Lesson A2
Mean statistics, Continuous Distributions, Probability density functions
GEO1001.2020

Clara García-Sánchez, Stelios Vitalis

Resources adapted from:
- David M. Lane et al. (http://onlinestatbook.com)
- Allen B. Downey et al. (https://greenteapress.com/wp/think-stats-2e/)

Lesson A2

Mean Statistics

Overview

- Central Tendency
- Variability
- Shape

Overview

- **Central Tendency**
- Variability
- Shape

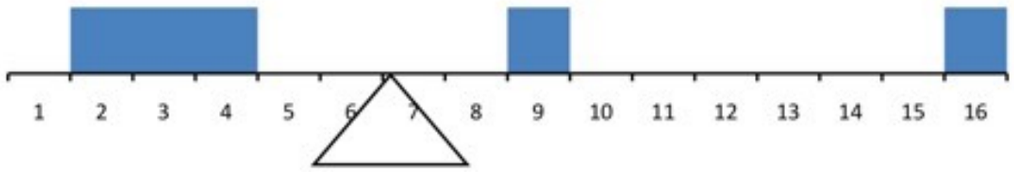
Central Tendency

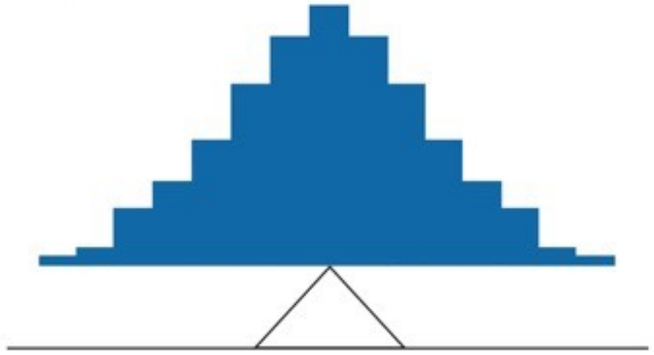
It is about determining where is the centre of the distribution. There are different ways to calculate it. The message after all is, how far your grade for example is from the centre of the distribution?

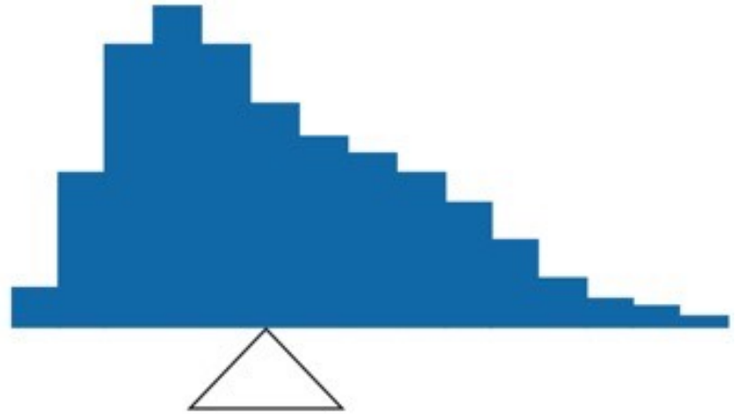
Student	Dataset A	Dataset B	Dataset C
You	3	3	3
John's	3	4	2
Maria's	3	4	2
Shareecia's	3	4	2
Luther's	3	5	1

Central Tendency

1. Balance scale







2. Smallest Absolute Deviation

Values	Absolute Deviations from 10
2	8
3	7
4	6
9	1
16	6
Sum	28

3. Smallest Squared Deviation

Values	Squared Deviations from 10
2	64
3	49
4	36
9	1
16	36
Sum	186

Central Tendency - Measures

1. Arithmetic mean (mean)

$$mean = \mu = \frac{\sum X}{N} = \frac{1}{n}\sum_i x_i$$

2. Median

Odd number of numbers: the median is the middle number

Even number of numbers: the median is the mean of the two middle numbers

When there are numbers with the same values —> 50th percentile formula

3. Mode

It is the most frequently occurring value in the list

The mode of continuous data is normally computed from a grouped frequency distribution

Overview

- Central Tendency
- **Variability**
- Shape

Variability - Measures

The term variability refers to how spread out is a distribution. It can also be referred as spread or dispersion of the distribution.

There are 4 frequently used measures of variability:

1. **Range**: simply the highest and the lowest number in the distribution.

2. **Interquartile Range (IQR)**: it is the range of the middle 50% of the scores in a distribution, as we saw before.

3. **Variance**: is the averaged squared difference of the scores from the mean.

$$variance = \sigma^2 = \frac{\sum (X - \mu)^2}{N} = \frac{1}{n} \sum_i (x_i - \mu)^2$$

4. **Standard deviation**: is simply the square root of the variance

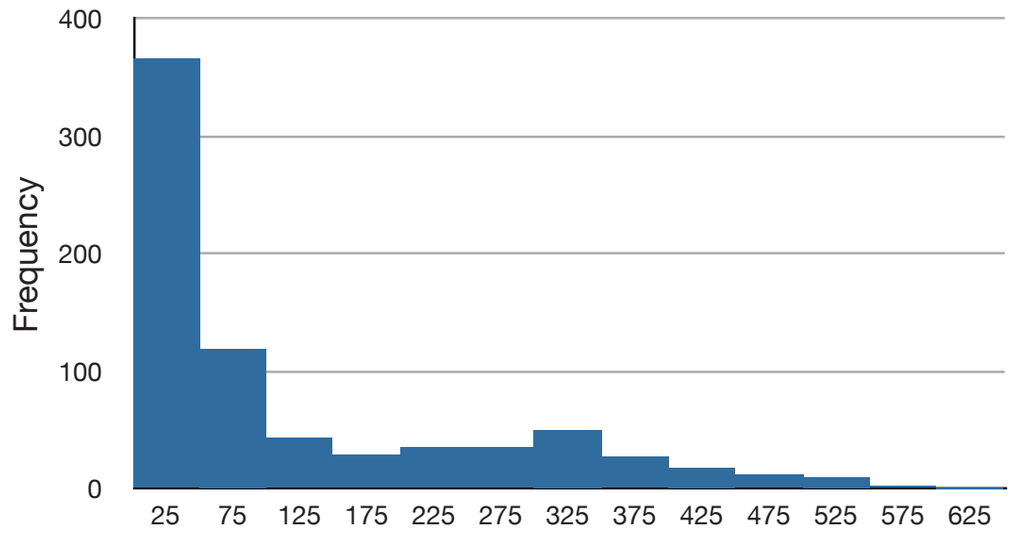
Overview

- Central Tendency
- Variability
- **Shape**

Shapes and distributions

We focused in two measures of shape distributions: skew and kurtosis (also know as 3rd and 4th order moments (1st and 2nd being σ and σ^2))

1. **Skewness**: distributions with + skew normally means mean>>median



$$\sigma^3 = \sum \frac{(X-\mu)^3}{\sigma^3}$$

2. **Kurtosis**: the value 3 is subtracted to define no kurtosis of a normal distribution, otherwise a normal distribution would have a kurtosis 3.

$$\sigma^4 = \sum \frac{(X-\mu)^4}{\sigma^4} - 3$$

Lesson A2

Probability Density Functions

Overview

- Probability Mass Functions (PMFs)
- Cumulative Density Functions (CDFs)
- Probability Density Functions (PDFs)
- Kernel Density Estimation
- The distribution Framework
- Moments
- Skewness

Overview

- **Probability Mass Functions (PMFs)**
- Cumulative Density Functions (CDFs)
- Probability Density Functions (PDFs)
- Kernel Density Estimation
- The distribution Framework
- Moments
- Skewness

Probability Mass Functions (PMFs)

It is a way to represent a distribution, which maps from each value its probability. **Probability** is a frequency expressed as a fraction of the sample size, n. To get from frequencies to probabilities, we divide through by n, which is called **normalization**.

- To plot a PMF you can use "pyplot.hist".
- By plotting PMF, instead of histogram, we can compare two distributions without being **mislead by the difference in sample size** (since PMFs are normalized)

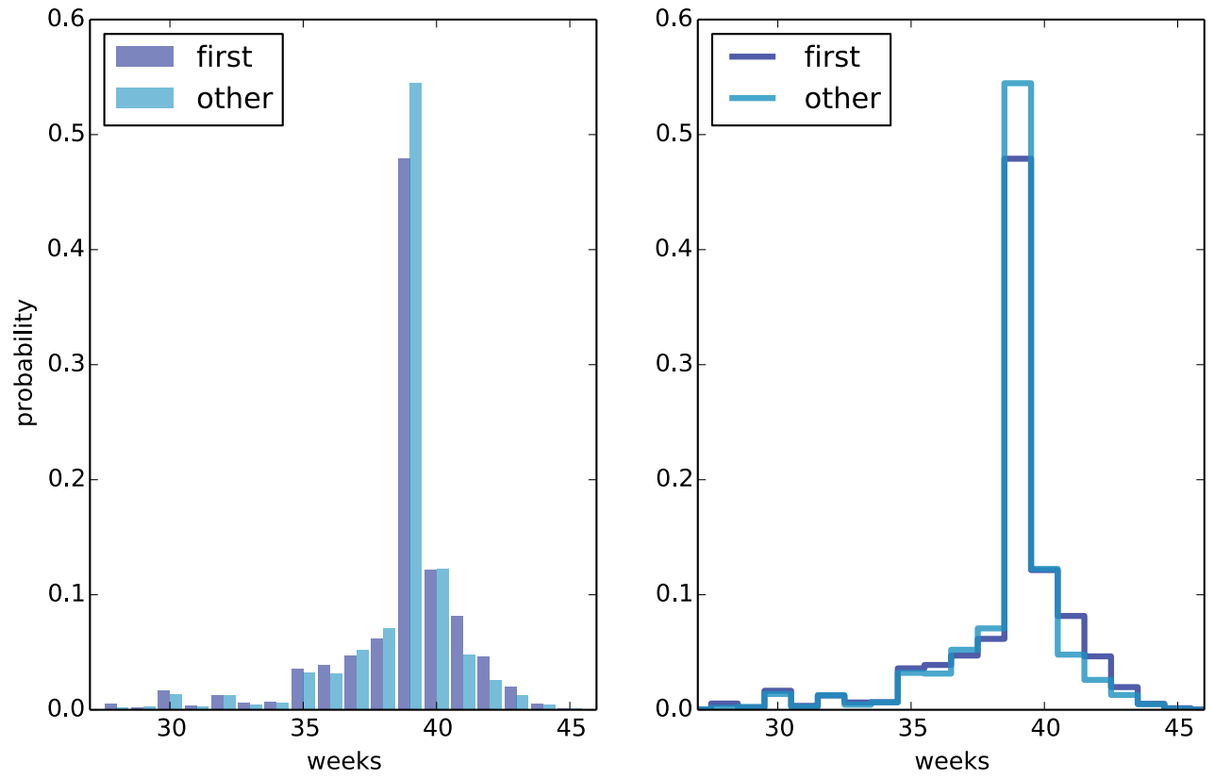


Figure 3.1: PMF of pregnancy lengths for first babies and others, using bar graphs and step functions.

Overview

- Probability Mass Functions (PMFs)
- **Cumulative Density Functions (CDFs)**
- Probability Density Functions (PDFs)
- Kernel Density Estimation
- The distribution Framework
- Moments
- Skewness

Cumulative Density Functions (CDFs)

The CDF is the function that maps from a value to its percentile rank.

It is a function of x, where x is any value that might appear in the distribution, and to evaluate the CDF(x) for a particular value of x, we compute the fraction of values in the distribution less than or equal to x.

Example:

- Suppose a sample [1,2,2,3,5].

- Some values of the CDF are:
CDF(0)=0, CDF(1)=0.2,
CDF(2)=0.6, CDF(3)=0.8,
CDF(4)=0.8, CDF(5)=1

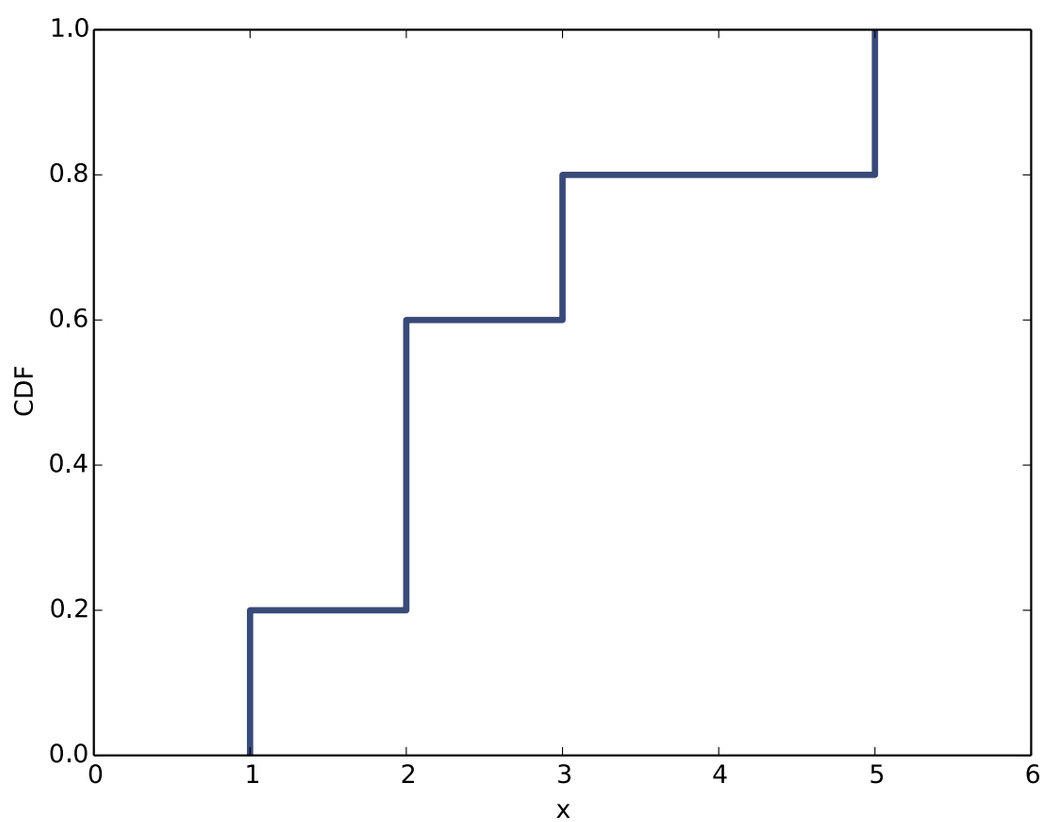


Figure 4.2: Example of a CDF.

Overview

- Probability Mass Functions (PMFs)
- Cumulative Density Functions (CDFs)
- **Probability Density Functions (PDFs)**
- Kernel Density Estimation
- The distribution Framework
- Moments
- Skewness

Probability Density Functions (PDFs)

The derivative of a CDF is called “probability density function”

$$\mathrm{PDF}_{expo}(x) = \lambda e^{-\lambda x} \longleftarrow CDF(x) = 1 - e^{-\lambda x}$$

Evaluating a PDF for a particular value of x is usually not that useful. The result is not a probability, but a probability density.

Probability density measures the probability per unit of x, but if you want to get the probability mass you need to integrate over x.

Overview

- Probability Mass Functions (PMFs)
- Cumulative Density Functions (CDFs)
- Probability Density Functions (PDFs)
- **Kernel Density Estimation**
- The distribution Framework
- Moments
- Skewness

Kernel Density Estimation

Kernel density estimation (KDE) is an algorithm that takes a sample and finds an appropriately smooth PDF that fits the data.

To compute it with python, you can use the library “scipy”

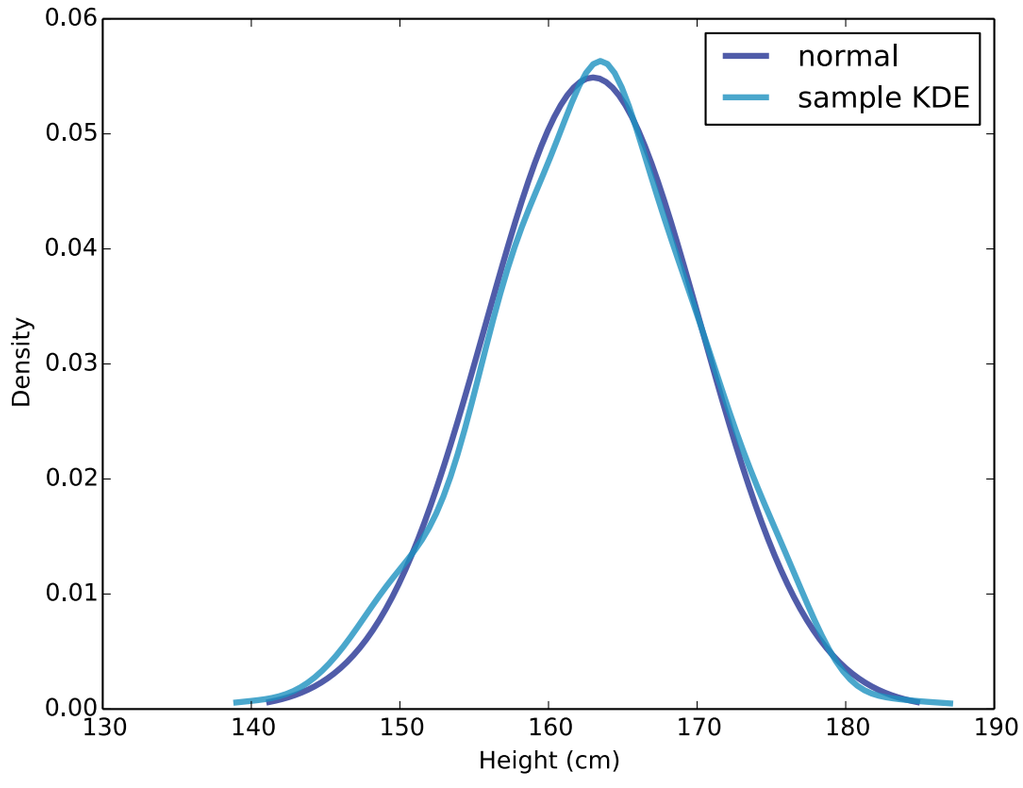


Figure 6.1: A normal PDF that models adult female height in the U.S., and the kernel density estimate of a sample with $n = 500$.

Kernel Density Estimation

Why KDE is useful?

- *Visualization*: during a project, the CDFs are usually the best visualisation of the data. However, for audiences, understanding PDFs is much easier than CDFs.

- *Interpolation*: an estimated PDF is a way to get from a sample to a model of the population. If you think the population distribution is smooth, then you can use KDE to interpolate the density values that don't appear in the sample data.

- *Simulation*: simulations are often based on the distribution of a sample, if the sample size is small, it might be appropriate to smooth the sample distribution with the KDE.

Overview

- Probability Mass Functions (PMFs)
- Cumulative Density Functions (CDFs)
- Probability Density Functions (PDFs)
- Kernel Density Estimation
- **The distribution Framework**
- Moments
- Skewness

The distribution Framework

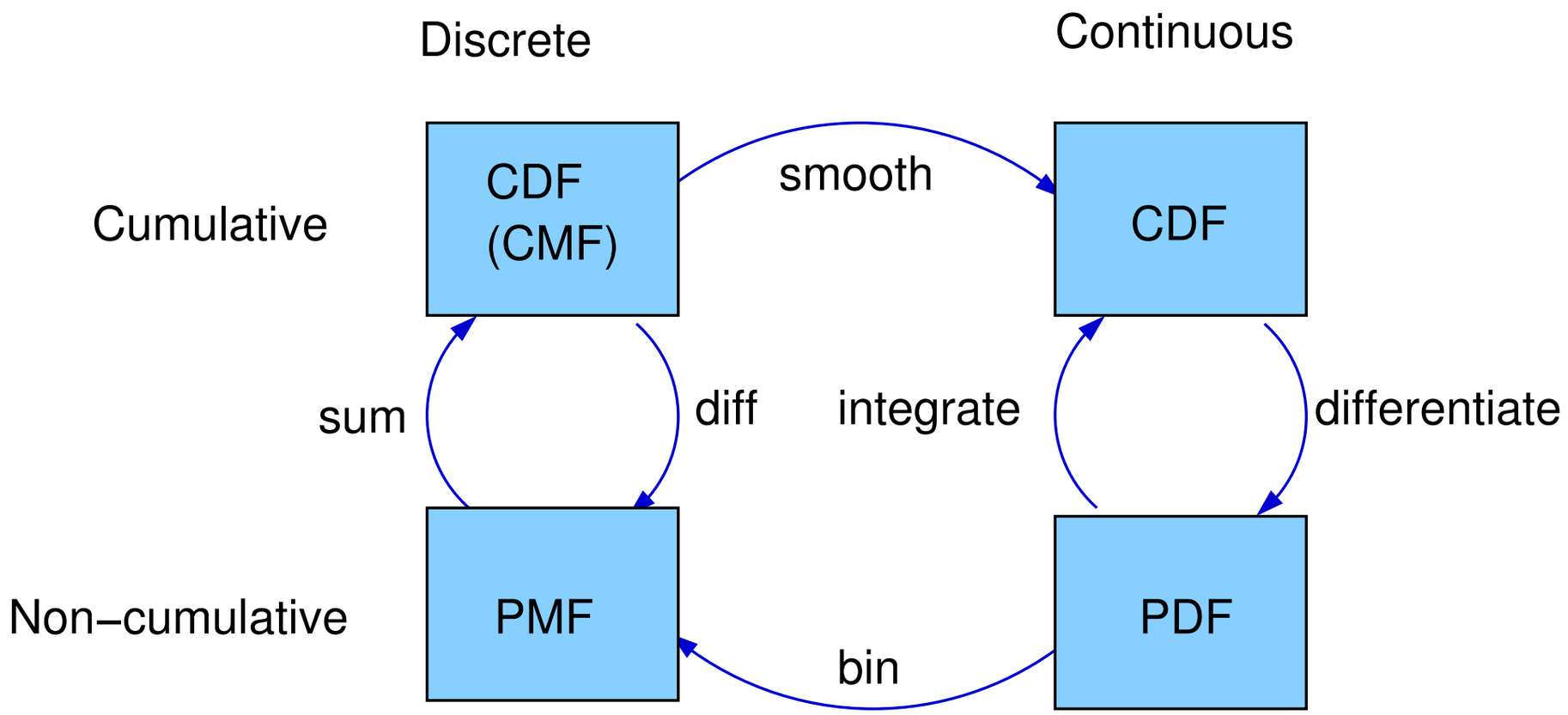


Figure 6.2: A framework that relates representations of distribution functions.

To get from discrete to continuous you can use different forms of smoothing:

1) Assume the data come from an analytic continuous distribution (like exponential or normal) and to estimate the parameters of that distribution.

2) Use kernel density

Practice (Optional)

Python practice to implement hist, pmf, and cdf using classes. This can be found in thinkStats book starting a page 80. This is optional for those who want to reinforce their python skills and probability concepts, no classes knowledge is required for this part of the course.

This practice won't be corrected or addressed in class, since it is optional.

Overview

- Probability Mass Functions (PMFs)
- Cumulative Density Functions (CDFs)
- Probability Density Functions (PDFs)
- Kernel Density Estimation
- The distribution Framework
- **Moments**
- Skewness

Moments

Whenever you take a data sample and reduce it to one number, that number is a statistic. We have seen several of them: mean, variance, median and IR.

Raw moments are kind of statistics, the kth raw moment of a sample is:

$$m'_k = \frac{1}{n}\sum_i x_i^k$$

E（期望）
= Σ pi * (xi)^k

因为(xi)^k要依据pi的值动态变化，
使乘积恒等于E，
类似于力学中的力矩，
故称(xi)^k是p的k阶矩

If k=1 the result is the sample mean.

The **central moments** are more useful, the kth central moment is:

$$m_k = \frac{1}{n}\sum_i (x_i - \mu)^k$$

If k=2, the result is the second central moment, which is actually the variance.

When you report a moment-based statistic it is important to think about the **units**!

Overview

- Probability Mass Functions (PMFs)
- Cumulative Density Functions (CDFs)
- Probability Density Functions (PDFs)
- Kernel Density Estimation
- The distribution Framework
- Moments
- **Skewness**

Skewness

We have heard already from **skewness**. It is a property that describes the shape of a distribution, if the distribution is symmetric around its central tendency, it is unscrewed. Otherwise it can be “right/left skewed”.

Skewness is the the third standardised moment, which means it has been normalised and it has no units.

$$m_3 = E\left[\left(\frac{x_i - \mu}{\sigma}\right)^3\right]$$

Negative **skewness** indicates skew to the left, and positive skew to the right.

In practice it is not good idea to compute it if there are outliers, since they have a large impact. A better way to do it is using **Pearson’s median skewness coefficient** which is a measure of the skewness based on the difference between the sample mean and the median:

$$g_p = 3(\bar{x} - m)/S$$

Practice

Download code exercise6-1.py and extract female heights. Using that distribution compute:

1) Compute and print median, mean, standard deviation, variance skewness, kurtosis
2) Plot a PMF, PDF and CDF of the female heights. You will need to code a function for PMF, while PDF and CDF have already functions defined within python libraries.

Lesson A2

Continuous Distributions

Overview

- Introduction
- The exponential distribution
- The Normal distribution
- Areas of Normal Distribution
- The lognormal distribution
- The Pareto distribution

Overview

- **Introduction**
- The exponential distribution
- The Normal distribution
- Areas of Normal Distribution
- The lognormal distribution
- The Pareto distribution

Introduction

All the distributions we have plotted based on data are called: "empirical distributions", because they are based on empirical observations with finite samples.

An alternative are the so called: **continuous distributions**, which are characterised by CDFs that is a continuous function (as opposed to a step function).

Many real world phenomena can be approximated by continuous distributions —> why this is important?

Discuss for 5min in the chat

Overview

- Introduction
- **The exponential distribution**
- The Normal distribution
- Areas of Normal Distribution
- The lognormal distribution
- The Pareto distribution

The Exponential distribution

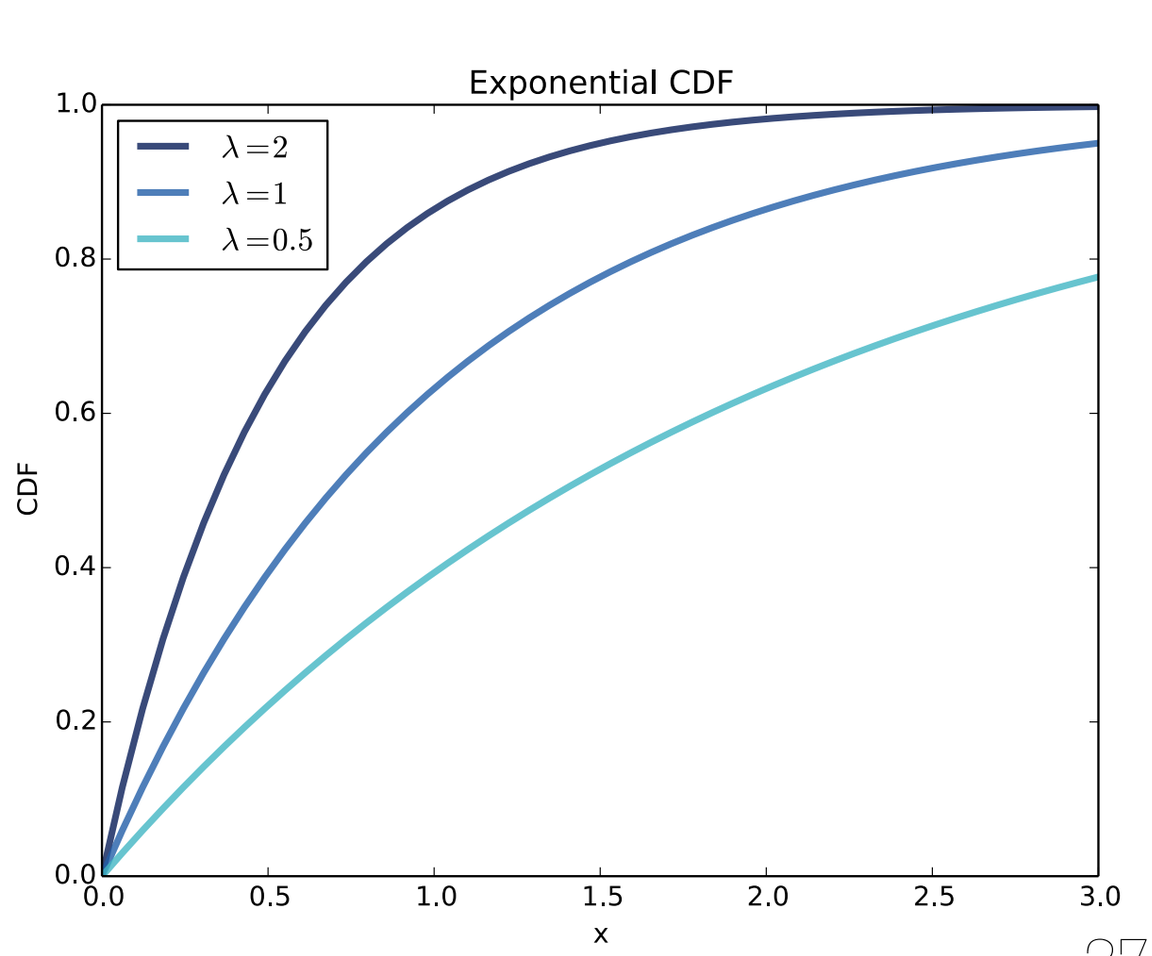
Exponential distributions come up when we look at a series of events and measure the times between events, which are called "interarrival times" —> if the events are equally likely to occur at any time, the distribution of inter arrival times tends to look like an exponential function

$$CDF(x) = 1 - e^{-\lambda x}$$

$$PDF(x) = \lambda e^{-\lambda x}$$

The parameter λ determines the shape of the distribution (λ=2 in fig)

$$\mu = \frac{1}{\lambda}; \sigma^2 = \frac{1}{\lambda^2}$$



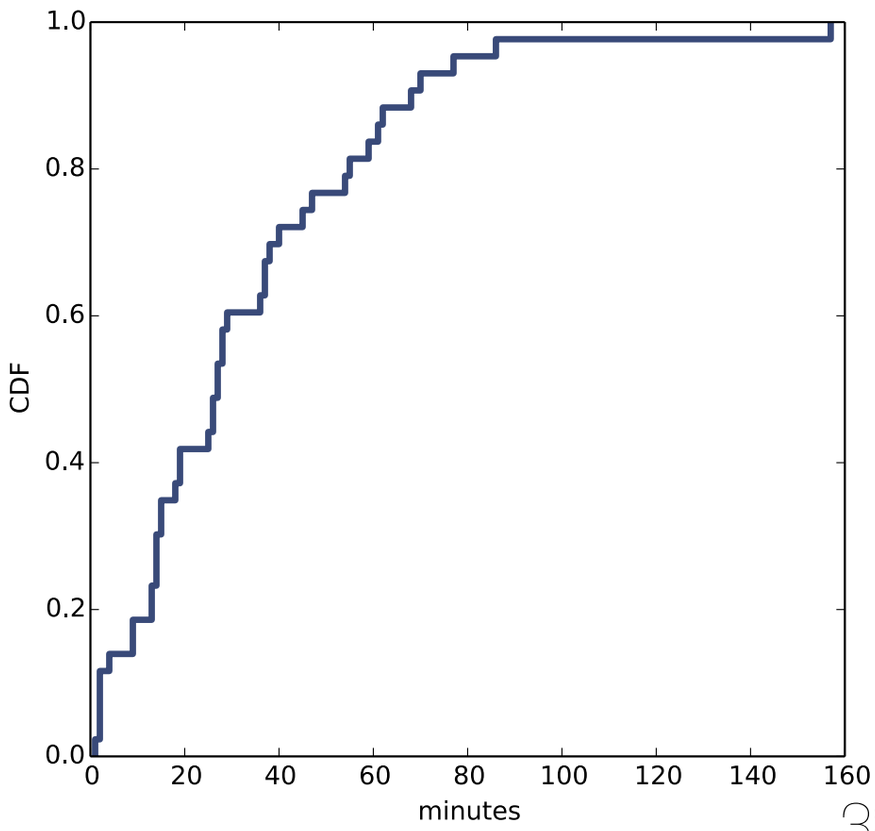
Practice

Let's look at the inter arrival time of births. Using the data downloaded from the repository, lets make this small program:

```
df = ReadBabyBoom()
diffs = df.minutes.diff()
cdf = thinkstats2.Cdf(diffs, label='actual')

thinkplot.Cdf(cdf)
thinkplot.Show(xlabel='minutes', ylabel='CDF')
```

1) Download Lecture2ContinuousDistributions.py
2) Comment the script with the specifications of what ReadBabyRoom(), diff(), thinkstats2.Cdf and thinkplot.Cdf functions do
3) Seems exponential, but how can we be sure? Find a way to proof it is or not exponential and plot the solution (think about complementary CDFs —> 1-CDF)



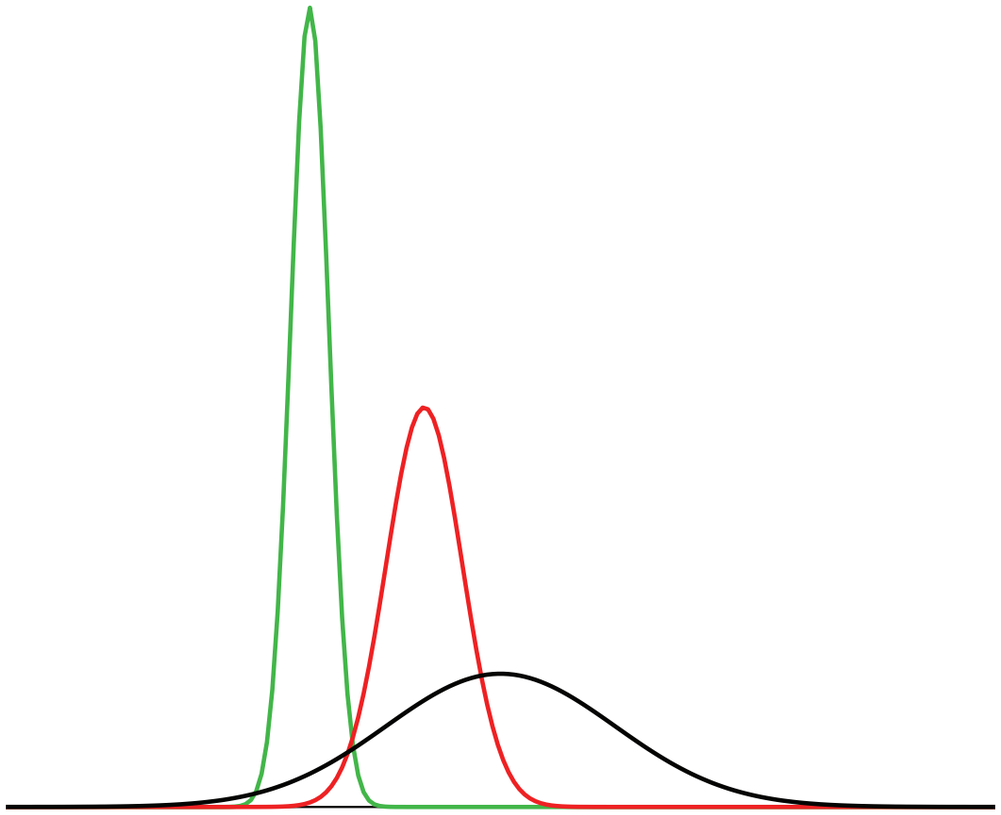
Overview

- Introduction
- The exponential distribution
- **The Normal distribution**
- Areas of Normal Distribution
- The lognormal distribution
- The Pareto distribution

The Normal distribution

The normal distribution is the most important and most widely used distribution in statistics:

- Also called “Gaussian distribution”



Which on has the largest mean?

Which one has the smallest standard deviation?

Figure 1. Normal distributions differing in mean and standard deviation.

The Normal distribution

The density of the normal distribution (the height for a given value on the x-axis) can be computed as:

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

Properties of normal distributions:

1. They are symmetric around their mean
2. The mean, median, and mode are equal
3. The are under the curve is equal to 1
4. They are defined by two parameters: μ, σ
5. They are denser in the centre than in the tails
6. 68% of the area is within one standard deviation of the mean
7. 98% is approximately within 2 standard deviations from the mean

Overview

- Introduction
- The exponential distribution
- The Normal distribution
- **Areas of Normal Distribution**
- The lognormal distribution
- The Pareto distribution

Areas under Normal distribution

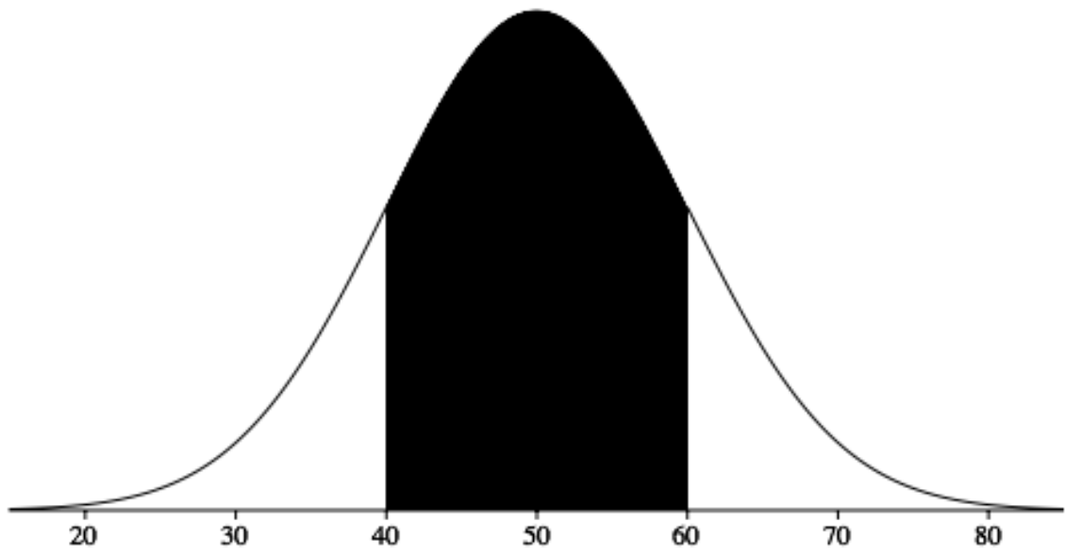


Figure 1. Normal distribution with a mean of 50 and standard deviation of 10. 68% of the area is within one standard deviation (10) of the mean (50).

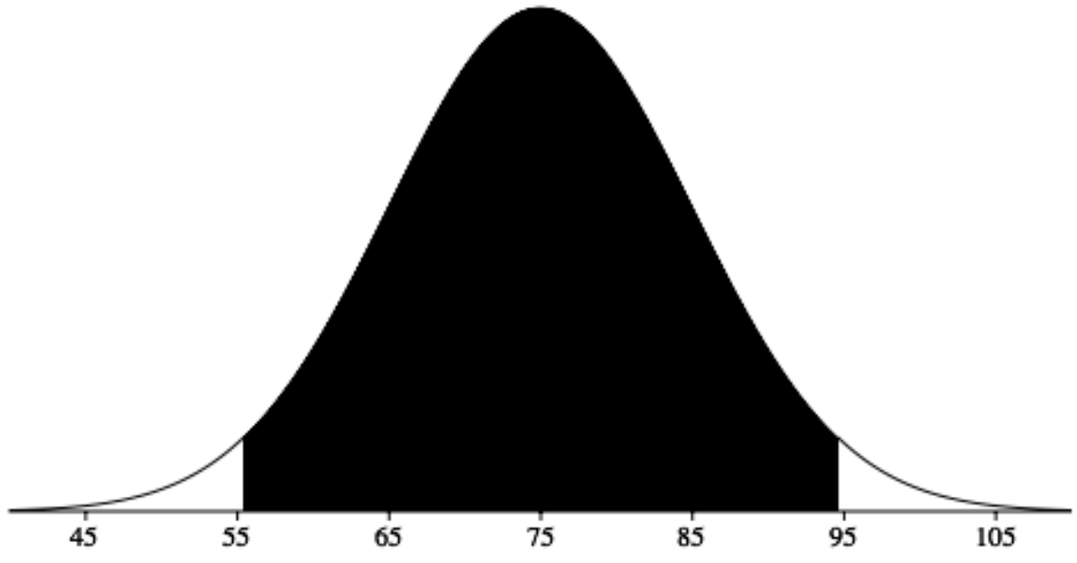


Figure 3. A normal distribution with a mean of 75 and a standard deviation of 10. 95% of the area is within 1.96 standard deviations of the mean.

Overview

- Introduction
- The exponential distribution
- The Normal distribution
- Areas of Normal Distribution
- **The lognormal distribution**
- The Pareto distribution

The Lognormal distribution

If the logarithms of a set of values have a normal distribution, the values have a **lognormal distribution**

$$CDF_{lognormal}(x) = CDF_{normal}(\log x)$$

If a sample is lognormal and you plot its CDF on a log-x scale, it will have the characteristic shape of a normal distribution:

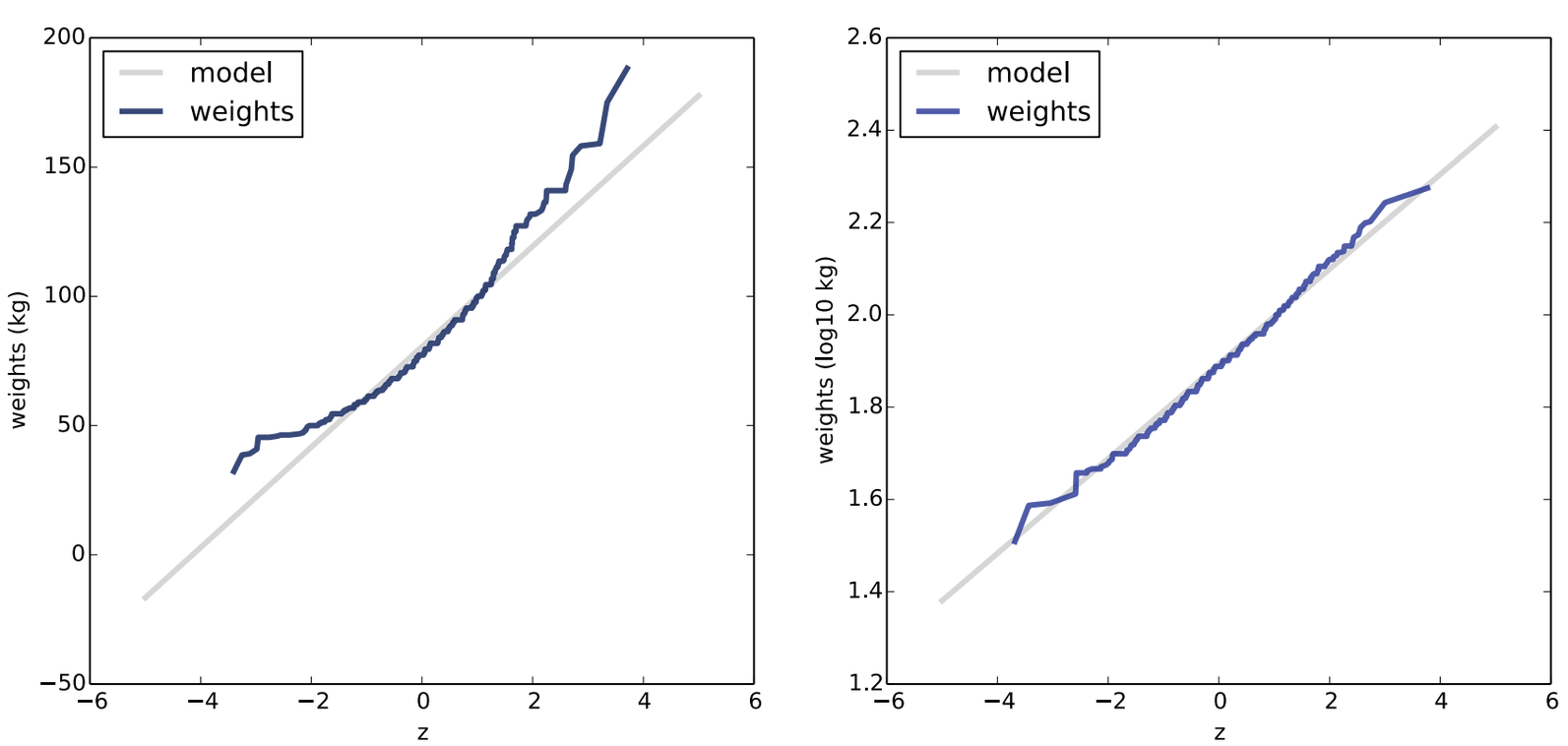


Figure 5.8: Normal probability plots for adult weight on a linear scale (left) and log scale (right).

Overview

- Introduction
- The exponential distribution
- The Normal distribution
- Areas of Normal Distribution
- The lognormal distribution
- **The Pareto distribution**

The Pareto distribution

The pareto distribution has been used to describe phenomena in the natural and social sciences, including sizes of cities and towns, sand particles and meteorites, forest fires or earthquakes.

$$CDF(x) = 1 - \left(\frac{x}{x_m}\right)^{-\alpha}$$

Where the parameters x_m and $\boldsymbol{\alpha}$ determine the location and shape of the distribution.

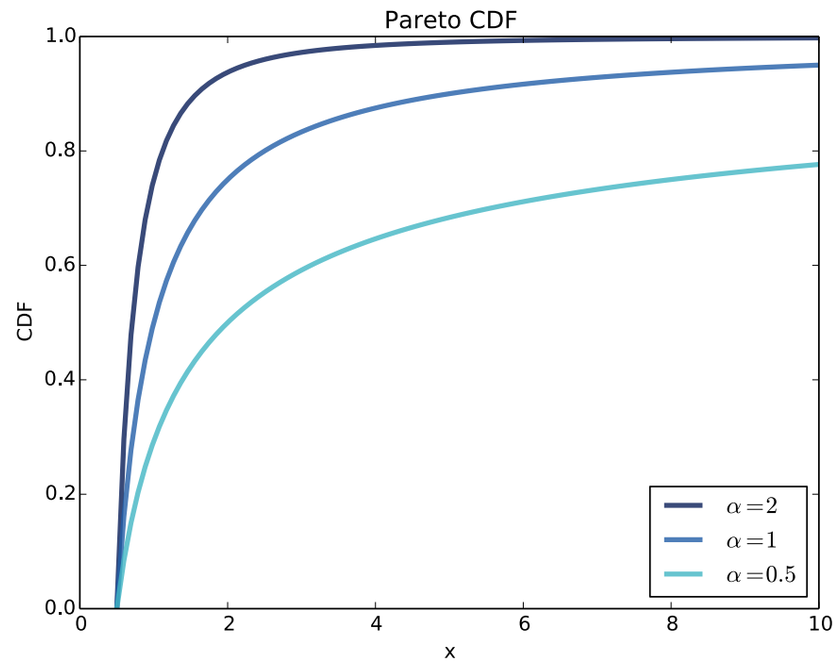


Figure 5.9: CDFs of Pareto distributions with different parameters.

Practice (optional)

Simple visual test that indicates whether an empirical distribution fits a Pareto distribution: on a log-log scale, the CCDF looks like a straight line

$$CDF(x) = 1 - \left(\frac{x}{x_m}\right)^{-\alpha}$$

Use the data in the repository (PEP_2012_PEPANNRES_with_ann.csv) and populations.py to check if the sizes of cities and towns follows a pareto distribution