Correlation *M&M §2.2*

References: A&B Ch 5,8,9,10; Colton Ch 6, M&M Chapter 2.2

Similarities between Correlation and Regression

- Both involve relationships between pair of numerical variables.
- Both: "predictability", "reduction in uncertainty"; "explanation".
- Both involve straight line relationships [can get fancier too].

Differences

Correlation	**Regression**
Symmetric	Directional
(doesn't matter which is on Y, which on X axis)	(matters which is on Y, which on X axis)
Chose n 'objects'; measure (X,Y) on each	(i) Choose n objects on basis of their X values; measure their Y; or
	(ii) Choose objects, (as with correlation); measure (X,Y)
	Regard X value as 'fixed'; .
	Can be extended to non-straight line relationships
	Can relate Y to multiple X variables.
Dimensionless (no units) (– 1 to + 1)	$\Delta Y/\Delta X$ units e.g., Kg/cm

Measures of Correlation

Loose Definition of Correlation:

Degree to which, in observed (x,y) pairs, y value tends to be larger than average when x is larger (smaller) than average; extent to which larger than average x's are associated with larger (smaller) than average y's

Pearson Product-Moment Correlation Coefficient

Context	**Symbol**	**Calculation**
sample of n pairs	r_{xy}	$\dfrac{\sum\{x_i - \bar{x}\}\{y_i - \bar{y}\}}{\sqrt{(\sum\{x_i - \bar{x}\}^2)(\sum\{y_i - \bar{y}\}^2)}}$
"universe" of all pairs	ρ_{xy}	$\dfrac{E\{(X - \mu_X)(Y - \mu_Y)\}}{\sqrt{E\{(X - \mu_X)^2\}\ E\{(Y - \mu_Y)^2\}}}$

Notes**:**
- ◦ ρ: Greek letter r, pronounced 'rho' ;
- ◦ E : Expected value ;
- ◦ μ: Greek letter 'mu'; denotes mean in universe.
- ◦ Think of r as an average product of scaled deviations [M&M p127 use n-1 because the two SDs involved in creating Z scores implicitly involve $1/\sqrt{(n-1)}$; result is same as above]

Spearman's (Non-parametric) **Rank Correlation Coefficient**

x -> rank	replace x's by their ranks (1=smallest to n=largest)
y -> rank	replace y's by their ranks (1=smallest to n=largest)
THEN	calculate Pearson correlation for n pairs of ranks

(see later)

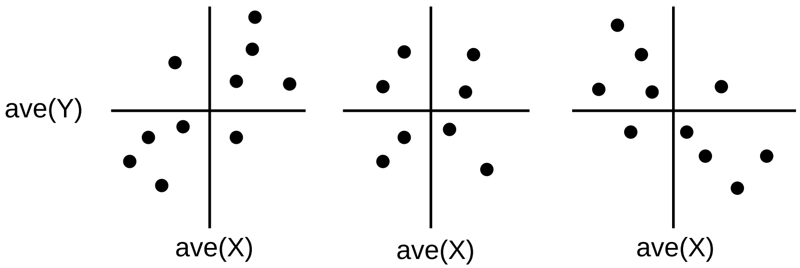
Correlation *M&M §2.2*

Correlation

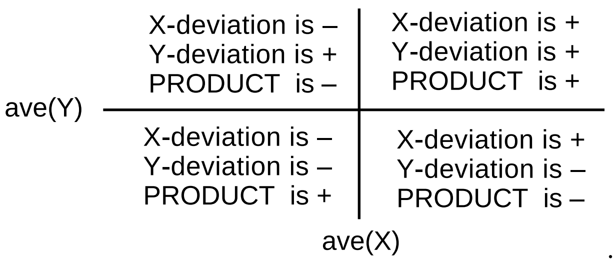
<u>Positive:</u> larger than ave. X's with larger than ave. Y's;
smaller than ave. X's with smaller than ave. Y's;

<u>Negative:</u> larger than ave. X's with smaller than ave. Y's;
smaller than ave. X's with larger than ave. Y's;

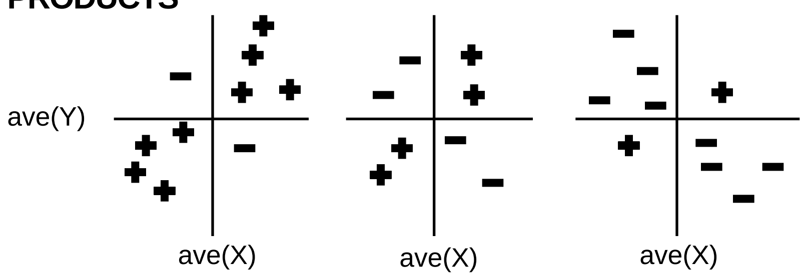
<u>None:</u> larger than ave. X's 'equally likely' to be coupled with larger as with smaller than ave. Y's



How r ranges from -1 (negative correlation) through 0 (zero correlation.) through +1 (positive correlation.) (r not tied to x or y scale)



PRODUCTS



ρ^2 is a measure of how much the variance of Y is reduced by knowing what the value of X is (or vice versa)

See article by Chatillon on "Balloon Rule" for visually estimating r. (cf. Resources for Session 1, course 678 web page)

$$\text{Var}(\,Y \mid X\,) = \text{Var}(\,Y\,) \times (\,1 - \rho^2\,)$$

ρ^2 called *"coefficient of determination"*

$$\text{Var}(\,X \mid Y\,) = \text{Var}(\,X\,) \times (\,1 - \rho^2\,)$$

Large ρ^2 (i.e. ρ close -1 or +1) - > close linear association of X and Y values; far less uncertain about value of one variable if told value of other.

If X and Y scores are standardized to have mean=0 and unit SD=1 it can be seen that ρ is like a "rate of exchange" ie the value of a standard deviation's worth of X in terms of PREDICTED standard deviation units of Y.

If we know observation is Z_X SD's from μ_X, then the least squares prediction of observation's Z_Y value (ie relative to μ_Y) is given by

$$\text{predicted } Z_Y = \rho \cdot Z_X$$

Notice the <u>regression towards mean</u>: ρ is always less than 1 in absolute value, and so the predicted Z_Y is closer to 0 (or equivalently make Y closer to μ_Y) than the Z_X was to 0 (or X was to μ_X).

Inferences re ρ *[based on sample of n (x,y) pairs]*

Naturally, the observed r in any particular sample will not exactly match the ρ in the population (i.e. the coefficient one would get if one included everybody). The quantity r varies from one possible sample of n to another possible sample of n. i.e. r is subject to sampling fluctuations about ρ.

1 A question all too often asked of one's data is whether there is evidence of a non-zero correlation between 2 variables. To test this, one sets up the null hypothesis that ρ is zero and determines the probability, calculated under this null hypothesis that $\rho = \mathbf{0}$, of obtaining an r more extreme than we observed. If the null hypothesis is true, r would just be "randomly different" from zero, with the amount of the random variation governed by n.

This discrepancy of r from **0** can be measured as $\frac{r\sqrt{n-2}}{\sqrt{1-r^2}}$ and should, if the null hypothesis of $\rho = \mathbf{0}$ is true, follow a t distribution with n-2 df.

[Colton's table A5 gives the smallest r which would be considered evidence that $\rho \neq \mathbf{0}$. For example, if n=20, so that df = 18, an observed correlation of 0.44 or higher, or between -0.44 and -1 would be considered statistically significant at the P=0.05 level (2-sided). **NB**: this t-test assumes that the pairs are from a Bivariate Normal distribution. ***Also, it is valid only for testing*** $\rho = 0$***, not for testing any other value of*** ρ.

JH has seen many the researcher scan a matrix of correlations, highlighting those with a small p-value and hoping to make something of them. But very often, that ρ *was non-zero was never in doubt; the more important question is how non-zero the underlying* ρ *really was. A small p-value (from maybe a feeble r but a large n!) should not be taken as evidence of an important* ρ*! JH has also observed several disappointed researchers who mistakenly see the small p-values and think they are the correlations! (the p-values associated with the test of* $\rho = 0$ *are often printed under the correlations)*

Interesting example where r ≠ 0, and not by chance alone!

1970 U.S. DRAFT LOTTERY during Vietnam War: See Moore and McCabe pp113-114, along with spreadsheet under Resources for Chapter 10, where the lottery is simulated using random numbers (Monte Carlo method)

2 Other common questions: given that r is based only on a sample, what interval should I put around r so it can be used as a (say 95%) confidence interval for the "true" coefficient ρ ?

Or (answerable by the same technique): one observes a certain r_1 ; in another population, one observes a value r_2 . Is there evidence that the ρ's in the 2 populations we are studying are unequal?

From our experience with the binomial statistic, which is limited to {0,n} or {0,1}, it is no surprise that the r statistic, limited as it is to {minus 1, plus 1}, also has a pattern of sampling variation that is not symmetric unless ρ is right in the middle, i.e. unless $\rho = 0$. The following transformation of r will lead to a statistic which is approximately normal even if the ρ('s) in the population(s) we are studying is(are) quite distant from 0:

$$\frac{1}{2}\ ln\left\{\frac{1+r}{1-r}\right\}$$ [where *ln* is log to the base e or natural log].

It is known as Fisher's transformation of r; the observed r, transformed to this new scale, should be compared against a Gaussian distribution with

$$\text{mean} = \frac{1}{2}\ ln\left\{\frac{1+\rho}{1-\rho}\right\} \text{ and SD} = \sqrt{\frac{1}{n-3}}.$$

Inferences re ρ *[continued...]*

***e.g. 2a: Testing* H_0: ρ = 0.5**

Observe r=0.4 in sample of n=20.

Compute $$\frac{\frac{1}{2}\ ln\{\frac{1+0.4}{1-0.4}\} - \frac{1}{2}\ ln\{\frac{1+0.5}{1-0.5}\}}{\sqrt{\frac{1}{n-3}}}.$$

and compare with Gaussian (0,1) tables. Extreme values of the standardized Z are taken as evidence against $\mathbf{H_0}$. Often, the alternative hypothesis concerning ρ is 1-sided, of the form ρ > **some quantity.**

e.g. 2b: *Testing* H_0: $\rho_1 = \rho_2$

r_1 & r_2 in independent samples of n_1 & n_2

Remembering that "variances add; SD's do not", compute the test statistic

$$\frac{\frac{1}{2}\ ln\{\frac{1+r_1}{1-r_1}\} - \frac{1}{2}\ ln\{\frac{1+r_2}{1-r_2}\} - [\ 0\ ]}{\sqrt{\frac{1}{n_1-3}+\frac{1}{n_2-3}}}.$$

and compare with Gaussian (0,1) tables.

e.g. 2c: 100(1–α)% CI for ρ from r=0.4 in sample of n=20.

By solving the double inequality

$$-z_{\alpha/2} \le \frac{\frac{1}{2}\ ln\{\frac{1+r}{1-r}\} - \frac{1}{2}\ ln\{\frac{1+\rho}{1-\rho}\}}{\sqrt{\frac{1}{n-3}}.} \le z_{\alpha/2}$$

so that the middle term is ρ , we can construct a CI for ρ**:**

$$\rho_{[High,\ Low]} = \frac{1+r-\{1-r\}\ e^{[\ \pm 2\ z_{\alpha/2}\ /\ Sqrt[n-3]\ ]}}{1+r+\{1-r\}\ e^{[\ \pm 2\ z_{\alpha/2}\ /\ Sqrt[n-3]\ ]}}$$

Worked e.g. 95% **CI(ρ)** based on r=0.55 in sample of n=12.

With α=0.05, $z_{\alpha/2}$ = 1.96, lower & upper bounds for ρ:

$$= \frac{1+0.55-\{1-0.55\}\ e^{[\ \pm 2 \bullet 1.96\ /\ \sqrt{9}\ ]}}{1+0.55+\{1-0.55\}\ e^{[\ \pm 2 \bullet 1.96\ /\ \sqrt{9}\ ]}}$$

$$= \frac{1.55-0.45\ e^{[\ \pm 2 \bullet 1.96\ /\ \sqrt{9}\ ]}}{1.55+0.45\ e^{[\ \pm 2 \bullet 1.96\ /\ \sqrt{9}\ ]}} = \frac{1.55-0.45\ e^{\pm 1.307}}{1.55+0.45\ e^{\pm 1.307}}$$

$$= \frac{1.55-0.45 \bullet 3.69}{1.55+0.45 \bullet 3.69},\ \frac{1.55-0.45\ /\ 3.69}{1.55+0.45\ /\ 3.69} = \mathbf{-0.04\ to\ 0.84}$$

This CI, which overlaps zero, agrees with the test of ρ =0 described above.

For if we evaluate $\frac{0.55\sqrt{12-2}}{\sqrt{1-0.55^2}}$, we get a value of 2.08, which is not as extreme as the tabulated $t_{10,0.05(2\text{-sided})}$ value of 2.23.

<u>Note</u>: There will be some slight discrepancies between the t-test of ρ =0 and the z-based CI's. The latter are only approximate. Note also that <u>both assume we have data which have a bivariate Gaussian distribution</u>.

(Partial) NOMOGRAM for 95% CI's for ρ

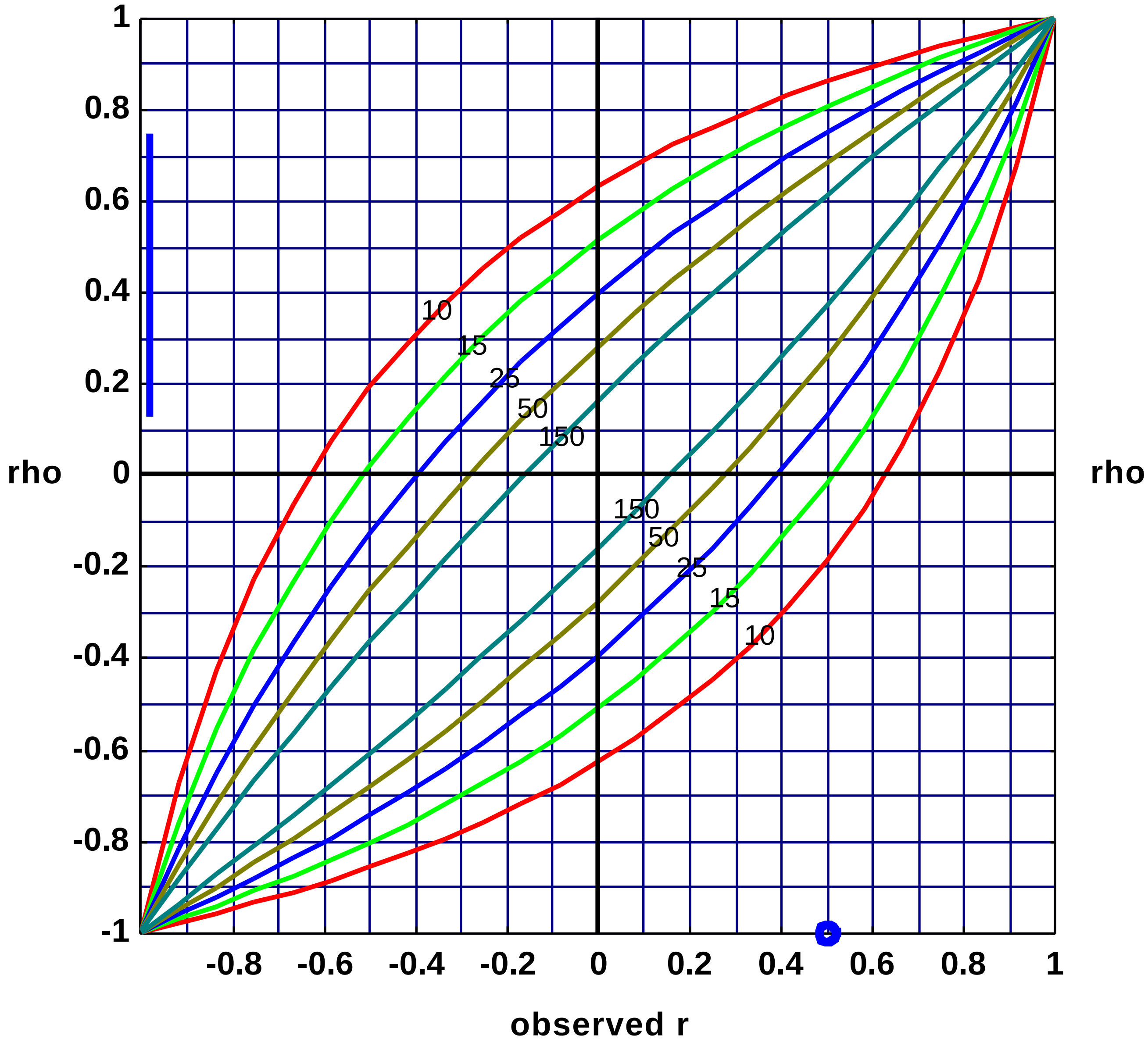
n = 10, 15, 25, 50, 150

It is based on Fisher's transformation of r. In addition to reading it vertically to get a CI for ρ (vertical axis) based on an observed r (horizontal axis), one can also use it to test whether an observed r is compatible with, or significantly different at the α = 0.05 level, from some specific ρ value, ρ_0 say, on the vertical axis: simply read across from $\rho = \rho_0$ and see if the observed r falls within the horizontal range appropriate to the sample size involved. Note that this test of a nonzero ρ is not possible via the t-test. Books of statistical tables have fuller nomograms.

Shown: CI if observe r=0.5 (o) with n=25.

Could aldo use nomogram to gauge the approx. 95% limits of variation for the correlation in a draft lottery. The n=366 is a little more than 2.44 times the n=150 here. So the (horizontal) variations around ρ = 0 should be only $1/\sqrt{2.44}$ or 64% as wide as those shown here for n=150. Thus the 95% range of r would be approx. -0.1 to +0.1. (since X and Y are uniform, rather than Gaussian, theory may be a little "off"). Observed r was -0.23.

{1+r-(1-r)Exp[±2z/Sqrt[n-3]]} / {1+r+(1-r)Exp[±2z/Sqrt[n-3]]}



Correlation *M&M §2.2*

Spearman's (Non-parametric) **Rank Correlation Coefficient**

How Calculated:

(i) replace x's and y's by their ranks (1=smallest to n=largest)

(ii) calculate Pearson correlation using the pairs of ranks.

Advantages

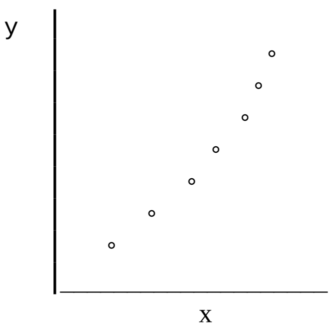
- Easy to do manually (if ranking not a chore);

  $$r_{Spearman} = 1 - \frac{6\sum d^2}{n\{n^2-1\}}$$

  { d = Δ between "X rank" & "Y rank" for each observation}

- Less sensitive to outliers (x -> rank
  ==> variance fixed (for a given n).
  Extreme $\{x_i - \bar{x}\}$ or $\{y_i - \bar{y}\}$ can exert considerable influence on $r_{Pearson}$.

- Picks up on non-linear patterns e.g. the $r_{Spearman}$ for the following data is 1, whereas the $r_{Pearson}$. is less.



Correlations -- obscured and artifactual

(i) Diluted / attenuated

(ii) Artifact

Examples:

(i) Diluted / attenuated / obscured

1 Relationship, in McGill Engineering students, between their first year university grades and their CEGEP grades

2 Relationship between heights of offspring and heights of their parents

X = average height of 2 parents
Y = height of offspring (ignore sex of offspring)

Galton's solution

'transmute' female heights to male heights

'transmuted' height = height × 1.08

(ii) Artifact / artificially induced

1. Blood Pressure of unrelated (male, female) 'couples'