statistics

FOR

DATA SCIENCE

UNIT-5

feedback/corrections: vibha@pesu.pes.edu

Vibha Masti

POWER OF A TEST

Power of test: probability of rejecting H_0 when it is false

Real / Research	H_0 is true	H_0 is false	
Reject H_0	Type I α	Correct $1-\beta$	← power
Fail to reject H_0	Correct $1-\alpha$	Type II β	← P(type II)

- when alternate (H_1) is true; chance of getting significant results
- find minimum sample size or magnitude of power
- how likely to correctly (predict H_1) if H_0 is false

Computing Power

1. Compute rejection region
   - null distribution
   - critical value
   - rejection region

2. Alternate hypothesis true
   - alternate distribution
   - Z-score under H_1 for critical point
   - P(reject H_0 | H_1 true)

Assume that a new chemical process has been developed that may increase the yield over that of the current process. The current process is known to have a mean yield of 80 and a standard deviation of 5, where the units are the percentage of a theoretical maximum. If the mean yield of the new process is shown to be greater than 80, the new process will be put into production.

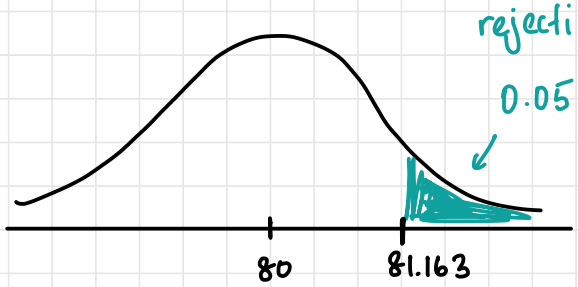
Q1. Let μ denote the mean yield of the new process. It is proposed to run the new process 50 times and then to test the hypothesis $H_0: \mu \leq 80$ versus $H_1: \mu > 80$ at a significance level of 5%.

$N = 50 \quad \alpha = 0.05 \quad \sigma = 5$

power of test? $\mu = 81$

$H_0: \mu \leq 80 \quad H_1: \mu > 80$

1. Compute rejection region



$Z_0 = 1.645$

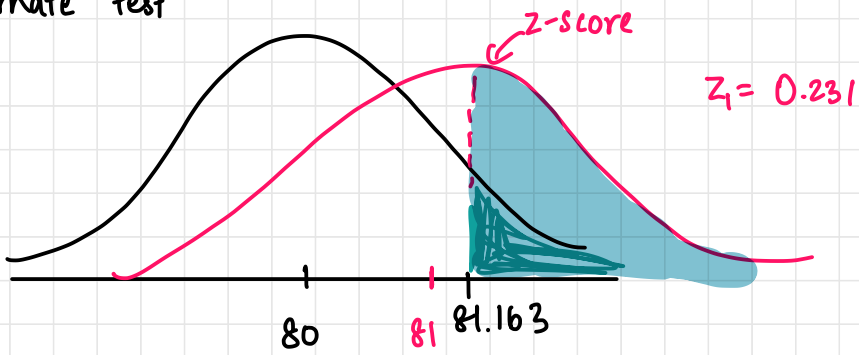
$\bar{X} \sim N\left(80, \left(\frac{5}{\sqrt{50}}\right)^2\right) \quad \frac{1}{\sqrt{2}}$

$\bar{X} \sim N(80, 0.707^2)$

p-value = 0.95

$x = 81.163$

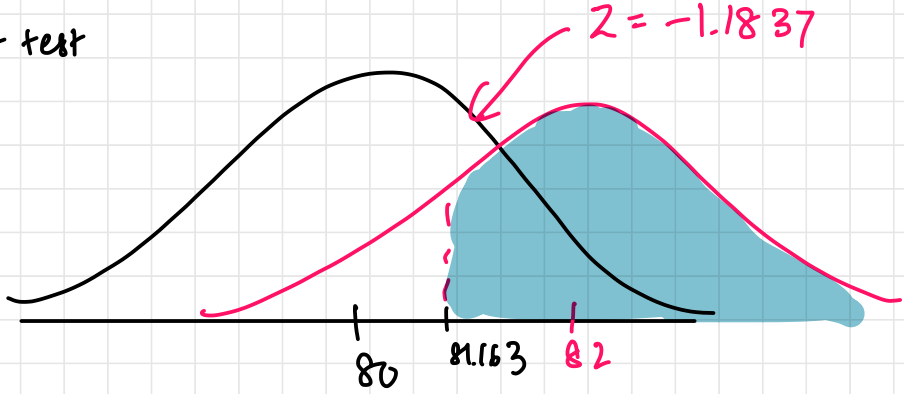
1. Alternate test



p-value = 0.409
= 40.9% → not great

Q2. Find the power of the 5% level test of
H_0 : μ ≤ 80 versus H_1 : μ > 80
for the mean yield of the new process under the alternative μ = 82, assuming n = 50 and σ = 5.

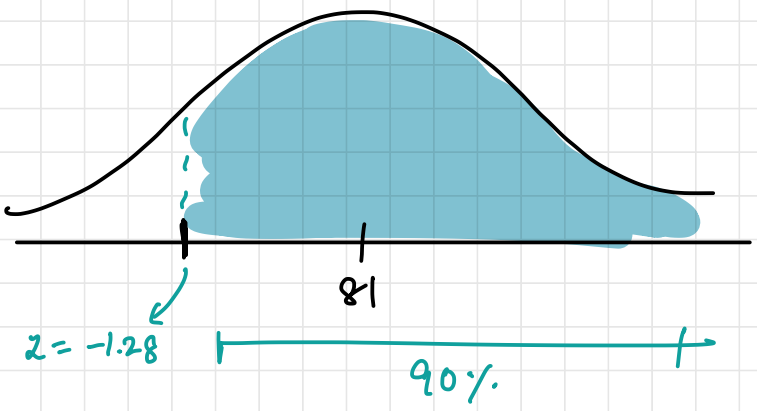
2. Alt test



p-value ≈ 0.88
= 88% power ✓

Q3. In testing the hypothesis H_0 : μ ≤ 80 versus H_1 : μ > 80 regarding the mean yield of the new process, how many times must the new process be run so that a test conducted at a significance level of 5% will have power 0.90 against the alternative μ = 81, if it is assumed that σ = 5?

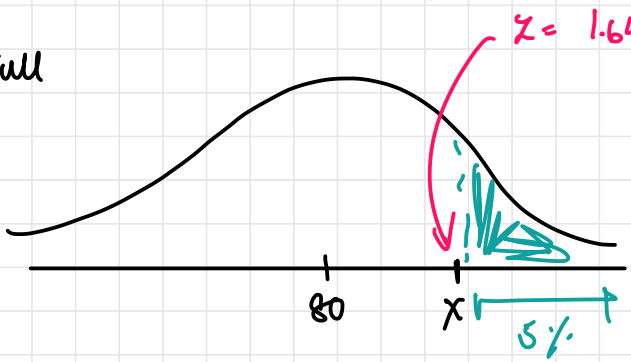
alternate :



Based on alternate

$$X = 81 + \frac{(-1.28) \times 5}{\sqrt{n}} \quad \text{—— (1)}$$

Null



$$x = 80 + \frac{(1.645)\times 5}{\sqrt{n}} \longrightarrow (2)$$

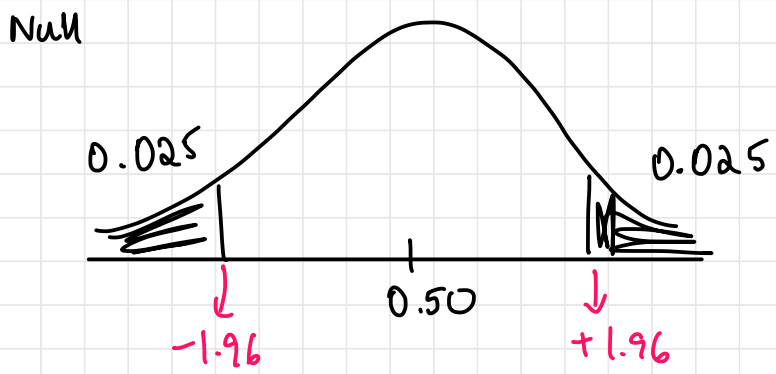
$$81 + \frac{(-1.28)\times 5}{\sqrt{n}} = 80 + \frac{(1.645)\times 5}{\sqrt{n}}$$

$$1 = \frac{5}{\sqrt{n}}(2.925)$$

$$n = (5 \times 2.925)^2 \approx 214$$

critical point $x = 80.56$

Q4. A pollster will conduct a survey of a random sample of voters in a community to estimate the proportion who support a measure on school bonds. Let p be the proportion of the population who support the measure. The pollster will test H_0 **: p = 0.50 versus** H_1 **: p ≠ 0.50** at the **5% level**. If 200 voters are sampled, **what is the power of the test if the true value of p is 0.55?**



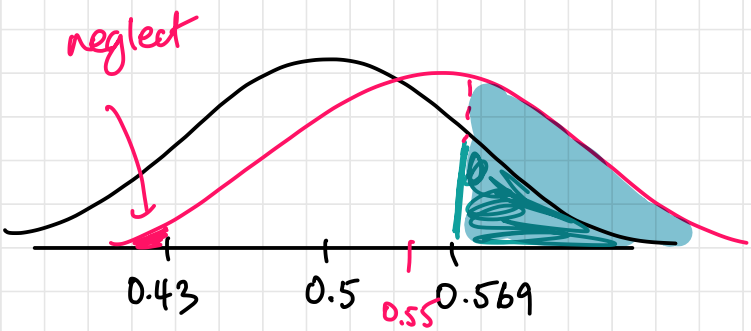
$z = -1.96$ $z = +1.96$

$$\sigma_{\hat{p}} = \sqrt{\frac{pq}{n}}$$

Rejection region

$$\hat{p} = 0.50 \pm 1.96 \times \sqrt{\frac{0.5 \times 0.5}{200}}$$

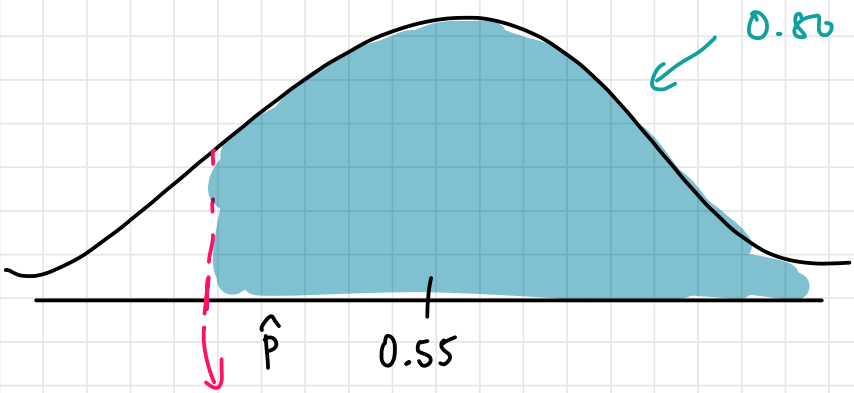
$= 0.569$ and 0.431



$p = 0.295 = 29.5\%$

Q5. A pollster will conduct a survey of a random sample of voters in a community to estimate the proportion who support a measure on school bonds.
Let p be the proportion of the population who support the measure. The pollster will test:
H_0 : **p = 0.50 versus** H_1 : **p ≠ 0.50**
at the **5% level**. **How many voters must be sampled so that the power will be 0.8 when the true value of p = 0.55?**

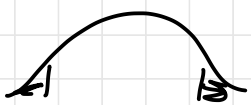
Alt.



$z = -0.842$

$$\hat{p} = 0.55 + \frac{(-0.842) \times \sqrt{0.5 \times 0.5}}{\sqrt{n}}$$

Null.

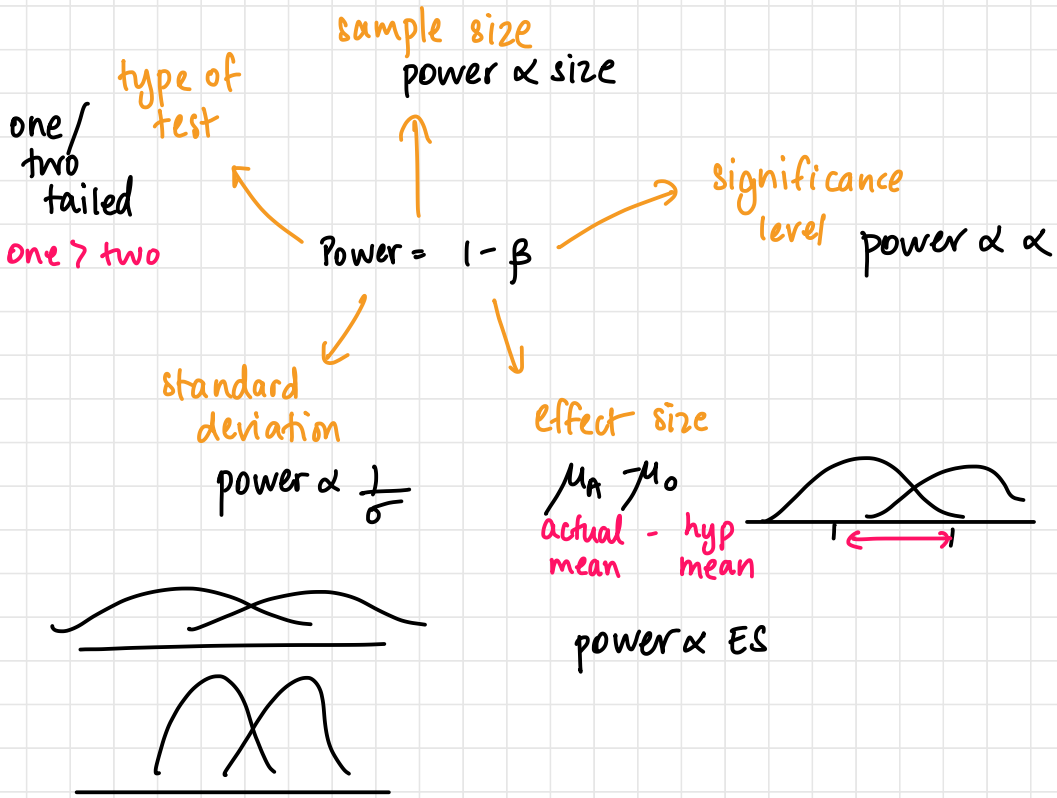


$$\hat{p} = 0.5 + \frac{(1.96 \times \sqrt{0.5 \times 0.5})}{\sqrt{n}}$$

$$0.05 = \frac{0.5}{\sqrt{n}} \times 2.802$$

$$\sqrt{n} = 28.02$$
$$n = 785$$

Factors Affecting Power of a Test



Independent Variable		Dependent Variable	
Input variable	Predictor variable	Output/ Response variable	Predicted variable
Controlled variable	Explanatory variable	Measured variable	Explained variable
Regressor	Manipulated variable	Regresand	Experimental variable

CORRELATION

strength of relation between 2 linear variables

Pearson Correlation Coefficient

$\bar{x}$ = mean of x's
$\bar{y}$ = mean of y's

s_x = std of x's
s_y = std of y's

$$r = \frac{1}{n-1} \sum_{i=1}^{n} \left(\frac{x_i - \bar{x}}{s_x} \right) \left(\frac{y_i - \bar{y}}{s_y} \right)$$

$$r = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n} (x_i - \bar{x})^2} \sqrt{\sum_{i=1}^{n} (y_i - \bar{y})^2}}$$

Q6- The Psychological tests of intelligence and of engineering ability were applied to 10 students. Here is a record of ungrouped data showing intelligence ratio (I.R) and engineering ratio (E.R) calculate the co-efficient of correlation?

Student	A	B	C	D	E	F	G	H	I	J
I.R	105	104	102	101	100	99	98	96	93	92
E.R	101	103	100	98	95	96	104	92	97	94

let $x = IR$, $y = ER$

$$r = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n} (x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n} (y_i - \bar{y})^2}}$$

$n = 10$

$\bar{x} = 99$

$\bar{y} = 98$

	170	36	25	9	4	1	0	1	9	36	49
$x - \bar{x}$		6	5	3	2	1	0	-1	-3	-6	-7
$y - \bar{y}$		3	5	2	0	-3	-2	6	-6	-1	-4
	140	9	25	4	0	9	4	36	36	1	16
Total	92	18	25	6	0	-3	0	-6	18	6	28

$$r = \frac{92}{\sqrt{170}\sqrt{140}} = 0.596$$

Pearson's coefficient

Ascombe's Quartet
- $\bar{x}, s_x, \bar{y}, s_y$
- r
- $\hat{\beta}_0, \hat{\beta}_1$

Confounding Variable

- var that influences ind & dep vars

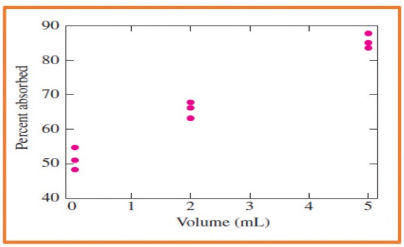
Q7.

- An environmental scientist is studying the rate of absorption of a certain chemical into skin.
- She places differing volumes of the chemical on different pieces of skin and allows the skin to remain in contact with the chemical for varying lengths of time.
- She then measures the volume of chemical absorbed into each piece of skin.
- She obtains the results shown in the following table.

Volume (ml)	Time (h)	Percent Absorbed
0.05	2	48.3
0.05	2	51.0
0.05	2	54.7
2.00	10	63.2
2.00	10	67.8
2.00	10	66.2
5.00	24	83.6
5.00	24	85.1
5.00	24	87.8

Correlation between Volume & Percent Absorbed

❖ Scatter Plot :



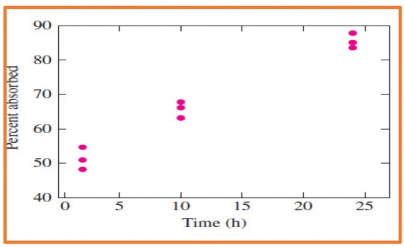
❖ Correlation , $r = 0.988$

❖ Positive Correlation

❖ Increasing the volume causes the percentage absorbed to increase.

Correlation between Time & Percent Absorbed

❖ Scatter Plot :



❖ Correlation , $r = 0.987$

❖ Positive Correlation

❖ Increasing the time that the skin is in contact with chemical causes the percentage absorbed to increase.

- No ∵ time & volume correlated
- cannot determine if time, vol or

Population Correlation

- random bivariate distribution (X & Y)
- r : sample correlation
- ρ : population correlation

Fischer Transformation:

$$w = \frac{1}{2} \ln\left(\frac{1+r}{1-r}\right) \qquad w \sim N(\mu_w, \sigma_w^2)$$

$$\mu_w = \frac{1}{2} \ln\left(\frac{1+\rho}{1-\rho}\right) \qquad \sigma_w^2 = \frac{1}{n-3}$$

$$e^{2\mu_w} = \frac{1+\rho}{1-\rho}$$

$$\frac{e^{2\mu_w}-1}{e^{2\mu_w}+1} = \frac{1+\rho-1+\rho}{1+\rho+1-\rho} = \rho$$

Q8.

- In a study of reaction times, the time to respond to a visual stimulus (x) and the time to respond to an auditory stimulus (y) were recorded for each of 10 subjects.
- Times were measured in ms.
- The results are presented in the following table.

x	161	203	235	176	201	188	228	211	191	178
y	159	206	241	163	197	193	209	189	169	201

- Find a 95% confidence interval for the correlation between the two reaction times.

$\bar{x} = 197.2$
$\bar{y} = 192.7$

$n = 10$
$r = 0.8159$

$$r = \frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i-\bar{y})^2}}$$

$$w = \frac{1}{2}\ln\left(\frac{1+r}{1-r}\right)$$

$$\mu_w = \frac{1}{2}\ln\left(\frac{1+\rho}{1-\rho}\right)$$

$$= \frac{1}{2}\ln(9.87) = 1.1444$$

$$\sigma_w = \sqrt{\frac{1}{n-3}} = \sqrt{\frac{1}{7}} = 0.378$$

find 95% CI for μ_w

$$w - 1.96\sigma_w \leq \mu_w \leq w + 1.96\sigma_w$$

$$0.4035 \leq \mu_w \leq 1.8853$$

$$\rho = \frac{e^{2\mu_w} - 1}{e^{2\mu_w} + 1} = 0.955$$

$$\rho = 0.383$$

95% CI: $0.383 \leq \rho \leq 0.955$

Q9.

- In a study of reaction times, the time to respond to a visual stimulus (x) and the time to respond to an auditory stimulus (y) were recorded for each of 10 subjects.
- Times were measured in ms.
- The results are presented in the following table.

x	161	203	235	176	201	188	228	211	191	178
y	159	206	241	163	197	193	209	189	169	201

i. Find the P – value for testing H_0: $\rho \leq 0.3$ versus H_1: $\rho > 0.3$

ii. Test the hypothesis H_0: $\rho \leq 0$ versus H_1: $\rho > 0$

$\bar{x} = 197.2$ $\quad n = 10$ $\quad w = 1.1444$

$\bar{y} = 192.7$ $\quad r = 0.8159$

(i) $\rho = 0.3 \Rightarrow \mu_w = \frac{1}{2} \ln\left(\frac{1+\rho}{1-\rho}\right) = 0.3095$

$$\sigma_w = \sqrt{\frac{1}{7}} = 0.378$$

$$W \sim N(0.3095, 0.378^2)$$

$$z = \frac{1.1444 - 0.3095}{0.378} = 2.21$$

P-value $= 0.0136$ $\quad \therefore \rho > 0.3$

(ii) $\rho = 0.0 \Rightarrow$ $\boxed{u = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}}}$

$r = 0.8159$

$u = 3.991$

p-value in between 0.001 & 0.005

$\therefore \rho > 0$

Regression Analysis

- relationships b/w dep & ind (estimation)

Least Squares Line

$$y_i = \hat{\beta}_1 x_i + \hat{\beta}_0$$

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n} (x_i - \bar{x})^2}$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

Q10. ❖The details pertaining to the no. of hours spent by students in preparing for an entrance exam and the marks scored (on a scale of (0 – 100) is provided in the following table.

Using these values,

i. Estimate the marks scored by a student who has spent 2.35 hours.

ii. Predict the marks that a student can score if he/she invests 20 hours.

SL No.	No. of hours spent	Marks Scored
1	6	82
2	10	88
3	2	56
4	4	64
5	6	77
6	7	92
7	0	23
8	1	41
9	8	80
10	5	59
11	3	47

x = hours y = marks

TODO

x	y	$x-\bar{x}$ X	$y-\bar{y}$ Y	$(x-\bar{x})^2$ X^2	XY
6	82	1.28	17.54	1.638	
10	88	5.28	23.55	27.878	
2	56	-2.72	-8.45	7.398	
4	64	-0.72	-0.45	0.518	
6	77	1.28	12.55	1.638	
7	92	-2.72	27.55	7.398	
0	23	-4.72	-41.45	22.278	
1	41	-3.72	-23.45	13.838	
8	80	3.28	15.55	10.758	
5	59	0.28	-5.45	0.078	
3	47	-1.72	-17.45	2.958	
					611.37

$\bar{x} = 4.72$

$\bar{y} = 64.45$

$$\hat{\beta}_1 = \frac{\sum XY}{\sum X^2} = 6.49$$

$$\hat{\beta}_0 = 33.77$$

$$\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y}) = \sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y}$$

$$\sum_{i=1}^{n} (x_i - x)^2 = \sum_{i=1}^{n} x_i^2 - n\bar{x}^2$$

$$\hat{\beta}_1 = r \frac{s_y}{s_x}$$

Goodness of Fit

$$\sum_{i=1}^{n} (y_i - \bar{y})^2 - \sum_{i=1}^{n} (y_i - \hat{y})^2 = \text{regression sos}$$

r^2

total sum of squares

error sum of squares

$$r^2 = \frac{\sum_{i=1}^{n} (y_i - \bar{y})^2 - \sum_{i=1}^{n} (y_i - \hat{y})^2}{\sum_{i=1}^{n} (y - \bar{y})^2}$$

coefficient of determination

Error Variance

residual

$$s^2 = \frac{\sum_{i=1}^{n} e_i^2}{n-2} = \frac{\sum (y_i - \hat{y})^2}{n-2}$$

$$s^2 = \frac{(1-r^2) \sum (y_i - \bar{y})^2}{n-2}$$

$$s_{\hat{\beta}_0} = s \sqrt{\left[\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2} \right]} \qquad s_{\hat{\beta}_1} = s \sqrt{\left[\frac{1}{\sum_{i=1}^{n} (x_i - \bar{x})^2} \right]}$$

uncertainties

Q 11. Problem: A chemical reaction is ran 12 times. The temperature and yield is recorded each time.

$\bar{x} = 65 \quad \bar{y} = 29.05 \quad \sum_{i-1}^{n}(x_i - \bar{x})^2 = 6032$

$\sum_{i-1}^{n}(y_i - \bar{y})^2 = 835.42$

$\sum_{i-1}^{n}(x_i - \bar{x})(y_i - \bar{y}) = 1988.4$ Compute the least squares estimates, error variance estimate.

Least squares estimate: $\hat{\beta}_1, \hat{\beta}_0$

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sum_{i=1}^{n}(x_i-\bar{x})^2} = \frac{1988.4}{6032} = 0.3296$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1\bar{x} = 29.05 - 0.3296 \times 65 = 7.626$$

$$r = \frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2 \times \sum_{i=1}^{n}(y_i-\bar{y})^2}} = \frac{1988.4}{\sqrt{6032 \times 835.42}} = 0.8857$$

$r^2 = 0.7846$

$1 - r^2 = 0.2154$

$$s^2 = \frac{\sum_{i=1}^{n}(y_i-\hat{y})^2}{n-2} = \frac{(1-r^2)\sum_{i=1}^{n}(y_i-\bar{y})^2}{n-2} = \frac{0.2154 \times 835.42}{10} = 17996$$

$s = 4.242$

Residual Plot

- e_i (residual $= y_i - \hat{y}$) vs y_i (fitted value)
- homoscedastic – vertical spread does not vary with fitted value – linear model
- heteroscedastic – spread varies – no linear model
- apply transformations

Transformations

- $\ln(y)$ vs x
- $\ln(y)$ vs $\ln(x)$
- y^b vs x^a
- $(\ln(y))^a$ vs $(\ln(x))^b$