Psych 2111
Prof. Matt Jones

Lecture 8: Probability and Estimation

Probability and Statistics
- Thought experiment
  - Know the population
  - Will sample *n* members at random
- What can we say about probability of sample statistics?
  - $p(M = x)$ for some value x
  - $p(M_1 > M_2)$
  - etc.
- All hypothetical
  - Based on assumptions about population
  - Worked out mathematically
  - Not based on any real data
- Logic behind all inferential statistics
  - *IF* the population has a certain distribution, what will the probabilities be?

Marbles
- Half red, half blue. p(red) =
- 2/3 red, 1/3 blue. p(red) =
- Probability is just a bag of marbles
  - Population = bag
  - Sample = drawing from bag
- Sample four marbles
  - p(2 reds, 2 blues) =

Probability Distributions
- Probability distribution
  - Distribution defined only by probabilities (not frequencies)
  - Can think of as infinite population
- Discrete vs. continuous probability distributions
  - Discrete: like histogram, but $p(x)$ instead of $f(x)$
- Probability is height of each bar
  - Continuous: density function
- Probability is area under the curve
- Random variable
  - Variable defined by probability distribution
  - Can take on one of many values
  - Each value (or range of values) has a probability

Expected Value

Expected value

Mean of a probability distribution

$$E(R) = \sum_x x \cdot p(R = x)$$

x	$p(x)$	$x \cdot p(x)$
0	.0625	0
1	.25	0.25
2	.375	0.75
3	.25	0.75
4	.0625	0.25
	$E(R)$ =	2.00

Properties of Expected Value

Multiplying by a fixed number

I give you \$3 per red: $E(3 \cdot R) = 3 \cdot E(R)$

$E(c \cdot R) = c \cdot E(R)$

Adding a fixed number

I give you \$1 per red, plus a bonus \$1: $E(R + 1) = E(R) + 1$

$E(R + c) = E(R) + c$

Sum of two random variables

$E(R + G) = E(R) + E(G)$

Relationship to mean

Expected value of any random variable equals its mean

Sampling from a Population

Sampling a single item ($n = 1$)

What is $E(X)$?

$$E(X) = \sum_x x \cdot p(x) = \mu$$

Sampling many items

What is $E\left(\sum_{sample} X\right)$?

$$\begin{aligned} E\left(\sum_{sample} X\right) &= E(X[1] + \ldots + X[n]) \\ &= E(X[1]) + \ldots + E(X[n]) \\ &= \mu + \mu + \ldots + \mu = n \cdot \mu \end{aligned}$$

What is $E(\text{mean}(X))$?

$$E\left(\frac{\sum_{sample} X}{n}\right) = \frac{E\left(\sum_{sample} X\right)}{n} = \frac{n \cdot \mu}{n} = \mu$$

$\rightarrow E(M) = \mu$

Unbiased estimators

Statistic a is <u>unbiased estimator</u> of parameter a if $E(a)$ = a

Sample mean is unbiased estimator of population mean

Biased Estimators: The Example of Variance

Population variance

$$\sigma^2 = \frac{\sum_{pop}(X-\mu)^2}{N}$$

$$= mean\left((X-\mu)^2\right)$$

$$= E\left((X-\mu)^2\right)$$

$(X\text{-}\mu)^2$ is unbiased estimator of σ^2

Many observations

$$E\left(\frac{\sum_{sample}(X-\mu)^2}{n}\right) = \sigma^2$$

Problem: We don't know μ

$$\frac{\sum_{sample}(X-M)^2}{n}?$$

Sample mean always shifted toward sample

$$\frac{\sum_{sample}(X-M)^2}{n} < \frac{\sum_{sample}(X-\mu)^2}{n}$$

$$E\left(\frac{\sum_{sample}(X-M)^2}{n}\right) < \sigma^2$$

$\frac{\sum_{sample}(X-M)^2}{n}$ is <u>biased</u> estimate of σ^2

Not a good definition for sample variance

Sample Variance

Goal: Define sample variance to be unbiased estimator of population variance

$E(s^2) = \sigma^2$

Problem: Obvious answer is biased

$$E\left(\frac{\sum_{sample}(X-M)^2}{n}\right) < \sigma^2$$

M is always closer to X than μ is

Solution:

$$s^2 = \frac{\sum_{sample}(X-M)^2}{n-1}$$

Unbiased: $E(s^2) = \sigma^2$

Sample standard deviation

$$s = \sqrt{\frac{\sum_{sample}(X-M)^2}{n-1}}$$

Take-away Points

Random variables and probability distributions

Expected value

Formula: $E(R) = \sum_x x \cdot p(R = x)$

Relationship to mean

Adding and multiplying

Samples and sample statistics as random variables

Biased and unbiased estimators

M is an unbiased estimator of μ: $E(M)$ = μ

$\frac{\sum_{sample} (X - M)^2}{n}$ is a biased estimator of s^2

Sample variance

Formula: $s^2 = \frac{\sum_{sample} (X - M)^2}{n-1}$

Unbiased estimator of population variance

Sample standard deviation

$s = \sqrt{\frac{\sum_{sample} (X - M)^2}{n-1}}$