Week ON1

Elmer V Villanueva

24 February 2020

SET YOUR WORKING DIRECTORY!

```
setwd("C:/Users/1/Dropbox/00 - Working Folder/Teaching/DPH112/2019-2020/Week ON1")
```

Expectations of this module

Modes of delivery

This module will be delivered via a mixture of online and face-to-face modes in Semester 2, 2020.

Online delivery

We will use ICE as the main platform for the delivery of online material. You are required to engage with ICE throughout the duration of this semester. Study materials and readings will be distributed here, and you will participate in quizzes and submit your assessments through the same platform.

All material will be in the form of RMD or PDF files. There will be no videos or audios available.

Face-to-face delivery

Once the university opens (06 April is the expected date), lectures will replace the online delivery mode. Information about times and rooms will be distributed closer to the date.

Required software

We will be using R in an RStudio shell. You need to download and install both software programs. The version of R we will be using is 3.6.2 "Dark and Stormy Night".

Assessments

There are four assessment components.

Quizzes (15%)

Quizzes will be released every Thursday at 1:00 PM. You will be required to take the quiz at that time and submit your answers within 20 minutes. The quiz covers materials distributed the previous Monday

Weekly problem sets (15%)

Weekly problem sets are released every Thursday at 3:00 PM and submitted every Monday at 3:55 PM. The problem sets cover the material distributed the previous Monday.

Assignments (30%)

Two assignments each worth 15% of the entire module mark will be released for submission on Saturday 18 April and Saturday 24 May. Each assignment will require a formal scientific report of no more than 500 words and the provision of R code.

Final Report (40%)

The final report is due on Saturday 20 June. The final report will require the submission of a formal scientific report of no more than 2,500 words and the provision of R code.

Attendance

Submission fo the quiz constitutes attendance for the week. If you fail to submit the quiz, you will be marked absent.

Readings

Required reading will be distributed via ICE. A textbook by Vittinghoff and colleagues has been uploaded onto ICE. This textbook is an easy introduction to linear models.

Module handbook

The module handbook, uploaded onto ICE, describes all the necessary policies relevant to the delivery of this module. You are required to read and understand the module handbook.

Reading

Read and understand Triola, Triola and Roy, Chapter 10-1.

Covariance and Correlation

Correlation analysis is concerned with measuring the strength of the relationship between numeric variables. When we compute measures of correlation from a set of data, we are interested in the degree of the correlation between variables. The concepts and terminology of correlation analysis originated with Galton, who first used the word correlation in 1888.

Covariance

Suppose we have observations on n subjects consisting of two numeric variables G and H (Table 1). We wish to measure both the *direction* and *strength* of this relationship. Correlation involving two variables implies a co-relationship between variables that puts them on an equal footing and does not distinguish between them by referring to one as the dependent and the other as the independent variable. Thus, there is no attempt to prescribe or interpret the causality of the association. For example, there may be an association between arm and leg length in humans, whereby individuals with longer arms generally have longer legs. Neither variable directly causes the change in the other. Rather, they are both in???uenced by other variables to which they both have similar responses.

Table 1. Notation for data used in correlation analysis

Observation Number	G	H
1	g_1	h_1
2	g_2	h_2
3	g_3	h_3

Observation Number	G	H
...	...	...
n	g_n	h_n

Example. The purpose of a study by Kwast-Rabben et al. [1] was to analyse somatosensory evoked potentials (SEPs) and their interrelations following stimulation of digits I, III, and V in the hand. The researchers wanted to establish reference criteria in a control population. Thus, healthy volunteers were recruited for the study. In the future this information could be quite valuable as SEPs may provide a method to demonstrate functional disturbances in patients with suspected cervical root lesion who have pain and sensory symptoms. In the study, stimulation below-pain-level intensity was applied to the fingers. Recordings of spinal responses were made with electrodes fixed by adhesive electrode cream to the subject???s skin. One of the relationships of interest was the relationship between a subject???s height (cm) and the peak spinal latency (Cv in msec) of the SEP. Data for 15 participants are given in Table 2.

Table 2. Height and spine SEP measurements (Cv) from stimulation of digit I for 15 subjects described in Kwast-Rabben et al.

Observation Number	Height (G)	Cv (H)
1	149	14.4
2	155	13.5
3	156	13.0
4	161	15.5
5	161	14.6
6	168	15.3
7	170	16.6
8	171	16.5
9	172	16.8
10	173	17.3
11	181	15.8
12	184	18.4
13	184	17.4
14	185	19.0
15	187	17.8

Let us produce a scatterplot.

```
HEIGHT <- c(149, 155, 156, 161, 161,
            168, 170, 171, 172, 173,
            181, 184, 184, 185, 187)
CV <- c(14.4, 13.5, 13.0, 15.5, 14.6,
        15.3, 16.6, 16.5, 16.8, 17.3,
        15.8, 18.4, 17.4, 19.0, 17.8)
KWASTSUB <- data.frame(HEIGHT, CV)
str(KWASTSUB)
```

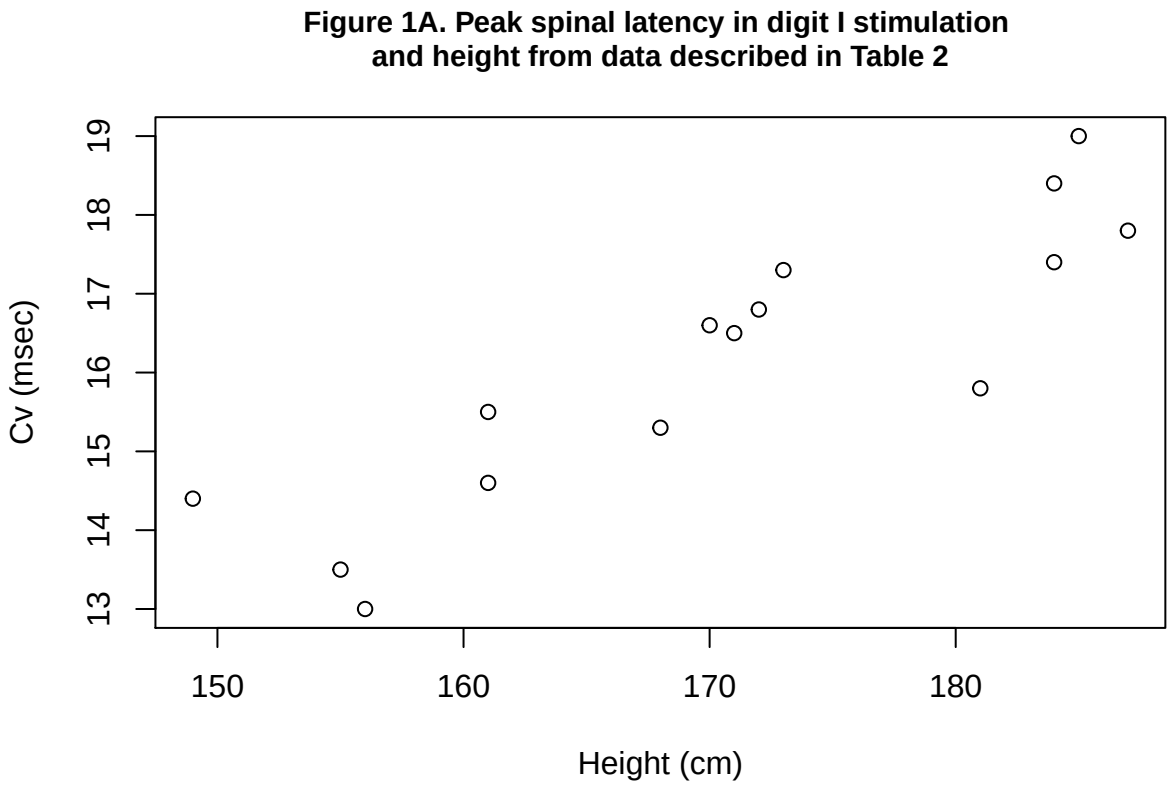
```
## 'data.frame':    15 obs. of  2 variables:
##  $ HEIGHT: num  149 155 156 161 161 168 170 171 172 173 ...
##  $ CV    : num  14.4 13.5 13 15.5 14.6 15.3 16.6 16.5 16.8 17.3 ...
```

```
head(KWASTSUB)
```

```
##   HEIGHT   CV
## 1    149 14.4
```

```
## 2    155 13.5
## 3    156 13.0
## 4    161 15.5
## 5    161 14.6
## 6    168 15.3
```

```
plot(HEIGHT, CV,
     main = "Figure 1A. Peak spinal latency in digit I stimulation \nand height from data described in T
     ylab = "Cv (msec)",
     xlab = "Height (cm)",
     cex.main = 0.9)
```

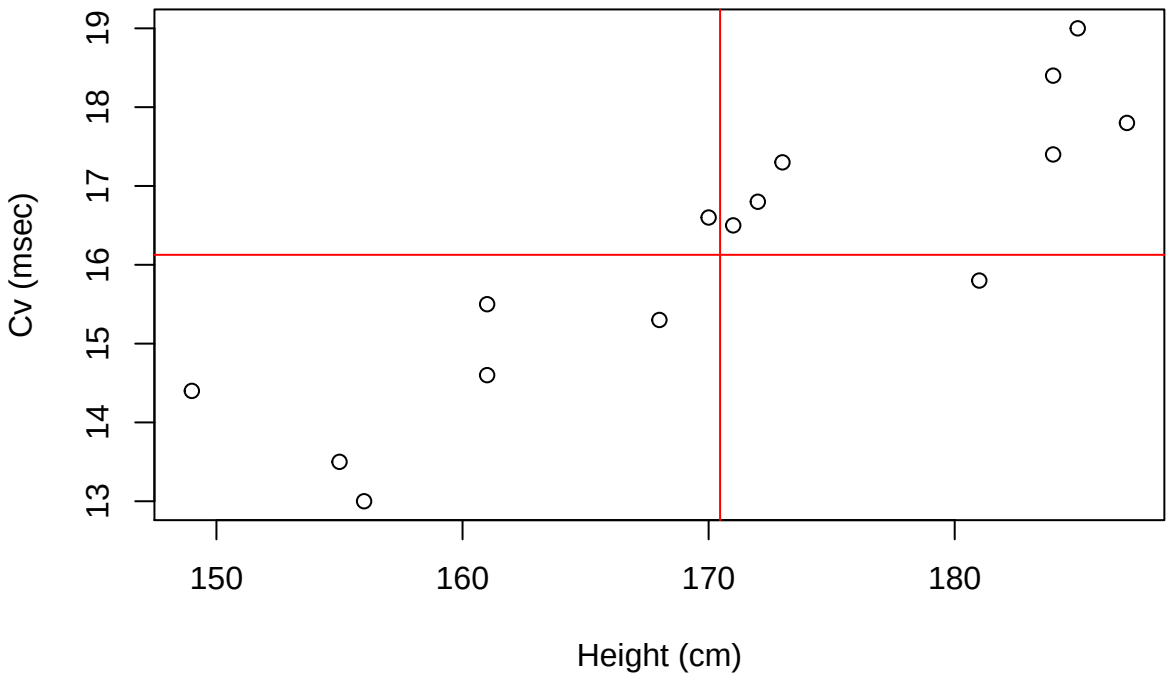


Note that since there is no expectation of a causal relationship between the two variables, it is just as appropriate for us to have plotted height against peak spinal latency.

Now, let us draw a vertical line at $\bar{g}$ and a vertical line at $\bar{h}$.

```
plot(HEIGHT, CV,
     main = "Figure 1B. Peak spinal latency in digit I stimulation \nand height from data described in T
     ylab = "Cv (msec)",
     xlab = "Height (cm)",
     cex.main = 0.9)
abline(h = mean(CV), v = mean(HEIGHT),
       col = c("red","red"))
```

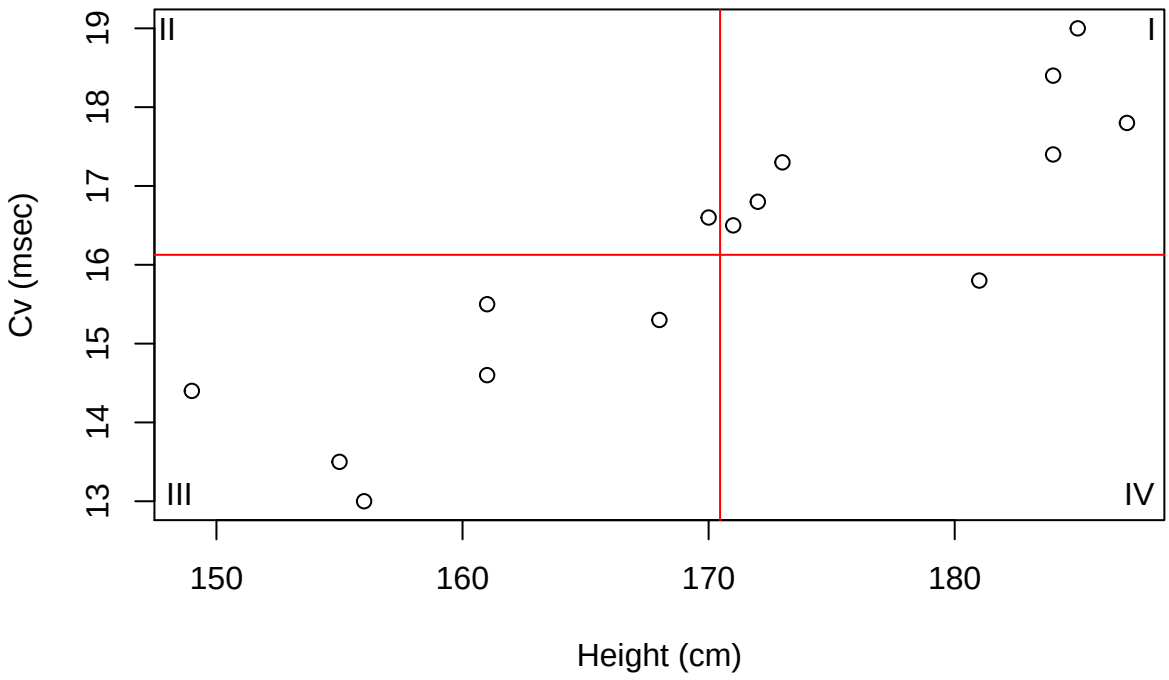
Figure 1B. Peak spinal latency in digit I stimulation and height from data described in Table 2



Note how the scatterplot is now divided into four quadrants:

```
plot(HEIGHT, CV,
     main = "Figure 1C. Peak spinal latency in digit I stimulation \nand height from data described in T
     ylab = "Cv (msec)",
     xlab = "Height (cm)",
     cex.main = 0.9)
abline(h = mean(CV), v = mean(HEIGHT),
       col = c("red","red"))
text(188, 19, "I")
text(148, 19, "II")
text(148.5, 13.1, "III")
text(187.5, 13.1, "IV")
```

Figure 1C. Peak spinal latency in digit I stimulation and height from data described in Table 2



For each point i in the graph, we will calculate the following: - $h_i - \bar{h}$, the deviation of each observation's peak spinal latency h_i from the mean peak spinal latency $\bar{h}$; - $g_i - \bar{g}$, the deviation of each observation's height g_i from the mean height $\bar{g}$; and - the product of the above two quantities $(h_i - \bar{h})(g_i - \bar{g})$

It is clear form Figure 1C that the quantity $h_i - \bar{h}$ is positive for every point in the first and second quadrants and is negative for every point in the third and fourth quadrants. Similarly, the quantity $g_i - \bar{g}$ is positive for every point in the first and fourth quadrants and is negative for every point in the second and third quadrants. These facts are summarized in Table 3.

Table 3. Algebraic signs of the quantities $(h_i - \bar{h})$ and $(g_i - \bar{g})$

Quadrant	$h_i - \bar{h}$	$g_i - \bar{g}$	$(h_i - \bar{h})(g_i - \bar{g})$
I	+	+	+
II	+	-	-
III	-	-	+
IV	-	+	-

If the relationship between G and H is positive (that is, as G increases H also increases), then there are more points in the first and third quadrants than in the second and fourth quadrants. In this case, the sum of the last column in Table 3 is likely to be positive because there are more positive than negative quantities. Conversely, if the relationship between G and H is negative (as G increases H decreases), then there are more points in the second and fourth quadrants than in the first and third quadrants. Hence the sum of the last column in Table 3 is likely to be negative. Therefore, the sign of the quantity

$$Cov(G,H) = \frac{\sum_{i=1}^{n}(g_i - \bar{g})(h_i - \bar{h})}{n-1}$$

which is known as the *covariance* between G and H, indicates the direction of the linear relationship between G and H. If $Cov(G, H) > 0$, then there is a positive relationship between G and H, but if $Cov(G, H) < 0$, then the relationship is negative.

Let us calculate the covariance of the data in Table 2 in R.

```
cov(HEIGHT, CV)
```

```
## [1] 18.8081
```

The value is positive indicating that as the peak spinal latency increases, so too does height.

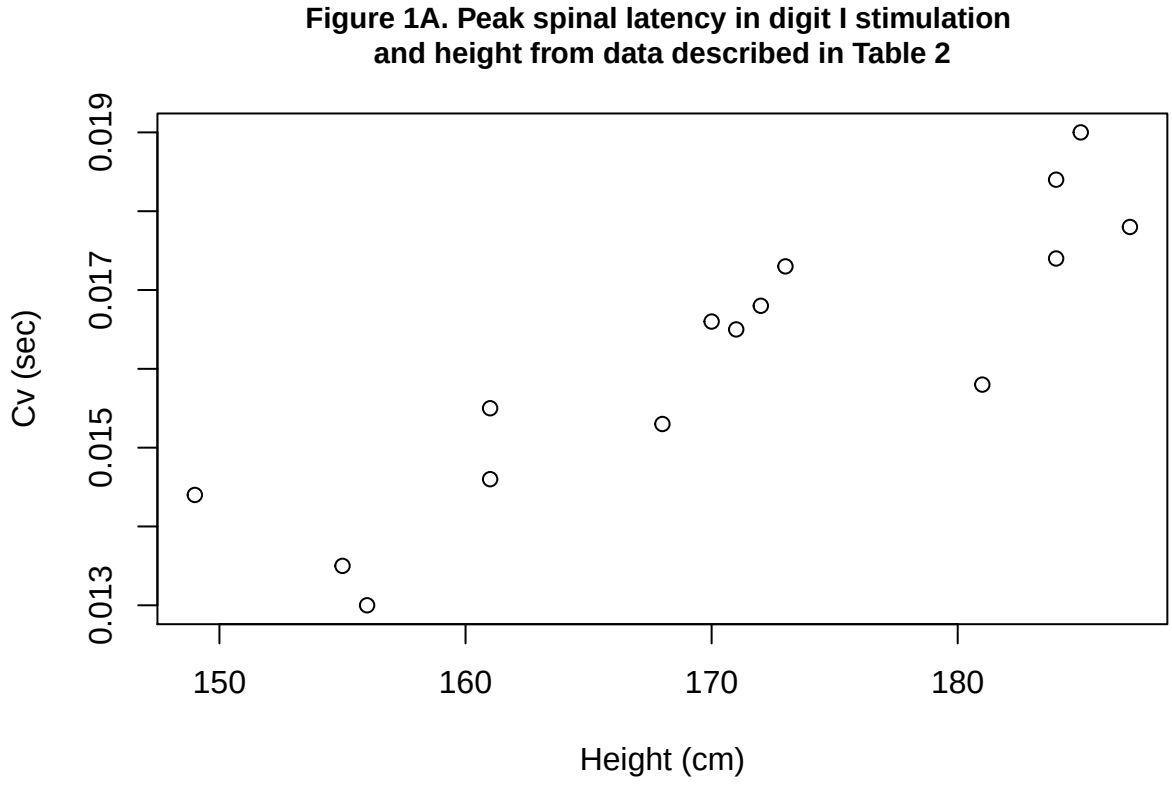
Unfortunately, $Cov(G, H)$ does not tell us much about the strength of such a relationship because it is affected by changes in the units of measurement.

Example. Consider what would happen if we measured peak spinal latency in seconds instead of milliseconds.

```
CV2 <- CV/1000
```

The spread of the points on the scatterplot is the same.

```
plot(HEIGHT, CV2,
     main = "Figure 1A. Peak spinal latency in digit I stimulation \nand height from data described in T
     ylab = "Cv (sec)",
     xlab = "Height (cm)",
     cex.main = 0.9)
```



However, when we calculate the covariance, we get a major difference.

```
cov(HEIGHT, CV2)
```

```
## [1] 0.0188081
```

Correlation

To avoid this disadvantage of the covariance, we standardise the data before computing the covariance. To standardise the H data, we first subtract the mean from each observation then divide by the standard deviation, that is, we compute $z_h = \frac{h_i - \bar{h}}{s_h}$. We know from DPH101 that the standardised variable $z_h \sim N(0, 1)$.

Similarly, we standardise the G data by subtracting the mean from each observation then divide by the standard deviation: $z_g = \frac{g_i - \bar{g}}{s_g}$.

If we do this, then the quantity $\dfrac{Cov(G, H)}{s_g s_h}$ is the covariance between the standardised G and standardised H data. This quantity has the special name of the *Pearson's product moment correlation coefficient* or $Cor(G, H)$ or ρ.

Like the covariance, the correlation coefficient is symmetric. That is, $Cor(G, H) = Cor(H, G)$. Unlike the covariance, however, the correlation coefficient is scale invariant. That is, the value of the quantity does not change if we change the units of measurement.

Example. Calculate the covariance of HEIGHT and CV, and HEIGHT and CV2. Then, calculate the correlation of HEIGHT and CV, and HEIGHT and CV2.

```
cov(HEIGHT, CV); cov(HEIGHT, CV2)
```

```
## [1] 18.8081
```

```
## [1] 0.0188081
```

```
cor(HEIGHT, CV); cor(HEIGHT, CV2)
```

```
## [1] 0.881347
```

```
## [1] 0.881347
```

Furthermore, $-1 \leq \rho \leq 1$.

The sign of $Cor(G, H)$ indicates the direction of the relationship between G and H. That is, $\rho > 0$ implies that G and H are positively related. Conversely, $\rho < 0$ implies that G and H are negatively related.

These properties make the ρ a useful quantity for measuring both the direction and the strength of the relationship between G and H. The magnitude of ρ measures the strength of the linear relationship between G and H. The closer ρ is to 1 or -1, the stronger is the relationship between G and H.

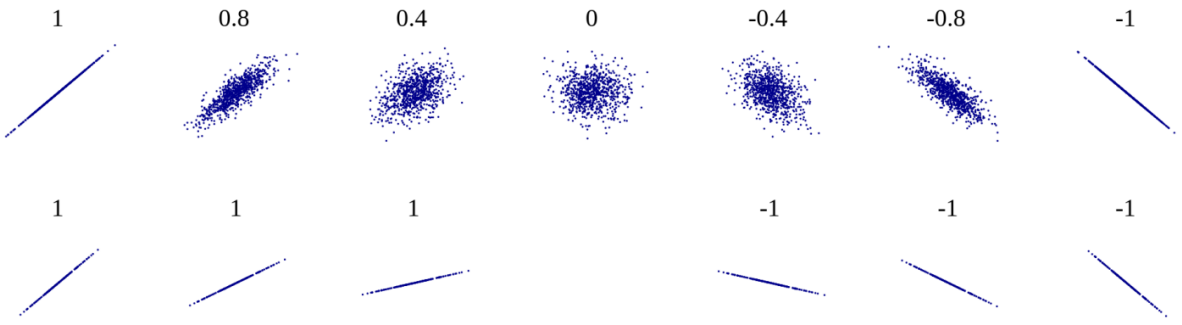


Figure 1:

Figure 2. Taxonomy of correlations. Top: Typical situations of a linear relationship of varying magnitudes of /rho. Bottom: Examples of perfect correlations.

Note that $\rho = 0$ does not necessarily mean that G and H are not related. It only implies that they are not *linearly* related because the correlation coefficient measures only *linear* relationships. In other words, ρ can still be zero when G and H are nonlinearly related.

Example. Consider the data in Table 4. First, calculate ρ. Then, produce a scatterplot of S against T.

Table 4. Data with a perfect relationship.

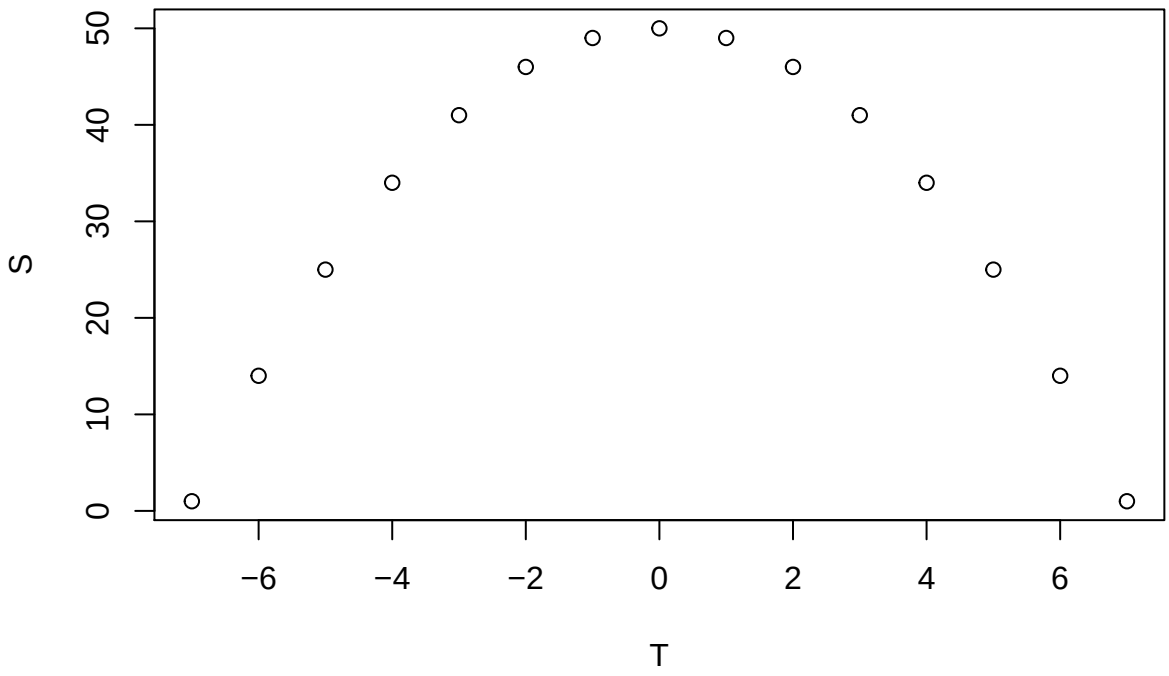
S	T
1	-7
14	-6
25	-5
34	-4
41	-3
46	-2
49	-1
50	0
49	1
46	2
41	3
34	4
25	5
14	6
1	7

```
S <- c(1, 14, 25, 34, 41, 46, 49, 50,
       49, 46, 41, 34, 25, 14, 1)
T <- c(-7:7)
cor(S, T)
```

```
## [1] 0
```

```
plot(T, S,
     main = "Figure 3. Graphical representation of data in Table 4.")
```

Figure 3. Graphical representation of data in Table 4.



The example above emphasises a point made consistently in DPH101. It is quite important that you learn to visualise your data before you perform your calculations. In the preceding example, depending merely on $\rho = 0$ would lead a naive analyst to conclude that there was no relationship between the two variables when a perfect nonlinear relationship does, in fact, exist.

The Bivariate Normal Distribution

Under the correlation model, G and H are assumed to vary together in what is called a joint distribution. If this joint distribution is a normal distribution, it is referred to as a *bivariate normal distribution.* Inferences regarding this population may be made based on the results of samples properly drawn from it. If, on the other hand, the form of the joint distribution is known to be nonnormal, or if the form is unknown and there is no justification for assuming normality, inferential procedures are invalid, although descriptive measures may be computed.

Assumptions about the bivariate normal distribution

The following assumptions must hold for inferences about the population to be valid when sampling is from a bivariate distribution.

1. For each value of G there is a normally distributed subpopulation of H values.
2. For each value of H there is a normally distributed subpopulation of G values.
3. The joint distribution of G and H is a normal distribution called the bivariate normal distribution.
4. The subpopulations of H values all have the same variance.
5. The subpopulations of G values all have the same variance.

Correlation coefficients in the bivariate normal distribution

The bivariate normal distribution five parameters, s_g, s_h, $\bar{g}$, $\bar{h}$, and ρ. The first four are, respectively, the standard deviations and means associated with the individual distributions. The remaining parameter, ρ, is the *population correlation coefficient* developed above and measures the strength of the linear relationship between G and H.

Hypothesis testing

We are usually interested in knowing if we may conclude that $\rho \neq 0$. That is, that G and H are linearly correlated. Since ρ is usually unknown, we draw a random sample from the population of interest, compute r (the *sample* correlation coefficient), the estimate of ρ, and test $H_0 : \rho = 0$ against the alternative $H_A : \rho \neq 0$.

The test statistic to be used is $t = r\sqrt{(n-2)/(1-r^2)}$ and is distributed as a Student's t disitrbution with $n-2$ degrees of freedom.

Example. The data from Kwast-Rabben et al. provided in Table 2 is a small subset of the entire set. In their study, Kwast-Rabben et al. colelcted data from 155 normal participants. These are given in full in Table 5.

Table 5. Height and spine SEP measurements (Cv) from stimulation of digit I for 155 subjects described in Kwast-Rabben et al.

ID	Height	Cv	ID	Height	Cv	ID	Height	Cv	ID	Height	Cv	ID	Height	Cv
1	149	14.4	32	164	16.0	63	170	17.0	94	179	17.6	125	187	17.8
2	149	13.4	33	164	16.0	64	170	16.4	95	179	17.8	126	187	19.3
3	155	13.5	34	165	15.7	65	171	16.5	96	179	16.1	127	188	17.5
4	155	13.5	35	165	16.3	66	171	16.3	97	179	16.0	128	188	18.0
5	156	13.0	36	165	17.4	67	171	16.4	98	179	16.0	129	189	18.0
6	156	13.6	37	165	17.0	68	171	16.5	99	179	17.5	130	189	18.8
7	157	14.3	38	165	16.3	69	172	17.6	100	179	17.5	131	190	18.3
8	157	14.9	39	166	14.1	70	172	16.8	101	180	18.0	132	190	18.6
9	158	14.0	40	166	14.2	71	172	17.0	102	180	17.9	133	190	18.8
10	158	14.0	41	166	14.7	72	172	17.6	103	181	18.4	134	190	19.2
11	160	15.4	42	166	13.9	73	173	17.3	104	181	16.4	135	191	18.5
12	160	14.7	43	166	17.2	74	173	16.8	105	181	15.8	136	191	18.5
13	161	15.5	44	167	16.7	75	174	15.5	106	181	18.8	137	191	19.0
14	161	15.7	45	167	16.5	76	174	15.5	107	181	18.6	138	191	18.5
15	161	15.8	46	167	14.7	77	175	17.0	108	182	18.0	139	194	19.8
16	161	16.0	47	167	14.3	78	175	15.6	109	182	17.9	140	194	18.8
17	161	14.6	48	167	14.8	79	175	16.8	110	182	17.5	141	194	18.4
18	161	15.2	49	167	15.0	80	175	17.4	111	182	17.4	142	194	19.0
19	162	15.2	50	167	15.5	81	175	17.6	112	182	17.0	143	195	18.0
20	162	16.5	51	167	15.4	82	175	16.5	113	182	17.5	144	195	18.2
21	162	17.0	52	168	17.3	83	175	16.6	114	182	17.8	145	196	17.6
22	162	14.7	53	168	16.3	84	175	17.0	115	184	18.4	146	196	18.3
23	163	16.0	54	168	15.3	85	176	18.0	116	184	18.5	147	197	18.9
24	163	15.8	55	168	16.0	86	176	17.0	117	184	17.7	148	197	19.2
25	163	17.0	56	168	16.6	87	176	17.4	118	184	17.7	149	200	21.0
26	163	15.1	57	168	15.7	88	176	18.2	119	184	17.4	150	200	19.2
27	163	14.6	58	168	16.3	89	176	17.3	120	184	18.4	151	202	18.6
28	163	15.6	59	168	16.6	90	177	17.2	121	185	19.0	152	202	18.6
29	163	14.6	60	168	15.4	91	177	18.3	122	185	19.6	153	182	20.0
30	164	17.0	61	170	16.6	92	179	16.4	123	187	19.1	154	190	20.0
31	164	16.3	62	170	16.0	93	179	16.1	124	187	19.2	155	190	19.5

Let's enter the data into a data frame.

```
HEIGHT.FULL <- c(149,149,155,155,156,156,157,157,158,158,160,160,161,161,161,
                 161,161,161,162,162,162,162,163,163,163,163,163,163,163,164,
                 164,164,164,165,165,165,165,165,166,166,166,166,166,167,167,
                 167,167,167,167,167,167,168,168,168,168,168,168,168,168,168,
                 170,170,170,170,171,171,171,171,172,172,172,172,173,173,174,
                 174,175,175,175,175,175,175,175,175,176,176,176,176,176,177,
                 177,179,179,179,179,179,179,179,179,179,180,180,181,181,181,
                 181,181,182,182,182,182,182,182,182,184,184,184,184,184,184,
                 185,185,187,187,187,187,188,188,189,189,190,190,190,190,191,
                 191,191,191,194,194,194,194,195,195,196,196,197,197,200,200,
                 202,202,182,190,190)
CV.FULL <- c(14.4,13.4,13.5,13.5,13.0,13.6,14.3,14.9,14.0,14.0,15.4,14.7,15.5,15.7,15.8,
             16.0,14.6,15.2,15.2,16.5,17.0,14.7,16.0,15.8,17.0,15.1,14.6,15.6,14.6,17.0,
             16.3,16.0,16.0,15.7,16.3,17.4,17.0,16.3,14.1,14.2,14.7,13.9,17.2,16.7,16.5,
             14.7,14.3,14.8,15.0,15.5,15.4,17.3,16.3,15.3,16.0,16.6,15.7,16.3,16.6,15.4,
             16.6,16.0,17.0,16.4,16.5,16.3,16.4,16.5,17.6,16.8,17.0,17.6,17.3,16.8,15.5,
             15.5,17.0,15.6,16.8,17.4,17.6,16.5,16.6,17.0,18.0,17.0,17.4,18.2,17.3,17.2,
             18.3,16.4,16.1,17.6,17.8,16.1,16.0,16.0,17.5,17.5,18.0,17.9,18.4,16.4,15.8,
             18.8,18.6,18.0,17.9,17.5,17.4,17.0,17.5,17.8,18.4,18.5,17.7,17.7,17.4,18.4,
             19.0,19.6,19.1,19.2,17.8,19.3,17.5,18.0,18.0,18.8,18.3,18.6,18.8,19.2,18.5,
             18.5,19.0,18.5,19.8,18.8,18.4,19.0,18.0,18.2,17.6,18.3,18.9,19.2,21.0,19.2,
             18.6,18.6,20.0,20.0,19.5)
KWAST.FULL <- data.frame(HEIGHT.FULL, CV.FULL)
str(KWAST.FULL)
```

```
## 'data.frame':    155 obs. of  2 variables:
##  $ HEIGHT.FULL: num  149 149 155 155 156 156 157 157 158 158 ...
##  $ CV.FULL    : num  14.4 13.4 13.5 13.5 13 13.6 14.3 14.9 14 14 ...
```

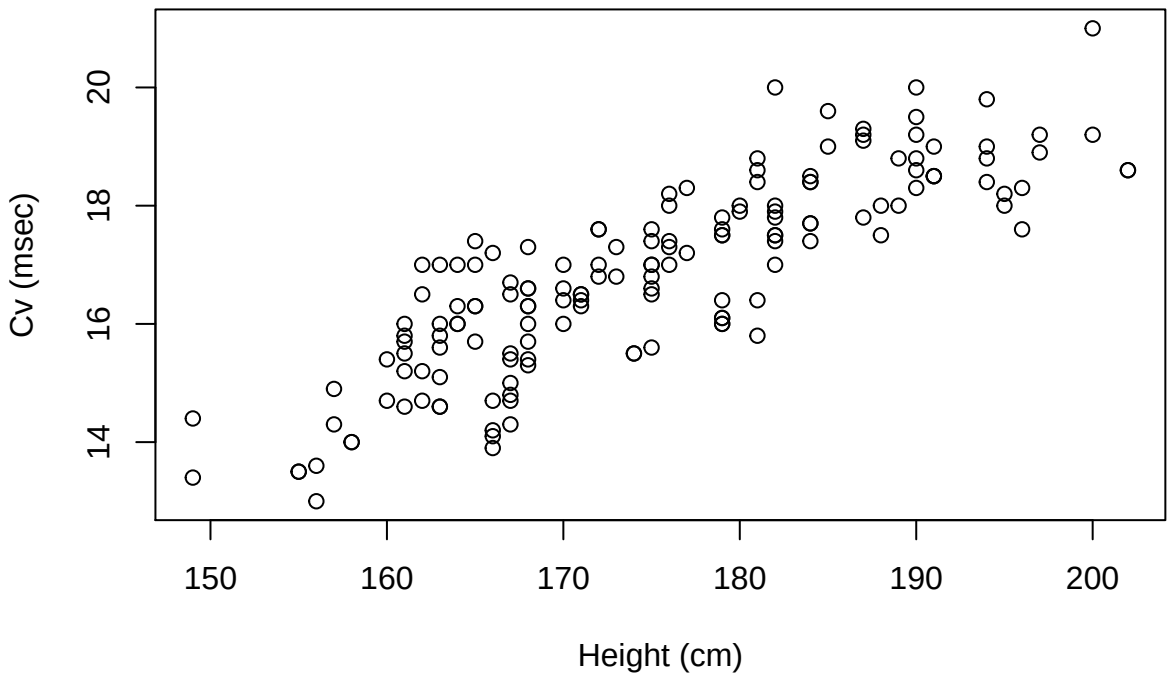
```
head(KWAST.FULL)
```

```
##   HEIGHT.FULL CV.FULL
## 1         149    14.4
## 2         149    13.4
## 3         155    13.5
## 4         155    13.5
## 5         156    13.0
## 6         156    13.6
```

We visualise the data using a scatterplot.

```
plot(HEIGHT.FULL, CV.FULL,
     main = "Figure 4. Peak spinal latency in digit I stimulation \nand height from 155 subjects in Kwas
     ylab = "Cv (msec)",
     xlab = "Height (cm)",
     cex.main = 0.9)
```

Figure 4. Peak spinal latency in digit I stimulation and height from 155 subjects in Kwast−Rabben et al.



Formally, we test the null hypothesis $H_0 : \rho = 0$ against the alternative hypothesis $H_A : \rho \neq 0$.

We can calculate r.

```
cor(HEIGHT.FULL, CV.FULL)
```

```
## [1] 0.8478829
```

We can use this value to calculate the test statistic.

$t = r\sqrt{(n-2)/(1-r^2)} = 0.8478829\sqrt{(155-2)/(1-0.8478829^2)} = 19.781.$

```
R.FULL <- cor(HEIGHT.FULL, CV.FULL); N.FULL <- 155
T.FULL <- R.FULL * sqrt((N.FULL-2)/(1-R.FULL^2))
T.FULL
```

```
## [1] 19.78133
```

Under $n - 2 = 155 - 2 = 153$ degrees of freedom, we can use this to arrive at a p-value.

```
dt(T.FULL, N.FULL-2)
```

```
## [1] 1.451553e-43
```

The p-value is <0.001.

The calculation of 95% confidence intervals is a multi-step process and will be described in later lectures.

Fortunately, we don't have to accomplish all of these by hand. In R, the function `cor.test()` will estimate the correlation coefficient, test the null hypothesis $H_0 : \rho = 0$ and provide confidence intervals, all within a single output.

```r
cor.test(HEIGHT.FULL, CV.FULL)
```

```
##
##  Pearson's product-moment correlation
##
## data:  HEIGHT.FULL and CV.FULL
## t = 19.781, df = 153, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.7967316 0.8869721
## sample estimates:
##       cor
## 0.8478829
```

References

[1] Kwast-Rabben O, Lileblius R, Heikkila H. Somatosensory evoked potentials following stimulation of digital nerves. *Muscle and Nerve* 2002;26:533-538.

THE END