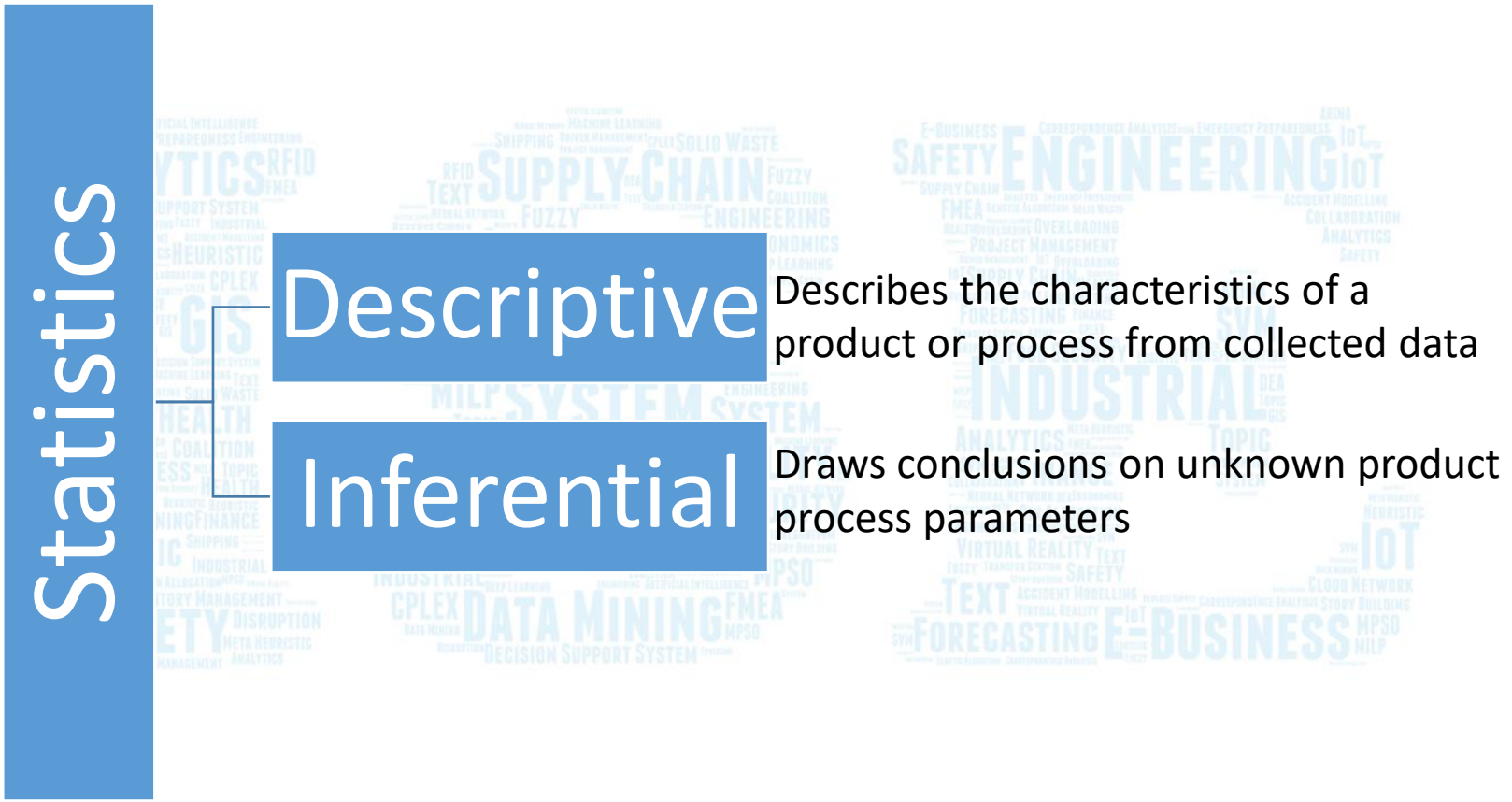
Probability Distributions and Statistical Concepts

Prof. Sayak Roychowdhury

IIT Kharagpur
Department of Industrial & Systems Engineering

Descriptive and Inferential Statistics



Describes the characteristics of a product or process from collected data

Draws conclusions on unknown product process parameters

Probability Distribution

- **Probability Distribution:** A mathematical model that relates the value of the **random variable** with the probability of occurrence of that value in the population.
- **Continuous Distributions:** When the random variable can be expressed on a continuous scale. E.g. $0 \leq x \leq 1$, cable diameter, length/width of a rectangle
- **Discrete Distributions:** When the random variable can take only certain values such as integers. e.g. # of nonconforming parts, # of defects on a circuit board, # of customers served within certain time limit

Probability Distribution

Discrete	Continuous
Probability Mass Function (pmf) $f_X(x) = \Pr[x = X]$	Probability Density Function (pdf) $f(x)$: $\Pr(a \le X \le b) = \int_a^b f(x)dx$
Cumulative Mass Function $\Pr[x \le X] = \sum_{x=0}^{x=X} f(x)$	Cumulative Density Function $\Pr[x \le X] = \mathrm{F(X)} = \int_{-\infty}^{X} f(t)dt$
e.g. Binomial, Poisson, Hypergeometric, Multinomial, Bernoulli, Negative Binomial	e.g. Normal, Exponential, Weibull, Beta, Gamma, Cauchy, Uniform

Probability Distribution: Application in QE

- Basis of ANOVA, t-test , Regression (Normal Distribution)
- Application in Sampling (Hypergeometric, Binomial, Poisson, Normal)
- Reliability Engineering (Weibull, Exponential, Normal)
- Application in control chart (Normal Distribution for Xbar, R chart, Binomial for p-chart, Poisson for u-chart)

Probability Distribution

- **Expected Value:** Long run average of the random variable for the experiment it represents. (Wiki)
  - Discrete case: $E(X) = \sum_{i=1}^{\infty} x_i * P(X = x_i)$
  - Continuous case: $E(X) = \int_{-\infty}^{+\infty} x * p(x)dx$
- **Variance:** The expected value of the squared deviation from the mean $\mu = E(X)$
  - $Var(X) = E(X - \mu)^2 = E\big(X - E(X)\big)^2 = E\big(X^2\big) - \big(E(X)\big)^2$
  - Discrete case: $\mathrm{Var}(X) = \sum_{i=1}^{\infty} (x_i - \mu)^2 * P(X = x_i)$
  - Continuous case: $\mathrm{Var}(X) = \int_{-\infty}^{+\infty} (x - \mu)^2 * p(x)dx$
- Ex. Probability of getting 1 bad apple is 0.8, 2 bad apples is 0.18 and 3 bad apples is 0.02 from a basket. What is the expected number of bad apples? What is the variance?

Basics

- $P(A) = \frac{N_A}{N}$
- $0 \leq P(A) \leq 1$
- $P(A \cup B) = P(A) + P(B) - P(A \cap B)$
- $P(A \cap B) = 0$ (when A and B are mutually exclusive)
- $P(A \cap B) = P(A|B) * P(B)$
- $P(A \cap B) = P(A) * P(B)$ when A and B are independent
- $P(A|B) = P(A)$, $P(B|A) = P(B)$ when A and B are independent

Population and Sample

- A **population** is the set of all items that possess a certain characteristic of interest.
- Example: Set of all cans of brand A of soup produced in a particular month (where average weight of the cans is the quantity of interest)
- A **sample** is a subset of population
- E.g. Average weight of all 50000 can produced in a particular month.
- A **parameter** is a characteristic of a population.
- A **statistic** is a characteristic of a sample used to make inferences on population parameters.

Hypergeometric

- Discrete probability distribution
- $P(d) = \frac{\binom{D}{d}\binom{N-D}{n-d}}{\binom{N}{n}}$
- $\mu = \frac{nD}{N} \quad \sigma = \sqrt{\frac{\frac{nD}{N}\left(1-\frac{D}{N}\right)(N-n)}{N-1}}$
- Sampling without replacement
- EXCEL =HYPGEOM.DIST(# nc in sample, sample size, #of nc in pop, pop size, TRUE=cumulative / False=prob. mass)
- A random sample of 4 insurance claims is selected from a lot of 12 that has 3 nonconforming claims. What is the probability that the sample will have 1 nc claim? Less than 3 nc claims?

Binomial

- Discrete probability distribution
- $P(d) = \frac{n!}{d!(n-d)!} p^d (1-p)^{n-d}$
- p = proportion, n = number in sample
- $\mu = np,\ \sigma = \sqrt{np\,(1-p)}$
- Sampling with replacement
- EXCEL: =BINOM.DIST(d, n, p, TRUE=cumulative / False=prob. Mass)
- A steady stream of income tax returns has 0.03 non-conforming. What is the probability of obtaining 2 nc units from a sample of 20?

Poisson

- Discrete probability distribution
- $P(x) = \frac{e^{-\lambda}\lambda^{x}}{x!}$
- x = count
- λ = average count, average number of events of a given classification occurring in a sample
- $\mu = \lambda \quad \sigma^2 = \lambda$
- EXCEL =POISSON.DIST(x , λ , TRUE=cumulative / False=prob. Mass)
- Average number of nc units is 1.6, what is the probability that a sample will contain 2 or fewer nc units?

Poisson

Tables of the Poisson Cumulative Distribution

The table below gives the probability of that a Poisson random variable X with mean = λ is less than or equal to x. That is, the table gives

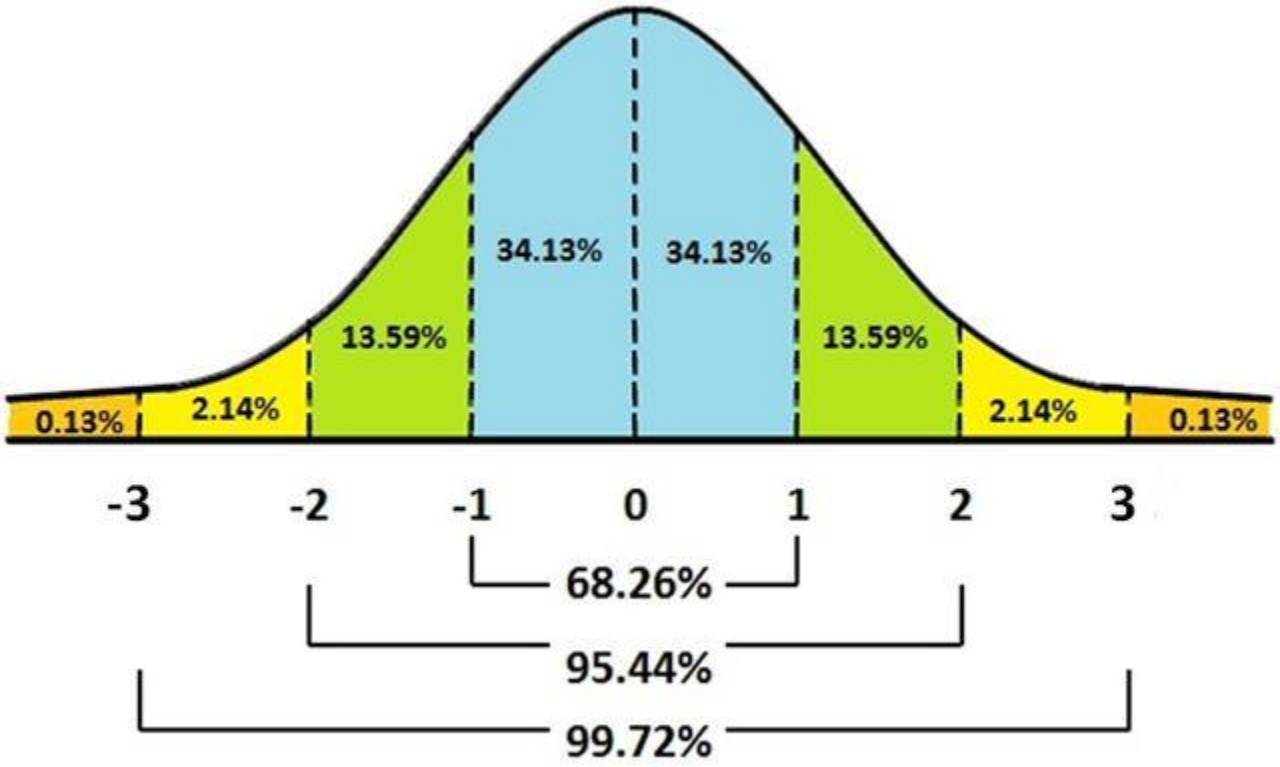
$$P(X \le x) = \sum_{r=0}^{x} \lambda^r \frac{e^{-\lambda}}{r!}$$

λ=		0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1.0	1.2	1.4	1.6	1.8
x=	0	0.9048	0.8187	0.7408	0.6703	0.6065	0.5488	0.4966	0.4493	0.4066	0.3679	0.3012	0.2466	0.2019	0.1653
	1	0.9953	0.9825	0.9631	0.9384	0.9098	0.8781	0.8442	0.8088	0.7725	0.7358	0.6626	0.5918	0.5249	0.4628
	2	0.9998	0.9989	0.9964	0.9921	0.9856	0.9769	0.9659	0.9526	0.9371	0.9197	0.8795	0.8335	0.7834	0.7306
	3	1.0000	0.9999	0.9997	0.9992	0.9982	0.9966	0.9942	0.9909	0.9865	0.9810	0.9662	0.9463	0.9212	0.8913
	4	1.0000	1.0000	1.0000	0.9999	0.9998	0.9996	0.9992	0.9986	0.9977	0.9963	0.9923	0.9857	0.9763	0.9636
	5	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.9999	0.9998	0.9997	0.9994	0.9985	0.9968	0.9940	0.9896
	6	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.9999	0.9997	0.9994	0.9987	0.9974
	7	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.9999	0.9997	0.9994
	8	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.9999
	9	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
λ=		2.0	2.2	2.4	2.6	2.8	3.0	3.2	3.4	3.6	3.8	4.0	4.5	5.0	5.5
x=	0	0.1353	0.1108	0.0907	0.0743	0.0608	0.0498	0.0408	0.0334	0.0273	0.0224	0.0183	0.0111	0.0067	0.0041
	1	0.4060	0.3546	0.3084	0.2674	0.2311	0.1991	0.1712	0.1468	0.1257	0.1074	0.0916	0.0611	0.0404	0.0266
	2	0.6767	0.6227	0.5697	0.5184	0.4695	0.4232	0.3799	0.3397	0.3027	0.2689	0.2381	0.1736	0.1247	0.0884
	3	0.8571	0.8194	0.7787	0.7360	0.6919	0.6472	0.6025	0.5584	0.5152	0.4735	0.4335	0.3423	0.2650	0.2017
	4	0.9473	0.9275	0.9041	0.8774	0.8477	0.8153	0.7806	0.7442	0.7064	0.6678	0.6288	0.5321	0.4405	0.3575
	5	0.9834	0.9751	0.9643	0.9510	0.9349	0.9161	0.8946	0.8705	0.8441	0.8156	0.7851	0.7029	0.6160	0.5289
	6	0.9955	0.9925	0.9884	0.9828	0.9756	0.9665	0.9554	0.9421	0.9267	0.9091	0.8893	0.8311	0.7622	0.6860
	7	0.9989	0.9980	0.9967	0.9947	0.9919	0.9881	0.9832	0.9769	0.9692	0.9599	0.9489	0.9134	0.8666	0.8095
	8	0.9998	0.9995	0.9991	0.9985	0.9976	0.9962	0.9943	0.9917	0.9883	0.9840	0.9786	0.9597	0.9319	0.8944
	9	1.0000	0.9999	0.9998	0.9996	0.9993	0.9989	0.9982	0.9973	0.9960	0.9942	0.9919	0.9829	0.9682	0.9462
	10	1.0000	1.0000	1.0000	0.9999	0.9998	0.9997	0.9995	0.9992	0.9987	0.9981	0.9972	0.9933	0.9863	0.9747
	11	1.0000	1.0000	1.0000	1.0000	1.0000	0.9999	0.9999	0.9998	0.9996	0.9994	0.9991	0.9976	0.9945	0.9890
	12	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.9999	0.9999	0.9998	0.9997	0.9992	0.9980	0.9955
	13	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.9999	0.9997	0.9993	0.9983
	14	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.9999	0.9998	0.9994
	15	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.9999	0.9998
	16	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.9999
	17	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000

Normal Distribution

- Continuous probability distribution
- $f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$
- EXCEL: =NORM.DIST(x, μ, σ, TRUE=cumulative / False=prob. Mass)
- Operating life of a mixer has a mean of 2200h, and standard dev. Of 120h. What is the probability that a single electric mixer will fail to operate at 1900h or less?

Standard Normal Curve



Interrelationship

- Hypergeometric can be approximated by
  - Binomial when $\frac{n}{N} \leq 0.1$
  - Poisson when $\frac{n}{N} \leq 0.1$, $p_0 \leq 0.10$ and $np_0 \leq 5$
  - Normal when $\frac{n}{N} \leq 0.1$
- Binomial can be approximated by
  - Poisson when $p \leq 0.10$ and $np \leq 5$
  - Normal when $p \sim 0.5$ and $n \geq 10$

Exponential Distribution

- Exponential Distribution: (continuous)
  - Prob. Density Function: $f(x,\lambda) = \begin{cases} \lambda e^{-\lambda x} & x > 0 \\ 0 & x \leq 0 \end{cases}$
  - Cumulative Density: $F(x,\lambda) = \begin{cases} 1 - e^{-\lambda x} & x > 0 \\ 0 & x \leq 0 \end{cases}$
  - Mean = $1/\lambda$, Variance = $\frac{1}{\lambda^2}$
  - In reliability, λ = failure rate, $1/\lambda$ = mean time to failure
  - Reliability at time t with mean life θ : $R_t = e^{-\frac{t}{\theta}}$
- EXCEL: =EXPON.DIST(x, λ, TRUE=cumulative / False=prob. Mass)

Insight on Exponential Distribution

- It is a probability distribution, that represents elapsed time between 2 events in a Poisson Point Process, i.e. where events occur continuously and randomly at a constant average rate.
- This distribution is *memoryless,* i.e. the distribution of waiting time does not depend on the time elapsed already.

  e.g. The time a store keeper must wait before arrival of a customer

Weibull Distribution

- Weibull Distribution: (continuous)
  - Prob. Density Function: $f(t,\theta,\beta) = \begin{cases} \left(\frac{\beta}{\theta}\right)\left(\frac{t}{\theta}\right)^{\beta-1} e^{-\left(\frac{t}{\theta}\right)^{\beta}} & t \geq 0 \\ 0 & t < 0 \end{cases}$
  - Cumulative Density: $F(t,\theta,\beta) = \begin{cases} 1 - e^{-\left(\frac{t}{\theta}\right)^{\beta}} & t \geq 0 \\ 0 & t < 0 \end{cases}$
  - In reliability θ = mean life
  - Reliability at time t with mean life θ, shape factor β : $R_t = e^{-\left(\frac{t}{\theta}\right)^{\beta}}$
  - Shape of cdf changes with β : $\beta = 1 \rightarrow$ Exponential; $\beta = 3.4 \rightarrow Normal$
- EXCEL: =WEIBULL.DIST(x, β, θ, TRUE=cumulative / False=prob. Mass)

Insight on Weibull Distribution

- A shape parameter of $\beta < 1$, indicates that the failure rate decreases over time. It represents the idea of infant mortality, or defective parts failing in the beginning of use and weeded out.
- A shape parameter of $\beta = 1$, reduces Weibull Distribution to Exponential Distribution. It suggests constant failure rate, which means random external events are causing the mortality.
- A shape parameter of $\beta > 1$, indicates that the failure rate increases over time. It represents the idea of aging, or parts that are more likely to fail with the increase of time.

Sampling Distribution

- An **estimator**, or **statistic** (which is a characteristic of a sample), is used to make inferences as to the corresponding **parameter**.
- For example, an estimator of sample mean is used to draw conclusions on the population mean. Similarly, a sample variance is an estimator of the population variance.
- Studying the behavior of these estimators through repeated sampling allows us to draw conclusions about the corresponding parameters.
- The **behavior of an estimator** in repeated sampling is known as the **sampling distribution** of the estimator, which is expressed as the probability distribution of the statistic.

Central Limit Theorem (CLT)

- **Definition:** If $x_1, \ldots, x_n$ are independent random variables with mean μ_i and variance σ_i^2, and if $y = x_1 + \cdots + x_n$, then the distribution of $\frac{y - \sum_{i=1}^{n} \mu_i}{\sqrt{\sum_{i=1}^{n} \sigma_i^2}}$ approaches the $N(0,1)$ distribution as n approaches infinity.

  (*Montgomery D.C., Introduction to Statistical Quality Control)*

- It implies that the sum of n independently distributed random variables is approximately normal, regardless of the distribution of the individual variables.

- If x_i are **independent and identically distributed** (IID), and distribution of each x_i does not depart radically from normal distribution, then CLT works quite well for $n \geq 3 \; or \; 4$. (common in SQC problems)

Central Limit Theorem

- Suppose that we have a population with mean μ and standard deviation σ. If random samples of size n are selected from this population, the following holds if the sample size is large:
- 1. The sampling distribution of the sample mean will be approximately normal.
- 2. The mean of the sampling distribution of the sample mean $\mu_{\bar{X}}$ will be equal to the population mean μ.
- 3. The standard deviation of the sample mean is given by$\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}}$, known as the standard error.

Important Sampling Distributions Derived from Normal Distribution

1. χ^2 distribution: If $x_1, .. x_n$ are standard normally and independently distributed then $y = x_1^2 + x_2^2 \ldots + x_n^2$ follow chi-squared distribution with n degrees of freedom.

2. t-distribution: If x is standard normal variable and y is chi-squared random variable with k degrees of freedom, and if x and y are independent then the random variable $t = \frac{x}{\sqrt{\frac{y}{k}}}$ is distributed as t with k degrees of freedom.

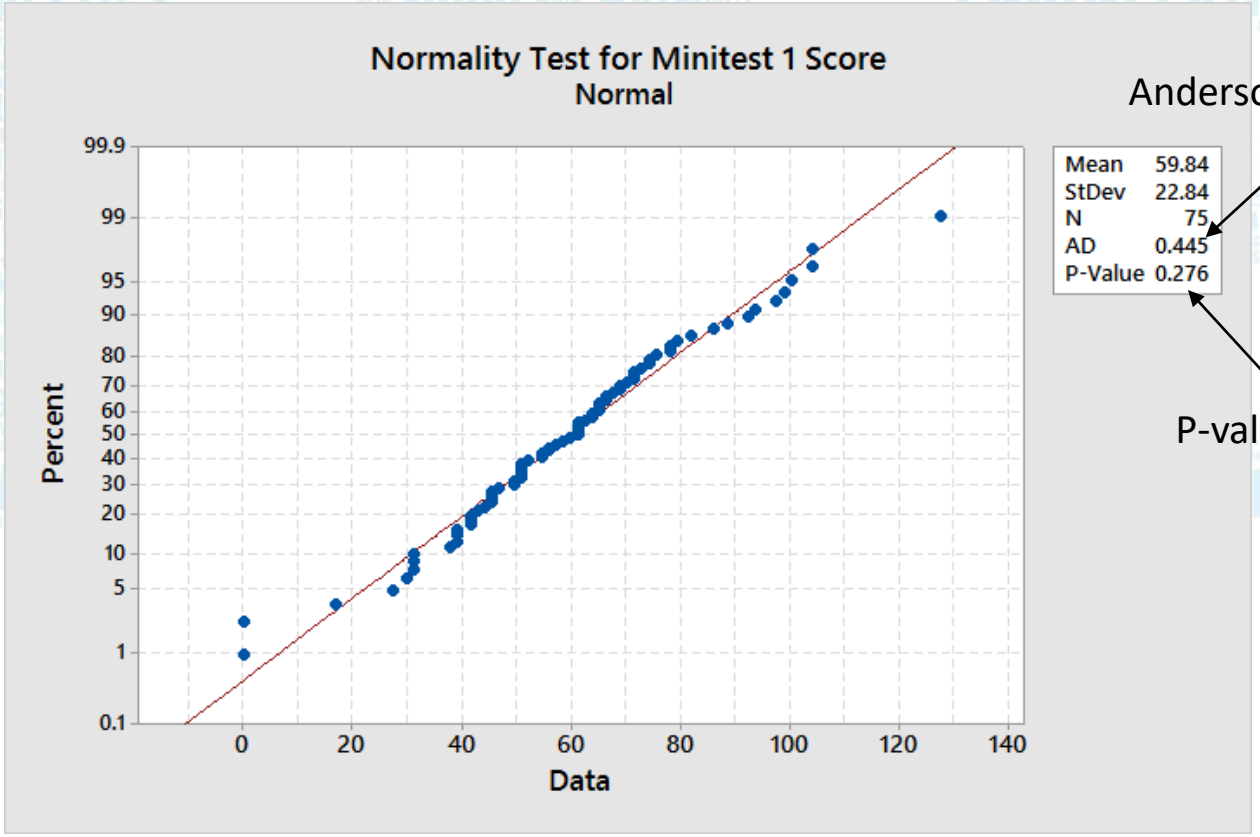
3. If w and y are two independent random chi-sq distributed variables with u and v degrees of freedom, then the ratio $F = \frac{\frac{w}{u}}{\frac{y}{v}}$ follows F distribution with (u, v) degrees of freedom

Tests of Normality

- Normal Probability Plot of Residuals
- Histogram
- Boxplot
- Skewness and Kurtosis
- Chi-sq. Test

Normal Probability Plot of Residuals

Minitab > Stat >Basic Statistics > Normality Tests>Select data > Ok



Anderson-Darling Test

P-value > 0.05, so?

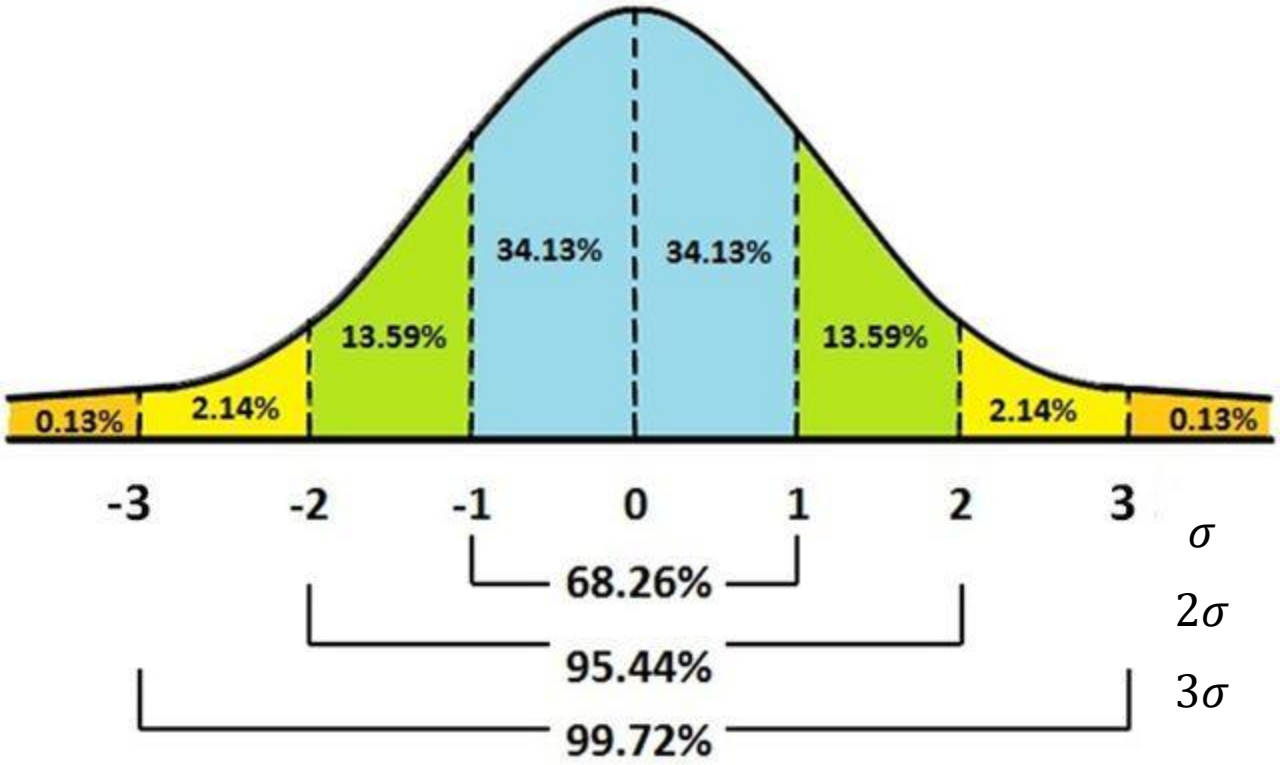
Normal Probability Plot of Residuals

- Order the data
- Rank the data i
- Calculate plotting position $PP = 100 * \frac{i-0.5}{n}$
- Plot the points on a Normal Probability Plot paper (Horizontal axis- PP, Vertical Axis- Data, or reverse) ***OR***
- Take $z_i = \frac{i-0.5}{n}$, make a column of $\Phi^{-1}(z_i)$. Plot data

(X_i) corresponding to rank i on x-axis, $\Phi^{-1}(z_i)$ on y-axis

- Fit a best fit line by observation
- Judgement on how close the points are to the straight line.

Normal Curve



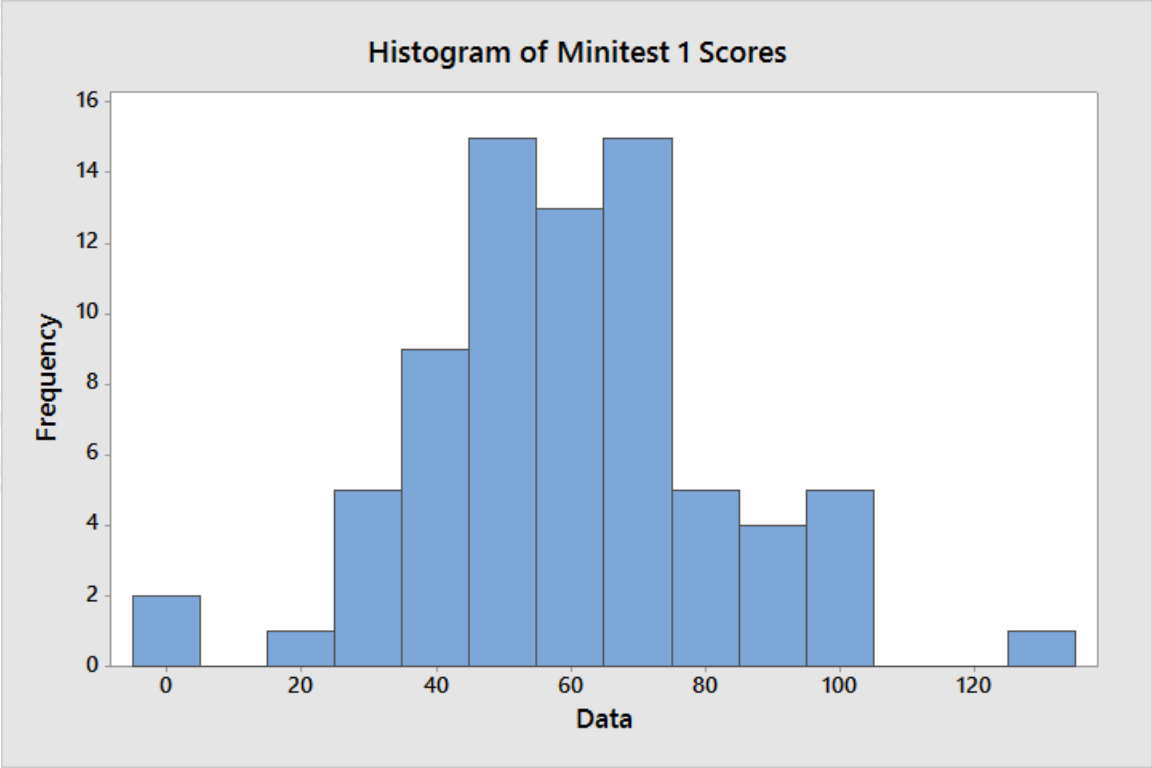
http://whatilearned.wikia.com/wiki/File:Normal_curve_probability.jpg

Histograms

- Tally grouped or ungrouped data
- Determine range $R = X_h - X_l$
- Determine cell (or bin) interval i (applying Sturges' rule is optional)
- Determine cell midpoints ($MP_l = X_l + \frac{i}{2}$)
- Determine cell boundaries (extra decimal place)
- Post cell/bins and the frequencies
- Plot (X-axis: midpoints, Y-axis: frequencies/Relative frequencies)

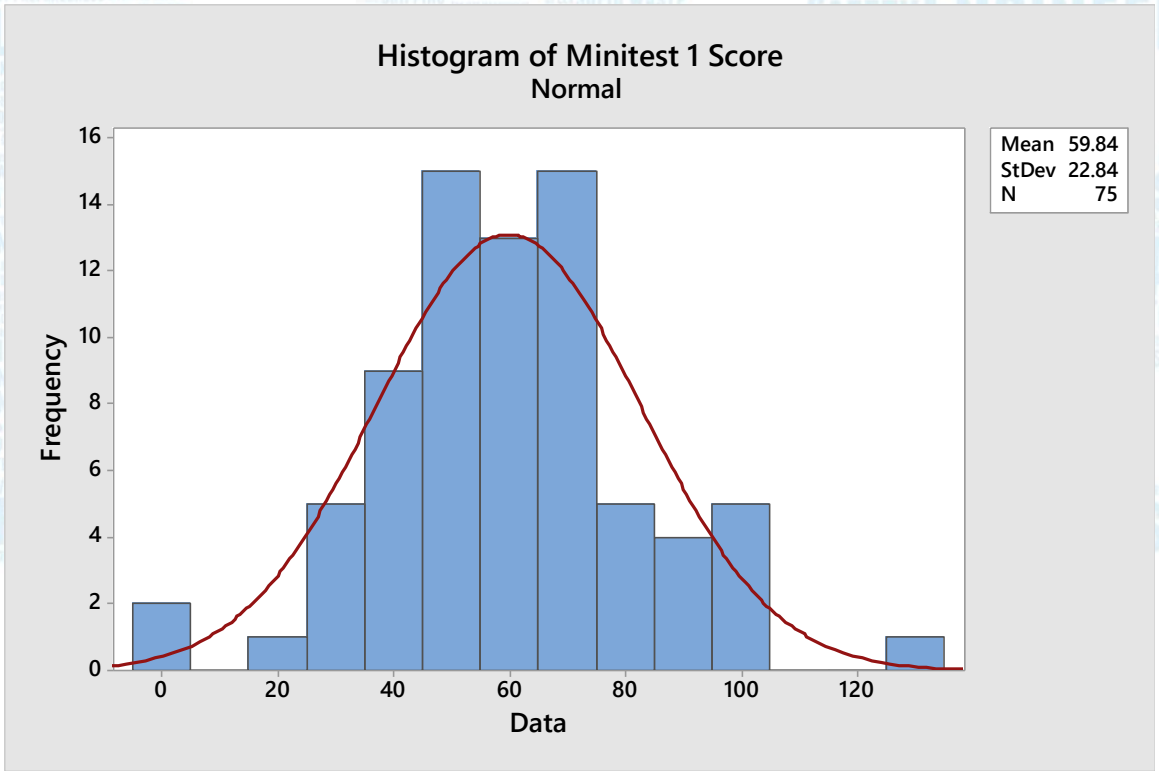
Histograms

Minitab > Graph > Histogram >Simple > Select Variable > OK



Histograms

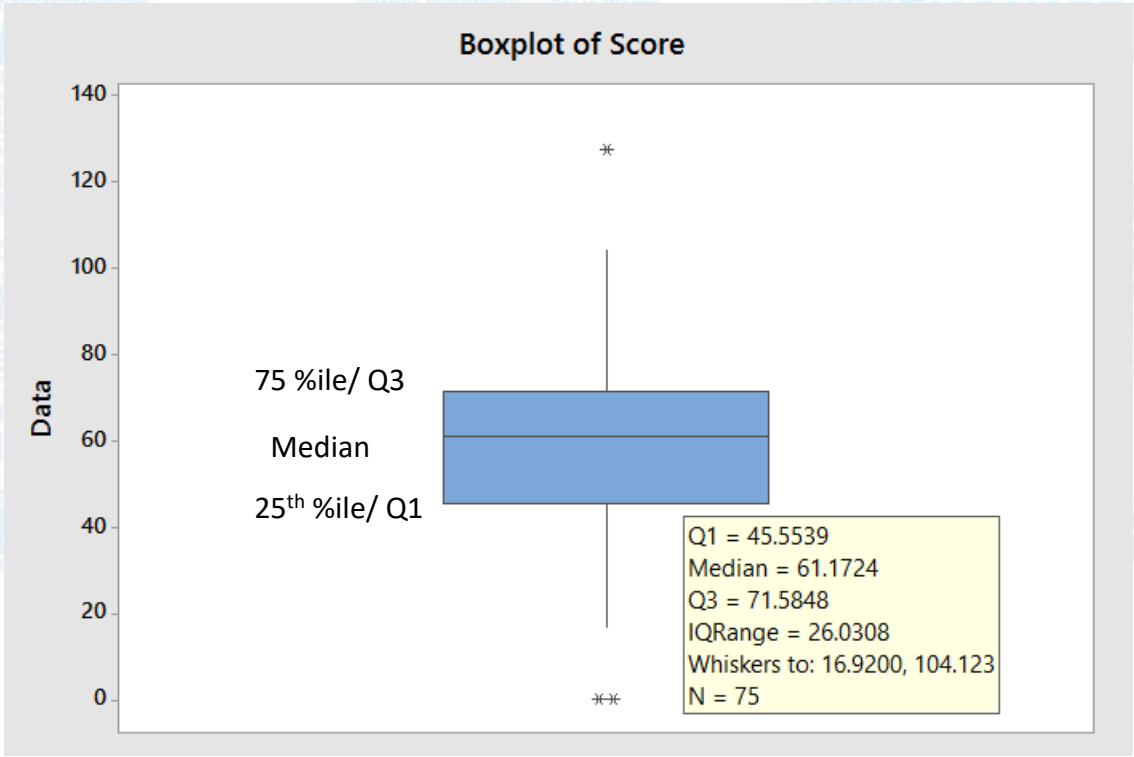
Minitab > Graph > Histogram >With Fit > Select Variable > OK



If the histogram closely follows normal curve,
then data is more likely to come from a normal distribution

Boxplot

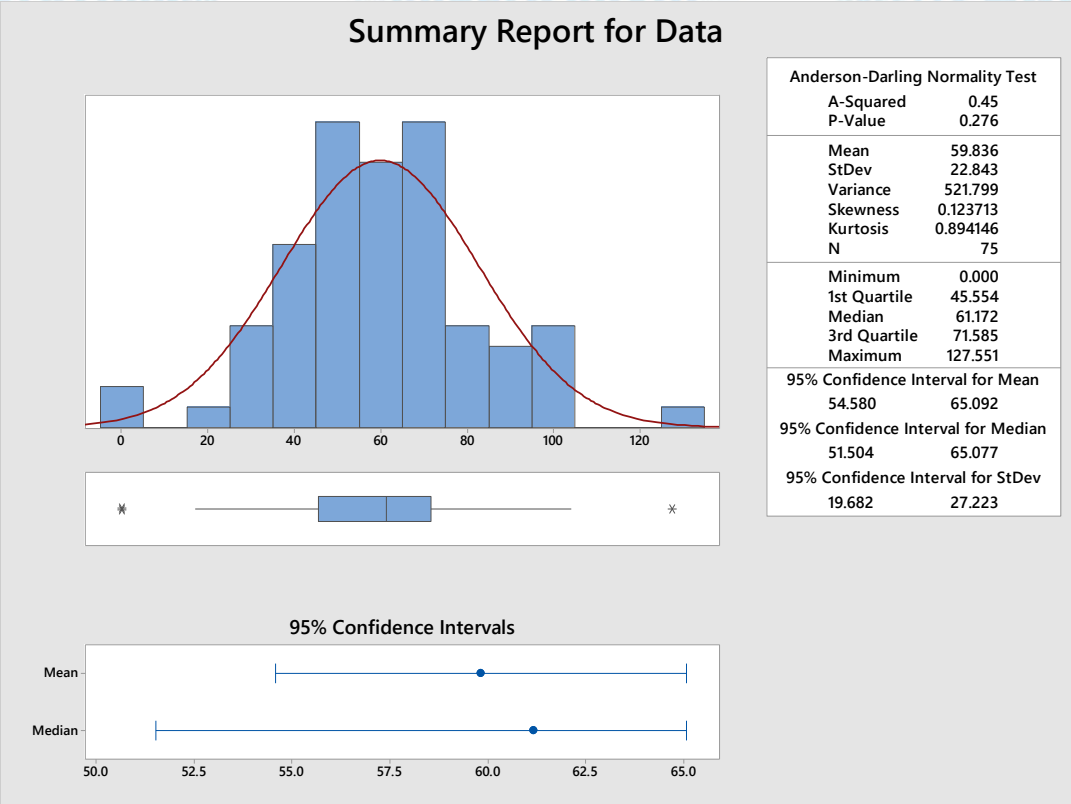
Minitab > Graph > Box plot > Simple > Select Variable



If the boxplot is symmetrical about median, then data is more likely to be normal

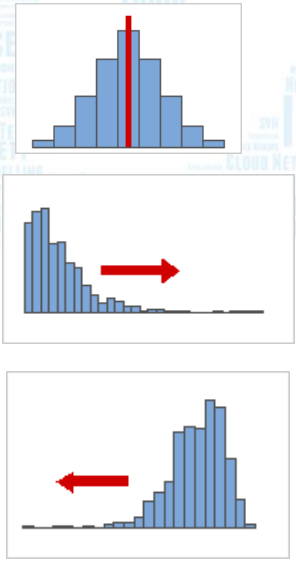
Graphical Summary

Minitab > Stat > Basic Statistics > Graphical Summary



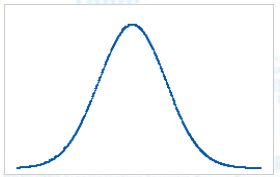
Skewness & Kurtosis: Skewness

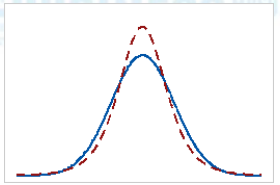
- Skewness: Determines the lack of symmetry of the data.
  - $a_3 = \frac{\frac{\sum_{i=1}^{n}(X_i - \bar{X})^3}{n}}{s^3}$ (for $X_1, \ldots, Xn$ data points, mean $\bar{X}$ and standard deviation s)
  - $a_3 = 0$ the data are symmetrical
  - $a_3 > 0$ skewed to right (1 is extreme)
    - e.g. Distribution of salary
  - $a_3 < 0$ skewed to left (-1 is extreme)
    - e.g. Life of light bulb

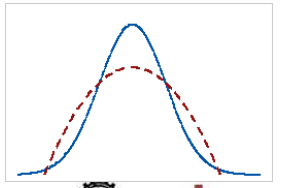


Skewness & Kurtosis: Kurtosis

- Kurtosis: Determines the peakedness of the data.
  - $a_4 = \frac{\frac{\sum_{i=1}^{n}(X_i - \bar{X})^4}{n}}{s^4}$ -3 (for $X_1, \ldots, Xn$ data points, mean $\bar{X}$ and standard deviation s)
  - $a_4 = 0$ normal distribution
  - $a_4 > 0$ more peaked (leptokurtic)
    - t-distribution
  - $a_4 < 0$ less peaked (platykurtic)
    - Beta distribution with shape parameters=2







Descriptive Statistics

Minitab > Stat > Basic Statistics > Display Descriptive Statistics >
Select Variables > Click Statistics and select Skewness and Kurtosis > Ok

Descriptive Statistics: Data

Variable	N	N*	Mean	SE Mean	StDev	Minimum	Q1	Median	Q3	Maximum	Skewness
Data	75	0	59.84	2.64	22.84	0.00	45.55	61.17	71.58	127.55	0.12

Variable	Kurtosis
Data	0.89

Hypothesis Testing

- Determine whether claims on product or process parameters are valid is the aim of Hypothesis Testing
- Hypothesis tests are based on sample data
- A standardized quantity is used as test statistic, based on point estimate.
- **Null hypothesis H_0** represents status quo or the circumstance being tested
- **Alternate hypothesis H_A** represents what we wish to prove or establish (Machine A is more accurate than machine B) or that which contradicts H_0.

Chi-Squared Test

- Goodness of fit test
- Create k bins for the data
- Calculate $\chi^2 = \sum_{i=1}^{k} (O_i - E_i)^2 / E_i$
  - O_i = Observed frequency in bin i
  - E_i = Expected frequency in bin i , can be calculated by $N * (F(UB_i) - F(LB_i))$, where F is cumulative density
- Check if $\chi^2 > \chi^2_{(1-\alpha)(k-c)}$ where k= number of non empty cells, c = number of distribution parameter + 1 (for theoretical distributions), c=1 otherwise.
- For normal dist, dof = k-3, for binomial dof = k-2
- https://www.itl.nist.gov/div898/handbook/eda/section3/eda35f.htm

Chi square Test Binomial Dist

A new casino game involves rolling 3 dice. The winnings are directly proportional to the total number of sixes rolled. Suppose a gambler plays the game 100 times, with the following observed counts:

Number of Sixes	Number of Rolls	Expected
0	48	
1	35	
2	15	
3	3	

Test whether the dice are fair.

Goodness of Fit Test for Normal Dist.

Bin	Observed Counts
(< -2.0)	6
(-2.0, -1.5)	6
(-1.5, -1.0)	18
(-1.0, -0.5)	33
(-0.5, 0.0)	38
(0.0, 0.5)	38
(0.5, 1.0)	28
(1.0, 1.5)	21
(1.5, 2.0)	9
(> 2.0)	3

Motivation of t-test: Example 1

- Single Sample (Statapault): 1 factor 1 level

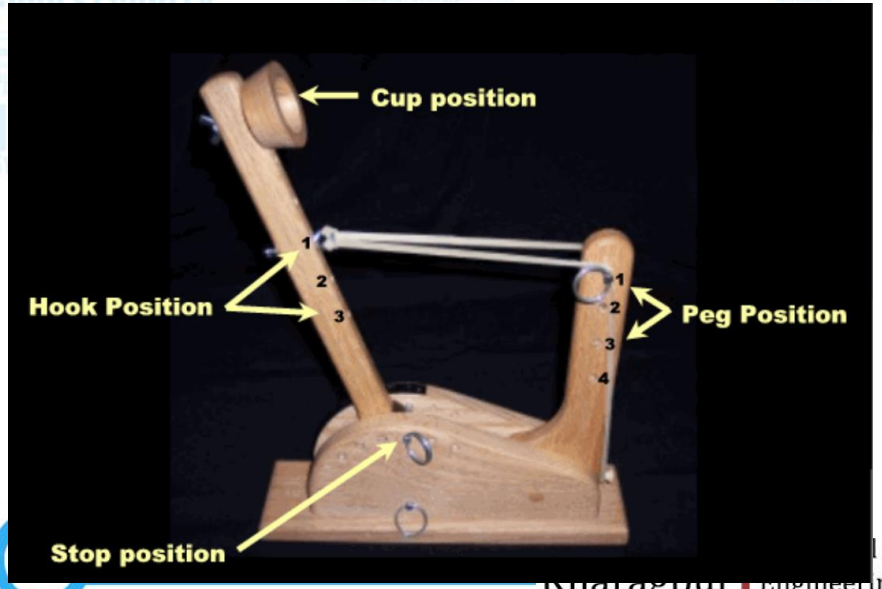
  Statapault at Hook position 1

  - Distance travelled in 5 shots (inches)

    11,13,12,10,11

  - Question: Is the mean distance travelled =12inches?

https://www.youtube.com/watch?v=eQptbZPpFI0



Motivation: Example 2

- Example 2: 1 factor 2 levels
  - Distance travelled in Hook position 1 (inches)
    11,13,12,10,11
  - Distance travelled in Hook position 2 (inches)
    17,14,13,15,15
  - Question: Does setting 2 travels more distance than setting 1?

Motivation: Example 3

- Example 3: 1 factor more than 2 levels
  - Distance travelled in hook position 1 (inches)
    11,13,12,10,11
  - Distance travelled in hook position 2 (inches)
    17,14,13,15,15
  - Distance travelled in hook position 3 (inches)
    19,17,21,23,18

  Question: Does setting significantly affect travelling distance?

Example 1

- Example 1: Single Sample (Statapault)
  - Distance travelled in 5 shots (inches)
    11,13,12,10,11
  - Question: Is the mean distance travelled =12inches?

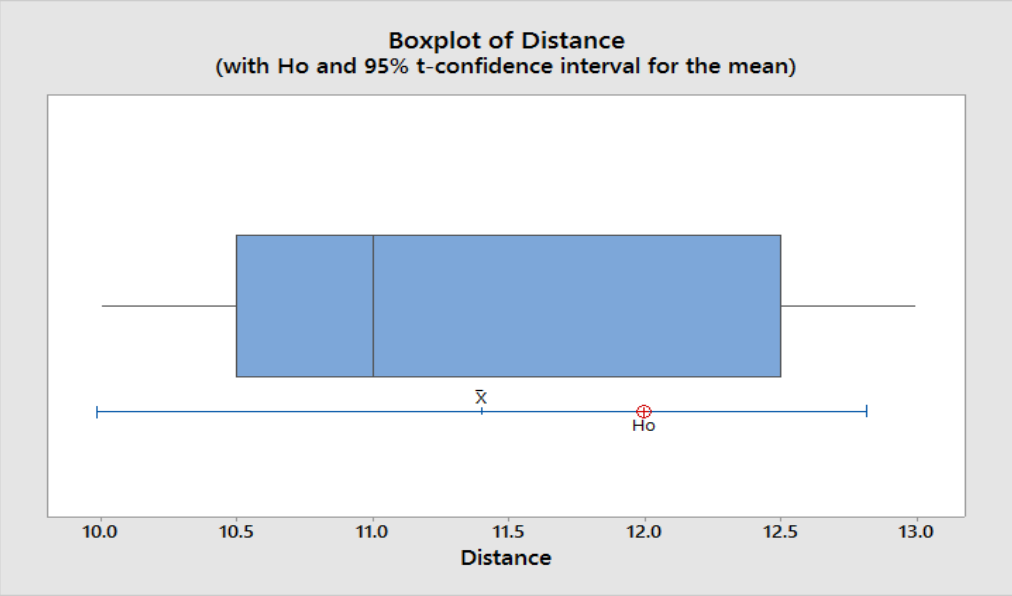
Example 1

Minitab: Stat-> Basic Statistics -> 1-sample t-test, Graphs: Select Boxplot

One-Sample T: Distance

Test of μ = 12 vs ≠ 12

Variable	N	Mean	StDev	SE Mean	95% CI	T	P
Distance	5	11.400	1.140	0.510	(9.984, 12.816)	-1.18	0.305



Example 1: Fail to reject Null hypothesis, i.e. we cannot say that population mean is different than 12. 95% Confidence interval (9.984,12.816)

Example 1: 1-sample T-test

- Null Hypothesis: $H_0: \mu = \mu_0$
- Alternate Hypothesis: $H_1: \mu \neq \mu_0$ (2 sided)

  $H_1: \mu > \mu_0$ (right tail)

  $H_1: \mu < \mu_0$ (left tail)
- Test Statistic $t_0 = \frac{\bar{y} - \mu_0}{s/\sqrt{n}}$

  where n= sample size,

  sample mean $\bar{y} = (\frac{1}{n})(\sum_{i=1}^{n} y_i)$

  sample standard deviation $s = \sqrt{\frac{\sum_{i=1}^{n} (y_i - \bar{y})^2}{n-1}}$

t is a random variable following t-distribution with ν degrees of freedom where $\nu = n - 1$

Example 1: 1-sample T-test

Procedure: 2 sided test: If $|t_0| \geq t_{\frac{\alpha}{2},n-1}$ the null hypothesis $H_0: \mu = \mu_0$ is rejected (conclusion $\mu \neq \mu_0$ *significant*)

1 sided test: For $H_1: \mu > \mu_0$, if $t_0 \geq t_{\alpha,n-1}$ then null hypothesis is rejected (conclusion $\mu > \mu_0$ *significant*)
For $H_1: \mu < \mu_0$, if $t_0 \leq -t_{\alpha,n-1}$ then null hypothesis is rejected rejected (conclusion $\mu < \mu_0$ *significant*)

Example 1: t-crit

$t_{0.025,4}$

t distribution critical values

	Upper-tail probability *p*							
df	.25	.20	.15	.10	.05	.025	.02	.01
1	1.000	1.376	1.963	3.078	6.314	12.71	15.89	31.82
2	0.816	1.061	1.386	1.886	2.920	4.303	4.849	6.965
3	0.765	0.978	1.250	1.638	2.353	3.182	3.482	4.541
4	0.741	0.941	1.190	1.533	2.132	2.776	2.999	3.747
5	0.727	0.920	1.156	1.476	2.015	2.571	2.757	3.365
6	0.718	0.906	1.134	1.440	1.943	2.447	2.612	3.143
7	0.711	0.896	1.119	1.415	1.895	2.365	2.517	2.998
8	0.706	0.889	1.108	1.397	1.860	2.306	2.449	2.896
9	0.703	0.883	1.100	1.383	1.833	2.262	2.398	2.821
10	0.700	0.879	1.093	1.372	1.812	2.228	2.359	2.764
11	0.697	0.876	1.088	1.363	1.796	2.201	2.328	2.718
12	0.695	0.873	1.083	1.356	1.782	2.179	2.303	2.681
13	0.694	0.870	1.079	1.350	1.771	2.160	2.282	2.650
14	0.692	0.868	1.076	1.345	1.761	2.145	2.264	2.624
15	0.691	0.866	1.074	1.341	1.753	2.131	2.249	2.602

Confidence Interval

- $100(1-\alpha)\%$ CI on the true mean μ is

$$\bar{y} - t_{\frac{\alpha}{2},n-1} * \frac{s}{\sqrt{n}} \leq \mu \leq \bar{y} + t_{\frac{\alpha}{2},n-1} * \frac{s}{\sqrt{n}}$$

Example 1: 1-sample Z-test (Known Variance)

- Null Hypothesis: $H_0: \mu = \mu_0$
- Alternate Hypothesis: $H_1: \mu \neq \mu_0$ (2 sided)

  $H_1: \mu > \mu_0$ (right tail)

  $H_1: \mu < \mu_0$ (left tail)
- Test Statistic $z_0 = \frac{\bar{y} - \mu_0}{\sigma/\sqrt{n}}$

  where n= sample size,

  sample mean $\bar{y} = (\frac{1}{n})(\sum_{i=1}^{n} y_i)$

Example 1: 1-sample Z-test (Known Variance)

Procedure: 2 sided test: If $|z_0| \geq z_{\frac{\alpha}{2}}$ the null hypothesis $H_0: \mu = \mu_0$ is rejected (conclusion $\mu \neq \mu_0$ *significant*) ($z_{0.025} = 1.96$),at $\alpha = 0.05$

1 sided test: For $H_1: \mu > \mu_0$, if $z_0 \geq z_\alpha$ then null hypothesis is rejected (conclusion $\mu > \mu_0$ *significant*)

For $H_1: \mu < \mu_0$, if $z_0 \leq -z_\alpha$then null hypothesis is rejected rejected (conclusion $\mu < \mu_0$ *significant*)

Confidence Interval

- $100(1-\alpha)\%$ CI on the true mean μ is

$$\bar{y} - z_{\frac{\alpha}{2}} * \frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{y} + z_{\frac{\alpha}{2}} * \frac{\sigma}{\sqrt{n}}$$

$$\bar{y} - 1.96 * \frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{y} + 1.96 * \frac{\sigma}{\sqrt{n}}$$

Example 2: 2-sample Z-test (Known Variance)

- Example 2: 2-sample
  - Distance travelled for cup-position 1 (inches)

    11,13,12,10,11
  - Distance travelled in cup-position 2 (inches)

    17,14,13,15,15
  - Question: Does setting 2 travels more distance than setting 1?

2-sample Z-test (Variance known)

- Assumptions:
  - $x_{11}, x_{12} \ldots, x_{1n_1}$ is a random sample from population 1 $\sim N(\mu_1, \sigma_1^2)$
  - $x_{21}, x_{22} \ldots, x_{2n_2}$ is a random sample from population 2 $\sim N(\mu_2, \sigma_2^2)$
  - The two populations are independent
  - Both populations are normal, if not, CLT applies

Based on the assumption we state that:

$$Z = \frac{\bar{x}_1 - \bar{x}_2 - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \sim N(0,1)$$

See **section 4.4** of Montogomery

2-sample Z-test (Variance known)

Null Hypothesis $H_0: \mu_1 - \mu_2 = \Delta_0$ (often $\Delta_0 = 0$)

Test Statistic

$$Z_0 = \frac{\bar{x}_1 - \bar{x}_2 - \Delta_0}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

Alternative Hypothesis $H_1: \mu_1 - \mu_2 \neq \Delta_0$ (2 sided)

Rejection criteria $|Z_0| > Z_{\frac{\alpha}{2}}$

P-value $P = 2[1 - \Phi(|Z_0|)]$

Confidence Interval:

$$\bar{x}_1 - \bar{x}_2 - Z_{\frac{\alpha}{2}}\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} \leq \mu_1 - \mu_2 \leq \bar{x}_1 - \bar{x}_2 + Z_{\frac{\alpha}{2}}\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

Example 2: 2-sample T-test

- Example 2: 2-sample
  - Distance travelled in setting 1 (inches)

    11,13,12,10,11
  - Distance travelled in setting 2 (inches)

    17,14,13,15,15
  - Question: Does setting 2 travels more distance than setting 1?

Example 2: 2-sample T-test

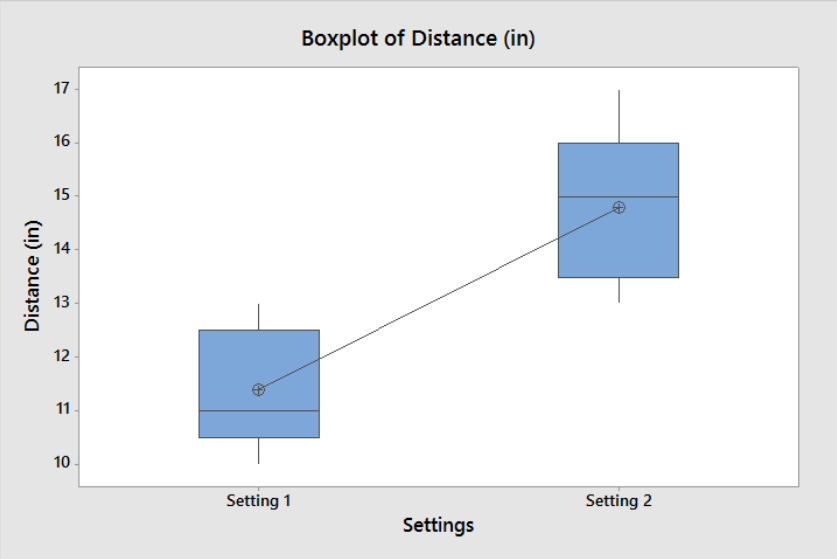
Minitab: Stat-> Basic Statistics -> 2-sample t-test,
Options: Select equal variance, difference <, Graphs: Select Boxplot

Two-Sample T-Test and CI: Distance (in), Settings

```
Two-sample T for Distance (in)

Settings    N   Mean  StDev  SE Mean
Setting 1   5  11.40   1.14     0.51
Setting 2   5  14.80   1.48     0.66


Difference = μ (Setting 1) - μ (Setting 2)
Estimate for difference:  -3.400
95% upper bound for difference:  -1.844
T-Test of difference = 0 (vs <): T-Value = -4.06  P-Value = 0.002  DF = 8
Both use Pooled StDev = 1.3229
```



Example 2: Null hypothesis Is rejected at $\alpha = 0.05$
Setting 2 provide significantly more distance than setting 1

Example 2: 2-sample T-test

- Null Hypothesis : $H_0: \mu_1 = \mu_2$ (samples came from the same distribution)
- Alternate Hypothesis: $H_1: \mu_1 \neq \mu_2$ or $\mu_1 > \mu_2$ or $\mu_1 < \mu_2$
- Test Statistic : $t_0 = \frac{\bar{y}_1 - \bar{y}_2}{s_p\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}}$, $s_p = \sqrt{\frac{(n_1-1)s_1^2+(n_2-1)s_2^2}{n_1+n_2-2}}$

s_1^2 and s_2^2 are sample variances and s_p is the pooled estimator of unknown sample variance σ .

- Number of degrees of freedom: $n_1 + n_2 - 2$
- Assumption: The two samples have *equal variance*

Example 2: 2-sample T-test

- **Procedure:** 2 sided test: If $|t_0| \geq t_{\frac{\alpha}{2},n_1+n_2-2}$ the null hypothesis $H_0: \mu_1 = \mu_2$ is rejected (conclusion $\mu_1 \neq \mu_2$ *significant*)
- **1 sided test:** For $H_1: \mu_1 > \mu_2$, if $t_0 \geq t_{\alpha,n_1+n_2-2}$ then null hypothesis is rejected (conclusion $\mu_1 > \mu_2$ *significant*)
- For $H_1: \mu_1 < \mu_2$, if $t_0 \leq -t_{\alpha,n_1+n_2-2}$ then null hypothesis is rejected rejected (conclusion $\mu_1 < \mu_2$ *significant*)

Example 2: 2-sample T-test

- When $\sigma_1^2 \neq \sigma_2^2$ (cannot assume equal variance) :

$$t_0 = \frac{\bar{y}_1 - \bar{y}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

- Degrees of freedom $\nu = \dfrac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\frac{\left(\frac{s_1^2}{n_1}\right)^2}{n_1 - 1} + \frac{\left(\frac{s_2^2}{n_2}\right)^2}{n_2 - 1}}$

Confidence Intervals

- $100(1-\alpha)\%$ CI on the difference in mean $\mu_1 - \mu_2$ is :
  - For equal variance case:

$$\bar{y}_1 - \bar{y}_2 - t_{\frac{\alpha}{2},n_1+n_2-2} s_p \sqrt{\frac{1}{n_1}+\frac{1}{n_2}} \leq \mu_1 - \mu_2 \leq$$

$$\bar{y}_1 - \bar{y}_2 + t_{\frac{\alpha}{2},n_1+n_2-2} s_p \sqrt{\frac{1}{n_1}+\frac{1}{n_2}}$$

  - For unequal variance case:

$$\bar{y}_1 - \bar{y}_2 - t_{\frac{\alpha}{2},\nu} \sqrt{\frac{s_1^2}{n_1}+\frac{s_2^2}{n_2}} \leq \mu_1 - \mu_2 \leq \bar{y}_1 - \bar{y}_2 + t_{\frac{\alpha}{2},\nu} \sqrt{\frac{s_1^2}{n_1}+\frac{s_2^2}{n_2}}$$

(ν= DOF in unequal variance case)

Errors

- Type I error: Null hypothesis is rejected when it is true.
- Type II error: Null hypothesis is *not* rejected when it is false.
  - $\alpha = P\{Type\ I\ error\} = P\{reject\ H_0 | H_0\ is\ true\}$
    - Also called **producer's risk**, probability that a good lot is rejected
  - $\beta = P\{Type\ II\ error\} = P\{fail\ to\ reject\ H_0 | H_0\ is\ false\}$
    - Also called **consumer's risk**, probability that a poor lot is accepted
  - Power= $1 - \beta$= $P\{reject\ H_0 | H_0\ is\ false\}$

P-Value

- $\alpha = 0.05$ (typically) Level of Significance of the test, probability of Type I error.
- P-value : Smallest level of significance that would lead to the rejection of H_0.
  - P-value is the probability that the test statistic would take on a value that is as extreme a value as the observed value of the statistic when H_0 is true.
  - Intuitively, smallest level of α at which the data are significant, gives an idea how significant the data are.
- At $\alpha = 0.05$, $p \leq 0.05$ shows significance (i.e. reject H_0)

Paired Comparison

Data for the Hardness Testing Experiment

Specimen	Tip 1	Tip 2
1	7	6
2	3	3
3	3	5
4	4	3
5	8	8
6	3	2
7	2	4
8	9	9
9	5	4
10	4	5

Paired Comparison

- When paired data are encountered.
- E.g. 2 machines measured tensile strengths of 8 specimens of fiber, test if the difference between measurements by 2 machines is significant
- 2 tips were used to test the hardness of 10 specimens of steel. test if the difference between measuremens by 2 tips is significant
- $y_{ij} = \mu_i + \beta_j + \epsilon_{ij} \quad i = 1,2\,;\; j = 1,2,..k$
- μ_i is the true mean response of i^{th} treatment
- β_j is the effect on response due to j^{th} specimen

Paired Comparison

- $d_j = y_{1j} - y_{2j}$ for $j = 1,2,..k$
- $\mu_d = E(d_j) = \mu_1 + \beta_j - \mu_2 - \beta_j = \mu_1 - \mu_2$
- $H_0: \mu_d = 0;\ H_a: \mu_d \neq 0$
- Test statistic $t_0 = \frac{\bar{d}}{\frac{S_d}{\sqrt{n}}}$ where $\bar{d} = (\frac{1}{n}) \sum_{j=1}^{n} d_j$
- $S_d = \sqrt{\frac{\sum_{j=1}^{n} (d_j - \bar{d})^2}{(n-1)}}$
- H_0 is rejected if $|t_0| > t_{\left(1-\frac{\alpha}{2}, n-1\right)}$
- Paired comparison design is a special case of “blocking”

Inferences about Variance

- Tests of hypotheses and confidence intervals for variances of normal distributions.
- Unlike the tests on means, the procedures for tests on variances are rather sensitive to the normality assumption
- Suppose we wish to test the hypothesis that the variance σ^2 of a normal population equals a constant σ_0^2, Stated formally,

$$H_0: \sigma^2 = \sigma_0^2$$
$$H_1: \sigma^2 \neq \sigma_0^2$$

Test statistic: $\chi_0^2 = \frac{SS}{\sigma_0^2} = \frac{(n-1)S^2}{\sigma_0^2}$

Null hypothesis is rejected if $|\chi_0^2| > \chi^2_{1-\frac{\alpha}{2},(n-1)}$

$100(1-\alpha)$ Confidence interval $\frac{(n-1)S^2}{\chi^2_{\frac{\alpha}{2},(n-1)}} \leq \sigma^2 \leq \frac{(n-1)S^2}{\chi^2_{1-\frac{\alpha}{2},(n-1)}}$

Inferences about Variance

- Testing the equality of the variances of two normal populations. If independent random samples of size n_1 and n_2 are taken from populations 1 and 2, respectively,
- $$H_0: \sigma_1^2 = \sigma_2^2$$
  $$H_1: \sigma_1^2 \neq \sigma_2^2$$

Test statistic: $F_0 = \frac{S_1^2}{S_2^2}$

Null hypothesis is rejected if $F_0 > F_{1-\frac{\alpha}{2},n_1-1,n_2-1}$

$100(1-\alpha)$ Confidence interval for the ratio of variances

$$\frac{S_1^2}{S_2^2} F_{\frac{\alpha}{2},n_1-1,n_2-1} < \frac{\sigma_1^2}{\sigma_2^2} < \frac{S_1^2}{S_2^2} F_{1-\frac{\alpha}{2},n_1-1,n_2-1}$$

Test of Variances

■ **TABLE 2.8**

Tests on Variances of Normal Distributions

Hypothesis	Test Statistic	Fixed Significance Level Criteria for Rejection
$H_0: \sigma^2 = \sigma_0^2$ $H_1: \sigma^2 \neq \sigma_0^2$	$\chi_0^2 = \dfrac{(n-1)S^2}{\sigma_0^2}$	$\chi_0^2 > \chi_{\alpha/2,n-1}^2$ or $\chi_0^2 < \chi_{1-\alpha/2,n-1}^2$
$H_0: \sigma^2 = \sigma_0^2$ $H_1: \sigma^2 < \sigma_0^2$		$\chi_0^2 < \chi_{1-\alpha,n-1}^2$
$H_0: \sigma^2 = \sigma_0^2$ $H_1: \sigma^2 > \sigma_0^2$		$\chi_0^2 > \chi_{\alpha,n-1}^2$
$H_0: \sigma_1^2 = \sigma_2^2$ $H_1: \sigma_1^2 \neq \sigma_2^2$	$F_0 = \dfrac{S_1^2}{S_2^2}$	$F_0 > F_{\alpha/2,n_1-1,n_2-1}$ or $F_0 < F_{1-\alpha/2,n_1-1,n_2-1}$
$H_0: \sigma_1^2 = \sigma_2^2$ $H_1: \sigma_1^2 < \sigma_2^2$	$F_0 = \dfrac{S_2^2}{S_1^2}$	$F_0 > F_{\alpha,n_2-1,n_1-1}$
$H_0: \sigma_1^2 = \sigma_2^2$ $H_1: \sigma_1^2 > \sigma_2^2$	$F_0 = \dfrac{S_1^2}{S_2^2}$	$F_0 > F_{\alpha,n_1-1,n_2-1}$