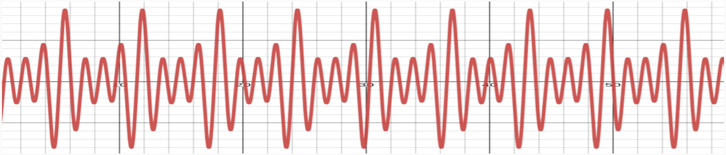
Basic Concepts



Prof. Fabrício Olivetti de França

Federal University of ABC

05 Februrary, 2024

Basic Concepts

Random Variable

A **random variable** is neither random nor a variable.

Random Variable

A ~~random variable~~ is a function $X : \Omega \to E$ where Ω is called the sample space and E is the set of possible outcomes.

Random Variable

For example, Ω can be the set of students at this graudate program and E is the age.

This function is used to extract properties of interest of random experiments.

Statistical Quantities

Let us pick a student at random, the *random variable* can be used to quantify this student somehow.

Following our example, we *convert* this student to its age.

Statistical Quantities

Applying this function to our **population** of students, we can obtain some information about them (e.g., how young is the youngest student?).

Sometimes is not possible to obtain data from the entire population, in those cases we can pick a **sample** and extrapolate from them.

With our quantified data, we can summarize our values applying an aggregation function called **statistical quantitties**.

Statistical Quantities

Some examples of statistical Quantities are:

- Mean
- Median
- Mode
- Variance
- Standard Deviation

Statistical Quantities

The measure can be of **central tendency** that returns the central point of your samples or of **spread**, showing the variation of your samples.

Measures of Central Tendency

Mean

The **mean** is given by:

$$\bar{X} = \frac{1}{N}\sum_{i=1}^{N} X_i$$

where N is the number of samples and X_i is the i-th sample. If we calculate the mean of the population we use the symbol μ.

Median

The **median** is the middle value when the list of samples is ordered. If the number of samples is even, we average the two middle values.

$$\begin{aligned} \text{Median}([1,2,3,4,5]) &= 3 \\ \text{Median}([1,2,3,4,5,6]) &= (3+4)/2 = 3.5 \end{aligned}$$

Mode

The **mode** is the most frequent value in the list of samples. If there are more than one most frequent value, the data is **multimodal**.

$$\mathrm{Mode}([1,1,1,2,2,3]) = 1$$
$$\mathrm{Mode}([1,1,1,2,2,3,3,3]) = [1,3]$$

Measures of spread

Range

The **range** is the difference between the maximum and minimum values:

$$\begin{aligned} \mathrm{Range}(X) &= \max(X) = \min(X) \\ \mathrm{Range}([1,2,3,4,5,6]) &= 6 - 1 = 5 \end{aligned}$$

Variance

The **variance** is the average squared *distance* between each sample and the central value:

$$s^2 = \frac{1}{N-1}\sum_{i=1}^{N}\left(X_i - \bar{X}\right)^2$$

$$s^2([1,2,3,4,5,6]) = 3.5$$

If we are calculating the variance for the entire population, we divide by N and use the symbol σ^2.

Variance $(N-1)$

Why do we divide by $N-1$ instead of N?

- We are estimating these quantities from a sample
- $\bar{X}$ is an estimation of μ
- If we pick many different samples and average the sample means:

$$\text{Mean}(\bar{X}) = \mu$$

Variance $(N-1)$

If we call σ_s^2 the variance of the sample (i.e., assume the sample is the population), we can pick many samples and average these variances:

$$\begin{aligned}
\sigma_s^2 &= \frac{1}{N}\sum_{i=1}^{N}\left(X_i - \bar{X}\right)^2 \\
&= \frac{1}{N}\sum_{i=1}^{N}\Big((X_i - \mu) - (\bar{X} - \mu)\Big)^2 \\
&= \frac{1}{N}\sum_{i=1}^{N}(X_i - \mu)^2 - \frac{2(\bar{X} - \mu)}{N}\sum_{i=1}^{N} X_i - \mu + \frac{(\bar{X} - \mu)^2}{N} N
\end{aligned}$$

Variance $(N-1)$

$$\begin{aligned}\bar{X}-\mu &= \frac{1}{N}\sum_{i=1}^{N} X_i - \mu \\ &= \frac{1}{N}\sum_{i=1}^{N}(X_i-\mu) \\ n(\bar{X}-\mu) &= \sum_{i=1}^{N} X_i - \mu\end{aligned}$$

Variance $(N-1)$

$$\begin{aligned}\sigma_s^2 &= \frac{1}{N}\sum_{i=1}^{N}(X_i-\mu)^2 - 2(\bar{X}-\mu)^2 + (\bar{X}-\mu)^2 \\ &= \frac{1}{N}\sum_{i=1}^{N}(X_i-\mu)^2 - (\bar{X}-\mu)^2\end{aligned}$$

Variance $(N-1)$

Com isso:

$$\begin{aligned}\mathrm{Mean}(\sigma_s^2) &= \mathrm{Mean}(\frac{1}{N}\sum_{i=1}^{N}(X_i-\mu)^2) - \mathrm{Mean}((\bar{X}-\mu)^2)\\ &= \sigma^2 - \frac{1}{n}\sigma^2\\ &= \frac{n-1}{n}\sigma^2\end{aligned}$$

Variance $(N-1)$

So, dividing by $N-1$ is a correction of our values to not understimate the estimation of the population variance.

Degrees of Freedom

This is linked to the concept of **degrees of freedom** (ν) and it is defined as the number of independent observations (N) minus the number of parameters used in the calculation.

So, for sample variance we have N observations and we use one parameter (the sample mean), leading to $\nu = N - 1$.

Standard Deviation

The sample **standard deviation** (s) or populational standard deviation (σ) is given by

$$s = \sqrt{\frac{1}{N-1}\sum_{i=1}^{N}\left(X_i - \bar{X}\right)^2}\sigma \quad = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(X_i - \mu)^2}$$

This measure of spread brings the variance to the same unit of our samples (instead of the square of differences).

Interquantile Range

The **interquantile range** is calculated as the difference between 75th and 25th percentiles of the sample.

The kth percentile is the value in which k% of the data falls below it.

The median is the 50th percentile as 50% of the data is below that value. To calculate the 25th and 75th percentiles we just calculate the median of the subset of the data below and above the median.

$$\begin{aligned}\mathrm{IQR}([1,1,2,4,5,6,7]) &= \mathrm{Median}([5,6,7]) - \mathrm{Median}([1,1,2]) \\ &= 6 - 1 = 5\end{aligned}$$

Why do we need all of them?

The difference between *mean*, *median*, *mode* becomes clear depending on how our values are **distributed**.

Data Distribution

Let us say we have a dice and our random variable X maps the face of the dice to its corresponding number.

The results of throwing a dice depends on many variables. The current position of the dice, the applied force, angle, etc.

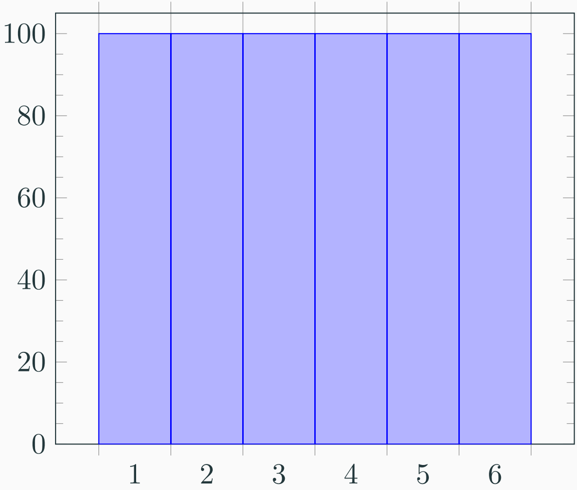
Data Distribution

In a perfect world, we could measure every variable and make an exact prediction of the result!

Imagining this perfect world, what would happen if we throw the dice in every possible start condition and counted the frequency of every outcome?

Data Distribution

We expect to see something like this:



The number of times we observe each value of the dice is constant.

Data Distribution

In relative frequency, we can say that each value happens $1/6$ of the time.

The frequencies are **uniformly distributed** among each possible value.

Data Distribution

The statistical quantities of the population of a dice throw are:

$$\begin{aligned} \bar{X} &= 3.5 \\ \text{Median} &= 3.5 \\ \text{Mode} &= [1, 2, 3, 4, 5, 6] \end{aligned}$$

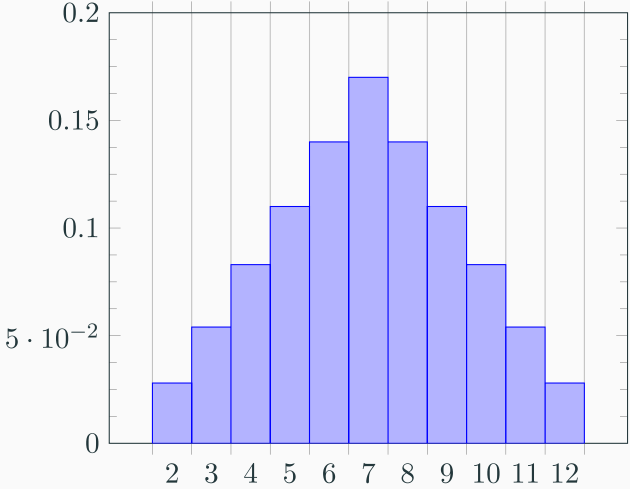
Notice how the mean and the median are the same! Since every value has the same frequency, the mode contains every value.

Data Distribution

Bartoš, František, et al. ”Fair coins tend to land on the same side they started: Evidence from 350,757 Flips.” arXiv preprint arXiv:2310.04153 (2023).

Data Distribution

Let us change our random variable to the sum of the values of the faces of two dice. This is what happens:



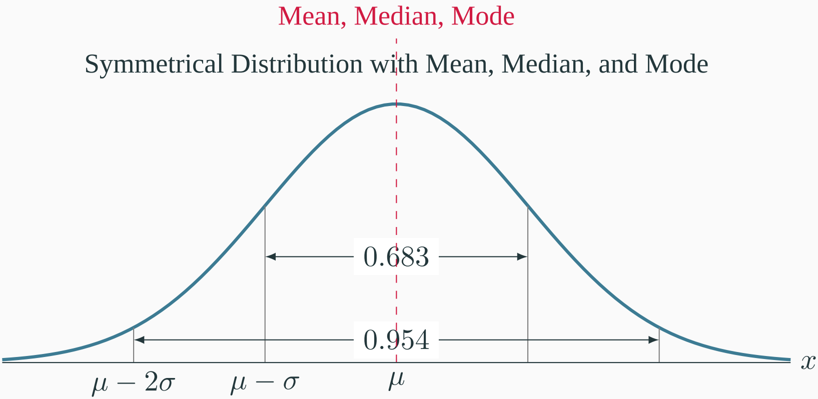
Data Distribution

The statistical quantities of the population of two dice throw are:

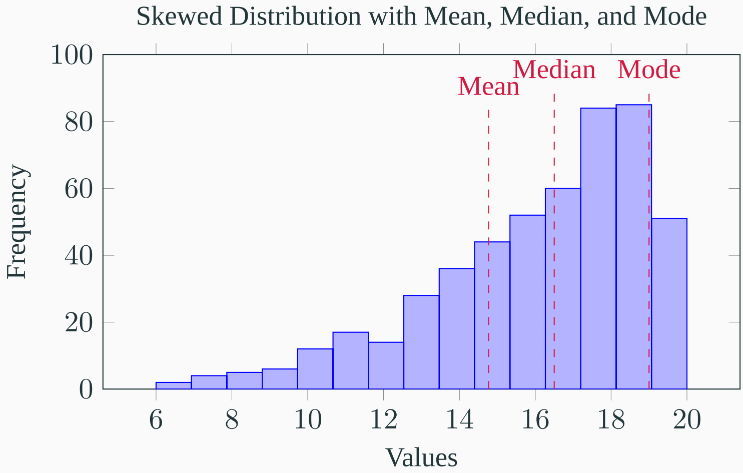
$$\begin{aligned} \bar{X} &= 7 \\ \text{Median} &= 7 \\ \text{Mode} &= 7 \end{aligned}$$

In this distribution, all the central measures coincide.

Data Distribution - Symmetrical



Data Distribution - Skewed / Asymmetrical



Data Distribution

From these plots we can see that:

- The **mean** is the point closest on average to every point in your sample. It is affected by unusually small or large values (outliers). Should we count the salary of a CEO to estimate the mean wage? Should we consider a student that never attended any class?
- The **median** says that 50% of the values are smaller or larger than that. A median wage will aleviate the influence of the CEO salary.
- The **mode** describes the most frequent values. What is the most frequent attendance count?

Data Distribution

Understanding how the data is distributed provides us with richer information.

A **probability distribution** is a function that maps events to their corresponding probabilities.

Data Distribution

The distribution of a fair coin is described by:

$$P(H) = 0.5$$
$$P(T) = 0.5$$

Data Distribution

The **probability function** P is the probability that a certain event occurs.

The **cumulative distribution function** cdf is the probability that a random variable X assumes a value less than or equal to x.

Data Distribution

The probability distribution can be **discrete** or **continuous**.

- **Discrete distribution:** when the set of events is finite or countable infinitely.
- **Continuous distribution:** when the set of evets is uncountable.

Data Distribution

If the event is discrete we describe the distribution with a **probability mass function (pmf)** $f(x)$ such that:

$$\sum_x f(x) = 1$$

Data Distribution

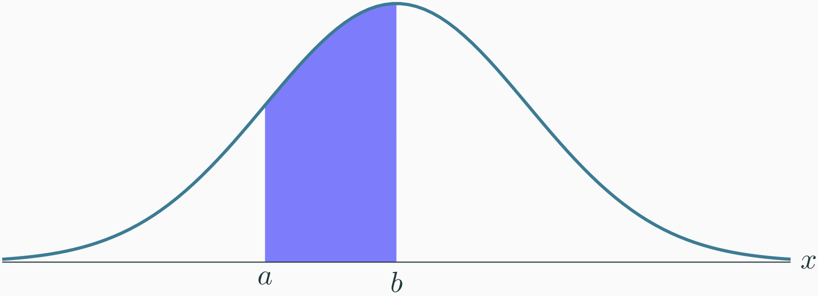
If the event is discrete we describe the continuous with a **probability density function (pdf)** that describes the probability of a range of values $a \leq x \leq b$:

$$P(a \leq x \leq b) = \int_a^b f(x)dx$$

Data Distribution

The *integral* of a range $[a, b]$ is the area of the function between those two points:

$$P(a \leq x \leq b) = \int_a^b f(x)dx$$



Parametrization

If the pmf or pdf depends on additional parameters, we will write $f(x \mid \mu, \sigma)$ as the value of the function is conditioned by the values of these parameters.

Another common notation is $f(x; \mu, \sigma)$.

Expected Value

The **expected value** or **expectation** is the weighted sum of each possible event by their corresponding probability:

$$E[f(x)] = \sum_x x f(x)$$

$$E[f(x)] = \int_{-\infty}^{\infty} x f(x) dx$$

This is equivalent to the population mean (μ).

Moments

The n-th **moment** of a distribution for $n > 1$ is:

$$E[(x-\mu)^n]$$

The first moment $n = 1$ is the expected value (mean, $E[x]$), the second moment $n = 2$ is the variance ($Var[x]$), the third moment $n = 3$ is a measure of skewness.

Discrete Distributions

Bernoulli

$$f(x \mid p) = \begin{cases} p & \text{if } x = 1 \\ 1 - p & \text{if } x = 0 \end{cases}$$

$$f(x \mid p) = p^x (1 - p)^x$$

$$f(x \mid p) = px + (1 - p)(1 - x)$$

- **Support:** $x \in 0, 1$
- Answers a yes/no question
- Probability that some experiment is successful ($x = 1$)
- Will a student attend class today?

Bernoulli

$$E[f(x \mid p)] = p$$

$$Var[f(x \mid p)] = p(1 - p)$$

Bernoulli

$$p = 0.5$$

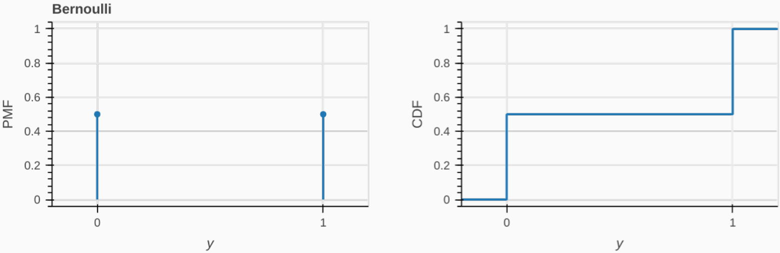


Figure 1: https://distribution-explorer.github.io/discrete/bernoulli.html

Binomial

$$f(x \mid n, p) = \binom{n}{x} p^x (1-p)^{n-x}$$

- **Support:** $x \in \mathbb{Z}, x \leq n$.
- How many success in a sequence of n independent trials with a probability of success p
- If I pick n students at random, how may will have attended class today?

Binomial

$$E[f(x \mid n, p)] = np$$
$$Var[f(x \mid n, p)] = np(1-p)$$

Binomial

$$n = 20, p = 0.5$$

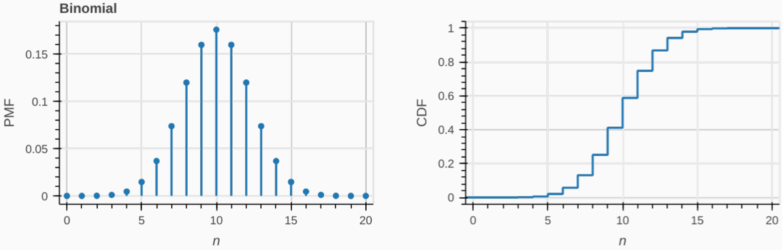


Figure 2: https://distribution-explorer.github.io/discrete/binomial.html

Binomial

$$f_{\text{Bernoulli}}(x \mid p) = f_{\text{Binomial}}(x \mid 1, p)$$

Poisson

$$f(x \mid \lambda) = \frac{\lambda^x e^{-\lambda}}{x!}$$

- **Support:** $x \in \mathbb{Z}, x \geq 0.$
- How many events will happen at a time frame if the rate of occurrence is λ?
- How many students will raise hands and ask questions within one hour?
- Assumes events are independent

Poisson

$$E[f(x \mid \lambda)] = \lambda$$
$$Var[f(x \mid \lambda)] = \lambda$$

Poisson

$$\lambda = 5$$

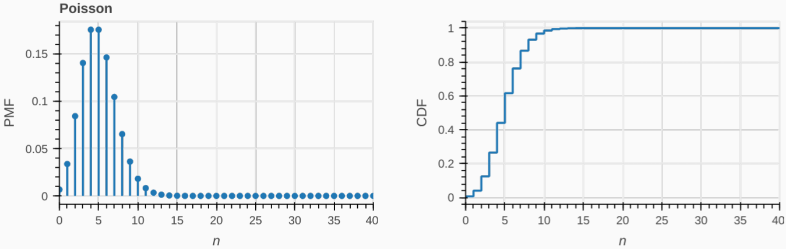


Figure 3: https://distribution-explorer.github.io/discrete/poisson.html

Multinomial

$$f(x_1, x_2, \ldots, x_n \mid p_1, p_2, \ldots, p_n, N) = \frac{N!}{x_1! x_2! \ldots x_n!} p_1^{x_1} p_2^{x_2} \ldots p_n^{x_n}$$

- **Support:** $x \in \mathbb{N}^n$.
- What is the probability of having a sequence of n outcomes given a different probability for each outcome
- $\sum_p p = 1$

Multinomial

$$
\begin{aligned}
E[f(x_i \mid p_i, N)] &= Np_i \\
Var[f(x_i \mid p_i, N)] &= Np_i(1 - p_i)
\end{aligned}
$$

https://distribution-explorer.github.io/multivariate_discrete/multinomial.html

Continuous Distributions

Gaussian

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

- **Support:** $x \in \mathbb{R}$.
- Symmetrical distribution, a central value is very likely and the neighbor valules are more rare
- Central limit theorem: average of many samples converge to Gaussian
- Also known as normal distribution
- The height of the students is likely normally distributed

Gaussian

$$E[f(x \mid \mu, \sigma)] = \mu$$
$$Var[f(x \mid \mu, \sigma)] = \sigma^2$$

Gaussian

$$\mu = 0, \sigma = 0.2$$

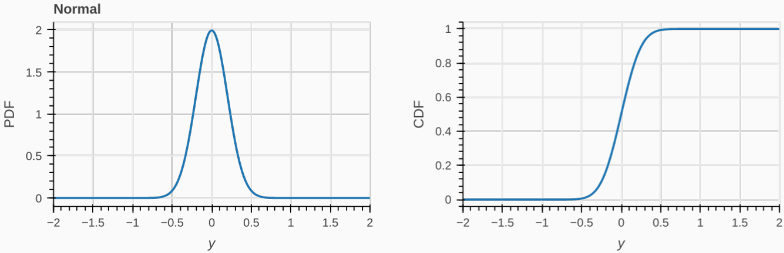


Figure 4: https://distribution-explorer.github.io/continuous/normal.html

Gamma

$$f(x \mid \alpha, \beta) = \frac{x^{\alpha-1} e^{-\beta x} \beta^{\alpha}}{\Gamma(\alpha)}$$

$$\Gamma(\alpha) = (\alpha - 1)!$$

- **Support:** $x \in \mathbb{R}, x \geq 0$.
- If events happen at the same rate (β), what is the probability of α events to happen.

Gamma

$$E[f(x \mid \alpha, \beta)] = \frac{\alpha}{\beta}$$

$$Var[f(x \mid \alpha, \beta)] = \frac{\alpha}{\beta^2}$$

Gamma

$$\alpha = \beta = 2$$

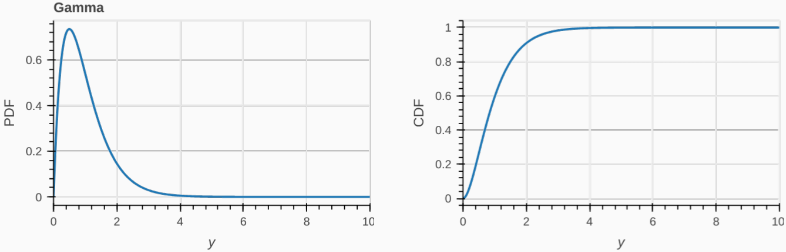


Figure 5: https://distribution-explorer.github.io/continuous/gamma.html

Pareto

$$f(x \mid x_{min}, \alpha) = \frac{\alpha}{x}\left(\frac{x_{min}}{x}\right)^{\alpha}$$

- **Support:** $x \in \mathbb{R}, x \geq x_{min}$.
- Low magnitude earthquakes are much more likely than high magnitude
- Long-tailed distributn
- The inverse of the Pareto distribution is the Power distribution

Pareto

$$E[f(x \mid x_{min}, \alpha)] = \begin{cases} \infty & \alpha \leq 1 \\ \frac{\alpha x_{min}}{\alpha - 1} & \alpha > 1 \end{cases}$$

$$Var[f(x \mid x_{min}, \alpha)] = \begin{cases} \infty & \alpha \leq 2 \\ \frac{\alpha x_{min}^2}{(\alpha - 1)^2 (\alpha - 2)} & \alpha > 2 \end{cases}$$

Pareto

$$x_min = 0.1, \alpha = 2$$

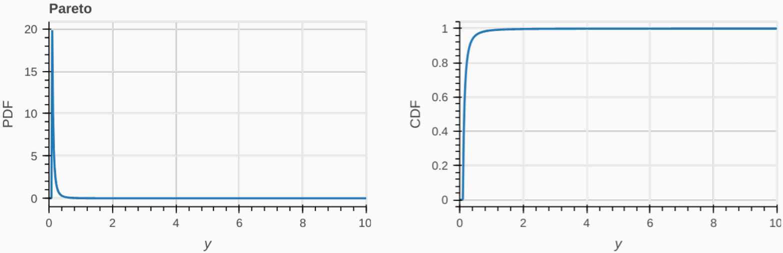


Figure 6: https://distribution-explorer.github.io/continuous/pareto.html

Generating our example data

Student attendance

Let us build an artificial data set of student attendance. We want to simulate the variables mentioned in the first lecture.

Initial Setup

```python
# All courses have 12 weeks,2 days a week
# To simplify,we will assume months with 28 days
# and will disconsider weekends
days = range(1, 29)
# Each term lasts three months and these
# months are always fixed
terms = [[2,3,4], [6,7,8], [10,11,12]]
years = [2021, 2022, 2023]

numStudents = 300
courses = 200
```

Weather information

```python
# estimated from SP averages taken
# from https://www.climatestotravel.com/climate/brazil
# sunny,rainy,cloudy
weather = [(1, 0, 0), (0, 1, 0), (0, 0, 1)]

# Probability of raining is the avg. number
# of rainy days for each month divided by 30 days
rainProb = { (k+1):(v/30) for k, v in
             enumerate([18,16,13,9,9,6,7,7,9,11,13,16])}
# sunny and cloudy probabilities are 80% and 20%
# of the remaining probability of rain
sunProb = {k:(1 - v)*0.8 for k, v in rainProb.items()}
cloudProb = {k:(1 - v)*0.2 for k, v in rainProb.items()}
```

Weather information

```
# avg. and std. of mm of rain,
qtyRain = {(k+1):(v/60) for k, v in
            enumerate([240, 215, 160, 75, 75, 55, 45,
              40, 80, 125, 145, 200])}
rateRain = {(k+1):v for k, v in
            enumerate([7, 7, 6, 4, 4, 3, 3, 3, 4, 5,
              5, 5])}

mmRain = rng.gamma(qtyRain[month], rateRain)
           if isRainy else 0
```

- The quantity of rain per timeframe (α) and rate of rain (β).

User information

```python
def getChildren(age):
    return np.round(rng.power(np.log(age)-3,1)*5)[0]

def getEnroll(age):
    return rng.poisson(age - 23)
```

- The number of children is the inverse of a Pareto distribution, a smaller value is more frequent
- The amount of time the student is enrolled is a poisson process. The distribution is centered at 0 for 23 years older

User information

```python
def hasWork(age):
    p = 1 / (1 + np.exp(-0.2*(age-25)))
    return rng.binomial(1, p)

def single(age):
    p = 1 / (1 + np.exp(-0.15*(age-25)))
    return rng.binomial(1, 1 - p)

ages = rng.poisson(4, numStudents) + 25
```

- `hasWork` and `single` follows a binomial distribution dependent of age (more on logistic function in the next lecture)
- age also follow a Poisson distribution with the central value at 29.

User-Date-Course relationship

```python
def sampleSpeed(dayInfo, studentInfo):
    mu = 40 + 5*dayInfo.isSunny - 5*dayInfo.isRainy
      - np.exp(0.2*dayInfo.mmRain)
      + 5*studentInfo.isSingle
      - 2*studentInfo.numChildren
      - 2*studentInfo.isSingle*studentInfo.numChildren
    return np.max([10, rng.normal(mu, 5)])
```

User-Date-Course relationship

```
def sampleAttend(dayInfo, studentInfo, ETA, hours, exam):
    y = (16.1 + 11.5*exam - 1.3*np.exp(0.4*hours)
           - 1.1*np.exp(0.3*ETA)
           - 3.2*studentInfo.numChildren)
        /(1 + 15.2*dayInfo.mmRain
        + 21.3*studentInfo.numChildren*dayInfo.isRainy)
    return rng.binomial(1, 1/(1 + np.exp(-y)))
```

User-Date-Course relationship

```python
def sampleGrade(row):
    y = 1.2*np.exp(-0.2*row.hoursWork_mean)
          - 3.2*np.log(1.5 + row.ETA_mean)
          + 3.2*np.log(1.2+row.numAttendence_max)
          - 1.1 * row.numChildren * (1 - row.isSingle)
      / (1.5 + 0.5*np.exp(0.9*row.age - 30))
    return np.round(np.clip(rng.normal(y, 0.5),0,10), 2)
```

Data generation

The whole script is available in Jupyter Notebook format at https://folivetti.github.io/courses/RegSimbolica/2024/EDA.ipynb.

Symbolic Models

- The symbolic models defined in the previous slides describe the parameters of each distribution.
- Our goal is to retrieve these expressions from a sampled data.
- The regression goal is to estimate a distribution parameter that fits the sampled data!!

Terminology learned today

- **random variable:** a function that maps one sample to a numerical quantity
- **population:** the entire data
- **sample:** a subset of the data
- **statistical quantity:** an aggregation function that summarizes the data
- **probability distribution:** maps events to probabilities

Terminology learned today (and more)

- **probability mass function (pmf):** probability of a discret event
- **probability density function (pdf):** probability of a event being between a range
- **cumulative distribution function (cdf):** probability that an event is less than or equal a value
- **quantile function:** the inverse of cdf, given a probability p what event will return $cdf(x) = p$?
- **expected value:** the center of the mass.
- **moment:** quantitative measure of the shape of the function.

Further reading

(click to follow the links):

- https://distribution-explorer.github.io/background/review_of_key_concepts.html
- https://en.wikipedia.org/wiki/Random_variable#Examples
- https://en.wikipedia.org/wiki/Sampling_(statistics)
- https://en.wikipedia.org/wiki/Statistic
- http://www.leahhoward.com/m163text/6-4.pdf
- https://web.stanford.edu/~peastman/statmech/interpretation.html
- Blitzstein, Joseph K., and Jessica Hwang. Introduction to probability. Crc Press, 2019.
- Bishop, C. “Pattern recognition and machine learning.” Springer google schola 2 (2006): 35-42.

Next lecture

- Introduction to Regression Analysis
- Confidence Intervals

Acknowledgments

- Thiago Ferreira Covões