Exploring Data

17 Jan 2012
Dr. Sean Ho

busi275.seanho.com

- ***HW1* due Thu 10pm**
- ***By Mon, send email to set proposal meeting***

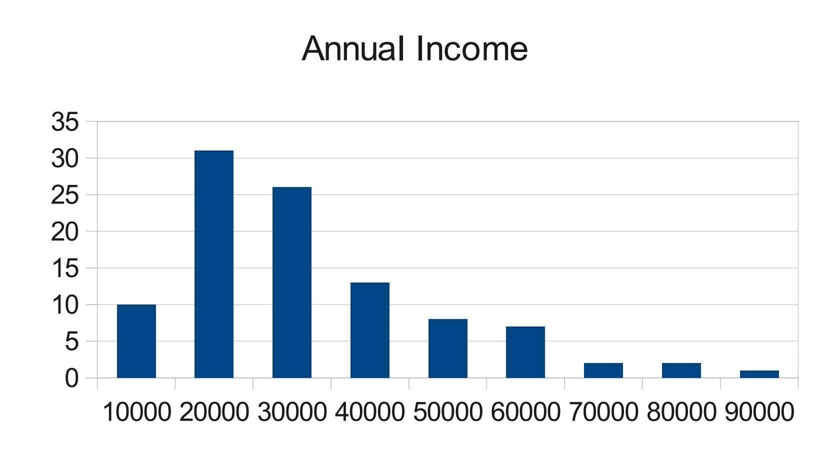
- ***For lecture, please download: 01-SportsShoes.xls***

Outline for today

- Charts
  - Histogram, ogive
  - Scatterplot, line chart
- Descriptives:
  - Centres: mean, median, mode
  - Quantiles: quartiles, percentiles
    - Boxplot
  - Variation: SD, IQR
    - CV, empirical rule, z-scores
- Probability
  - Venn diagrams
  - Union, intersection, complement

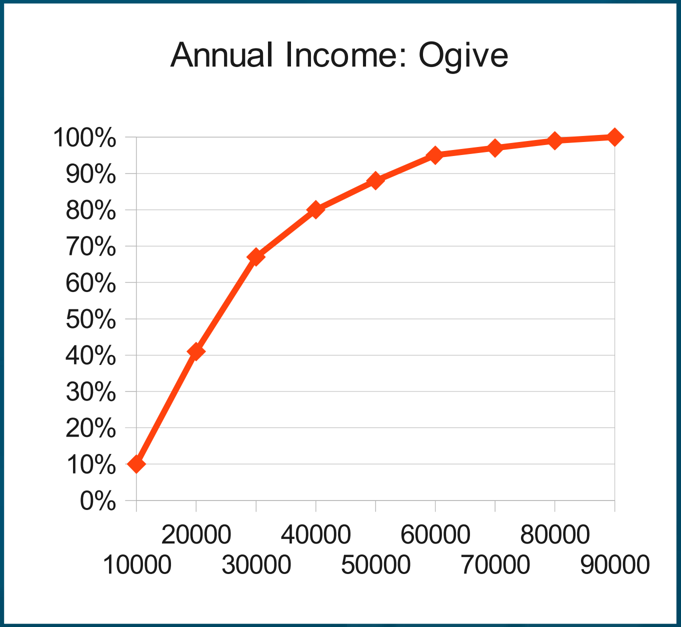
Quantitative vars: histograms

- For quantitative vars (scale, ratio), must group data into classes
  - e.g., length: 0-10cm, 10-20cm, 20-30cm… (class width is 10cm)
  - Specify class boundaries: 10, 20, 30, …
- How many classes? for sample size of n, use k classes, where $2^k \geq n$
- Can use FREQUENCY() w/ column chart, or
- Data > Data Analysis > Histogram



Cumulative distrib.: ogive

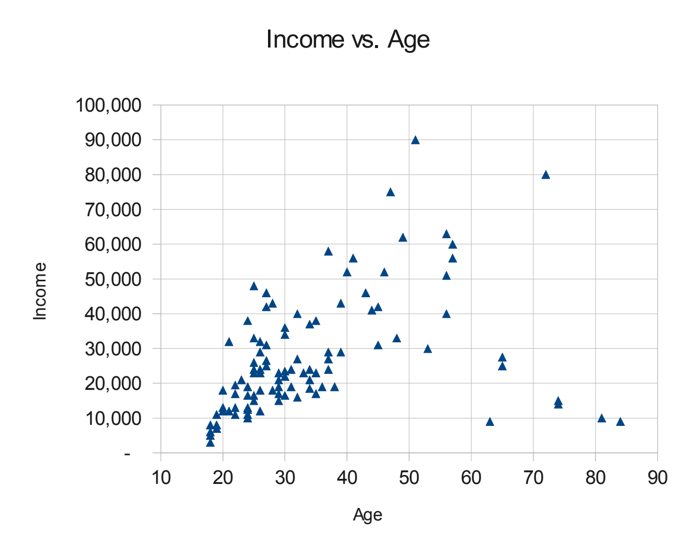
- The ogive is a curve showing the cumulative distribution on a variable:
  - Frequency of values equal to or less than a given value
- Compute cumul. freqs.
- Insert > Line w/Markers



- Pareto chart is an ogive on a nominal var, with bins sorted by decreasing frequency
  - Sort > Sort by: freq > Order: Large to small

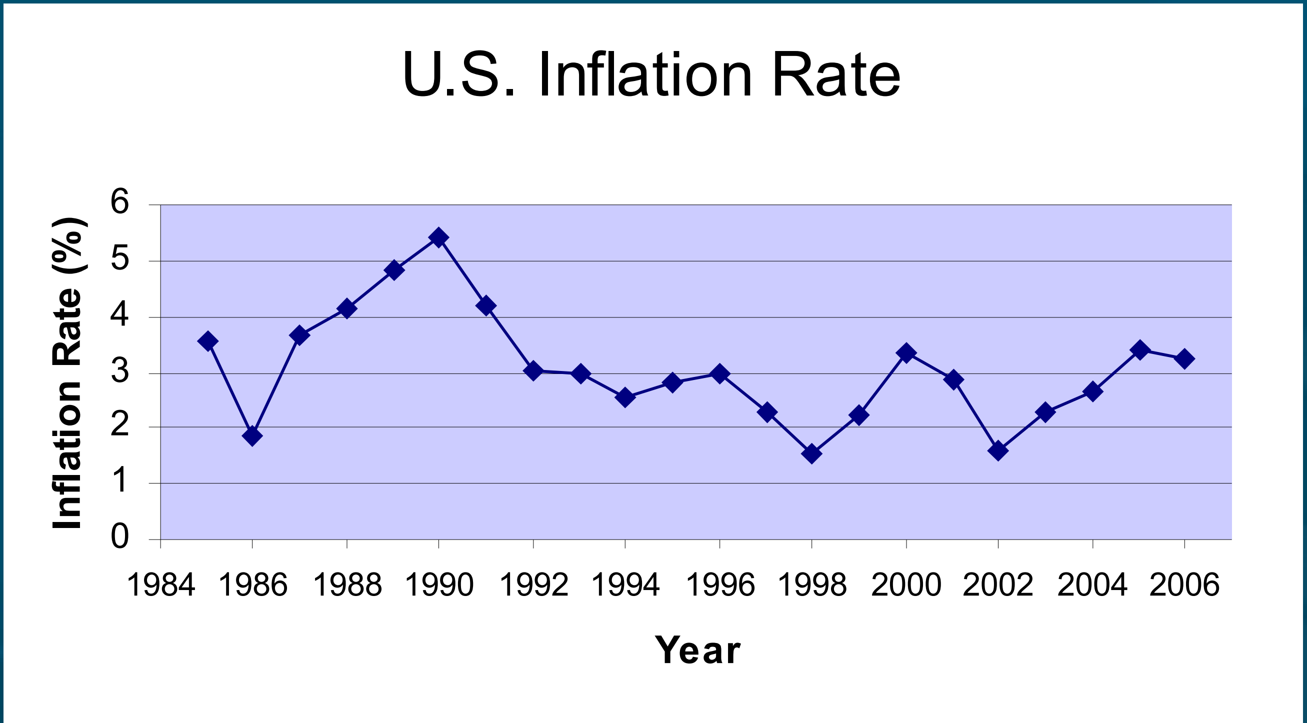
2 quant. vars: scatterplot

- Each participant in the dataset is plotted as a point on a 2D graph
  - (x,y) coordinates are that participant's observed values on the two variables
- Insert > XY Scatter
- If more than 2 vars, then either
  - 3D scatter (hard to see), or
  - Match up all pairs: matrix scatter



Time series: line graph

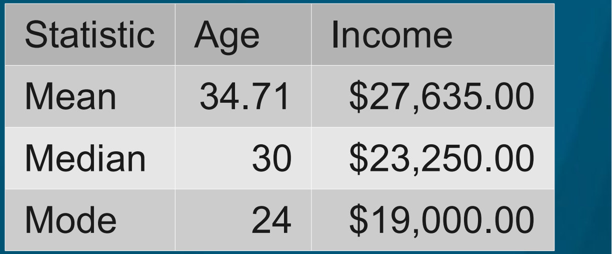
- Think of time as another variable
  - Horizontal axis is time
- Insert > Line > Line



Outline for today

- Charts
    - Histogram, ogive
    - Scatterplot, line chart
- Descriptives:
    - Centres: mean, median, mode
    - Quantiles: quartiles, percentiles
        - Boxplot
    - Variation: SD, IQR
        - CV, empirical rule, z-scores
- Probability
    - Venn diagrams
    - Union, intersection, complement

Descriptives: centres



Statistic	Age	Income
Mean	34.71	$27,635.00
Median	30	$23,250.00
Mode	24	$19,000.00

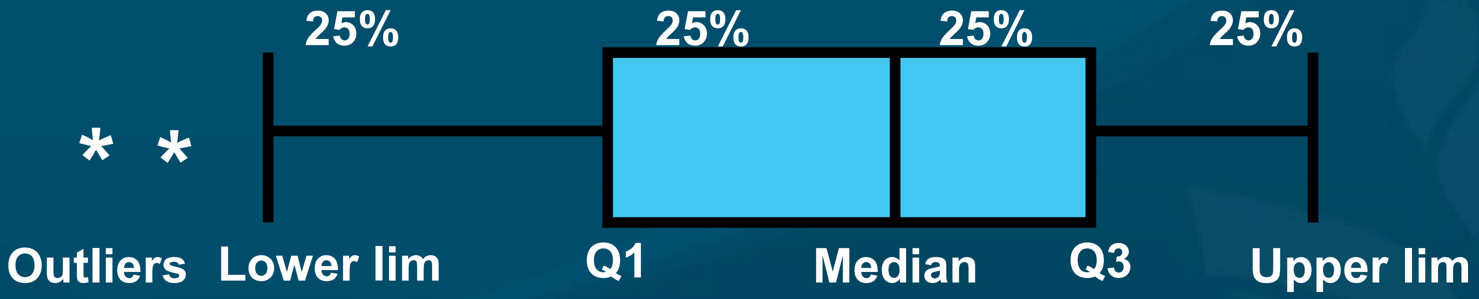
- Visualizations are good, but numbers also help:
  - Mostly just for quantitative vars
- Many ways to find the “centre” of a distribution
  - Mean: AVERAGE()
    - Pop mean: μ ;   sample mean: $\bar{x}$
    - What happens if we have outliers?
  - Median: line up all observations in order and pick the middle one
  - Mode: most frequently occurring value
    - Usually not for continuous variables

Descriptives: quantiles

- The first quartile, Q_1, is the value ¼ of the way through the list of observations, in order
  - Similarly, Q_3 is ¾ of the way through
  - What's another name for Q_2?
- In general the p^th^ percentile is the value p% of the way through the list of observations
  - Rank = (p/100)n: if fractional, round up
    - If exactly integer, average the next two
  - Median = which percentile?
- Excel: QUARTILE(data, 3), PERCENTILE(data, .70)

Box (and whiskers) plot

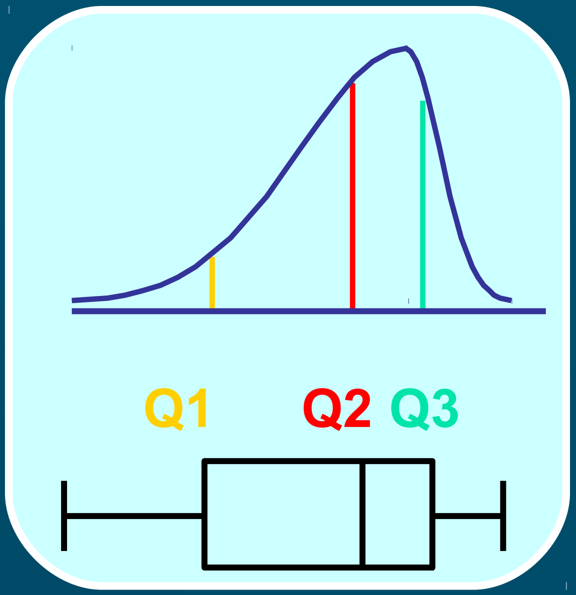
- Plot: median, Q_1, Q_3, and upper/lower limits:
  - Upper limit = $Q_3 + 1.5(IQR)$
  - Lower limit = $Q_1 - 1.5(IQR)$
- IQR = interquartile range = $(Q_3 - Q_1)$
- Observations outside the limits are considered outliers: draw as asterisks (*)



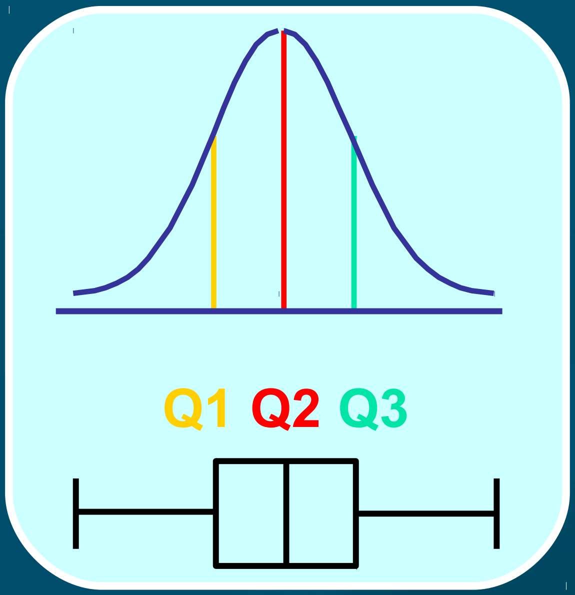
- Excel: try tweaking bar charts

Boxplots and skew

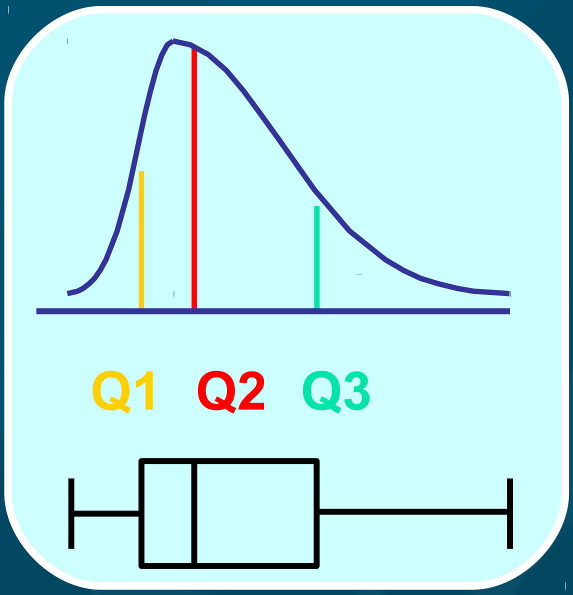
Left-Skewed



Symmetric



Right-Skewed

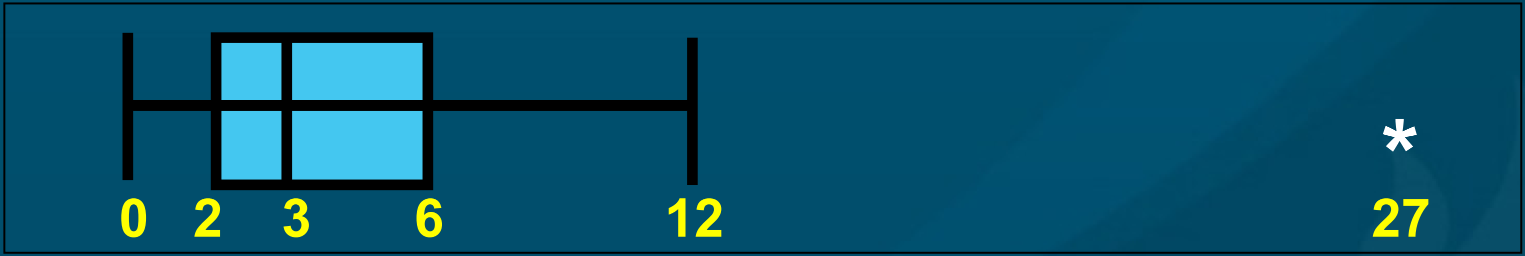


Boxplot Example

- Data:

Min		Q_1			Q_2			Q_3		Max
0	2	2	2	3	3	4	5	6	11	27

- Right skewed, as the boxplot depicts:



$$\text{Upper limit} = Q_3 + 1.5\,(Q_3 - Q_1)$$
$$= 6 + 1.5\,(6 - 2) = 12$$

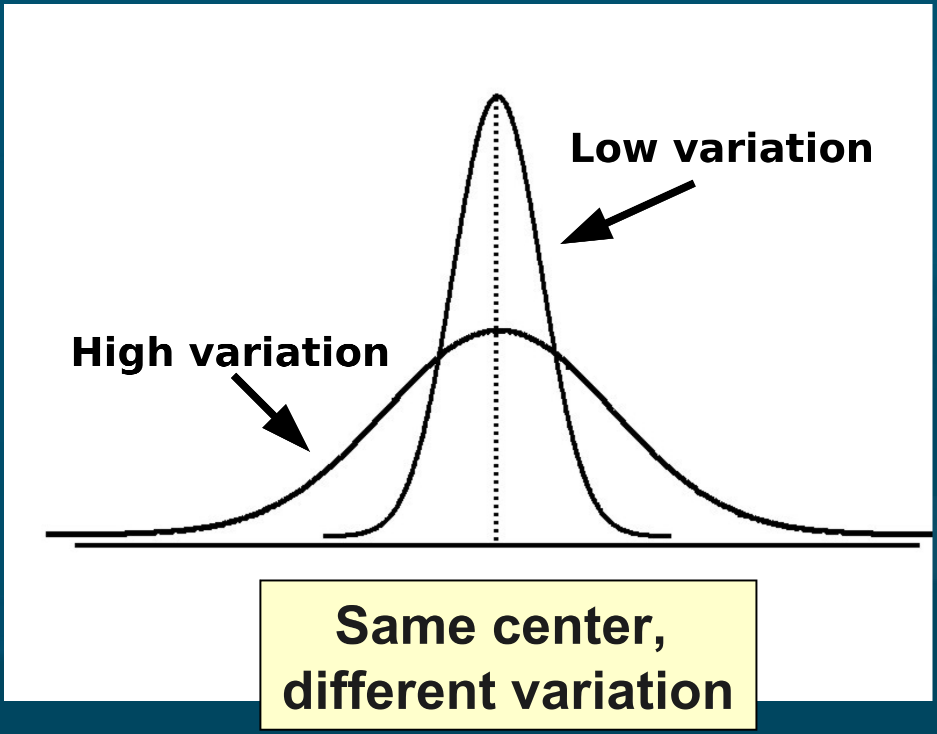
27 is above the upper limit so is shown as an outlier

Outline for today

- Charts
  - Histogram, ogive
  - Scatterplot, line chart
- Descriptives:
  - Centres: mean, median, mode
  - Quantiles: quartiles, percentiles
    - Boxplot
  - Variation: SD, IQR
    - CV, empirical rule, z-scores
- Probability
  - Venn diagrams
  - Union, intersection, complement

Measures of variation

- Spread (dispersion) of a distribution:
  are the data all clustered around the centre,
  or spread all over a wide range?



Same center,
different variation

Range, IQR, standard deviation

- Simplest: range = max – min
  - Is this robust to outliers?
- IQR = Q_3 – Q_1 (“too robust”?)
- Standard deviation:
  - Population: $\sigma = \sqrt{\frac{\sum_{i=1}^{n}\left(x_i - \mu\right)^2}{n}}$
  - Sample: $s = \sqrt{\frac{\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2}{n-1}}$
  - In Excel: STDEV()
- Variance is the SD w/o square root

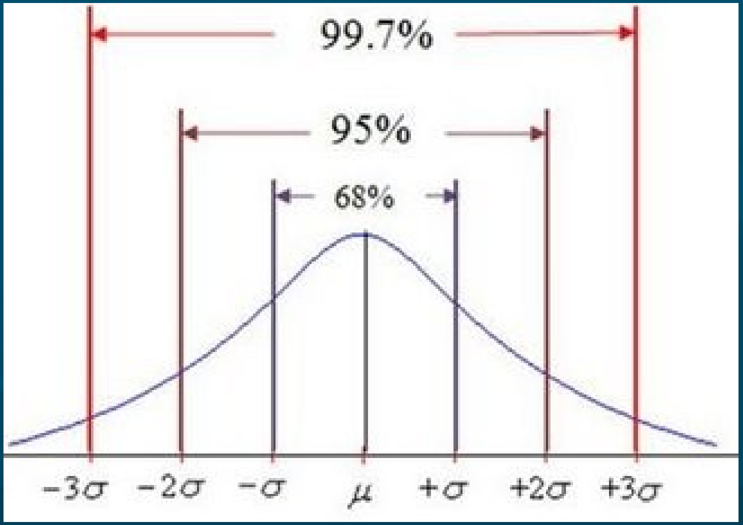
	Pop.	Samp.
Mean	μ	$\bar{x}$
SD	σ	s

Coefficient of variation

- Coefficient of variation: SD relative to mean
  - Expressed as a percentage / fraction
- e.g., Stock A has avg price $\overline{x}=\$50$ and $s=\$5$
  - CV = $s / \overline{x}$ = 5/50 = 10% variation
- Stock B has $\overline{x}=\$100$ same standard deviation
  - CV = $s / \overline{x}$ = 5/100 = 5% variation
- Stock B is less variable relative to its average stock price

SD and Empirical Rule

- Every distribution has a mean and SD, but for most “nice” distribs two rules of thumb hold:
- Empirical rule: for “nice” distribs, approximately
  - 68% of data lie within ±1 SD of the mean
  - 95% within ±2 SD of the mean
  - 99.7% within ±3 SD



SD and Tchebysheff's Theorem

- For any distribution, at least $(1-1/k^2)$ of the data will lie within k standard deviations of the mean
  - Within $(\mu \pm 1\sigma)$: $\geq(1-1/1^2) = 0\%$
  - Within $(\mu \pm 2\sigma)$: $\geq(1-1/2^2) = 75\%$
  - Within $(\mu \pm 3\sigma)$: $\geq(1-1/3^2) = 89\%$

z-scores

- Describes a value's position relative to the mean, in units of standard deviations:
  - $z = (x - \mu)/\sigma$
- e.g., you got a score of 35 on a test: is this good or bad? Depends on the mean, SD:
  - $\mu=30$, $\sigma=10$: then $z = +0.5$: pretty good
  - $\mu=50$, $\sigma=5$: then $z = -3$: really bad!

Outline for today

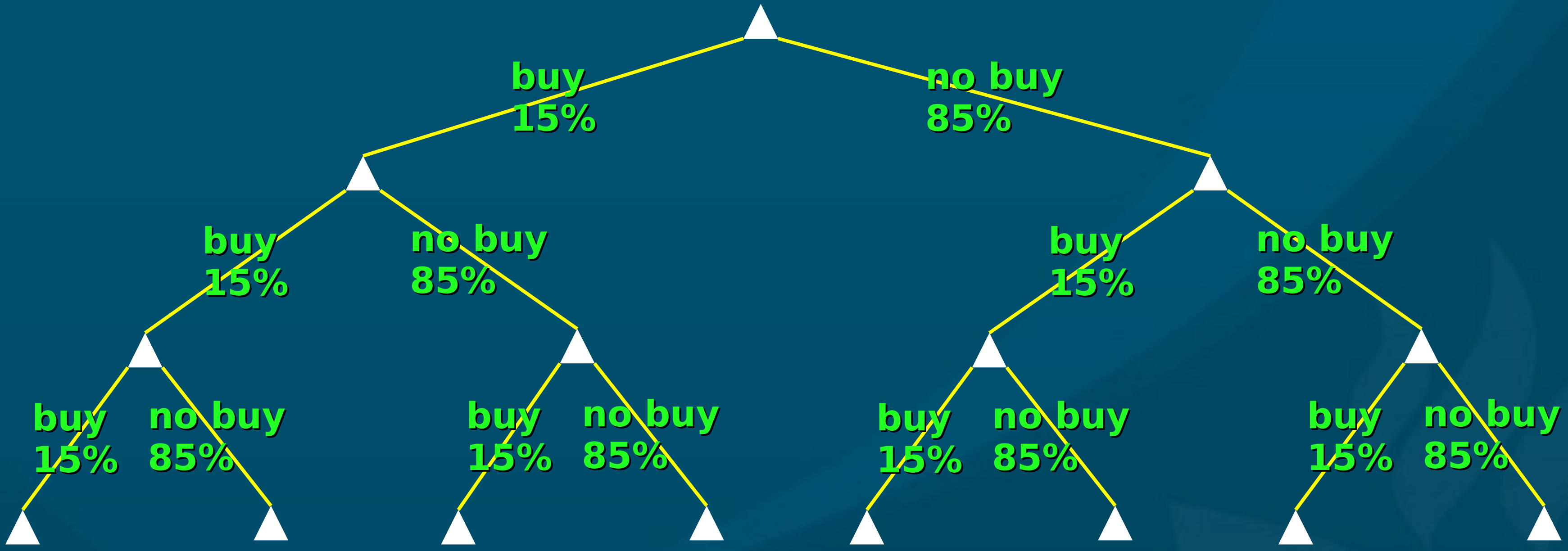
- Charts
  - Histogram, ogive
  - Scatterplot, line chart
- Descriptives:
  - Centres: mean, median, mode
  - Quantiles: quartiles, percentiles
    - Boxplot
  - Variation: SD, IQR
    - CV, empirical rule, z-scores
- Probability
  - Venn diagrams
  - Union, intersection, complement

Probability

- Chance of a particular event happening
- e.g., in a sample of 1000 people, say 150 will buy your product:
  - ⇒ the probability that a random person from the sample will buy your product is 15%
  - Experiment: pick a random person (1 trial)
  - Possible outcomes: {“buy”, “no buy”}
  - Sample space: {“buy”, “no buy”}
  - Event of interest: A = {“buy”}
  - P(A) = 15%

Event trees

- Experiment: pick 3 people from the group
- Outcomes for a single trial: {“buy”, “no buy”}
- Sample space: {BBB, BBN, BNB, BNN, NBB, …}



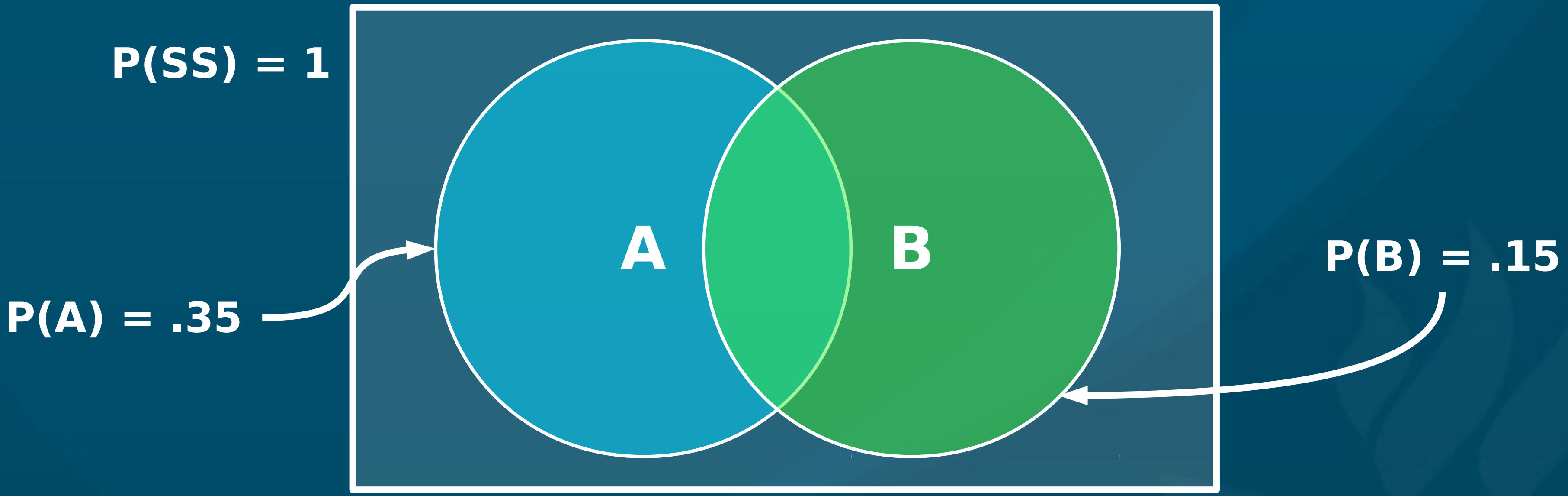
P(BNB)

= (.15)(.85)(.15)

- Event: A = {at least 2 people buy}: P(A) = ?

Venn diagrams

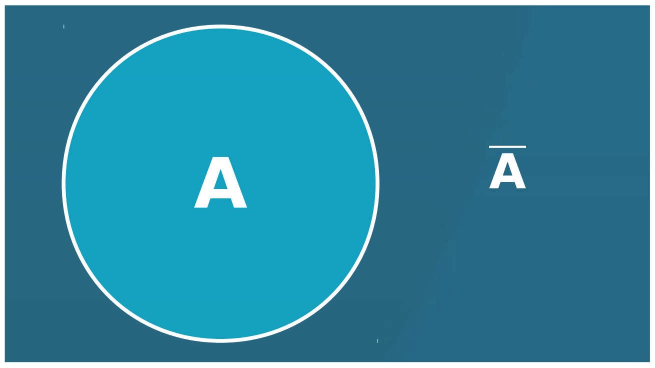
- Box represents whole sample space
- Circles represent events (subsets) within SS
- e.g., for a single trial:

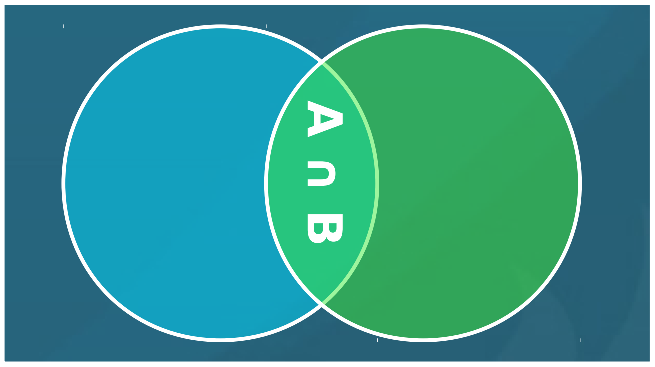


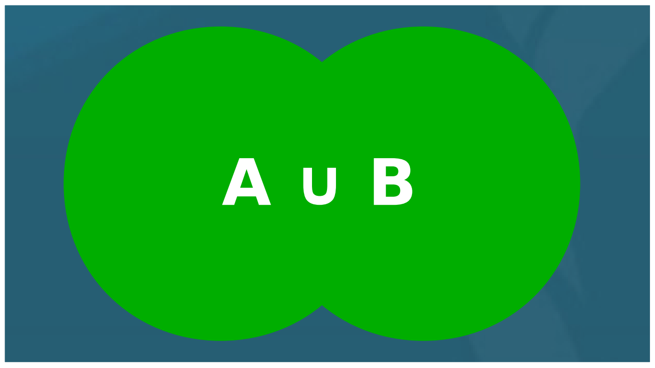
- A = “clicks on ad”
- B = “buys product”

Venn: set theory

- Complement: $\bar{A}$
  = “does not click ad”
  - $P(\bar{A}) = 1 - P(A)$
- Intersection: A ∩ B
  = “clicks ad and buys”
- Union: A ∪ B
  = “either clicks ad or buys”

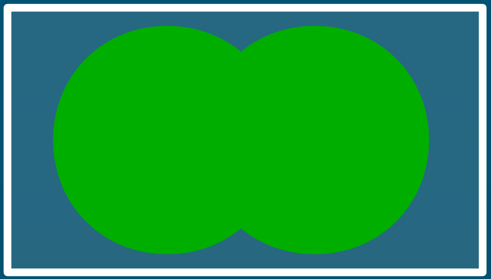






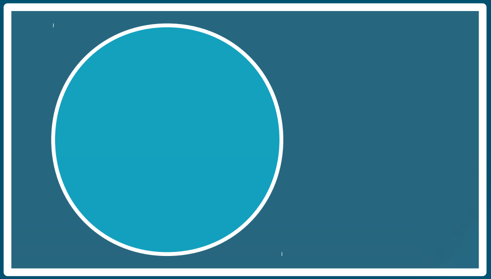
Addition rule: A ∪ B

$P(A \cup B)$



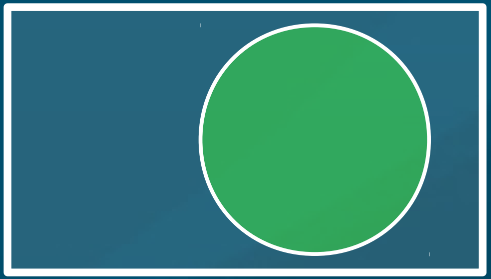
$=$

$P(A)$



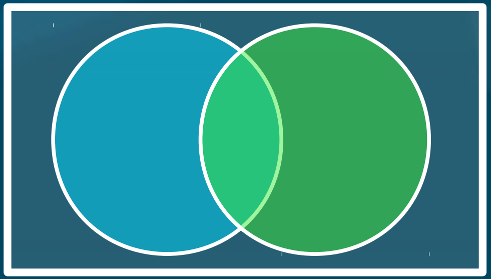
$+$

$P(B)$



$-$

$P(A \cap B)$



Addition rule: example

- 35% of the focus group clicks on ad:
  - P(?) = .35
- 15% of the group buys product:
  - P(?) = .15
- 45% are “engaged” with the company: either click ad or buy product:
  - P(?) = .45
- ⇒ What fraction of the focus group buys the product through the ad?
  - P(A ∪ B) = P(A) + P(B) – P(A ∩ B)

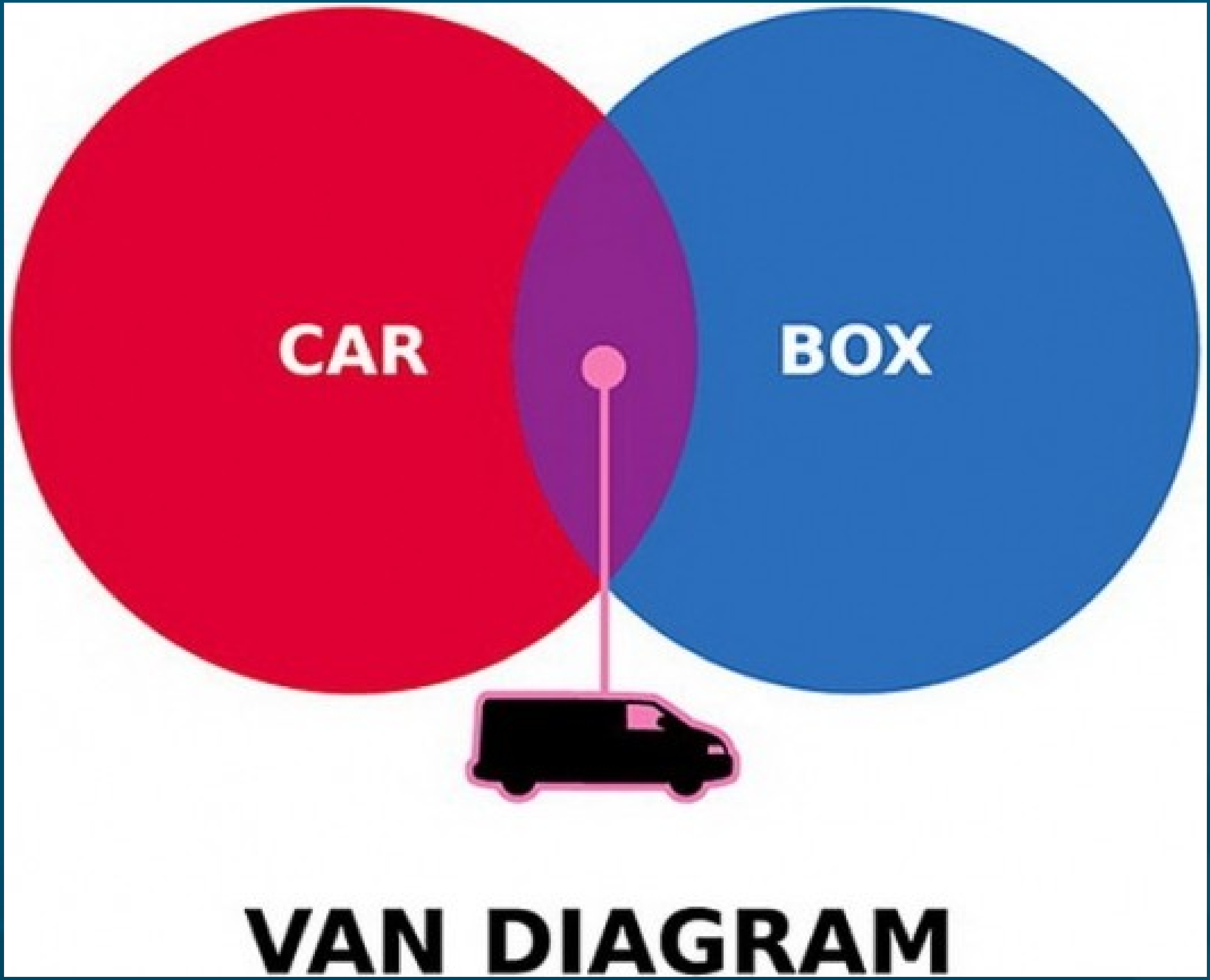
    ? = ? + ? - ?

Mutual exclusivity

- Two events A and B are mutually exclusive if the intersection is null: $P(A \cap B) = 0$
  - i.e., an outcome cannot satisfy both A and B simultaneously
- e.g., A = male, B = female
- e.g., A = born in Alberta, B = born in BC

- If A and B are mutually exclusive, then the addition rule simplifies to:
  - $P(A \cup B) = P(A) + P(B)$

Yep!



TODO

- HW1 (ch1-2): due online, this Thu 19Jan
  - Text document: well-formatted, complete English sentences
  - Excel file with your work, also well-formatted
  - HWs are to be individual work
- Get to know your classmates and form teams
  - Email me when you know your team
- Discuss topics/DVs for your project
  - Find existing data, or gather your own?
- Schedule proposal meeting during 23Jan - 3Feb