Review

IMGD 2905

What are two main types of data for game analytics?

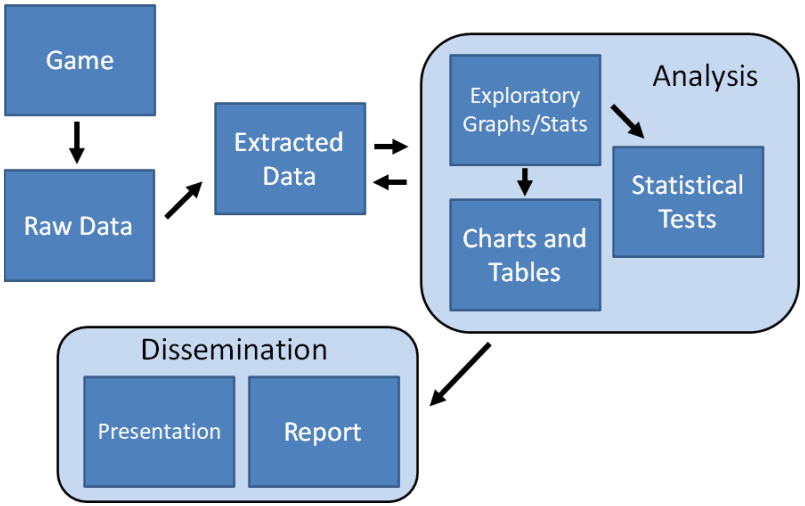
What are two main types of data for game analytics?

- Quantitative – objective data from the game, often from instrumention (code to write/log data), typically players playing the game
- Qualitative – subjective evaluation, typically from players during or after gameplay

What steps are in the game analytics pipeline?

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



What is population versus sample?

What is population versus sample?

- Population – all members of group pertaining to study
  - Typically want *parameter* of this group
- Sample – part of population selected for analysis
  - Typically compute *statistic* to estimate *parameter*

What is probability sampling?

What is probability sampling?

- Probability sampling – selecting members from the population group while considering the likelihood of selection
  - Likelihood as part of population

What is a variable in statistics?

What is a variable in statistics?

- Any characteristics that can be measured, classified or counted
  - Examples: age, eye color, income, high score, kill-death ratio, vehicle type
  - e.g., time spent in competitive mode in *Starcraft 2*
  - e.g., vehicle choice in *Grand Theft Auto* (GTA)

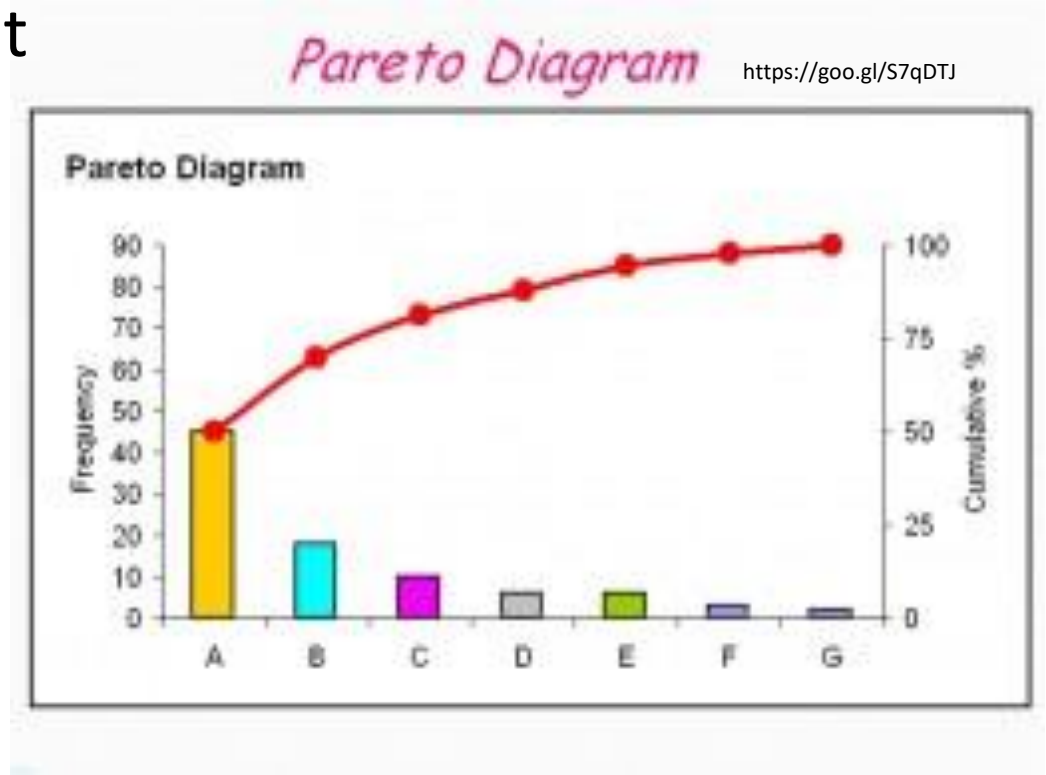
Player	Hours	Champ
A	2	Leona
B	7.5	Teemo

- Variables in columns
- Independent variable is inherent in population, versus dependent variable that want to assess

What is a Pareto chart? When used?

What is a Pareto chart? When used?

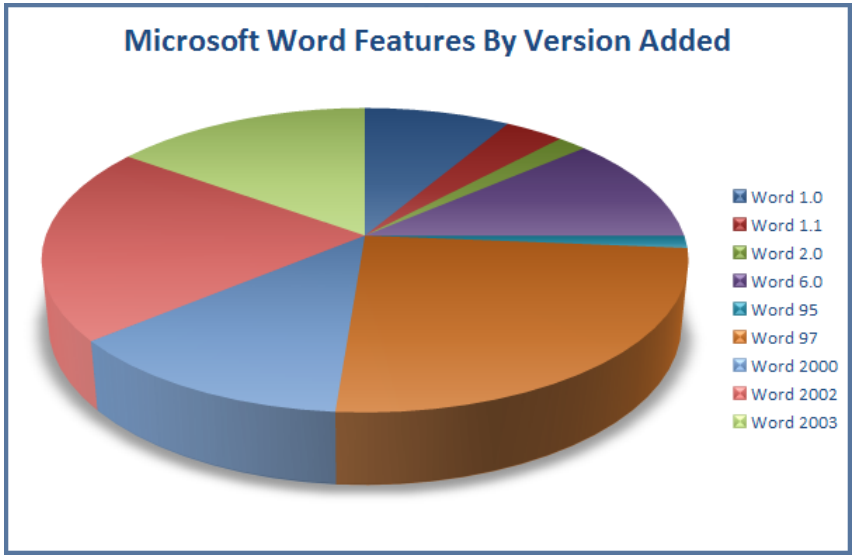
- Bar chart, arranged most to least frequent
- Line showing cumulative percent
- Helps identify most common, quantify relative amounts



When should you *not* use pie chart?

When should you *not* use pie chart?

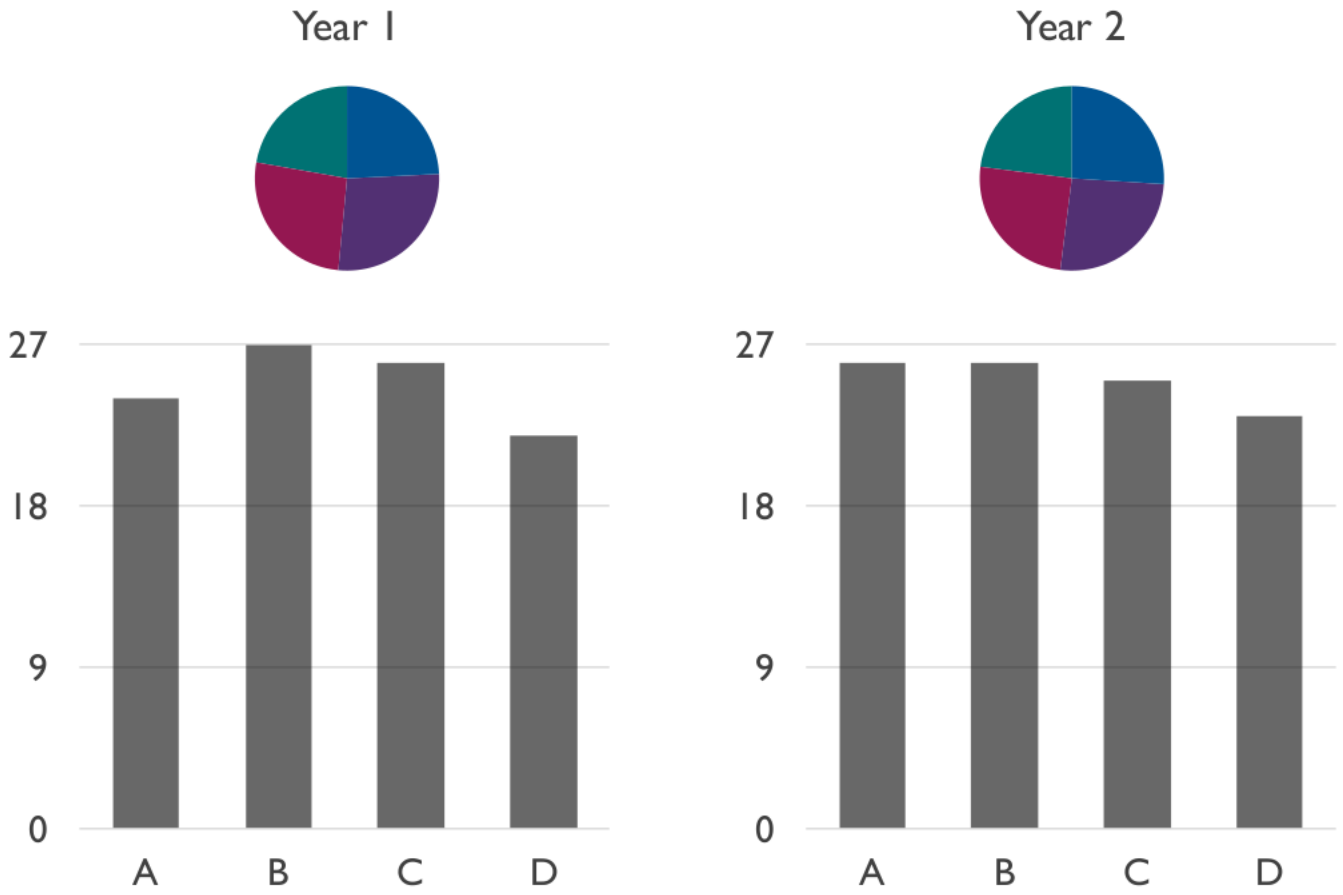
- When too many slices (more than 3)



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

When should you *not* use pie chart?

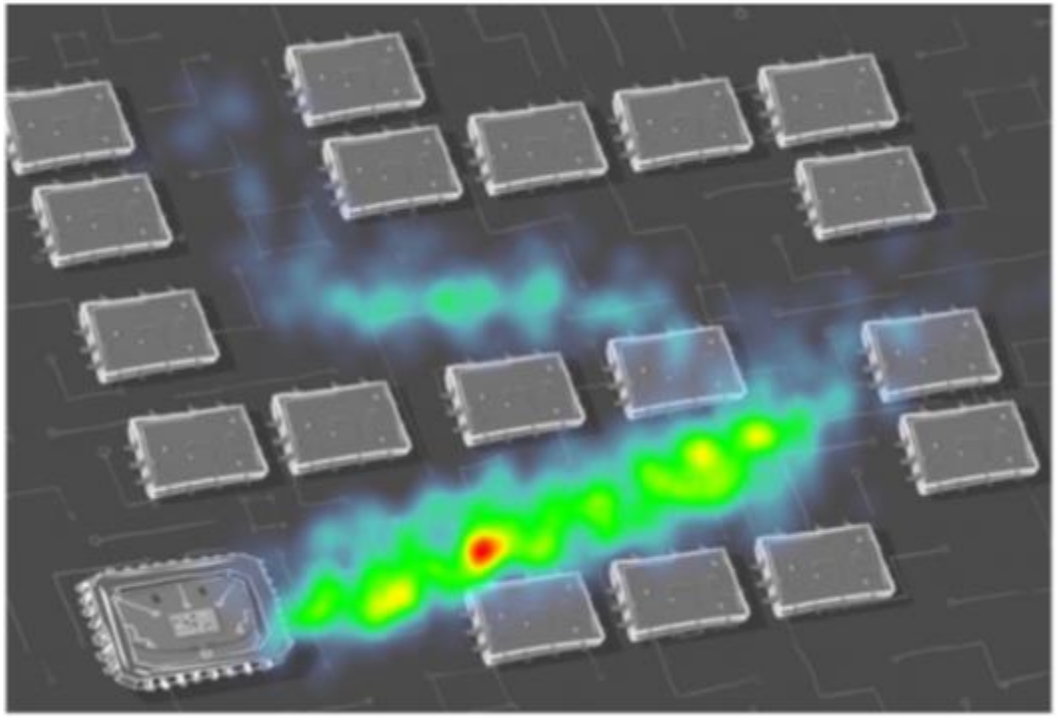
- (Often) when comparing pies



What is a heat map? Describe an example

What is a heat map? Describe an example

- Map where data represented as colors
  - Typically, greater values → brighter intensity colors

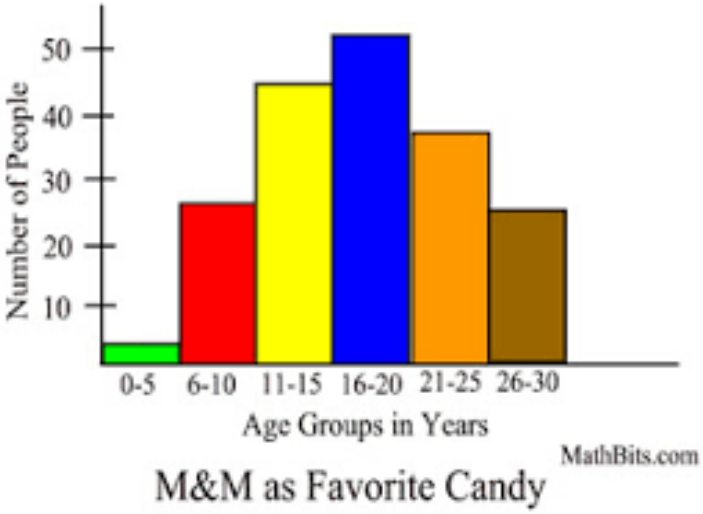


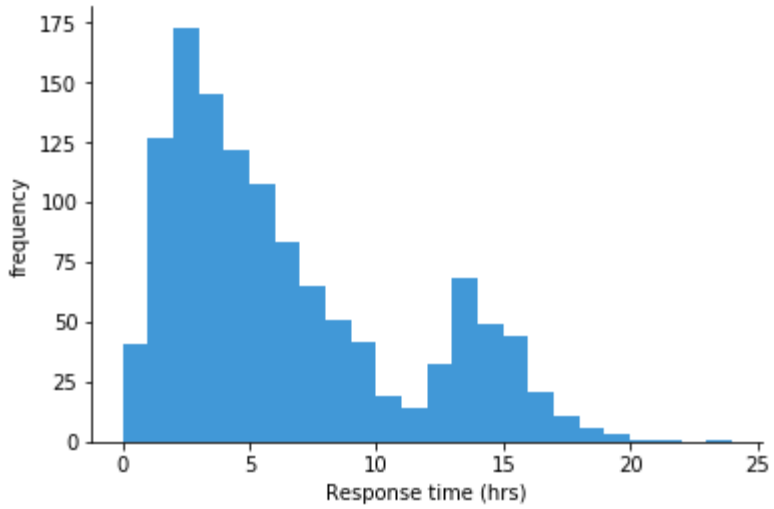
Provide three guidelines for good charts

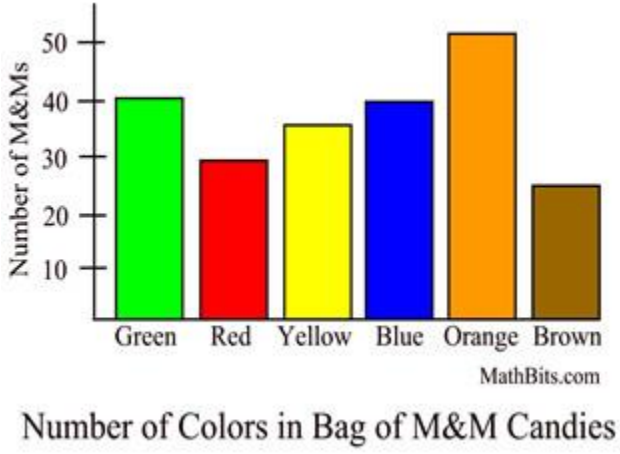
Provide three guidelines for good charts

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

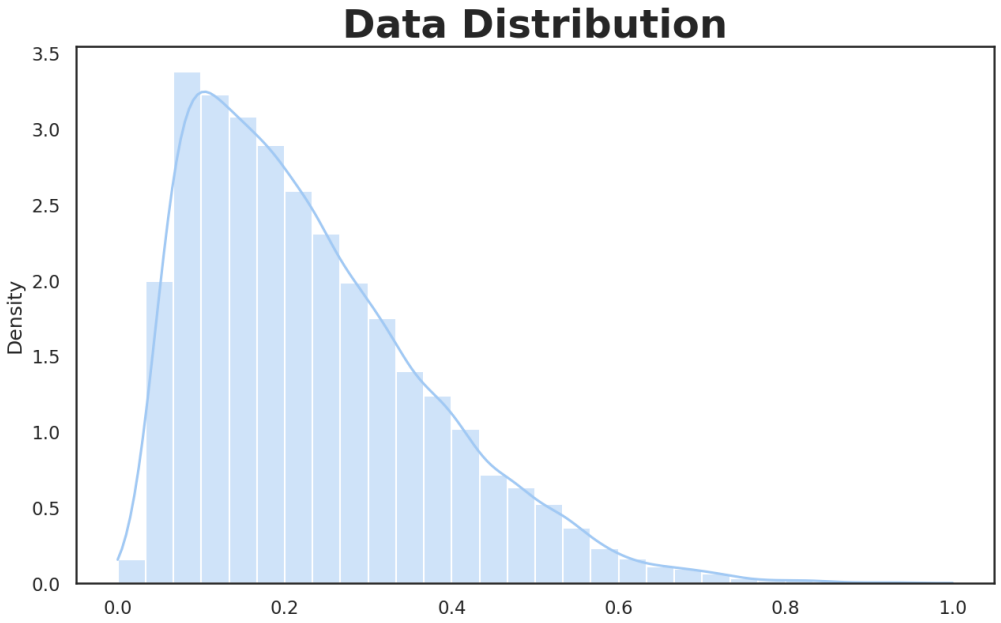
Which Measure of Central Tendency to Use? Why?



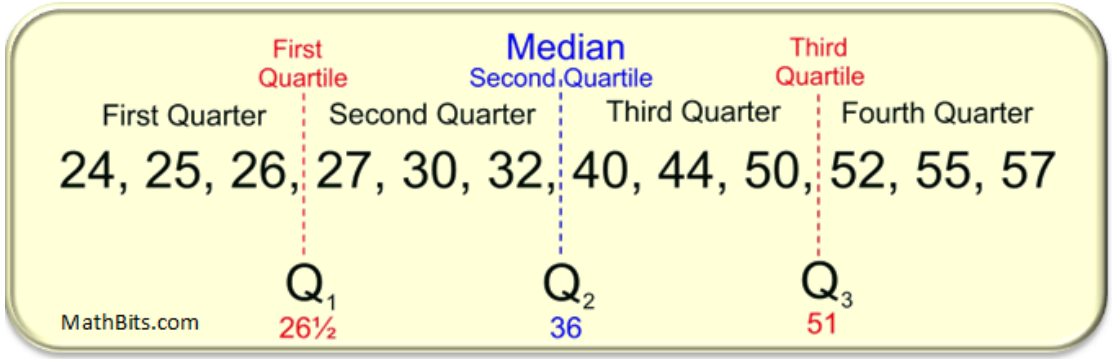




What are Quartiles?



Three values that divide population into four equal sized groups

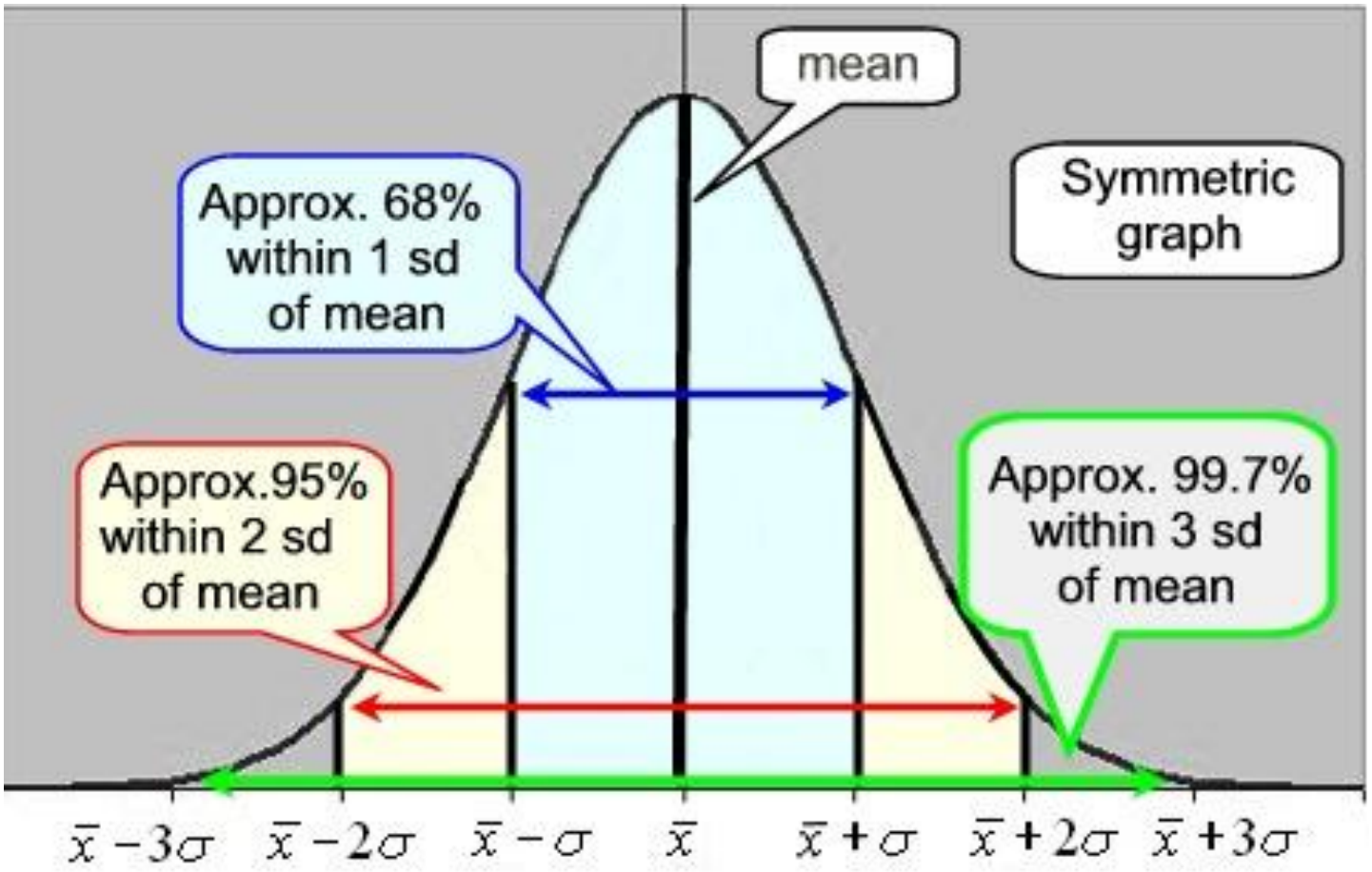


Describe how to Compute Variance

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).

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

What is Mendenhall's Empirical Rule?

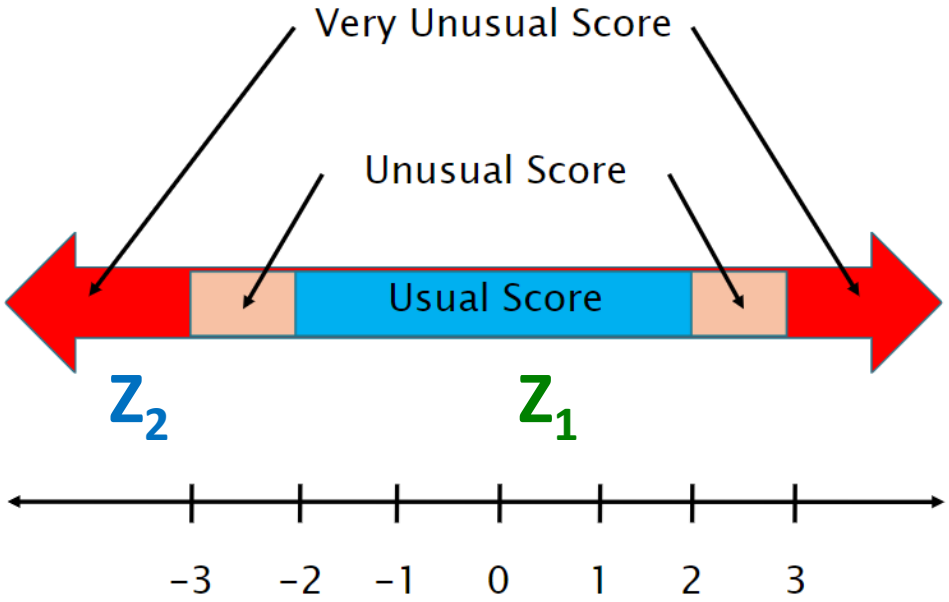


https://mathbitsnotebook.com/Algebra1/StatisticsData/normalgrapha.jpg

What can you interpret from Z-score?

$Z_1 = \mathbf{0.5}$? $Z_2 = \mathbf{-3.2}$?

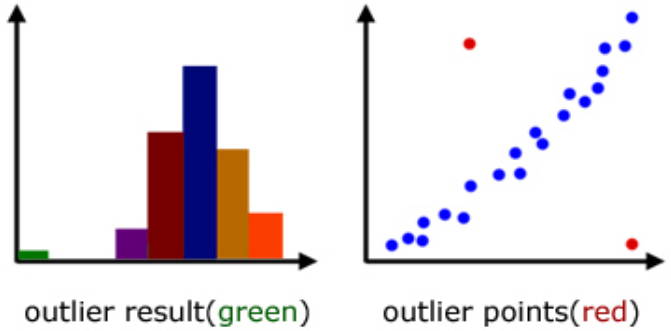
- How "unusual" a score is
- Where (above or below) score is relative to the average (mean) and spread (std dev)



$$z = \frac{X - \bar{X}}{s}$$

Groupwork

- Rank *measures of dispersion* by sensitivity to outliers
  - CoV
  - Range
  - Std Dev
  - Semi-interquartile Range



http://www.a-levelmathstutor.com/images/statistics/outliers-graph01.jpg

https://web.cs.wpi.edu/~imgd2905/d23/groupwork/4-outlier-effect/handout.html

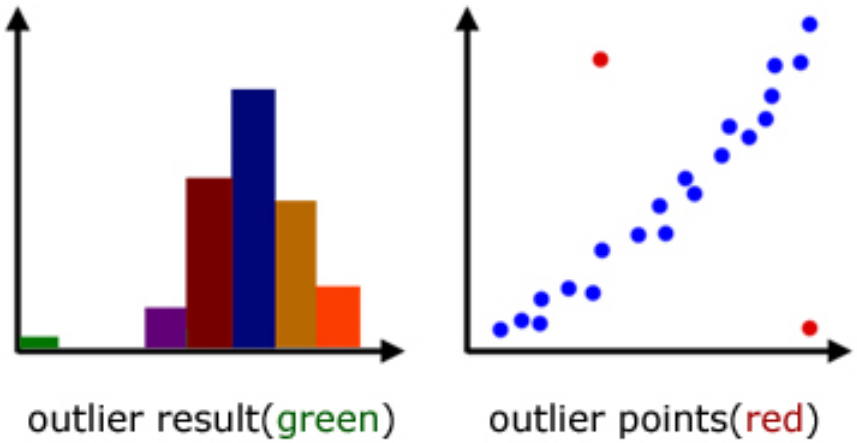
Ranking of Affect by Outliers?

Measure of Dispersion

- Range
- Standard Deviation
- Coefficient of Variation
- Semi-interquartile Range

Most to Least

?



http://www.a-levelmathstutor.com/images/statistics/outliers-graph01.jpg

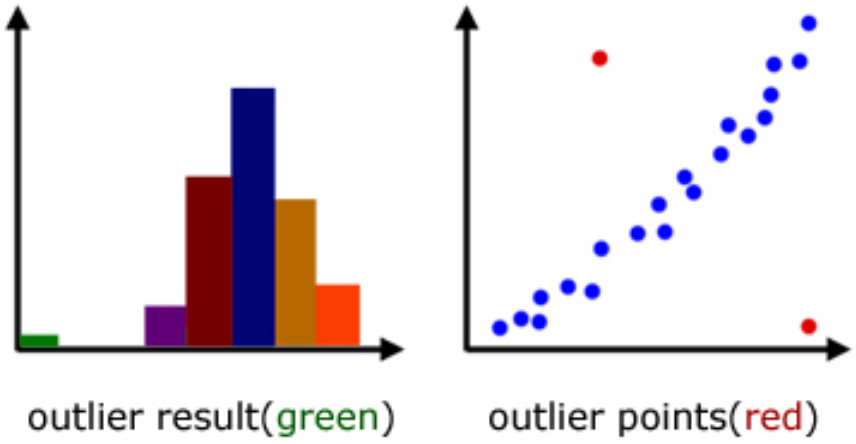
Ranking of Affect by Outliers?

Measure of Dispersion

- Range
- Standard Deviation
- Coefficient of Variation
- Semi-interquartile Range

Most to Least

- Range — susceptible
- Variance
  - Standard Deviation
  - Coefficient of Variation
- SIQR — resistant



http://www.a-levelmathstutor.com/images/statistics/outliers-graph01.jpg

Only for quantitative data!
categorical can't quantify spread since no 'distance'
Instead, give categories for given percentile of samples
e.g., "90% of samples are in 3 categories" (Pareto chart)

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 Champion in LoL: events {Shen, Teemo, Leona, …} (all possible Champions listed)

What Numeric Values do Probabilities take?
(Hint: we had two rules)

- Probabilities must be between 0 and 1 (but often written/said as percent)
- Probabilities of set of *exhaustive*, *mutually exclusive* events must add up to 1

Probability

- Draw 1 card. What is the probability drawing a Jack?

P(J)
= 2 favorable outcomes
/
5 total outcomes
= 2/5

- ○ 1/5
- ○ 2/5
- ○ 3/5
- ○ 50/50
- ○ I don't know

Poll 1!

https://web.cs.wpi.edu/~imgd2905/d23/polls.html

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

- ○ 2/5
- ○ 4/25
- ○ 1/10
- ○ 4/5
- ○ I don't know

Poll 2!

https://web.cs.wpi.edu/~imgd2905/d23/polls.html

Probability

- Draw 3 cards *simultaneously*. What is the probability of not drawing at least 1 King?

$P(K') \times P(K' \mid K') \times P(K' \mid K'K')$
$= 3/5 \times 2/4 \times 1/3$
$= 6/60$
$= 1/10$

- ○ 3/5
- ○ 8/125
- ○ 0.01
- ○ 1/10
- ○ I don't know

Poll 3!

https://web.cs.wpi.edu/~imgd2905/d23/polls.html

Probability

- Draw 1 card. What is the probability of drawing a King or a Queen?

P(K or Q)
= P(K) + P(Q)
= 2/5 + 1/5
= 3/5

- ○ 1/5
- ○ 2/5
- ○ 3/5
- ○ 4/5
- ○ I don't know

Poll 4!

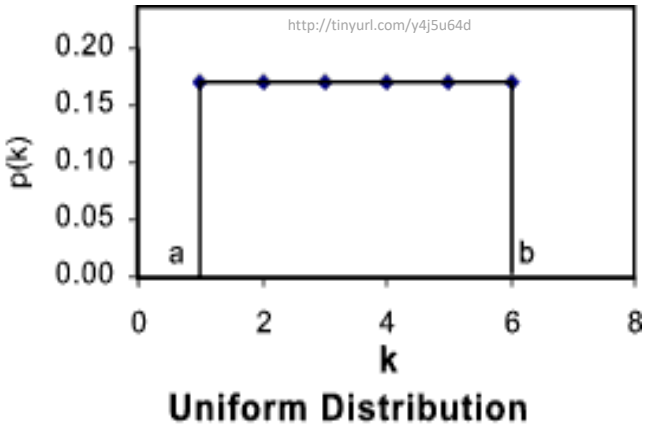
https://web.cs.wpi.edu/~imgd2905/d23/polls.html

What Kind of Probability Distribution is:

- Rolling one 6-sided dice (d6)? Can you draw it?

What Kind of Probability Distribution is:

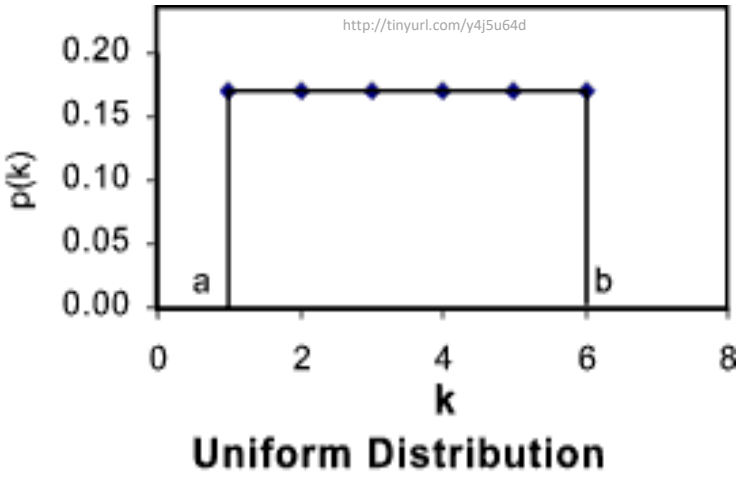
- Rolling one 6-sided dice (d6)? Can you draw it?
  - Uniform (or “square”)



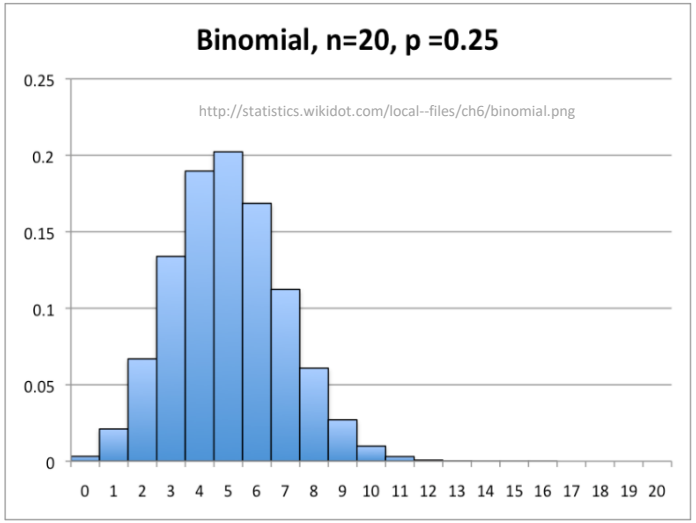
- Number of 1s when rolling 20 4-sided dice (d4)? Can you draw it?

What Kind of Probability Distribution is:

- Rolling one 6-sided dice (d6)? Can you draw it?
  - Uniform (or “square”)



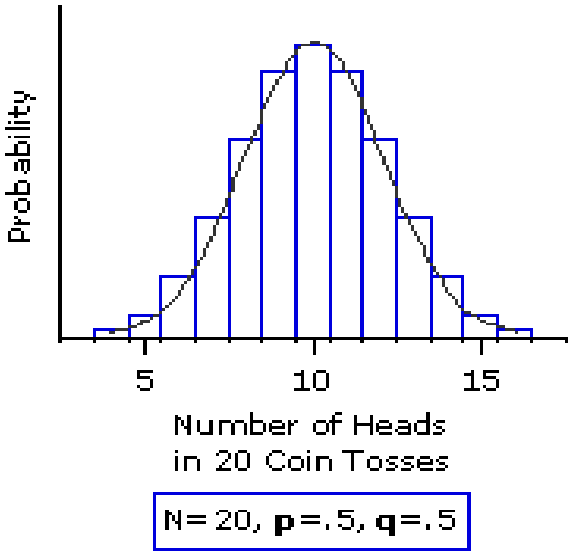
- Number of 1s when rolling 20 4-sided dice (d4)? Can you draw it?
  - Binomial



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

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

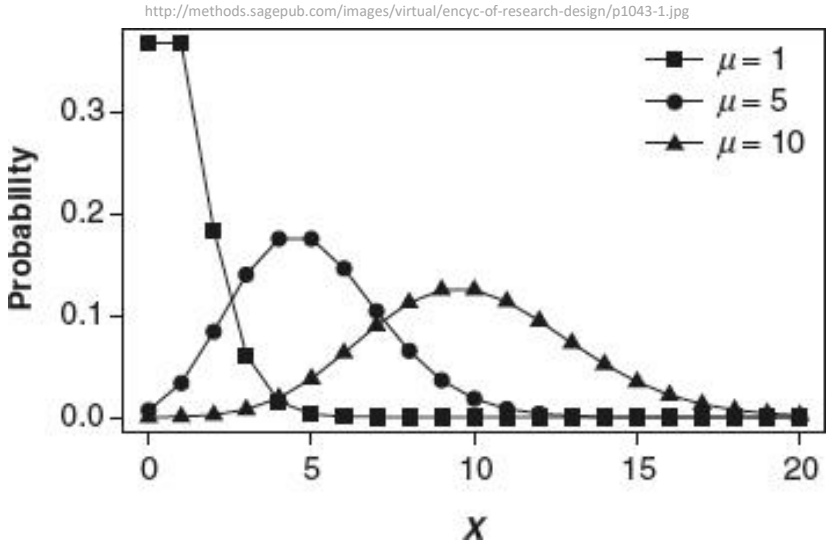


http://www.vassarstats.net/textbook/f0603.gif

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

http://methods.sagepub.com/images/virtual/encyc-of-research-design/p1043-1.jpg



Phrase people use is "random arrivals"

Expected Value

What is the formula for expected value?

$\mu_x = E(X) = ???$

Expected Value

What is the formula for expected value?

$$\mu_x = E(X) = x_1P(x_1) + x_2P(x_2) + \ldots + x_nP(x_n)$$

Expected Value

What is the formula for expected value?

Toss: Flip 2 coins
Each Head gives 1 point
2 Tails → bust, turn over

$$\mu_x = E(X) = x_1P(x_1) + x_2P(x_2) + \ldots + x_nP(x_n)$$

Expected Value

What is the formula for expected value?

Toss: Flip 2 coins
Each Head gives 1 point
2 Tails → bust, turn over

$$\mu_x = E(X) = x_1P(x_1) + x_2P(x_2) + \ldots + x_nP(x_n)$$

What is the expected value after 1 toss?

Poll 1!

Expected Value

What is the formula for expected value?

Toss: Flip 2 coins
Each Head gives 1 point
2 Tails → bust, turn over

$$\mu_x = E(X) = x_1P(x_1) + x_2P(x_2) + \ldots + x_nP(x_n)$$

What is the expected value after 1 toss?

$$E(X) = 0*P(TT) + 1*P(HT) + 1*P(TH) + 2*P(HH)$$
$$= 0 + 1/4 + 1/4 + 2/4$$
$$= 4/4$$
$$= 1$$

Expected Value

What is the formula for expected value?

Toss: Flip 2 coins
Each Head gives 1 point
2 Tails → bust, turn over

$$\mu_x = E(X) = x_1P(x_1) + x_2P(x_2) + \ldots + x_nP(x_n)$$

What is the expected value after 1 toss?

$E(X) = 0*P(TT) + 1*P(HT) + 1*P(TH) + 2*P(HH)$
$= 0 + 1/4 + 1/4 + 2/4$
$= 4/4$
$= 1$

$E(X) = x_1P(bust) + x_2P(bust')$

After 2 tosses?

Expected Value

What is average if don't bust?

Poll 2!

Toss: Flip 2 coins
Each Head gives 1 point
2 Tails → bust, turn over

Expected Value

What is average if don't bust?

A = HT + TH + HH = (1 + 1 + 2) / 3 = 4/3

What is the expected value after 1 toss?

Poll 3!

Toss: Flip 2 coins
Each Head gives 1 point
2 Tails → bust, turn over

Expected Value

What is average if don't bust?

A = HT + TH + HH = (1 + 1 + 2) / 3 = 4/3

What is the expected value after 1 toss?

E(X) = P(TT) * 0 + (1- P(TT)) * 4/3

= ¾ * 4/3

= 1

Toss: Flip 2 coins
Each Head gives 1 point
2 Tails → bust, turn over

Expected Value

What is average if don't bust?

A = HT + TH + HH = (1 + 1 + 2) / 3 = 4/3

What is the expected value after 1 toss?

E(X) = P(TT) * 0 + (1- P(TT)) * 4/3

= ¾ * 4/3

= 1

2 tosses?

Poll 4!

Toss: Flip 2 coins
Each Head gives 1 point
2 Tails → bust, turn over

Expected Value

What is average if don't bust?

A = HT + TH + HH = (1 + 1 + 2) / 3 = 4/3

What is the expected value after 1 toss?

E(X) = P(TT) * 0 + (1- P(TT)) * 4/3

= ¾ * 4/3

= 1

2 tosses?

E(X) = $(1\text{-}P(TT))^2$ * (4/3 * 2)

= ¾ * ¾ * 8/3

= 1.5

3 tosses?

Toss: Flip 2 coins
Each Head gives 1 point
2 Tails → bust, turn over

Expected Value

What is average if don't bust?

A = HT + TH + HH = (1 + 1 + 2) / 3 = 4/3

What is the expected value after 1 toss?

E(X) = P(TT) * 0 + (1- P(TT)) * 4/3

= ¾ * 4/3

= 1

2 tosses?

E(X) = $(1\text{-}P(TT))^2$ * (4/3 * 2)

= ¾ * ¾ * 8/3

= 1.5

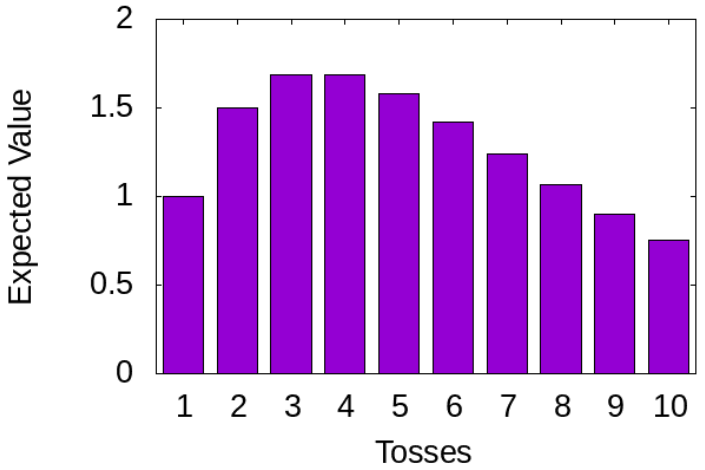
3 tosses?

E(X) = $(1\text{-}P(TT))^3$ * (4/3 * 3)

= ¾ * ¾ * ¾ * 12/3

= 1.6875

Toss: Flip 2 coins
Each Head gives 1 point
2 Tails → bust, turn over

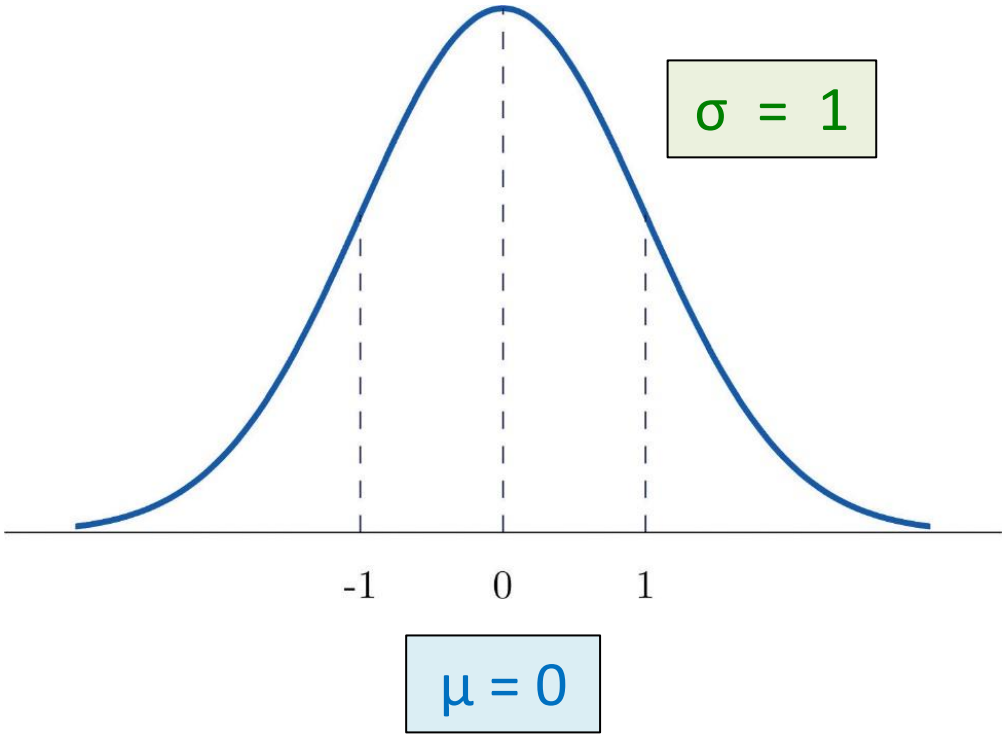


BEST_BOT?

What is the Standard Normal Distribution?

What is the Standard Normal Distribution?

- Normal distribution
- Mean $\mu = 0$
- Std dev $\sigma = 1$



What is the Central Limit Theorem?

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

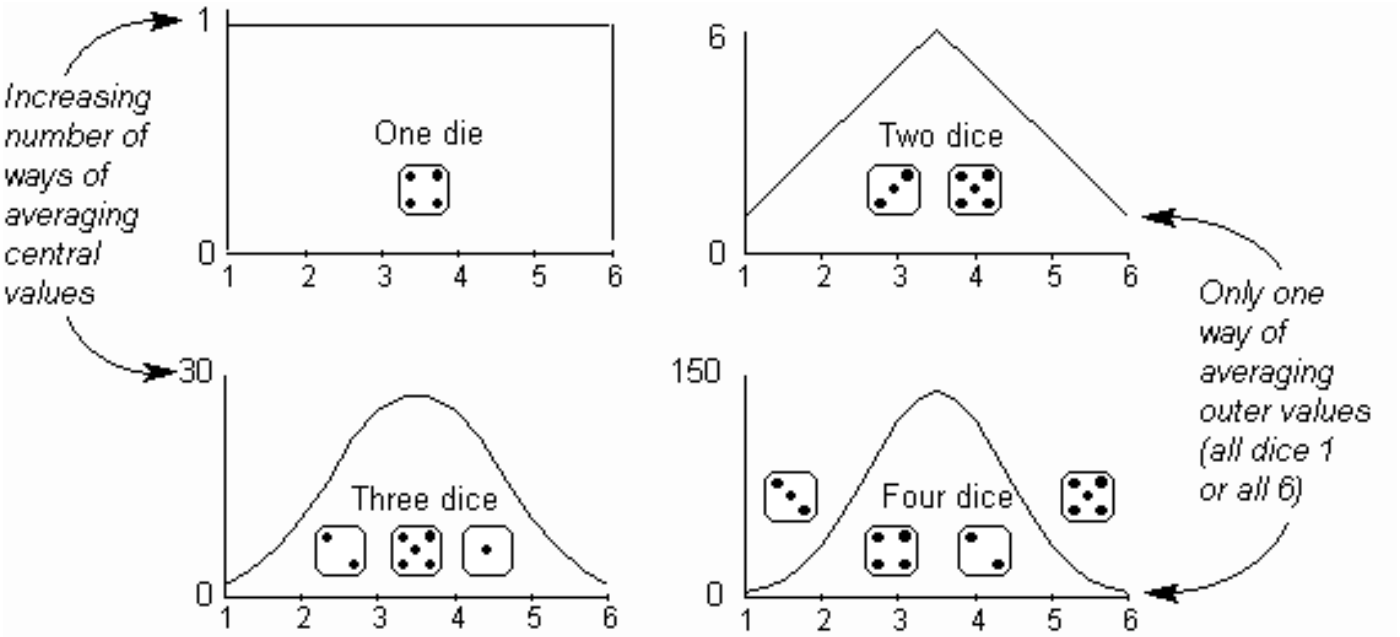
→ What is Distribution shape?

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 many is "large enough"?



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

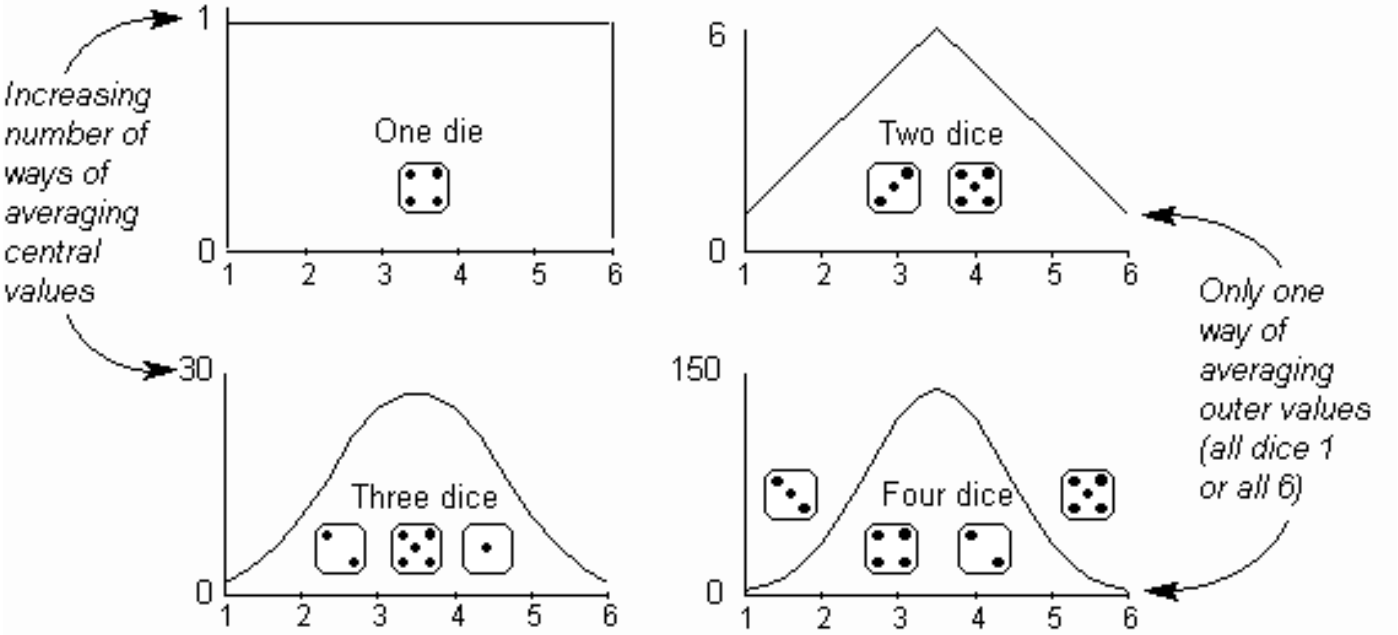
What is the Central Limit Theorem?

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

→ Distributed Normally

How many is "large enough"?

- 30
- (15)



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

Does underlying distribution matter?

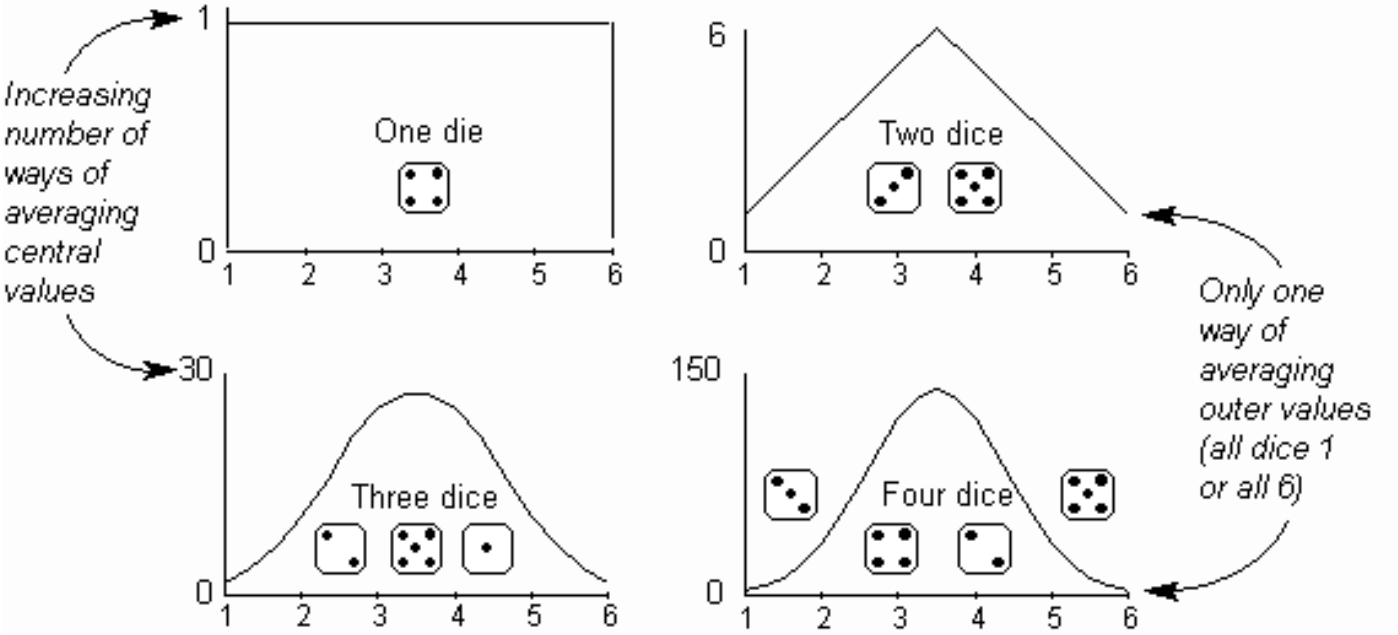
What is the Central Limit Theorem?

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

→ Distributed Normally

How many is "large enough"?

- 30
- (15)



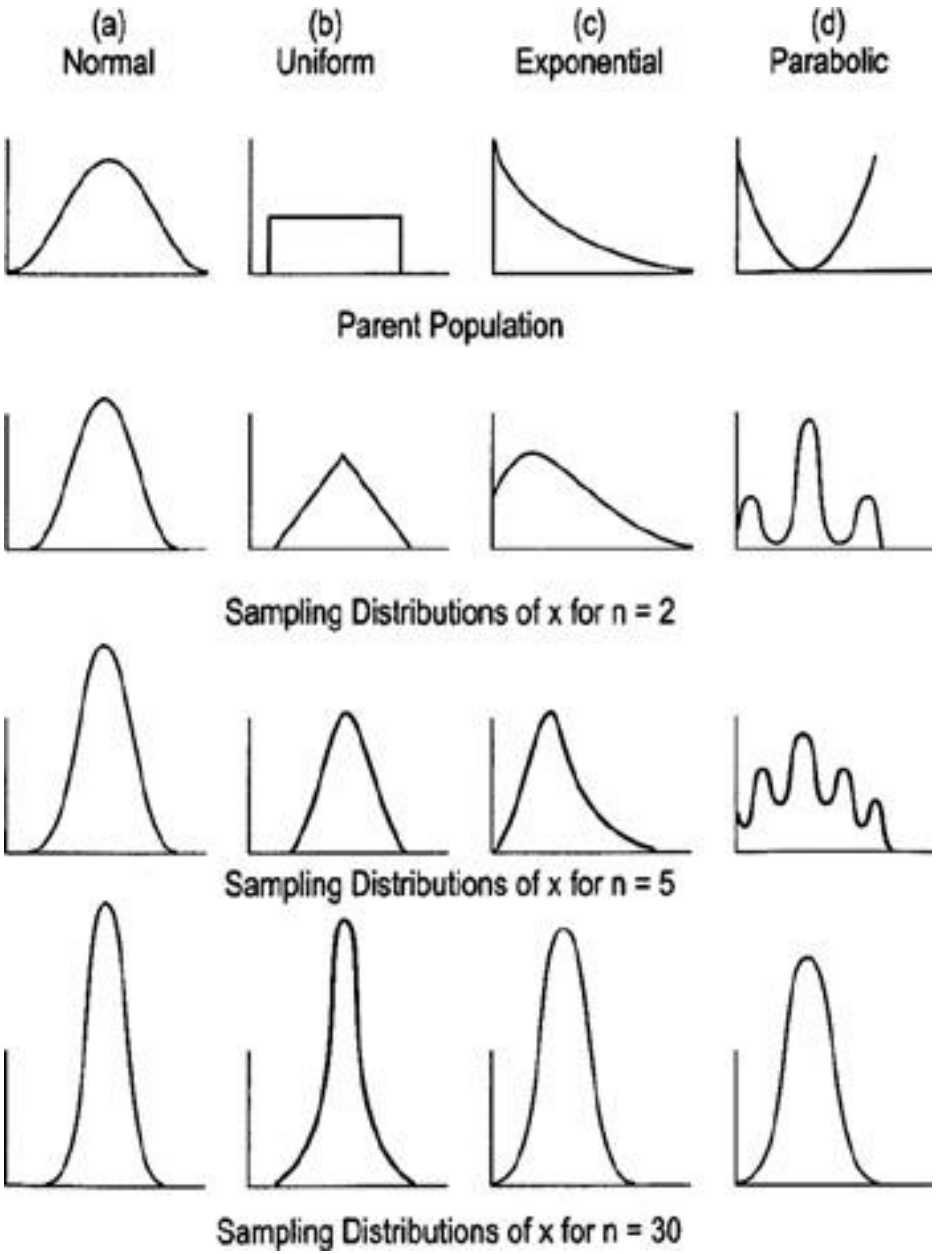
http://home.ubalt.edu/ntsbarsh/Dice_001.gif

Does underlying distribution matter?

- No

(see next slide)

Underlying Distribution does not Matter



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

Why do we care?

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

Sampling Error

- What is sampling error?

Sampling Error

- What is sampling error?
  - Error from estimating population parameters from sample statistics

- Size of error is based on what two main factors?

Sampling Error

- What is sampling error?
  - Error from estimating population parameters from sample statistics

- Size of error is based on what two main factors?
  - Population variance (e.g., σ)
  - Sample size (N)

Statistic versus Sample Size (N)

- Suppose wanted to know likelihood that WPI student played *Hearthstone*
  - Ask N people, count "yes" and divide by N
- Ask 1 person?
- Ask 2 people?
- Ask 100 people?
- What does graph

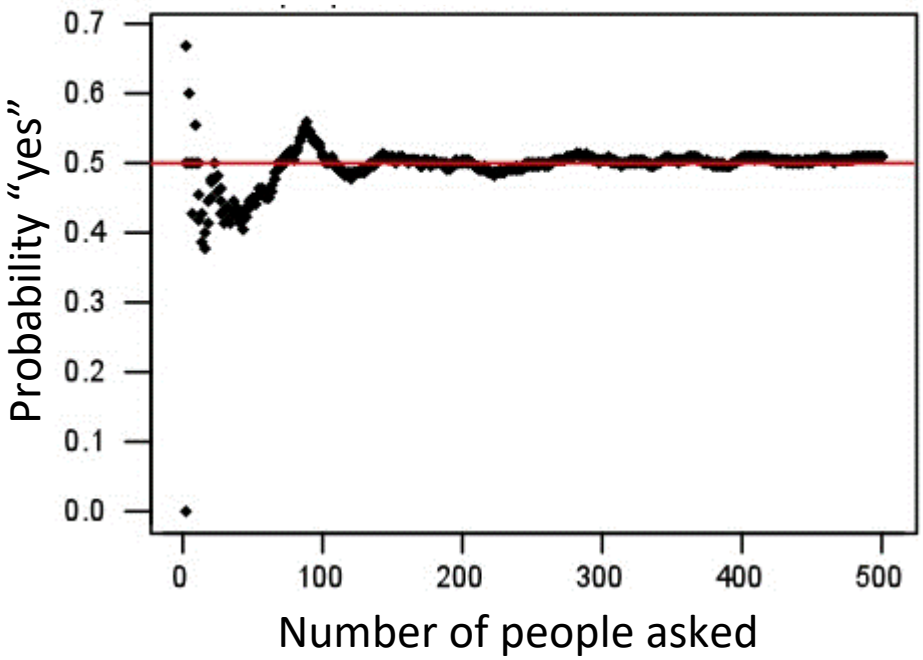
of "yes" probability

versus N people look

like?

Statistic versus Sample Size (N)

- Suppose wanted to know likelihood that WPI student played *Hearthstone*
  - 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?



Groupwork

https://web.cs.wpi.edu/~imgd2905/d23/groupwork/8-review/handout.html

Confidence Intervals

- What is a confidence interval? Give an example

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?

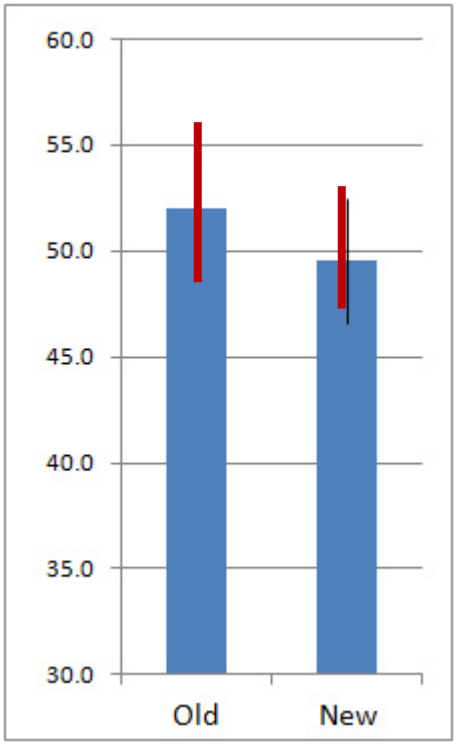
Confidence Intervals

- 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)

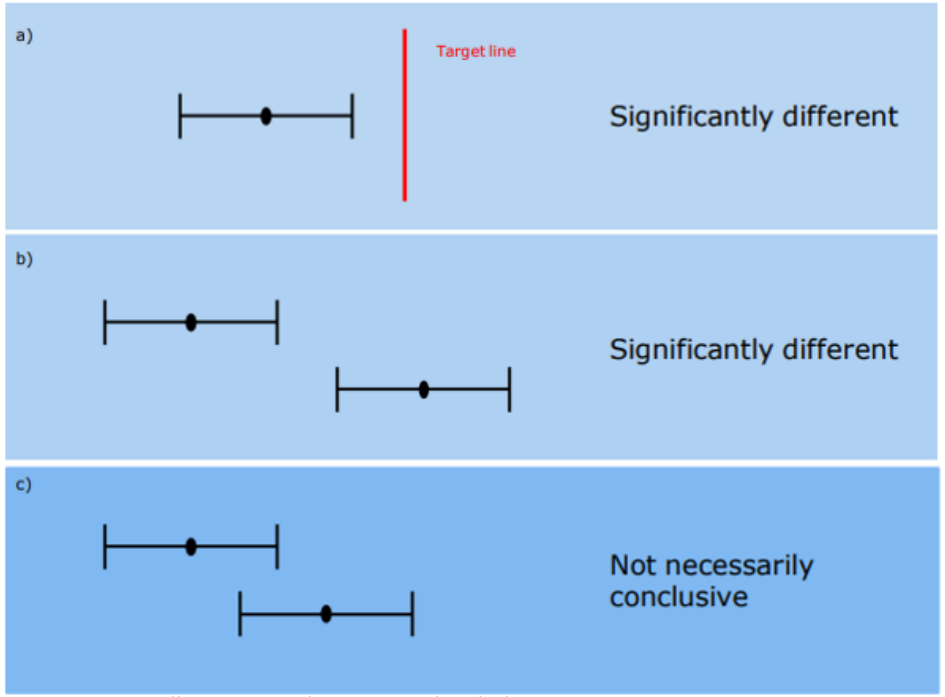
$$\overline{X} \pm t \frac{s}{\sqrt{n}}$$

Interpreting Confidence Intervals

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



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



https://emedia1.nhs.wales/PublicHealthWales/cache/file/DF7D68CA-6C7F-42BA-87FFCBD9FF0668BB.png

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

Hypothesis Testing

- Studio has new model for Hero
- Want to see if played more often
- Steps?

Hypothesis Testing

- Studio has new model for Hero
- Want to see if played more often
- Steps?

1. Set hypotheses, pick α, decide N
2. Gather data
3. Compute sample mean
4. Test (compute p value)
5. Analyze results to accept or reject

Hypothesis Testing

- What is the Null Hypothesis?

- What is the Alternate Hypothesis?

Hypothesis Testing

- What is the Null Hypothesis?
  - The measured statistic is the same as the population parameter (e.g., $\bar{x} = \mu$)
- What is the Alternate Hypothesis?
  - Contrary to null hypothesis (e.g., there *is* a difference in the two ($\bar{x}$!= μ))
- Which do we test and *why*?

Hypothesis Testing

- What is the Null Hypothesis?
  - The measured statistic is the same as the population parameter (e.g., $\bar{x} = \mu$)
- What is the Alternate Hypothesis?
  - Contrary to null hypothesis (e.g., there *is* a difference in the two ($\bar{x}$!= μ))
- Which do we test and *why*?
  - Test Null
  - Data can only reject hypothesis, not prove

  → Reject Null

Hypothesis Testing

- Gathered “new” 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?

Hypothesis Testing

- Gathered “new” 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?
  - CANNOT reject H_0
- What does that mean?

Hypothesis Testing

- Gathered “new” 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 “new” mean and population mean (at 0.01 significance)

Regression

- What is the purpose 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

$$Y = mX + b$$

Regression

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

  value = 32670 + 35.04 *x* (square feet)

- How do you *interpret* the model? How can you use it?

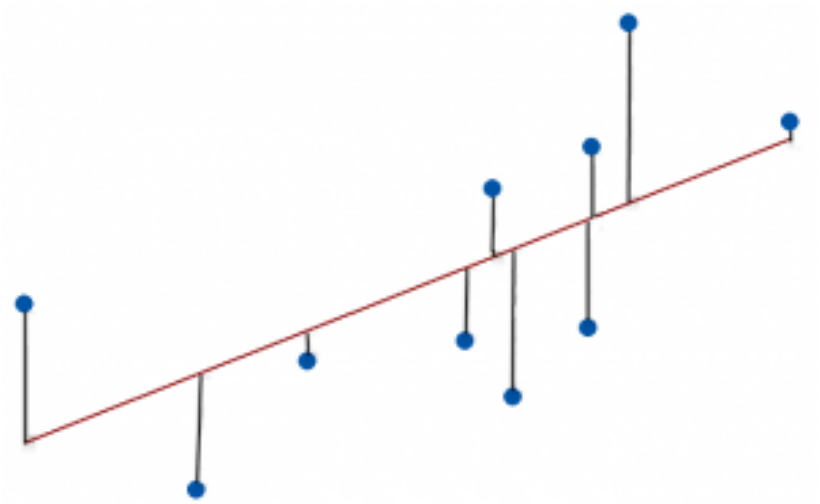
  1. Intercept is 32670. So, base house value is $33k.
  2. Slope is 35.04. So, every square foot increases house value by $35
  3. Given square feet, predict value: 1800 sq feet

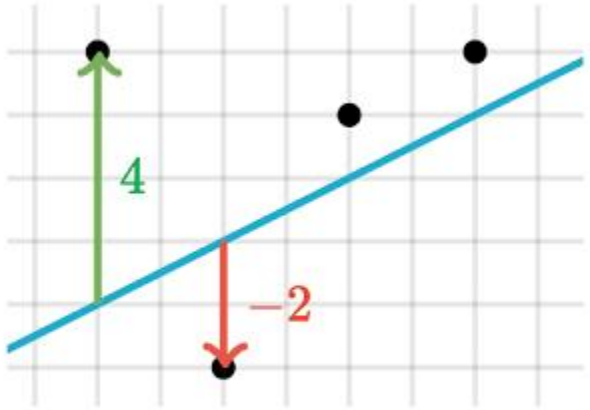
  value = 32670 + 35.04 x (1800) = $95,742

What are Residuals?

What are Residuals?

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





What is Residual Analysis?

What is Residual Analysis?

https://www.qualtrics.com/support/stats-iq/analyses/regression-guides/interpreting-residual-plots-improve-regression/

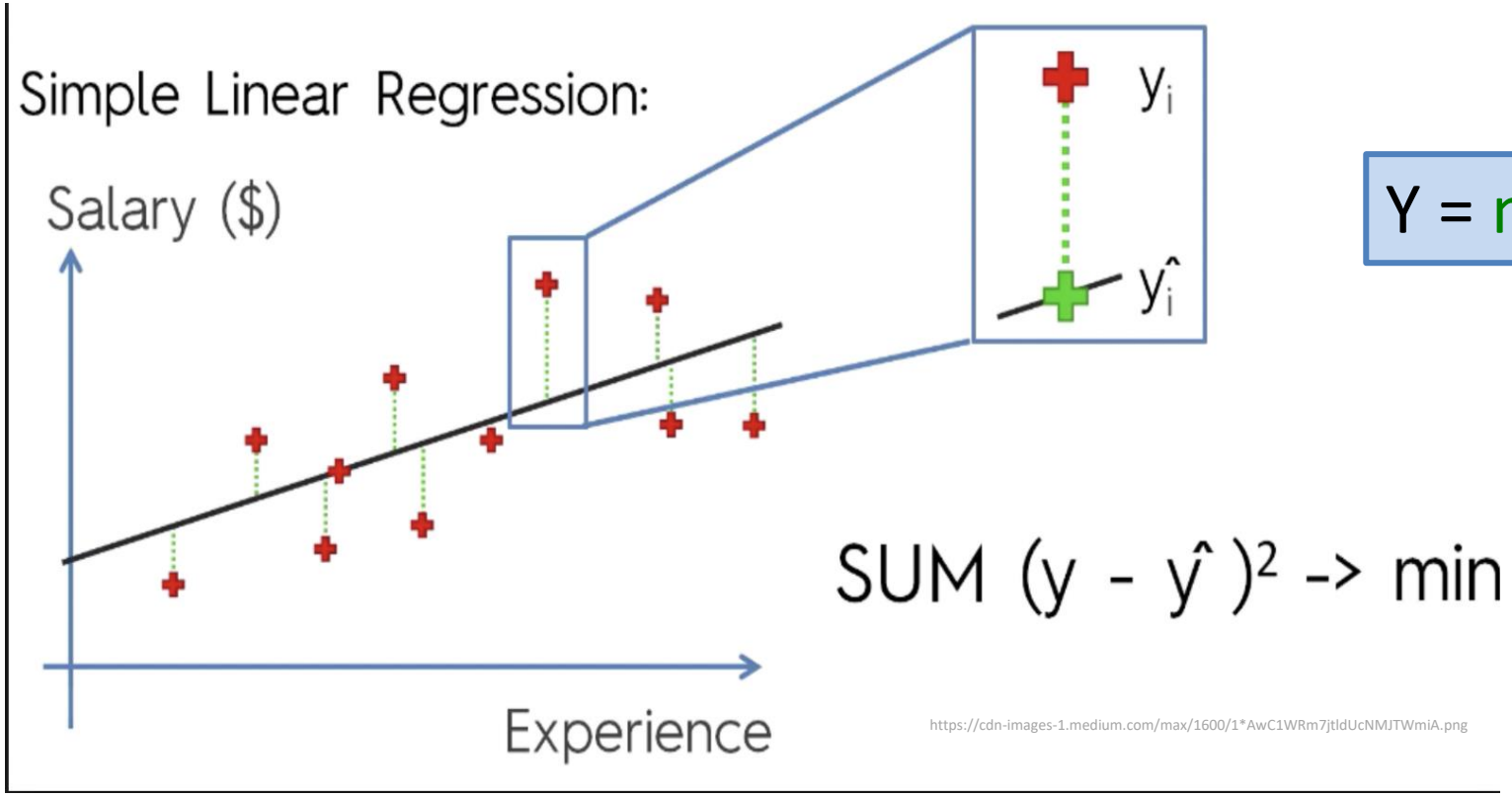
Note that we've colored in a few dots in orange so you can get the sense of how this transformation works.

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

What is a Least Squares Line?

What is a Least Squares Line?

- Line that minimizes sum squared error



$Y = mX + b$

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

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

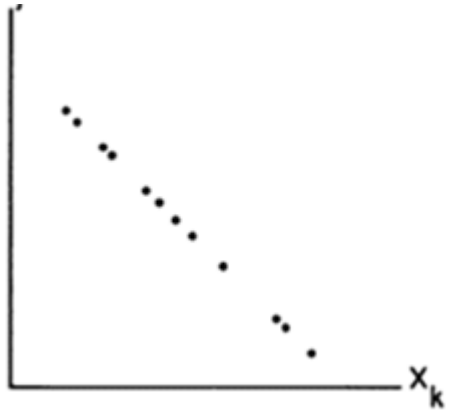
- Proportion of variance in the dependent variable predictable by the independent variable
- Fraction (percentage) of variance explainable by model

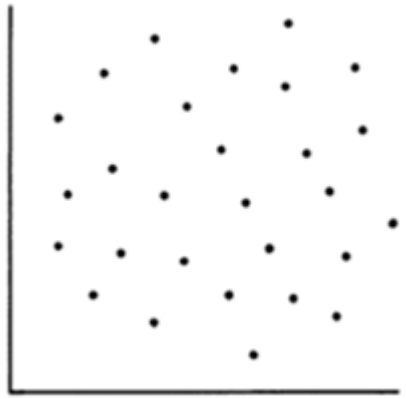
$$R^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SST}$$

What is the value of R^2? of R?

$R^2 = 1$

$R = -1$



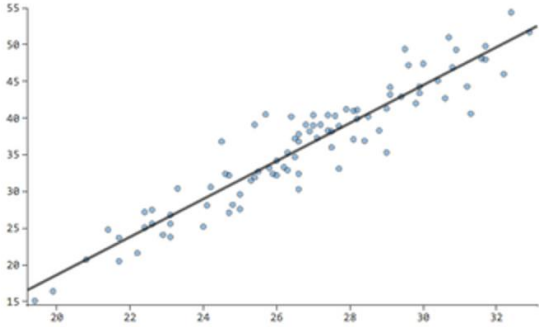


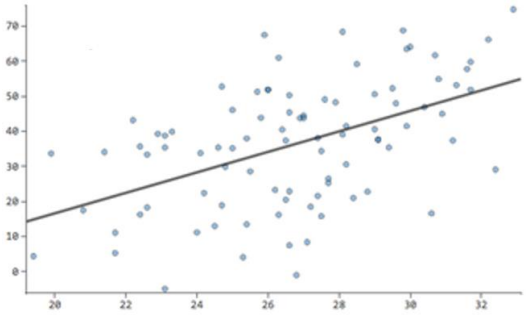
$R^2 = 0$

$R = 0$

$R^2 = 0.8$

$R = 0.9$



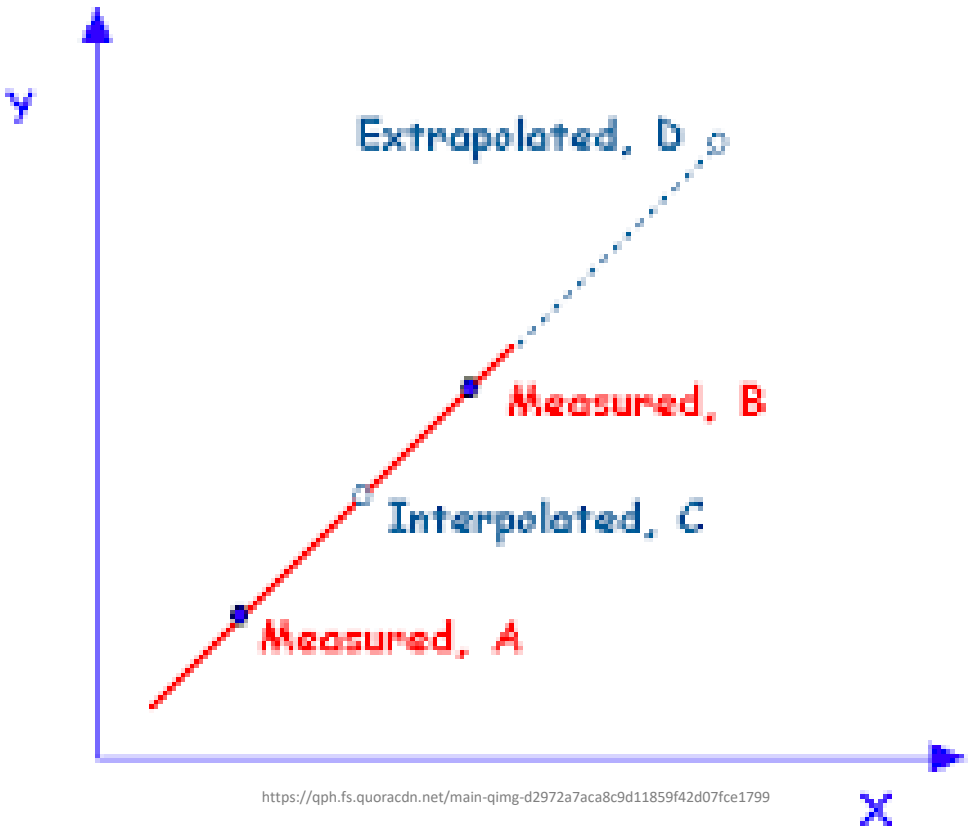


$R^2 = 0.2$

$R = 0.4$

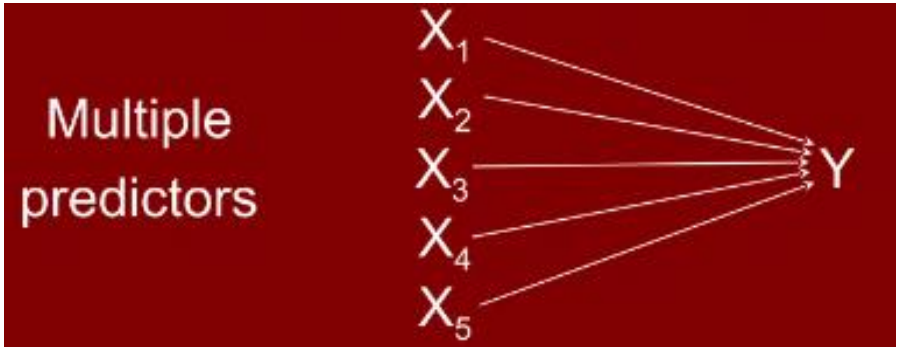
What is Interpolation? Extrapolation?

- Prediction
  - Interpolation – within measured X-range
  - Extrapolation – outside measured X-range



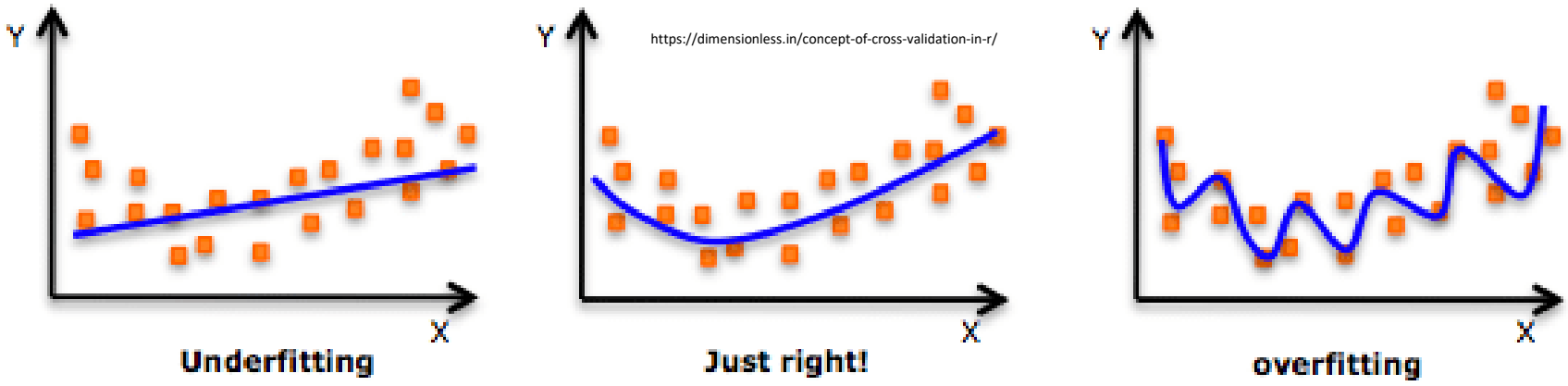
What is Multiple Linear Regression?

- Use several independent variables to predict dependent variable



$$Y = b_0 + b_1X_1 + b_2X_2 + b_3X_3 .. b_nX_n$$

In modeling, what is an overfit? Underfit?

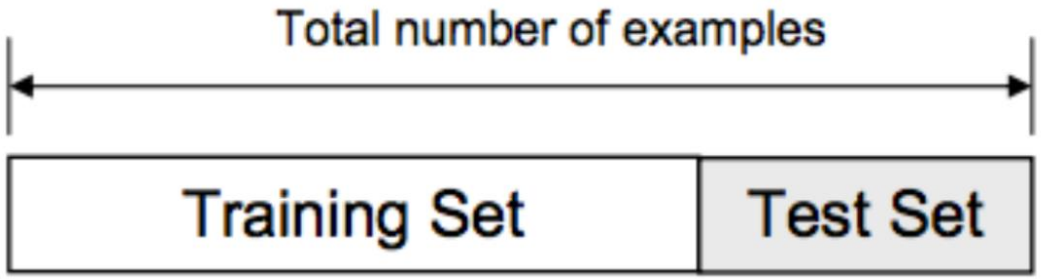


An example of overfitting, underfitting and a model that's "just right!"

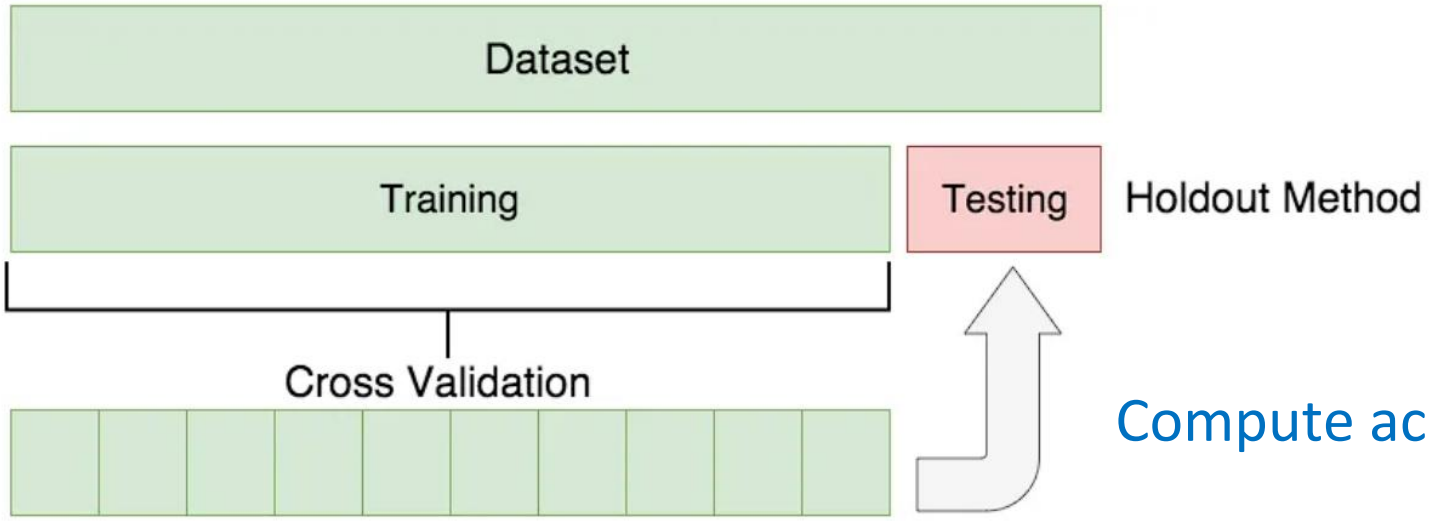
- Overfit – model fits the observed data too well, failing to generalize to unseen data
- Underfit – model is too simple to capture underlying complexity

Test → Cross Validation

Cross Validation (1 of 2)

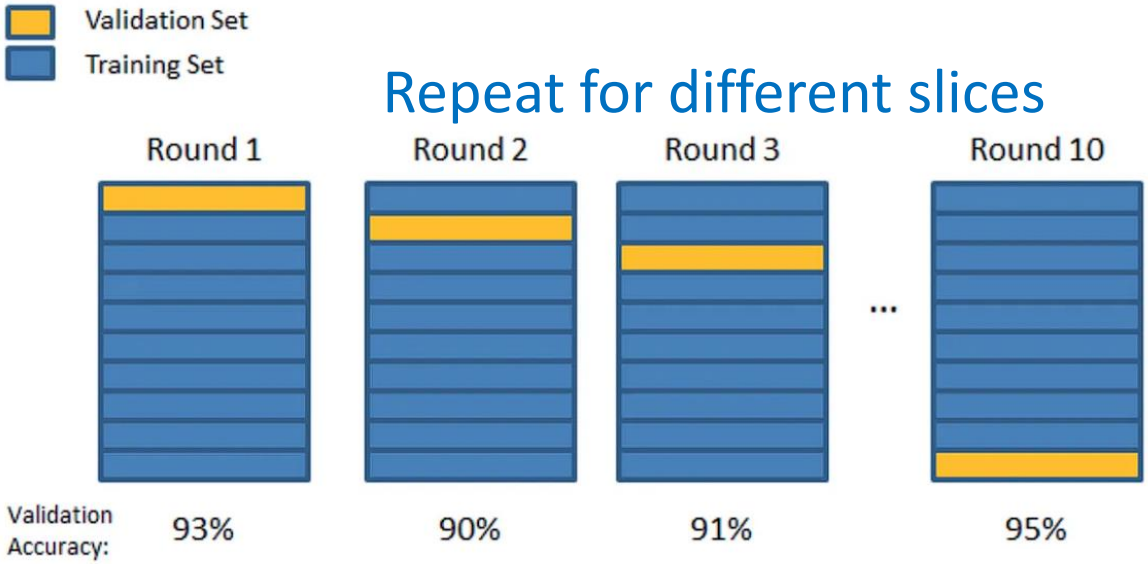


Use to build model



Compute accuracy

Cross Validation (2 of 2)



- Overfit and Underfit will both have lower accuracy than “just right”