Sampling Variability of a Proportion

Zaw Myo Tun

IRD Global

Descriptive analysis

- Summarising data
  - Frequency, mean, standard deviation, etc
- Estimation
  - Confidence intervals

Statistical Inference

We want to use results from our sample to draw valid conclusions about the population of interest.

Statistical inference

- How do we select our sample?
  - In an unbiased way
  - **Randomly sampled**

- How many individuals should we sample?
  - Lots if possible
  - **But** numbers do not eliminate bias

Statistical inference

- Assume randomly selected
  - Confidence intervals
  - Significance tests

- Account for sampling variation

Categorical data

↓

Variables with 2 categories (binary)

alive/dead
infected/not infected
smoker/non-smoker

↓

Proportion (percentage)

proportion died
proportion infected

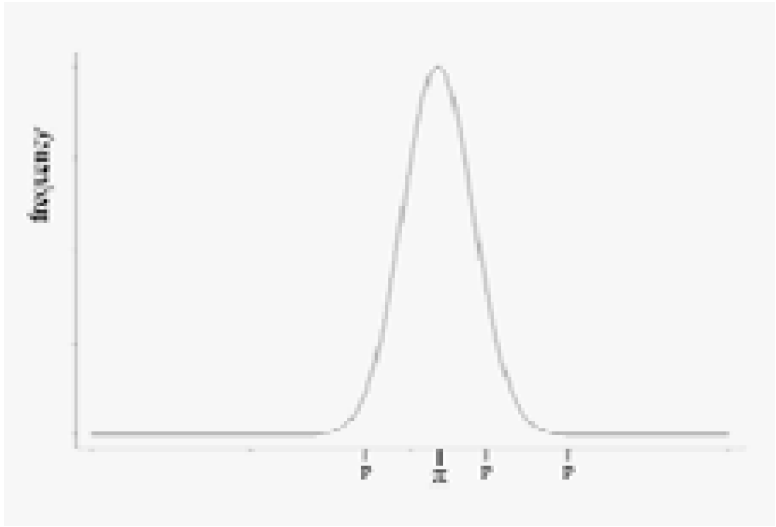
Proportion who smoke 127/335 = 0.379 or 37.9%

Sampling distribution

π = proportion of smokers in the population of interest ('true' proportion)

p = proportion of smokers in the sample

Use p to say something about π



‘Average error’ in p is called the STANDARD ERROR: SE(p)

SE(p) = $$\sqrt{\frac{\pi \times (1-\pi)}{n}}$$

SE(%) = $$\sqrt{\frac{\pi \times (100-\pi)}{n}}$$

Note the influence of n on the SE.

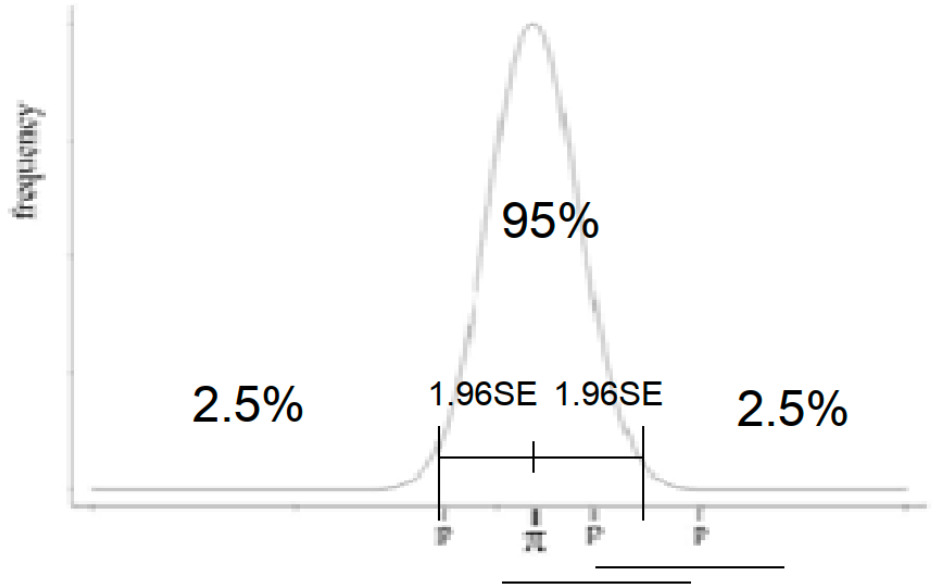
Estimate π with p

$$SE(p) = \sqrt{\frac{p \times (1-p)}{n}}$$

Eg, Survey of 335 men, 127 smoked

p = 127/335 = 0.379 or 37.9%

$$SE(p) = \sqrt{\frac{0.379 \times (1-0.379)}{335}} = 0.0265$$

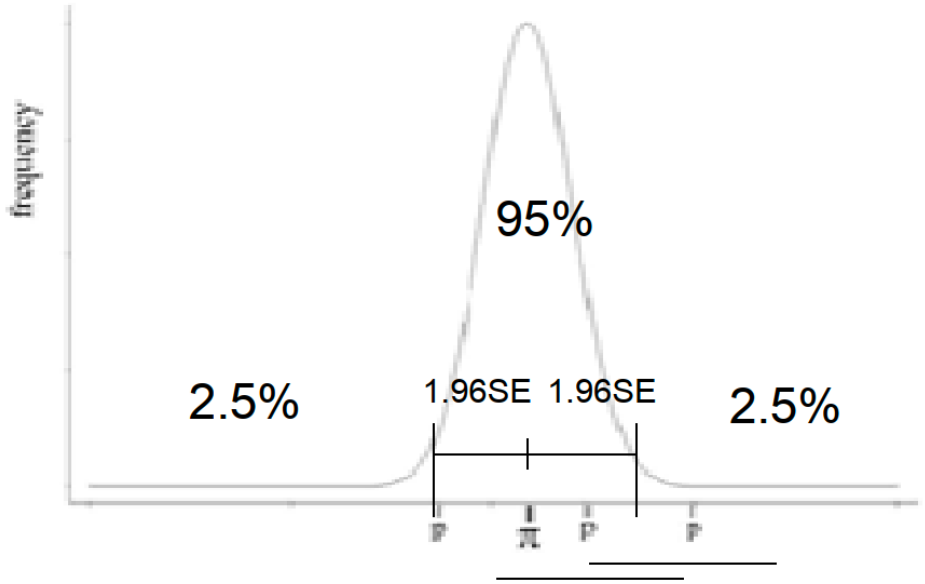


95% confidence interval for π

Graph shows the Sampling Distribution of p

We use properties of a specific statistical distribution known as the Normal Distribution to construct a confidence of π

95% confidence interval for π



p - 1.96x $\sqrt{\frac{p \times (1-p)}{n}}$ to p + 1.96x $\sqrt{\frac{p \times (1-p)}{n}}$

p ± 1.96 x SE (p)

For percentages, the 95%CI is

p - 1.96x $\sqrt{\frac{p \times (100-p)}{n}}$ to p + 1.96x $\sqrt{\frac{p \times (100-p)}{n}}$

Example

- The 95% confidence interval for the true population proportion of smokers in Karachi

$$0.379 \pm 1.96 \times 0.0265$$

$$0.327 \text{ to } 0.431$$

or 32 7% to 43 1% .7% to 43.1%

- There is a 95% chance that this interval includes the true proportion/percentage
- So, our best estimate of the percentage who smoke is 37.9% but we're quite confident that the true percentage lies between 32.7% and 43.1%

What is a ’good’ confidence interval?

Influence of sample size (n)

Summary

- When a sample is taken, random error (sampling variability) occurs. Have to take account of this when interpreting results.
- When measuring a proportion (or percentage) in a sample, this is our best and only estimate of the true proportion (or percentage) in the population.
- A confidence interval gives us two limits which we are reasonably sure include the true proportion.

Summary

- An important factor in determining the width of the CI is the size of the sample (n).
- 95% CIs are the most frequently used.

- General formula: estimate ± 1.96 x SE(estimate)
- Proportion: $p \pm 1.96 \times \sqrt{\{p(1-p)/n\}}$
- Percentage: $\% \pm 1.96 \times \sqrt{\{\% (100-\%)/n\}}$