Lecture 5:
Statistical Inference

Heidi Perry, PhD

Hack University

heidiperryphd@gmail.com

3/8/2016

Overview

Sampling

- A **parameter** is a number that describes a population (e.g. μ and σ in normal distribution.) It is impossible to know without measuring the whole population.
- A **statistic** is a number computed from a sample.
- Statistical inference provides a way to estimate the population parameter from the sample statistics and characterize the uncertainty.

Introduction to Inference

Make a statement about something that is *not observed*, and characterize uncertainty about that statement. Before making an inference:

1. Identify and describe the population.
2. Describe the sampling process.
3. Describe a model for the population, complete with assumptions.

Example: A simple linear model

$$y = \beta_0 + \beta_1 x + \epsilon$$

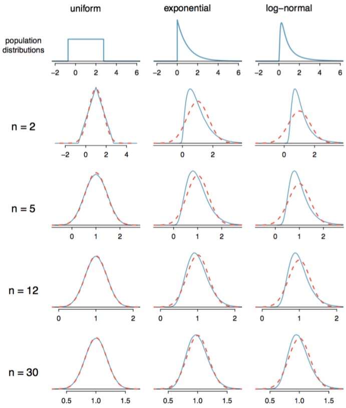
x, y are features of population; β_0 , β_1 describe the relationship,

ϵ is random, making this a statistical model

Central Limit Theorem

Central Limit Theorem

The mean of a large number (> 30) of independent, identically distributed variables will be approximately normal, for all underlying distributions.



Graphic from [Diez, 2016]

Standard Error

Standard error of an estimate

The standard deviation associated with an estimate. It describes the uncertainty associated with the estimate.

Given n independent observations from a population with standard deviation σ, the standard error of the sample mean is:

$$SE = \frac{\sigma}{\sqrt{n}}$$

Since we do not generally have the population standard deviation σ, we use the sample standard deviation s to estimate the standard error.

$$SE \approx \frac{s}{\sqrt{n}}$$

Exercise

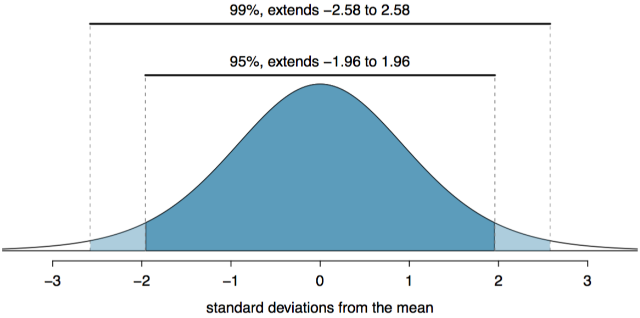
Foundations for statistical inference - Sampling distributions.

Confidence Intervals

- A confidence interval gives a range of possible values of a **population parameter** with a given level of confidence that the parameter is in the range.
- To use the normal distribution in defining a confidence interval, the sample distribution must be nearly normal:
  - The sample observations are independent (a simple random sample consisting of under 10% of the population can be assumed to be independent).
  - The sample size is large (≥ 30 is a good rule of thumb).
  - The population distribution is note strongly skewed (the larger the sample size, the more skew is okay).
- In a confidence interval, $z^* \times SE$ is the **margin of error**.

Confidence Intervals

95% Confidence Interval: point estimate $\pm 1.96 \times SE$
99% Confidence Interval: point estimate $\pm 2.58 \times SE$
Generally, z^* chosen such that the area between $-z^*$ and z^* corresponds to the confidence level.



Graphic 4.10 in [Diez, 2016]

Exercise

Foundations for statistical inference - Confidence intervals

Hypothesis Testing

- Specify the null (H_0) and the alternate (H_A) hypothesis.
- Choose a sample.
- Assess the evidence.
- Draw conclusions.

p-value

p-value provides an estimate of how often the obtained result would occur by chance, if in fact the null hypothesis is true.

A result is statistically significant if it is unlikely to have occurred by chance alone.

Significance Level of a Test

1. The cut-off of what we consider to be "unlikely".
2. Commonly chosen to be $\alpha = 0.05$.
3. If p-value $< \alpha$, we reject the null hypothesis and accept the alternate hypothesis. If p-value $> \alpha$, we fail to reject the null hypothesis.

		Test Conclusion	
		do not reject H_0	reject H_0 in favor of H_A
Truth	H_0 true	okay	Type 1 Error
	H_A true	Type 2 Error	okay

Exercise

Inference for numerical data

References

David Diez, Christopher Barr, & Mine Çetinkaya-Rundel (2015)
OpenIntro Statistics, OpenIntro

Recommended Reading

OpenIntro Statistics, Chapters 4-6
Data Science from Scratch, Chapter 7
Art of Data Science, Chapter 6

Articles about *p*-values and *p*-hacking:

Statisticians Found One Thing They Can Agree On: Its Time To Stop Misusing P-Values
Statisticians issue warning over misuse of P values
I Fooled Millions into Thinking Chocolate Helps Weight Loss. Here's How.
You can't trust what you read about nutrition
Science Isn't Broken
Not Even Scientists Can Easily Explain P-values