Class01-02

February 1, 2021

1 Data

Data is involved in most every conversation, and in every stage, of a scientfiic investigation.

The scientific method is made up of 3 steps: (i) propose a hypothesis (ii) deduce a set of events you would observe if you hypothesis were true, and (iii) devise experments that generate **data**, or observe a phenomena and collect **data** supporting or refuting your hypothesis. This last step is crucial to proving your hypothesis and almost completely depends on **data**.

Understanding the components of data, and how to manipulate them, is an essential skill of statistics and data science (SDS).

A misunderstanding of how to store and use data can lead to mistakes that impact decision-making and X. For example, the UK Public Health Department (Public Health Endlgand) reported 0 deaths due to COVID-19 for a week because they collected data in an excel spread sheet which has a limited data capacity. A subtler data issue appeared in understanding data related to healthcare management([article])(https://science.sciencemag.org/content/366/6464/447). Researchers studied an algorithm that was meant to assign scores to patients based on the amounf of medical care they needed. A patient who needs more medical care, who is sicker, should recieve a higher score. But the algorithm used the amount of money spent on patients as a proxy for their health, and because Black patients often have less access to healthcare they recieved smaller scores. The result: This algorothm, using data on healthcare expenditures, assigned White patients higher scores than Black patients, providing them more help.

The method we use to collect **data**—called sampling—is another important issue. When we select observations to include in a dataset we want them, intuitively, to represent all possible observations. If a systematic method for selecting observations creeped into how we select observations then this could hurt the validity of any conclusions we draw from our data.

A lot of time should be invested in planning how to sample, store, and structure data.

One more data issue that we should address now: whether the word data is plural or singular, whether we should write "data are" or "data is". The statistical new outlet FiveThirtyEight addressed this question here. They polled readers of their stories asking them if it is more appropriate to write "data" is or "data are" and found, among 1,894 responses that 1,279 (68%) felt it was better to write "data is". So "data is" it is.

Our goals for this class will be to learn about: * Observations, variables, and a data matrix * Data collection and sampling * Experiments vs Observational studies

1.1 Observations, variables and the data matrix (data frame)

1.1.1 Observations and variables

An **observation** (O) is a set of measurements or values derived from an object under study. Individual characteristics of an observation are called **variables**. We can write an observation as

$$O = \{v_0 : \text{value}_0, v_1 : \text{value}_1, v_2 : \text{value}_2, \cdots\}$$

where O is an observation, v_i is the i^{th} variable that contains a corresponding value (value_i)

We can represent an observation in Python using a dictionary In Python3, a dictionary is made up of **keys** and their corresponding **values**.

Lets build a single observation in Python3 using a dictionary. Suppose we want to record data on the number of times prof m's three dogs: Banjo, Fiddle, and Kazoo ask for a minature milk bone (there favorite) throughout the day. At the end of the first day we find Banjo asked 10 times (probably an underestimate), Fiddle asked 3 times, and Akzoo asked 5 times.

[1]:
```
# Observation 1
milkBoneObservation01 = {"Banjo":10, "Fiddle":3, "Kazoo":5}
print(milkBoneObservation01)
```

```
{'Banjo': 10, 'Fiddle': 3, 'Kazoo': 5}
```

A dictionary is created with "curly" brackets ({ and }). Inside the curley brackets you can place a set of keys:values seperated by commas.

Above we have three keys ("Banjo", "Fiddle", and "Kazoo"). In Python, an object enclosed in quotes is called a **string**.

These keys are linked to values. To be specific, the keys above are linked to integers (another special type in Python).

Dictionaries are useful for structuring a lot of different type of data. In our example, we strucutred a single observation, the number of minature milkbones requested by Banjo, Fiddle, and Kazoo with a dictionary.

We can access that data in Python like this:

[2]:
```
BanjosNumberOfRequests = milkBoneObservation01['Banjo']
print("Banjo asked for a minature milk bone {:d} times today".
 ↪format(BanjosNumberOfRequests))
```

```
Banjo asked for a minature milk bone 10 times today
```

We extracted the number of requests by Banjo for a milkbone from our dictionary by writing `milkBoneObservation01['Banjo']`. In general, we can extract the value corresponding to a specific key from a dictionary using this code `dictionary[key]`.

1.1.2 Dataframes

Often, many observations with their associated variables are typically organized into a **data matrix**—sometimes called a ***dataframe***.

Data was collected on heart failure patients and organized into a **dataframe**. Data on 299 patients with heart failure was collected. Researchers studied 13 different traits of patients with heartfailure to determine if specific characteristics of a patient were more, or less, likely to lead to death.

Below, each row represents a patient with heart failure (an observation), and each column stores a different variable (age, whether they are anaemic, the patient's creatine levels, etc.).

As a statistician, you will often not receive data in such a nice format. It is the statistician's job to format data so that it can be clearly communicated to others.

[3]:
```
import pandas as pd # Pandas is a python module used to manipualte data.
heartfailuredata = pd.read_csv("https://archive.ics.uci.edu/ml/
 ↪machine-learning-databases/00519/heart_failure_clinical_records_dataset.csv")␣
 ↪# read in a data file in CSV format.

heartfailuredata.head(10) # print the first 10 rows in the data matrix (data␣
 ↪frame)
```

[3]:

	age	anaemia	creatinine_phosphokinase	diabetes	ejection_fraction	high_blood_pressure	platelets	serum_creatinine	serum_sodium	sex	smoking	time	DEATH_EVENT
0	75.0	0	582	0	20	1	265000.00	1.9	130	1	0	4	1
1	55.0	0	7861	0	38	0	263358.03	1.1	136	1			
2	65.0	0	146	0	20	0	162000.00	1.3	129	1			
3	50.0	1	111	0	20	0	210000.00	1.9	137	1			
4	65.0	1	160	1	20	0	327000.00	2.7	116	0			
5	90.0	1	47	0	40	1	204000.00	2.1	132	1			
6	75.0	1	246	0	15	0	127000.00	1.2	137	1			
7	60.0	1	315	1	60	0	454000.00	1.1	131	1			
8	65.0	0	157	0	65	0	263358.03	1.5	138	0			
9	80.0	1	123	0	35	1	388000.00	9.4	133	1			

1	0	6	1
2	1	7	1
3	0	7	1
4	0	8	1
5	1	8	1
6	0	10	1
7	1	10	1
8	0	10	1
9	1	10	1

2 Data collection and sampling

When a scientist (like yourself) formulates a hypothesis—a question or statement that can be tested, measured, or otherwise supported or denied—they define a population. A **population** is the collection of all possible observations used to address a hypothesis. It is normally impossible to study an entire population. Instead, statisticians and data scientists study a **sample**—a subset of the observations that make up the entire population—from the population.

In the above data frame of 299 heart failure patients, the population would be every single patient who has heart failure. This is, unfortunately, a massive set of people. The 299 heart failures we collected represents a sample from the population of all heart failure patients.

As another example, lets consider the following research question:

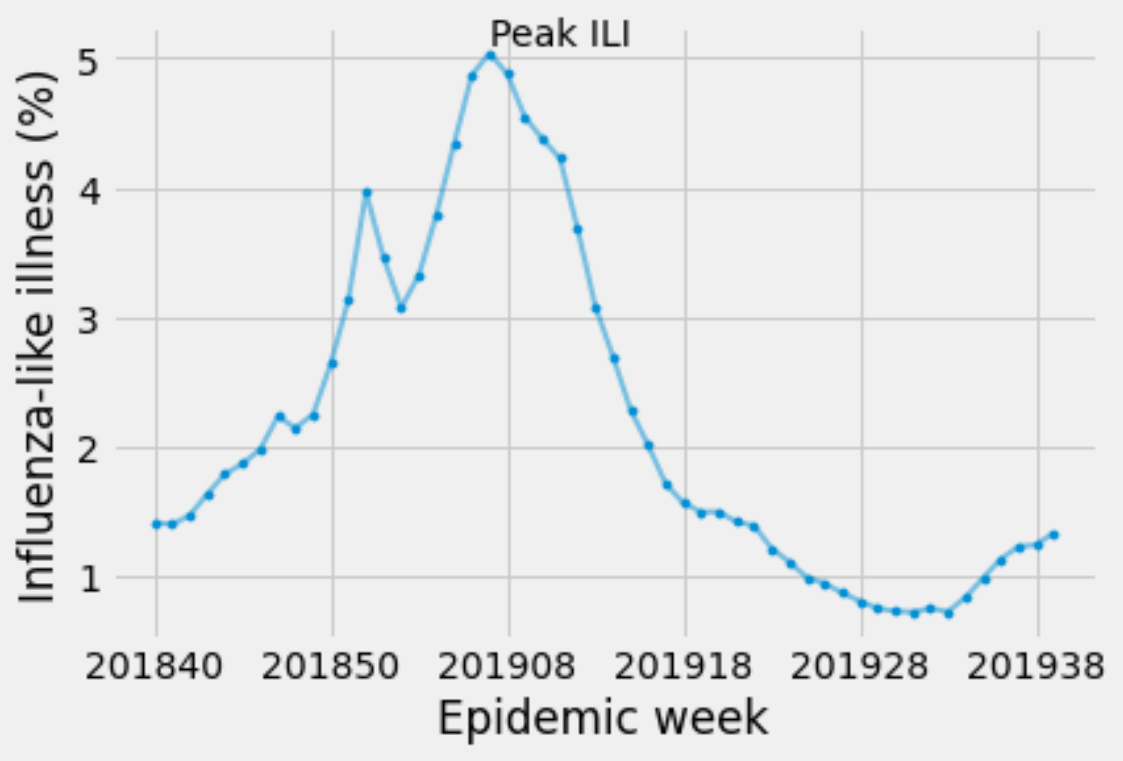
During Flu season, in what epidemic week is the proportion of influenza-like illness likely to peak in the United States?

We will collect data from the Centers for Disease Control (CDC) on proportions of influenza-like illness (ILI) over past seasons in the US from the 2010/2011 season up to the 2019/2020 season.

[21]:
```
# download data from dr.m's github repo (how convienant)
sampleIliData = pd.read_csv("https://raw.githubusercontent.com/mcandrewlab/
 ↪cdcilidata/master/ilidata_cdc_us.csv")

season1819 = sampleIliData[sampleIliData.season=='2018/2019'] # pick a single␣
 ↪season

# quick look at a single flu season
plt.style.use("fivethirtyeight")
fig,ax = plt.subplots()

ax.plot( season1819.modelweek,season1819.wili, lw=2,alpha=0.5 )
ax.scatter( season1819.modelweek,season1819.wili,s=10 )

ax.set_xlabel("Epidemic week")
ax.set_ylabel("Influenza-like illness (%)")

ax.set_xticks(season1819.modelweek[::10])
```

```
ax.set_xticklabels(season1819.epiweek[::10])

peak = season1819.sort_values("wili").iloc[-1:]
ax.text( x = peak.modelweek, y = peak.wili, s="Peak ILI", ha="left", va="bottom"␣
 ↪)
```

[21]:
```
Text(436    2562
Name: modelweek, dtype: int64, 436    5.03689
Name: wili, dtype: float64, 'Peak ILI')
```



In the 2018/2019 infleunza season, influenza-like illness (ILI) peaked during the 8th week of 2019—19 weeks from the start of the season.

Our unit of observation is a single influenza season, defined as values of ILI from the week 40 of one year until the 20th week of the next year. We can define two variables: the epidemic week where ILI is highest and the number of weeks from the beginning of the season where ILI peaks. And we can build a data matrix (data frame) where rows are observations (influenza seasons) and columns are the epidemic week (ew) and number of weeks from the beginning of the season (from40) where ILI peaks.

[23]:
```
seasonPeaks = pd.read_csv("https://raw.githubusercontent.com/mcandrewlab/
 ↪cdcilidata/master/seasonpeaks.csv")
print(seasonPeaks)
```

```
      season      ew  from40
0  2010/2011  201105      17
1  2011/2012  201211      23
2  2012/2013  201252      12
3  2013/2014  201352      12
4  2014/2015  201452      12
5  2015/2016  201610      22
6  2016/2017  201706      18
7  2017/2018  201805      17
8  2018/2019  201907      19
9  2019/2020  201952      12
```

The influenza seasons from 2010 - 2020 are a subset of all influenza seasons that have occurred in the US. They are a **sample** from the **population** of all US influenza seasons.

2.1 Sampling strategies

When we take a sample from a population, we want to ensure the sample is **representative** and has as little **bias** as possible. A sample is **representative** when the sample reflects the characteristics of the larger population you wish to study. A representative sample allows you to generalize findings from your sample to your population.

There are many ways in which a sample can be un-representative of the population under study. One way is when a population is samples so that some observations are more likely to be chosen than others. This is called **sampling bias**.

For example, suppose you want to study the life expectancy of men and woman in the US. To collect a sample, you post on ad on Reddit that asks participants to reply to this post with their date of birth, sex, and GPS coordinates of their place of residence.

Will this sample be biased? If yes, in what ways?

What type of common bias is called **response-bias**. Response bias occurs when some members of a population are more probable to respond and be included in the sample than others. For example, in our above sampling scheme, only those people who have a Reddit account could respond to our survey.
Are those who have Reddit accounts representative of all men and women in the US? Likely not.

There are methods for drawing sample from a population of interest that tries to control for bias. The goal of any sampling strategy is to collect a sample that best represents the population.

Lets define a population as a **set**—or collection of items—and assign it the variable P. We can take a sample S from our population

$$P = \{p_0, p_1, p_2, p_3, p_4, \cdots, p_{N-1}\} \tag{1}$$

and define a sample as a subset, or smaller (at most the same) collection of items, from P. That is,

$$S = \{s_0, s_1, s_2, \cdots, s_M\} \tag{2}$$

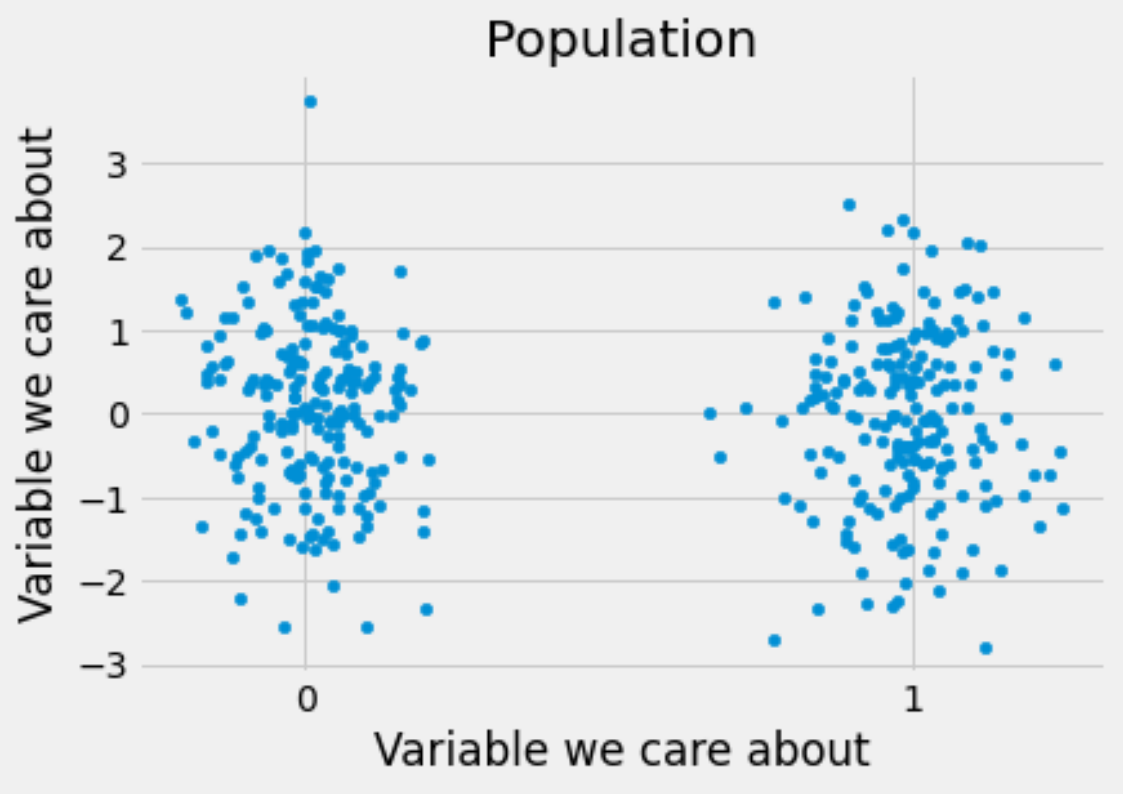
where each item in S is picked from our population P and the number of observations in S is smaller than the number of observations in P

2.1.1 Simple random sampling

A simple random sample (SRS) chooses each observation from the population (P) with equal probability to build a sample (S). When you sample there is no consideration for any variables measured in the population. Suppose we could see our entire population P, and further, we could measure a single variable that takes two different values: 0 or 1. (see figure)

[89]:
```
x0 = np.random.normal(0,0.1,200)
_0 = np.random.normal(0,1,200)
pop0 = pd.DataFrame({'Variable we measured':x0, 'Variable we care about':_0})
pop0["f"]=0

x1 = np.random.normal(1,0.1,200)
_1 = np.random.normal(0,1,200)
pop1 = pd.DataFrame({'Variable we measured':x1, 'Variable we care about':_1})
pop1["f"]=1

population = pop0.append(pop1)

fig,ax = plt.subplots()
ax.scatter(x=population["Variable we measured"],y=population["Variable we care␣
 ↪about"],s=20)
ax.set(xlabel="Variable we care about",ylabel="Variable we care about")

ax.set_xticks([0,1])
ax.set_yticks(np.arange(-3,3+1))
ax.set_title("Population")
```

[89]:
```
Text(0.5, 1.0, 'Population')
```



The population above has in total 400 observations (400 dots). We are interested in sampling this population to study a single variable we care about.

[103]:
```
srs = population.sample(60)

fig,axs = plt.subplots(1,2)

ax = axs[0]
ax.scatter(x=population["Variable we measured"],y=population["Variable we care␣
 ↪about"],s=20,alpha=0.20)
ax.scatter(x=srs["Variable we measured"],y=srs["Variable we care␣
 ↪about"],s=30,color="red")

ax.set(xlabel="",ylabel="")

ax.set_xticks([0,1])
ax.set_yticks(np.arange(-3,3+1))
ax.set_ylim(-3,3)

ax.set_title("Population")

ax=axs[1]
```

```
srs["v"]=np.random.normal(0,1,60)
ax.scatter(x=srs["v"],y=srs["Variable we care about"],s=30,color="red")

ax.set_xticks([])
ax.set_yticks(np.arange(-3,3+1))
ax.set_ylim(-3,3)

ax.set_title("Simple random sample")
```

[103]: `Text(0.5, 1.0, 'Simple random sample')`

We will take a simple random sample (S) of 60 observations (red dots) from the total population (blue dots). It appears that, for values of the variable we care about, the red dots are representative of the entire population. Conclusions we draw from our sample will likely generalize to the entire population.

But a simple random sample is not always representative of a population. What if observations behaved differently dependent on if they are classified into one of two groups, and further we are not aware of the variable we care about differs by this classification.

[118]:
```
x0 = np.random.normal(0,0.1,200)
_0 = np.random.normal(-1.5,0.5,200)
pop0 = pd.DataFrame({'Variable we measured':x0, 'Variable we care about':_0})
pop0["f"]=0
```

```
x1 = np.random.normal(1,0.1,100)
_1 = np.random.normal(1.5,0.5,100)
pop1 = pd.DataFrame({'Variable we measured':x1, 'Variable we care about':_1})
pop1["f"]=1

population = pop0.append(pop1)


srs = population.sample(60)

fig,axs = plt.subplots(1,2)

ax = axs[0]
ax.scatter(x=population["Variable we measured"],y=population["Variable we care␣
 ↪about"],s=20,alpha=0.20)
ax.scatter(x=srs["Variable we measured"],y=srs["Variable we care␣
 ↪about"],s=30,color="red")

ax.set(xlabel="",ylabel="")

ax.set_xticks([0,1])
ax.set_yticks(np.arange(-3,3+1))
ax.set_ylim(-3,3)

ax.set_title("Population")

ax=axs[1]
srs["v"]=np.random.normal(0,1,60)
ax.scatter(x=srs["v"],y=srs["Variable we care about"],s=30,color="red")

ax.set_xticks([])
ax.set_yticks(np.arange(-3,3+1))
ax.set_ylim(-3,3)

ax.set_title("Simple random sample")
```

[118]: Text(0.5, 1.0, 'Simple random sample')

A simple random sample here does look representative. But a **stratified sample** may be more appropriate. A stratified random sample (STRS) draws observations from the population based on which strata (group) they fall into. A STRS is appropriate when: - The population has groups of data that may be interesting in their own right. For example, a sample of heart failure patients stratified by diabetic and non-diabetic status. - It may be convenient to sample by strata. For example, to collect data on influenza-like illness (ILI), a group of hospitals records the number of patients who enter the hospital and the number who are positive for ILI. The strata here are hospitals and it is more convenient to ask them to collect data than having the CDC hire additional people to collect this data.

A **convenience sample** is a (typically) non-probabilistic sampling strategy where the scientists collects observations from their sample that are easies to collect. Many times a convenience sample is biased—one where the probability of sampling observations is not uniform in the population, and is also often not representative of the population.

2.2 Experimental versus Observational data

Two main ways we can classify data are whether they are experimental or observational. If data is purposely influenced by the scientific team, the data is **experimental**. If instead the data is collected without interference from the scientific team, the data is **observational data**.

For example, a clinical trial enrolls patients with heart failure. Patient are screened, and if included in the trial, they are assigned to two groups: (i) patients in the device group receive a transapical valve replacement and (ii) patients in the control group are given state of the art medical management. This is a simplified description of the PARTNER trial.

An example of an observational dataset is the Framingham Heart study (FHS). The FHS, started in 1948, followed over 5,000 men and women to identify characteristics that contribute to heart disease. Because researchers did not influence the participants, instead following them and recording observations, this study is considered observational.

2.2.1 Experimental data

Experiments, and so experimental data, have several advantages over observational data when they are feasible. Experimental data can select the observations that most represent the population of interest—called **control**.

Experimental designs can also **randomize**—or choose with a given probability one of many conditions to give to participants—to control for factors the experiment is unable to measure. Randomization is a cornerstone of experimental designs. By randomizing participants to different groups (sometimes called treatments) you can reduce the chances observed differences between groups is due to systematic characteristics of your sampled observations, called **allocation bias**.

For example, suppose we ran two studies and measured the difference in the proportion of myocardial infarction (MI) between a control and treatment group. In the first study we let patients choose whether they are included in the control or treatment groups. In the second study, each patient is assigned treatment or control at random with probability $1/2$. When we measure the difference in proportion of MI, it will be difficult to determine in the first study if the observed difference is because of the treatment or because of some other characteristic of patients who chose the treatment or control group. By randomly assigning patients to treatment or control, the second study reduces the chances the difference in the proportion of MI is due to unmeasured properties of the participants, increasing the chances the observed difference is due to the treatment.

2.2.2 Observational data

Observational studies record information on samples from a population. Observational studies can be either **retrospective** or **prospective**. A **retrospective** study collects past data on samples that fit the population under study. A **prospective** study identifies samples that fit the population and then records data on these samples at future time points.

Researchers who conduct an observational study should be aware of **selection bias**. **Selection bias** occurs when the researcher chooses samples that do not properly represent the population and the outcome of interest, changing the outcome of the study in one direction or the other. For example, suppose we decide to study the nutritional value of foods purchased in homes to average life expectancy of the residents. We sample 60 different homes and record the results. If we do not record any additional information on the homes it may have been the case homes were selected that are in different socioeconomic classes, in locations with different environmental factors known to impact health, or made up of members born in different eras, and so had different views on health.

Below is data collected by the World Health Organization (WHO) on the life expectancy of US men and women, and stored in their **Global Health Observatory (GHO)**. Researchers at the WHO have collected data from government birth and death certificates, health systems, surveys, and research organizations and made this data available to the public. Because the WHO has simply observed this data and has not influenced samples included in the data, the GHO is observational data.

[119]:
```
whodata = pd.read_csv("https://apps.who.int/gho/athena/api/GHO/WHOSIS_000001?
 ↪format=csv")

usaLifeExp = whodata[whodata.COUNTRY=="USA"]
usaLifeExpMaleFemale = usaLifeExp[usaLifeExp.SEX.str.contains("MLE","FMLW")]

fig,ax=plt.subplots()
sns.scatterplot(x="YEAR",y="Numeric", hue="SEX",data=usaLifeExpMaleFemale,ax=ax)
ax.set(xlabel="Year", ylabel="Average Life Expectancy (yrs)")
ax.set_xticks(np.arange(2000,2016+1,2))
```

[119]:
```
[<matplotlib.axis.XTick at 0x119a99ed0>,
 <matplotlib.axis.XTick at 0x119a99490>,
 <matplotlib.axis.XTick at 0x119ae7f90>,
 <matplotlib.axis.XTick at 0x119afdb50>,
 <matplotlib.axis.XTick at 0x119b084d0>,
 <matplotlib.axis.XTick at 0x119b08a10>,
 <matplotlib.axis.XTick at 0x119b088d0>,
 <matplotlib.axis.XTick at 0x119b0f510>,
 <matplotlib.axis.XTick at 0x119b0fa50>]
```