What you should know before you collect data

BAE 815 (Fall 2017)
Dr. Zifei Liu
Zifeiliu@ksu.edu

- Types and levels of study
- Descriptive statistics
- Inferential statistics
- How to choose a statistical test
- Cross-validation
- Uncertainty analysis
- Sensitivity analysis

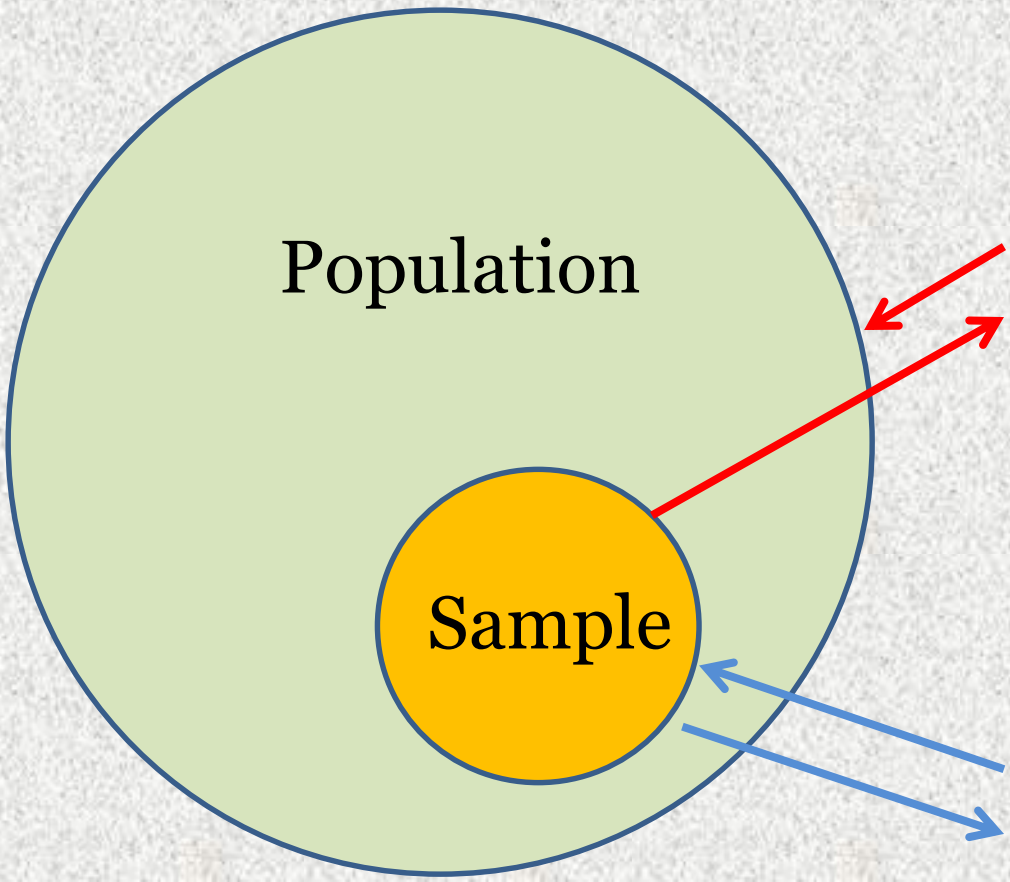
Types of study (research design)

- Descriptive (e.g., case-study, observational)
  - No control over extraneous variables
  - Leaves cause-effect relationship ambiguous
- True experimental
  - Manipulate one variable and measure effects on another
  - Higher internal validity
- Semi-experimental (e.g. field experiment, quasi-experiment)
  - Suffer from the possibility of contamination
  - Higher external validity than lab experiments
- Survey
- Review, meta-analysis

Internal vs. external validity

- Internal validity:
  - Is the experiment free from confounding?
  - The degree to minimizes systematic error
- External validity:
  - How well can the results be generalized to other situations?
  - Representativeness of sample

Types of statistics



Inferential statistics:
make inference about the population from samples

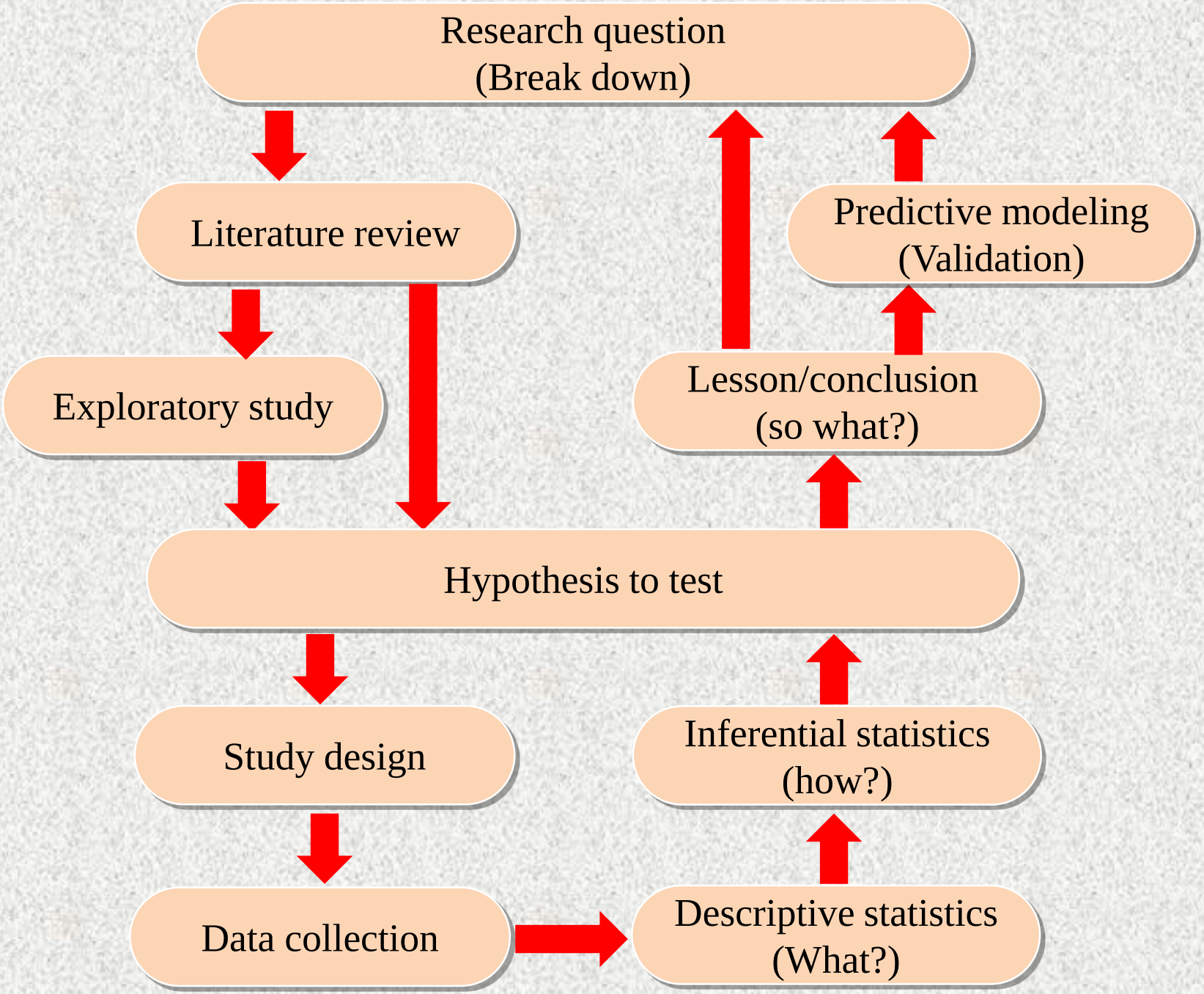
- confidence intervals
- hypothesis test
  - compare groups
  - relationship

Descriptive statistics:
summarize and describe features of data

- measures of location: mean and median;
- measures of variability: range, variance

Levels of data analysis

- Exploratory research seeks to generate a hypotheses by looking for potential relations between variables.
  - lack knowledge of the direction and strength of the relation.
  - minimize type II error (the probability of missing a real effect).
- Confirmatory research tests a hypotheses, which are usually derived from a theory or the results of previous studies.
  - minimize type I error (the probability of falsely reporting a coincidental result as meaningful).
- Predictive modeling

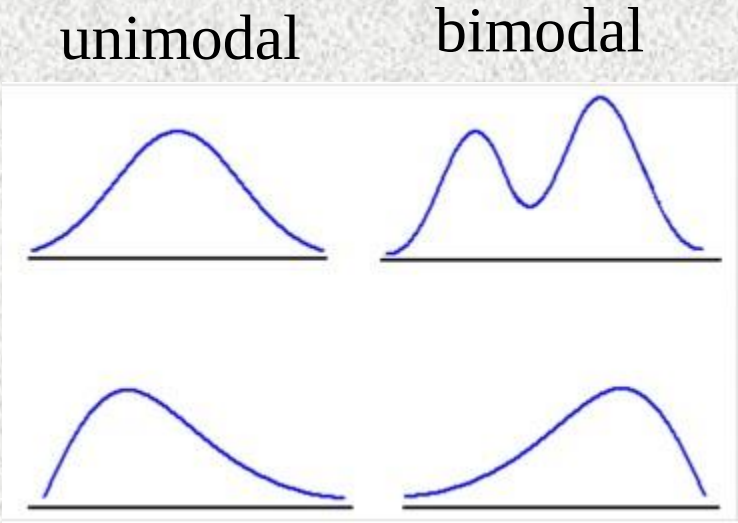

Research question
(Break down)
Literature review
Exploratory study
Predictive modeling
(Validation)
Lesson/conclusion
(so what?)
Hypothesis to test
Study design
Inferential statistics
(how?)
Data collection
Descriptive statistics
(What?)

Descriptive statistics

Check frequency distribution of your data

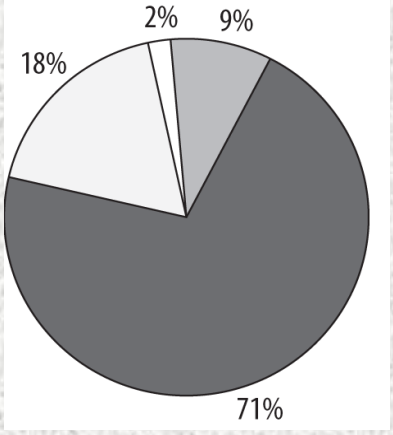
- Modality
- Symmetry
- Central tendency
  - mean, median, mode
- Dispersion or variation:
  - range, min to max
  - standard deviation
  - interquartile range, IQR=Q3-Q1

  Q1 and Q3 are defined as the 25th and 75th percentiles)

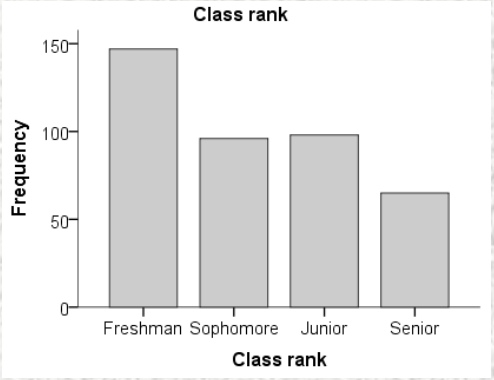


Categorical variable

Pie chart

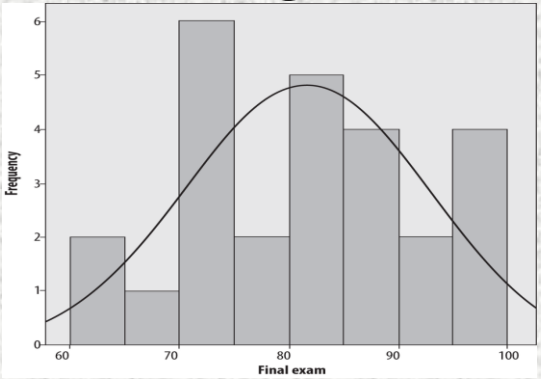


Bar chart

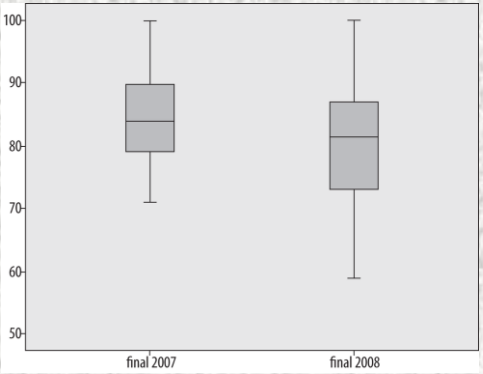


Quantitative variable

Histogram



Boxplot (unimodal)



Frequency distribution of one variable

61,64,68,

70,70,71,73,74,74,76,79,

80,80,83,84,84,87,89,89,89,

90,92,95,95,98

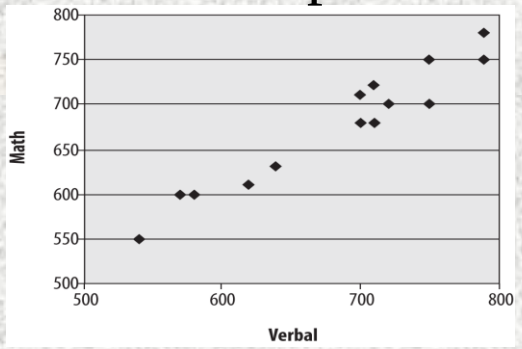
Stem-and-leaf display

Stem	Leaf
6	148
7	00134469
8	003447999
9	02558

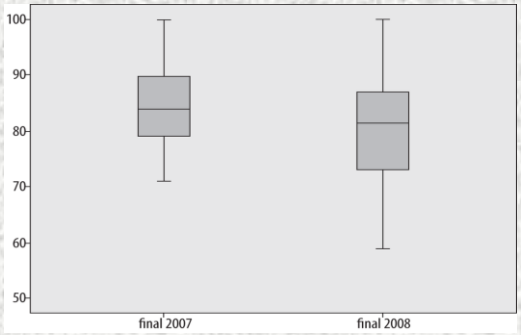
Similar to a histogram on its side, but it has the advantage of showing the actual data values.

Stem-and-leaf display

Two quantitative variables: Scatter plot



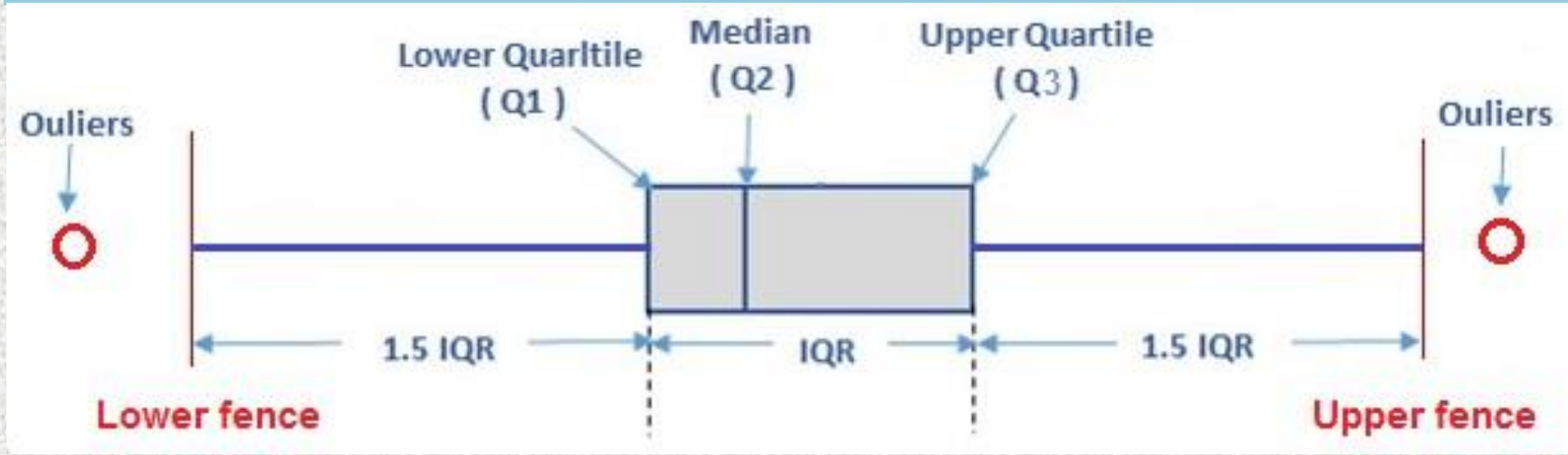
One categorical and one quantitative variable: Boxplot



Crosstabulation: work for both categorical and quantitative variables

A	B			Total
	B1	B2	B3	
A1				
A2				
A3				
Total				

Relationship between two variables



Outliers:

- Values < Q1-1.5IQR, or Values > Q3+1.5IQR

Extreme values

- Values < Q1-3IQR, or Values > Q3+3IQR

Median and IQR are robust statistics that are less affected by outliers.

Outliers and extreme values

- t-test
  - determine if a difference exists between the means of two groups.
- ANOVA (Analysis of Variance)
  - comparing three or more groups for statistical significance.
  - generalize the t-test to more than two groups.
- Regression Analysis
  - assess if change in one variable predicts change in another variable.
- Factor Analysis
  - describe variability among observed, correlated variables in terms of a potentially lower number of unobserved variables called factors.
  - groups similar variables into dimensions.

Inferential statistics

Nonparametric vs. parametric tests

Nonparametric tests

- Fewer assumptions (e.g. normality, homogeneity of variance).
- Less powerful

Possible reasons to use nonparametric tests

- Your area of study is better represented by the median
- You have a very small sample size
- You have ordinal data, or outliers that you can’t remove

Choosing a statistical test

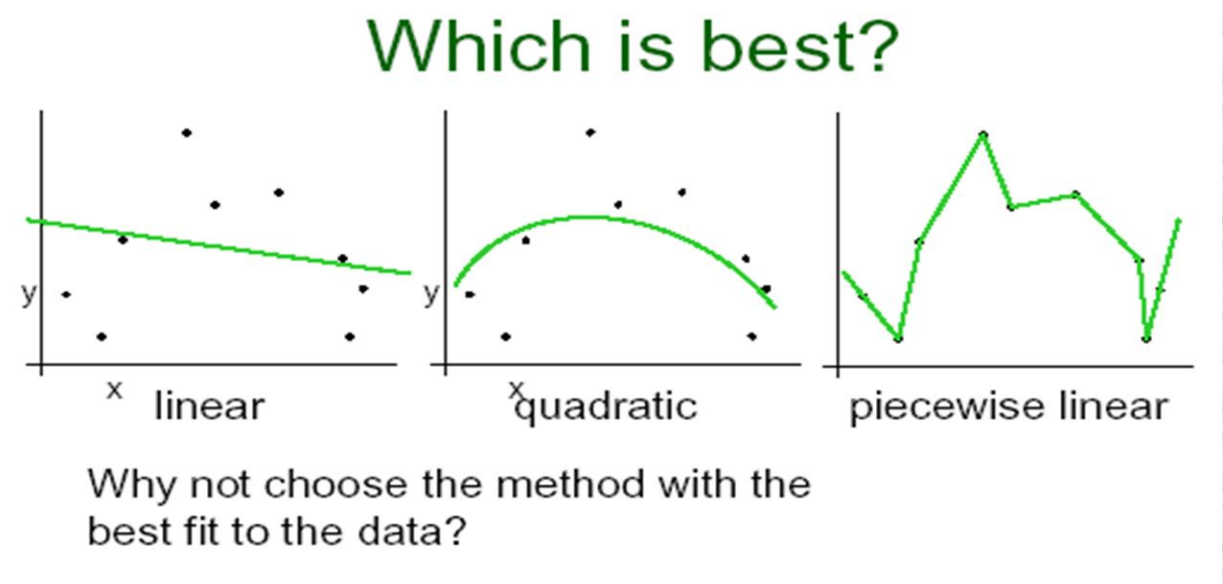
	Test of comparison	Test of correlation
Categorical data	Chi-square, Sign test	Chi-square, Fisher's Exact
Quantitative data (Nonparametric)	Wilcoxon, Mann Whitney, Friedman, Kruskal-Wallis	Spearman's r, Kendall Tau (<20 rankings)
Quantitative data (Parametric)	t-test, ANOVA	Pearson'r

- Provide exact p value, e.g. “p=.028” instead of “p<.05”, or “P<.03”.

- “P<.001” instead of “p=.00”, or “p=.0007584”.

- When your p-value is greater than the alpha (e.g. 0.05)
  - Call it ‘non-significant’ and write it up as such.
  - Learn from non-significance; check the power of your experimental design; make suggestions for future directions

The nice way to report p values



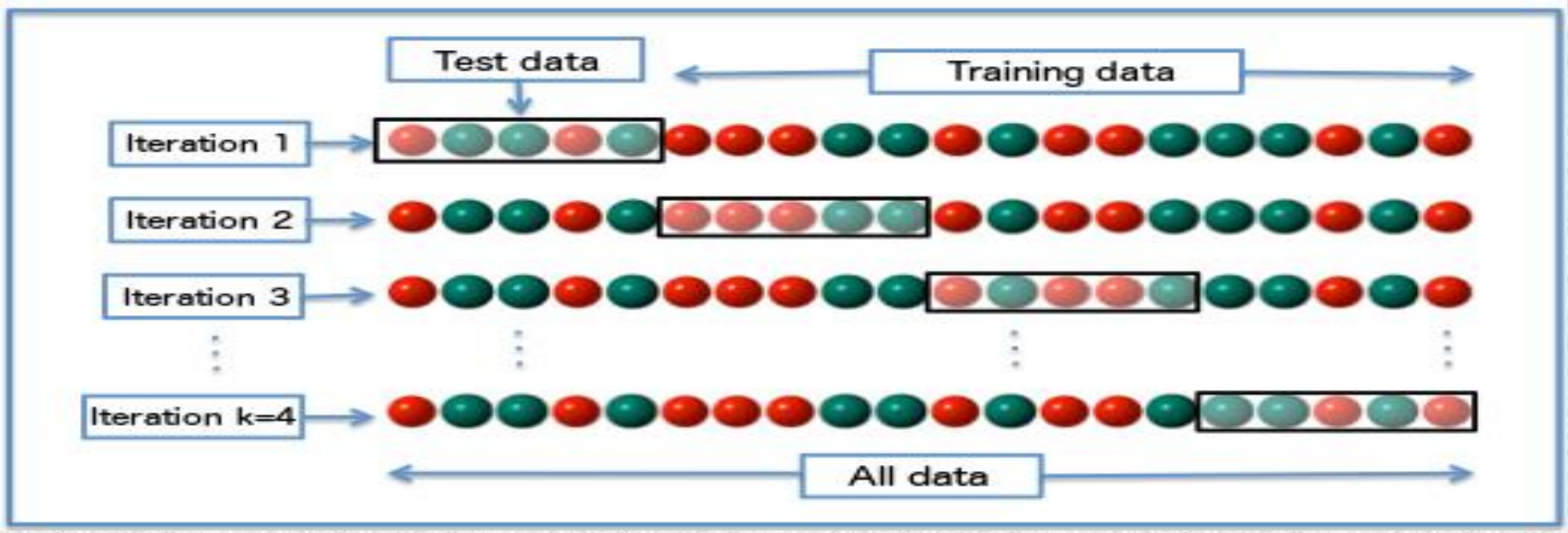
Cross-validation is a way to estimate prediction error.

- prevent overfitting (overfitting models will have high r^2, but will perform poorly in predicting out-of-sample cases)
- compare different model algorithms
- estimate prediction error

Cross-validation for predictive modeling

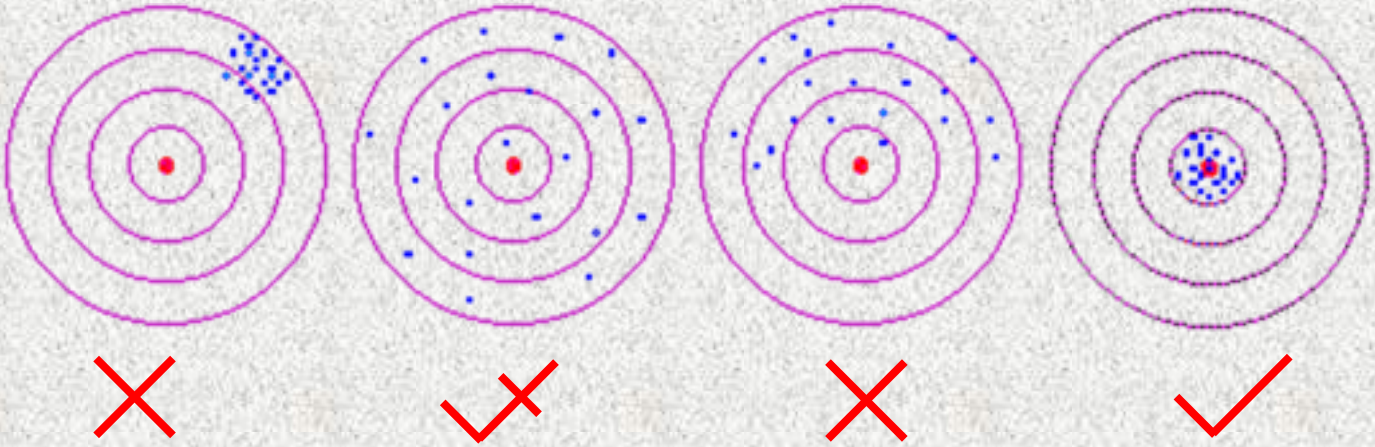
Cross-validation

- k-fold cross-validation
- Leave-p-out
- Leave-one-out

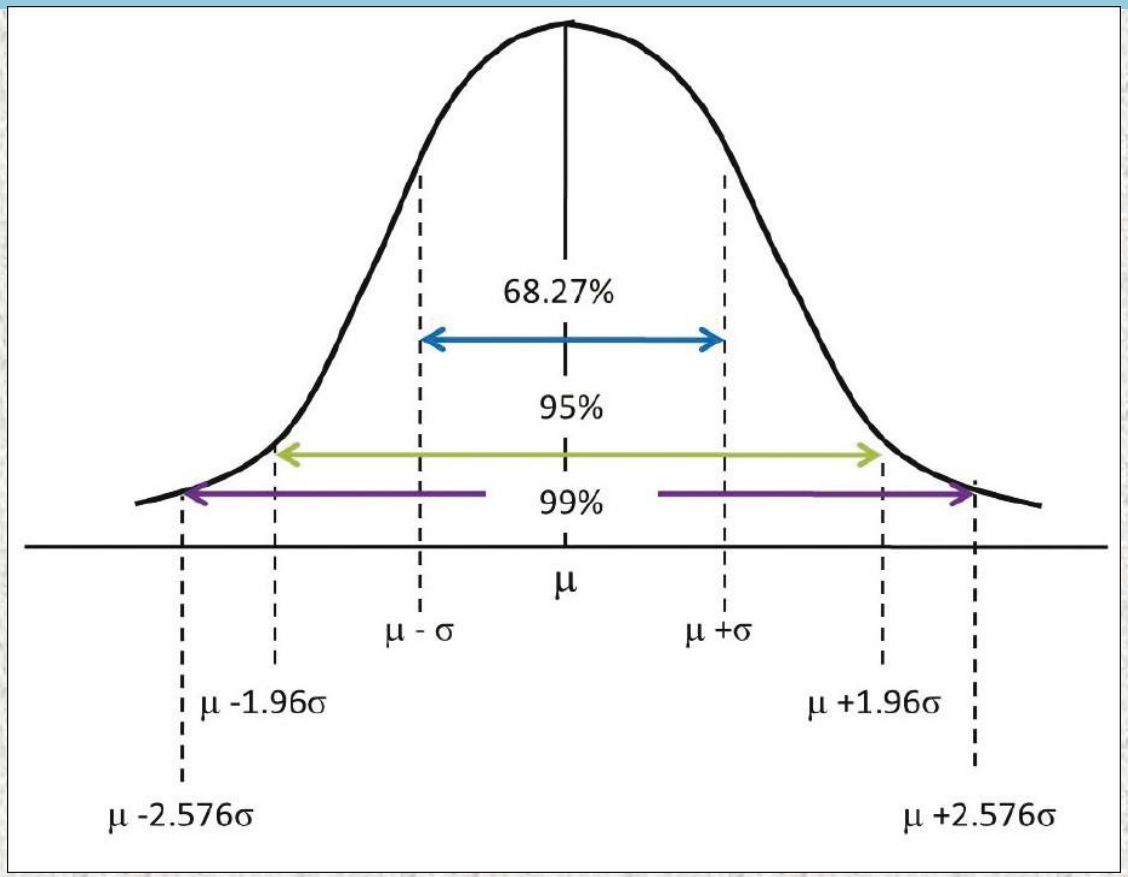


Validity and reliability of data

Is the test measuring what you think it's measuring?



- Validity: the extent to which a test measures what it is supposed to measure; affected by systematic error/bias
- Reliability: the extent to which a test is repeatable and yields consistent scores; affected by random error/bias



For 95% confidence level, margin of error =1.96 $\frac{Std}{\sqrt{n}}$

Sample size and the margin of error

Detection limit (instrument or method)

- Detection limit is the smallest value of measurement that is significantly different from blank
- When measurements are under detection limit, report so as such.
- Methods of treating data below detection limit
  - E.g. USEPA (2000): if the undetected data are less than 15% of the total, use half the detection limit for those values.

Propagation of errors

If z=cx, then $\Delta z = c\Delta x$

If z=x±y, then $\Delta z = \sqrt{\Delta x^2 + \Delta y^2}$

If z=xy or x/y, then $\frac{\Delta z}{z} = \sqrt{(\frac{\Delta x}{x})^2 + (\frac{\Delta y}{y})^2}$

In general,

If z=F(x,y,…), then $\Delta z = \sqrt{(\frac{\partial z}{\partial x}\Delta x)^2 + (\frac{\partial z}{\partial y}\Delta y)^2 + \cdots}$

Uncertainty analysis

- Errors should be specified to one, or at most two significant figures.

- The most precise column in the error should also be the most precise column in the mean value.

4.432 ± 0.00265 should be 4.432 ± 0.003

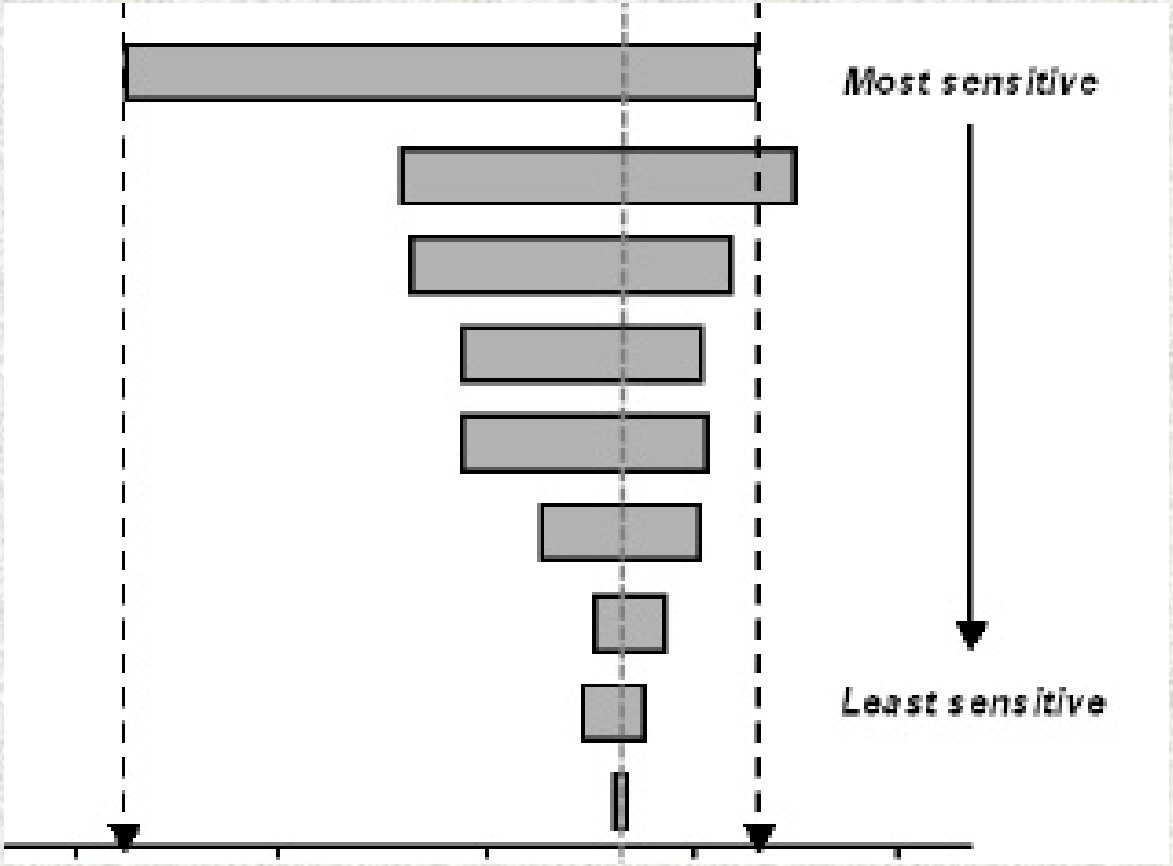
4.432 ± 0.1 should be 4.4 ± 0.1

Significant figures

What model inputs are more important to predictions?

- Partial derivative of the output with respect to an input factor
- Linear regression

Tornado diagram



Sensitivity analysis

Primary and secondary data

- Primary data: first hand data that is collected for a specific purpose.
- Secondary data: second hand data have been collected for some other purpose; may be abstracted from existing published or unpublished sources.
  - may be out of date or inaccurate.
  - should be carefully and critically examined before they are used.
  - may need proper adjustment for new purpose.

- Simple random
- Stratified random
  - random sample from each strata
- Systematic
  - select elements at regular intervals through an ordered list
- Cluster
  - sample groups rather than individuals
- Convenience

Sampling method