Probability & Visualization

Python & Statistics Bootcamp

NaLette M. Brodnax

The Institute for Quantitative Social Science
Harvard University

June 5, 2018

Set up

Data

1. Go to https://nmbrodnax.github.io/python-stats/ and click on *Probability & Visualization*
2. Download the *County Demographics 2016* file and save it in your `bootcamp` directory
3. For your project, choose any other dataset, download it, and save it in your `bootcamp` directory

Jupyter notebook

1. Launch Jupyter from the command line: `jupyter notebook`, or from the Anaconda Navigator graphical interface
2. Navigate to the browser where your notebook is running
3. Create a new Python 3 notebook called `Visualization`

Working with Pandas

Statistics

Descriptive Statistics

Visualization

The pandas package

Pandas provides high-performance data manipulation and analysis. It was designed to allows users to load, prepare, manipulate, model, and analyze data.

Features

- A DataFrame object, which is similar to a two-dimensional array but allows different data types
- Tools for loading data from files of different formats
- Routines for merging, joining, and reshaping data
- Label-based slicing, indexing and subsetting

Data structures in pandas

series – one-dimensional, labeled array; it can be created from various inputs such as an array or dictionary

data frame – two-dimensional, labeled tabular structure with columns of the same or of different types

panel – three-dimensional, size-mutable array; rarely used

Working with pandas

```
import pandas as pd
import numpy as np

s = pd.Series()
print(s)
```

Working with pandas

```
import pandas as pd
import numpy as np

s = pd.Series()
print(s)
```

```
df = pd.read_csv("county_demographics_2016.csv")
print(df[:10])
```

pandas: useful features

- Reference by label

```
med_inc = df['median_income']
print(med_inc[:10])
```

- View a subset of rows

```
print(med_inc.head())
```

- Compute summary statistics

```
print(df.describe())
```

Activity: Use the `.isnull()` method to determine the number of rows of missing observations in median income.

Working with Pandas

Statistics

Descriptive Statistics

Visualization

Statistics

Statistics is the science of collecting, organizing, summarizing, analyzing, and drawing conclusions from data

- Descriptive statistics involves summarizing and presenting data
- Inferential statistics involves generalizing from samples to populations, performing estimations and hypothesis tests, determining relationships among variables, and making predictions

Why probability?

Deterministic thinking: 0 or 1, known with certainty

Probabilistic thinking: 0 to 1, known with uncertainty

- Events have some chance of occurring
- Use probability to quantify uncertainty
- Can be based on on what we observe or what we believe based on prior information

Probability

Probability is the chance of an event occurring, denoted $P(E)$

- Classical – all outcomes are equally likely to occur (e.g., coin flip)
- Empirical – outcomes may not be equally likely (e.g., World Cup winner) so we estimate the probability by observing how frequently those outcomes occur

Probability

Probability is the chance of an event occurring, denoted $P(E)$

- Classical – all outcomes are equally likely to occur (e.g., coin flip)
- Empirical – outcomes may not be equally likely (e.g., World Cup winner) so we estimate the probability by observing how frequently those outcomes occur

Probability Rules

- A given probability must fall between 0 and 1
- A probability and its complement must sum to 1
- The probabilities for each outcome must sum to 1

Probability

Probability is the chance of an event occurring, denoted $P(E)$

- Classical – all outcomes are equally likely to occur (e.g., coin flip)
- Empirical – outcomes may not be equally likely (e.g., World Cup winner) so we estimate the probability by observing how frequently those outcomes occur

Probability Rules

- A given probability must fall between 0 and 1
- A probability and its complement must sum to 1
- The probabilities for each outcome must sum to 1

Q: What are some examples of classical and empirical probability events?

Working with Pandas

Statistics

Descriptive Statistics

Visualization

Descriptive statistics

A variable is the unit of description we use to describe data

- characteristic or attribute that can assume different values
- called a **random variable** when values are determined by chance

Descriptive statistics

A variable is the unit of description we use to describe data

- characteristic or attribute that can assume different values
- called a **random variable** when values are determined by chance

Ways to describe data

- Qualitative (categorical) or quantitative (numerical)
- Discrete (countable) or continuous
- Type of measurement
  - nominal – name, category, label
  - ordinal – ordered in some way
  - interval – ordered and we can measure the differences
  - ratio – zero has a true meaning and we can calculate ratios across populations

Describing data

Measures of central tendency

- Mean
- Median
- Mode

Measures of spread

- Variance
- Standard deviation

Activity: describing data

1. Download the dataset you would like to use for your project from https://nmbrodnax.github.io/python-stats/ on the *Probability & Visualization* page
2. Confirm that the dataset is saved in the same directory as your notebook (in the `bootcamp` folder
3. Review the pandas documentation for `.isnull()`, `.notnull()`, and `.fillna()`
4. Create a data frame object for your dataset
5. Create a subset with three quantitative variables
6. Are your variables missing observations? If so, how many?
7. Compute the following descriptive statistics for each variable: mean, variance, and standard deviation

Working with Pandas

Statistics

Descriptive Statistics

Visualization

The matplotlib package

Matplotlib is a powerful visualization library

- built on numpy
- works well with many operating systems
- creates many types of outputs, including: png, jpeg, eps, pdf, and tiff

Working with matplotlib

Load matplotlib

```
import numpy as np
import pandas as pd
import matplotlib as mpl
import matplotlib.pyplot as plt
```

Create a figure

```
x = np.linspace(0, 10, 100)

fig1 = plt.figure()
plt.plot(x, np.sin(x))
plt.plot(x, np.cos(x))

plt.show()
```

matplotlib: useful features

Embed figures into your Jupyter notebook

```
%matplotlib inline
```

Specify a style

```
plt.style.use('classic')
```

Save a figure

```
fig1.savefig('myfig.png')
```

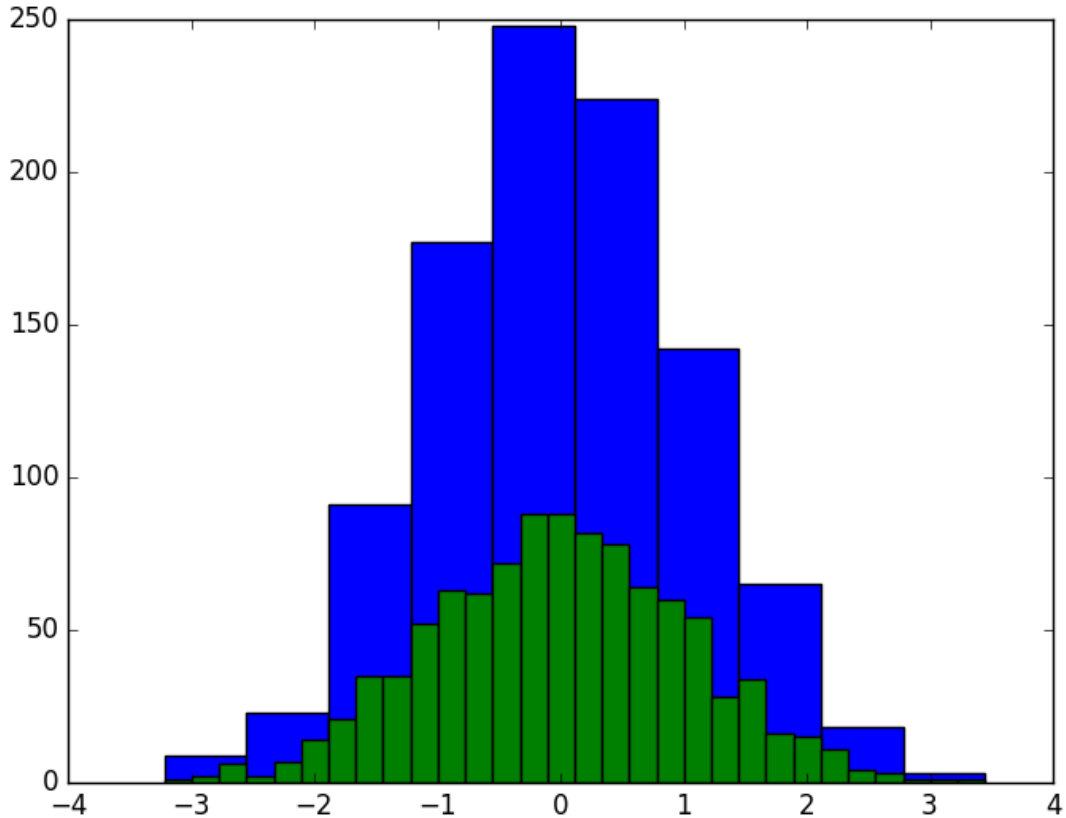
Add labels

```
fig2 = plt.figure()
plt.plot(x, np.sin(x))

plt.title("A Sine Curve")
plt.xlabel("x")
plt.ylabel("sin(x)")
```

Plot a histogram

```
data = np.random.randn(1000)
fig3 = plt.figure()
plt.hist(data)
plt.hist(data, bins=30)
fig3.savefig('myhist.png')
```



Probability distributions

A **probability distribution** is a list of values that a random variable can take and the corresponding probabilities of the values based on those frequencies

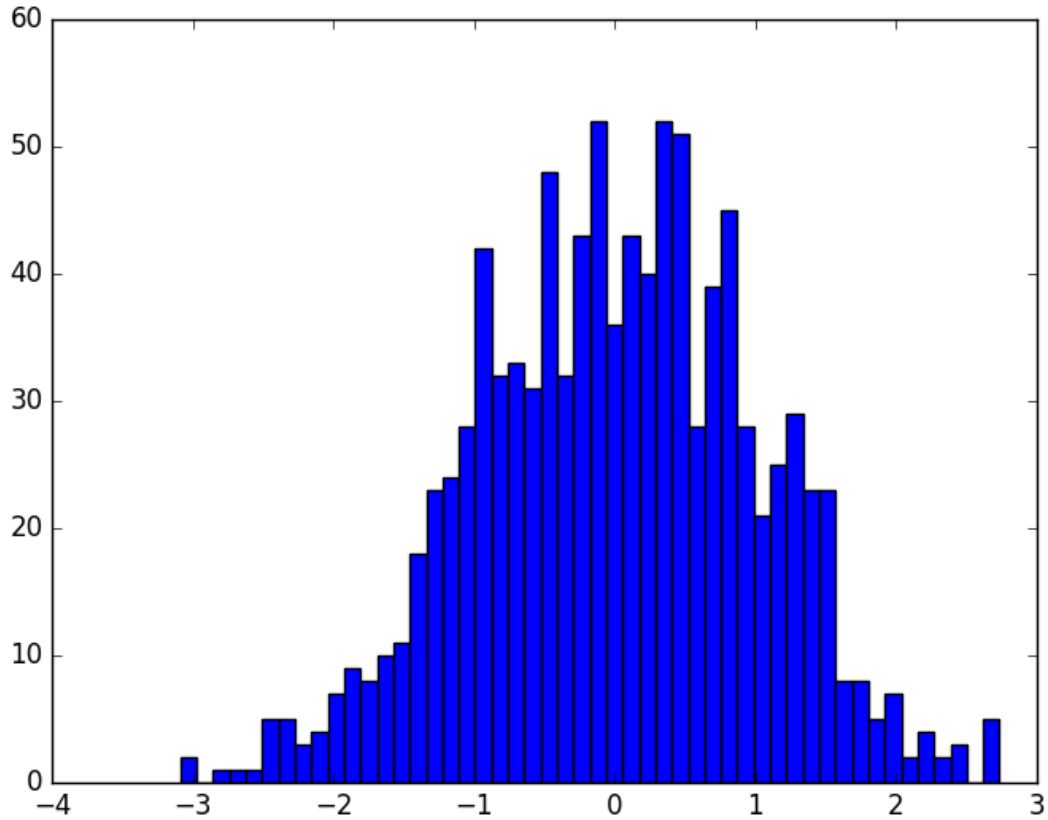
Standard Normal Distribution

- continuous, bell-shaped, and symmetric
- mean = median = mode
- 99% of area falls within three standard deviations
- characterized by the function

$$y = \frac{e^{-(X-\mu)^2/(2\sigma^2)}}{\sigma\sqrt{2\pi}} \tag{1}$$

Plot the normal distribution

```
norm = np.random.standard_normal(1000)
fig4 = plt.figure()
plt.hist(norm, bins=50)
fig4.savefig('normal.png')
```



Questions?