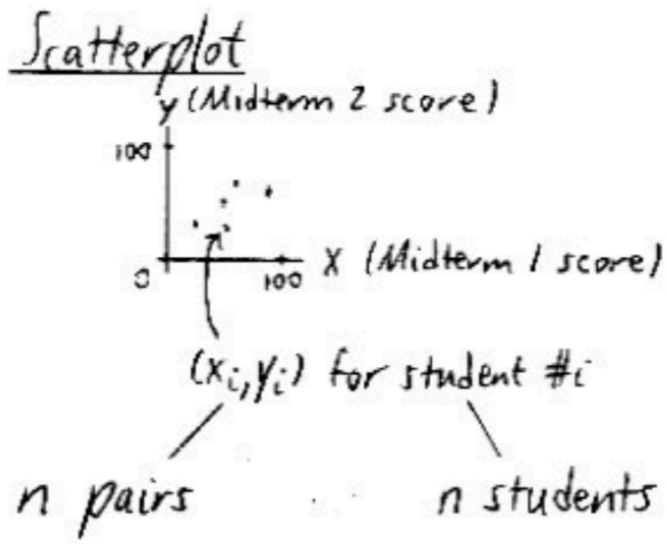
CORRELATION and REGRESSION

LESSON 43: CORRELATION and REGRESSION

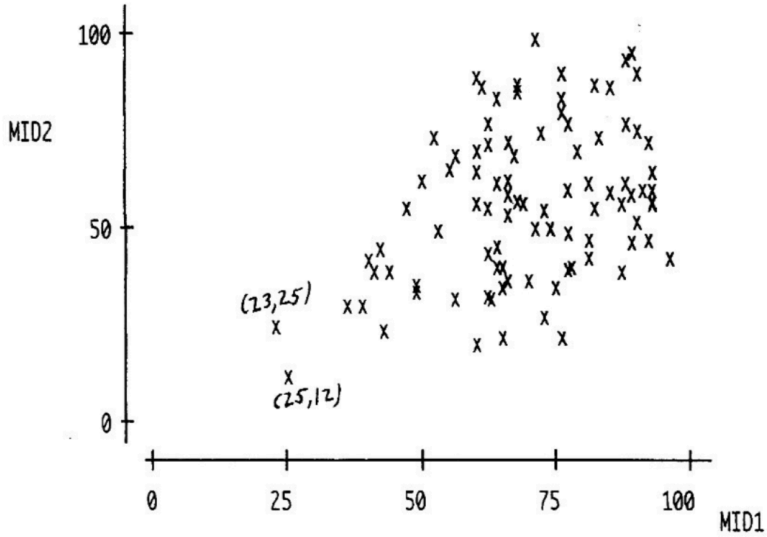
What is the Relationship Between Two Quantitative Variables? How Strong is It?

PART A: SCATTERPLOTS

A scatterplot is used to graph paired data, which is bivariate (involving two variables).



Below is a **scatterplot** where Midterm 2 scores are plotted against Midterm 1 scores in a calculus class using an old version of MINITAB.



PART B: CORRELATION

ρ (the Greek letter "rho") is the linear correlation coefficient for a population.

r is the linear correlation coefficient for a sample.

- These measure the **direction** and **strength** of the **linear relationship** between two variables.
- They were developed by Karl Pearson, for whom they are sometimes named.

The properties below apply to both ρ and r.

Properties of r

Assume that we have paired data of the form (x_i, y_i).

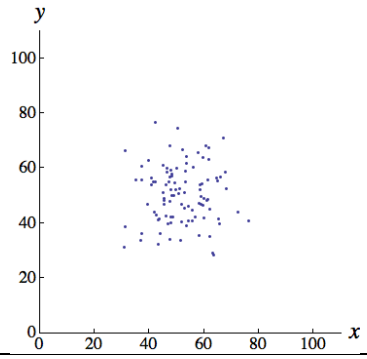
- $-1 \le r \le 1$
- $|r|$ measures the **strength of the linear relationship** between x and y. The higher $|r|$ is, the stronger it is.
- The **sign** of r measures the **direction of the linear relationship** between x and y.
  - § If $r > 0$, then y tends to **increase** as x increases. Think: $\nearrow$
  - § If $r < 0$, then y tends to **decrease** as x increases. Think: $\searrow$
  - § If $r = 0$, then there is **no linear relationship** between x and y.

We can test H_0: $\rho = 0$.

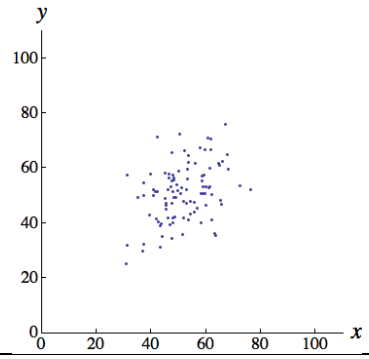
LINEAR CORRELATION COEFFICIENTS: r and ρ

In these plots, X and Y are normally distributed random variables with mean 50 and standard deviation 10. r is the sample linear correlation coefficient. (If you are considering population data, use ρ, the population linear correlation coefficient.)

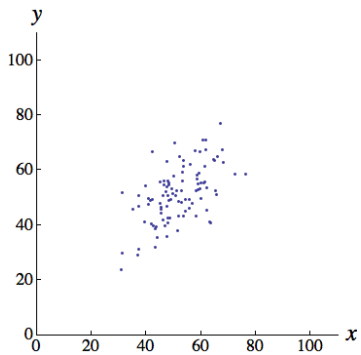
$r = 0.00$



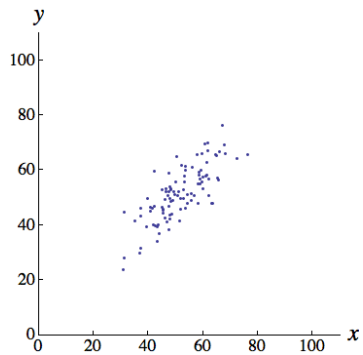
$r = 0.40$



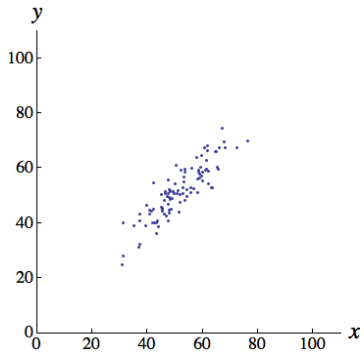
$r = 0.60$



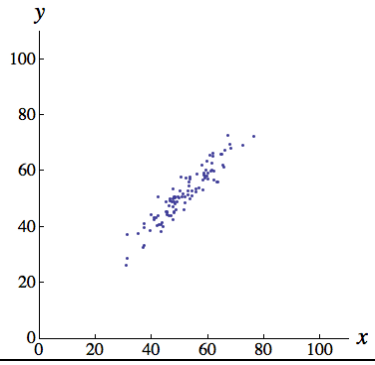
$r = 0.80$



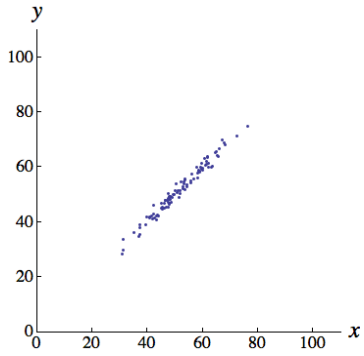
$r = 0.90$



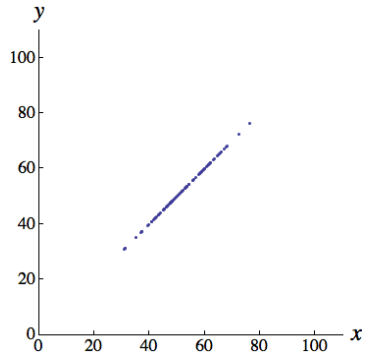
$r = 0.95$



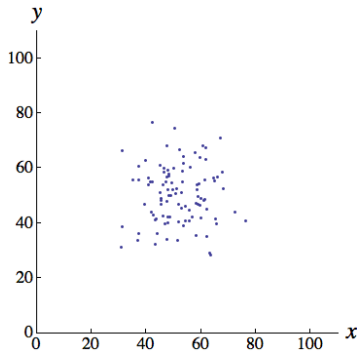
$r = 0.99$



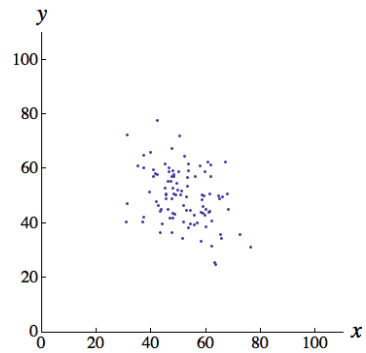
$r = 1.00$



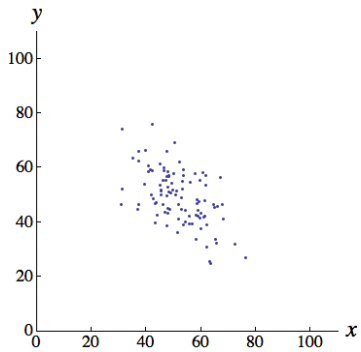
$r = 0.00$



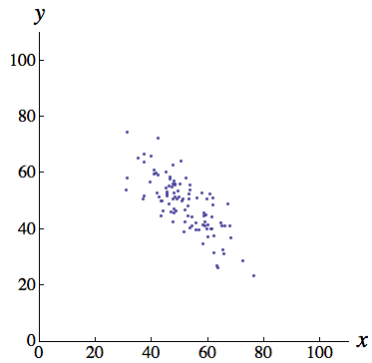
$r = -0.40$



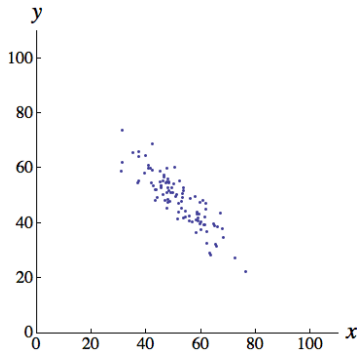
$r = -0.60$



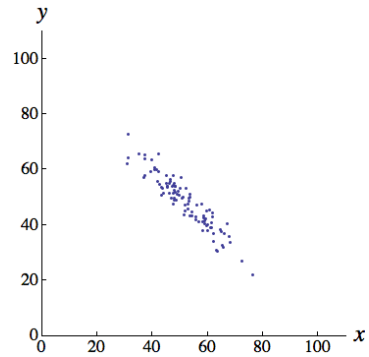
$r = -0.80$



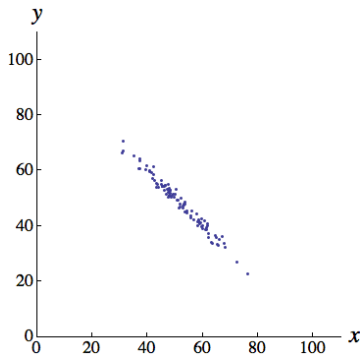
$r = -0.90$



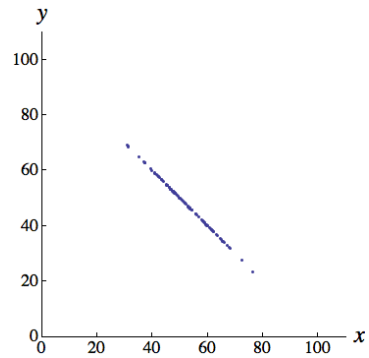
$r = -0.95$



$r = -0.99$

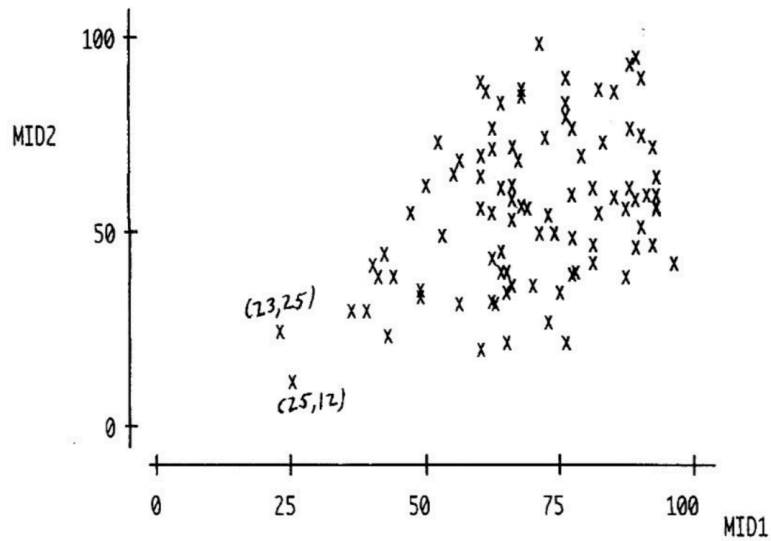


$r = -1.00$



Graphs produced using *Mathematica*.

In the scatterplot of calculus scores below, $\rho = 0.4$, which was low because the Midterm 2 material was very different from the Midterm 1 material!



Warnings

• **Correlation does not imply causality.** We earlier discussed how **statistical dependence** does not imply causality in Lesson 41. Now, consider a scatterplot of test scores vs. number of pets, with income level as a potential **confounding variable**.

• *r* might not pick up a strong **nonlinear** relationship between two variables. For example, *r* could be close to 0 for the (smoothed-out) scatterplot below; see also https://en.wikipedia.org/wiki/Correlation_and_dependence#/media/File:Correlation_examples2.svg

• **Averaging tends to inflate** $|r|$. For example, *r* is likely to be higher if $\left(\text{average Midterm 1 score, average Midterm 2 score}\right)$ is plotted for dorms of students instead of individual students.

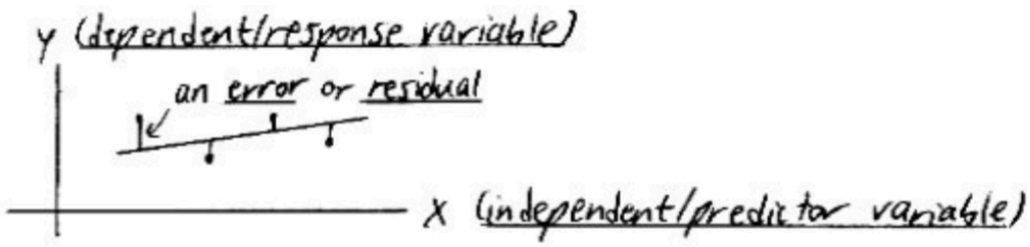
• **Use SDs to avoid distortions.** To help standardize the visual sense of correlation, it is helpful to make one SD on the *y*-axis the same physical length as one SD on the *x*-axis.

PART C: LINEAR REGRESSION

If there is a **strong enough linear correlation** between x and y (maybe test $H_1: \rho \neq 0$), then we may want to find the least squares regression line.

• See:
https://en.wikipedia.org/wiki/Correlation_and_dependence#/media/File:Pearson_Correlation_Coefficient_and_associated_scatterplots.png

• This is the line that "best fits" a scatterplot in the sense that the sum of squared errors (or residuals) is **minimized**.



See:
https://en.wikipedia.org/wiki/Linear_regression#/media/File:Linear_least_squares_example2.png
and
https://www.statisticshowto.datasciencecentral.com/wp-content/uploads/2014/11/least-squares-regression-line.jpg

• Remember that Slope-Intercept Form for a line is given by: $y = mx + b$. Consider rewriting this as $y = b + mx$, where m is the slope.

• The "true" best line for **population data** is denoted by: $y = \beta_0 + \beta_1 x$
β_0 and β_1 are **population parameters**.

• The best line for **sample data** is denoted by: $\hat{y} = b_0 + b_1 x$
b_0 and b_1 are **sample statistics** computed using data and formulas.

• This analysis is **sensitive to outliers**.

• The same regression line may fit very different-looking scatterplots!
https://en.wikipedia.org/wiki/Linear_regression#/media/File:Anscombe's_quartet_3.svg

PART D: INTERPRETING SLOPE AS MARGINAL CHANGE

Example 1 (Interpreting Slope as Marginal Change)

Let's say we have paired sample data for midterm scores in a class.

Let x = Midterm 1 score.
Let y = Midterm 2 score.

Let's say the **least squares regression line** for the sample is given by:

$$\hat{y} = 10 + 1.2x \text{ (in points)}$$

where $\hat{y}$ is the predicted Midterm 2 score based on a Midterm 1 score of x.

According to this regression line, for every **1-point increase** in the Midterm 1 score (x), there is a **1.2-point increase** in the Midterm 2 score (y).

For instance, a student who scores 20 points on Midterm 1 would be predicted to get $\hat{y} = 10 + 1.2(20) = 34$ points on Midterm 2, while a student who scores 21 points on Midterm 1 would be predicted to get $\hat{y} = 10 + 1.2(21)$, or $34 + 1.2 = 35.2$ points on Midterm 2.

§

PART E: REGRESSION TO THE MEAN

In terms of SDs, a person should expect to score closer to the mean on Midterm 2 than he/she did on Midterm 1. For example:

Midterm 1	Midterm 2
$\mu_1 = 50$	$\mu_2 = 50$
$\sigma_1 = 10$	$\sigma_2 = 10$
$\times 70$	$\times 60$
50 —	50 —

The idea is that, if a student does extremely well on Midterm 1, then it is likely that "errors" such as luck worked in his/her favor. It is also expected that the student will not be as lucky on Midterm 2.

People who don't understand this may **overestimate** the corrective value of punishing poor performance and also **overestimate** the detriment caused by rewarding superior performance. People may simply be witnessing regression to the mean.

The "Sports Illustrated" jinx, which claims that athletes who appear on the cover of the magazine one year then tend to decline in their performance the following year, is simply **regression to the mean**.

PART F : r^2, THE COEFFICIENT OF DETERMINATION

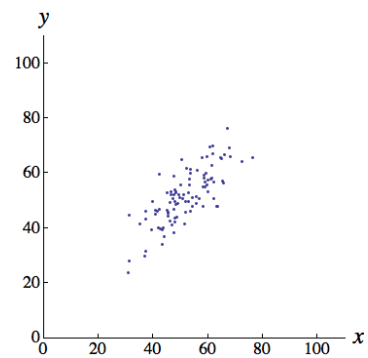
In percent form, r^2, the coefficient of determination, gives us the percent of the **variance in *y*** that is **accounted for** by *x* and the regression line.

Example 2 (Coefficient of Determination)

Let's say we have paired sample data for midterm scores in a class.

Let *x* = Midterm 1 score.
Let *y* = Midterm 2 score.

In the scatterplot below, $r = 0.80$.



$r^2 = (0.80)^2 = 0.64$, so 64% of the **variance** in the Midterm 2 scores is **accounted for** by the Midterm 1 scores and the regression line.

§

If $r = 1$, then $r^2 = 1$ and 100% of the **variance** in the Midterm 2 scores would be **accounted for** by the Midterm 1 scores and the regression line.