PROBABILITY DISTRIBUTIONS AND STATISTICAL CONCEPTS

07.09.22

Statistics → Descriptive, Inferential

Descriptive Statistic → Describes the characteristics of a product or process from collected data.

Inferential Statistic → 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 Distribution → when the random variable can be expressed on a continuous scale. Eg- $0 \le x \le 1$, cable diameter, length/width of a rectangle.

Discrete Distribution → when the random variable can take only certain values, such as integers. Eg - # of nonconforming parts, # of defects on a circuit board, # customers served within certain time limit.

Discrete Probability Distribution →

Probability Mass Function (pmf): $f_x(x) = Pr[x = X]$

Cumulative Mass Function (cmf): $Pr[x \le X] = \sum_{x=0}^{X} f(x)$

Eg → Binomial, Poisson, Hypergeometric, Multinomial, Bernoulli, Negative Binomial.

Continuous Probability Distribution →

Probability Density Function (pdf): $f(x): Pr(a \le x \le b) = \int_a^b f(x)\,dx$

Cumulative Density Function (cdf): $Pr[x \le X] = F(x) = \int_{-\infty}^{X} f(t)\,dt$

Eg → Normal, Exponential, Weibull, Beta, Gamma, Cauchy, Uniform.

Expected Value →

Long run average of the random variable for the experiment it represents.

Discrete Case: $$E(X) = \sum_{i=1}^{\infty} x_i \times P(X = x_i)$$

Continuous Case: $$E(X) = \int_{-\infty}^{\infty} x \times p(x)\,dx$$

Variance →

The expected value of the squared deviation from the mean $\mu = E(X)$.

$$Var(X) = E(X-\mu)^2 = E(X-E(X))^2 = E(X^2) - [E(X)]^2$$

Discrete Case: $$Var(X) = \sum_{i=1}^{\infty} (x_i - \mu)^2 \times P(X = x_i)$$

Continuous case: $$Var(X) = \int_{-\infty}^{\infty} (x-\mu)^2 \times p(x)\,dx$$

Q. 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?

→ $P(x=1) = 0.8$, $P(x=2) = 0.18$, $P(x=3) = 0.02$

$E(X) = 1 \times 0.8 + 2 \times 0.18 + 3 \times 0.02 = 1.22$ (Ans)

$Var(X) = (1-1.22)^2 \times 0.8 + (2-1.22)^2 \times 0.18 + (3-1.22)^2 \times 0.02$

$= 0.2116$ (Ans)

Basics →

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

Population & Sample →

Population: The set of all items that possess a certain characteristic of interest.

Eg: 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.)

Sample: A subset of population

Eg: Average weight of all 50000 can produced in a particular month.

Parameter: Characteristic of a population

Statistic: A characteristic of a sample used to make inferences on population parameters.

Hypergeometric Distribution →

→ Discrete Probability Distribution

→ $$P(d) = \frac{\binom{D}{d}\binom{N-D}{n-d}}{\binom{N}{n}}$$

→ $$\mu = \frac{nD}{N}$$

→ $$\sigma = \sqrt{\frac{\frac{nD}{N}\left(1-\frac{D}{N}\right)(N-n)}{N-1}}$$

→ Sampling without replacement

→ EXCEL = HYPGEOM.DIST(# nc in sample, sample size, # of nc in pop, pop size, TRUE = cumulative / FALSE = prob. mass)

Q. 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?

→ $$P(1) = \frac{\binom{4}{1}\binom{8}{2}}{\binom{12}{3}} = \frac{4 \times 28}{220} = \frac{28}{55} = 0.509 \text{ (Ans)}$$

$$P(0) + P(1) + P(2) = 1 - P(3) = 1 - \frac{\binom{4}{3}\binom{8}{0}}{\binom{12}{3}} = 1 - \frac{4 \times 1}{220} = \frac{54}{55}$$

$$= 0.982 \text{ (Ans)}$$

Binomial Distribution:

→ Discrete Probability distribution

→ $P(d) = \frac{n!}{d!(n-d)!} p^d (1-p)^{n-d}$

→ p = proportion, n = no. of sample

→ $\mu = np$, $\sigma = \sqrt{np(1-p)}$

→ Sampling with replacement

→ EXCEL = BINOM.DIST(d, n, p, TRUE = cumulative / FALSE = prob. mass)

Q. 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?

→ $d = 2$, $n = 20$, $p = 0.03$

$$P(2) = \frac{20!}{2!\,18!} (0.03)^2 (0.97)^{18} = 0.0988 \text{ (Ans)}$$

Poisson Distribution:

→ Discrete probability distribution

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

→ x = count, λ = avg count, avg no. of events of a given classification occuring in a sample.

→ $\mu = \lambda$, $\sigma^2 = \lambda$

→ EXCEL = POISSON.DIST(x, λ, TRUE = cumulative / FALSE = prob. mass)

Q. Avg no. of nc units is 1.6, what is the probability that a sample will contain 2 or fewer nc units?

→ $\lambda = 1.6$

$$P(x \le 2) = P(0) + P(1) + P(2)$$
$$= e^{-\lambda} + \frac{e^{-\lambda}\lambda}{1!} + \frac{e^{-\lambda}\lambda^2}{2!}$$
$$= e^{-1.6}\left[1 + 1.6 + \frac{(1.6)^2}{2}\right]$$
$$= 0.7834 \text{ (Ans)}$$

✱ We can use the tables of the Poisson Cumulative Distribution.

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)

Q. Operating life of a mixer has a mean of 2200 h and standard deviation of 120h. What is the probability that a single electric mixer will fail to operate at 1900h or less?

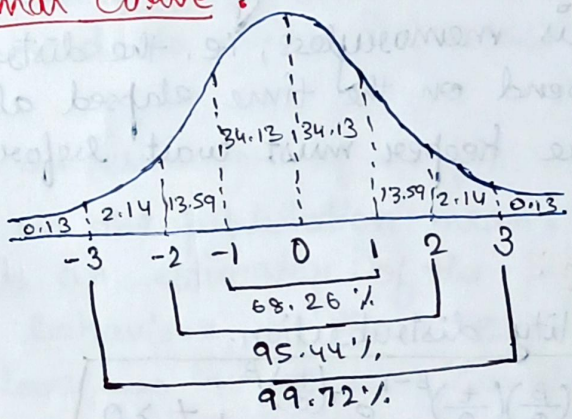
→ $\mu = 2200h$, $\sigma = 120h$

$$P(x \le 1900h) = \int_0^{1900h} \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} dx$$

$$= \frac{1}{120\sqrt{2\pi}} \int_0^{1900} e^{-\frac{(x-2200)^2}{2\times120\times120}} dx$$

$$= 6.2097 \times 10^{-3} \text{ (Ans)}$$

Standard Normal Curve:



Interrelationship:

→ Hypergeometric can be approximated by
- Binomial when $\frac{n}{N} \le 0.1$
- Poisson when $\frac{n}{N} \le 0.1$, $p_0 \le 0.1$ and $np_0 \le 5$
- Normal when $\frac{n}{N} \le 0.1$

→ Binomial can be approximated by
- Poisson when $p \le 0.1$ and $np \le 5$
- Normal when $p \sim 0.5$ and $n \ge 10$

Exponential Distribution:

→ Continuous Probability distribution

→ 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 = $\frac{1}{\lambda}$, Variance = $\frac{1}{\lambda^2}$

→ In reliability, λ = failure rate, $\frac{1}{\lambda}$ = mean time to failure.

→ Reliability at time t with mean life θ: $R_t = e^{-t/\theta}$

→ EXCEL = EXPON.DIST (x, λ, TRUE = cumulative / FALSE = prob. mass)

→ It is a probability distribution, that represents elapsed time b/w 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.
Eg- The time a store keeper must wait before arrival of a customer.

Weibull Distribution:

→ Continuous Probability distribution.

→ pdf: $f(t, \theta, \beta) = \begin{cases} \left(\frac{\beta}{\theta}\right)\left(\frac{t}{\theta}\right)^{\beta-1} e^{-\left(\frac{t}{\theta}\right)^{\beta}}, & t \geq 0 \\ 0, & t < 0 \end{cases}$

→ cdf: $F(t, \theta, \beta) = \begin{cases} 1 - e^{-\left(\frac{t}{\theta}\right)^{\beta}}, & t \geq 0 \\ 0, & t < 0 \end{cases}$

→ 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/FALSE)
(same func as before)

→ 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 behaviour of these estimators through repeated sampling allows us to draw conclusions about the corresponding parameters.

→ The behaviour 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 r.v. with mean μ_i and variance σ_i^2, and if $y = x_1 + \ldots + x_n$, then the distribution of $\frac{y - \sum_{i=1}^{n} \mu_i}{\sqrt{\sum_{i=1}^{n} \sigma_i^2}}$ approaches $N(0,1)$ distribution as n approaches infinity.

→ It implies that the sum of n independently distributed r.v. is approximately normal, regardless of the distribution of the individual variable.

→ If x_i are independent & 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.

→ 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 $\boxed{\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}}$, known as standard error.

Important Sampling Distributions derived from normal distribution:

1. **χ^2 distribution:** If $x_1, \ldots, x_n$ 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 $t = \frac{x}{\sqrt{y/k}}$ is distributed as t with k d.o.f.

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

Tests of Normality:

1. Normal Probability Plot of Residuals
2. Histogram
3. Boxplot
4. Skewness and Kurtosis
5. Chi-sq Test

Normal Probability Plot of Residuals:

1. Order the data
2. Rank the data i
3. Calculate plotting position $PP = 100 \times \frac{i - 0.5}{n}$
4. Plot the points on a normal probability plot paper

OR

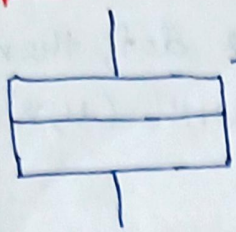
5. Take $z_i = \frac{i-0.5}{n}$, make a column of $\Phi^{-1}(z_i)$. Plot data (x_i) corresponding to rank i on x-axis, $\Phi^{-1}(z_i)$ on y-axis.
6. Fit a least fit line by observation.
7. Judgement on how close the points are to the straight line.

Histograms:

1. Tally grouped or ungrouped data.
2. Determine range $R = x_n - x_1$
3. Determine cell interval i
4. Determine cell midpoints $(MP_1 = x_1 + \frac{i}{2})$
5. Determine cell boundaries (extra decimal place).
6. Post cell and the frequencies.
7. Plot (x-axis – midpoints, Y-axis – freq / relative freq.)

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

Boxplot:



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

Skewness & Kurtosis:

1. Skewness: Determines the lack of symmetry of the data.

$$a_3 = \frac{\frac{\Sigma (x_i - \bar{x})^3}{n}}{s^3}$$

→ $a_3 = 0$ → Data is symmetrical.

→ $a_3 > 0$ → skewed to right (1 is extreme).
eg: Distribution of salary.

→ $a_3 < 0$ → skewed to left (-1 is extreme)
eg: Life of light bulb.

2. Kurtosis: Determines the peakedness of the data.

→ $$a_4 = \frac{\Sigma (x_i - \bar{x})^4}{n \cdot s^4} - 3$$

→ $a_4 = 0$ - normal distribution

→ $a_4 > 0$ - more peaked (leptokurtic)
eg: t-distribution

→ $a_4 < 0$ - less peaked (platykurtic)
eg: Beta distribution with shape parameters = 2.

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 or that which contradicts H_0.

Chi-squared Test:

→ Goodness of fit test
→ Create k bins for the data
→ Calculate $\chi^2 = \sum (O_i - E_i)^2 / E_i$

O_i → observed freq. in bin i
E_i → Expected freq. in bin i → $N \times (F(UB_i) - F(LB_i))$
where F is cumulative density.

→ check if $\chi^2 > \chi^2_{(1-\alpha)(k-c)}$ where k = no. of non-empty cells,
c = no. of distribution parameter + 1, c = 1 otherwise
→ for normal dist, dof = k − 3, for binomial dof = k − 2

8. A new casino game involves rolling 3 dice. The winnings are directly proportional to the total no. of sixes rolled. Suppose a gambler plays the game 100 times, with the following observed counts:

No. of sixes	No. of Rolls
0	48
1	35
2	15
3	3

Test whether the dices are fair.

→

No. of Sixes	No. of Rolls	Expected	$\frac{(f_o - f_e)^2}{f_e}$
0	48	$100 \times {}^3C_0 (\frac{5}{6})^3 = 57.87$	1.683
1	35	$100 \times {}^3C_1 (\frac{1}{6})(\frac{5}{6})^2 = 34.72$	0.015
2	15	$100 \times {}^3C_2 (\frac{1}{6})^2(\frac{5}{6}) = 6.94$	9.361
3	3	$100 \times {}^3C_3 (\frac{1}{6})^3 = 0.46$	14.025
			$\chi^2 = 25.084$

$\chi^2_{\frac{0.05}{2}, 2} = 7.38$

$\chi^2 \geq \chi^2_{crit}$ → reject H_0 → dices are not fair.

Example 1:

Single Sample - Distance travelled in 5 shots - 11, 13, 12, 10, 11

Question → Is the mean distance travelled = 12 inches?

↳ **1 sample T-test →**

→ Null Hypothesis: $H_0 : \mu = \mu_0$

→ Alternative Hypothesis: $H_1 : \mu \neq \mu_0$ (2 sided)

→ Test Statistic $\boxed{t_0 = \frac{\bar{y} - \mu_0}{s/\sqrt{n}}}$ $n \rightarrow$ sample size

→ Sample mean $\boxed{\bar{y} = \frac{1}{n} \sum y_i}$

→ Sample Standard Deviation $\boxed{s = \sqrt{\frac{\sum (y_i - \bar{y})^2}{n-1}}}$

t is a random variable following t-distribution with ν dof where $\boxed{\nu = n-1}$

→ 2 sided Test:

If $\boxed{|t_0| \geq t_{\frac{\alpha}{2}, n-1}}$ the null hypothesis is rejected.

→ 1 sided Test:

For $H_1 : \mu > \mu_0$, if $\boxed{t_0 \geq t_{\alpha, n-1}}$ then H_0 is rejected.

For $H_1 : \mu < \mu_0$, if $\boxed{t_0 \leq -t_{\alpha, n-1}}$ then H_0 is rejected.

→ Confidence Interval:

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

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

↳ 1 sample z-test (Known Variance) →

→ Null Hypothesis: $H_0: \mu = \mu_0$

→ Alternate Hypothesis: $H_1: \mu \neq \mu_0$

→ Test Statistic: $\boxed{z_0 = \frac{\bar{y} - \mu_0}{\sigma/\sqrt{n}}}$ $n \rightarrow$ sample size.

→ sample mean: $\boxed{\bar{y} = \frac{1}{n}\sum y_i}$

→ 2 sided Test:

If $\boxed{|z_0| \geq z_{\frac{\alpha}{2}}}$, the H_0 is rejected $\boxed{z_{0.025} = 1.96, \alpha = 0.05}$

→ 1 sided Test:

For $H_1: \mu > \mu_0$, if $\boxed{z_0 \geq z_\alpha}$, H_0 is rejected.

For $H_1: \mu < \mu_0$, if $\boxed{z_0 \leq -z_\alpha}$, H_0 is rejected.

→ Confidence Interval:

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

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

Example 2:

2 sample – dist travelled (pos 1) – 11, 13, 12, 10, 11 – dist travelled
(pos 2) – 17, 14, 13, 15, 15

Question → Does setting 2 travels more distance than setting 1?

↳ 2 sample z-test (known variance) →

→ Assumption:

→ $x_{11}, x_{12}, \ldots, x_{1n_1}$ – random sample from population 1 $\sim N(\mu_1, \sigma_1^2)$

→ $x_{21}, x_{22}, \ldots, x_{2n_2}$ – random sample from population 2 $\sim N(\mu_2, \sigma_2^2)$

→ Two populations are independent

→ Both populations are normal, if not, CLT applies.

→ Based on the assumption we state that:

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

→ Null Hypothesis: $H_0: \mu_1 - \mu_{02} = \Delta_0$ (often $\Delta_0 = 0$)

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

→ Alternate Hypothesis: $H_1: \mu_1 - \mu_2 \neq \Delta_0$

→ Rejection criteria: $|z_0| > z_{\frac{\alpha}{2}}$. $z_{0.025} = 1.96$

→ P-value: $P = 2[1 - \Phi(|z_0|)]$

→ Confidence Interval:

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

↳ 2 sample T-test →

→ Null Hypothesis: $H_0: \mu_1 = \mu_2$

→ Alternate Hypothesis: $H_1 = \mu_1 \neq \mu_2$

→ Test Statistic: $$t_0 = \frac{\bar{y}_1 - \bar{y}_2}{S_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}, \quad S_p = \sqrt{\frac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{n_1+n_2-2}}$$

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

→ dof: $n_1 + n_2 - 2$

→ Assumption: They both have equal variance.

→ 2 sided test: If $|t_0| \geq t_{\frac{\alpha}{2}, n_1+n_2-2}$, H_0 is rejected.

→ 1 sided test:

For $H_1: \mu_1 > \mu_2$, if $t_0 \geq t_{\alpha, n_1+n_2-2}$, H_0 is rejected.

For $H_1: \mu_1 < \mu_2$, if $t_0 \leq -t_{\alpha, n_1+n_2-2}$, H_0 is rejected.

→ when $\sigma_1^2 \neq \sigma_2^2$, $$t_0 = \frac{\bar{y}_1 - \bar{y}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

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

→ Confidence Interval:

100(1-α)% CI on the diff in mean $\mu_1 - \mu_2$ is:

→ For equal variance —

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

→ For unequal variance —

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

Errors:

1. Type I error: Null hypothesis is rejected when it is true.
2. Type II error: Null hypothesis is accepted when it is false.

→ $\alpha = P\{\text{Type I error}\} = P\{\text{reject } H_0 \mid H_0 \text{ is true}\}$
also called producer's risk, probability of a good lot rejected.

→ $\beta = P\{\text{Type II error}\} = P\{\text{fail to reject } H_0 \mid H_0 \text{ is false}\}$
also called consumer's risk, probability of a bad lot accepted.

→ $\text{Power} = 1-\beta = P\{\text{reject } H_0 \mid H_0 \text{ is false}\}$

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.

→ It 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:

→ When paired data are encountered.

→ $\boxed{y_{ij} = \mu_i + \beta_j + \varepsilon_{ij}}$, $i = 1,2$, $j = 1,2,\ldots,k$

μ_i → True mean response of i^{th} treatment

β_j → Effect on response due to j^{th} specimen

→ $\boxed{d_j = y_{1j} - y_{2j}}$, $j = 1,2,\ldots,k$

→ $\boxed{\mu_d = E(d_j) = \mu_1 + \beta_j - \mu_2 - \beta_j = \mu_1 - \mu_2}$

→ $H_0: \mu_d = 0$, $H_A: \mu_d \neq 0$

→ Test Statistic: $\boxed{t_0 = \frac{\bar{d}}{s_d/\sqrt{n}}}$, $\boxed{\bar{d} = \frac{1}{n}\sum d_j}$

$\boxed{s_d = \sqrt{\frac{\sum (d_j - \bar{d})^2}{n-1}}}$

→ $\boxed{|t_0| > t_{(1-\frac{\alpha}{2}, n-1)}}$ → H_0 rejected.

→ Paired comparison design is a special case of "blocking".

Ex:

Sl. no.	x_1	x_2	$d_j (x_1 - x_2)$
1	7	6	1
2	3	3	0
3	3	5	-2
4	4	3	1
5	8	8	0
6	3	2	1
7	2	4	-2
8	9	9	0
9	5	4	1
10	4	5	-1

$\bar{d} = \frac{1}{10} \times (-1) = -0.1$

$s_d^2 = \frac{13 - \frac{1}{10}}{9} = \frac{13 - 0.1}{9} = \frac{12.9}{9} = 1.43$

$t_0 = \frac{-0.1}{1.43/\sqrt{10}} = -3.162$

$t_{0.025,9} = 2.262$

$|t_0| > t_{0.025,9}$

↳ rejected. H_0.

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,

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

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

H_0 is rejected if $\boxed{|\chi_0^2| > \chi^2_{1-\frac{\alpha}{2},(n-1)}}$

$100(1-\alpha)\%$ CI →

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

→ 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: $\boxed{F_0 = \frac{S_1^2}{S_2^2}}$

H_0 is rejected if $\boxed{F_0 > F_{1-\frac{\alpha}{2},n_1-1,n_2-1}}$

$100(1-\alpha)\%$ CI → ratio of variances

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

11.09.22

SPC TOOLS AND VISUAL METHODS

Perspectives:

1. Designing a new product, process, service.
2. Maintaining prescribed quality.
3. Improving an existing product, process, service.

"What cannot be measured, cannot be managed." — Peter Drucker.

Outline:

1. Learn about the process: Ask for the SOP
2. Investigate the issues: C&E, check sheet, Pareto sheet.
3. Tools for control / Improvement.

Magnificent seven:

1. Histogram or stem & Leaf Plot
2. Check Sheet
3. Pareto Chart
4. Cause and effect diagram.
5. Defect concentration diagram
6. Scatter diagram
7. Control Chart.

Statistical Process Control (SPC):

If a product is to meet or exceed customer expectations, generally it should be produced by a process that is stable or repeatable. More precisely, the process must be capable of operating with little variability around the target or nominal dimensions of the product's quality characteristics. Statistical Process Control (SPC) is a powerful collection of problem-solving tools useful in achieving process stability and improving capability through the reduction of variability.

Stable system of chance causes:

In the framework of statistical quality control, the natural variability is often called a "stable system of chance causes". A process that is operating with only chance causes of variation present is said to be in statistical control. In other words, the chance causes are an inherent part of the process.

Assignable causes of Variation:

Other kinds of ~~variation~~ variability usually arises from these sources: improperly adjusted or controlled machines, operator errors, or defective raw material. It is generally large compared to background noise and usually represents an unacceptable level of process performance. These sources of variability that are not part of the chance cause pattern are assignable causes of variation. A process that is operating in the presence of assignable causes is called to be an out-of-control process.

When the process is in control, most of the production will fall between the lower and upper specification limits (LSL/USL).

When the process is out of control, a higher proportion of the process lies outside of these specifications.

Objective of Statistical Process Control:

To quickly detect the occurence of assignable causes of process shifts so that investigation of the process and corrective action may be undertaken before many non conforming units are manufactured.

SPC: Statistical Process Control

SQC: Statistical Quality Control

ASQ: American Society for Quality

7 Quality-control Tools (7-QC):

1. Cause & Effect Diagram
2. Check Sheet
3. Control Chart
4. Histogram
5. Pareto chart
6. Scatter Plot
7. Stratification

7 Supplemental Tools (7-SUPP):

1. Data Stratification
2. Defect Maps
3. Event logs
4. Process flowcharts
5. Progress centres
6. Randomization
7. Sample size determination.

Graphical and Visual Tools:

1. Standard Operating Procedure – Learn about the process, also new SOP once the change is implemented.
2. Check Sheet
3. Cause and Effect Diagram (Fishbone diagram)
4. Pareto Charts

} Identify the problems and the extent of causes

5. Quality Function deployment – Voice of customer to design of product, process
6. Value stream mapping – Efficiency of Process

Standard Operating Procedure (SOP):

→ A step by step guide compiled by an organization to help workers carry out complex routine operations.
→ SOPs aim to achieve efficiency, quality output and uniformity of performance while reducing miscommunication and failure to comply with industry regulations.
→ It should follow 4 C's, Clear, Complete, Concise, Courteous & Correct.
→ SOP helps smoothing the transition process from one worker to another.
→ The sections may include:
- Purpose/objective
- Scope
- Responsibilities
- Accountability
- Procedure.

Check Sheet:

→ When to use?
→ when data can be observed and collected repeatedly
→ when collecting data on the frequency of patterns, problems, defects, issues
→ Production process
→ How to use?
→ Decide what problems are observed. Define the problems.
→ Decide the duration and length of data collection.
→ Design a form, so that data can be recorded simply by ~~producing~~ putting check marks, 'x's or numbers.
→ Test for a short trial period and then implement in appropriate situations.

Cause & Effect Diagram (Fishbone Diagram):

→ When to use?
→ To identify possible cause of a problem.
→ when the problem is too complicated to be resolved by experts in specific fields.

→ How to use?
  → Agree on a problem statement.
  → Brainstorm major categories of causes of the problem. The generic categories are:
    • Methods • Machines • Manpower • Materials • Measurement • Environment
  → Write the causes as branches of the main problem.
  → Ask 'why' for each of the major causes, then write sub-causes as the branches of the main causes. Generate deeper levels of causes.

Pareto Chart:

→ When to use?
  → To analyze frequency of problems or causes of a problem.
  → When there are multiple problems in a process and you want to focus on important few.
  → Communicating distribution of problems/causes.
→ How to use?
  → Collect data on number of occurences of problems/causes.
  → Order the data in descending order of frequencies.
  → Create a bar chart with the frequency (or %) on y-axis, problem category labels on x-axis.
  → From the top of the highest bar, draw a line digram using cumulative frequency of problem categories.

Defect Concentration Diagram:

→ Picture of a unit showing all the relevant views and the associated defects.
→ Defects are color-coded- helps ~~identify~~ identifying the source of the defects.

Quality Function Deployment (House of Quality):

→ A tool "To satisfy or even delight the customers, QFD is an essential tool" – ASQ.

→ Focused on "voice of the customer"

→ Tool to design quality product incorporating customer needs.

→ Evaluates competitors on two perspectives: customers' and technical

→ QFD cuts down on time that would otherwise be spent on product redesign.

→ QFD is also used to create training programs, hire new employees, establish supplier development criteria and improve services.

→ Needs a cross-functional team for data collection and analysis.

Four Phases →

Performance Requirements | System Design | Process Design | Process Control

Lean Management:

→ "Value" is something for which your customer is willing to pay. These are called value-adding activities.

→ Everything else falls under the category of "waste".

→ Taichi Ohno, an architect of Toyota production system conceived this idea of lean, and devoted his carreer in elimenting waste from production process.

→ 7 types of waste: Transport, Inventory, Motion, Waiting, Overproduction, Over-processing, Defects.

→ Pure waste: Any activity that does not bring value and damages efficiency.

→ Necessary waste: Activities that our customer does not want to pay for but is necessary to provide value for the end product.

Value Stream Mapping (VSM):

→ Representation of the flow of material and information from supplier to customer through your organization.

→ It enables you to see where the delays are in the process, if there is bottleneck, excersive inventory or other restraints.

→ You create your current state map and work towards producing your ideal state map.

VARIABLE CONTROL CHART PROCESS CAPABILITY

12.09.22

Sampling:

→ It is not always possible to measure quality characteristics of each item in a population.
→ Samples are used to provide information about process or product characteristics at a fraction of cost.
→ Necessary for destructive tests
→ A sampling design is a procedure by which the observations in a sample are chosen from the population.
→ An element is an object for which data are gathered.
→ A sampling unit is an individual element or a collection of elements from a population.
→ A sampling frame is a list of sampling units.

Sampling Errors:

1. Random Variation – Inherent sampling variability, e.g. due to instrument, people, etc.
2. Misspecification – Happens in opinion polling, customer satisfaction survey, incorrect listing of sampling frame.
3. Non responses – Happens in sample surveys, cases where measurements not possible.

Sampling Methods:

1. Simple Random Sampling
2. Stratified Sampling
3. Cluster Sampling

Simple Random Sampling:

→ A sample of size n is chosen from a fixed population of size N. In SRS, each possible sample of size n has equal chance of being selected.

→ Random number tables may be used for sampling.

→ In estimating population mean μ by the sample mean $\bar{x}$, the variance of the estimator is given by

$$\hat{\sigma}^2_{\bar{x}} = \frac{\frac{s^2}{n}(N-n)}{N} \qquad (s^2 \text{ sample variance})$$

$\frac{N-n}{N}$ is called finite population correction factor.

→ Precision is inverse of variance.

Stratified Random sample:

→ Useful when the population is heterogeneous, eg. production from multiple machines, under multiple operators, samples from different geographical regions.

→ They are obtained by obtained by separating the elements of the population in nonoverlapping groups (strata).

→ Proportional allocation of sample size, for k strata, let N_i be the population size of the i^{th} strata, and $\sum N_i = N$.

Sample size from each strata: $n_i = \frac{n N_i}{N}$, $i = 1, 2, \ldots, k$

sample mean and variance of estimator are given by

$$\bar{x}_{st} = \frac{1}{N} \sum N_i \bar{x}_i$$

$$Var(\bar{x}_{st}) = \frac{1}{N^2} \sum N_i^2 \left(\frac{(N_i - n_i)}{N_i}\right)\left(\frac{s_i^2}{n_i}\right), \quad i = 1, 2, \ldots, k$$

where $\bar{x}_i = \frac{1}{n_i} \sum_{j=1}^{n_i} x_{ij}$, $\quad s_i^2 = \sum_{j=1}^{n_i} \frac{(x_{ij} - \bar{x}_i)^2}{n_i - 1}$

Cluster sampling:

→ When a sampling frame is not available or obtaining samples from all segments of the population is not feasible due to geographical reasons, cluster sampling is used.
→ Population is divided into groups of elements, called clusters.
→ Clusters are randomly selected and a census data is obtained.
→ Sampling error maybe reduced by choosing many small clusters rather than choosing a large clusters.

Ex:
A researcher wants to conduct a study to judge the performance of sophomore's in business education across the India.
By using cluster sampling, the researcher can club the universities from each city into one cluster. These clusters then define all the sophomore student population in India. Next, either using simple random sampling or systematic random sampling, randomly pick clusters for the research study. Subsequently, by using simple or systematic sampling, the sophomore's from each of these selected clusters can be chosen on whom to conduct the research study.

How to choose sample size?

→ Bound on error estimation on <u>population mean</u>:
Let there is $(1-\alpha)$ probability that the difference b/w the estimated mean and the actual mean is not greater than B (tolerable error bound).

$$B = Z_{\frac{\alpha}{2}}\sigma_{\bar{x}} = Z_{\frac{\alpha}{2}}\left(\frac{\sigma}{\sqrt{n}}\right) \Rightarrow \boxed{n = Z^2_{\frac{\alpha}{2}}\left(\frac{\sigma}{B}\right)^2}$$

Q. An analyst wishes to estimate the average bore size of a large casting. Based on historical data, it is estimated that the standard deviation of the bore size is 4.2 mm. If it is desired to estimate with a probability of 0.95 the average bore size to within 0.8 mm, find the appropriate sample size.

→ $\sigma = 4.2\text{mm}$, $\alpha = 0.05$, $B = 0.8\text{ mm}$

$$n = z^2_{\frac{\alpha}{2}} \left(\frac{\sigma}{B}\right)^2 = z^2_{0.025} \left(\frac{4.2}{0.8}\right)^2 = (1.96)^2 \times \left(\frac{4.2}{0.8}\right)^2 = 105.8841 \approx 106 \text{ (Ans)}$$

→ Bound on error estimation on population proportion:
Let there is $(1-\alpha)$ probability that the difference b/w the estimated proportion $\hat{p}$ and the actual proportional p is not greater than B (tolerable error bound).
Eg: proportion of satisfied customers, prop of non-conforming

$$B = z_{\frac{\alpha}{2}} \sigma_{\hat{p}} = z_{\frac{\alpha}{2}} \sqrt{\frac{p(1-p)}{n}}$$

$$\Rightarrow \boxed{n = z^2_{\frac{\alpha}{2}} \frac{p(1-p)}{B^2}}$$

Either put $p = \hat{p}$ (sample proportion) or $p = 0.5$ for conservative estimate.

Q. we want to estimate with a probability of 0.90 the proportion of non-conforming tubes to within 4%. How large a sample should be chosen if no prior information is available on the process?

→ $\alpha = 0.10$, $p = 0.5$, $B = 0.04$

$$n = z^2_{\frac{\alpha}{2}} \frac{p(1-p)}{B^2} = z^2_{\frac{0.10}{2}} \frac{p(1-p)}{B^2} = (1.645)^2 \frac{(0.5)^2}{(0.04)^2}$$

$$= 422.816 \approx 423 \text{ (Ans)}$$

→ Estimating difference b/w 2 population means:

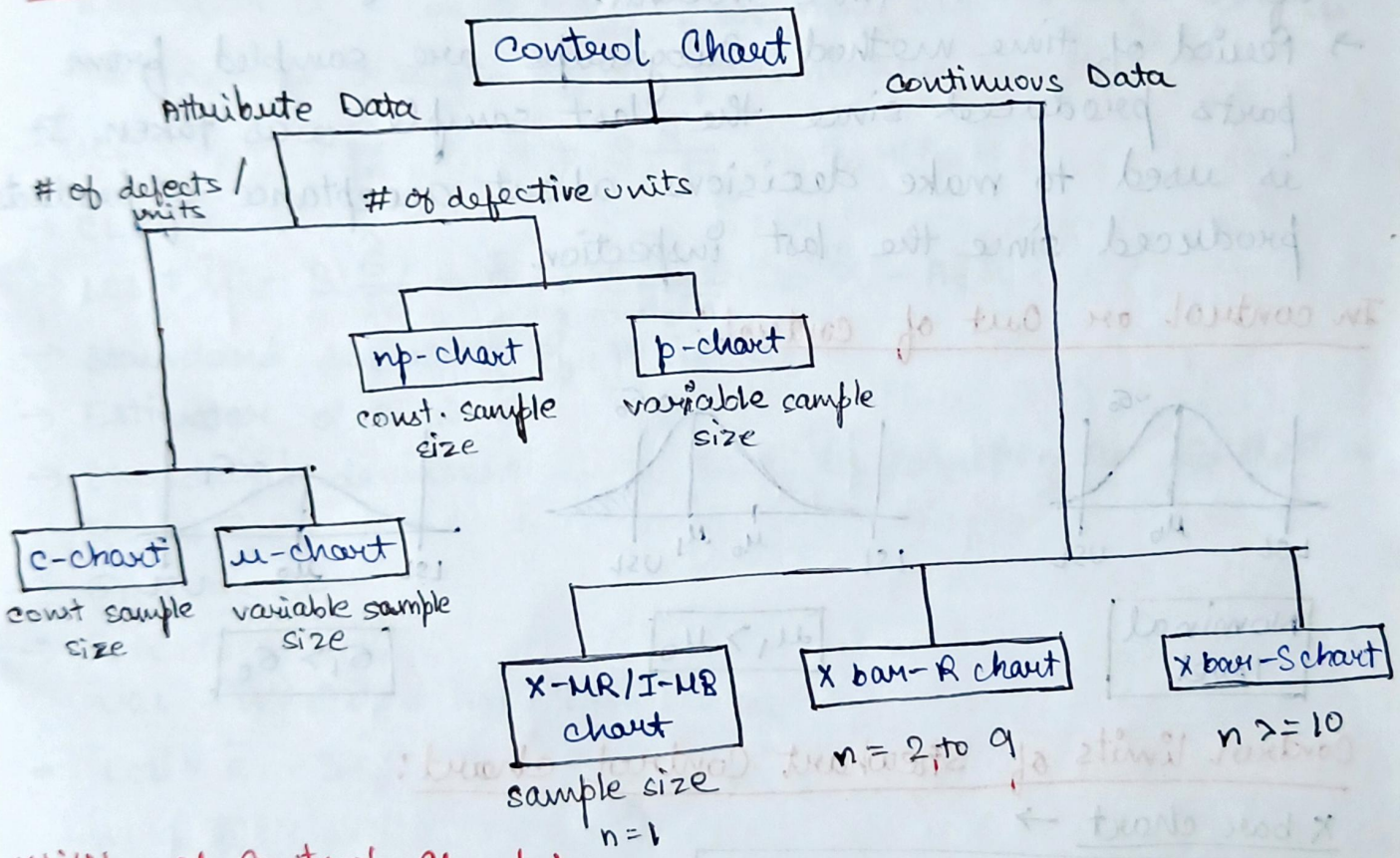
$$n = Z_{\frac{\alpha}{2}}^2 \left(\frac{\sigma_1^2 + \sigma_2^2}{B^2} \right)$$

where B is the tolerance of error for estimating the diff in population means with sample means.

→ Estimating diff b/w 2 population proportions:

$$n = Z_{\frac{\alpha}{2}}^2 \left(\frac{p_1(1-p_1) + p_2(1-p_2)}{B^2} \right)$$

Control Charts:



Utility of Control Charts:

→ Control charts are proven techniques to improve productivity.
→ Effective in defect identification and prevention.
→ Control charts ~~do~~ prevent unnecessary process adjustments
→ Diagnostic information.
→ Process capability information.

Rational Subgroups:

→ Sampling procedure to ensure that the variation within the group is only due to chance causes.

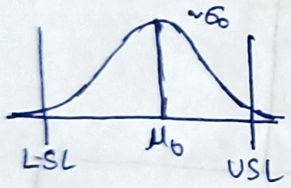
→ Lots from which the subgroups are chosen should be homogeneous, e.g. same machines, same operator, same mold cavity, etc.

→ Items of any one subgroup should be produced under essentially same conditions.

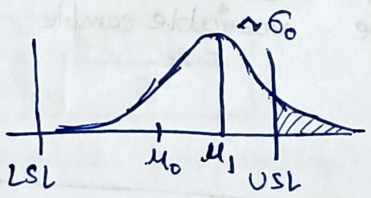
→ Instant time method: Parts in the subgroup are chosen in the same time-instant. The next subgroup is picked after a certain time interval.

→ Period of time method: Subgroups are sampled from parts produced since the last sample was taken. It is used to make decisions about acceptance of products produced since the last inspection.

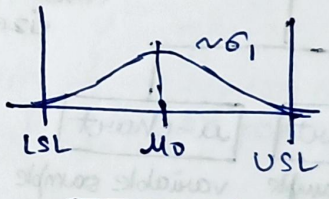
In control or Out of control:



Nominal Level



$\mu_1 > \mu_0$



$\sigma_1 > \sigma_0$

Control limits of Shewhart Control chart:

x bar chart →

$$UCL_{\bar{x}} = \bar{\bar{x}} + 3\sigma_{\bar{x}} \approx \bar{\bar{x}} + A_2\bar{R}$$

$$LCL_{\bar{x}} = \bar{\bar{x}} - 3\sigma_{\bar{x}} \approx \bar{\bar{x}} - A_2\bar{R}$$

$CL_{\bar{x}} = \bar{\bar{x}}$ Center line

R chart →

$$UCL_{\bar{R}} = \bar{R} + 3\sigma_{\bar{R}} \approx D_4\bar{R}$$

$$LCL_{\bar{R}} = \bar{R} - 3\sigma_{\bar{R}} \approx D_3\bar{R}$$

Derivation:

→ $\bar{\bar{x}} = \frac{1}{m}\sum_{i=1}^{m}\bar{x}_i = \frac{1}{mn}\sum_{i=1}^{m}\sum_{j=1}^{n}x_{ij}$ Centre line

→ UCL & LCL these are 3 sigma

→ $x \sim N(\mu, \sigma)$, $\bar{x} \sim N(\mu, \frac{\sigma}{\sqrt{n}})$

→ $UCL = \bar{\bar{x}} + 3\sigma_{\bar{x}} = \bar{\bar{x}} + 3\frac{\sigma}{\sqrt{n}}$

→ $LCL = \bar{\bar{x}} - 3\sigma_{\bar{x}} = \bar{\bar{x}} - 3\frac{\sigma}{\sqrt{n}}$

→ Relative Range $W = \frac{R}{\sigma}$ a random variable.

→ Parameters of distribution of W depend on sample size n

→ Mean of W is d_2

→ Estimator of σ is $\hat{\sigma} = \frac{R}{d_2}$, we may use $\hat{\sigma} = \frac{\bar{R}}{d_2}$ as $\bar{R}$ is the average range of m preliminary samples.

→ $UCL = \bar{\bar{x}} + 3\frac{\hat{\sigma}}{\sqrt{n}} = \bar{\bar{x}} + 3\frac{\bar{R}}{d_2\sqrt{n}} = \bar{\bar{x}} + A_2\bar{R}$

→ $CL = \bar{\bar{x}}$

→ $LCL = \bar{\bar{x}} - 3\frac{\hat{\sigma}}{\sqrt{n}} = \bar{\bar{x}} - 3\frac{\bar{R}}{d_2\sqrt{n}} = \bar{\bar{x}} - A_2\bar{R}$

→ Standard deviation of W is d_3

→ Estimator of σ is $\hat{\sigma} = R/d_2$

→ Standard deviation of R can be written as $\sigma_R = d_3\sigma$ as $R = W\sigma$

→ $\hat{\sigma}_R = d_3\frac{\bar{R}}{d_2}$

→ For R chart,

$UCL = \bar{R} + 3\hat{\sigma}_R = \bar{R} + \frac{3d_3\bar{R}}{d_2} = D_4\bar{R}$

$LCL = \bar{R} - 3\hat{\sigma}_R = \bar{R} - \frac{3d_3\bar{R}}{d_2} = D_3\bar{R}$

$CL = \bar{R}$

Revised Control limits:

Discard out of control samples with assignable causes. Revised control limits are calculated as below, for total no. of samples m and no. of defective samples d:

$$\bar{\bar{x}}_{new} = \frac{m\bar{\bar{x}} - \Sigma_d \bar{x}_d}{m-d} = \bar{x}_0$$

$$\bar{R}_{new} = \frac{m\bar{R} - \Sigma_d R_d}{m-d} = \bar{R}_0$$

$$\sigma_0 = \frac{\bar{R}_0}{d_2}$$

$$UCL_{\bar{x}} = \bar{x}_0 + A\sigma_0 \; ; \; LCL_{\bar{x}} = \bar{x}_0 - A\sigma_0$$

$$UCL_{\bar{R}} = D_2\sigma_0 \; ; \; LCL_{\bar{R}} = D_1\sigma_0$$

X-bar and R chart:

→ $\bar{x}$ chart monitors b/w sample variability, R chart monitors within sample variability.

→ To design $\bar{x}$-R chart, the following must be specified:
- Sample size
- Control limit width
- Frequency of sampling.

→ $\bar{x}$ chart is capable to signal moderate to large process shifts (2σ or larger)

→ R chart is relatively insensitive to shift in process standard deviation for small samples, eg: n=5

Error in making inference:

→ Type I error → This error results from inferring a process is out of control when it is not. It is denoted by α. This happens due to chance causes, when a control charts falls outside control limits. For 3σ limits, probability for type I error is 0.0027.

→ Type II error → This error results from inferring a process is in control when it is out of control. It is denoted by β. This can happen when the process mean or the process variability or both have changed.

Process capability:

Measurement of the common cause variation / system quality:

→ 6σ derive fairly measures common cause variation.

→ 18.4 hours (you can pretty much count on range less than that).

x bar and S chart:

→ It is occasionally desirable to monitor process standard deviation directly, rather than indirectly as done in R chart.

→ $\bar{x}$ and s chart are preferable when

- The sample size n is moderately large for $n > 10$ or 12.
- The sample size is variable.

→ The unbiased estimator of population variance σ^2 is sample variance s^2

$$s^2 = \frac{\sum (x_i - \bar{x})^2}{n-1}$$

→ The sample sd estimates $c_4\sigma$, sd of s is $\sigma\sqrt{1-c_4^2}$

→ Since $E(s) = c_4\sigma$ the center line is $c_4\sigma$. The 3σ sigma limits of the s-chart is given by

$$UCL = c_4\sigma + 3\sigma\sqrt{1-c_4^2}$$

$$CL = c_4\sigma$$

$$LCL = c_4\sigma - 3\sigma\sqrt{1-c_4^2}$$

X bar and S chart with sample estimators:

→ Consider $\frac{\bar{s}}{c_4}$ as an unbiased estimator of σ

→ For m preliminary samples with sd s_i, the average of m standard deviation is given by $\bar{s} = \frac{1}{m}\sum s_i$.

$$UCL = \bar{s} + \frac{3\bar{s}}{c_4}\sqrt{1-c_4^2} = B_4\bar{s}$$

$$CL = \bar{s}$$

$$LCL = \bar{s} - \frac{3\bar{s}}{c_4}\sqrt{1-c_4^2} = B_3\bar{s}$$

$$B_3 = 1 - \frac{3}{c_4}\sqrt{1-c_4^2}$$

$$B_4 = 1 + \frac{3}{c_4}\sqrt{1-c_4^2}$$

→ Control limit for corresponding $\bar{X}$, chart is given by

$$UCL_{\bar{x}} = \bar{\bar{x}} + \frac{3\bar{s}}{c_4\sqrt{n}} = \bar{\bar{x}} + A_3\bar{s}$$

$$CL_{\bar{x}} = \bar{\bar{x}}$$

$$LCL_{\bar{x}} = \bar{\bar{x}} - \frac{3\bar{s}}{c_4\sqrt{n}} = \bar{\bar{x}} - A_3\bar{s}$$

Probability of an Xbar False Alarm:

Q. Assume that the subgroups are rational and the system is under control, what is the probability of a false alarm on the next subgroup from the x bar chart?

→ CLT → $\bar{X} \sim N\left[\mu, \frac{\sigma}{\text{sqrt}(n)}\right]$

$$1 - P_r\left\{LCL_{xbar} = \mu - \frac{3\sigma}{\sqrt{n}} \le X_{bar} \le UCL_{xbar} = \mu + \frac{3\sigma}{\sqrt{n}}\right\}$$

$$= 1 - P_r\{-3 \le Z \le 3\} = 1 - 2P_r\{Z \le -3\} = 0.0027$$

Average run length in control $\frac{1}{0.0027} = 370.4$.

Average Run length (ARL):

→ To measure the performance of a control chart, ARL is used.

→ ARL denotes the no. of samples, on average, required to detect and out of control signal.

→ If P_d is the probability that a process is out of control then run length is 1 with probability P_d, 2 with probability $(1-P_d)P_d$, 3 with $(1-P_d)^2 P_d$. Hence

$$ARL = \sum_{j=1}^{\infty} j(1-P_d)^{j-1} P_d = \frac{P_d}{[1-(1-P_d)]^2} = \frac{1}{P_d}$$

→ For a process in control P_d is α (probability of type I error)

→ For an in-control process, ARL should be as large as possible.

→ For an out of control process $P_d = 1-\beta$. β is probability of type II error. $ARL = \frac{1}{1-\beta}$

→ For an out of control process, ARL should be as small as possible.

Analysis of Pattern in Control Chart:

→ Process should be investigated when there is non-random pattern in control chart.

→ Most sample averages are below centre line.

→ Continuous rise (run-up) or fall (run-down)

→ A run length of 8 or consecutive 8 points above or below centre line may indicate out-of-control.

→ Cycles are another type of pattern.

→ A mixture pattern when most points are near control limits, result of two or more distributions.

→ A shift in the process may occur due to introduction of new operator, material, machine, inspection method, etc.

→ A trend occurs when there is continuous deterioration of tools. In chemical processes they occur due to settling or seperation of components.

→ Stratification is when points cluster artificially near centre line. This may happen when rational subgrouping is not done.

Rules of Identifying Out of Control Points:

1. If single point plots are outside control limits.
2. If 2 out of 3 consecutive points plots fall outside 2σ warning limits on the same side of centre line.
3. If 4 out of 5 consecutive points fall beyond 1σ limit on the same side of centre line.
4. If 9 or more consecutive points fall on one side of centre line.
5. If 6 or more consecutive points steadily increases or decreases.

Process Capability:

→ **Assumption:**

→ The quality characteristic has normal distribution.

→ Process is in statistical control.

→ **Process Capability Ratio (PCR):** $\boxed{\hat{C}_p = \frac{USL - LSL}{6\hat{\sigma}}}$

→ $\boxed{P = \left(\frac{1}{C_p}\right) 100}$ gives the % of the specification band used by the process.

→ $\boxed{C_{pk} = \min\left(C_{pu} = \frac{USL - \mu}{3\sigma}, C_{pl} = \frac{\mu - LSL}{3\sigma}\right)}$ is used to determine if the process is centered.

→ More $C_{pk} < C_p$, there is right shift of the process.

Type I and Type II Error in CC:

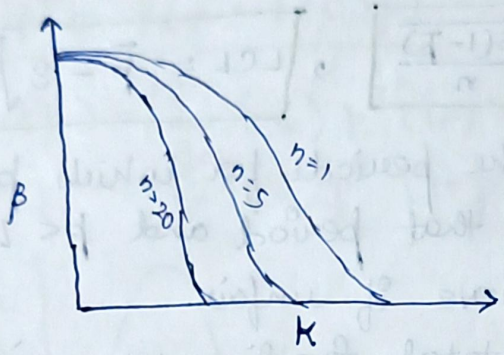
→ **Type I error:** Detecting a shift in process when there is no shift (False alarm). Wider the control limits, lower is the probability of Type I error.

→ **Type II error (β):** Not detecting a shift in process when there is a shift.

→ Closer the control limits, smaller is the probability of Type II error.

→ Decreases with increase in sample size.
→ OC curve is used to as a visual tool to analyze the change in process parameter.
→ Sample size should be chosen judiciously, so that a balance in probabilities of Type I and Type II error is maintained.

OC curve for $\bar{x}$ chart:



→ Consider the process mean has shifted from μ_0 to $\mu_0 + k\sigma$

→ $\beta = P(UCL \le \bar{x} \le LCL \mid \mu = \mu_1 = \mu_0 + k\sigma)$.

→ $$\boxed{\beta = \Phi\left(\frac{UCL - (\mu_0 + k\sigma)}{\frac{\sigma}{\sqrt{n}}}\right) - \Phi\left(\frac{LCL - (\mu_0 + k\sigma)}{\frac{\sigma}{\sqrt{n}}}\right)}$$

→ $$\beta = \Phi\left(\frac{\mu_0 + \frac{L\sigma}{\sqrt{n}} - (\mu_0 + k\sigma)}{\frac{\sigma}{\sqrt{n}}}\right) - \Phi\left(\frac{\mu_0 - \frac{L\sigma}{\sqrt{n}} - (\mu_0 + k\sigma)}{\frac{\sigma}{\sqrt{n}}}\right)$$

→ $$\boxed{\beta = \Phi(L - k\sqrt{n}) - \Phi(-L - k\sqrt{n})}$$

ATTRIBUTE CONTROL CHARTS, I-MR CHART

13.09.22

P- chart: (fraction of non-conforming)

Step 1 → Obtain the total fraction of nonconforming units or systems using 25 rational subgroups each of size n.

Step 2 → Calculate trial limits:

$$UCL = \bar{p} + 3\sqrt{\frac{\bar{p}(1-\bar{p})}{n}}, \quad LCL = \bar{p} - 3\sqrt{\frac{\bar{p}(1-\bar{p})}{n}}$$

Step 3 → Identify all the periods for which p = fraction non-conforming in that period and $p < LCL_{trial}$ or $p > UCL_{trial}$. Investigate, remove if unfair.

Step 4 → Calculate the total fraction nc using remaining.

New $\bar{p}$ = "process capability". Calculate revised limits.

Step 5 → Plot the fraction nonconforming p_j, for each period j and alert designated local authority if out-of-control signals occur.

np - chart: (# of non-conforming)

→ Subgroup size needs to be constant

→ Calculate trial limits:

$$UCL = n\bar{p} + 3\sqrt{n\bar{p}(1-\bar{p})}, \quad LCL = n\bar{p} - 3\sqrt{n\bar{p}(1-\bar{p})}$$

Variable Sample Size for non-conforming attribute Control Charts:

1. **Variable Control limits**:

$$UCL = \bar{p} + 3\sqrt{\frac{\bar{p}(1-\bar{p})}{n_i}} \qquad LCL = \bar{p} - 3\sqrt{\frac{\bar{p}(1-\bar{p})}{n_i}}$$

2. **Control limits based on average sample size**:

$$UCL = \bar{p} + 3\sqrt{\frac{\bar{p}(1-\bar{p})}{\bar{n}}} \qquad LCL = \bar{p} - 3\sqrt{\frac{\bar{p}(1-\bar{p})}{\bar{n}}}$$

where $\bar{n} = \frac{\sum n_i}{m}$ is the avg sample size.

3. Standardized control charts:

$$Z_i = \frac{p_i - \bar{p}}{\sqrt{\frac{\bar{p}(1-\bar{p})}{n_i}}}$$ with control limits +3 and -3.

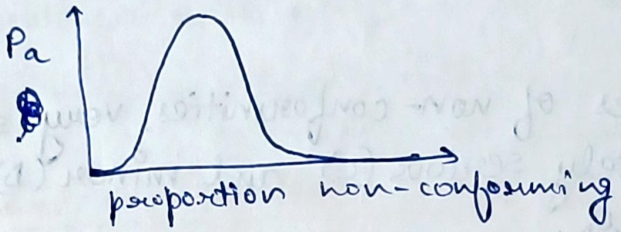
OC - Curve for p-chart:

→ Probability of Type II error.

→ $\beta = P(\hat{p} \le UCL \mid p) - P(\hat{p} \le LCL \mid p) = P(D \le nUCL \mid p) - P(D \le nLCL \mid p)$

→ $P(x \le nUCL \mid p) = \sum_{0}^{nUCL} \binom{n}{x} p^x (1-p)^{1-x}$

→ $P(x \le nLCL \mid p) = \sum_{0}^{nLCL} \binom{n}{x} p^x (1-p)^{1-x}$



α curve p-chart

C - chart: (count of non-conformities)

Step 1: Collect data

Step 2: Subgroup size is one inspected unit.

Step 3: Calculate trial limits

$UCL_{trial} = \bar{c} + 3\sqrt{\bar{c}}$, $LCL_{trial} = \max\{\bar{c} - 3\sqrt{\bar{c}}, 0\}$

Step 4: Find out of control signals. Remove if unfair. Revise limits. Revised $\bar{c}$ is process capability.

Step 5: Plot and local authority investigates if out-of-control signals occur (can act).

$CL_{trial} = \bar{c}$ (avg. count of non-conformities)

U- chart: (count of non-conformities / unit)

Step 1: Collect data for m periods. $\bar{u} = \frac{\sum c_i}{\sum n_i}$ where c_i is the count of defects in each subgroup.

Step 2: Calculate trial limits for i^{th} sample:

$$UCL_{trial} = \bar{u} + 3\sqrt{\frac{\bar{u}}{n_i}}, \quad LCL_{trial} = \max\left(\bar{u} - 3\sqrt{\frac{\bar{u}}{n_i}}, 0\right)$$

Step 3: Find out of control signals. Remove if unfair.

Step 4: Revise limits. Revised $\bar{u}$ is process capability.

Step 5: Plot & local authority investigates if out-of-control signals occur. (can act).

Demerit Chart:

→ Determine 4 classes of non-conformities, very serious (A_0), serious (B) moderately serious (C) and minor (D). Assign weights ω_A, ω_B, ω_C and ω_D.

→ For sample size n, let c_A, c_B, c_C and c_D denote the total no. of defects of each class.

→ Determine standard non-conformities per unit $\bar{u}_A, \bar{u}_B, \bar{u}_C, \bar{u}_D$

→ $D = \omega_A c_A + \omega_B c_B + \omega_C c_C + \omega_D c_D$

→ Demerits per unit is given by $U = \frac{D}{n}$

→ $CL = \bar{U} = \omega_A \bar{u}_A + \omega_B \bar{u}_B + \omega_C \bar{u}_C + \omega_D \bar{u}_D$

→ $\sigma_{0u} = \sqrt{\frac{\omega_A^2 \bar{u}_A + \omega_B^2 \bar{u}_B + \omega_C^2 \bar{u}_C + \omega_D^2 \bar{u}_D}{n}}$

→ $UCL = \bar{U} + 3\sigma_{0u}$, $LCL = \bar{U} - 3\sigma_{0u}$

Chart Comparison:

Methods	Advantages	Disadvantages
xbar & R charting	Uses fewer inspections, gives greater sensitivity.	Requires 2 or more charts for single type of unit
p-charting	Requires only go-no-go data, intuitive.	Requires many more inspections, less sensitive.
Demerit charting	Addresses diff b/w non-conforming	Requires more inspections (less than p), less sensitive.
u-charting	Relatively simple version of demerit charting.	Requires more inspections (less than p), less sensitive.
c-charting	Simple(u with n=1)	Requires more inspections (less than p), less sensitive.
np-charting	Simpler (eq. to p just w.o. ÷n)	Same as p plus can't have variable n.

I-MR (Individual, Moving Range) Chart:

→ For some cases, rate of production is low, it is not feasible for a sample size to be greater than 1.
→ When testing process is destructive and the cost of item is high, then sample size might be chosen to be 1.
→ If every unit is inspected, sample size is 1.
→ Data comes slowly, so samples with elements produced at long intervals creates problem for rational subgrouping
→ Control chart for individual measurements.
→ Useful for detection of system stability.

→ $R_i = |x_i - x_{i-1}|$ (moving range)

→ Trial limits –

$$UCL = \bar{x} + 3 \times \frac{\overline{MR}}{d_2} \quad (\text{for } n=2, d_2 = 1.128)$$

$$CL = \bar{x}$$

$$LCL = \bar{x} - 3 \times \frac{\overline{MR}}{d_2}$$

$$UCL_{\bar{R}} = D_4 \overline{MR}$$

$$CL_{\bar{R}} = \overline{MR}$$

$$LCL_{\bar{R}} = 0$$

→ Revised limits –

$\sigma_0 = 0.8865\, MR_0$, $x_0 = \bar{x}_{new}$

$UCL = \bar{x}_0 + 3\sigma_0$

$LCL = \bar{x}_0 - 3\sigma_0$

$UCL_R = 3.6866\,\sigma_0$

$LCL_R = 0$

→ Researchers have indicated that MR chart can't really provide additional info on variability, MR values are not independent.

Limitations of Shewhart Charts:

→ In shewhart chart, the plotted point represents info corresponding to observation only.
→ It doesn't use info from prev. observations.
→ This makes them insensitive to small shifts.
→ They are less useful in phase II.
→ Warning limits & patterns can be useful, but they reduce the simplicity of the control charts.

CUSUM Chart and EWMA Chart:

→ To detect small shifts, CUSUM (cumulative sum) and EWMA (Exponentially weighted moving average) charts are used as alternative to Shewhart control chart.
→ They are good alternatives for phase II monitoring.
→ Sometimes process needs to be monitored when sample size $n=1$, both CUSUM and EWMA works well in this situation.
→ EWMA charts are particularly robust against non-normality.

CUSUM Chart:

→ Plots the quantity $C_i = \sum_{j=1}^{i} (\bar{x}_j - \mu_0)$

where $\bar{x}_j$ is the avg. of the j^{th} sample.
μ_0 is the target for process mean.
Also applicable for $n = 1$.

→ So CUSUM charts are particular useful in chemical and process industries and discrete part manufacturing, where frequently subgroup size is 1.

→ There are 2 ways to represent CUSUM charts, tabular method and V-mask method.

→ $C_i^+ = \max\{0, x_i - (\mu_0 + K) + C_{i-1}^+\}$ (Upper CUSUM)

→ $C_i^- = \max\{0, (\mu_0 + K) - x_i + C_{i-1}^-\}$ (Lower CUSUM)

→ $C_0^+ = C_0^- = 0$

→ $C_i^+ > H$ or $C_i^- > H$ indicates the process mean has shifted.

→ K is called reference value, allowance or slack value, H is called decision interval.

→ $K = k\sigma$, typically halfway b/w μ_0 and out of control mean that we want to detect.

→ $K = \frac{|\mu_1 - \mu_0|}{2} = k\sigma$

→ $H = h\sigma$

→ K and H are choosen to provide good ARL performance.

→ Generally $k = 0.5$ and $h = 4$ or 5 are chosen.

EWMA chart:

→ Control chart to detect small shift in the process, ideally used with individual observations.

→ The exponentially weighted moving average is defined as

$$z_i = \lambda x_i + (1-\lambda) z_{i-1} \quad (z_0 = \mu_0 \text{ (target) or } \bar{\bar{x}})$$

→ λ should lie b/w 0.05, 0.25 (use smaller λ for smaller shifts)

→ Limits –

$$UCL = \bar{\bar{x}} + L\sigma\sqrt{\frac{\lambda}{2-\lambda}\left[1-(1-\lambda)^{2i}\right]}$$

$$CL = \bar{\bar{x}}$$

$$LCL = \bar{\bar{x}} - L\sigma\sqrt{\frac{\lambda}{2-\lambda}\left[1-(1-\lambda)^{2i}\right]}$$

Use μ_0 (target mean) in place of $\bar{\bar{x}}$ if given

→ Usually L is taken to be 3. For a steady state process $\sqrt{\frac{\lambda}{2-\lambda}\left[1-(1-\lambda)^{2i}\right]}$ becomes $\sqrt{\frac{\lambda}{2-\lambda}}$

→ σ can be estimated by process standard deviation or $\frac{\bar{R}}{d_2}$ if $\bar{R}$ can be obtained. Sometimes process history is used for estimation.

→ EWMA is often used with Shewhart Chart, so that the combined chart can detect small shifts and large shifts quickly enough.

→ EWMA chart is fairly robust against non-normal distribution, compared to shewhart charts for individual measurement.