Relationships Between Variables: Correlation

Even variables have relationships.
Kinds of Variables

• categorical - naming things/people or putting them into categories or groups (nominal variables) (grouping variables)

• ranking - putting things/people in order (ordinal variables)

• numeric - measuring things/people (interval and ratio variables)
## How to Assess Relationships Between Variables

<table>
<thead>
<tr>
<th>first variable (IV)</th>
<th>second variable (DV)</th>
</tr>
</thead>
<tbody>
<tr>
<td>categorical</td>
<td>categorical</td>
</tr>
<tr>
<td>numeric</td>
<td>numeric</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>categorical</th>
<th>numeric</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>chi-square</td>
<td>Pearson correlation</td>
</tr>
<tr>
<td></td>
<td>phi coefficient (2x2)</td>
<td>(multiple regression if more than one IV)</td>
</tr>
<tr>
<td></td>
<td>t-test (2 levels of IV)</td>
<td>R²</td>
</tr>
<tr>
<td></td>
<td>point-biserial correlation</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Cohen’s d</td>
<td></td>
</tr>
<tr>
<td></td>
<td>ANOVA (&gt;2 levels of IV)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>eta-squared</td>
<td></td>
</tr>
<tr>
<td></td>
<td>(beyond the scope of this course - but logistic regression)</td>
<td></td>
</tr>
</tbody>
</table>
Are College GPAs Related to SAT Scores?

• James Tresselt, Psyc 497, Fall 1995

• a random sample (!) of freshman admitted to Coastal (CCC) in Fall 1990 and who were still present at the end of Spring 1991

• data obtained from the Office of Institutional Research (now called something else)

• tresselt.txt
the variables
- gender - obvious (categorical, 2 levels, how many subjects are there altogether?)
- SATV - SAT Verbal scores (numeric, on 200-800 scale, so interval or ratio?)
- SATQ - SAT Quantitative score (numeric, on 200-800 scale)
- SATT - SAT Total score (SATQ + SATQ, numeric, on 400-1600 scale)
- GPA91 - grade point average at end of Spring 1991 semester
- HSGPA - high school GPA (I made these up to replace class ranks)
- orient - did the student attend freshman orientation? (categorical, 2 levels, dummy coded, 0=no; I made these up out of thin air!)
- sex - a dummy coded version of Gender (0=female, 1=male)
- dummy (0/1) coding - for a dummy coded variable, what does the mean mean?
Correlation Matrix

- correlation coefficients
  - -1 to +1 scale with 0 meaning no relationship
  - assumes a linear relationship
- correlation between SATV and SATQ
  - two numeric variables
  - Pearson r (or Pearson product-moment correlation)
  - assess significance (is it different from 0?) using the table at the website, df = n - 2
  - or use the cor.test() function in R
- correlation between GPA91 and orient
  - a numeric and a dummy coded categorical variable
  - point-biserial correlation
  - assess significance using the table, or use cor.test()
  - this is equivalent to doing a t-test
- to be continued...

\[
r = \frac{n(\Sigma xy) - (\Sigma x)(\Sigma y)}{\sqrt{n(\Sigma x^2 - (\Sigma x)^2)(\Sigma y^2 - (\Sigma y)^2)}}
\]

formula for Pearson r (just in case you want to do them by hand!)
• correlation between orient and sex
  • two dummy coded (0/1) categorical variables
  • **phi coefficient**
  • assess *approximate* significance using the table
  • use a chi-square table for exact significance
    • chi-squared = phi-squared * n, df = 1
    • critical value for alpha = .05 is 3.8415
    • or just do a chisq.test() in R
  • phi coefficients are always positive
• correlation between a 1/2/3/... coded categorical variable and anything else
  • this is **NOT** dummy coding
  • these correlation are **MEANINGLESS**
  • (although R will do them because it doesn’t know 1/2/3 are not numbers unless you tell it)
  • we don’t have an example of that in this dataset - what would be an example?
Point-Biserial and t-test Are Equivalent

> t.test(SATQ ~ Gender, data=TR, var.eq=T)  # two-tailed with pooled variance

Two Sample t-test

data:  SATQ by Gender
t = -2.2, df = 249, p-value = 0.02873
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
  -40.41464  -2.23351
sample estimates:
mean in group Female  mean in group Male
  443.0827          464.4068

> cor.test(TR$SATQ, TR$sex)

Pearson's product-moment correlation

data:  TR$SATQ and TR$sex
t = 2.2, df = 249, p-value = 0.02873
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
  0.01451045 0.25749825
sample estimates:
cor
  0.1380816

What if there are more than two levels of the IV? Or more than one IV?
A (Kind Of) Way To Get a “Correlation” With More Than Two Categories

> summary(aov(GPA91~Gender*factor(orient), data=TRES))

<table>
<thead>
<tr>
<th></th>
<th>Df</th>
<th>Sum Sq</th>
<th>Mean Sq</th>
<th>F value</th>
<th>Pr(&gt;F)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gender</td>
<td>1</td>
<td>6.61</td>
<td>6.615</td>
<td>10.215</td>
<td>0.00157**</td>
</tr>
<tr>
<td>factor(orient)</td>
<td>1</td>
<td>4.36</td>
<td>4.363</td>
<td>6.738</td>
<td>0.01000*</td>
</tr>
<tr>
<td>Gender:factor(orient)</td>
<td>1</td>
<td>0.42</td>
<td>0.417</td>
<td>0.644</td>
<td>0.42301</td>
</tr>
<tr>
<td>Residuals</td>
<td>247</td>
<td>159.95</td>
<td>0.648</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

---

Signif. codes:  0 ‘****’ 0.001 ‘***’ 0.01 ‘**’ 0.05 ‘.’ 0.1 ‘ ’ 1

> 6.61 + 4.36 + 0.42 + 159.95  # total SS
[1] 171.34

> 171.34 - 159.95  # explained variability
[1] 11.39

> 11.39 / 171.34  # eta-squared or R-squared
[1] 0.06647601
Graphics

• most people prefer to look at pictures
• numeric variables - scatterplot
• grouped data - boxplots, profile plot
• two categorical variables - bar graphs
  • bar graphs are for frequencies!
  • if you insist on using bar graphs for means, add error bars!
Scatterplot

plot(GPA91 ~ SATT, data=TR)

$r = +0.33$
Scatterplots and Correlation (Pearson $r$)

Scatter Plots & Correlation Examples

- **Perfect Positive Correlation**
  - $r = 1$

- **Highly Positive Correlation**
  - $r = 0.8$

- **Low Positive Correlation**
  - $r = 0.3$

- **No Correlation**
  - $r = 0$

- **Low Negative Correlation**
  - $r = -0.3$

- **Highly Negative Correlation**
  - $r = -0.8$

- **Perfect Negative Correlation**
  - $r = -1$
Boxplots

boxplot(GPA91 ~ Gender, data=TR)
What's the problem here?
Alternative to a Scatterplot

GPA91~cut(SATT,4), data=TRES

SATT

GPA91

(559,742) (742,925) (925,1.11e+03) (1.11e+03,1.29e+03)
Bar Graph

`barplot(table(TR$orient, TR$Gender), beside=T, legend=T, axis.lty=1, ylim=c(0,100))`
What is this then?

Mileage by No. Cylinders and No. Gears

- error bars can be SEMs, SDs, or CIs - which are they here?
- the problem with error bars - unless the distribution is normal, they can be misleading

I borrowed this off the Internet, but this was drawn with R.
Data for the previous graph

> mtcars

<table>
<thead>
<tr>
<th></th>
<th>mpg</th>
<th>cyl</th>
<th>disp</th>
<th>hp</th>
<th>drat</th>
<th>wt</th>
<th>qsec</th>
<th>vs</th>
<th>am</th>
<th>gear</th>
<th>carb</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mazda RX4</td>
<td>21.0</td>
<td>6</td>
<td>160.0</td>
<td>110</td>
<td>3.90</td>
<td>2.620</td>
<td>16.46</td>
<td>0</td>
<td>1</td>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td>Mazda RX4 Wag</td>
<td>21.0</td>
<td>6</td>
<td>160.0</td>
<td>110</td>
<td>3.90</td>
<td>2.875</td>
<td>17.02</td>
<td>0</td>
<td>1</td>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td>Datsun 710</td>
<td>22.8</td>
<td>4</td>
<td>180.0</td>
<td>93</td>
<td>3.85</td>
<td>2.320</td>
<td>18.61</td>
<td>1</td>
<td>1</td>
<td>4</td>
<td>1</td>
</tr>
<tr>
<td>Hornet 4 Drive</td>
<td>21.4</td>
<td>6</td>
<td>258.0</td>
<td>110</td>
<td>3.08</td>
<td>3.215</td>
<td>19.44</td>
<td>1</td>
<td>0</td>
<td>3</td>
<td>1</td>
</tr>
<tr>
<td>Hornet Sportabout</td>
<td>18.7</td>
<td>8</td>
<td>360.0</td>
<td>175</td>
<td>3.15</td>
<td>3.440</td>
<td>17.02</td>
<td>0</td>
<td>0</td>
<td>3</td>
<td>2</td>
</tr>
<tr>
<td>Valiant</td>
<td>18.1</td>
<td>6</td>
<td>225.0</td>
<td>105</td>
<td>2.76</td>
<td>3.460</td>
<td>20.22</td>
<td>1</td>
<td>1</td>
<td>3</td>
<td>1</td>
</tr>
<tr>
<td>Duster 360</td>
<td>14.3</td>
<td>8</td>
<td>360.0</td>
<td>245</td>
<td>3.21</td>
<td>3.570</td>
<td>15.84</td>
<td>0</td>
<td>0</td>
<td>3</td>
<td>4</td>
</tr>
<tr>
<td>Merc 240D</td>
<td>24.4</td>
<td>4</td>
<td>146.7</td>
<td>62</td>
<td>3.69</td>
<td>3.190</td>
<td>20.00</td>
<td>1</td>
<td>0</td>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td>Merc 230</td>
<td>22.8</td>
<td>4</td>
<td>140.8</td>
<td>95</td>
<td>3.92</td>
<td>3.440</td>
<td>22.90</td>
<td>1</td>
<td>0</td>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td>Merc 280</td>
<td>19.2</td>
<td>6</td>
<td>167.6</td>
<td>123</td>
<td>3.92</td>
<td>3.440</td>
<td>18.30</td>
<td>1</td>
<td>0</td>
<td>4</td>
<td>2</td>
</tr>
<tr>
<td>Merc 280C</td>
<td>17.8</td>
<td>6</td>
<td>167.6</td>
<td>123</td>
<td>3.92</td>
<td>3.340</td>
<td>18.90</td>
<td>1</td>
<td>0</td>
<td>4</td>
<td>2</td>
</tr>
<tr>
<td>Merc 450SE</td>
<td>16.4</td>
<td>8</td>
<td>275.8</td>
<td>180</td>
<td>3.07</td>
<td>4.070</td>
<td>17.40</td>
<td>0</td>
<td>0</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>Merc 450SL</td>
<td>17.3</td>
<td>8</td>
<td>275.8</td>
<td>180</td>
<td>3.07</td>
<td>3.730</td>
<td>17.60</td>
<td>0</td>
<td>0</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>Merc 450SLC</td>
<td>15.2</td>
<td>8</td>
<td>275.8</td>
<td>180</td>
<td>3.07</td>
<td>3.780</td>
<td>18.00</td>
<td>0</td>
<td>0</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>Cadillac Fleetwood</td>
<td>10.4</td>
<td>8</td>
<td>472.0</td>
<td>205</td>
<td>2.93</td>
<td>5.520</td>
<td>17.98</td>
<td>0</td>
<td>0</td>
<td>3</td>
<td>4</td>
</tr>
<tr>
<td>Lincoln Continental</td>
<td>10.4</td>
<td>8</td>
<td>460.0</td>
<td>215</td>
<td>3.00</td>
<td>5.420</td>
<td>17.82</td>
<td>0</td>
<td>0</td>
<td>3</td>
<td>4</td>
</tr>
<tr>
<td>Chrysler Imperial</td>
<td>14.7</td>
<td>8</td>
<td>440.0</td>
<td>230</td>
<td>3.23</td>
<td>5.365</td>
<td>17.32</td>
<td>0</td>
<td>0</td>
<td>3</td>
<td>4</td>
</tr>
<tr>
<td>Fiat 128</td>
<td>32.4</td>
<td>4</td>
<td>78.7</td>
<td>66</td>
<td>4.08</td>
<td>2.200</td>
<td>19.47</td>
<td>1</td>
<td>1</td>
<td>4</td>
<td>1</td>
</tr>
<tr>
<td>Honda Civic</td>
<td>30.4</td>
<td>4</td>
<td>75.7</td>
<td>52</td>
<td>4.93</td>
<td>1.615</td>
<td>18.52</td>
<td>1</td>
<td>1</td>
<td>4</td>
<td>1</td>
</tr>
<tr>
<td>Toyota Corolla</td>
<td>33.9</td>
<td>4</td>
<td>71.1</td>
<td>65</td>
<td>4.22</td>
<td>1.835</td>
<td>19.90</td>
<td>1</td>
<td>1</td>
<td>4</td>
<td>1</td>
</tr>
<tr>
<td>Toyota Corona</td>
<td>21.5</td>
<td>4</td>
<td>120.1</td>
<td>97</td>
<td>3.70</td>
<td>2.465</td>
<td>20.01</td>
<td>1</td>
<td>0</td>
<td>3</td>
<td>1</td>
</tr>
<tr>
<td>Dodge Challenger</td>
<td>15.5</td>
<td>8</td>
<td>381.0</td>
<td>150</td>
<td>2.76</td>
<td>3.520</td>
<td>16.87</td>
<td>0</td>
<td>0</td>
<td>3</td>
<td>2</td>
</tr>
<tr>
<td>AMC Javelin</td>
<td>15.2</td>
<td>8</td>
<td>384.0</td>
<td>150</td>
<td>3.15</td>
<td>3.435</td>
<td>17.30</td>
<td>0</td>
<td>0</td>
<td>3</td>
<td>2</td>
</tr>
<tr>
<td>Camaro Z28</td>
<td>13.3</td>
<td>8</td>
<td>350.0</td>
<td>245</td>
<td>3.73</td>
<td>3.040</td>
<td>15.41</td>
<td>0</td>
<td>0</td>
<td>3</td>
<td>4</td>
</tr>
<tr>
<td>Pontiac Firebird</td>
<td>19.2</td>
<td>8</td>
<td>400.0</td>
<td>175</td>
<td>3.08</td>
<td>3.845</td>
<td>17.05</td>
<td>0</td>
<td>0</td>
<td>3</td>
<td>2</td>
</tr>
<tr>
<td>Fiat X1-9</td>
<td>27.3</td>
<td>4</td>
<td>79.0</td>
<td>66</td>
<td>4.08</td>
<td>1.935</td>
<td>18.90</td>
<td>1</td>
<td>1</td>
<td>4</td>
<td>1</td>
</tr>
<tr>
<td>Porsche 914-2</td>
<td>26.0</td>
<td>4</td>
<td>120.3</td>
<td>91</td>
<td>4.43</td>
<td>2.140</td>
<td>16.70</td>
<td>0</td>
<td>1</td>
<td>5</td>
<td>2</td>
</tr>
<tr>
<td>Lotus Europa</td>
<td>30.4</td>
<td>4</td>
<td>95.1</td>
<td>113</td>
<td>3.77</td>
<td>1.513</td>
<td>16.90</td>
<td>1</td>
<td>1</td>
<td>5</td>
<td>2</td>
</tr>
<tr>
<td>Ford Pantera L</td>
<td>15.8</td>
<td>8</td>
<td>351.0</td>
<td>264</td>
<td>4.22</td>
<td>3.170</td>
<td>14.50</td>
<td>0</td>
<td>1</td>
<td>5</td>
<td>4</td>
</tr>
<tr>
<td>Ferrari Dino</td>
<td>19.7</td>
<td>6</td>
<td>145.0</td>
<td>175</td>
<td>3.62</td>
<td>2.770</td>
<td>15.50</td>
<td>0</td>
<td>1</td>
<td>5</td>
<td>6</td>
</tr>
<tr>
<td>Maserati Bora</td>
<td>15.0</td>
<td>8</td>
<td>301.0</td>
<td>335</td>
<td>3.54</td>
<td>3.570</td>
<td>14.60</td>
<td>0</td>
<td>1</td>
<td>5</td>
<td>8</td>
</tr>
<tr>
<td>Volvo 142E</td>
<td>21.4</td>
<td>4</td>
<td>121.0</td>
<td>109</td>
<td>4.11</td>
<td>2.780</td>
<td>18.60</td>
<td>1</td>
<td>1</td>
<td>4</td>
<td>2</td>
</tr>
</tbody>
</table>

Data means and standard deviations:

> with(mtcars, tapply(mpg, list(gear, cyl), mean))

```
4 6 8
3 21.50 19.75 15.05
4 26.925 19.75 NA
5 28.200 19.70 15.40
```

> with(mtcars, tapply(mpg, list(gear, cyl), sd))

```
4 6 8
3 NA 2.333452 2.774395
4 4.80736 1.552417 NA
5 3.1127 NA 0.5656854
```

mtcars \{datasets\}

Motor Trend Car Road Tests

Description

The data was extracted from the 1974 Motor Trend US magazine, and comprises fuel consumption and 10 aspects of automobile design and performance for 32 automobiles (1973–74 models).
Profile Plot

![Graph showing the relationship between mtcars$engine_cyl and mean of mtcars$mpg across different values of mtcars$gear. The graph illustrates the trend where the mean mpg increases with the number of engine cylinders, with distinct markers for different number of gears.]
Why Are Variables Related ("Correlated")

"Correlation does not necessarily imply causation."

True (randomized, designed) experiments are the best way to find causation.