

DS100: Data Science Fundamental Week One



By Harris Chow, Listree Inc.

DATA & ANALYST

- ❖ Data: All the possible descriptions, measurements of any events happened in our universe. It could be from an isolated event, it could be from a sequence of events.
- ❖ Analyst: A person lead us to the answers of the questions on the data, via logical analysis

Data Analyst Skill Lists

- ❖ Mathematics
 - ❖ Summation, Derivative, Integration, etc
- ❖ Descriptive/Inferential Statistics
 - ❖ Describe the Distribution, Conclude the Distribution
- ❖ Statistical Modeling
 - ❖ Regressions, Ensemble of Regressions

Data Analyst Skill Lists

- ❖ Programming Skills
 - ❖ Python, Scala, SQL like, R, Java, SAS, Matlab, etc
- ❖ Communication Skills
 - ❖ A Good Story Presenter*
- ❖ Personality
 - ❖ Curiosity, Self Motivation

AROUND DATA

- ❖ Population: The whole collection of objects you interested in. The size of a defined population probably be infinite.
- ❖ Sample: A subset of a defined population.
- ❖ Sample Space: All possible outcomes of the measure/ experiment randomly conducted.
- ❖ Sample Data: An observation of a sample, all possible values of the observations are subsets of the sample space.

Type of Data

- ❖ Categorical
 - ❖ Nominal: natural categories
 - ❖ Ordinal: ordering categories
- ❖ Numerical
 - ❖ Deseret: counts
 - ❖ Continuous: measures

Example of Data Type

- ❖ Nominal: gender, phone branch, type of programming language, etc.
- ❖ Ordinal: employ evaluation level, service satisfaction level, etc.
- ❖ Deseret: count of passenger in PGV/day, ebola case per day in west african countries, etc.
- ❖ Continuous: Weight, Length, etc

Probability Distribution

- ❖ Each time we measure/observe a variable, the outcome is not determined
- ❖ Categorize it as a random variable X (categorical/numerical)
- ❖ The outcomes follow certain pattern, which is described by probability distribution function
- ❖ A set s (a subset of the sample space) or A numerical range r (an interval of the sample space)

$$P(X \in s), \quad P(X \in r), \quad \textit{respectively}$$





































Random Sample*

- ❖ Random Sampling offers the collected data could represent the whole population at large scale. It is essential when collecting observational data points.

Probability Density/Cumulative Distribution Function

- ❖ P.D.F v.s. C.D.F.
- ❖ Categorical:
 - ❖ Density: the likelihood of a unique category (this category is mutually exclusive to other possible categories)
 - ❖ Cumulative: the probability of the category in a collection of categories
 - ❖ The total probability of one observation in all category is 1.

Cont.,

| | | | | | |
|---|--|---|---|---|---|
|  |  |  |  |  |  |
|  |  |  |  |  |  |
|  |  |  |  |  |  |
|  |  |  |  |  |  |
|  |  |  |  |  |  |
|  |  |  |  |  |  |

Cont.,

- ❖ Numerical (discrete)
 - ❖ Density: the probability of the variable equals a unique possible value
 - ❖ Cumulative: the probability of the variable in a collection of unique possible values (since it is numerical now, we may follow a order for the collection of values, i.e. $X \leq$ 'a value')
 - ❖ The total probability of one observation in all possible values is 1.

Cont.,

- ❖ Poisson Distribution
- ❖ lamda: expected event occurrence frequency in a given time interval
- ❖ k: occurrence of event in given a time interval

$$f(x = k, \lambda) = P(x = k) = \frac{\lambda^k e^{-\lambda}}{k!}$$

$$F(x \leq k, \lambda) = P(x \leq k) = \sum_{i=0}^k e^{-\lambda} \frac{\lambda^i}{i!}$$

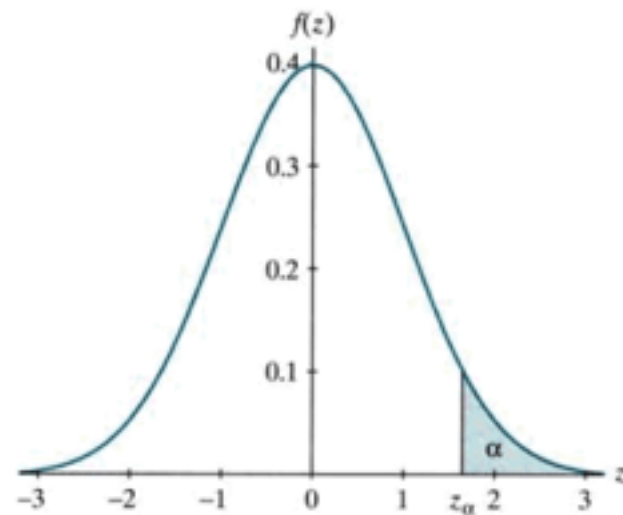
Cont.,

- ❖ Numerical (continuous)
 - ❖ Density: the trend of the probability of the variable falls on a specific num value
 - ❖ Cumulative: the probability of the variable falls in a certain num interval
 - ❖ The probability of the variable falls in its sample space is 1.
 - ❖ The area under the density curve is 1.

Cont.,

❖ Normal Distribution

TABLE Vb: The Normal Distribution



$$P(Z > z_\alpha) = \alpha$$

$$P(Z > z) = 1 - \Phi(z) = \Phi(-z)$$

| z_α | 0.00 | 0.01 | 0.02 | 0.03 | 0.04 | 0.05 | 0.06 | 0.07 | 0.08 | 0.09 |
|------------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|
| 0.0 | 0.5000 | 0.4960 | 0.4920 | 0.4880 | 0.4840 | 0.4801 | 0.4761 | 0.4721 | 0.4681 | 0.4641 |
| 0.1 | 0.4602 | 0.4562 | 0.4522 | 0.4483 | 0.4443 | 0.4404 | 0.4364 | 0.4325 | 0.4286 | 0.4247 |
| 0.2 | 0.4207 | 0.4168 | 0.4129 | 0.4090 | 0.4052 | 0.4013 | 0.3974 | 0.3936 | 0.3897 | 0.3859 |
| 0.3 | 0.3821 | 0.3783 | 0.3745 | 0.3707 | 0.3669 | 0.3632 | 0.3594 | 0.3557 | 0.3520 | 0.3483 |
| 0.4 | 0.3446 | 0.3409 | 0.3372 | 0.3336 | 0.3300 | 0.3264 | 0.3228 | 0.3192 | 0.3156 | 0.3121 |
| 0.5 | 0.3085 | 0.3050 | 0.3015 | 0.2981 | 0.2946 | 0.2912 | 0.2877 | 0.2843 | 0.2810 | 0.2776 |
| 0.6 | 0.2743 | 0.2709 | 0.2676 | 0.2643 | 0.2611 | 0.2578 | 0.2546 | 0.2514 | 0.2483 | 0.2451 |
| 0.7 | 0.2420 | 0.2389 | 0.2358 | 0.2327 | 0.2296 | 0.2266 | 0.2236 | 0.2206 | 0.2177 | 0.2148 |
| 0.8 | 0.2119 | 0.2090 | 0.2061 | 0.2033 | 0.2005 | 0.1977 | 0.1949 | 0.1922 | 0.1894 | 0.1867 |
| 0.9 | 0.1841 | 0.1814 | 0.1788 | 0.1762 | 0.1736 | 0.1711 | 0.1685 | 0.1660 | 0.1635 | 0.1611 |

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

$$F(x, \mu, \sigma) = \int_{-\infty}^x f(x, \mu, \sigma) dx$$

Descriptive Statistics

- ❖ Central Tendency: It is an estimate of the centers of the distribution
- ❖ Dispersion: it is an estimate of the spreading of the value around its center.

Cont.,

- ❖ Median
- ❖ P_k , kth-Percentile: the numerical value from the distribution that $P\%$ of all possible values from the distribution is smaller than P_k (k is an integer)
- ❖ 5 frequently used Percentiles:
 - ❖ 0%(min), 25%(Q1), 50%(Q2, Median), 75%(Q3), 100%(max)
- ❖ Range: $P_{100} - P_0$; Interquartile Range: $Q3 - Q1$

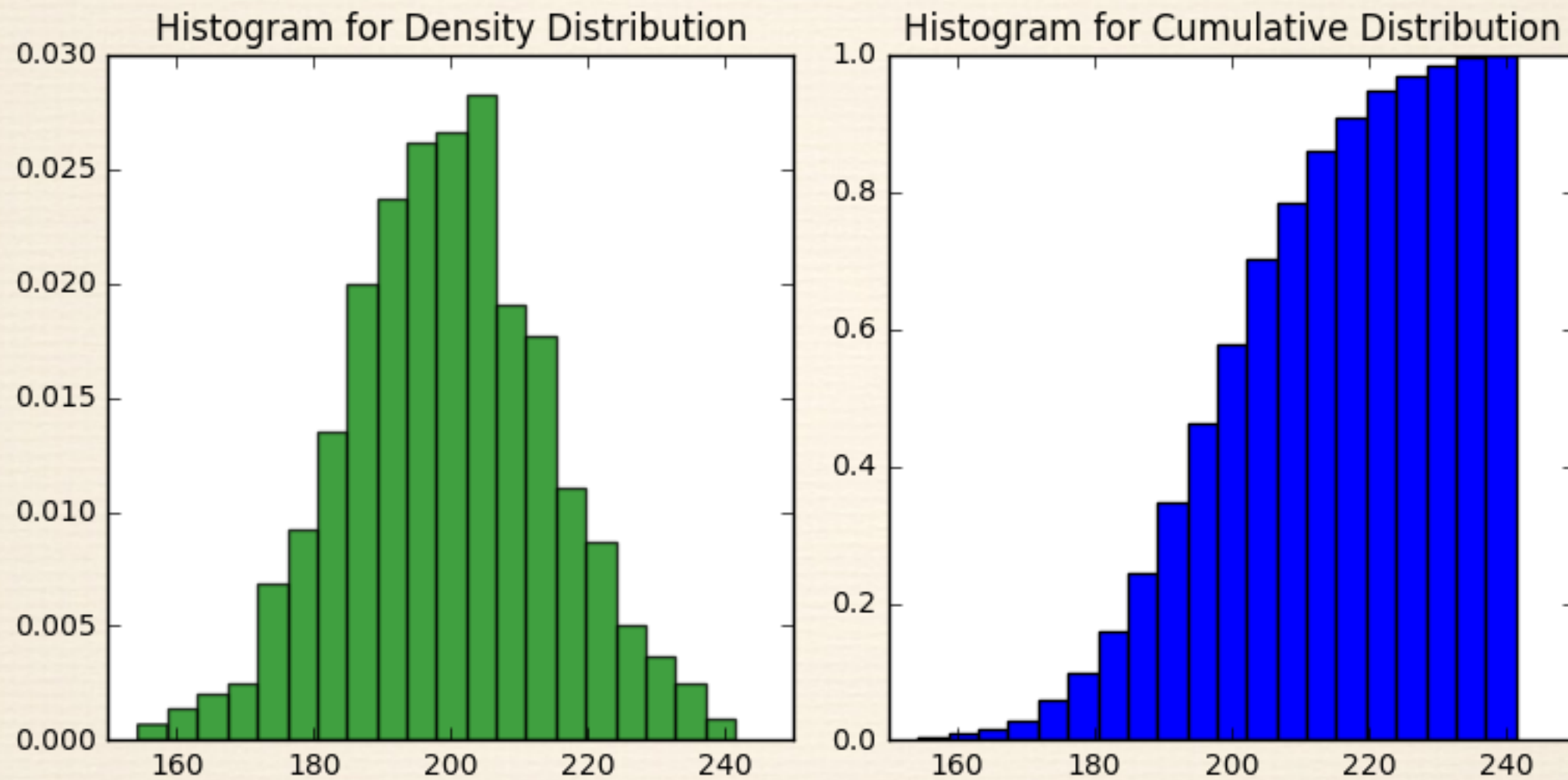
Cont.,

- ❖ Mode: the value with highest frequency
- ❖ Mean: numerical average of all possible value
- ❖ Variance: numerical average of square deviation to the mean
- ❖ Standard Deviation: square root of Variance
- ❖ Mean Absolute Deviation: numerical average of absolute deviation to the mean

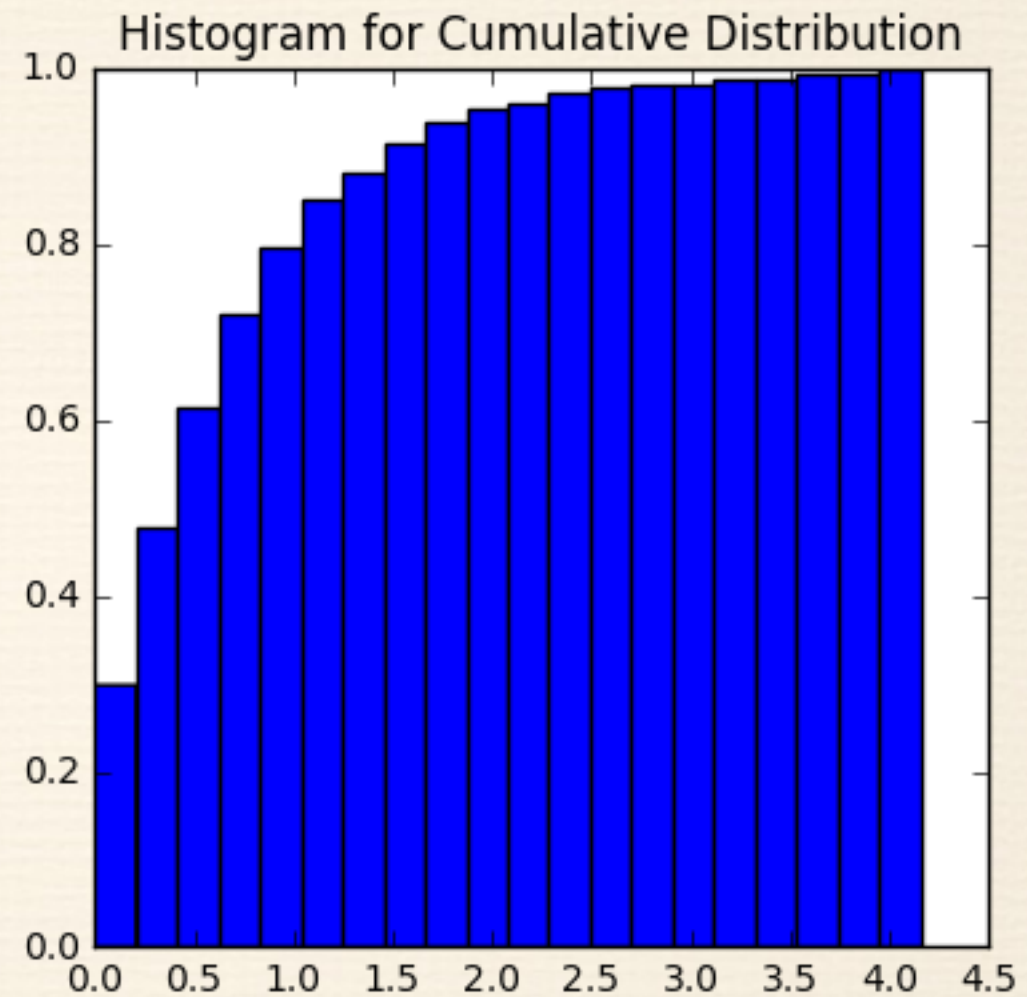
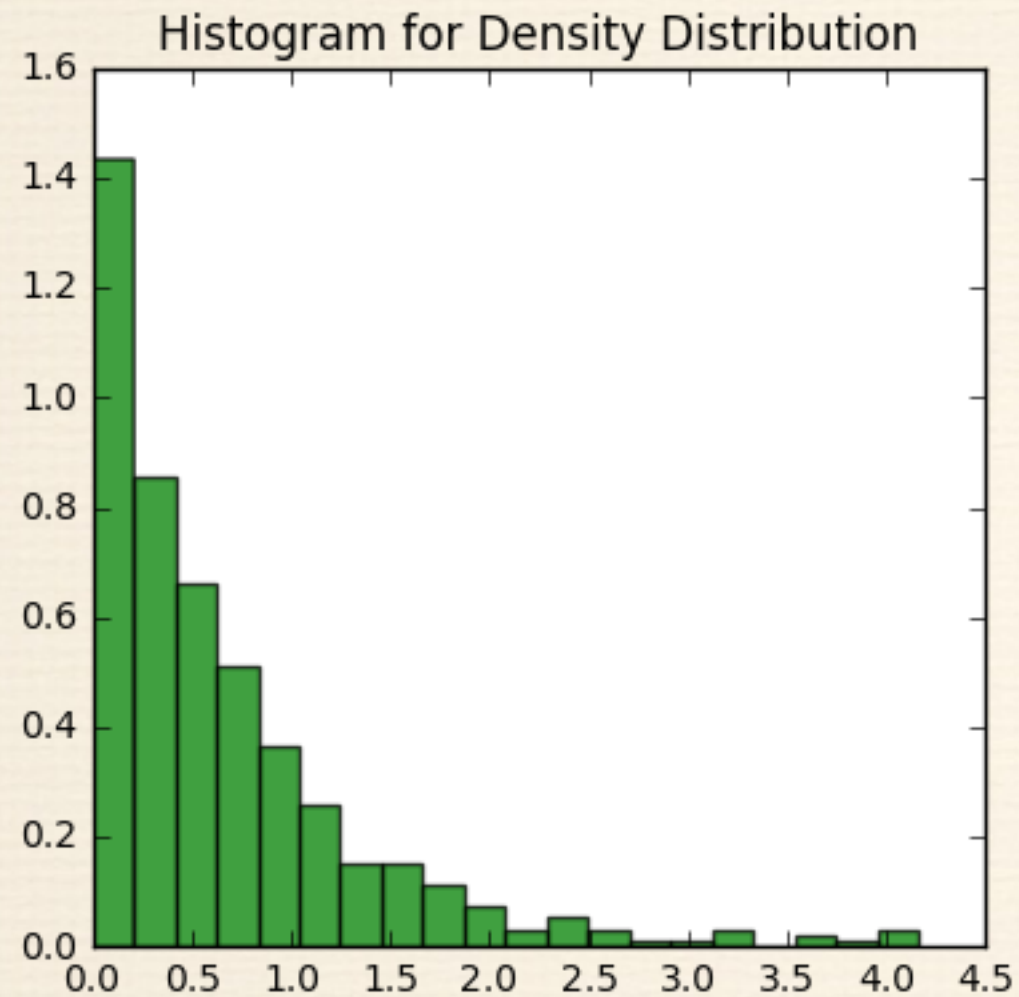
Shape of P.D.F.

- ❖ For Continuous Random Variable
 - ❖ Single Mode Symmetric
 - ❖ Single Mode Non Symmetric
 - ❖ Longer tail on right side
 - ❖ Longer tail on left side
 - ❖ Bi Mode: mixed distribution
 - ❖ Multi Mode: complex mixed distribution
 - ❖ No Mode: white noise

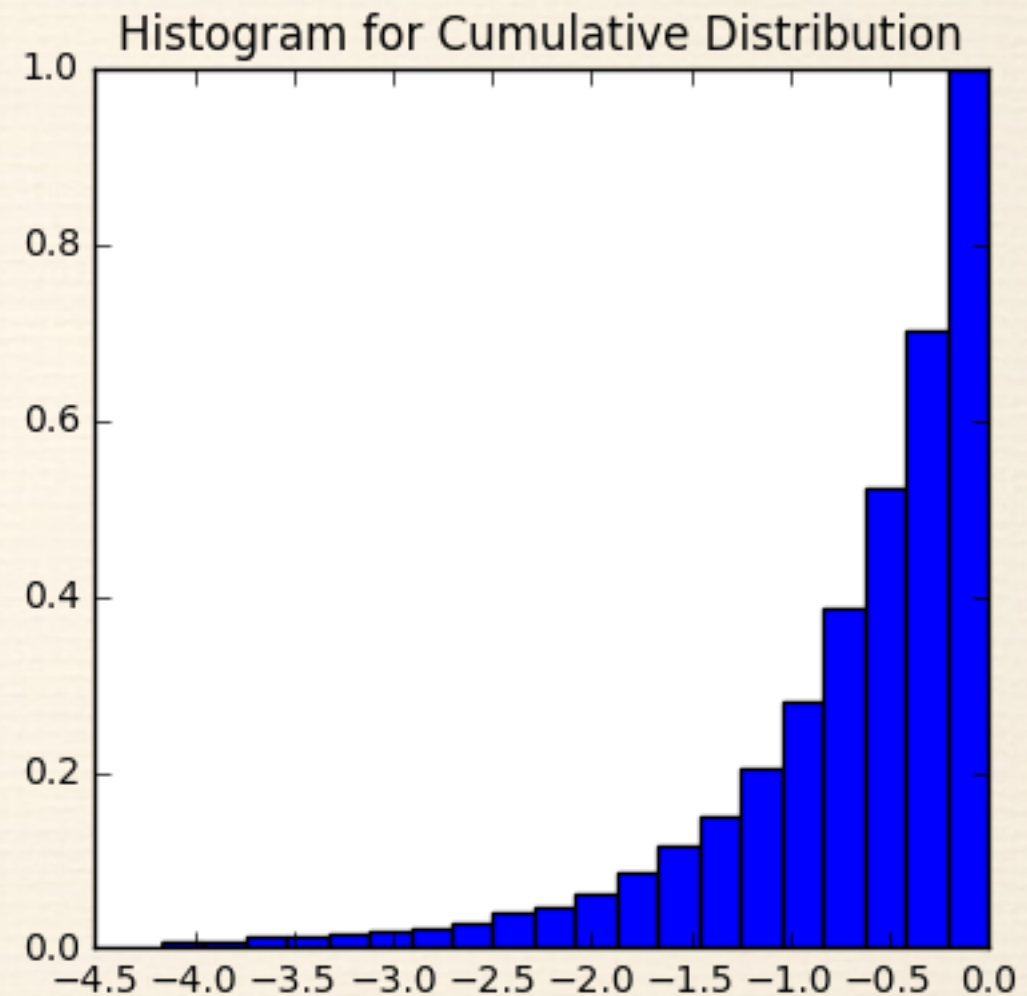
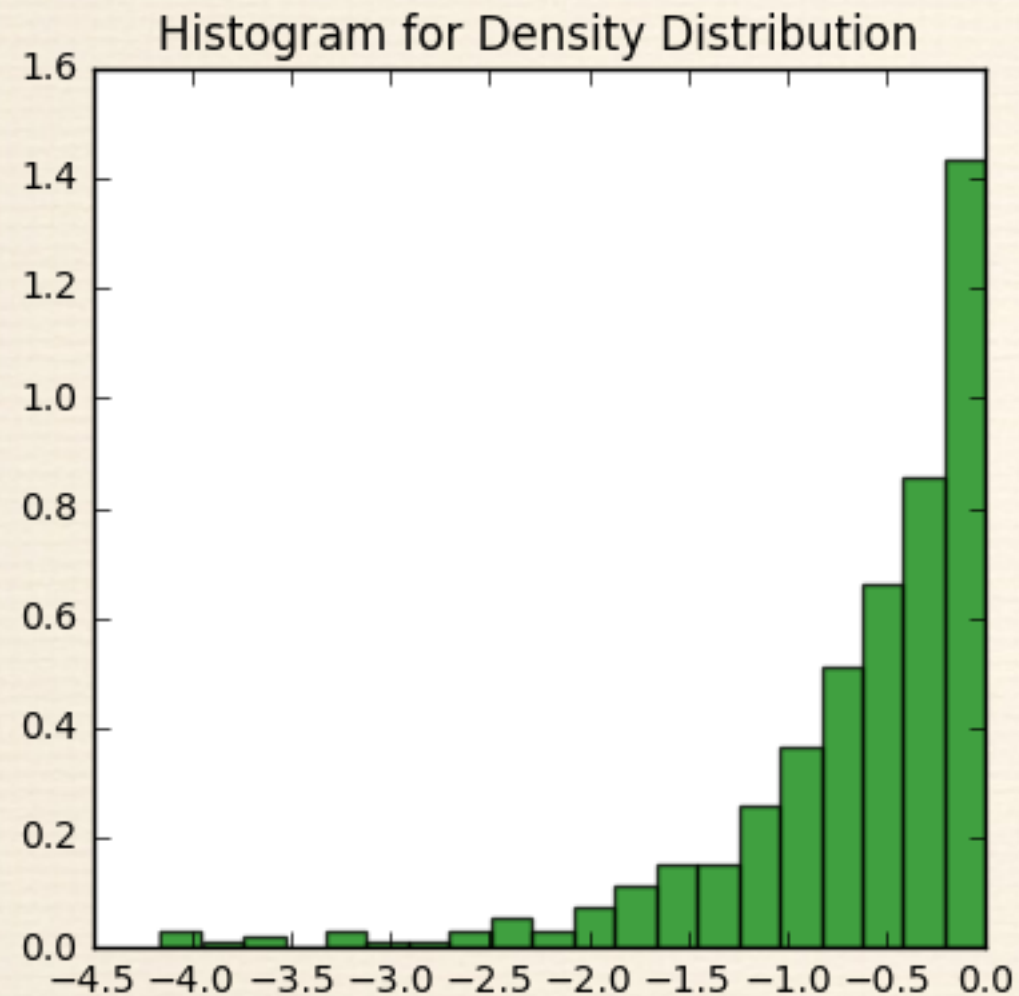
Single Mode Symmetric



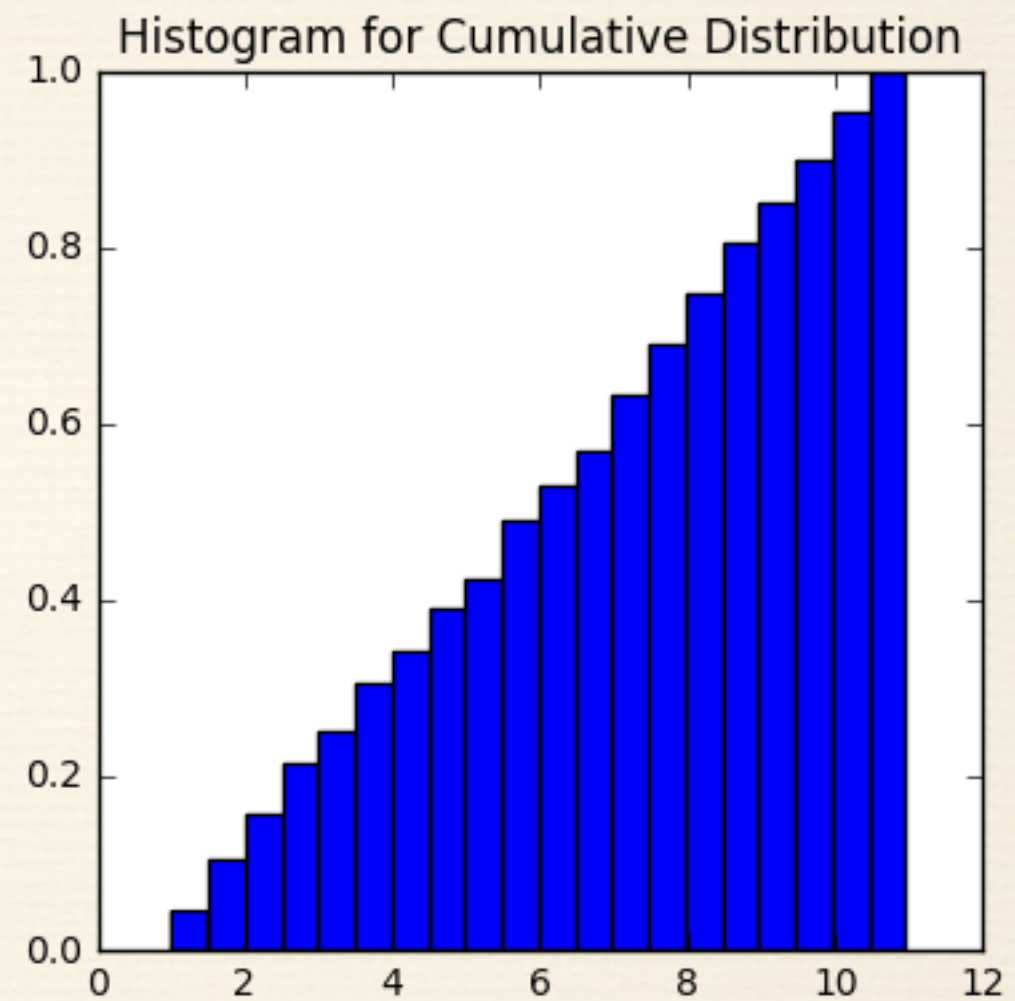
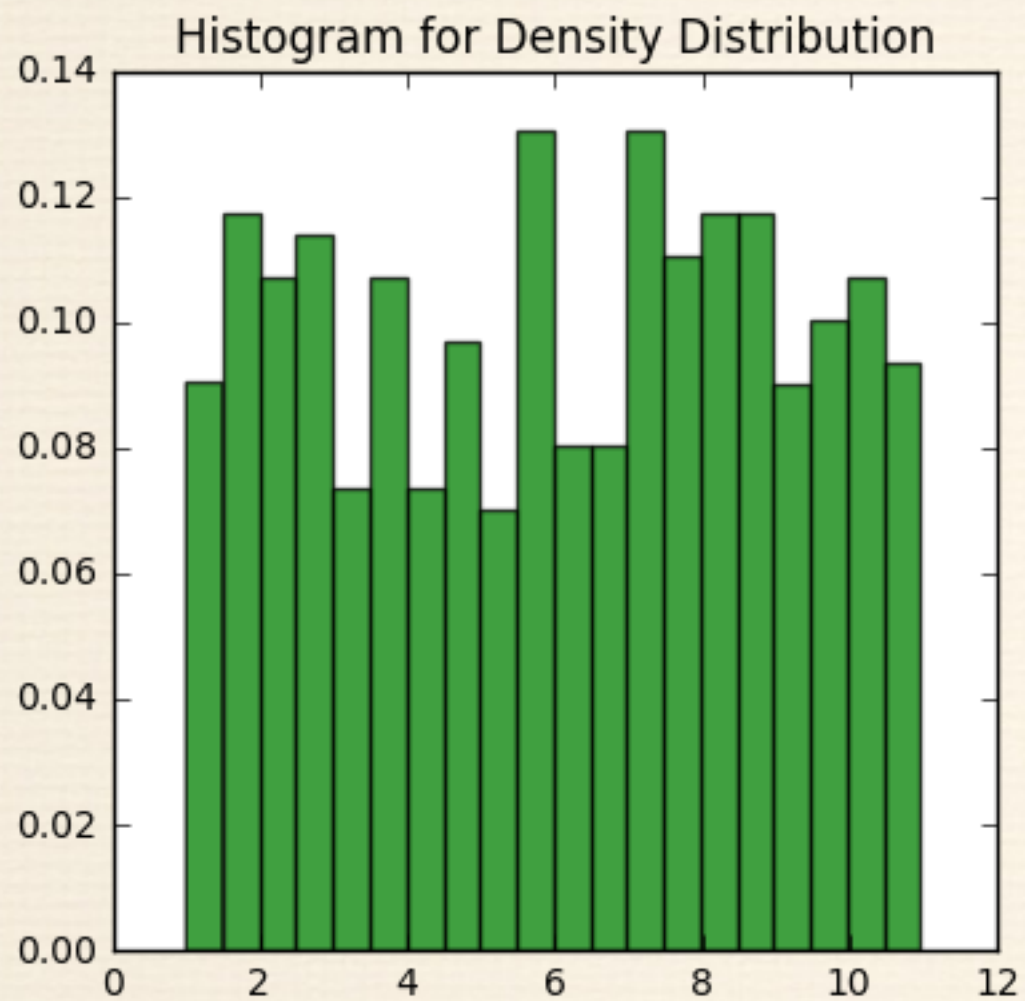
Single Mode Skew to Right



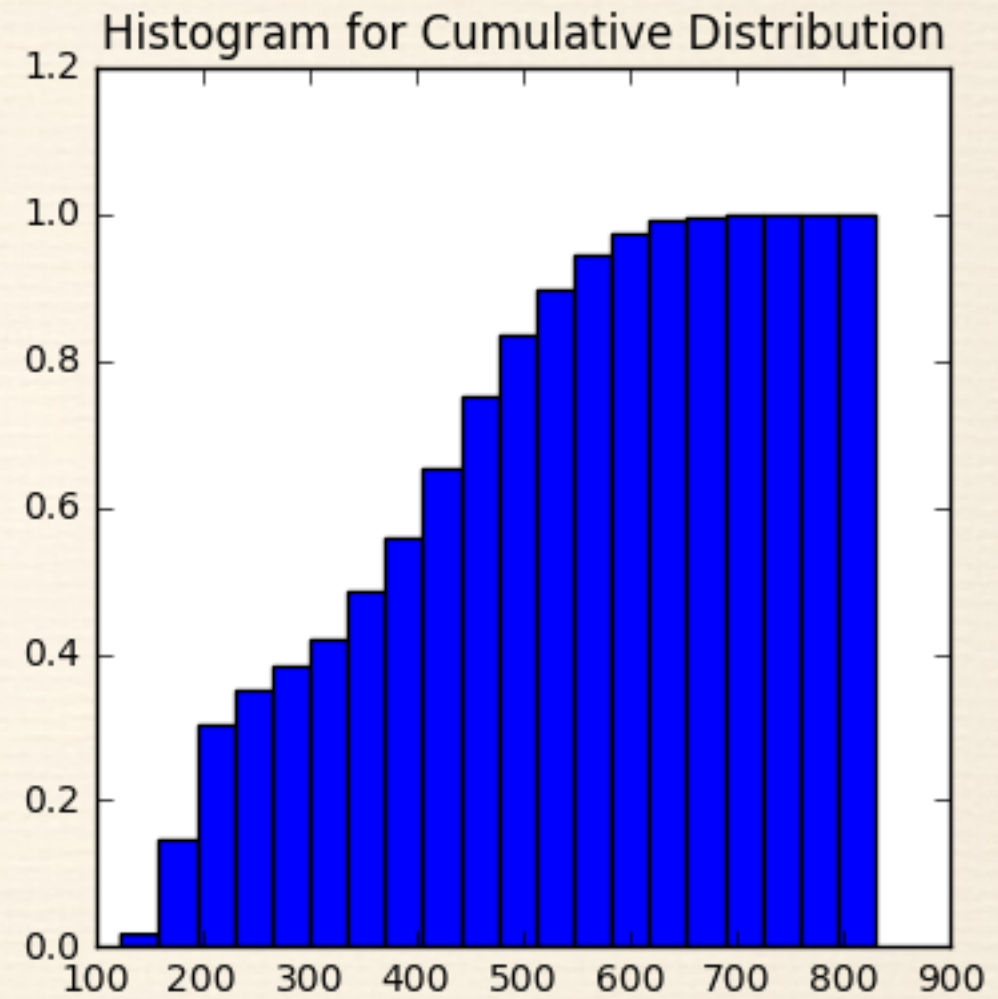
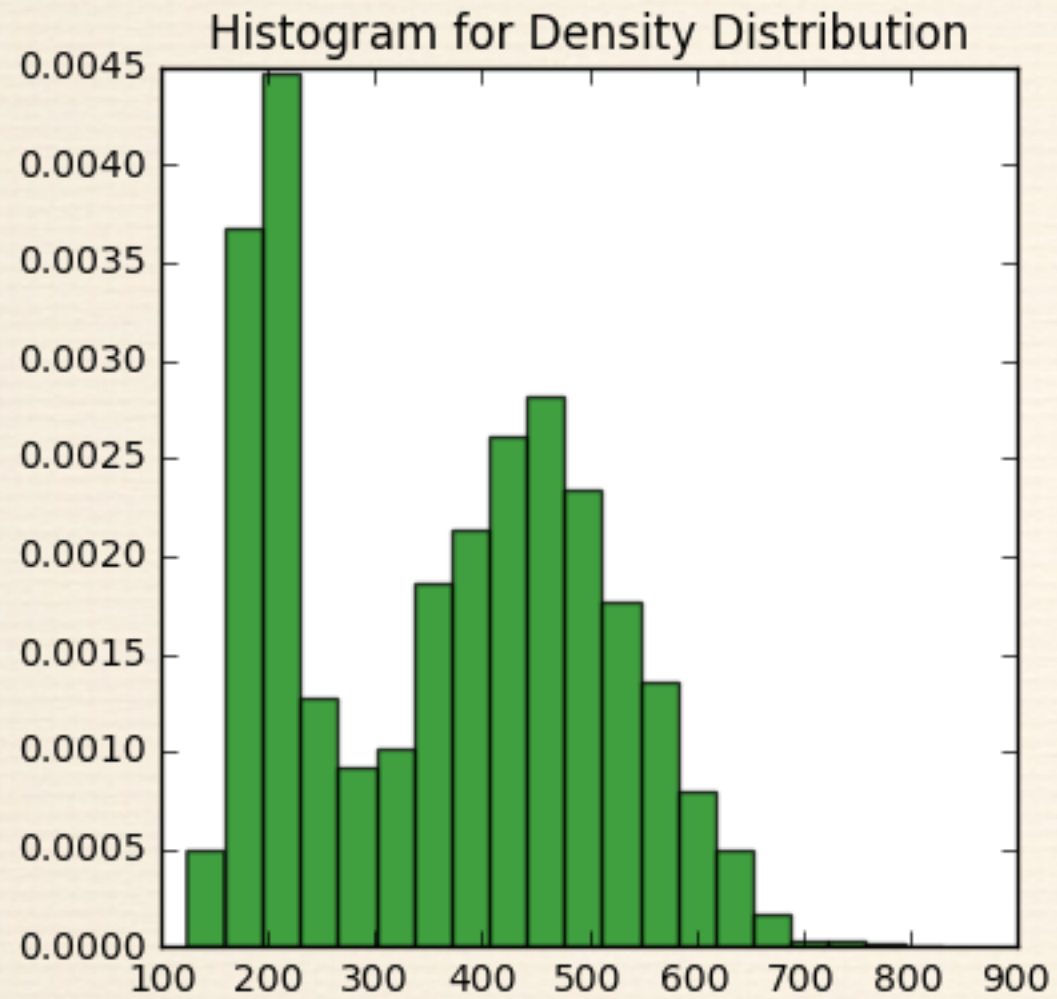
Single Mode Skew to Left



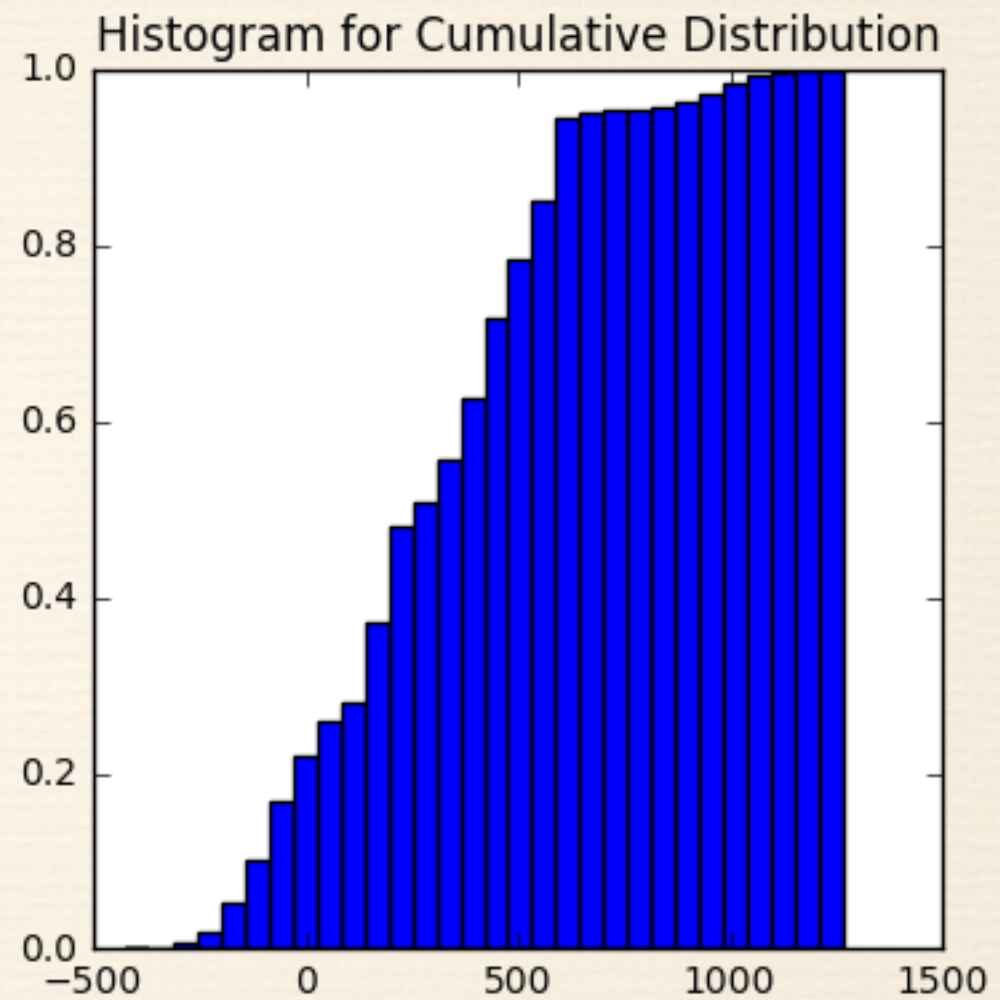
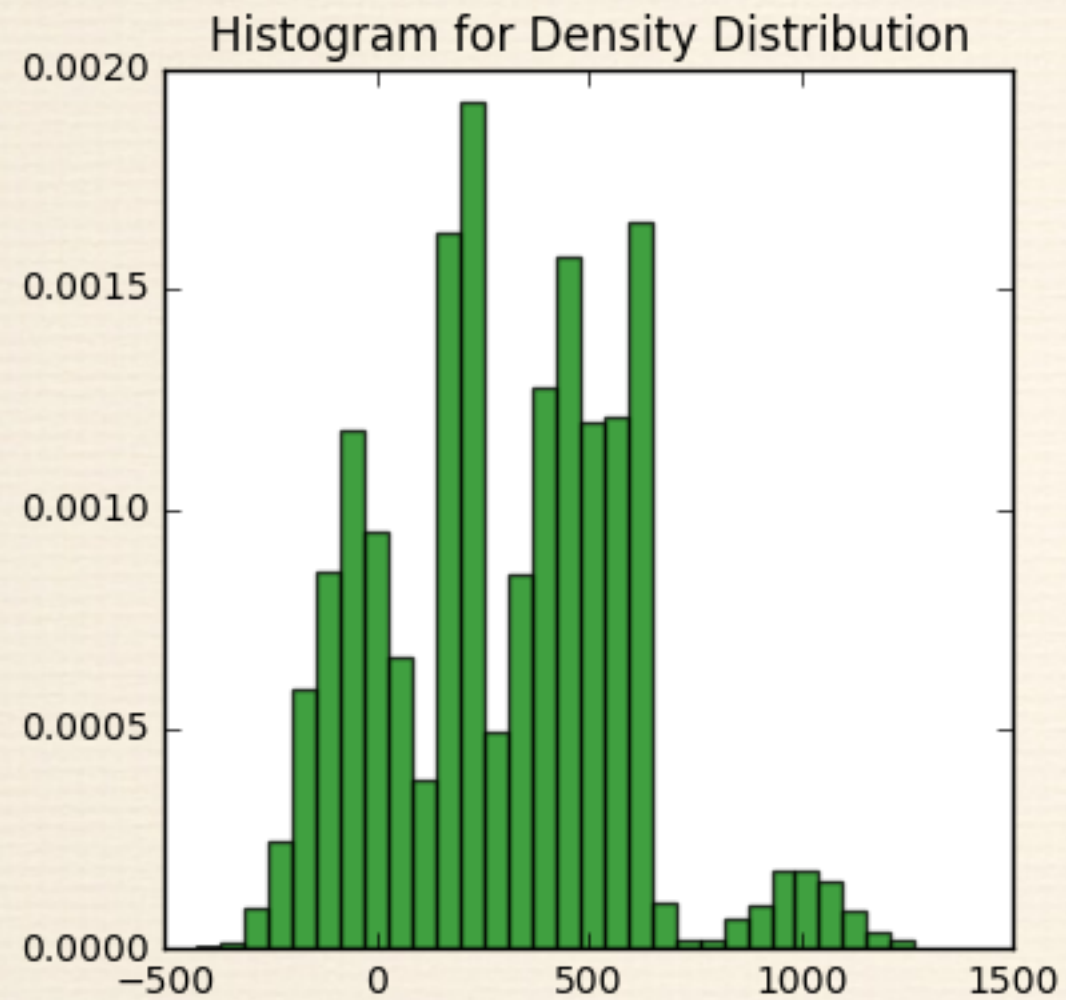
No Mode Uniform



Bi-Mode



Multi Mode



Some Notes

- ❖ histogram
- ❖ percentiles/box plot
- ❖ mean (sensitive to extreme value) v.s. median (not sensitive)

Inferential Statistics

- ❖ from sample to population
- ❖ make estimate, conduct statement
- ❖ Given a Statistic,
 - ❖ estimate its value in a confidence range
 - ❖ Compare the observed statistic to other statistics or constants, to conduct conclusion on it/them

dummy example 1

- ❖ the average height of man, in what range?
- ❖ is the the average height of man larger than the average height of woman? how sure you are?

Remark

- ❖ Statistic for a random variable is a random variable
- ❖ Chase the distribution of a statistic, we would have answers for above questions...

dummy example 2

- ❖ assume data from r.v.s. are independent, identical distribution
- ❖ sample mean follow normal distribution
- ❖ assume data from normal distributed ind, ide distribution
- ❖ sample variance follow chi-square distribution

Confidence Interval

- ❖ Inference on the population parameter
- ❖ Sample Statistics provide the estimate
- ❖ Derive the Probability Distribution of the Sample Statistics to provide interval estimate

Instance

*estimate the population mean μ
sample mean \bar{X} is the statistics*

An estimation is from an observed sample, \bar{x}

derive the distribution of \bar{X} $F(\bar{X})$

A confidence interval of α level for the μ then is :

$$\{\bar{x} - \text{Percentile}_{\alpha/2}F(\bar{X}), \bar{x} + \text{Percentile}_{1-\alpha/2}F(\bar{X})\}$$

Hypothesis Test

- ❖ Inference on the population parameter
- ❖ Sample Statistics provide the estimate
- ❖ Derive the Probability Distribution of the Sample Statistics to provide the evidence how far the estimate depart from our hypothesis and if the evidence is stronger than our desire, we may make a statement for the parameter

Instance

$$H_0 : \mu = 0(\mu_0); H_a : \mu \neq 0(\mu_a)$$

sample mean \bar{X} is the statistics

An estimation is from an observed sample, \bar{x}

derive the distribution of \bar{X} $F(\bar{X})$

the departure likelihood from evidence to hypothesis

is measured by $1 - F(|\bar{x} - \mu_0|)$

If above measure is smaller than our setting $\alpha/2^$,*

we would reject H_0

**it is called significance level,*

** normally α is choose to be 0.05 or 0.1*

Reality

- ❖ Formulas are beautiful?
- ❖ Assumptions are hidden
- ❖ Assumptions are fragile
- ❖ We want big data

Big Data 1

- ❖ Non parametric approach
- ❖ Re-sampling procedure to re-form random samples (bootstrap)
- ❖ Brute force simulate the distribution of the statistics
- ❖ With even larger data, we may have major proportion of the population

Big Data 2

- ❖ Outlier

- ❖ extreme values, when sample is small, maybe a handful of such observations

- ❖ Rare Event

- ❖ may also present themselves as extreme values or small likelihood events, when sample is large, maybe thousands of such observations

Practice

- ❖ Install Anaconda Python Environment
 - ❖ <https://www.continuum.io/downloads>
- ❖ Install lib: plotly
 - ❖ <https://plot.ly/python/getting-started/>
 - ❖ <https://plot.ly/python/reference/#scatter>
- ❖ Set up jupyter notebook
 - ❖ with in anaconda
- ❖ Matplotlib reference
 - ❖ <http://matplotlib.org/contents.html>

Practical Suggestion

- ❖ Avoid windows
- ❖ iOS is DS friendly
- ❖ Linux is DS friendly
 - ❖ if no Linux installed, VirtualBox to set up a virtual machine (email me if you wish)
 - ❖ <https://www.virtualbox.org/wiki/Downloads>
 - ❖ <https://www.ubuntu.com/download/desktop>