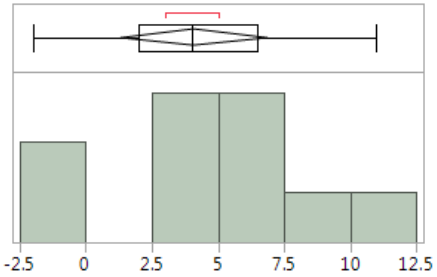


Hypothesis that $\mu_d = 0$ cannot be rejected
If pairing is not used

PAIRED

Distributions

d



Summary Statistics

Mean	4.1
Std Dev	3.8715486
Std Err Mean	1.2242912
Upper 95% Mean	6.8695391
Lower 95% Mean	1.3304609
N	10

Test Mean

Hypothesized Value	0
Actual Estimate	4.1
DF	9
Std Dev	3.87155

t Test

Test Statistic	3.3489
Prob > |t|	0.0085*
Prob > t	0.0043*
Prob < t	0.9957

Reject hypothesis that $\mu_d = 0$!

Class Lab #2: Testing for a Mean Difference in Two Populations

- Room temperature data from VAV locations 3-06 and 4-06 were collected for the 1st half of 2017 and are contained in file "VAV 3-06 4-06 1stHalf_stacked.xls". (An unstacked version of the data is contained in file "VAV 3-06 4-06 1stHalf.xls".) The question was asked as to whether the mean temperatures from the two locations were the same or not since they represented data from identical locations except for different floors.

  a. Use histograms, plots and boxplots to visually compare the location temperatures. What visual comparisons can you make?

  b. Test the hypothesis that the means are the same using the two sample t-test procedure and your software of choice. (Variances appear to be the same)

  c. Create a confidence interval for the mean difference and check to see that it aligns with your test result in b.

  d. Using the data in "VAV 3-06 4-06 1stHalf.xls", visually explore temp variation and differences, then test the hypothesis that the means are the same using a paired comparison approach.

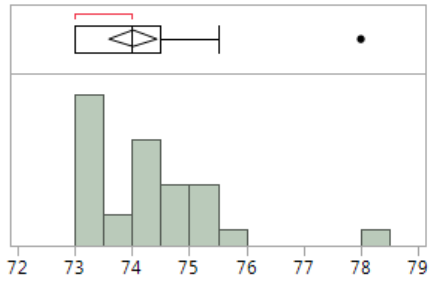
Class Lab #2 Solutions

a. Using data set VAV 3-06 4-06 1st Half
Menu->Analyze->Distribution
Drag Rm.Temp.VAVCO 3-06 to the "Y,Columns" box
Drag Rm.Temp.VAVCO 4-06 to the "Y,Columns" box
Click "OK"

Using data set VAV 3-06 4-06 1st Half_stacked
Menu->Analyze->Fit Y by X
Drag RmTemp to the "Y,Response" box
Drag VAV Unit to the "X,Factor" box
Click "OK" to get a visual plot of the data by VAV unit
Click the red triangle -> Display Options -> Boxplots to add boxplots

Distributions

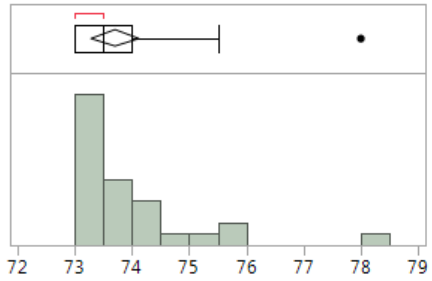
Rm.Temp.VAVCO 3-06



Quantiles		
100.0%	maximum	78
99.5%		78
97.5%		78
90.0%		75
75.0%	quartile	74.5
50.0%	median	74
25.0%	quartile	73
10.0%		73
2.5%		73
0.5%		73
0.0%	minimum	73

Summary Statistics	
Mean	74.017241
Std Dev	1.0977474
Std Err Mean	0.2038466
Upper 95% Mean	74.434802
Lower 95% Mean	73.599681
N	29

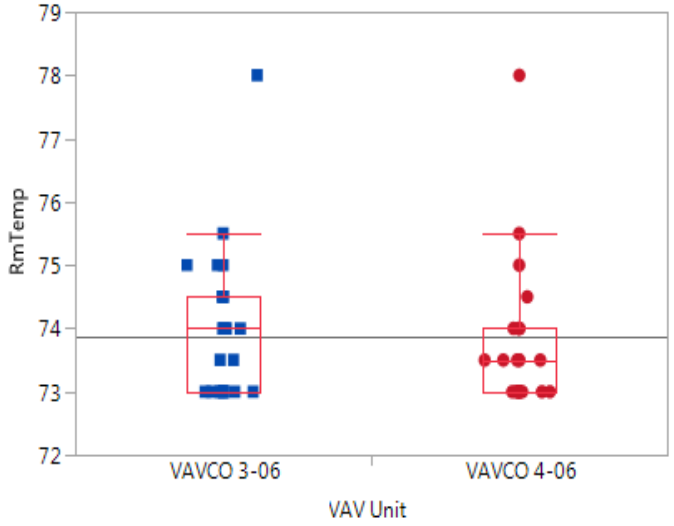
Rm.Temp.VAVCO 4-06



Quantiles		
100.0%	maximum	78
99.5%		78
97.5%		78
90.0%		75.5
75.0%	quartile	74
50.0%	median	73.5
25.0%	quartile	73
10.0%		73
2.5%		73
0.5%		73
0.0%	minimum	73

Summary Statistics	
Mean	73.706897
Std Dev	1.114172
Std Err Mean	0.2068966
Upper 95% Mean	74.130705
Lower 95% Mean	73.283088
N	29

Oneway Analysis of RmTemp By VAV Unit



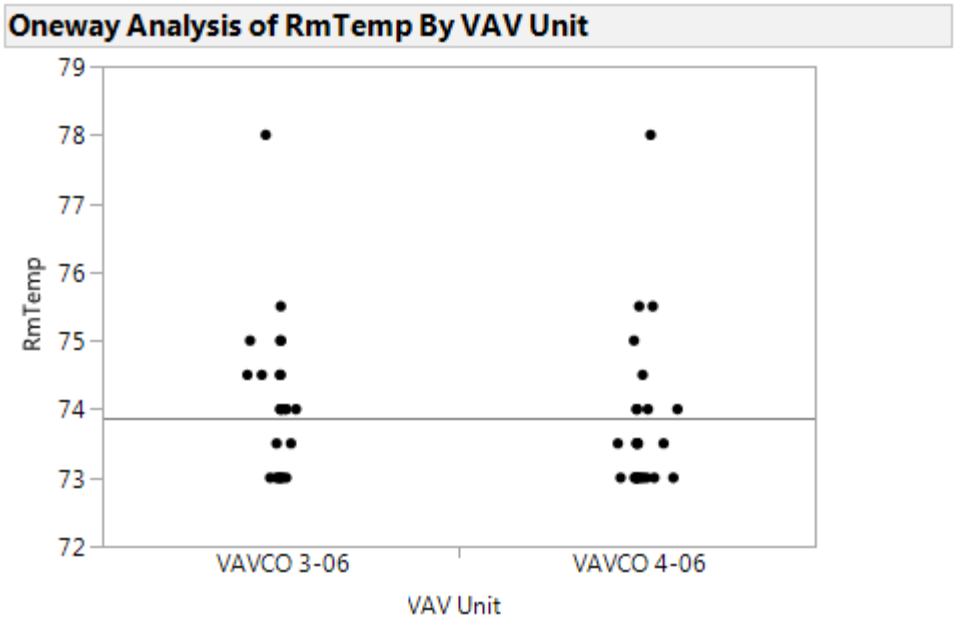
PhaseTwo Analytics

Class Lab #2 Solutions

Within the prior Fit Y by X report window:
Click the red triangle -> Means/ANOVA/ Pooled t

With a p-value of .2899 you cannot reject H_0 that the Means are the same

Notice that the confidence interval for μ_d contains 0 Reaffirming the conclusion in b.



Oneway Anova

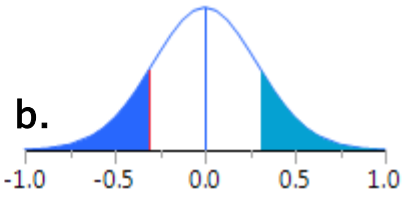
Pooled t Test

VAVCO 4-06- VAVCO 3-06
Assuming equal variances

Difference	-0.31034	t Ratio	-1.06851
Std Err Dif	0.29045	DF	56
Upper CL Dif	0.27149	Prob > |t|	0.2899
Lower CL Dif	-0.89218	Prob > t	0.8551
Confidence	0.95	Prob < t	0.1449

c.

b.

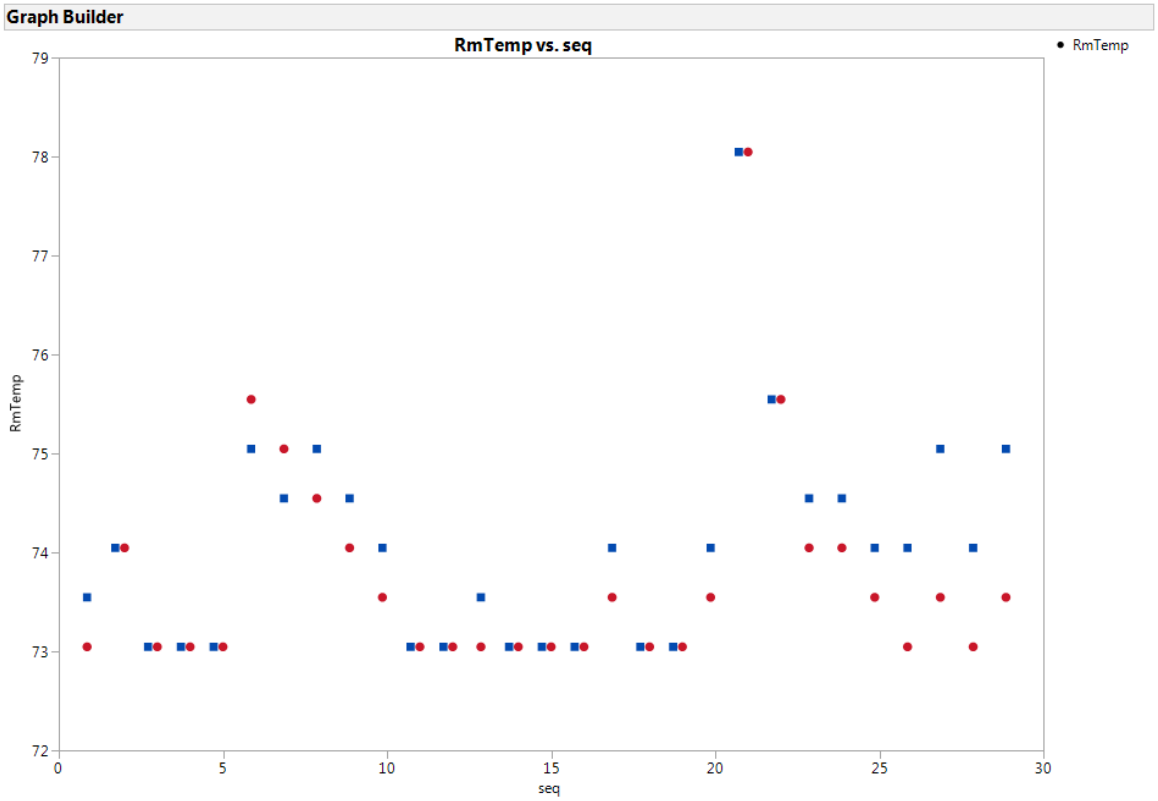


PhaseTwo Analytics

Class Lab #2 Solutions

d. Plotting the data in sequential order shows revealing behavior
For each time period the RmTemp of VAVCO 3-06 is generally
Equal to or Greater then the RmTemp of VAVCO 4-06

Menu->Graph->Graph Builder
Drag RmTemp to the y axis
Drag Seq to the x-axis
Click on Jitter ->Centered Grid



We will examine temp differences next

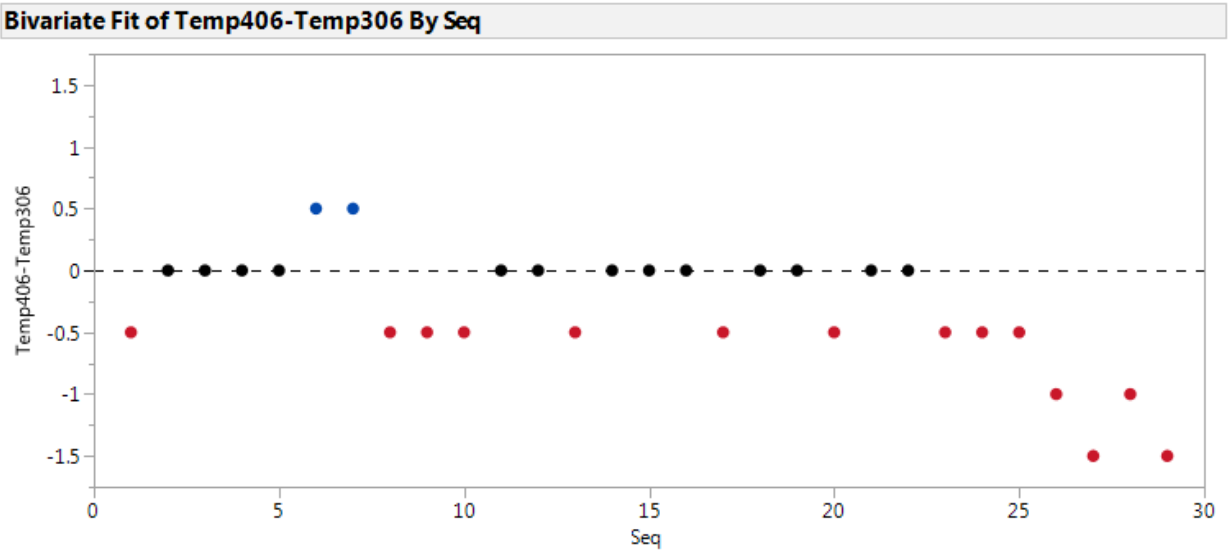
Class Lab #2 Solutions

d. In the dataset VAV 3-06 4-06 1stHalf.xls create the variable Temp306-Temp406 which is the difference in room temp on each day when the units began operation

Note that over the time period studied, temperature varied from 73 to 78 but the room temps were measured At the same time on each day offering an opportunity for a paired comparison of room temps

Seq	Start.time of VAVCO 3-06	Rm.Temp.VAVCO 3-06	Start.time of VAVCO 4-06	Rm.Temp.VAVCO 4-06	Temp406-Temp306
1	03/15/2017 7:00 AM	73.5	03/15/2017 7:01 AM	73	-0.5
2	04/03/2017 7:00 AM	74	04/03/2017 7:00 AM	74	0
3	04/04/2017 7:04 AM	73	04/04/2017 7:04 AM	73	0
4	04/05/2017 7:01 AM	73	04/05/2017 7:01 AM	73	0
5	04/06/2017 7:03 AM	73	04/06/2017 7:08 AM	73	0
6	05/01/2017 7:04 AM	75	05/01/2017 7:04 AM	75.5	0.5
7	05/02/2017 7:02 AM	74.5	05/02/2017 7:02 AM	75	0.5
8	05/03/2017 5:39 AM	75	05/03/2017 5:39 AM	74.5	-0.5

To plot the differences:
Menu-> Analyze-> Fit Y by X
Drag Temp406-Temp306 to Y,Response
Drag Seq to X,Regressor



PhaseTwo Analytics

Class Lab #2 Solutions

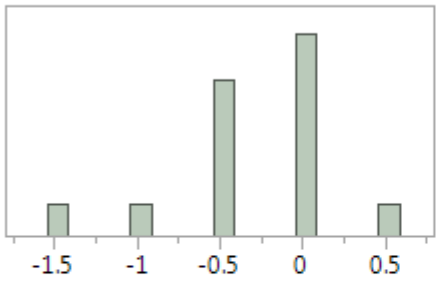
d. In the dataset VAV 3-06 4-06 1stHalf.xls create the variable Temp306-Temp406 which is the difference in room temp on each day when the units began operation
Now use the paired comparison t-test on the differenced data

To run a t-test on the differences:
Menu->Analyze->Distribution
Drag Temp406-Temp306 to Y-columns
Click OK
Click the red triangle on the Temp406-Temp306 bar
Click on Test Mean and enter 0 for specified mean

Alternatively,
Menu->Analyze->Specialized Modeling->Matched Pairs
Drag Rm.Temp.VAVCO 3-06 to "Y,Paired Response"
Drag Rm.Temp.VAVCO 4-06 to "Y,Paired Response"
Click OK

Distributions

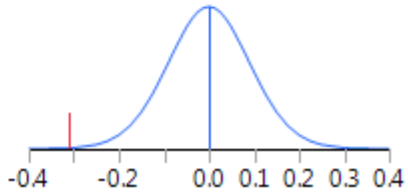
Temp406-Temp306



Summary Statistics	
Mean	-0.310345
Std Dev	0.4894201
Std Err Mean	0.090883
Upper 95% Mean	-0.124179
Lower 95% Mean	-0.49651
N	29

Test Mean	
Hypothesized Value	0
Actual Estimate	-0.3103
DF	28
Std Dev	0.48942

t Test	
Test Statistic	-3.4148
Prob > |t|	0.0020*
Prob > t	0.9990
Prob < t	0.0010*



Using a paired comparison the mean difference
Is shown to be significantly different from 0

This was hidden without pairing!

Determining the Statistical Significance of Regression Coefficients

A Brief Diversion

PhaseTwo Analytics

Statistical Significance of Regression Coefficients

- Objective … we will derive the test procedure for a regression coefficient (a statistic)
  - Review how regression coefficients are determined
  - Show the form of the coefficient estimates
  - Show why they follow an approx Normal distribution
  - Determine the mean of the coefficient
  - Determine the s.e of the coefficient
  - As before,

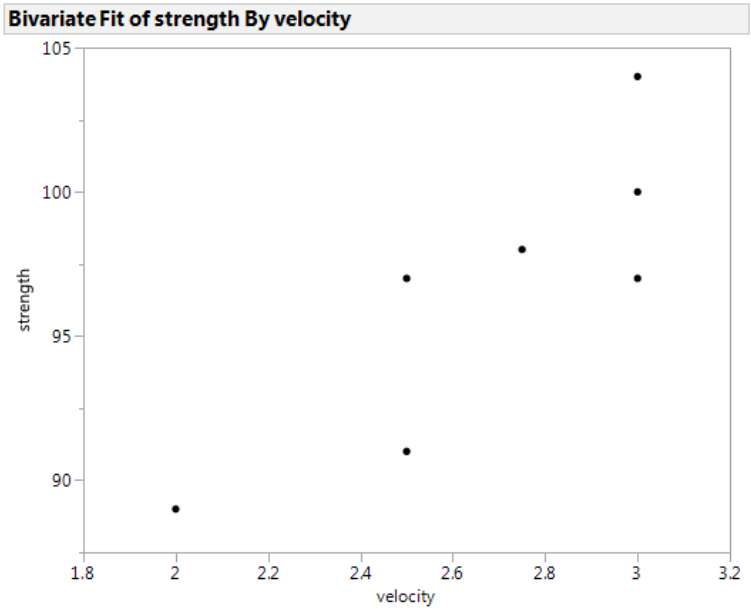
$$\frac{\text{statistic} - \text{mean(statistic)}}{\text{s.e. (statistic)}}$$

Will form a t-statistic that will be the basis for the test of significance for the coefficient

Determining the Statistical Significance of Regression Coefficients

- Let's examine data from an inertia welding process
  - X is the rotational velocity of the workpiece
  - Y is the breaking strength of the weld
  - There appears to be a linear relationship between velocity and strength
  - The predictive model has the form $Y = \beta_0 + \beta X + \varepsilon$ where ε is random error
  - ε is distributed $N(0, \sigma^2)$

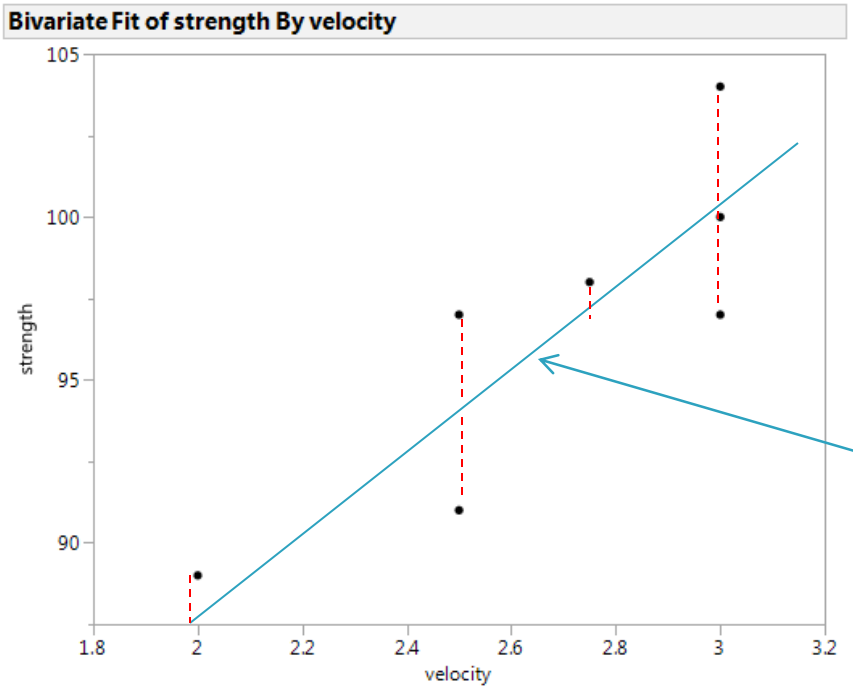
	velocity	strength
1	2	89
2	2.5	97
3	2.5	91
4	2.75	98
5	3	100
6	3	104
7	3	97



Determining Regression Coefficients Least Squares

- Least squares regression mathematically identifies the line through the data that minimizes the sum of the squared 'errors' … i.e. distances of the Y_is from the line $\hat{Y}_i$ where $\hat{Y}_i = b_0 + bx_i$
- SSE = $\sum_{i=1}^{n} (\hat{Y}_i - Y_i)^2$
- SSE/(n−2) provides an estimate of σ^2

Why n-2 ????



Testing Regression Coefficients

- The least squares solution yields:
  - $b = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^{n}(X_i - \bar{X})^2}$
  - $= \left(\frac{1}{\sum_{i=1}^{n}(X_i - \bar{X})^2}\right)\left(\sum_{i=1}^{n}(X_i - \bar{X})(Y_i)\right)$ which is a weighted **sum of Yi** and
  - $b0 = \bar{Y} - b\,\bar{X}$
  - Since b is a weighted sum of the y_i … the CENTRAL LIMIT THEOREM will apply and **b will be approx Normally distributed with increasing n**

Recall

$$t = \frac{\text{statistic} - \text{mean(statistic)}}{\text{s.e.(statistic)}}$$

will follow a t distribution

Testing Regression Coefficients
Testing $H_0: \beta = 0$

- It can be shown (see the appendix):
- $E[b] = \beta$ and
- s.d. (b) $= \sigma / \sqrt{\sum_{i=1}^{n} (X_i - \bar{X})^2}$
- So, s.e. (b) $= s / \sqrt{\sum_{i=1}^{n} (X_i - \bar{X})^2}$ it follows that
- $t = \dfrac{b - \beta}{s / \sqrt{\sum_{i=1}^{n} (X_i - \bar{X})^2}}$ follows a t distribution with n-2 df and is

  used to test $H_0: \beta = 0$ vs $H_a: \beta \neq 0$

Testing Regression Coefficients – Testing Intercept $H_0: \beta_0 = 0$?

- Recall that the least squares estimate of β_0 is: $b_0 = \bar{Y} - b\,\bar{X}$
- Since b follows a normal dist and b_0 just adds $\bar{Y}$ which also follows a Normal dist, then it follows that b_0 follows a Normal dist (approx)
- It can be shown (see the Appendix)
- $E[b_0] = \beta_0$ (so b_0 is an unbiased estimator of β_0)
- And, $s.e.[b_0] = s\sqrt{\frac{\bar{X}^2}{\sum_{i=1}^{n}(X_i - \bar{X})^2}}$
- Thus, $t = \frac{b_0 - \beta_0}{s/\sqrt{\frac{\bar{X}^2}{\sum_{i=1}^{n}(X_i - \bar{X})^2}}}$ provides a t test of the hypothesis that $\beta_0 = 0$

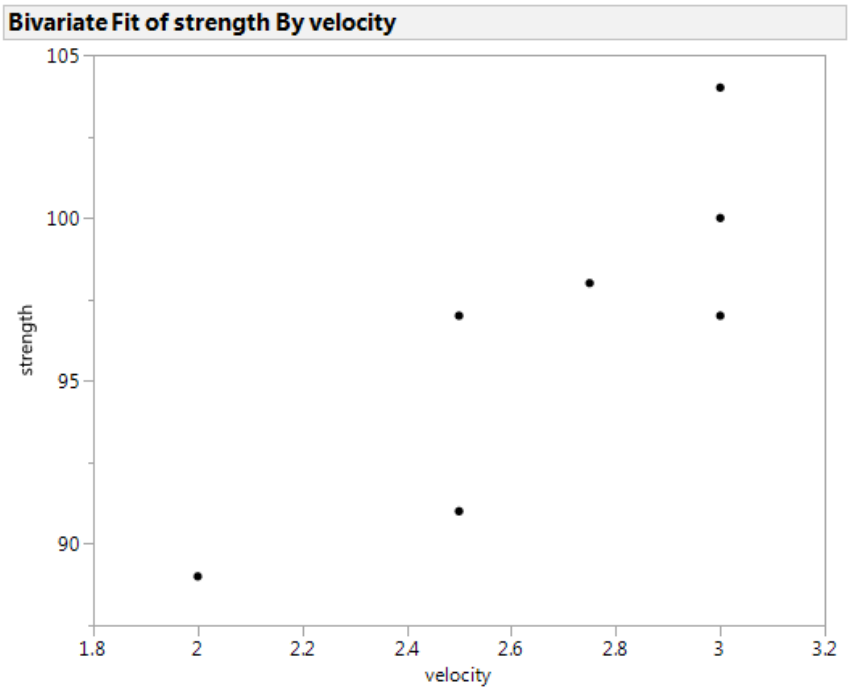
Example 8: Welding data regression

- Fit a line using JMP, Python & R to see how these hypothesis tests occur using software.

  a. Plot Strength vs Velocity to visually explore the relationship

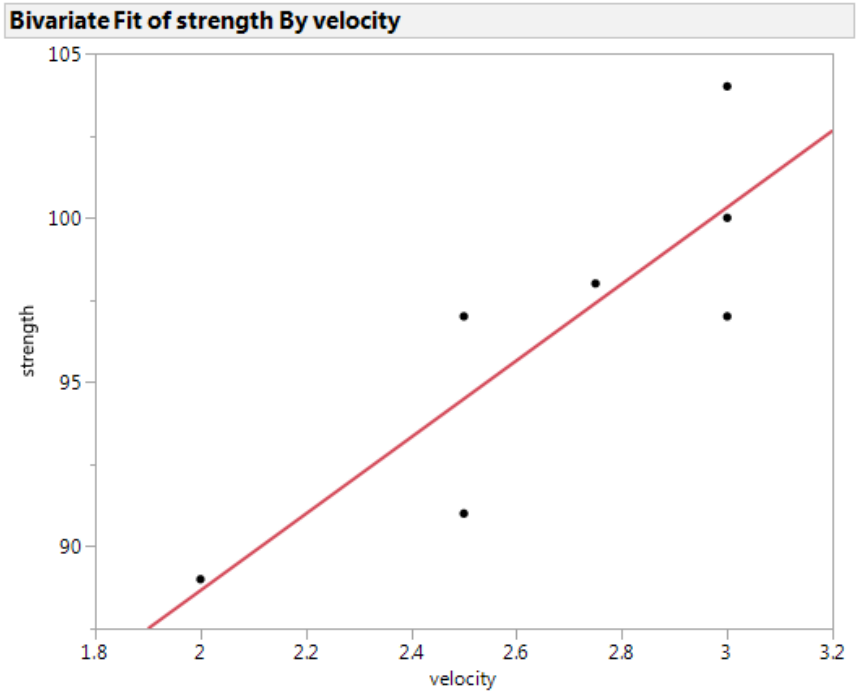
Menu->Analyze-> Fit Y by X
Drag the Strength Column to the "Y,Response" box
Drag the Velocity Column to the "X,Factor" box
Click "OK"

	velocity	strength
1	2	89
2	2.5	97
3	2.5	91
4	2.75	98
5	3	100
6	3	104
7	3	97



Fitting a Regression Line

In the previous window
Click the red triangle -> Fit Line



Linear Fit

strength = 65.340426 + 11.659574*velocity

Summary of Fit

RSquare	0.723443
RSquare Adj	0.668131
Root Mean Square Error	2.953541
Mean of Response	96.57143
Observations (or Sum Wgts)	7

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Ratio
Model	1	114.09726	114.097	13.0794
Error	5	43.61702	8.723	Prob > F
C. Total	6	157.71429		0.0153*

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	65.340426	8.707432	7.50	0.0007*
velocity	11.659574	3.223948	3.62	0.0153*

ANOVA
testing for differences in k means

- ANOVA is used to answer the question "Are the means of these k groups the same or not?"
- The null hypothesis is $\mu_1 = \mu_2 = \ \ldots = \mu_k$
- The alternative hypothesis is that at least one mean is different
- The standard assumptions are independence, normality of the samples within each group.
- Equality of variance is also assumed but be reminded of the robustness work of Box in the discussion earlier

Conceptual Model

- k groups of observations
- n_i observations in group i
- y_{ij} is the j^{th} observation in group i
- The model is:

$$y_{ij} = \eta + \tau_i + \varepsilon_{ij} \qquad i = 1,k \quad j = 1, n_i \quad \varepsilon_{ij} \sim N(0, \sigma^2)$$

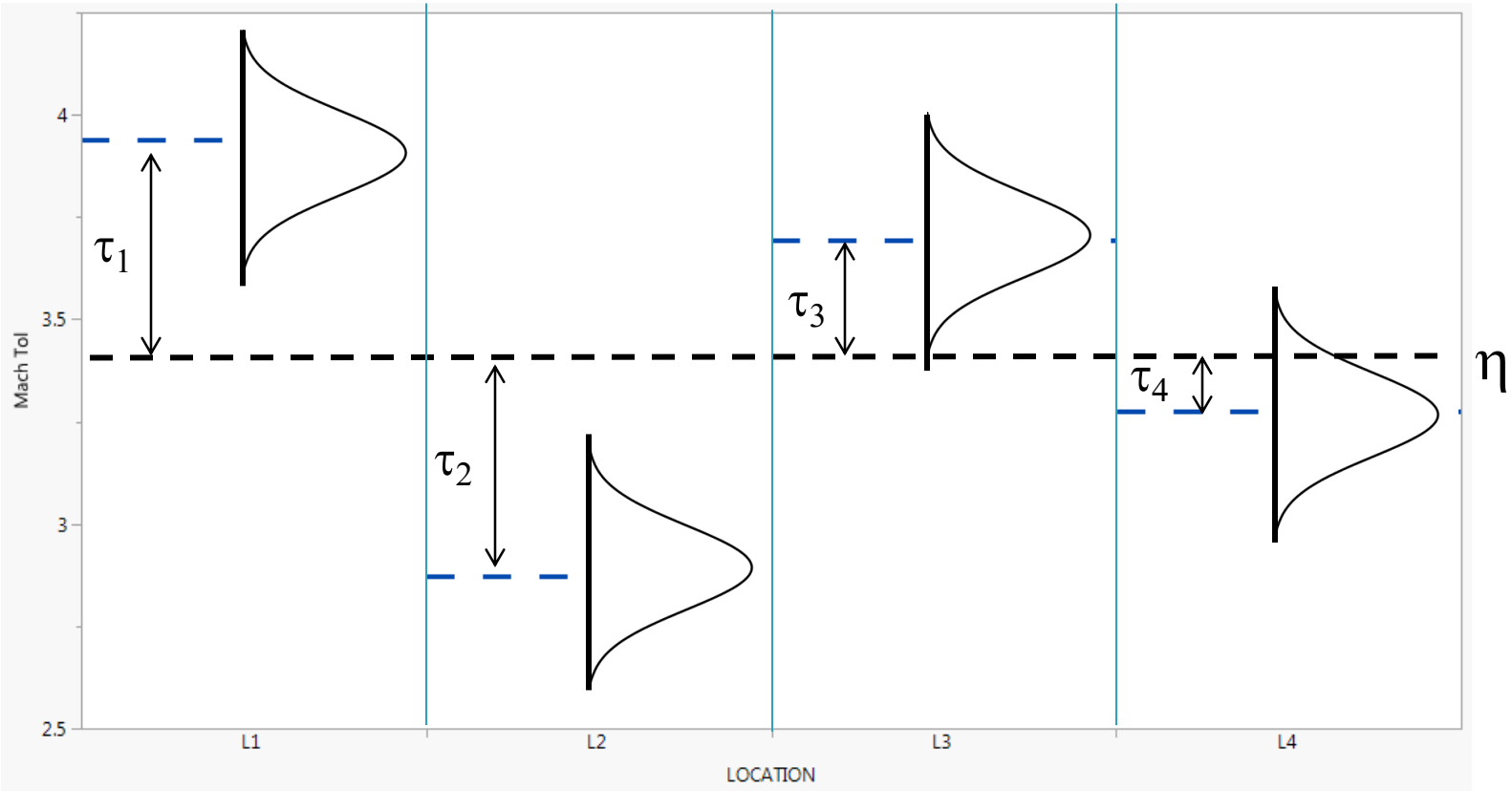
overall mean — η

effect of group i — τ_i

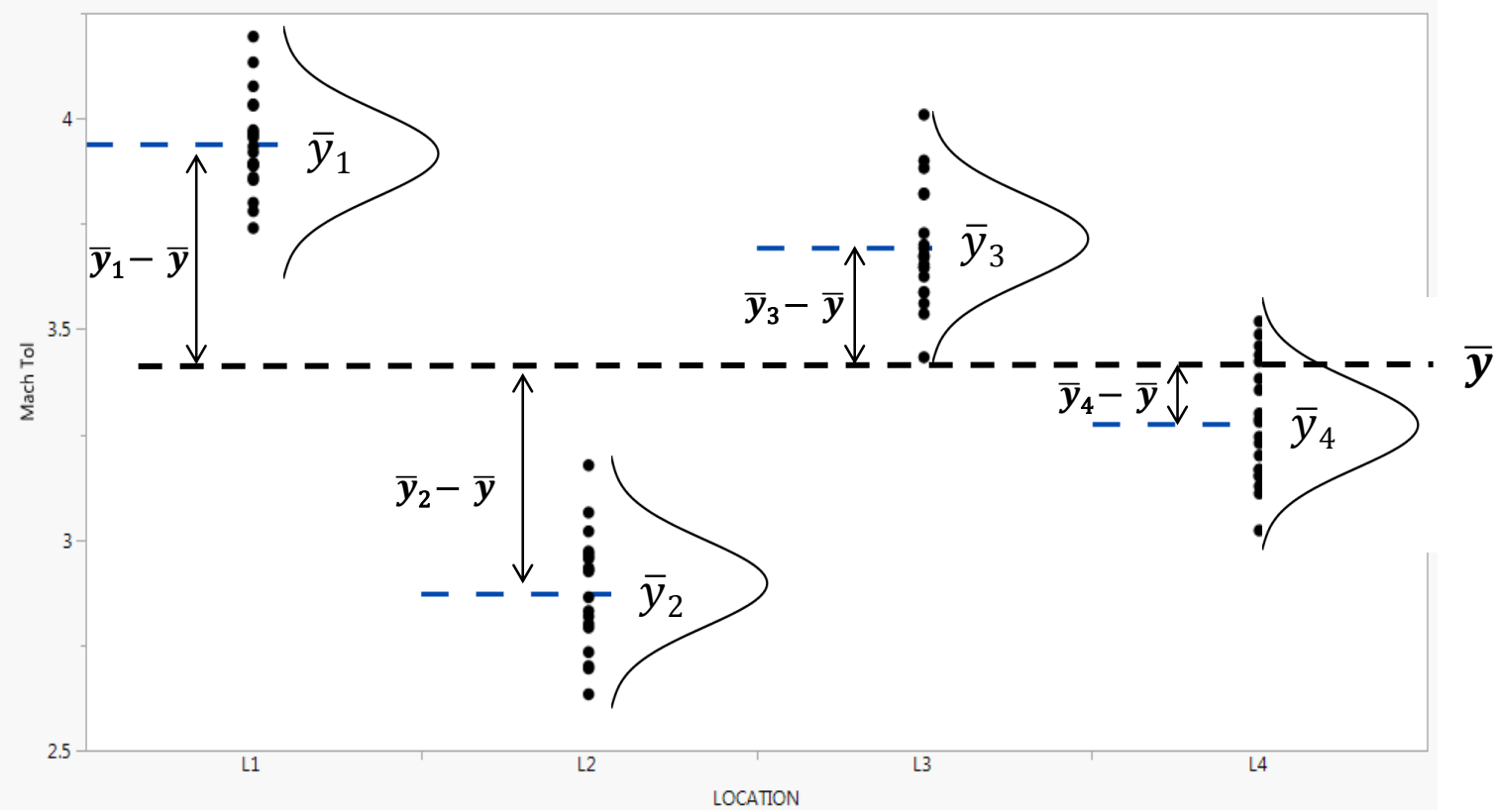
random error — ε_{ij}

Graphical Representation of the Conceptual Model Machining Tolerance Example

- A part is machined at 4 locations and a tolerance is measured
- Question: Is the mean tolerance the same for each location?



Realization of the Conceptual Model based on Gathered Machining Tolerance Data



ANOVA TECHNIQUE

- The total "Sum of Squares" (SST) of the data can be partitioned into components
  - The variability, or Sum of Square, associated with the "model" … **SSM**
  - And, the variability due to error … **SSE**
- SST = SSM + SSE
- Here's how it's done:

ANOVA TECHNIQUE

- $\text{SST} = \sum_{i=1}^{k}\sum_{j=1}^{n_i}(y_{ij} - \bar{y})^2$

- By rewriting $(y_{ij} - \bar{y})^2$ as $((y_{ij} - \bar{y}_i) + (\bar{y}_i - \bar{y}))^2$, completing the square and realizing the cross products sum to zero

$$\sum_{i=1}^{k}\sum_{j=1}^{n_i}(y_{ij} - \bar{y})^2 = \sum_{i=1}^{k}\sum_{j=1}^{n_i}(\bar{y}_i - \bar{y})^2 + \sum_{i=1}^{k}\sum_{j=1}^{n_i}(y_{ij} - \bar{y}_i)^2$$

- $\text{SST} = \text{SSM} + \text{SSE}$

ANOVA TABLE

SOURCE	df	SUM OF SQUARES	MEAN SQUARE	EXPECTED MEAN SQUARE	F - TEST
MODEL	K-1	SSM = $\sum_{i=1}^{k}\sum_{j=1}^{n_i}(\bar{y}_i - \bar{y})^2$	S_M^2 = SSM/K-1	$\sum_{i=1}^{k} n_i\tau_i^2/(k-1) + \sigma^2$	$F = \frac{S_M^2}{S_E^2}$
ERROR	N-K	SSE= $\sum_{i=1}^{k}\sum_{j=1}^{n_i}(y_{ij} - \bar{y}_i)^2$	S_E^2 = SSE/N-K	σ^2	
TOTAL	N-1	SST= $\sum_{i=1}^{k}\sum_{j=1}^{n_i}(y_{ij} - \bar{y})^2$			

The hypothesis that $\mu_1 = \mu_2 = \ldots = \mu_k$, implies that all $\tau_i^2 = 0$

and the F -ratio will follow an F distribution with (k−1) and (N−k) df

ANOVA TABLE

- The hypothesis that $\mu_1 = \mu_2 = \ldots = \mu_k$, implies that all $\tau_i^2 = 0$
- If $\tau_i^2 = 0$ for all i, then F follows an F distribution with (k-1) and (N-k) df
- If one or more of the τ_i^2 do not equal 0, then the expected value of the F ratio will be greater than 1 and large values are evidence that the Null Hypothesis should be rejected
- The Null Hypothesis is rejected if the p-value associated with the F ratio is less than α (where α is typically .05)

Eg. : ANOVA of Machining Tolerance Data

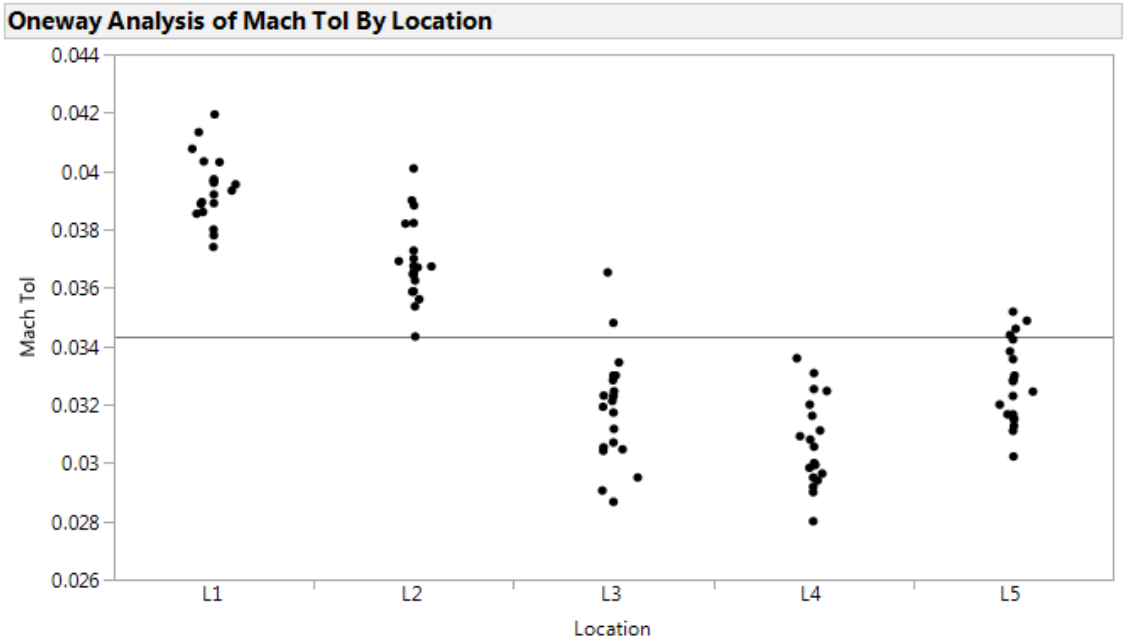
- The file **MachineTol.xlsx** contains machining tolerance data measured on a part at 5 locations L1–L5. The tolerance measurement is provided in the variable MachTol. Conduct an analysis of variance of MachTol to determine if the population means associated with locations L1–L5 are the same or are different.

  a. Plot MachTol by Location.

  b. Show the results of an ANOVA and provide the conclusion to the hypothesis test.

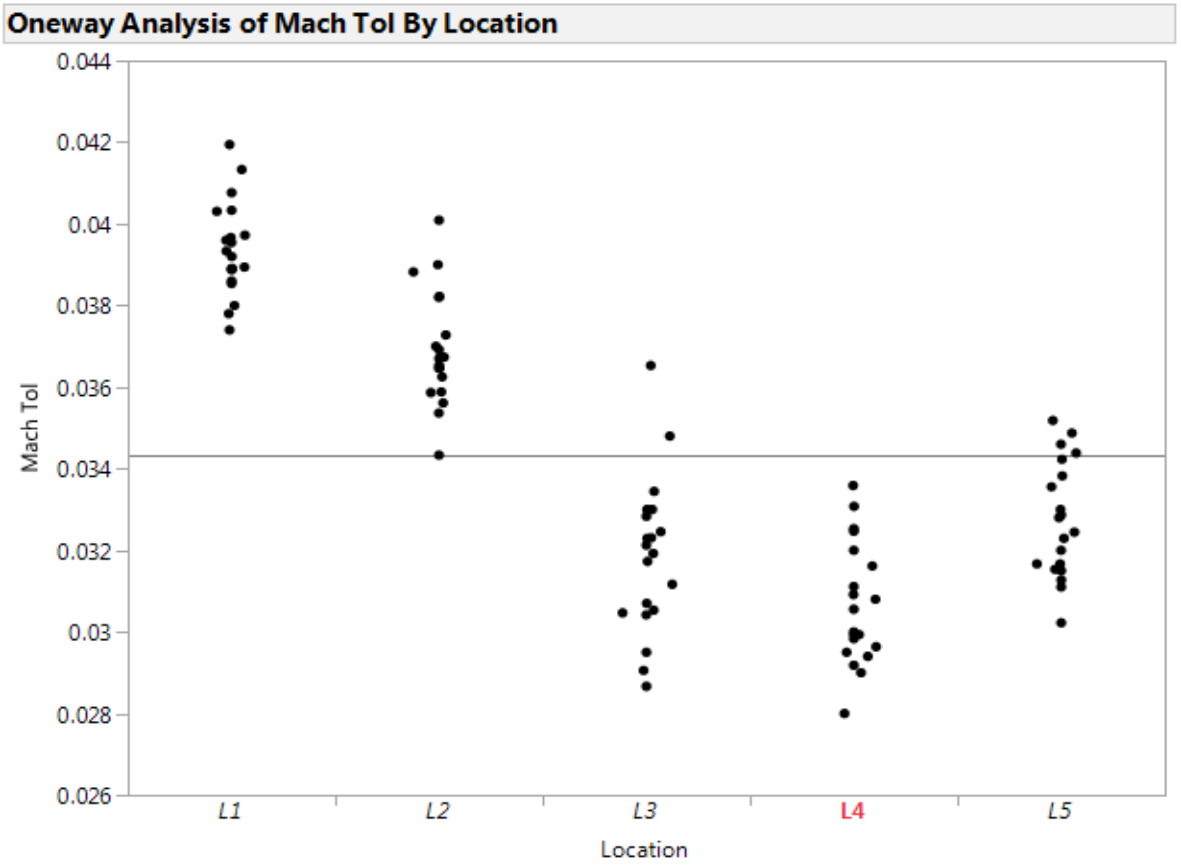
a. Plot machine tolerance(MachTol) by Location

Menu-> Analyze-> Fit Y by X
Drag Column MachTol to the “Y,Response” bo
Drag Column Location to the “X,Factor” box
Click OK



b. ANOVA of Machining Tolerance Data

In the previous window …
Click on the red triangle-> Means/ANOVA



Oneway Anova

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Ratio	Prob > F
Location	4	0.00108920	0.000272	123.4103	<.0001*
Error	95	0.00020961	2.206e-6		
C. Total	99	0.00129881			

Multiple Comparisons Methods to Compare Means

- ANOVA determines if the means are different … or not
- ANOVA *doesn't* tell you the nature of the differences
- Care must be taken when doing multiple, pairwise comparisons (tests) because for a fixed error rate on any test, the combined error rate over all comparisons can be much higher!
  - Consider 5 groups being compared G1, G2, G3, G4, G5
  - There are 10 pairwise comparisons
  - If each test has a Type I error rate of .05 then the chances of getting no Type I errors is $(.95)^{10} = .599$
  - So, the probability of committing at least 1 Type one error is around .4
  - The more comparisons done, the more likely that a Type I error is made
- We will consider 3 multiple comparison options
  - 1. **Each pair** (students t) 2. **All pairs** (Tukey-Kramer HSD) 3. **With best** (Hsu MCD)

Each Pair, Student's t

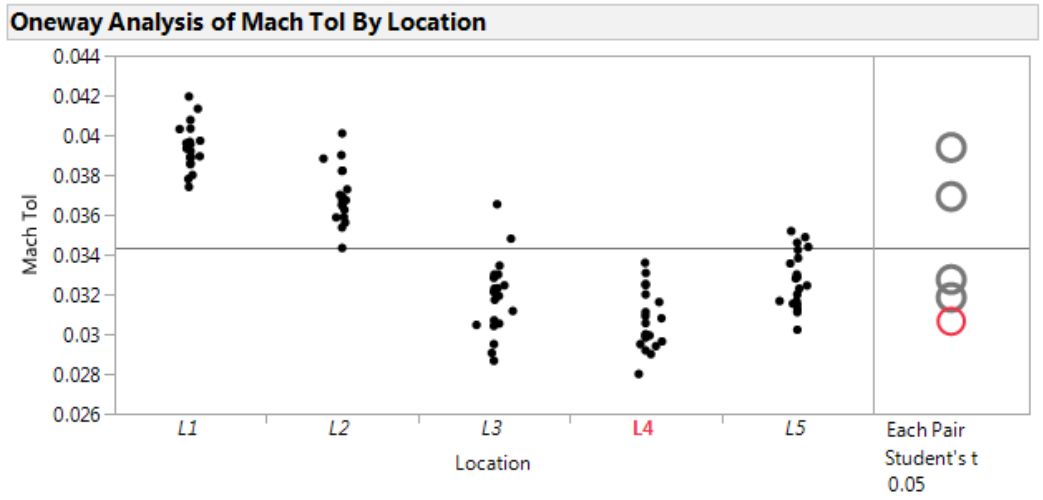
- The "**Each Pair" , Student's t** option just applies a t-test to every pair.
- No adjustment to reduce Type I error is made
- Using this can lead to a high Type One error rate

Each Pair, Student's t

Machine Tolerance Data

Each Pair, t-test, .05

To conduct the "Each Pair" multiple comparison:
In the ANOVA report window …
click the red triangle->compare means->"each pair-students t



Means Comparisons

Comparisons for each pair using Student's t

Confidence Quantile

t	Alpha
1.98525	0.05

LSD Threshold Matrix

Abs(Dif)-LSD

	L1	L2	L5	L3	L4
L1	-0.00093	0.00153	0.00570	0.00661	0.00780
L2	0.00153	-0.00093	0.00324	0.00414	0.00534
L5	0.00570	0.00324	-0.00093	-0.00003	0.00116
L3	0.00661	0.00414	-0.00003	-0.00093	0.00026
L4	0.00780	0.00534	0.00116	0.00026	-0.00093

Positive values show pairs of means that are significantly different.

Connecting Letters Report

Level		Mean
L1	A	0.03938575
L2	B	0.03692415
L5	C	0.03275091
L3	C	0.03184722
L4	D	0.03065391

Levels not connected by same letter are significantly different.

All pairs (Tukey-Kramer HSD)

- Tukey-Kramer is similar to “Each Pair” but the alpha level of each comparison is adjusted to be smaller
- The adjustment made to alpha yields an “experiment-wise” Type I error of .05 (versus the “pair-wise” .05 error).
- The trade off is that for many levels, the alpha level for any one comparison is more conservative.
- With a lower alpha, pairwise differences are less apparent.

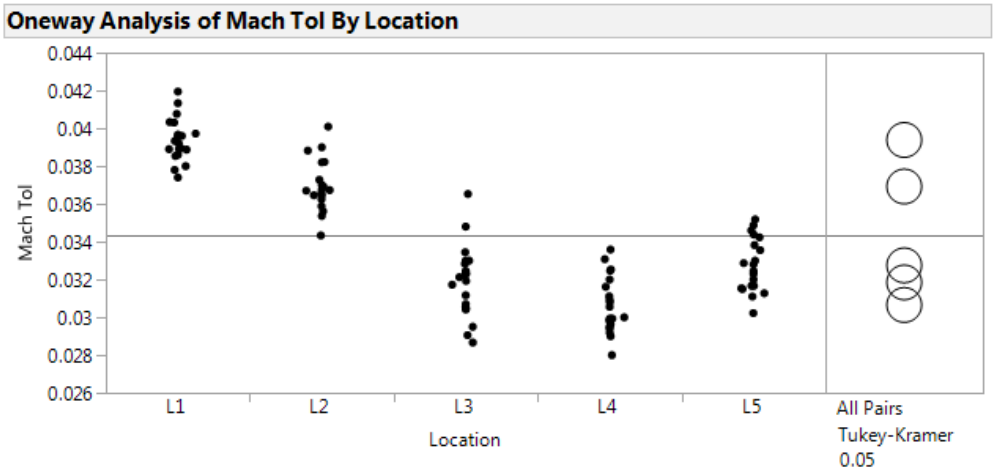
All pairs (Tukey–Kramer HSD)

Machine Tolerance Data

All Pairs

Tukey/Kramer HSD

To conduct the "All Pairs– Tukey–Kramer HSD" multiple comparison:
In the ANOVA report window …
click the red triangle–>compare means–>"all pairs– Tukey HSD"



Means Comparisons

Comparisons for all pairs using Tukey-Kramer HSD

Confidence Quantile

q*	Alpha
2.78086	0.05

HSD Threshold Matrix

Abs(Dif)-HSD

	L1	L2	L5	L3	L4
L1	-0.00131	0.00116	0.00533	0.00623	0.00743
L2	0.00116	-0.00131	0.00287	0.00377	0.00496
L5	0.00533	0.00287	-0.00131	-0.00040	0.00079
L3	0.00623	0.00377	-0.00040	-0.00131	-0.00011
L4	0.00743	0.00496	0.00079	-0.00011	-0.00131

Positive values show pairs of means that are significantly different.

Connecting Letters Report

Level					Mean
L1	A				0.03938575
L2		B			0.03692415
L5			C		0.03275091
L3			C	D	0.03184722
L4				D	0.03065391

Levels not connected by same letter are significantly different.

Comparison With "Best" (Hsu MCD)

- Compare With Best adjusts alpha but under the assumption that only two comparisons will be made … comparing the two largest or comparing the two smallest
  - So, either "Bigger is Better" or "Smaller is Better"
  - It's a smaller alpha adjustment than the Tukey-Kramer so it's more sensitive to differences

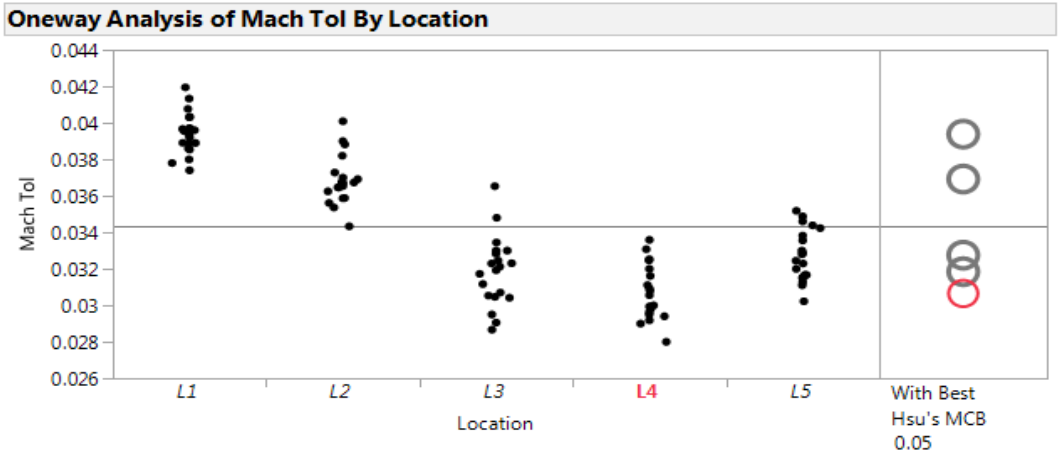
With Best (Hsu MCD)

Machine Tolerance Data

Compare With Best

Hsu MCD

To conduct the "With Best-Hsu MCD" multiple comparison:
In the ANOVA report window …
click the red triangle->compare means->"With Best-Hsu MCD"



Means Comparisons

Comparisons with the best using Hsu's MCB

Comparisons with Max and Min

Level	with Max p-Value	with Min p-Value
L1	1.0000	<.0001*
L2	<.0001*	<.0001*
L5	<.0001*	<.0001*
L3	<.0001*	0.0217*
L4	<.0001*	0.9999

LSD Threshold Matrix

Mean[i]-Mean[j]-LSD

	L1	L2	L5	L3	L4
L1	-0.00103	0.00143	0.00561	0.00651	0.00770
L2	-0.00349	-0.00103	0.00314	0.00405	0.00524
L5	-0.00766	-0.00520	-0.00103	-0.00012	0.00107
L3	-0.00857	-0.00611	-0.00193	-0.00103	0.00016
L4	-0.00976	-0.00730	-0.00313	-0.00222	-0.00103

If a column has any positive values, the mean is significantly less than the max.

Mean[i]-Mean[j]+LSD

	L1	L2	L5	L3	L4
L1	0.00103	0.00349	0.00766	0.00857	0.00976
L2	-0.00143	0.00103	0.00520	0.00611	0.00730
L5	-0.00561	-0.00314	0.00103	0.00193	0.00313
L3	-0.00651	-0.00405	0.00012	0.00103	0.00222
L4	-0.00770	-0.00524	-0.00107	-0.00016	0.00103

If a column has any negative values, the mean is significantly greater than the min.

Because of its increased Sensitivity, this method concludes that L4 produced a superior tolerance result over L3

This could not be concluded using the Tukey HSD

Class Lab #3: ANOVA to determine if lubricant formulations have a significant effect on extrusion friction.

- **Problem description: Low friction can lead to slippage in an extrusion process that produces surface damage. A series of 5 tests were developed to determine if specific lubricant formulations led to a significant change in friction. The data are provided in the data set "Friction.xlsx".**

  a. **Plot the friction results over tests to see if there is any visual evidence of differences in mean friction from one test to another**

  b. **Is there a particular test that appears visually superior?**

  c. **Conduct an analysis of variance on the data to determine if there is a significant difference in mean frictions**

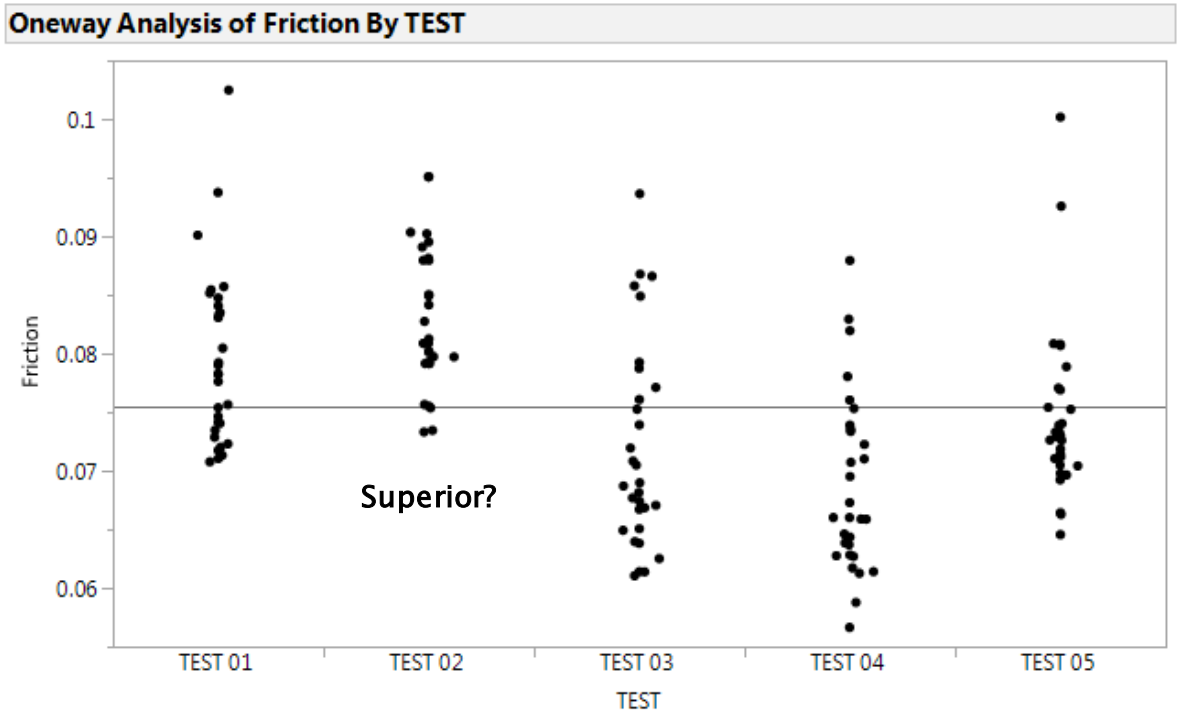
  d. **Conduct a multiple comparison test to determine if any of the tests produced a statistically superior result (i.e. significantly higher mean friction)**

  e. **Would the Tukey-Kramer method have given a different result?**

Class Lab #3: ANOVA to determine if lubricant formulations have a significant effect on extrusion friction.

a. Plot the friction results over tests to see if there is any visual evidence of differences in mean friction from one test to another

To plot Friction over TESTs:
Menu: Analyze-> Fit Y by X
Drag the Friction Column to the "Y,Response" box
Drag the TEST Column to the "X,Grouping" box
Click "OK"

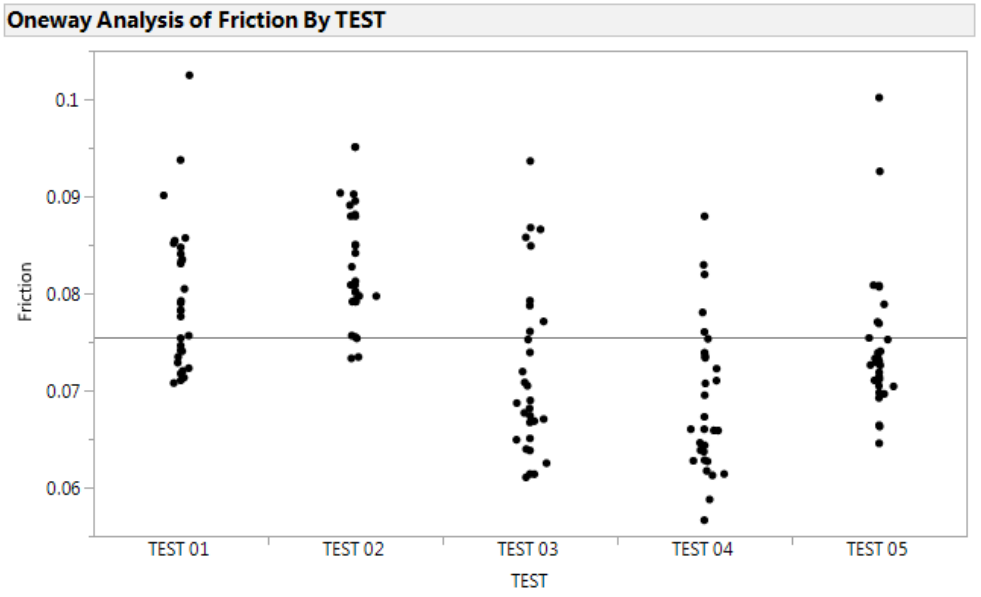


Class Lab #3 Solutions

C. Conduct an analysis of variance on the data to determine if there is a significant difference in mean frictions

To conduct the ANOVA:
In the Fit Y by X report window:
Click the red triangle->Means/Anova

Su



Oneway Anova

Analysis of Variance

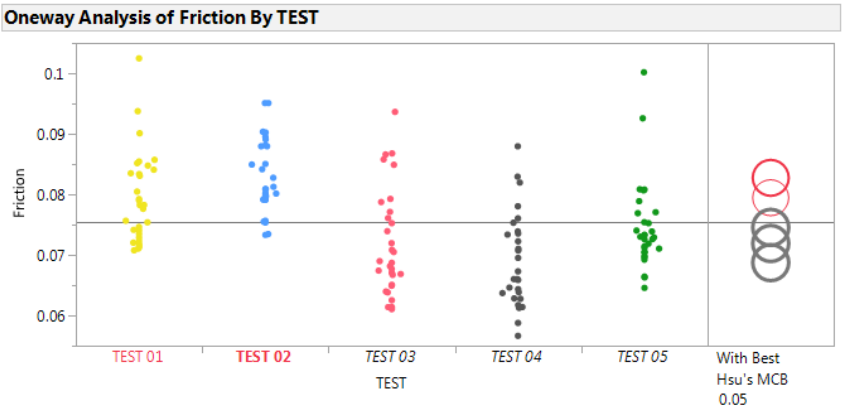
Source	DF	Sum of Squares	Mean Square	F Ratio	Prob > F
TEST	4	0.00382856	0.000957	17.1848	<.0001*
Error	145	0.00807606	0.000056		
C. Total	149	0.01190462			

The ANOVA indicates a significant difference in mean friction

Class Lab #3 Solutions

d. Test for superior/inferior results use the **Hsu MCD test**

To conduct the Hsu multiple comparison:
In the ANOVA report window …
Click on the red triangle->Compare Means->With Best,Hsu MCD



Oneway Anova

Means Comparisons

Comparisons with the best using Hsu's MCB

Comparisons with Max and Min

Level	with Max p-Value	with Min p-Value
TEST 02	0.9978	<.0001*
TEST 01	0.1314	<.0001*
TEST 05	<.0001*	0.0055*
TEST 03	<.0001*	0.1476
TEST 04	<.0001*	0.9972

LSD Threshold Matrix

Mean[i]-Mean[j]-LSD

	TEST 02	TEST 01	TEST 05	TEST 03	TEST 04
TEST 02	-0.00420	-0.00091	0.00399	0.00664	0.00980
TEST 01	-0.00748	-0.00420	0.00070	0.00336	0.00652
TEST 05	-0.01239	-0.00910	-0.00420	-0.00154	0.00161
TEST 03	-0.01504	-0.01175	-0.00685	-0.00420	-0.00104
TEST 04	-0.01820	-0.01491	-0.01001	-0.00736	-0.00420

If a column has any positive values, the mean is significantly less than the max.

Mean[i]-Mean[j]+LSD

	TEST 02	TEST 01	TEST 05	TEST 03	TEST 04
TEST 02	0.00420	0.00748	0.01239	0.01504	0.01820
TEST 01	0.00091	0.00420	0.00910	0.01175	0.01491
TEST 05	-0.00399	-0.00070	0.00420	0.00685	0.01001
TEST 03	-0.00664	-0.00336	0.00154	0.00420	0.00736
TEST 04	-0.00980	-0.00652	-0.00161	0.00104	0.00420

If a column has any negative values, the mean is significantly greater than the min.

APPENDIX

Additional hypothesis test example … σ^2 unknown

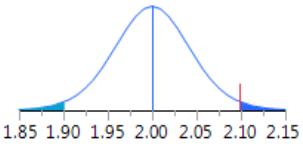
- % elongation of metal samples is measured to check to see if the mean elongation has increased from its historical 2% level. σ^2 is not known
- Conduct a hypothesis test of $H_0: \mu \leq \mu_0$ versus $H_a: \mu > \mu_0$

- From the data we calculate:
  - $\bar{X} = 2.099$
  - $s = .1686$
  - n= 16 (so df = 15)
  - If $\alpha = .05$ then $t_\alpha(\nu) = 1.75$

  - So, $t = \frac{\bar{X}-\mu_0}{s/\sqrt{n}} = \frac{2.099-2}{.1686/4} = 2.348$
  - Since $t > t_{.05}(15)$, we reject $H_0: \mu \leq \mu_0$
  - JMP provides a p-value of .0167

Test Mean	
Hypothesized Value	2
Actual Estimate	2.09875
DF	15
Std Dev	0.16856

	t Test
Test Statistic	2.3434
Prob > |t|	0.0333*
Prob > t	0.0167*
Prob < t	0.9833



	Elongation
1	2.42
2	2.09
3	2.17
4	1.91
5	2.5
6	1.99
7	2.19
8	2.08
9	1.99
10	2.11
11	1.97
12	2.05
13	2.22
14	1.96
15	1.97
16	1.96

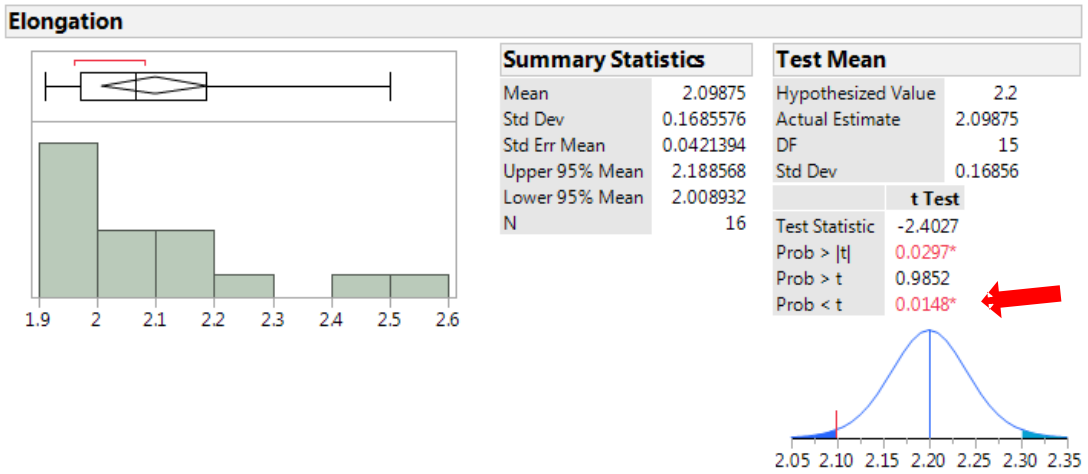
Additional hypothesis test example … σ^2 unknown

- % elongation of metal samples is measured to check to see if the mean elongation has decreased from its historical **2.2%** level. σ^2 is not known
- Conduct a hypothesis test of $H_0: \mu \geq \mu_0$ **vs** $H_a: \mu < \mu_0$

- Manually, $t = \frac{\bar{x}-\mu_0}{s/\sqrt{n}} = \frac{2.099-2.2}{.1686/4} = -2.4$

- Since $t < -t_{.05}(15) = -1.75$ we reject $H_0: \mu \geq \mu_0$

- The easier solution:



PhaseTwo Analytics

Determining the sample size needed to achieve desired β (with additional math)

- $\beta = P(\text{Accept } H_0 \mid \mu = \mu_1)$
- $= P\left(\frac{\bar{X}-\mu_0}{\sigma/\sqrt{n}} < z_\alpha \mid \mu = \mu_1\right)$
- $= P\left(\bar{X} < \mu_0 + z_\alpha \sigma/\sqrt{n} \mid \mu = \mu_1\right)$
- $= P\left(\frac{\bar{X}-\mu_1}{\sigma/\sqrt{n}} < \frac{(\mu_0-\mu_1)+z_\alpha \sigma/\sqrt{n}}{\sigma/\sqrt{n}} \mid \mu = \mu_1\right)$
- $= P\left(Z < \frac{(\mu_0-\mu_1)}{\sigma/\sqrt{n}} + z_\alpha\right) = \beta$ **if**
- $\frac{(\mu_0-\mu_1)}{\sigma/\sqrt{n}} + z_\alpha = -z_\beta$ solving for n
- $$n = \left[\frac{(z_\alpha + z_\beta)\sigma}{(\mu_1 - \mu_0)}\right]^2$$

Testing Regression Coefficients (with additional math)

- The least squares solution is:
  - $b = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^{n}(X_i - \bar{X})^2} = \left(\frac{1}{\sum_{i=1}^{n}(X_i - \bar{X})^2}\right)\left(\sum_{i=1}^{n}(X_i - \bar{X})(Y_i) - \bar{Y}\sum_{i=1}^{n}(X_i - \bar{X})\right)$ where $\sum_{i=1}^{n}(X_i - \bar{X}) = 0$
  - $= \left(\frac{1}{\sum_{i=1}^{n}(X_i - \bar{X})^2}\right)\left(\sum_{i=1}^{n}(X_i - \bar{X})(Y_i)\right)$ which is a weighted sum of Yi and
  - $b0 = \bar{Y} - b\,\bar{X}$
  - By rewriting the 1st eqn., b can be seen to be a linear combination sums of the y_i … so the CENTRAL LIMIT THEOREM APPLIES and b will be approx Normally distributed with increasing n!!!
  - What is the E [b] ??? What is V[b] ??? What does this do for me???
  - Recall $t = \frac{\text{statistic} - \text{mean(statistic)}}{\text{s.e.(statistic)}}$ would follow a t distribution

PhaseTwo Analytics

Testing Regression Coefficients
Deriving E[b]

- $E[b] = \left(\frac{1}{\sum_{i=1}^{n}(X_i - \bar{X})^2}\right)\left(\sum_{i=1}^{n}(X_i - \bar{X})\,E[Y_i]\right) =$

$$= \left(\frac{1}{\sum_{i=1}^{n}(X_i - \bar{X})^2}\right)\sum_{i=1}^{n}(X_i - \bar{X})(\beta_0 + \beta Xi)$$

$$= \left(\frac{1}{\sum_{i=1}^{n}(X_i - \bar{X})^2}\right)\sum_{i=1}^{n}(X_i - \bar{X})(\beta_0 + \beta Xi - \beta\bar{X} + \beta\bar{X})$$

$$= \left(\frac{1}{\sum_{i=1}^{n}(X_i - \bar{X})^2}\right)[\beta_0 \cancelto{0}{\sum_{i=1}^{n}(X_i - \bar{X})} + \beta\sum_{i=1}^{n}(X_i - \bar{X})^2 + \beta\bar{X}\cancelto{0}{\sum_{i=1}^{n}(X_i - \bar{X})}]$$

$$= \left(\frac{1}{\sum_{i=1}^{n}(X_i - \bar{X})^2}\right)\beta\sum_{i=1}^{n}(X_i - \bar{X})^2 = \beta$$

PhaseTwo Analytics

Testing Regression Coefficients – Testing Intercept $\beta_0 = 0$?

- Recall that the least squares estimate of β_0 is: $b_0 = \bar{Y} - b\,\bar{X}$
- Since b follows a normal dist and b_0 just adds $\bar{Y}$ which also follows a Normal dist, then it follows that b_0 follows a Normal dist (approx)
- $E[b_0] = E[\bar{Y}] - \bar{X}\,E[b] = E[\bar{Y}] - \beta\bar{X}$
- $E[y_i] = \beta_0 + \beta x_i + 0$ (since $y_i = \beta_0 + \beta x_i + \varepsilon_i$) and $\varepsilon_i \sim N(0, \sigma^2)$
- Thus, $E[\bar{Y}] = \beta_0 + \beta\,\bar{X}$ and it follows then that
- $E[b_0] = \beta_0 + \beta\,\bar{X} - \beta\bar{X} = \beta_0$ (so b_0 is an unbiased estimator of β_0)
- Next, lets find $V[b_0]$

Testing Regression Coefficients – $\beta_0 = 0$?

- $y_i = \beta_0 + \beta\, x_i + \varepsilon_i$ and $\varepsilon_i \sim N(0, \sigma^2)$
- So $V[yi] = \sigma^2$ and therefore $V[\bar{Y}] = \sigma^2/n$
- So, $V[b_0] = \sigma^2/n + \bar{X}^2 v[b]$
- $$= \sigma^2/n + \bar{X}^2 \frac{\sigma^2}{\sum_{i=1}^{n}(X_i - \bar{X})^2}$$
- $$= \sigma^2 \left[\frac{1}{n} + \frac{\bar{X}^2}{\sum_{i=1}^{n}(X_i - \bar{X})^2}\right]$$
- so, $s.e.[b_0] = s \sqrt{\frac{\bar{X}^2}{\sum_{i=1}^{n}(X_i - \bar{X})^2}}$
- Meaning $t = \frac{b_0 - \beta_0}{s/\sqrt{\frac{\bar{X}^2}{\sum_{i=1}^{n}(X_i - \bar{X})^2}}}$ provides a t test of the hypothesis that $\beta_0 = 0$