Sample size

Sample size

Issues of concern

- Representativeness
  - How well a sample represents the population

    Determined by the sampling method
- Accuracy
  - How close the sample statistics are to the population parameters

    Determined by the sample size

Sample size

Some terms

- Parameter
  - Summary measure that defines a population

    Population mean μ, proportion π, standard deviation σ
- Statistic
  - Summary measure that defines a sample

    Sample mean "x-bar", proportion p, standard deviation s

Determining size

Arbitrary method

- Pseudo rule-of-thumb
  - Specifies a percentage of the population as the sample size
  - Simple to use
  - No basis for the method
  - Inefficient

Determining size

Conventional approach

- Assumes the existence of a “conventional” sample size
  - Has no scientific basis
  - Previous model might be faulty
  - May result in too large or too small samples
  - May result in too accurate or not accurate enough sample sizes

Determining size

Cost basis approach

- Begins with a “budget” constraint
  - Ignores the value of information
  - May yield too large or too small samples
  - May be sensible to use cost in a comparative sense to select one from among several alternative sample sizes

Determining size

Statistical approach

- Utilizes statistics and probability as the foundation of the sample size
  - Has sound scientific basis
  - Efficient samples
  - Samples with desired accuracy

Variability

Explained

- The measure of variability for normal distribution is its standard deviation (σ for populations, s for samples)

$$\sigma = \sqrt{\frac{\sum (x_i - \mu)^2}{N}} \qquad s = \sqrt{\frac{\sum (x_i - x\ )^2}{n-1}}$$

Variability

Explained

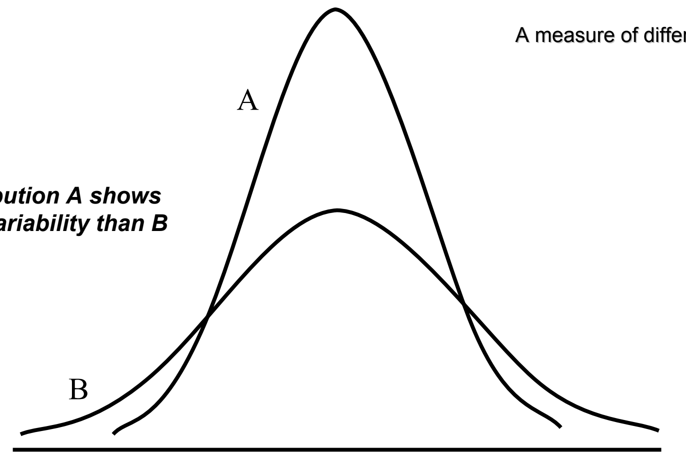
- A measure of the similarities in the value of the elements

  The greater the standard deviation the greater the differences among the elements of the distribution

Variability

A measure of differences

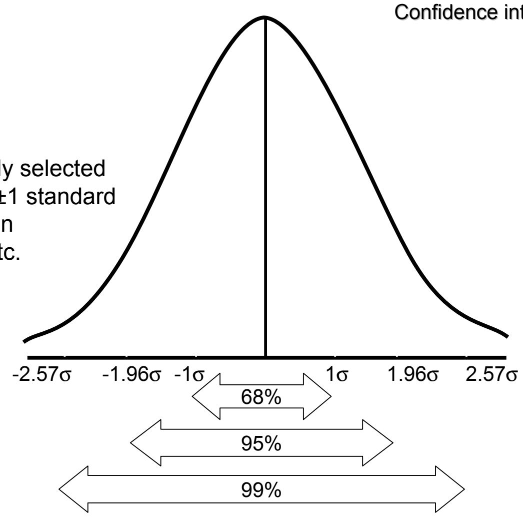
Distribution A shows less variability than B



Normal distribution

Confidence intervals

The value of a randomly selected element will fall within ±1 standard deviation from the mean with 68% probability, etc.



Sampling distribution

Distribution of a sample statistic

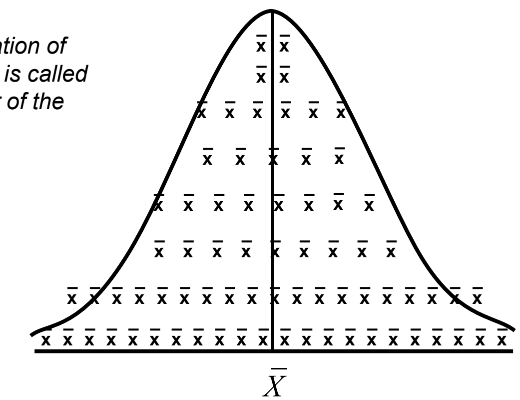
- If all possible samples of a given size were drawn from a population, their sample statistic, say mean, will have a distribution. This is called the sampling distribution of the statistic (mean or proportion)

Sampling distribution

Distribution of sample means

Standard deviation of sample means is called "standard error of the mean"

$\sigma_{\bar{x}}$



Sampling distribution

Central limit theorem

- As the sample size increases, the distribution of the sample statistic approaches normal with the following properties:

$$\bar{X} = \mu \qquad \sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$$

Sample size

Revisited

- Given the previous information and:
  - e   tolerable error
  - z   level of confidence coefficient
  - σ   population standard deviation

$$e = z.\sigma_{\bar{x}} \qquad \sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} \qquad e = \frac{z.\sigma}{\sqrt{n}}$$

$$e^2 = \frac{z^2.\sigma^2}{n} \quad \xrightarrow{\text{Yields}} \quad n = \frac{z^2.\sigma^2}{e^2}$$

Sample size

An example

- Given that:
  - s = 1000
  - e = 100
  - 95 % confidence (z = 1.96 or apprx= 2)
- The sample size:

$$\frac{\mathbf{1000^2 \;.\; 2^2}}{\mathbf{100^2}} \mathbf{= 400}$$

Sample size

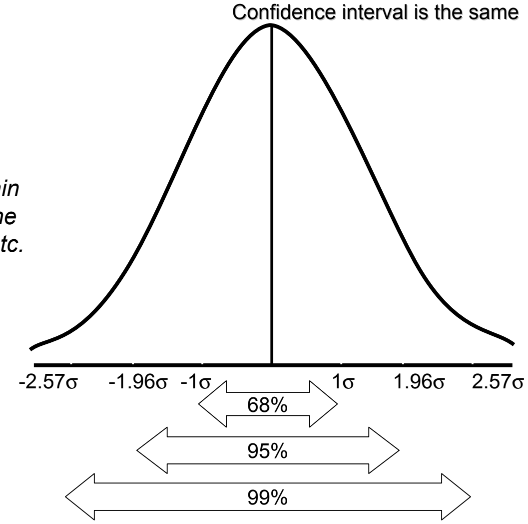
What about proportions

- Sampling distribution of the proportion has similar properties to that of the mean

Normal distribution

Confidence interval is the same

The value of a randomly selected element will fall within ±1 standard deviation from the mean with 68% probability, etc.



Sampling distribution

Dealing with proportions

- The same behavior of the sampling distribution will apply:

$$\sigma_p = \sqrt{\frac{\pi(1-\pi)}{n}}$$

And the sample size becomes:

$$n = \frac{z^2\ \pi(1-\pi)}{e^2}$$

Sampling distribution

Dealing with proportions

- Unlike continuous distributions with an infinite range of variability binomial distributions have a maximum level of variability

Male
Female

Estimating variability

Dealing with proportions

- When dealing with proportions we can use maximum variability (50-50 split) if we have no other information about the population.
  - Assume the worst case which will yield a sample size to accommodate even the most diverse populations
  - If information exists about the population proportion use that

Estimating variability

Dealing with continuous variables

- Standard deviation can assume any positive value. No maximum exists.
- Estimate using a sensible approach
  - Previous studies may provide information
  - Initial sampling to estimate standard deviation
  - One-sixth rule-of-thumb

Estimating variability

Estimate from the sample

- Take a sample and use its standard deviation as an unbiased estimator of the population standard deviation

$$s = \sqrt{\frac{\Sigma(x_i - \bar{x})^2}{n - 1}} \qquad n = \frac{z^2 . s^2}{e^2}$$

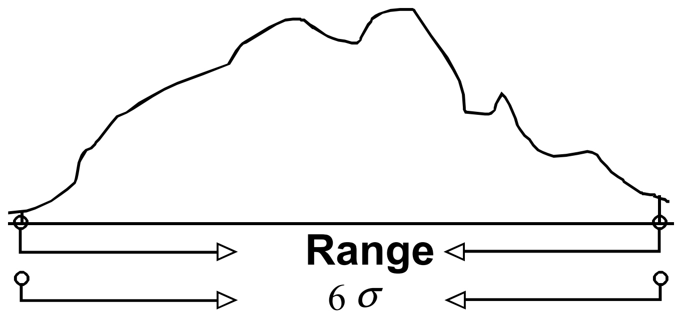
Estimating variability

One-sixth rule-of-thumb

- Many distributions manifest similar behavior to the normal distribution.
  - Approximately 99% of the total population will lie within "3 standard deviations
  - Consequently, the value of the standard deviation will be about 1/6 of the range
  - Estimate the minimum and the maximum values and divide the difference by six to develop an estimate of the standard deviation.

Estimating variability

One-sixth rule-of-thumb



σ = Range / 6