STATISTICAL ESTIMATION AND PREDICTION

CSCI 452: Data Mining

Still to Come Today

- Statistical Estimation and Prediction
- Measuring Standard Error, Confidence Intervals
- Simple Linear Regression
- Evaluating LM using Testing Set
- More R Basics

Statistical Analysis Motivation

- If estimation and prediction are considered to be data mining tasks, then statistical analysts have been performing data mining for over a century!
- These slides:
  - Examination of some traditional methods of estimation and prediction based on statistical analysis

Univariate Statistical Analysis

- Univariate: analyzing *one* variable a time
  - Point estimation for population means
  - Confidence interval estimation for population means
  - …
- Multivariate: analyzing *more than one* variable

Churn Dataset

- From UCI Machine Learning Repository
- In D2L
- 3333 records (customers)
- 20 predictor variables
- 1 target variable: churn (whether or not customer left the company)

Churn variables

- *State:* categorical (50 states + DC)
- *Account Length:* integer (how long account has been active)
- *Area Code:* categorical
- *Phone Number:* (can be used for customer ID)
- *International Plan:* binary (yes or no)
- *Voice Mail Plan:* binary
- *Number of Voice Mail Messages:* integer
- *Total Day Minutes:* continuous (minutes of day calls by customer)
- *Total Day Calls:* integer
- *Total Day Charge:* continuous

Churn variables (cont.)

- Total Evening Minutes
- Total Evening Calls
- Total Evening Charge
- Total Night Minutes
- Total Night Calls
- Total Night Charge
- Total International Minutes
- Total International Calls
- Total International Charge
- Number of Calls to Customer Server: integer
- Churn: binary (whether or not customer has left the company)

Desired Results

- We would like our findings from analyzing the Churn dataset to be applicable to *all* customers (the population), not just the subset of 3333 customers in the dataset (the sample).
- Sample needs to be *representative* of the population
  - If not, (sample characteristics deviate systematically from the population characteristics), statistical inference should not be applied.

Vocabulary

- Parameter: a characteristic of the population
  - *Example*: mean number of customer service calls, of all phone customers
- Statistic: a characteristic of the sample
  - *Example*: mean number of customer service calls, for customers in the sample
    - *3333 customers in Churn sample, mean is 1.563*
    - *???? Customers in population, mean is ????*
- Values of population parameters are usually *unknown*.

Symbols

	Sample Statistic	…Estimates…	Population Parameter	
Mean	$\bar{x}$	➔	μ	Pronounced "myu"
Standard Deviation	s	➔	σ	"sigma"
Proportion	p	➔	π	"pi"

Estimation

- Point Estimation: the use of a single known value of a statistic to estimate the population parameter
  - *Examples:*
    - Using sample mean to estimate the population mean
    - Using the sample 27^{th} percentile to estimate the population 27^{th} percentile
    - … basically any sample statistic to estimate a population parameter

How confident are we in our estimates?

- Point estimates will "almost always" have some error: <u>sampling error</u>
  - *Example:* Distance between the observed sample mean and the unknown population mean

$$\left|\bar{x} - \mu\right|$$

- Since the true value of the parameter are usually unknown, the value of the sampling error is usually unknown.

How close is the point estimate?

- Point estimates have no measure of confidence in their accuracy.
  - Estimate may be *close* to the value of the target parameter (<u>small sampling error</u>)
  - Estimate may be *far* from the value of the target parameter (<u>large sampling error</u>)

Confidence Interval Estimation

- Confidence Interval Estimate: interval produced from a point estimate, with an associated confidence level specifying the probability that the interval contains the parameter
- *General Form:*
  - *point estimate* $\pm$ *margin of error*
  - "margin of error" as a measure of precision for the estimate

t-interval

- *t*-interval for the population mean:

$$\underbrace{\bar{x}}_{\text{Sample mean}} \pm \underbrace{t_{\alpha/2}\left(\frac{s}{\sqrt{n}}\right)}_{\text{Margin of error}}$$

- *t*-interval may be used, when either:
  1. Population is <u>normal</u>
  2. Sample size is large

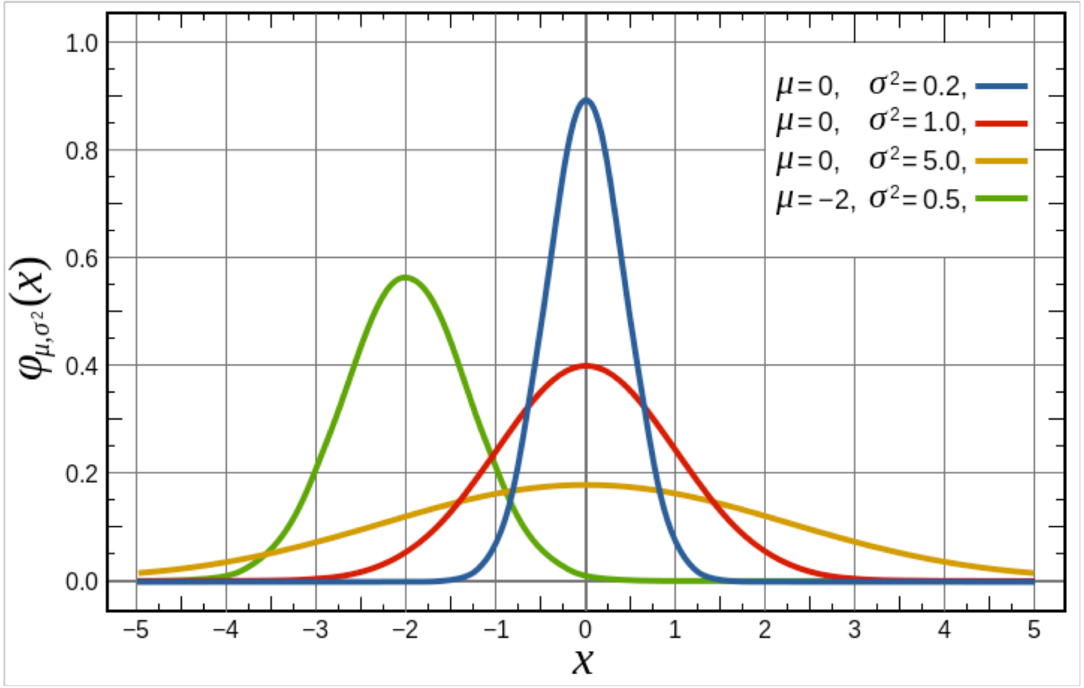
Normal Population Distribution

- Probability Theory
- Normal Distribution also called Gaussian Distribution
  - continuous probability distribution
  - “bell curve” shape

$$f(x,\mu,\sigma) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

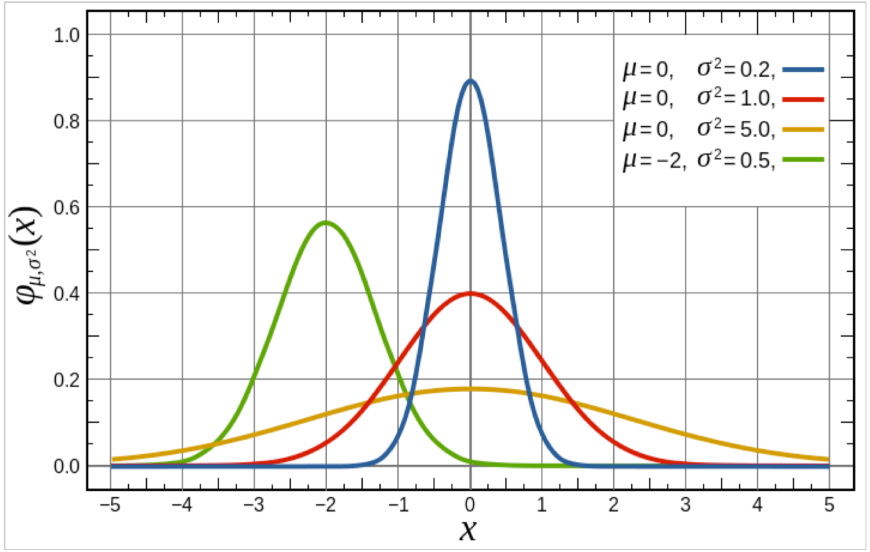
Different shapes depending on specified μ and σ

Normal Population Distribution



"Standard" normal distribution when $\mu=0$ and $\sigma=1$

Normal Population Distribution



- *Notice*: value of the normal distribution approaches zero when x is more than a few standard deviations away from the mean

t-interval

When will margin of error be small?

- *t*-interval for the population mean:

$$\bar{x} \pm t_{\alpha/2}\left(\frac{s}{\sqrt{n}}\right)$$

Sample mean

Margin of error

- *t*-interval may be used, when either:
  1. Population is <u>normal</u>
  2. Sample size is large

Standard Error

Can expect small standard error, whenever:

1. Large sample size
2. Variance is small

- <u>Standard Error:</u> how much you expect a value averaged from several measurements to vary from the true population value
  - Standard deviation divided by root of sample size $\frac{s}{\sqrt{n}}$

- <u>Standard Deviation:</u> how much you expect an individual measurement to vary from the average

t-value

$$\bar{x} \pm t_{\alpha/2}\left(\frac{s}{\sqrt{n}}\right)$$

t-value multiplier

- Dependent on:
  1. sample size
  2. desired confidence level
- Specified by analyst (usually with 95% confidence level)
  - $\alpha = 1 - .95 = .05$

t-values

- http://www.statisticsmentor.com/tables/table_t.htm
- Interested in:
  - Two-tailed α=.05 scores
  - Df ("degrees of freedom") = sample size – 1
    - $df = n - 1$

Churn Example 1

95% *t*-interval for the mean number of customer service calls for all customers:

$$\bar{x} \pm t_{\alpha/2}\left(\frac{s}{\sqrt{n}}\right)$$

$$1.563 \pm 1.96(1.315/\sqrt{3333})$$

$$1.563 \pm 0.045$$

$$(1.518, 1.608)$$

- Reduce sample size to 28
- R

Churn in R

- Loading Churn
- Calculating mean, sd
- Looking up *t*-value
- Performing a one sample *t*-test

Churn Example 2

- Let's only select customers who have:
  - Enrolled in the International Plan
  - Enrolled in the VoiceMail Plan
  - >= 200 day minutes
- Reduces sample from 3333 to 28 customers
  - Still large enough to construct the confidence interval

$$\bar{x} \pm t_{\alpha/2}\left(\frac{s}{\sqrt{n}}\right)$$

$$1.607 \pm 2.051(1.892 / \sqrt{28})$$

$$1.607 \pm 0.733$$

$$(0.874, 2.340)$$

Churn 2 in R

- Selecting subset of customers

How to Reduce the Margin of Error?

- Margin of Error is function of:
  1. *t*-value (depends on confidence level and sample size)
  2. Sample standard deviation (characteristic of data)
  3. *n*, the sample size
- To decrease margin of error:
  1. NO: decrease confidence level
  2. YES: increase sample size

References

- *Data Mining and Business Analytics in R*, 1st edition, Ledolter
- *An Introduction to Statistical Learning*, 1st edition, James et al.
- *Discovering Knowledge in Data*, 2nd edition, Larose et al.