Math 316

Section 12.1 Introduction

See **Book Table 12.1.1, Figure 12.1.1** on page 512. Our goal is to come up with something like **Figure 12.4.3** on page 542. We're interested in cause and effect, in this case, how the dosage of amphetamine affects food consumption of rats. In this example, there are three distinct dosages, with several observations per dosage.

A second version of this type of problem is in **Figure 12.1.2** on page 513. In this example there is a continuous range of temperature (the cause) and a continuous range in dissolved oxygen (the effect). We want to find a linear relationship (a straight line) as shown in **Figure 12.3.8** on page 533.

Section 12.2 The Correlation Coefficient

So we want to find straight lines that fit some data. Let's look at a simpler example, which includes less data: **Table 12.2.1** and **Figure 12.2.1** on page 513. In this section we look at the correlation coefficient r which measures how well a straight line would fit the data. In Section 12.3 we find that line, and later we discuss what that line would be useful for.

From the book:

The correlation coefficient, r

$$r = \frac{1}{n-1}\sum_{i=1}^{n}\left(\frac{x-\bar{x}}{s_x}\right)\left(\frac{y-\bar{y}}{s_y}\right)$$

Before we starting finding r values, here is a bit of information about what it means. The value of r is always between -1 and 1. In general:

r	Relationship between x and y	How well line fits data
> 0	Positive: As x goes up, so does y. As x goes down, so does y.	Larger $\lvert r \rvert$ means a better fit of the line to the data.
< 0	Negative: As x goes up, y goes down. As x goes down, y goes up.	
$= 1$	Positive	Line fits data perfectly.
$= -1$	Negative	
$= 0$	None	It doesn't.

See **Figure 12.2.3** on page 516 for some possibilities.

Notation in the above (at the bottom of the previous page) formula:

x_i is one of the X values.

y_i is the corresponding Y value.

s_x is the standard deviation of all of the X values.

s_y is the standard deviation of all of the Y values.

Let's look at this in **Table 12.2.2** on page 515, and then **Example 12.2.2** on page 516. Usually we don't have to do so much work—the value of r is usually just given to us. And **we could use Excel** to find r. So the question is whether there is a relationship between a snake's weight and length. (I would guess so.) If there were no linear relationship, then we would have $r = 0$. If there were a perfect linear relationship, then either it would be $r = 1$ (as weight increases, so does length) or $r = -1$ (as weight increases, length decreases, which would be strange). So the null and alternative hypotheses are:

H_0: $\rho = 0$ (there is not a significant correlation between X and Y)

H_A: $\rho \neq 0$ (there is a correlation between X and Y)

Wait, what is this ρ? Where did r go?

ρ is correlation coefficient for the entire *population*

r is correlation coefficient found for the *sample*, our best guess/estimate for ρ

So the closer r is to 0, the closer we expect ρ to be to 0, and thus the more likely that H_0 is actually true. That is, the farther ρ is from 0, the farther from 0 we expect ρ to be, and thus more like that H_A is true. So how far from 0 does r need to be for us to reject H_0 and accept H_A?

As before, we find a test statistic and a corresponding P value. So yet another test and table at the back at the book? Not exactly. The test statistic:

$$t_s = r\sqrt{\frac{n-2}{1-r^2}} = \sqrt{n-2}\sqrt{\frac{r^2}{1-r^2}} \quad \text{and} \quad df = n-2.$$

As usual, we see larger sample size n would result in larger t_s (and smaller P), and thus make it more likely we would accept H_A and conclude that there is a relationship between the x values and the y values.

Some values of r:

If r is	then r^2 is	and $\sqrt{r^2 \,/\, 1-r^2}$ is
0.0	0	0
± 0.2	0.04	0.2041
± 0.4	0.16	0.4364
± 0.6	0.36	0.7500
± 0.8	0.64	1.3300
± 0.9	0.81	2.0647
$\rightarrow \pm 1.0$	$\rightarrow 1.00$	$\rightarrow \infty$

Let's look at **Figure 12.2.4** on page 518 and **Example 12.2.4** on page 519.

See **Figure 12.2.5** on page 521 for possible effects of outliers on correlation coefficient.

Sometimes one variable is a function of the other (e.g. energy expenditure is a function of body weight), and sometimes there is simply a relationship between the two variables (e.g. length and weight are positively correlated). In finding r, it doesn't matter whether or not there is cause and effect, merely whether there is correlation.

Similar to in the past, rather than just testing whether $\rho = 0$ or not (based on how different ρ is from 0), we could also find a confidence interval for ρ. If the confidence interval contains 0, then we don't reject $H_0: \rho = 0$ and if the confidence interval does not contain 0, then we do reject $H_0: \rho = 0$ and accept $H_A: \rho \neq 0$. We'll not cover that, but you might want to take a quick look at the discussion in the book on pages 519 – 520.

If time, let's look at what r would be if the data were exactly on a line $y = mx + b$.

As a final note, notice the **alternative formula for r** at the bottom of page 515.