PUBH 7430
Lecture 2

Erika Helgeson

Division of Biostatistics
University of Minnesota School of Public Health

Notes adapted from material provided by Drs. Julian Wolfson and Ashley Petersen

Announcements

- TA office hour doodle poll (Friday 6PM)
- **Group Project/ First week check-in survey (Monday 6PM, 5 points)**
- Assignment 0 (Before class Tuesday)

Quick Review

Syllabus

Quick Review

Types of studies with background correlation

- Longitudinal studies
  - Repeated measurements on the same individuals
- Studies with ‘clusters’ of observations
  - Examples: cluster randomized trials; studies of family units, schools, hospitals, social networks, etc
- Geospatial studies
  - Expect observations closer together to be more strongly correlated (e.g. pollution and oil spill study)
- Studies can be a combination!
  - Longitudinal study of kids within schools

Quick Review

The Four Skills: How to handle correlated data.

1. **Recognition:** Identify situations where data may be correlated.
2. **Description:** Visualize and summarize correlated data in an informative way.
3. **Modeling/estimation:** Choose and fit statistical models which account for correlation and respond to the scientific question.
4. **Inference/interpretation:** Correctly interpret the results of analyses; understand their assumptions and the potential consequences if they are violated.

Today's topic: Independence, dependence, and covariance

Independence

Two events A and B are said to be **independent** if

P(A occurs and B occurs)$=$ P(A occurs)$\times$ P(B occurs)

Stats Example

Probability of getting two "heads" when flipping a fair-sided coin twice.
$\text{P(two "heads")} = \text{P}(\text{flip}_1 = \text{"heads"}, \text{flip}_2 = \text{"heads"})$
$= \text{P}(\text{flip}_1 = \text{"heads"})(\text{flip}_2 = \text{"heads"}) = 1/4$

Independence

Two events A and B are said to be **independent** if

P(A occurs and B occurs)$=$ P(A occurs)$\times$ P(B occurs)

Everyday Examples

- Ate bananas and attended a birthday party
- Ate bananas and flew on a plane
- Ate bananas and went to a movie

Independence

We commonly refer to events A and B as "correlated" if they are **not independent**:

$$P(A \text{ occurs and } B \text{ occurs}) \neq P(A \text{ occurs}) \times P(B \text{ occurs})$$

Stats Example

Change the coin flip experiment

- Flip the coin once.
- If the first flip is **tails**, flip the coin again and record outcome of second flip.
- If the first flip is **heads**, flip the coin two more times. Record second outcome as heads if heads come up in second or third flip, and as tails if no heads come up.

P("two heads") = P(flip_1="heads")× P(flip_2="heads" or flip_3="heads") = $1/2 \times 3/4 = 3/8$

Independence

We commonly refer to events A and B as "correlated" if they are **not independent**:

$$P(A \text{ occurs and } B \text{ occurs}) \neq P(A \text{ occurs}) \times P(B \text{ occurs})$$

Everyday Examples

- Ate cake and attended a birthday party
- Ate peanuts and flew on a plane
- Ate popcorn and went to the movies

Conditioning and independence

An equivalent definition of independence is

$$P(A \text{ occurs} | B \text{ occurs}) = P(A \text{ occurs})$$

where $P(A|B)$ is read as "probability of A, given B".

Stats Example, cont'd.

P(second outcome= "tails" | flip_1="tails") = 3/4
P(second outcome= "tails")= 5/8
P(second outcome= "tails" | flip_1="tails") $\neq$ P(second outcome= "tails")

Conditioning and independence

An equivalent definition of independence is

$$P(A \text{ occurs}|B \text{ occurs}) = P(A \text{ occurs})$$

where $P(A|B)$ is read as "probability of A, given B".

Everyday example, cont'd.

Knowing whether I ate bananas on a certain day doesn't help you predict if I went to a birthday party

Conditional independence

Events A and B are said to be **conditionally independent** given that event C occurs if

$$P(A, B|C) = P(A|C) \times P(B|C)$$

or equivalently

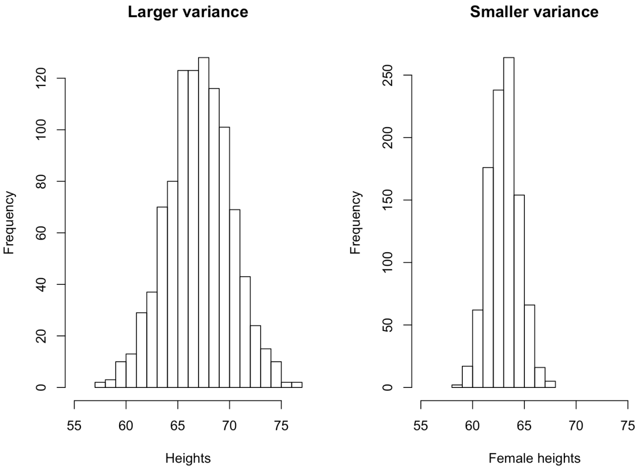
$$P(A|B, C) = P(A|C)$$

Example

Asthma risk of two (unrelated) people is independent given (i.e., conditional on) where they live.

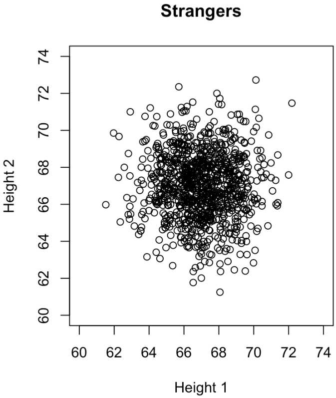
Variance

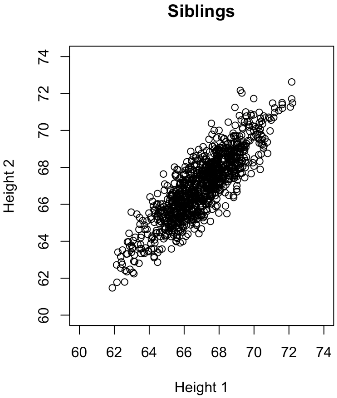
$$Var(Y) = E[(Y - E[Y])^2]$$



Exercise!

- You sample pairs (X, Y) of heights from (1) strangers and (2) siblings
- *Q: For which group is $Var(X + Y)$ larger?*
- *Q: For which group is $Var(X - Y)$ larger?*





Variance of sum

Variance of difference

Covariance

Let X and Y be two random variables (possibly dependent). What is the variance of $X + Y$ And $X - Y$?

$$Var(X+Y) = Var(X) + Var(Y) + 2[E(XY) - E(X)E(Y)]$$

$$Var(X-Y) = Var(X) + Var(Y) - 2[E(XY) - E(X)E(Y)]$$

We give the part in red a special name, the **covariance**:

$$Cov(X,Y) = E(XY) - E(X)E(Y),$$

which tells us how the variables "move" together

Back to our example

Covariance – properties

$$\begin{aligned} Cov(X, Y) &= E(XY) - E(X)E(Y) \\ &\equiv E[(X - E(X))(Y - E(Y))] \end{aligned}$$

Properties of covariance

- $Cov(X, X) = Var(X)$
- $Cov(aX, bY) = a \cdot b \cdot Cov(X, Y) \neq Cov(X, Y)$
- $Cov(a + X, b + Y) = Cov(X, Y)$
- Covariance is **symmetric**: $Cov(X, Y) = Cov(Y, X)$

Covariance and independence

Recall: If X and Y are **independent**, we have

$$P(X = x, Y = y) = P(X = x) \times P(Y = y)$$

which implies

$$E(XY) = E(X) \times E(Y)$$

Hence...

$$\begin{aligned} Cov(X, Y) &= E(XY) - E(X)E(Y) \\ &= E(X)E(Y) - E(X)E(Y) \\ &= 0 \end{aligned}$$

In words: **If X and Y are independent, then their covariance is zero.**

Covariance – properties

But, though the symmetry is tempting:

If the covariance of two random variables is zero, it does **NOT** follow that they are independent!

(Example to come ...)

Consequences of correlation

Many statistics require us to calculate the **variance of an average**:

- 95% confidence interval for the mean:

$$\bar{Y} \pm 1.96 \times \sqrt{Var(\bar{Y})}$$

- t statistic for regression coefficient:

$$\frac{\hat{\beta}}{\sqrt{Var(\hat{\beta})}}$$

Covariance

General formula for the variance of a sum:

$$Var(Y_1 + \cdots + Y_n) = \sum_{i=1}^{n} Var(Y_i) + \sum_{i \neq j} Cov(Y_i, Y_j)$$

- If we **erroneously assume** that $Y_1, \ldots, Y_n$ are independent, then we will "naively" calculate $Var(Y_1 + \cdots + Y_n) = \sum_{i=1}^{n} Var(Y_i)$
- "Naive" variance will be too large or too small depending on value of $\sum_{i \neq j} Cov(Y_i, Y_j)$

Covariance

$$Var(Y_1 + \cdots + Y_n) = \sum_{i=1}^{n} Var(Y_i) + \sum_{i \neq j} Cov(Y_i, Y_j)$$

- In practice, positive covariances are more common than negative, hence "naive" variance (assuming independence) of a sum typically **underestimates** true variance $\Rightarrow$ **incorrect statistical inference.**
- Note that

$$Var(Y_1 - Y_2) = Var(Y_1) + Var(Y_2) - 2Cov(Y_1, Y_2)$$

so "naive" variance of a difference usually **overestimates** the true variance.

Consequences of ignoring correlation (sums)

For estimates involving **sums** (such as total effect of treatment)

$$Var(Y_1 + Y_2) = Var(Y_1) + Var(Y_2) + 2Cov(Y_1, Y_2)$$

Fill in the blanks Ignoring correlation leads to:

- _______ (larger/smaller) estimates of variance.
- Too _______ (wide/narrow) of confidence intervals
- Too _______ (large/small) of p-values

Consequences of ignoring correlation (differences)

For estimates involving **differences** (such as change over time)

$$Var(Y_1 - Y_2) = Var(Y_1) + Var(Y_2) - 2Cov(Y_1, Y_2)$$

Fill in the blanks Ignoring correlation leads to:

- _______ (larger/smaller) estimates of variance.
- Too _______ (wide/narrow) of confidence intervals
- Too _______ (large/small) of p-values

Exercise

- Consider a randomized trial of a placebo and a drug where each subject receives a treatment at two different timepoints
- **Scenario A: Each subject took the same treatment at both timepoints**
- **Scenario B: Each subject took the placebo at one timepoing and drug at the other**
- You analyze the data to estimate the treatment effect *ignoring the correlation* between outcomes on the same subject
- **Q: What goes wrong in each scenario? Are the confidence intervals too wide or too narrow? Are the p-values too large or small?**

Consequences of ignoring correlation

Pairwise correlation

Pearson's correlation coefficient

The **correlation coefficient** between X and Y is defined by

$$\rho(X, Y) = \frac{Cov(X, Y)}{\sqrt{Var(X)}\sqrt{Var(Y)}} = \frac{E[(X - E[X])(Y - E[Y])]}{\sqrt{Var(X)}\sqrt{Var(Y)}}$$

Notes

- Correlation is only defined between **pairs** of random variables.
- $\rho(X, Y)$ ranges between -1 and 1.
- **Formally**, X and Y are said to be **correlated** if $\rho(X, Y) \neq 0$.
- X and Y independent implies $\rho(X, Y) = 0$ (but not the reverse!)

Sample correlation

We can **estimate** the correlation between X and Y by computing the **sample correlation** between observations $x_1, \ldots, x_n$ from X and $y_1, \ldots, y_n$ from Y:

$$\hat{\rho}(X, Y) = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

Software commands

R

```
> cor(age,height)
```

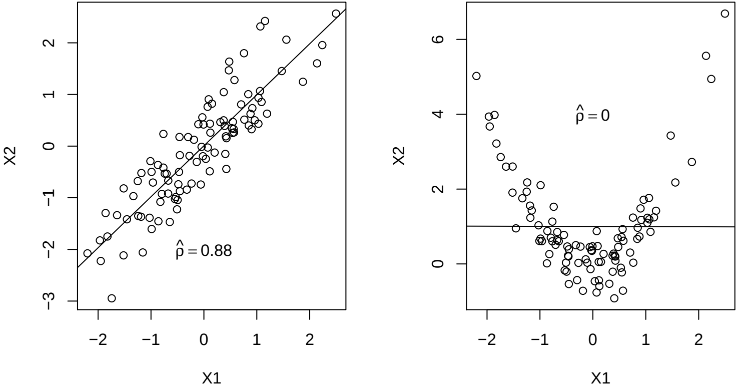
SAS

```
proc corr data = "xxxxxxxxx";
  var age height;
run;
```

If more than two variables are provided, above commands generally compute **all pairwise correlations**.

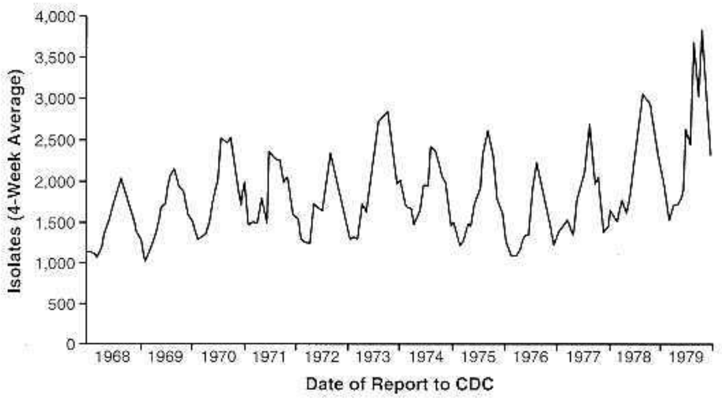
Correlation coefficient: caveats

ρ indicates strength of **linear relationship**, may "miss" other types:



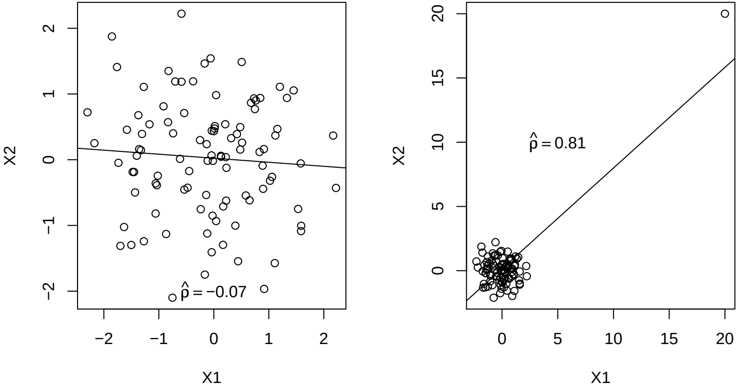
Correlation coefficient: caveats

ρ indicates strength of **linear relationship**, may "miss" other types:



Correlation coefficient: caveats

ρ can be affected dramatically by outliers



Correlation coefficient: alternatives

Question

Can we devise a way to measure dependence which is less sensitive to outliers?

Consider **ranking** the data, eg.

$$\begin{array}{cccccc} X & 4 & 5 & 1 & 18 & -3 \\ rank(X) & 3 & 4 & 2 & 5 & 1 \end{array}$$

Spearman's ρ

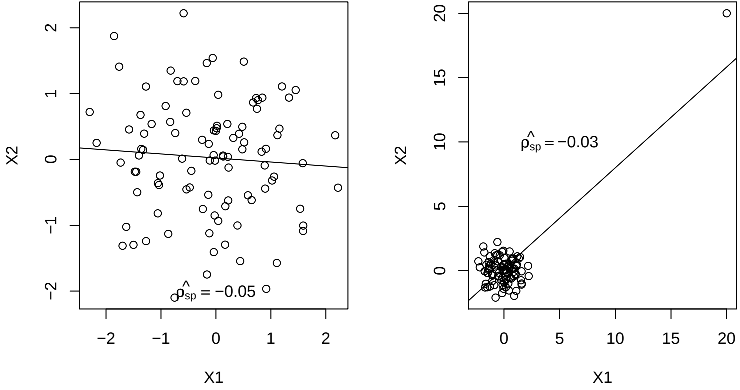
Once the original data $x_1, \ldots x_n$ and $y_1, \ldots, y_n$ have been ranked, yielding $r(x_1), \ldots, r(x_n)$ and $r(y_1), \ldots, r(y_n)$, we can compute the usual correlation coefficient on the ranks:

$$\hat{\rho}_{sp} = \frac{\frac{1}{n}\sum_{i=1}^{n}(r(x_i) - \bar{r}(\mathbf{x}))(r(y_i) - \bar{r}(\mathbf{y}))}{\sqrt{\frac{1}{n}\sum_{i=1}^{n}(r(x_i) - \bar{r}(\mathbf{x}))^2}\sqrt{\frac{1}{n}\sum_{i=1}^{n}(r(y_i) - \bar{r}(\mathbf{x}))^2}}$$

ρ_{sp} is known as **Spearman's** ρ.

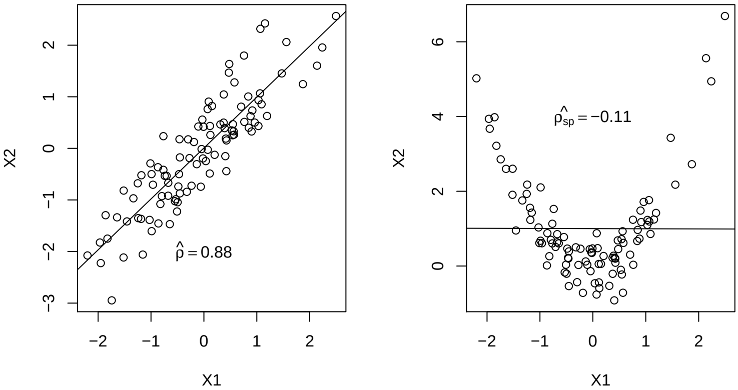
Spearman's ρ: features

Indicates strength of linear relationship between **ranks**, hence tests to what degree original data are **monotone functions** of each other.



Spearman's ρ: features

Indicates strength of linear relationship between **ranks**, hence tests to what degree original data are **monotone functions** of each other.

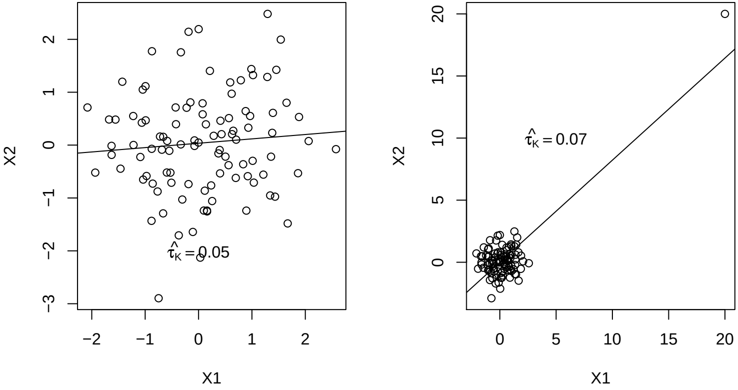


More fun with ranks: Kendall's τ

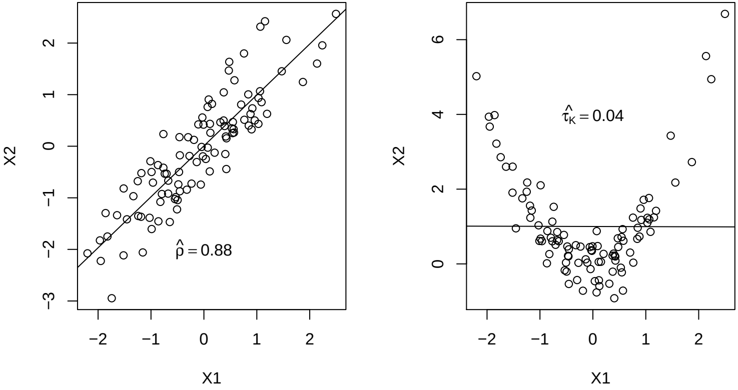
- Consider data as pairs $(x_1, y_1), \ldots, (x_n, y_n)$
- For every $i \neq j$, the pairs (x_i, y_i) and (x_j, y_j) may be:
  - *Concordant:* $x_i < x_j$ and $y_i < y_j$ or $x_i > x_j$ and $y_i > y_j$
  - *Discordant:* $x_i < x_j$ and $y_i > y_j$ or $x_i > x_j$ and $y_i < y_j$
  - *Neither:* $x_i = x_j$ or $y_i = y_j$
- Define **Kendall's** τ as

$$\tau_K = \frac{(\#\text{ of concordant pairs}) - (\#\text{ of discordant pairs})}{\frac{1}{2}n(n-1)}$$

Kendall's τ: examples



Kendall's τ: examples



Points to remember

- A and B are independent if $P(A, B) = P(A)P(B)$
  - equivalently $P(A|B) = P(A)$
- If X and Y are independent $\Rightarrow$ Cov(X,Y)=0
- **But** if Cov(X,Y)=0 $\not\Rightarrow$ X and Y are independent
- Ignoring correlation can lead to incorrect inference
- Strength of relationships
  - Pearson's correlation coefficient
  - Spearman's ρ
  - Kendall's τ

Assignment 0

- Your task: Download and explore a longitudinal dataset.
- Don't worry about getting the "right" answer; the goal is to familiarize you with longitudinal data and get you thinking about the analysis issues that arise.
- You won't be handing anything in, but come to next Tuesday's class prepared to discuss the dataset and your analysis.