# 9. Bivariate covariance and correlation

LPO.8800: Statistical Methods in Education Research

Sean P. Corcoran



## Last time: Tests for comparing two groups

- Confidence interval and test for difference in two population means ($\mu_1$ and $\mu_2$)
- Confidence interval and test for difference in two population proportions ($\pi_1$ and $\pi_2$)
- Independent vs. dependent samples
- Paired sample t-test
- Statistical power for two-sample tests

## Introduction

Most everything we have done thus far relates to *univariate* distributions:

- Descriptive statistics ($e.g., \bar{x}, s^2$, centiles)
- Transformations (e.g., $z$-scores, log)
- Probability distributions (e.g., normal, binomial, uniform) and population parameters ($e.g., \mu, \sigma^2$)
- Random sampling and inferences about population parameters (confidence intervals, hypothesis testing)



## Introduction

The most interesting applications of statistics examine the relationship between two or more variables:

- Do women earn less than men?
- Does education increase earnings?
- Does smoking cause lung cancer?
- Do students learn more in small classes than in large ones?
- Does lower birth weight increase the risk of poor health outcomes later in life?

These are examples of **bivariate analyses**. In the last lecture we saw one example of a bivariate analysis, where one variable (the group identifier) was binary. The second variable was either continuous or binary.

## Introduction

Today we will turn to *bivariate* distributions and measures of association, or covariance.



## Review of univariate probability distributions

A probability distribution for $X$ assigns probabilities to values of $X$ or to intervals of values. PDFs come in discrete and continuous types, depending on the nature of the random variable.

The PDF $f(x_i)$ for a discrete random variable $X$ provides the probability that $X = x_i$ for possible values of $X$:

$$f(x) = P(X = x_i)$$

The PDF $f(x)$ for a continuous random variable $X$ provides the probability that $X$ is between certain values. E.g., between $a$ and $b$:

$$P(a \leq X \leq b) = \int_a^b f(x)dx$$

# Review of univariate probability distributions

The probabilities in a discrete PDF are nonnegative and must sum to one. If there are $N$ possible outcomes indexed by $i$:

$$\sum_{i=1}^{N} P(X = x_i) = 1$$

The area under a continuous PDF (i.e., the integral) must equal one:

$$P(-\infty \leq X \leq +\infty) = \int_{-\infty}^{+\infty} f(x)dx = 1$$



# Review of univariate probability distributions

Two important features of a PDF are its **expected value** (mean) and **variance**, the first and second population *moments*.

For a discrete random variable $X$ with $N$ unique outcomes indexed by $i$, the expected value of $X$ is:

$$E(X) = \mu_x = \sum_{i=1}^{N} x_i P(x_i)$$

The variance of $X$ is:

$$Var(X) = \sigma_x^2 = E(x - E(X))^2 = \sum_{i=1}^{N} (x_i - E(X))^2 P(x_i)$$

# Review of univariate probability distributions

For a continuous random variable $X$, the expected value of $X$ is:

$$E(X) = \mu_x = \int_{-\infty}^{+\infty} xf(x)dx$$

The variance of $X$ is:

$$Var(X) = \sigma_x^2 = E(x - E(X))^2 = \int_{-\infty}^{+\infty} (x - E(X))^2 f(x)dx$$



# Joint probability distributions

A **joint probability distribution** for two random variables $X$ and $Y$ assigns probabilities to values of $(X, Y)$ or to intervals of values. Like univariate PDFs, joint PDFs are discrete or continuous.

The PDF $f(x, y)$ for a pair of discrete random variables $X$ and $Y$ provides the probability that $X = x$ and $Y = y$ for possible values $x$ and $y$:

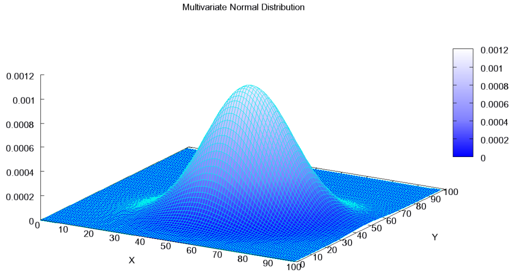
$$f(x, y) = P(X = x \text{ and } Y = y)$$

The PDF for continuous random variables $X$ and $Y$ provides the probability that $X$ is between $a$ and $b$ and $Y$ is between $c$ and $d$:

$$\int_a^b \int_c^d f(x, y)dydx$$

## Example: bivariate normal distribution

An example joint PDF: bivariate normal distribution






## Joint probability distributions

The probabilities in a discrete joint PDF are nonnegative and must sum to one. If there are $N$ possible outcomes for $X$ indexed by $i$, and $M$ possible outcomes for $Y$ indexed by $j$:

$$\sum_{i=1}^{N}\sum_{j=1}^{M} P(X = x_i \text{ and } Y = y_j) = 1$$

The area under a continuous joint PDF (i.e., the double integral) must equal one:

$$P(-\infty \leq X \leq +\infty \text{ and } -\infty \leq Y \leq +\infty) = \int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} f(x,y)dydx = 1$$

## Measures of association

Bivariate probability distributions allow us to think about how two variables are related in the population:

- *Direction*: are the variables positively correlated, negatively correlated, or uncorrelated?
- *Shape*: is the relationship between the two variables *linear* or *nonlinear*?
- *Strength*: is the relationship between the two variables strong, weak, or moderate?

The joint PDF tells us nothing about whether the relationship is a *causal* one or not.



## Covariance

The population **covariance** between two random variables $X$ and $Y$ is:

$$Cov(X,Y) = \sigma_{xy} = E\left[(x - E(X))(y - E(Y))\right]$$

For a discrete joint PDF:

$$Cov(X,Y) = \sigma_{xy} = \sum_x \sum_y (x - E(X))(y - E(Y)) f(x,y)$$

For a continuous joint PDF:

$$Cov(X,Y) = \sigma_{xy} = \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} (x - E(X))(y - E(Y)) f(x,y) dydx$$

## Covariance

The covariance is like a weighted average of products: $X$'s deviation from its mean multiplied by $Y$'s deviation from its mean.

- If $Y$ tends to be higher than average when $X$ is higher than average, these products will tend to be positive (a **positive covariance**).
- If $Y$ tends to be *lower* than average when $X$ is higher than average, these products will tend to be negative (a **negative covariance**).

The weights come from the joint probability of $X$ and $Y$.



## Covariance

Some facts about covariance:

- The magnitude of the covariance depends on the units of $X$ and $Y$
- $Cov(X, Y) = E(XY) - E(X)E(Y)$
- If $X$ and $Y$ are independent, $Cov(X, Y) = 0$
- Independence implies zero covariance but the converse is not true.
- Covariance is a measure of *linear* association
- If $Y = X$, $Cov(X, X) = Var(X)$. (The covariance of a variable with itself is just its variance).

# Correlation

**Correlation** is a standardized, or unit-free measure of covariance:

$$Corr(X,Y) = \rho_{xy} = E\left[\left(\frac{x - E(X)}{\sigma_x}\right)\left(\frac{y - E(Y)}{\sigma_y}\right)\right]$$

$$= \frac{Cov(X,Y)}{\sigma_x \sigma_y} = \frac{\sigma_{xy}}{\sigma_x \sigma_y}$$



# Correlation

Some facts about correlation:

- Like covariance, $\rho$ is a measure of *linear* association
- $-1 \leq \rho_{xy} \leq 1$
  - $\rho_{xy} = 1$ is a perfect positive correlation
  - $\rho_{xy} = -1$ is a perfect negative correlation
  - $\rho_{xy} = 0$ is no correlations
- $\rho_{xy}$ requires both $\sigma_x$ and $\sigma_y$ to be positive (i.e., not zero).

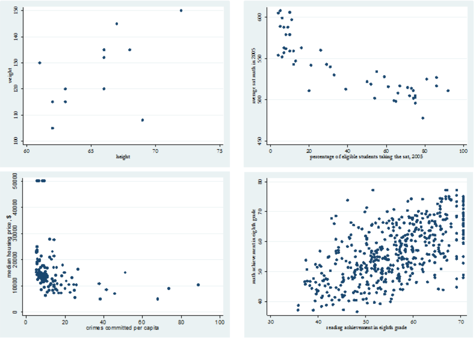
# Scatter diagrams

The easiest way to see how two variables are associated using data is via a *scatter diagram* or *scatterplot*.

- In Stata: `scatter` *yvar xvar*
- Appropriate for variables that are at least interval measured
- Scatter diagrams can provide a sense of *direction* of relationship (if any), *linearity*, and *strength* of association



# Scatter diagrams



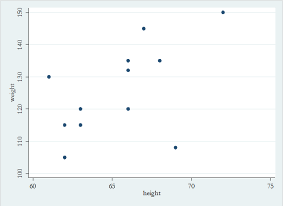
# Scatter diagrams

Often with scatter diagrams there is a natural **response** or **outcome**, and **explanatory** variable. We may have in mind a theory in which variation in the response is (at least in part) explained by variation in the explanatory variable.

- Denote the outcome as $Y$ and explanatory variable as $X$.
- Some use the terms "dependent" and "independent" variables. I prefer *not* to use these, given these terms' other meanings in statistics.

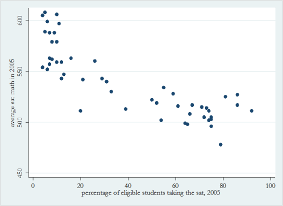


# Scatter diagram 1: height and weight




- Positive association
- Mostly linear–a line would describe this relationship well
- Moderately strong association

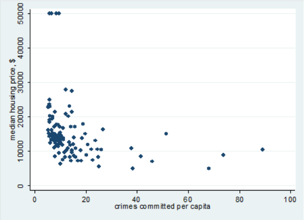
## Scatter diagram 2: percent taking SAT and scores




- Negative association
- Quite linear
- Strong association

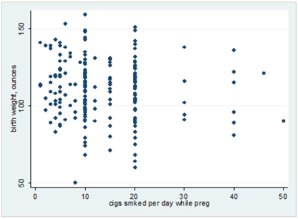


## Scatter diagram 3: crime rate and median house price




- Negative association
- *Nonlinear*
- Strong nonlinear association

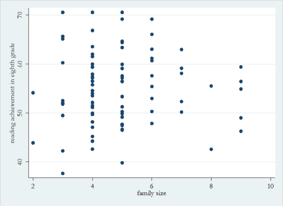
# Scatter diagram 4: maternal smoking and birthweight




- Cigarettes per day while pregnant, and birthweight (ounces)
- Zero to negative association
- Linearity not obvious
- Weak association

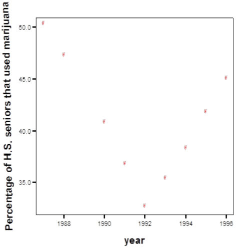


# Scatter diagram 5: family size and NELS achievement




- 8th grade reading scores in West region
- Zero to negative association
- Linearity not obvious
- Weak association

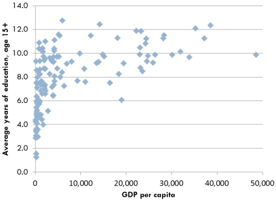
# Scatter diagram 6: marijuana usage and time




- No *linear* association
- Strong *nonlinear* association / time trend

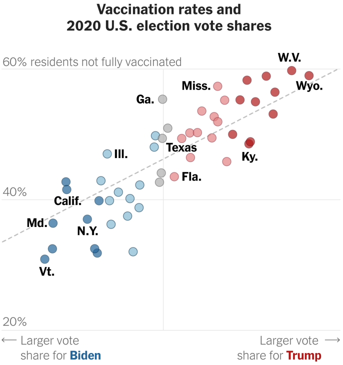


# Scatter diagram 7: GDP per capita and education



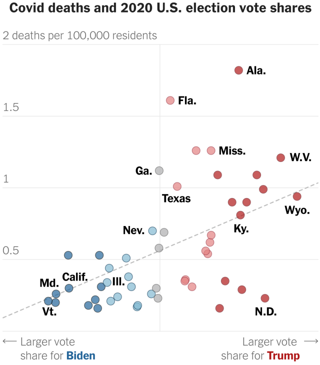

- Positive association
- *Nonlinear* relationship
- Moderately strong association

# Scatter diagram 8: Trump vote share and vaccination



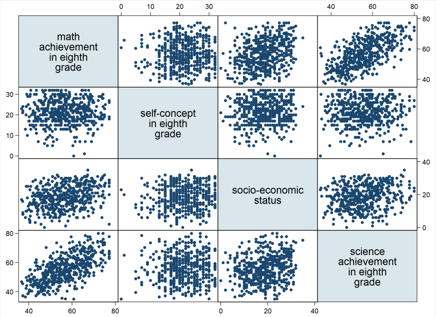



# Scatter diagram 9: Trump vote share and COVID deaths



# Scatterplot matrix

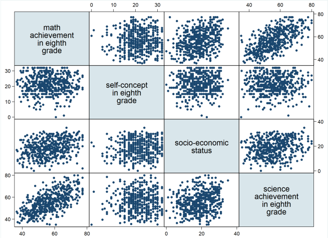
`graph matrix` *varlist* is a useful command for visualizing the bivariate associations between two or more variables. For example:






# Scatterplot matrix

The horizontal axes in each column apply to the variable named in that column; the vertical axes apply to the variable name in that row. Thus, for example, the scatterplot in cell (3,1) is the same as the scatterplot in cell (1,3), but with the axes flipped.



# Binscatter

With large datasets scatterplots may be difficult to read. The user-written command `binscatter` groups the x-axis into bins and computes the mean for each bin. The command has lots of options, including a fitted line.

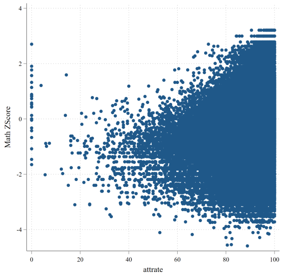
`binscatter` *yvar xvar*, `options`

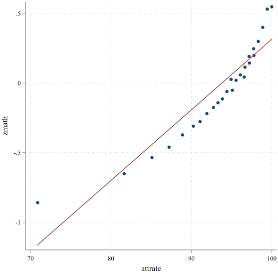
You can change the number of bins (default is 20), whether the fitted line is linear or quadratic or connected, etc.



# Binscatter

Math scores and annual school attendance in a large dataset ($N > 420,000$)



## Sample covariance and correlation

The sample analog to the population covariance is:

$$s_{xy} = \frac{\sum_{i=1}^{N} (x_i - \bar{x})(y_i - \bar{y})}{n-1}$$

The sample analog to the population correlation is:

$$r_{xy} = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{s_x s_y (n-1)} = \frac{s_{xy}}{s_x s_y}$$

$r_{xy}$ is known as the sample *correlation coefficient* or the Pearson product moment correlation. The sample $s_{xy}$ and $r_{xy}$ are estimators of the population parameters $\sigma_{xy}$ and $\rho_{xy}$.



## Covariance and correlation—calculation

| | Height ($x_i$) | Weight ($y_i$) | $x_i - \bar{x}$ | $y_i - \bar{y}$ | $(x_i - \bar{x}) \times (y_i - \bar{y})$ |
|---|---|---|---|---|---|
| | 69 | 108 | 3.58 | -17.83 | -63.90 |
| | 61 | 130 | -4.42 | 4.17 | -18.40 |
| | 68 | 135 | 2.58 | 9.17 | 23.68 |
| | 66 | 135 | 0.58 | 9.17 | 5.35 |
| | 66 | 120 | 0.58 | -5.83 | -3.40 |
| | 63 | 115 | -2.42 | -10.83 | 26.18 |
| | 72 | 150 | 6.58 | 24.17 | 159.10 |
| | 62 | 105 | -3.42 | -20.83 | 71.18 |
| | 62 | 115 | -3.42 | -10.83 | 37.01 |
| | 67 | 145 | 1.58 | 19.17 | 30.35 |
| | 66 | 132 | 0.58 | 6.17 | 3.60 |
| | 63 | 120 | -2.42 | -5.83 | 14.10 |
| Mean | 65.42 | 125.83 | 0.00 | 0.00 | 23.74 |
| Sum | | | 0.00 | 0.00 | 284.85 |
| SD | 3.32 | 14.24 | | | |

## Covariance and correlation—calculation

$$s_{xy} = \frac{\sum_{i=1}^{N} (x_i - \bar{x})(y_i - \bar{y})}{n-1}$$

$$s_{xy} = \frac{284.85}{12-1} = 25.895$$

$$r_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{s_x s_y (n-1)}$$

$$r_{xy} = \frac{284.85}{(3.32)(14.24)(12-1)} = 0.548$$



## Covariance and correlation in Stata

To obtain correlation coefficients in Stata, use `corr` *yvar xvar*. The result is a correlation matrix.

- Be aware of how Stata handles missing values:
  - **listwise deletion** means observations are not used if *any* of the listed variables in the command are missing.
  - **pairwise deletion** means correlations of pairs of variables are considered in isolation.
- `pwcorr` *yvar xvar* uses pairwise deletion. `corr` uses listwise deletion.

To obtain the covariance in Stata, use `corr` with `cov` option. The result is a **variance-covariance matrix**.

# Correlation coefficient in Stata

Examples:

```
. corr height weight
(obs=12)

             |   height   weight
-------------+------------------
      height |   1.0000
      weight |   0.5486   1.0000
```

```
. corr height weight, cov
(obs=12)

             |   height   weight
-------------+------------------
      height |  10.9924
      weight |  25.8939  202.697
```



# Correlation coefficient in Stata

Examples:

```
. corr satv satm pertak
(obs=51)

             |     satv     satm   pertak
-------------+---------------------------
        satv |   1.0000
        satm |   0.9715   1.0000
      pertak |  -0.8827  -0.8372   1.0000
```

```
. corr famsize achrdg08 if region==4
(obs=93)

             |  famsize achrdg08
-------------+------------------
     famsize |   1.0000
    achrdg08 |   0.0260   1.0000
```

```
. corr year marij
(obs=16)

             |     year    marij
-------------+------------------
        year |   1.0000
       marij |   0.3963   1.0000
```

# Correlation coefficient in Stata

Examples:

```
. corr achmat08 achrdg08
(obs=500)

             | achmat08 achrdg08
-------------+------------------
    achmat08 |   1.0000
    achrdg08 |   0.5947   1.0000

. pwcorr achmat08 achrdg08 achmat10 achrdg10

             | achmat08 achrdg08 achmat10 achrdg10
-------------+------------------------------------
    achmat08 |   1.0000
    achrdg08 |   0.5947   1.0000
    achmat10 |   0.8489   0.5919   1.0000
    achrdg10 |   0.5803   0.7538   0.6531   1.0000
```



# Correlation coefficient

What is a "strong correlation?" It depends on the context. (How strong would you expect the correlation to be? Is there a theoretical reason why the correlation should be particularly strong or weak?)

- Rule of thumb ("Cohen's scale") based on the absolute value $|r_{xy}|$:
  - $|r_{xy}| < 0.1$: zero to weak correlation
  - $0.1 < |r_{xy}| < 0.3$: weak to moderate correlation
  - $0.3 < |r_{xy}| < 0.5$: moderately strong correlation
  - $|r_{xy}| > 0.5$: strong correlation

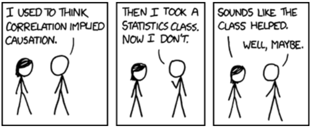
The correlation coefficient itself is *ordinal*. An increase in correlation from 0.1 to 0.2 is not equivalent to an increase from 0.4 to 0.5.

Try "guess the correlation:"
`https://istats.shinyapps.io/guesscorr/`

# Correlation vs. causation

Important: *correlation does not imply causation!*

- *Correlation* means two variables move together.
- *Causation* means that change in one variable is causing change in the other.

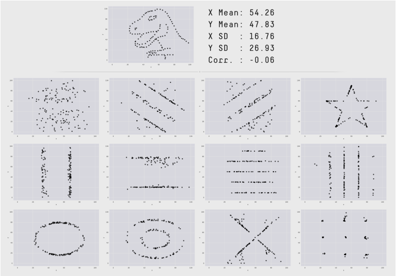






# The importance of visualizing your data

Never trust summary statistics alone! All of the datasets used below have the same $\bar{x}, \bar{y}, s_x, s_y$, and $r_{xy}$.




Recommended source:
`https://www.autodesk.com/research/publications/same-stats-different-graphs`

# Correlation coefficient—special cases

The Pearson product moment correlation can be applied to any pair of interval-measured variables. However, when one or both of the variables is dichotomous, the correlation coefficient can be expressed in alternative ways.

- **"Point biserial" correlation**: when one variable is dichotomous ($x$) and the other is continuous ($y$).
- **"Phi coefficient"**: when both variables are dichotomous.

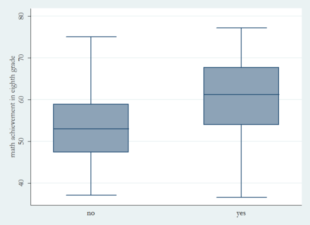
The following slides simply show how $r_{xy}$ can be written when one or both variables are dichotomous. In practice, continue to use `corr` or `pwcorr` in Stata. The alternative formulas may be useful in some cases.



# Point biserial correlation

$$r_{xy} = \frac{(\bar{y}_1 - \bar{y}_0)s_x}{s_y}$$

$\bar{y}_1$ is the mean of $y$ for observations where $x = 1$, and $\bar{y}_0$ is the mean of $y$ for observations where $x = 0$. Consider the relationship between enrollment in advanced math and math achievement, in the NELS:



# Point biserial correlation

```
. sum achmat08 advmath

    Variable |        Obs        Mean    Std. Dev.       Min        Max
-------------+---------------------------------------------------------
    achmat08 |        500    56.59102    9.339608      36.61       77.2
    advmath8 |        491    .4602851    .4989286          0          1

. sum achmat08 advmath if advmath~=.

    Variable |        Obs        Mean    Std. Dev.       Min        Max
-------------+---------------------------------------------------------
    achmat08 |        491    56.73473    9.342372      36.61       77.2
    advmath8 |        491    .4602851    .4989286          0          1
```

```
. tabstat achmat08, by(advmath) stat(mean sd n)

Summary for variables: achmat08
     by categories of: advmath8 (advanced math t

advmath8 |      mean        sd         N
---------+------------------------------
      no |  53.57491  8.098492       265
     yes |  60.43982  9.358113       226
---------+------------------------------
   Total |  56.73473  9.342372       491
----------------------------------------
```



# Point biserial correlation

$$r_{xy} = \frac{(\bar{y}_1 - \bar{y}_0)s_x}{s_y}$$

$$r_{xy} = \frac{(60.45 - 53.57)(0.499)}{9.342} = 0.367$$

```
. corr achmat08 advmath
(obs=491)

             | achmat08 advmath8
-------------+------------------
    achmat08 |   1.0000
    advmath8 |   0.3666   1.0000
```

## Phi coefficient

When the two variables $x$ and $y$ are dichotomous, we can calculate the Pearson correlation coefficient (also referred to as a "Phi coefficient") from a 2 x 2 frequency table:

| | Variable 2: | |
|---|---|---|
| Variable 1: | 0 | 1 |
| 0 | A | B |
| 1 | C | D |

$$r_{xy} = \phi = \frac{AD - BC}{\sqrt{(A+B)(C+D)(A+C)(B+D)}}$$



## Phi coefficient

Example:

| | Gender: | |
|---|---|---|
| Advanced math: | Male | Female |
| No | 25 | 36 |
| Yes | 22 | 21 |

$$\phi = \frac{AD - BC}{\sqrt{(A+B)(C+D)(A+C)(B+D)}}$$

$$\phi = \frac{(25 * 21) - (36 * 22)}{\sqrt{(25+36)(22+21)(25+21)(36+22)}} = -0.101$$

# Phi coefficient

| advanced math taken in eighth grade | gender male | female | Total |
|---|---|---|---|
| no | 25 | 36 | 61 |
| | 53.19 | 63.16 | 58.65 |
| yes | 22 | 21 | 43 |
| | 46.81 | 36.84 | 41.35 |
| Total | 47 | 57 | 104 |
| | 100.00 | 100.00 | 100.00 |

```
. corr advmath gender if region==1
(obs=104)
```

| | advmath8 | gender |
|---|---|---|
| advmath8 | 1.0000 | |
| gender | -0.1007 | 1.0000 |



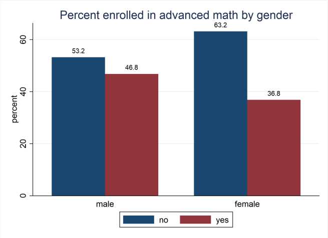
# Phi coefficient

Other ways to observe an association between two dichotomous variables:

- **Clustered bar graph** (a bar graph by group), where the height of the bar is the percentage of cases equal to one within each group
- Contingency table (or **crosstabulation**) with row and column percentages

# Sample clustered bar graph

```
graph bar if region==1, over(advmath8) over(gender) asyvars percentages
blabel(bar, format(%3.1f)) title(Percent enrolled in advanced math by gender)
graphregion(fcolor(white))
```




The `asyvars percentages` options ensure the bar heights represent percentages within group (and sum to 100). Here, girls appear less likely to be enrolled in advanced math.



# 2x2 crosstabulation

```
tabulate gender advmath if region==1, row
```

| Key |
|---|
| *frequency*<br>*row percentage* |

| gender | advanced math taken in eighth grade<br>no | <br>yes | Total |
|---|---|---|---|
| male | 25<br>53.19 | 22<br>46.81 | 47<br>100.00 |
| female | 36<br>63.16 | 21<br>36.84 | 57<br>100.00 |
| Total | 61<br>58.65 | 43<br>41.35 | 104<br>100.00 |

The `row` option reports percentages that sum to 100 in each row. Here again, girls appear less likely to be enrolled in advanced math.

# Spearman rank correlation

Spearman's rank correlation (sometimes called "rho") can be used with ordinal-measured variables—where the size of the difference between $x$ and its mean is not meaningful—or in cases where the underlying relationship is nonlinear.



# Spearman rank correlation

Spearman's correlation depends on how variables *rank* in their respective distributions.

- Rank each variable $x$ and $y$ in its respective distribution, in ascending order $1, ..., n$
- For an observation $i$, $d_i$ is the *difference* between $i$'s ranking for $x$ and $i$'s ranking for $y$: $d_i = rank(x_i) - rank(y_i)$

$$r_{s,xy} = 1 - \frac{6 \sum_{i=1}^{n} d_i^2}{n^3 - n}$$

# Spearman rank correlation

A few notes about Spearman's correlation:

- When $x_i$ and $y_i$ are identically ranked, $d_i = 0$. If $d_i = 0$ for *all cases*, then $r_{s,xy} = 1$ (a perfect positive correlation).

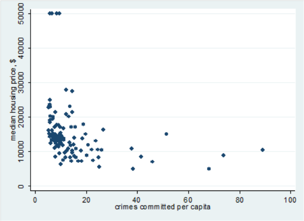
- The further apart $x_i$ and $y_i$ are in their ranks, the larger is $d_i$ and $r_{s,xy}$ gets closer to -1 (a perfect negative correlation).

$$r_{s,xy} = 1 - \frac{6\sum_{i=1}^{n} d_i^2}{n^3 - n}$$



# Spearman rank correlation in Stata

The Spearman rank correlation is obtained in Stata with the command `spearman` *yvar xvar*. For example, consider the nonlinear relationship between median house price and crime:



# Spearman rank correlation in Stata

```
. spearman price crime

 Number of obs =       506
Spearman's rho =       -0.5564

Test of Ho: price and crime are independent
    Prob > |t| =       0.0000

. sum price crime

    Variable |        Obs        Mean    Std. Dev.       Min        Max
-------------+---------------------------------------------------------
       price |        506    22511.51    9208.856       5000      50001
       crime |        506    3.611536    8.590247       .006     88.976

. corr price crime
(obs=506)

             |  price   crime
-------------+------------------
       price | 1.0000
       crime | -0.3879  1.0000
```



# Variable transformations and correlation

Consider the linear transformations of variables $x$ and $y$:

$$x_1 = a + (b * x_0)$$

$$y_1 = c + (d * y_0)$$

What happens to the correlation between two variables $x$ and $y$ when one (or both) is transformed by a linear function?

# Variable transformations and correlation

Re-write the correlation coefficient as follows.

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{s_x s_y (n-1)} = \frac{\sum_{i=1}^{n}(\frac{x_i - \bar{x}}{s_x})(\frac{y_i - \bar{y}}{s_y})}{n-1} = \frac{\sum_{i=1}^{n} z_{xi} z_{yi}}{n-1}$$

How is a $z$-score affected by a linear transformation?

$$z_{x_1} = \frac{a + (b * x_0) - a - (b * \bar{x_0})}{|b| * s_{x_0}} = \frac{b * (x_0 - \bar{x_0})}{|b| * s_{x_0}} = \frac{b}{|b|} z_{x_0}$$

When $b$ is positive, the $z$-score of the transformed variable $x_1$ is the same as the original $z$-score. When $b$ is *negative*, the new $z$-score is the inverse of the original.



# Variable transformations and correlation

How transformations affect the correlation between $x$ and $y$ thus depend only on the multiplicative factor applied to $x$, $y$, or both:

- If both multiplicative factors are positive ($b > 0$ and $d > 0$) or both multiplicative factors are negative ($b < 0$ and $d < 0$), then the correlation between the new, transformed variables is the same as the correlation between the original variables.

- If one multiplicative factor is positive and the other is negative ($b > 0$ and $d < 0$ or ($b < 0$ and $d > 0$), then the correlation between the new, transformed variables is the negative of the correlation between the original variables.

# Variable transformations and covariance

Plug the transformed values of $x_0$ and $y_0$ into the covariance:

$$s_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{n-1}$$

$$s_{xy,new} = \frac{\sum_{i=1}^{n}(a + bx_{0i} - a - b\bar{x}_0)(c + dy_{0i} - c - d\bar{y}_0)}{n-1}$$

$$= \frac{\sum_{i=1}^{n}(bx_{0i} - b\bar{x}_0)(dy_{0i} - d\bar{y}_0)}{n-1}$$

$$= \frac{\sum_{i=1}^{n} bd(x_{0i} - \bar{x}_0)(y_{0i} - \bar{y}_0)}{n-1}$$

$$= bds_{xy,old}$$



# Variable transformations and covariance

That is, the covariance of the two transformed variables is equal to the covariance of the original (untransformed) variables times $bd$.

## Hypothesis tests about $\rho$

The sample correlation $r_{xy}$ can be used to estimate the population correlation coefficient $\rho$. If we know the sampling distribution of $r_{xy}$, we can construct confidence intervals for $\rho$ or conduct hypothesis tests.

Under $H_0 : \rho = 0$, the following test statistic has a $t$-distribution with $n-2$ degrees of freedom:

$$t = \frac{|r_{xy}|\sqrt{n-2}}{\sqrt{1-r_{xy}^2}}$$

Find $p$ and then determine whether $p < \alpha$. In Stata, use `pwcorr` with the `sig` option. Note this $t$ statistic is $|r-0|$ divided by the standard error



## Hypothesis tests about $\rho$

Example:

```
. pwcorr height weight, sig

             |   height   weight
-------------+------------------
      height |   1.0000
             |
             |
      weight |   0.5486   1.0000
             |   0.0648
             |

. display (0.5486*sqrt(10))/sqrt(1-0.5486^2)
2.0749393

. display 2*ttail(10,2.0749393)
.06474521
```

0.0648 is the $p$-value for the two-sided hypothesis test with $H_0 : \rho = 0$.

# Hypothesis tests about $\rho$

Assumptions:

- Independent random sample from the population
- $x$ and $y$ follow a bivariate normal distribution
- If sample size is large enough, normality is not required and test statistic follows an approximate $t$ distribution.



# Hypothesis tests about $\rho$

Note the pwcorr includes a bonferroni option that adjusts $p$-values for multiple hypotheses:

```
. pwcorr price mpg headroom weight gear_ratio, sig bonf

             |    price      mpg headroom   weight gear_r~o
-------------+---------------------------------------------
       price |   1.0000
             |
             |
         mpg |  -0.4686   1.0000
             |   0.0003
             |
    headroom |   0.1145  -0.4138   1.0000
             |   1.0000   0.0025
             |
      weight |   0.5386  -0.8072   0.4835   1.0000
             |   0.0000   0.0000   0.0001
             |
  gear_ratio |  -0.3137   0.6162  -0.3779  -0.7593   1.0000
             |   0.0650   0.0000   0.0090   0.0000
```

$p$-values have been multiplied by the number of pairwise tests $g(g-1)/2$, where $g$ is the number of variables. Above: 10

# Hypothesis tests about $\rho$

Confidence intervals for $\rho$ are not part of the `corr` command. It is not as straightforward as using the $t$ test statistic above to construct the interval, since this is the distribution under the null of no correlation.

When $|\rho| > 0$, the sampling distribution is actually quite *skewed*, even with large samples. This is because the correlation coefficient is constrained to be between -1 and $+1$.



# Fisher's $z$

Fisher showed that a transformation of $r$ has an approximate normal distribution (the inverse hyperbolic tangent transformation):

$$z' = 0.5 \ln \left[ \frac{1+r}{1-r} \right]$$

The standard error of $z'$ is $1/\sqrt{n-3}$. You can calculate a $(1-\alpha)\%$ confidence interval using $z'$ and then transform back into the correlation scale.

This is not a built-in Stata function, but see the user-written commands `corrci` and `corrcii` and Cox (2008).

Cox (2008): `https://journals.sagepub.com/doi/pdf/10.1177/1536867X0800800307`

# Simulating draws from a bivariate normal in Stata

Example of 100 draws of $X$ and $Y$ from a bivariate normal distribution with $\mu_x = \mu_y = 0$, $\sigma_x^2 = \sigma_y^2 = 1$ and $\rho_{xy} = 0.5$:

```
matrix C = (1 0.5 \ 0.5 1)
drawnorm x y, n(100) corr(C)
```

The `corr` option in `drawnorm` takes a matrix which tells Stata how the two simulated variables are correlated. Note the covariance in this case is also 0.5 since $\rho = \sigma_{xy}/\sigma_x\sigma_y$ and $\sigma_x^2 = \sigma_y^2 = 1$.

$$C = \begin{bmatrix} 1.0 & 0.5 \\ 0.5 & 1.0 \end{bmatrix}$$