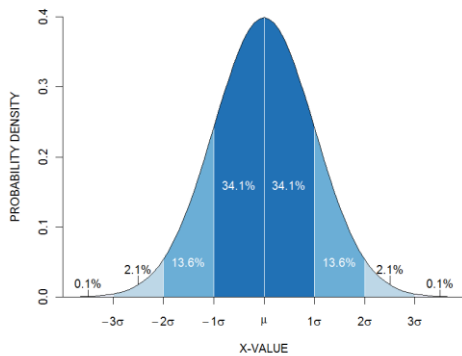


SPM8 for Basic and Clinical Investigators

Basic Statistics



Variance

$$\text{VAR} = \frac{\sum (X - u)^2}{N}$$

where: X = value

u = sample mean

N = sample size

Standard deviation

$$\text{SD} = \sqrt{\frac{\sum (X - u)^2}{N}}$$

where: X = value

u = sample mean

N = sample size

Standard deviation

$$\text{Unbiased SD} = \sqrt{\frac{\sum (X - u)^2}{N - 1}}$$

where: X = value

u = sample mean

N = sample size

Standard error

$$SE = \frac{SD}{\sqrt{N}}$$

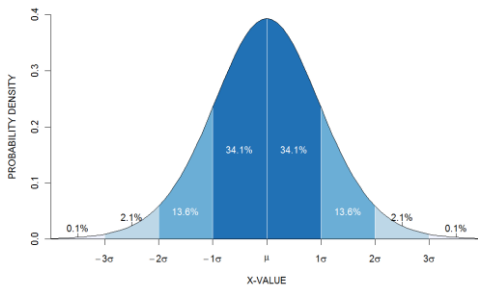
where: X = value
 u = sample mean
 N = sample size

Confidence interval

$$CI = u \pm t_{1-a/2} \frac{SD}{\sqrt{N}}$$

where: $t_{1-a/2}$ = t value for 100(1-a)% confidence
 SD = standard deviation
 u = sample mean
 N = sample size

t distribution



Confidence interval

$$CI = u \pm t_{1-a/2} \frac{SD}{\sqrt{N}}$$

where: $t_{1-a/2}$ = t value for 100(1-a)% confidence
 SD = standard deviation
 u = sample mean
 N = sample size

Descriptive statistics

$$VAR = \frac{\sum (X - u)^2}{N}$$

$$SD = \sqrt{\frac{\sum (X - u)^2}{N}}$$

$$SE = \frac{SD}{\sqrt{N}}$$

$$CI = u \pm t_{1-a/2} \frac{SD}{\sqrt{N}}$$

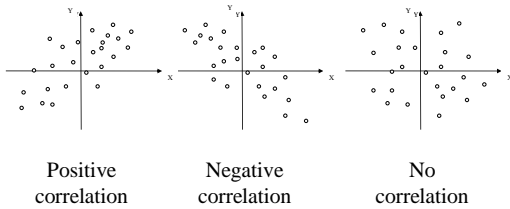
Correlation and Regression

- Is there a relationship between x and y ?
- What is the strength of this relationship
 - Pearson's r
- Can we describe this relationship and use it to predict y from x ?
 - Regression
- Is the relationship we have described statistically significant?
 - F- and t-tests
- Relevance to SPM
 - GLM

Relationship between x and y

- Correlation describes the strength and direction of a linear relationship between two variables
- Regression tells you how well a certain independent variable predicts a dependent variable
- CORRELATION and CAUSATION
 - In order to infer causality: manipulate independent variable and observe effect on dependent variable

Scattergrams



Variance vs. Covariance

- Do two variables change together?

Variance ~ $DX * DX$

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

Covariance ~ $DX * DY$

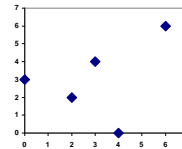
$$\text{cov}(x, y) = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{n}$$

Covariance

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

- When $X \uparrow$ and $Y \uparrow$: $\text{cov}(x, y) = \text{pos.}$
- When $X \downarrow$ and $Y \downarrow$: $\text{cov}(x, y) = \text{neg.}$
- When no constant relationship: $\text{cov}(x, y) = 0$

Example Covariance



x	y	$x_i - \bar{x}$	$y_i - \bar{y}$	$(x_i - \bar{x})(y_i - \bar{y})$
0	3	-3	0	0
2	2	-1	-1	1
3	4	0	1	0
4	0	1	-3	-3
6	6	3	3	9
$\bar{x}=3$	$\bar{y}=3$			$\Sigma=7$

$$\text{cov}(x, y) = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{n} = \frac{7}{5} = 1.4$$

What does this number tell us?

Example of how covariance value relies on variance

High variance data				Low variance data			
Subject	x	y	x error * y error	x	y	x error * y error	
1	101	100	2500	54	53	9	
2	81	80	900	53	52	4	
3	61	60	100	52	51	1	
4	51	50	0	51	50	0	
5	41	40	100	50	49	1	
6	21	20	900	49	48	4	
7	1	0	2500	48	47	9	
Mean	51	50		51	50		
Sum of x error * y error :			7000			28	
Covariance:			1166.67	Covariance:		4.67	

Pearson's R

$$-\infty \leq \text{COV}(x, y) \leq \infty$$

- Covariance does not really tell us anything
 - Solution: standardise this measure
- Pearson's R: standardise by adding std to equation:

$$r_{xy} = \frac{\text{COV}(x, y)}{s_x s_y}$$

Basic assumptions

- Normal distributions
- Variances are constant and not zero
- Independent sampling – no autocorrelations
- No errors in the values of the independent variable
- All causation in the model is one-way (not necessary mathematically, but essential for prediction)

Pearson's R: degree of linear dependence

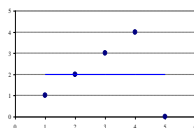
$$\text{cov}(x, y) = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{n} \rightarrow r_{xy} = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{n s_x s_y}$$

$-1 \leq r \leq 1$

$$r_{xy} = \frac{\sum_{i=1}^n Z_{x_i} * Z_{y_i}}{n}$$

Limitations of r

- r is actually
 - r = true r of whole population
 - = estimate of r based on data
- r is very sensitive to extreme values:



In the real world...

- r is never 1 or -1
- interpretations for correlations in psychological research (Cohen)

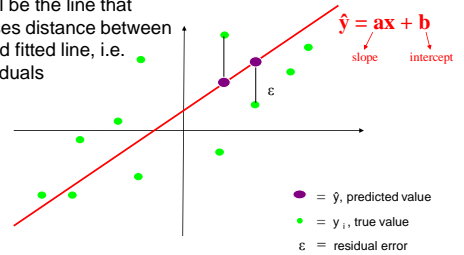
Correlation	Negative	Positive
Small	-0.29 to -0.10	00.10 to 0.29
Medium	-0.49 to -0.30	0.30 to 0.49
Large	-1.00 to -0.50	0.50 to 1.00

Regression

- Correlation tells you if there is an association between x and y but it doesn't describe the relationship or allow you to predict one variable from the other.
- To do this we need REGRESSION!

Best-fit Line

- Aim of linear regression is to fit a straight line, $\hat{y} = ax + b$, to data that gives best prediction of y for any value of x
- This will be the line that minimises distance between data and fitted line, i.e. the residuals



Least Squares Regression

- To find the best line we must minimise the sum of the squares of the residuals (the vertical distances from the data points to our line)

Model line: $\hat{y} = ax + b$ a = slope, b = intercept

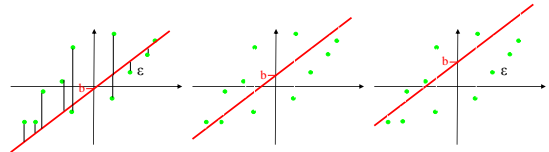
Residual (ϵ) = $y - \hat{y}$

Sum of squares of residuals = $\sum (y - \hat{y})^2$

- we must find values of a and b that minimise $\sum (y - \hat{y})^2$

Finding b

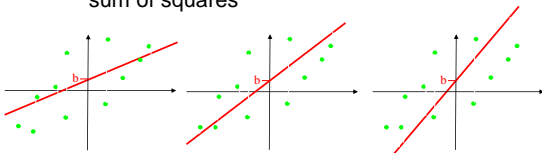
- First we find the value of b that gives the min sum of squares



- Trying different values of b is equivalent to shifting the line up and down the scatter plot

Finding a

- Now we find the value of a that gives the min sum of squares

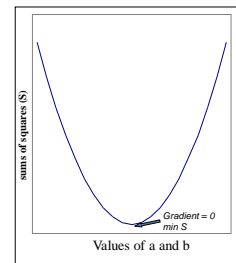


- Trying out different values of a is equivalent to changing the slope of the line, while b stays constant

Minimizing sums of squares

- Need to minimize $\sum (y - \hat{y})^2$
- $\hat{y} = ax + b$
- so need to minimize: $\sum (y - ax - b)^2$

- If we plot the sums of squares for all different values of a and b we get a parabola, because it is a squared term



- So the min sum of squares is at the bottom of the curve, where the gradient is zero.

Computation

- So we can find a and b that give min sum of squares by taking partial derivatives of $\sum(y - ax - b)^2$ with respect to a and b separately
- Then we solve these for 0 to give us the values of a and b that give the min sum of squares

The solution

- Doing this gives the following equations for a and b:

$$a = \frac{r s_y}{s_x}$$

r = correlation coefficient of x and y
 s_y = standard deviation of y
 s_x = standard deviation of x

- You can see that:
 - A low correlation coefficient gives a flatter slope (small value of a)
 - Large spread of y, i.e. high standard deviation, results in a steeper slope (high value of a)
 - Large spread of x, i.e. high standard deviation, results in a flatter slope (high value of a)

The solution cont.

- Our model equation is $\hat{y} = ax + b$
- This line must pass through the mean so:

$$\bar{y} = a\bar{x} + b \iff b = \bar{y} - a\bar{x}$$

- We can put our equation into this giving:

$$b = \bar{y} - \frac{r s_y}{s_x} \bar{x}$$

r = correlation coefficient of x and y
 s_y = standard deviation of y
 s_x = standard deviation of x

- The smaller the correlation, the closer the intercept is to the mean of y

Back to the model

- We can calculate the regression line for any data, but the important question is: How well does this line fit the data, or how good is it at predicting y from x?

How good is our model?

- Total variance of y: $s_y^2 = \frac{\sum(y - \bar{y})^2}{n - 1} = \frac{SS_y}{df_y}$

- Variance of predicted y values (\hat{y}):

$$s_{\hat{y}}^2 = \frac{\sum(\hat{y} - \bar{y})^2}{n - 1} = \frac{SS_{\text{pred}}}{df_{\hat{y}}}$$

This is the variance explained by our regression model

- Error variance:

$$s_{\text{error}}^2 = \frac{\sum(y - \hat{y})^2}{n - 2} = \frac{SS_{\text{er}}}{df_{\text{er}}}$$

This is the variance of the error between our predicted y values and the actual y values, and thus is the variance in y that is NOT explained by the regression model

How good is our model cont.

- Total variance = predicted variance + error variance

$$s_y^2 = s_{\hat{y}}^2 + s_{\text{er}}^2$$

- Conveniently, via some complicated rearranging

$$s_{\hat{y}}^2 = r^2 s_y^2$$

$$\Downarrow$$

$$r^2 = s_{\hat{y}}^2 / s_y^2$$

- so r^2 is the proportion of the variance in y that is explained by our regression model

How good is our model cont.

- Insert $r^2 s_y^2$ into $s_y^2 = s_y^2 + s_{er}^2$ and rearrange to get:

$$\begin{aligned} s_{er}^2 &= s_y^2 - r^2 s_y^2 \\ &= s_y^2 (1 - r^2) \end{aligned}$$

- From this we can see that the greater the correlation the smaller the error variance, so the better our prediction

Is the model significant?

- i.e. do we get a significantly better prediction of y from our regression equation than by just predicting the mean?

- F-statistic:

$$F_{(df_y, df_{er})} = \frac{s_y^2}{s_{er}^2} \stackrel{\text{complicated rearranging}}{=} \dots = \frac{r^2 (n-2)^2}{1-r^2}$$

- And it follows that:

$$\text{(because } F = t^2) \quad t_{(n-2)} = \frac{r(n-2)}{\sqrt{1-r^2}}$$

So all we need to know are r and n !