intro to statistics
- quantitative technology for empirical science; logic and methodology for measurement of uncertainty and for an examination of that uncertainty
- uncertainty is key word
- stats become necessary when observations are variable
- aims:
  - estimate values of important parameters (things we want to know about)
  - test hypoth about these parameters
  - also about good scientific practice
- a sample of convenience is a collection of individuals that happen to be available at the time
  - e.g. cats falling out of windows shows the higher the fall the less injured they are - what we got was biased data as the data measured were only alive cats, and not dead ones falling from high windows

variables and data
- have to ensure data is not biased
- variable is a characteristic measured on individuals drawn from a population under study
- data are measurements of one or more variables made on a collection of individuals
- types
  - response (dependent)
  - explanatory (independent, predictor)
  - one major use is to relate one variable to another, by examining associations between variables and differences between groups
  - we try to predict or explain a response variable from an explanatory variable

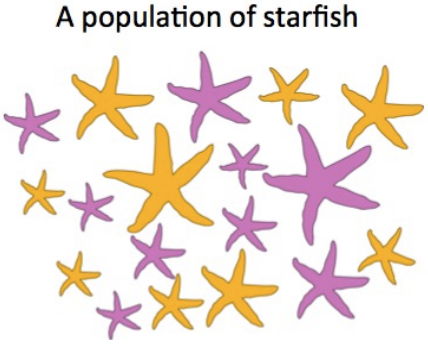
populations and samples
- population - total number of individuals that are used to summarise a group of measurements
  - e.g. mean median standard deviation standard error
- sample
  - much smaller set of individuals from the population
  - an attempt to make representation of population
  - not usually possible or feasible to measure every single individual in entire population

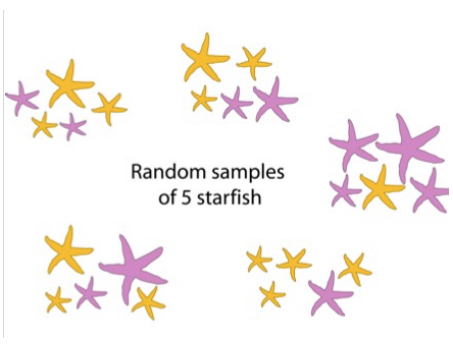
Populations and Samples

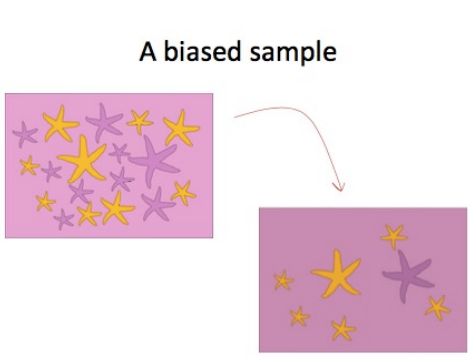
Populations <-> Parameters

Samples <-> Statistics<-> Estimates

parameters and statistics
- parameter
  - is summary describing population
    - e.g means, measures of variation, measures of relationship
    - the truth if you were able to measure all individuals in population
- statistic (estimate)
  - approximation or estimate of the truth
  - subject to error
  - if we could measure every person in population, then we would know the parameter without error, but this is rarely possible
  - we used estimates based on incomplete data (samples) to estimate true values
  - use statistics to determine how good estimates are





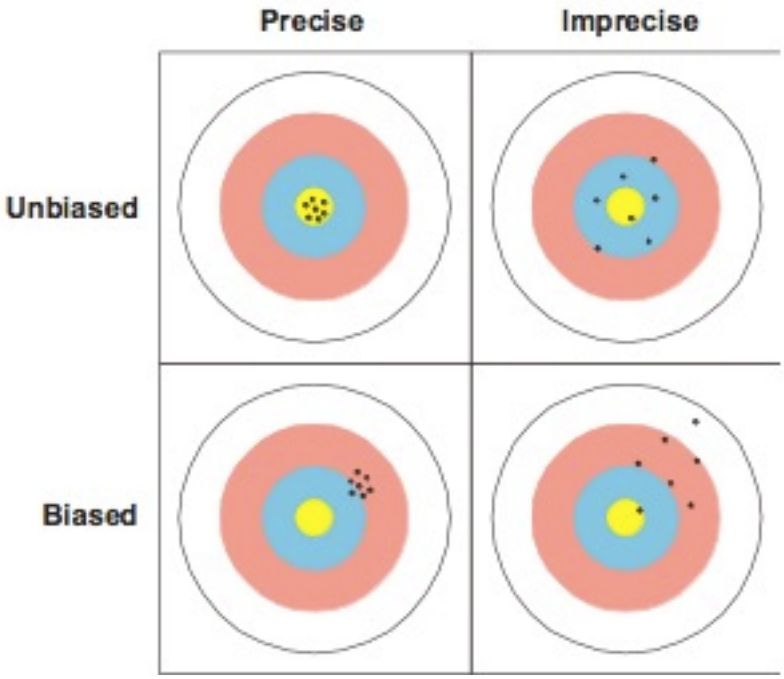


bias: systematic discrepancy between estimates and true population characteristic (value)

example:

1936 US presidential election
- did a poll to figure out who would win
- 2.4mil respondents
- based on questionairre mailed to 10mil people, chosen from telephone books and club lists
- predicted landon wins: landen 57% over roosevelt 43%
- issues as: only people who could afford phones and clubs would respond therefore wealthy people only voted = biased towards one side as wealthy people not representative of whole population

volunteer bias
- volunteers for a study are likely to be different, on average from population
- e.g.
  - volunteers for sex studies are more likely to be open about sex
  - volunteers for medical studies may be sicker than general population

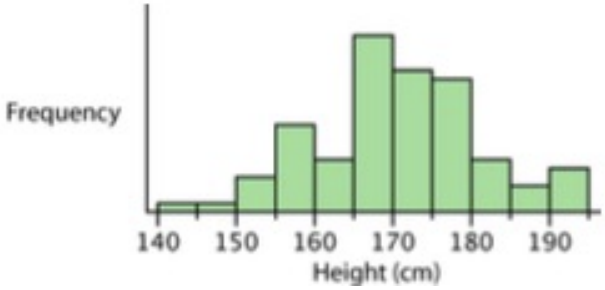


properties of good sample - IN EXAM
- independent selection of individuals
- random selection of individuals
- sufficiently large to ensure random selection can occur

random sampling
- in a random sample, each member of population has an equal and independent chance of being selected
- one way to random sample: give each individual a number, randomly choose numbers
- population parameters are constants whereas estimates (what you measure in experiment) are random variables, changing from one random sample to next from same population
  - e.g. sampling 100 on day 1, 100 on day 2, 100 on day 3, does that add bias? different people?

sampling error
- difference between estimate (what you measure in experiment) and population parameter being estimated caused by chance
- difference between sampling result and true body size in population = sampling error
- larger samples, smaller the sampling error - as sample is larger so capturing more of overall population - more likely to be representative

describing data
- two most common and important descriptions of data:
  - location (or central tendency) - tells us about average or typical individual
    - e.g. mean median mode
  - spread (or variation) - tells us how variable the measurements are from individual to individual



mean: the centre - the estimate of the middle - average then divide by amount - this is the normal one to use

The data:

18 28 24 25 36 14 34

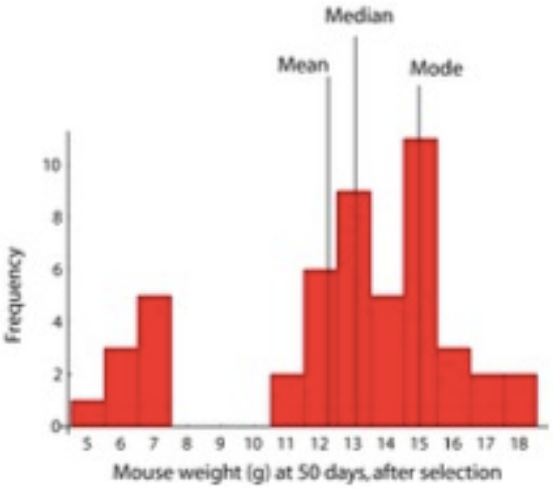
can be put in order:

14 18 24 25 28 34 36

Median is 25.

median: middle measurement in a set of ordered data

mode: most frequent measurement



measures of width (variation)

range: maximum minus minimum

- poor measure of distribution width
  - small samples tend to give lower estimates of range than large samples
  - biased estimator

variance in population

$$\sigma^2 = \frac{\sum_{i=1}^{N} (Y_i - \mu)^2}{N}$$

N is the number of individuals in the population.
μ is the true mean of the population.

Variance of a sample

$$s^2 = \frac{\sum_{i=1}^{n} (Y_i - \bar{Y})^2}{n-1}$$

n is the sample size