Chapter 8
Assessing Relationships with Correlation and Regression

Section 8.1
Correlation

David W. Gerbing

The School of Business
Portland State University

- Correlation
  - Relationships and the Scatter Plot
  - Correlation and Scatterplots
  - Inference for the Correlation Coefficient
  - Appendix: Basis of the Correlation Coefficient

8.1a
Relationships and the Scatter Plot

Relationship Between Variables

A relationship is positive or negative

- **Relationship** of two variables: As the values of one variable increase, the values of the other variable tend to either systematically increase, or systematically decrease
- **Positive relationship**: As values of one variable increase, the values of the other variable tend to increase
  - Food quality increases, customer satisfaction increases
  - Occupancy rate increases, needed staff increases
- **Negative (inverse) relationship**: As values of one variable increase, the values of the other variable tend to decrease
  - Price decreases, sales volume increases
  - Time brushing teeth increases, cavities decrease

The Scatterplot

Graphical representation of the scatterplot

- Unlike a categorical variable, a continuous variable has many possible numerical values, requiring a numerical axis to plot
- **Scatterplot**: Plot of the pairs of values for two different variables for each observation (e.g, people, companies), with one value scaled on the horizontal axis and the other value scaled on the vertical axis
- Each point on the scatterplot represents the values of the two variables for a single observation
  - Height and weight of one person
  - Gross and net income of one company
- For example, consider measurements of Height and Weight for 10 adult men, found at:

  `http://web.pdx.edu/~gerbing/data/bodyfat10.csv`

Scatterplot and Regression: Adult Height and Weight

Coordinates of one point, the data for one person:

The man from the data set with a Height of 69.50 inches and a Weight of 187.75 lbs

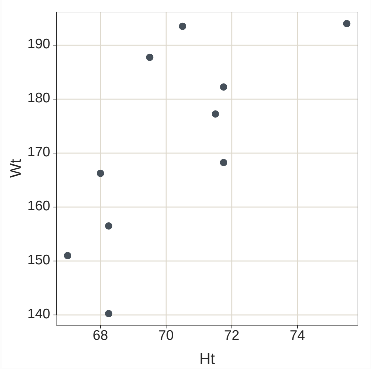
Scatterplot: Male Adult Height and Weight

Height (x) and Weight (y) for ten men

```
     Ht     Wt
1  71.75 182.25
2  71.75 168.25
3  75.50 194.00
4  68.25 140.25
5  68.25 156.50
6  69.50 187.75
7  70.50 193.50
8  71.50 177.25
9  67.00 151.00
10 68.00 166.25
```

- Use the `lessR` `Plot()` function or `sp()`, with two variables

```
> Plot(Ht,Wt)
```



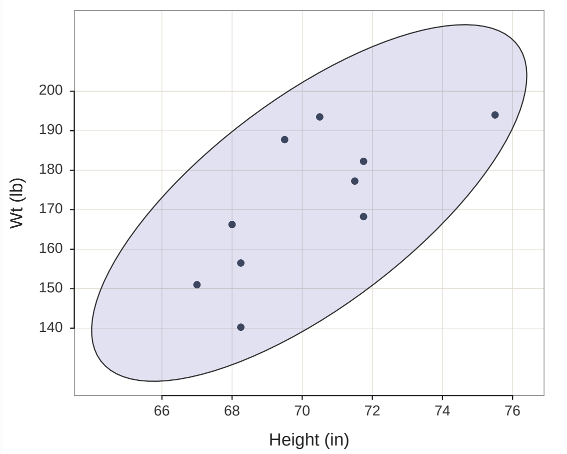
The Data Ellipse

Better visualization of the relationship

- Presuming that each variable is normally distributed, the graphical rendition of the correlation coefficient is the width of the ellipse that contains *most* of the points in the scatterplot
- **.95 Data Ellipse**: Estimated ellipse that contains 95% of the points in a scatterplot of two normally distributed variables for the entire population
  - For any one sample, *approximately* 95% of the points are within the ellipse
  - The option `ellipse=TRUE` with the `lessR` function Plot yields the 0.95 data ellipse

```
> Plot(Ht, Wt,
             xlab="Height (in)", ylab="Wt (lb)",
             ellipse=TRUE)
```

Scatter Plot with Data Ellipse



Scatterplot: Adult Height and Weight, Conclusion

Interpret the scatterplot

- Height and Weight appear to be related
- **Interpretation**: The relation is positive, as Height increases, Weight also tends to increase
- The relationship is a tendency, not a perfect linear relationship
    - For a given value of Height, there are many possible values of Weight, but larger Heights are more often associated with larger Weights
    - If you know a person's Height, then a better, though not perfect, forecast can generally be made of the person's Weight than if the person's Height was not known

8.1b
Example Scatterplots

Correlation Coefficient

Numerical index of the strength of a linear relationship

- The statistic that reflects the strength of a linear (straight-line) relationship is the correlation coefficient
- The sample or population correlation varies between -1 and 1, which respectively indicate perfect $-$ or $+$ linear relationship
- Notation of the correlation of variables x and y
    - Sample: r_{xy}, or simply r
    - Population: corresponding Greek letter for r, rho, written as ρ_{xy}, or simply ρ
- The sign of the correlation indicates the direction of the relationship
    - A positive sign indicates a positive relationship
    - A negative sign indicates a negative relationship
    - Zero indicates no relationship

Scatterplot for No Relation

When the correlation is zero

- **No relation** between x and y: As x increases, y can increase *or* decrease with equal probability
- For no relation, for any value of X, half of the values of y will be above the mean of y and half of the values will be below
- As illustrated in the next slide, the ellipse for a scatterplot of no relation is so wide that the ellipse becomes a circle[1]

[1] More generally, depending on how the axes are scaled, the "circle" may become an ellipse with no tilt

Sample Correlation from a Population with $\rho = 0.00$

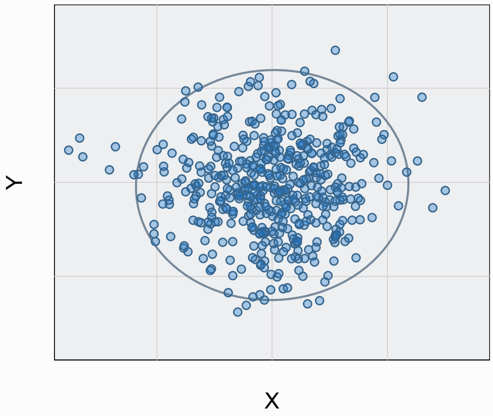


Figure: Sample correlation of 0.01, n=500.

Positive Relationships of Varying Strengths

As x increases, y tends to increase

- As the strength of the relationship between x and y increases, the scatter in the scatterplot decreases
- So, as strength or magnitude of the relationship increase, the ellipse that contains most of the points becomes increasingly narrow
- The following figures illustrate different relationships, for correlations of increasing strength from 0.25, 0.50, 0.75, 0.95 to the perfect 1.00

Sample Correlation from a Population with $\rho = 0.25$

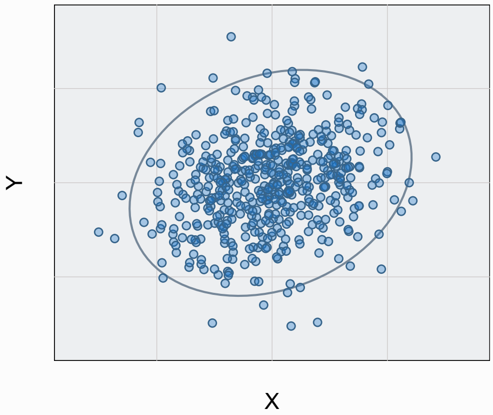


Figure: Sample correlation of 0.22, n=500.

Sample Correlation from a Population with $\rho = 0.50$

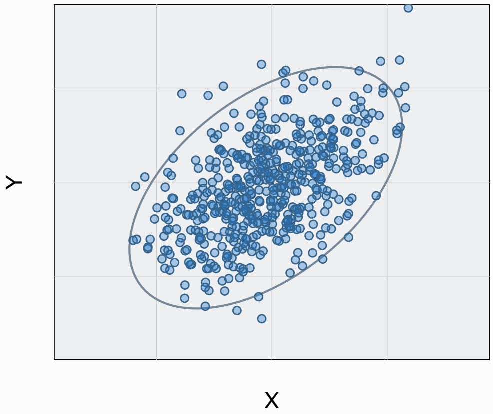


Figure: Sample correlation of 0.51, n=500.

Sample Correlation from a Population with $\rho = 0.75$

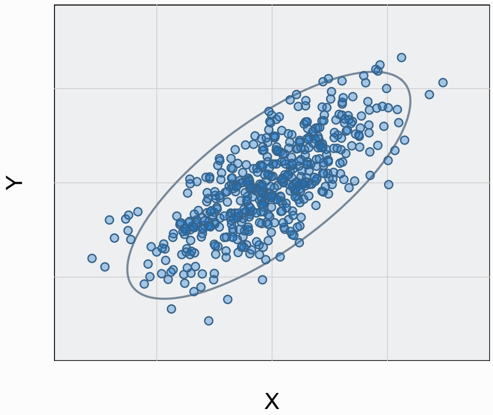


Figure: Sample correlation of 0.74, n=500.

Sample Correlation from a Population with $\rho = 0.95$

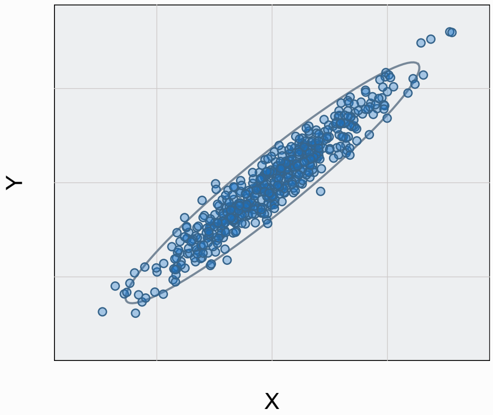


Figure: Sample correlation of 0.96, n=500.

Sample Correlation from a Population with $\rho = 1.00$

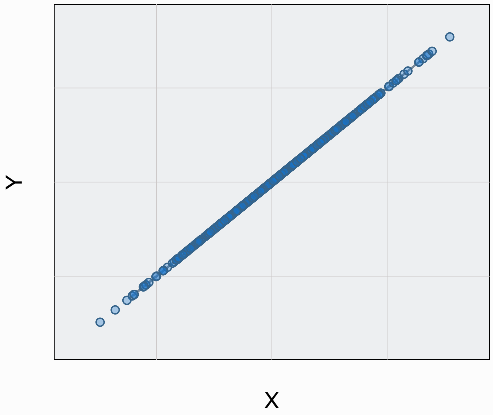


Figure: Sample correlation of 1.00, n=500.

Negative Relationships of Varying Strengths

As x increases, y tends to decrease

- As the strength of the relationship between x and y increases, the scatter in the scatterplot decreases
- So, as strength or magnitude of the relationship increase, the ellipse that contains most of the points becomes increasingly narrow
- The following slides successively illustrate different relationships, for correlations of increasing strength from −0.25, −0.50, −0.75, −0.95 to the perfect −1.00

Sample Correlation from a Population with $\rho = -0.25$

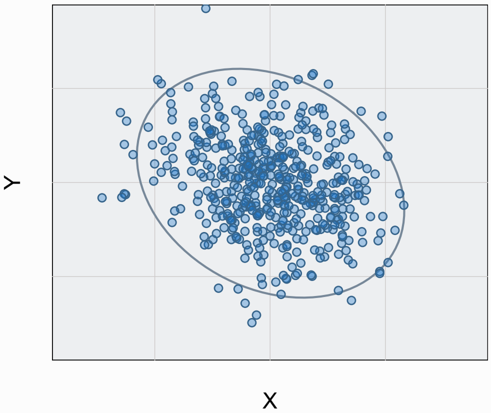


Figure: Sample correlation of -0.24, *n*=500.

Sample Correlation from a Population with $\rho = -0.50$

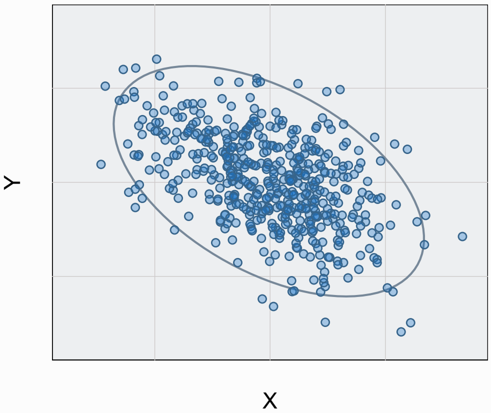


Figure: Sample correlation of -0.50, *n*=500.

Sample Correlation from a Population with $\rho = -0.75$

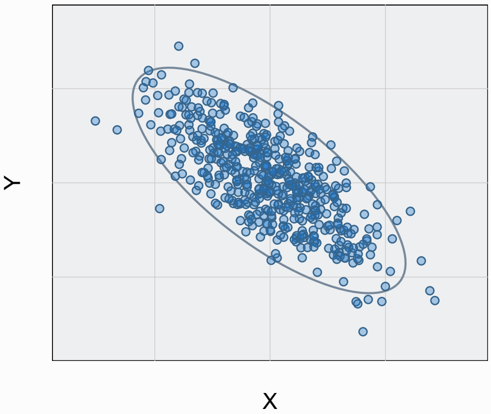


Figure: Sample correlation of -0.73, *n*=500.

Sample Correlation from a Population with $\rho = -0.95$

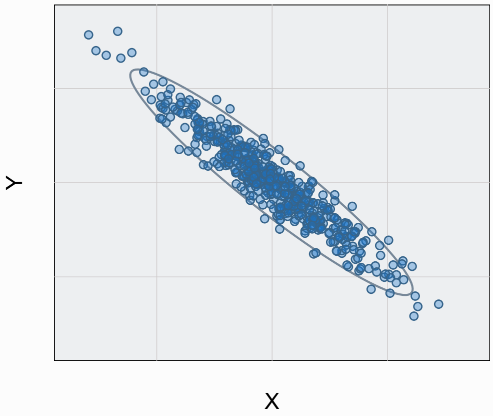


Figure: Sample correlation of -0.94, n=500.

Sample Correlation from a Population with $\rho = -1.00$

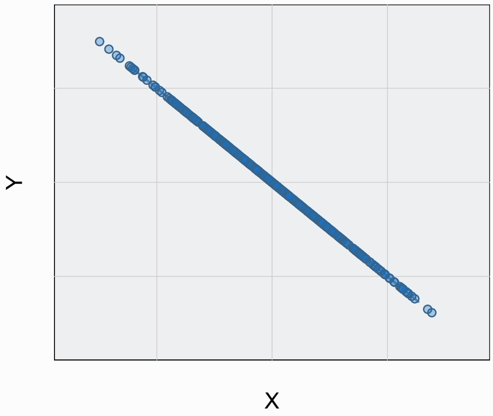
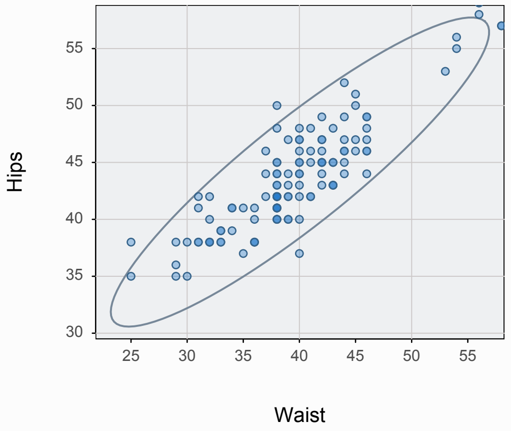


Figure: Sample correlation of -1.00, n=500.

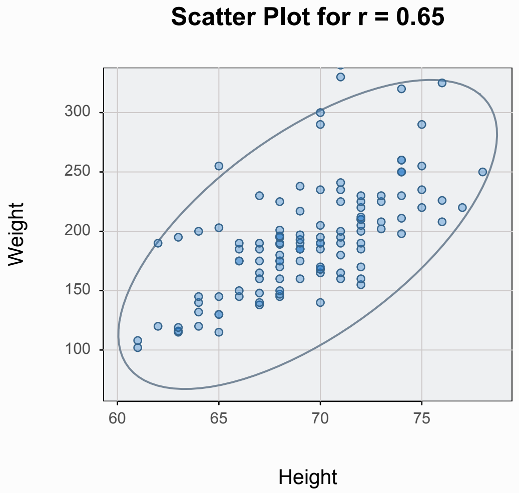
Correlation of Adult Waist and Hips

n = 122


Scatter Plot for r = 0.91

Correlation of Adult Height and Weight

n = 122



8.1c
Inference for the Correlation Coefficient

Correlation: Inferential Analysis

Moving from the sample to the population

- **Key Concept**: The analysis of *any* sample statistic, such as the mean or correlation, should always be augmented with an inferential analysis, confidence interval and/or hypothesis test
- The question of interest is always: How close is the sample statistic to the underlying population value?
- **Hypothesis Test** of the correlation: The null hypothesis is usually no relationship, a population correlation, ρ, of zero
  - ○ Null hypothesis, $H_0 : \rho = 0$
  - ○ Alternative hypothesis, $H_1 : \rho \neq 0$
- **Confidence Interval** of the correlation: Range of values that likely contain the true population correlation, ρ
- The confidence interval estimates the value of ρ at the specified level of confidence

Correlation: Standard Error, Degrees of Freedom

For a population correlation of zero, $\rho = 0$

- Inference requires a degrees of freedom to identify the relevant t-distribution to establish the 95% range of variation
- For the correlation coefficient, subtract 1 from the sample size, n, for each variable, so, for r, $df = n - 2$
- The basis of inference is the standard error of the corresponding statistic, here the standard deviation of the sample correlation, r, over usually hypothetical multiple samples
- To perform statistical inference for the population correlation, ρ
  - When $\rho = 0$, the standard error: $s_r = \sqrt{\dfrac{1 - r^2}{df}}$
  - For H_0 of $\rho = 0$, base hypothesis test on: $t_r = \dfrac{r - 0}{s_r}$
  - 95% Confidence interval: ρ is likely in $r \pm (t_{.025})(s_r)$

R/lessR: Correlation

```
> Plot(Ht,Wt)     or     sp(Ht,Wt)
> Correlation(Ht,Wt) or cr(Ht,Wt), no scatterplot
```

```
Number of paired values with neither missing, n: 10
Number of cases (rows of data) deleted: 0

Sample Covariance: cov = 33.435

Sample Correlation of Ht and Wt: r = 0.715

Alternative Hypothesis:
  True correlation is not equal to 0
  t-value: 2.892,  df: 8,  p-value: 0.020

95% Confidence Interval of Population Correlation
  Lower Bound: 0.155     Upper Bound: 0.927
```

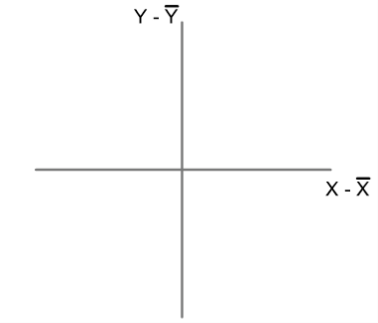
Correlation: Interpretation of Inferential Results

Always describe results in non-technical language

- Hypothesis Test
  - Statistical decision: p-value $< \alpha = 0.05$, so reject the null hypothesis that $\rho = 0$
  - **Interpretation**: For variables Ht and Wt, a relationship is detected
- Confidence Interval
  - Result: The 95% confidence interval for the correlation varies from 0.16 to 0.93
  - **Interpretation**: With 95% confidence, for Ht and Wt the population correlation is somewhere between a weak relationship of 0.16 to a strong relationship of 0.93, so as Ht increases, on average, so does Wt

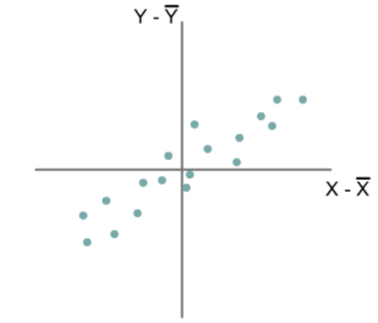
Appendix
Basis of the Correlation Coefficient

Logic of Correlation/Covariance Coefficient



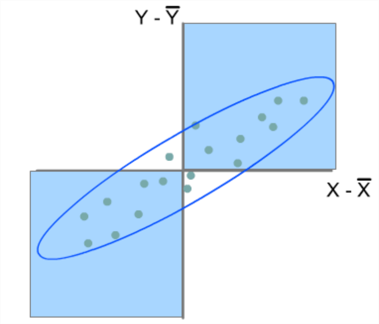
- The logic of the correlation coefficient is based on the analysis of the mean deviated variables, $x_i - \bar{x}$ and $y_i - m$

Logic of Correlation/Covariance Coefficient



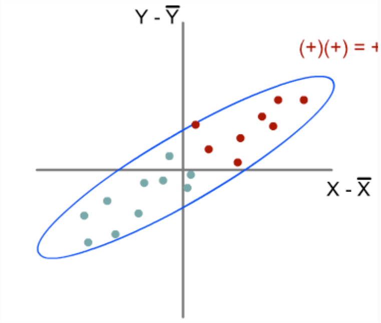
- The mean of a mean deviated variable is 0, so the scatterplot of the mean deviated variables is centered over $< 0, 0 >$

Logic of Correlation/Covariance Coefficient



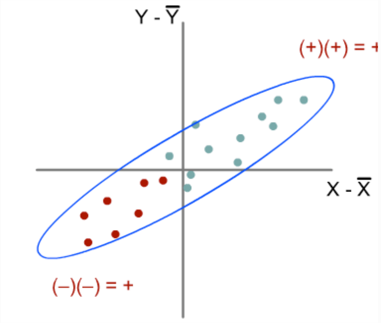
- For a positive linear relationship, most of the points in the (mean deviated) scatterplot fall in these two marked quadrants

Logic of Correlation/Covariance Coefficient



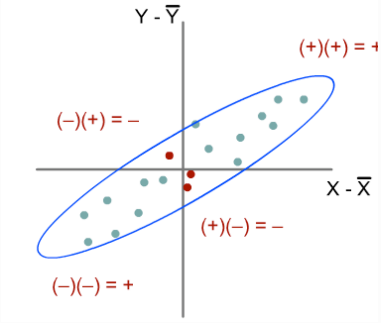
- All points in the upper-right quadrant have + coordinates on both axes, and the product of two + numbers is +

Logic of Correlation/Covariance Coefficient



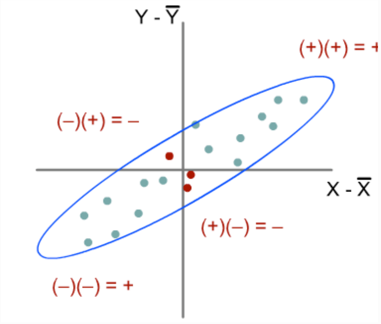
- All points in the lower-right quadrant have − coordinates on both axes, and the product of two − numbers is +

Logic of Correlation/Covariance Coefficient



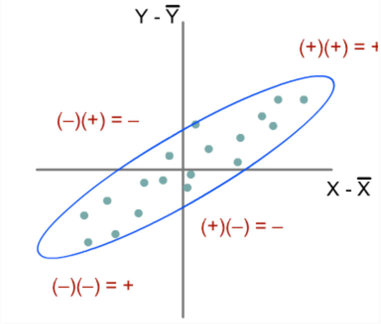
- All points in the two remaining quadrants have a $+$ coordinate on one axis, and a $-$ coordinate on the other

Logic of Correlation/Covariance Coefficient



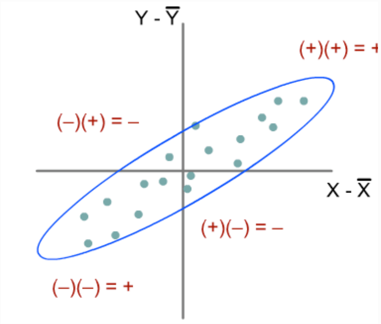
- Multiplying $-$ and $+$ mean deviated values yields a $-$ number

Logic of Correlation/Covariance Coefficient



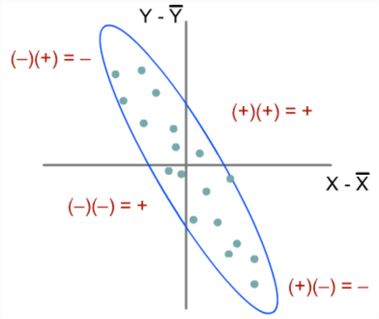
- **Cross-product**: Product of the two mean deviated coordinates for each point
- The cross-product is the fundamental concept underlying the correlation coefficient

Logic of Correlation/Covariance Coefficient



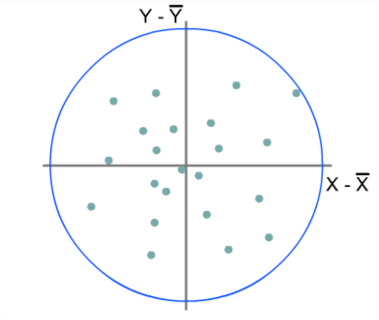
- Cross-products tend to be $+$ for a positive relation, so the sum of all cross-products for a $+$ relation is also $+$

Logic of Correlation/Covariance Coefficient



- Similarly, for a negative (inverse) relationship, summing the cross-products yields a negative sum

Logic of Correlation/Covariance Coefficient



- For no relationship, cross-products sum to approximately zero in a sample

Sample Covariance

The formula

- Sum of the cross-products of the mean deviations for variables x and y indicates the strength of the linear relationship
- The sum is confounded by sample size, so to remove divide the sum by degrees of freedom to get a mean
- Sample **Covariance** of x and y: Average sum of the cross-products of the corresponding deviation scores

$$s_{xy} = \frac{\sum(x_i - \bar{x})(y_i - m)}{n-1}$$

- The covariance is **Scale Dependent**, its value depends on the underlying measurement scales, such as feet versus inches

Sample Correlation

The formula

- **Scale Independence**: A change of the measurement scale of one or both variables does *not* change the value of the resulting statistic
- Obtain scale independence by standardizing x and y, dividing the deviation scores by their respective standard deviations
- Sample Correlation, which varies from -1 to 0 to 1:

$$r_{xy} \;=\; \frac{\sum \dfrac{(x_i - \bar{x})}{s_x}\dfrac{(y_i - m)}{s_y}}{n-1} \;=\; \frac{\sum z_x z_y}{n-1} \;=\; \frac{s_{xy}}{s_x s_y}$$

- Will obtain the same correlation if Height is measured in inches or cm because Ht in inches or Ht in cm are the same values *after* standardization

Illustration of Scale Dependence

The data

- Consider measurements of Height and Weight for 10 adult men, with Height measured in both inches and centimeters, found at:

  `http://lessRstats.com/data/CovScaleDep.csv`

```
   Ht.in  Ht.cm     Wt
1  71.75 182.24 182.25
2  71.75 182.25 168.25
3  75.50 191.77 194.00
4  68.25 173.35 140.25
5  68.25 173.35 156.50
6  69.50 176.53 187.75
7  70.50 179.07 193.50
8  71.50 181.61 177.25
9  67.00 170.18 151.00
10 68.00 172.72 166.25
```

- 2.54 centimeters to each inch, so Ht.cm = 2.54 * Ht.in

Illustration of Scale Dependence

Scale dependent covariance vs. scale independent correlation

- Obtain both the covariance and the correlation from the `lessR` function `Correlation` applied to two variables

```
> Correlation(Ht.in, Wt)
Sample Covariance: s = 33.435

Sample Correlation: r = 0.715

> Correlation(Ht.in, Wt)
Sample Covariance: s = 84.942

Sample Correlation: r = 0.715
```

- The covariance based on cm is 2.54 times larger than for inches

$$84.942 = 33.435 * 2.54$$

- The correlation is the same for both analyses, $r = 0.715$

- The End