Correlation

IS381 - Statistics and Probability with R

Jason Bryer, Ph.D.

December 1, 2025

Correlation

Correlation is a measure of the relationship between two variables. The correlation can range from -1 indicating a "perfect" negative relationship to 1 indicating a "perfect" positive relationship. A correlation of 0 indicates no relationship.

Population Correlation

For a population, the correlation is defined as the ratio of the covariance to the product of the standard deviations, and is typically denoted using the Greek letter rho (ρ), is defined as:

$$\rho = \frac{cov(X, Y)}{\sigma_X \sigma_Y}$$

The standard deviation (σ) is equal to the square root of the variance ($\sigma = \sqrt{\frac{\Sigma(x_i - \bar{x})^2}{n-1}}$). What is new here is the covariance. Like variance, we are interested in deviations from the mean except now in two dimensions.

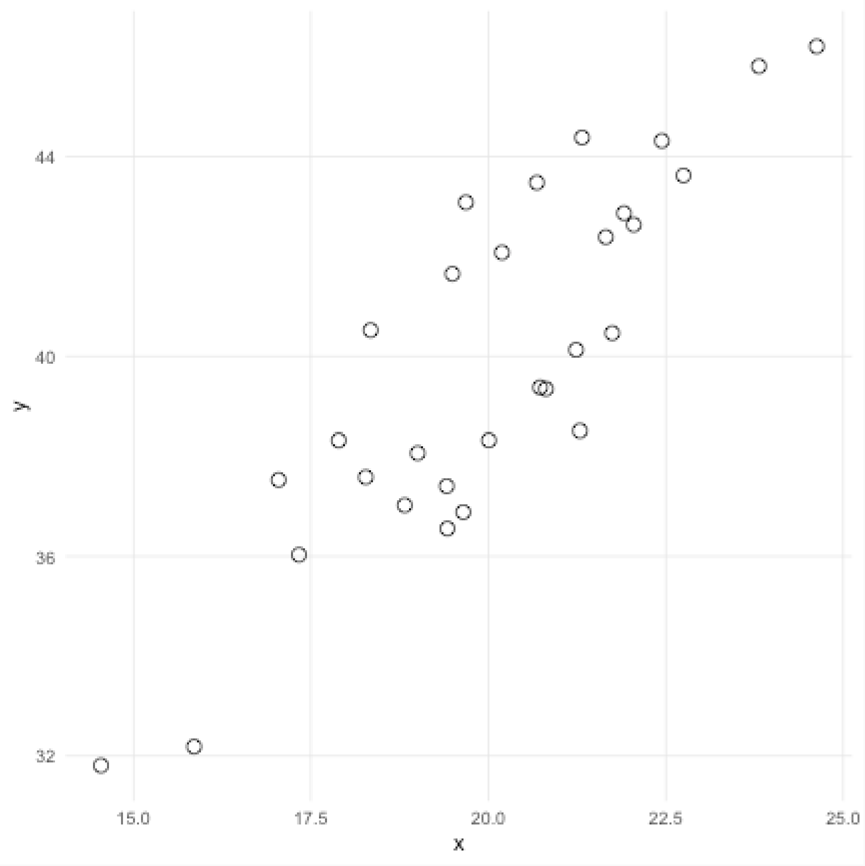
Covariance

The formula for the covariance is:

$$cov_{xy} = \frac{\Sigma(x_i - \bar{x})(y_i - \bar{y})}{n - 1}$$

Covariance (Simulated Example)

```r
mean_x <- 20
mean_y <- 40
sd_x <- 2
sd_y <- 3
n <- 30
rho <- 0.8
set.seed(2112)
df <- mvtnorm::rmvnorm(
    n = n,
    mean = c(mean_x, mean_y),
    sigma = matrix(
        c(sd_x^2, rho * (sd_x * sd_y),
          rho * (sd_x * sd_y), sd_y^2), 2, 2)) |>
    as.data.frame() |>
    dplyr::rename(x = V1, y = V2) |>
    dplyr::mutate(
        x_deviation = x - mean(x),
        y_deviation = y - mean(y),
        cross_product = x_deviation * y_deviation)
```

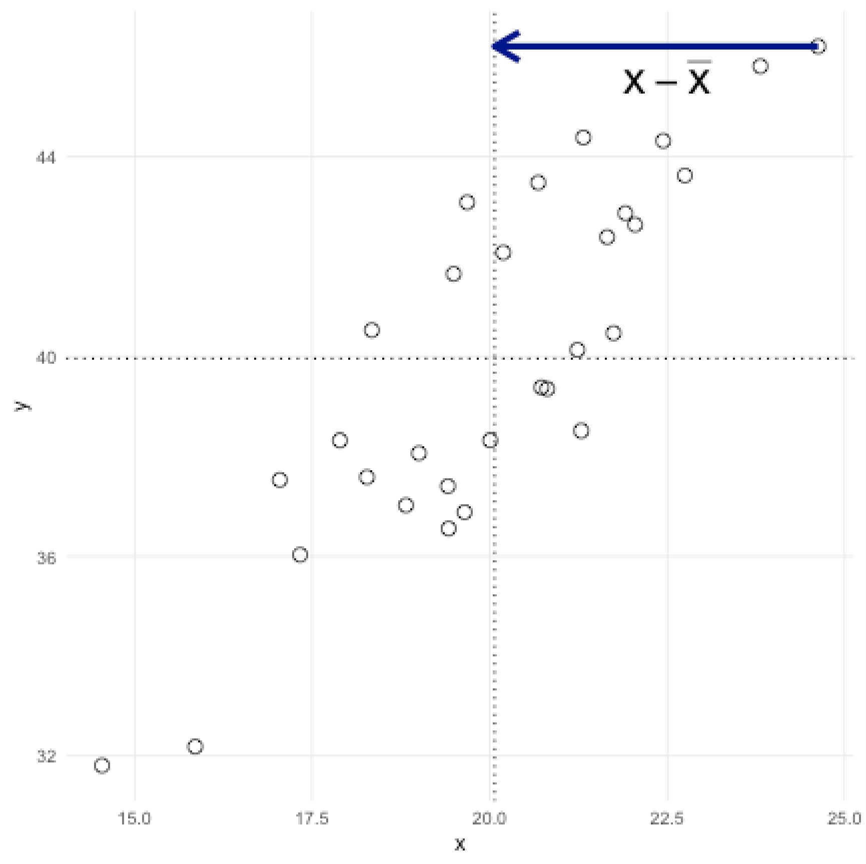


Covariance $(x - \bar{x})$

$$cov_{xy} = \frac{\Sigma(\mathbf{x_i} - \bar{\mathbf{x}})(y_i - \bar{y})}{n-1}$$

Consider the point with the largest *x* and *y*-value.

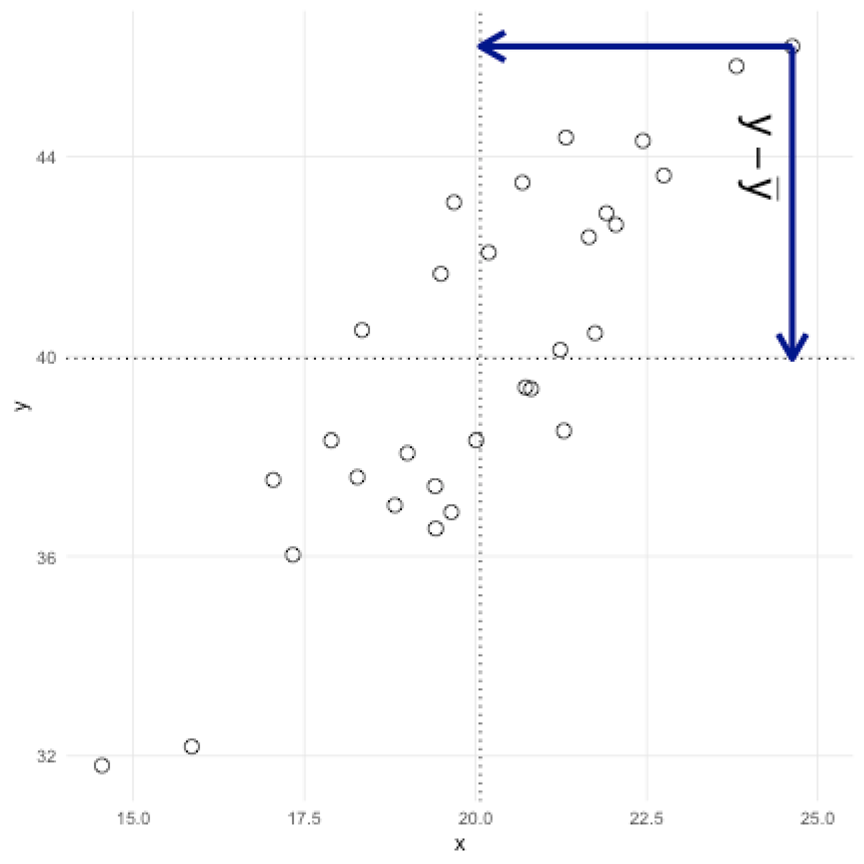
First step in the numerator is to subtract the mean of *x* ($\bar{x}$) from the *x*-value.



Covariance $(y - \bar{y})$

$$cov_{xy} = \frac{\Sigma(x_i - \bar{x})(\mathbf{y_i} - \bar{\mathbf{y}})}{n-1}$$

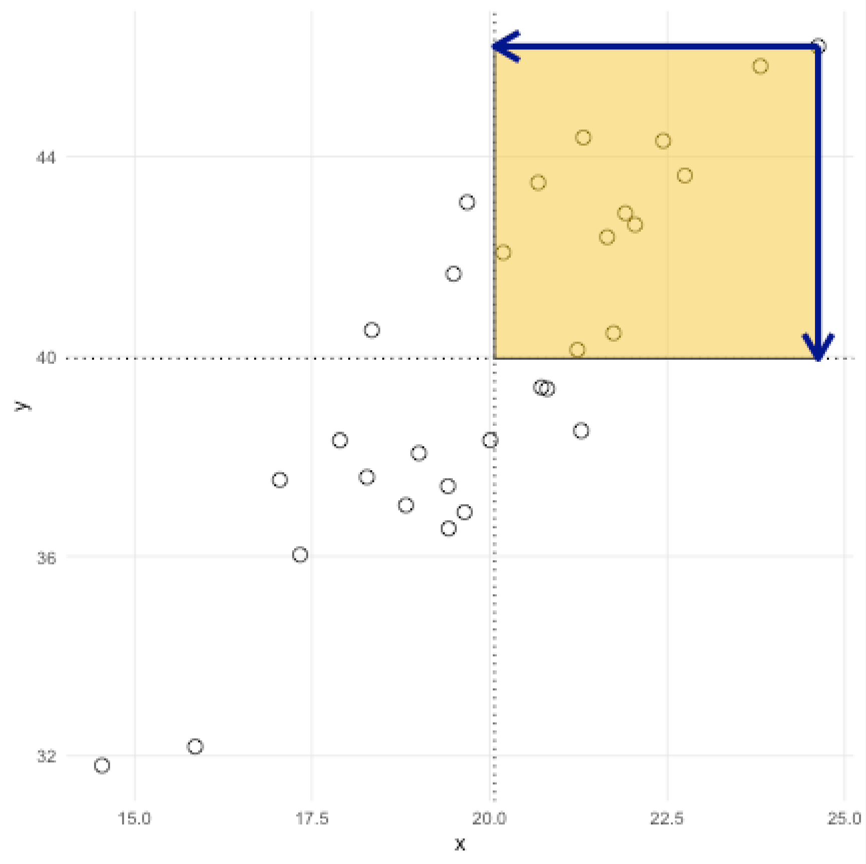
Second step in the numerator is to subtract the mean of *y* ($\bar{y}$) from the *y*-value.



Covariance

$$cov_{xy} = \frac{\Sigma(\mathbf{x_i} - \bar{\mathbf{x}})(\mathbf{y_i} - \bar{\mathbf{y}})}{n - 1}$$

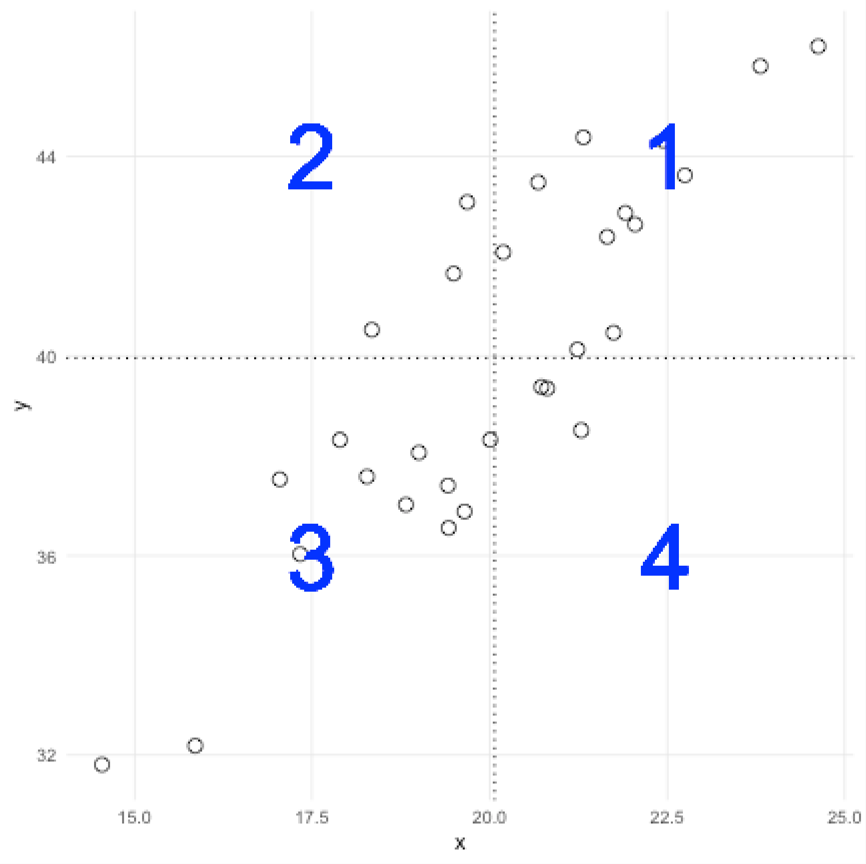
For the first observation, its contribution to the sum is simply the area of the rectangle. We call each of these areas the cross product (i.e. $x_i - \bar{x})(y_i - \bar{y})$).

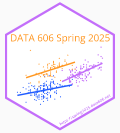


Covariance (quadrants)

$$cov_{xy} = \frac{\Sigma(\mathbf{x_i} - \bar{\mathbf{x}})(\mathbf{y_i} - \bar{\mathbf{y}})}{n - 1}$$

We can divide the plot into four quadrants split at $\bar{x}$ and $\bar{y}$.





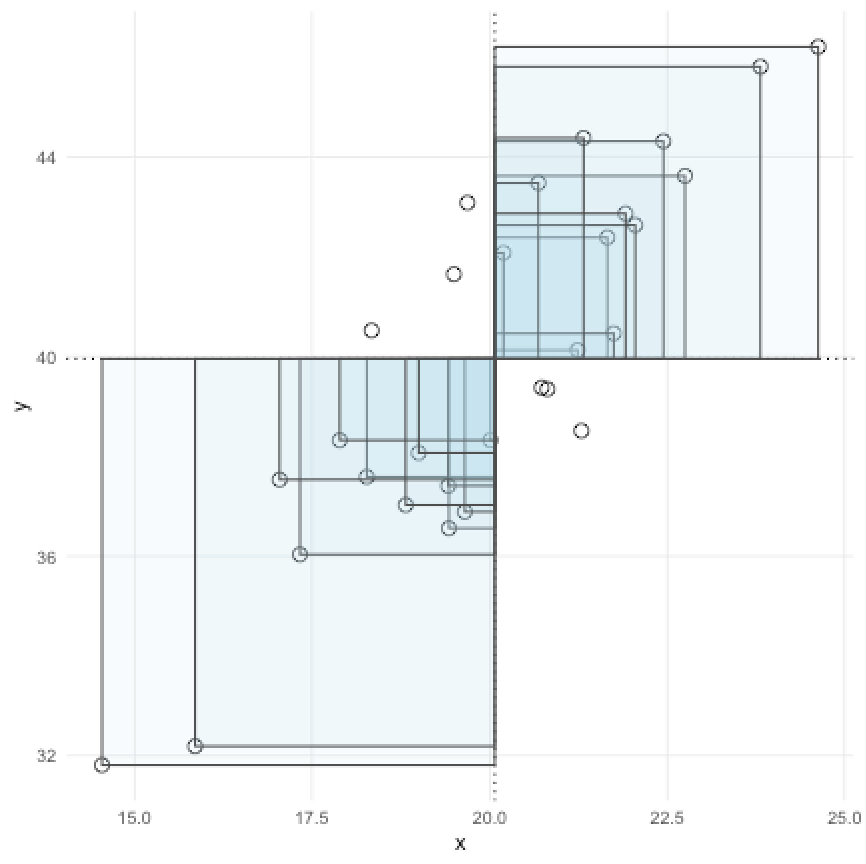
Covariance (positive cross products)

$$cov_{xy} = \frac{\Sigma(x_i - \bar{x})(y_i - \bar{y})}{n-1}$$

For observations in quadrant 1, $x - \bar{x}$ is **positive** and $y - \bar{y}$ is **positive** so the cross product is **positive.**

For observations in quadrant 3, $x - \bar{x}$ is **negative** and $y - \bar{y}$ is **negative** so the cross product is **positive.**

Hence, all observations in quadrants 1 and 3 contribute **positively** to the sum of cross products.



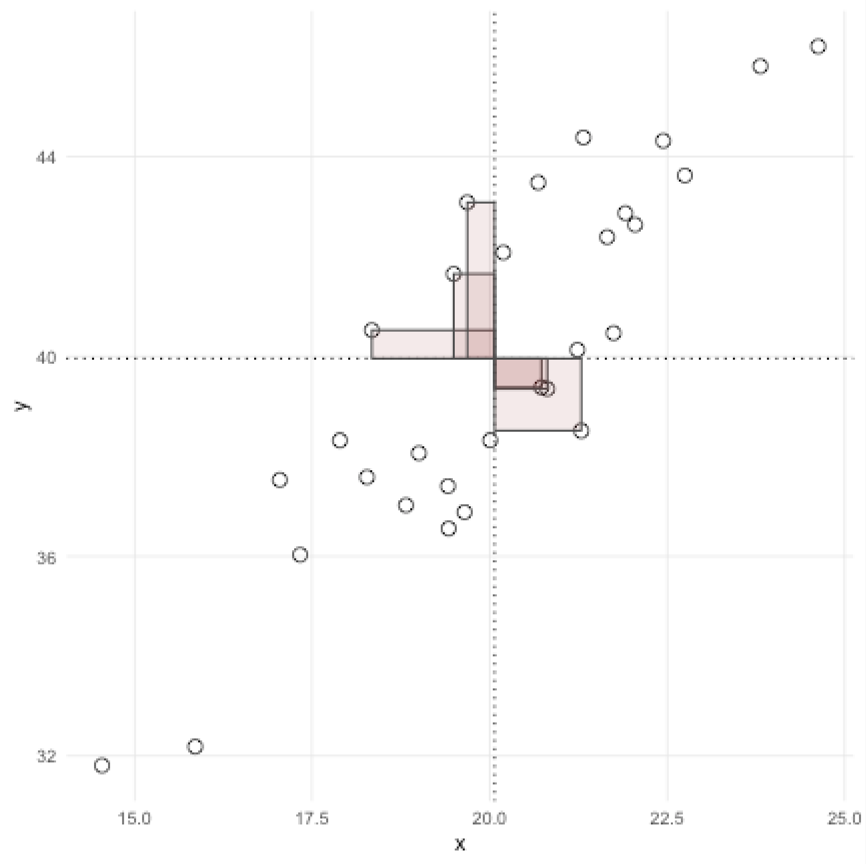
Covariance (negative cross products)

$$cov_{xy} = \frac{\Sigma(x_i - \bar{x})(y_i - \bar{y})}{n-1}$$

For observations in quadrant 2, $x - \bar{x}$ is **negative** and $y - \bar{y}$ is **positive** so the cross product is **negative**

For observations in quadrant 2, $x - \bar{x}$ is **positive** and $y - \bar{y}$ is **negative** so the cross product is **negative**

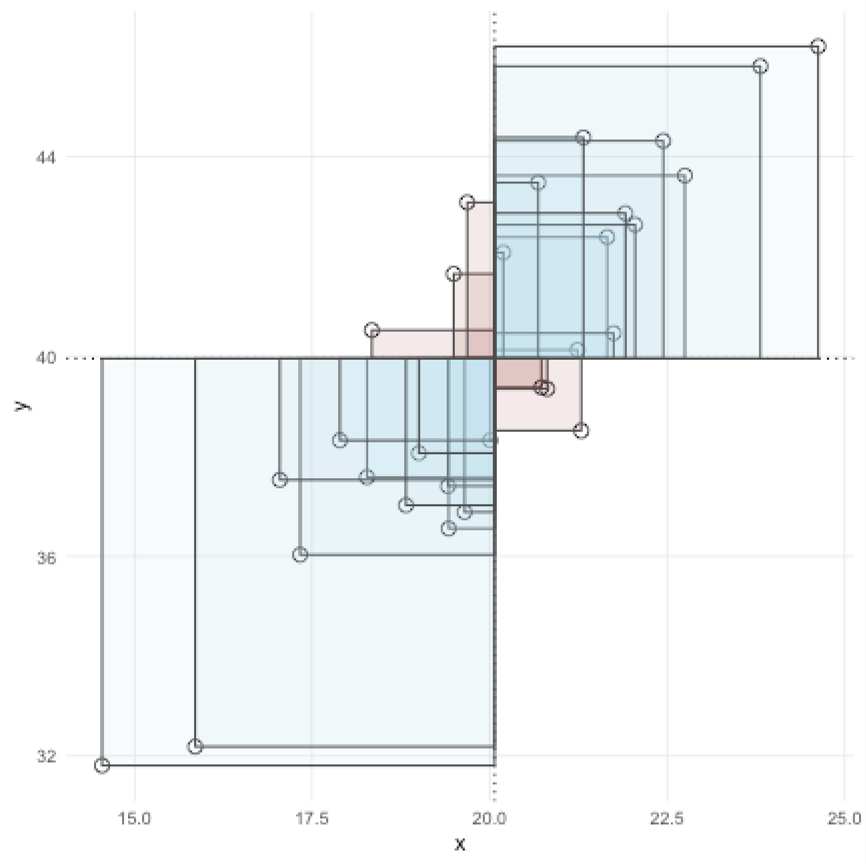
Hence, all observations in quadrants 2 and 4 contribute **negatively** to the sum of cross products.



Covariance

$$cov_{xy} = \frac{\Sigma(x_i - \bar{x})(y_i - \bar{y})}{n - 1}$$

The covariance is then the ratio of positive cross products to negative cross products.

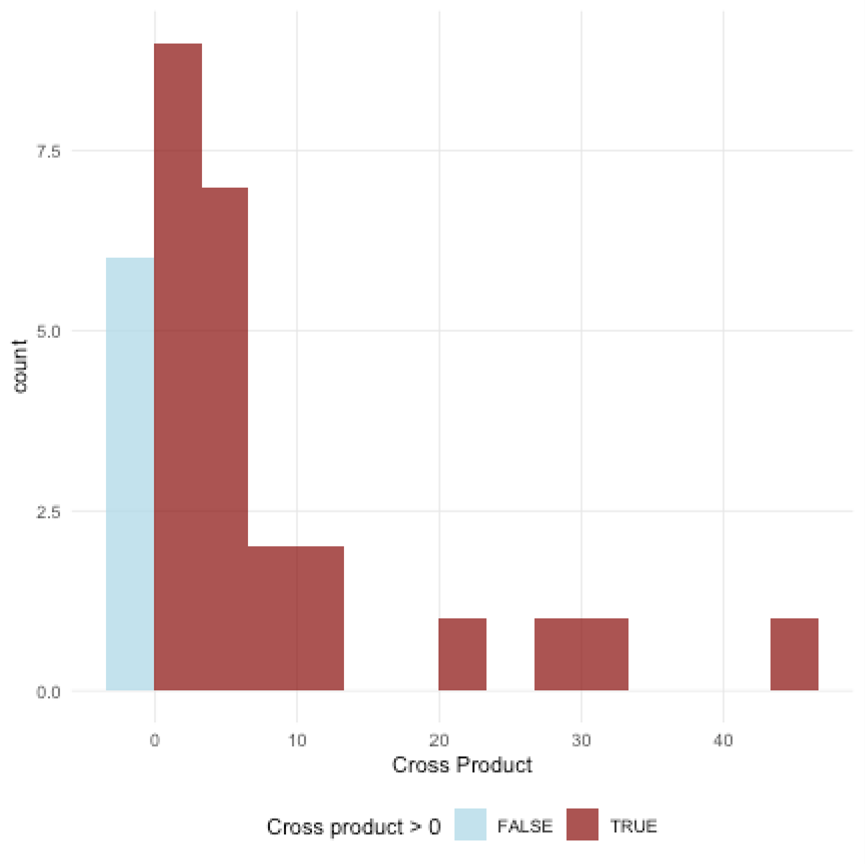


Covariance

$$cov_{xy} = \frac{\Sigma(x_i - \bar{x})(y_i - \bar{y})}{n - 1}$$

The covariance is then the ratio of positive cross products to negative cross products.

Which can be more easily seen by looking at a histogram of cross products.



Sample Correlation

Putting it all together we get...

$$r_{xy} = \frac{\frac{\sum_{i=1}^{n}\left(X_i - \overline{X}\right)\left(Y_i - \overline{Y}\right)}{n-1}}{s_x s_y}$$

Interestingly, if we have standardized scores (i.e. *z*-scores where mean = 0 and standard deviation = 1), we can simplify the correlation calculation...

$$r_{xy} = \frac{\frac{\sum_{i=1}^{n}\left(X_i - \overline{X}\right)\left(Y_i - \overline{Y}\right)}{n-1}}{s_x s_y} = \frac{\frac{\sum_{i=1}^{n}(X_i - 0)(Y_i - 0)}{n-1}}{1 \times 1} = \frac{\sum_{i=1}^{n} X_i Y_i}{n-1}$$

Try the following shiny application with various correlations.

```
VisualStats::regression_shiny()
```

Example: SAT Scores

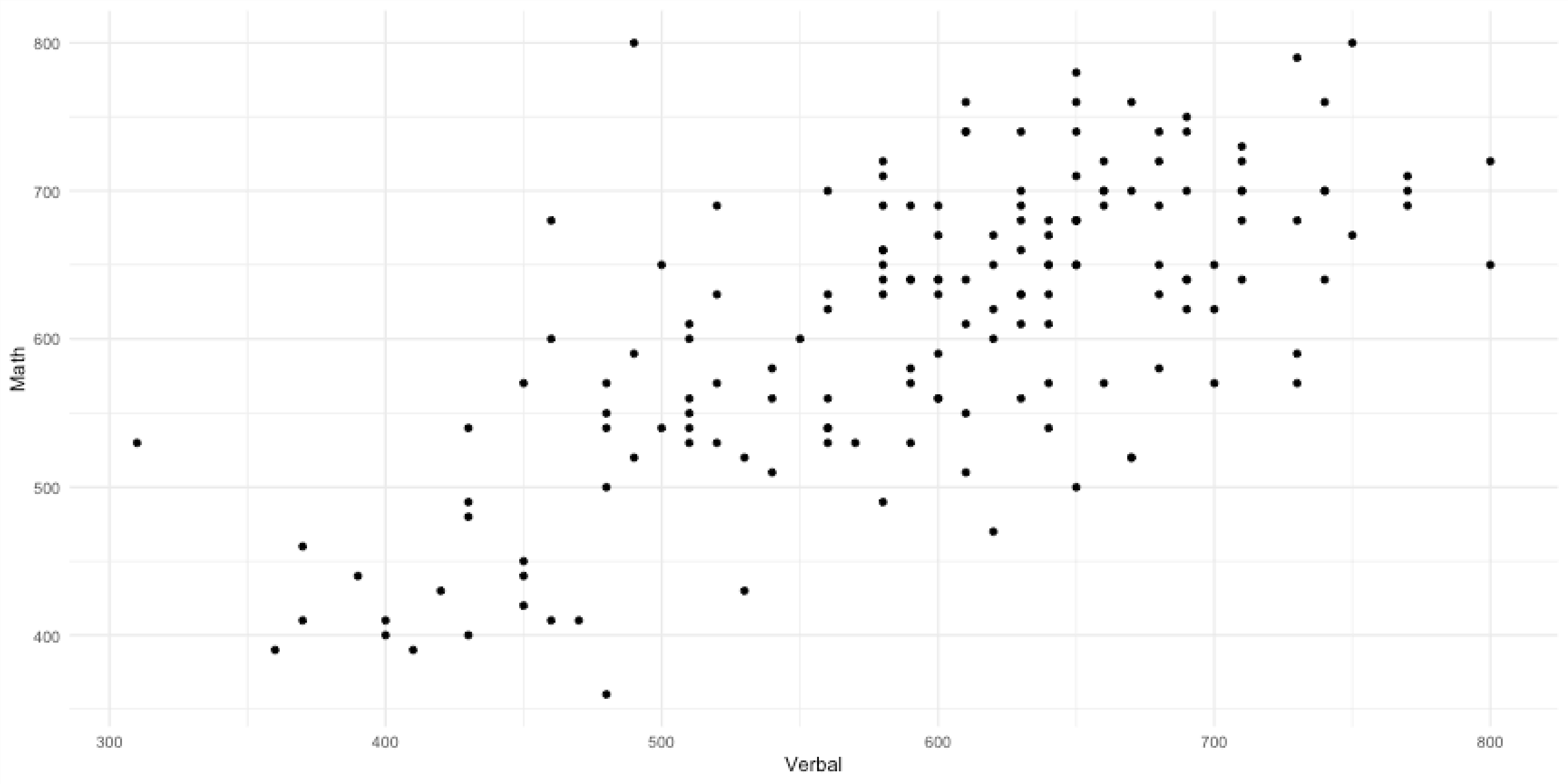
What is the correlation between SAT math and verbal scores?

To begin, we read in the CSV file and convert the `Verbal` and `Math` columns to integers. The data file uses `.` (i.e. a period) to denote missing values. The `as.integer` function will automatically convert those to `NA` (the indicator for a missing value in R). Finally, we use the `complete.cases` eliminate any rows with any missing values.

```
sat <- read.csv('SAT_scores.csv', stringsAsFactors=FALSE)
names(sat) <- c('Verbal','Math','Sex')
sat$Verbal <- as.integer(sat$Verbal)
sat$Math <- as.integer(sat$Math)
sat <- sat[complete.cases(sat),]
```

Scatter Plot

The first step is to draw a scatter plot. We see that the relationship appears to be fairly linear.



Descriptive Statistics

Next, we will calculate the means and standard deviations.

```
( verbalMean <- mean(sat$Verbal) )
```

```
## [1] 596.2963
```

```
( mathMean <- mean(sat$Math) )
```

```
## [1] 612.0988
```

```
( verbalSD <- sd(sat$Verbal) )
```

```
## [1] 99.5199
```

```
( mathSD <- sd(sat$Math) )
```

```
## [1] 98.13435
```

```
( n <- nrow(sat) )
```

```
## [1] 162
```

Covariance

The population correlation, rho, is defined as $\rho_{xy} = \frac{\sigma_{xy}}{\sigma_x \sigma_y}$ where the numerator is the *covariance* of *x* and *y* and the denominator is the product of the two standard deviations.

The sample correlation is calculated as $r_{xy} = \frac{Cov_{xy}}{s_x s_y}$

The covariates is calculated as $Cov_{xy} = \frac{\sum_{i=1}^{n} \left(X_i - \overline{X}\right)\left(Y_i - \overline{Y}\right)}{n-1}$

```
(cov.xy <- sum( (sat$Verbal - verbalMean) * (sat$Math - mathMean) ) / (n - 1))
```

```
## [1] 6686.082
```

Or we can use the built-in `cov` function.

```
cov(sat$Verbal, sat$Math)
```

```
## [1] 6686.082
```

Covariance (cont.)

$$r_{xy} = \frac{\frac{\sum_{i=1}^{n} \left(X_i - \overline{X}\right)\left(Y_i - \overline{Y}\right)}{n-1}}{s_x s_y}$$

```
cov.xy / (verbalSD * mathSD)
```

```
## [1] 0.6846061
```

Or we can use the built-in `cor` function.

```
cor(sat$Verbal, sat$Math)
```

```
## [1] 0.6846061
```

Using z-Scores

Calculate z-scores (standard scores) for the verbal and math scores.

$$z = \frac{y - \overline{y}}{s}$$

```
sat$Verbal.z <- (sat$Verbal - verbalMean) / verbalSD
sat$Math.z <- (sat$Math - mathMean) / mathSD
head(sat)
```

```
##   Verbal Math Sex    Verbal.z      Math.z
## 1    450  450   F -1.47002058 -1.65180456
## 2    640  540   F  0.43914539 -0.73469449
## 3    590  570   M -0.06326671 -0.42899113
## 4    400  400   M -1.97243268 -2.16131016
## 5    600  590   M  0.03721571 -0.22518889
## 6    610  610   M  0.13769813 -0.02138665
```

Correlation

Calculate the correlation manually using the z-score formula:

$$r = \frac{\sum z_x z_y}{n - 1}$$

```
r <- sum( sat$Verbal.z * sat$Math.z ) / ( n - 1 )
r
```

```
## [1] 0.6846061
```

We can see that this matches the correlation using the unstandardized values.

```
cor(sat$Verbal, sat$Math)
```

```
## [1] 0.6846061
```

And to show that the units don't matter, calculate the correlation with the z-scores.

```
cor(sat$Verbal.z, sat$Math.z)
```

```
## [1] 0.6846061
```

Is the correlation different than zero?

Just because we have a non-zero correlation does not necessarily mean the correlation is *statistically* different from zero. We can conduct a null hypothesis test where:

- H_0: The correlation is zero.
- H_A: The correlation is not equal to zero.

The `cor.test` function will perform that null hypothesis test providing both the *p*-value and confidence interval.

```
cor.test(sat$Verbal.z, sat$Math.z)
```

```
##
##     Pearson's product-moment correlation
##
## data:  sat$Verbal.z and sat$Math.z
## t = 11.88, df = 160, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to
## 95 percent confidence interval:
##  0.5930107 0.7587098
## sample estimates:
##       cor
## 0.6846061
```

The following Shiny application will allow for estimating the sampling distribution for varying correlations between -1 and 1. Be sure to note the relationship of sample size to the confidence interval, especially when the population correlation is zero.

```
shiny::runApp(paste0(find.package('VisualStats'), '/shiny/simulated_correlation'))
```

One Minute Paper

1. What was the most important thing you learned during this class?
2. What important question remains unanswered for you?

https://forms.gle/N8WjTAysfKbGLptLA