Chapter I: Introduction

1. What is statistics

Statistics is a formal system that can be used for two things:

1. Describing a given set of data (descriptive statistics).
2. Making inferences and decisions motivated by the analysis of the data (inferential statistics)

In statistics we call a function that can be applied to a given data sample a 'statistic'. E.g. the mean of the values in sample X is a statistic on X.

2. Basic terms of statistics

Generally, when statistics is used in everyday life, we want to analyze features of a great collection of entities without having to look at each and every of its items. Therefore we draw a certain number of random items from the collection and analyze them in order to make an educated guess on the whole collection.

The collection of items is generally called the 'population', while the items drawn from the population are called 'samples'. The distinction between population and sample is crucial to understanding statistics and we have to ensure that they are never confused. We execute a statistic on a sample in order to infer what this statistic might look like on the whole population. However, there is always a certain probability that this estimation is damn wrong. We will need means to calculate this probability in order to know whether we can 'trust' our sample statistic.

When we use a statistic on a data collection, we calculate usually one or several values. Those values are called parameters of the data collection.

Train of thought in statistics

1. I want to know a parameter of a huge population, e.g. the mean. But because the population is that large (or even infinitely large) and I do not have infinite time, I cannot calculate the mean on the whole population.
2. I cover my eyes and grab e.g. 100 data values from my population. I do my best to make this process at pure random.
3. I simply calculate the mean on my drawn data sample. Because I drew randomly from my population, I expect the sample to show to some extent the same characteristics as my population. Therefore I use my sample mean as an estimation of the population mean.
4. Damn! What if I drew unrepresentative values from my population? After all, I do have two left hands, and may just have used extreme values from the population and the sample mean isn't anywhere near the population mean.
5. I sure ain't gonna stop reading this script before I know how to be sure of my estimation.

Chapter II: Probability Theory

1. Set Notation

Sets are one way to express probability. Let's start with the notion of an event.

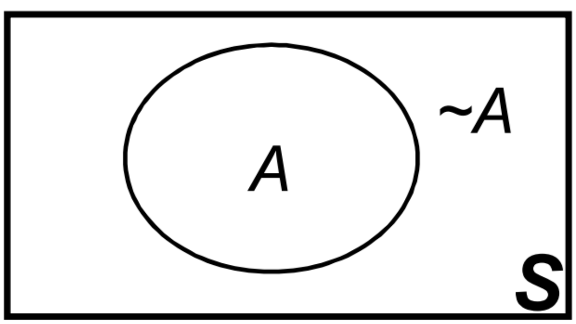
An **event** is the one, definitive outcome of a simple experiment. For example, the following experiments will have events as there outcomes:

- Select a single card from a set of normal playing cards (resulting event: the card XYZ was drawn)
- flip a fair coin 8 times (resulting event: the coin showed X times heads)
- throw two fair dice (resulting event: the first die shows number X, the second die shows number Y.)

All the possible events that can result from a simple experiment are called the sample space S. We can imagine the sample space as a room whose floor is completely covered by a mosaic, where each stone of the mosaic stands for an event. The room is showed (without mosaic) to the right.

The mosaic room (sadly without mosaic)

One stone of the mosaic is called elementary event. For example, if we are in a room that stands for selecting a single card from a standard deck of cards, there is one stone that stands for the elementary event “The drawn card is diamond 7”.



One class and its complement

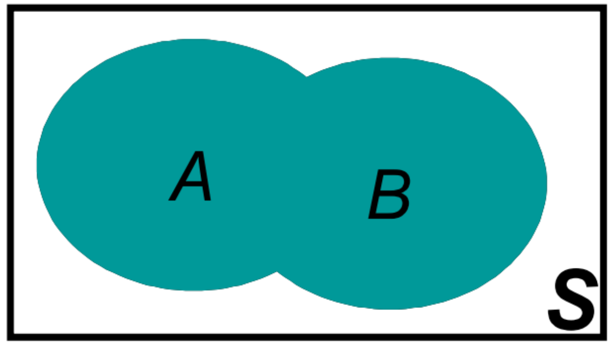
The amount of all elementary events (which we call N), describes the whole room S.

The next thing we want to do is to create classes. Let's say we want to have all diamond-cards in one group. We take a piece of chalk and sorround all mosaic stones that refer to the event of drawing a diamond card by a line. We can give this class any name we want, e.g. 'Frank'. But for no reason, we will rather stick to 'A' here. Everything that is not in class A gets the name 'not A' or '~A'. This 'not-class' is called the complement of A.

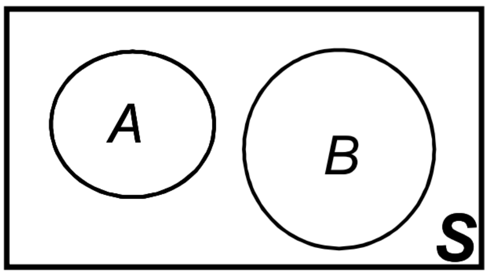
As we still have chalk left, we create another class. This time, we surround all mosaic stones that represent the event of drawing a 7 from the deck of cards. While doing that, we realize that there is also a diamond 7, which already belongs to class A. Therefore our lines will intersect at some point and the card diamond 7 belongs now to two classes: A and (as we most creatively call it) B. Such groups are called 'crosscutting'.

Thanks to our newly created classes, we can now describe a lot of mosaic stones very shortly:

A	All elements in class A (all diamond-cards)
B	All elements in class B (all 7's)
$A \cap B$	All elements that are both in class A and B (all cards that are diamond and 7, i.e. the diamond 7
$A \cup B$	All elements that are in class A or B (all cards that are diamond as well as all cards that are 7's, the blue area to the right)



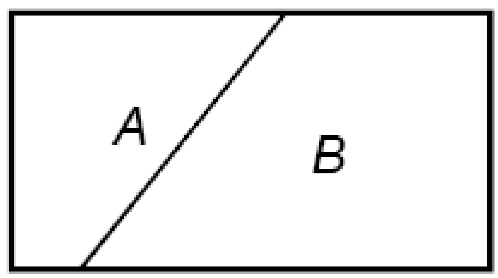
The union of A and B

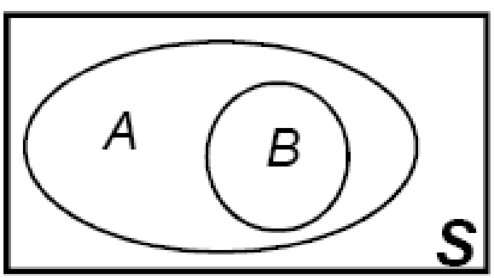


Mutually exclusive classes

But event classes do not always have to intersect: If B were e.g. all cards with the symbol heart, then we would get an image like the one on the left. If we now ask, what the elements in $A \cap B$ are, we do not get any element. This result is called the 'null-set' or $\emptyset$. Any time that two classes have no common elements, they are called 'disjoint' or 'mutually exclusive'.

The groups on the right are an especially peculiar set of mutually exclusive classes, as they are at the same time exclusive and cover the whole sample space S. Groups like that are called exhaustive classes.





One last thing: What if two classes do not only intersect but are contained wholly in one another? In this scenario, B is called a 'subset of A', which means that every element of B is also an element of A. The notation is $B \subseteq A$.

The term 'proper subset' means the same except for one little addition: At least one member of class A is not contained in B, i.e. A is larger and comprises more elements than B (the situation shown in the graphic). The notation is $B \subset A$.

2. A first approximation of probability

As a first approach, we can use the term probability as being equal to 'the relative frequency of an event in the context of all possible events'. Normally, as a variable for that value, we use P(A), which is the probability of A.

There are some rules to get you started off:

Rule	**Explanation**
$P(A) \geq 0$ for all A	The probability of any event is greater then 0, because otherwise there would be no chance at all that this event could occur – and thus it would not be one of the possible events.
$P(S) = 1$	The probability of all possible events equals one, because one of all possible alternatives is going to happen for sure.
$P(A_1 \cup A_2 \cup \ldots) = P(A_1) + P(A_2) + P(\ldots)$ for mutually exclusive events A	If two or more events are mutually exclusive, you can simply add up their separate probabilities in order to get their conjoined probability.
$P(A) = 1 - P(\sim A)$	$P(A) + P(\sim A) = P(S) = 1$, because P(A) and P(~A) are exhaustive.
$P(\emptyset) = 0$	The empty set is the complement of S (thus ~S). $P(S) = 1$, therefore $P(\sim S) = 1 - P(S) = 0$
$P(A \cup B) = P(A) + P(B) - P(A \cap B)$	Think of two intersecting classes in a Venn-diagram. Do not confuse calculating with sets and calculating with their respective probabilities! If you add up P(A) and P(B), you count the probability of those elements that are in both groups twice. Therefore you have to subtract $P(A \cap B)$.
If two events are independent, the following holds because B has no influence on A: $P(A\|B) = P(A)$ This means that $P(A \cap B) = P(A) * P(B)$	$P(A\|B) = P(A \cap B) / P(B)$, and as P(A|B) and P(A) are interchangeable also: $P(A) = P(A \cap B) / P(B)$ immediately leading to the second term.
If $P(A\|B) = P(A)$ holds (the events are independent), then also: $P(B\|A) = P(B)$	Symmetry: If A is independent of B, then B is also independent of A.
Positive Association: $P(A \cap B) > P(A) * P(B)$ Negative Association: $P(A \cap B) < P(A) * P(B)$	

Now let's get our hands at some numbers! If all the events in S have the same probability, then we can simply say that the probability of class A is the number of elements in A divided by the total number of elements:

$$P(A) = \frac{n(A)}{N}$$

Example 1:

What is the probability to select a card of the class 'symbol spades' when drawing randomly from a normal deck of playing cards?

$$P(SymbolSpades)=\frac{NumberOfCardsWithSpades}{TotalNumberOfCards}=\frac{13}{52}=.25$$

Example 2:

What is the probability to draw a card that belongs to both the classes 'being an ace' and 'symbol heart'?

$$P(Ace \cup Heart)=P(Ace)+P(Heart)-P(Ace \cap Heart)=\frac{1}{13}+\frac{1}{4}-\frac{1}{52}=\frac{4}{13}$$

3. The law of large numbers

Bernouilli's theorem states that if you make enough observations at random, the relative frequency with which an event A is observed equals its probability. This is what we intuitively expect: Say we have got three balls, of which three are blue and one is yellow. That means that

$$P(DrawingTheYellowBall)=.25$$

Stubbern, we draw a ball with closed eyes, note the color and put it back. After we have done this three times, we calculate

$$\frac{YellowBallsDrawn}{AllBallsDrawn}$$

and may be very upset, because it doesn't equal .25 . But as we repeat this process 100 or 1000 times, we realize that the calculated values gets closer and closer to the expected .25 . Bernouilli proved that this difference to the probability becomes arbitrarily small if accordingly many draws are made.

Still, do remember that this arbitrarily small probability still exists. There is a very small probability that in the previous experiment, your calculated result is 0.6. Or 1.0. That is very unlikely, but not impossible, thus we can only give <u>approximations</u> of probability by conducting experiments and counting.

4. Conditional Probabilities

Whenever we talk about conditional probabilities, we think of a situation where one parameter is already given, and we want to estimate another. Think about the following: A bucket full of balls is marked with A, B or with both. We draw one ball randomly with closed eyes. A friend tells us, that the ball drawn has an A on it – what is the probability that it also has a B?

We would write the following: $$P(B|A)=\frac{P(A \cap B)}{P(A)}$$

The first part is read "The probability of B, given A". In other words, you are asking "What proportion of elements in B are also in A?".

We can say some more things about conditional probabilities:

$$P(A|B)=P(B|A),\ if\ P(A)=P(B)$$

5. Bayes' Theorem

Bayes' theorem is a way to 'translate' one conditional probability into another – or, more precise, how we can define P(A|B) with P(B|A):

$$P(A|B)=\frac{P(B|A)\cdot P(A)}{P(B)}$$

Example:

There are three baskets with balls in them.

- In the first basket, there are 90% white balls and 10% black balls
- In the second basket, there are 50% white balls and 50% black balls
- In the third basket, there are 10% white balls and 90% black balls

You close your eyes and another person draws a ball from any of the three baskets and tells you the color – it is black. What is the probability that the ball was drawn from basket 1, 2 or 3, respectively?

Remember, we want to put that question into terms of conditional probabilities. Thus we could express the questions above as:

$$P(Basket_1|BlackBall)$$
$$P(Basket_2|BlackBall)$$
$$P(Basket_3|BlackBall)$$

However, what is given above is the probability of getting a black ball, given a specific basket, which we can rewrite as the following:

$$P(BlackBall|Basket_1)=0.1$$
$$P(BlackBall|Basket_2)=0.5$$
$$P(BlackBall|Basket_3)=0.9$$

If we know $P(Basket_1), P(Basket_2), P(Basket_3)$ and $P(BlackBall)$, then Bayes' theorem gives us the way how to calculate the result:

$$P(Basket_1|BlackBall)=\frac{P(BlackBall|Basket_1)\cdot P(Basket_1)}{P(BlackBall)}$$

Well, $P(Basket_i)$ is easy: We have three baskets, with equal probability to be selected, therefore $P(Basket_1)=P(Basket_2)=P(Basket_3)=0.33$.

Next, P(BlackBall) can be calculated from

$$P(BlackBall)=\sum_i P(BlackBall|Basket_i)\cdot P(Basket_i)=0.1\cdot\frac{1}{3}+0.5\cdot\frac{1}{3}+0.9\cdot\frac{1}{3}=(0.1+0.5+0.9)\cdot\frac{1}{3}=0.5$$

Now we can insert those two values in the term above:

$$P(Basket_1|BlackBall)=\frac{P(BlackBall|Basket_1)\cdot P(Basket_1)}{P(BlackBall)}=\frac{0.1\cdot\frac{1}{3}}{0.5}=0.067$$

And this is the probability that the black ball was drawn from Basket 1.

Bayes' theorem can be extended to two or more balls that are drawn from the baskets. For example, when we want to calculate the probability that two black balls were drawn from a specific basket H_i :

$$P(H_i|BB)=\frac{P(B|H_i)\cdot P(H_i|B)}{\sum_j P(B|H_j)P(H_j|B)}$$

Bayes' theorem is useful where something new is learned about the situation, which changes one's belief about the probabilities of the respective hypotheses.

However, do always remember that this theorem expects the events to be mutually exclusive and exhaustive, which is not always the case.

6. Additional fun with Bayes' theorem

So far, we have only used the theorem to calculate probabilities. However, it turns out that it can also be used to infer on dependence and exclusiveness of variables.

Look at the following diagrams. We can immediately tell the values of the respective parameters, just by looking at them:

$$P(A)=.5$$
$$P(B)=.5$$
$$P(A\cap B)=.25$$
$$P(A|B)=\frac{.25}{.5}=.5$$
$$P(B|A)=\frac{.25}{.5}=.5$$

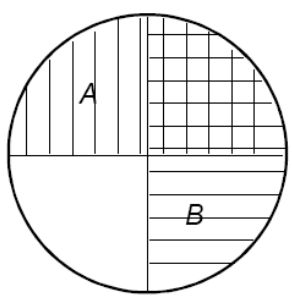


Diagram 1

$$P(A)=.5$$
$$P(B)=.5$$
$$P(A\cap B)=.125$$
$$P(A|B)=\frac{.125}{.5}=.25$$
$$P(B|A)=\frac{.125}{.5}=.25$$

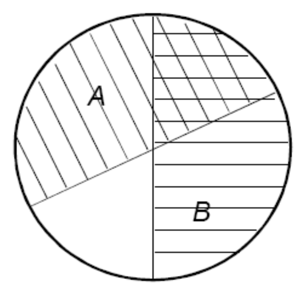


Diagram 2

As you might recall, the following holds for independent A and B:

$$P(A|B)=P(A)\wedge P(B|A)=P(B)$$

Check it! The condition obviously holds for diagram 1, but not for diagram 2. That means that A and B are independent in the first case whereas they are dependent on each other in the second case.

As the first diagram also shows, independent does not mean the same as mutually exclusive. The classes in diagram 1 are independent, but not mutually exclusive. Similarly, diagram 3 shows that classes that are mutually exclusive do not have to be independent:

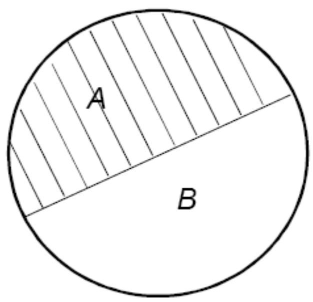


Diagram 3

$P(A)=.5$

$P(B)=.5$

$P(A \cap B)=0$ (because they are mutually exclusive)

$P(A|B)=\frac{0}{.5}=0$

$P(B|A)=\frac{0}{.5}=0$ (because by knowing A (or B) we do not get any additional information on B (or A).

Still, $P(A) \neq P(A|B)$, which means that the two classes are not independent.

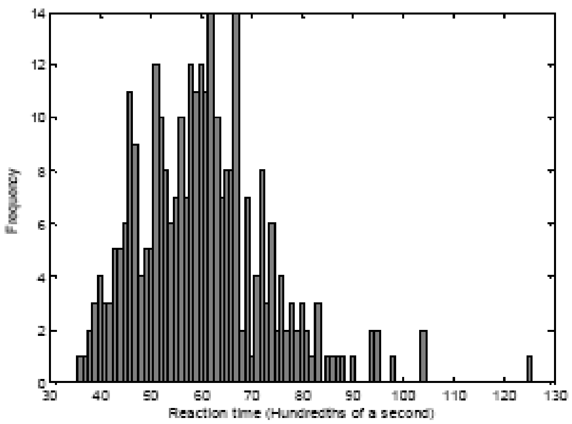
Chapter III: Describing and Exploring Data

Suppose we have conducted an experiment in a person had to respond to a given stimulus as fast as he could. We measured the time he needed to respond under 6 different conditions. This gave the following table:

Condition	Reaction Times, in 100ths of a Second																
	40	41	47	38	40	37	38	47	45	61	54	67	49	43	52	39	46
1	47	45	43	39	49	50	44	53	46	64	51	40	41	44	48	50	42
	90	51	55	60	47	45	41	42	72	36	43	94	45	51	46	52	
	52	45	74	56	53	59	43	46	51	40	48	47	57	54	44	56	47
2	62	44	53	48	50	58	52	57	66	49	59	56	71	76	54	71	104
	44	67	45	79	46	57	58	47	73	67	46	57	52	61	72	104	
	73	83	55	59	51	65	61	64	63	86	42	65	62	62	51	62	72
3	55	58	46	67	56	52	46	62	51	51	61	60	75	53	59	56	50
	43	58	67	52	56	80	53	72	62	59	47	62	53	52	46	60	
	73	47	63	63	56	66	72	58	60	69	74	51	49	69	51	60	52
4	72	58	74	59	63	60	66	59	61	50	67	63	61	80	63	60	64
	64	57	59	58	59	60	62	63	67	78	61	52	51	56	95	54	
	39	65	53	46	78	60	71	58	87	77	62	94	81	46	49	62	55
5	59	88	56	77	67	79	54	83	75	67	60	65	62	62	62	60	58
	67	48	51	67	98	64	57	67	55	55	66	60	57	54	78	69	
	66	53	61	74	76	69	82	56	66	63	69	76	71	65	67	67	55
6	65	58	64	65	81	69	69	63	68	70	80	68	63	74	61	85	125
	59	61	74	76	62	83	58	72	65	61	95	58	64	66	66	72	

A datatable, on which no operations have been performed so far is called 'raw data'. The disadvantage of such a table is that we get no overview of the data and it is thus very hard to interpret it.

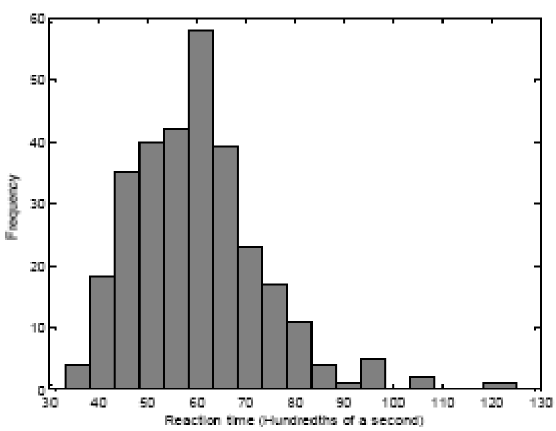
A first thing we can do to order the results is to create a frequency table: We count how often each result occured. Such a frequency table can be easily plotted: We create a bar for each reaction time value and adjust its height according to the respective frequency.



First bar plot of the frequency table

Reaction Time, in 100ths of a Second	Frequency	Reaction Time, in 100ths of a Second	Frequency
36	1	71	4
37	1	72	8
38	2	73	3
39	3	74	6
40	4	75	2
41	3	76	4
42	3	77	2
43	5	78	3
44	5	79	2
45	6	80	3
46	11	81	2
47	9	82	1
48	4	83	3
49	5	84	0
50	5	85	1
51	12	86	1
52	10	87	1
53	8	88	1
54	6	89	0
55	7	90	1
56	10	91	0
57	7	92	0
58	12	93	0
59	11	94	2
60	12	95	2
61	11	96	0
62	14	97	0
63	10	98	1
64	7	99	0
65	8	...	...
66	8	...	...
67	14	104	2
68	2	...	...
69	7	125	1
70	1		

The frequency table

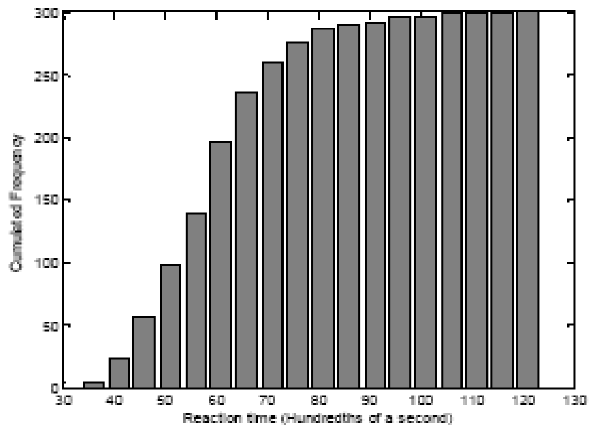


Second, grouped bar plot

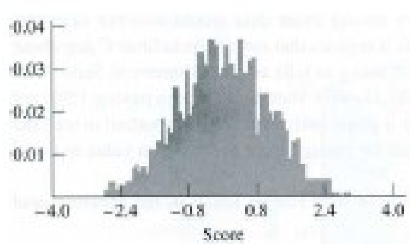
To really get a broad overview on what is going on, the first bar plot might still be a little too fine-grained. We can create another one that represents the same data, but uses one bar for several values.

Another way to display the same information is the cumulative frequency distribution: The frequencies of values in the frequency table are added up in a manner that now every value gives the frequency of one specific reaction time *and those below*. Again, this table can be plotted in a bar diagram.

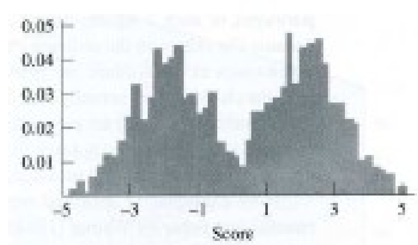
Interval	Midpoint	Frequency	Cumulative Frequency	Interval	Midpoint	Frequency	Cumulative Frequency
35–39	37	7	7	85–89	87	4	291
40–44	42	20	27	90–94	92	3	294
45–49	47	35	62	95–99	97	3	297
50–54	52	41	103	100–104	102	2	299
55–59	57	47	150	105–109	107	0	299
60–64	62	54	204	110–114	112	0	299
65–69	67	39	243	115–119	117	0	299
70–74	72	22	265	120–124	122	0	299
75–79	77	13	278	125–129	127	1	300
80–84	82	9	287				



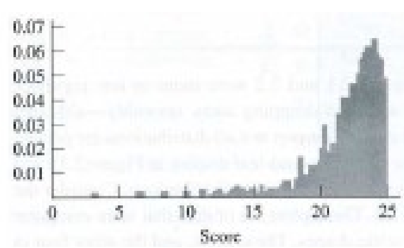
1. Shapes of distributions



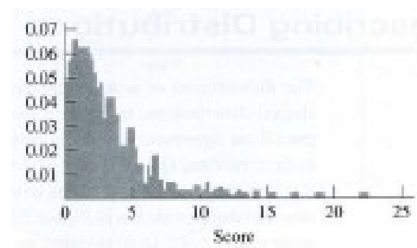
Unimodal



Bimodal



negatively skewed



positively skewed

2. Measures of Central Tendency

1) **Mode**: The score in the distribution that is most common. In the first bar plot, it would be the value corresponding to the highest bar.
2) **Median**: The score that corresponds to the point on or under which 50% of all scores fall if they are arranged in numerical order.
3) **Mean**: The sum of the scores divided by the number of scores: $\bar{X} = \frac{\sum X}{N}$

3. Measures of Variability

1) **Range**: The distance from the lowest to the highest score (suffers a lot from outliers)
2) **Interquartile Range**: Discard the upper and lower 25 % of the data in numerical order and give the range of the remains. This technique suffers less from outliers.
3) **Variance**: The averaged squared deviation from the mean:

   $$s_x^2 = \frac{\sum (\bar{X} - X)^2}{N-1}$$

   The variance of a sample is called s^2 , while the variance of a population is called σ^2 . We will come to this difference again later.

The descriptive measures are a good means to give a first description of data. However, they will get even more handy when we try to estimate population statistics by sample statistics.

4. Creating and modifying distribution plots

One commonly used type of distribution is the frequency distribution. Whenever we can group elementary events into classes, we can also create a frequency distribution, that displays the frequency of events within each class.

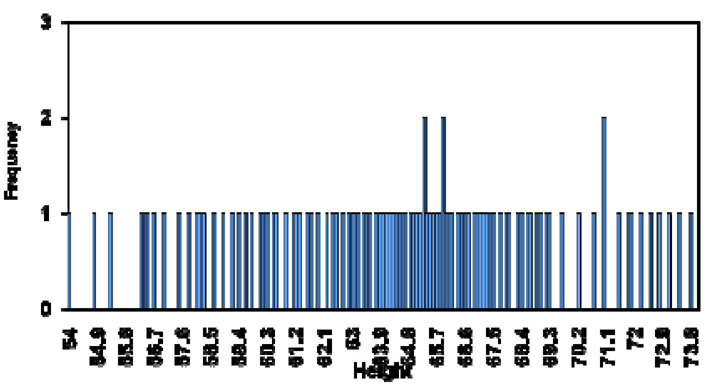
Obviously, those frequency distributions are important to infer on the probabilities of events belonging to the classes – if a certain class has a high frequency its events seem to be more likely to happen than those of a class with lower frequency.

The classes of a frequency distribution have to be mutually exclusive and exhaustive.

If the classes weren't exclusive, we would not be able to infer from the frequencies on the probabilities, because the frequencies counted would be higher than the total number of elementary events.

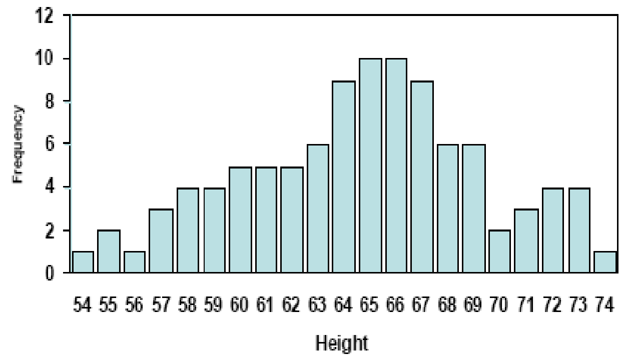
If the classes weren't exhaustive we would not get a complete overview of all possible classes.

But how are classes created? Suppose we measure a variable like height. This variable can be measured arbitrarily precise, e.g. in tens or hundreds of inches. If you decide to measure in hundreds of inches, you already create classes that are 1/100-inch of size each, in which you put all heights that are *approximately* the same.



The plot on the left shows the result of a measurement with such classes. As you can see, the classes are as fine-grained that there are hardly two individuals that both belong in the same class. The plot is not very informative, because it doesn't show tendencies of larger groups. The same problem occurs when we have only few groups: E.g. only two groups are too coarse-grained to show tendencies within the data.

For the plot above, it turns out that a grouping into one-inch classes gives a nice overview on the data.



Probability distributions are very similar to frequency distributions, but display the probabilities of events belonging to one class. Like in frequency distributions, the classes have to be mutually exclusive and extensive. This means that the following holds:

- $P(A)=\frac{freq(A)}{N}$ (the probability of an event A is the frequency of that event, divided by the total number of events)
- $P(S)=\sum_i P(A_i)=1$ (the cumulative probability of the whole sample space is the sum of the probabilities of all classes, where i is the number of classes. The probability of the whole sample space is always 1.

5. Random Variables

Up to now, we have simply displayed whole populations or samples. We will now have a look at a means to tune this output to a certain parameter we want to investigate.

In order to do that, we introduce 'Random Variables'. These variables stand for certain functions, and those functions select a subset of our population that should be investigated.

Let's look at a particular distribution to make that a bit clearer. Imagine you are flipping a coin three times. As with all coins, the result of each coin toss can be either Heads (H) or Tails (T). Therefore, we can get the following results:

TTT	HTT	THT	TTH
HHT	HTH	THH	HHH

Now, out of nothing, we create a random variable X. Let X be a function that tells us for each event in the population above how many times heads showed up. That's easy, the function would do the following:

X(TTT) = 0	X(HTT) = 1	X(THT) =1	X(TTH) = 1
X(HHT) = 2	X(HTH) = 2	X(THH) = 2	X(HHH) = 3

The next thing we are going to do is to select subsets from all events with the aid of the variable. For example, we can easily describe all events where no head was shown as:

P(X=0) = P(TTT) = 1/8 (this is the probability for any one given sequence of three flips, $(1/2)^3$.

Similarly, we do for:

P(X=1) = P(HTT or THT or TTH) = 3/8

P(X=2) = P(HHT or HTH or THH) = 3/8

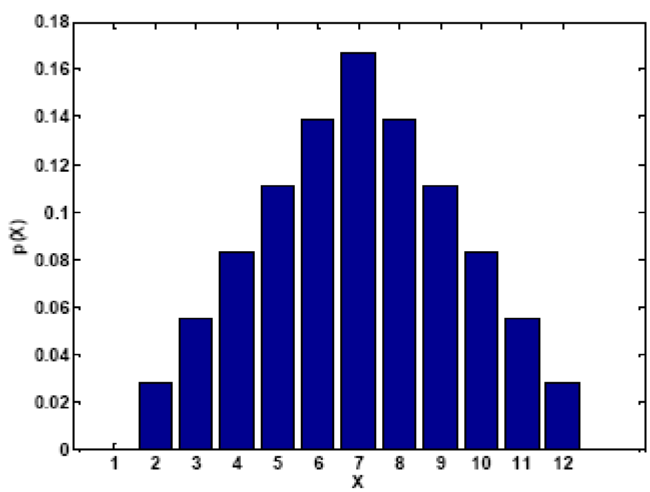
P(X=3) = P(HHH) = 1/8

Like this, any subset of events can be specifically looked at with the help of the variable function. You can even go further and ask for P(X>1), the probability that 'more than one head showed up', and so forth.

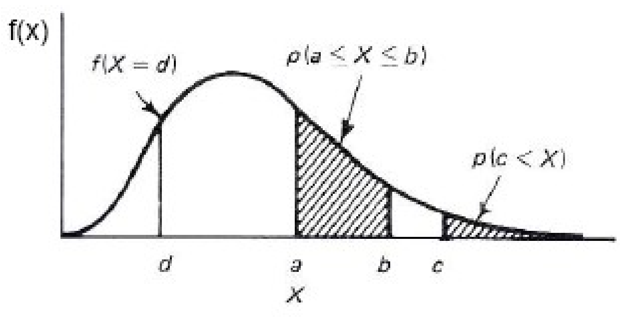
Random variables can be either discrete or continuous. The first means, that the variable can take on only certain specific numbers and not others. In our example before X was discrete, because it could take on number from 0 to 3, but not e.g. the value 2.5 . The latter means that the variable can take on all possible values and its function returns a meaningful value for all of them.

6. Probability density functions

Probability density functions are generally a means to calculate the probability of a certain value being assigned to a random variable. However, the word means two different things, depending on whether we are talking about discrete or continuous distributions.



The graph on the right clearly denotes a discrete function, as specific values are assigned specific probabilities, and other values are not. As a matter of fact, P(X) denotes the probability that X dots show up when two dice are thrown.



The next graph however is clearly continuous. We can imagine that the random variable X can take on any given number, like e.g. 3.8172152, and this number would be represented somehow in the picture. You can imagine the creation of such a distribution like that (although that is not very scientifically accurate): You start with a distribution that looks discrete, like the one above. You however know that each of your 12 classes represents an infinite number of values, and that your plot might not be accurate enough with only such a limited amount of classes. Then you realize that you are able to put in more classes, e.g. 24. Thus you can say what the probability is that X falls e.g. in the interval 7.5 – 8. Then you want to make distinctions even more fine-grained and create 100 classes, 10000, and finally, infinitely many.

It's only then that you realize you can't calculate the probability of one specific value anymore – you have infinitely many classes which means that each class is infinitely small, i.e. strictly speaking it is too small for any values to fall into it. Therefore the function f at any given value X is zero. (By the way, now is the appropriate time to grab your head and cry out loud „What have I done?!?“).

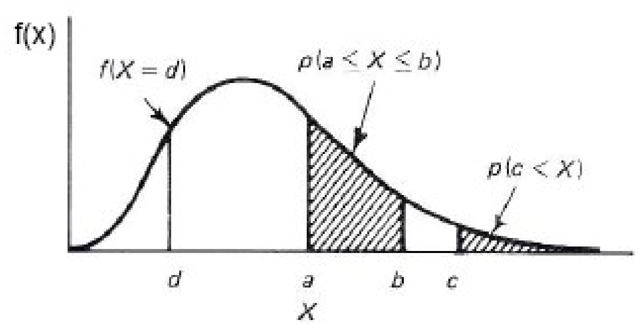
But there's a way out: You can integrate over any given interval in order to get real values again. If you, for example, want to know the probability that your variable X takes on any value between a and b, that is, $p(a \leq X \leq b)$, you can integrate over your probability density function from a to b:

$$\int_a^b f(x)\,dx$$

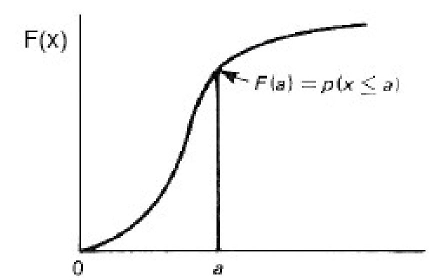
7. Cumulative distribution functions

The cumulative distribution function is directly related to the probability density function, in that it shows for any value a, what the probability is that the random variable X takes on a value smaller or equal to a. Or to put it more formally:

$$F(a) = p(X \leq a) = \int_{-\infty}^{a} f(x)\,dx$$



The probability density function



The cumulative distribution function (CDF)

It can be easier to use the CDF, because you do not need to integrate any longer. If you want to know the value of $p(X \leq a)$, you simply look up F(a) and you are done. Of course, if you want to know the value of $p(X > a)$, you can also simply calculate 1 - F(a) and you are done.

For any value range like $p(a \leq X \leq b)$, the term is F(b) – F(a).