

# MBC 638

---

LIVE SESSION WEEK 10

# Agenda

---

Topic
Introduction
Final Review: Video 10 and Beyond
Process Improvement Project Rubric Review
Review of Upcoming Assignments and Open Question

## 10.2 Summary of DMAIC

**Define** ← Most important tool in define is the problem definition worksheet

- Categorized data (discrete and continuous)
- Descriptive statistics, measuring central tendency (mean, median, and mode) ←
- Measure of dispersion (range, standard deviation, and variance) ←
  - $\text{Variance} = (\text{standard deviation})^2$  ←
- Tools for displaying/presenting data
  - Charts and graphs (pie, line, trend, bar, etc.)
- Pareto principle (80-20 rule)
- Frequency distribution and histograms

### Define (cont.)

← A good way to start

- Process maps and thought-process maps
- SIPOC (Supply, Input, Process, Output, Customer)
- Affinity diagram (brainstorming)
- Fishbone diagram
- Calculate SQL (sigma quality level)
- Problem definition worksheet ←
  - ROI: How much is your problem worth?

## 10.2 Summary of DMAIC

### Measure

- Quantitative and qualitative measurements
- Importance of operational definition
- Minimizing variation in measurement system
- Calculated kappa ( $k$  value) to evaluate discrete measurement system
- Sampling distribution of sample mean becomes normal as sample gets larger
- Area under normal curve = 1
- $z = \frac{x - \mu}{\sigma}$  ←
- Sample size driven by level of confidence, margin of error, and standard deviation ←

- Different formulas for continuous and discrete data

We also completed data measurement plan and data Stratification tree in measure.

Histograms are also a useful tool in measure.

# 10.2 Summary of DMAIC

## Analyze

- Inferential statistics: drawing conclusion on population based on sample
- Confidence interval and hypothesis testing
  - Confidence interval = range of values
  - Write  $H_0$  and  $H_a$
  - $H_a$ : wants, concerns
  - Equality statement goes in null
  - "If  $p$  is low then  $H_0$  must go"
- Chi-square test for independence

A hypothesis test  
On discrete data  
Tells you if there is a relationship  
Does not tell of strength or direction of relationship

## 10.2 Summary of DMAIC

- Simple linear regression: only one input; continuous data
- Multiple linear regression: multiple input variables; can include categorical/indicator variables
  - Use inputs with low  $p$ -value
- Correlation measures strength of linear relationship between  $x$  and  $y$
- $R^2$  = coefficient of determination
  - Measures variability in  $y$  that is accounted for by including particular  $x$  in model
- Practical, graphical, analytical
- Time series analysis
  - Models: first-order autoregressive, moving average, exponential smoothing

} Regression can show strength and direction of relationship between input and output

## 10.2 Summary of DMAIC

### Improve

- Make sure analysis has led to root cause
- Streamline customer and business value-added steps
  - Eliminate non-value-added steps
- Pilot solution: do an experiment

### Control

- Monitor critical inputs ( $x$ 's)
- Use control chart to understand signal vs. noise
  - Only react to signals
- Use appropriate control chart for type of data
  - $\bar{X}$ -R,  $\bar{X}$ -s, IMR, c, u, p, np
- Continual process improvement a journey, not a destination

# 10.3 Questions You Should Be Able to Answer

## Define

- What does *DMAIC* stand for?
- What is the difference/relationship between standard deviation and variance? How are they related?
- Give me an example of a measure of location, a.k.a. measure of central tendency.
- Give me an example of measure of dispersion.
- What is the difference between discrete data and continuous data?
- Fishing line sold per year: Is that continuous or discrete data?
- Name two things you can learn from plotted data.

Define, Measure, Analyze, Improve, Control

Variance = standard deviation squared  
Std dev=3, Variance=9

Mean, Mode, Median

Range, Standard deviation, Variance

## Measure

- How could you visually display variation?
- Sample size is a function of what three things?
- In order to determine SQL (sigma quality level) for your process, what do you need to determine first?
- What is the difference between repeatability and reproducibility?
- If you want to increase your level of confidence, what do you need to do to your sample size?

Sample Size Formula for Continuous Data

$$n = \left( \frac{z * \hat{\sigma}}{E} \right)^2$$



# 10.3 Questions You Should Be Able to Answer

## Analyze

- "If  $p$  is low,  $H_0$  must go." Lower than what?
- What data would be considered inappropriate for a regression model?
- What does variation do to cycle time?
- What is the difference between  $R$  and  $R^2$ ?
- What is a type 1 error?
- What is a confidence interval?
- When would you calculate the  $t$ -test statistic vs. the  $Z$  test statistic?
- What is a residual?

A confidence interval is an estimate of a parameter consisting of an interval of numbers based on a point estimate, together with a confidence level specifying the probability that the interval contains the parameter.

If you reject a null hypothesis that  $\mu = 8$  at an alpha of .01, does your 99% confidence interval include 8?

Alpha, If you have an alpha of .10 at what  $p$  would you reject???? Anything lower (.09, .08, etc.)

$R$  = The correlation coefficient is a measure for quantifying the linear relationship between two quantitative variables.

$R^2$  = The coefficient of determination is a measure of the variability in  $Y$  that can be accounted for by  $X$ , measure the goodness of fit of the regression equation to the data.

		The actual state of things (what actually happened)	
		$H_0$ is true	$H_0$ is false
Fail to reject $H_0$		Correct conclusion	Type II error, beta risk, or consumer's risk
Reject $H_0$		Type I error, alpha risk, or producer's risk	Correct conclusion

Rejecting  $H_0$  when it is true

## 10.3 Questions You Should Be Able to Answer

### Analyze (cont.)

- The correlation coefficient can take on any value in what range? **-1 to 1, remember this is r**
- If your R value is equal to zero, what does that mean?
- Name three models that can aid in the analysis of time series data.
- What is it called when you have correlation between successive values of a time series? **Autocorrelation**
- When the variability in your  $y$  increases, the correlation coefficient gets closer to what number? **0**
- What if the seasons contribute to the variation in your time series data?
  - What might you do to account for that in your predictive model?

Don't forget to look at your scatter plots

# 10.3 Questions You Should Be Able to Answer

## Improve

- List two ways that regression can be useful.
- What does a Pareto show you?
- How can you tell if a particular input variable is significant enough to include in your regression equation?

P is LOW

No patterns in the data points or points outside the limits

If all the points are within the limits it tells you the process variation is stable and predictable.

## Control

- Name two ways you can tell if your process is in control.
- What can a range chart tell you?
- What kind of control chart would be most appropriate to use when you are measuring data from a service center, counting the lost calls per day?
- What type of control chart is appropriate for continuous data?
- When the normal functioning of a process is disturbed by some unpredictable event, what kind of variation is added to the common cause variation found in a control chart?

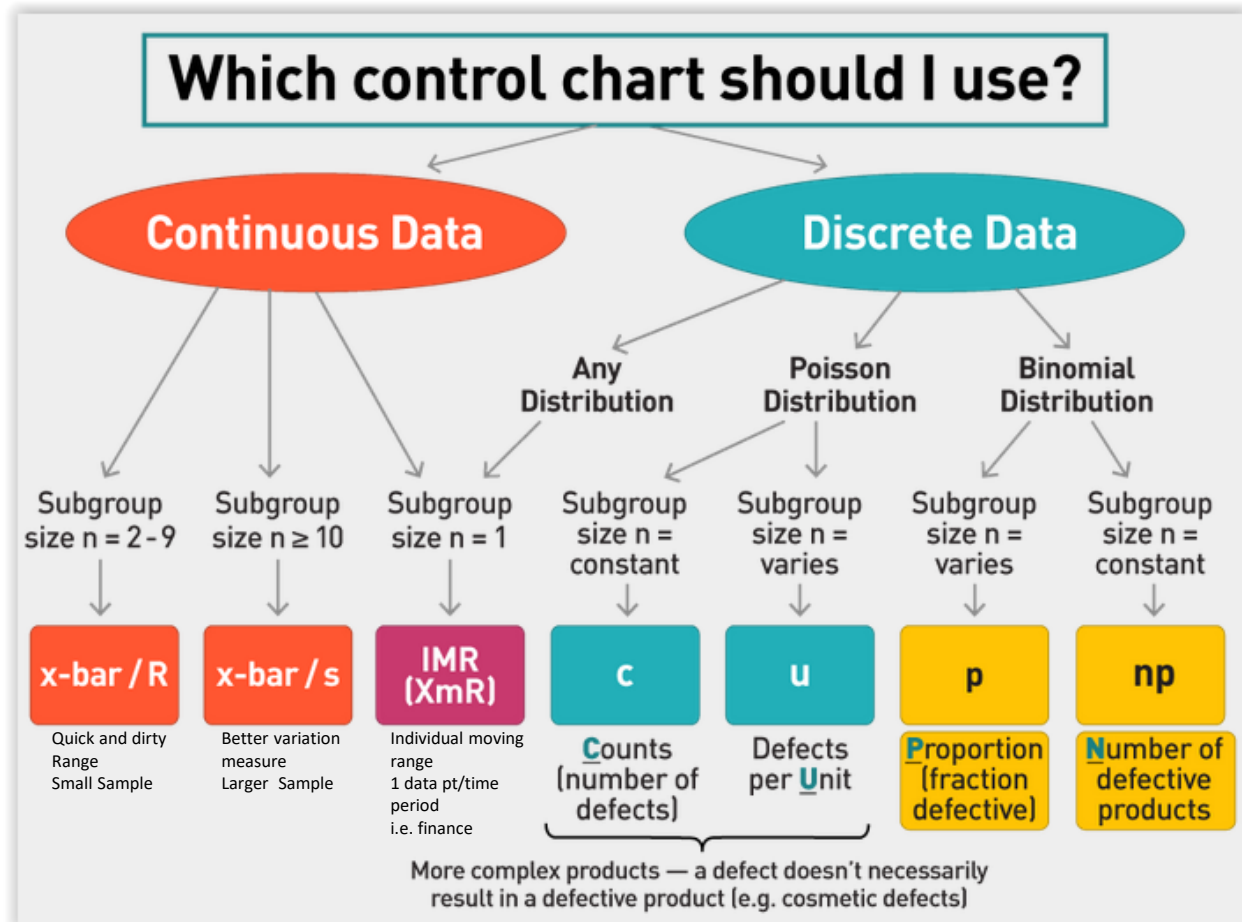
# 8.2 Control Chart Introduction and Types Available

If you take 4 measurements per hour on the length of a part? What chart?  
X bar/R bar

If you collect data on the number of visible dings on your part and you look at 3 every day? What chart?  
C chart

If you collect data on the number of visible scratches on your part and you look at a different number of parts every day? What chart?  
U chart

Every month you collect data on your electric bill? What chart?  
IMR/XMR chart

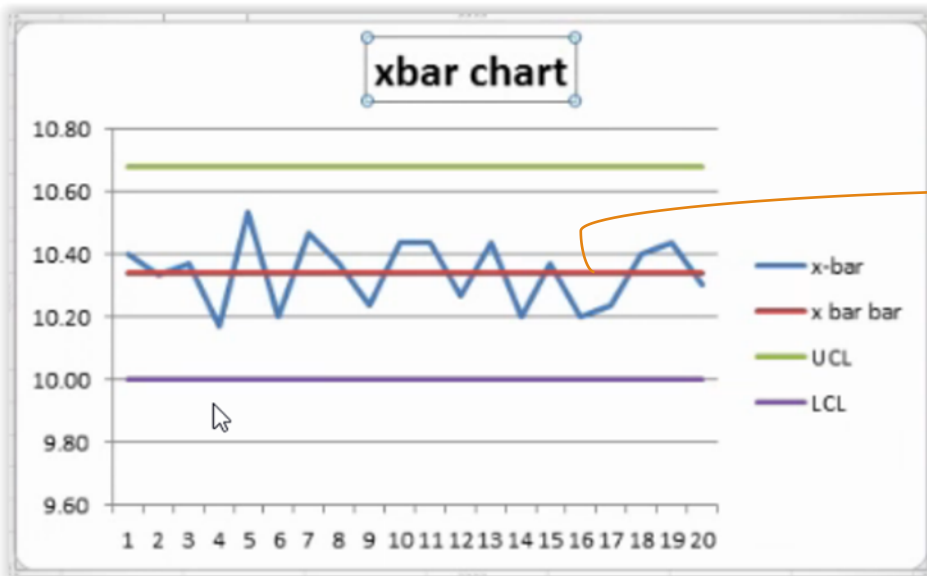


# 8 Control Chart Interpretation



Always do range chart first, needs to be in control.

The process variation is stable and predictable.



X bar is the average measurement or dimension of a part.

The process is consistent and stable, although that may not mean you are meeting the requirements.

What coefficient would you look at if comparing forecasting models in terms of which has the strongest relationship between the input and output?

## 9.5 Three Time Series Models

Run a regression to create an equation to predict next value

### Autoregressive Model: AR(1)

- Takes advantage of linear relationship between successive values of time series
- **First-order autoregressive model**
  - Linear regression equation:  $y_t = \beta_0 + \beta_1 y_{t-1} + \varepsilon_t$
  - $y_t$  = output at time  $t$
  - Example: March;  $t = 3$
  - $y_3 = \beta_0 + \beta_1 y_{3-1} + \varepsilon_3$
  - I.e., predicting March data by using February data

Takes advantage of relationship between successive values.  
Looks back 1 time period.

## Time Series Models

1. First-order autoregressive model, a.k.a. AR(1)
2. Moving average forecast model
3. Exponential smoothing model

Uses weighting on actual and forecasted previous value to predict next value

### Exponential Smoothing Model

- Best suited for forecasting time series without seasonal variation
- Unlike moving average model, data values not all weighted equally
- Forecasting equation:  $\hat{y}_t = w y_{t-1} + (1 - w) \hat{y}_{t-1}$ 
  - $\hat{y}_t$  = estimate of  $y$  at time period  $t$
  - Example:  $t$  = February
  - $w$  = smoothing constant
  - Your choice, pick any number between 0 and 1
  - Example:  $w = 80\%$

Uses all the data available and does not weight them equally.

Average the last K number of values to predict next value

### Moving Average Model

A.k.a. "rolling average method"; smooths out short-term fluctuations

- Uses average of last several values of time series to forecast next value;  $k$  = number of values in span
  - Example: Monthly data, span ( $k$ ) = 3
  - I.e., use average of values from January, February, and March to predict April value
- Can look back more than one time period
- Disadvantage: If, say,  $n = 100$  and  $k = 5$ , forecast overlooks 95% of available data

This is easy to hand calculate just by averaging the last K number of values to predict the next.

Answer: correlation coefficient, R, the higher the R the better, the stronger the relationship between input and output

How do you determine if your measurement system is repeatable and reproducible?

Calculate Kappa, define operational definitions, have multiple people collect the same data, have the same person collect the same data on different days

# Highlights: Video Segment 1.9

- **Kappa(K) : is an index that can be used to determine if your measurement system(tool) is good for discrete data is good in terms of reproducibility (between people) and repeatability(the same person's ratings).**
- **This acts as a flag that the measurement system needs to be reevaluated if it is not producing reproducible and or repeatable results. This means your results may not be valid.**

	Is it Good or Bad?	Is it Good or Bad?	Did you agree?
Peanut #	Your answer	Your fellow inspector's answers	yes/no
1	G	G	TRUE
2	B	B	TRUE
3	G	B	FALSE
4	B	B	TRUE
5	B	G	FALSE
6	G	G	TRUE
7	B	B	TRUE
8	G	G	TRUE
9	G	B	FALSE
10	B	B	TRUE
11	B	G	FALSE
12	B	B	TRUE
13	B	G	FALSE
14	G	G	TRUE
15	B	B	TRUE
16	G	G	TRUE
17	B	B	TRUE
18	B	G	FALSE
19	B	B	TRUE
20	B	B	TRUE
Totals	20	20	
Percent Good	7	9	
Percent Bad	13	11	
Percent Agreed			14
Percent Good	0.35	0.45	
Percent Bad	0.65	0.55	
Percent Agreed			0.70

Calculate Kappa:			
$K = (P \text{ observed} - P \text{ chance}) / (1 - P \text{ chance}) =$			
P Observed	0.70		Note:
P Chance	$(.35 \times .45) + (.65 \times .55) =$	0.515	good x good + bad x bad
$K = (.70 - .515) / (1 - .515) =$		0.381443299	
Is your measurement system good?			
If $K > .7$ then the system is good, my K value is .38, therefore it is not a good measurement system in terms of reproducibility.			

# Sample Size

## Sample Size Formula for Continuous Data

$$n = \left( \frac{z^* \hat{\sigma}}{E} \right)^2$$

- If you want a higher confidence level or be able to see a smaller change ....Z goes up, so N gets bigger
- If you have higher variability, sigma is higher, so N is larger
- If you have a higher margin of error you are willing to accept, N gets smaller
- If you have a smaller margin of error, then your N gets bigger



# Final Review

---

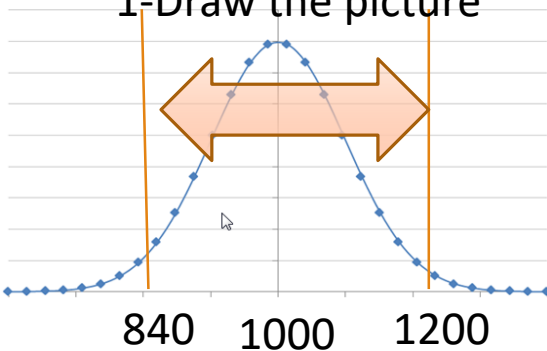
- **Regression**
  - How do you write down the formula from Excel Output?
  - How to use a regression equation to predict the output(y)?
  - How do you tell which variables are useful to have in your regression?
- **Correlation Coefficient R vs Coefficient of Determination R<sup>2</sup> – what do they represent?**
- **When is data not appropriate for regression? What are residuals?**
- **Causation vs Correlation**
- **Z calculation for the probability of a value falling between A and B**
- **Time Series - Autocorrelation and R<sup>2</sup>**
- **List of Statistical tools: Correlation, regression, hypothesis testing, scatter plots, process control charts, chi-square testing, etc etc.**
- **Basic ways to describe data and Calculate: mean, median, mode, range, standard deviation, variance**
- **Sample size formula and manipulation impacts**
- **Margin of error and confidence intervals**
- **Process Control charts**
- **How can you determine if your measurement system is repeatable and reproducible?**
- **Hypothesis testing – at what alpha do you reject, at what p-value do you reject**

## Quiz 2 Prep Question 1: Practice with Z calculations

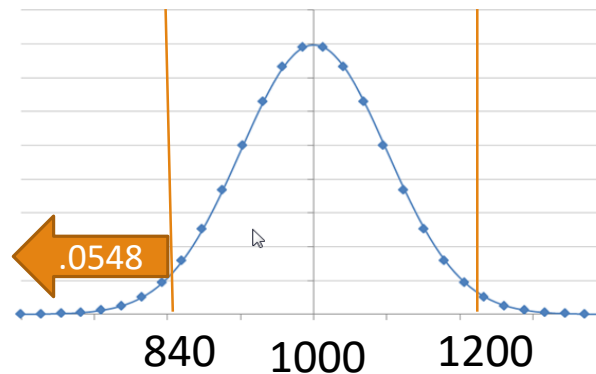
The distribution of weekly incomes of supervisors at the ABC Company follows the normal distribution, with a mean of \$1000 and a standard deviation of \$100.

What percent of the supervisors have a weekly income between \$840 and \$1200?

1-Draw the picture



2-Think about what you are calculating related to the picture



$$Z = \frac{X - \mu}{\sigma}$$

$$Z = \frac{840 - 1000}{100} = -1.6$$

Look up in tables,  $p = .0548$

Or in Excel

`=NORM.DIST(840,1000,100,TRUE)`

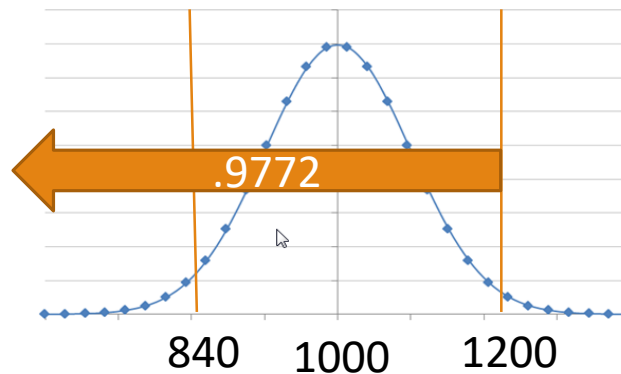
$$Z = \frac{X - \mu}{\sigma}$$

$$Z = \frac{1200 - 1000}{100} = 2$$

Look up in tables,  $p = .9772$

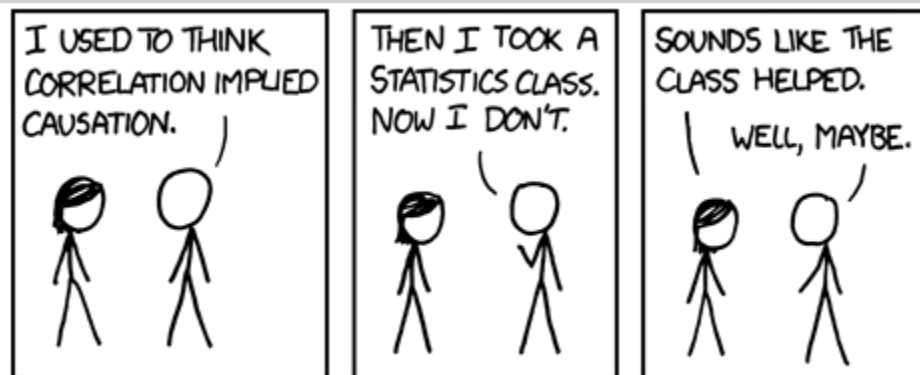
Or in Excel

`=NORM.DIST(1200,1000,100,TRUE)`



$.9772 - .0548 = .9224$ , so 92.24% have a weekly income between \$840 and \$1200

# Correlation vs Causation



From [xkcd](#), a comic by Randall Munroe

Correlation refers to the degree in which two measurements tend to vary together. Take the correlation between ice cream sales and drowning deaths. As ice cream sales increase, so do drowning deaths. Does that mean selling ice cream causes people to drown? Probably not. More likely is that people swim more and eat more ice cream the hotter it gets, so both are driven by the outside temperature.

Strong correlation doesn't mean cause and effect relationship....

## Correlation has different causes

- the first caused the second
- the second caused the first
- Confounding factor– interference by a third variable distorts the association being studied between two other variables, because of a strong relationship with both of the other variables
- Common Cause – like the ice cream example
- Coincidence

## Highlights: Video Segment 6.7:Correlation

### Two Indices

1. Correlation coefficient ( $r$ )
2. Coefficient of determination ( $r^2$ )

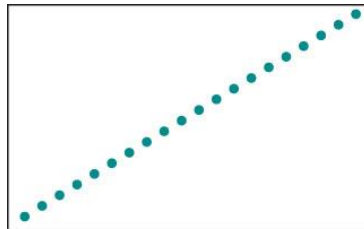
### Correlation Coefficient ( $r$ )

- $-1 < r$
- $-1$  = perfect negative correlation
- $1$  = perfect positive correlation
- $0$  = no relationship
- Rule of thumb:  $r$  value of  $\sim \pm 0.7$  desired
- Indicates meaningful relationship

Scatterplots provide a visual description of the relationship between two quantitative variables. The *correlation coefficient* is a numerical measure for quantifying the linear relationship between two quantitative variables.

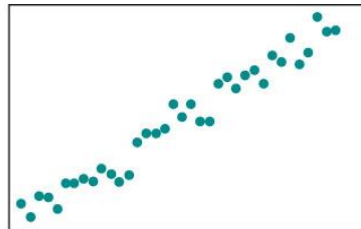
If the variability decreases, what does your correlation coefficient get closer to?  
What does a correlation coefficient  $r = -.72$  mean?

# Properties of $r$



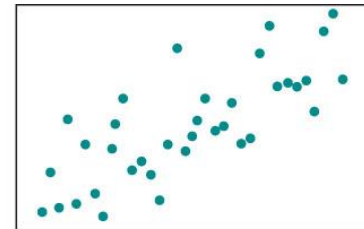
Perfect positive linear relationship,  $r = 1$

(a)



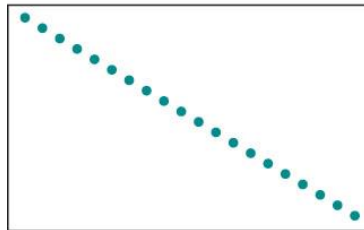
Strong positive linear relationship,  $r = 0.9$

(b)



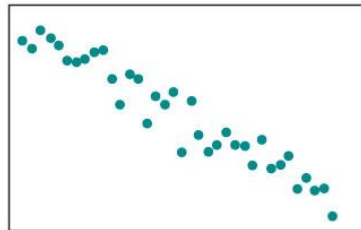
Moderate positive linear relationship,  $r = 0.5$

(c)



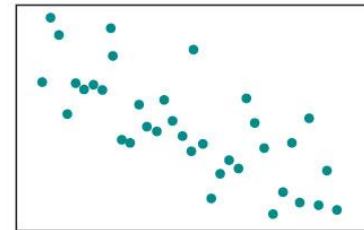
Perfect negative linear relationship,  $r = -1$

(d)



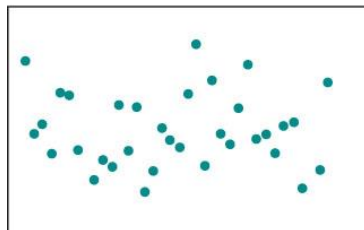
Strong negative linear relationship,  $r = -0.9$

(e)



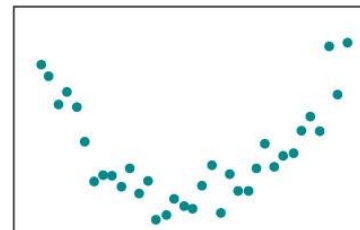
Moderate negative linear relationship,  $r = -0.5$

(f)



No apparent linear relationship,  $r = 0$

(g)



Nonlinear relationship but no linear relationship,  $r = 0$

(h)

Answer: 1 or -1

Answer: There is a moderate negative correlation between two factors

What if you performed a linear regression analysis on successive values of a time series analysis and you see autocorrelation.....what might your  $r^2$  ?

## Highlights: Video Segment 6.7:Correlation

### Two Indices

1. Correlation coefficient ( $r$ )
2. Coefficient of determination ( $r^2$ )

### Coefficient of Determination ( $r^2$ )

- Correlation coefficient squared
- Measure of the percentage of variability in  $y$  that can be accounted for by  $x$ 
  - Trying to find an input  $x$  that is influencing our output  $y$
  - $x$  will not explain all of  $y$
  - Recall: There is variability in everything we do.
- Metric for whether input  $x$  is really contributing to output

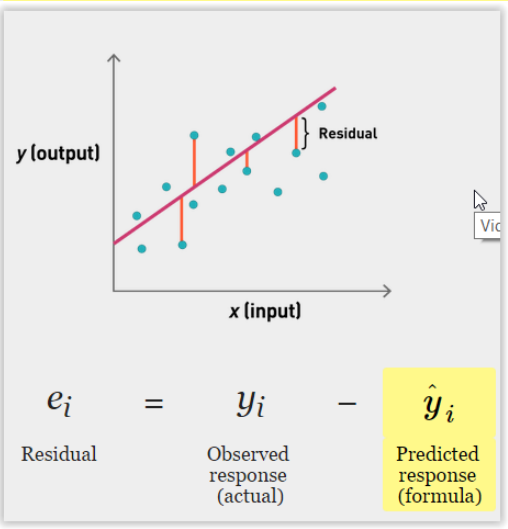
Measures the goodness of fit of the regression equation to the data. We interpret  $r^2$  as the proportion of the variability in  $y$  that is accounted for by the linear relationship between  $y$  and  $x$ . The values that  $r^2$  can take are  $0 \leq r^2 \leq 1$ .

Answer: correlation,  $r$  would be closer to 1 or -1, which would mean  $r^2$  would be close to 1

# Highlights: Video Segment 6.9:Residuals and Other Warnings

## What Is a Residual?

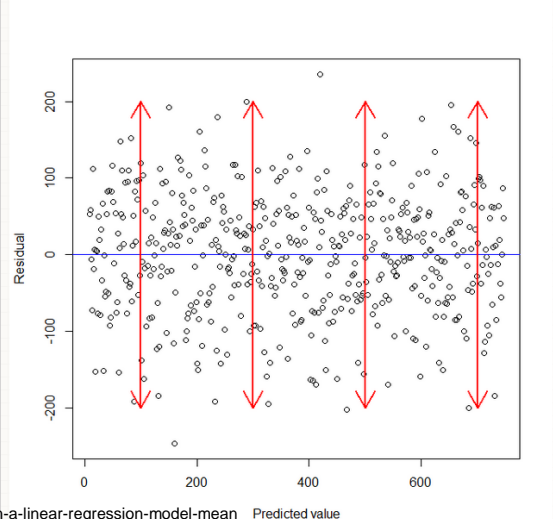
- Synonymous to error; should be random
  - The distance between actual data point and the line determined by linear equation
  - Determined by the difference between observed and predicted values of  $y$ 
    - Ideally, points fall on regression line (i.e., perfect model)
    - Error would then be zero (rare).
- When plotted, a random series of points around a zero reference with no evidence of a pattern



## Assumptions of Regression

1. Residuals are independent.
2. Residuals are normally distributed with a mean of zero.
  - The regression line will sometimes be high or low (i.e., over- or underpredicting).
3. There are equal variances ( $\sigma^2$ ) of  $y$ .

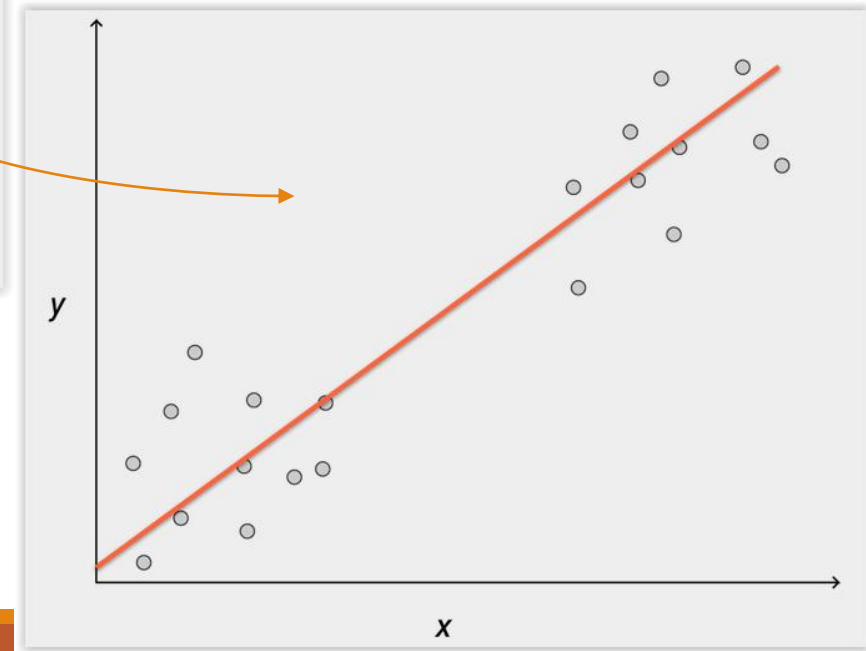
It means that when you plot the individual error against the predicted value, the variance of the error predicted value should be constant. See the red arrows in the picture below, the length of the red lines (a proxy of its variance) are the same.



# Highlights: Video Segment 6.9:Residuals and Other Warnings

## Other Points of Interest

- Certain data is inappropriate for a regression analysis:
  - Residuals form a pattern.
  - Large outliers are present.
  - "Clumped" data appears linear.
- Avoid extrapolating outside data.
- Beware of lurking variables, or Simpson's paradox.
- A strong correlation does not mean causation.





# Highlights: Video Segment 6.4:Regression Intro

## Simple Linear Regression: Equation

- Tries to create best-fitting line through plot
- Describes the relationship between two variables

$$\hat{y} = b_0 + b_1 x$$

Output variable  
(predicted  
response)

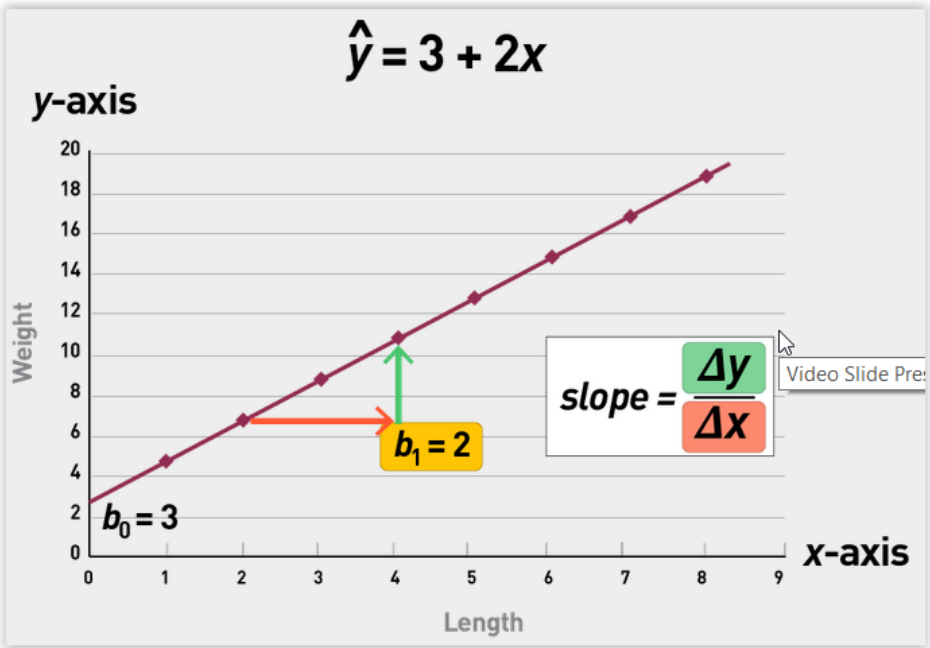
$b_0$   $y$ -intercept

$b_1$  Slope  $x$  Input variable

- **$y$ -intercept:** where the line crosses the  $y$ -axis
- **Slope:** how slanted the line is

Slope

1 Variable (aka 1 X) in simple linear regression



# Regression Breakout Review from Week 9

[illegible]

# Agenda

---


Topic
Introduction
Final Review: Video 10 and Beyond
Process Improvement Project Rubric Review
Review of Upcoming Assignments and Open Question

## Process Improvement Project – Feedback –

Content Requirements	Possible Points	Points Earned	Comments
<b>Project</b>			
A) An executive summary is provided in the storyboard format including: Is the storyboard presented in 1 PowerPoint slide? Follows DMAIC? Are tools/graphs/charts used and clearly visible? Do they support findings and conclusions Are arrows, call-out boxes, etc. used to summarize, highlight questions and key learnings? Are expected results clear? And next steps noted?	5		
B) Is it a cohesive presentation opening with the business process and problem statement? The back-up slides (5-15) detail and support the storyboard content.	2		
C) Was the success measure clearly identified, operationally defined and baseline identified? (Was the data identified as continuous or discrete, includes SQL?)	3		
D) Was the data measurement plan or data stratification tree included?	1		
E) Was the data collection method identified?	1		
F) Was there rationale for the sample size taken? Use of the formula? Is there any reference to measurement error and how to minimize?	1		
G) Are at least 5 different tools and techniques clearly identified? Are the tools linked/ pertinent to the data analysis?	5		
H) Does the data analysis clearly tie to the problem conclusion? Is the "discovery" clear to the reader?	2		
<b>Total possible 100 points</b>	20	0.00	

# Review of Upcoming Assignments: Thursday Section

1. Process Improvement Project due 12/10 Midnight EST
2. Final Exam, 12/13, 9PM EST
  - Time limit: 90 mins
  - Don't leave any blank
3. BONUS QUESTION for .5 pts EXTRA on your final grade
  - **Describe at least one thing you will do differently at work or home, now that you have taken this course.**
  - Minimum three sentences in an email sent to [lsgill@syr.edu](mailto:lsgill@syr.edu), subject line: BONUS QUESTION YOUR NAME SECTION DAY and TIME
  - Due Midnight EST on 12/13, **no late submissions will be accepted**

	December 2018						
	Sunday	Monday	Tuesday	Wednesday	Thursday	Friday	Saturday
Week #10	2	3	4	5	6	7	8
	<b><u>Homework #6 Due:</u></b> 1. Time Series Problem posted in Excel				Live Class #10 		
Week #11	9	10	11	12	13	14	15
		<b><u>Process Improvement Project DUE</u></b>			Live Class #11: FINAL EXAM		

Other notes: I will send out a final grade report similar to the interim that shows your final grade, you will have 24 hrs. after receiving it to let me know if you have any questions or if I made any mistakes etc., before the grade becomes final. I expect to be sending those out by 12/21 at the latest, hopefully 12/17

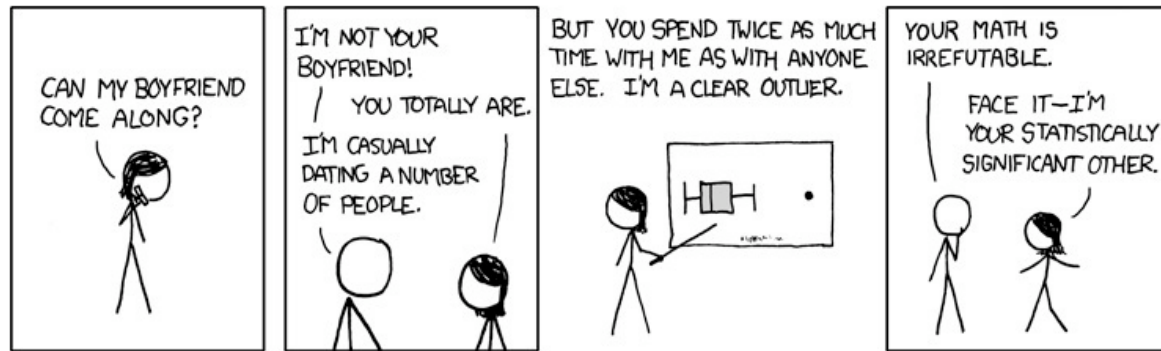
# Review of Upcoming Assignments: Sunday Section

1. Process Improvement Project due 12/13 Midnight EST
2. Final Exam, 12/16, 6:30PM EST
  - Time limit: 90 mins
  - Don't leave any blank
3. BONUS QUESTION for .5 pts EXTRA on your final grade
  - **Describe at least one thing you will do differently at work or home, now that you have taken this course.**
  - Minimum three sentences in an email sent to [lsgill@syr.edu](mailto:lsgill@syr.edu), subject line: BONUS QUESTION YOUR NAME SECTION DAY and TIME
  - Due Midnight EST on 12/16, **no late submissions will be accepted**

	December 2018						
	Sunday	Monday	Tuesday	Wednesday	Thursday	Friday	Saturday
Week #10	9	10	11	12	13	14	15
	Live Class #10 				<u>Process</u> <u>Improvement</u> <u>Project DUE</u>		
Week #11	16	17	18	19	20	21	22
	Live Class #11: FINAL EXAM						

Other notes: I will send out a final grade report similar to the interim that shows your final grade, you will have 24 hrs. after receiving it to let me know if you have any questions or if I made any mistakes etc., before the grade becomes final. I expect to be sending those out by 12/21 at the latest, hopefully 12/17

# Thank you and A Few Last Jokes



Special thanks to the incomparable [xkcd](#) for producing such wonderful comics.

Copyright 2003 by Randy Glasbergen.  
[www.glasbergen.com](http://www.glasbergen.com)



"I'm the consultant they brought in to create some new statistical buzzwords."

