Data Science 100
Principles & Techniques of Data Science

Slides by:
Deborah Nolan
deborah_nolan@berkeley.edu
Announcements for Today

- The class has been enlarged and the wait list is operating.
- If you are a graduate student not on the waitlist, try to get on it ASAP.
- Annotated slides are added after class
- HW 2 will be released tonight and due 11:59 Wednesday Sep 11
- Office hours are found at http://ds100.org/fa19/calendar
Topics for Today

- How to solve probability problems
- Review random variables, probability distribution, expectation and variance
- Review Error, Loss, and Risk and the Relationship between the Data and the “World”
- An Example
How do we solve probability problems?
Basic Approaches

- Symmetry and Analogy
- Counting and equally likely
- Trees and conditional probability
Recall our group of 10 mothers

<table>
<thead>
<tr>
<th>Number of Children</th>
<th>Count</th>
<th>Proportion</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>2</td>
<td>20%</td>
</tr>
<tr>
<td>2</td>
<td>4</td>
<td>40%</td>
</tr>
<tr>
<td>3</td>
<td>3</td>
<td>30%</td>
</tr>
<tr>
<td>4+</td>
<td>1</td>
<td>10%</td>
</tr>
</tbody>
</table>

- Select a mother at random from the 10, record her #kids
- Do not replace
- Repeat for a total of 3 samples
Recall our group of 10 mothers

<table>
<thead>
<tr>
<th>Number of Children</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4+</th>
</tr>
</thead>
<tbody>
<tr>
<td>Count</td>
<td>2</td>
<td>4</td>
<td>3</td>
<td>1</td>
</tr>
<tr>
<td>Proportion</td>
<td>20%</td>
<td>40%</td>
<td>30%</td>
<td>10%</td>
</tr>
</tbody>
</table>

What is the chance the second mom selected has 1 child?
Symmetry & Analogy

- Urn with 10 marble one for each mother, indistinguishable except for the # written on it.
- Box with 10 indistinguishable tickets, except for the # on it.
- Deck of 10 indistinguishable cards, except for the # on the flip side.
Symmetry & Analogy

- Draw marbles from well mixed urn
- Select tickets from well mixed box
- Deal cards from top of well shuffled deck
- Deal cards from bottom of well shuffled deck

These 4 scenarios are all equivalent to choosing mothers to participate in a survey.
Symmetry & Analogy

- Chance the second draw is 1

\[ \frac{2}{10} \]

1\(^{st}\) draw & 2\(^{nd}\) draw have same chance of 1 or Ace
Counting

- 10 people named A, B, C, D, E, F, G, H, I, J
- With values 1, 1, 2, 2, 2, 2, 3, 3, 3, 4
- Number of Combinations of first and second draws
- Number of Combinations where the second draw is 1
- Since each combination is equally likely, we take the ratio of these two counts to get the probability
Counting

- Chance the second draw is 1

\[ \text{# comb of 2 moms? \# comb w/ 2nd mom has 1 child} \]

\[ \frac{18}{40} = \frac{9}{20} \]

\[ 10 \times 9 = 90 \]

\[ \text{A, B, B, A} \]
\[ \text{C, A, C, B} \]
\[ \text{J, A, J, B} \]

\[ \frac{78 \times 2}{18 \text{ total}} \]
Tree and Conditioning

Two step process.

Only need to track whether card is 1 or not.

If you know the result of the first draw, compute the conditional chance of the second draw.
Tree and Conditioning

- Chance the second draw is 1

Probabilities go on branches.
Probabilities from the same node sum to 1.
Probabilities are conditional on the information about the path on the tree that was traveled to get there.
Multiply down the tree to get the chance of the final result.

\[
\frac{2}{90} + \frac{16}{90} = \frac{18}{90} = \frac{2}{10}
\]
Many approaches to figuring out probabilities

- Get good at one
- But be flexible and try multiple approaches

FUN PROBLEM: There are 3 cards, one has a circle on both sides, one has a square on both sides, and the third has a circle on one side and square on the other. Mix them up and place one card on the table. It displays a circle. What’s the chance there is a circle on the reverse side?
Go To www.yellkey.com/easy To Register your Answer

Mix Up
Pick 1
Put on Table
We see What’s the chance circle on the other side?

\[
\frac{\frac{1}{6} + \frac{1}{6} + \frac{1}{6}}{\frac{1}{6} + \frac{1}{6} + \frac{1}{6}} = \frac{2}{3}
\]
Working formally with Random Variables
0-1 Random Variables

- In discussion yesterday, you worked with random variables that take on the 0 or 1 values.
- We will start with it as an example.

\[
X = 0 \text{ with prob } 1 - p \quad \text{or} \quad X = 1 \text{ with prob } p
\]
Examples?

Medical Trials
Games of Chance
Survey Results
Random occurrences in Nature
Probability Distribution

\[
\begin{array}{c|cc}
  x & 0 & 1 \\ 
  P(x) & 1-p & p \\
\end{array}
\]

\[
E(X) = 0 \times (1-p) + 1 \times p = p
\]

\[
\text{Var}(X) = E((X-p)^2) = (0-p)^2(1-p) + (1-p)^2p = p^2(1-p) + p(1-p)^2 = p(1-p)[p+1-p] = p(1-p)
\]
Expected Value and Variance

$$E(X) = p$$

$$Var(X) = p(1-p)$$

For what value of $p$ is $\text{Var}(X)$ largest?
More Generally, Expected Value and Variance of a Discrete RV

Probability Distribution

\[ E(X) = \sum_{j=1}^{m} x_j P_j = \mu \]

\[ \text{Var}(X) = \sum_{j=1}^{m} (x_j - \mu)^2 P_j = \sigma^2 \]
More Generally

\[
\mathbb{E}(aX + b) = a\mathbb{E}(X) + b
\]

\[
\text{Var}(aX + b) = a^2 \text{Var}(X)
\]

\[
\mathbb{E}[aX + b - (a\mathbb{E}(X) + b)]^2
\]

\[
= \mathbb{E}[aX - a\mathbb{E}(X)]^2
\]

\[
= \sum_j (aX_j - a\mathbb{E}(X))^2 \rho_j
\]

\[
= a^2 \sum_j (X_j - \mathbb{E}(X))^2 \rho_j
\]

\[
= a^2 \text{Var}(X)
\]
Sums of 0-1 Random Variables

\[ X_i = 0 \text{ with prob } 1 - p \]

\[ = 1 \text{ with prob } p \quad \text{for } i = 1, \ldots, n \]

Examples?

- SRS moms
- 1 child or not
- 0 more than 1
- Spin roulette wheel
- 1 lands 17
- 0 other #

Draw what up? # of trials

\[ \#17 \]

\[ \frac{38}{36} \]

\[ p = \frac{1}{38} \]
Expected Value

$$E(X_1 + \cdots + X_n) = E(X_1) + \cdots + E(X_n)$$

for both ind and dep $X_i$ is

\[ = np \]
Variance

If Independent

\[ \text{Var}(X_1 + \cdots + X_n) = \text{Var}(X_1) + \cdots + \text{Var}(X_n) = np(1-p) \]

If From a Simple Random Sample

\[ \text{Var}(X_1 + \cdots + X_n) = \frac{np(1-p)}{N-n} \]

\[ \text{N is pop size} \]

\[ \text{With large populations there is little difference between SRS & independent draws} \]

\[ \text{Finite pop correction factor} \]
Concrete: $n = 4$ and $Y = X_1 + X_2 + X_3 + X_4$ and the $X$s are independent with same chance of 0 or 1 (knowing the value of $X_1$ doesn’t change $X_2$ distribution).

Given $n = 4$, $Y = 2$ can be achieved by the following combinations of $X_1$, $X_2$, $X_3$, and $X_4$:

- 1100
- 1010
- 1001
- 0110
- 0101
- 0011

There are 6 ways to achieve $Y = 2$. The probability of each combination is the same since each $X$ is independent with a chance of 0 or 1. Therefore, the probability of $Y = 2$ is:

$$P(Y = 2) = \frac{6}{6} = 1$$

Recall $4! = 4 \times 3 \times 2 \times 1 = 24$, and $0! = 1$. The probability of $Y = 2$ is given by the binomial distribution:

$$P(Y = 2) = \binom{4}{2} p^2 (1-p)^2$$

where $p = \frac{1}{38}$. Thus,

$$P(Y = 2) = \binom{4}{2} \left(\frac{1}{38}\right)^2 \left(1 - \frac{1}{38}\right)^2$$

Roulette wheel:

$$p = \frac{1}{38}$$

$X_i = 1$ if $i \leq 17$, 0 otherwise.
Probability Distribution

$n$ independent 0-1 variables

\[ Y = X_1 + \ldots + X_n \]

\[ P(X_i = 1) = p \]

\[ P(Y = k) = \binom{n}{k} p^k (1 - p)^{n-k} \text{ for } k = 1, \ldots, n \]

\( \text{Binomial Dist}(n, p) \)
Fun Problem Related to HW:

- Roll a fair die 5 times.
- \( N_E = \) number of evens,
- \( N_P = \) number of primes
- \( N_1 = \) number of 1s

\[
P(N_1 = 1, N_P = 2, N_E = 2) = \]

\[
\text{Chance: } \frac{1}{6} \times \frac{1}{2} \times \frac{1}{2} \times \frac{1}{3} \times \frac{1}{3}
\]

\[
= \left( \frac{1}{6} \right) \left( \frac{1}{2} \right)^2 \left( \frac{1}{3} \right)^2
\]

\[
\Rightarrow
\]
5!  # Arrangements of 5 unique things

1! 2! 2!

The E's are interchangeable.

The T's are interchangeable.

\[
P(N_I = 1, N_T = 2, N_E = 2) = \frac{5!}{1! 2! 2!} \left( \frac{1}{6} \right)^1 \left( \frac{1}{3} \right)^2 \left( \frac{1}{2} \right)^2
\]

This is a trinomial distribution.

The parameters are: \( n = 5 \) \( (p_1, p_2, p_3) \)

\( \frac{1}{6} \quad \frac{1}{3} \quad \frac{1}{2} \)
Skip Ahead to Slide 53

The next 7 slides are a review of Error, Loss, & Risk

Summary Statistics as Estimators of Population Parameters
Data Life Cycle

Generalization

Get Data from the World and Generalize Data Findings to the world

Data Design/Generation
The Simple Random Sample

- Suppose we have a population with $N$ subjects
- We want to sample $n$ of them
- The SRS is a random sample where every unique subset of $n$ subjects has the same chance of appearing in the sample
- This means each person is equally likely to be in the sample
Empirical (Data)

DATA: $x_1, x_2, \ldots, x_n$

The sample that we have to work with

Model (World)

Random Variables: $X_1, X_2, \ldots, X_n$

Probability distribution from, e.g., a SRS from the population
Empirical (Data)

DATA: $x_1, x_2, \ldots, x_n$

Summary statistic that minimizes the empirical risk

$$\frac{1}{n} \sum_{i=1}^{n} l(x_i - c)$$

Model (World)

Random Variables: $X_1, X_2, \ldots, X_n$

Probability parameter that minimizes the Risk

$$\mathbb{E}l(X - c)$$
Empirical (Data)

DATA: $x_1, x_2, \ldots, x_n$

Summary statistic that minimizes the empirical risk

For $l_2$ loss, $\bar{x}$ minimizes the average loss

Model (World)

Random Variables: $X_1, X_2, \ldots, X_n$

Probability parameter that minimizes the Risk

For $l_2$ loss, $\mathbb{E}(X)$ minimizes the average loss
Empirical (Data)

Connect the sample average and expected value:

$$E(\bar{X}) = E(X)$$

The expected value of a sample average from a SRS is unbiased.

Its variability is quantifiable – the sampling error.

Model (World)

$\bar{X}$ is a random variable.
Data Life Cycle

Probability Samples give us Representative Data where the sample average is well behaved and an accurate estimate of the population average.
An Example: Wait Time for a Repair
Data Life Cycle

Question Formulation

? 

Generalization

Data Design/Generation

Data Analysis
Question

What is the typical wait time for a PG&E repair?

Context

PG&E must report to a utilities commission about its service record.
How might we/they focus this question?

The Question gives focus to the Population that we want to study
What is the Population of Interest?

What is the Sampling Frame?
Scenario: Administrative Data
The Data

\(x_1, x_2, \ldots, x_n\) every wait time over a 3 month period

Can we provide a summary statistic?
Why is the sample median such a desirable summary?
Summarizing the Data

DATA: $x_1, x_2, ..., x_n$ where $n$ is 1665 for our data

ERROR: $x_1 - c, x_2 - c, ..., x_n - c$

LOSS: $l: \mathbb{R} \to \mathbb{R}^+$
Minimize the Average $L_1$ Loss

$$\frac{1}{n} \sum_{i=1}^{n} l(x_i - c) = \frac{1}{n} \sum_{i=1}^{n} |x_i - c|$$
Minimize the Average Absolute Error

\[ \frac{1}{n} \sum_{i=1}^{n} |x_i - c| \]
Probability Samples give us Representative Data where the sample median is a good estimate of the population.
Where does Probability Sampling Come into this Problem?
Probabilistic Behavior of the Median

- Not as simple to work with as the mean
- We need to make more assumptions about the underlying probability distribution of $X$
- In many circumstances the sample median is well-behave and close to the median($X$)
HW 2
Introduction
2016 Presidential Election

- Outcome took many by surprise
- Most polls were predicting Clinton victory was 90%
- FiveThirtyEight said 70% and a couple of days before indicated that Trump had a chance to win
- Now that the election has passed, we have the opportunity to see the world (voters who voted in the election)
Population: Pennsylvania voters

We have a record on the
# Trump votes
# Clinton votes
# Other votes

We can simulate the polls to see the sampling distribution of:

\[
\frac{(# \text{T votes} - # \text{C votes})}{\text{Total Votes Sampled}}
\]
Population: Pennsylvania voters

We can introduce a little bias

Simulate the polls to see the sampling distribution of the biased sampling frame:

\[
\frac{(# T \text{ votes} - # C \text{ votes})}{\text{Total Votes Sampled}}
\]