BTS | Barcelona Technology School

Statistical Foundations for DS MBDS 2019

$Alex\ Kumenius - Business\ Intelligence\ Data\ Scientist$
$Date : October\ 2019$

PROBABILITY

We use **Probability** to build tools to describe and understand apparent **randomness**. We often frame probability in terms of a **random process** giving rise to an **outcome**.

Probability The **Probability** of an *outcome* is the **proportion of times** the outcome would **occur** if we observed the **random process** an infinite number of times. **Probability** is defined as a ***proportion***, and it always takes **values** between **0 and 1 (inclusively)**. It may also be displayed as a **percentage** between **0% and 100%**.

Examples :

Rolling a die or a ***flipping a coin*** is a **random process** and each give rise to an **outcome**.

```
Roll a die --> 1,2,3,4,5, or 6
    Flip a Coin --> H or T
```

Probability can be illustrated by rolling a die many times. Let $\hat{p}_n$ be the **proportion** of outcomes that are 1 after the first n rolls. As the number of rolls increases, $\hat{p}_n$ will converge to the probability of rolling a 1, $p = 1/6$. The tendency of $\hat{p}_n$ to stabilize around p is described by the **Law of Large Numbers**.

Above we write "p as the probability of rolling a 1". But we can also write this probability as :

$$P(rolling\ \ a\ \ 1) \quad or \quad P(1)$$

LAW OF LARGE NUMBERS

As more observations are collected, the proportion $\hat{p}_n$ of occurrences with a particular outcome converges to the probability p of that outcome.

In [ ]:
```
import random
import operator

random.seed()

ROLLED = {i: 0 for i in range(1, 7)}
ITERATIONS = int(input('How many times would you like to roll the dice? '))

def probability():
    print("Calculation of probability: ")
    for key, count in ROLLED.items():
        print("\t{}: {:.2f}".format(key, count*100./ITERATIONS*1.))

for _ in range(ITERATIONS):
    ROLLED[random.randint(1, 6)] += 1

probability()
```

To find the **most rolled**, and **least rolled** of the die, you can use a custom operator on **ROLLED** dictionary:

In [ ]:
```
# Most rolled
max(ROLLED.items(), key=operator.itemgetter(1))
```

In [ ]:
```
# Least rolled
min(ROLLED.items(), key=operator.itemgetter(1))
```

Let's plot the **Law of Large Numbers** showing this convergence for n die rolls.

In [ ]:
```
import numpy as np
from matplotlib import pyplot as plt
from pylab import rcParams

rolls = np.random.randint(1, 7, 2)
data = []
for i in range(rolls.size):
    data.append(rolls[:i + 1].mean())
rcParams['figure.figsize'] = 10, 5
plt.gca().yaxis.grid(True)
plt.xlabel("Die Rolls")
plt.ylabel("Mean")

plt.plot(data)
plt.show()
```

However, even if a behavior is not truly random, modeling its behavior as a random process can still be useful.

DISJOINT or MUTUALLY EXCLUSIVE OUTCOMES

Two **outcomes** are called **disjoint** or if they cannot both happen. The terms **disjoint** and **mutually exclusive** are ***equivalent*** and ***interchangeable***.

For instance, if we roll a die, the outcomes 1 **and** 2 are disjoint since they **cannot** both occur.
On the other hand, the outcomes 1 and *"rolling an **odd** number"* are **not** disjoint since both occur if the outcome of the roll is a 1.

Calculating the probability of disjoint outcomes is easy.

When rolling a die, the outcomes 1 **and** 2 are disjoint, and we compute the probability that one of these outcomes will occur by **adding** their **separate probabilities**:

$$P(1 \; or \; 2) = P(1) + P(2) = 1/6 + 1/6 = 1/3$$

EXERCISE - 2.1

What about the probability of rolling a **1, 2, 3, 4, 5, or 6**?

Addition Rule

The Addition Rule **guarantees the accuracy** of this approach when the **outcomes** are disjoint.

If A_1 and A_2 represent two disjoint outcomes, then the probability that one of them occurs is given by :

$$P(A_1 \; or \; A_2) \;=\; P(A_1) \;+\; P(A_2)$$

If there are many disjoint outcomes A_1, ..., A_k, then the probability that one of these outcomes will occur is :

$$P(A_1) \;+\; P(A_2) \;+\; \cdots \;+\; P(A_k)$$

Statisticians rarely work with individual outcomes and instead consider sets or collections of outcomes.

Let A represent the event where a die roll results in 1 **or** 2 and B represent the event that the die roll is a 4 **or a** 6.
We write A as the set of outcomes {1, 2} and B = {4, 6}.

These sets are commonly called events. Because A and B have ***no elements in common***, they are disjoint events.

The **Addition Rule** applies to both disjoint outcomes and disjoint events. The **probability** that one of the **disjoint events** A or B occurs is the sum of the separate probabilities:

$$P(A) \;=\; P(1 \;\; or \;\; 2) \;=\; P(1) \;+\; P(2) \;=\; \frac{1}{6} + \frac{1}{6} \;=\; \frac{2}{6} \;=\; \frac{1}{3}$$

$$P(B) \;=\; P(4 \;\; or \;\; 6) \;=\; P(4) \;+\; P(6) \;=\; \frac{1}{6} + \frac{1}{6} \;=\; \frac{2}{6} \;=\; \frac{1}{3}$$

$$P(A \;\; or \;\; B) \;=\; P(A) \;+\; P(B) \;=\; \frac{1}{3} + \frac{1}{3} \;=\; \frac{2}{3}$$

Probabilities when Events are NOT Disjoint

Let's consider calculations for two events that are **not** disjoint in the context of a regular deck of 52 cards.

2♣	3♣	4♣	5♣	6♣	7♣	8♣	9♣	10♣	J♣	Q♣	K♣	A♣
2♢	3♢	4♢	5♢	6♢	7♢	8♢	9♢	10♢	J♢	Q♢	K♢	A♢
2♡	3♡	4♡	5♡	6♡	7♡	8♡	9♡	10♡	J♡	Q♡	K♡	A♡
2♠	3♠	4♠	5♠	6♠	7♠	8♠	9♠	10♠	J♠	Q♠	K♠	A♠

EXERCISE - 2.2

What is the probability that a randomly selected card is a diamond?

EXERCISE - 2.3

What is the probability that a randomly selected card is a face card ?

Venn diagrams is a diagram that depict all possible logical relations between a finite collection of different sets.

Venn diagrams are useful when outcomes can be categorized as ***"in"*** or ***"out"*** for two or *three variables, attributes, or random processes*.

In [ ]:
```
# Library
import matplotlib.pyplot as plt
#!pip3 install matplotlib_venn
from matplotlib_venn import venn2
%matplotlib inline
```

In [ ]:
```
# First way to call the 2 group Venn diagram:
venn2(subsets = (10, 9, 3), set_labels = ('Diamond 0.250', 'Face cards 0.231'), alpha=.4)
plt.show()
```

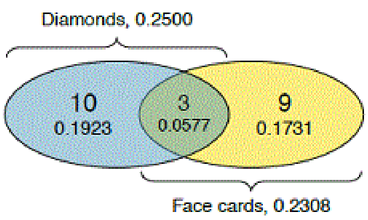
*There are also **30 cards** that are neither diamonds not face cards*

.

The **Venn diagrams** uses a circle to represent diamonds and another to represent face cards.

If a card is both a diamond and a face card, it falls into the intersection of the circles. If it is a diamond but not a face card, it will be in part of the left circle that is not in the right circle (and so on).

The total number of cards that are diamonds is given by the total number of cards in the diamonds circle: $10 + 3 = 13$. The probabilities are also shown (e.g. $10/52 = 0.1923$).



EXERCISE - 2.4

How do we compute $P(A$ or $B)$?

Let A represent the event that a *randomly selected card* is a diamond and B represent the event that it is a face card.

Events A and B are not disjoint – the cards $J(D)$, $Q(D)$, and $K(D)$ fall into both *categories* – so we **cannot** use the **Addition Rule** for disjoint events.

General Addition Rule

If A and B are any two **events**, ***disjoint or not***, then the **probability** that at least one of them will occur is :

$$P(A \quad or \quad B) \quad = \quad P(A) \quad + \quad P(B) \quad - \quad P(A \quad and \quad B)$$

where $P(A$ and $B)$ is the probability that both **events** occur.

SOLUTION - 2.4 (Part II)

*When we write **"or"** in statistics, we mean **"and/or"** unless we explicitly state otherwise. Thus, A or B occurs means A, B or both A and B occur.*

Probability Distributions

A **Probability Distribution** is a table of all **disjoint outcomes** and their associated probabilities.

A **Probability Distribution** is a list of the **possible outcomes** with corresponding probabilities that satisfies three rules: 1. The ***outcomes listed*** must be **disjoint**. 2. Each **probability** must be **between 0 and 1**. 3. The **probabilities must total 1**.

We have talked about the importance of plotting data to provide a quick summaries. So, we can also summarized in a bar plot **Probability Distributions**.

Plotting Probability Distribution Histogram (https://towardsdatascience.com/histograms-and-density-plots-in-python-f6bda88f5ac0)

Show a Bar plot with **US household** dataset and the **Probability Distribution** for the sum of two dice.

In [ ]:
```
import seaborn
from scipy.stats import binom
import matplotlib.pyplot as plt
import warnings
warnings.filterwarnings("ignore", message="Numerical issues were encountered ")

# binomial discrete random variable set
data = binom.rvs(n=17,p=0.7,loc=0,size=1010)

#univariate distribution of observations
ax = seaborn.distplot(data,
             kde=True,
             color='darkblue',
             hist_kws={"linewidth": 22,'alpha':0.77})
ax.set(xlabel='Binomial',ylabel='Frequency')
plt.show()
```

If the outcomes are numerical and discrete, it is usually convenient to make a bar plot. The **heights** represent the probabilities of outcomes.

Complement of an Event

A **set** of all possible outcomes is called the **Sample Space** (S) for rolling a die. Rolling a die produces a value in the set {1, 2, 3, 4, 5, 6}.

We often use the **Sample Space** to examine the scenario where an event does not occur.

The **Complement**, represents all outcomes in our sample space that are **not** in A. The complement is denoted by A^c.

A **complement** of an **event** A is constructed to have two very important properties: 1. every possible outcome **not** in A is in A^c, and 2. A and A^c are **disjoint**.

Property (1) implies :

$$P(A \quad or \quad A^c) = 1$$

If the outcome is **not** in A, it must be represented in A^c.

We use the **Addition Rule** for disjoint events to apply **Property (2)** :

$$P(A \quad or \quad A^c) = P(A) + P(A^c)$$

Combining **properties 1 and 2** yields a very useful relationship between the probability of an event and its complement.

Complement The complement of event A is denoted A^c, and A^c represents all outcomes **not** in A. A and A^c are mathematically related:

$$P(A) + P(A^c) = 1,$$

i.e.

$$P(A) = 1 - P(A^c)$$

Multiplication Rule for Independent Processes

Just as variables and observations / cases can be **Independent**, random processes can be Independent, too. Two processes are Independent if knowing the outcome of one provides *no useful information* about the outcome of the other.

For instance, flipping a coin and rolling a die are two independent processes – knowing the coin was heads does not help determine the outcome of a die roll. On the other hand, ***stock prices usually move up or down together****, so they are* ***not independent***.

Multiplication Rule for Independent Processes If A and B represent events from ***two different*** and ***independent*** processes, then the probability that both A and B occur can be calculated as the **product** of their **separate probabilities**:

$$P(A \quad or \quad B) = P(A) \quad x \quad P(B)$$

Similarly, if there are k events A_1, ..., A_k from k independent processes, then the probability they all occur is :

$$P(A_1) \quad x \quad P(A_2) \quad x \ldots x \quad P(A_k)$$

Rolling two dice.

We want to determine the probability that both will be 1.

Suppose one of the dice is red and the other blue. If the outcome of the red die is a 1, it provides no information about the outcome of the blue die. We first calculated the probability of both cases:

- $1/6$ of the time the red die is a 1, and
- $1/6$ of those times the blue die will also be 1.

Because the rolls are independent, the probabilities of the corresponding outcomes can be multiplied to get the final answer:

$$(\frac{1}{6}) \ x \ (\frac{1}{6}) \quad = \quad \frac{1}{36} = 0.028$$

.

This can be *generalized* to many independent processes.

CONDITIONAL PROBABILITY

We call a **Conditional Probability** because we computed the probability under a condition:

e.g.: Computing the probability a *teen attended college* based on the condition that at *least one parent has a college degree*.

The general formula for **Conditional Probability** :

Conditional Probability The **Conditional Probability** of the outcome of interest A **given** a condition B is computed as the following:

$$P(A|B) = \frac{P(A\ \ and\ \ B)}{P(B)}$$

* **Condition** is denoted with a vertical bar **"|"**, read as **given**.

CASE STUDY 2.CS

Compute the probability a $teen$ attended college based on the condition that at least one $parent$ has a college degree</i>.

*The $family\ college$ dataset contains a sample of **792 cases** with **two variables**, $teen$ and $parents$*

- *The $teen$ variable is either $college$ or not, where the $college$ label means the teen went to college immediately after high school.*
- *The $parents$ variable takes the value $degree$ if at least one $parent$ of the teenager **completed** a college degree.*

In [ ]:
```
import os
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import timeit
import random
import warnings
warnings.filterwarnings("ignore", message="Numerical issues were encountered ")
```

In [ ]:
```
# loading family_college dataset
family_college = pd.read_csv('D:\\Documents\\EureCat\\Eurecat 2019\\BTS\\Datasets\\family_college.
csv',
                             encoding='utf-8', sep=',', index_col=0)
```

In [ ]:
```
# Check dataset dimension/shape
family_college.shape
```

In [ ]:
```
# Varaible's data types
family_college.dtypes
```

In [ ]:
```
family_college.head()
```

*We considered only those **cases** that met the **condition**, $parents$ degree, and then we computed the **ratio** of those **cases** that **satisfied** our **outcome of interest**, the $teenager$ attended college.*

In [ ]:
```
# Compute a simple cross-tabulation of two (or more) factors.

pd.crosstab(family_college.teen, family_college.parents, margins=True, margins_name="Total")
```

EXERCISE - 2.5

If at least one $parent$ of a $teenager$ completed a college $degree$, what is the chance the $teenager$ attended college right after high school?

In [ ]:
```
# Compute PROPORTIONS - Percentage, a simple cross-tabulation of two (or more) factors.

pd.crosstab(family_college.teen, family_college.parents, margins=True, margins_name="Total", normalize='columns')
```

EXERCISE - 2.6

*A teenager is randomly selected from the **sample** and she **did not attend** college right after high school.*
What is the probability that at least one of her $parents$ has a college $degree$?

In [ ]:
```
pd.crosstab(family_college.teen, family_college.parents, margins=True, margins_name="Total", normalize='index')
```

Marginal and Joint Probabilities

*In any Contingency Table summary, the **totals** represent **Marginal Probabilities** for the **sample**, which are the probabilities based on a single variable -- $P(A)$ without regard to **any** other variables. Consequently a probability of outcomes for **two or more variables or processes** -- $P(A, B)$ is called a **Joint Probability**.*

> ***Marginal and Joint Probabilities** If a **probability** is based on a single variable, it is a **Marginal Probability**. The **probability** of outcomes for two or more variables or processes is called a **Joint Probability**.*

*We use **Table Proportions / Contingency Table** to summarize Joint Probabilities for the sample. These proportions are computed by dividing each count in the table by the **table's total**, to obtain the proportions.*

In [ ]:
```
pd.crosstab(family_college.teen, family_college.parents, margins=True, margins_name="Total", normalize=True)
```

General Multiplication Rule might not be Independent

***General Multiplication Rule** for events or processes that might **not be independent**.*

> *If A and B represent **two outcomes or events**, then:*
> $$P(A \ and \ B) = P(A|B) \ \ x \ \ P(B)$$
> ** The vertical bar "|" is read as **given**. * It is useful to think of A as the outcome of interest and B as the condition.*

*This **General Multiplication Rule** is simply a **rearrangement** of the definition for **Conditional Probability** equation.*

Sum of Conditional Probabilities

Let A_1, ..., A_k represent all the **disjoint outcomes** for a variable or process. Then if B is an event, possibly for another variable or process, we have:

$$P(A_1|B) \ + \ \ldots + \ P(A_k|B) = 1$$

The rule for complements also holds when an event and its complement are conditioned on the same information:

$$P(A|B) = 1 - P(A^c|B)$$

* The vertical bar "|" is read as **given**.

Independence considerations in conditional probability

If two events are independent, then knowing the outcome of one should provide no information about the other.

*We can show this is mathematically true using **conditional probabilities**.*

EXERCISE - 2.7

Let X and Y represent the outcomes of rolling two dice.

1. *What is the probability that the first die, X, is 1?*
2. *What is the probability that both X and Y are 1?*
3. *Use the formula for conditional probability to compute $P(Y = 1|X = 1)$.*
4. *What is $P(Y = 1)$? Is this different from the answer from part (3)? Explain.*

*We can show that the conditioning information has **no influence** by using the **Multiplication Rule** for **independence processes** :*

$$P(Y=1|X=1) = \frac{P(Y=1 \ and \ X=1)}{P(X=1)}$$
$$= \frac{P(Y=1) \ x \ P(X=1)}{P(X=1)}$$
$$= P(Y=1)$$