

Panel 1

Goal: Investigate 2 numeric variables:

- Graph - Scatter Plot
- Best-fit line (least-square regression line)  $y = mx + b$
- Estimate correlation coeff.  $r$
- With formulas
 
$$S_{xx} = \sum x^2 - \frac{(\sum x)^2}{n}$$

$$S_{yy} = \sum y^2 - \frac{(\sum y)^2}{n}$$

$$S_{xy} = \sum xy - \frac{(\sum x)(\sum y)}{n}$$

$$r = \frac{S_{xy}}{\sqrt{S_{xx} S_{yy}}}, \quad m = \frac{S_{xy}}{S_{xx}}, \quad b = \bar{y} - m\bar{x}$$

Panel 2

pre-test	post-test
1	2
2	4
3	4
4	7
5	8

sketch

$y = mx + b$   
 $-11x + 13$

$r = 0.9022$  (strong pos. relation)

Predict  $y$  if  $x = 9$  (say):  $y = 11 \cdot 9 + 13 = 112$

scatter plot  $\Rightarrow$

Simple linear regression results  
 Dependent Variable: posttest  
 Independent Variable: pretest  
 pre-test = 1.3 + 1.1 post-test  
 Sample size: 5  
 R Squared Coefficient: 0.8122  
 R Squared: 0.7995303  
 Estimate of error standard deviation: 1.8182

Parameter	Estimate	Std. Err.	Alternative DF	T-Stat	P-Value
Intercept	1.3	1.0861897	<=	1.2183188	0.3090
pre-test	1.1	0.2274988	<=	4.8378488	0.0070

Analysis of Variance table for regression model

Source	DF	SS	MS	F-Stat	P-Value
Model	1	12.1	12.1	11.706376	0.0418
Error	3	2.8	0.9333333		
Total	4	15.2			

Panel 3

For ques:

x	y	x <sup>2</sup>
3	4	9
2	4	4
1	1	1
6		

$\bar{x} = \frac{6}{3} = 2$   
 $\bar{y} = \frac{13}{3} = 4.33$

Find  $S_{xx} \rightarrow 14$

Given that  $S_{xx} = 6$ ,  $S_{yy} = 5$ ,  $S_{xy} = 4$ , find  $m, b, r$

$m = \frac{4}{6} = \frac{2}{3} = 0.67$   
 $b = 4.33 - 0.67 \cdot 2 = 2.99$   
 $\bar{y} = m\bar{x}$

$r = \frac{4}{\sqrt{6 \cdot 5}} = \frac{4}{\sqrt{30}} = 0.73$

Panel 4

Suppose  $y = 3x + 2$ . Predict  $y$  if  $x = 4$ .

$r = 0.037$

$y = 3 \cdot (4) + 2 = 14$

Prediction is not reliable because  $r \approx 0$

Draw best-fit line and estimate  $y = mx + b$   
 $m = \frac{2}{1} = 2, \quad b = -2$

Panel 5

I. Basic analysis of 1 variable ✓ mean, median, ..., histogram or freq. distr., std. dev., variance, quantiles, percentiles

II. 2 variables + associations ✓ count table + Chi-Square test

III: Hypothesis Testing + Estimation

Regressions, linear

First: some probability theory

Ex: Flip coin once, what is the probability that H is up?  $\frac{1}{2}$

fair  $\frac{1}{2}$  good  $\frac{1}{2}$

Panel 6

Ex: Flip <sup>one</sup> coin twice. Prob of at least one Tail?  $\frac{3}{4} = 0.75$

All possibilities: TT, TH, HT, HH

2 ways to assign prob.:  $\rightarrow$  thinking and looking all possible outcomes  
 $\rightarrow$  experimentation

Ex: Roll a die.  $P(\text{even}) = \frac{3}{6} = \frac{1}{2}$

Ex: Roll two dice + record the sum.  
 $P(\text{sum is at least } 5)$

Panel 7

Get organized:

sum	1	2	3	4	5	6
1	2	3	4	5	6	7
2	3	4	5	6	7	8
3	4	5	6	7	8	9
4	5	6	7	8	9	10
5	6	7	8	9	10	11
6	7	8	9	10	11	12

$P(\text{sum} = 10) = \frac{3}{36} = \frac{1}{12} = \#$

$P(\text{sum} = 5) = \frac{4}{36} = \frac{1}{9}$

$P(\text{sum at least } 5) = \frac{20}{36} = \frac{5}{9}$

opposite  $\frac{6}{36}$

Panel 8

Principles of Probabilities

$P(E)$  stands for probability that event E occurs

- $P(E)$  is between 0 and 1
- $P(E) = 0$  means E is impossible  
 $P(E) = 1$  means E will happen for sure
- $P(\text{everything}) = 1$
- $P(E) = 1 - P(\text{opposite event})$

Panel 9

Probabilities can be found by counting + organizing, or by experimentation

Ex: E is the event of a random person on the street making between \$30K and \$50K. Find  $P(E) = 0.225 + 0.069$

Salary	Salary	relat %	probabilities
0-10K	6.9	0.069	
10-20K	49.5	0.495	
20-30K	22.5	0.225	
30-40K	6.9	0.069	
40-50K	5.5	0.055	
>50	9.5	0.095	

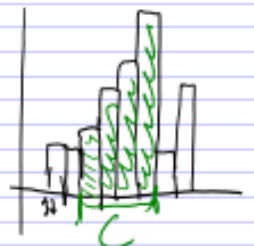
Panel 10

Using Histograms: Study to record how tall people are

Height	Frequency	probabilities
57-58.5	1	0.025
58.5-60.1	3	0.075
60.1-61.7	6	0.15
61.7-63.3	8	0.2
63.3-64.9	11	0.275
64.9-66.4	7	0.175
66.4-68.0	4	0.1
<b>Total</b>	<b>40</b>	

$P(A) = \frac{5}{40}$   
 $P(B) = \frac{10}{40}$   
 $P(C) = 1 - \frac{5}{40} - \frac{10}{40} = \frac{25}{40}$

Panel 11

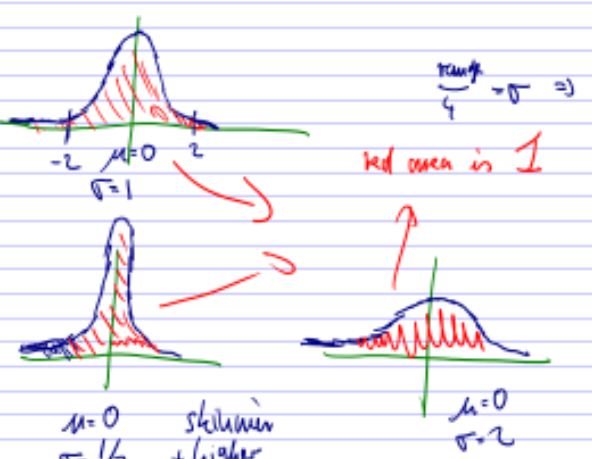


$P(C) = \frac{\text{area of } C}{\text{total area}}$

Most distributions are bell-shaped or normal. One of these bell-shaped distributions will be called Standard Normal Dist (mean = 0, standard deviation = 1)

Panel 12

Standard Normal Distribution



$\frac{\text{range}}{4} = \sigma \Rightarrow \text{range} = 4$   
 red area is 1  
 $\mu = 0, \sigma = \frac{1}{2}$  (skinnier + higher)  
 $\mu = 0, \sigma = 2$

Panel 13

Want to compute probabilities of standard normal distributions

$P(x > 1) = \frac{\text{Shaded area}}{\text{Total area}}$

$= \frac{\text{Shaded}}{1}$

$P(x > 1.00) = 0.2420$

Know: 95% of data is between -2 and 2

$P(x > 2) = 0.0228$

Panel 14

$P(x > 0.53) = 0.2981$

$P(x < 0.53) = 0.7019$

$-0.2981$

z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
0.0	.5000	.4960	.4920	.4880	.4840	.4801	.4761	.4721	.4681	.4641
0.1	.4602	.4562	.4522	.4483	.4443	.4404	.4364	.4325	.4286	.4247
0.2	.4207	.4168	.4129	.4090	.4051	.4013	.3974	.3936	.3897	.3859
0.3	.3821	.3783	.3745	.3707	.3669	.3632	.3594	.3557	.3520	.3483
0.4	.3446	.3409	.3372	.3336	.3300	.3264	.3228	.3192	.3156	.3121
0.5	.3085	.3050	.3015	.2981	.2946	.2912	.2877	.2843	.2809	.2776
0.6	.2743	.2709	.2676	.2643	.2611	.2578	.2546	.2514	.2483	.2451
0.7	.2420	.2389	.2358	.2327	.2296	.2266	.2236	.2206	.2177	.2148
0.8	.2118	.2088	.2059	.2030	.2001	.1972	.1943	.1914	.1886	.1857
0.9	.1828	.1799	.1770	.1742	.1714	.1685	.1657	.1629	.1601	.1573
1.0	.1545	.1517	.1489	.1462	.1434	.1406	.1379	.1351	.1324	.1297
1.1	.1270	.1243	.1216	.1189	.1162	.1135	.1108	.1081	.1054	.1027
1.2	.1000	.0974	.0948	.0922	.0896	.0870	.0845	.0819	.0793	.0768