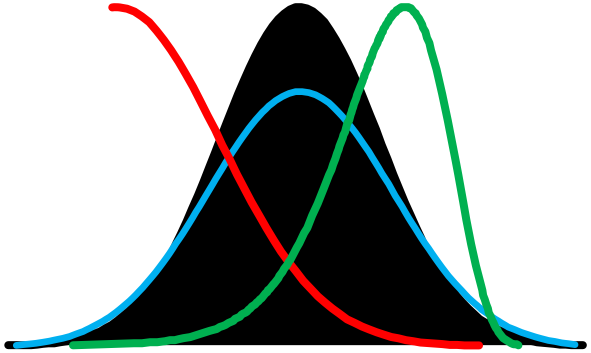
Continuous Improvement Toolkit

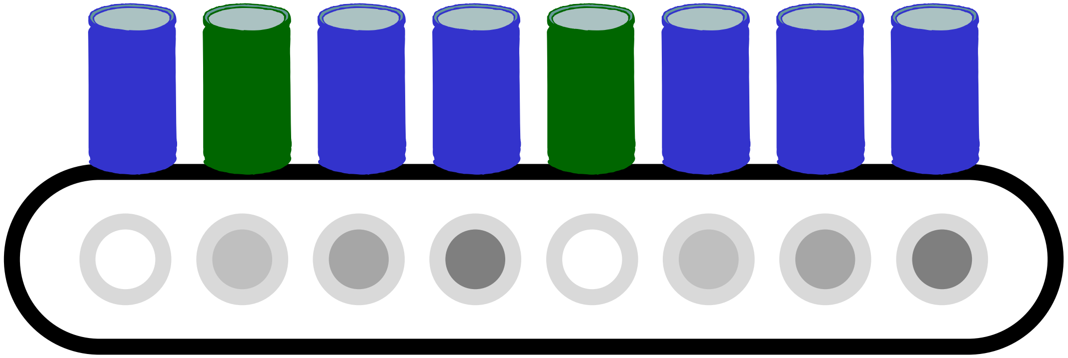
Probability Distributions



The Continuous Improvement Map

- Probability Distributions

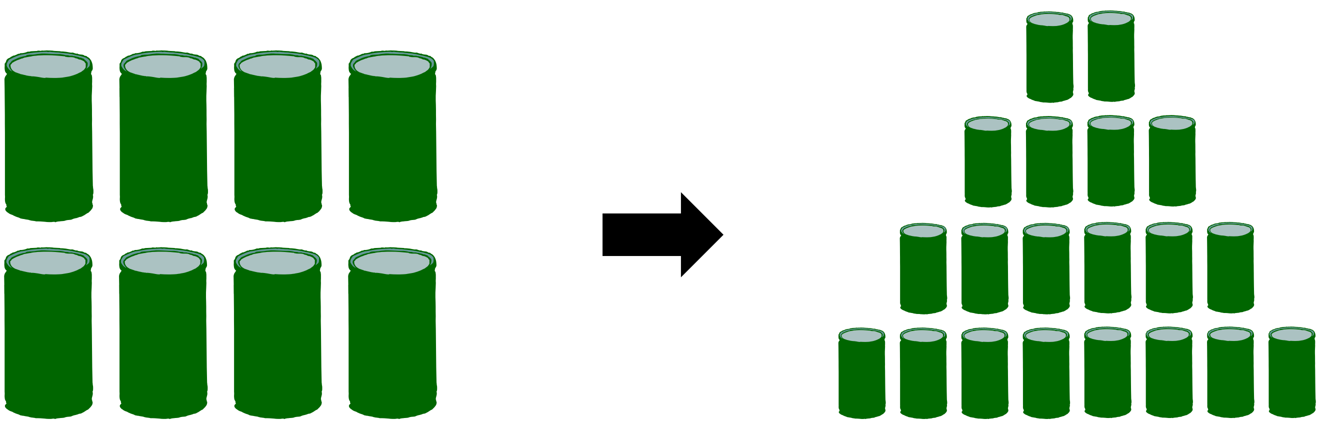
- Most improvement projects and scientific research studies are conducted with **sample** data rather than with data from an entire **population**.



- Probability Distributions

What is a **Probability Distribution?**

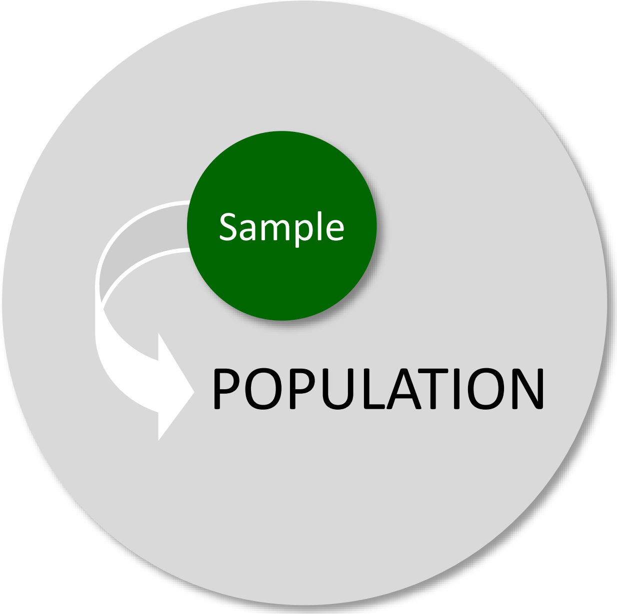
- It is a way to shape the sample data to make predictions and draw conclusions about an entire population.
- It refers to the **frequency** at which some events or experiments occur.



Can height

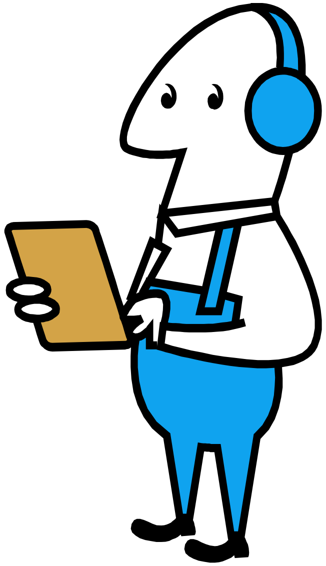
- Probability Distributions

- Helps finding all the possible values a random variable can take between the minimum and maximum possible values.
- Used to model real-life events for which the outcome is uncertain.
- Once we find the appropriate model, we can use it to make **inferences** and **predictions**.



- Probability Distributions

- **Line managers** may use probability distributions to generate sample plans and predict process yields.
- **Fund managers** may use them to determine the possible returns a stock may earn in the future.
- **Restaurant mangers** may use them to resolve future customer complaints.
- **Insurance managers** may use them to forecast the uncertain future claims.

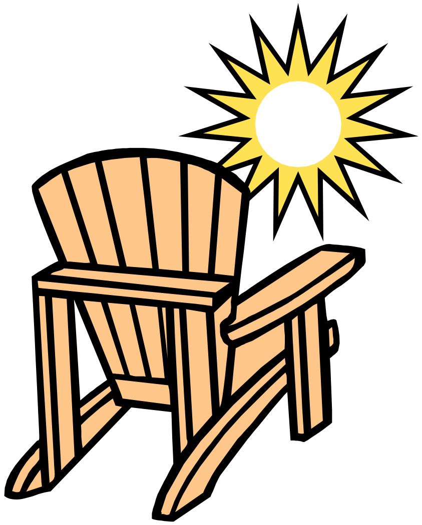


- Probability Distributions

- **Probability runs on a scale of 0 to 1.**
- If something could never happen, then it has a probability of 0.
  - For example, it is impossible you could breathe and be under water at the same time without using a tube or mask.

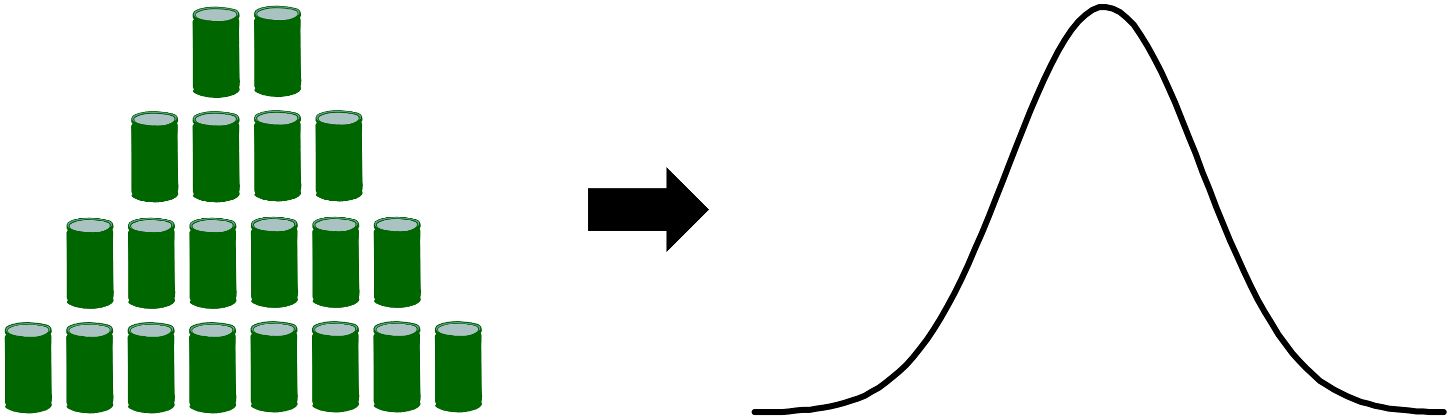
- Probability Distributions

- If something is certain to happen, then it has a probability of 1.
  - For example, it is certain that the sun will rise tomorrow.



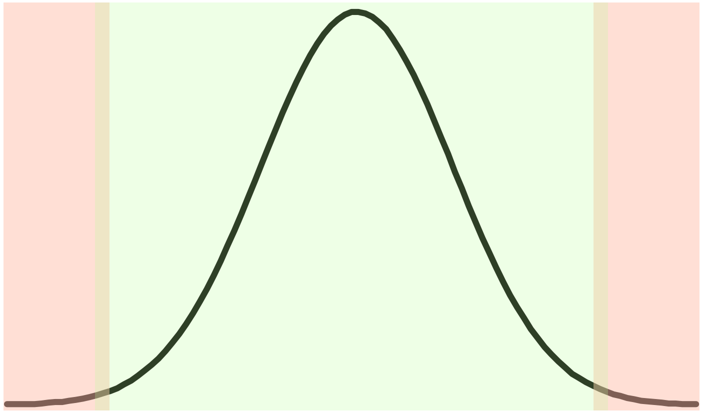
- Probability Distributions

- You might be certain if you examine the whole **population**.
- But often times, you only have **samples** to work with.
- To draw conclusions from sample data, you should compare values obtained from the sample with the theoretical values obtained from the probability distribution.



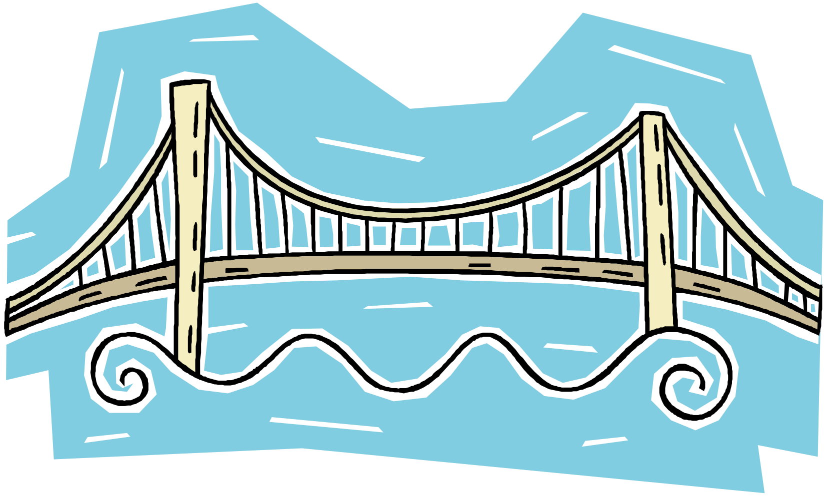
- Probability Distributions

- There will always be a **risk** of drawing false conclusions or making false predictions.
- We need to be sufficiently confident before taking any decision by setting **confidence levels**.
  - Often set at **90** percent, **95** percent or **99** percent.



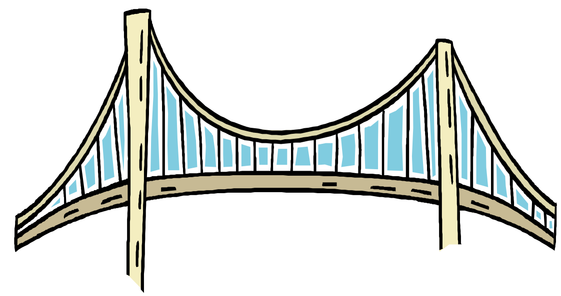
- Probability Distributions

- Many probability distribution can be defined by **factors** such as the mean and standard deviation of the data.
- Each probability distribution has a **formula**.



- Probability Distributions

- There are different shapes, models and classifications of probability distributions.



- **They are often classified into two categories:**
  - Discrete.
  - Continuous.

Binomial	Uniform
Hypergeometric	Exponential
Poisson	Normal
Binary	Student's t

- Probability Distributions

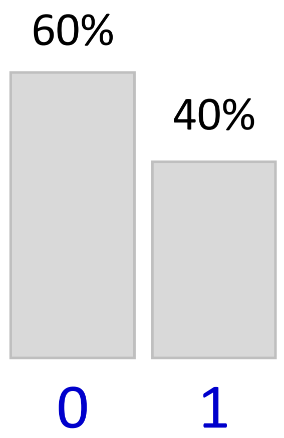
Discrete Probability Distribution:

- A **Discrete Probability Distribution** relates to discrete data.
- It is often used to model uncertain events where the possible values for the variable are either attribute or countable.
- The two common discrete probability distributions are **Binomial** and **Poisson** distributions.

- Probability Distributions

Binary Distribution:

- A discrete probability distribution that takes only two possible values.
- There is a probability that one value will occur and the other value will occur the rest of the time.
- Many real-life events can only have two possible outcomes:
  - A product can either pass or fail in an inspection test.
  - A student can either pass or fail in an exam.
  - A tossed coin can either have a head or a tail.
- Also referred to as a **Bernoulli distribution**.



- Probability Distributions

Binomial Distribution:

- A discrete probability distribution that is used for data which can only take one of two values, i.e.:
  - Pass or fail.
  - Yes or no.
  - Good or defective.
- It allows to compute the **probability of the <u>number of successes</u>** for a given **<u>number of trials</u>**.
  - Each is either a success or a failure, given the probability of success on each trial.
- **Success** could mean anything you want to consider as a positive or negative outcome.

Defective
Pass
Pass
Pass
Fail
Fail
Pass
Pass
Pass
Fail

- Probability Distributions

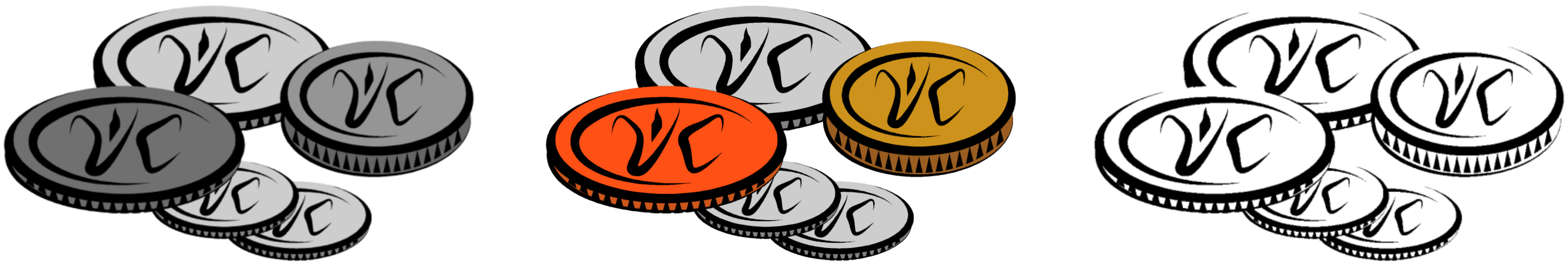
Binomial Distribution:

- Assume that you are tossing a coin 10 times.
- You will get a number of heads between 0 and 10.
- You may then carry out another 10 trials, in which you will also have a number of heads between 0 and 10.
- By doing this many times, you will have a data set which has the shape of the **binomial distribution**.
- Getting a head would be a **success** (or a hit).
- The number of tosses would be the **trials**.
- The probability of success is 50 percent.

- Probability Distributions

Binomial Distribution:

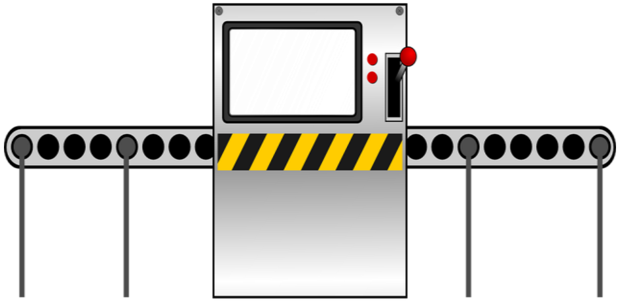
- The binomial test requires that each trial is **independent** from any other trial.
- In other words, the probability of the second trial is not affected by the first trial.



- Probability Distributions

Binomial Distribution:

- **This test has a wide range of applications, such as:**
  - Taking 10 samples from a large batch which is 3 percent defective (as past history shows).
  - Asking customers if they will shop again in the next 12 months.
  - Counting the number of individuals who own more than one car.
  - Counting the number of correct answers in a multi-choice exam.



- Probability Distributions

Binomial Distribution:

- **The binomial distribution is appropriate when the following conditions apply:**
  - There are only two possible outcomes to each trial (success and failure).
  - The number of trials is fixed.
  - The probability of success is identical for all trials.
  - The trials are independent (i.e. carrying out one trial has no effect on any other trials).

- Probability Distributions

Binomial Distribution:

- The probability of ‘r’ successes P(r) is given by the **binomial formula**:

$$P(r) = n!/(r!(n-r))! * p^r(1-p)^{n-r}$$

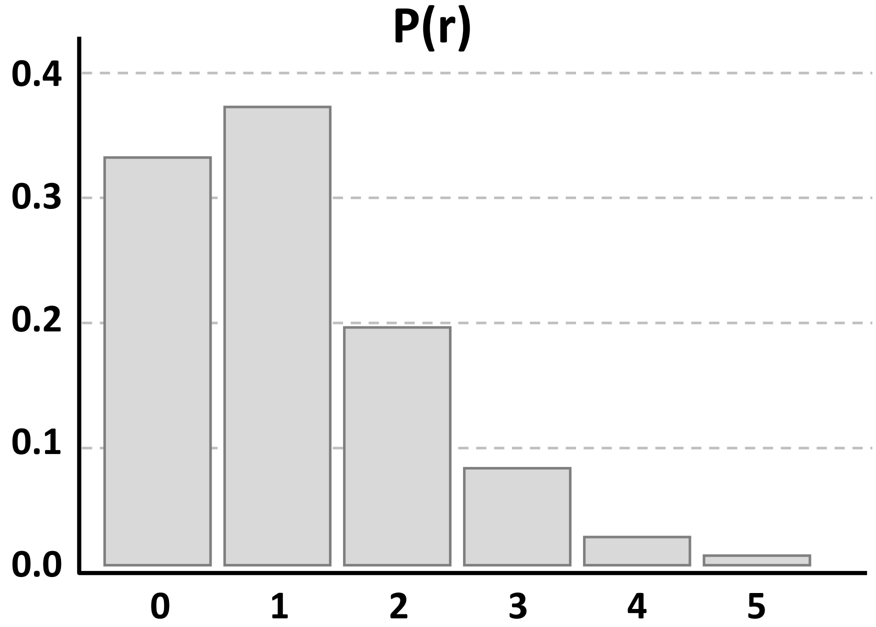
p: probability of success
n: number of independent trials
r: number of successes in the n trials

The binomial distribution is fully defined if we know both ‘n’ & ‘p’

- Probability Distributions

Binomial Distribution:

- The data can be plotted on a graph.



The exact shape of a particular distribution depends on the values of ‘n’ and ‘p’

- Probability Distributions

Binomial Distribution – Example:

- In a sample of **20** drawn from a batch which is **5%** defective, what is the probability of getting exactly **3** defective items?

$$P(r) = n!/(r!(n-r))! * p^r(1-p)^{n-r}$$

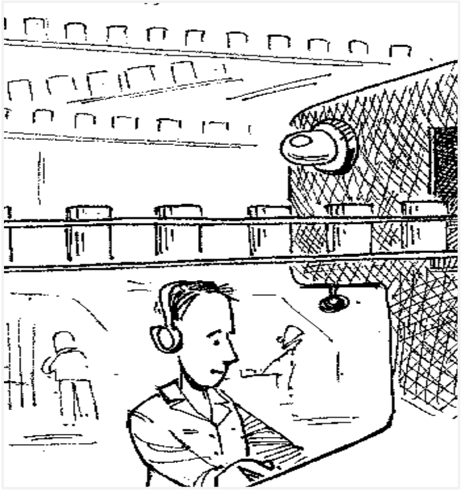
$$P(r) = 20!/(3!(20 - 3)!) * 0.05^3(1 - 0.05)^{20-3}$$

$$P(r) = 1{,}140 * 0.05^3 * (0.95)^{17} = \mathbf{0.060}$$

- Probability Distributions

Binomial Distribution – Example:

- When sampling, we commonly want to accept a batch if there are (say) **1** or less defective in the sample, and reject it if there is **2** or more.
- In order to determine the probability of acceptance, the individual probabilities for **0** and **1** defectives are summed:
  - P (1 or less) = P(0) + P(1)
    = 0.358 + 0.377 = 0.735

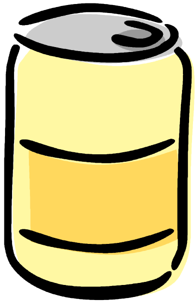


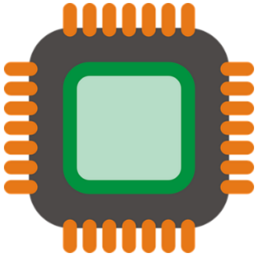
- So the probability of rejection is:
  - 1 – 0.735 = 0.265

- Probability Distributions

Binomial Distribution – Other Examples:

- What is the probability of obtaining exactly **2** heads in the **5** tosses?
- What is the probability that in a random sample of **10** cans there are exactly **3** defective units, knowing that on average there is a **5%** defective product?
- What is the probability that a random sample of **4** units will have exactly **1** unit is defective, knowing that that process will produce **2%** defective units on average?





- Probability Distributions

Binomial Distribution:

- In **Excel**, you may calculate the binomial probabilities using the **BINOM.DIST** function. Simply write:

=BINOM.DIST(number of successes, number of trials, probability of success, FALSE)

- Probability Distributions

Binomial Distribution:

- Suppose that we want to know the chance of getting exactly **4** heads out of **10** tosses.
- Instead of using the binomial formula, we might skip straight to the Excel formula.
- In Excel, we simply write:

=BINOM.DIST(4,10,0.5,FALSE)

- The result value will be **0.205078125**.
- This means that there is a **20.5%** chance that **10** coin tosses will produce exactly **4** heads.

- Probability Distributions

Hypergeometric Distribution:

- A very similar to the binomial distribution.
- The only difference is that a hypergeometric distribution **does not use replacement** between trials.
- It is often used for samples from relatively small populations.

- Probability Distributions

Hypergeometric Distribution:

- Assume that there are 5 gold coins and 25 silver coins in a box.
- You may close your eyes and draw 2 coins without replacement.
- By doing this many times, you will have a data set which has the shape of the hypergeometric distribution.
- You can then answer questions such as:
  - What is the probability that you will draw one gold coin?

Note that the probability of success on each trial is not the same as the size of the remaining population will change as you remove the coins.

- Probability Distributions

Poisson Distribution:

- It is not always appropriate to classify the outcome of a test simply as pass or fail.
- Sometimes, we have to count the number of defects where there may be several defects in a single item.
- The **Poisson Distribution** is a discrete probability distribution that specifies the probability of a certain number of occurrences over a specified interval.
  - Such as time or any other type of measurements.

Defects
14
8
11
2
13
17
11
9
12

- Probability Distributions

Poisson Distribution:

- With the Poisson distribution, you may examine a unit of product, or collect data over a specified interval of time.
- You will have a number of successes (zero or more).
- You may then repeat the exercise, in which you will also have a number of successes.
- By doing this many times, you will have a data set which has the shape of the Poisson distribution.

- Probability Distributions

Poisson Distribution:

- This test has a wide range of applications, such as:
  - Counting the number of defects found in a single product.
  - Counting the number of accidents per year in a factory.
  - Counting the number of failures per month for a specific equipment.
  - Counting the number of incoming calls per day to an emergency call center.
  - Counting the number of customers who will walk into a store during the holidays.

- Probability Distributions

Poisson Distribution:

- **The Poisson distribution is appropriate when the following conditions apply:**
  - Occurrences take place within a defined interval or area of opportunity (per sample, per unit, per hour, etc.).
  - Occurrences take place at a constant rate, where the rate is proportional to the length or size of the interval.
  - The likelihood of occurrences is not affected by which part of the interval is selected (e.g. the time of day).
  - Occurrences take place randomly, singly and independently of each other.

- Probability Distributions

Poisson Distribution:

- The probability of ‘r’ occurrences is given by the **Poisson formula**:

$$P(r) = \lambda^{r} e^{-\lambda} / r!$$

λ: average or expected number of occurrences in a specific interval
r: number of occurrences

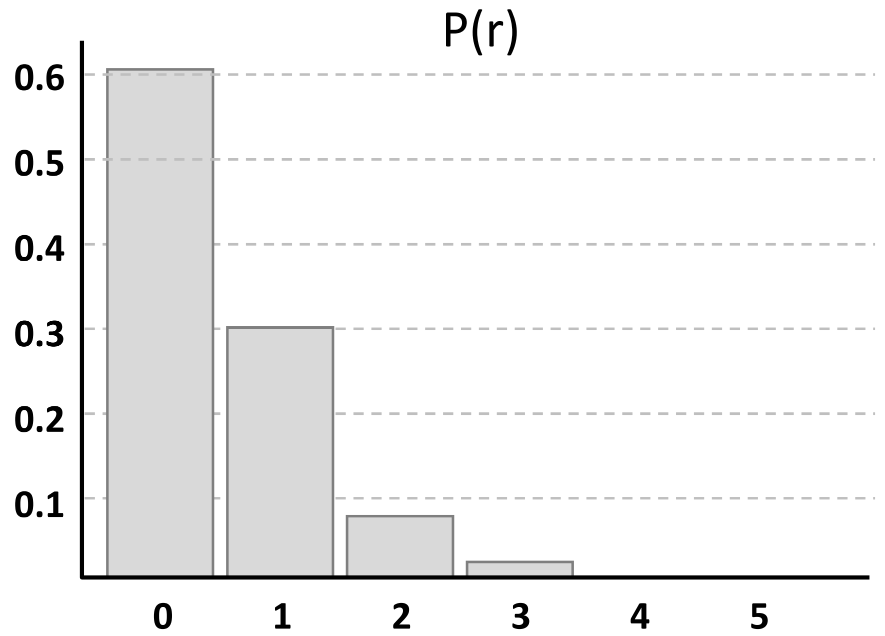
The Poisson distribution is fully defined if we know the value of λ

- Probability Distributions

Poisson Distribution:

- The data can be plotted on a graph.

r	P(r)
0	0.607
1	0.303
2	0.076
3	0.013
4	0.002
5	<0.001
6	0



The exact shape of a particular distribution depends solely on the value of λ

- Probability Distributions

Poisson Distribution – Example:

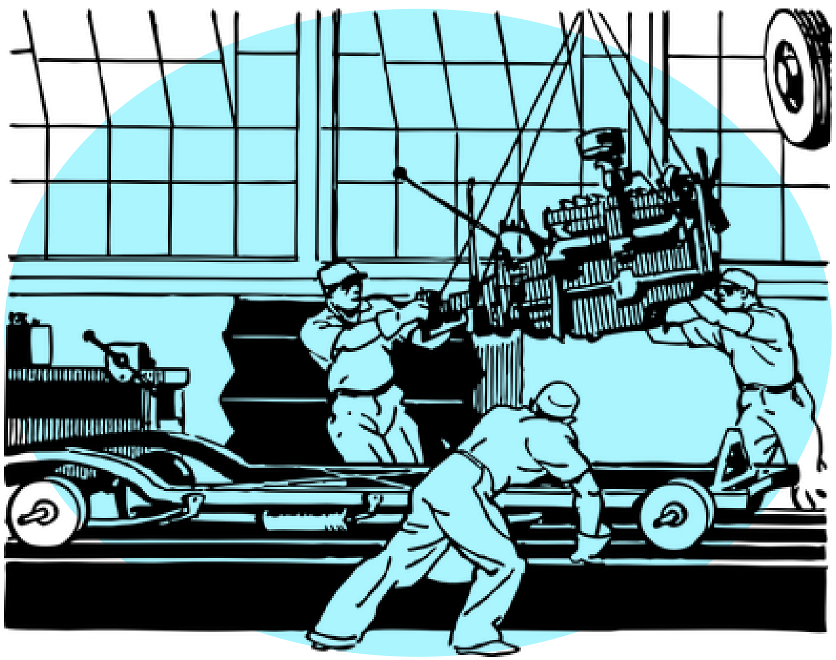
- What is the probability of at least **1** accident taking place in a given week?
- In other words, what is the probability of **1** or more accidents taking place?
- We can add the probabilities above **0** or use the complement method.
- P(**1** or more accidents) = **1** – P (**0** accidents)
- = 1 – 0.607 = **0.393**

r	P(r)
0	0.607
1	0.303
2	0.076
3	0.013
4	0.002
5	<0.001
6	0

- Probability Distributions

Poisson Distribution – Other Example:

- What is the probability of assembling **1** part with less than **3** defects, knowing that a study has determined that on average there are **3** defects per assembly.



- Probability Distributions

Poisson Distribution:

- In **Excel**, you may calculate the Poisson probabilities using the **POISSON.DIST** function. Simply write:

=POISSON.DIST(number of successes, expected mean, FALSE)

- Probability Distributions

Continuous Distributions:

- Relates to continuous data.
- Can take any value and can be measured with any degree of accuracy.
- The commonest and the most useful continuous distribution is the **normal distribution**.
- There are other continuous probability distributions that are used to model non-normal data.

Uniform

Exponential

Normal

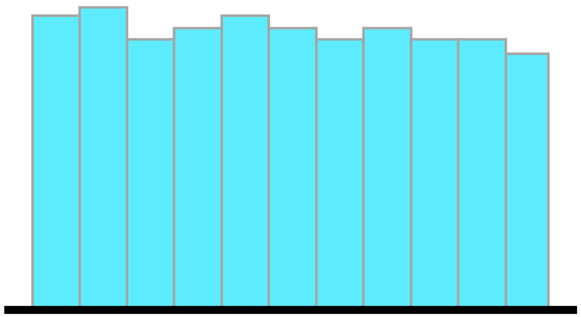
Student's t

Chi-square

- Probability Distributions

Uniform Distribution:

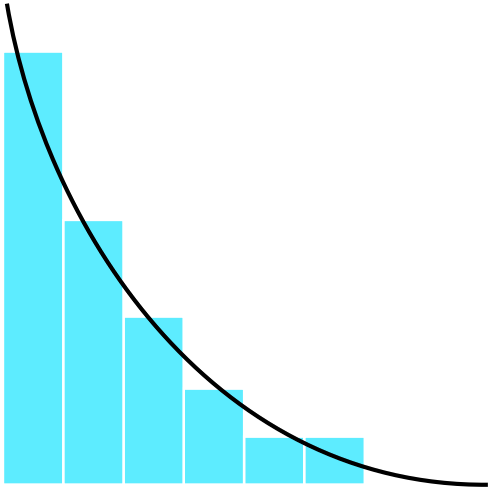
- All the events have the exact same probability of happening anywhere within a fixed interval.
- When displayed as a graph, each bar has approximately the same height.
- Often described as the **rectangle distribution**.
- Does not occur often in nature.
- However, it is important as a reference distribution.



- Probability Distributions

Exponential Distribution:

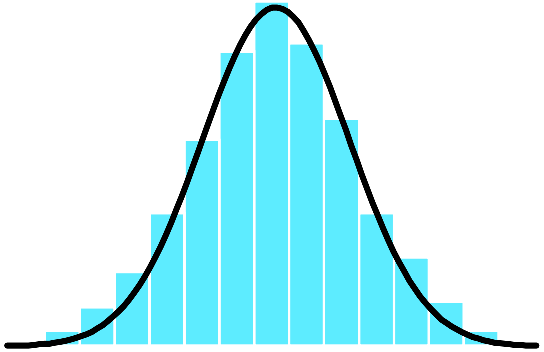
- Often used in quality control and reliability analysis.
- Mainly concerned with the amount of time until some specific event occurs.
  - Calculating the probability that a computer part will last less than six year.
- Its shape is highly skewed to the right.
  - There is a much greater area below the mean than above it.



- Probability Distributions

Normal Distribution:

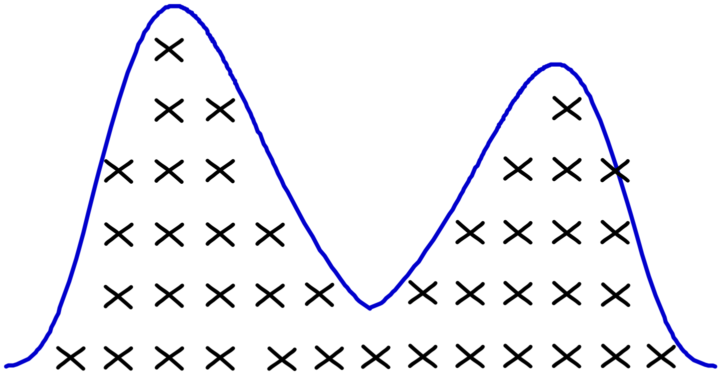
- A symmetrical probability distribution.
- Most results are located in the middle and few are spread on both sides.
- Has the shape of a bell.
- Can entirely be described by its **mean** and **standard deviation**.
- Normality is an important assumption when conducting statistical analysis so that they can be applied in the right manner.



- Probability Distributions

Bimodal Distribution:

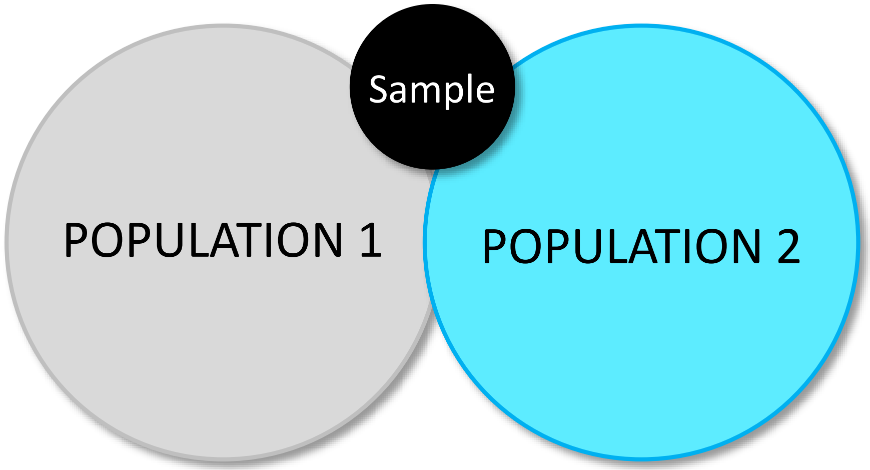
- It has two modes (or peaks).
- Two values occur more frequently than the other values.
- It can be seen in traffic analysis where traffic peaks during the morning rush hour and then again in the evening rush hour.



- Probability Distributions

Bimodal Distribution:

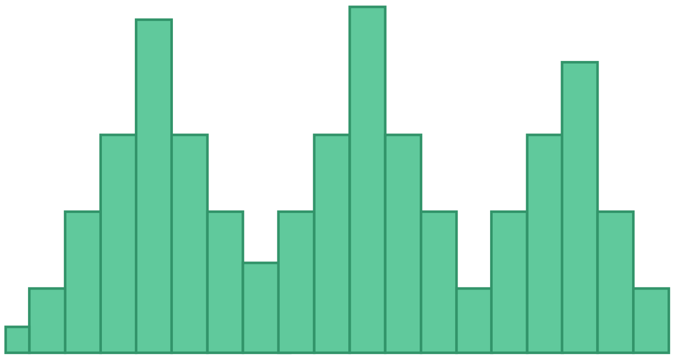
- It may also result if the observations are taken from two different populations.
  - This can be seen when taking samples from two different shifts or receiving materials from two different suppliers.
  - There is actually one mode for the two data sets.



- Probability Distributions

Bimodal Distribution:

- The bimodal is considered a **Multimodal Distribution** as it has more than one peak.
- This may indicate that your sample has several patterns of response, or has been taken from more than one population.
- Multimodal distributions can be seen in daily water distribution as water demand peaks during different periods of the day.



- Probability Distributions in Minitab

- There are different shapes, models and classifications of probability distributions.
- It is always a good practice to know the distribution of your data before proceeding with your analysis.
- Once you find the appropriate model, you can then perform your statistical analysis in the right manner.
- **Minitab** can be used to find the appropriate probability distribution of your data.

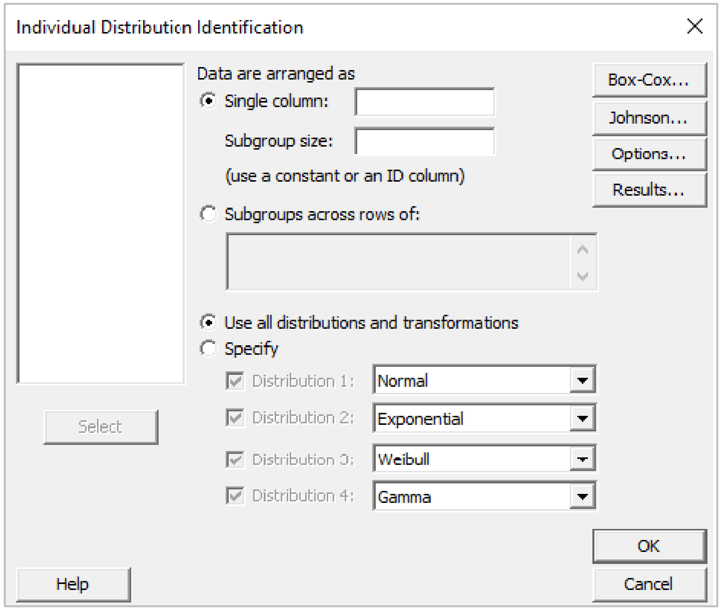
Probability Distributions

- Probability Distributions in Minitab

- You may use the **Individual Distribution Identification** in confirm that a particular distribution best fits your current data.
- It allows to easily compare how well your data fit various different distributions.
- Let's look at an example where a hospital is seeking to detect the **presence of high glucose levels** in patients at admission.

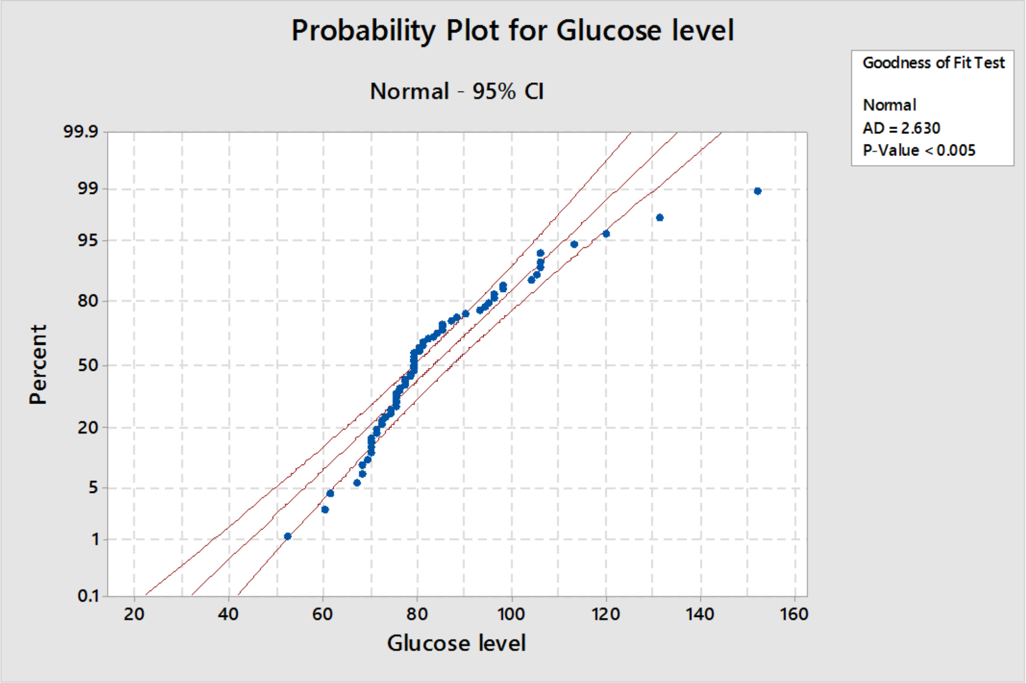
- Probability Distributions in Minitab

- Remember to copy the data from the Excel worksheet and paste it into the Minitab worksheet.
- To find out the probability distribution that best fit the data:
  - Select **Stat > Quality Tools > Individual Distribution Identification**.
  - Specify the column of data to analyze.
  - Specify the distribution to check the data against.
  - Then click OK.



- Probability Distributions in Minitab

- The resulted graph if only the normal distribution has been selected:

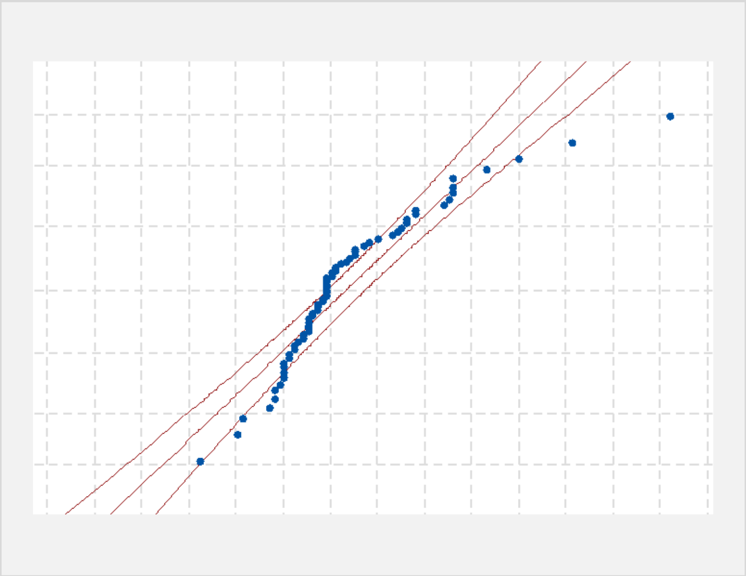


Minitab can also create a series of graphs that provide most of the same information along with probability plots

- Probability Distributions in Minitab

- A given distribution is a good fit if the data points approximately follow a straight line and the p-value is greater than **0.05**.

In our case, the data does not appear to follow a straight line



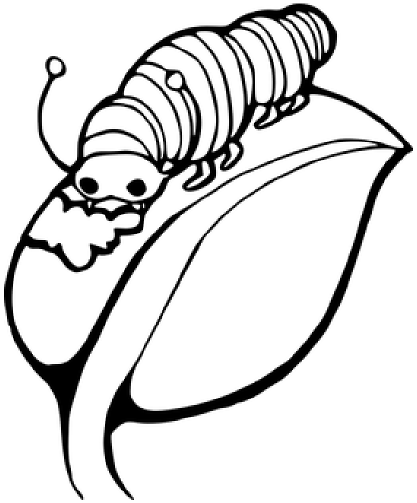
A good place to start is to look at the highest p-values in the session window

Goodness of Fit Test

Distribution	AD	P	LRT P
Normal	2.630	<0.005	
Box-Cox Transformation	0.862	0.025	
Lognormal	1.495	<0.005	
3-Parameter Lognormal	1.006	*	0.048
Exponential	21.000	<0.003	
2-Parameter Exponential	9.300	<0.010	0.000
Weibull	3.831	<0.010	
3-Parameter Weibull	1.889	<0.005	0.000
Smallest Extreme Value	5.993	<0.010	
Largest Extreme Value	0.884	0.022	
Gamma	1.814	<0.005	
3-Parameter Gamma	1.975	*	1.000
Logistic	1.712	<0.005	
Loglogistic	1.073	<0.005	
3-Parameter Loglogistic	0.653	*	0.048
Johnson Transformation	0.246	0.749	

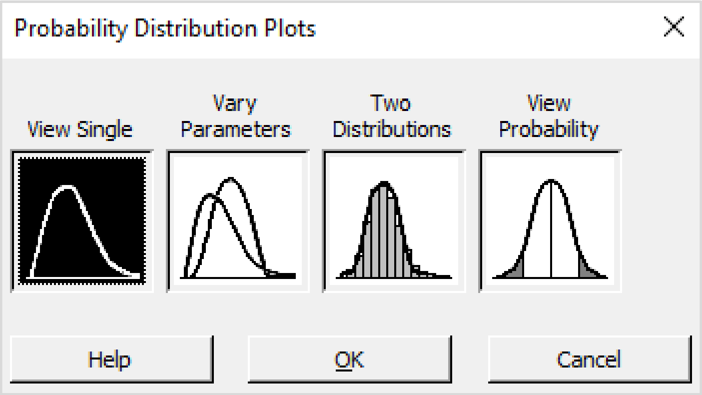
- Probability Distributions in Minitab

- You may transform your non-normal data using the **Box-Cox** or **Johnson** transformation methods so that it follows a normal distribution.
- You can then use the transformed data with any analysis that assumes the data follow a normal distribution.



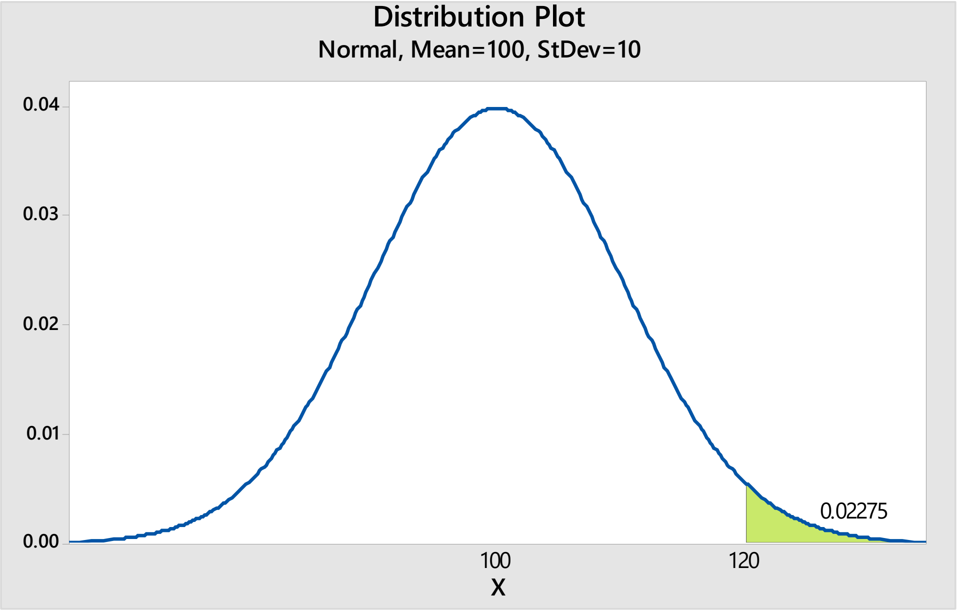
- Probability Distributions in Minitab

- You can also use the **Probability Distribution Plots** to clearly communicate probability distribution information in a way that can be easily understood by non-experts.
- Select **Graph > Probability Distribution Plot**, and then choose one of the following options:
  - **View Single** to display a single probability distribution plot.
  - **Vary Parameters** to see how changing parameters will affect the distribution.
  - **Two Distributions** to compare the shape of curves based on different parameters.
  - **View Probability** to see where target values fall in a distribution.



- Probability Distributions in Minitab

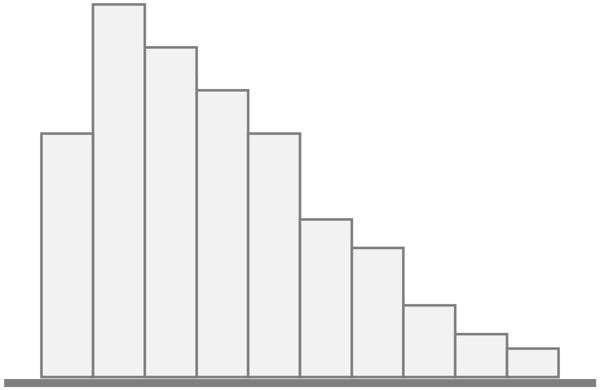
- Here is an example of a process with a mean of **100**, a standard deviation of **10** and an upper specification limit of **120**.
- The following screenshot shows the shaded area under the curve that is above the upper specification limit:



- Probability Distributions

Further Information:

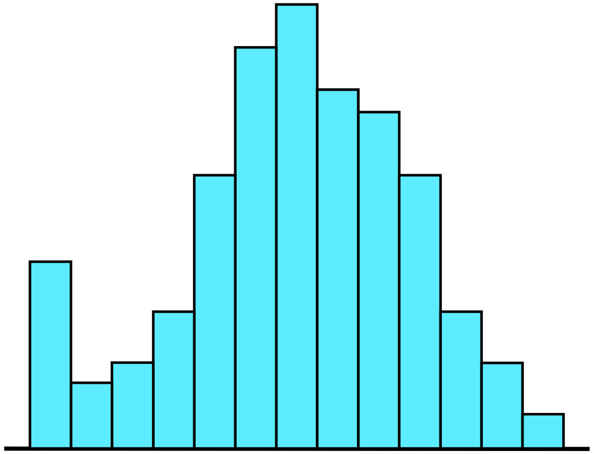
- A **truncated distribution** is a probability distribution that has a single tail.
- Can be truncated on the right or left.
- Occurs when there is no ability to know about or record data below a threshold or outside a certain range.



- Probability Distributions

Further Information:

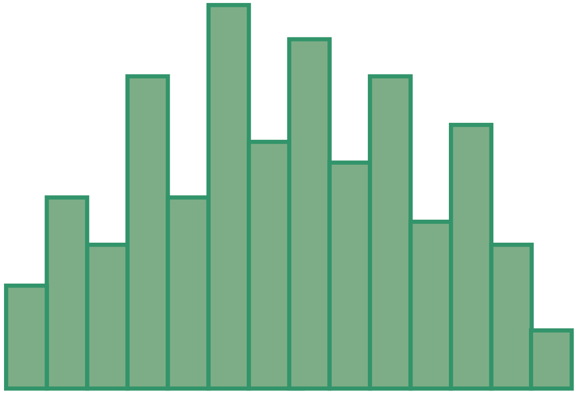
- An **edge peak distribution** is where there is an out of place peak at the edge of the distribution.
- This usually means that the data has been collected or plotted incorrectly, unless you know for sure your data set has an expected set of outliers.



- Probability Distributions

Further Information:

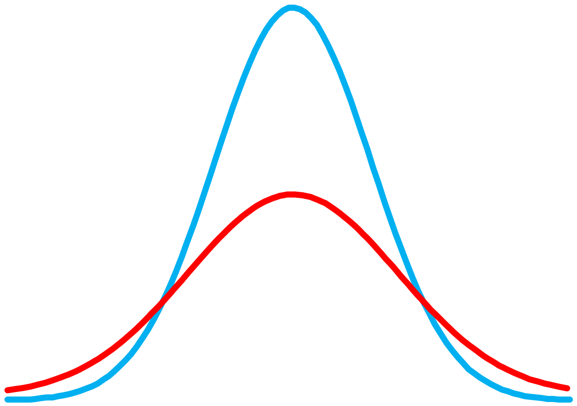
- A **comb distribution** is so-called because the distribution looks like a comb, with alternating high and low peaks.
- A comb shape can be caused by rounding off.
- Or if more than one person is recording the data or more than one instrument is used.



- Probability Distributions

Further Information:

- **Student's T-Distribution:**
  - A bell shaped probability distribution that is symmetrically shaped.
  - Generally much flatter and wider than the normal distribution (has more probability in the tails).
  - There is a t-distribution for each sample size of **n > 1**.
  - Appropriate when constructing confidence intervals based on small samples (n<30) from populations with unknown variances.



- Probability Distributions

Further Information:

❑ Student's T-Distribution:

- Defined by a single parameter, the **degrees of freedom** (df).
- The degrees of freedom refers to the number of independent observations in a set of data.

df	0.1	0.05	0.02	0.01	...
1	6.314	12.706	31.821	63.657	...
15	1.753	2.131	2.602	2.947	...
30	1.697	2.042	2.457	2.750	...
∞	1.645	1.960	2.327	2.576	...

Critical values of the t-distribution

- It is equal to the number of sample observations minus one.
  - The distribution of the t-statistic from samples of size **8** would be described by a t-distribution having **7** degrees of freedom.

- Probability Distributions

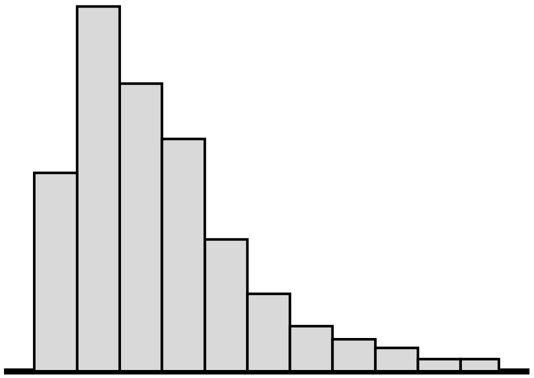
Further Information:

- **F-Distribution:**
  - A continuous probability distribution that will help in the testing of whether two normal samples have the same variance.
  - Used to compute probability values in the analysis of variance (ANOVA) and the validity of a multiple regression equation.
  - An example of a positively skewed distribution is household incomes.

- Probability Distributions

Further Information:

- **F-Distribution:**
  - **Has two important properties:**
    - It’s defined only for positive values.
    - It’s not symmetrical about its mean. instead, it’s positively skewed.
  - The peak of the distribution happens just to the right of zero.
  - The higher the **F-value** after that point, the lower the curve.



- Probability Distributions

Further Information:

- **F-Distribution:**
  - Like the student's t-distribution, the probability is determined by the **degrees of freedom**.
  - Unlike the Student's t-distribution, the F-distribution is characterized by two different types of degrees of freedom:
    - Numerator (**dfn**).
    - Denominator (**dfd**).
  - The shape depends on the values of the numerator and denominator degrees of freedom.