Ch3: Exploring Your Data with Descriptives

- **HW1** due tonight 10pm
- Download and open "02-SportsShoes.xls"

15 Sep 2011
BUSI275
Dr. Sean Ho

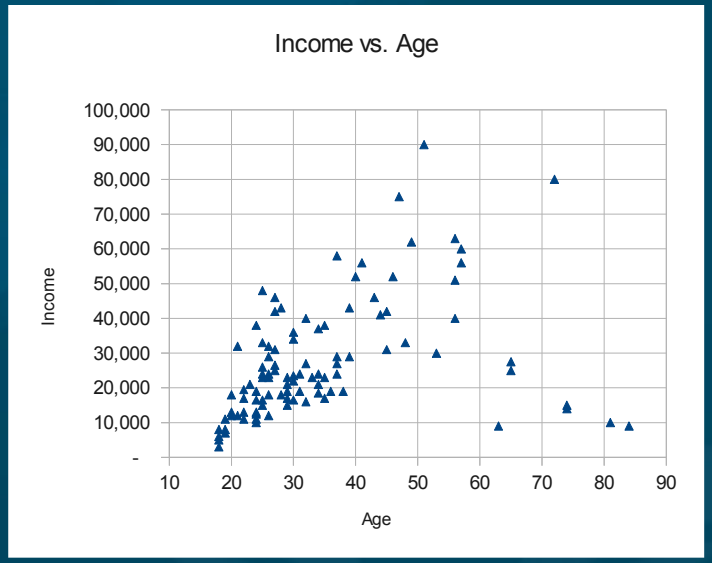
TRINITY WESTERN UNIVERSITY

Outline for today

- Exploring data with charts: line, scatter
- Exploring data with descriptives:
  - Measures of Centre
    - Mean, median, mode
  - Quartiles, percentiles, boxplot
  - Measures of Variation
    - Range, IQR
    - Standard deviation, variance, coef of var
  - Empirical Rule and z-scores

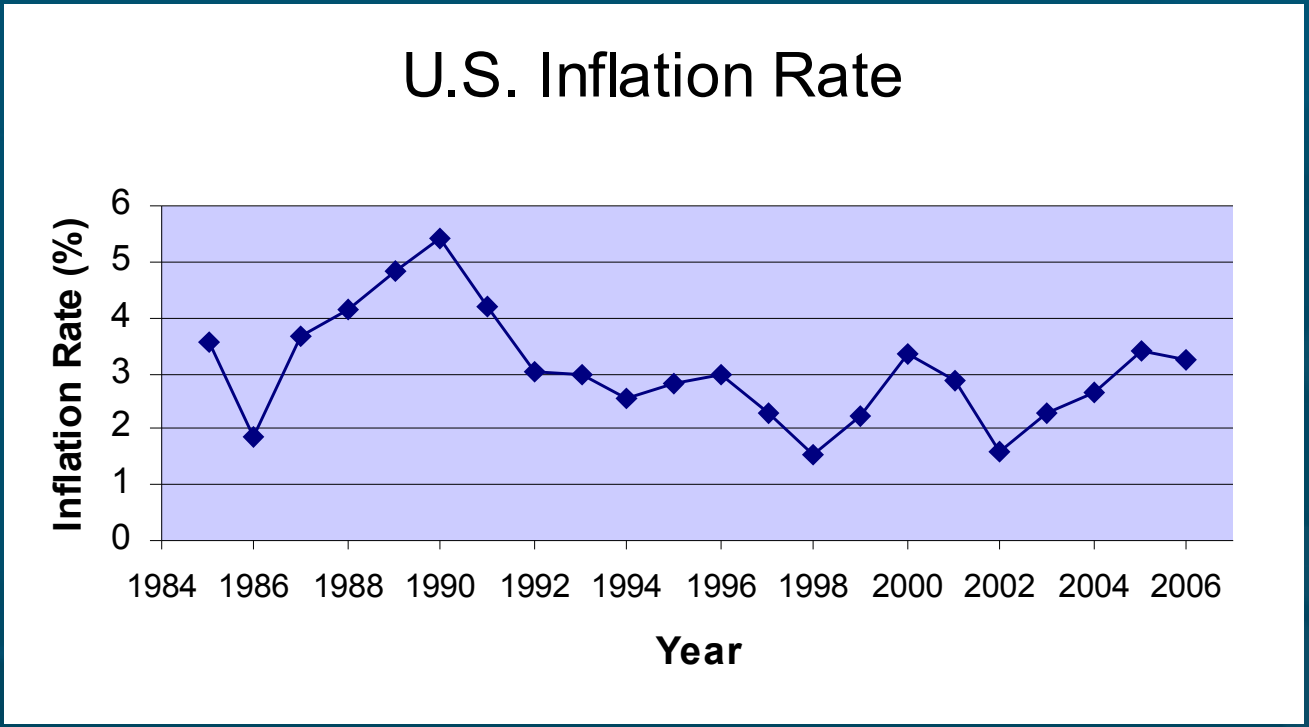
2 quant. vars: scatterplot

- 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



Descriptives: centres

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

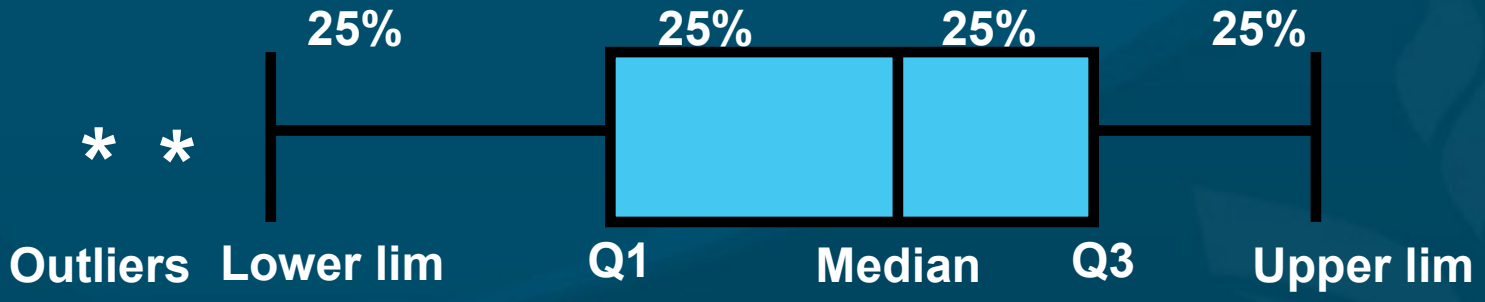
- Visualizations are good, but numbers also help:
  - Mostly just for quantitative vars
- Many ways to find the "centre" of a distribution
  - Mean: AVERAGE()
    - Pop mean: μ ; sample mean: $\bar{x}$
    - What happens if we have outliers?
  - Median: line up all observations in order and pick the 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

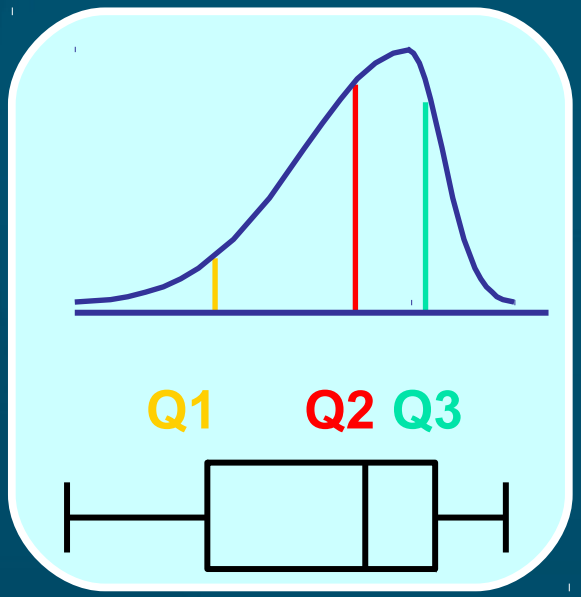
- Plot: median, Q_1, Q_3, and upper/lower limits:
  - Upper limit = $Q_3 + 1.5(IQR)$
  - Lower limit = $Q_1 - 1.5(IQR)$
- IQR = interquartile range = $(Q_3 - Q_1)$
- Observations outside the limits are considered outliers: draw as asterisks (*)



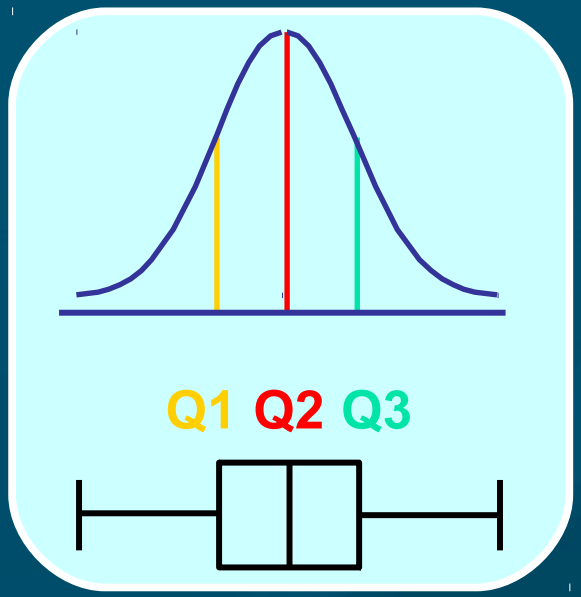
- Excel: try tweaking bar charts

Boxplots and skew

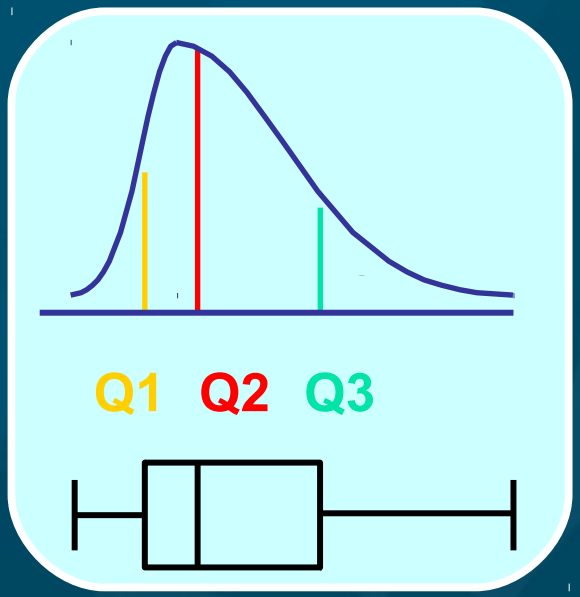
Left-Skewed



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:

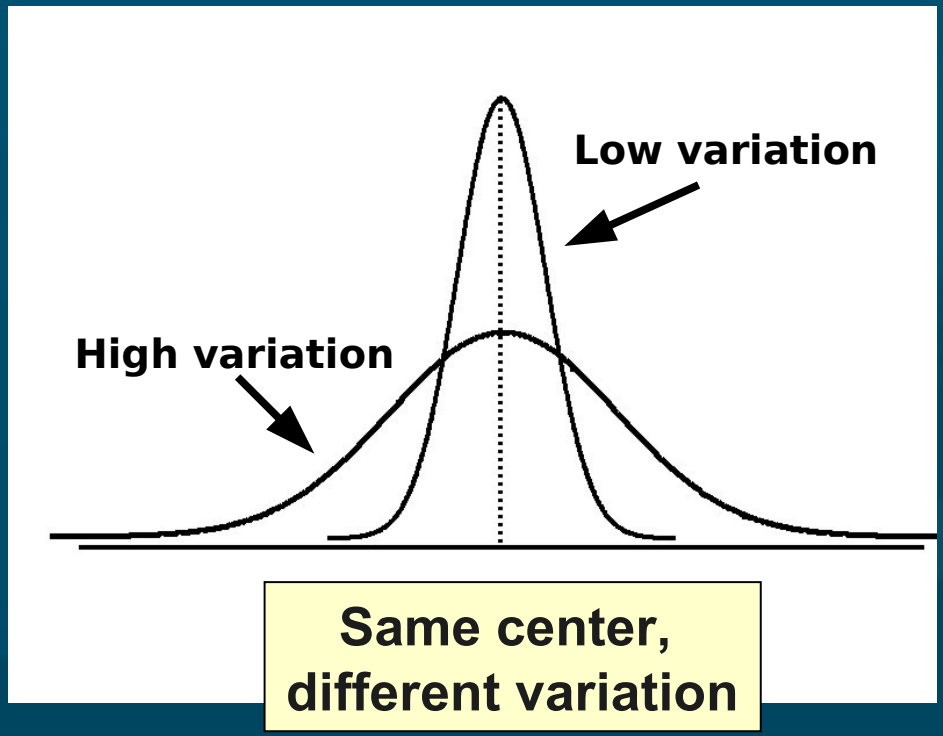
Upper limit = $Q_3 + 1.5\,(Q_3 - Q_1)$

$= 6 + 1.5\,(6 - 2) = 12$

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

Measures of variation

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



Range, IQR, standard deviation

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

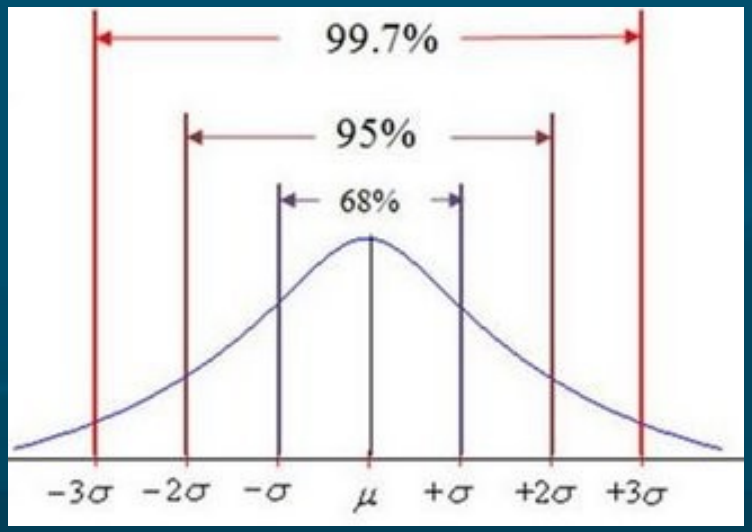
	Pop.	Samp.
Mean	μ	$\bar{x}$
SD	σ	s

Coefficient of variation

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




NausicaaDistribution

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!

TODO

- HW1 (ch1-2): due tonight at 10pm
  - Format as a clear, neat document
  - Also upload your Excel spreadsheet
  - HWs are to be individual work
- Get to know your classmates and form teams
  - Email me when you know your team
  - You can come up with a good name, too
- Discuss topics/variables you are interested in
  - Find existing data, or gather your own?