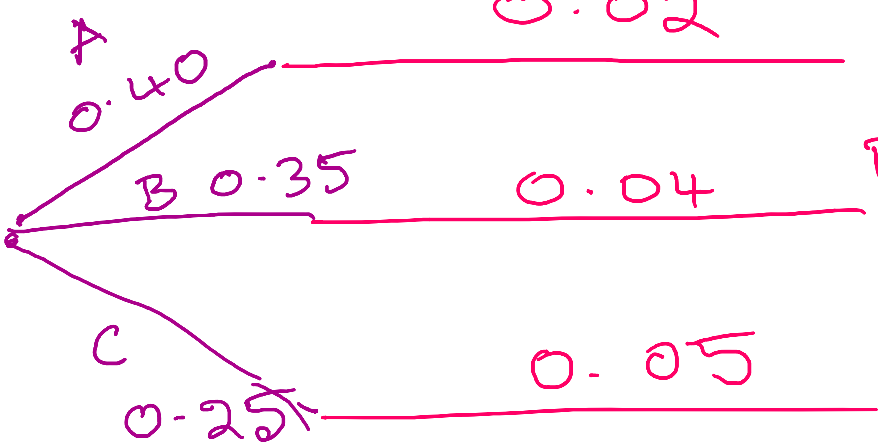
L- 5: Descriptive and inferential statistics

Agenda

- Quick Review of the topics covered in previous class
- Normal Distribution
- Sampling
- Testing of Hypothesis

Highest



$E/A = \dfrac{0.4 \times 0.02}{0.0345} = 0.232$

$E/B = \dfrac{0.35 \times 0.04}{''} = 0.406$

$E/C = \dfrac{0.25 \times 0.05}{''} = 0.362$

Total probability

$= P(E/A)\,P(E) + P(E|B)\,P(B) + P(E|C)\,P(C)$

$= (0.02)(0.4) + (0.04)(0.35) + (0.05)(0.25)$

$= 0.0345$

Example

Technicians regularly make repairs when breakdowns occur on an automated production line. Janak, who services 20% of the breakdowns, makes an incomplete repair 1 time in 20.Tarun ,who services 60% of the breakdowns ,makes an incomplete repair 1 time in 10 Gautham, who services 15% of the breakdowns, makes an incomplete repair 1 time in 10 and Prasad ,who services 5% of the breakdowns, makes an incomplete repair 1 time in 20.For the next problem with the production line diagnosed as being due to an initial repair that was incomplete, what is the probability that this initial repair was made by Janak?

Solution

Let A be the event that the initial repair was incomplete

B_1 that the repair was made by Janak

B_2 that it was made by Tarun ,

B_3 that it was made by Gautham,

B_4 that it was made by Prasad,

$$P(B_1/A) = \frac{P(B_1)P(A/B_1)}{P(B_1)P(A/B_1)+P(B_2)P(A/B_2)+P(B_3)P(A/B_3)+P(B_4)P(A/B_4)}$$

$$= \frac{(0.20)(0.05)}{(0.20)(0.05)+(0.60)(0.10)+(0.15)(0.10)+(0.05)(0.05)}$$

$$= 0.114$$

Problem

On the average, five cars arrive at a particular car wash every hour. Let X count the number of cars that arrive from 10AM to 11AM. (mean = 5).What is the probability that no car arrives during this period?

Poisson dist. $\lambda = 5$

$$P(x) = \frac{e^{-\lambda} \lambda^{x}}{x!}$$

$$P(x = 0) = \frac{e^{-5} 5^{0}}{0!} = e^{-5}$$

Problem

Suppose the car wash is in operation from 8AM to 6PM, and we let Y be the number of customers that appear in this period.(lambda = 50).

What is the probability that there are between 48 and 50 customers, inclusive?

$\lambda = 5$,

$P(48 \le x \le 50) = P(48) + P(49) + P(50)$

$= \dfrac{e^{-5}\, 5^{48}}{48!} + \dfrac{e^{-5} \cdot 5^{49}}{49!} + \dfrac{e^{-5} \cdot e^{50}}{50!}$

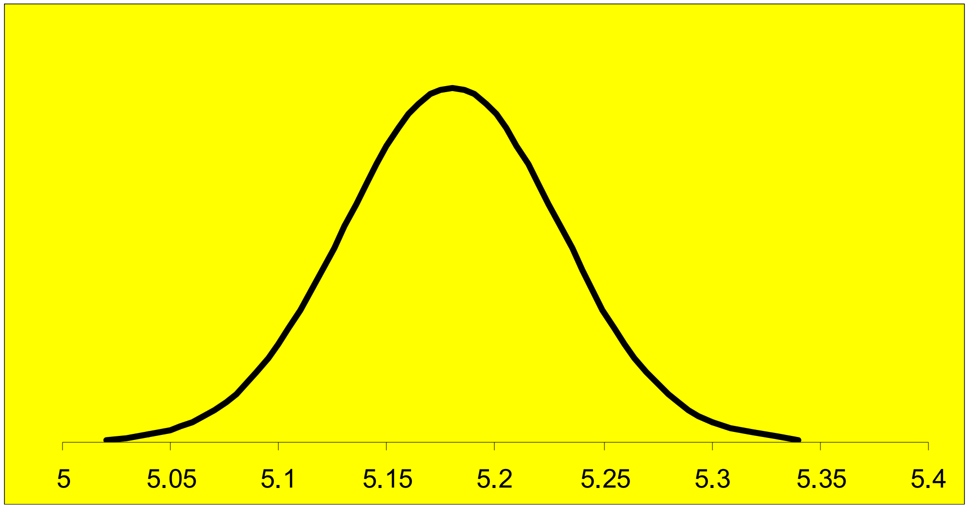
Normal distribution

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

$$, \quad -\infty < x < \infty$$

Normal Distribution

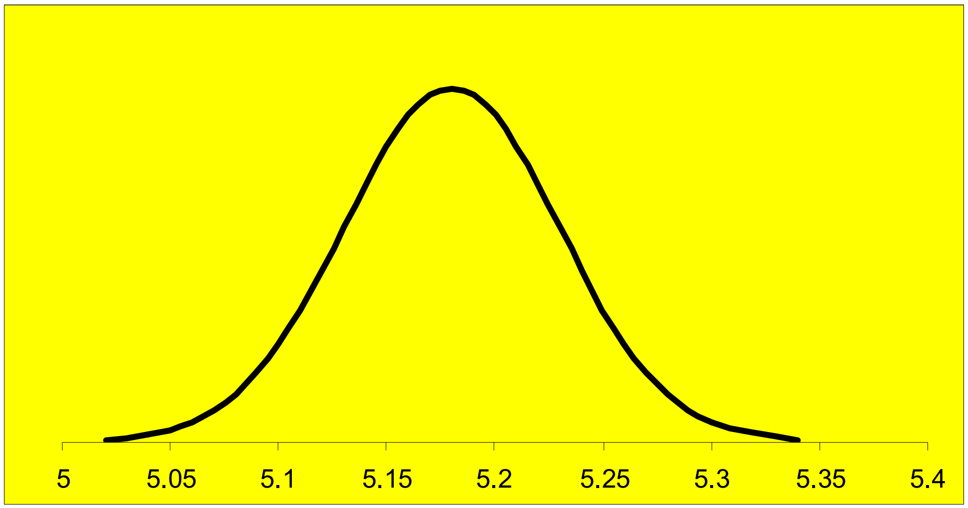
Probability density function - *f*(X)



$$f(X) = \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{-1/2(X-\mu)^2}{\sigma^2}}$$

Normal Distribution

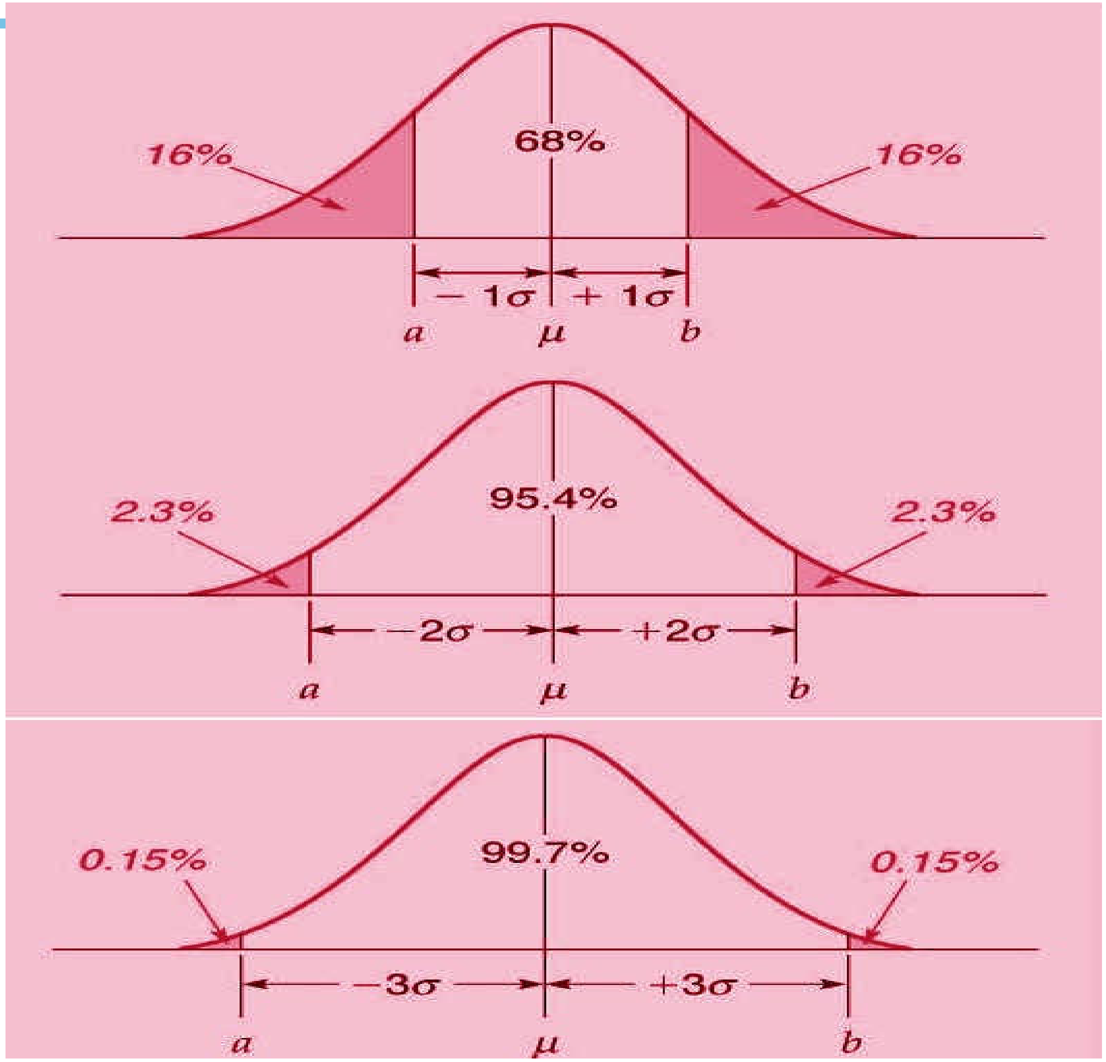
Probability density function - *f*(X)



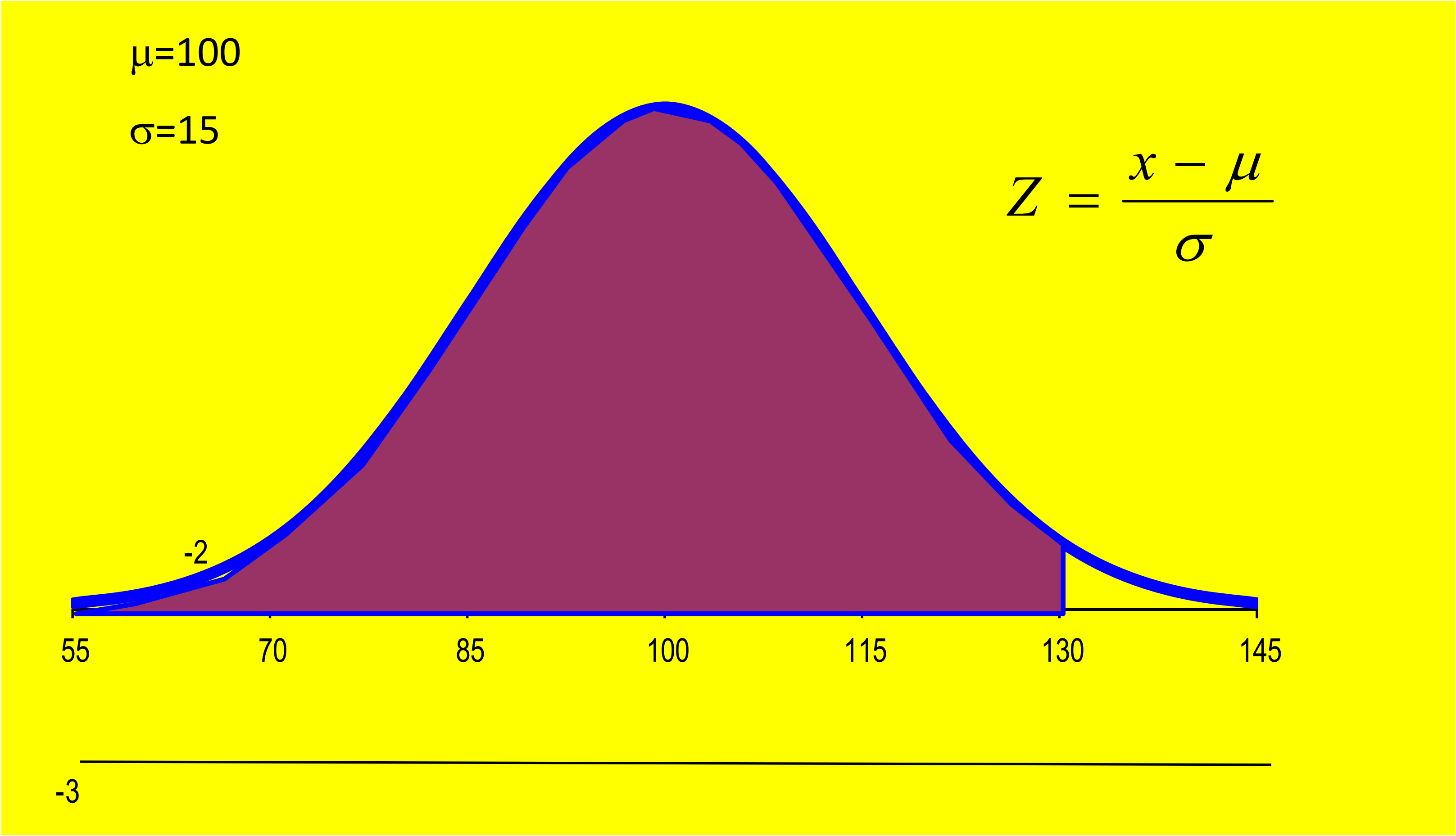
Normal curve
or
Gaussian
curve

Three Common Areas Under the Curve

Three Normal distributions with different areas



Standard Normal Distribution



How to find

$$P(2 < x < 5)$$

$$= \int_{x=2}^{5} f(x)\, dx$$

$$= \int_{x=2}^{5} \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} dx$$

Integration ????

Note

Since the normal density cannot be integrated in between every pair of limits a and b , probabilities relating to normal distributions are usually obtained from special tables (see tables)

$$P(x_1 \le x \le x_2) = \int_{x_1}^{x_2} f(x)\, dx$$

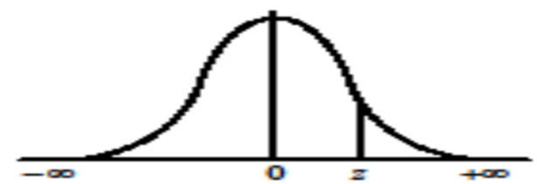
Let $\frac{x-\mu}{\sigma} = z$ ie $dx = \sigma\, dz$

$$= \int_{z_1}^{z_2} \frac{1}{\cancel{\sigma}\sqrt{2\pi}} e^{-z^2/2} \cdot \cancel{\sigma}\, dz$$

$$= \frac{1}{\sqrt{2\pi}} \int_{z_1}^{z_2} e^{-z^2/2}\, dz \longrightarrow F(z)$$

$$= F(z_2) - F(z_1)$$

NORMAL DISTRIBUTION TABLE



	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
.0	.5000	.5040	.5080	.5120	.5160	.5199	.5239	.5279	.5319	.5359
.1	.5398	.5438	.5478	.5517	.5557	.5596	.5636	.5675	.5714	.5753
.2	.5793	.5832	.5871	.5910	.5948	.5987	.6026	.6064	.6103	.6141
.3	.6179	.6217	.6255	.6293	.6331	.6368	.6406	.6443	.6480	.6517
.4	.6554	.6591	.6628	.6664	.6700	.6736	.6772	.6808	.6844	.6879
.5	.6915	.6950	.6985	.7019	.7054	.7088	.7123	.7157	.7190	.7224
.6	.7257	.7291	.7324	.7357	.7389	.7422	.7454	.7486	.7517	.7549
.7	.7580	.7611	.7642	.7673	.7704	.7734	.7764	.7794	.7823	.7852
.8	.7881	.7910	.7939	.7967	.7995	.8023	.8051	.8078	.8106	.8133
.9	.8159	.8186	.8212	.8238	.8264	.8289	.8315	.8340	.8365	.8389
1.0	.8413	.8438	.8461	.8485	.8508	.8531	.8554	.8577	.8599	.8621
1.1	.8643	.8665	.8686	.8708	.8729	.8749	.8770	.8790	.8810	.8830
1.2	.8849	.8869	.8888	.8907	.8925	.8944	.8962	.8980	.8997	.9015
1.3	.9032	.9049	.9066	.9082	.9099	.9115	.9131	.9147	.9162	.9177
1.4	.9192	.9207	.9222	.9236	.9251	.9265	.9279	.9292	.9306	.9319
1.5	.9332	.9345	.9357	.9370	.9382	.9394	.9406	.9418	.9429	.9441
1.6	.9452	.9463	.9474	.9484	.9495	.9505	.9515	.9525	.9535	.9545
1.7	.9554	.9564	.9573	.9582	.9591	.9599	.9608	.9616	.9625	.9633
1.8	.9641	.9649	.9656	.9664	.9671	.9678	.9686	.9693	.9699	.9706
1.9	.9713	.9719	.9726	.9732	.9738	.9744	.9750	.9756	.9761	.9767
2.0	.9772	.9778	.9783	.9788	.9793	.9798	.9803	.9808	.9812	.9817
2.1	.9821	.9826	.9830	.9834	.9838	.9842	.9846	.9850	.9854	.9857
2.2	.9861	.9864	.9868	.9871	.9875	.9878	.9881	.9884	.9887	.9890
2.3	.9893	.9896	.9898	.9901	.9904	.9906	.9909	.9911	.9913	.9916
2.4	.9918	.9920	.9922	.9925	.9927	.9929	.9931	.9932	.9934	.9936
2.5	.9938	.9940	.9941	.9943	.9945	.9946	.9948	.9949	.9951	.9952
2.6	.9953	.9955	.9956	.9957	.9959	.9960	.9961	.9962	.9963	.9964
2.7	.9965	.9966	.9967	.9968	.9969	.9970	.9971	.9972	.9973	.9974
2.8	.9974	.9975	.9976	.9977	.9977	.9978	.9979	.9979	.9980	.9981
2.9	.9981	.9982	.9982	.9983	.9984	.9984	.9985	.9985	.9986	.9986
3.0	.9987	.9987	.9987	.9988	.9988	.9989	.9989	.9989	.9990	.9990
3.1	.9990	.9991	.9991	.9991	.9992	.9992	.9992	.9992	.9993	.9993
3.2	.9993	.9993	.9994	.9994	.9994	.9994	.9994	.9995	.9995	.9995
3.3	.9995	.9995	.9995	.9996	.9996	.9996	.9996	.9996	.9996	.9997
3.4	.9997	.9997	.9997	.9997	.9997	.9997	.9997	.9997	.9997	.9998

Normal distribution will take on a value

1) to the left of $z = -1.78$

$z = -1.78$

2) to the right of $z = -1.45$

3) corresponding to $-0.80 \le z \le 1.53$

4) to the left of $z = -2,52$ and to the right of $z = 1.83$

Normal distribution will take on a value

1) to the left of $z = -1.78$

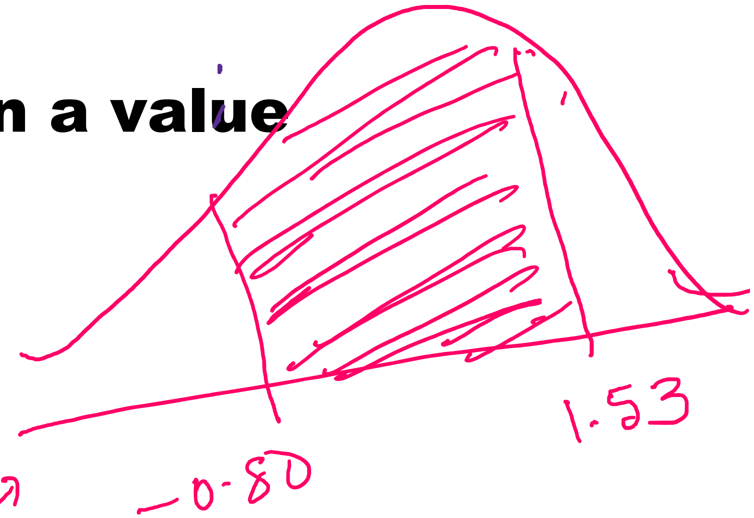
2) to the right of $z = -1.45$

$z = -1.45$

3) corresponding to $-0.80 \leq z \leq 1.53$

4) to the left of $z = -2,52$ and to the right of $z = 1.83$

Normal distribution will take on a value



1) to the left of z = -1.78

2) to the right of z = -1.45

3) corresponding to $-0.80 \le z \le 1.53$

ie $F(1.53) - F(-0.80)$

4) to the left of z = -2,52 and to the right of z = 1.83

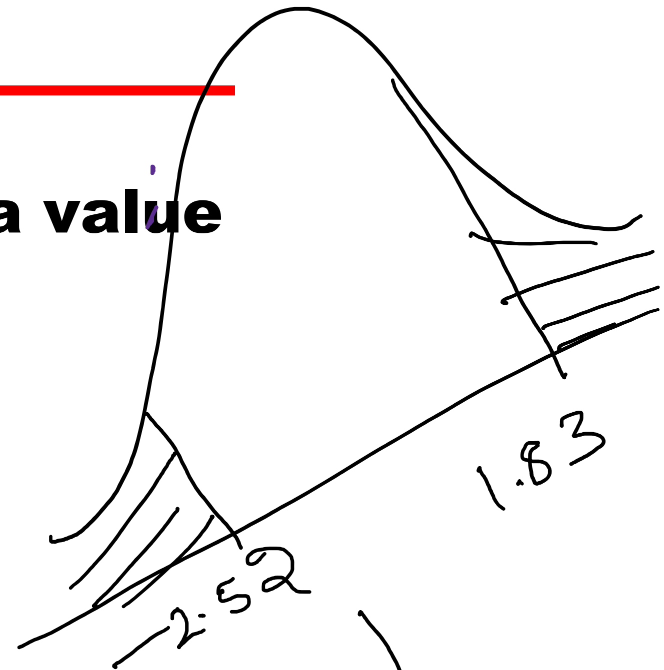
Normal distribution will take on a value

1) to the left of z = -1.78

2) to the right of z = -1.45

3) corresponding to -0.80 ≤ z ≤ 1.53

4) to the left of z = -2,52 and to the right of z = 1.83



$F(-1.83)$

$F(-2.52) + (1 - F(1.83))$

$= F(-2.52) +$

Calculation of probabilities using a normal distribution

Problem

The mean and standard deviation of a normal variate are 8 and 4 respectively

Find 1) $P[5 \leq X \leq 10]$

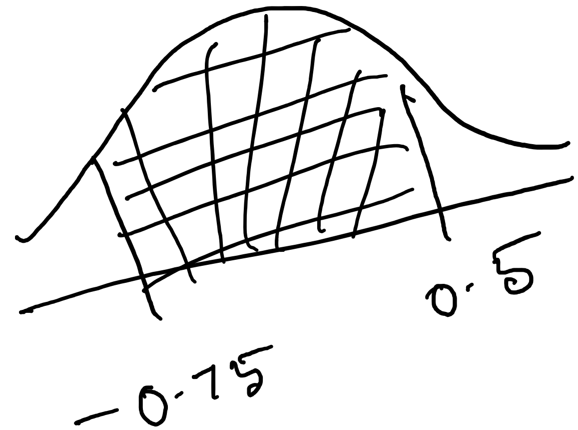
2) $P[X \geq 5]$

<u>Solution</u>

1) $\mu = 8$

$\sigma = 4$

We know that $Z = \dfrac{X-\mu}{\sigma} = \dfrac{X-8}{4}$



When X=5 $Z = \dfrac{5-8}{4} = -0.75$

$= F(0.5) - F(-0.75)$

When X=10 $Z = \dfrac{10-8}{4} = 0.5$

$=$

$P[\,5 \le X \le 10\,] = P[\,-0.75 \le Z \le 0.5\,]$

$$= F(0.5) - F(-0.75)$$

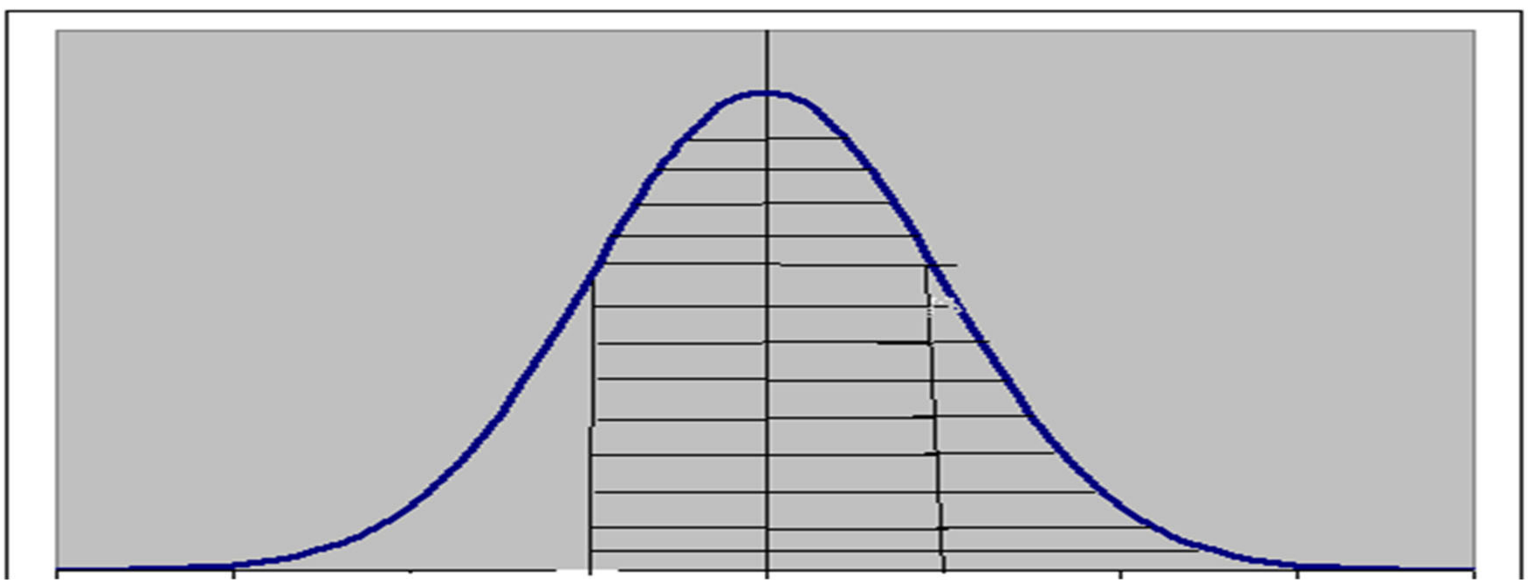
$$= 0.6915 - .22663 = 0.4649$$

2)P [X≥ 5] = P [Z ≥ -0.75] = 1- F (-0.75)

= F(0.75)

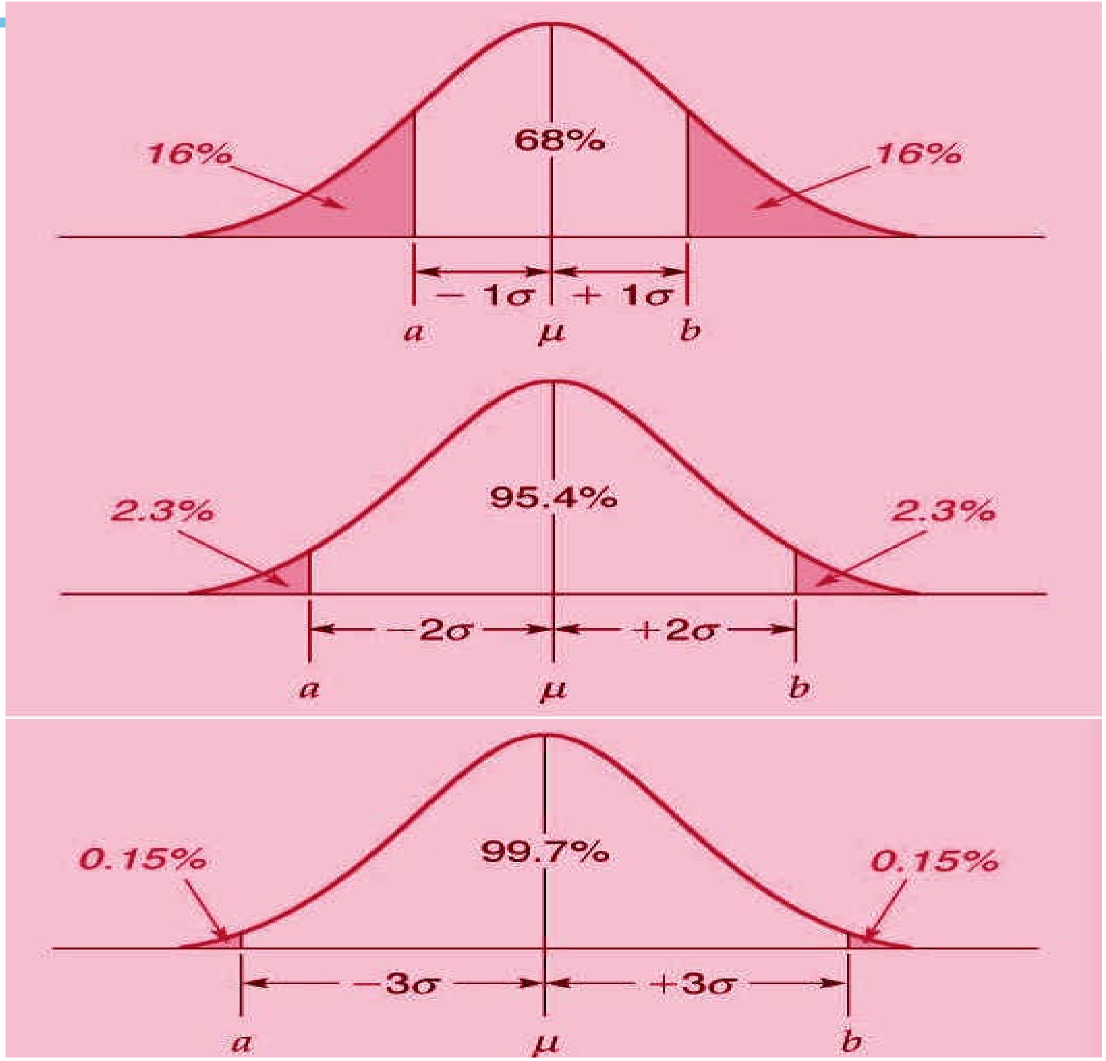
~~0.7734~~

= 0.7734



Three Common Areas Under the Curve

Three Normal distributions with different areas



Example:

In a test conducted on 1000 candidates, the average score is 42 with a S.D of 24. Assuming normal distribution, find

a) no of candidates whose score exceeds 58

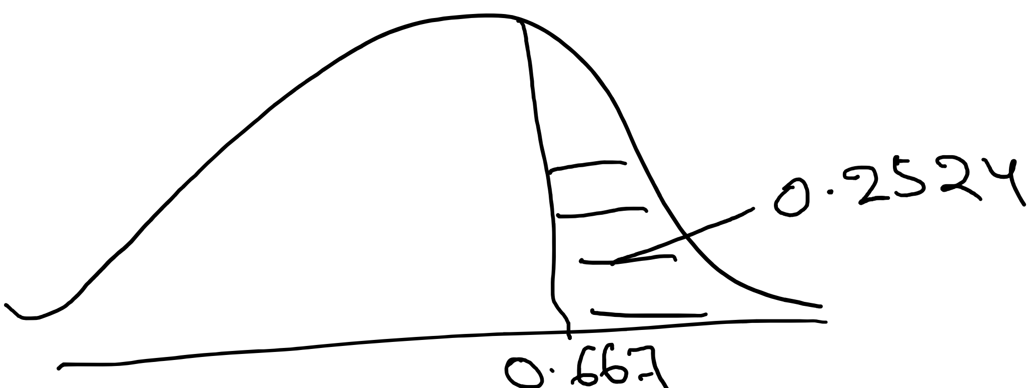
b) no of candidates whose scores lies b/w 30 and 66

a) Score exceeds 58 i.e. $x > 58$

$P(x > 58)$

$$z = \frac{x - \mu}{\sigma} = \frac{58 - 42}{24} = 0.667$$

$P(z > 0.667)$



1000×0.2524

$\downarrow$

252 candidates

b) $P(30 \le x \le 66)$

$$\frac{30-42}{24} = -0.5 \qquad \frac{66-42}{24} = 1$$

$P(-0.5 \le z \le 1)$

$= F(1) - F(-0.5)$

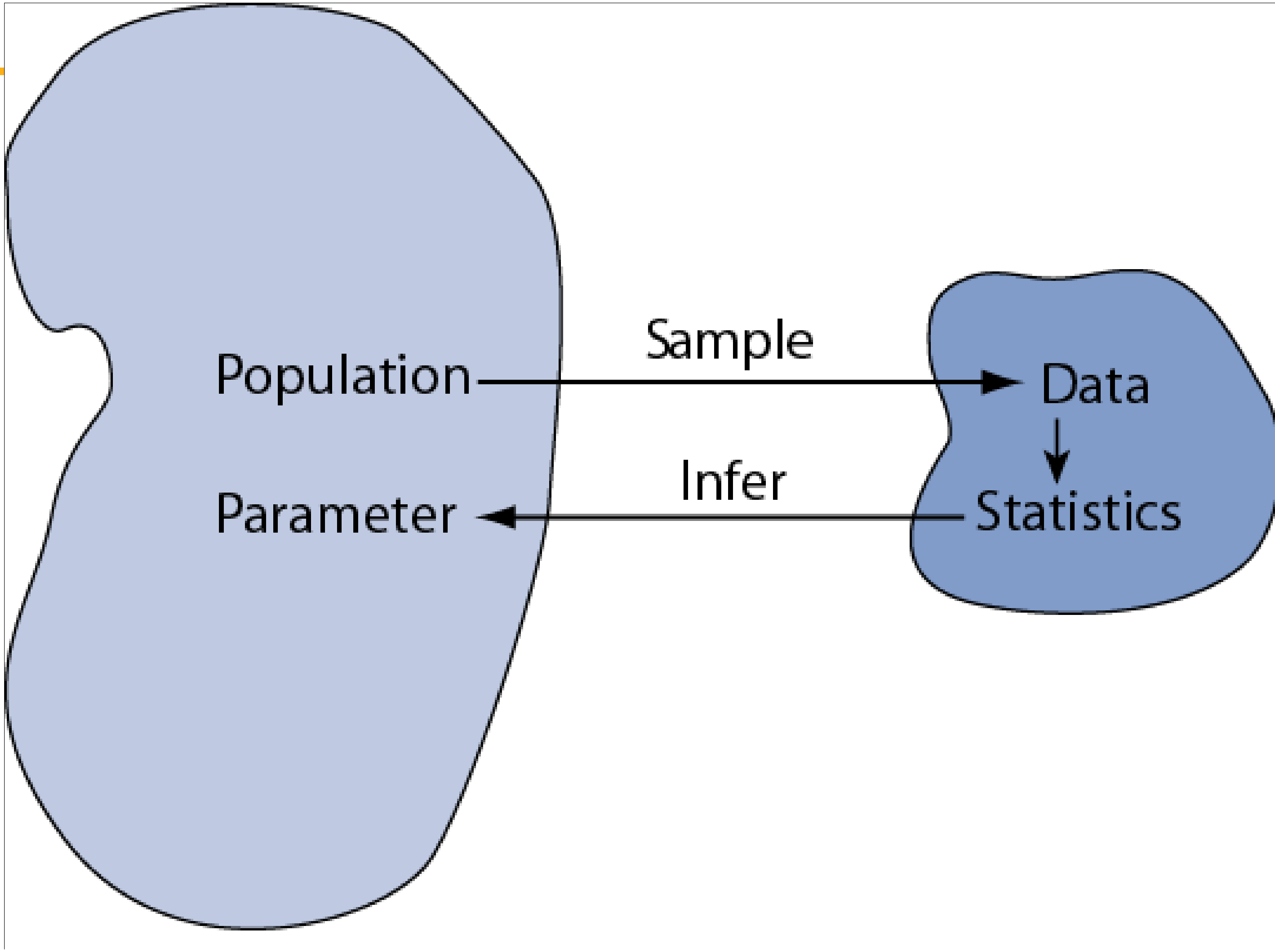
$= 0.5328$

$0.5328 \times 1000 = 532.8$

533

Inferential Statistics

- Sampling
- Sample
- Random sampling
- Central Limit theorem


Sample
Population
Data
Infer
Parameter
Statistics

Statistical Inferences

Theory of statistical inference is divided into two major areas

- Estimation

- Tests of hypothesis

Hypothesis Testing

Goal:

Make statement(s) regarding unknown population parameter values based on sample data

Hypothesis Testing

- Is also called *significance testing*
- Tests a claim about a parameter using evidence (data in a sample

Example

Drug company has new drug, wishes to compare it with current standard treatment

Federal regulators tell company that they must demonstrate that new drug is better than current treatment to receive approval

Firm runs clinical trial where some patients receive new drug, and others receive standard treatment

Numeric response of therapeutic effect is obtained (higher scores are better).

Parameter of interest: $m_{\text{New}} - m_{\text{Std}}$

Hypothesis Testing Steps

- Null and alternative hypotheses
- Test statistic
- P-value and interpretation
- Significance level (optional)

Example

Null hypothesis $H_{0:}\ \mu = 170$

The **alternative hypothesis** can be either $H_{1:}\ \mu > 170$ (**one-sided test**)

or

$H_{1:}\ \mu \neq 170$ (**two-sided test**)

Test Statistic

Use this statistic to test the problem:

$$z_{\text{stat}} = \frac{\bar{x} - \mu_0}{SE_{\bar{x}}}$$

where $\mu_0 \equiv$ population mean assuming H_0 is true

and $SE_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$

Example

A. Hypotheses:

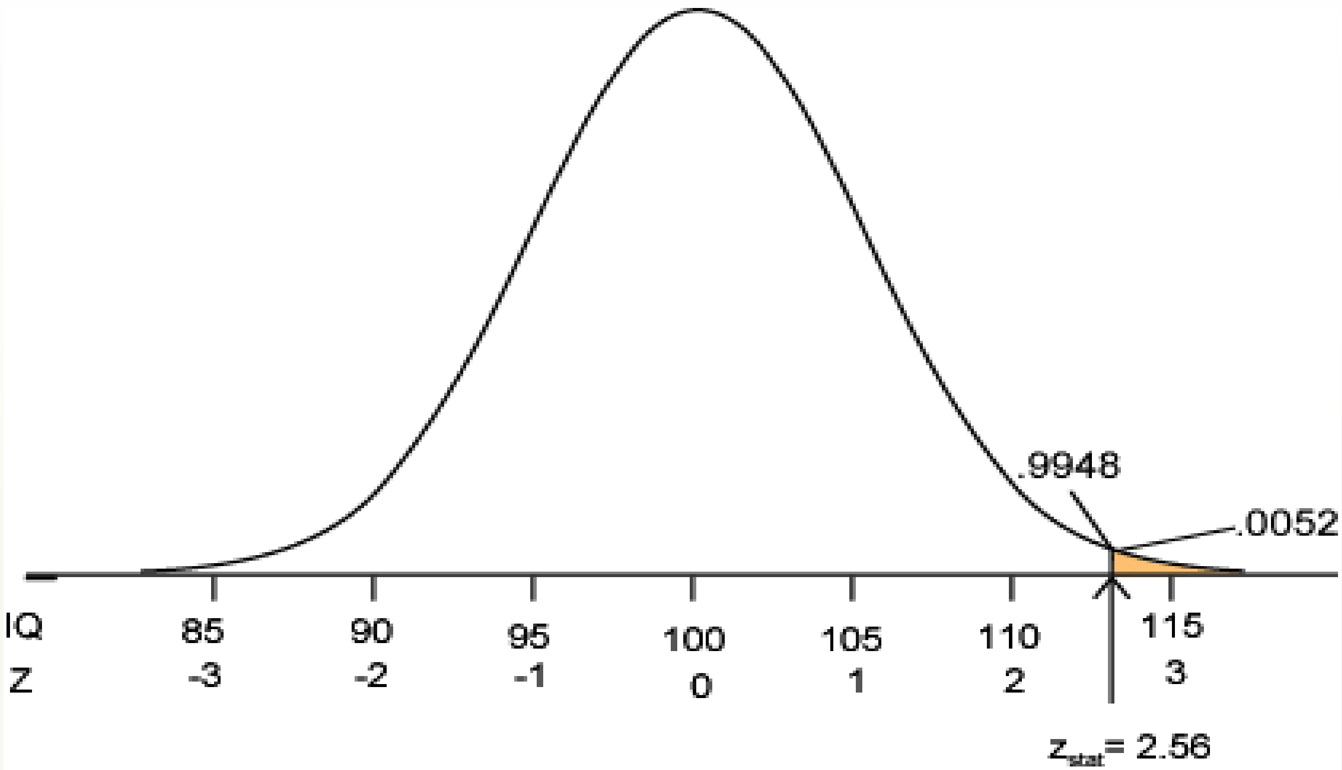
H_0: μ = 100 versus
H_a: μ > 100 (one-sided)
H_a: μ ≠ 100 (two-sided)

B. Test statistic:

$$SE_{\bar{x}} = \frac{\sigma}{\sqrt{n}} = \frac{15}{\sqrt{9}} = 5$$

$$z_{\text{stat}} = \frac{\bar{x} - \mu_0}{SE_{\bar{x}}} = \frac{112.8 - 100}{5} = 2.56$$

C. *P*-value: $P = \Pr(Z \geq 2.56) = 0.0052$



$P = .0052 \Rightarrow$ it is unlikely the sample came from this null distribution $\Rightarrow$ strong evidence against H_0

Hypothesis Testing

Test Result – True State	H_0 True	H_0 False
H_0 True	Correct Decision	Type I Error
H_0 False	Type II Error	Correct Decision

$$\alpha = P(Type\ I\ Error) \quad \beta = P(Type\ II\ Error)$$

- Goal: Keep α, β reasonably small

Problem

It is claimed that a random sample 49 tyres has a mean life of 15200 kms. This sample was drawn from a population whose mean is 15150 kms and a standard deviation of 1200kms. Test the significance at 0.05 level.

Solution:

1. Null hypothesis $H_0 : \mu = 15200$

2. Alternate hypothesis $H_1 : \mu \neq 15200$

3. Level of significance $\alpha = 0.05$

4. critical region :- This is a two tailed test (large sample). So reject H_0 if $(Z_{cal} = Z) < - Z_{\frac{\alpha}{2}}$ or $(Z = Z_{cal}) > Z_{\frac{\alpha}{2}}$

Here $\alpha = 0.05$

$$\frac{\alpha}{2} = \frac{0.05}{2}$$

$$= 0.025$$

From table we get

$\therefore \; Z_{\frac{\alpha}{2}} = 1.96$

i.e; if

$Z_{cal} = Z < -1.96$ or $Z_{cal} > 1.96$ we reject null hypothesis.

6. Computation :

Test statistic

$$Z_{cal} = Z = \frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}}} = \frac{15200 - 15150}{\frac{1200}{\sqrt{49}}}$$

$$= 0.2916$$

7. Decesion:

Since $Z_{cal} = 0.2916 < 1.96$ we accept the mull hypothesis.

Problem

A trucking firm is suspicious of the claim that the average life time of certain tyres is at least 28,000 miles. To check the claim, the firm puts 40 of these tyres on its trucks and get a mean life of 27,463miles with a standard deviation of 1,348 miles. What can it conclude if the probability of Type I error is to be at most 0.01

Solution

1.Null hypothesis : H_0 : $\mu \geq 28{,}000$ miles

2. Alternate hypothesis: H_1: $\mu < 28{,}000$ miles

3. Level of significance: $\alpha = 0.01$

4. Critical region

This is a left tailed test (large sample)

If $Z = Z_{cal} < - Z_{\alpha}$ we reject null hypothesis

If $Z = Z_{cal} < - Z_{\alpha} = - Z_{0.01} = -2.33$ we reject null hypothesis

5.Computation

Test statistic

$$Z = \frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}}} = \frac{27,463 - 28,000}{\frac{1,348}{\sqrt{40}}} = -2.52$$

6.Conclusion

Since $Z = Z_{cal} = -2.52 < -2.33$, we reject null hypothesis at level of significance 0.01. In other words the trucking firm's suspicion that $\mu < 28,000$ miles is confirmed.

Hypothesis concerning one mean (small sample)

Procedure

1. Null hypothesis $H_0 : \mu = \mu_0$

2. Alternate Hypothesis $H_1 : \mu \neq \mu_0$ (Two tailed test)

Or

$H_1 : \mu > \mu_0$ (Right tailed test)

Or

$H_1 : \mu < \mu_0$ (left tailed test)

3. Level of significance : α

4. Critical region

For two tailed test $H_1 : \mu \neq \mu_0$

Reject H_0 if $t < -t_{\frac{\alpha}{2}}$ or

$t > t_{\frac{\alpha}{2}}$ with (n-1) degrees of freedom

For right tailed test $H_1 : \mu > \mu_0$

Reject H_0 if $t > t_{\alpha}$ with (n-1) degrees of freedom

For left tailed test $H_1 : \mu < \mu_0$

Reject H_0 if $t < -t_{\alpha}$ (n-1) degrees of freedom

5. Test statistic

$$t = \frac{\bar{x} - \mu}{\frac{s}{\sqrt{n}}} \text{ with (n-1) degrees of freedom}$$

6. Calculation

7. Decision

A random sample of 6 steel beams has a mean compressive strength of 58,392 p.s.i (pounds per square inch) with a standard deviation of 648 p.s.i . use this information at the level of significance $\alpha = 0.05$ to test

whether the true average compressive strength of steel from which the sample came is 58,000 p.s.i

Thanks