Panel 1

Goal: Investigate 2 numeric variables:
Is there a linear relation between them.

1. Scatter plot
2. Correlation coefficient
3. Compute least square regression line
4. Make predictions

Formulas:

$$S_{xx} = \sum x^2 - \frac{(\sum x)^2}{n}$$

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

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

$$r = \frac{S_{xy}}{\sqrt{S_{xx} \cdot S_{yy}}}$$

$$m = \frac{S_{xy}}{S_{xx}} \qquad b = \bar{y} - m\bar{x}$$

1

Panel 2

Highest year of school completed, father x	Highest year of school completed y	x^2	y^2	xy
12	12	144	144	144
15	16	225	256	240
5	7	25	49	35
16	19	256	361	304
48	54	650	810	723

$$S_{xx} = 650 - \frac{48^2}{4} = 74$$

$$S_{yy} = 810 - \frac{54^2}{4} = 81$$

$$S_{xy} = 723 - \quad = 75$$

$$r = \frac{75}{\sqrt{74 \cdot 81}} = 0.968 \text{ close linear relation}$$

$$m = \frac{S_{xy}}{S_{xx}} = \frac{75}{74} = 1.01$$

$$b = \bar{y} - m\bar{x} = \frac{54}{4} - 1.01 \cdot \frac{48}{4} = 1.38$$

$y = mx + b$

$y = 1.01x + 1.38$

2

Panel 3

Simple linear regression results:
Dependent Variable: y
Independent Variable: x
y = 1.3378378 + 1.0135135 x
Sample size: 4
R (correlation coefficient) = 0.96873032
R-sq = 0.93843844
Estimate of error standard deviation: 1.5790007

Parameter estimates:

Parameter	Estimate	Std. Err.	Alternative	DF	T-Stat	P-Value
Intercept	1.3378378	2.3898776	≠ 0	2	0.57175547	0.6252
Slope	1.0135135	0.1835551	≠ 0	2	5.5215763	0.0313

Analysis of variance table for regression model:

Source	DF	SS	MS	F-stat	P-value
Model	1	76.013514	76.013514	30.487805	0.0313
Error	2	4.9864865	2.4932432		
Total	3	81			

$y = 1.01x + 1.337$

Suz dad has 14 years of school.

⇒ Predict

$y = 1.01 \cdot 14 + 1.337$

$= 15.477$

prediction is good because $r \approx 1$

3

Panel 4

Use GSS data to find out if HOURS OF TV WATCHING is related to AGE and predict how many hours a 31 year old person watches TV.

AGE = x

TV = y

$r = 0.15$

$m = 0.023$

$s = 1.95$

$y = 0.023 \cdot 31 + 1.95 = 2.59$

prediction is NOT believable because r close to 0!

4

Panel 5

Predict Life expectancy for a country with literacy rate of 75%.

Lit. Rate: x

Life Exp: y

$r = 0.843$

Model: $y = 0.77 \cdot x + 38.5$

Predicted y for $x = 75$ is $y = 66.77$

Sample size: 107
R (correlation coefficient) = 0.84359974
R-sq = 0.71166052
Estimate of error standard deviation: 5.4208087

Parameter estimates:

Parameter	Estimate	Std. Err.	Alterna
Intercept	38.469986	1.8770674	
Slope	0.37040223	0.023008831	

Analysis of variance table for regression model

Source	DF	SS	MS	F-stat
Model	1	7615.2864	7615.2864	259.1540
Error	105	3085.4426	29.385167	
Total	106	10700.729		

Predicted values:

X value	Predicted y	s.e.(Pred. y)	95% C.I.
75	66.250153	0.52964236	(65.19997,

5

Panel 6

I. Basic analysis of 1 variable ← mean, mode, median, std dev, Q_1, Q_3, minimum, range

II. 2 vars + their relation

ordinal → contingency tables, chi-Square test, p-values

numeric → linear regression

III. Hypothesis Testing and Estimation.

first. Probability Theory

Ex: Flip coin once. What is the probability that H is up?

$P(\text{heads}) = 50\%$ if coin is fair.

6

Panel 7

Prob. can be assigned experimentally or by counting.

Ex: Flip coin 1000 times ⇒ 537 H and 463 T

⇒ P(H) ≈ 50%

P(at least one T in flipping one coin twice) = 3/4

All possibilities: TT, TH, HT, HH

P(at least one T in flipping 2 coins simult.) = 2/3

All possible: TT, HT, HH

7

Panel 8

Ex: Roll a die. P(even) = 3/6 = 1/2 = 0.5 = 50%

Ex: Roll two dice + record the sum.

P(sum is at least 5) =

(1,6), (2,6), (3,5), ...

Get organized

8

Panel 9

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

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

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

$P(\text{sum at least } 5) = \frac{36}{36} - \frac{6}{36} =$

$\frac{30}{36}$

9

Panel 10

Principles of Probabilities

1. $P(E)$ is always between 0 and 1
2. $P(E) = 0$ means E won't happen for sure
   $P(E) = 1$ means E will happen, we know.
3. $P(\text{Everything}) = 1$
4. $P(\text{Event}) = 1 - P(\text{opposite event})$

10

Panel 11

Probabilities can be found by counting and plotting organized, or by experimentation

P(person's salary is between $30-$50K) = 0.22 + 0.069 = 0.29

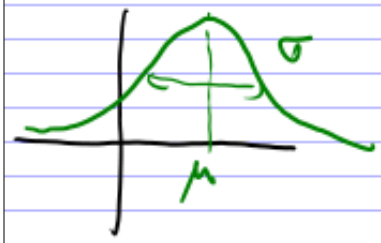
Survey: Salary	valid %	probabilities
10-20K	6.9	0.069
20-30K	49.5	0.495
30-40K	22.1	0.221
40-50K	6.9	0.069
50-60K	5.5	0.055
>60	9.5	0.095

Want to use frequency distributions to work out probabilities!

11

Panel 12

Standard Normal Distribution



Want to use these normal distributions to compute probabilities!

Next

Wed: Quiz with phone!

12