

INTRODUCTION TO STATISTICS AND COMPUTATION WITH DATA

Rituparna Sen, notes by Ramdas Singh

Second Semester

List of Symbols

$\hat{\theta}$, the estimate of a variable θ .

\bar{x} , the mean of the observations x_i .

x_M , the mediam of the observations x_M .

Contents

1	AN INTRODUCTION TO STATISTICS	1
1.1	Fundamental Elements of Statistics	1
1.2	Types of Statistics and Data	1
1.2.1	Descriptive Statistics	1
1.2.2	Inferential Statistics	2
1.2.3	Type of Data	2
1.3	Collecting Data	2
1.3.1	Sampling	2
1.3.2	Sources of Error due to Flawed Sampling	2
2	REPRESENTING QUANTITATIVE DATA	3
2.1	Graphical Measures	3
2.2	Numerical Measures	4
2.2.1	Variability or Spread	5
2.2.2	Measures of Relative Standing	6
3	PROBABILITY	7
3.1	Probabilistic Terms	7
3.1.1	Random Variables	7
3.1.2	Mean and Variance	7
3.2	Distributions	7
3.2.1	Discrete Distributions	7
3.2.2	Continuous Distributions	8
3.3	Normality	8
3.3.1	Sampling Distribution	9
3.3.2	Law of Large Numbers	9
3.4	Convergence of Random Variables	9
3.5	The Method of Moments	10
	Appendices	11
A	Appendix	13
	Index	15

Chapter 1

AN INTRODUCTION TO STATISTICS

January 2nd.

Commonly referred to as the science of data, *statistics* involves collecting, summarizing, presenting, and interpreting data. Randomness and variability in data necessitate the use of statistics. If a process is deterministic, there is no real need of statistics.

Often in probability, we are given the probability of getting heads in a single coin toss and we may be tasked to find the probability of 4 heads appearing in 10 tosses. In contrast, statistics starts with observing 4 heads appearing in 10 tosses, and utilises this data to determine the probability of a head. If we are to determine the price of a house in city, factors to look for may include the city itself, the specific location and area, the kind of house, the square footage, the age of the house, and the change with time. Despite all this, there is still an element of randomness; it is not a true deterministic quantity.

1.1 Fundamental Elements of Statistics

We first discuss some fundamental elements.

- An *experimental unit*. It may be a singular item such a coin toss, or a single house as in the previous example(s).
- The *population*. The set of all experimental units. We may note that studying all experimental units is a population may not be possible.
- A *census* studies all the units in a population.
- A *sample* of the population. A subset of the population. It is ideally chosen in a way that represents the entire population. A true representative sample may not always be possible without the the subset being the entire population.
- A *variable*. For each experimental unit in the sample, we record the data on several variables. In the case of the prices of houses, the factors discussed are the variables.
- *Univariate* and *multivariate* samples. As expected, a univariate sample has only one variable per unit, and a multivariate one has multiple variables per unit. Two variables per unit in the sample is also often referred to as a *bivariate* sample.

1.2 Types of Statistics and Data

1.2.1 Descriptive Statistics

Often, in statistics, we use pictures, tables, and summary numbers to describe the data. R Studio (or just R) may be used to handle descriptive statistics.

1.2.2 Inferential Statistics

Here, we make statements about the population based on our sample observations. In the case of coin tossing, let us look at the population of infinite coin tosses, where the probability of a heads is an unknown p . A census of this population is not possible, so we may take a sample of 10 tosses. Suppose we get X heads in this sample. We know that X follows a binomial distribution as $X \sim \text{Bin}(10, p)$. Here, we describe a new variable called the *estimate*, \hat{p} . In this case, we find that $\hat{p} = X/10$. This is how we work in statistics; we deal with an unknown variable of the population, say θ , by looking at a sample and describing an estimate $\hat{\theta}$.

We may also be tasked to find the *measure of reliability*; how reliable our estimate is. One such measure may be $|\hat{\theta} - \theta| < \delta$ where we target to make δ as small as possible.

Example 1.1. Suppose 1000 cola consumers participate in a blind taste test among 2 brands, A and B , and are asked their preference. To know which kind is preferred universally, let us begin by asking the following:

- Describing the population. In this case, it is all the cola consumers.
- Describing the sample. In this case, it is our chosen 1000 cola consumers.
- The variable of interest. Whether people prefer brand A or brand B .
- Our inference. The preference in the sample is extended to all the cola consumers.

1.2.3 Type of Data

Qualitative data, or categorical or nominal or ordinal data, is data with no numerical value representation of it. In reference to the previous example, preference between A and B is a qualitative piece of data. Other such examples include choice of elective courses of a student, the gender of a person, a preference of a cricket team, etc..

Quantitative data on the other hand has a numerical value which interests us in statistics. Examples include the age of a person, the semestral marks of a student, the salary of a worker, the cost of books, etc.. Quantitative data is also divided into two parts, discrete and continuous.

1.3 Collecting Data

January 7th.

To collect data to perform statistics on, one may choose of the following ways to do so; a most basic source for sampling is a *published source*. In a *designed experiment*, we select experimental units and administer some treatment on each one. In the medical field, these medical experiments are called clinical trials. An *observational study* may also be conducted.

1.3.1 Sampling

Two kinds of simple random sampling exist; one is *simple random sampling without replacement* (SR-SWOR), and the other is *simple random sampling with replacement* (SRSWR). Here, the experimental units are chosen sequentially at random with or without replacement. In *cluster sampling*, the population is divided into smaller groups known as clusters. Experimental units are then randomly selected among these clusters to form a sample. In *convenience sampling*, which is a non-probability sampling, the sample is drawn from that part of the population which is close to hand.

1.3.2 Sources of Error due to Flawed Sampling

A major source of error is *selection bias*; some parts of the population are deliberately left out when choosing the sample. It is an error of bias and not of randomness. For example, online surveys leave out people without access to internet. While this may be at fault of the person sampling the population, a *non-response bias* occurs when the population does not respond. A *response bias* also exists where the population does not reflect the true value. An error in noting down values or measuring samples may also occur, known as a *measurement error*.

Chapter 2

REPRESENTING QUANTITATIVE DATA

The most one can possibly do with qualitative data is to pictorially represent it via tables and charts.
January 9th.

For qualitative data, frequency tables, bar charts, pie charts, and ogives are the most appropriate way to pictorially represent them. Quantitative data, on the other hand, have more variety in terms in descriptive statistics. They can be represented in graphs such as a dot plot, stem and leaf plot, histogram, and even a box and whisker plot.

Plots can also be categorized as skewed and symmetric. A *symmetric plot* is (roughly) the same as its image under reflection about some vertical line. We typically look for symmetry of the population from which the data is a sample. A *skewed plot* is one where a peak in the plot occurs. A peak towards higher values is termed a left skewed plot, while a peak towards lower values is termed a right skewed plot. We can also have a *bimodal distribution* where 2 peaks occur, or even a *multimodal distribution*. More peaks suggests a mixture in the population.

Another thing to look out for are *outliers*. These are a few data points that are very different from the rest of the data. When such an outlier occurs, we are generally provoked to investigate it. A reason for an outlier could be a mistake in recording of data, which can be removed by fixing it. Another possible reason for an outlier is that the data point(s) come from a different distribution. In such a case, it is best to drop such outliers. Outliers may also occur purely due to the population being highly skewed; in this case, such outliers are expected by one.

2.1 Graphical Measures

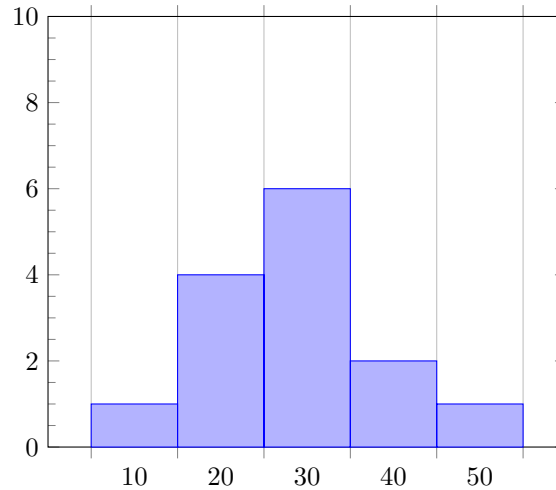
A *stem and leaf plot*. It is typically used when there are too many data points to draw a meaningful dot plot. It also has the added advantage of no loss of data. One may even look at it as a horizontal

Table 2.1: Key: 1|1 = 11

Stem	Leaf
1	3
2	2 4 8 9
3	1 2 6 6 7 8
4	3 7
5	2

histogram, with the number of values per row representing the ‘height’ of the bars.

A *histogram*. For the same data as the previous example, we use bars instead to categorize classes. However, there is a loss of individual data points here.



A histogram.

A general rule of thumb exists for determining the number of classes. When there are less than 25 observations, we choose 5–6 classes; for 25 to 50 observations, 7–14 classes are enough. For more than 50 observations, one may even choose to use about 15–20 classes. Another thing to note about the histogram is that the area under the curve for a particular region represents the proportion of that region in the sample.

A box and whisker plot. Again using the same data, a box and whisker plot utilises other aspects of the data such as the maximum value, the minimum value, the median value, the first quartile end value, and the third quartile end value. In our case, these are 52, 12, 24, 28, and 38 respectively (*A box and whisker plot is to be added here later*).

A scatter plot. The plots discussed before were for univariate data. For bivariate data, a scatter plot is helpful. Here, we plot one variable against the other by representing them as 2-dimensional points. We also discuss the linear relationship in this case.

2.2 Numerical Measures

We first discuss some central values.

- The *mean*. It is the average value (the arithmetic mean) of the sample.
- The *median*. It is that value the splits the data in half; half the data points are below this value, while the other half are above it.
- The *mode*. It is that value which occurs with the highest frequency. Often, for larger samples, a modal class is more meaningful.

Note that the mean and median do not make sense for qualitative data. In contrast to this, the mode can be used for both qualitative and quantitative data. The median is sometimes used in qualitative data when it is ordinal.

Example 2.1. For a set of data points x_1, \dots, x_n , the mean is defined as $\bar{x} = (\sum x_i) / n$. Show that, for $a = \bar{x}$, the sum of squared deviations of the data points from the value a is minimized, that is, the value $\sum (x_i - a)^2$ is minimized.

Proof. We have

$$\begin{aligned} \sum (x_i - a)^2 &= \sum (x_i - \bar{x} + \bar{x} - a)^2 = \sum (x_i - \bar{x})^2 + \sum (\bar{x} - a)^2 + 2 \sum (x_i - \bar{x})(\bar{x} - a) \\ &= \sum (x_i - \bar{x})^2 + n(\bar{x} - a)^2 + 2(\bar{x} - a) \sum (x_i - \bar{x}) \\ &= \sum (x_i - \bar{x})^2 + n(\bar{x} - a)^2. \end{aligned}$$

Now $(\bar{x} - a)^2 \geq 0$; it is 0 if and only if $a = \bar{x}$. ■

January 16th.

We also have another result that says that $\sum |x_i - a|$ is minimized when the value of $a = \text{median}(x_1, \dots, x_n)$.

2.2.1 Variability or Spread

We also are interested in the *variability* or the *spread*; how far the observations or data points are from each other or from the centre.

- One such measure is the *range* defined as $\max - \min$. The range is a simple and intuitive measure of spread. Between different samples from the same population, the value of the range can be very different. It depends on only two values.
- Another measure is the *variance*. For n x_i observations, we have

$$\sigma^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2 \quad (2.1)$$

as the variance; it is the average squared distance of each point from the centre. Dividing by $n-1$ makes the measure ‘good’ as it is unbiased for the population variance.

- If we define x_M to be the median of the x_i ’s, we can define another measure: *MAD from the median*.

$$\text{MAD} = \frac{1}{n} \sum_{i=1}^n |x_i - x_M|. \quad (2.2)$$

MAD is short for *mean absolute deviation*.

- The *standard deviation* may be defined as the square root of the variance.
- The *interquartile range* or IQR. It is a robust measure; applicable if there are a few very high or very low values. It is defined as

$$\text{IQR} = Q_3 - Q_1 \quad (2.3)$$

where Q_3 is the third quartile, the value for which 75% of the data resides below it, and Q_1 is the first quartile, the value for which 25% of the data resides below it.

Example 2.2. Let us work with the following dataset—

40, 60, 65, 65, 65, 68, 68, 70, 70, 70, 70, 70, 70, 74, 75, 75, 90, 95

For this dataset, the median is the $\frac{1}{2}(n+1)$ th observation, or the 9.5th observation: $70 + 0.5(70 - 70) = 70$.

Q_1 will be the $\frac{1}{4}(n+1)$ th observation, or the 4.75th observation: $65 + 0.75(65 - 65) = 65$.

Q_3 will be the $\frac{3}{4}(n+1)$ th observation, or the 14.25th observation: $74 + 0.25(75 - 74) = 74.25$.

The Chebyshev Inequality

The *Chebyshev inequality* can be summarised as

$$P\left(\left|\frac{X - \mu}{\sigma}\right|\right) \leq \frac{1}{k^2} \quad (2.4)$$

where μ is the population mean and σ is the population standard deviation. Along with this, we also adopt the convention \bar{X} for the sample mean, and s for the sample standard deviation. We can also infer that 2 standard deviations worth of values around the mean contain, at the very least, three-fourths of the values.

$$P(\mu - 2\sigma < X < \mu + 2\sigma) \geq \frac{3}{4}. \quad (2.5)$$

Note that this is true for all distributions. Similarly, we have $P(\mu - 3\sigma < X < \mu + 3\sigma) \geq \frac{8}{9}$. If we have \bar{X} and s , then we can say that $\frac{8}{9}$ ths of the population values will lie between $\bar{X} - 3s$ and $\bar{X} + 3s$, approximately. In practice, the coverage is more. For example, if we have the normal distribution, 95% of the values will lie in between $\mu \pm 2\sigma$. Chebyshev guarantees that 75% of the values will lie here for any distribution.

2.2.2 Measures of Relative Standing

These measure look at the ‘position’ of a particular value relative to other, or in comparison to others.

- A *rank* is simply ordering of the population, and numbering them in order. The number assigned is the rank.
- A *percentile* or *quartile* also exists; it is defined as

$$\text{Quartile of value } a = \frac{\#\{x_i \leq a\}}{n}. \quad (2.6)$$

- The *z-score* measures how many standard deviations above the mean is the value a —

$$\text{z-score of } a = \frac{a - \bar{x}}{s}. \quad (2.7)$$

Detecting Outliers

The z-score is helpful in detecting outliers; if the absolute value of the z-score of a value a is above 3, there is a high likelihood it is an outlier.

For a boxplot, a good detection system is to check for those values greater than $Q_3 + 1.5\text{IQR}$ or less than $Q_1 - 1.5\text{IQR}$.

Chapter 3

PROBABILITY

January 23rd.

We first discuss the *simple random sampling*. From a set of N elements, a subset of size n is selected such that each subset of size n has equal probability of being the chosen subset. With replacement, this probability is $\frac{1}{N^n}$, and without replacement, this probability is $\frac{1}{\binom{N}{n}}$. In practice, this is done iteratively. There are ways to achieve such a sampling; some of these include physically mixing up and making a draw, using a generator, random number tables, or software.

3.1 Probabilistic Terms

3.1.1 Random Variables

For a probability space (Ω, \mathcal{F}, P) , a *random variable* is simply a function $X : \Omega \rightarrow \mathbb{R}$. For a given subset $A \subseteq \mathbb{R}$, the probability $P(X \in A)$ is defined as $P(X \in A) = P(X^{-1}(A))$.

We also define the *cumulative distribution function* of X as

$$F_X(a) = P(X \leq a) = P(X \in (-\infty, a]). \quad (3.1)$$

If the random variable is discrete, that is, it takes countably many values, then a *probability mass function* can be defined as $p_X(x_i) = P(X = x_i)$. If the random variable is continuous, that is, it takes all values in an interval, then a *probability density function* is defined as $\int_a^b f_X(x)dx = F_X(b) - F_X(a)$.

3.1.2 Mean and Variance

The mean of a random variable is defined as

$$\mu = EX = \begin{cases} \int_{-\infty}^{\infty} xf_X(x)dx & \text{if } X \text{ is continuous,} \\ \sum_i x_i p_X(x_i) & \text{if } X \text{ is discrete.} \end{cases} \quad (3.2)$$

The variance of a random variable is defined as

$$\sigma^2 = E[(X - \mu)^2] = E[X^2] - \mu^2 = \begin{cases} \int_{-\infty}^{\infty} (x - \mu)^2 f_X(x)dx & \text{if } X \text{ is continuous,} \\ \sum_i (x_i - \mu)^2 p_X(x_i) & \text{if } X \text{ is discrete.} \end{cases} \quad (3.3)$$

3.2 Distributions

3.2.1 Discrete Distributions

- *Binomial distribution*. It consists of n trials, all independent, which result in outcome 0 or 1. The probability of success in each trial is the same, p . Here, X is the number of successes.

$$p_X(k) = \binom{n}{k} p^k (1-p)^{n-k}, \quad EX = np, \quad \text{Var}X = np(1-p).$$

- *Multinomial distribution.* The random variable is a vector from $\Omega \rightarrow \mathbb{R}^k$. There are n trials, all independent, with k possible outcomes in each trial. Each outcome has probability p_i , with $\sum_{i=1}^k p_i = 1$. The random variable is written as $X = (X_1 \dots X_k)$, where X_i is the number of times outcome i is observed.
- *Geometric distribution.* A trial is conducted, independently, until a success instead of a failure is observed. X is the number of such trials required.

$$p_X(k) = p \cdot (1-p)^{k-1}.$$

- *Poisson distribution.* The random variable here has the probability mass function defined as

$$p_X(k) = e^{-\lambda} \frac{\lambda^k}{k!}.$$

- *Hypergeometric distribution.* N objects exist, with M of the marked. n are sampled, and X is defined as the number of marked ones picked. The probability is given by

$$p_X(k) = \frac{\binom{M}{k} \binom{N-M}{n-k}}{\binom{N}{n}}.$$

3.2.2 Continuous Distributions

We shall define the probability mass functions here.

- *Uniform distribution.* It has a finite support only. It is given by

$$f_X(x) = \begin{cases} \frac{1}{b-a} & \text{if } a < x < b, \\ 0 & \text{if otherwise.} \end{cases}$$

- *Beta distribution.* It also have a finite support.

$$f_X(x) = \begin{cases} \frac{x^{\alpha-1}(1-x)^{\beta-1}}{B(\alpha, \beta)} & \text{if } 0 < x < 1, \\ 0 & \text{if otherwise.} \end{cases}$$

- *Exponential distribution.* It has a ‘rate’ denoted by λ , and a support of \mathbb{R}^+ , with probability density function

$$f_X(x) = \begin{cases} \lambda e^{-\lambda x} & \text{if } x > 0, \\ 0 & \text{if otherwise.} \end{cases}$$

- *Gamma distribution.* Takes two parameters λ and α , with support of \mathbb{R}^+ , with its function defined as

$$f_X(x) = \begin{cases} \frac{e^{-\lambda x} x^{\alpha-1} \lambda^\alpha}{\Gamma(\alpha)} & \text{if } x > 0, \\ 0 & \text{if otherwise.} \end{cases}$$

- *Normal distribution.* Takes two parameters μ and σ , which are the mean and the standard deviation respectively. It is defined as

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

3.3 Normality

January 28th. We ask if the assumption is reasonable. There are several ways to achieve this;

- Check if the histogram is symmetric and unimodal (even roughly).

- Check if $\bar{x} \pm s$, $\bar{x} \pm 2s$, and $\bar{x} \pm 3s$ contains close to 67%, 95%, and 99%, respectively, of the data under normality. These are called the *empirical percentages*.
- Check if the ratio of the IQR to the standard deviation is close to 1.3.
- Check if the skewness is close to 0 and kurtosis is close to 3.
- For datapoints x_1, x_2, \dots, x_n , we rearrange them such that $x_{(1)} < x_{(2)} < \dots < x_{(n)}$. This is known as *ranked data*. For this rearrangement, we make a normal probability plot. If the distribution is normal, then we would have $X_i = \mu + \sigma Z_i$, where $Z_i \sim N(0, 1)$. Here,

$$\Phi(z) = P(Z \leq z) \approx \Phi_n(z) = \frac{\#\{Z_i \leq z\}}{n}. \quad (3.4)$$

In this case, we have $Z_{(j)} \approx \Phi^{-1}(\frac{j}{n})$. We make a *QQ plot*, plotting the ranked datapoints $x_{(j)}$ against $\Phi^{-1}(\frac{j}{n})$. A *PP plot* can also be made, where we plot $\Phi(x_j)$ against $\frac{j}{n}$. Both of these are types of normal probability plots.

3.3.1 Sampling Distribution

A *parameter* is defined as a finite dimensional object, most commonly scalar, that is a function of the population, mean, variance, skewness, median, etc. An *estimator* is a statistic which is any function of the sample values used to estimate a particular parameter.

3.3.2 Law of Large Numbers

Let X_1, X_2, \dots be a sequence of random variables, with $EX_i = \mu$ and $\text{Var}X_i = \sigma^2 < \infty$. Let $\bar{X}_n = \frac{1}{n} \sum_{i=1}^n X_i$. For $\varepsilon > 0$, the *weak law of large numbers* states that

$$\lim_{n \rightarrow \infty} P\left(\left|\frac{\bar{X}_n - \mu}{\mu}\right| > \varepsilon\right) = 0. \quad (3.5)$$

3.4 Convergence of Random Variables

1. Convergence in probability: We say a sequence of random variables $\{Y_n\}_{n \in \mathbb{N}}$ converges to Y if, for a given $\varepsilon > 0$,

$$\lim_{n \rightarrow \infty} P(|Y_n - Y| > \varepsilon) = 0. \quad (3.6)$$

2. Almost sure convergence: We say a sequence of random variables $\{Y_n\}_{n \in \mathbb{N}}$ converges almost surely to Y if

$$P(\{\omega \in \Omega : Y_n(\omega) \rightarrow Y(\omega)\}) = 1. \quad (3.7)$$

3. Convergence in distribution: If $Y_n \sim F_n$ and $Y \sim F$, then we say $Y_n \rightarrow Y$ if, for every continuity point x of F , we have

$$\lim_{n \rightarrow \infty} F_n(x) = F(x). \quad (3.8)$$

Example 3.1. If X_1, X_2, \dots are independent and identically distributed with mean μ and variance σ^2 , we can define

$$Y_n = \frac{(\frac{1}{n} \sum_{i=1}^n X_i - \mu)}{\sigma/\sqrt{n}} \quad (3.9)$$

and we will have $Y_n \rightarrow Y \sim N(0, 1)$. This is the *central limit theorem*. It is very useful because if we have large n , then we can simply use the standard normal distribution as an approximation of the actual distribution of the sample mean. The theorem applies regardless of the distribution of the X_i 's.

3.5 The Method of Moments

For X_1, \dots, X_n independent and identically distributed, the method of moments is useful for estimation.

Example 3.2. The estimator for μ is \bar{X} . We wish to estimate σ^2 . Note that $\bar{X} = EX_1 = \mu$, and $EX_1^2 = \sigma^2 + \mu^2$ which gives us $\overline{X^2} = E[\frac{1}{n}(X_1^2 + \dots + X_n^2)] = \sigma^2 + \mu^2$. Thus, the estimator for σ^2 may be given as

$$\hat{\sigma}^2 = \overline{X^2} - \bar{X}^2 = \frac{1}{n} \sum (X_i - \bar{X})^2 = \frac{n-1}{n} s^2. \quad (3.10)$$

This is the estimator for σ^2 .

Appendices

Chapter A

Appendix

Extra content goes here.

Index

- Beta distribution, 8
- bimodal distribution, 3
- Binomial distribution, 7
- bivariate, 1
- box and whisker plot, 4
- census, 1
- central limit theorem, 9
- Chebyshev inequality, 5
- cluster sampling, 2
- convenience sampling, 2
- cumulative distribution function, 7
- designed experiment, 2
- empirical percentages, 9
- estimate, 2
- estimator, 9
- experimental unit, 1
- Exponential distribution, 8
- Gamma distribution, 8
- Geometric distribution, 8
- histogram, 3
- Hypergeometric distribution, 8
- interquartile range, 5
- MAD from the median, 5
- mean, 4
- mean absolute deviation, 5
- measure of reliability, 2
- measurement error, 2
- median, 4
- mode, 4
- multimodal distribution, 3
- Multinomial distribution, 8
- multivariate, 1
- non-response bias, 2
- Normal distribution, 8
- observational study, 2
- outliers, 3
- parameter, 9
- percentile, 6
- Poisson distribution, 8
- population, 1
- PP plot, 9
- probability density function, 7
- probability mass function, 7
- published source, 2
- QQ plot, 9
- Qualitative data, 2
- Quantitative data, 2
- quartile, 6
- random variable, 7
- range, 5
- rank, 6
- ranked data, 9
- response bias, 2
- sample, 1
- scatter plot, 4
- selection bias, 2
- simple random sampling with replacement, 2
- simple random sampling without replacement, 2
- simple random sampling, 7
- skewed plot, 3
- spread, 5
- standard deviation, 5
- statistics, 1
- stem and leaf plot, 3
- symmetric plot, 3
- Uniform distribution, 8
- Univariate, 1
- variability, 5
- variable, 1
- variance, 5
- weak law of large numbers, 9
- z-score, 6