MTH208a: Worksheet 21

Statistical and Algorithmic Bias

Last week we discussed Statistical Paradoxes and correlations. We witnessed, how important it is to know the quality of your data and to know which variables are missing.

Today we discuss a different aspect: bias

Types of Bias

The types of bias can roughly be categorized as:

- Selection Bias
- Survivorship Bias
- Omitted variable bias
- Recall bias
- Observer bias
- Funding bias

Example 1

Consider estimating the average number of tweets in a day by tracking the tweets from 1pm - 3pm.

Any problem with this?

Example 1

Consider estimating the average number of tweets in a day by tracking the tweets from 1pm - 3pm.

Any problem with this?

This is selection bias.

When we collect data, we hope the *sample* collected accurately represents the *population.* Here we are only considering data from one part of the day!

Example 2

Group 18 decides to survey all IIT Kanpur people on their favorite choice of food. They choose to make their survey available using a QR code.

Any problem with this?

Example 2

Group 18 decides to survey all IIT Kanpur people on their favorite choice of food. They choose to make their survey available using a QR code.

Any problem with this?

QR codes are new-age phenomenon and not everyone (especially older people) would be familiar with how to use it.

This is again selection bias.

Example 3

Suppose you want to see how much does a 21 year old get on a 10th class Maths exam. You make all your classmates take the exam and calculate the average marks as a guess.

Example 3

Suppose you want to see how much does a 21 year old get on a 10th class Maths exam. You make all your classmates take the exam and calculate the average marks as a guess.

You *sample* is not representative of your *population*.

Example 4

We enroll for a gym membership and attend for a few days. We see the same faces of many people who are fit, motivated and exercising everyday whenever we go to gym. After a few days we lose motivation whereas we see all the other people keep staying motivated! We conclude that we have a weaker resolve than other people :-(

Example 4

We enroll for a gym membership and attend for a few days. We see the same faces of many people who are fit, motivated and exercising everyday whenever we go to gym. After a few days we lose motivation whereas we see all the other people keep staying motivated! We conclude that we have a weaker resolve than other people :-(

What we didn't see was that many of the people who had enrolled for gym membership had also stopped turning up and we didn't see them.

Example 4

We enroll for a gym membership and attend for a few days. We see the same faces of many people who are fit, motivated and exercising everyday whenever we go to gym. After a few days we lose motivation whereas we see all the other people keep staying motivated! We conclude that we have a weaker resolve than other people :-(

What we didn't see was that many of the people who had enrolled for gym membership had also stopped turning up and we didn't see them.

This is surivorship bias.

Where else have we seen this bias?

Types of data collection

There are two main ways of obtaining data:

- Experimental data
- Observational data

Types of data collection

There are two main ways of obtaining data:

- Experimental data
- Observational data

When collecting experimental data, we can safeguard against unknown behaviors and pursue to generate *representative data.*

In observational data, we do not have the above luxury.

Why is handling bias important?

Simple: if your data is biased, your conclusions will be biased!

Example 4

Please do the following: Google search -

“greatest musician of all time”

Do you see any problems with the results? Repeat with “greatest actors of all time”.

Example 5

When YouTube launched their video upload app for iOS, between 5-10 percent of videos uploaded by users were upside-down. Why?

Example 5

When YouTube launched their video upload app for iOS, between 5-10 percent of videos uploaded by users were upside-down. Why?

About 8-9% of the world population is left-handed people. Often, left handed people rotate their phone clock-wise instead of anti-clockwise for right-handed people.

YouTube had an algorithm that detected a horizontal video and flipped it to display it vertically. This algorithm neglected that left-handed people rotated phones differently!

Exercise

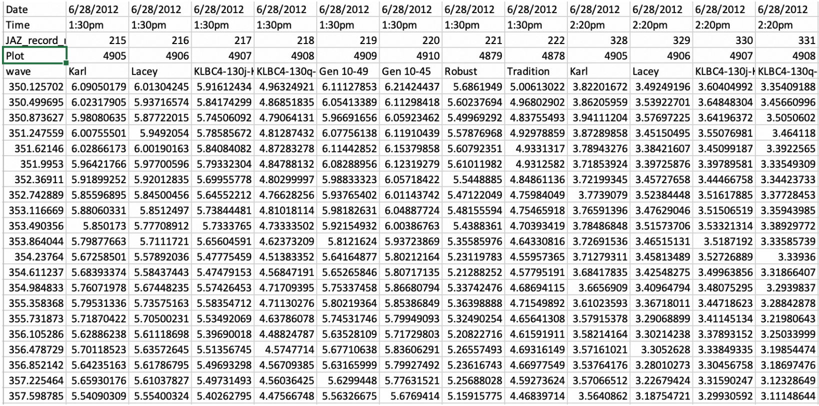
An experiment was done in an agriculture lab. Several variety of Barley were plotted in different plots of land. A JAZ spectrometer was taken to measure spectral reflectance from plots of Barley. The goal was to find

which wavelengths can differentiate between barley?

The machine essentially checks how much light is reflected back from the plants. Measurements were taken by graduate students over the span of two days from morning till afternoon.

Exercise

You are given a `LowRepeated.csv` dataset in your repository. This data looks like:



Date	6/28/2012	6/28/2012	6/28/2012	6/28/2012	6/28/2012	6/28/2012	6/28/2012	6/28/2012	6/28/2012	6/28/2012	6/28/2012	6/28/2012
Time	1:30pm	1:30pm	1:30pm	1:30pm	1:30pm	1:30pm	1:30pm	1:30pm	2:20pm	2:20pm	2:20pm	2:20pm
JAZ_record_r	215	216	217	218	219	220	221	222	328	329	330	331
Plot	4905	4906	4907	4908	4909	4910	4879	4878	4905	4906	4907	4908
wave	Karl	Lacey	KLBC4-130j-K	KLBC4-130q-	Gen 10-49	Gen 10-45	Robust	Tradition	Karl	Lacey	KLBC4-130j-K	KLBC4-130q-
350.125702	6.09050179	6.01304245	5.91612434	4.96324921	6.11127853	6.21424437	5.6861949	5.00613022	3.82201672	3.49249196	3.60404992	3.35409188
350.499695	6.02317905	5.93716574	5.84174299	4.86851835	6.05413389	6.11298418	5.60237694	4.96802902	3.86205959	3.53922701	3.64848304	3.45660996
350.873627	5.98080635	5.87722015	5.74506092	4.79064131	5.96691656	6.05923462	5.49969292	4.83755493	3.94111204	3.57697225	3.64196372	3.5050602
351.247559	6.00755501	5.9492054	5.78585672	4.81287432	6.07756138	6.11910439	5.57876968	4.92978859	3.87289858	3.45150495	3.55076981	3.464118
351.62146	6.02866173	6.00190163	5.84084082	4.87283278	6.11442852	6.15379858	5.60792351	4.9331317	3.78943276	3.38421607	3.45099187	3.3922565
351.9953	5.96421766	5.97700596	5.79332304	4.84788132	6.08288956	6.12319279	5.61011982	4.9312582	3.71853924	3.39725876	3.39789581	3.33549309
352.36911	5.91899252	5.92012835	5.69955778	4.80299997	5.98833323	6.05718422	5.5448885	4.84861136	3.72199345	3.45727658	3.44466758	3.34423733
352.742889	5.85596895	5.84500456	5.64552212	4.76628256	5.93765402	6.01143742	5.47122049	4.75984049	3.7739079	3.52384448	3.51617885	3.37728453
353.116669	5.88060331	5.8512497	5.73844481	4.81018114	5.98182631	6.04887724	5.48155594	4.75465918	3.76591396	3.47629046	3.51506519	3.35943985
353.490356	5.850173	5.77708912	5.7333765	4.73333502	5.92154932	6.00386763	5.4388361	4.70393419	3.78486848	3.51573706	3.53321314	3.38929772
353.864044	5.79877663	5.7111721	5.65604591	4.62373209	5.8121624	5.93723869	5.35585976	4.64330816	3.72691536	3.46515131	3.5187192	3.33585739
354.23764	5.67258501	5.57892036	5.47775459	4.51383352	5.64164877	5.80212164	5.23119783	4.55957365	3.71279311	3.45813489	3.52726889	3.33936
354.611237	5.68393374	5.58437443	5.47479153	4.56847191	5.65265846	5.80717135	5.21288252	4.57795191	3.68417835	3.42548275	3.49963856	3.31866407
354.984833	5.76071978	5.67448235	5.57426453	4.71709395	5.75337458	5.86680794	5.33742476	4.68694115	3.6656909	3.40964794	3.48075295	3.2939837
355.358368	5.79531336	5.73575163	5.58354712	4.71130276	5.80219364	5.85386849	5.36398888	4.71549892	3.61023593	3.36718011	3.44718623	3.28842878
355.731873	5.71870422	5.70500231	5.53492069	4.63786078	5.74531746	5.79949093	5.32490254	4.65641308	3.57915378	3.29068899	3.41145134	3.21980643
356.105286	5.62886238	5.61118698	5.39690018	4.48824787	5.63528109	5.71729803	5.20822716	4.61591911	3.58214164	3.30214238	3.37893152	3.25033999
356.478729	5.70118523	5.63572645	5.51356745	4.5747714	5.67710638	5.83606291	5.26557493	4.69316149	3.57161021	3.3052628	3.33849335	3.19854474
356.852142	5.64235163	5.61786795	5.49693298	4.56709385	5.63165999	5.79927492	5.23616743	4.66977549	3.53764176	3.28010273	3.30456758	3.18697476
357.225464	5.65930176	5.61037827	5.49731493	4.56036425	5.6299448	5.77631521	5.25688028	4.59273624	3.57066512	3.22679424	3.31590247	3.12328649
357.598785	5.54090309	5.55400324	5.40262795	4.47566748	5.56326675	5.6769414	5.15915775	4.46839714	3.5640862	3.18754721	3.29930592	3.11148644

Exercise

- Date is mm/dd/yyyy
- Time of data collection
- Plot is the code for the plot of land (not useful to you right now)
- Wave (vertically down) is the wavelength
- Karl, Lacey, ... are names of the variety of Barley.

Notice, that there are repeated measurements as well.

GOAL: can you find a wavelength that can truly differentiate between the varieties of Barley?