CAP 5768 – Intro to Data Science

Lecture 3: Statistics – Part 2

Oge Marques, PhD

Professor

College of Engineering and Computer Science

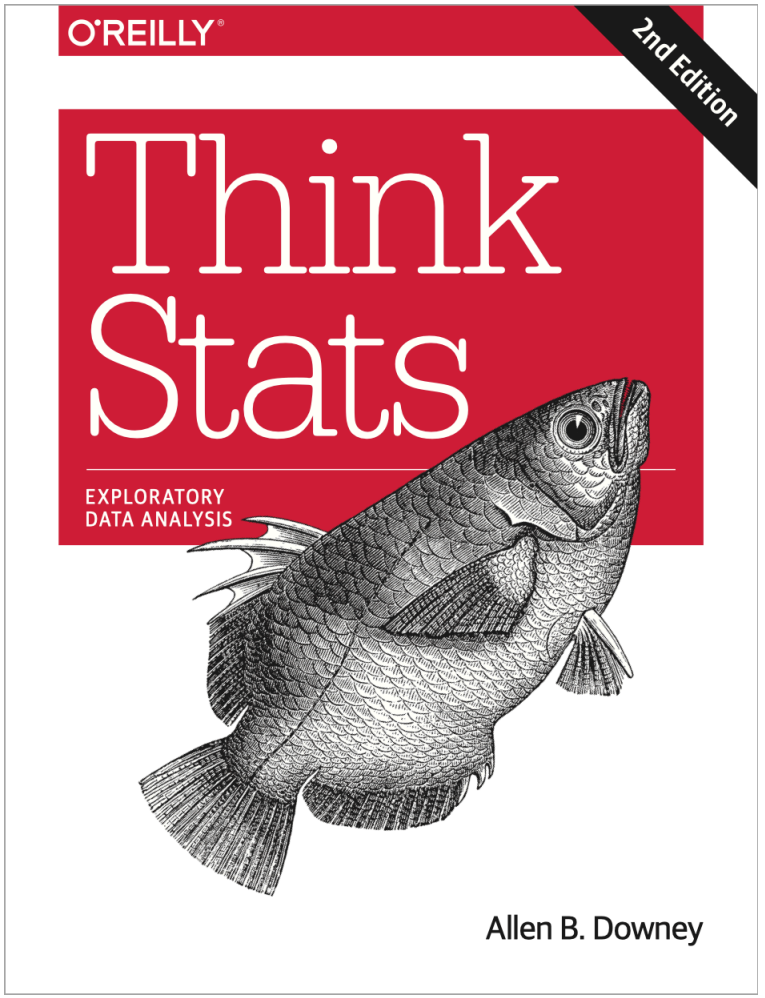
College of Business

FAU FLORIDA ATLANTIC UNIVERSITY

Highlights from

"Think Stats 2/e"

by Allen B. Downey



Chapter 5: Modeling distributions

Exponential distribution

Normal (Gaussian) distribution

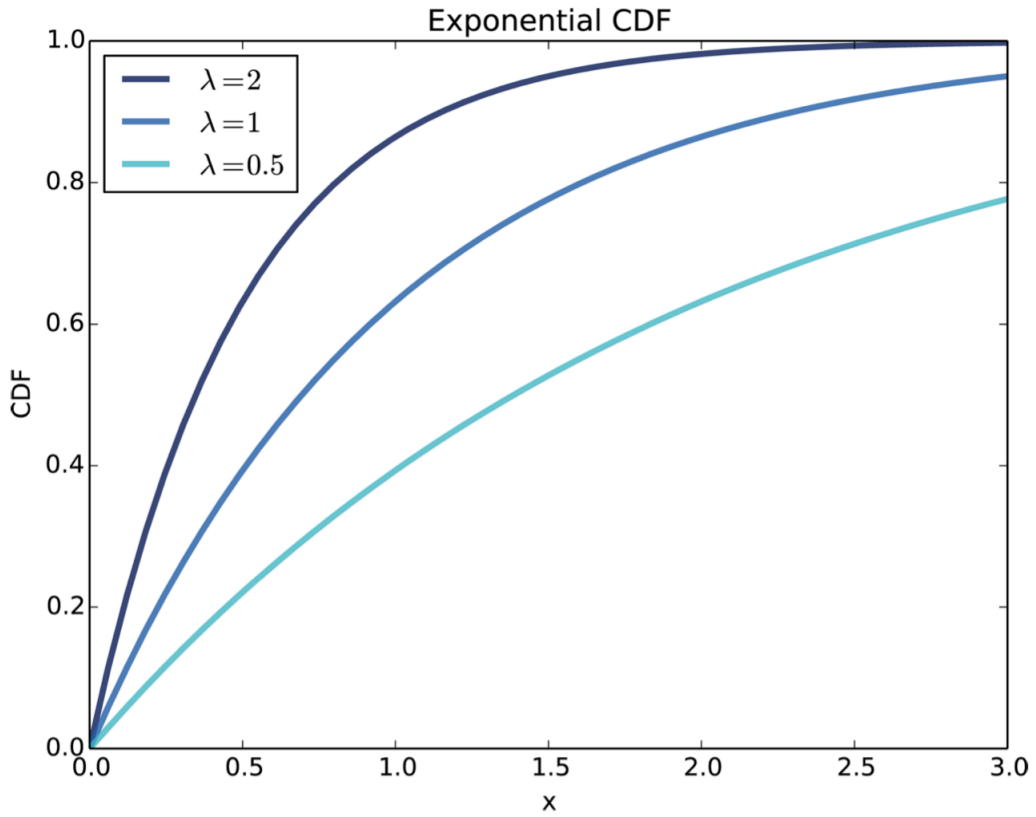
Lognormal distribution

Pareto distribution

Transition: from *empirical* to *analytic* distributions

- **Empirical distributions** are based on *empirical observations*, which are necessarily finite samples.
- The alternative is an **analytic distribution**, which is characterized by a CDF that is a *mathematical function*.
- **Analytic distributions can be used to model empirical distributions**.
  - In this context, a model is a simplification that leaves out unneeded details.

Exponential distribution



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

Exponential distribution

- In the real world, exponential distributions come up when we look at a series of events and measure the times between events, called **interarrival times**.
- If the events are equally likely to occur at any time, the distribution of interarrival times tends to look like an exponential distribution.
- Example: interarrival time of births
  - On December 18, 1997, 44 babies were born in a hospital in Brisbane, Australia.
  - The time of birth for all 44 babies was reported in the local paper.
  - The complete dataset is in a file called *babyboom.dat*, in the ThinkStats2 repository.
  - CDF (and CCDF) plot(s): see next slide

Complementary CDF (CCDF): 1 – CDF(x)

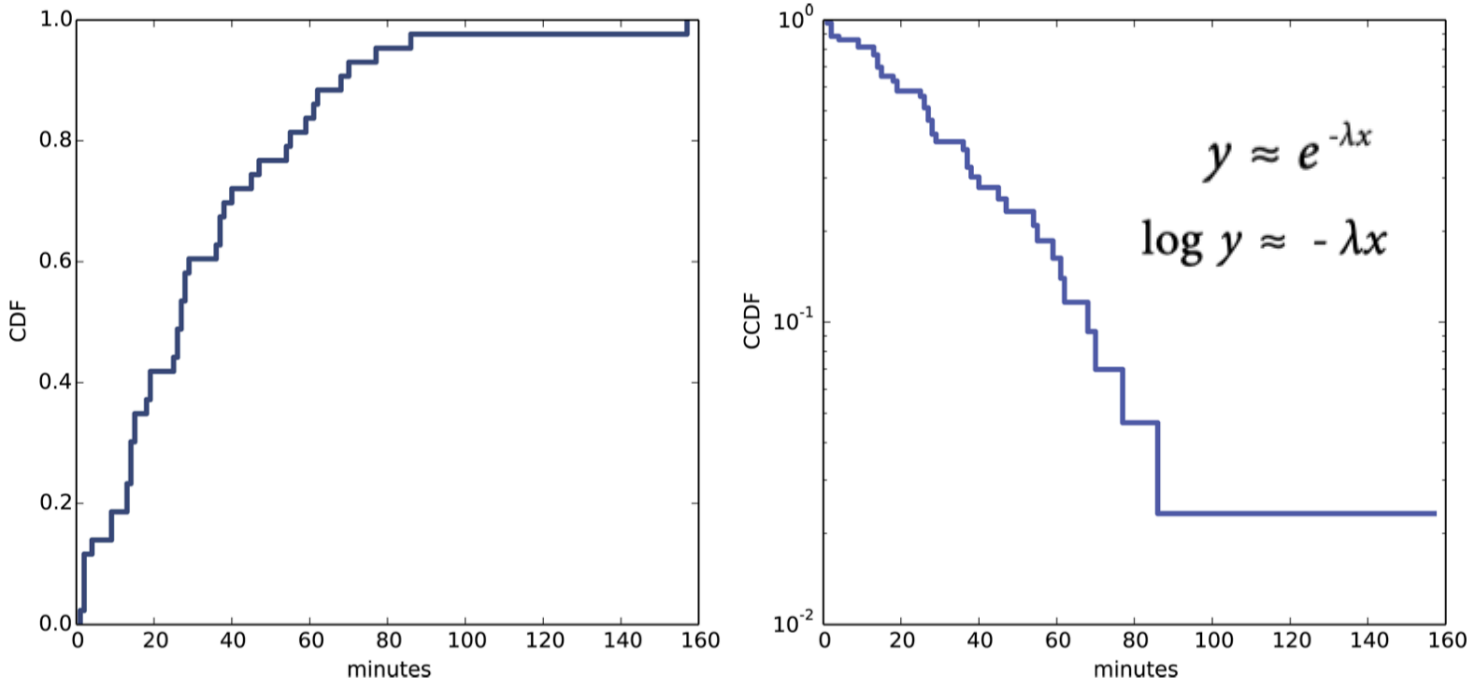


Figure 5-2. CDF of interarrival times (left) and CCDF on a log-y scale (right)

The mean of an exponential distribution is 1 / λ , so the mean time between births is 32.7 minutes.

Normal (Gaussian) distribution

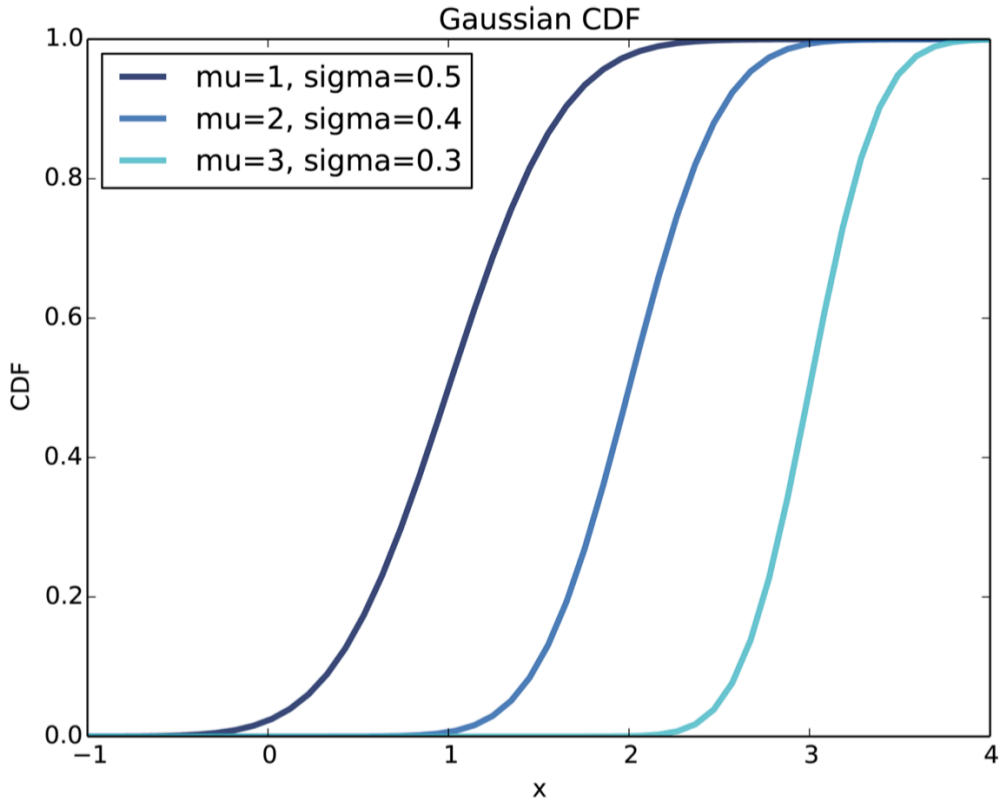
- The normal distribution, also called Gaussian, is commonly used because it describes many phenomena, at least approximately.
- Central Limit Theorem
  - if we add up **n** values from almost any distribution, the distribution of the sum converges to normal as **n** increases.
- The Central Limit Theorem explains the prevalence of normal distributions in the natural world.
  - Many characteristics of living things are affected by genetic and environmental factors whose effect is additive.
  - The characteristics we measure are the sum of a large number of small effects, so their distribution tends to be normal.

Normal (Gaussian) distribution

- The normal distribution is characterized by two parameters: the mean, μ, and standard deviation σ.
  - The normal distribution with $\mu = 0$ and $\sigma = 1$ is called the **standard normal distribution**.
- Its CDF is defined by an integral that does not have a closed form solution, but there are algorithms that evaluate it efficiently.
  - One of them is provided by SciPy: **scipy.stats.norm** is an object that represents a normal distribution; it provides a method, **cdf**, that evaluates the standard normal CDF.

Normal (Gaussian) distribution

Examples for different values of mu and sigma



Normal (Gaussian) distribution

- Distribution of baby weights in the NSFG dataset
  - The normal distribution is a good model for this dataset, so if we summarize the distribution with the parameters
    $\mu = 7.28$
    and
    $\sigma = 1.24$,
    the resulting error (difference between the model and the data) is small.

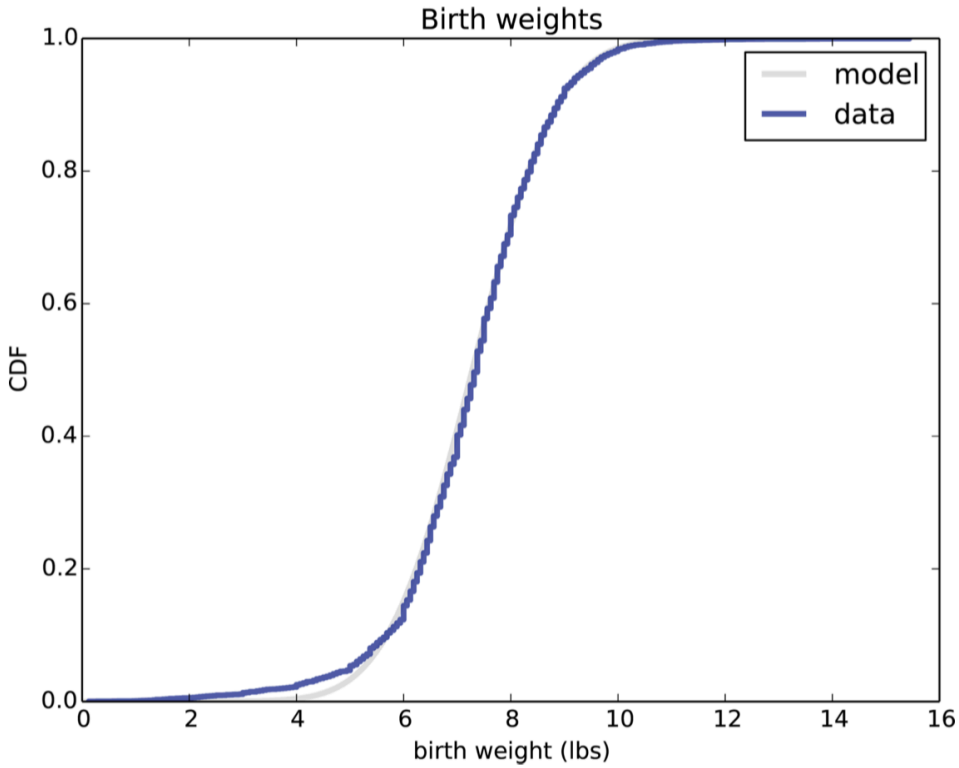


Figure 5-4. CDF of birth weights with a normal model

Normal probability plot

- A normal probability plot is a visual test for normality.
- How to generate a normal probability plot:
  1. Sort the values in the sample.
  2. From a standard normal distribution ($\mu = 0$ and $\sigma = 1$), generate a random sample with the same size as the sample, and sort it.
  3. Plot the sorted values from the sample versus the random values.
- If the distribution of the sample is approximately normal, the result is a straight line with intercept mu and slope sigma.

Normal probability plot: examples

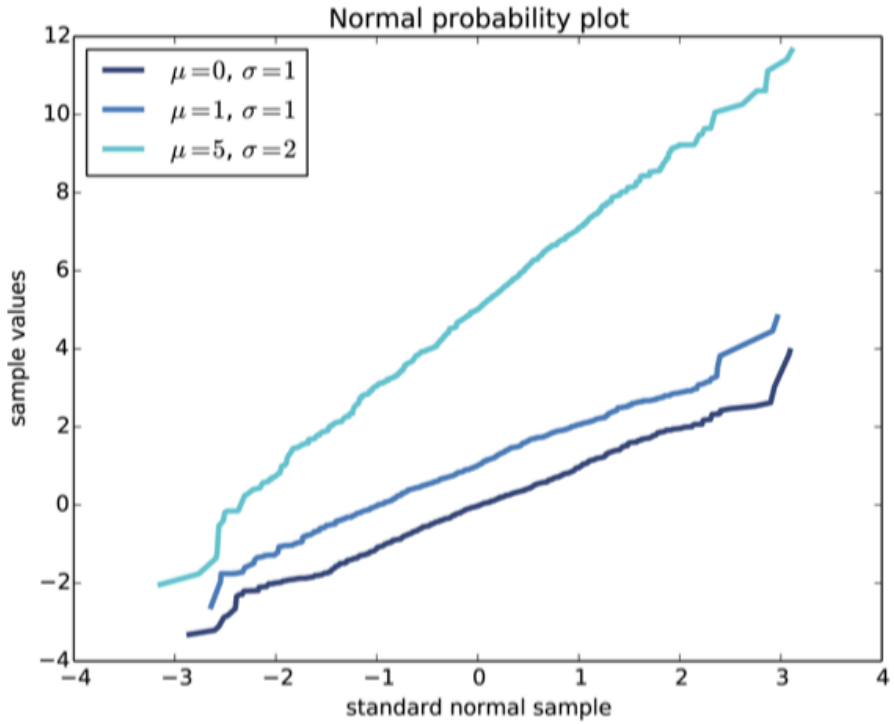
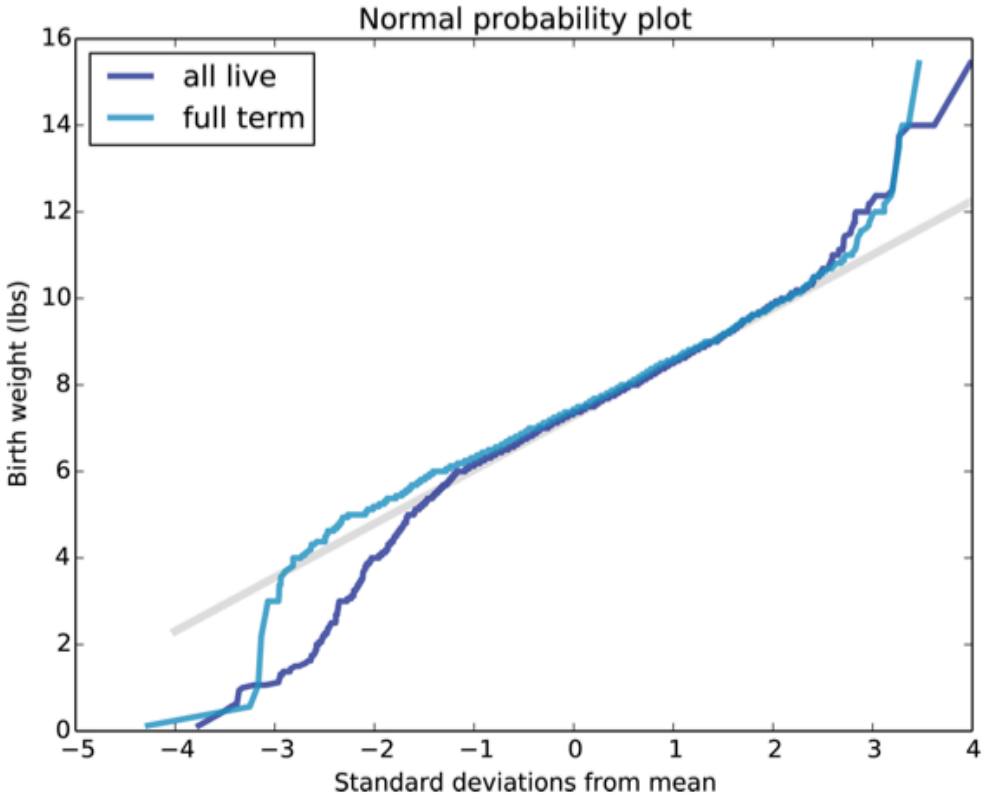


Figure 5-5. Normal probability plot for random samples from normal distributions

Normal probability plot: birth weights (NSFG)

- Results for all live births, and also for full term births (pregnancy length greater than 36 weeks).
- Both curves match the model near the mean and deviate in the tails.
- The heaviest babies are heavier than what the model expects, and the lightest babies are lighter.



The lognormal distribution

- If the logarithms of a set of values have a normal distribution, the values have a **lognormal distribution**.
- The CDF of the lognormal distribution is the same as the CDF of the normal distribution, with log x substituted for x.

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

- The parameters of the lognormal distribution are usually denoted μ and σ.
  - NB: *these parameters are <u>not</u> the mean and standard deviation!*
- If a sample is approximately lognormal and you plot its CDF on a log-x scale, it will have the characteristic shape of a normal distribution.
  - To test how well the sample fits a lognormal model, you can make a normal probability plot using the log of the values in the sample.

Dataset

- The National Center for Chronic Disease Prevention and Health Promotion conducts an annual survey as part of the Behavioral Risk Factor Surveillance System (BRFSS).
- In 2008, they interviewed 414,509 respondents and asked about their demographics, health, and health risks.
- Among the data they collected are the weights in kilograms of 398,484 respondents.
- The repository for the book contains **CDBRFS08.ASC.gz**, a fixed-width ASCII file that contains data from the BRFSS, and **brfss.py**, which reads the file and analyzes the data.

The adult weight distribution: CDFs

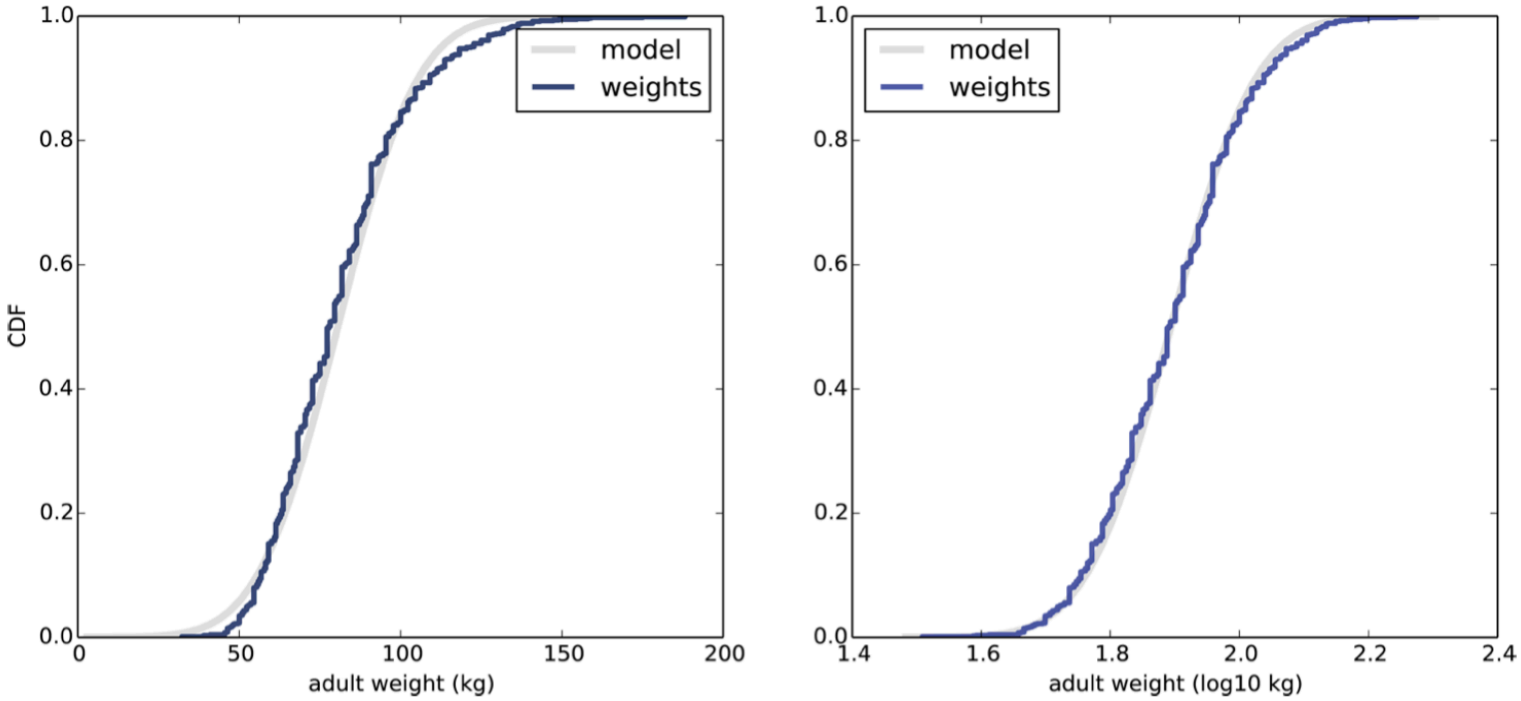


Figure 5-7. CDF of adult weights on a linear scale (left) and log scale (right)

The adult weight distribution: normal probability plots

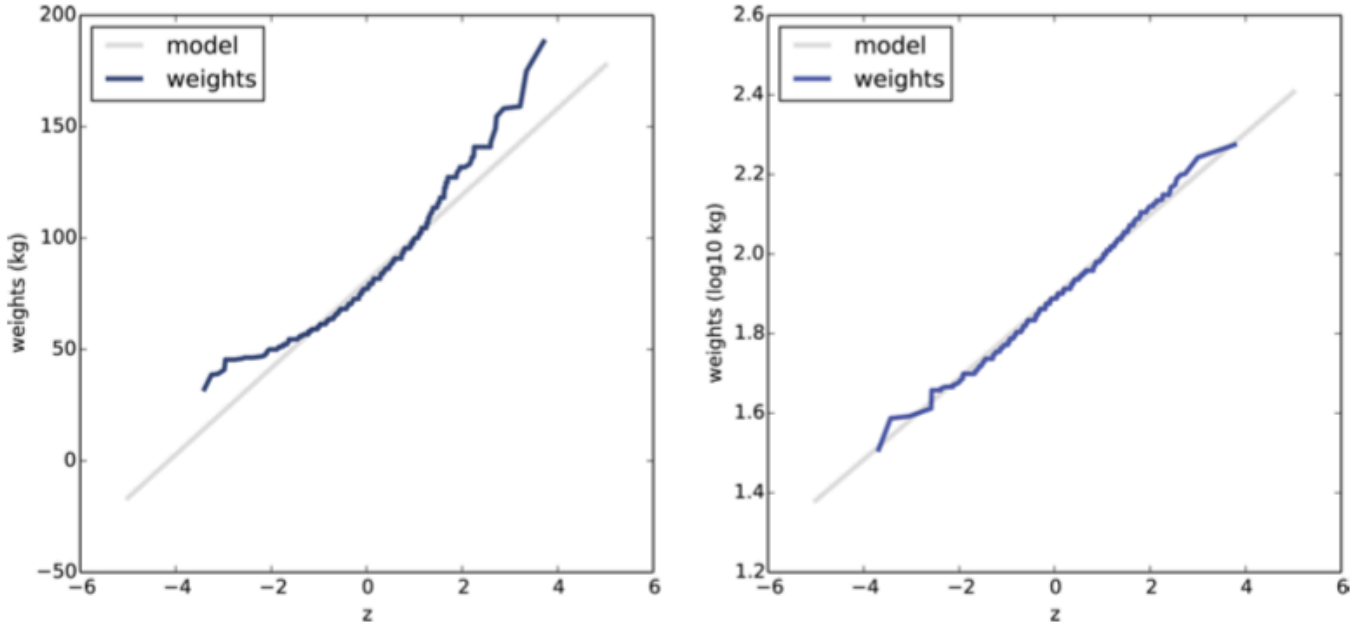


Figure 5-8. Normal probability plots for adult weight on a linear scale (left) and log scale (right).

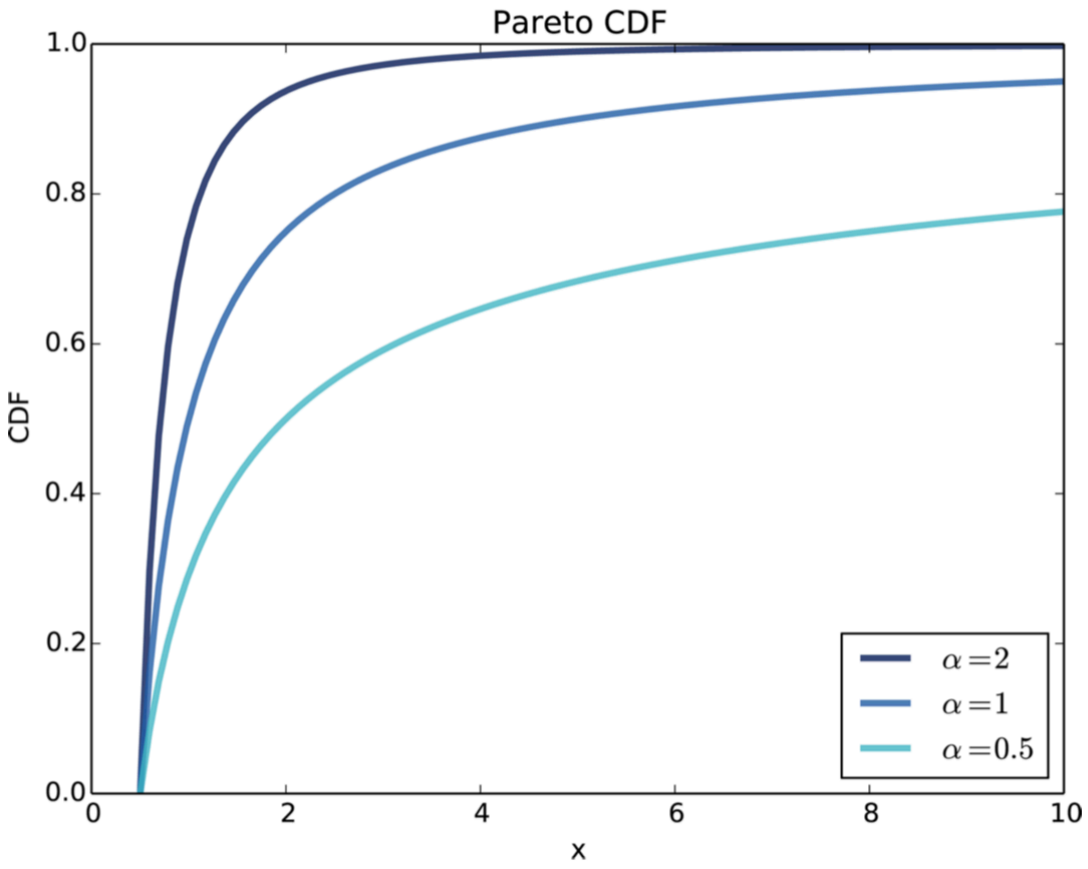
The Pareto distribution

- The Pareto distribution is named after the economist Vilfredo Pareto, who used it to describe the distribution of wealth.
- Since then, it has been used to describe phenomena in the natural and social sciences including sizes of cities and towns, sand particles and meteorites, and forest fires and earthquakes.

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

- The parameters x_m and α determine the location and shape of the distribution. **x_m is the minimum possible value**.

Pareto distribution



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

The Pareto distribution

- There is a 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.

$$y \approx \left(\frac{x}{x_m}\right)^{-\alpha} \qquad \log y \approx -\alpha(\log x - \log x_m)$$

- If you plot log y versus log x, it should look like a straight line with slope $-\alpha$ and intercept $\alpha \log x_m$.

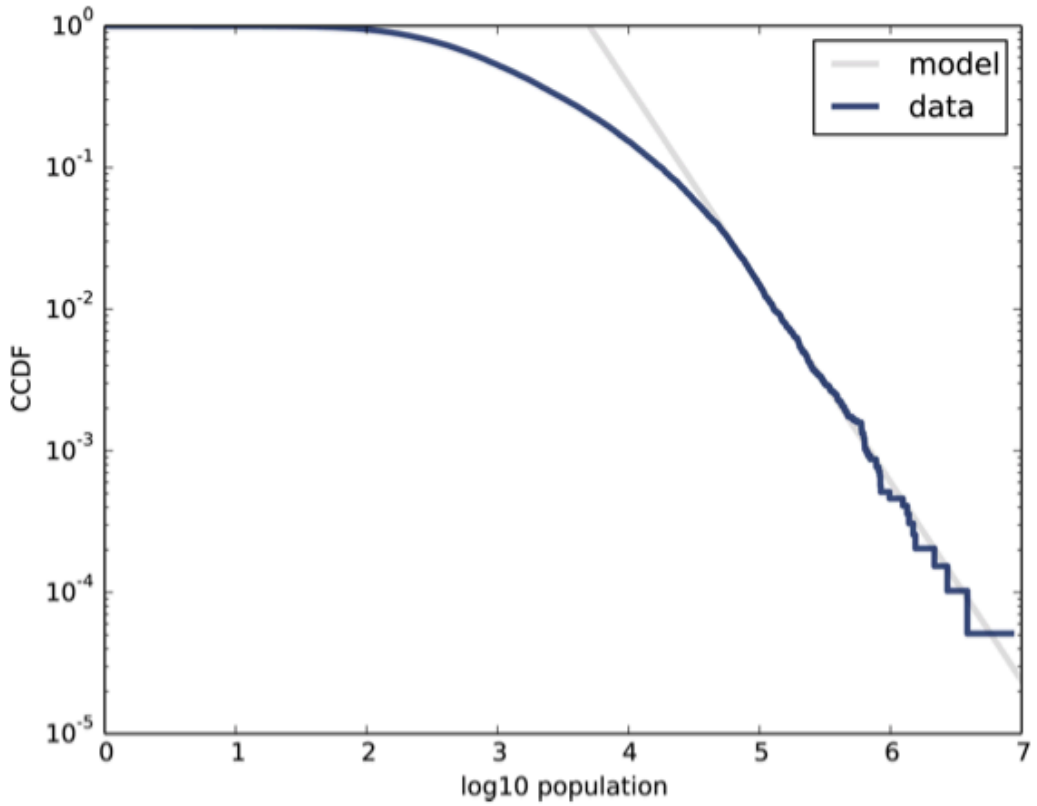
The Pareto distribution: example (sizes of cities and towns)

- The US Census Bureau publishes the population of every incorporated city and town in the United States.

- The data is in the repository for the Think Stats book in a file named **PEP_2012_PEPANNRES_with_ann.csv**.
  - The repository also contains populations.py, which reads the file and plots the distribution of populations.

The Pareto distribution: example (sizes of cities and towns)

- The CCDF of populations on a log-log scale.
  - The largest 1% of cities and towns, below 10^{-2} , fall along a straight line.
  - So we could conclude, as some researchers have, that the tail of this distribution fits a Pareto model.



The Pareto distribution: example (sizes of cities and towns)

- On the other hand, a lognormal distribution also models the data well.

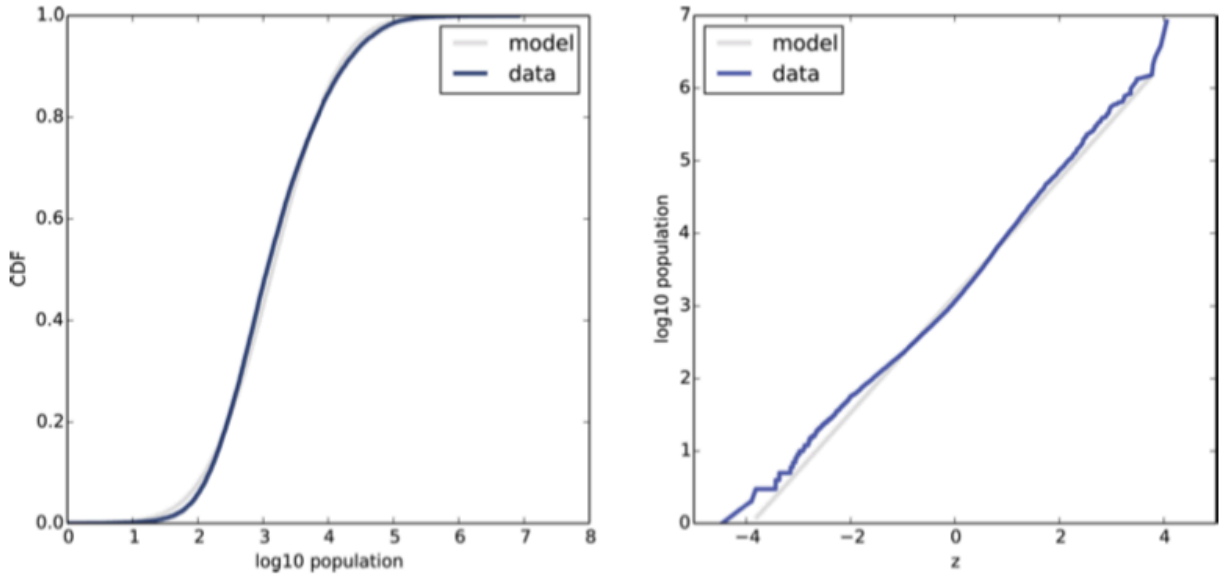


Figure 5-11. CDF of city and town populations on a log-x scale (left), and normal probability plot of log-transformed populations (right)

The Pareto distribution: example (sizes of cities and towns)

- Conclusion
  - Neither model is perfect.
    - The Pareto model only applies to the largest 1% of cities, but it is a better fit for that part of the distribution.
    - The lognormal model is a better fit for the other 99%.
  - Which model is appropriate depends on which part of the distribution is relevant.

Why model?

- Like all models, analytic distributions are abstractions, which means they leave out details that are considered irrelevant.
  - For example, an observed distribution might have measurement errors or quirks that are specific to the sample; analytic models smooth out these idiosyncrasies.
- Analytic models are also a form of data compression.
  - When a model fits a dataset well, a small set of parameters can summarize a large amount of data.
- It is sometimes surprising when data from a natural phenomenon fit an analytic distribution, but these observations can provide insight into physical systems.
- Also, analytic distributions lend themselves to mathematical analysis.

Words of wisdom

- All models are imperfect.
- Data from the real world never fit an analytic distribution perfectly.
- There are always differences between the real world and mathematical models.
- Models are useful if they capture the relevant aspects of the real world and leave out unneeded details.
- But what is “relevant” or “unneeded” depends on what you are planning to use the model for.

Chapter 6: Probability Density Functions (PDFs)

Kernel Density Estimation (KDE)

The distribution framework

Raw and central moments

Skewness and Pearson’s median skewness coefficient

PDFs (Probability Density Functions)

- The derivative of a CDF is called a probability density function, or PDF.

- Examples
  - Exponential

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

  - Normal

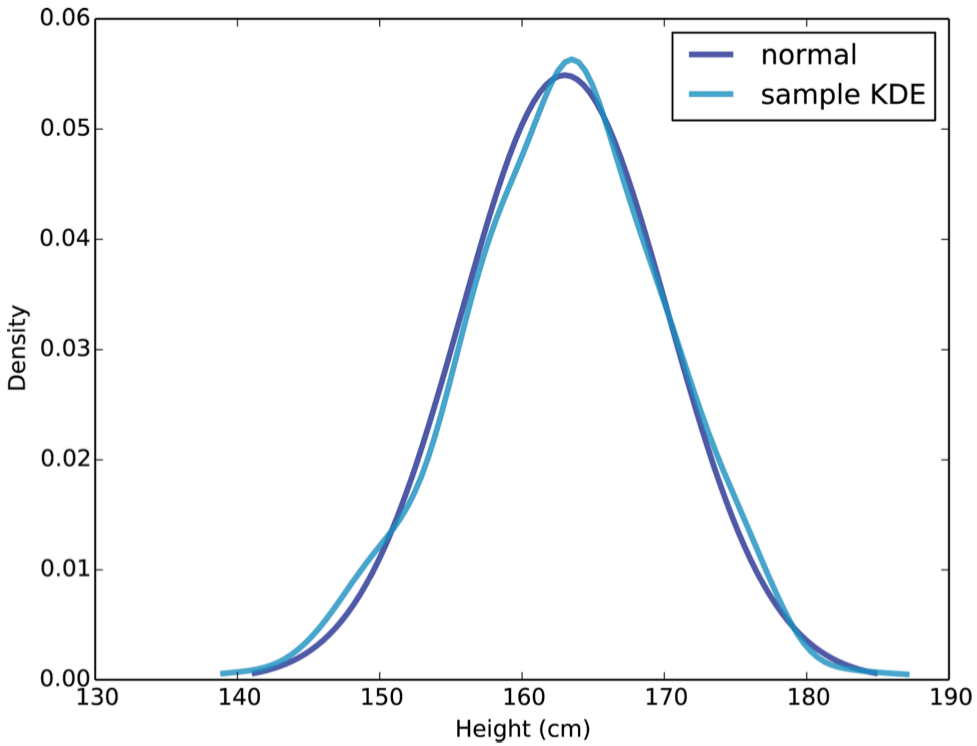
$$\mathrm{PDF}_{normal}(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left[-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2 \right]$$

Kernel Density Estimation (KDE)

- An algorithm that takes a sample and finds an appropriately smooth PDF that fits the data.

Kernel Density Estimation (KDE)

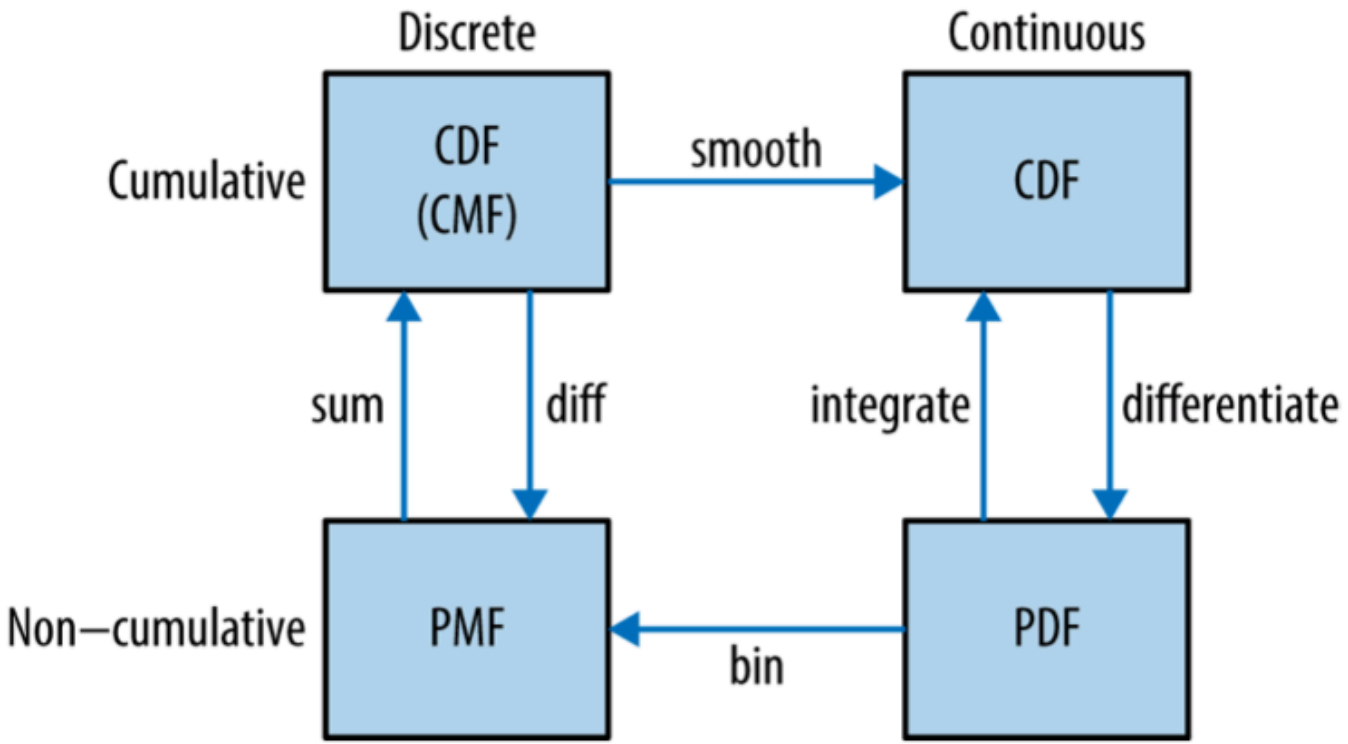
- Example:
  - Normal density function and KDE based on a sample of 500 random heights.
- The estimate is a good match for the original distribution.



Kernel Density Estimation (KDE)

- Estimating a density function with KDE is useful for several purposes:
  - Visualization
    - During the exploration phase of a project, CDFs are usually the best visualization of a distribution. After you look at a CDF, you can decide whether an estimated PDF is an appropriate model of the distribution. If so, it can be a better choice for presenting the distribution to an audience that is unfamiliar with CDFs.
  - Interpolation
    - An estimated PDF is a way to get from a sample to a model of the population. If you have reason to believe that the population distribution is smooth, you can use KDE to interpolate the density for values that don't appear in the sample.
  - 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 using KDE, which allows the simulation to explore more possible outcomes, rather than replicating the observed data.

The distribution framework



The distribution framework

- PMFs represent the probabilities for a discrete set of values.
  - To get from a PMF to a CDF, you add up the probability masses to get cumulative probabilities.
  - To get from a CDF back to a PMF, you compute differences in cumulative probabilities.
- A PDF is the derivative of a continuous CDF (i.e., a CDF is the integral of a PDF).
  - Remember that a PDF maps from values to probability densities; to get a probability, you have to integrate.
- To get from a discrete to a continuous distribution, you can perform various kinds of *smoothing*.
  - One form of smoothing is to assume that the data come from an analytic continuous distribution (like exponential or normal) and to estimate the parameters of that distribution.
  - Another option is kernel density estimation.
- The opposite of smoothing is *discretizing*, or *quantizing*.
  - If you evaluate a PDF at discrete points, you can generate a PMF that is an approximation of the PDF.
  - You can get a better approximation using numerical integration.
- To distinguish between continuous and discrete CDFs, one could call a discrete CDF “cumulative mass function” (CMF).

Moments

- Raw moment

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

Raw moments are just sums of powers.

The first raw moment is the mean.
The other raw moments don't mean much.

```
def RawMoment(xs, k):
    return sum(x**k for x in xs) / len(xs)
```

- Central moment

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

Central moments are powers of distances from the mean.

The first central moment is approximately 0.
The second central moment is the variance.

```
def CentralMoment(xs, k):
    mean = RawMoment(xs, 1)
    return sum((x - mean)**k for x in xs) / len(xs)
```

Standardized moments

- Standardized moments are ratios of central moments, with powers chosen to make the dimensions cancel.

```
def StandardizedMoment(xs, k):
    var = CentralMoment(xs, 2)
    std = np.sqrt(var)
    return CentralMoment(xs, k) / std**k
```

- The third standardized moment is skewness.

```
def Skewness(xs):
    return StandardizedMoment(xs, 3)
```

Skewness

- Skewness is a property that describes the shape of a distribution.
  - If the distribution is symmetric around its central tendency, it is *unskewed*.
  - If the values extend farther to the right, it is “right skewed”
  - If the values extend left, it is “left skewed.”

- This use of “skewed” does not have the usual connotation of “biased.”
  - Skewness only describes the shape of the distribution; it says nothing about whether the sampling process might have been biased.

Skewness

- Pearson's median skewness coefficient is a measure of skewness based on the difference between the sample mean and median

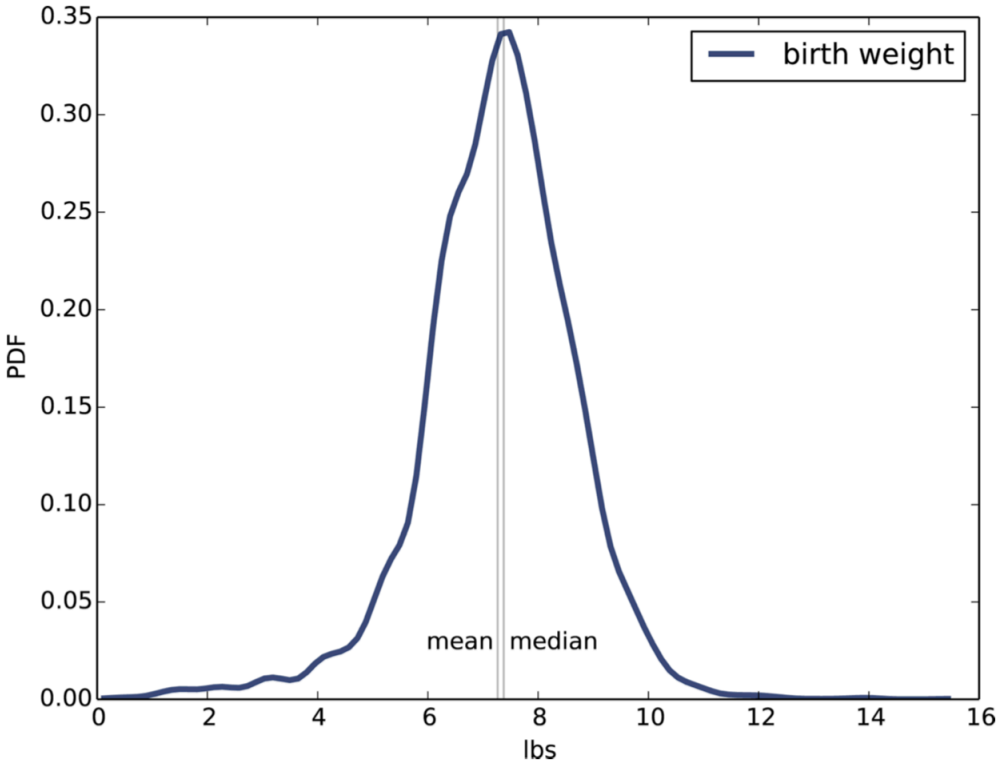
$$g_p = 3(\bar{x} - m) / S$$

Where $\bar{x}$ is the sample mean, m is the median, and S is the standard deviation.

- This statistic is robust, which means that it is less vulnerable to the effect of outliers.

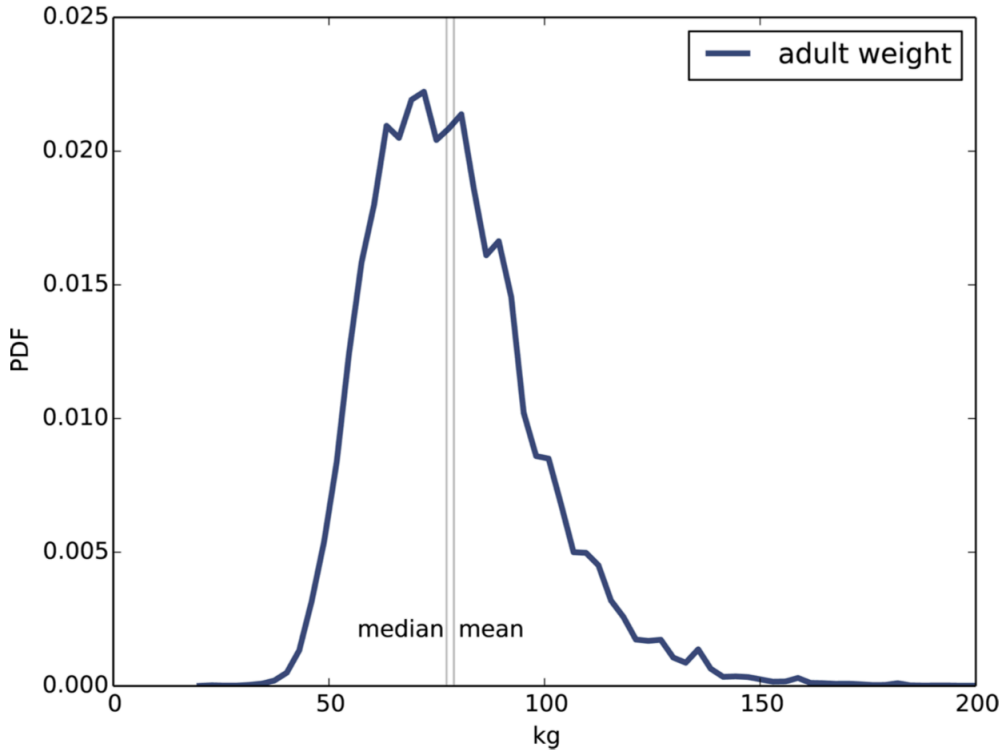
Skewness – example

- The left tail appears longer than the right, so we suspect the distribution is skewed left.
- The mean, 7.27 lbs., is a bit less than the median, 7.38 lbs., so that is consistent with left skew.
- Both skewness coefficients are negative:
  - sample skewness = -0.59
  - Pearson's median skewness = -0.23.



Skewness – example

- The distribution appears skewed to the right.
- The mean, 79.0, is bigger than the median, 77.3.
  - Sample skewness = 1.1
  - Pearson's median skewness = 0.26.



Chapter 8: Estimation

The estimation game

Guess the variance

Sampling distributions

Sampling bias

Exponential distributions

The estimation game

- Given (raw) data and (occasionally) additional hints, you should be able to estimate the data distribution (and its parameters).
- Which estimator is best?
  - It depends on the circumstances (for example, whether there are outliers) and on what the goal is.
  - Are you trying to minimize errors, or maximize your chance of getting the right answer?

The estimation game

- Example 1

  - I think of a distribution, and you have to guess what it is.
  - I'll give you two hints: it's a normal distribution, and here's a random sample drawn from it:

    ```
    [-0.441, 1.774, -0.101, -1.138, 2.975, -2.138]
    ```

    What do you think is the mean parameter, μ, of this distribution?

Possible solution: Compute the sample mean as an estimate of μ.
In this example, it would be reasonable to guess μ = 0.155.

The estimation game

- Example 2 (outliers)
  - I think of a distribution, and you have to guess what it is.
  - I'll give you two hints: It's a normal distribution, and here's a sample that was collected by *an unreliable surveyor who occasionally puts the decimal point in the wrong place*.

```
[-0.441, 1.774, -0.101, -1.138, 2.975, -213.8]
```

Possible solution: Compute the sample mean as an estimate of μ.

In this case, μ = -35.12.

Alternative methods:

1. Identify and discard outliers, and then compute the sample mean of the rest.
2. Use the median as an estimator.

Mean Squared Error (MSE)

- If there are no outliers, the sample mean minimizes the mean squared error (MSE).
- If we play the game many times, and each time compute the error xbar - μ , the sample mean minimizes

$$MSE = \frac{1}{m}\Sigma\,(\bar{x} - \mu)^2$$

Where m is the number of times you play the estimation game, not to be confused with n, which is the size of the sample used to compute $\bar{x}$.

Root Mean Squared Error (RMSE)

- The root mean squared error (RMSE) can be used as a measure of how well we did.
  - The lower the RMSE, the better.

- Minimizing MSE is a nice property, but it's not always the best strategy.
  - Example: Suppose we are estimating the distribution of wind speeds at a building site. If the estimate is too high, we might overbuild the structure, increasing its cost. But if it's too low, the building might collapse.
  - Because cost as a function of error is not symmetric, minimizing MSE is not the best strategy.

Guessing the variance

- Back to Example 1…
  - I think of a distribution, and you have to guess what it is.
  - I'll give you two hints: it's a normal distribution, and here's a random sample drawn from it:

```
[-0.441, 1.774, -0.101, -1.138, 2.975, -2.138]
```

What do you think is the variance, σ^2, of my distribution?

Possible solution: Compute the sample variance. $S^2 = \frac{1}{n}\Sigma (x_i - \bar{x})^2$

Better solution: Compute the variance using an unbiased estimator. $S_{n-1}^2 = \frac{1}{n-1}\Sigma (x_i - \bar{x})^2$

Sampling distributions

- Variation in the estimate caused by random selection is called **sampling error**.
- Example: the average weight of the adult female gorillas in a wildlife preserve.
  - Having weighed 9 female gorillas, you might find xbar = 90 kg and sample standard deviation, S = 7.5 kg.
  - The sample mean is an unbiased estimator of μ, and in the long run it minimizes MSE.
  - So if you report a single estimate that summarizes the results, you would report 90 kg.

How confident should you be in this estimate?

Sampling distributions

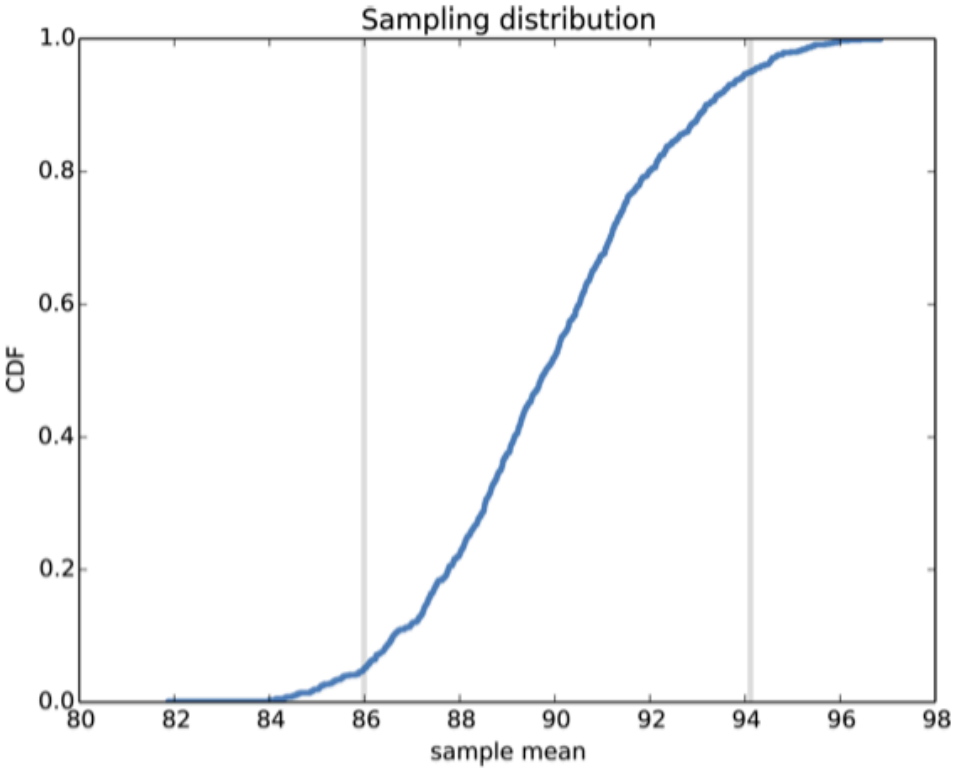
- To quantify sampling error we can simulate the sampling process with hypothetical values of μ and σ, and see how much xbar varies.
- Since we don't know the actual values of μ and σ in the population, we'll use the estimates xbar and S.
- So the question we answer is:
  - "If the actual values of μ and σ were 90 kg and 7.5 kg, and we ran the same experiment many times, how much would the estimated mean, xbar, vary?"

Sampling distributions

```
def SimulateSample(mu=90, sigma=7.5, n=9, iters=1000):
    xbars = []
    for j in range(iters):
        xs = np.random.normal(mu, sigma, n)
        xbar = np.mean(xs)
        xbars.append(xbar)
    return xbars

xbars = SimulateSample()
```

- This distribution is called the sampling distribution of the estimator. It shows how much the estimates would vary if we ran the experiment over and over.
- The mean of the sampling distribution is pretty close to the hypothetical value of μ, which means that the experiment yields the right answer, on average.
- After 1,000 tries, the lowest result is 82 kg, and the highest is 98 kg. This range suggests that the estimate might be off by as much as 8 kg.



Sampling distributions

- There are two common ways to summarize the sampling distribution:
  - Standard error (SE)
    - A measure of how far we expect the estimate to be off, on average.

      For each simulated experiment, we compute the error, xbar - μ , and then compute the root mean squared error (RMSE). In this example, it is roughly 2.5 kg.

  - Confidence interval (CI)
    - A range that includes a given fraction of the sampling distribution.

      For example, the 90% confidence interval is the range from the 5th to the 95th percentile. In this example, the 90% CI is (86, 94) kg.

Sampling distributions: caution!

- Do not confuse standard error with standard deviation!
  - Standard deviation describes variability in a measured quantity
    - In this example, the standard deviation of gorilla weight is 7.5 kg.
  - Standard error describes variability in an estimate
    - In this example, the standard error of the mean, based on a sample of 9 measurements, is 2.5 kg.
  - One way to remember the difference is that, as sample size increases, standard error gets smaller; standard deviation does not.

Sampling distributions: caution!

- People often think that there is a 90% probability that the actual parameter, μ, falls in the 90% confidence interval.
  - Sadly, that is not true.
  - If you want to make a claim like that, you have to use Bayesian methods.
- The sampling distribution answers a different question: it gives you a sense of how reliable an estimate is by telling you how much it would vary if you ran the experiment again.

Sampling distributions: caution!

- Confidence intervals and standard errors only quantify sampling error; that is, error due to measuring only part of the population.

- The sampling distribution does not account for other sources of error, notably **sampling bias** and **measurement error**.

- Example: estimating the average weight of women in the city where you live *using phone calls* may introduce **both** sampling bias and measurement error.

Exponential distributions and MLE

- Back to the guessing game…
  - I think of a distribution, and you have to guess what it is.
  - I'll give you two hints: it's an exponential distribution, and here's a random sample drawn from it:

```
[5.384, 4.493, 19.198, 2.790, 6.122, 12.844]
```

What do you think is the parameter, λ, of this distribution?

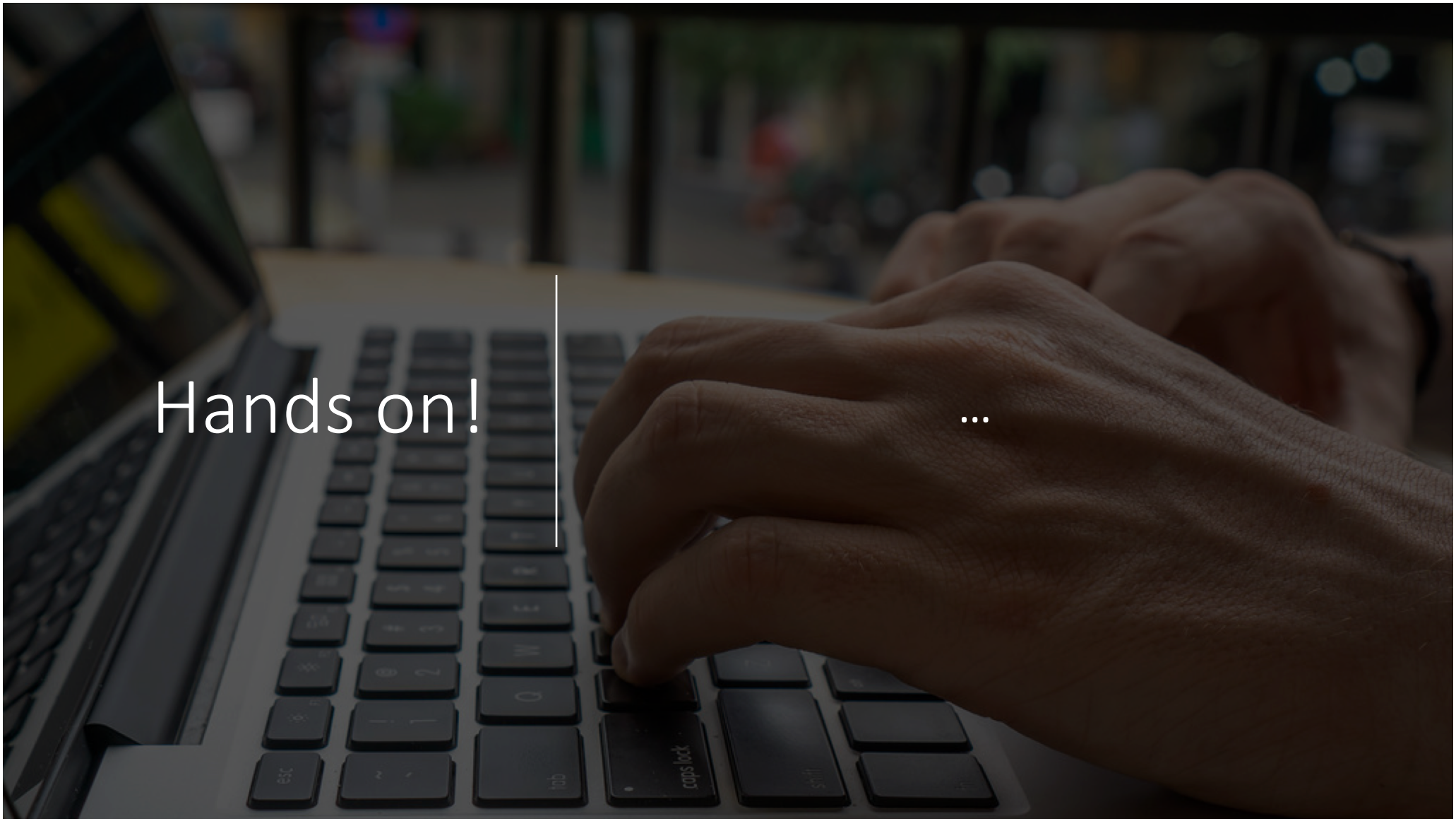
- L is an estimator of λ. $L = 1/\bar{x}$
- And not just any estimator; it is also the maximum likelihood estimator.
- If we want to maximize our chance of guessing λ exactly, L is the way to go.

Exponential distributions and MLE

- Since the sample mean (xbar) is not robust in the presence of outliers, so we expect L to have the same problem.

- We can choose an alternative based on the sample median.

$$L_m = \ln(2)/m$$


Hands on!
…

Assignment 4: Statistics – Part 2

- Goals:
  - To practice the computation and displaying of representative statistical distributions.
  - To expand upon the prior experience of manipulating, summarizing, and visualizing small datasets.
  - To compute moments and skewness measures.
  - To increase our statistical analysis skills.
  - To estimate the parameters of a distribution and propose a model that explains the underlying data.

Concluding remarks