Data Science 100
Principles & Techniques of Data Science

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

- We have asked for permission to increase the class size to enroll about 100 people from the wait list.... Stay tuned
- We will try using Google forms today
- Slides and notes from lecture available online at http://ds100.org/fa19
- HW 1 is due 11:59 Wednesday Sep 4
- Office hours are found at http://ds100.org/fa19/calendar
We will give a simple example.

Data Life Cycle

Question Formulation

These 2 pieces are crucial

Generalization

Most of our focus is on this part of the cycle.
START SIMPLE

QUESTION:
What is the typical family size (children only)?
### DATA:

Survey of all 250 students enrolled in Data 100 in Fall 2017 and asked their family size

<table>
<thead>
<tr>
<th></th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4+</th>
</tr>
</thead>
<tbody>
<tr>
<td>Counts</td>
<td>61</td>
<td>131</td>
<td>44</td>
<td>14</td>
</tr>
<tr>
<td>Percent</td>
<td>24%</td>
<td>52%</td>
<td>18%</td>
<td>6%</td>
</tr>
</tbody>
</table>
START SIMPLE

ANALYSIS:

<table>
<thead>
<tr>
<th></th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4+</th>
</tr>
</thead>
<tbody>
<tr>
<td>Counts</td>
<td>61</td>
<td>131</td>
<td>44</td>
<td>14</td>
</tr>
<tr>
<td>Percent</td>
<td>24%</td>
<td>52%</td>
<td>18%</td>
<td>6%</td>
</tr>
</tbody>
</table>

Can we provide a summary statistic?

Bar chart is a good visual summary.

For now, we ignore the 4+ and treat it as 4.

How about the mean?
DETOUR: Why is the sample mean such a desirable summary?
Summarizing the Data

DATA: $x_1, x_2, ..., x_n$ where $n$ is 250 in our example

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

LOSS: $l: R \rightarrow R^+$

The loss function maps errors to the nonnegative values. It represents the 'cost' of an error.

We want a single numeric summary of our data: $c$

We want $c$ to be close to our data.

So, we look at the error between an observation and $c$: $x_i - c$

If $c$ is 2 and $x_i$ is 2 then the error is 0
If $x_i$ is 4 then it is 2
Summarizing the Data

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

AKA EMPIRICAL RISK

Minimize the empirical risk

We want to \( \min_{C} \frac{1}{n} \sum_{i=1}^{n} l(x_i - c) \)

We need to specify the loss function to do this.
Minimize the Average Loss

$$\frac{1}{n} \sum_{i=1}^{n} l(x_i - c) = \frac{1}{n} \sum_{i=1}^{n} (x_i - c)^2$$

This is $l_2$ loss.
We also call it squared error.

Before we minimize

We give a short refresher about sums and averages

It is the most commonly used loss function because it has several useful properties
Refresher

\[
\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i
\]

Recall the sample mean

\[
\frac{1}{n} \sum_{i=1}^{n} (ax_i + b) = a\bar{x} + b
\]

\[
= \frac{1}{n} \sum_{i=1}^{n} ax_i + \frac{1}{n} \sum_{i=1}^{n} b
\]

\[
= a \frac{1}{n} \sum_{i=1}^{n} x_i + \frac{1}{n} nb
\]

\[
= a \bar{x} + b
\]

We will use this property several times
Minimize the Average Loss

\[
\frac{1}{n} \sum_{i=1}^{n} (x_i - c)^2 = \frac{1}{n} \sum_{i=1}^{n} \left( x_i - \bar{x} + \bar{x} - c \right)^2
\]

A simple approach that does not involve calculus

\[
= \frac{1}{n} \sum_{i=1}^{n} \left[ (x_i - \bar{x})^2 + 2(\bar{x} - c)(x_i - \bar{x}) + (\bar{x} - c)^2 \right]
\]

\[
= \frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})^2 + 2(\bar{x} - c) \frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x}) + (\bar{x} - c)^2
\]

\[
= 0
\]

\[
\frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x}) = \frac{1}{n} \sum_{i=1}^{n} x_i - \frac{1}{n} \sum_{i=1}^{n} \bar{x}
\]

\[
= \frac{\sum_{i=1}^{n} x_i}{n} - \bar{x}
\]

\[
= \frac{\sum_{i=1}^{n} x_i - n\bar{x}}{n}
\]

\[
= \frac{\sum_{i=1}^{n} x_i}{n} - \bar{x}
\]
We have

$$\frac{1}{n} \sum_{i=1}^{n} (x_i - c)^2 = \frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})^2 + (\bar{x} - c)^2$$

To minimize w.r.t. $c$, there is no choice.

The minimum is when $c = \bar{x}$
The Sample Average Minimizes Empirical Risk

\[
\frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})^2 \leq \frac{1}{n} \sum_{i=1}^{n} (x_i - c)^2
\]

This is the Sample Variance
Data Life Cycle

Let's step back and consider the question

Question Formulation

In some cases we do not want/need to Generalize

When we have data about all of the individuals that interest us, we don't need to generalize further.

Data Design/Generation

Data Analysis
Consider the Question Carefully

What is the typical family size (children only)?

How well can we measure this?

What are we trying to measure?

Families come in all different shapes and sizes.

Suppose we are most interested in the number of children a woman gives birth to.
Focus the Question

From Female Fertility Perspective:

Some Questions
WHERE
WHEN
WHO
HELP
WHAT

US 2016

#births

Females 40-44

(we may be interested in comparing # children a woman has today to 20 or 50 years ago)

Our Question

Some Questions
WHERE
WHEN
WHO
HELP
WHAT

US 2016

#births

Females 40-44

(we may be interested in comparing # children a woman has today to 20 or 50 years ago)
Focus the Question

The Question gives focus to the Population that we want to study
Data Life Cycle

In order to generalize from data to the population of interest, our sample needs to look like the population.
How Well Does our Data 100 class represent the group of interest?

- Mothers of children at UC Berkeley
- Measure the mothers via the children
- Mothers who are 40-44 in 2014

How might these characteristics impact the estimate of the number of children a US woman bears in her lifetime in 2014?

Bias up, Bias Down, Not impact

A mother w/ 4 children has more chances of getting into the sample than a mother w/ 1 child. This is called **Size Biased Sampling**.
According to Pew Research Center

How might this impact the Data 100 average?

The data used in these analyses are designed to assess women’s fertility, and as such a “mother” is here defined as any woman who has given birth. However, many women who do not bear their own children are indeed mothers.

http://www.pewsocialtrends.org/2015/05/07/family-size-among-mothers/
The individuals we want to study
Population of Interest

Sampling Frame

Access the Population through the Frame

Some individuals might not even be in the population — see e.g., the title Report Discussion.

These individuals are not reachable through the sampling frame.
Population of Interest

Sampling Frame

The Sample is a subset of the Frame

How we select the individuals from the sampling frame matters
Sample = Sampling Frame = Population

Scenario: Census

All individuals are studied
Sampling Frame = Population

Scenario: Access to all members of the Population when sampling

We often assume away the difference between the frame & population
Sampling Frame = Sample

Scenario:
Administrative Data

With Admin Data we have access to all in our Frame.
Population of Interest

Sampling Frame

Most Common Scenario

SAMPLE
How are the data generated?

- What is the population of interest?
- What is the sampling frame?
- How are the data generated?

We will turn our focus to this question.
DETOUR:
1. The simple random sample
2. Why is a probability sample so desirable?
Sampling Frame = Population

Scenario: Access to all members of the Population when sampling

HOW IS THE SAMPLE TAKEN?
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

There are $\binom{N}{n}$ possible samples of size $n$ from $N$

Recall that

$$\binom{N}{n} = \frac{N!}{h! (N-h)!}$$

Convince yourself of this with a simple example
The Advantages of a SRS

- Representative: The sample tends to look like the population
- Statistics based on the sample tend to be close to statistics based on the population
- We can provide typical deviations of sample statistics from population values.
- AND MORE...

\[
\begin{align*}
N &= 4 \\
n &= 2 \\
\binom{4}{2} &= \frac{4!}{2!2!} = \frac{4 \times 3}{2 \times 1} = 6
\end{align*}
\]

Possible samples of size 2:

- \((A, B)\)
- \((A, C)\)
- \((A, D)\)
- \((B, C)\)
- \((B, D)\)
- \((C, D)\)

6 samples of size 2
Start Simple

➢ Suppose our population contains only 10 mothers and we take a *Simple Random Sample* of 3 for our survey.

<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>

There are \( \binom{10}{3} \) possible samples

\[
\frac{10!}{3! \cdot 7!} = \frac{10 \times 9 \times 8}{3 \times 2 \times 1} = 120
\]
Formal Set Up

One way to think about taking the sample:
Write each mother’s value on a ticket.
Put the tickets in an urn.
Mix.
Draw one at a time.

Without Replacement

<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>

\[ X_1 \text{ The number of children for the first mother chosen} \]
\[ X_2 \text{ The number of children for the second mother chosen} \]
\[ X_3 \text{ The number of children for the third mother chosen} \]
Formal Set Up

$$X_1$$: The number of children for the first mother chosen

<table>
<thead>
<tr>
<th>Number of Children</th>
<th>Count</th>
<th>Percent</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>

<table>
<thead>
<tr>
<th>Probability Distribution</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4+</th>
</tr>
</thead>
<tbody>
<tr>
<td>$$P(X_1 = x)$$</td>
<td>20%</td>
<td>40%</td>
<td>30%</td>
<td>10%</td>
</tr>
</tbody>
</table>

$$P(X_1 = 1) = \text{Chance drawn a 1 from the urn}$$

$$= \frac{2}{10} \quad \text{# 1's}$$

$$= \frac{9}{10} \quad \text{# tickets}$$
X_1 \text{ The number of children for the first mother chosen}

What is the expected value of X_1?

\[
E(X_1) = \sum_{j=1}^{4} x_j P(X_1 = x_j)
\]

\[
= 1 \times \frac{2}{10} + 2 \times \frac{4}{10} + 3 \times \frac{3}{10} + 4 \times \frac{1}{10}
\]

\[
= 2.3
\]
$X_2$ number of children for the 2$^{nd}$ mother chosen

<table>
<thead>
<tr>
<th>$x$</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4+</th>
</tr>
</thead>
<tbody>
<tr>
<td>$P(X_2 = x)$</td>
<td>$\frac{3}{10}$</td>
<td>$\frac{4}{10}$</td>
<td>$\frac{3}{10}$</td>
<td>$\frac{1}{10}$</td>
</tr>
</tbody>
</table>

For example, we could draw the two tickets and then swap them.

By Symmetry, $P(X_1 = 1) = P(X_2 = 1)$
$X_2$ number of children for the 2$^{nd}$ mother chosen

Counting Way

- # pairs with 1 for 2$^{nd}$ mom
- # pairs (order matters)
  - 10 ways to pick 1$^{st}$ mom
  - 9 ways to choose 2$^{nd}$ mom

<table>
<thead>
<tr>
<th>ABCDEFGH</th>
<th>J I J</th>
<th>moms</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 2 2 2 2 3 3 3 4</td>
<td></td>
<td>values</td>
</tr>
</tbody>
</table>

- (A, B) (B, A) 2
- (C, A) (C, B) $\frac{2 \times 7}{2} = 7$
- (J, A) (J, B) $2 + 2 \times 8 = 18$

$\frac{18}{90} = \frac{2}{10}$
DETOUR CONTINUED: Why is the expected value a desirable summary of a probability distribution?
Random Variables

Random Variables: $X_1, X_2, ..., X_n$

Random ERROR: $X_1 - c, X_2 - c, ..., X_n - c$

LOSS: $l: R \rightarrow R^+$

Use $L_2$ loss again

$\sqrt{(X - c)^2}$

Now find the Expected Value of the loss

AKA RISK $\mathbb{E}(X - c)^2$

In General $X - c$

is the error

It is a random variable
Summarizing the Probability Distribution

EXPECTED LOSS:

AKA RISK \[ \mathbb{E}[l(X - c)] = \mathbb{E}[(X - c)^2] \]

Minimize the risk \[ \mathbb{E}(X - \mathbb{E}(X) + \mathbb{E}(X) - c)^2 \]

Like before we add and subtract \( \mathbb{E}(X) \)
Properties of Expected Value

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

\[ E(aX + b) = \sum_{j=1}^{m} (a x_j + b) P(X = x_j) \]

\[ = a \sum_{j=1}^{m} x_j P(X = x_j) + b \sum_{j=1}^{m} P(X = x_j) \]

\[ = a E(X) + b \]
Minimize the Risk

\[ \mathbb{E}[(X - c)^2] = \mathbb{E}(X - \mu + \mu - c)^2 \]

\[ = \sum_{j=1}^{m} (x_j - \mu + \mu - c)^2 \cdot P_j \]

\[ = \sum_{j=1}^{m} (x_j - \mu)^2 \cdot P_j + 2(\mu - c) \sum_{j=1}^{m} (x_j - \mu) \cdot P_j + \sum_{j=1}^{m} (\mu - c)^2 \cdot P_j \]

\[ = \mathbb{E}(X - \mu)^2 + (\mu - c)^2 \]

To simplify the writing, let \( \mathbb{E}(X) = \mu \)

Min for \( c = \mu \)
The Expected Value Minimizes Risk

\[
\mathbb{E}[X - \underbrace{\mathbb{E}(X)}_{\mu}]^2 \leq \mathbb{E}[(X - c)^2]
\]

\[
\mathbb{E}(x - \mu)^2 = \sum_{j=1}^{m} (x_j - \mu)^2 P(x_j)
\]

This side is the Variance
Probability Samples give us Representative Data where the sample average is well behaved and an accurate estimate of the population average.
Sampling Frame = Sample

Scenario: Administrative Data
Can we make up for no Probability Sample with Big Data?
Sample and Population Averages

The gap between these is based on three things:

- Data **quality** measure (the correlation between the sampling technique and the response)
- Data **quantity** measure (how big is the sample relative to the population)
- **Problem difficulty** measure (how variable is the response)

Meng 2018, Annals Applied Probability
Sample and Population Averages

- Probabilistic sampling ensures high data quality by eliminating selection bias and confounding.

- When combining data sources for population inferences, those relatively tiny but higher quality sources should be given far more weights than suggested by their sizes.

Active Area of Research Area
Large Administrative Data vs Small SRS

The admin data set is very accurate but biased.

The bias may be small enough to not matter. If it isn't, it's a problem.

The possible values the sample may have could be for a SRS of n=400.
Data Life Cycle

- Precise
  - Question Formulation
- Generalization
  - $E({\bar{X}}) = E(L(x)) = \mu$
  - in the world

Data Design/Generation
- $x$ in our sample

Data Analysis