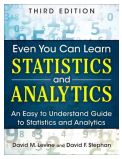
IMGD 2905

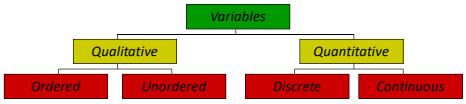
Presenting Data

Chapter 2



Types of Variables

- Qualitative (Categorical) variables
  - Can have states or subclasses
    - e.g., rank: [platinum, diamond, gold]
  - Can be ordered or unordered
    - e.g., bronze, silver, gold → ordered
    - e.g., support, tank, jungler → unordered
- Quantitative (Numeric) variables
  - Numeric levels
  - Discrete or continuous
    - e.g., gold per minute, deaths, character level
    - e.g., kills + assists / deaths ratio, win percentage



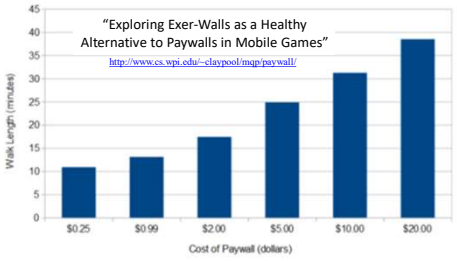
2

Outline

- Types of Charts (next)
- Guidelines for Charts
- Common Mistakes

Categorical: Bar Chart

- Chart containing rectangles ("bars") where length represents count, amount, or percent
- Better than table for comparing numbers

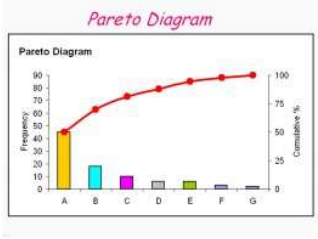


Note: bars could be sideways, too

Demo: imgdpops.xlsx

Categorical: Pareto Chart

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

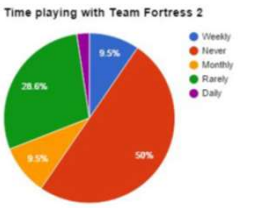


Sort.
New column for percent [=B2/SUM(B$2:B$12)]
New column for running [=SUM(D$2:D2)]
Note: $ "locks" value in (e.g., B$12 versus B12)
Insert combo plot

Demo: imgdpops.xlsx

Categorical: Pie Chart

- Wedge-shaped areas ("pie slices") – represent count, amount or percent of each category from whole
- Best if few slices since quantifying "size" of pie difficult
- Comparing pies also difficult



"The Effects of Latency and Jitter on a First Person Shooter: Team Fortress 2"
http://www.cs.wpi.edu/~claypool/iqp/tf2/

Demo: imgdpops.xlsx

Categorical: Cross-Classification Table

- Multi-column table that presents count or percent for 2+ categorical variables
  - Good for comparison across multi-categorical data

Class rank * Do you live on campus? Crosstabulation

Count

		Do you live on campus?		
		Off-campus	On-campus	Total
Class rank	Freshman	37	100	137
	Sophomore	42	48	90
	Junior	90	8	98
	Senior	62	1	63
Total		231	157	388

Demo: grades.xlsx

Insert Pivot Chart
Select Major through Grade
Drag Majors to Axis
Drag Grade to Axis
Drag Grade to Values

Numeric: Frequency Distribution

Poll class!

- Groups of numeric values and frequency
- e.g., Survey of Champion "skins" bought with RP
  - 1, 2, 1, 0, 3, 4, 0, 1, 1, 1, 2, 2, 3, 2, 3, 2, 1, 4, 0, 0
  - Cluster into groups
  - Report frequency per group

Skins	Freq.	Percent
0	4	20%
1	6	30%
2	5	25%
3	3	15%
4	2	10%

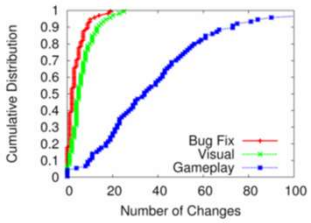
- May include percentage
- Typically equal size
  - Sometimes ends are open (for extremes)
- Bin size/number variable
  - Too many and not readable
  - Guide:
    - 100 or less 7-10
    - 101-200 11-15
    - 200+ 13-20

Cumulative Distribution

- Cumulative amount of data with value or less
- Easy to see min, max, median
- Compare shapes of distributions

Demo: lol-patches.xlsx

Select Banrate data
Sort low to high
New column for percent [=ROW()/42]
Select column → paste down all
Select both columns
Insert → Scatter plot with lines

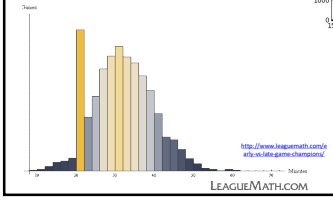


"Nerfs, Buffs and Bugs - Analysis of the Impact of Patching on League of Legends"
http://www.cs.wpi.edu/~claypool/papers/lol-crawler/

Histogram

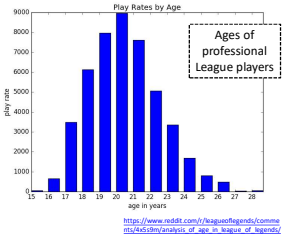
- Bar chart for grouped numerical data
  - No (or small) gaps btwn adjacent bars

https://www.mathsisfun.com/data/images/bar-chart-vs-histogram.gif



http://www.leaguemath.com/early-vs-late-game-champions/

LEAGUEMATH.COM



https://www.reddit.com/r/leagueoflegends/comments/4x5c9m/analysis_of_age_in_league_of_legends/

Demo: grades.xlsx

Select GPA data
Insert → Statistics Chart → Histogram
Can adjust bins, overflow/underflow

Stem and Leaf Display

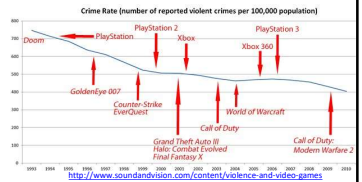
- "Histogram-lite" for analysis w/out software
  - e.g., exam scores: 34, 81, 75, 51, 82, 96, 55, 66, 95, 87, 82, 88, 99, 50, 85, 72

9| 6 5 9
8| 1 2 7 2 8 5
7| 5 2
6| 6
5| 1 5 0
4|
3| 4

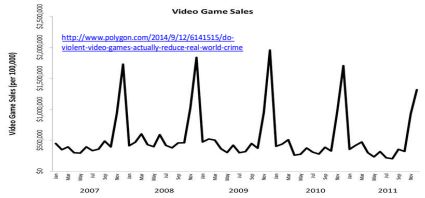
11

Time Series Plot

- Associate data with date
- Line graph with dates (proportionally spaced!)



http://www.soundandvision.com/content/violence-and-video-games



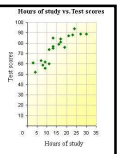
http://www.polygon.com/2014/9/12/6141515/do-violent-video-games-actually-reduce-real-world-crime

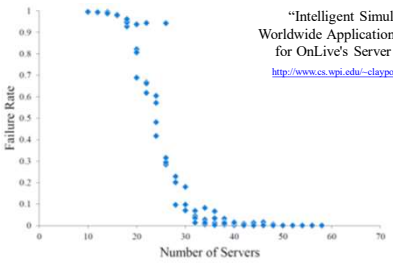
Demo: majors.xlsx

Sel. year and majors
Insert → Line Chart
→ More Line Charts

Scatter Plot

- Two numerical variables, one on each axis
- Reveal patterns in relationship
- Setup "right" models (later)





"Intelligent Simulation of Worldwide Application Distribution for OnLive's Server Network"
http://www.cs.wpi.edu/~claypool/mqp/onlive/

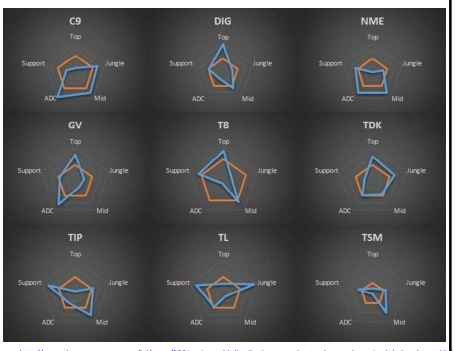
Demo: lol-rates.xlsx

Select two of {win, pick, ban}
Insert → scatter plot

Radar Plot

Gold compared to average, LoL NA teams, by role

- Also called "star charts" or "kiviat plots"
- Good for quick visual compare, especially when axes unequal



Demo: lol-rates.xlsx

Select top line {win, pick, ban} + 1 row num
Insert → Other → Radar scatter plot

http://www.thescoreesports.com/lol/news/2561-using-gold-distribution-to-understand-team-dynamic-global-na-lcs-and-lc

14

Many More Charts!

https://en.wikipedia.org/wiki/Chart

- Bubble
- Waterfall
- Tree
- Gap
- Polar
- Violin
- Candlestick
- Kagi
- Gantt
- Nolan
- Pert
- Smith
- Skyline
- Vowel
- Nomogram
- Natal

- If common chart effective for message, use
- Learn/use other charts as needed

Game Analytics Charts

Gunter Wallner and Simone Kriglstein. "An Introduction to Gameplay Data Visualization", *Game Research Methods*, pages 231-250, ETC Press, ISBN: 978-1-312-88473-1, 2015.
http://dl.acm.org/citation.cfm?id=2812792

- Player choices (e.g., build units)
- Density of activities (e.g., where spend time on map)
- Movement through levels

Player Choices – Pie-Chart

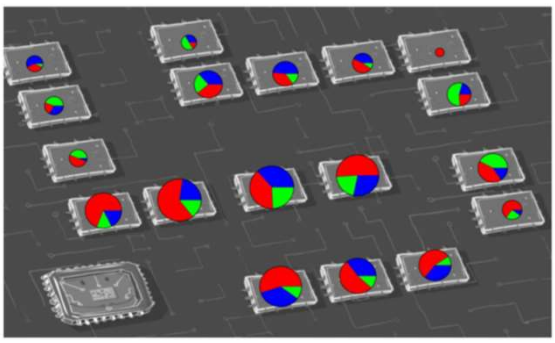


Figure 1. Pie-charts show which types of towers have been built on the different building lots. The radius of the pie-chart is proportional to the number of towers built (Kayali, et al., 2014).

(Custom game, comparative study)

Player Location – Heat Map (1 of 2)

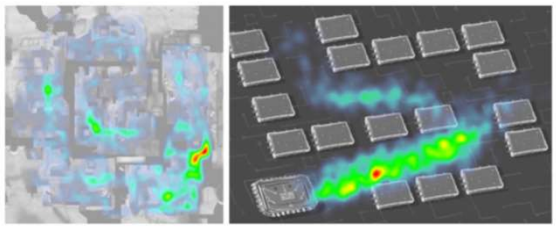
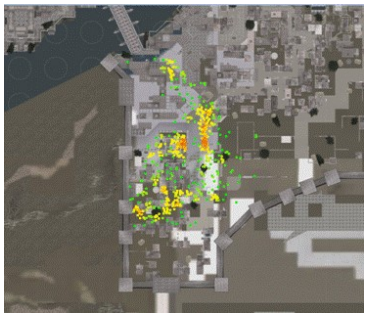


Figure 2. (a) Heatmap of death locations on the Team Fortress 2 map Goldrush. (b) Heatmap showing locations where players of a tower defense game collected coins dropped by defeated enemies (Kayali, et al., 2014).

Player Location – Heat Map (2 of 2)



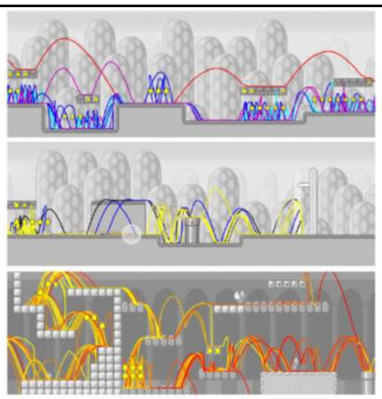
http://www.gamasutra.com/blogs/JonathanDankoff/20140320/213624/Game_Telemetry_with_DNA_Tracking_on_Assassins_Creed.php

Assassin's Creed

Where play testers failed

Result: Make red areas easier

Movement (1 of 2)



(game: *Infinite Mario*, clone of Super Mario Bros.)

Figure 4. Examples of path visualizations coupled with color-coding to communicate additional information. Top: color coding reflects the reported expertise of players obtained through a pre-game survey. Middle: colors depict the state in which the player's character currently resides in. Bottom: the color-gradient reflects physiological data measured in the form of galvanic skin response (Mirza-Babaei, et al., 2014).

Movement (2 of 2)

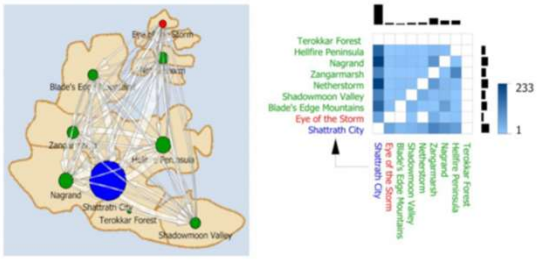


Figure 5. Left: Player movement between regions, cities, and battlegrounds on the World of Warcraft continent Outland. Right: Corresponding matrix view with cells colored according to the number of players moving from one area to another.

Player Behavior - Node-link

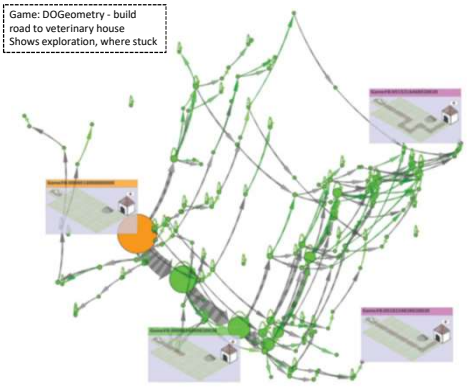
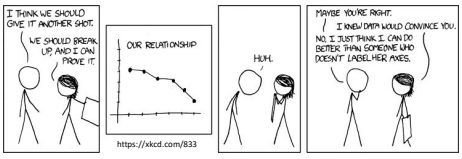


Figure 6. Node-link visualization of player behavior in an abstract puzzle game (Wallner and Kriglstein, 2012b).

Outline

- Types of Charts (done)
- Guidelines for Charts (next)
  - Again, "art" not "rules". Learn with experience. Recognize good/bad when see it.
- Common Mistakes

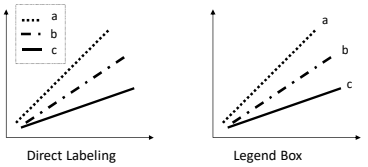


https://xkcd.com/833

Guidelines for Good Charts (1 of 5)

- Require minimum effort from reader
  - Perhaps *most* important metric
  - Given two, can pick one that takes less reader effort

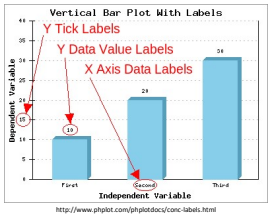
e.g.,



24

Guidelines for Good Charts (2 of 5)

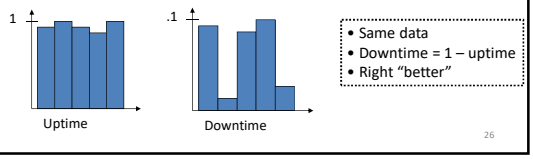
- Maximize information
  - Make self-sufficient
  - Key words in place of symbols
    - e.g., "Gold IV" and not "Player A"
    - e.g., "Daily Games Played" not "Games Played"
  - Axis labels as informative as possible
    - e.g., "Game Time (seconds)" not "Game Time"
  - Help by using captions (or title, if stand-alone)
    - e.g., "Game time in seconds versus player skill in total hours played"



25

Guidelines for Good Charts (3 of 5)

- Minimize ink (1 of 2)
  - Maximize information-to-ink ratio
  - Too much unnecessary ink makes chart cluttered, hard to read
    - e.g., no gridlines unless needed to help read
  - Chart that gives easier-to-read for same data is preferred



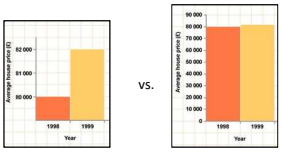
26

Guidelines for Good Charts (3 of 5)

- Minimize ink (2 of 2)

Guidelines for Good Charts (4 of 5)

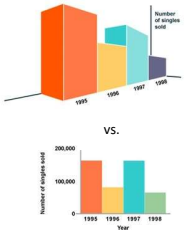
- Use commonly accepted practices
  - Present what people expect
  - e.g., origin at (0,0)
  - e.g., independent (cause) on x-axis, dependent (effect) on y-axis
  - e.g., x-axis scale is linear
  - e.g., increase left to right, bottom to top
  - e.g., scale divisions equal
- Departures are permitted, but require extra effort from reader → so use sparingly!



28

Guidelines for Good Charts (5 of 5)

- Avoid ambiguity
  - Show coordinate axes
    - at right angles
  - Show origin
    - usually at (0,0)
  - Identify individual curves and bars
    - With key/legend or label
  - Do not plot multiple variables on same chart
    - Single y-axis



29

Checklist for Good Charts

- Axes
  - Are both axes labeled?
  - Are the axis labels self-explanatory and concise?
  - Are the scale and divisions shown on both axes?
  - Are the min and max ranges appropriate?
  - Are the units indicated?
- Lines/Curves/Points
  - Is the number of lines/curves reasonably small?
  - Are curves labeled?
  - Are all symbols clearly distinguishable?
  - Is a concise, clear legend provided?
  - Does the legend obscure any data?
- Information
  - If the y-axis is variable, is an indication of spread (error bars) shown?
  - Are grid lines required to read data (if not, then remove)?
- Scale
  - Are units increasing left to right (x-axis) and bottom to top (y-axis)?
  - Do all charts use the same scale?
  - Are the scales contiguous?
  - Is bar chart order systematic?
  - Are bars appropriate width, spacing?
- Overall
  - Does the whole chart add information to reader?
  - Are there no curves/symbols/text that can be removed and still have the same information?
  - Does the chart have a title or caption (not both)?
  - Is the chart self-explanatory and concise?
  - Do the variables plotted give more information than alternatives?
  - Is chart referenced and discussed in any accompanying report?

Describing Chart in Report & Presentation

- "Formula"
  - Describe all axes
    - E.g., "The x-axis is time since game began, in seconds"
  - Describe data sets/trendlines
    - E.g., "The blue dots are the average maze completion time"
  - **Then** provide message
    - E.g., "Notice how the red bar is higher than the blue, indicating that …"
- Example on Web page

http://web.cs.wpi.edu/~imgd2905/d17/samples/analysis-example.html

Guidelines for Good Charts (Summary)

- For each chart, go over "checklist"
- The more "yes" answers, the better
  - Remember, while guidelines, art and not science
  - So, may consciously decide not to follow these guidelines if better without them → but have good reason!
- In practice, takes several trials before arriving at "best" chart
- Want to present message the most: accurately, simply, concisely, logically
- Accompany with description! Text or verbal
  - Remember, audience/reader has not seen! – Make sure to introduce

32

Outline

- Types of Charts (done)
- Guidelines for Charts (done)
- Common Mistakes (next)

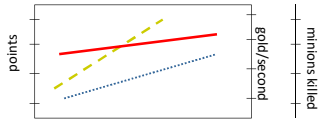
Common Mistakes (1 of 6)

- Presenting too many alternatives on one chart
- Guidelines
  - More than 5 to 7 messages is too many
    - (Maybe related to the limit of human short-term memory?)
  - Line chart with 6+ curves
  - Column chart with 10+ bars
  - Pie chart with 8+ components
  - Each cell in histogram fewer than 5 values

34

Common Mistakes (2 of 6)

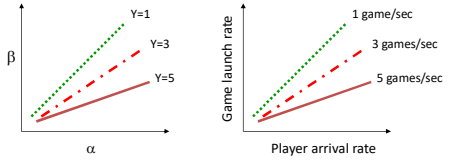
- Presenting many y-variables on single chart
  - Better to make separate graphs
  - Plotting many y-variables saves space, but better to requires reader to figure out relationship
  - Sometimes, space constraints (e.g., journal/conference papers),
    - So may "bend" but better to remove than "break"



35

Common Mistakes (3 of 6)

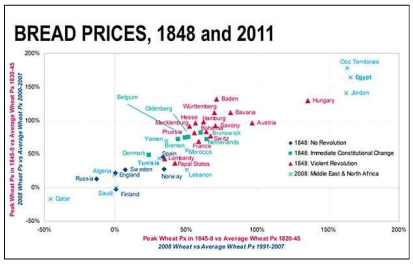
- Using symbols in place of text
- More difficult to read symbols than text
- Reader must flip through report to see symbol mapping to text
  - Even if "save" writers time, really "wastes" it since reader is likely to skip!



36

Common Mistakes (4 of 6)

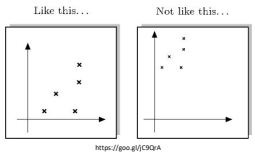
- Placing extraneous information on chart
  - Goal to convey message, so extra information distracting
  - e.g., Using gridlines only when exact values needed
  - e.g., Showing "per-user" data when only average user data needed



37

Common Mistakes (5 of 6)

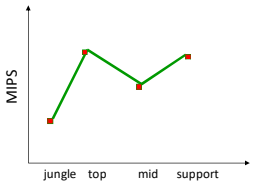
- Selecting scale ranges improperly
  - Most prepared by automatic rules
    - Give good first-guess
  - But
    - May include outlying data points, shrinking body
    - May have endpoints hard to read since on axis
    - May place too many (or too few) tics
  - In practice, (almost) always over-ride scale values



38

Common Mistakes (6 of 6)

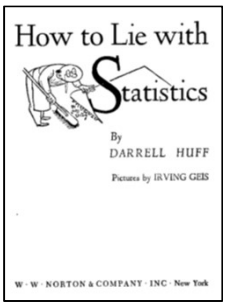
- Using line chart instead of column chart
  - Lines joining successive points signify that they can be approximately interpolated
  - If don't have meaning, should not use line chart



- No linear relationship between champion types
- Instead, use column chart

39

Misleading Charts



There are three kinds of lies: lies, damned lies, and statistics. –Disraeli

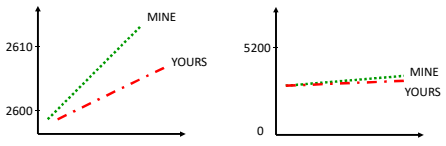
Statistical thinking will one day be as necessary for efficient citizenship as the ability to read and write. –H. G. Wells

It ain't so much the things we don't know that get us in trouble. It's the things we know that ain't so. –Artemus Ward

Round numbers are always false. –Samuel Johnson

Non-Zero Origins to Emphasize (1 of 3)

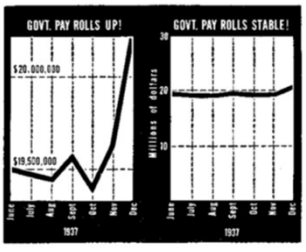
- Normally, both axes meet at origin
- By moving and scaling, can magnify (or reduce!) difference



Which graph is better?

41

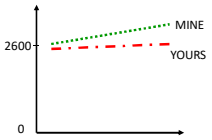
Non-Zero Origins to Emphasize (2 of 3)



Dun's Review, 1938

Non-Zero Origins to Emphasize (3 of 3)

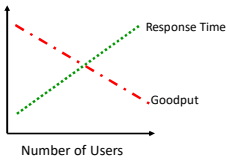
- Choose scale so that vertical height of highest point is at least ¾ of the horizontal offset of right-most point
  - Three-quarters rule
- (And represent origin as 0,0)



43

Using Double-Whammy Graph

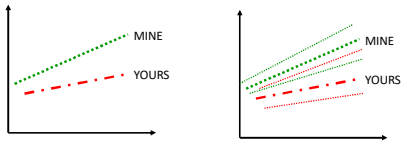
- Two curves can have twice as much impact
  - But if two metrics are related, knowing one predicts other … so use one!



44

Plotting Quantities without Measure of Spread

- When random quantification, representing mean (or median) alone (or single data point!) not enough

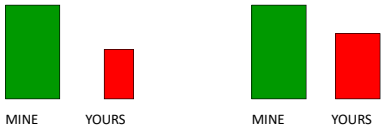


(Worse) (Better)

45

Pictograms Scaled by Height

- If scaling pictograms, do by area not height since eye drawn to area
  - e.g., twice as good → doubling height quadruples area

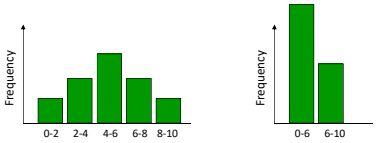


(Worse) (Better)

46

Using Inappropriate Cell Size in Histogram

- Getting cell size "right" always takes more than one attempt
  - If too large, all points in same cell
  - If too small, lacks smoothness

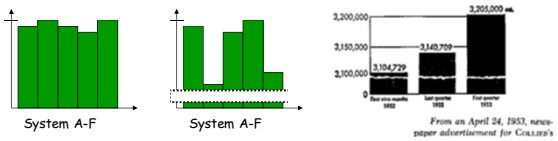


Same data. Left is "normal" and right is "exponential"

47

Using Broken Scales in Column Charts

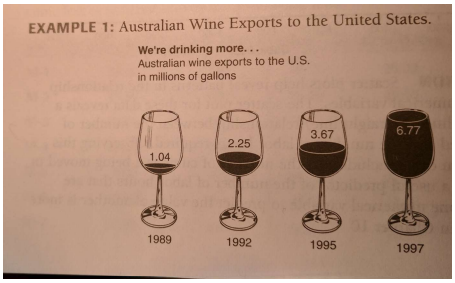
- By breaking scale in middle, can exaggerate differences
  - May be trivial, but then looks significant
  - Similar to "zero origin" problem



48

Pictorial Games (1 of 2)

- Can deceive as easily as can convey meaning



49

Pictorial Games (2 of 2)

- Can deceive as easily as can convey meaning