

TODO – Go into some details here

# Chapter 1

**anecdotal evidence** Evidence, often personal, that is collected casually rather than by a well-designed study.

**population** A group we are interested in studying. “Population” often refers to a group of people, but the term is used for other subjects, too.

**cross-sectional study** A study that collects data about a population at a particular point in time.

**cycle** In a repeated cross-sectional study, each repetition of the study is called a cycle.

**longitudinal study** A study that follows a population over time, collecting data from the same group repeatedly.

**record** In a dataset, a collection of information about a single person or other subject.

**respondent** A person who responds to a survey. sample The subset of a population used to collect data. **representative** A sample is representative if every member of the population has the same chance of being in the sample.

**oversampling** The technique of increasing the representation of a subpopulation in order to avoid errors due to small sample sizes.

**raw data** Values collected and recorded with little or no checking, calculation or interpreta‐ tion.

**recode** A value that is generated by calculation and other logic applied to raw data.

**data cleaning** Processes that include validating data, identifying errors, translating between data types and representations, etc.

# Chapter 2

**distribution**  
The values that appear in a sample and the frequency of each.  
**histogram**  
A mapping from values to frequencies, or a graph that shows this mapping.  
**frequency**  
The number of times a value appears in a sample.  
**mode**  
The most frequent value in a sample, or one of the most frequent values.  
**normal distribution**  
An idealization of a bell-shaped distribution; also known as a Gaussian distribution.  
**uniform distribution**  
A distribution in which all values have the same frequency.  
**tail**  
The part of a distribution at the high and low extremes.  
**central tendency**  
A characteristic of a sample or population; intuitively, it is an average or typical value.  
**outlier**  
A value far from the central tendency.  
**spread**  
A measure of how spread out the values in a distribution are.  
**summary statistic**  
A statistic that quantifies some aspect of a distribution, like central tendency or spread.  
**variance**  
A summary statistic often used to quantify spread.  
**standard deviation**  
The square root of variance, also used as a measure of spread.  
**effect size**  
A summary statistic intended to quantify the size of an effect like a difference be‐tween groups.  
**clinically significant**  
A result, like a difference between groups, that is relevant in practice.

# Chapter 3

**Probability mass function (PMF)**  
a representation of a distribution as a function that maps from values to probabil‐ities.  
**probability**  
A frequency expressed as a fraction of the sample size.  
**normalization**  
The process of dividing a frequency by a sample size to get a probability.  
**index**  
In a pandas DataFrame, the index is a special column that contains the row labels.

# Chapter 4

**percentile rank**  
The percentage of values in a distribution that are less than or equal to a given value.  
**percentile**  
The value associated with a given percentile rank.  
**cumulative distribution function (CDF)**  
A function that maps from values to their cumulative probabilities. CDF(x) is the fraction of the sample less than or equal to x.  
**inverse CDF**  
A function that maps from a cumulative probability, p, to the corresponding value.  
**median**  
The 50th percentile, often used as a measure of central tendency.  
**interquartile range**  
The difference between the 75th and 25th percentiles, used as a measure of spread.  
**quantile**  
A sequence of values that correspond to equally spaced percentile ranks; for ex‐ample, the quartiles of a distribution are the 25th, 50th and 75th percentiles.  
**replacement**  
A property of a sampling process. “With replacement” means that the same value can be chosen more than once; “without replacement” means that once a value is

# Chapter 5

**empirical distribution**  
The distribution of values in a sample.  
**analytic distribution**  
A distribution whose CDF is an analytic function.  
**model**  
A useful simplification. Analytic distributions are often good models of more com‐plex empirical distributions.  
**interarrival time**  
The elapsed time between two events.  
**complementary CDF**  
A function that maps from a value, x, to the fraction of values that exceed x, which  
is 1 - CDF(x).  
**standard normal distribution**  
The normal distribution with mean 0 and standard deviation 1.  
**normal probability plot**  
A plot of the values in a sample versus random values from a standard normal  
distribution.

# Chapter 6

**Probability density function (PDF)**  
The derivative of a continuous CDF, a function that maps a value to its probability  
density.  
**Probability density**A quantity that can be integrated over a range of values to yield a probability. If the  
values are in units of cm, for example, probability density is in units of probability  
per cm.  
**Kernel density estimation (KDE)**An algorithm that estimates a PDF based on a sample.  
**discretize**To approximate a continuous function or distribution with a discrete function. The  
opposite of smoothing.  
**raw moment**A statistic based on the sum of data raised to a power.  
**central moment**A statistic based on deviation from the mean, raised to a power.  
**standardized moment**A ratio of moments that has no units.  
**skewness**A measure of how asymmetric a distribution is.  
**sample skewness**A moment-based statistic intended to quantify the skewness of a distribution.  
**Pearson’s median skewness coefficient**A statistic intended to quantify the skewness of a distribution based on the median,  
mean, and standard deviation.  
**robust**A statistic is robust if it is relatively immune to the effect of outliers.

# Chapter 7

**scatter plot**A visualization of the relationship between two variables, showing one point for  
each row of data.  
**jitter**Random noise added to data for purposes of visualization.  
**saturation**Loss of information when multiple points are plotted on top of each other.  
**correlation**A statistic that measures the strength of the relationship between two variables.  
**standardize**To transform a set of values so that their mean is 0 and their variance is 1.  
**standard score**A value that has been standardized so that it is expressed in standard deviations  
from the mean.  
**covariance**A measure of the tendency of two variables to vary together.  
**rank**The index where an element appears in a sorted list.  
**randomized controlled trial**An experimental design in which subjects are divided into groups at random, and  
different groups are given different treatments.  
**treatment group**A group in a controlled trial that receives some kind of intervention.  
**control group**A group in a controlled trial that receives no treatment, or a treatment whose effect  
is known.  
**natural experiment**An experimental design that takes advantage of a natural division of subjects into  
groups in ways that are at least approximately random

# Chapter 8

**estimation**The process of inferring the parameters of a distribution from a sample.  
**estimator**A statistic used to estimate a parameter.  
**mean squared error (MSE)**A measure of estimation error.  
**root mean squared error (RMSE)**The square root of MSE, a more meaningful representation of typical error mag‐  
nitude.  
**maximum likelihood estimator (MLE)**An estimator that computes the point estimate most likely to be correct.  
**bias (of an estimator)**The tendency of an estimator to be above or below the actual value of the parameter,  
when averaged over repeated experiments.  
**sampling error**Error in an estimate due to the limited size of the sample and variation due to chance.  
**sampling bias**Error in an estimate due to a sampling process that is not representative of the  
population.  
**measurement error**Error in an estimate due to inaccuracy collecting or recording data.  
**sampling distribution**The distribution of a statistic if an experiment is repeated many times.  
**standard error**The RMSE of an estimate, which quantifies variability due to sampling error (but  
not other sources of error).  
**confidence interval**An interval that represents the expected range of an estimator if an experiment is  
repeated many times.

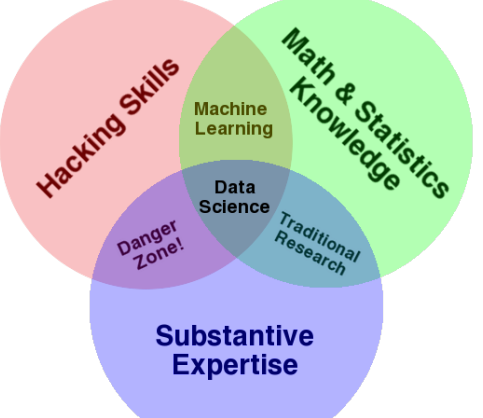
# Chapter 9

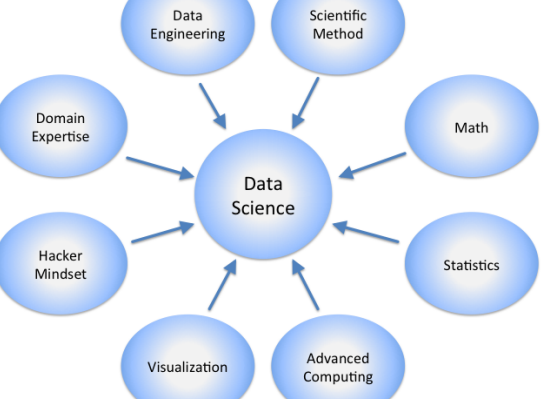
**hypothesis testing**The process of determining whether an apparent effect is statistically significant.  
**test statistic**A statistic used to quantify an effect size.  
**null hypothesis**A model of a system based on the assumption that an apparent effect is due to  
chance.  
**p-value**The probability that an effect could occur by chance.  
**statistically significant**An effect is statistically significant if it is unlikely to occur by chance.  
**permutation test**A way to compute p-values by generating permutations of an observed dataset.  
**resampling test**A way to compute p-values by generating samples, with replacement, from an observed dataset.  
**two-sided test**A test that asks, “What is the chance of an effect as big as the observed effect, positive  
or negative?”  
**one-sided test**A test that asks, “What is the chance of an effect as big as the observed effect, and  
with the same sign?”  
**chi-squared test**A test that uses the chi-squared statistic as the test statistic.  
**false positive**The conclusion that an effect is real when it is not.  
**false negative**The conclusion that an effect is due to chance when it is not.  
**power**The probability of a positive test if the null hypothesis is false.

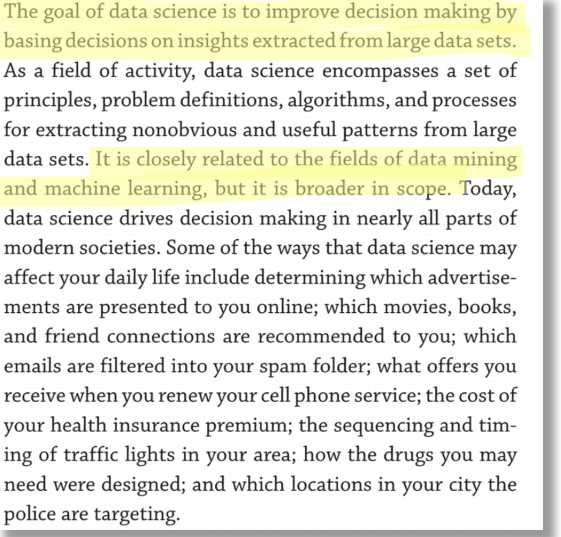
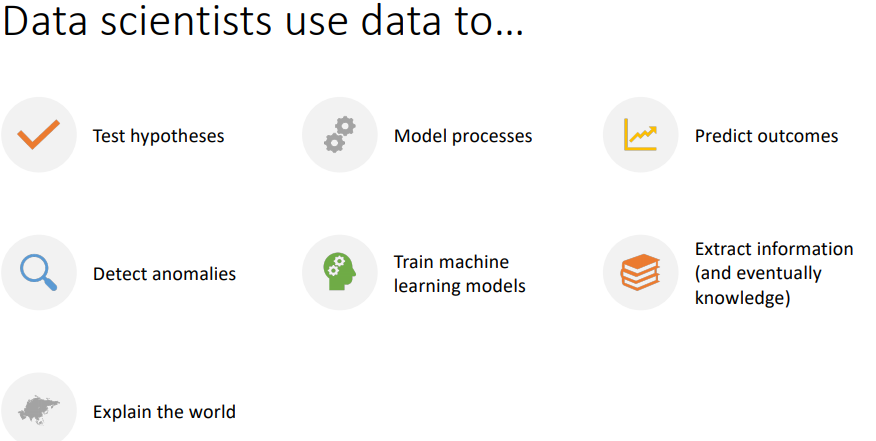
# Slide Notes

## Data Science Big Picture

What is data science - “Data science […] is perhaps the best label we have for the cross-disciplinary set of skills that are becoming increasingly important in many applications across industry and academia.”







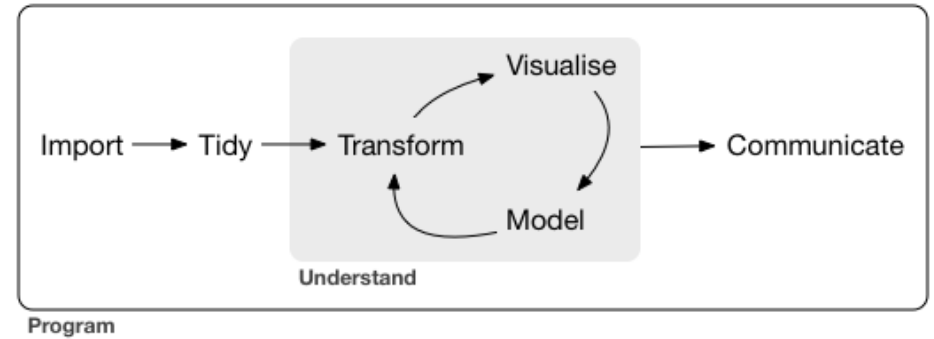
What is the recipe for success in data science?

Math/Statistics knowledge, hacking/coding skills, subject-mater expertise

What is driving Data science

* Enormous availability of raw data
* Open source tools
* Easy access to code and datasets
* Numerous use cases
* Faster/ubiquitous computing platforms
* Lower barriers to enter

Data Science workflow



• Import: Take data stored in a file, database, or web API, and load it into a data frame.

• Tidy: Store it in a consistent form that matches the semantics of the dataset with the way it is stored. When your data is tidy, each column is a variable, and each row is an observation.

• Transform: Narrow in on observations of interest, create new variables that are functions of existing variables, and calculate a set of summary statistics.

• Tidying + Transforming = Data Wrangling

• Visualize: Fundamental step for human analysis.

• Model: Build the ability to make predictions and inferences based on the data.

• Communicate: Share results, get feedback, promote reproducibility.

• Programming: The “glue” that keeps everything together.

## Exploratory Data Analysis (EDA) in context

[…] “The process I use when I start working with a dataset”:

• Importing and cleaning the data

• Single variable explorations: start by examining one variable at a time, finding out what the variables mean, looking at distributions of the values, and choosing appropriate summary statistics.

• Pair-wise explorations: to identify possible relationships between variables, look at tables and scatter plots, and compute correlations and linear fits.

• Multivariate analysis: if there are apparent relationships between variables, use multiple regression to add control variables and investigate more complex relationships.

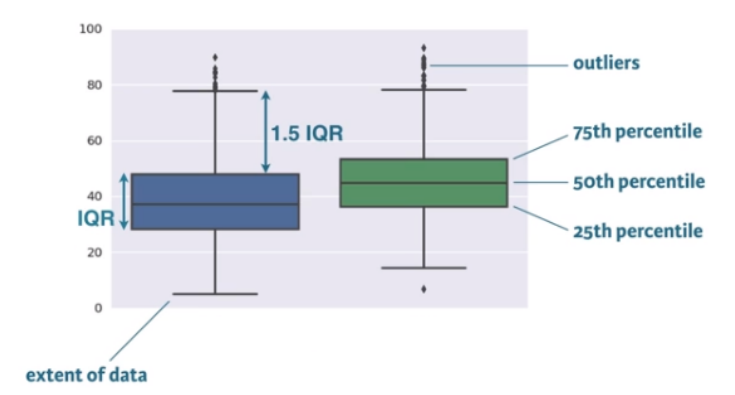
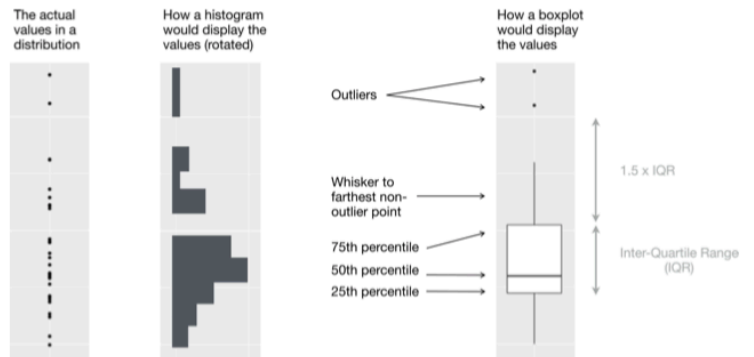
• Estimation and hypothesis testing: When reporting statistical results, it is important to answer three questions: How big is the effect? How much variability should we expect if we run the same measurement again? Is it possible that the apparent effect is due to chance?

• Visualization: During exploration, visualization is an important tool for finding possible relationships and effects. Then if an apparent effect holds up to scrutiny, visualization is an effective way to communicate results.

### Percentile based statistics

* Median = 50th percentile
  + A measure of central tendency of the distribution
* Interquartile Range (IQR) = the difference between the 75th and 25th percentiles
  + A measure of the spread of the distribution

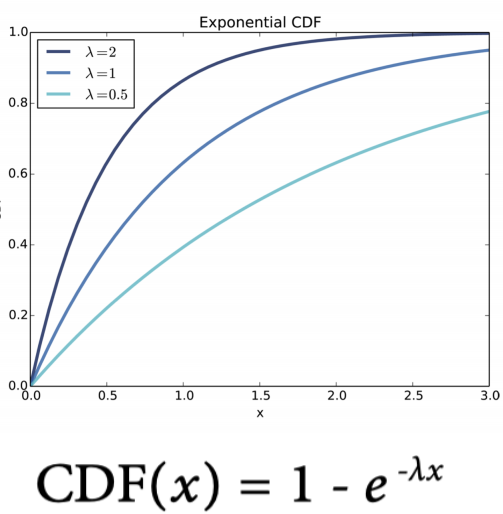
### Box and whisker plots



## Transition: from empirical to analytic distributions

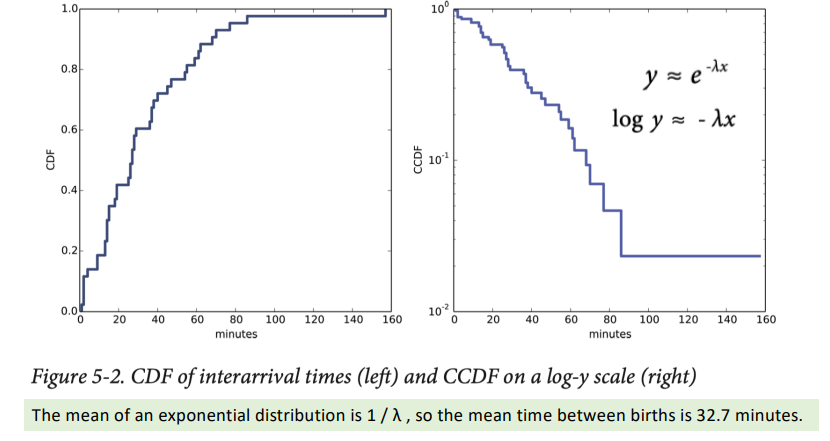
* **Empirical distributions** are based on *empirical observations*, which are necessarily finite samples.
* The alternative is an **analytic distribution**, which is characterized by a CDF that is a *mathematical function*.
* **Analytic distributions can be used to model empirical distributions**.
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

## Exponential distribution

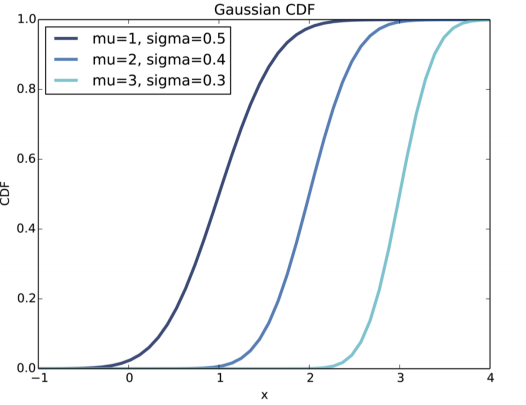


* In the real world, exponential distributions come up when we look at a series of events and measure the times between events, called interarrival times.
* If the events are equally likely to occur at any time, the distribution of interarrival times tends to look like an exponential distribution.
* 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)

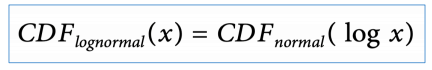


## Normal (Gaussian) distribution

* 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.
* The normal distribution is characterized by two parameters: the mean, μ, and standard deviation σ.
  + The normal distribution with μ = 0 and σ = 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 CD

## The lognormal distribution

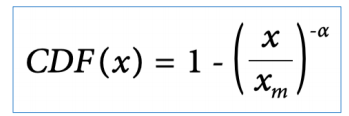
* If the logarithms of a set of values have a normal distribution, the values have a lognormal distribution.
* The CDF of the lognormal distribution is the same as the CDF of the normal distribution, with log x substituted for x.



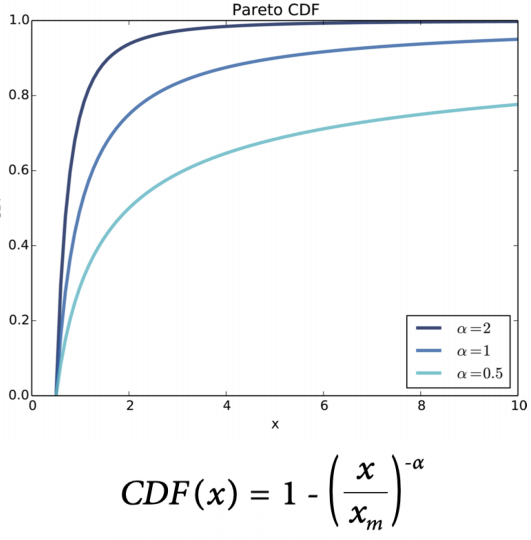
* The parameters of the lognormal distribution are usually denoted μ and σ.
  + NB: these parameters are not the mean and standard deviation!
* If a sample is approximately lognormal and you plot its CDF on a log-x scale, it will have the characteristic shape of a normal distribution.
  + To test how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

## The Pareto distribution

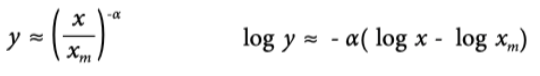
* The Pareto distribution is named after the economist Vilfredo Pareto, who used it to describe the distribution of wealth.
* Since then, it has been used to describe phenomena in the natural and social sciences including sizes of cities and towns, sand particles and meteorites, and forest fires and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



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



* 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.



* If you plot log y versus log x, it should look like a straight line with slope -α and intercept α log xm

## Why Model?

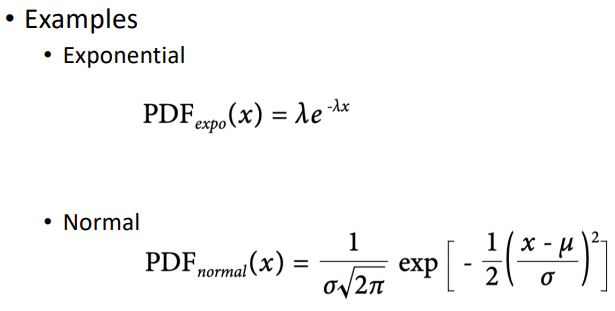
* Like all models, analytic distributions are abstractions, which means they leave out details that are considered irrelevant.
  + For example, an observed distribution might have measurement errors or quirks that are specific to the sample; analytic models smooth out these idiosyncrasies.
* Analytic 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

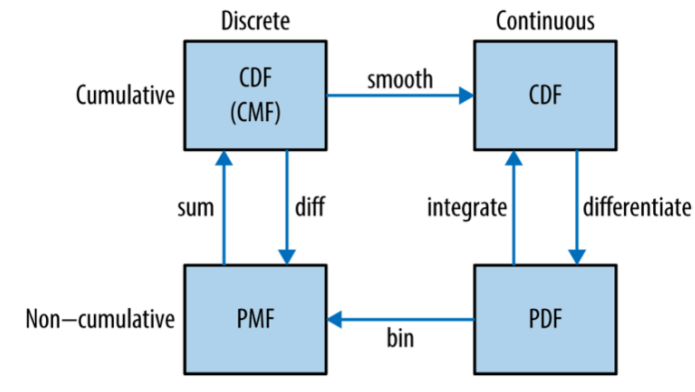
* All models are imperfect.
* Data from the real world never fit an analytic distribution perfectly.
* There are always differences between the real world and mathematical models.
* Models are useful if they capture the relevant aspects of the real world and leave out unneeded details.
* But what is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

## PDFs (Probability Density Functions)

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



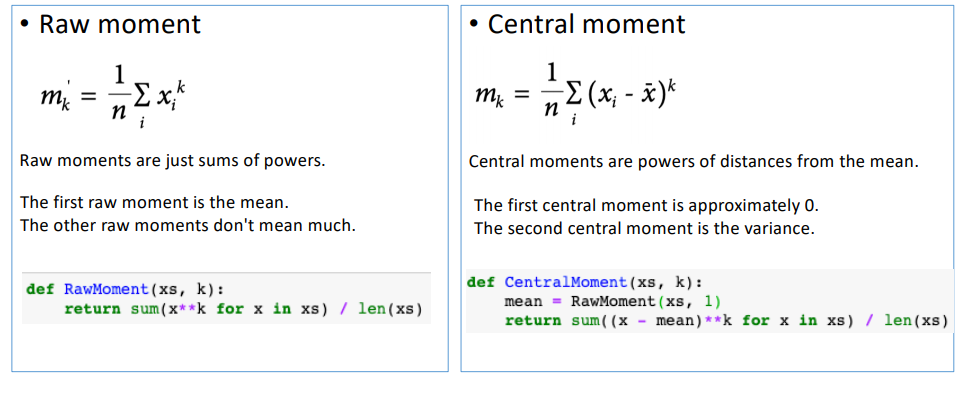
## The distribution framework



* 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).

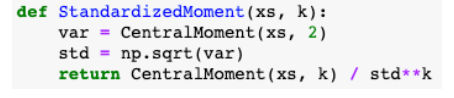
## Moments

Raw moment, central moment



## Standardized moments

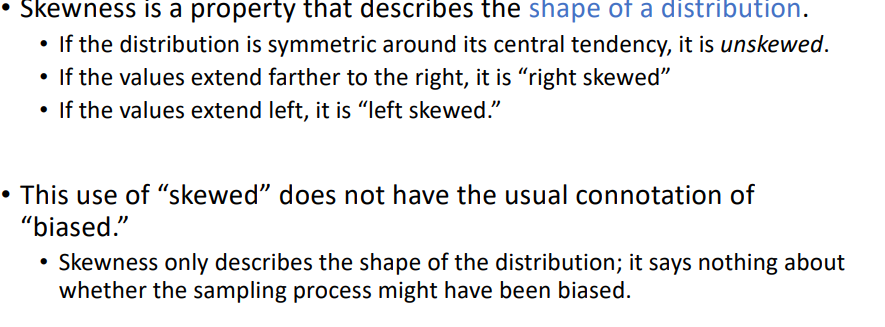
* Standardized moments are ratios of central moments, with powers chosen to make the dimensions cancel.

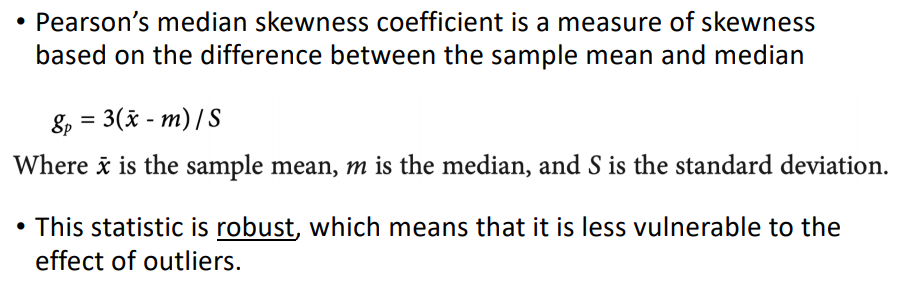


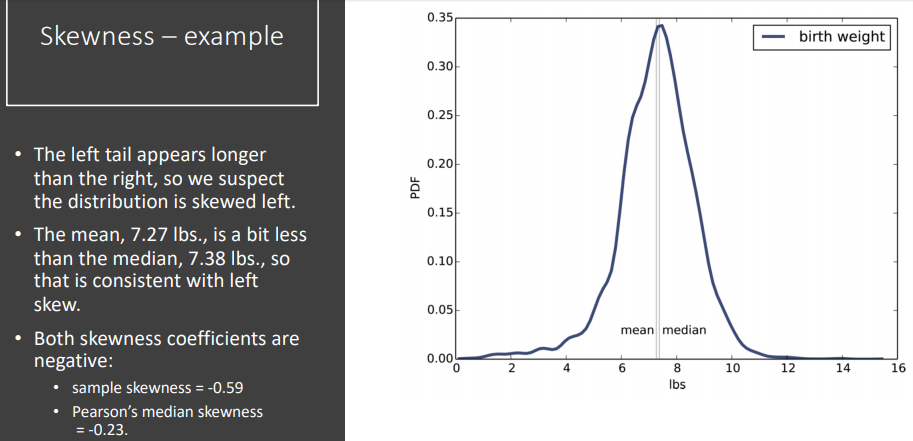
* The third standardized moment is skewness.

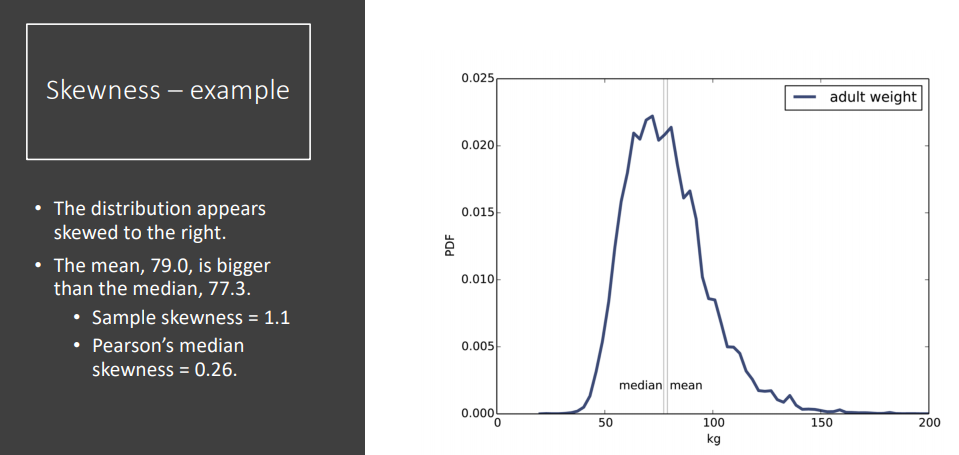


## Skewness



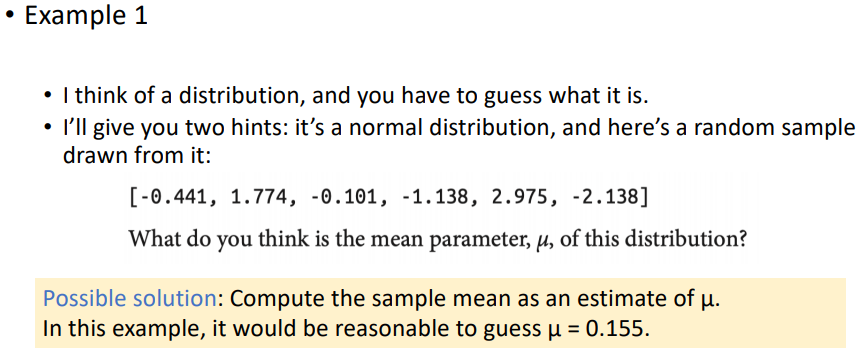


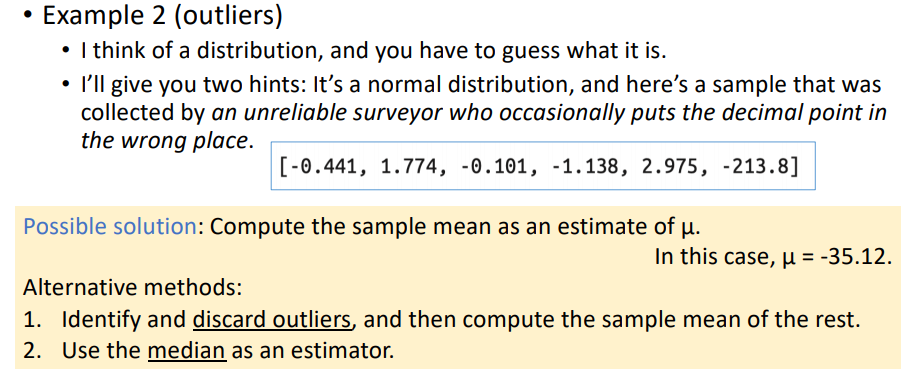




## The estimation game

* 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?

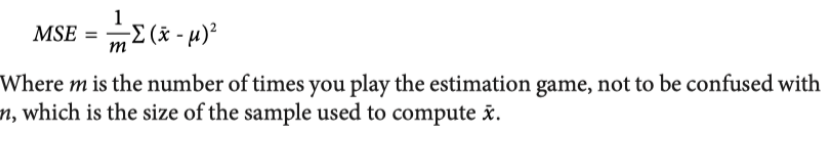




## Mean Squared Error (MSE)

If there are no outliers, the sample mean minimizes the mean squared error (MSE).

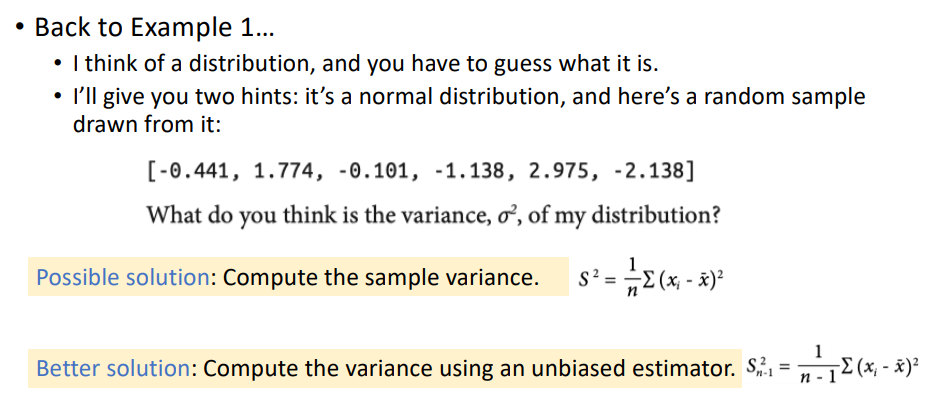
• If we play the game many times, and each time compute the error xbar - μ , the sample mean minimizes



## 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

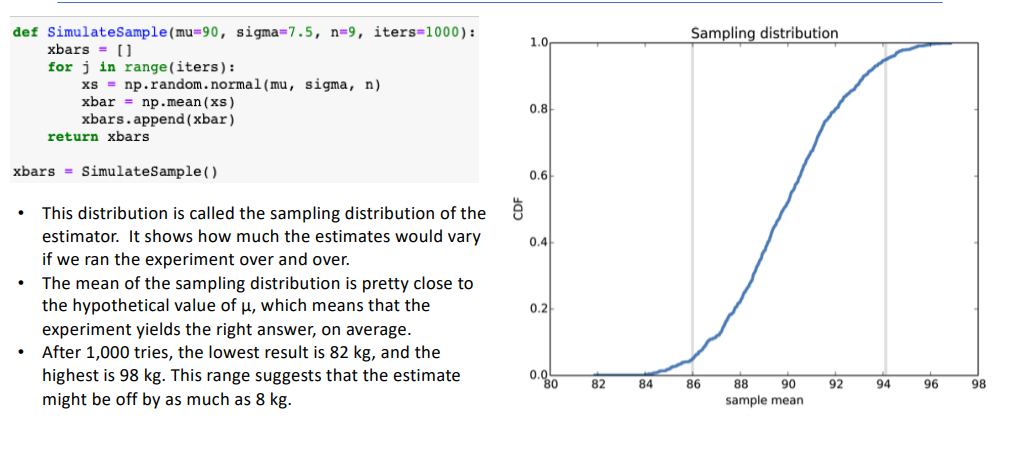


## Sampling distributions

* 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.

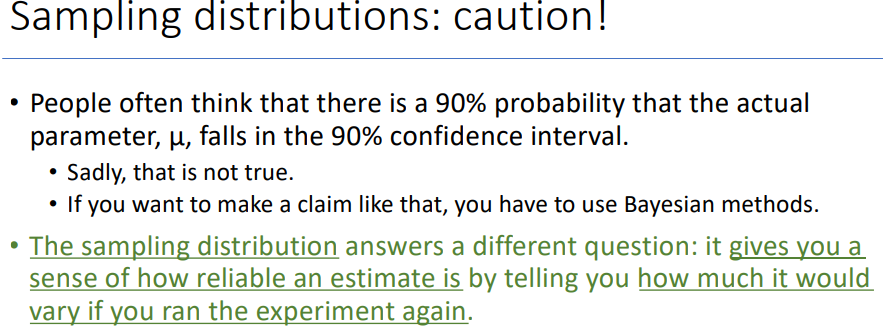
How confident should you be in this estimate?

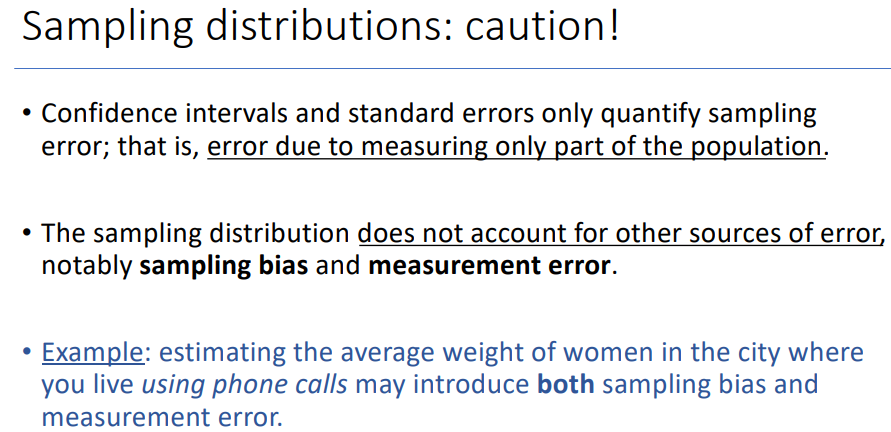
* To quantify sampling error we can simulate the sampling process with hypothetical values of μ and σ, and see how much xbar varies.
  + Since we don’t know the actual values of μ and σ in the population, we’ll use the estimates xbar and S.
  + So the question we answer is:
* “If the actual values of 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?”



There are two common ways to summarize the sampling distribution:

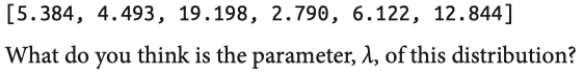
* Standard error (SE)
  + A measure of how far we expect the estimate to be off, on average.
  + For each simulated experiment, we compute the error, xbar - μ , and then compute theroot mean squared error (RMSE). In this example, it is roughly 2.5 kg.
* Confidence interval (CI)
  + A range that includes a given fraction of the sampling distribution.
  + For example, the 90% confidence interval is the range from the 5th to the 95thpercentile. In this example, the 90% CI is (86, 94) kg.
* 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.



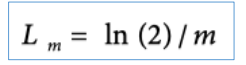


## Exponential distributions and MLE

* 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:



* + L is an estimator of λ. 
  + 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.
* 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.



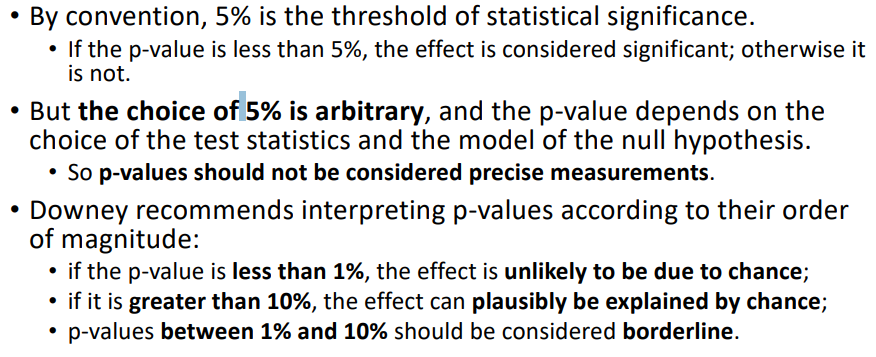
## Classical hypothesis testing

* Scope: going from “apparent effects” to “rigorous hypothesis testing”
* The fundamental question we want to address is whether the effects
* we see in a sample are likely to appear in the larger population.
* Example:
  + In the NSFG sample we see a difference in mean pregnancy length for first babies and others.
* We would like to know if that effect reflects a real difference for women in the U.S., or if it might appear in the sample by chance.
* The goal of classical hypothesis testing is to answer the question:
* “Given a sample and an apparent effect, what is the probability of seeing such an effect by chance?”
* The logic of the process is similar to a proof by contradiction.
  + To prove a mathematical statement, A, you assume temporarily that A is false.
  + If that assumption leads to a contradiction, you conclude that A must actually be true.

### Steps

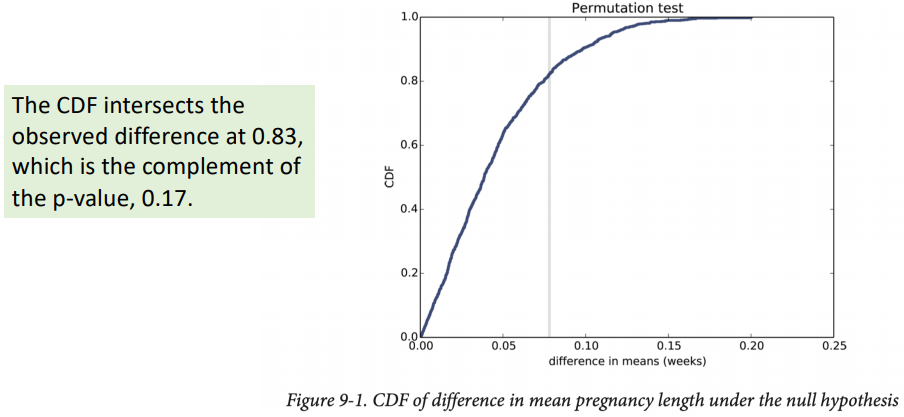
1. Quantify the size of the apparent effect by choosing a **test statistic**.
2. Define a **null hypothesis**, which is a model of the system based on the assumption that the apparent effect is not real.
3. Compute a **p-value**, which is the probability of seeing the apparent effect if the null hypothesis is true.
4. **Interpret the result.**  
   If the p-value is low, the effect is said to be **statistically significant**, which means that it is unlikely to have occurred by chance.   
   In that case we infer that the effect is more likely to appear in the larger population.

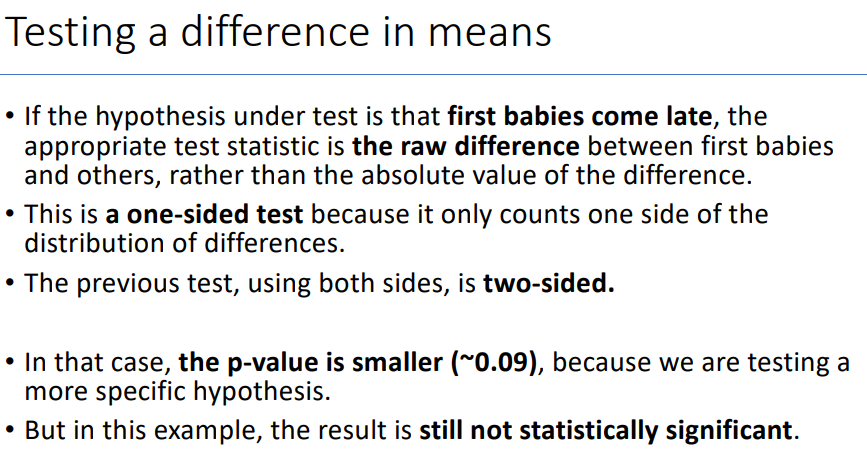
### Interpreting the result

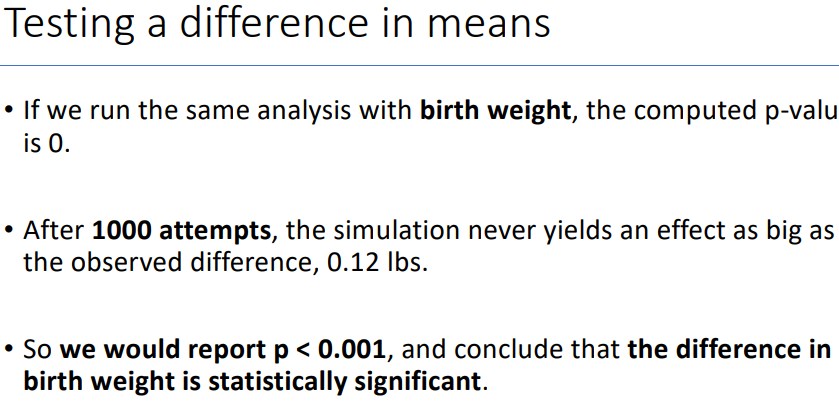


## Testing a difference in means

* Hypothesis: the pregnancy length is longer for first babies.
* Test statistic: difference in mean pregnancy length between 2 groups.
* Null hypothesis: the distributions for the two groups are the same.
* One way to model the null hypothesis is by permutation; that is, we can take values for first babies and others and shuffle them, treating the two groups as one big group.
* If you run the code, you’ll see that the computed p-value is ~0.17, which means that we expect to see a difference as big as the observed effect about 17% of the time.
* So this effect is not statistically significant.

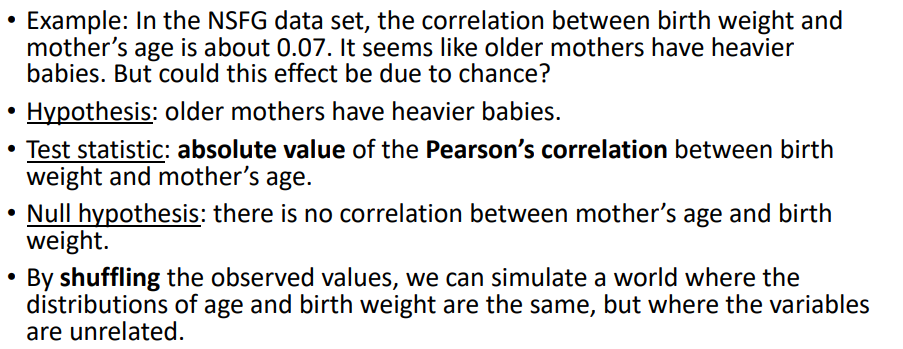


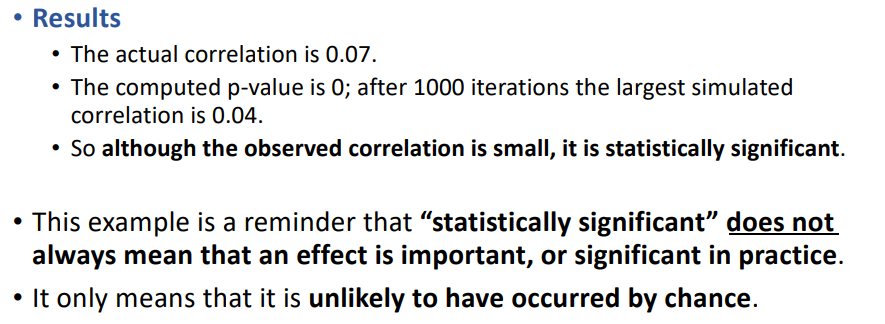




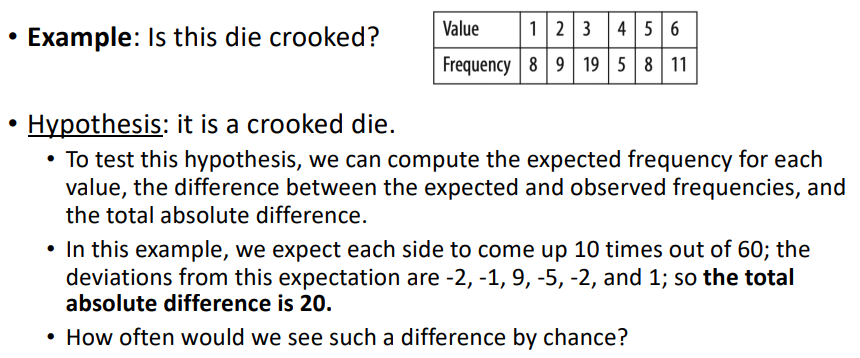
## Difference in standard deviation

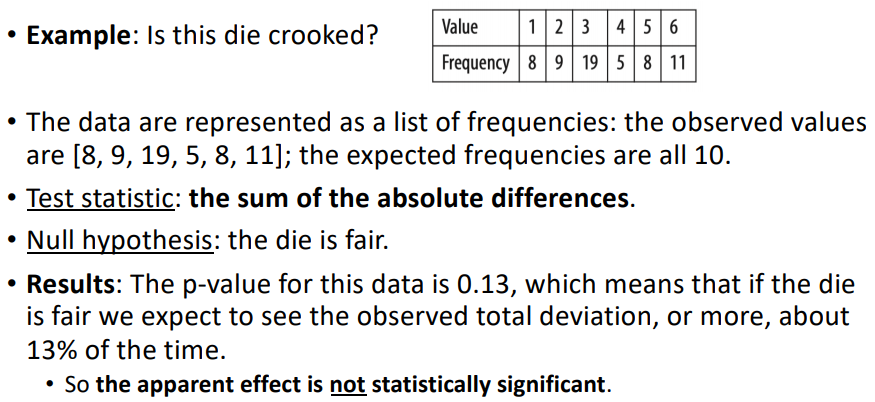
## Testing correlation





## Testing proportion





## Chi-squared tests

