6.S085 Statistics for Research Projects **IAP 2014**

Lecture 1: January 21

Lecturer: Ramesh Sridharan and George Chen *Notes by: William Li*

Disclaimer: *These notes have not been subjected to the usual scrutiny reserved for formal publications. They may be distributed outside this class only with the permission of the Instructor.*

1.1 Motivating Examples

- polling probability: extreme statistical anomalies with random polling
- URL: `http://www.dailykos.com/story/2010/06/29/880179/-Research-2000-Problems-in-plain-sight)`
- iris recognition: probability of match of iris (used to find the subject of famous National Geographic cover)

1.2 Introduction: Some Concepts in Statistics for Research

1.2.1 Some Definitions

Probability: have model/"truth", want "what kind of data will this give me?"

Statistics: have data, want to find the underlying model/"truth"

Bayesian: hidden model is random

Frequentist: hidden model is fixed but unknown ("there is some fixed value of the model that exists")

This class focuses on classical frequentist methods.

1.2.2 Types of Data

Categorical: red/blue, yes/no

Ordinal: anything that can be ordered: disagree/neutral/agree, etc.

Continuous: numerical, any values

Discrete: we will "lump these in" with one of the other models (categorical, ordinal, continuous)

1.2.3 Random Variables

Working definition: a quantity that takes on random values

Examples: Height of a randomly chosen student; temperature in January

Probability distributions, i.e. for random variable x, $P(x)$

Empirical distribution: based on observed data

Example: Suppose we observe $(1, 1, 3, 4, 7, 8, 8)$; then $p(1) = 2/8$; $p(4) = 1/8$; $p(7) = 2/8$

Expectation: the "average value" that a random variable takes, $E[x] = \sum_a a \cdot P(a)$

Expectation is linear, therefore:

$E[ax + by] = aE[x] + bE[y]$

Example:

```
x    y    x+y
------------
1    3    4
2    4    6
5    3    8
4    3    7
3    4    7
```

(all rows are equally likely)

$E[x] = 15/5 = 3$

$E[y] = 17/5 = 3.4$

$E[x + y] = 32/5 = 6.4$

What if we scramble x and y separately?

```
x    y    x+y
------------
1    3    4
2    3    5
3    3    6
4    4    8
5    4    9
```

$E[x]$, $E[y]$, $E[x + y]$ are the same

No matter how independent/dependent x and y are, expectation is always linear

1.2.4 Variance

$var[x] = \sum_a p(a) \cdot (a - E[x])^2$

$var[ax] = a^2 var[x]$

$var[x + y] = var[x] + var[y]$ IF x, y are independent

standard deviation: $\sqrt{var[x]}$

1.2.5 Notation

μ_x: mean of r.v. x

σ_x: standard deviation of r.v. x

σ_x^2: variance of r.v. x

1.3 Exploratory Analysis

1.3.1 Visualization

When you get some data, you often want to see what's going on by visualizing the data.

Histogram: count frequency of data

```
4
3    x
2 x  x     x
1 x  x  x  x  x
------------------
  1  2  3  4  5
```

Boxplots:

```
          median        end of data that falls in outlier threshold
|-----[     |      ]--------|   .   .   .
   quartile  quartile (75%)          outliers
-----------------------------------------
             x
```

Cumulative Distribution Function: for random variable x, $f(a) = P(x \leq a)$

- The CDF is a monotonically increasing function

For a discrete distribution, it might look something like this:

```
4/4|         o----
3/4|      o--.
2/4|   o--.
1/4| ._.
0/4 o------------
```

Scatter Plot: visualizing two random variables

```
  |
  |     x           x    x
```

```
y |        x   x   x   x
  |          x    x
  |   x        x
  |
  ----------------------------
               x
```

Recommendation: visualize data before jumping into the analysis (you will catch things that you otherwise wouldn't see)

Examples:

- Is your data multimodal? If so, then using the mean to summarize it is not helpful
- Skew: long tail to the right ("right skew"); long tail to the left("left skew") – this will pull the mean further in that direction than the median

The median can be more robust to high values

1.3.2 Quantitative Measures

Sample mean: $\hat{\mu}_x = \frac{1}{n}\sum\limits_i x_i$

Sample Variance: $\sigma_x^2 = \frac{1}{n-1}\sum\limits_i (x_i - \hat{\mu}_x)^2$ (note the $n-1$ in denominator)

Median: 50% of the data is below this value

Mode: most common value

Range: largest - smallest

Is the sample mean a good approximation of the true mean?

- x_i are random
- They have fixed but unknown mean μ_x
- We compute $\hat{\mu}_x$

Insight: $\hat{\mu}_x$ is also a random variable

$$
\begin{aligned}
E[\hat{\mu}_x] &= E[\frac{1}{n}\sum x_i] && (1.1)\\
&= \frac{1}{n}\sum_i \mu_x && (1.2)\\
&= \mu_x && (1.3)
\end{aligned}
$$

Sample variance: why do we have this $n-1$ term?

We underestimate the sample variance because the $(x_i - \hat{\mu}_x)$ term is too small

$E[\hat{\sigma}_x^2] = \sigma_x^2$

Bias: how "wrong" a quantity is

1.3.3 Anscombe's Quartet

Consider 4 datasets with (x, y) pairs:

- same mean in x and y
- same standard deviation in x and y
- same correlation between x and y

same mean in x, same mean in y, same std dev in x and y, same correlation – but four very different datasets!

Question: how many summary statistics do you need to summarize a dataset?

1.3.4 Gaussian/Normal Distribution

$p(x) = a \cdot e^{(x-\mu)^2}$

A Gaussian distribution is very concentrated around its mean

We only need the mean and variance to characterize it

Probability of being within one standard deviation of the mean $\approx 68\%$

Two SDs: 95%

Three SDs: 99%

1.3.5 Bernoulli Distribution

binary random variable:

$Pr(x = 0) = 1 - p$

$Pr(x = 1) = p$

If x is Bernoulli:

$x\ Ber(p)$

$E[x] = p$

$var[x] = p(1 - p)$

1.3.6 Binomial Distribution

sum of n independent and identically distributed (i.i.d.) Bernoulli random variables parameters: n (number of Bernoulli r.v.'s) and p (probability of 1 in Bernoulli r.v.)

If b is binomial

$b\ B$ ("b is distributed as B")

$E[b] = E[\sum x_i] = np$

$var[b] = np(1-p)$

1.3.7 Chi-squared Distribution

Represented as χ^2

If $x_1, ...x_n$, $x_i\ N(0,1)$

$y = \sum x_i^2$, $y\ \chi^2(n)$

parameters: n (degrees of freedom)

1.3.8 Standard Normal

$x\ N(\mu, \sigma^2)$

$y = \frac{x-\mu}{\sigma}$

$y = N(0,1)$

1.4 Endnote: 2004 Election

Bush won all 15 poorest states but only won 36% of the "poor vote"

Kerry won 9 out of 11 of the richest states but only won 38% of the "rich vote"

There is a confounding factor: rate at which

weak dependence on income in Connecticut, a rich state

strong dependence on income in Mississippi, a poor state

Simpson's paradox