biologist visiting once, at any given hour, will meet only 2 fishermen, one fishing for one hour and one fishing for 8 hours.

The arithmetic mean of the sample = (8+1)/2 = 4.5 hours

The harmonic mean of the sample = 1/((0.125+1)/2) =1.778

This type of mean is appropriate for samples where the probability of being included in the sample is a function of the value being measured.

Summary

The true population mean is denoted μ, the greek letter mu.

The sample estimate of the mean is denoted $\overline{Y}$, and is called "y-bar".

Remember always that our sample estimate of the mean is just one of many possible samples. Each sample is an attempt to estimate the true population mean. Our sample may be one of the good ones, pretty close to the true population mean, or it may be one of the not so good ones. We won't really know.

How good is our estimate from the sample? This depends on how good our sample is and on how much variability there is in the population.

Our best guarantee of getting a good sample is to sample at random. This should at least give us a representative sample of the population.

Variability is the other big problem in sampling, so we need to estimate how variable the population is, our next topic.

Applications for other types of means

Arithmetic mean – the usual case

Geometric mean – Used as a transformation for some non-normal distributions, particularly the negative binomial, a strongly skewed distribution.

Harmonic mean – Used for particular cases where the probability of being sampled is an inverse function of the variable of interest.

SAS example (#1a) from Freund & Wilson (1997) Table 1.1

```
PROC UNIVARIATE DATA=HouseSales PLOT; VAR SP;
    TITLE4 'Proc Univariate of house sales price'; RUN;
```

See SAS output for results

Probability distributions

PROBABILITY – a measure of the likelihood of the occurrence of some event

An event can be any outcome (e.g. verbal, mathematical or graphical)

Some rules of Probability

If an event (A) is certain to occur, the probability is 1 (one, unity), so P(A) = 1

If the event is certain to NOT occur, the probability is 0 (zero, null), so P(A) = 0

The probability of an event will always range be between 0 and 1 (inclusive). $0 \leq P(A) \leq 1$

The sum of the probability of all possible events, where the events are mutually exclusive, is one (1).

Where a number of mutually exclusive events are denoted A_i, for i = 1, 2, ..., r, $\Sigma P(A_i) = 1$ when summed across all of the possible events

Note that for truly continuous variables the probability of a given number is zero (0).

The Binomial Distribution

First example of a distribution

A binomial distribution consists of a set of binomial observations or Bernoulli trials. These are observations with two possible, mutually exclusive, outcomes.

Examples of binomial observations or Bernoulli trials, where each observation takes one of two possible values.

- Have you taken EXST7005? Yes, No
- A given fish is: Male, Female
- A given fish is: Tagged, Untagged
- A coin toss is: Heads, Tails

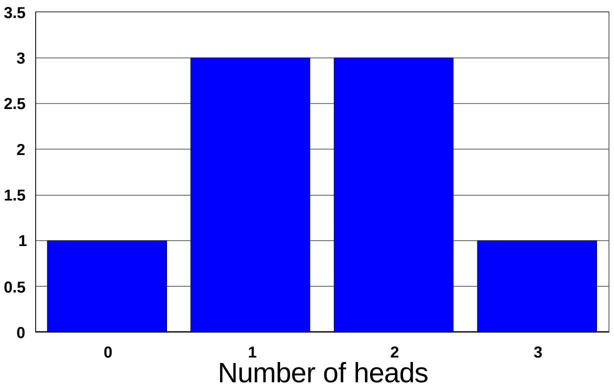
Our first experiment, toss three coins, each coin having two possible outcomes. For the three coins together there are 8 possible outcomes. We are interested in the distribution of the outcomes.

Outcome	Coin 1	Coin 2	Coin 3	Frequency of heads
1	Tail	Tail	Tail	0
2	Tail	Tail	Head	1
3	Tail	Head	Tail	1
4	Head	Tail	Tail	1
5	Tail	Head	Head	2
6	Head	Tail	Head	2
7	Head	Head	Tail	2
8	Head	Head	Head	3

Note that each event is equally likely and mutually exclusive. Prepare a frequency table of the results.

Number of Heads	frequency (f)	relative frequency (r.f.)	Probability
0	1	1/8	P(0)=0.125
1	3	3/8	P(1)=0.375
2	3	3/8	P(2)=0.375
3	1	1/8	P(3)=0.125
Total	8	1	1

James P. Geaghan Copyright 2012



This chart represents the distribution of all possible outcomes of tossing 3 coins, each with a binomial outcome. This is the binomial distribution.

Probability defined: If an event can occur in "n" mutually exclusive and equally likely ways, and if "m" of these ways hold the attribute "A", the probability of the occurrence of "A" will be the ratio of "m" to "n".

Where A is some particular attribute

n is the number of possible outcomes (Trials)

m is the number of ways A can occur (Successes)

Then P(A) = m / n , or the number of successes over the number of trials

Example from our 3 coins, find the probability of event A where A is the attribute "2 heads".

n = 8 possible outcomes (HHH, HHT, HTH, THH, TTH, THT, HTT, TTT)

the outcomes are equally likely.

m = 3 outcomes with the chosen attribute (2 heads)

P(A) = m / n = $^3/_8$ = 0.375

Working with Probabilities

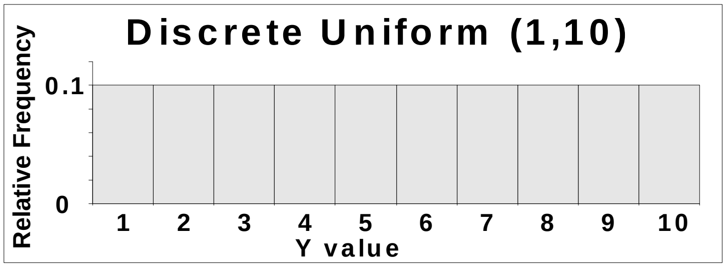
We will be working first with Probability Distributions, similar to the bar charts we examined earlier.

The probability will be the proportion of the area under the graph between given limits.

The probability is the relative frequency of occurrence of observations within the set limits.

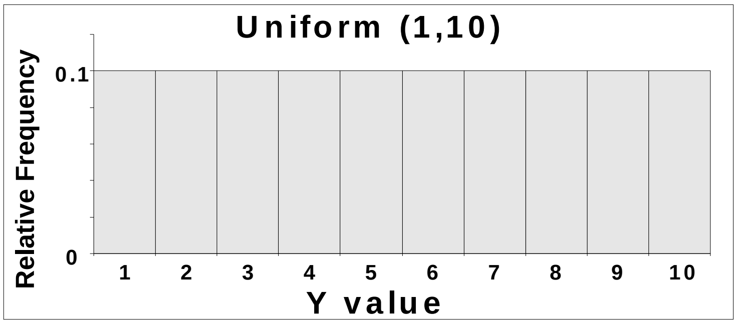
The Uniform Distribution

A uniform distribution is a distribution where every outcome has an equal probability of occurrence.



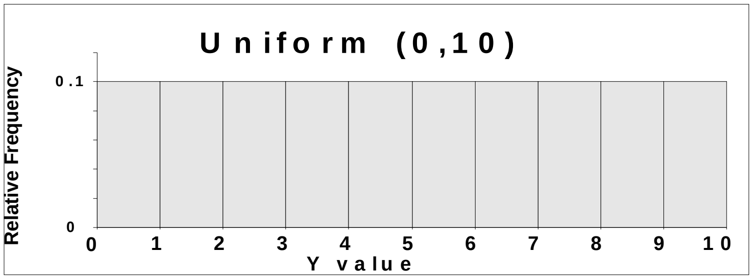
We will consider two similar uniform distributions, discrete and continuous.

Discrete Uniform Distribution (1, 10)

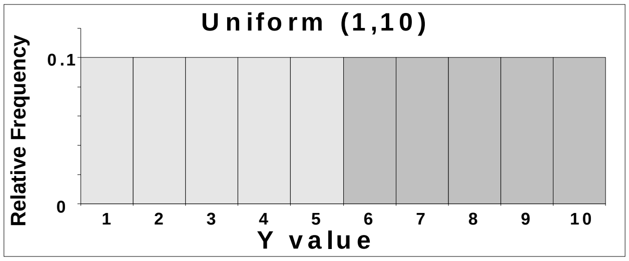


The probability in each cell is $^1/_{10} = 0.1$.

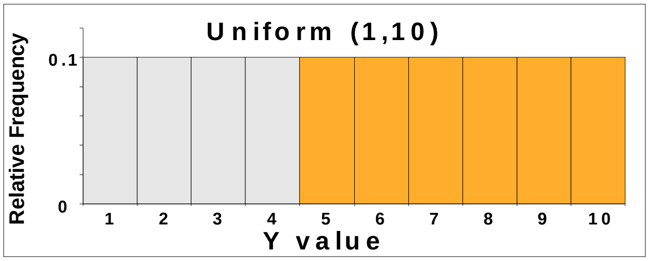
Continuous Uniform (0, 10): The probability between any two integers is $^1/_{10} = 0.1$



Finding Probabilities from the DISCRETE Uniform Distribution



Find $P(Y_i) > 5$, Note that 5 itself is excluded.



Find $P(Y_i) \geq 5$, Now 5 is included.

Find the following probabilities for a discrete Uniform (1,10) distribution.

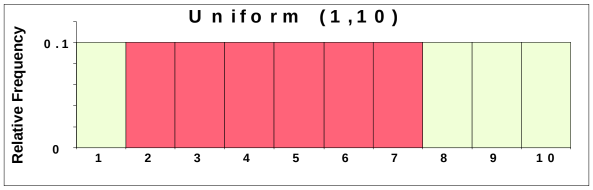
a) $2 \leq P(Y_i) \leq 7$

b) $P(Y_i) = 9$

c) $P(Y_i) \geq 9$

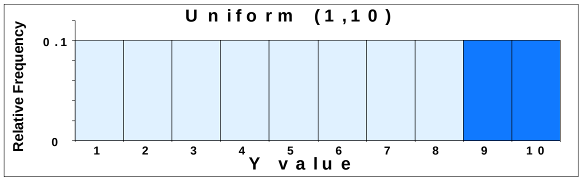
d) $P(Y_i) > 10$

Finding Probabilities from the Uniform Distribution



a) $2 \leq P(Y_i) \leq 7$

b) $P(Y_i) = 9$ This is the probability of a single cell.



c) $P(Y_i) \geq 9$

d) $P(Y_i) > 10$ This probability is zero

Find the following probabilities for a discrete Uniform (1,10) distribution.

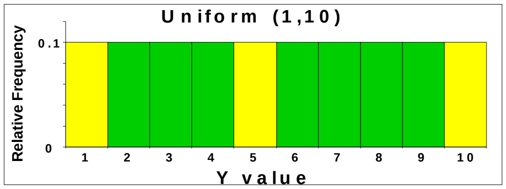
a) $2 \leq P(Y_i) \leq 4$ OR $6 \leq P(Y_i) \leq 9$

This type of statement is true if either of the two statements is true, so the individual probabilities are added

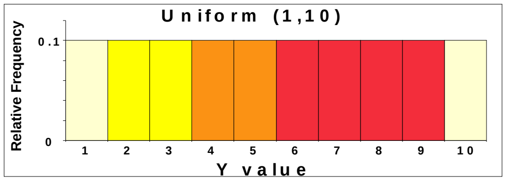
b) $2 \leq P(Y_i) \leq 5$ AND $4 \leq P(Y_i) \leq 9$

This type of statement is true only if BOTH of the statements are true; we determine the area overlap between the two statements

Finding Probabilities from the Uniform Distribution



$2 \leq P(Y_i) \leq 4$ OR $6 \leq P(Y_i) \leq 9$



$2 \leq P(Y_i) \leq 5$ AND $4 \leq P(Y_i) \leq 9$

Finding Probabilities from the CONTINUOUS Uniform Distribution

Find the following probabilities for a continuous Uniform (0,10) distribution.

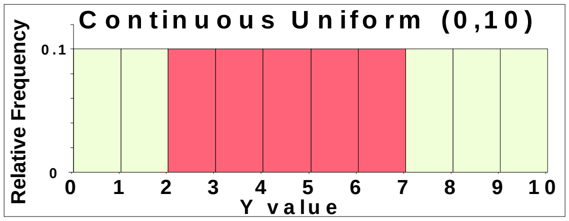
This is a little trickier because we don't just count cells; we consider the range between limits.

a) $2 \leq P(Y_i) \leq 7$

b) $P(Y_i) = 9$

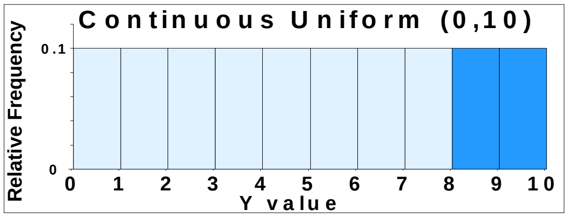
c) $P(Y_i) \geq 8$

The Continuous Uniform Distribution



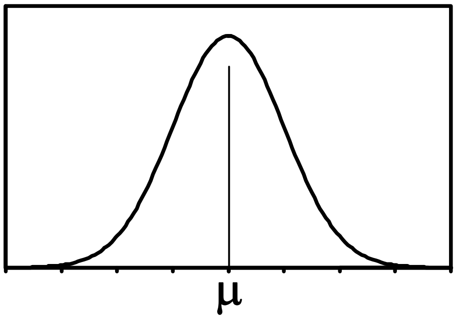
a) $2 \leq P(Y_i) \leq 7$

b) $P(Y_i) = 9$ This probability is zero



c) $P(Y_i) \geq 8$

The Normal Distribution: N(μ, σ^2)



The equation for the Normal Distribution is: $f(y) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(Y-\mu)^2}{2\sigma^2}}$

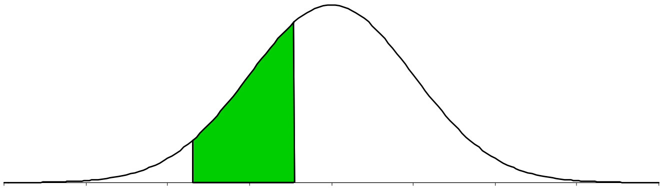
Note that there are two separate and distinct parameters, mu (μ) and sigma (σ)

Characteristics of the Normal Distribution

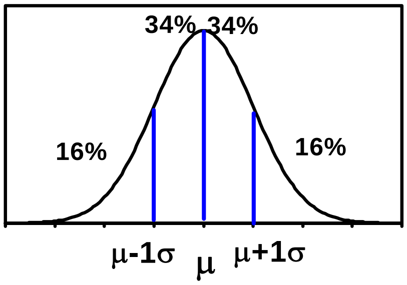
For a variable distributed normally N(μ, σ^2)

- The distribution is symmetric about the mean
- The distribution has only two parameters

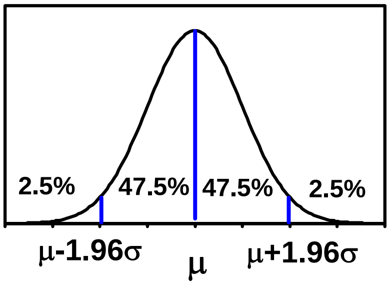
The probability that a random observation will fall within specified limits is given by the area under the curve between those limits.



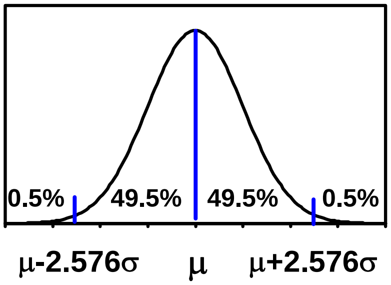
The middle 68% of the distribution is included in the interval $\mu \pm 1\sigma$



The middle 95% of the distribution is included in the interval $\mu \pm 1.96\sigma$



The middle 99% of the distribution is included in the interval $\mu \pm 2.576\sigma$



Examples from the Normal Distribution

The knowledge of the ranges on the previous pages allows us to make some probability statements from the normal distribution.

Suppose we are examining the height (in inches) of adult males. For the particular population of interest, the mean is μ = 5' 10" = 70"

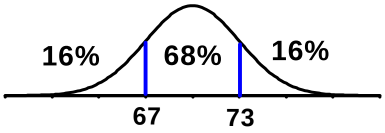
The standard deviation, σ = 3"

The middle 68% of the population of all individuals is between what limits?

From out previous discussion we know that 68% fall between $\mu \pm 1\sigma$.

$\mu \pm 1\sigma = 70 \pm 1(3)$

So, the lower limit is 70 – 3 = 67 and the upper limit is 70 + 3 = 73,



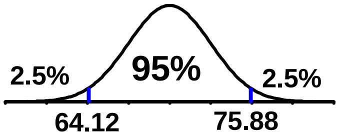
and we can state that $P(67 \le Y \le 73) = 0.68$

95% of individuals are included in what interval? From our previous discussion we know that 95% fall between $\mu \pm 1.96\ \sigma$.

$\mu \pm 1.96\sigma = 70 \pm 1.96(3)$

So, the limits are 70 – 5.88 = 64.12 and 70 + 5.88 = 75.88

$P(64.12 \le Y \le 75.88) = 0.95$

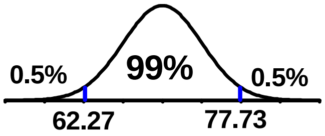


99% of all individuals are included in what interval? From out previous discussion we know that 99% fall between $\mu \pm 2.576\ \sigma$.

$\mu \pm 2.576\sigma = 70 \pm 2.576(3)$

So, the limits are 70 – 7.73 = 62.27 and 70 + 7.73 = 77.73

$P(62.27 \le Y \le 77.73)=0.99$



The empirical rule

Sometimes refered to as the three sigma rule, it states that approximately 68% and 95% of the observations are within one and two standard deviations of the mean, respectively.

Nearly all of the observations (99.74%) will be within 3 standard deviation units of the mean.

Other distributions

Previously mentioned were the

Binomial (a discrete distribution)

Where π is the true population probability of an event (estimated in a sample by "*p*"), and where n is the sample size;

- Mean = $n\ \pi$ (for a sample, Mean = np)
- Variance = $n\ \pi(1-\pi)$ (for a sample, Var = $np(1-p)$)

note that the variance is less than the mean

Uniform (can be either discrete, but most of our distributions will be continuous)

- Mean = (Max + Min)/2
- Variance (continuous) = $(\text{Max–Min})^2/12$ **Corrected**
- Variance (discrete) = $((\text{Max–Min}+1)^2–1)/12$

Normal (a continuous distribution)

- Mean μ
- Variance σ^2

the variance and mean are two distinct parameters

Poisson – a discrete distribution

- Mean = λ
- Variance = λ

a single parameter describes both variance and the mean

Negative binomial – a discrete distribution with a parameter *k* that provides an index of dispersion.

- Mean = μ
- Variance = $\mu + k\mu^2$

the variance is greater than the mean

Log normal – a continuous distribution.

The logarithm of the values in this distribution are normally distributed.

Standard normal – a normal distribution with mean = 0 and variance =1

The distributions that we will be most concerned with are the normal and the standard normal.

Measures of dispersion

Our first major objective is to develop the concepts needed to understand hypothesis testing. We will primarily test hypotheses about means, but variances can also be tested. Testing means will require a measure of the dispersion or variability in the data set, so testing both means and variances requires knowledge of variance.

The following presents some measures of variation or variability among the elements (observations) of a data set

- Range – difference between the largest and smallest observation

This is a rough estimator which does not use all of the information in the data set.

- Interquartile range – difference between the third and first quartile ($Q_3 - Q_1$)

Recall that the first quartile (Q_1) is the value that has one quarter of the observations with lesser values and the third quartile has three quarters of the observations with lesser values. This may be a better measure of variability than the range in most situations because the range can be influenced by a single unusually large or unusually small value. However, this measure also does not use all of the information in the data set.

- Variance – the "average" squared deviation from the mean,

The Population Variance is σ^2 (called "sigma squared")

This is a parameter, and therefore a constant

The variance is given by $\sigma^2 = \frac{\sum_{i=1}^{N}(Y_i - \mu)^2}{N}$ where *N* is the size of the population

S^2 is the Sample Variance (called "S-squared").

This is a statistic, and therefore a variable