The Center of Applied Data Science

python™

Distributions

Descriptive Statistics

In [ ]:

```
import seaborn as sns
import matplotlib.pyplot as plt
import numpy as np
import pandas as pd

from scipy.stats import norm, poisson, uniform, skew, kurtosi
```

In [ ]:

```
sns.__version__ >= '0.9.0'
np.__version__ >= '1.15.4'
pd.__version__ >= '0.23.4'
```

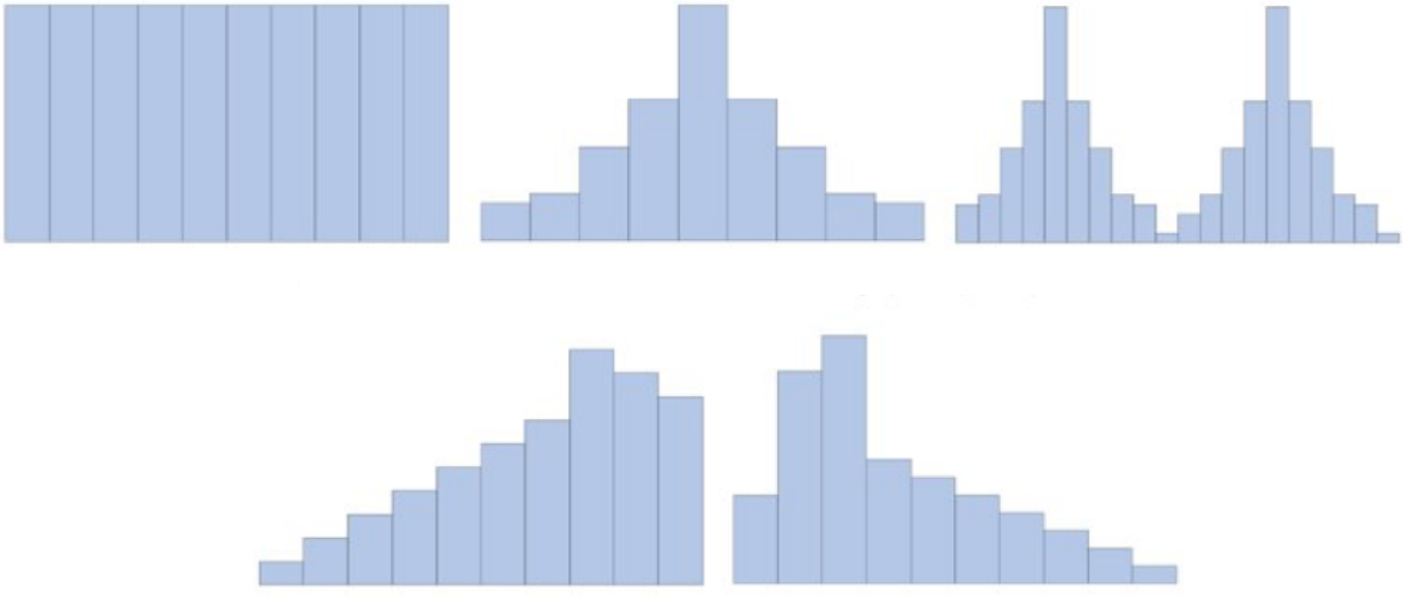
Content Outline

1. **Characterizing Distributions**
    - The Normal Distribution
    - Skewness
    - Kurtosis
2. **Other Common Distributions**
    - The Poisson Distribution
    - The Uniform Distribution

1. Characterizing Distributions

Recall that in **exploratory data analysis**, we examined a single variable by visual inspection and/or numerical summaries with the goal of summarizing the main characteristics of the data at hand. In this section on distributions, we will shift our focus towards classifying the various distributions that exist and how their general behavior can be extrapolated to explain why our data behaves the way it does.

The figure below shows the various forms in which a single variable can be distributed as observed using a histogram. Each form gives us different information regarding the variable's behavior, which may potentially change how we use the information in a business setting.



To encapsulate the motivation behind studying the general properties of common data distributions, consider the following example:

Example

Run the following code segment to load the `Height`, `Weight`, `Salary` and `HumanLongevity` data sets and plot their respective histograms.

In [ ]:

```
# create a list of data frames
files = ["Height", "Weight", "Salary", "HumanLongevity"]
df_ls = [pd.read_csv(f"../data/{i}.csv") for i in files]
```

In [ ]:

```
# concatenate this list of data frames with keys
cat = pd.concat([df_ls[i].iloc[:,-1] for i in range(4)], keys
```

In [ ]:

```
# provide column names, reset index
cat2 = cat.reset_index().drop("level_1", axis = 1)
cat2.columns = ["variable", "value"]
```

In [ ]:

```
# use seaborn facet grid to create a grid of plots
g = sns.FacetGrid(cat2, col = "variable", sharex = False, sha
g.map(sns.distplot, "value")
```

In [ ]:

```
g = sns.FacetGrid(cat2, col = "variable", sharex = False)
g.map(sns.boxplot, "value", order = cat2.variable.unique())
```

From the histograms, we see **two** main things:

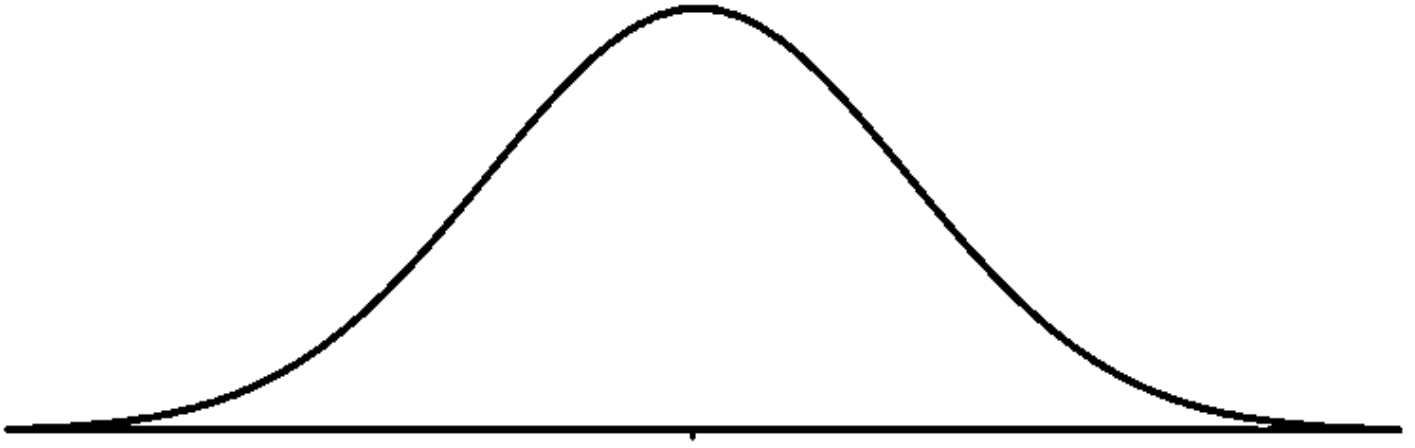
1. The general shape of each distribution is roughly similar in the sense that they all resemble a *mountain* of sorts. The location of the peaks vary from one variable to the next - `Height` and `Weight` seem to have the peak closer to the middle of the data, whereas the peaks for `HumanLongevity` and `Salary` are further to the right and left respectively. This feature is called **skewness**.
2. The heights towards the tail ends of the graphs are different. Though they may look similar in the plots, note that the y-axes are of *different scale* for the different variables. This feature is called **kurtosis**.

Before we take a deeper look at each of these features and how they impact our analysis, we need to first familiarize ourselves with the concept of **normality**.

The Normal Distribution:

A **Normal** distribution holds the following properties:

1. It forms a **bell-shaped** curve.
2. It is **symmetric** about the center.



The normal distribution is particularly interesting because one major assumption for most statistical modeling techniques is **normality**, i.e. the tendency for a set of data to be normally distributed. As such, having a variable that isn't normally distributed may limit the choice of tools we have at our disposal to model the data.

Mathematically, the bell curve of a normal distribution is given by the formula:

$$f(\mu|\sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

$$f(\mu\,|\,\sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

where $\mu\mu$ is the **mean** and $\sigma^2\sigma^2$ is the **variance**. One special case of this distribution is the **standard normal** distribution, where $\mu = 0\mu = 0$ and $\sigma = 1\sigma = 1$.

In [ ]:

```
x = np.linspace(-6, 6, 100000)
plt.plot(x, norm.pdf(x))
```

Adjusting the values of the **parameters** $\mu\mu$ and $\sigma\sigma$ modifies the shape of the distribution by varying the center and strech of the plot as illustrated below:

In [ ]:

```
x = np.linspace(-20, 20, 100000)
for i in (0, 2, 4):
    plt.plot(x, norm.pdf(x, loc = i, scale = 3), label = f"$\
plt.legend()
plt.title('Effect of Changing the Mean of a Normal Distributi
```

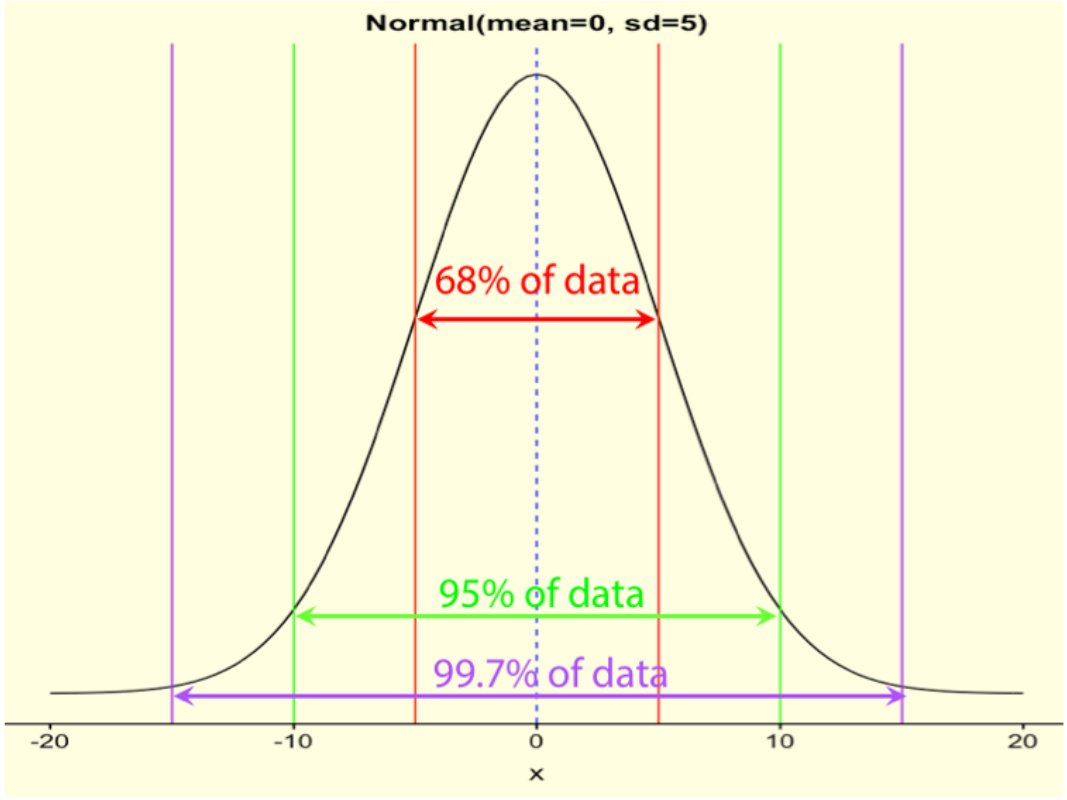
In [ ]:

```
x = np.linspace(-15, 15, 100000)
for i in (1, 3, 5):
    plt.plot(x, norm.pdf(x, loc = 0, scale = i), label = f"$\
plt.legend()
plt.title('Effect of Changing the Variance of a Normal Distri
```

As we can see, increasing/decreasing the mean shifts the axis of symmetry for the curve to the right/left, whereas increasing/decreasing the variance/standard deviation streches/contracts the curve. In practice, many real-world measurable variables such as the height and weight of people, IQ, and error in physical measurements are *approximately normal* and can be modeled using a normal distribution using by choosing appropriate parameter values.

One notable property of the normal distribution is that the distribution of data within its range is prescribed in terms of the standard deviation. This property is known as the **Empirical Rule**: *For a data set that is normally distributed, the following hold true*:

1. 68% of the data lies within **one** standard deviation of the mean.
2. 95% of the data lies within **two** standard deviations of the mean.
3. 99.7% of the data lies within **three** standard deviations of the mean.



Example

The battery life of a cell phone is normally distributed with a mean of 40 hours of audio playback with a standard deviation of 1.5 hour. What percentage of these cell phones have battery life:

1. between (37,43)
2. less than 44.5
3. At least 41.5
4. less than 40
5. More than 40

```
1. 95%
2. 99.85%
3. 16%
4. 50%
5. 50%
```

Exercise

Distribution of blood pressure can be approximated as a normal distribution with mean 85 mm. and standard deviation 20 mm. What is the percentage of the people who have blood pressure:

1. between (65, 105)
2. less than 125
3. At least 85
4. More than 85
5. At least 65

In [ ]:

```
1
```

Challenging Exercise

Zack takes the SAT and his best friend Nick takes the ACT. Zack's SAT math score is 590, and Nick's ACT math score is 27. SAT math scores in the county are normally distributed, with a mean of 500 and a standard deviation of 100. ACT math scores in the county are also normally distributed, with a mean of 18 and a standard deviation of 6. Assuming that both tests measure the same kind of ability, who has the better score?

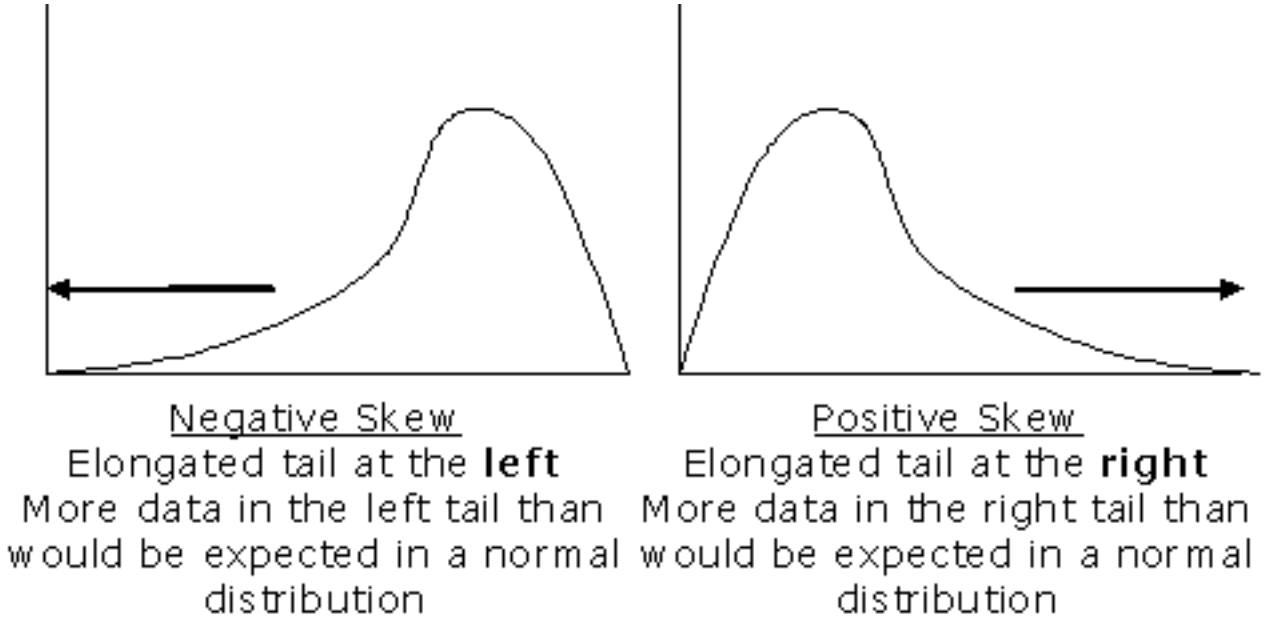
In [ ]:

```
1
```

Skewness

Skewness is a measure of **asymmetry** in the distribution of a given set of data. A given variable can be classified as either:

1. **Symmetric:** The distribution of data is mirrored evenly along both sides of the median. Also known as *unskewed*.
2. **Negatively Skewed:** The distribution of data is heavier to the right side of the graph, with the left side tapering off in an elongated tail. Also known as *skewed to the left*.
3. **Positively Skewed:** The distribution of data is heavier to the left side of the graph, with the right side tapering off in an elongated tail. Also known as *skewed to the right*.



In cases where the degree of skewness is minute, visual inspection can be difficult to carry out. To make the process of identifying the skewness of a distribution easier, most statistical texts refer to the following *rule of thumb*:

1. If the distribution is **symmetric**, the mean is **equal to** the median.
2. If the distribution is **negatively skewed**, the mean is **less than** the median.
3. If the distribution is **positively skewed**, the mean is **greater than** the median.

Using this rule of thumb, let's examine the distributions in the data sets we loaded earlier.

Guided Exercise

Overlay vertical lines on the plots to indicate where the **mean** and **median** lies for each data set. Based on the result, what is the skewness of each variable?

In [ ]:

```
g = sns.FacetGrid(cat2, col = "variable", sharex = False, sha
g.map(sns.distplot, "value")
```

In [ ]:

```
summ_cat2 = cat2.groupby("variable").agg(['mean', 'median'])

g = sns.FacetGrid(cat2, col = "variable", sharex = False, sha
g.map(sns.distplot, "value")
ax = g.axes[0]
i=0
for axis in ax:
    axis.axvline(x = summ_cat2['value']['mean'][files[i]], c
    axis.axvline(x = summ_cat2['value']['median'][files[i]],
    i=i+1
plt.legend()
plt.show()
```

We see that `Height` and `Weight` are symmetric, with `HumanLongevity` being negatively skewed and salary being positively skewed. The first two variables are known to be normally distributed, whereas the skewness visible in our human longevity and salary data can be attributed to advancement in medical care over the years and a widening income gap between lower and upper class society.

To get a better picture of the distribution for each variable, let's add on boxplots to see how much of an impact extreme values have on each set of data.

Guided Exercise

1. Add vertical lines for Q1, Q3, and the lower and upper fences to the above histograms to show the presence of outliers (if any).
2. Generate relevant boxplots for comparison with the histograms.

In [ ]:

```
q1 = lambda x: x.quantile(.25)
q3 = lambda x: x.quantile(.75)
lf = lambda x: q1(x) - 1.5 * iqr(x)
uf = lambda x: q3(x) + 1.5 * iqr(x)

q1.__name__ = "q1"
q3.__name__ = "q3"
lf.__name__ = "lf"
uf.__name__ = "uf"
summ_cat2 = cat2.groupby("variable", sort = False).agg(['mean
```

In [ ]:

```
# use Seaborn colour palette RGB values

current_palette = sns.color_palette()
g = sns.FacetGrid(cat2, col = "variable", sharex = False, sha
g.map(sns.distplot, "value")
ax = g.axes[0]
i = 0
for axis in ax:
    j = 0
    for statistic in summ_cat2['value'].columns.values:
        axis.axvline(x = summ_cat2['value'][statistic][i], c
        j += 1
    i += 1

plt.legend(loc=0)
```

In [ ]:

```
g = sns.FacetGrid(cat2, col = "variable", sharex = False)
g.map(sns.boxplot, "value")
```

One interesting point to note is that though the rule of thumb works quite easily, it can **fail** in certain cases (most notably with multimodal distributions). Various formulas have been developed over the years to compute skewness with the most common one being the **Adjusted Fisher–Pearson Standardized Moment Coefficient**, G_1 G_1:

$$G_1 = \frac{\sqrt{n(n-1)}}{n-2}\left[\frac{\frac{1}{n}\Sigma(x_i - \bar{x})^3}{\left(\frac{1}{n}\Sigma(x_i - \bar{x})^2\right)^{\frac{3}{2}}}\right]$$

$$G_1 = \frac{\sqrt{n(n-1)}}{n-2}\left[\frac{\frac{1}{n}\Sigma(x_i - \bar{x})^3}{\left(\frac{1}{n}\Sigma(x_i - \bar{x})^2\right)^{\frac{3}{2}}}\right]$$

Here x_i x_i represents the *i*-th data point, $\bar{x}$ $\bar{x}$ is the sample mean, and n n is the sample size. For the sake of brevity, we will automate evaluation of this formula using the `skew` method for Pandas objects and the `scipy.stats.skew` function.

Guided Exercise

Using the `skew` method for Pandas objects, compute the skewness of the `HumanLongevity` and `Salary` data.

In [ ]:

```
cat.skew(level = 0)
```

Equivalently, use the `skew()` function from `scipy.stats` :

In [ ]:

```
cat2.groupby("variable").agg(lambda x: skew(x, bias = False))
```

If a skewness value of **greater than 1** is obtained in either direction (positive/negative), we say that the distribution is **highly skewed**. A skewness value of 0 represents a **perfectly symmetric** distribution, which in the case of real-world data is close to impossible to observe.

In [ ]:

```
1 cat.head()
```

Exercise

Using `scipy.stats.skew`, compute the skewness of the `Height` and `Weight` data.

In [ ]:

```
1
```

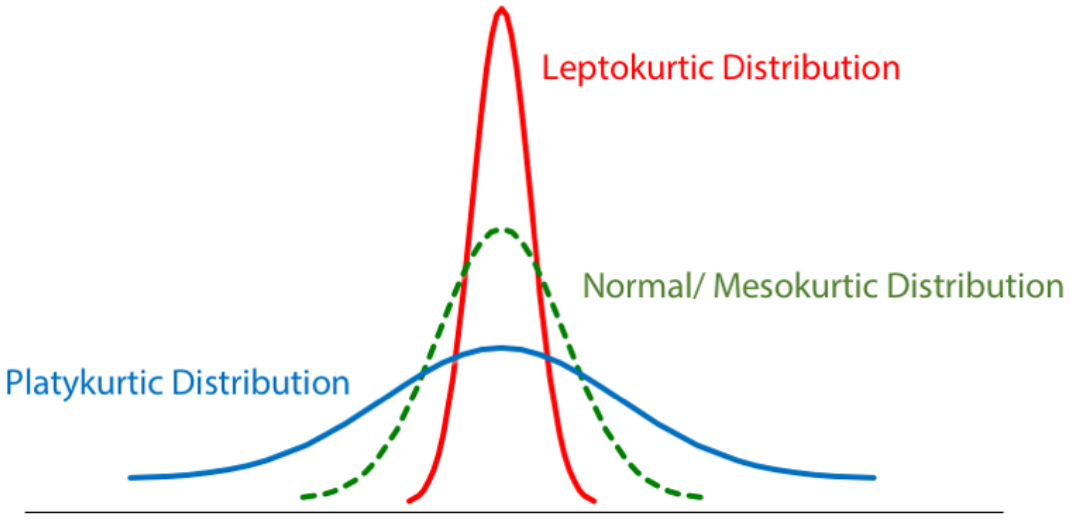
In [ ]:

```
1
```

As expected, the data for `Height` and `Weight` are not perfectly symmetrical. In practice, most distributions you encounter will have *some* degree of skewness, but we consider it to be symmetrical if the value is close to 0.

Kurtosis

Kurtosis is a measure of *tailedness*, i.e. how heavily the data is saturated in the tails of the distribution as opposed to its center. Visually, kurtosis manifests as either a *shorter* or *taller* distribution along the y-axis. This appearance is mainly due to the tails being *fatter* or *thinner* respectively.



Kurtosis uses the Normal distribution as a **point of reference**, i.e. it measures how much thinner/fatter the tails of a distribution are *compared to* a Normal distribution.

1. If the kurtosis is **negative** (< 0), we say the distribution is **platykurtic**. This means its tails are **thinner** than that of a Normal distribution. Visually, platykurtic distributions appear **shorter** in height compared to a Normal distribution.
2. If the kurtosis value is **zero** ($= 0$), we say the distribution is **mesokurtic**. This means the thickness of its tails are **identical/similar** to that of a Normal distribution.
3. If the kurtosis value is **positive** (> 0), we say the distribution is **leptokurtic**. This means its tails are **thicker** than that of a Normal distribution. Visually, leptokurtic distributions appear **taller** in height compared to a Normal distribution.

In Python, we can utilize the `scipy.stats.kurtosis` to compute the kurtosis of a given set of data.

Guided Exercise

Using `scipy.stats.kurtosis`, compute the kurtosis of the `HumanLongevity` and `Salary` data.

In [ ]:

```
cat.kurtosis(level = 0)
```

As outliers are commonly seen in the tails, kurtosis can be used as a rough indicator of the presence of outliers - a higher kurtosis value indicates thicker tails, i.e. we are more likely to encounter outliers.

Exercise

Using `scipy.stats.kurtosis`, compute the kurtosis of the `Height` and `Weight` data.

In [ ]:

```

```

In [ ]:

```

```

Additionally, kurtosis can also be seen in **Normal Quantile-Quantile Plots (Q-Q plots)** - any deviation from the shape of a normal distribution will show as points straying from a diagonal line. In most cases, we use Q-Q plots to validate assumptions of normality before proceeding with advanced statistical analysis.

In [ ]:

```
from statsmodels.graphics.gofplots import qqplot
fg = qqplot(cat["Salary"], fit = True, line = '45')
```

For comparison, let's take a look at the Q-Q plot for height, which is approximately normal:

In [ ]:

```
fg = qqplot(cat["Height"], fit = True, line = '45')
```

Other Common Distributions

The Poisson Distribution

If a variable models the number of times an event occurs in a **fixed interval of time or space**, that variable has **Poisson Distribution**. The Poisson distribution is used to describe the distribution of rare events in a large population.

A Poisson distribution:

1. Takes only non-negative numbers.
2. Is defined by the mean, $\lambda\lambda$.
3. Each occurrence is independent of the other occurrences.
4. The occurrences in each interval can range from zero to infinity.
5. The mean number of occurrences must be constant throughout the experiment.

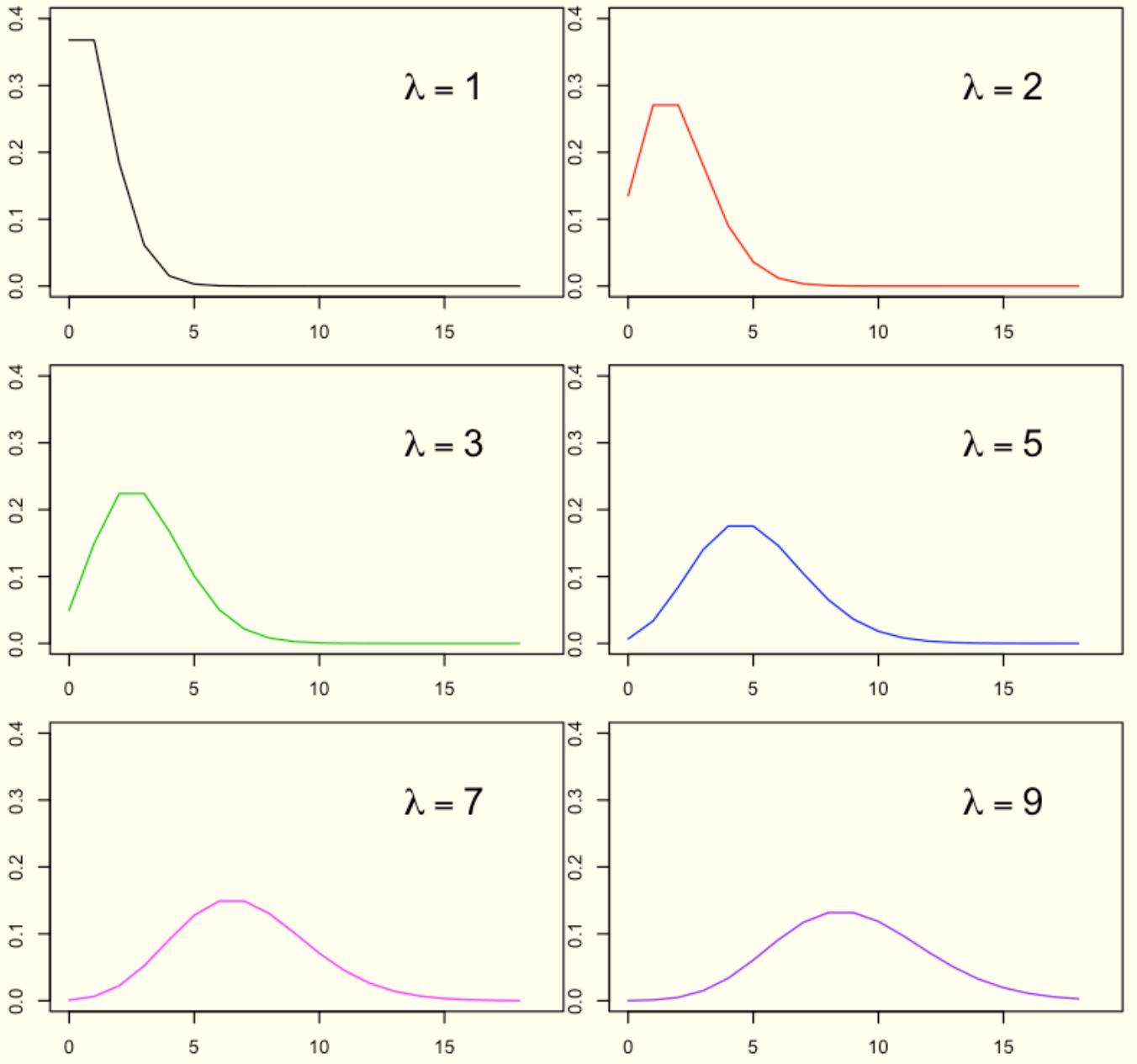
For example, if a variable measures the number of:

- Typos on a *printed page*
- Patients who enter an emergency room in *one hour*
- Customers at a Maybank ATM at Mid Valley Megamall in *10-minute intervals*
- Surface defects on a *new refrigerator*
- Repairs needed on *10 miles of highway*
- Bankruptcies that are filed in *a month*
- Arrivals at a car wash in *one hour*
- Network failures per *day*
- File server virus infection at a data center during a *24-hour period*
- Airbus 330 aircraft engine shutdowns per *100,000 flight hours*
- Asthma patient arrivals in a *given hour at a walk-in clinic*
- Work-related accidents over a *given production time*
- Birth, deaths, marriages, divorces, suicides, and homicides over a *given period of time*
- Customers who call to complain about a service problem per *month*
- Visitors to a web site per *minute*
- Calls to consumer hot line in a *5-minute period*
- Telephone calls per *minute* in a small business

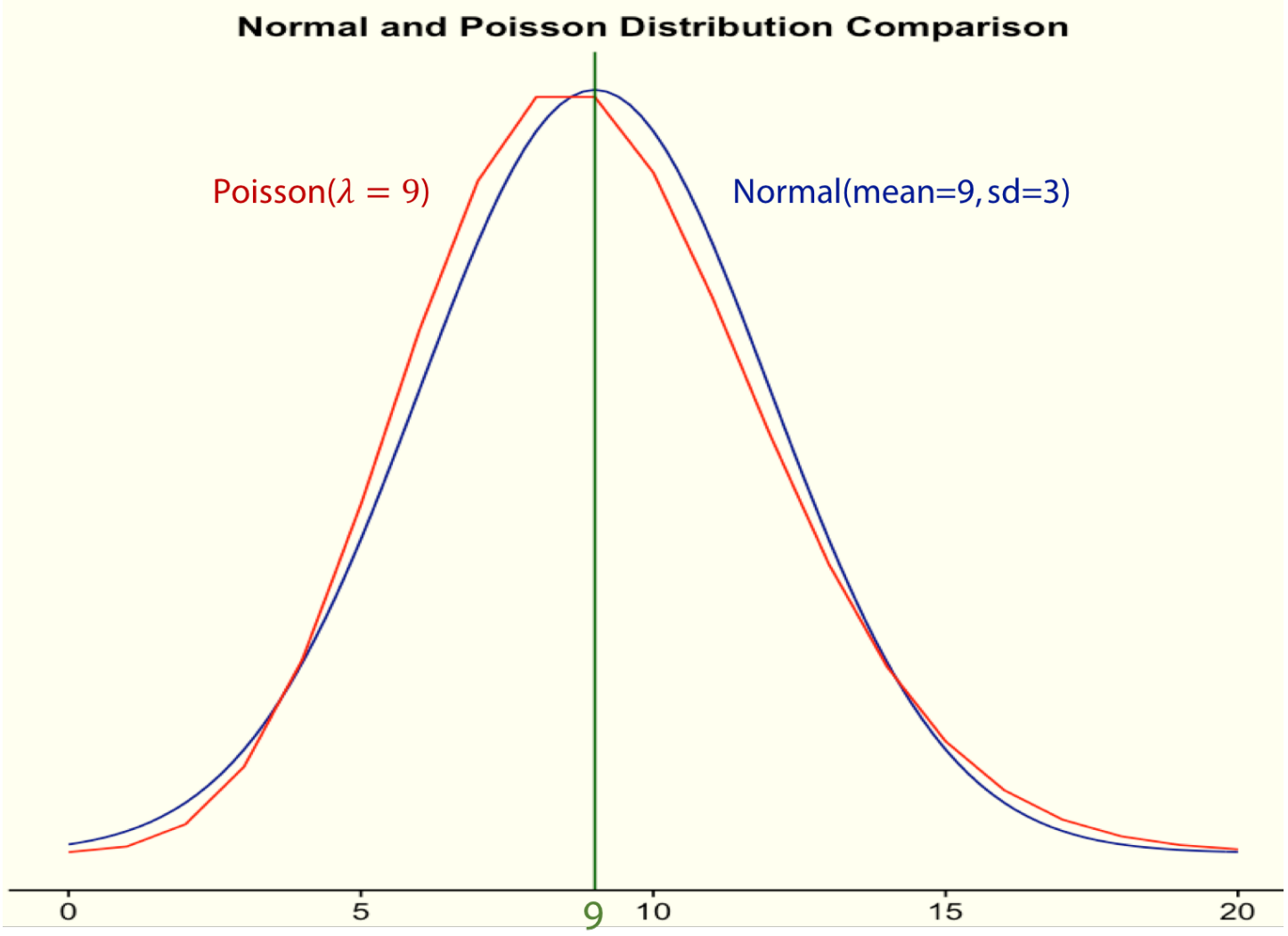
then we say that the variable is Poisson distributed.

If a variable is poisson distributed with mean $\lambda\lambda$, its standard deviation will be $\sqrt{\lambda}\sqrt{\lambda}$.

The following figure compares poisson distributions for different means ($\lambda\lambda$). It shows when the mean of the distribution increases, the plot moves to the right, its standard deviation increases and its height decreases.



The figure below compares a Normal distribution with $\mu = 9$ $\mu = 9$ and $\sigma = 3$ $\sigma = 3$ and a Poisson distribution with $\lambda = 9$ $\lambda = 9$:



We can see that the Poisson distribution is quite similar to a Normal one, albeit a little more skewed.

Example

A bank is interested in studying the number of people who use the ATM located outside its office late at night. On average, 1.3 customers walk up to the ATM during any 10 minute interval between 9pm and midnight. Here $\lambda_{10} = 1.3$ $\lambda_{10} = 1.3$.

Guided Exercise

A website receives hits at the rate of 150 per hour. What is the distribution of the number of the calls per hour for this website? Complete the following code to draw this distribution.

In [ ]:

```
x = np.linspace(100, 200, 101)
plt.plot(x, poisson.pmf(x, mu = 150))
```

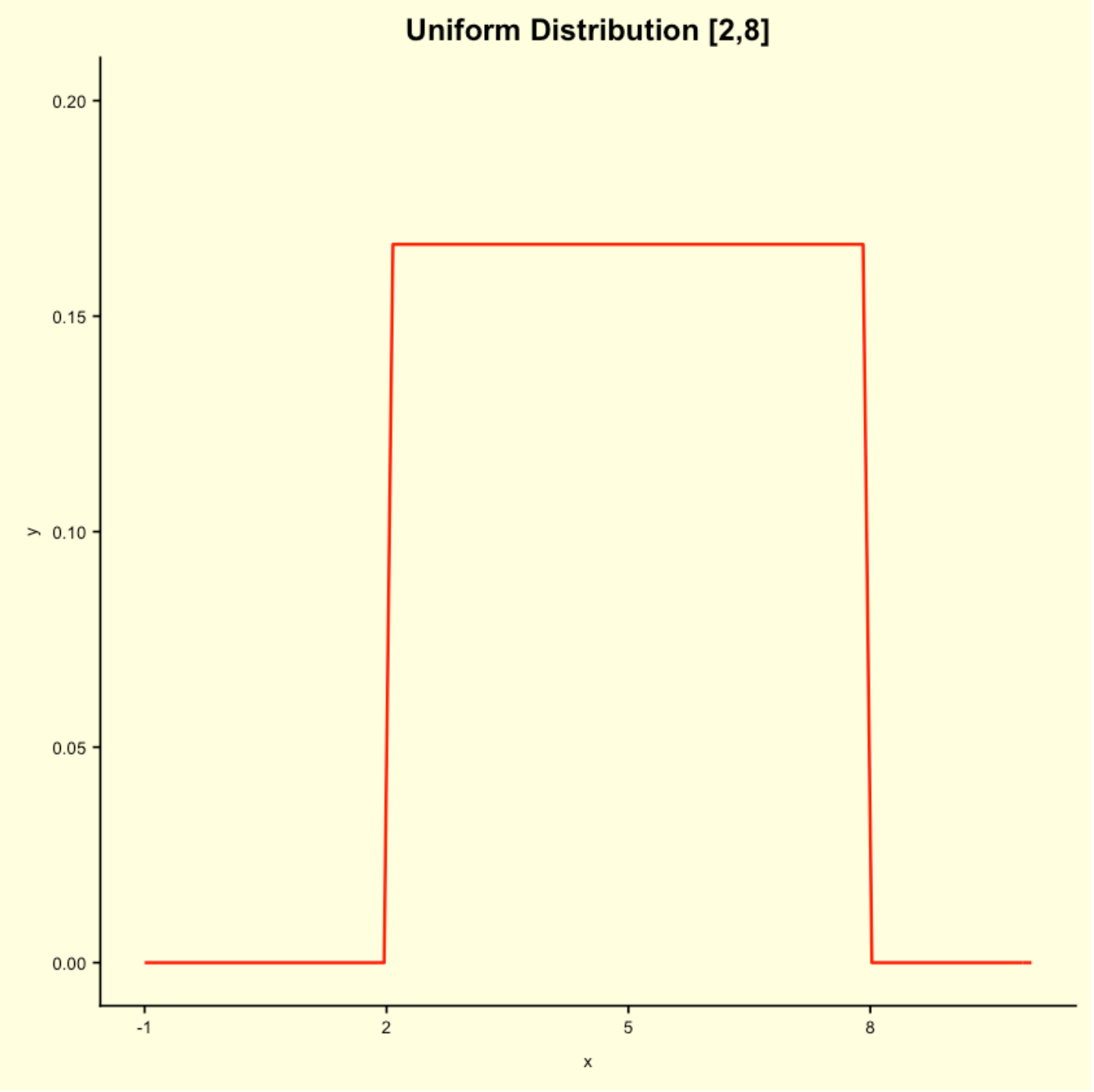
The Uniform Distribution

A variable has **Uniform Distribution** if its distribution plot looks like a rectangle or is **heavily multimodal**. This variable has the range between [a,b] but there is no information that would allow us to expect that one outcome is more likely than the others.

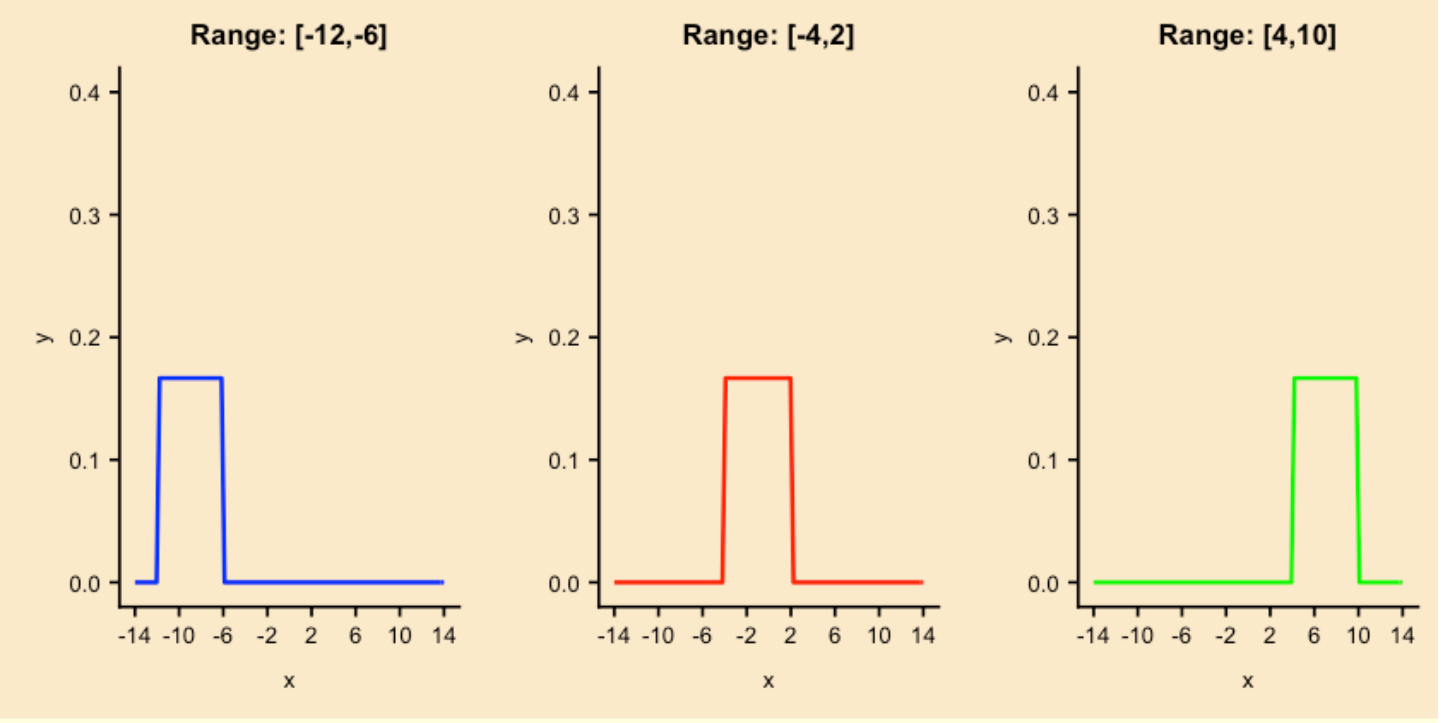
For example, if x represents:

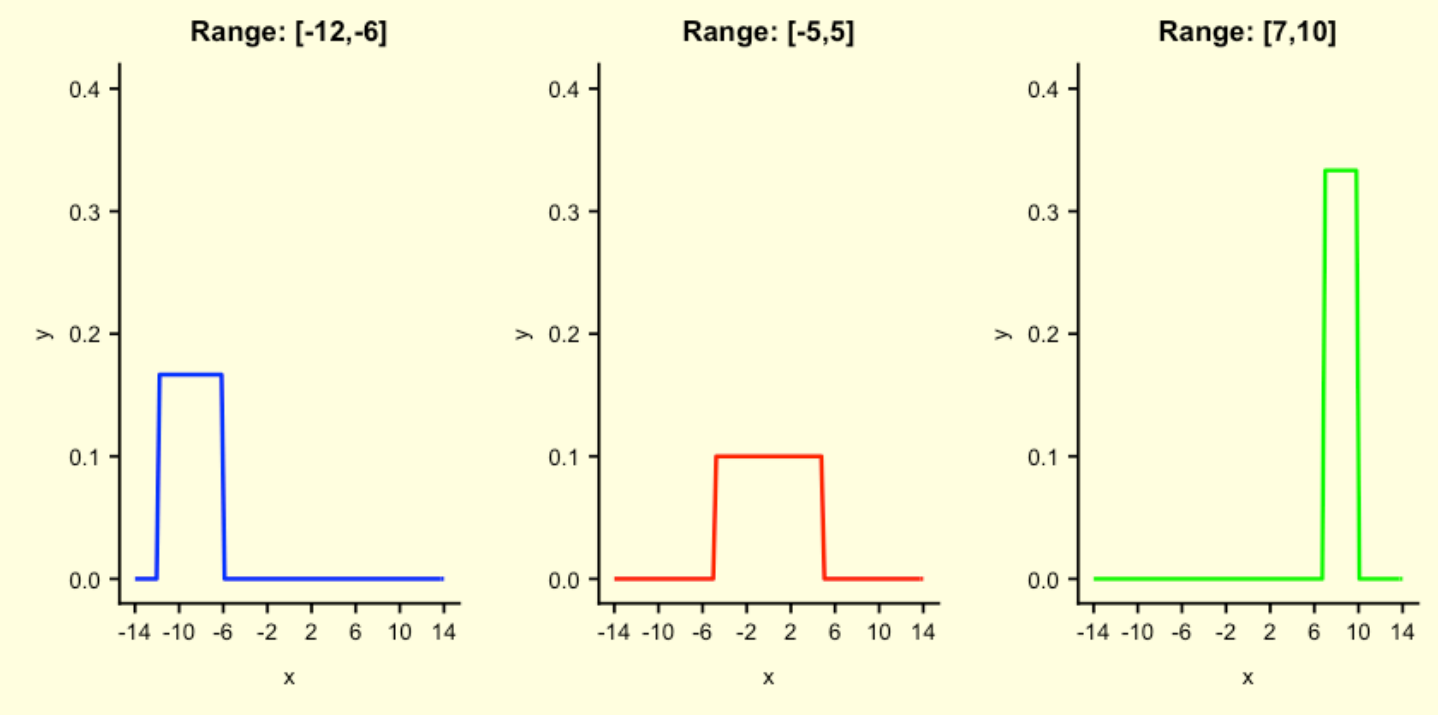
- Month of birth of a large group of people
- The day of the week of the hottest day of a year
- The last digit of the ID number
- The number that comes up from the roll of a fair die

The following figure shows uniform distribution plot in range [0,9]:



Unlike the Normal and Poisson distributions, the Uniform distribution has **no parameters**, and is instead defined on a **fixed interval**. The figure below illustrates this for a variety of different ranges:





Example

Suppose in a quiz there are 60 participants. A question is given to all of them and the time allowed to answer it is 25 seconds. The response time of each student can be any number between 0 (immediately) to 25 seconds. We can assume the response time is uniformely distributed since any respond time from 0 to and including 25 seconds is equally likely.

The following code draws the distribution of the response time in the above example:

In [ ]:

```
x = np.linspace(-10, 30, 60)
plt.plot(x, uniform.pdf(x, loc = 0, scale = 25))
plt.title('Uniform Distribution of Response Time [0,25]')
```

Guided Exercise

Load the file `Smile.csv` which contains the smiling times (in seconds) of 55 individuals in seconds, of an twelve-week old baby. Compute the five figure summary and draw a histogram with vertical lines to indicate the mean, median, first and third quartiles, and the upper and lower fences.

In [ ]:

```
smile = pd.read_csv("../data/Smile.csv")
```

In [ ]:

```
smile.info()
```

In [ ]:

```
smile.describe()
```

In [ ]:

```
summ_smile = smile.agg([q1, q3, 'mean', 'median', lf, uf])
summ_smile
```

In [ ]:

```
ax = plt.subplot()
smile.hist(bins = 30, ax = ax)

for i in range(len(summ_smile.index.values)):
    index = summ_smile.index.values[i]
    ax.axvline(summ_smile.loc[index, "Smile"], label = index,

ax.legend()
```

Additional Reading

1. Nancy R. Tague (2005). The Quality Toolbox. Summary of Histogram section available here (http://asq.org/learn-about-quality/data-collection-analysis-tools/overview/histogram2.html).

The Center of Applied Data Science

python™

Distributions

Descriptive Statistics

In [ ]:

In [ ]:

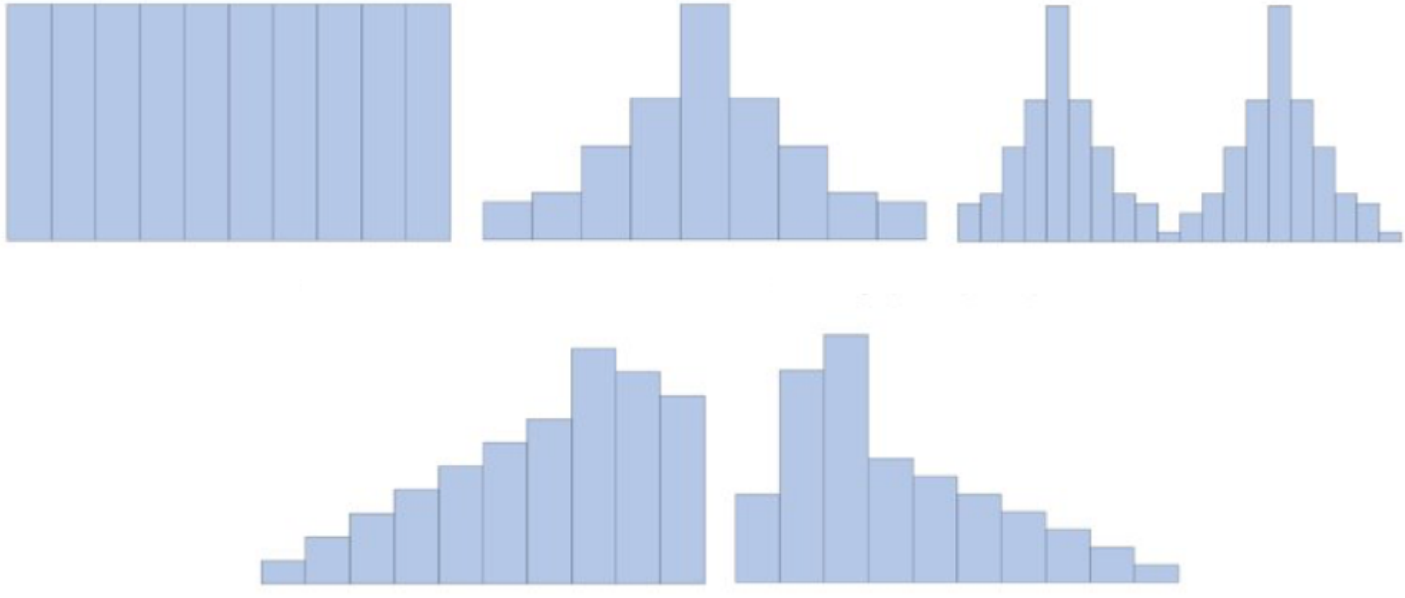
Content Outline

1. **Characterizing Distributions**
    - The Normal Distribution
    - Skewness
    - Kurtosis
2. **Other Common Distributions**
    - The Poisson Distribution
    - The Uniform Distribution

1. Characterizing Distributions

Recall that in **exploratory data analysis**, we examined a single variable by visual inspection and/or numerical summaries with the goal of summarizing the main characteristics of the data at hand. In this section on distributions, we will shift our focus towards classifying the various distributions that exist and how their general behavior can be extrapolated to explain why our data behaves the way it does.

The figure below shows the various forms in which a single variable can be distributed as observed using a histogram. Each form gives us different information regarding the variable's behavior, which may potentially change how we use the information in a business setting.



To encapsulate the motivation behind studying the general properties of common data distributions, consider the following example:

Example

Run the following code segment to load the `Height`, `Weight`, `Salary` and `HumanLongevity` data sets and plot their respective histograms.

In [ ]:

In [ ]:

In [ ]:

In [ ]:

In [ ]:

From the histograms, we see **two** main things:

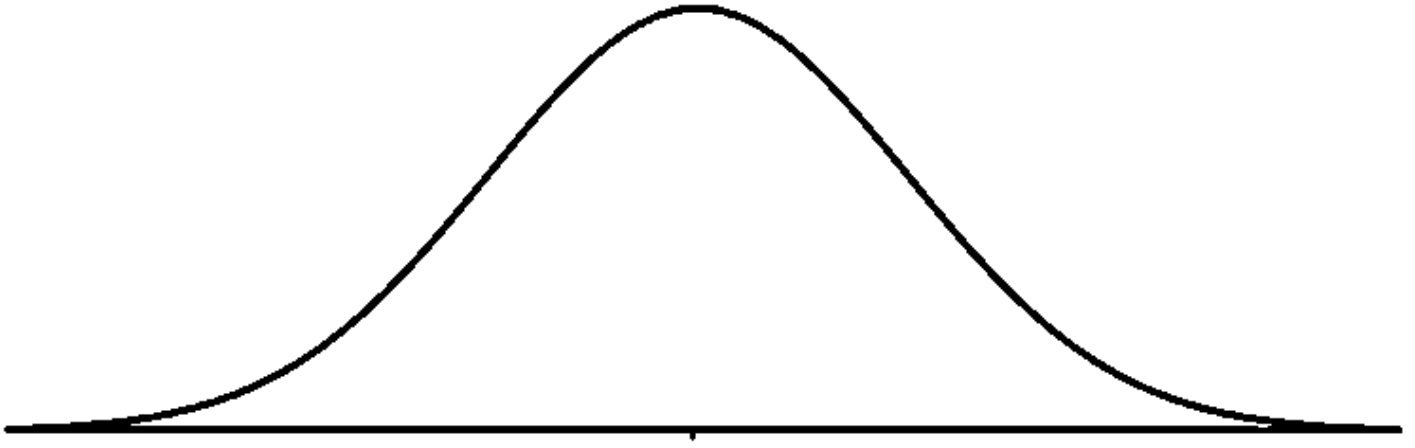
1. The general shape of each distribution is roughly similar in the sense that they all resemble a *mountain* of sorts. The location of the peaks vary from one variable to the next - `Height` and `Weight` seem to have the peak closer to the middle of the data, whereas the peaks for `HumanLongevity` and `Salary` are further to the right and left respectively. This feature is called **skewness**.
2. The heights towards the tail ends of the graphs are different. Though they may look similar in the plots, note that the y-axes are of *different scale* for the different variables. This feature is called **kurtosis**.

Before we take a deeper look at each of these features and how they impact our analysis, we need to first familiarize ourselves with the concept of **normality**.

The Normal Distribution:

A **Normal** distribution holds the following properties:

1. It forms a **bell-shaped** curve.
2. It is **symmetric** about the center.



The normal distribution is particularly interesting because one major assumption for most statistical modeling techniques is **normality**, i.e. the tendency for a set of data to be normally distributed. As such, having a variable that isn't normally distributed may limit the choice of tools we have at our disposal to model the data.

Mathematically, the bell curve of a normal distribution is given by the formula:

$$f(\mu \mid \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

where μ is the **mean** and σ^2 is the **variance**. One special case of this distribution is the **standard normal** distribution, where $\mu = 0$ and $\sigma = 1$.

In [ ]:

Adjusting the values of the **parameters** μ and σ modifies the shape of the distribution by varying the center and strech of the plot as illustrated below:

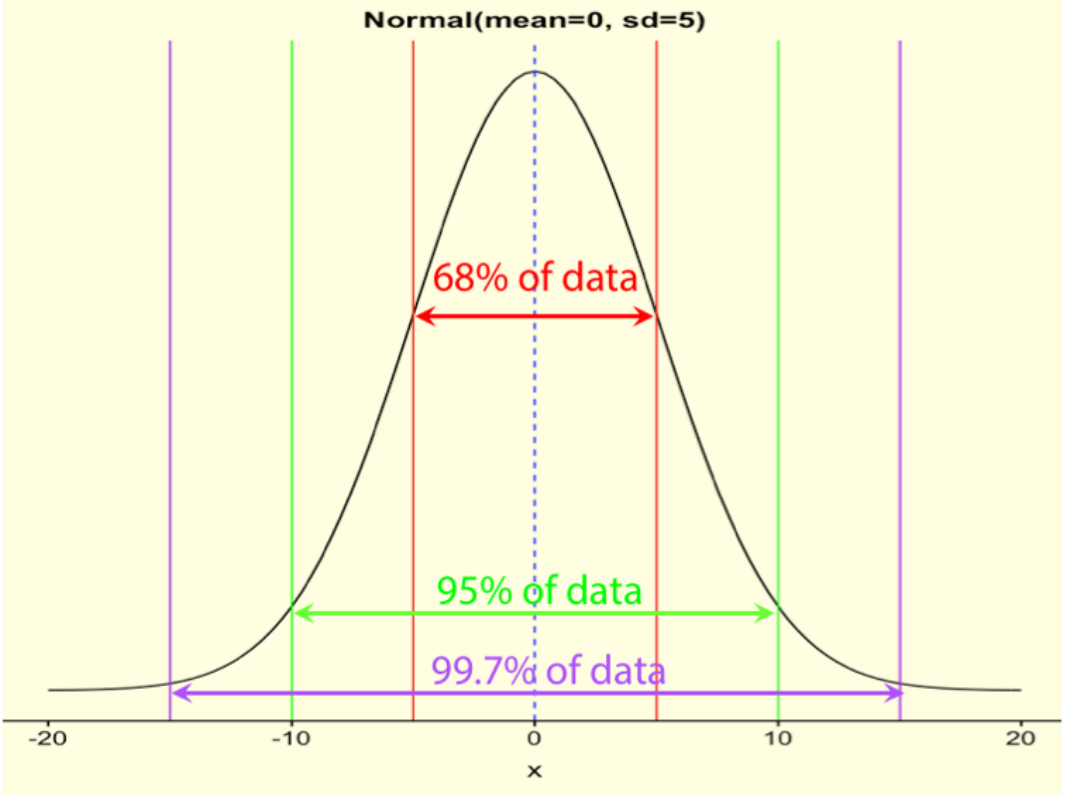
In [ ]:

In [ ]:

As we can see, increasing/decreasing the mean shifts the axis of symmetry for the curve to the right/left, whereas increasing/decreasing the variance/standard deviation streches/contracts the curve. In practice, many real-world measurable variables such as the height and weight of people, IQ, and error in physical measurements are *approximately normal* and can be modeled using a normal distribution using by choosing appropriate parameter values.

One notable property of the normal distribution is that the distribution of data within its range is prescribed in terms of the standard deviation. This property is known as the **Empirical Rule**: *For a data set that is normally distributed, the following hold true*:

1. 68% of the data lies within **one** standard deviation of the mean.
2. 95% of the data lies within **two** standard deviations of the mean.
3. 99.7% of the data lies within **three** standard deviations of the mean.



Example

The battery life of a cell phone is normally distributed with a mean of 40 hours of audio playback with a standard deviation of 1.5 hour. What percentage of these cell phones have battery life:

1. between (37,43)
2. less than 44.5
3. At least 41.5
4. less than 40
5. More than 40

Exercise

Distribution of blood pressure can be approximated as a normal distribution with mean 85 mm. and standard deviation 20 mm. What is the percentage of the people who have blood pressure:

1. between (65, 105)
2. less than 125
3. At least 85
4. More than 85
5. At least 65

In [ ]:

Challenging Exercise

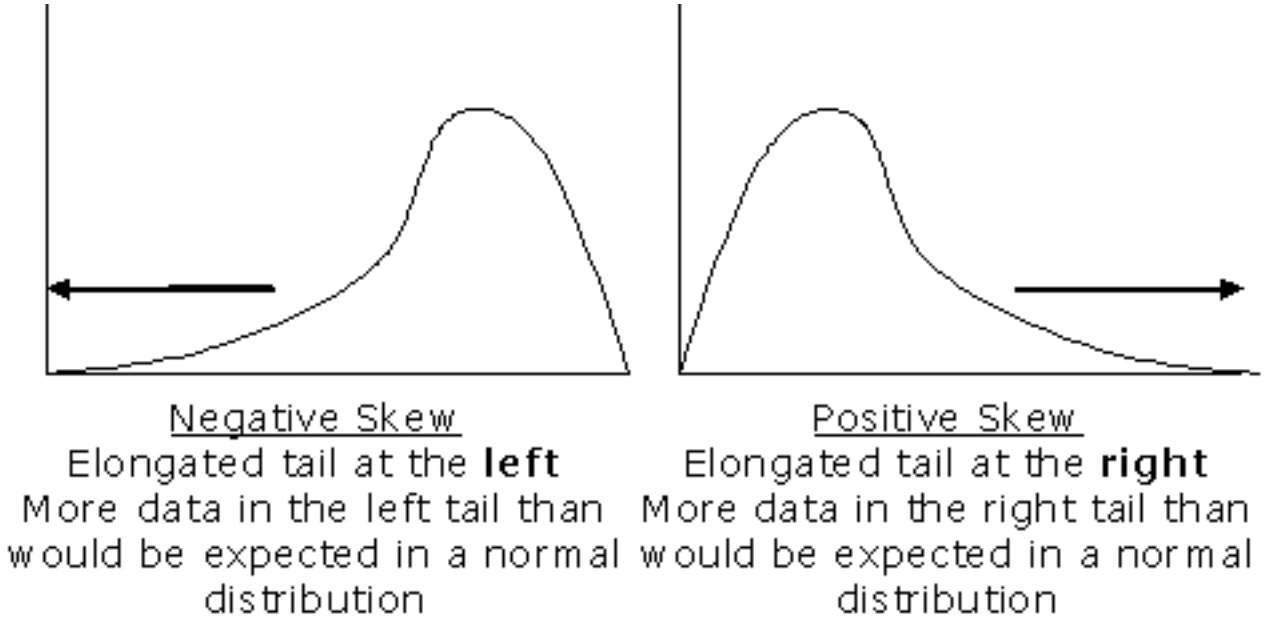
Zack takes the SAT and his best friend Nick takes the ACT. Zack's SAT math score is 590, and Nick's ACT math score is 27. SAT math scores in the county are normally distributed, with a mean of 500 and a standard deviation of 100. ACT math scores in the county are also normally distributed, with a mean of 18 and a standard deviation of 6. Assuming that both tests measure the same kind of ability, who has the better score?

In [ ]:

Skewness

Skewness is a measure of **asymmetry** in the distribution of a given set of data. A given variable can be classified as either:

1. **Symmetric:** The distribution of data is mirrored evenly along both sides of the median. Also known as *unskewed*.
2. **Negatively Skewed:** The distribution of data is heavier to the right side of the graph, with the left side tapering off in an elongated tail. Also known as *skewed to the left*.
3. **Positively Skewed:** The distribution of data is heavier to the left side of the graph, with the right side tapering off in an elongated tail. Also known as *skewed to the right*.



In cases where the degree of skewness is minute, visual inspection can be difficult to carry out. To make the process of identifying the skewness of a distribution easier, most statistical texts refer to the following *rule of thumb*:

1. If the distribution is **symmetric**, the mean is **equal to** the median.
2. If the distribution is **negatively skewed**, the mean is **less than** the median.
3. If the distribution is **positively skewed**, the mean is **greater than** the median.

Using this rule of thumb, let's examine the distributions in the data sets we loaded earlier.

Guided Exercise

Overlay vertical lines on the plots to indicate where the **mean** and **median** lies for each data set. Based on the result, what is the skewness of each variable?

In [ ]:

In [ ]:

We see that `Height` and `Weight` are symmetric, with `HumanLongevity` being negatively skewed and salary being positively skewed. The first two variables are known to be normally distributed, whereas the skewness visible in our human longevity and salary data can be attributed to advancement in medical care over the years and a widening income gap between lower and upper class society.

To get a better picture of the distribution for each variable, let's add on boxplots to see how much of an impact extreme values have on each set of data.

Guided Exercise

1. Add vertical lines for Q1, Q3, and the lower and upper fences to the above histograms to show the presence of outliers (if any).
2. Generate relevant boxplots for comparison with the histograms.

In [ ]:

In [ ]:

In [ ]:

One interesting point to note is that though the rule of thumb works quite easily, it can **fail** in certain cases (most notably with multimodal distributions). Various formulas have been developed over the years to compute skewness with the most common one being the **Adjusted Fisher–Pearson Standardized Moment Coefficient**, G_1:

$$G_1 = \frac{\sqrt{n(n-1)}}{n-2}\left[\frac{\frac{1}{n}\Sigma(x_i - \bar{x})^3}{\left(\frac{1}{n}\Sigma(x_i - \bar{x})^2\right)^{\frac{3}{2}}}\right]$$

Here x_i represents the *i*-th data point, $\bar{x}$ is the sample mean, and n is the sample size. For the sake of brevity, we will automate evaluation of this formula using the `skew` method for Pandas objects and the `scipy.stats.skew` function.

Guided Exercise

Using the `skew` method for Pandas objects, compute the skewness of the `HumanLongevity` and `Salary` data.

In [ ]:

Equivalently, use the `skew()` function from `scipy.stats`:

In [ ]:

If a skewness value of **greater than 1** is obtained in either direction (positive/negative), we say that the distribution is **highly skewed**. A skewness value of 0 represents a **perfectly symmetric** distribution, which in the case of real-world data is close to impossible to observe.

In [ ]:

Exercise

Using `scipy.stats.skew`, compute the skewness of the `Height` and `Weight` data.

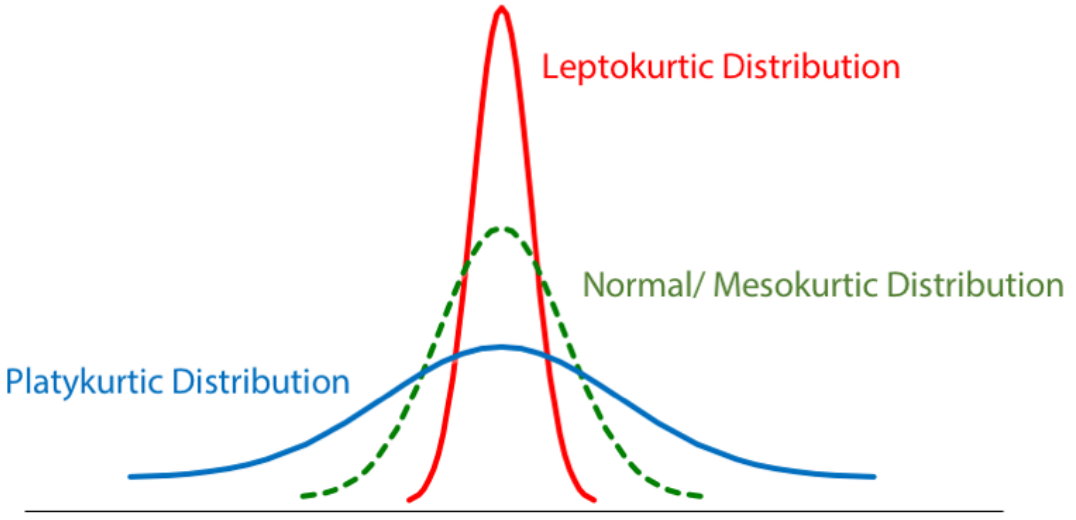
In [ ]:

In [ ]:

As expected, the data for `Height` and `Weight` are not perfectly symmetrical. In practice, most distributions you encounter will have *some* degree of skewness, but we consider it to be symmetrical if the value is close to 0.

Kurtosis

Kurtosis is a measure of *tailedness*, i.e. how heavily the data is saturated in the tails of the distribution as opposed to its center. Visually, kurtosis manifests as either a *shorter* or *taller* distribution along the y-axis. This appearance is mainly due to the tails being *fatter* or *thinner* respectively.



Kurtosis uses the Normal distribution as a **point of reference**, i.e. it measures how much thinner/fatter the tails of a distribution are *compared to* a Normal distribution.

1. If the kurtosis is **negative** (< 0), we say the distribution is **platykurtic**. This means its tails are **thinner** than that of a Normal distribution. Visually, platykurtic distributions appear **shorter** in height compared to a Normal distribution.
2. If the kurtosis value is **zero** ($= 0$), we say the distribution is **mesokurtic**. This means the thickness of its tails are **identical/similar** to that of a Normal distribution.
3. If the kurtosis value is **positive** (> 0), we say the distribution is **leptokurtic**. This means its tails are **thicker** than that of a Normal distribution. Visually, leptokurtic distributions appear **taller** in height compared to a Normal distribution.

In Python, we can utilize the `scipy.stats.kurtosis` to compute the kurtosis of a given set of data.

Guided Exercise

Using `scipy.stats.kurtosis`, compute the kurtosis of the `HumanLongevity` and `Salary` data.

In [ ]:

As outliers are commonly seen in the tails, kurtosis can be used as a rough indicator of the presence of outliers - a higher kurtosis value indicates thicker tails, i.e. we are more likely to encounter outliers.

Exercise

Using `scipy.stats.kurtosis`, compute the kurtosis of the `Height` and `Weight` data.

In [ ]:

In [ ]:

Additionally, kurtosis can also be seen in **Normal Quantile-Quantile Plots (Q-Q plots)** - any deviation from the shape of a normal distribution will show as points straying from a diagonal line. In most cases, we use Q-Q plots to validate assumptions of normality before proceeding with advanced statistical analysis.

In [ ]:

For comparison, let's take a look at the Q-Q plot for height, which is approximately normal:

In [ ]:

Other Common Distributions

The Poisson Distribution

If a variable models the number of times an event occurs in a **fixed interval of time or space**, that variable has **Poisson Distribution**. The Poisson distribution is used to describe the distribution of rare events in a large population.

A Poisson distribution:

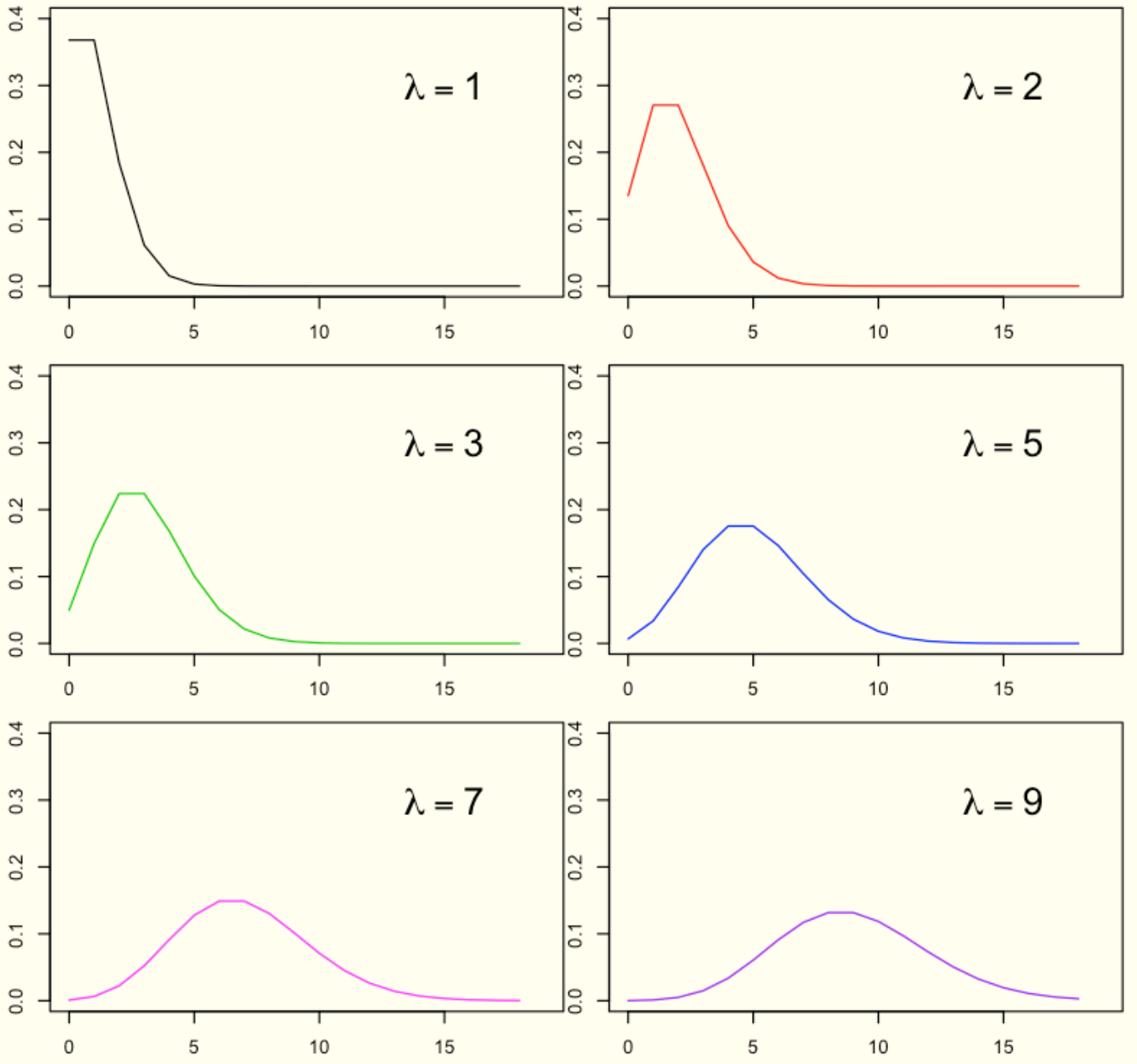
1. Takes only non-negative numbers.
2. Is defined by the mean, λ.
3. Each occurrence is independent of the other occurrences.
4. The occurrences in each interval can range from zero to infinity.
5. The mean number of occurrences must be constant throughout the experiment.

For example, if a variable measures the number of:

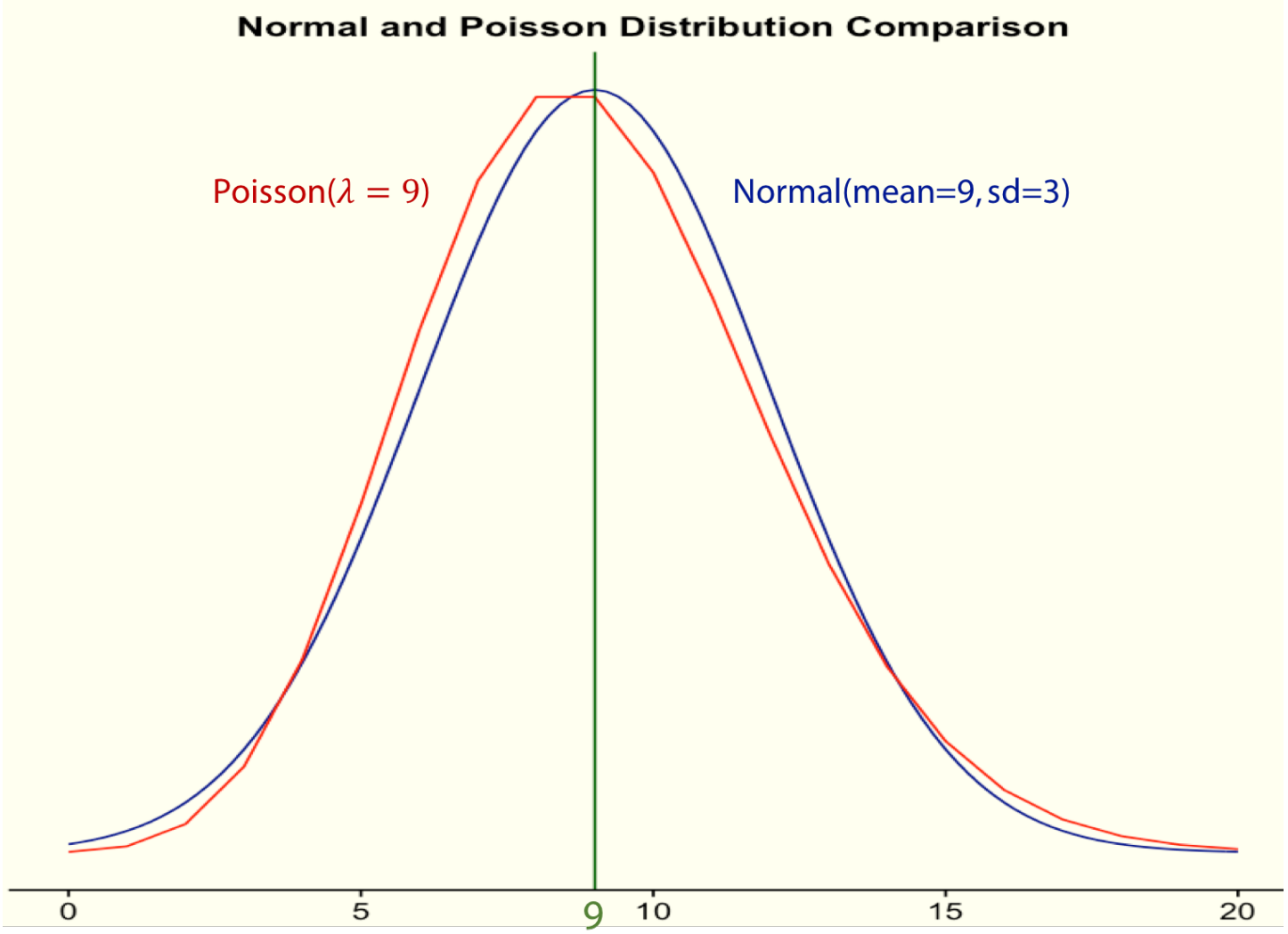
- Typos on a *printed page*
- Patients who enter an emergency room in *one hour*
- Customers at a Maybank ATM at Mid Valley Megamall in *10-minute intervals*
- Surface defects on a *new refrigerator*
- Repairs needed on *10 miles of highway*
- Bankruptcies that are filed in *a month*
- Arrivals at a car wash in *one hour*
- Network failures per *day*
- File server virus infection at a data center during a *24-hour period*
- Airbus 330 aircraft engine shutdowns per *100,000 flight hours*
- Asthma patient arrivals in a *given hour at a walk-in clinic*
- Work-related accidents over a *given production time*
- Birth, deaths, marriages, divorces, suicides, and homicides over a *given period of time*
- Customers who call to complain about a service problem per *month*
- Visitors to a web site per *minute*
- Calls to consumer hot line in a *5-minute period*
- Telephone calls per *minute* in a small business

then we say that the variable is Poisson distributed.

If a variable is poisson distributed with mean λ, its standard deviation will be $\sqrt{\lambda}$. The following figure compares poisson distributions for different means (λ). It shows when the mean of the distribution increases, the plot moves to the right, its standard deviation increases and its height decreases.



The figure below compares a Normal distribution with $\mu = 9$ and $\sigma = 3$ and a Poisson distribution with $\lambda = 9$:



We can see that the Poisson distribution is quite similar to a Normal one, albeit a little more skewed.

Example

A bank is interested in studying the number of people who use the ATM located outside its office late at night. On average, 1.3 customers walk up to the ATM during any 10 minute interval between 9pm and midnight. Here $\lambda_{10} = 1.3$.

Guided Exercise

A website receives hits at the rate of 150 per hour. What is the distribution of the number of the calls per hour for this website? Complete the following code to draw this distribution.

In [ ]:

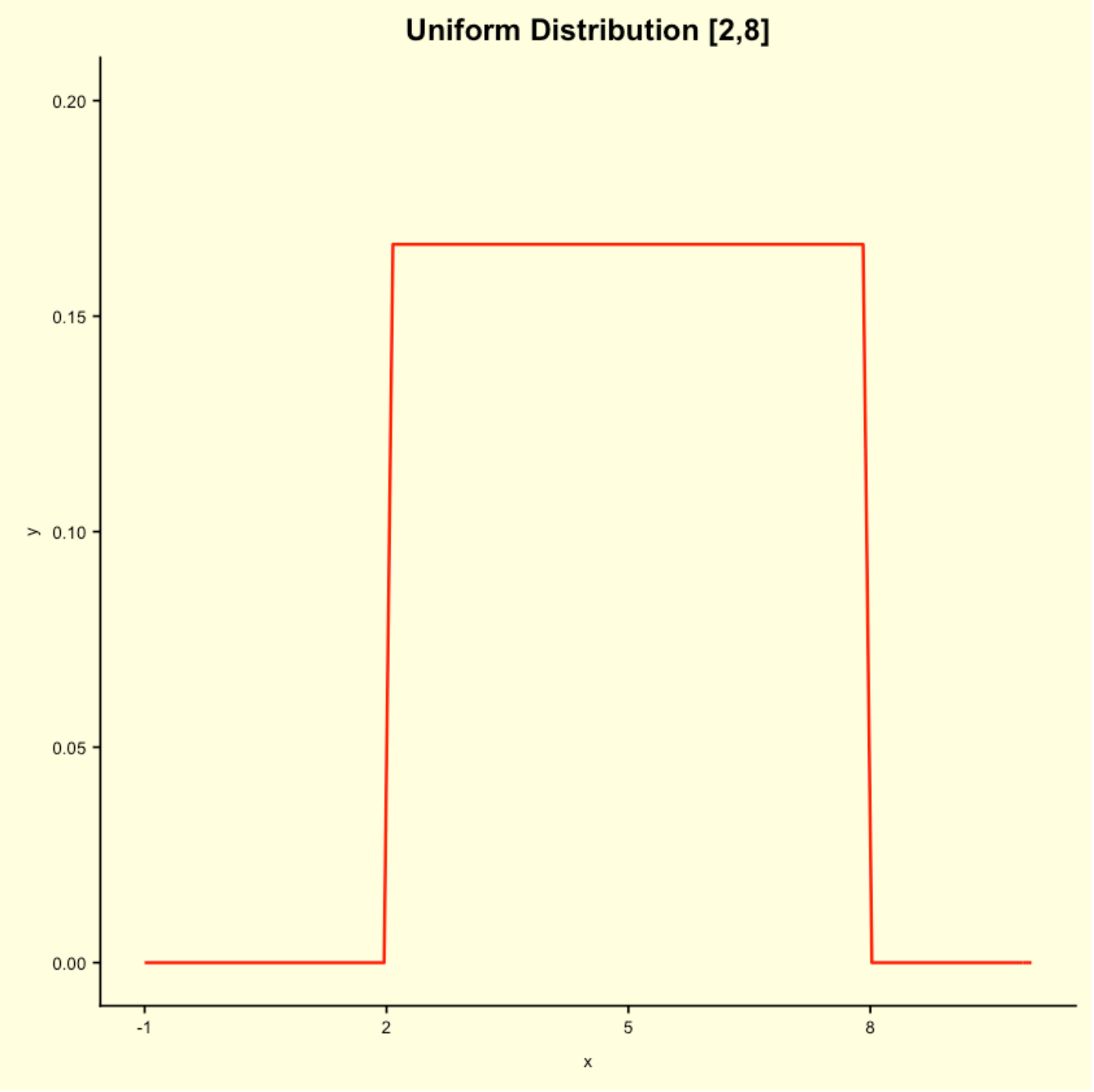
The Uniform Distribution

A variable has **Uniform Distribution** if its distribution plot looks like a rectangle or is **heavily multimodal**. This variable has the range between [a,b] but there is no information that would allow us to expect that one outcome is more likely than the others.

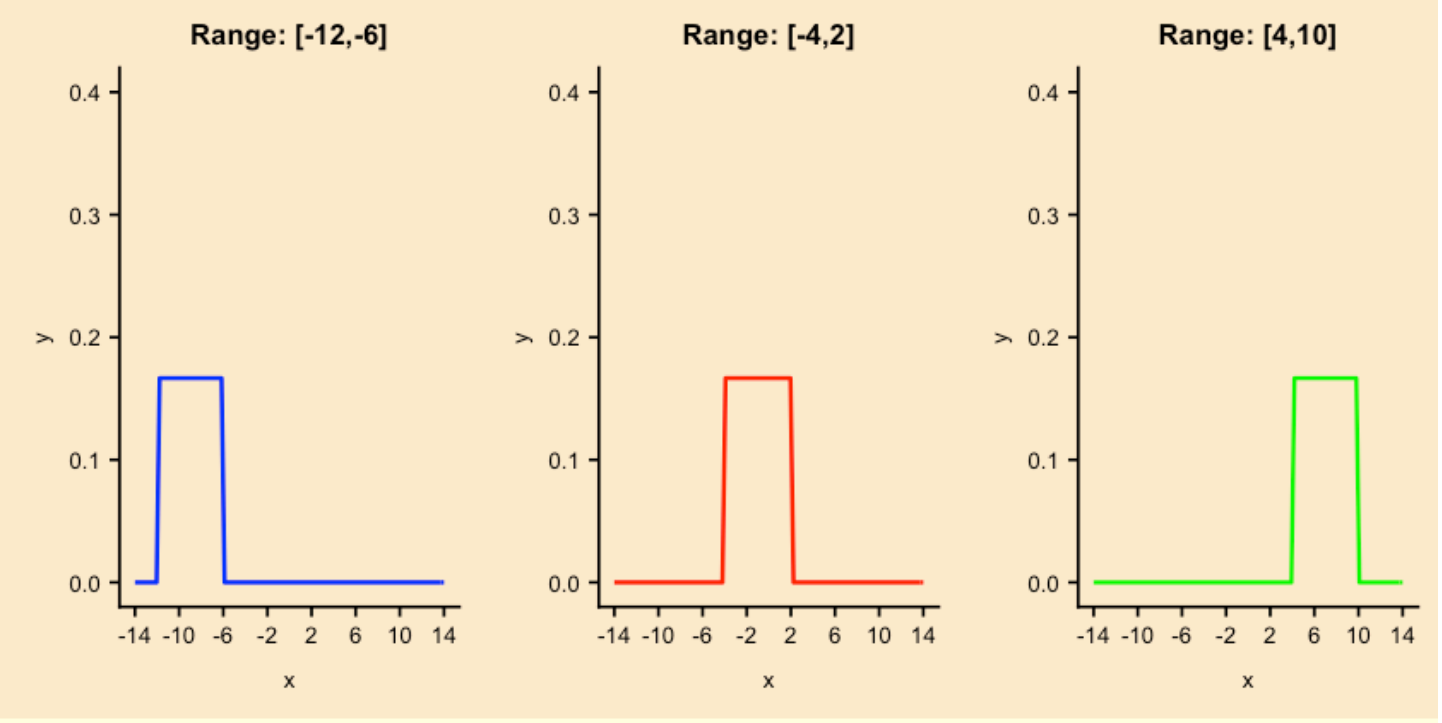
For example, if x represents:

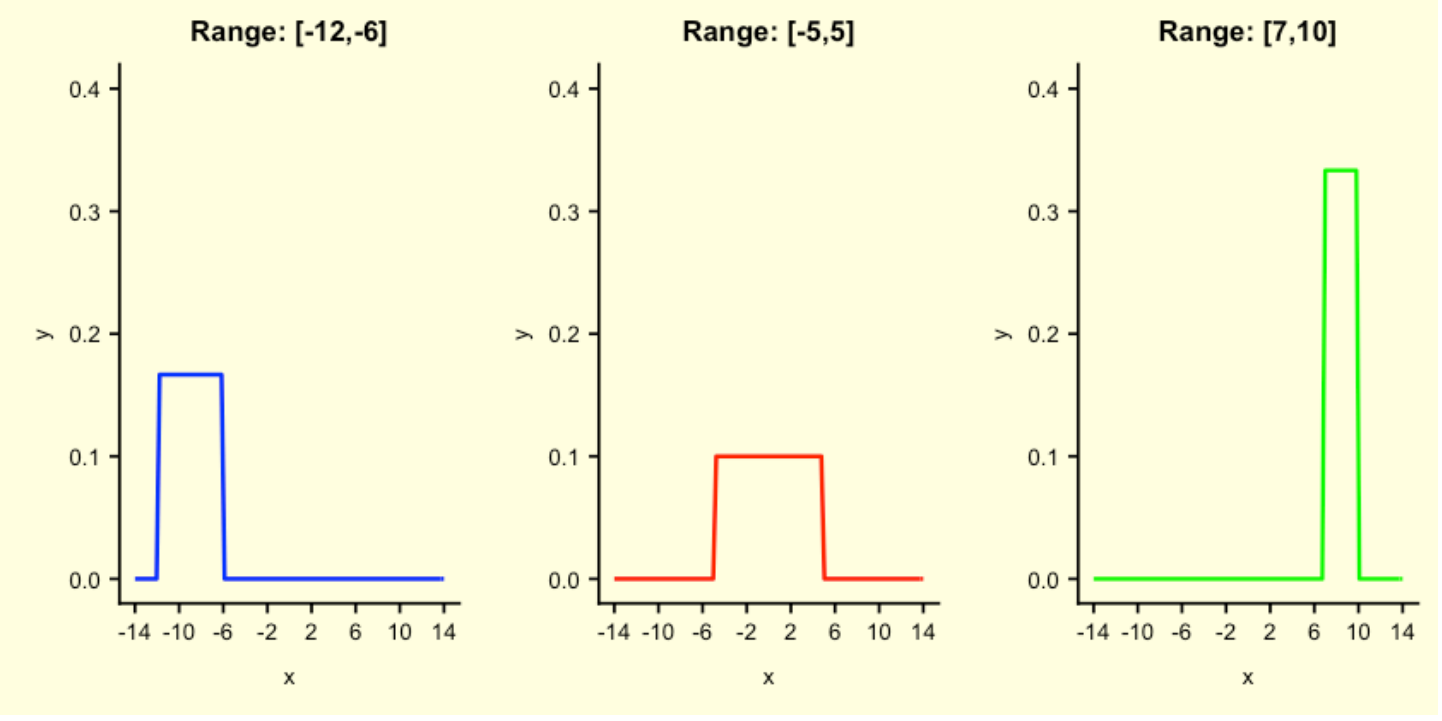
- Month of birth of a large group of people
- The day of the week of the hottest day of a year
- The last digit of the ID number
- The number that comes up from the roll of a fair die

The following figure shows uniform distribution plot in range [0,9]:



Unlike the Normal and Poisson distributions, the Uniform distribution has **no parameters**, and is instead defined on a **fixed interval**. The figure below illustrates this for a variety of different ranges:





Example

Suppose in a quiz there are 60 participants. A question is given to all of them and the time allowed to answer it is 25 seconds. The response time of each student can be any number between 0 (immediately) to 25 seconds. We can assume the response time is uniformely distributed since any respond time from 0 to and including 25 seconds is equally likely.

The following code draws the distribution of the response time in the above example:

In [ ]:

Guided Exercise

Load the file `Smile.csv` which contains the smiling times (in seconds) of 55 individuals in seconds, of an twelve-week old baby. Compute the five figure summary and draw a histogram with vertical lines to indicate the mean, median, first and third quartiles, and the upper and lower fences.

In [ ]:

In [ ]:

In [ ]:

In [ ]:

In [ ]:

Additional Reading

1. Nancy R. Tague (2005). The Quality Toolbox. Summary of Histogram section available here (http://asq.org/learn-about-quality/data-collection-analysis-tools/overview/histogram2.html).