Clustering and Prediction

Probability and Statistics for Data Science

CSE594 - Spring 2016

But first,

One final useful statistical technique from Part II

Confidence Intervals

Motivation: p-values tell a nice succinct story but neglect a lot of information.

Estimating a point, approximated as normal (e.g. error or mean)

$$\hat{\mu} = \bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i. \qquad \mathrm{SE}_{\bar{x}} = \frac{s}{\sqrt{n}}$$

$$Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \qquad \left[\overline{x} - 1.96\frac{\sigma}{\sqrt{n}},\ \overline{x} + 1.96\frac{\sigma}{\sqrt{n}}\right]$$

find CI% based on standard normal distribution (i.e. CI% = 95, z = 1.96)

Resampling Techniques Revisited

The bootstrap

- What if we don’t know the distribution?

Resampling Techniques Revisited

The bootstrap

- What if we don't know the distribution?
- *Resample* many potential distributions based on the observed data and find the range that CI% of the data fall in (e.g. mean).

Resample: for each *i* in *n* observations, put all observations in a hat and draw one (all observations are equally likely).

Clustering and Prediction

(now back to our regularly scheduled program)

I. Probability Theory

II. Discovery: Quantitative Research Methods

III. Clustering and Prediction

(now back to our regularly scheduled program)