

Data and Statistics

Why?

- Management involves collection and analysis of data
- If either activity is not done correctly, your results will be misleading
- Proper data collection and analysis requires a knowledge of statistics, which is based on laws of probability
- Probability can be tricky, do not trust your intuition
- [Let's Make a Deal Paradox](#)

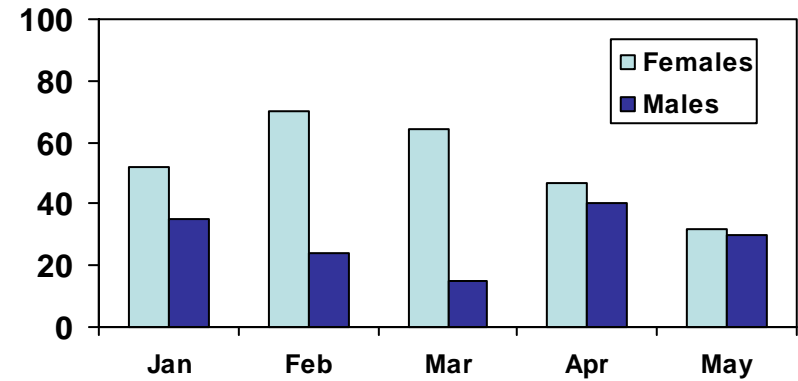
Types of data

- Determines analytical approach
- Nominal
 - Non-ordered categories
 - Examples: sex, eye color
- Ordinal
 - Ordered categories
 - Differences among categories is relative; absolute differences unimportant
 - Other types of data can be reduced to this type by binning
 - Examples: class grades, clothes sizes (S, M, L, XL)
- Interval
 - Ordered with a constant scale
 - No true zero
 - Examples: temperature (C and F but not K), compass bearings, time of day
- Ratio
 - Ordered with constant scale
 - True zero
 - Examples: length, mass, counts, temperature (K)

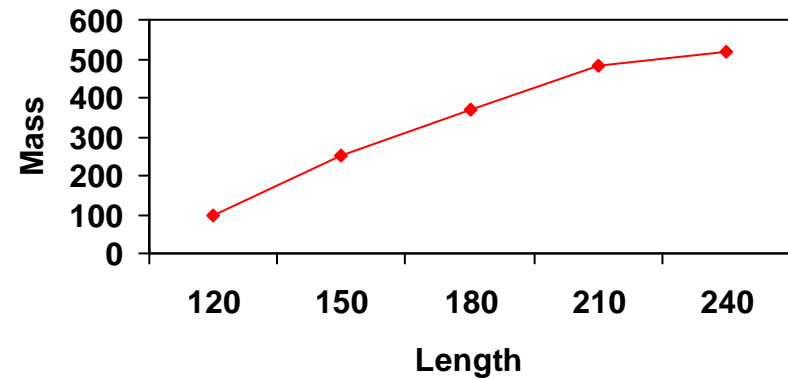
Discrete vs. continuous variables

- Discrete
 - Categorical, only certain values are valid
 - Example: Counts
- Continuous
 - All values possible
 - Can be made discrete by binning
 - Examples: temperature, length
- Differences are reflected in graphs

- Discrete variables are graphed as histograms



- Continuous variables are graphed as lines



Terms

- Population

- A group of things about which inferences are made, existing in a specific area delimited in space and time
- Has certain characteristics
 - central tendency (mean, median mode)
 - dispersion (variance)
- Characteristics describing the population are parameters and are denoted by Greek letters
 - μ , σ^2

Terms

- Sample
 - A subset of individuals of the population that are measured to characterize the entire population through inference
 - Used to generate estimates of population attribute, usually denoted by a “hat” above the symbol, e.g.,
 \hat{N}
 - Sample characteristics are statistics and are denoted by Latin letters

\bar{x}

Difference between sample and population

- Ann Landers survey
- Asked readers to respond to question: “If you had it to do over again, would you have children?”
- 70% said they would not
- Is that result valid?
- No: relied on voluntary response, not sampling
- Survey conducted by university using random sampling indicated 90% would have children

Characterizing dispersion of data

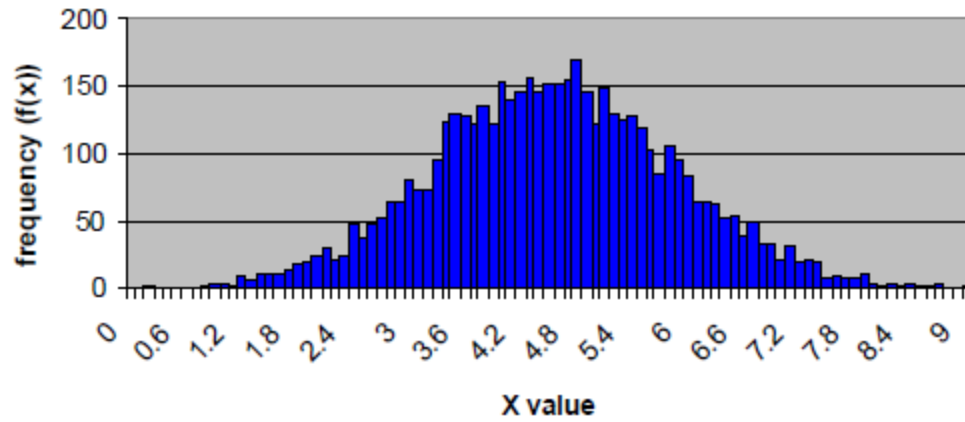
- Formula for sample variance:
- Why that formula?

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

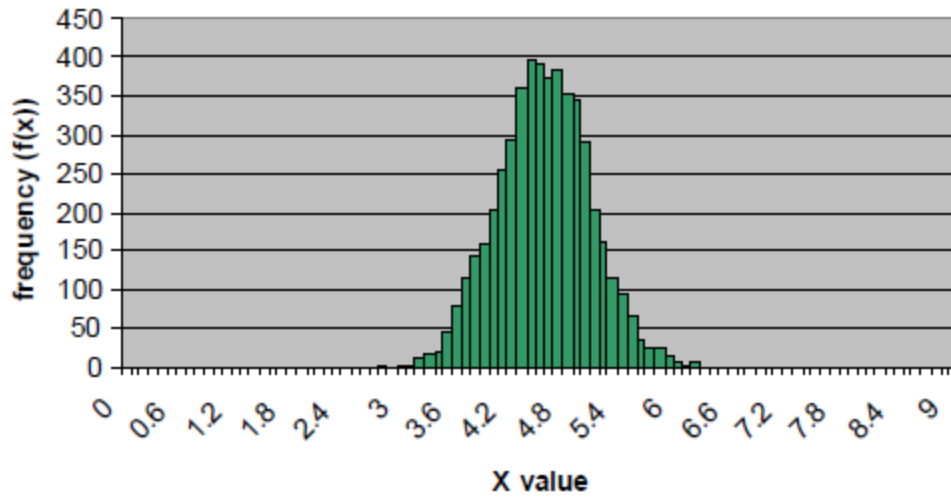
Confidence limits

- The percent confidence you have that the true value of the parameter is within a certain range of values surrounding the estimate
- Based on standard deviation and the underlying statistical distribution
- For a normal distribution,
 - mean $\pm 1s$ includes approximately the central 68% of observations
 - mean $\pm 2s$ includes approximately the central 95% of observations
 - mean $\pm 3s$ includes approximately the central 99% of observations

$N = 5,000$ $\mu = 4.5$, $\sigma = 1.284$

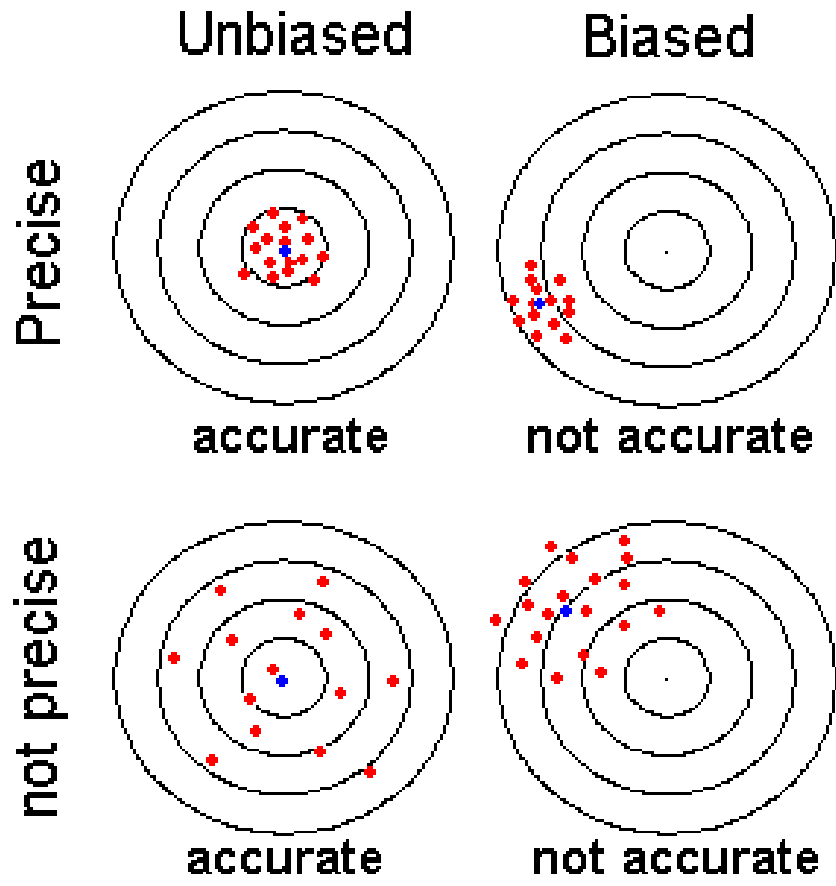


$N = 5,000$ $\mu = 4.5$, $\sigma = 0.5$



Bias and precision

- Accuracy—nearness to true value, measured as mean deviation
- Precision—low variation (similarity of repeated measurements), measured as sampling variance or standard error
- Bias—consistent deviation of an estimator from truth



More on Sampling

Standard errors, Central Limit
Theorem, Bootstrapping

Standard Error Revisited

- Property of the sampling distribution
 - Distribution of means (derived from estimates)
 - Not distribution of data
- Can be estimated from data
 - Variance
 - Sample size

Central Limit Theorem

- As sample size increases, the sampling distribution (distribution of means) approaches a normal distribution, no matter what the underlying distribution of data is
- http://onlinestatbook.com/stat_sim/sampling_dist/index.html
- Why is this important
- Many of the statistics we estimate assume a normal distribution

For instance

- Estimating required sample size
- How many samples must I measure to have a xx% chance of coming within E units of the true mean?
- Clutch size in grouse nests
 - Want to be within ± 2 eggs of true mean 95% of the time
- $n = (t^2 s^2) / E^2$
 - E is specified by researcher
 - s^2 can be obtained by a preliminary sample
 - $t?$
 - # of standard errors on either side of mean enclosed with confidence
 - a function of a normal distribution

Bootstrapping

- Simulation of many sampling efforts from the data of one sampling effort
- Assumes only that the data are representative of the population
- Allows examination of the properties of the samples
 - bias
- Allows estimation of variance when it can't be reliably estimated with one sample

How to bootstrap

- Take sample of size n
- Select a random sample (w/ replacement) of those data
- Repeat that resampling many times to generate sampling distribution
- Jackknife is a similar procedure but samples without replacement

Sampling

- Replication
- Randomization
- Stratification

Replication

- Sample multiple times
- Permits estimation of variation in data
- Affects precision

Randomization

- Pick sampling locations randomly
- Along with replication reduces the effects of *unknown* lurking variables

Stratification

- Allocate samples among identifiable strata that may influence the variable of interest
- Reduces the effects of *known* lurking variables
- Increases precision

- Stratify what you can, randomize what you cannot
- When in doubt, stratify

Last resort

- Systematic sampling
- Weakest of sampling methods