Review

IMGD 2905

What are two main sources for data for game analytics?

What are two main sources for data for game analytics?

- Quantitative – instrumented game
- Qualitative – subjective evaluation

What steps are in the game analytics pipeline?

What steps are in the game analytics pipeline?

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

What is population versus sample?

What is population versus sample?

- Population – all members of group pertaining to study

- Sample – part of population selected for analysis
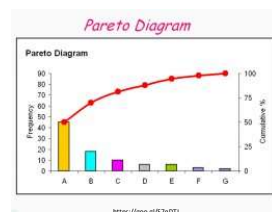
What is probability sampling?

What is probability sampling?

- Probability sampling - sampling considering likelihood of selection
  - Consider likelihood as part of population

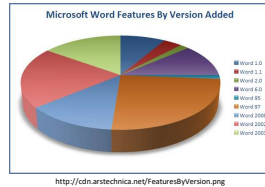## What is a Pareto chart? When used?

What is a Pareto chart? When used?

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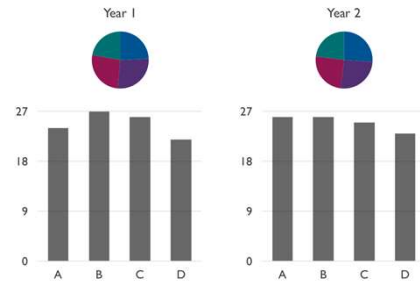
- Helps identify most common, relative amounts

Pareto Diagram

https://goo.gl/57qDTJ

When should you *not* use pie chart?

When should you *not* use pie chart?

- When too many slices

Microsoft Word Features By Version Added

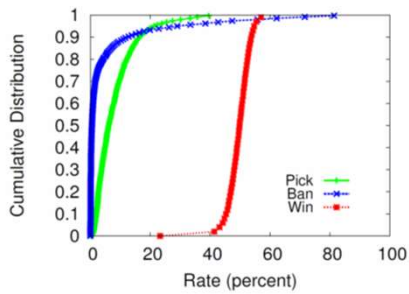http://cdn.arstechnica.net/FeaturesByVersion.png

When should you *not* use pie chart?

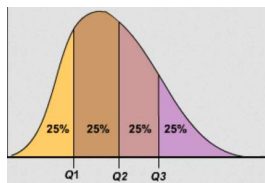- (Often) when comparing pies

Which Measure of Central Tendency to Use? Why?

What are Quartiles?

What are Quartiles?

Three values that divide population into four equal groups

25% 25% 25% 25%

Q1 Q2 Q3

First Quartile Median Second Quartile Third Quartile

First Quarter Second Quarter Third Quarter Fourth Quarter

24, 25, 26, 27, 30, 32, 40, 44, 50, 52, 55, 57

Q₁ 26½ Q₂ 36 Q₃ 51

MathBits.com

Describe how to Compute Variance

Describe how to Compute Variance

1. Compute mean
2. Compute how far each sample value is from mean. Square this.
3. Add these up.
4. Divide by number of samples.

Describe what Standard Deviation is in Words

## Describe what Standard Deviation is in Words

- "The 'average' of how far each sample point is from the mean"

Empirical Rule

- 1000 data points
- Mean of 50
- Standard deviation of 10

- Between 40-60?
- How many points are between 40-60?
- How many points are between 20-80?

Empirical Rule

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

- 1000 data points
- Mean of 50
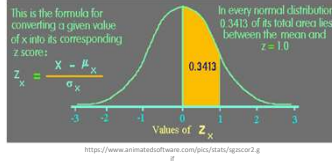- Standard deviation of 10
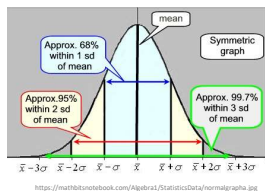
- Between 40-60?
  – About 950 (95%)
- How many points are between 40-60?
  – About 700 (68%)
- How many points are between 20-80?
  – Nearly all (99.7%), so only about 3 outside

Z-Score

- 1000 data points
- Mean of 50
- Standard deviation of 10

- My data point is a 75. What is it's Z-score?

- Your data point is a 10. What is it's Z-score?

Z-Score

- 1000 data points
- Mean of 50
- Standard deviation of 10

- My data point is a 75. What is it's Z-score?
  (75 - 50) / 10 = 2.5
- Your data point is a 10. What is it's Z-score?
  (10 - 50) / 10 = -4.0

Rank the Following High to Low in Susceptibility to Outliers

Measure of Variation   **Most to Least**
- Semi-interquartile Range
- Range
- Coefficient of Variation

Rank the Following High to Low in Susceptibility to Outliers

Measure of Variation   **Most to Least**
- Semi-interquartile Range   - Range
- Range   - Coefficient of Variation
- Coefficient of Variation   - Semi-interquartile Range

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

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., picking champion in LoL: events {Darius, Leona, Fizz, …} (all possible Champions listed)

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

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

- Classical (theory)
  - e.g., equal likelihood d6, so P(1) = 1/6th
- 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

Probability

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

Probability

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

P(2J) = P(J) x P(J | J)
= 2/5 x 1/4
= 1/10

Probability

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

Probability

- Draw 3 cards. 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

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
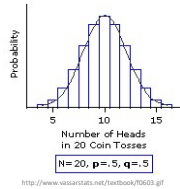
Number of Heads in 20 Coin Tosses
N=20, p=.5, q=.5
http://www.vassarstats.net/textbook/f0603.gif

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

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
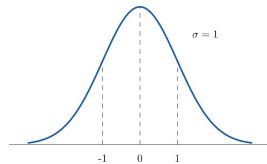4. Events occur singly (not simultaneously)
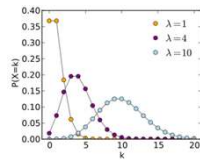
Phrase people use is "random arrivals"

What is the Standard Normal Distribution?

What is the Standard Normal Distribution?

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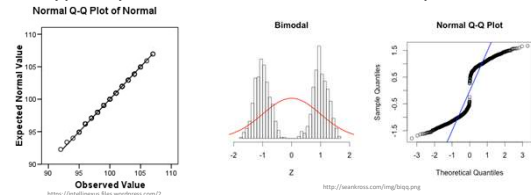- Std dev σ = 1

What is the Probability Distribution for number of heads?

- For flipping one coin?
  - Square
- For flipping two coins?

What is a Quantile-Quantile Plot?

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
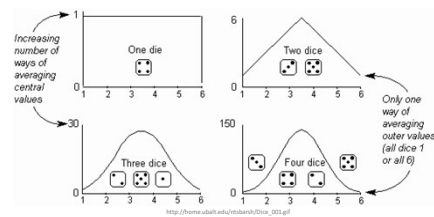
What is the Central Limit Theorem?

- Given population
- If take a large enough sample size
- What does probability of sample means look like?
→ What is the Distribution shape?

What is the Central Limit Theorem?

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

How big is "enough"?

What is the Central Limit Theorem?

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

How big is "enough"?
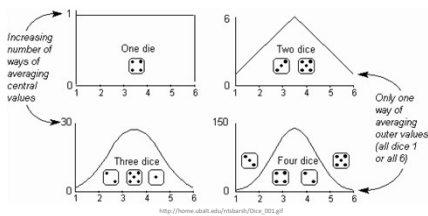- 30
- (15)

Does underlying distribution matter?

What is the Central Limit Theorem?

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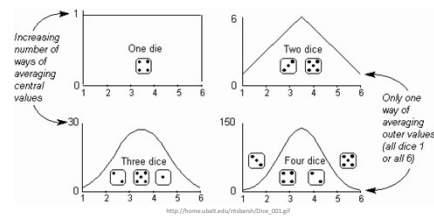- What does probability of sample means look like?
→ Distributed Normally

How big is "enough"?
- 30
- (15)

Does underlying distribution matter?
- No

(see next slide)

Underlying Distribution does not Matter

Why do we care?
→ Can apply rules (e.g., empirical rule) to Normal Distributions!

Sampling Error

- What is the sampling error?

Sampling Error

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

- The *size* of the error is based on what two main factors?

Sampling Error

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

- The *size* of the error is based on what two main factors?
  - Population variance
  - Sample size (N)

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?

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?

Confidence Intervals

- What is a confidence interval? Give an example
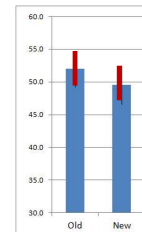
Confidence Intervals

- What is a confidence interval? Give an example
  - Range of values with specific certainty that population parameter is within
  - 95% confidence interval for 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 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 (number of items in sample) (standard deviation)

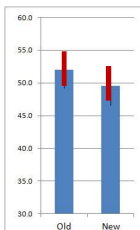$$\overline{X} \pm t \frac{s}{\sqrt{n}}$$

Interpreting Confidence Intervals

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

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