Review

IMGD 2905

1

What are two main sources for data for game analytics?

2

What are two main sources for data for game analytics?

- Quantitative – instrumented game
- Qualitative – subjective evaluation

3

What steps are in the game analytics pipeline?

4

What steps are in the game analytics pipeline?

- Game (instrumented)
- Data (collected from *players* playing game)
- Extracted data (e.g., from scripts)
- Analysis
  - Statistics, Charts, Tests
- Dissemination
  - Report
  - Talk, Presentation

5

What is population versus sample?

6

What is population versus sample?

- Population – all members of group pertaining to study
- Sample – part of population selected for analysis

7

What is probability sampling?

8

What is probability sampling?

- Probability sampling – sampling considering likelihood of selection
  - Likelihood as part of population

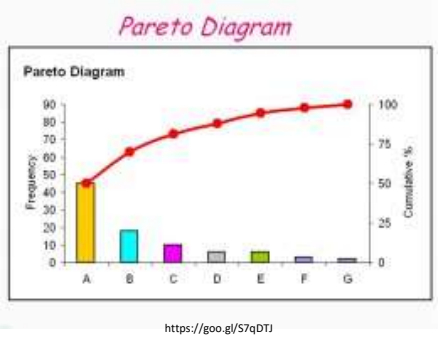
9

What is a Pareto chart? When used?

10

What is a Pareto chart? When used?

- Bar chart, arranged most to least frequent
- Line showing cumulative percent
- Helps identify most common, relative amounts



https://goo.gl/S7qDTJ

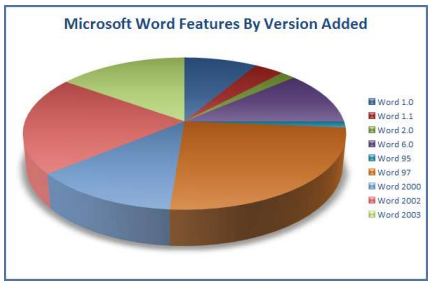
11

When should you *not* use pie chart?

12

When should you *not* use pie chart?

- When too many slices

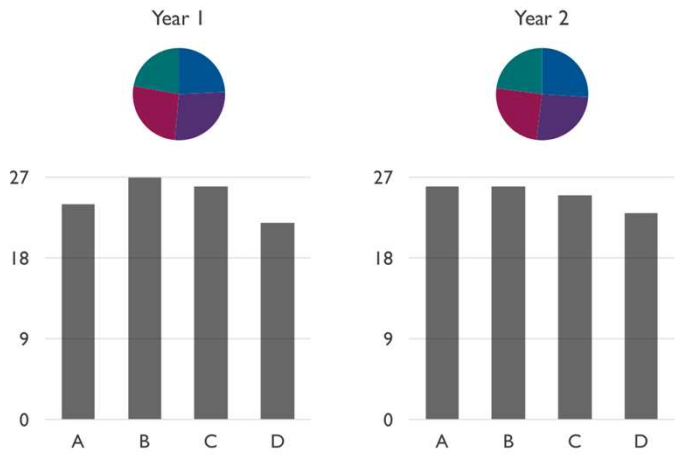


http://cdn.arstechnica.net/FeaturesByVersion.png

13

When should you *not* use pie chart?

- (Often) when comparing pies



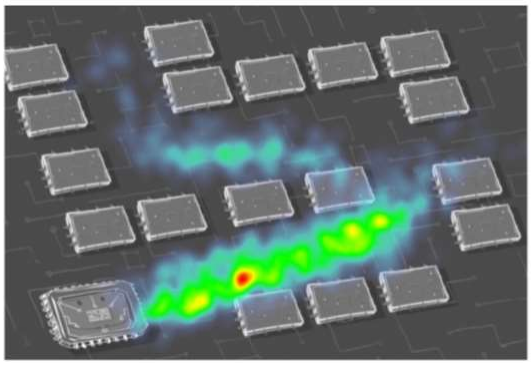
14

What is a heat map? Describe an example

15

What is a heat map? Describe an example

- Map where data represented as colors



16

Provide three guidelines for good charts

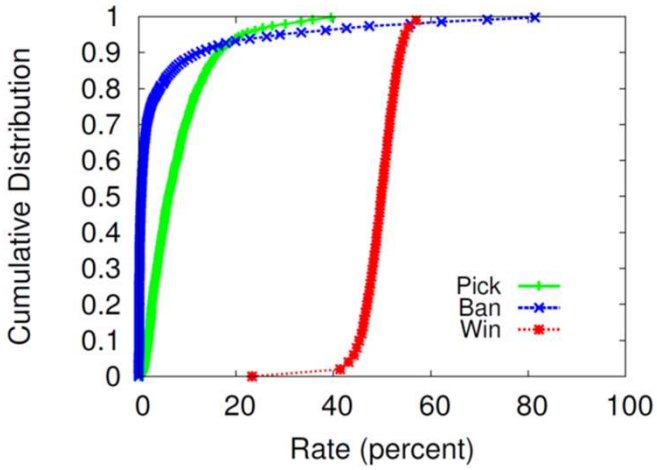
17

Provide three guidelines for good charts

1. Require minimum effort from reader
2. Maximize information
3. Minimize ink
4. Use commonly accepted practices
5. Avoid ambiguity

18

Which Measure of Central Tendency to Use? Why?

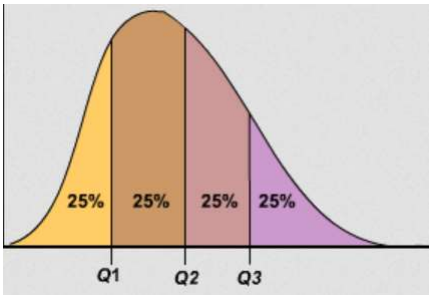


19

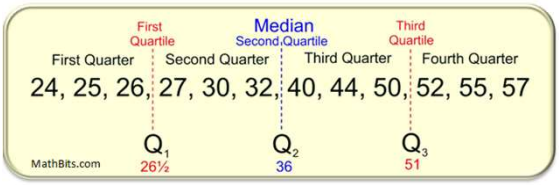
What are Quartiles?

20

What are Quartiles?



Three values that divide population into four equal sized groups



21

Describe how to Compute Variance

22

Describe how to Compute Variance

1. Compute mean.
2. Take a sample and compute how far it is from mean. Square this.
3. Repeat #2 for each sample.
4. Add up all.
5. Divide by number of samples (-1).

$$\text{Sample Variance} = s^2 = \frac{\Sigma(X - \overline{X})^2}{n - 1}$$

23

In Probability, what is an Exhaustive Set of Events? Give an Example.

24

In Probability, what is an Exhaustive Set of Events? Give an Example.

- A set of all possible outcomes of an experiment or observation

- e.g., coin: events {heads, tails}
- e.g., d6: events {even number, odd number}
- e.g., picking Hero in HOTS: events {LiLi, Ana, Malfurion, …} (all possible Champions listed)

25

Broadly, What are 3 Ways to Assign Probabilities? Give examples.

26

Broadly, What are 3 Ways to Assign Probabilities? Give examples.

- Classical (theory)
  - e.g., equal likelihood d6, so P(1) = $1/6^{th}$
- Empirical (by measurement/observation)
  - Probability of 1 min service rate at DD by observing service rates for 1 hour
- Subjective (hunch – sometimes guided by a bit of theory)
  - Probability of Iceland winning World Cup by deep analysis of teams and competition

27

Probability

- Draw 2 cards simultaneously. What is the probability of drawing 2 Jacks?

28

Probability

- Draw 2 cards simultaneously. What is the probability of drawing 2 Jacks?

P(2J) = P(J) x P(J | J)

= 2/5 x 1/4

= 1/10

29

Probability

- Draw 3 cards simultaneously. What is the probability of not drawing at least one King?

30

Probability

- Draw 3 cards simultaneously. What is the probability of not drawing at least one King?

P(K') x P(K' | K') x P(K' | K'K')
= 3/5 x 2/4 x 1/3
= 6/60
= 1/10

31

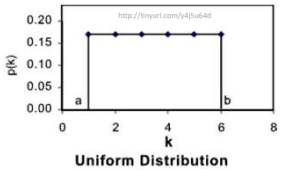
What Kind of Probability Distribution is:

- Rolling one 6-sided dice (d6)?

32

What Kind of Probability Distribution is:

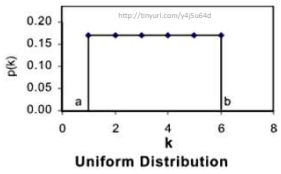
- Rolling one 6-sided dice (d6)?
  - Uniform (or "square")
- The number of 1's when rolling 20 d4's?

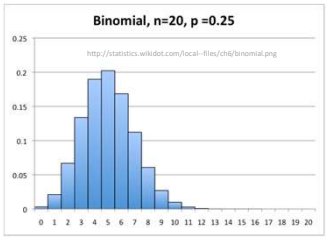


33

What Kind of Probability Distribution is:

- Rolling one 6-sided dice (d6)?
  - Uniform (or "square")
- The number of 1's when rolling 20 d4's?
  - Binomial
- The number of people that buy glazed donuts every 5 minutes at DD?

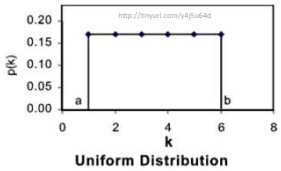


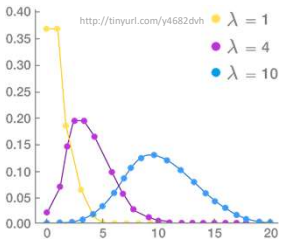


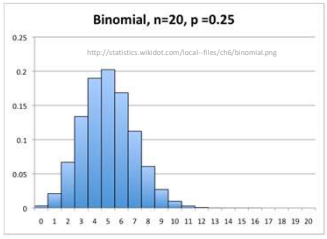
34

What Kind of Probability Distribution is:

- Rolling one 6-sided dice (d6)?
  - Uniform (or "square")
- The number of 1's when rolling 20 d4's?
  - Binomial
- The number of people that buy glazed donuts every 5 minutes at DD?
  - Poisson







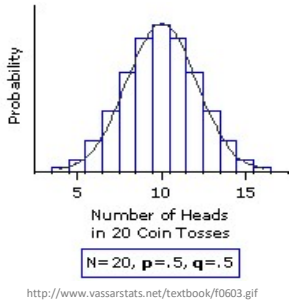
35

What are the characteristics of an experiment with a binomial distribution of outcomes?

36

What are the characteristics of an experiment with a binomial distribution of outcomes?

- Experiment consists of n independent, identical trials
- Each trial results in only success or failure (probability p for success for each)
- Random variable of interest (X) is number of successes in n trials



http://www.vassarstats.net/textbook/f0603.gif

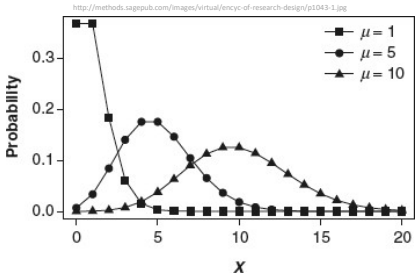
37

What are the characteristics of an experiment with a Poisson distribution of outcomes?

38

What are the characteristics of an experiment with a Poisson distribution of outcomes?

1. Interval (e.g., time) with units
2. Probability of event same for all interval units
3. Number of events in one unit independent of others
4. Events occur singly (not simultaneously)
5. Random variable of interest (X) is number of events that occur in an interval



Phrase people use is "random arrivals"

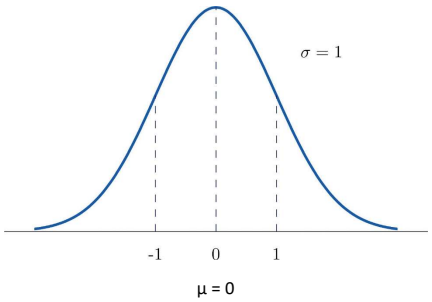
39

What is the Standard Normal Distribution?

40

What is the Standard Normal Distribution?

- Normal distribution
- Mean μ = 0
- Std dev σ = 1



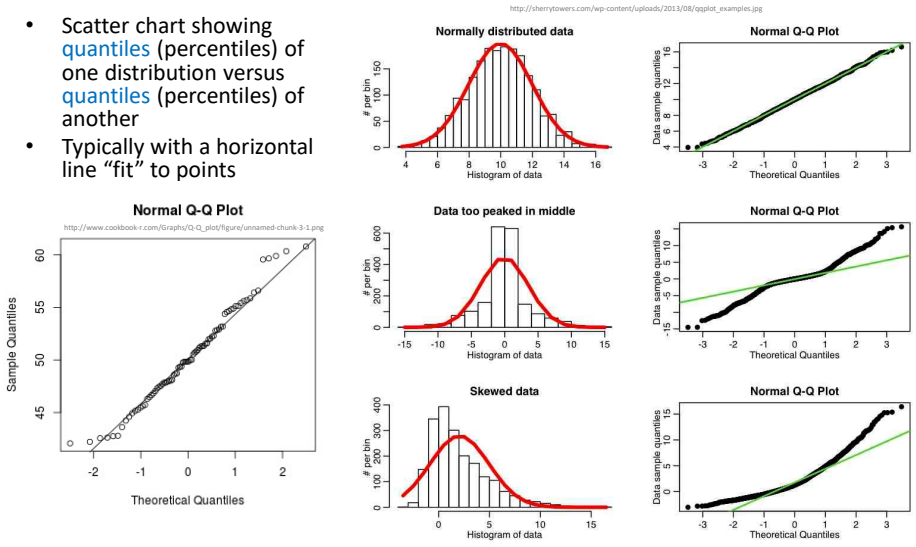
41

What is a Quantile-Quantile Plot?

42

What is a Quantile-Quantile Plot?

- Scatter chart showing quantiles (percentiles) of one distribution versus quantiles (percentiles) of another
- Typically with a horizontal line "fit" to points



How to read? → On line, distributions are similar

43

What is the Central Limit Theorem?

- Given population
- If take large enough sample size
- What does probability of sample means look like?

→ What is Distribution shape?

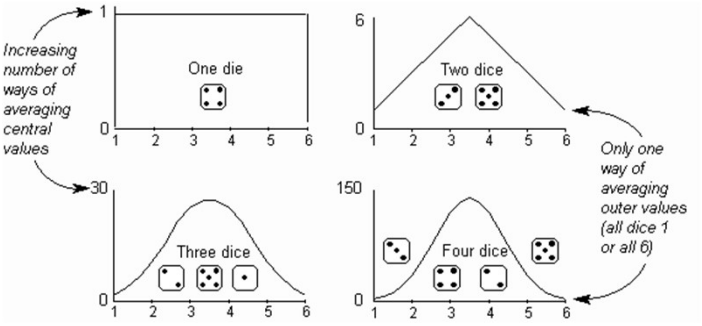
44

What is the Central Limit Theorem?

- Given population
- If take large enough sample size
- What does probability of sample means look like?

→ Distributed Normally

How big is "enough"?



http://home.ubalt.edu/ntsbarsh/Dice_001.gif

45

What is the Central Limit Theorem?

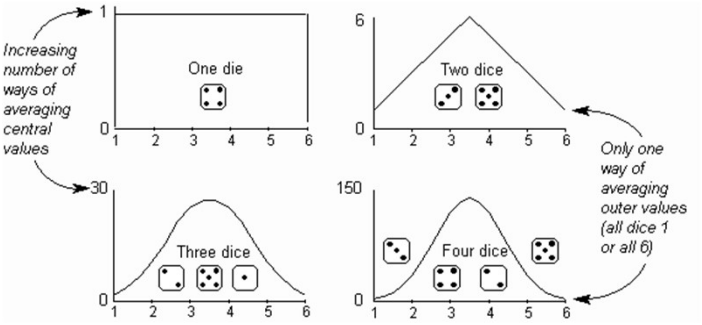
- Given population
- If take large enough sample size
- What does probability of sample means look like?

→ Distributed Normally

How big is "enough"?

- 30
- (15)

Does underlying distribution matter?



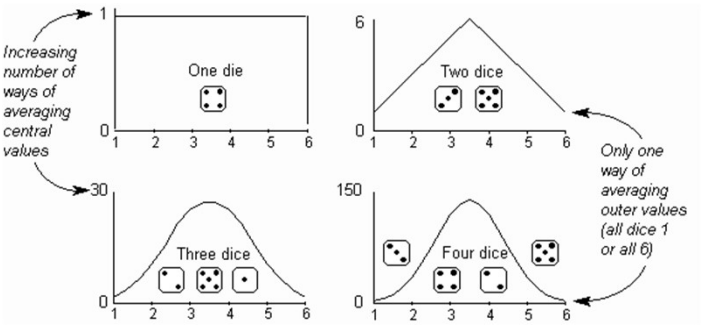
http://home.ubalt.edu/ntsbarsh/Dice_001.gif

46

What is the Central Limit Theorem?

- Given population
- If take large enough sample size
- What does probability of sample means look like?

→ Distributed Normally



http://home.ubalt.edu/ntsbarsh/Dice_001.gif

How big is "enough"?

- 30
- (15)

Does underlying distribution matter?

- No

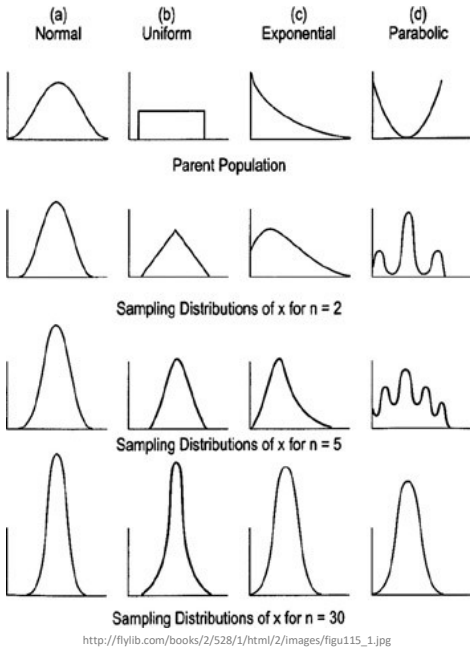
(see next slide)

47

Underlying Distribution does not Matter

Why do we care?

→ Can apply rules (e.g., empirical rule) to Normal Distributions!



http://flylib.com/books/2/528/1/html/2/images/figu115_1.jpg

48

Sampling Error

- What is sampling error?

49

Sampling Error

- What is sampling error?
  - Error from estimating population parameters from sample statistics
- *Size* of error is based on what two main factors?

50

Sampling Error

- What is sampling error?
  - Error from estimating population parameters from sample statistics
- *Size* of the error is based on what two main factors?
  - Population variance
  - Sample size (N)

51

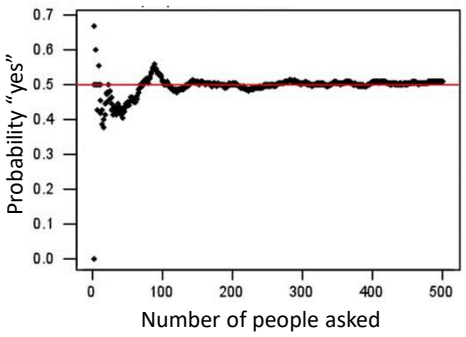
Statistic versus Sample Size

- Suppose wanted to know likelihood that WPI student played *Heroes of the Storm*
  - Ask N people, count "yes" and divide by N
- Ask 1 person?
- Ask 2 people?
- Ask 100 people?
- What does graph of "yes" probability versus N people look like?

52

Statistic versus Sample Size

- Suppose wanted to know likelihood that WPI student played *Heroes of the Storm*
  - Ask N people, count "yes" and divide by N
- Ask 1 person?
- Ask 2 people?
- Ask 100 people?
- What does graph of "yes" probability versus N people look like?



53

Confidence Intervals

- What is a confidence interval? Give an example

54

Confidence Intervals

- What is a confidence interval? Give an example
  - Range of values with specific certainty that population parameter is within
  - 95% confidence interval for mean time to complete a level in Super Mario: [1.25 minutes, 1.75 minutes]
- What is the *size* of confidence interval based on?

55

Confidence Intervals

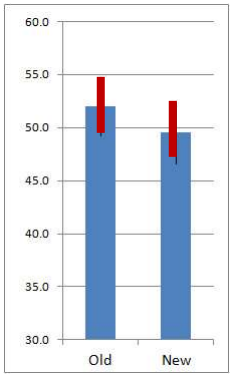
- What is a confidence interval? Give an example
  - Range of values with specific certainty that population parameter is within
  - 95% confidence interval for mean time to complete a level in Super Mario: [1.25 minutes, 1.75 minutes]
- What is the *size* of confidence interval based on?
  - Confidence (1-α)
  - Standard error (N, number of items in sample) (standard deviation)

$$\overline{X} \pm t\frac{s}{\sqrt{n}}$$

56

Interpreting Confidence Intervals

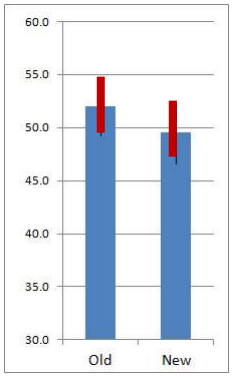
- Assume bars are conference intervals
- Interpret difference in *old* versus *new*



57

Interpreting Confidence Intervals

- Assume bars are conference intervals
- Interpret difference in *old* versus *new*



- Large overlap
- No statistically significant difference (at given α level)

Helpful hint: ignore sample means. Think about population means for Old and New

58

Hypothesis Testing

- What is the Null Hypothesis?
- What is the Alternate Hypothesis?

59

Hypothesis Testing

- What is the Null Hypothesis?
  - No significance difference between measured value and population parameter
- What is the Alternate Hypothesis?
  - Contrary to null hypothesis (i.e., there *is* a difference)
- Which do we test and *why*?

60

Hypothesis Testing

- What is the Null Hypothesis?
  - No significance difference between measured value and population parameter
- What is the Alternate Hypothesis?
  - Contrary to null hypothesis (i.e., there *is* a difference)
- Which do we test and *why*?
  - Test Null
  - Data can only reject hypothesis, not prove
  - → Reject Null

61

Hypothesis Testing

- Gathered data, computed sample mean, created Null hypothesis (H_0), chose significance ($\alpha = 0.01$)
- Calculate p-value = 0.05
- Make inference: CAN or CANNOT reject H_0?

62

Hypothesis Testing

- Gathered data, computed sample mean, created Null hypothesis (H_0), chose significance (α = 0.01)
- Calculate p-value = 0.05
- Make inference: CAN or CANNOT reject H_0?
  - CANNOT reject H_0
- What does that mean?

63

Hypothesis Testing

- Gathered data, computed sample mean, created Null hypothesis (H_0), chose significance (α = 0.01)
- Calculate p-value = 0.05
- Make inference: CAN or CANNOT reject H_0?
  - CANNOT reject H_0
- What does that mean?
  - May be no difference between sample mean and population mean (at 0.01 significance)

64

Regression

- What is the role of regression in data analytics?

65

Regression

- What is the role of regression in data analytics?
  - To predict an unobserved value from a mathematical model
- What is simple linear regression?

66

Regression

- What is the role of regression in data analytics?
  - To predict an unobserved value from a mathematical model
- What is simple linear regression?
  - A linear model relating two variables/factors
  - m is slope, b is y-intercept

$$Y = mX + b$$

67

Regression

- If the market value of a house can be represented by the model:

  value = 32673 + 35.036 *x* (square feet)
- How do you interpret the model? How can you use it?

68

Regression

- If the market value of a house can be represented by the model:

  value = 32673 + 35.036 *x* (square feet)

- How do you interpret the model? How can you use it?
  - Intercept is 32673. Base house value is $32k.
  - Slope is 35.036. Every square foot increases house value by $35
  - Given square feet, predict value: 1800 sq feet

  value = 32,673 + 35.036 x (1800) = $95,737.80

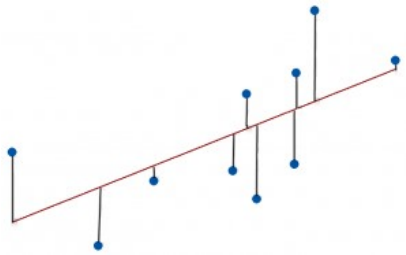
69

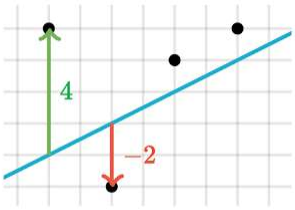
What are Residuals?

70

What are Residuals?

- A residual is difference between observed value and predicted value
- Vertical distance between a data point and **regression line**





71

What is Residual Analysis?

72

What is Residual Analysis?

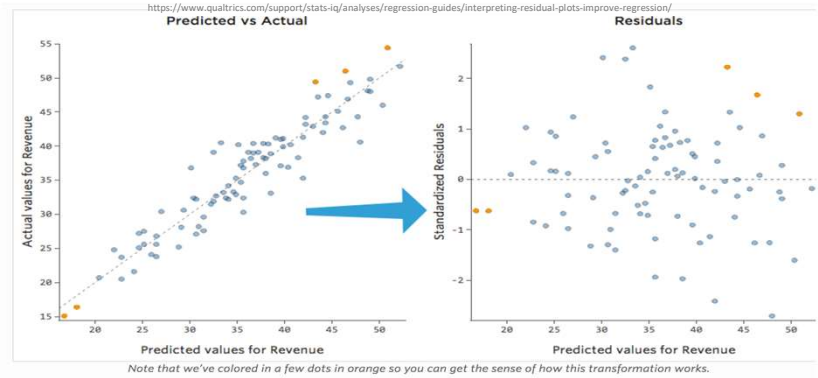


Chart residuals on vertical axis
and independent variable on horizontal axis.
No pattern? → Linear ok

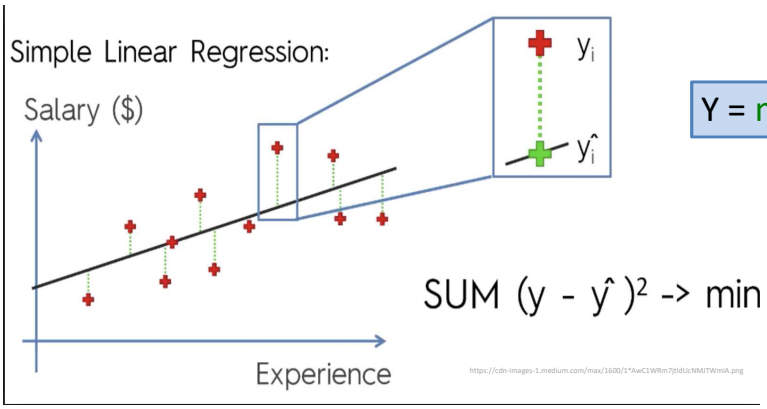
73

What is Least Squares Line?

74

What is Least Squares Line?

- Line that minimizes sum squared error



$Y = mX + b$

75

What is the Coefficient of Determination (R^2)?

76

What is the Coefficient of Determination (R^2)?

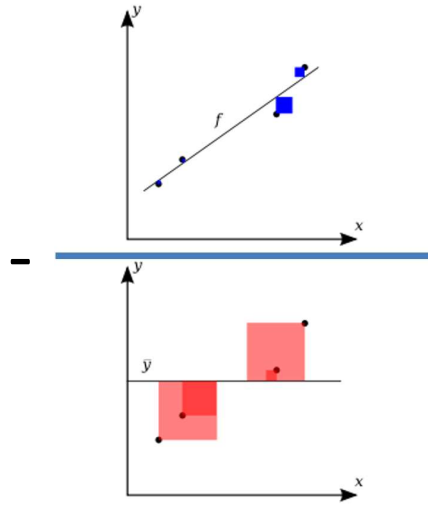
- Proportion of variance in the dependent variable predictable by the independent variable
- Percentage of variance explainable by model

$$R^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SST}$$

77

What is the Coefficient of Determination (R^2)?

$R^2 = 1 -$



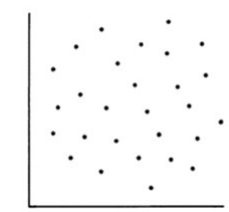
Variation in observed data model cannot explain (error)

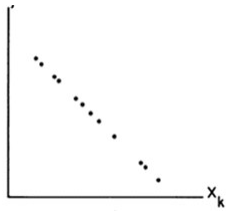
Total variation in observed data

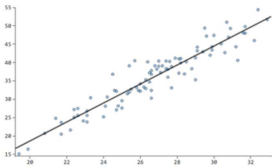
https://upload.wikimedia.org/wikipedia/commons/thumb/8/86/Coefficient_of_Determination.svg/400px-Coefficient_of_Determination.svg.png

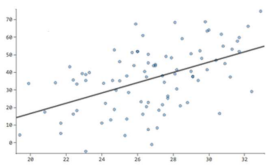
78

What is a the value of R^2?





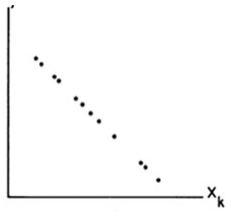




79

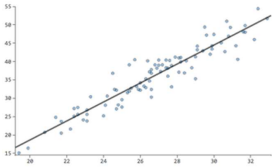
What is a the value of R^2?

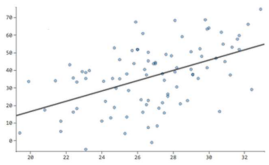
$R^2 = 0$



$R^2 = 1$

$R^2 = 0.8$





$R^2 = 0.2$

80