Statistics: Estimation and Sampling

Logit

2016

TABLE OF CONTENTS

Intro to Estimation and Sampling

In statistics, we look at data to better understand the properties of a population (of people, animals, manufactured goods, etc.) But we can't always see every person, animal, or other element from a population.

With estimation, we try to fit the existing data to a model that hopefully explains the larger population.

With sampling, we consider the ways in which we collect a selection of data from a larger population, and we consider workarounds for having small samples.

But first, let's go over some of the fundamental summary statistics that describe the properties of any population or sample.

Summary Statistics

Summary Statistics

Summary statistics describe the most fundamental information about the distribution of a data set. Things like the mean, median, standard deviation, and inter-quartile range give us information about the approximate "center" and spread of a data set. Plotting a 1-dimensional data set with histograms, box plots, and violin plots are all part of Exploratory Data Analysis (EDA).

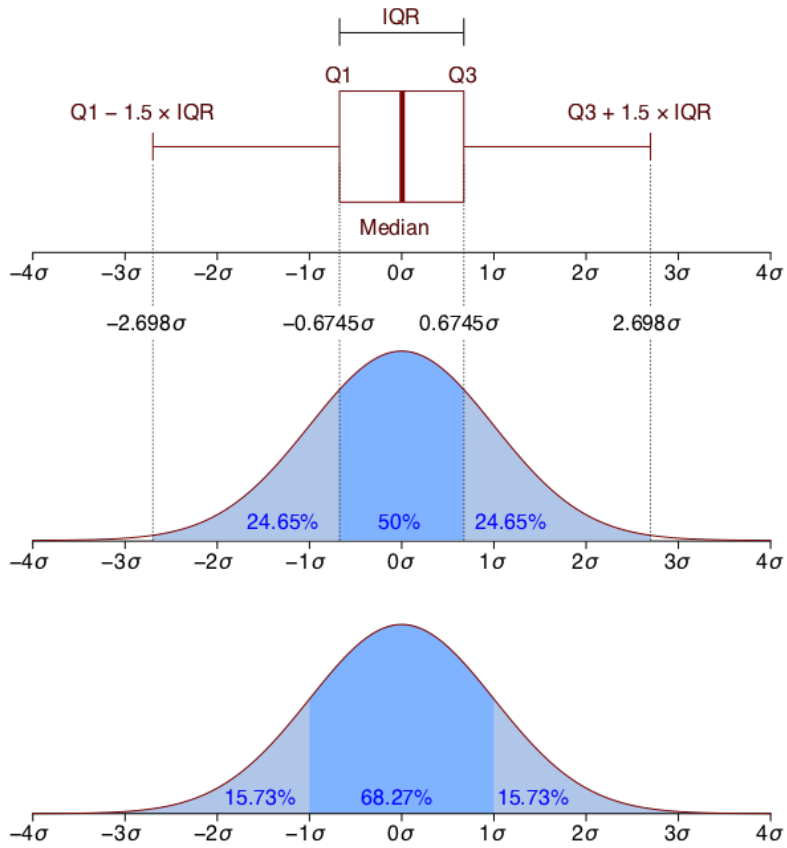

Image Source

Arithmetic Mean

The arithmetic mean or average of a data set is defined as the sum of the data divided by the number of elements in the data set. More formally, if we have a set of points $\{x_1, \ldots, x_n\}$, then the mean $\overline{x}$ is

$$\overline{x} = \frac{x_1 + \cdots + x_n}{n} = \frac{1}{n} \sum_{i=1}^{n} x_i$$

The arithmetic mean sits at the "balance point" of the data set.

Image Source

Arithmetic Mean

What is the arithmetic mean of the following set?

$$\{27, 83, 35, 51\}$$

Image Source

Arithmetic Mean

What is the arithmetic mean of the following set?

$$\{27, 83, 35, 51\}$$

$$\begin{aligned}\overline{x} &= \frac{27 + 83 + 35 + 51}{4} \\ &= 49\end{aligned}$$

Image Source

Median

The median of a data set is the "middle" value, when the set is ordered from least to greatest. With the median, 50% of the population lies above it, and 50% lies below. More formally, if we have an ordered set

$$\{x_1, \ldots, x_n\}$$

then if n is odd, the median is $x_{\frac{n+1}{2}}$. If n is even, the median is $\frac{1}{2}\left(x_{\frac{n}{2}} + x_{\frac{n}{2}+1}\right)$.

The median is better than the mean as a measure of central tendency when dealing with more "skewed" distributions.

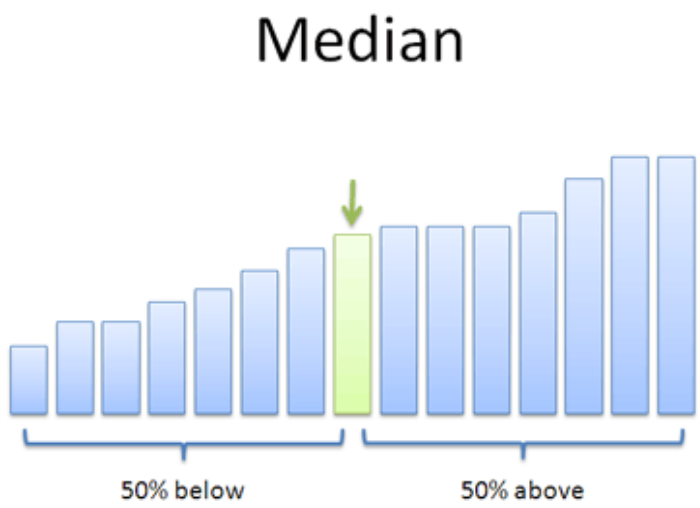

Image Source

Median

What is the median of the following set?

$$\{89, 73, 84, 91, 88, 77\}$$

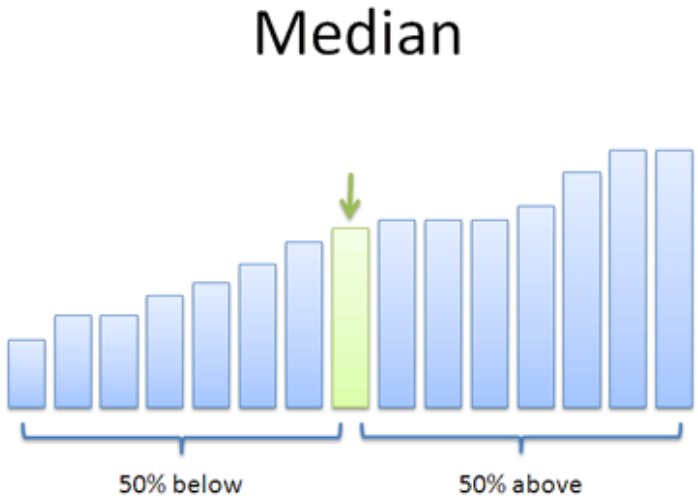

Image Source

Median

What is the median of the following set?

$$\{89, 73, 84, 91, 88, 77\}$$

First, let's put the elements in order.

$$\{73, 77, 84, 88, 89, 91\}$$

Then, we can take the average of the two middle terms.

$$\frac{84 + 88}{2} = \boxed{86}$$

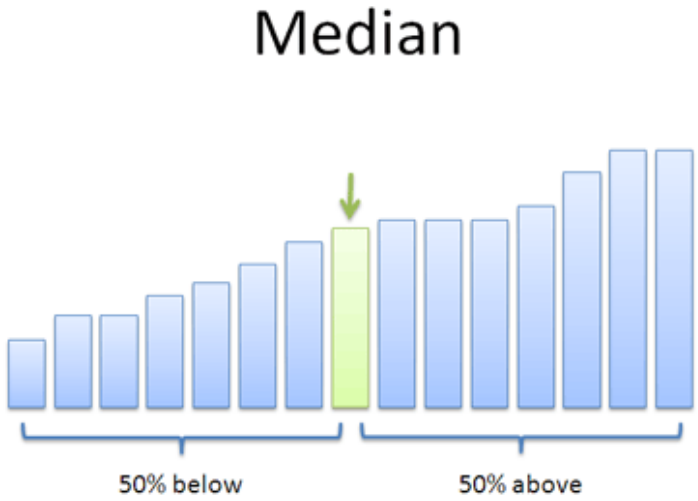

Image Source

Mode

The mode of a data set is the most common element of that set. In a symmetric distribution with one peak ("unimodal" distribution), the mode often coincides with the mean and median.

For instance, the mode of

$$\{1, 3, 6, 6, 6, 7, 7, 12, 12, 17\}$$

is 6.

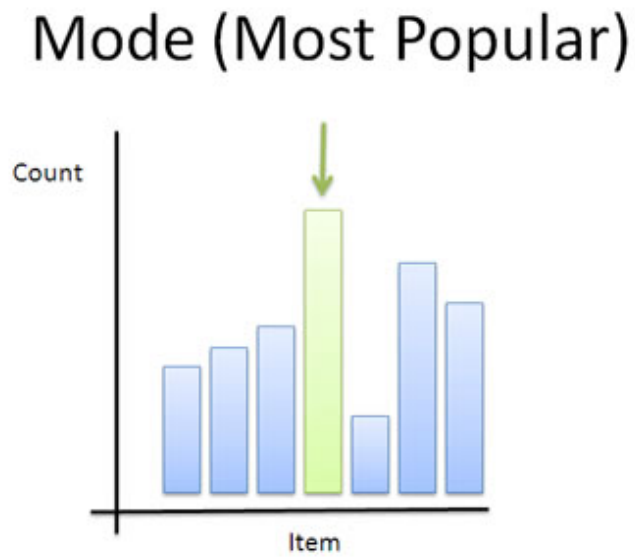

Image Source

Variance and Standard Deviation

Variance is a measure of the "spread" of a probability distribution (or the spread of a set of data points). A larger variance corresponds with a more spread-out distribution, with a greater proportion of the probability "weights" located further away from the mean. Standard deviation, another measure of spread, is just the square root of the variance.

Variance and Standard Deviation

In general, if μ is the mean (expected value) of a random variable X, the variance is defined as

$$\text{Var}(X) = \mathbb{E}[(X - \mu)^2] = \mathbb{E}[X^2] - (\mathbb{E}[X])^2$$

For a discrete collection of data points X, this can also be defined as

$$\text{Var}(X) = \sigma^2(X) = \sum_{i=1}^{n} \frac{1}{n}(x_i - \mu)^2$$

When dealing with a sample of data, the unbiased sample variance s^2 is

$$s^2(X) = \sum_{i=1}^{n} \frac{1}{n-1}(x_i - \overline{x})^2$$

We define s^2 this way because the estimate of the population variance is biased by a factor of $\frac{n-1}{n}$.

Variance and Standard Deviation

What are the unbiased sample variance and standard deviation for the following set?

$$\{3, 5, 7, 9, 11\}$$

Variance and Standard Deviation

What are the unbiased sample variance and standard deviation for the following set?

$$\{3, 5, 7, 9, 11\}$$

Calculate the sample mean:

$$\overline{x} = \frac{3+5+7+9+11}{5} = 7$$

Unbiased sample variance:

$$\begin{aligned} s^2 &= \sum_{i=1}^{n} \frac{1}{n-1}(x_i - \overline{x})^2 \\ &= \frac{1}{4}\left[(3-7)^2 + (5-7)^2 + (7-7)^2 + (9-7)^2 + (11-7)^2\right] \\ &= 10 \end{aligned}$$

Unbiased standard deviation:

$$s = \sqrt{10}$$

68 - 95 - 99.7 Rule

For normal / Gaussian distributions with mean μ and standard deviation σ, the 68 - 95 - 99.7 rule can help us remember that

$$68\% \text{ of data} \leftarrow \text{within } 1\sigma \text{ of } \mu$$
$$95\% \text{ of data} \leftarrow \text{within } 2\sigma \text{ of } \mu$$
$$99.7\% \text{ of data} \leftarrow \text{within } 3\sigma \text{ of } \mu$$

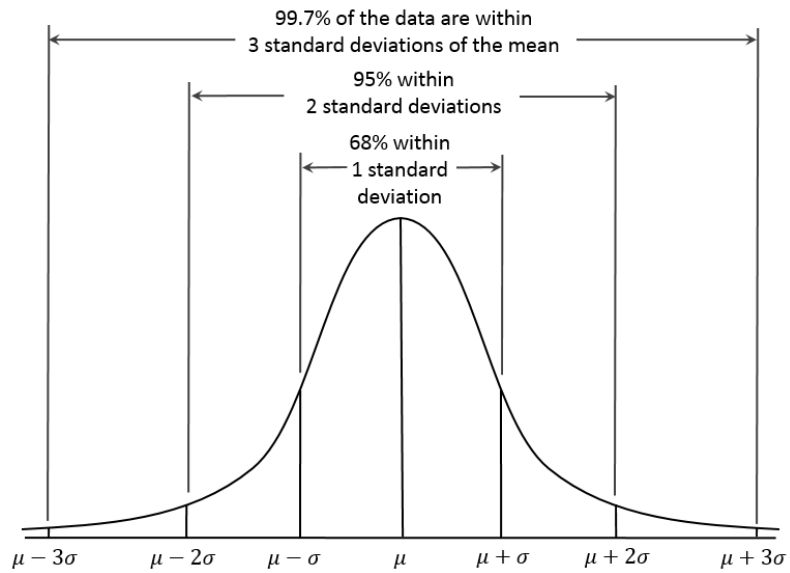

Image Source

Quartiles

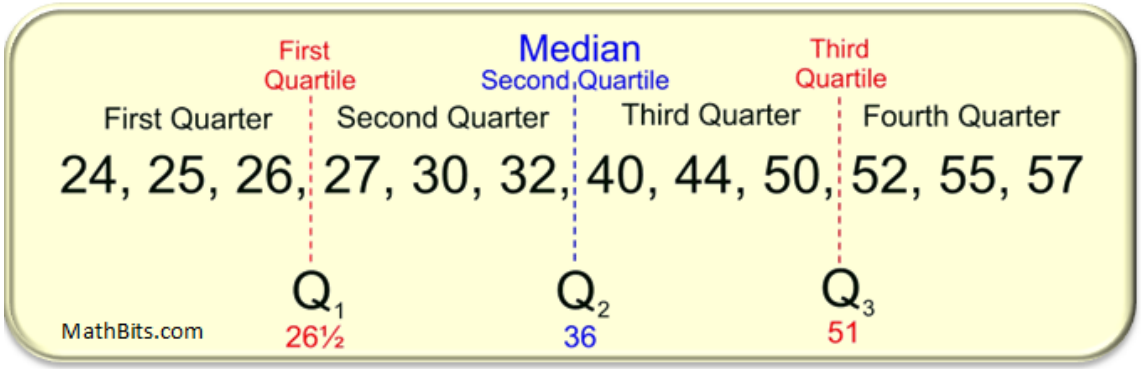

Image Source

Quartiles offer us another way to consider measures of central tendency and spread in the distribution of data. When the elements of the set are put in order, the set can be divided into 4 equal quarters. From this, we can calculate the inter-quartile mean, inter-quartile range, and draw box-and-whisker plots.

Inter-Quartile Mean

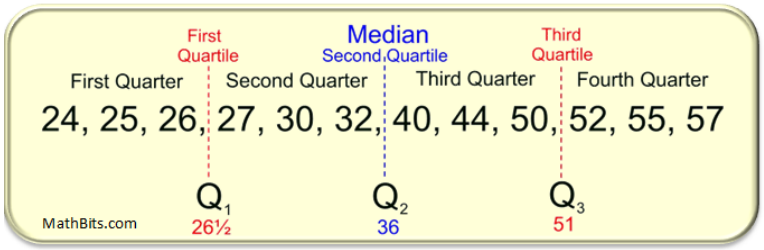

Image Source

The inter-quartile mean is calculated by taking the arithmetic mean of data between the first and third quartiles. This provides a measure of central tendency that is less affected by outliers. For this data set:

$$\{\cancel{24}, \cancel{25}, \cancel{26}, 27, 30, 32, 40, 44, 50, \cancel{52}, \cancel{55}, \cancel{57}\}$$

the inter-quartile mean is the average of the elements not crossed out, which is

$$\frac{27 + 30 + 32 + 40 + 44 + 50}{6} = 31.166\ldots$$

Inter-Quartile Mean

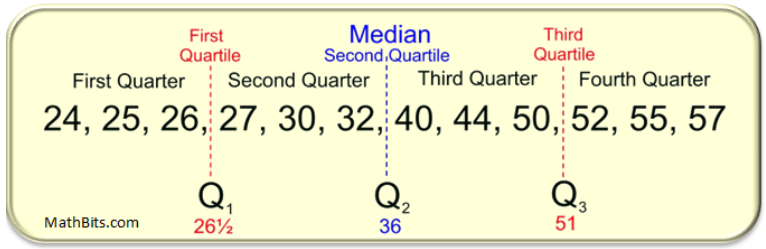

Image Source

The inter-quartile range (IQR) is calculated by taking the difference between the first and third quartiles. IQR is a measure of spread in our data, analogous to variance. For this data set,

$$\text{IQR} = Q_3 - Q_1 = 51 - 26.5 = 24.5$$

Box Plot

Box plots offer a way to visualize the quartiles and overall distribution of a data set. The "box" part of the box plot marks off the median, along with the first and third quartiles. 50% of the data lies between the first and third quartiles. Box plots are also known as box-and-whisker plots, where the "whiskers" at the ends either mark the maximum and minimum, or 1.5 times the inter-quartile range (IQR) past the first or third quartiles.

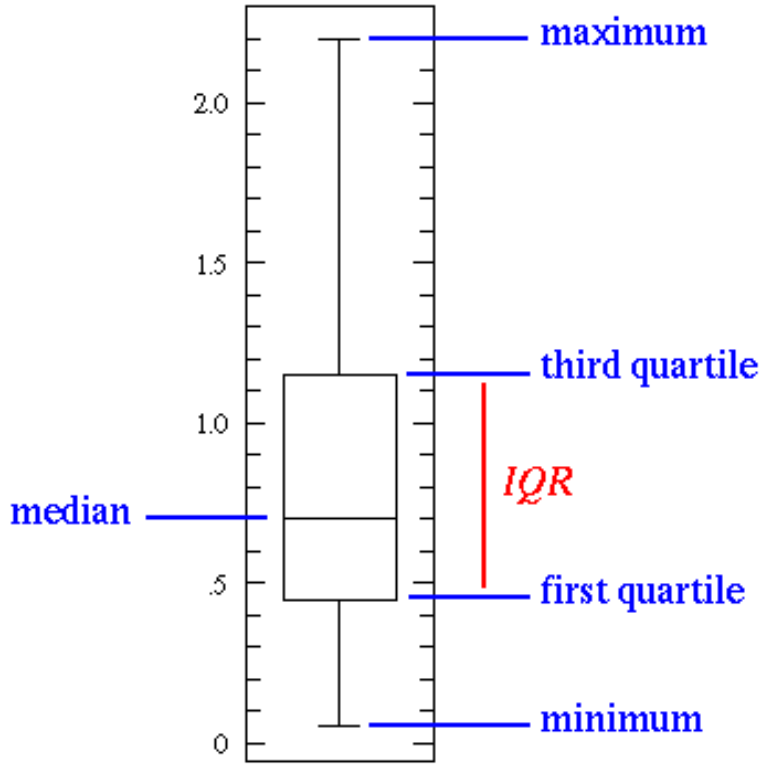

Image Source

Box Plot

Here, you can see how using quartiles compares with using standard deviation as a means of describing the distribution of data.

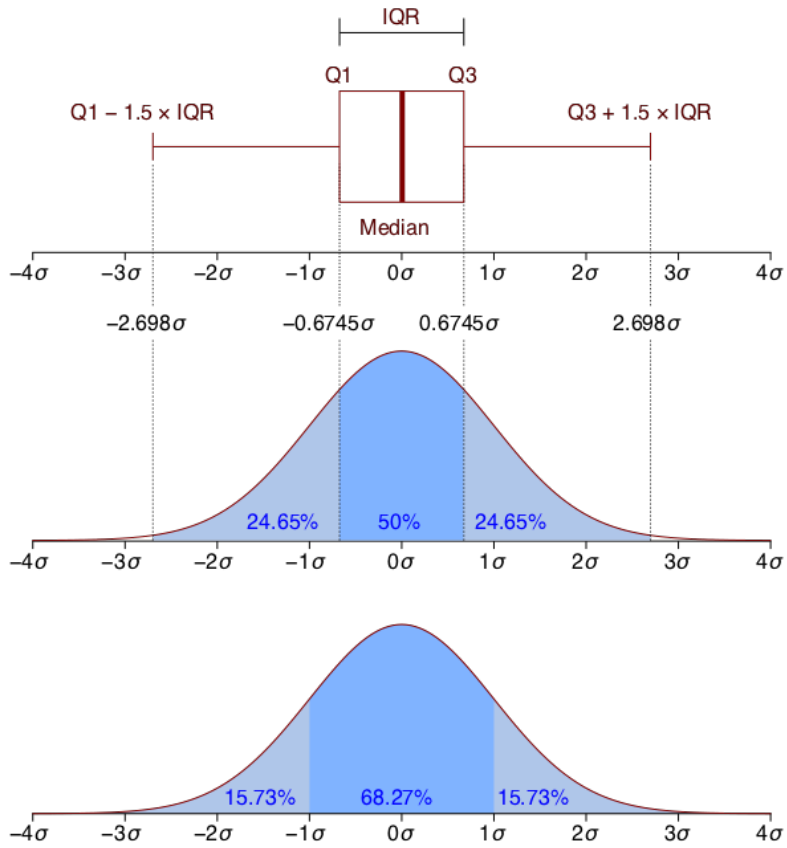

Image Source

Violin Plot

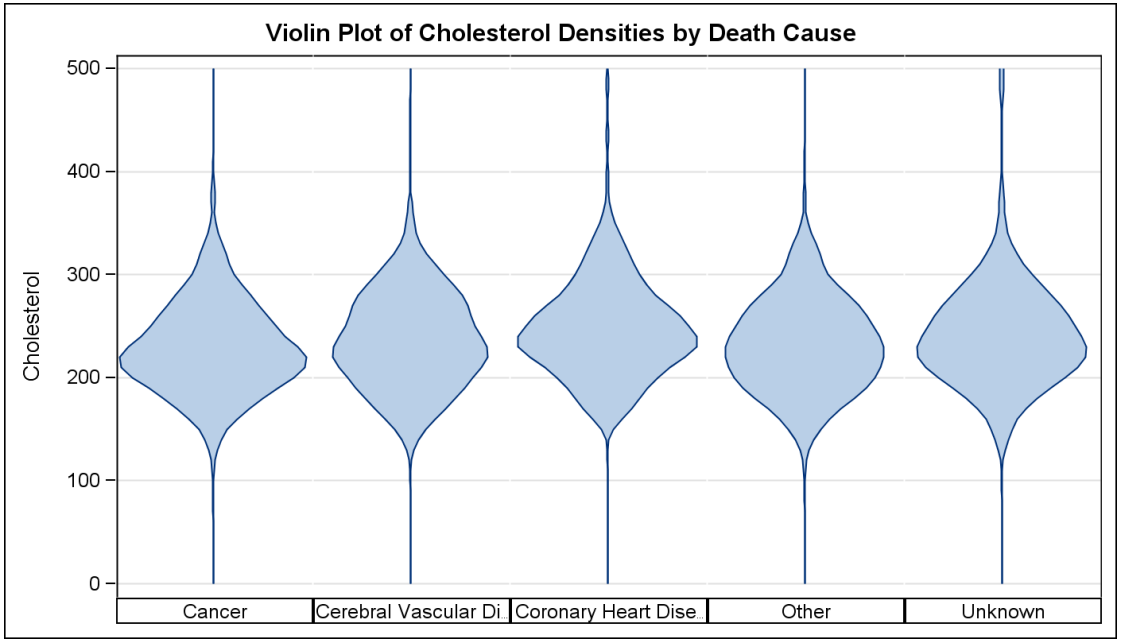

Image Source

Violin plots are an alternative to box-and-whisker plots that provide more detail about the distribution of data.

Histograms

Histograms, or bar charts, take the data and display the number of data points that fit into each bin. The bins are the sets represented by the width of each bar in the chart.

At right is a histogram showing how many tips were received at a restaurant, grouped by the tip amount. The bin width is \$1.00, so the first bar in the histogram represents tips received with amounts between \$0.50 and \$1.50.

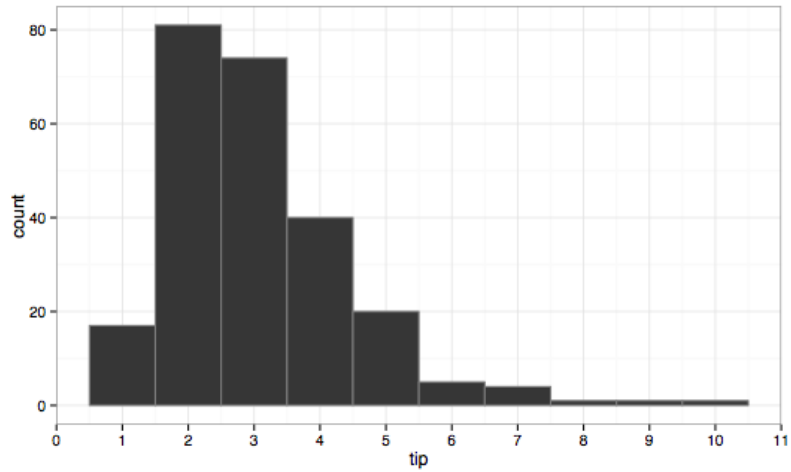

Image Source

Histograms

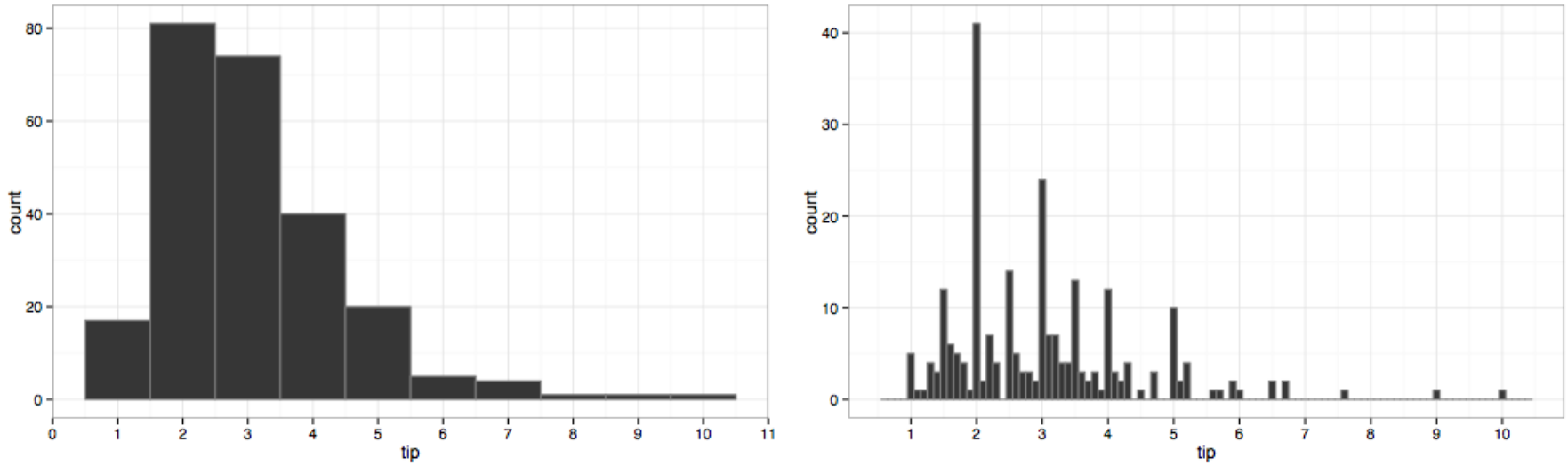

Image Source

Bin width can have a significant effect on a histogram. At left, the bin width is \$1.00, while at right, the bin width is \$0.10. Notice the spikes every 50 cents.

Estimation

Problems Requiring Estimation

- You know the heights for some group of American women. What is the probability that any given American woman is over 6' tall?
- You work on an assembly line that manufactures gears, and you have sampled some of those gears and checked them for defects. You want to know the likelihood that a given batch of gears will contain more than one defective gear.
- A pizza shop keeps track of how many pizzas are ordered every day. What are the chances that the pizza shop runs out of materials on a given day?

Expected Value

As you'll recall from the lecture on probability, the expected value of a random variable is a kind of "weighted average" for all possible values that variable can take. The expected value is what we would tend to see on average after doing an experiment a large number of times. We can also define expected value in terms of a sample of data.

Expected Value

For a discrete set of data points in a collection X, the expected value is

$$\mathbb{E}[X] = \sum_{i=1}^{n} \frac{1}{n} x_i$$

where $\{x_i\}$ is the set of values in the collection X.

Parametric vs. Nonparametric

Parametric models make assumptions about the distribution of the population, including that there is a finite set of parameters. This can lead to problems if we choose the wrong distribution. We will discuss the following parametric methods for estimation:

- Method of Moments (MOM)
- Maximum Likelihood Estimator (MLE)
- Maximum a Posteriori (MAP)

Nonparametric models are not dependent on assumptions about the shape of the population distribution, or its parameters. These models tend to have less power and are usually less interpretable than parametric models. We will discuss:

- Kernel Density Estimation (KDE)

Moments

In probability and statistics, moments give us numerical information about the shape of a distribution. Here, we will describe the central moments, which are moments about the mean.

- First moment: $\mathbb{E}[X]$. This is the mean, which is a measure of central tendency.
- Second moment: $\mathbb{E}[X^2] - (\mathbb{E}[X])^2$. This is the variance, which is a measure of spread.
- Third moment: skewness, which is a measure of how much the distribution is shifted either to the left or right:

$$\mathbb{E}\left[\left(\frac{X-\mu}{\sigma}\right)^3\right]$$

- Fourth moment: kurtosis, which describes the sharpness of the peak of the distribution curve and the size of the tails:

$$\mathbb{E}\left[\left(\frac{X-\mu}{\sigma}\right)^4\right]$$

Skewness and Kurtosis

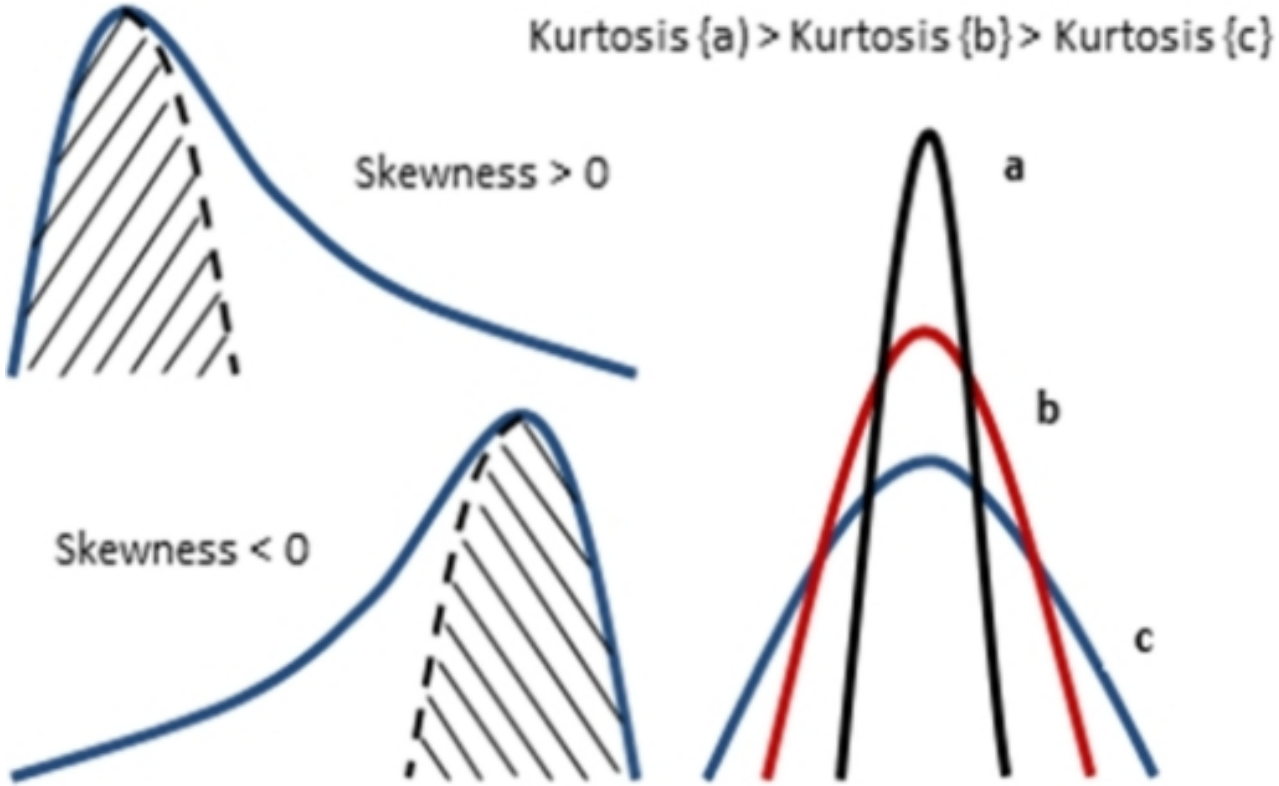

Image Source

Method of Moments (MOM)

The Method of Moments (MOM):

- Assume the data comes from a specific type of probability distribution (Normal, Poisson, Binomial, etc.).
- Derive equations relating the parameters of the distribution to the moments (mean, variance, skewness, etc.).
- Compute the moments from sample data to give us estimates for the parameters.

MOM Example: Pizza Shop

Based on a year's worth of data, a pizza shop sells on average 200 pizzas on any given Friday night. How can we model pizza sales to estimate the probability that the pizza shop may run out of materials on a given Friday?

MOM Example: Pizza Shop

- Since we are modeling sales in a set time frame (a Friday night), this scenario can be modeled with a Poisson distribution.
- We know the first moment $\mathbb{E}[X]$ directly gives us the parameter for a Poisson distribution:

$$\mathbb{E}[X] = \lambda$$

- Plugging this in, we get $\lambda = 200$, so the probability mass function for pizza sales should be

$$\mathbb{P}(k \text{ pizzas sold}) = \frac{\lambda^k e^{-k}}{k!} = \frac{200^k e^{-200}}{k!}$$

MOM Example: Height

Suppose we have height measurements for a random sample of 500 American women. Based on these measurements, what is the probability that any given American woman is over 6' tall?

MOM Example: Height

Since we are modeling height from a collection of people, we can model this with a normal / Gaussian distribution. We know the first moment $\mathbb{E}[X]$ directly gives us an estimate for the population mean, and the second moment $\mathbb{E}[X^2]$ can help us find the variance estimate:

$$\mathbb{E}[X] = \frac{1}{n}\sum_{i=1}^{n} X_i = \hat{\mu}$$

$$\mathbb{E}[X^2] = \frac{1}{n}\sum_{i=1}^{n} X_i^2 = \hat{\mu}^2 + \hat{\sigma}^2$$

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n} (X_i - \hat{\mu})^2$$

Maximum Likelihood Estimator (MLE)

Another approach to estimating parameters is the Maximum Likelihood Estimator (MLE). With MLE, we have a set of data points

$$\{x_1, x_2, \ldots, x_n\}$$

and we want to determine a parameter (or collection of parameters) θ. These determine a density function f which we will use to define a likelihood function L. We want to find parameter values θ that maximize L:

$$L(\theta; x_1, \ldots, x_n) = f(x_1, \ldots, x_n|\theta) = \prod_{i=1}^{n} f(x_i|\theta)$$

We can then take the logarithm of this likelihood function to give us the log-likelihood function ℓ. Logarithms are convenient here because they turn products into sums.

$$\ell(\theta; x_1, \ldots, x_n) = \ln L(\theta; x_1, \ldots, x_n) = \sum_{i=1}^{n} \ln f(x_i|\theta)$$

MLE Example: Pizza Shop

Suppose we have a problem we would like to solve with a Poisson distribution, like the pizza shop problem we saw before. What would it look like if we tried to use the maximum likelihood estimator here?

MLE Example: Pizza Shop

If $\{x_1, \ldots, x_n\}$ is the data set giving us the number of pizzas ordered on previous Fridays, then the likelihood function L is

$$L(\lambda) = \prod_{i=1}^{n} \left(\frac{\lambda^{x_i}}{(x_i)!} e^{-\lambda} \right)$$

The log-likelihood function ℓ is

$$\begin{aligned} \ell(\lambda) &= \ln \left[\prod_{i=1}^{n} \left(\frac{\lambda^{x_i}}{(x_i)!} e^{-\lambda} \right) \right] \\ &= \sum_{i=1}^{n} [x_i \ln(\lambda) - \ln(x_i!) - \lambda] \\ &= \ln(\lambda) \sum_{i=1}^{n} [x_i] - \sum_{i=1}^{n} [\ln(x_i!)] - n\lambda \end{aligned}$$

MLE Example: Pizza Shop

How do we maximize the log-likelihood function ℓ? We will take the derivative with respect to λ and set it equal to 0.

$$\frac{d\ell(\lambda)}{d\lambda} = \frac{d}{d\lambda}\left[\ln(\lambda)\sum_{i=1}^{n}[x_i] - \sum_{i=1}^{n}[\ln(x_i!)] - n\lambda\right] = \frac{1}{\lambda}\sum_{i=1}^{n}[x_i] - n$$

Setting $\frac{d\ell(\lambda)}{d\lambda} = 0$:

$$\frac{1}{\lambda}\sum_{i=1}^{n}[x_i] - n = 0 \quad \Rightarrow \quad \frac{1}{\lambda}\sum_{i=1}^{n}[x_i] = n$$

Therefore

$$\hat{\lambda} = \frac{1}{n}\sum_{i=1}^{n} x_i = \hat{\mu}$$

The parameter estimate $\hat{\lambda}$ is just the same as the sample mean $\hat{\mu}$.

Maximum a Posteriori (MAP) Estimation

The maximum likelihood estimator (MLE) has a weakness - it does not factor in our prior beliefs about the parameter values. Using Bayes' theorem, we can take our prior beliefs and convert them into a posterior probability. Bayes' theorem states:

$$\mathbb{P}(\theta|X) = \frac{\mathbb{P}(X|\theta)\mathbb{P}(\theta)}{\mathbb{P}(X)}$$

We can use this to get a posterior probability for the Maximum a Posteriori (MAP) estimate of our parameters, $\hat{\theta}_{MAP}$.

Maximum a Posteriori (MAP) Estimation

Finding the MAP estimate for parameters, like MLE, involves maximizing a function with respect to the parameters θ.

$$\begin{aligned}\hat{\theta}_{MAP} &= \text{argmax}_\theta \mathbb{P}(\theta|X) \\ &= \text{argmax}_\theta \frac{\mathbb{P}(X|\theta)\mathbb{P}(\theta)}{\mathbb{P}(X)} \\ &= \text{argmax}_\theta \mathbb{P}(X|\theta)\mathbb{P}(\theta) \\ &= \textit{argmax}_\theta \prod_{x_i \in X} \mathbb{P}(x_i|\theta)\mathbb{P}(\theta)\end{aligned}$$

We drop the denominator in the third line since it does not depend on θ.

Kernel Density Estimator (KDE)

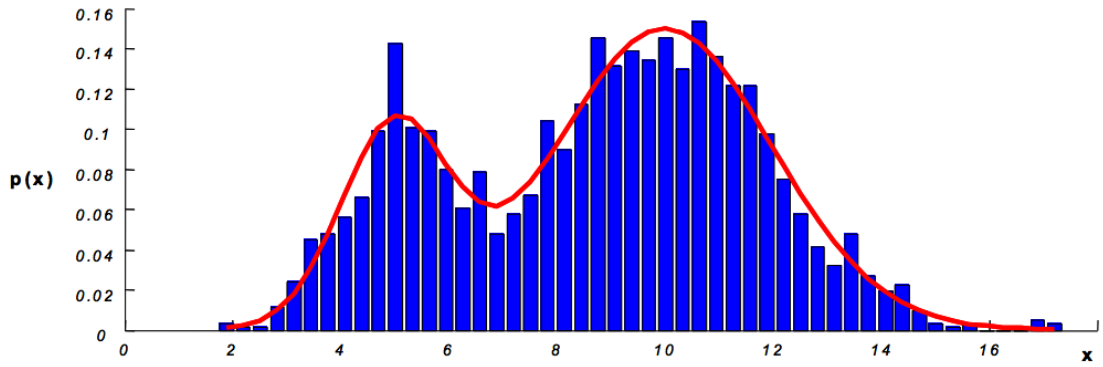

Image Source

The last three estimation methods were all parametric methods that assumed the data came from a certain distribution. The Kernel Density Estimator (KDE) is a method that is nonparametric and does not make these assumptions. In essence, KDE takes the histogram of existing data points and tries to "smooth" it out using a kernel.

Kernel Density Estimator (KDE)

If $(x_1, \ldots, x_n)$ is our set of sample data, then the kernel density estimator is

$$p_{KDE}(x) = \frac{1}{nh} \sum_{i=1}^{n} k(\frac{x - x_i}{h})$$

where k is the kernel function, and h is the bandwidth, which determines the "smoothness" of the estimation. It's common to use a Gaussian kernel:

$$k(u) = \frac{1}{\sqrt{2\pi}} e^{-\frac{u^2}{2}}$$

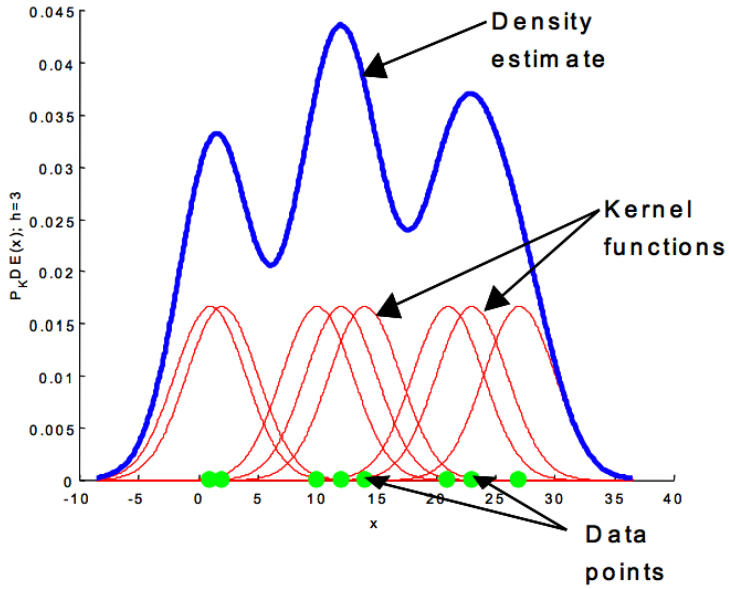

Image Source

Kernel Density Estimator (KDE)

Choosing the bandwidth h should be done carefully.

The gray curve on the right represents the true density function. The other curves have been fitted with KDE using a sample of 100 points. The red curve comes from a small bandwidth ($h = 0.05$) while the green curve comes from a large bandwidth ($h = 2$). The black curve comes closest with an intermediate bandwidth ($h = 0.337$).

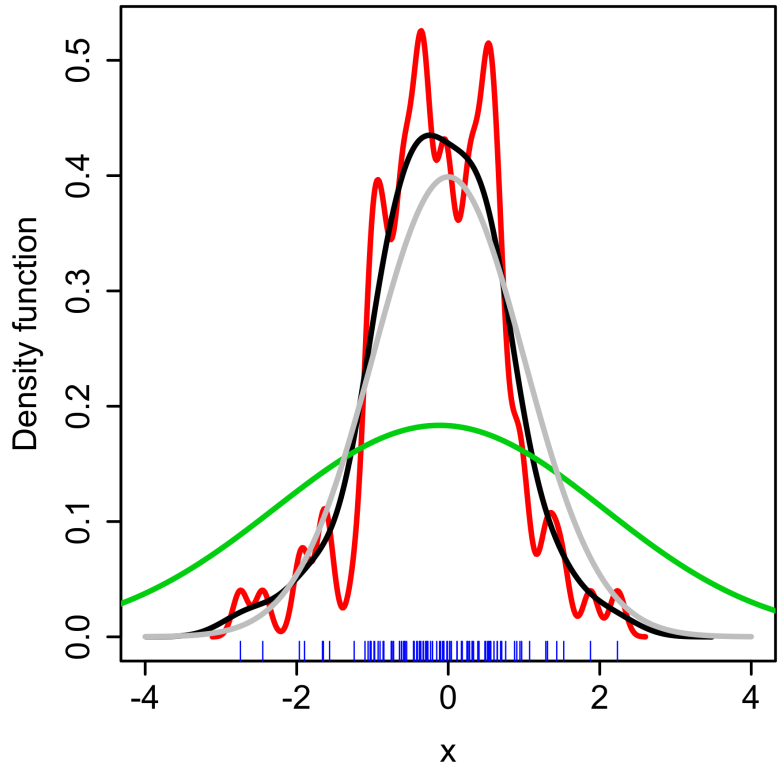

Image Source

Kernel Density Estimator (KDE)

A good rule of thumb for bandwidth (when using KDE with a Gaussian kernel) is

$$h = \left(\frac{4\hat{\sigma}^5}{3n}\right)^{\frac{1}{5}} \approx 1.06\hat{\sigma}n^{-1/5}$$

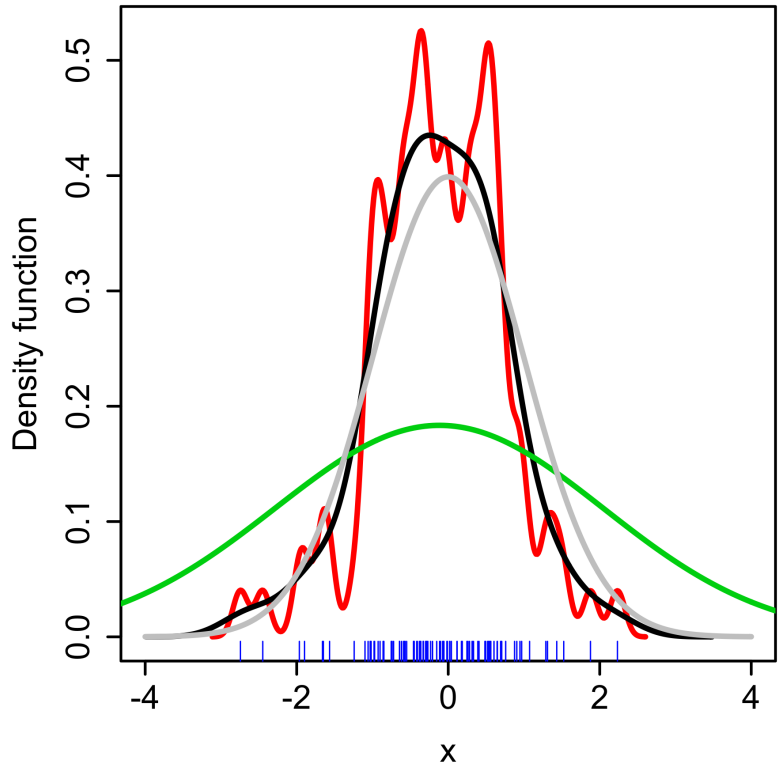

Image Source

Sampling

Intro to Sampling

Sometimes, we would like to know the properties and distribution of a large population, but all the necessary data is not available. To deal with this, we turn to sampling to give us a quantitative way of estimating the larger population using data from a subset of the population.

Intro to Sampling

It is important that a sample is somehow representative of the overall population. A biased sample will skew our estimates and give us an incorrect picture of the larger population.

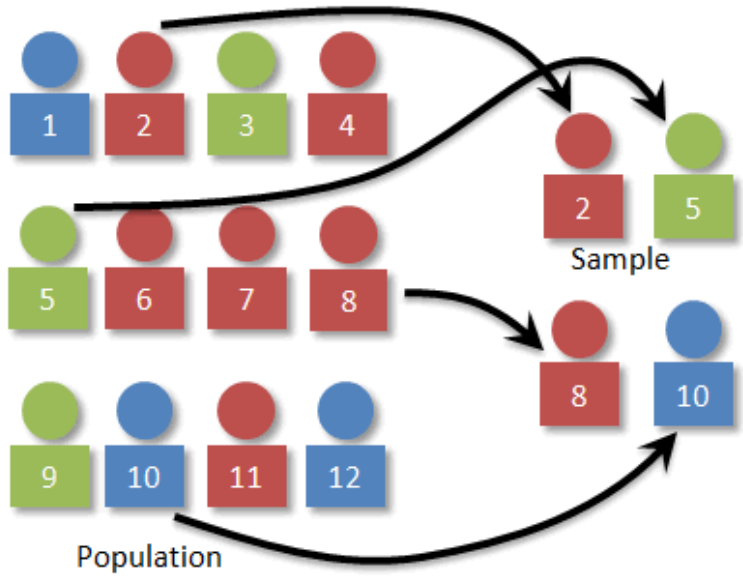

Image Source

Sampling Methods

- Simple random sampling.
- Cluster sampling.
- Stratified sampling.
- Systematic sampling.

Image Source

Simple Random Sampling

With simple random sampling, as the same suggests, we choose samples from the population at random. This is done in the hopes of getting a representative sample.

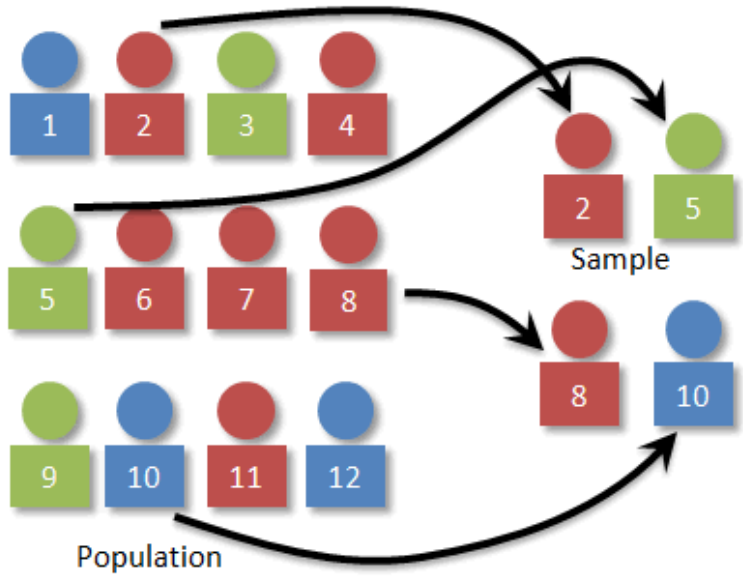

Image Source

Cluster Sampling

With cluster sampling, we break the population up into groups (clusters), not necessarily by any shared characteristic within clusters. A subset of these clusters is chosen at random, and this becomes our sample.

This is useful when the population is already split into groups, but the groups themselves are representative of the population as a whole.

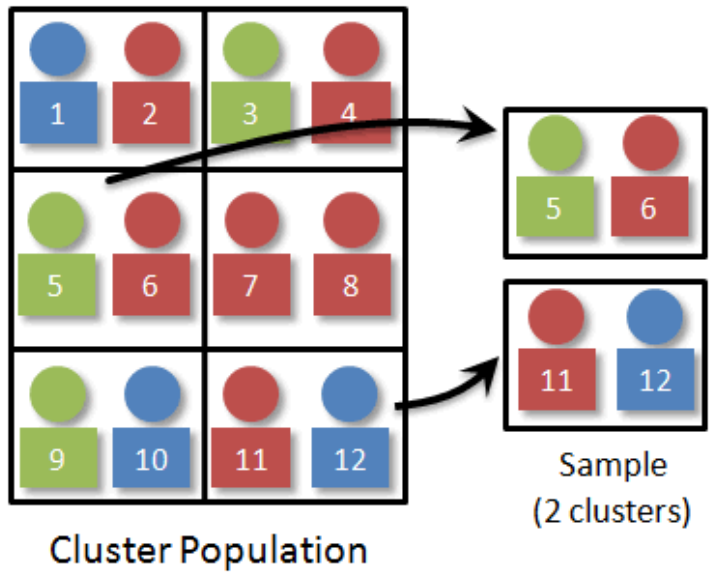

Image Source

Stratified Sampling

With stratified sampling, we first organize the population by some shared characteristic. In the example to the right, the strata are organized by color. A proportional number are chosen from each strata at random.

This can be used to ensure that our sample is proportionally representative of the population.

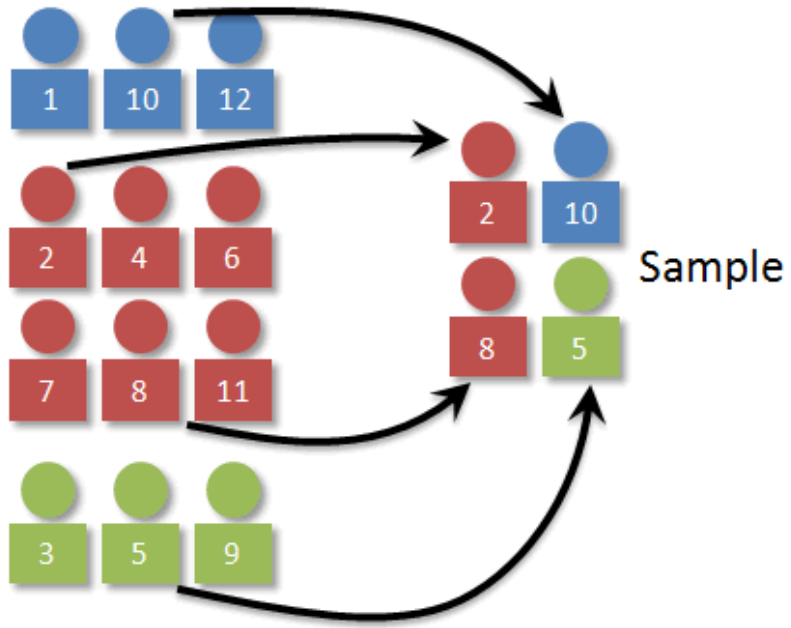

Image Source

Systematic Sampling

With systematic sampling, the population is put in some order. From this ordering, every k^{th} element is chosen. For instance, in the example on the right, element 2 is chosen along with every 3^{rd} element after that.

This is useful when the population is already lined up and numbered, but not necessarily ordered in a meaningful way.

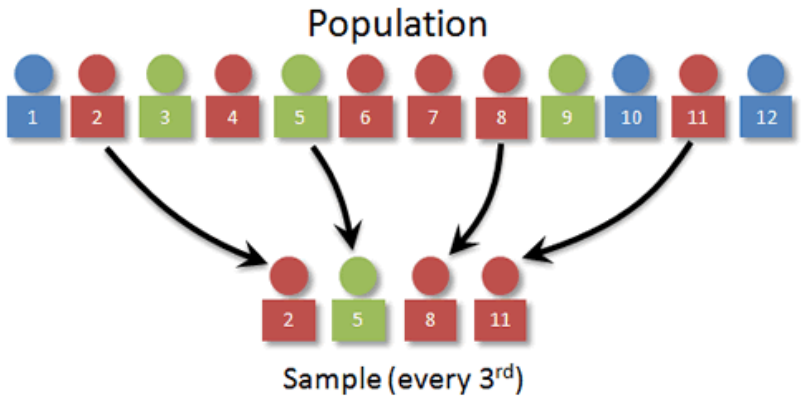

Image Source

Central Limit Theorem

According to the Central Limit Theorem (CLT), given a collection of independent and identically distributed (i.i.d.) random variables $\{X_1, \ldots, X_n\}$, their average distribution should converge to a normal / Gaussian distribution as $n \to \infty$.

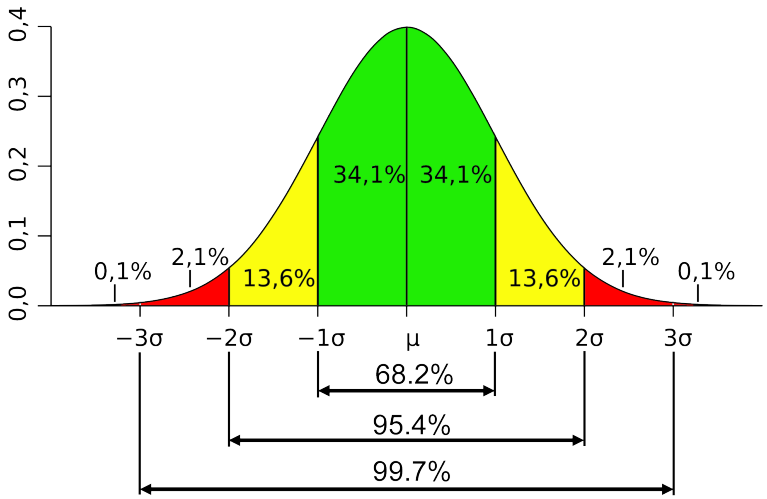

Image Source

Central Limit Theorem

If each X_i has mean μ and standard deviation σ, then the average distribution $\overline{X}$ has the following distribution:

$$\overline{X} \sim \text{Normal}\left(\mu, \frac{\sigma^2}{n}\right)$$

From this, we can calculate a Z score:

$$Z = \frac{\overline{X} - \mu}{\sigma/\sqrt{n}}$$

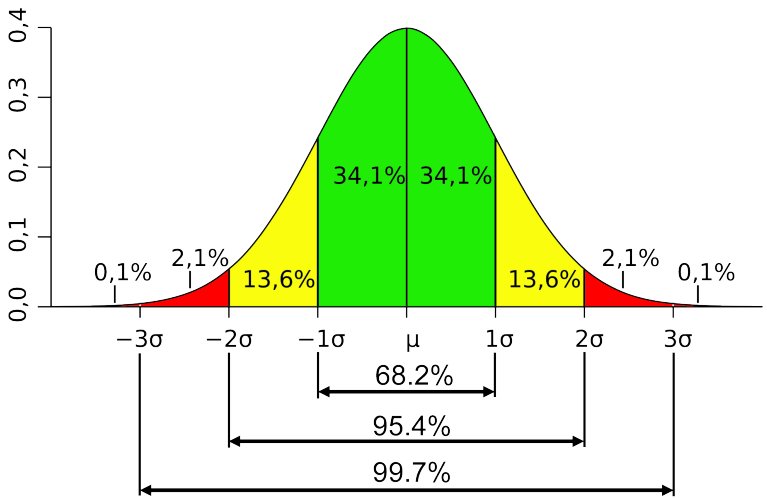

Image Source

Confidence Interval

We can establish confidence intervals for parameter estimates of the distribution. For instance, a 95% confidence interval for the estimate of the mean is everything within 1.96 standard deviations of the mean estimate:

$$\overline{x} \pm 1.96 \frac{s}{\sqrt{n}}$$

where $\overline{x}$ is the sample mean, s is the sample standard deviation, and n is the sample size.

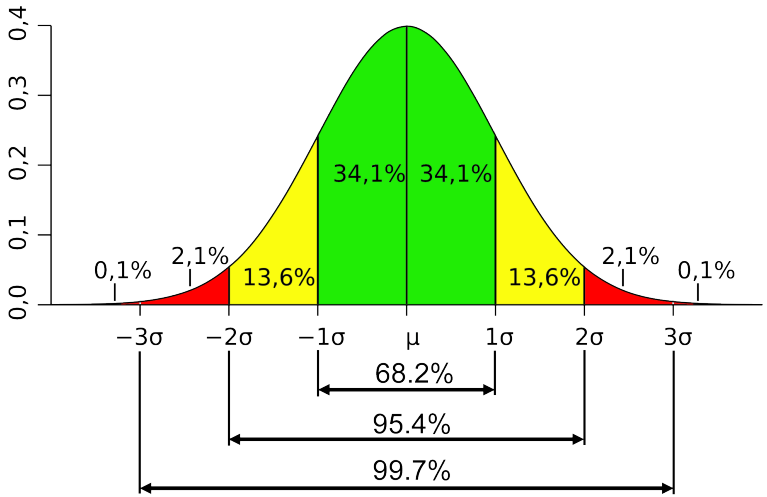

Image Source

Resampling

Resampling is the process of drawing repeated samples from our data. It is a topic we will cover at length later in the course. Some resampling methods include:

- Bootstrapping
- Cross-Validation
- Jackknifing
- Permutation Tests

Bootstrapping

Sometimes, we are limited by having a relatively small sample. We can resample from our original sample by bootstrapping. It can help us determine things like the standard error of a coefficient estimate or a confidence interval for a coefficient estimate, but it can also be applied to a wide range of machine learning methods.

The name comes from "pulling oneself up by one's bootstraps", as we use our sample data to create more samples.

Image Source

Bootstrapping

Bootstrapping should be done when

- The sample size is too small, so we need to kind of "artificially" get more data.
- The distribution is either unknown or too complicated.
- We want to estimate the variance of a parameter, like the sample mean.

Image Source

Bootstrapping

How to bootstrap:

- Start with your sample of size n.
- Pick n items from the sample, with replacement. This means there will likely be duplicates. This is one bootstrapped sample.
- Repeat this B times for B bootstrap samples.
- Use these bootstrapped samples for estimating parameters like variance.

Image Source

Bootstrapping for Sample Variance

Suppose we want to find the sample variance for the estimated mean of a population. We can create B bootstrapped datasets $D_1, D_2, \ldots, D_B$. Let

$$\hat{\mu}_j = \text{ mean of } D_j$$

which are the means of each bootstrapped sample. Then the sample variance for the estimated mean $\hat{\mu}$ is

$$\text{Var}[\hat{\mu}] = \frac{\sum_{b=1}^{B}(\hat{\mu}_b - \hat{\mu}_t)^2}{B-1}$$

where

$$\hat{\mu}_t = \frac{\sum_{b=1}^{B} \hat{\mu}_b}{B}$$

Cross-Validation

With cross-validation, we randomly split our data set into pieces and sample based on those splits of the data. Cross-validation can be either

- validation-set approach (split into 2 pieces).
- k-fold cross-validation (split into k pieces).
- leave-one-out cross-validation (split into individual data points).

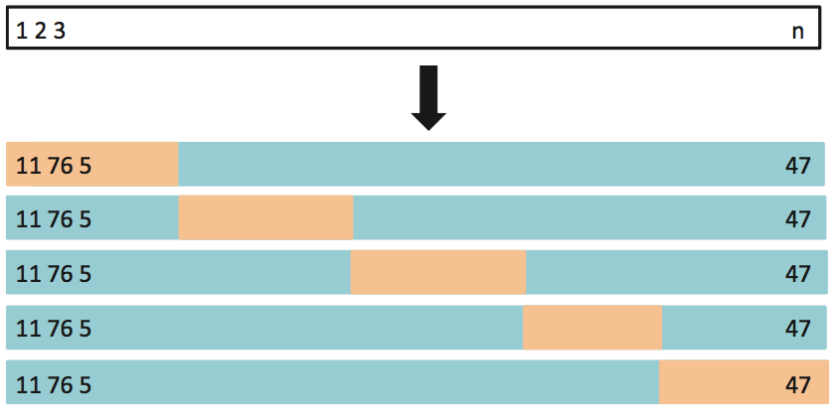

Image Source

Cross-Validation

As we will see in the lectures on machine learning in weeks 3-5, cross-validation can be used to divide the data into "training" and "test" sets. This can be used to prevent a model from "over-fitting" the data (fitting the noise more than the underlying relationship).

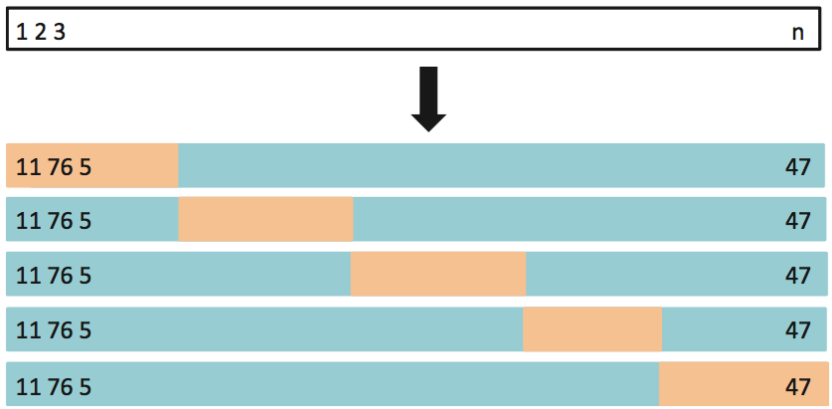

Image Source

Jackknifing

Jackknifing or jackknife resampling is used for estimating parameters by considering all possible subsets of the data set that have 1 element left out. Jackknifing gets its name because, like a Boy Scout's jackknife, it is a simple but rugged tool that can be used in plethora of problems.

Image Source

Jackknifing

For example, we can estimate the sample mean based on any i^{th} element left out of the data:

$$\overline{x}_i = \frac{1}{n-1} \sum_{j \neq i}^{n} x_j$$

We can combine these jackknife sample mean estimates into one estimator:

$$\overline{x}_{(.)} = \frac{1}{n} \sum_{k=1}^{n} \overline{x}_k$$

From this, we can get an estimate for the variance of the sample mean

$$\mathbb{V}\text{ar}(\overline{x}) = \frac{n-1}{n} \sum_{i=1}^{n} (\overline{x}_i - \overline{x}_{(.)})^2$$

Permutation Tests

Permutation tests are statistical significance tests, often used with hypothesis tests, where the distribution of a test statistic is obtained by considering rearrangements of the labels of observed data points.

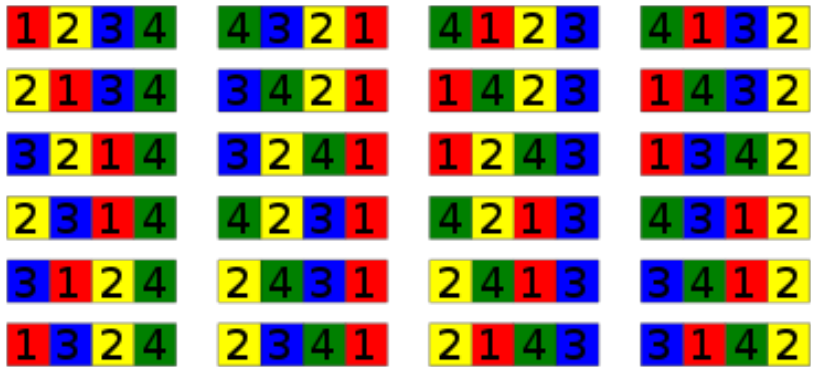

Image Source

Permutation Tests

For instance, suppose we want to test the hypothesis that men and women have the same height.

We could take a random sample of men and women, consider every possible permutation (or, for computability's sake, a subset of all permutations) where we mix up the labels on the data. In this case, we would randomly re-label data points as male or female, and see how the distribution of permutations compares with our original sample.