Statistics and Data Analysis
Lecture 1

Lecture by Dr. Zohar Yakhini
Typeset by Steven Karas

2016-11-06
Last edited 22:57:18 2016-12-11

NOTE: I arrived late due to traffic, and missed the first 15 minutes.

1 Simpsons Paradox

Not really a paradox, but two/three different ways of describing the truth. For example, look at freethrow shots in basketball:

	<1.7m	1.7-1.9	>1.9	Total
Women	4/6	4/6	8/9	16/21
Men	1/2	1/2	23/27	25/31

The women beat the men in each category, yet overall they are worse.

2 Administrative

Tuesday 1830-2100, 15 minutes break. 13 week semester including recitation. Office hours monday 10-12 (C.121). zohar.yakhini@idc.ac.il or zohar.yakhini@gmail.com. 4 homeworks including theoretical and practical on weeks 2,4,7,10. Homework due after two weeks. Each HW worth 18 points. Exam worth 28 points. Individual work unless stated otherwise. Any language acceptable.

Matlab licenses may be available through IDC. GNU Octave is a popular alternative. Python with Scipy/Numpy/Pandas is also possible.

3 Syllabus

1. Intro and review of probability theory
2. Distribution functions
3. Statistical independence
4. Data presentation/visualization
5. Binomial distribution and Central Limit Theorem
6. Statistical inference, confidence intervals, p-values, hypothesis testing
7. Correlation measures
8. Hypergeometric distribution
9. Ranked lists, Wilcoxon rank sum, mHG
10. Multiple testing, Bonferroni, and FDR corrections
11. Entropy and information theory, KL and distribution distances

4 Probability theory and statistics

Probability: given a model, what should we expect to observe? Statistics: given observations, what can we infer about the model?

4.1 Random Variables

Random variable maps from the sample space to some number.

4.2 Probability Distribution Functions

Describes the values a RV can take on, and it's corresponding probability, or density.[1]

Discrete PDFs: $p(x) = P(X = x)$. Assigns the probability that a given discrete value will be observed.

Continuous PDFs: $P(X \in I) = \int_I f(x)dx$, where the function itself is $f(x)$.

Cumulative distribution function: $F(x) = P(X \leq x)$

All probability distribution functions should sum or integrate to 1. $\sum_{x \in X} p(x) = 1$

Given two fair six sided dice, each combination of events has discrete probability $\frac{1}{36}$.

4.2.1 Mean

Aka expected value, but that's a bad name. This is the weighted average value of a RV: $E(X) = \mu = \sum_{x \in X} xp(x)$.

4.2.2 Variance

Average squared deviation between realization of a RV (or PDF) and its mean.[2]

$$V(X) = \sigma^2 = E[(X - E(X))^2] = E[X^2] - E[X]^2$$

4.2.3 Standard Deviation

σ. If you were to observe a RV, it would be more or less this far away from the mean.

4.2.4 Linearity

Linearity is defined as $g(X) = aX + b$ for some constant values a, b.

[1] Scipy refers to these as Probability Mass Functions for discrete distributions, and Probability Density Functions for continuous distributions

[2] full derivation on slide 22

Mean

$$E[aX+b] = \sum_{x \in X} (ax+b)p(x)$$

$$= a\sum_{x \in X} xp(x) + b\sum_{x \in X} p(x) = a\mu + b$$

Variance

$$V[aX+b] = \sum_{x \in X} \left((ax+b) - (a\mu+b)\right)^2 p(x)$$

$$= a^2 \sum_{x \in X} (x-\mu)^2 p(y) = a^2\sigma^2$$

Standard Deviation

$$\sigma_{aX+b} = |a|\sigma$$

Moreover, the sum of linear mean functions is also linear: $E(X+Y) = E(X) + E(Y)$

4.2.5 Coeffecient of Variance

$CV(X) = \frac{\sigma}{\mu}$. More or less the confidence that an observation is close to the mean.

4.2.6 Mode

Most frequently observed RV value. $\max(y \in Y)$

4.2.7 Tchebysheff's theorem

Suppose Y is any random variable with mean μ and stddev σ. Thus: $P(\mu - b\sigma \leq Y \leq \mu + b\sigma) \geq 1 - (\frac{1}{b^2})$ for all $b > 0$. Colloquially, the likelihood of a RV to be within some distance of its mean is strictly bound by some constant multiple of the stddev.

Proof Remember that $V = Var(X) = E[(X-\mu)^2]$. We want to consider specifically values within some distance of the mean. Because they must have greater variance, we relax the equivalence as such: $V \geq b^2 P(|X-\mu| \geq b)$. By limiting our sample space to those within distance b, we can replace x by b with the same equivalence as previously: $\sum xp(x) \geq b^2 \sum p(x)$. $P(|X-\mu| \geq b) \leq \frac{V}{b^2}$ which implies $P(|X-\mu| \geq b \cdot t) \leq \frac{1}{t^2}$ for some t.

5 Common Distributions

5.1 Discrete Uniform Distribution

Rolling a single fair six sided die, flipping a coin, etc.

$$f(x) = \frac{1}{b-(a-1)} \quad a \leq x \leq b$$

Details on slide 28.

5.2 Bernoulli Distribution

Trial with binary outcome. Details on slide 29. Coin tossing, etc.

Binomial experiment Repetitions of n identical, independent Bernoulli trials. p is constant for each trial. Takes two parameters: number of trials and p: $P(Y = k) = \binom{n}{k} p^k (1-p)^{n-k}$. It's important to note that $\sum_{k}^{n} P(Y = k) = 1$.

$$2^n = \sum_{k=0}^{n} \binom{n}{k} 1^k \cdot 1^{n-k}$$

$$(x+y)^n = \sum_{k=0}^{n} \binom{n}{k} x^k \cdot y^{n-k}$$

Mean Linear expectations is quick way to prove $E(n,p) = n \cdot p$

Variance See slides. Linear properties makes proof easy

Standard deviation see slides. Linear properties makes proof easy

6 p values

Given a trial, the p value is the probability of the null hypothesis giving the result of the trial. This can be hacked in many ways.