elements_of_probability

October 11, 2021

1 DATA 5600: Introduction to Regression and Machine Learning for Analytics

1.1 Some Brief Notes on Basic Probability Concepts

Author: Tyler J. Brough Updated: October 4, 2021

[ ]:
```
import numpy as np
from scipy import stats
import matplotlib.pyplot as plt

plt.rcParams['figure.figsize'] = [10, 8]
```

1.2 Probability Distributions

These notes are based upon readings from the following books:

- *Doing Bayesian Data Analysis* by John Kruschke.
- *Mathematical Statistics with Applications* by Wackerly, Mendenhall, Scheaffer
- *Introduction to Probability and Mathematical Statistics* by Bain & Engelhardt
- *The Cambridge Dictionary of Statistics, 4th Edition*

A ***probability distribution*** is the list of all possible outcomes and their corresponding probabilities.

Notes: * In class I said that this was a *"mapping from the event space to the probability space"*

- The distribution can be represented by a graph, a table, or a formula.
- Sometimes a distinction is made been the probability *density* and the probability *distribution*, the latter being when the random variable falls at or below some particular value.
- We will use the terms interchangably and explicitly refer to the latter as the ***cumulative distribution function*** or simply the ***CDF***.

1.2.1 The Discrete Probability Density Function (Probability Mass Function)

If the set of all possible values of a random variable, X, is a countable set, $x_1, x_2, \ldots, x_n$, or $x_1, x_2, \ldots$, then X is called a **discrete random variable**. The function

$$f(x) = P[X = x] \quad x = x_1, x_2, \ldots$$

that assigns the probability to each possible value of x will be called the **discrete probability density function** (discrete PDF).

Some Common Examples

Example 1 The values of the discrete pdf of a roll of a fair die can be given by the following table.

x	1	2	3	4	5	6
f(x)	1/6	1/6	1/6	1/6	1/6	1/6

Example 2 When tossing a coin with unknown probability of heads (success). The pdf is given by the ***Bernoulli distribution function***.

$$f(x; \theta) = \theta^x (1 - \theta)^{1-x} \quad x = \{0, 1\}$$

Note:

- $P(X = 0) = 1 - \theta$
- $P(X = 1) = \theta$

A jar contains 30 green jelly beans and 20 purple jelly beans. What is the probability of drawing a single green? A single purple?

$$P(\text{drawing a single green}) = 30/50 = 0.6$$

$$P(\text{drawing a single purple}) = 20/50 = 0.4$$

We can confirm this in `Python` as follows:

[ ]:
```
stats.binom(1, 0.6).pmf(1)
```

[ ]:
```
stats.binom(1, 0.6).pmf(0)
```

Example 3 When tossing a coin n times and counting the number of heads the pdf is given by the ***Binomial distribution function***.

$$f(x; n, \theta) = \binom{n}{x} \theta^x (1 - \theta)^{n-x}$$

A jar contians 30 green jelly beans and 20 purple jelly beans. Suppose 10 jelly beans are selected at random from the jar. Find the probability of obtaining exactly five purple jelly beans if they are selected with replacement.

[ ]:
```
stats.binom(10, 0.4).pmf(5)
```

Note: often the definitions of 'success' and 'failure' are arbitrary. In this case they should be symmetric.

1.2.2 The Cumulative Distribution Function (CDF)

The **cumulative distribution function** (CDF) of a random variable X is defined for any real x by

$$F(x) = P[X \leq x]$$

Example 4 A jar contians 30 green jelly beans and 20 purple jelly beans. Suppose 10 jelly beans are selected at random from the jar. What is the probability of getting 4 or fewer green jelly beans?

[ ]:
```
stats.binom(10, 0.6).cdf(4)
```

Check this against the following and make sure it makes sense to you based on the definitions of a discrete random variable and the CDF given above:

[ ]:
```
p = 0
for i in range(5):
    p += stats.binom(10, 0.6).pmf(i)

print(f"{p : 0.10f}")
```

Q: did you get the same answer? Why or why not?

1.2.3 The Continuous Probability Density Function

A random variable X is called a **conintuous random variable** if there is a function $f(x)$ called the **probability density function** (pdf) of X, such that the CDF can be represented as

$$F(X) = \int_{-\infty}^{x} f(t)dt$$

Example 5 Suppose the research department of a steel manufacturer believes that one of the company's rolling machines is producing sheets of steel of varying thickness. The thickness is represented by a random variable following a uniform random variable with values between 150 and 200 millimeters. Any sheets less than 160 millimeters must be scrapped because they are unacceptable to buyers.

- **a.** Calculate and interpret the mean and standard deviation of x, the thickness of the sheets produced by this machine.
- **b.** Graph the probability distribution of x, and show the mean on the horizontal axis. Also show the 1- and 2-standard deviation intervals around the mean.
- **c.** Calculate the fraction of steel sheets produced by this machine that have to be scrapped.

Solution

a.

To calculate the mean and standard deviation for x, we substitute 150 and 200 millimeters for c and d, respectively, in the formulas for uniform random variables. Thus,

$$\mu = \frac{c+d}{2} = \frac{150+200}{2} = 175 \text{ millimeters}$$

and

$$\sigma = \frac{d-c}{\sqrt{12}} = \frac{200-150}{\sqrt{12}} = \frac{50}{3.464} = 14.43 \text{ millimeters}$$

b.

[ ]:
```
x = np.linspace(150, 200, 1000)
y = stats.uniform.pdf(x, loc=150, scale=50)
plt.plot(x, y, lw = 2.0, color='darkblue', alpha=0.8)
plt.fill_between(x, y, facecolor='orange', alpha=0.5)
plt.title(f"The Uniform Distribution")
plt.show()
```

c.

To find the fraction of steel sheets produced by the machine that have to be scrapped, we must find the probability that x, the thickness, is less than 160 millimeters. We need to calculate the area under the frequency function $f(x)$ points $x = 150$ and $x = 160$. Therefore, in this case $a = 150$ and $b = 160$.

We have

$$\begin{aligned} P(x < 160) &= P(150 < x < 160) \\ &= \frac{b-a}{d-c} = \frac{160-150}{200-150} = \frac{10}{50} = \frac{1}{5} = 0.20 \end{aligned}$$

That is, 20% of all the sheets made by this machine must be scrapped.

Of course, we can also simply use the Uniform CDF.

[ ]:
```
start = 150
width = 50
stats.uniform(loc=start, scale=width).cdf(160)
```

Example 6 Suppose the length of time (in hours) between emergency arrivals at a certain hospital is modeled as an *exponential distribution* with $\lambda = 2$. What is the probability that more than 5 hours pass without an emergency arrival?

Solution

The probability we want is the area under the curve to right of 5. To find this probability we use the CDF function and the complement rule.

$$P(X > 5) = 1 - P(X \leq 5) = 1 - .917915 = .082085$$

[ ]:
```
## Set the rate parameter
 = 2.0

## Use the complement rule
1 - stats.expon(scale= ).cdf(5)
```

1.2.4 The Normal Distribution

A random variable X follows the **normal distribution** with mean μ and variance σ^2 if it has the pdf

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

for $-\infty < x < \infty$, where $-\infty < \mu < \infty$ and $0 < \sigma < \infty$. This is denoted by $X \sim N(\mu, \sigma^2)$

- **NB:** parse this as: "the random variable X is normally distributed with expected value μ and variance σ^2"
- A special case is called the **Standard Normal Distribution** denoted by $X \sim N(0, 1)$ where $\mu = 0$ and $\sigma^2 = 1$

Some textbooks adopt a special notation for the standard normal with the standarized random variable $z = \frac{(x-\mu)}{\sigma}$

$$\Phi(z) = \frac{1}{\sqrt{2\pi}} e^{-z^2/2} \qquad -\infty < z < \infty$$

Example 7 Let X represent the lifetime of months of a battery, and assume that approximately $X \sim N(60, 36)$. The fraction of batteries that will fail within a four-year warranty period is given by

$$P[X \leq 48] = \Phi\left(\frac{48 - 60}{6}\right)$$

$$= \Phi(-2)$$

$$= 0.0228$$

If one wished to know what warranty period would correspond to 5% failures, then

$$P[X \leq x_{0.05}] = \Phi\left(\frac{x_{0.05} - 60}{6}\right) = 0.05$$

which means that $(x_{0.05} - 60)/6 = -1.645$, and $x_{0.05} = -1.645(6) + 60 = 50.13$ months.

[ ]:
```
## Get the z value that corresponds to the 5th percentile
## the `ppf` method is the "point percentile function", otherwise known as the inverse CDF

stats.norm.ppf(0.05)
```

[ ]:
```
  = 0.05
  = 60
  = 6
x5 =   + stats.norm.ppf( ) *
x5

print(f"\033[1m\nThe warranty that corresponds with 5% failures is {x5 : 0.2f} months\n")
```

1.2.5 The Central Limit Theorem

Let $X_1, X_2, \ldots X_n$ be independent and identically distributed (*iid*) random variables with $E(X_i) = \mu$ and $V(X_i) = \sigma^2 < \infty$. Define

$$Z_n = \sqrt{n}\left(\frac{\bar{X} - \mu}{\sigma}\right) \quad \text{where} \quad \bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

Then the distribution function of Z_n converges to a standard normal distribution as $n \to \infty$.

That is, $Z_n \xrightarrow{d} Z \sim N(0, 1)$ as $n \to \infty$.
