Statistics in Python

David Arroyo Menéndez

March 1, 2019

Statistics: Summary

- Descriptives
- Manipulating Data
- Matplotlib
- Distributions
- Statistics Tests
- Logistic Regression
- Principal Component Analysis

Descriptives

Source!

```
$ python3 scipy-descriptives.py
```

Manipulating Data

Pandas is for dataframes

```
$ python3 pandas/pandas-10min.py
$ python3 pandas/creating-dataframe.py
$ python3 pandas/creating-dataframa-from-arrays.py
$ python3 pandas/manipulating-data.py
$ python3 pandas/remove-rows-with-nan.py
$ python3 pandas/handle-missing-data.py
$ python3 pandas/data-analysis/pd-diabetes.py
```

Numpy is algebra is for arrays

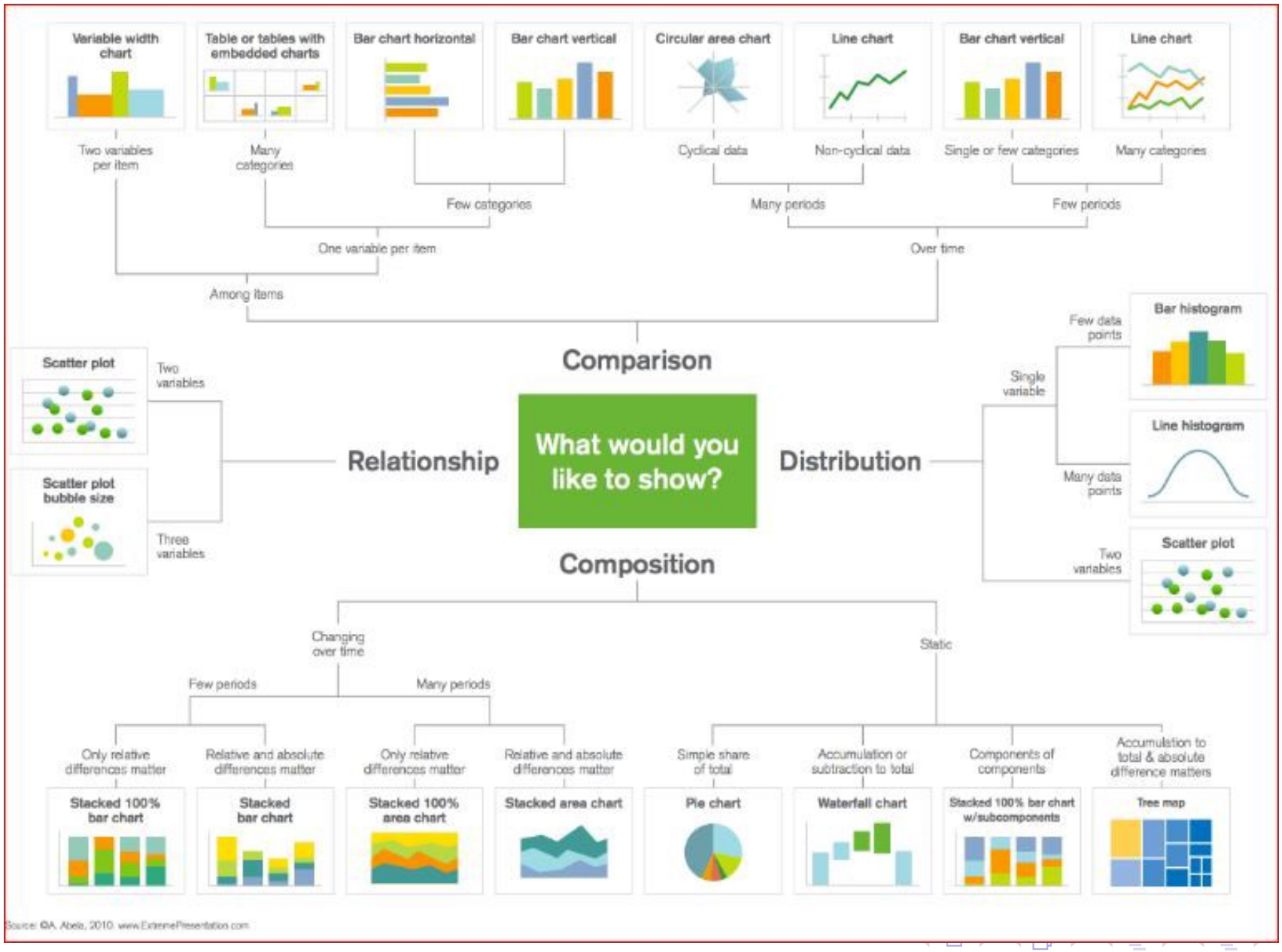
```
$ python3 numpy/reject-outliers.py
```

Matplotlib

You can display statistics with matplotlib

```
$ python3 barchart_demo.py
$ python3 boxplot-example2.py
$ python3 boxplot-example.py
$ python3 colorbar_basics.py
$ python3 image_demo.py
$ python3 pie_features.py
$ python3 plot_3D.py
$ gimp surface3d_frontpage.png &
$ python3 pyplot_text.py
$ python3 scatter-example.py
$ python3 stackplot_demo.py
$ python3 subplot.py
$ python3 unicode_minus.py
```

Matplotlib. What would you like to show?

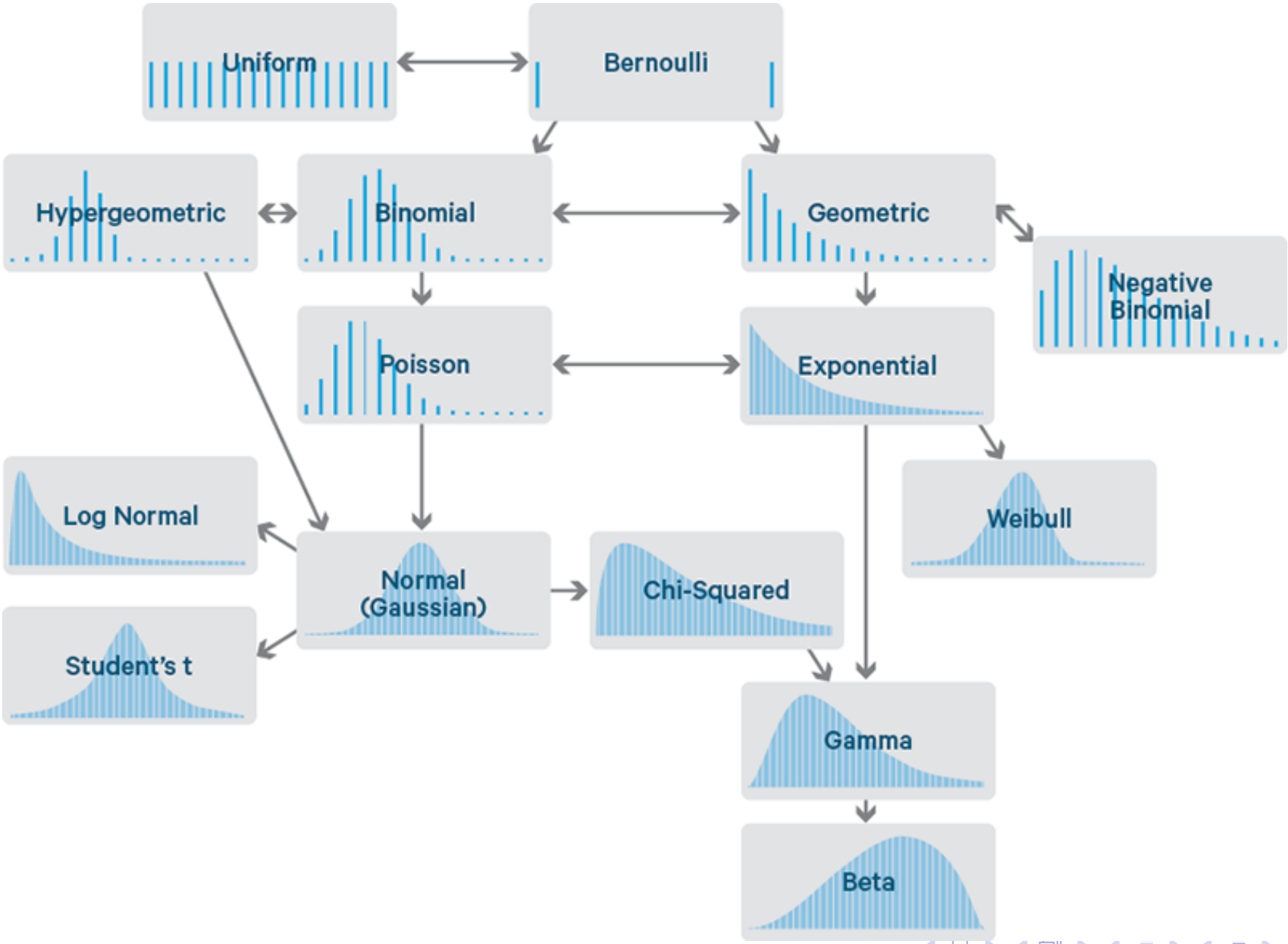


Distibutions in Statistics

An histogram trends to be a continuos line in a table, we can draw a distribution with this trend.

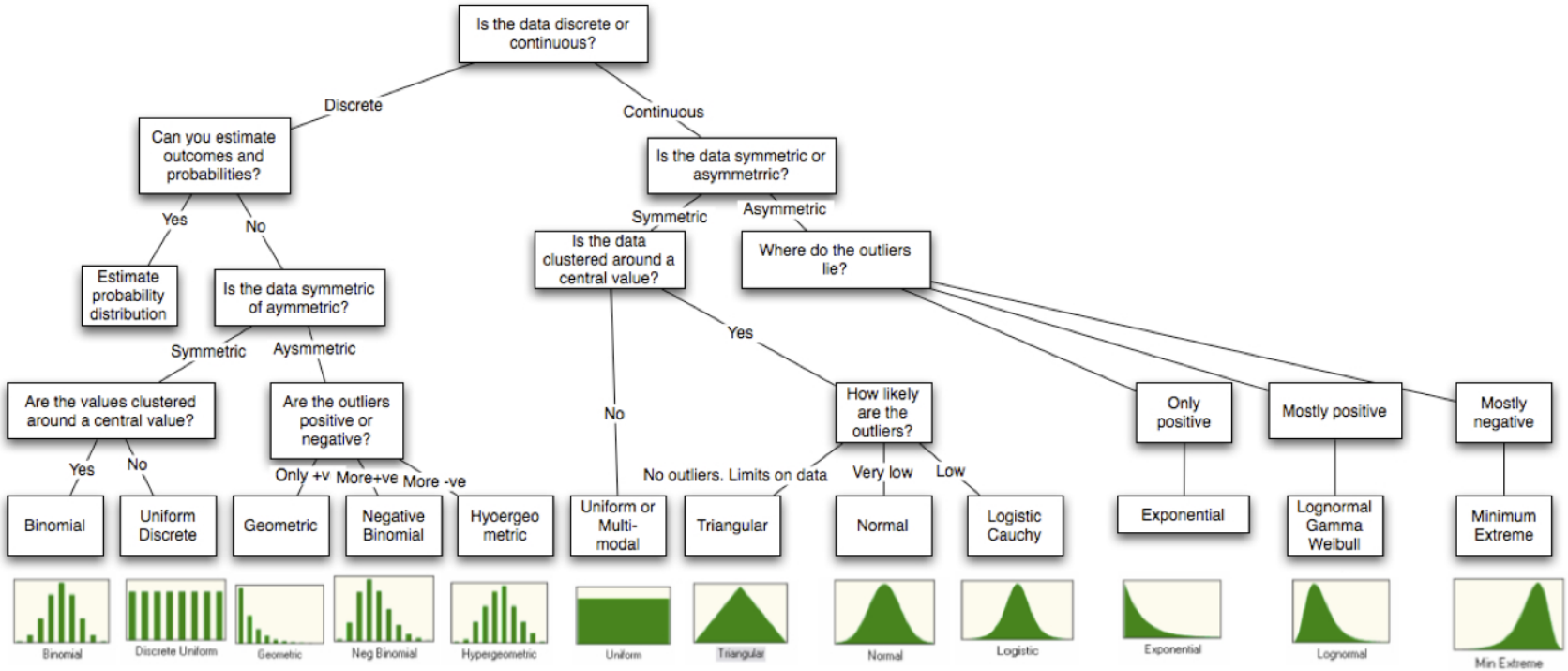
```
$ python3 bernoulli.py
$ python3 plot_normal.py
$ python3 poisson.py
$ python3 binomial.py
$ python3 exponential-distribution.py
```

Distributions in Statistics (II)

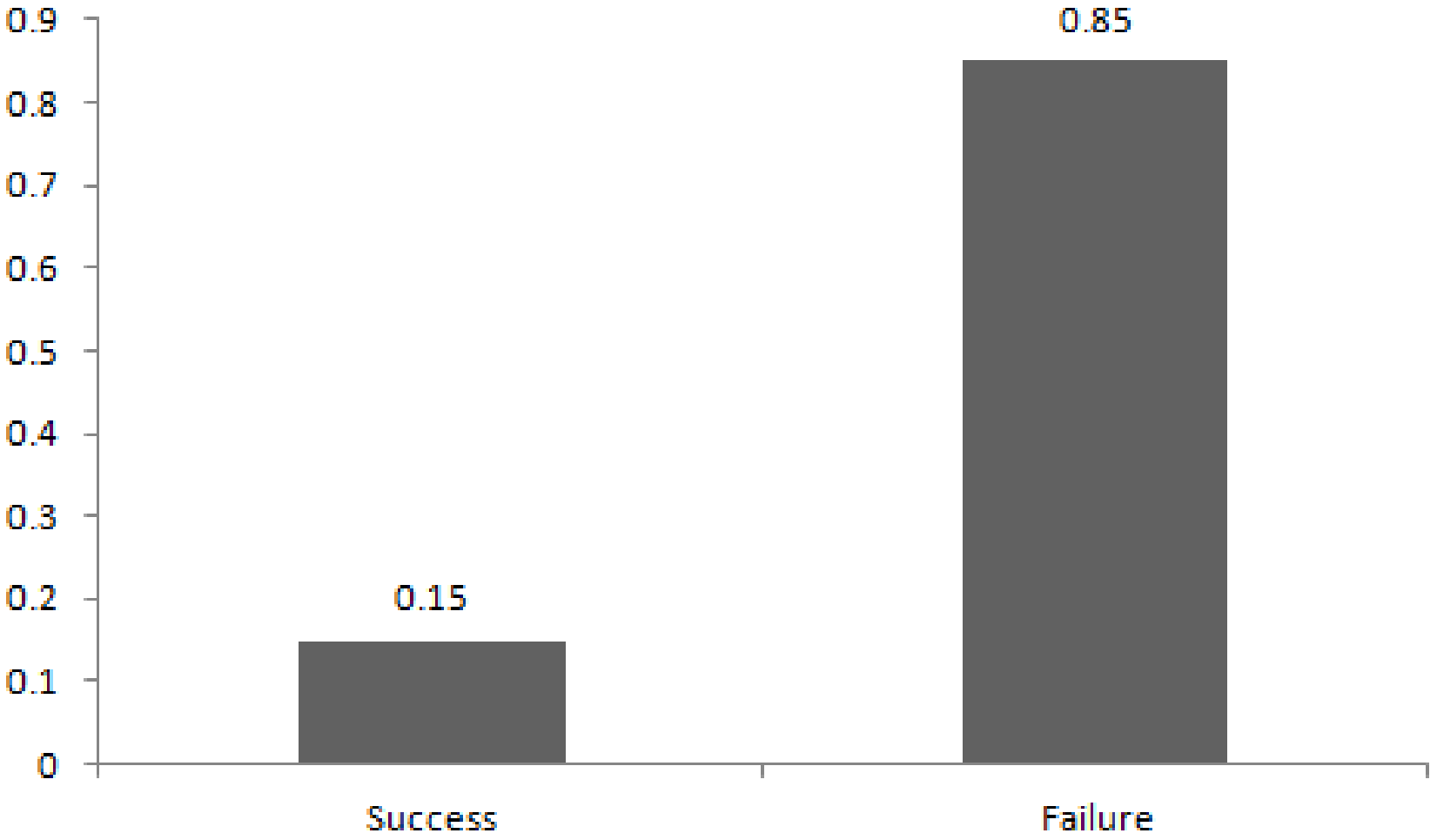


Distributions in Statistics (III)

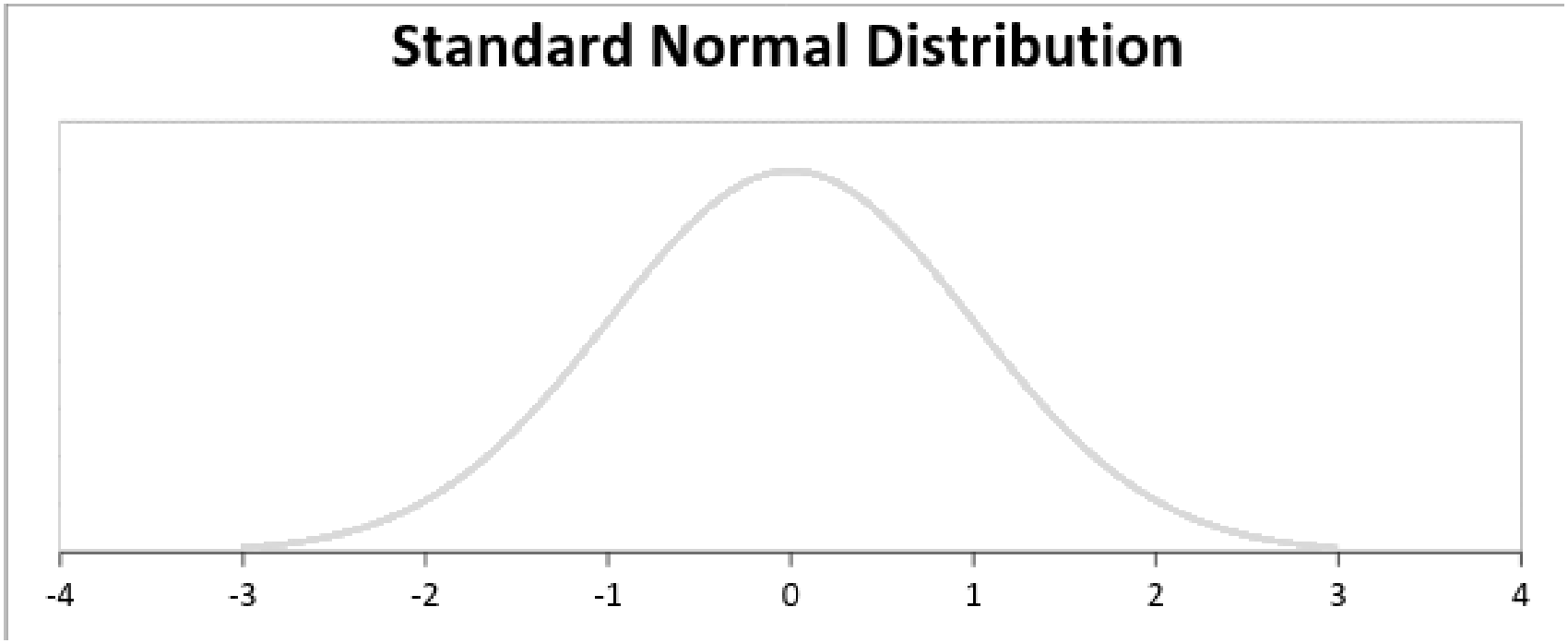
Figure 6A.15: Distributional Choices



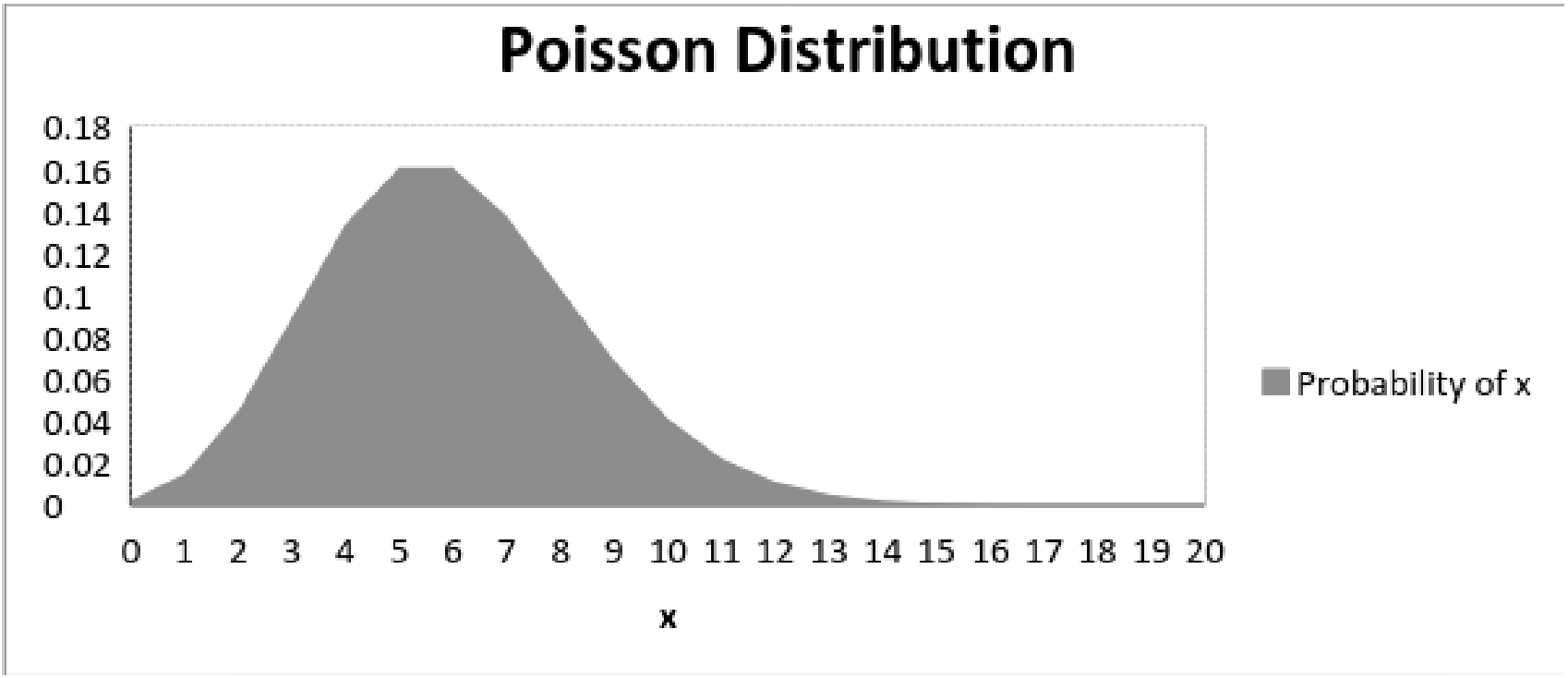
Bernoulli Distribution



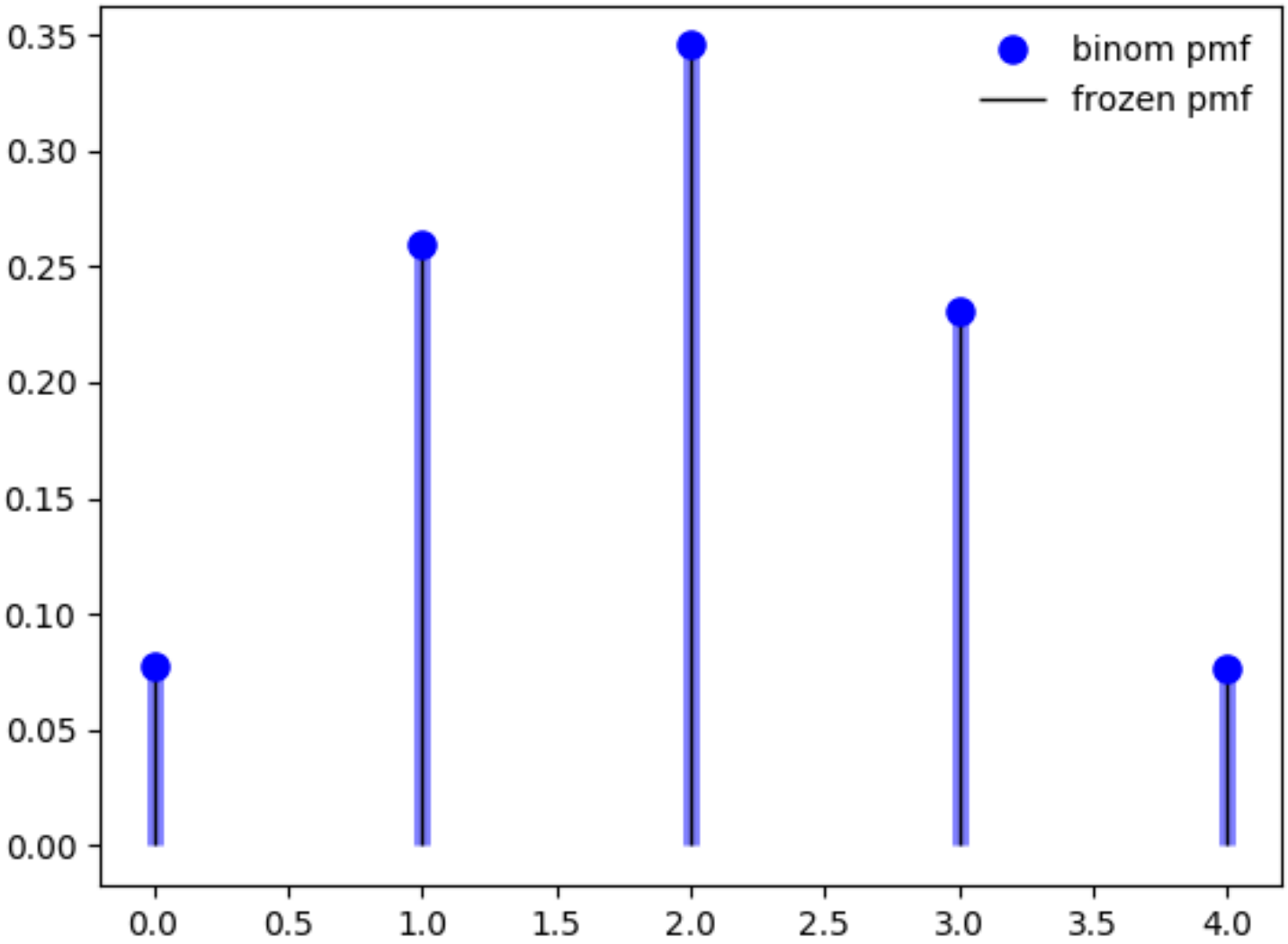
Normal Distribution



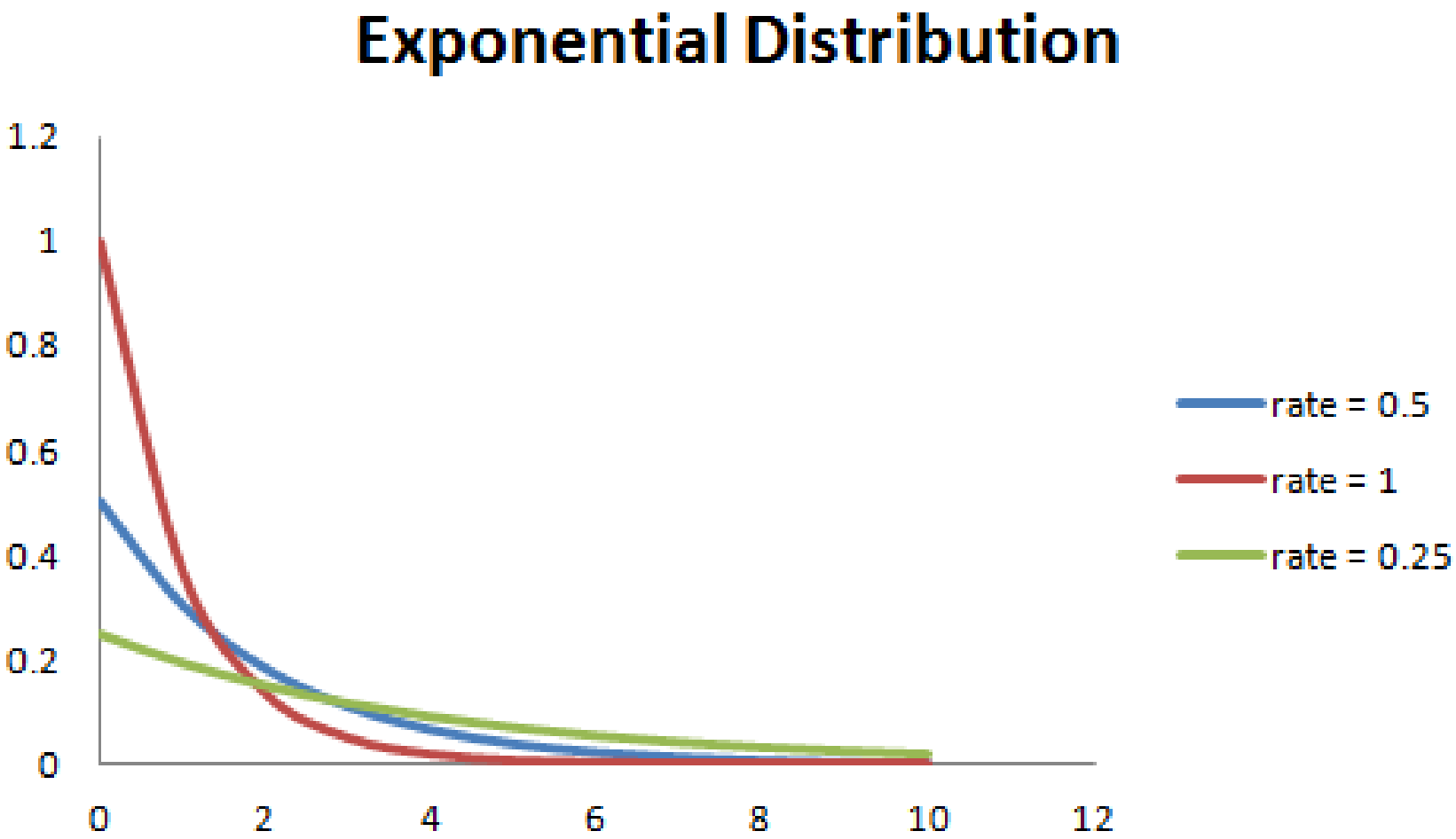
Poisson Distribution



Binomial Distribution



Exponential Distribution



Moments in a Distribution

Moment number	Name	Measure of	Formula
1	Mean	Central tendency	$\bar{X} = \frac{\sum_{i=1}^{N} X_i}{N}$
2	Variance (Volatility)	Dispersion	$\sigma^2 = \frac{\sum_{i=1}^{N} (X_i - \bar{X})^2}{N}$
3	Skewness	Symmetry (Positive or Negative)	$Skew = \frac{1}{N} \sum_{i=1}^{N} \left[\frac{(X_i - \bar{X})}{\sigma} \right]^3$
4	Kurtosis	Shape (Tall or flat)	$Kurt = \frac{1}{N} \sum_{i=1}^{N} \left[\frac{(X_i - \bar{X})}{\sigma} \right]^4$

Where X is a random variable having N observations (i = 1,2,…,N).

Statistics Tests / Choice Models

To see a result from a hipothesis you can use tests:

```
$ python scipy-special-tests.py
$ python discrete-choice-models.py
$ python pearson.py # for testing non-correlation
$ python fisher.py
```

Multivariate Analisis. Choosing a model (I)

Variable dependiente	Variable(s) explicativas	Ejemplo	Modelos paramétricos	Condiciones de validez	Otras soluciones
Una variable cuantitativa	Una variable cualitativa (= factor) con dos niveles	Efecto de la contaminación (sí / no) en la concentración de un elemento de traza en una planta	ANOVA unifactorial con dos niveles	1 ; 2 ; 3 ; 4	Prueba de Mann-Whitney
	Una variable cualitativa con k niveles	Efecto del sitio (4 fábricas) en la concentración de un elemento de traza en una planta	ANOVA unifactorial	1 ; 2 ; 3 ; 4	Prueba de Kruskal-Wallis
	Varias variables cualitativas con varios niveles	Efectos combinatorios del sitio (4 fábricas) y las especies de plantas sobre la concentración de un compuesto en el tejido de la planta	ANOVA multifactorial (diseños factoriales)	1 ; 2 ; 3 ; 4	
	Una variable cuantitativa	Efecto de la temperatura sobre la concentración de una proteína	Regresión lineal simple; modelos no lineales (depende de la forma de la relación entre las variables dependiente y explicativa)	1 - 3	Regresión no paramétrica(*); Regresión cuantil; Árboles de clasificación y de regresión(*); K Vecinos Más Próximos(*)
	Varias variables cuantitativas	Efecto de la concentración de diversos contaminantes sobre la biomasa de las plantas	Regresión lineal múltiple; modelos no lineales	1 - 6	Regresión PLS (*); K Vecinos Más Próximos(*)
	Mezcla de variables cualitativas y cuantitativas	Efectos combinatorios del sexo y la edad en la glucemia asociada a un tipo de diabetes	ANCOVA	1 - 6	Regresión PLS(*); Regresión cuantil; Árboles de clasificación y de regresión(*); K Vecinos Más Próximos(*)

Multivariate Analisis. Choosing a model (II)

Varias variables cuantitativas	Variable(s) cualitativa(s) y/o cuantitativa(s)	Efecto de una matriz de variables ambientales sobre el transcriptoma	MANOVA	1 ; 4 ; 7 ; 8	Análisis de Redundancia; Regresión PLS(*)
Una variable cualitativa	Variable(s) cualitativa(s) y/o cuantitativa(s)	Efecto de la dosis en la supervivencia / muerte de ratones individuales	Regresión logística (binomial u ordinal o multinomial)	5 ; 6	PLS-DA(*); Análisis Discriminante(*); Árboles de clasificación y de regresión(*); K Vecinos Más Próximos(*)
Una variable de frecuencia (con muchos ceros)	Variable(s) cualitativa(s) y/o cuantitativa(s)	Efectos de la dosis en el número de necrosis en ratones	Regresión log-lineal (Poisson)	5 ; 6	

Multivariate Analisis. Choosing a model (II)

Logistic Regression

Scikit is your friend

```
$ python3 scikit/logistic-regression/logistic-function.py
$ python3 scikit/logistic-regression/data-using-pandas.py
```

Principal Component Analysis

Scikit is your friend

```
$ python3 scikit/pca-choosing-components.py --csv="scikit/feat
$ python3 scikit/pca-features.py
```

Montecarlo Statistics Methods

It's a statistic game where the players is betting.

```
$ python3 statistics/montecarlo/bettor.py
$ python3 statistics/montecarlo/doublebettor.py
$ python3 statistics/montecarlo/bettor-statistics.py
$ python3 statistics/montecarlo/dalambert.py
```