So far: Explored one variable at a time

Is there a relationship between these 2 variables?

Visual relationship with a scatterplot

To generate a scatterplot, need the actual matched pairs of X_i & Y_i values.

Baseball: Team Wins & Team OPS

OPS: On base Plus Slugging a measure of team’s ability to get on base and hit for power
Perceived Orientation vs. Stimulus Angle

probability of responding “counter-clockwise of vertical” plotted as a function of angle from vertical

probability ("counter-clockwise")

\[ \mu = -1.32 \]

\[ \sigma = 1.40 \]

\[ sO = 30.00 \]

\[ nV30C100_1_tilt.mat \]

Visual Acuity vs Age in Human Infants

Correlation & Linear Regression

• correlation & linear regression are statistical methods that are used to assess the association between variables

• Correlation: measure the linear association between 2 variables

• Regression: estimates the “best-fitting” line that relates a predictor variable, X, to a criterion variable, Y.
  - used to understand how Y changes when X varies

https://www.sciencedirect.com/topics/nursing-and-health-professions/eye-chart
Examples of Correlational Research

- effectiveness of fluoridated water in reducing cavities
- association between smoking and lung cancer
- false claim of link between autism & measles, mumps, & rubella (MMR) inoculations – proposed that MMR causes autism
- drinking coffee associated with longer life
- claim that orchestra conductors live longer

Correlation & Linear Regression

Regression line: the "line of best fit"; represents best prediction of Yi for a given value of Xi

Correlation measures the strength & direction of linear association

Correlation depends on fit, not slope
Perfect Linear Relation \((r = 1\) or \(r = -1)\)

- **Perfect positive correlation**: \(r = 1.0\)
- **Perfect negative correlation**: \(r = -1.0\)

Common measure of correlation — Pearson \(r\) — varies between -1 & +1

No linear association \((r = 0)\)

Vertical & horizontal lines indicate means of \(x\) and \(y\).

Correlations & deviations around the means

- **\(r = 0.7\)**
- **\(r = -0.7\)**

Vertical & horizontal lines indicate means of \(x\) and \(y\).

Note that points do not fall within each quadrant equally.

Linear vs. curvilinear trends/relations

- **Monotonic**
  - **Linear**
  - **Curvilinear**

- **Non-monotonic**
  - **Curvilinear**

- **Monotonic trend**: as \(X\) increases, \(Y\) increases or decreases without reversal
  - might not be a straight line!
- **Non-monotonic trend**: as \(X\) increases, \(Y\) changes direction at least once
- **Linear relationship**: best fit line is straight
- **Curvilinear**: best fit “line” is not straight (non-linear)
- **Correlation is only sensitive to linear relations!**
4 (x,y) data sets plotted with a reference (best-fitting) line

Question: which set has the highest correlation?

The correlation is the same in all 4 sets! $r = 0.82$

Correlation is a measure of linear association and is insensitive to non-linear association.

Correlation is a measure of linear association between variables

**Covariance**

- Measures the degree to which 2 variables vary together

\[
\text{COV}_{XY} = \frac{\sum_{i=1}^{N}(X_i - \bar{X})(Y_i - \bar{Y})}{N - 1}
\]

- where N is the number of observations.
- Covariance depends on sum of products of deviation scores
  - Covariance is positive when deviation scores have same sign
  - Covariance is negative when deviation scores have opposite signs
Green points drive COV<sub>XY</sub> in positive direction. Blue points drive COV<sub>XY</sub> in negative direction. Net COV<sub>XY</sub> differs from zero.

Green points drive COV<sub>XY</sub> in positive direction. Blue points drive COV<sub>XY</sub> in negative direction. Note that net COV<sub>XY</sub> is zero.

Covariance is a measure of association (but is influenced by spread of X & Y)

Covariance = 5

Covariance = 250

Similar (X,Y) associations but very different covariances due to differences in (X,Y) spread.

Covariance = 5

Covariance = 250

Same data as in previous slide, but replotted to highlight difference in (X,Y) spread.
Correlation vs. Covariance

COV(X,Y) is affected by variances of X and Y. So changing the units of measures generally changes covariance.

Correlation Coefficient, r

- $r = \frac{\text{COV}(X,Y)}{S_X S_Y}$ Standardizes the covariance so we can compare correlations obtained with data sets with very different (X,Y) spreads.

- $s_X$ and $s_Y$ are the standard deviations of X and Y respectively.
- $r$ is the Pearson Product-Moment Correlation Coefficient
- Note that $r$ has no “units”

Correlation vs. Covariance

Correlation, $r$, is less affected than COV(X,Y) by variances of X and Y.

Correlation vs. Covariance

Correlation, $r$, is less affected than COV(X,Y) by variances of X and Y. So changing units changes covariance but not $r$.
**Pearson r**

- **r** ranges from -1 to 1
  - **r** = +1 when **X** and **Y** are perfectly positively correlated
  - **r** = -1 when **X** and **Y** are perfectly negatively correlated
  - **r** = 0 when there is no linear relationship between **X** and **Y**

- **in Psychology,**
  - Weak: 0.1-0.3
  - Moderate: 0.3-0.5
  - Strong: >0.5

---

**What does r mean?**

- Data: 20 pairs of **X, Y** scores

<table>
<thead>
<tr>
<th>Score</th>
<th>X</th>
<th>Y</th>
</tr>
</thead>
<tbody>
<tr>
<td>20</td>
<td>71.02058</td>
<td>100.201137</td>
</tr>
<tr>
<td>30</td>
<td>62.31119</td>
<td>52.953707</td>
</tr>
<tr>
<td>40</td>
<td>179.31734</td>
<td>163.077298</td>
</tr>
<tr>
<td>50</td>
<td>110.02782</td>
<td>83.582028</td>
</tr>
<tr>
<td>60</td>
<td>172.73203</td>
<td>142.985011</td>
</tr>
<tr>
<td>70</td>
<td>128.27301</td>
<td>138.997126</td>
</tr>
<tr>
<td>80</td>
<td>81.07347</td>
<td>58.119052</td>
</tr>
<tr>
<td>90</td>
<td>65.82096</td>
<td>53.981494</td>
</tr>
<tr>
<td>100</td>
<td>60.2459</td>
<td>65.250576</td>
</tr>
<tr>
<td>110</td>
<td>123.27456</td>
<td>108.558325</td>
</tr>
<tr>
<td>120</td>
<td>135.13302</td>
<td>152.882385</td>
</tr>
<tr>
<td>130</td>
<td>68.73133</td>
<td>65.628767</td>
</tr>
<tr>
<td>140</td>
<td>22.89659</td>
<td>7.949241</td>
</tr>
<tr>
<td>150</td>
<td>182.69514</td>
<td>173.079349</td>
</tr>
<tr>
<td>160</td>
<td>14.85593</td>
<td>17.623536</td>
</tr>
<tr>
<td>170</td>
<td>77.4901</td>
<td>55.465981</td>
</tr>
<tr>
<td>180</td>
<td>72.90529</td>
<td>85.641115</td>
</tr>
<tr>
<td>190</td>
<td>104.4358</td>
<td>95.790047</td>
</tr>
<tr>
<td>200</td>
<td>174.46641</td>
<td>178.942469</td>
</tr>
</tbody>
</table>

- IQR = 62
- IQR = 83

---

**Plot of Y scores**

- Each **Y** is paired with 1 **X**
- Knowing value of **X** tells us something about value of **Y**
- Relation between **Y & X** is not perfect (**r** = 1)

---

**Plot of Y vs. X scores**

- **r** = 0.95

---

**residuals (errors)**

- Residuals represent the part of Y score that is NOT “accounted for” or “explained by” or “associated with” X
- Variance of residuals is much less than variance of original Y scores

---

When X & Y are correlated, knowing value of one variable reduces uncertainty about value of other variable.

\[
 r^2 = \text{proportion of variance in Y that is accounted for by X} \\
 \text{VAR}(Y) = 2500 \\
 \text{VAR(Residuals)} = 243.75 \\
 1 - \frac{243.75}{2500} = 0.9025 \\
 \text{proportion of variance not explained by } X \\
 \frac{2500 - 243.75}{2500} = 0.9025 \\
 r^2 = 0.9^2 = 0.81
Decomposing weight variance into “explained” and “unexplained” parts

\[ R^2 = 0.77^2 = 0.59 \]

VARIANCE(weight) = 865.4
VARIANCE(predicted weights) = 514.1
VARIANCE(errors) = 351.3

\[
\text{proportion of variance(weight) that is } \text{ accounted for by height} \\
\frac{\text{VARIANCE(predicted weights)}}{\text{VARIANCE(weight)}} = \frac{514.1}{865.4} = 0.59 = R^2
\]

\[
\text{IQR(weight)} = 40.8 \quad \text{IQR(predicted weights)} = 34.2 \quad \text{IQR(errors)} = 22.7
\]

Knowing X reduces uncertainty about Y

\[ r \] is related to how much uncertainty about Y is reduced by knowing X

\[ r = 0 \]

This looks like a perfect correlation. Why is \( r = 0 \)?
r varies across samples
calculate r for many samples of data (n = 20 per sample)

Sample r (n = 20; true-r = 0.4)

- population r = 0.4
- mean (sample r) = 0.39
- range (sample r) = [-0.45, 0.88]

Each sample r is an estimate of the population r. Some estimates are good, others are not so good.

Can we quantify the uncertainty of our estimate?

Confidence Interval

A confidence interval is a range of values, calculated from the sample observations, that are believed, with a particular probability, to contain the true population parameter. A 95% confidence interval, for example, implies that were the estimation process repeated again and again, then 95% of the calculated intervals would be expected to contain the true parameter value.

— B.S. Everitt, Dictionary of Statistics

Confidence Interval

- **95% Confidence Interval**
  - an interval estimate of the value of a population parameter (e.g., r)
  - calculated from data in your sample
  - the interval varies across samples
  - in the long run, the interval contains the true population value 95% of the time

- in our simulation with true population r = 0.4:
  - we obtained one (X,Y) data set with a correlation r = 0.37
  - for that sample, our statistical software calculates CI_{95%} = [-0.08, 0.70]
    - we estimate that the true value of r is between -0.08 & 0.70
    - note that our CI does contain the true population r = 0.4
Confidence Interval

- **95% Confidence Interval**
  - an *interval estimate* of the value of a population parameter (e.g., r)
  - calculated from data in your sample
  - the interval varies across samples
  - in the long run, the interval contains the true population value 95% of the time
- in our simulation with true population r=0.4:
  - we obtained one (X,Y) data set with a correlation r = 0.37
  - for that sample, our statistical software calculates CI95% = [-0.08, 0.70]
  - we estimate that the true value of r is between -0.08 & 0.70
  - note that our CI does contain the true population r=0.4
- Note that we can adjust the width of our interval
  - 99% CI: contains true population value 99% of the time
  - 90% CI: contains true population value 90% of the time
  - Question: For our data set, which CI would be widest: 99%, 95%, or 90%?
  - Hint: What do you think is the 100% CI?

Confidence Interval Examples

"Margin of error" in polls

1. What is the margin of error anyway?

Because surveys only talk to a sample of the population, we know that the result probably won't exactly match the "true" result that we would get if we interviewed everyone in the population. The margin of sampling error describes how close we can reasonably expect a survey result to fall relative to the true population value. A margin of error of plus or minus 3 percentage points at the 95% confidence level means that if we fielded the same survey 100 times, we would expect the result to be within 3 percentage points of the true population value 95 of those times.


Confidence Interval Examples

(90% Confidence Interval)

https://www.texasgateway.org/resource/807-confidence-intervals-real-world

Other Types of Correlations

- Two continuous variables
  - Pearson Product-Moment Correlation Coefficient (r)
- Two ranked/ordinal variables
  - Spearman’s correlation coefficient for ranked data (rho (ρ) or r_)
- One dichotomous & one continuous variable
  - e.g., correct/incorrect answer on mc question and exam total score
  - Point-biserial Correlation (rpb): Calculate Pearson’s r but call it rpb
- Two dichotomous variables
  - e.g. age (teens/seniors) and coin purse ownership (yes/no)
  - Calculate Pearson’s r but called in rφ
Correlations with ranked data

- Spearman's correlation coefficient, $r_s$
  - useful when observations have been replaced by their numerical ranks
    - e.g., changing applicants' exam scores to ranks

<table>
<thead>
<tr>
<th>english</th>
<th>math</th>
<th>english.rank</th>
<th>math.rank</th>
</tr>
</thead>
<tbody>
<tr>
<td>56</td>
<td>66</td>
<td>2.0</td>
<td>7</td>
</tr>
<tr>
<td>75</td>
<td>70</td>
<td>8.0</td>
<td>9</td>
</tr>
<tr>
<td>45</td>
<td>40</td>
<td>1.0</td>
<td>1</td>
</tr>
<tr>
<td>71</td>
<td>60</td>
<td>7.0</td>
<td>4</td>
</tr>
<tr>
<td>61</td>
<td>65</td>
<td>4.5</td>
<td>6</td>
</tr>
<tr>
<td>64</td>
<td>56</td>
<td>6.0</td>
<td>2</td>
</tr>
<tr>
<td>58</td>
<td>59</td>
<td>3.0</td>
<td>3</td>
</tr>
<tr>
<td>80</td>
<td>77</td>
<td>10.0</td>
<td>10</td>
</tr>
<tr>
<td>76</td>
<td>67</td>
<td>9.0</td>
<td>8</td>
</tr>
<tr>
<td>61</td>
<td>63</td>
<td>4.5</td>
<td>5</td>
</tr>
</tbody>
</table>

Spearman’s rank order correlation

Same as Pearson $r$ correlation for ranks, not actual $(x,y)$ values.
Why use $r_s$ if you have the original values?

$r_s$ is more robust to some types of extreme points

$r = -0.05$
$r_s = -0.05$

Changing only 2 out of 100 points markedly affects $r$ but not $r_s$
...but not to all types of extreme scores

Correlations with ranked data

- Spearman’s correlation coefficient, \( r_s \), calculated with the same formula that is used to calculate Pearson’s \( r \)
  - formula applied to ranks, not actual \((X,Y)\) values
- \( r_s \) is much more resistant/robust than \( r \) to extreme/outlier data points
- However, interpretations of \( r \) and \( r_s \) differ:
  - \( r \) is an index of the strength/direction of linear relation
  - \( r_s \) is an index of the strength/direction of monotonic relation

\[ r = 0.22 \]
\[ r = 0.78 \]
\[ r_s = 0.37 \]
\[ r_s = 0.83 \]

In this case, including extreme point alters \( r_s \) too. Why?

\( r \) measures linear association between \( X,Y \) ranks
\( r_s \) measures monotonic association between \( X,Y \) scores

\( r_s \) is sensitive to extreme scores affecting monotonicity

Including Northern Ireland data point significantly alters monotonicity of the \( X,Y \) association
**Other Types of Correlations**

- Two continuous variables
  - Pearson Product-Moment Correlation Coefficient ($r$)
- Two ranked/ordinal variables
  - Spearman’s correlation coefficient for ranked data (rho ($\rho$) or $r_i$)
- One dichotomous & one continuous variable
  - e.g., correct/incorrect answer on mc question and exam total score
  - **Point-biserial Correlation ($r_{pb}$):** Calculate Pearson’s $r$ but call it $r_{pb}$
- Two dichotomous variables
  - e.g., age (teens/seniors) and coin purse ownership (yes/no)
  - Calculate Pearson’s $r$ but called in $r_\phi$ (phi)

**Point-biserial Correlation $r_{pb}$**

- example: analyzing an item on a multiple-choice test

- **X variable:**
  - answer for one multiple-choice question
  - “incorrect” (0) or “correct” (1)
- **Y variable:** total score on remaining questions
- Correlate X & Y: $r_{pb} = 0.46$
- re-code “incorrect” & “correct” responses with other numbers (e.g., -10 & 10):
  - magnitude $r_{pb}$ is not changed by coding scheme for X variable
  - (sign of $r_{pb}$ can change)

### Example Table

<table>
<thead>
<tr>
<th>ADD diagnosis</th>
<th>Needs Remedial English Course</th>
<th>No</th>
<th>Yes</th>
</tr>
</thead>
<tbody>
<tr>
<td>No</td>
<td>187</td>
<td>22</td>
<td></td>
</tr>
<tr>
<td>Yes</td>
<td>74</td>
<td>19</td>
<td></td>
</tr>
</tbody>
</table>

**$r_\phi$ (phi) measures association in 2 x 2 contingency tables**

- study of 302 school children
  - each assessed for ADD & need for remedial English classes
  - each child provides 2 binary measures
- $r_\phi$ measures association between binary variables
  - code each variable as 0 (NO) or 1 (Yes)
  - calculate Pearson $r$ on zeros and ones
- $r_\phi = 0.133$
How do we know if a correlation is “real”?

• \( r \) estimates the population correlation

• \( r \) varies across samples
  - a true pop \( r \) of 0.4 can yield a sample \( r \leq 0 \)
  - and a true pop \( r \) of 0.0 can yield a sample \( r \) that is not zero

Sample \( r \) (\( n = 20; \) true-\( r = 0.4 \))

95% of values lie between -0.04 & 0.68

How do we know if a correlation is “real”?

• and a true pop \( r \) of 0.0 can yield a non-zero sample \( r \)

• how can we decide if our observed sample correlation occurred just by chance (and true \( r = 0 \))?

• one strategy: assume true \( r = 0 \) and calculate the probability of getting our sample \( r \) (or something even bigger) just by chance

Sample \( r \) (\( n = 20; \) true-\( r = 0 \))

95% of values lie between -0.44 & 0.38

2.5% of values > 0.38

Null Hypothesis Testing Logic

• Null Hypothesis: true correlation is zero \( H_0: r = 0 \)
• Alternative Hypothesis: true correlation is non-zero \( H_1: r \neq 0 \)
• Assume \( H_0 \) is true:
  - what is the probability of obtaining an \( r \) that is at least as big as the one we found in our sample?
  - if that probability is very low (e.g., less than 5%), then our \( r \) is unusual (assuming \( H_0 \) is true)
    ‣ therefore we may reject \( H_0 \) in favour of \( H_1 \) (i.e., the true \( r \neq 0 \))
  - if that probability is not low, then our \( r \) is not unusual (assuming \( H_0 \) is true)
    ‣ our observed correlation is NOT unlikely when true \( r = 0 \)
    ‣ we do not have sufficient evidence to reject \( H_0 \)
    ‣ do not “accept” \( H_0 \); simply fail to reject it

Null Hypothesis Testing (Example)

Question: are these correlations larger than we would expect when true population \( r = 0 \)?

Try to answer question by estimating probability of getting \( r \) this large when true correlation is zero
Null Hypothesis Testing (Example)

- Randomly scramble order of X,Y values
  - *so each X paired RANDOMLY with a Y*
  - Scrambling means there is no X,Y association

- Calculate r for our scrambled sample
  - expect it to be zero **on average**
  - but it will vary... not always exactly zero

- Repeat this process many times, record r's for all scrambled samples

- Because of random pairing, large values of ±r occur only by chance

---

Null Hypothesis Testing (Example)

- Randomly scramble order of X,Y values
  - *so each X paired RANDOMLY with a Y*
  - Scrambling means there is no X,Y association

- Calculate r for our scrambled sample
  - expect it to be zero **on average**
  - but it will vary... not always exactly zero

- Repeat this process many times, record r's for all scrambled samples

- Because of random pairing, large values of ±r occur only by chance

- Estimate probability of getting r that is at least as extreme as our sample r
  - observed r for data set 1: r = 0.82
  - observed r for data set 2 r = 0.63
Null Hypothesis Testing (Example)

- Our example used a PERMUTATION test
- Other kinds of tests can be used to evaluate null hypothesis (true pop r = 0)
- But the logic is similar:
  - assume NULL HYPOTHESIS IS TRUE (i.e., true pop r is zero)
  - calculate probability of getting a sample r that is at least as extreme as yours
  - if the probability is small, then observed r is unusual WHEN THE NULL HYPOTHESIS IS TRUE
  - and therefore you may reject HO in favour of alternative (i.e., true pop r ≠ 0)
- we say the correlation was statistically significant ($r=0.82, p<0.05$) or ($r=0.63, p<0.05$)
- N.B. You can make mistakes!
  - observed r may just be unusual (i.e., true pop r may really be zero)
  - we can estimate the probability of making these mistakes
  - in our example the probability of this error, referred to as alpha, was .05 or 5%. Why?
- Our alpha = 5% because we defined an “unusual” r as being outside the 95% boundaries

Factors that affect correlation

- Nonlinearity
- Extreme observations
- Range restrictions
- Heterogeneous subsamples

Restricted range of X variable

- Restricted range of X variable can obscure a linear association between X & Y.
Restricted Range (Example)

Restricted range can obscure a **curvilinear** association.

**Heterogeneous Subsamples**

- Simpson (1951):
  - A statistical association observed in a population can be attenuated and even reversed within subgroups that make up that population.
- Consequently, correlations calculated on populations consisting of heterogeneous subgroups may be misleading.

**Simpson’s Paradox**

Correlation for population is positive

Correlations within sub-groups are negative

Overall correlation is approximately zero

Strong, opposite correlations in 2 sub-groups of tennis players
Correlation ≠ Causation

- reading skill is correlated with shoe size
- over last 2 centuries, price of bread in Great Britain is correlated with sea level in Venice
- rpb: BMW owners have higher incomes than Ford owners
- number of pirates is correlated with global temperature
- ...even strong correlations do not mean X causes Y

Correlation (summary)

- Correlation: is a measure of the association between 2 variables
  - Pearson $r$: index of strength & direction of linear association
    - $r = \text{COV}_{xy}/(s_x s_y)$; varies between -1 & +1
    - measures how much our uncertainty about the value of $Y$, is reduced by knowing the value of $X$,
    - $r$ is an estimate of the population correlation that varies across samples
    - a 95% Confidence Interval of $r$, which varies from sample to sample, contains the true population parameter 95% of the time
  - Spearman’s $r_s$: index of monotonicity of association
    - equivalent to Pearson’s $r$ for ranked data
    - generally more robust than $r$ to extreme (i.e., “high-leverage”) data points
  - Correlations affected by: nonlinearity, extreme scores, range restrictions, heterogeneous samples
- CORRELATION ≠ CAUSATION
- Null Hypothesis Testing is a method for making decisions about value of population $r$
  - $p$ values are probability of getting our data when null hypothesis is true

Pirates Do Not Cause Global Warming
Correlation ≠ Causation