

Chapter 14: Inference for Correlation

DSCC 462

Computational Introduction to Statistics

Anson Kahng

Fall 2022

Final Project Announcements

- Teams are due by Friday: Use the Google Sheet (see Announcements) to sign up / find teammates
- Final project description is up on Blackboard
 - Setup: Ad dataset, company wants to understand their data...
 - Some open-ended questions, some less so (room for interpretation!)
- Datasets will be released once teams are formed (Friday)

Plan for Today

- Introduce tools that allow us to go past our previous assumptions of independence when comparing random variables
- In particular, learn how to infer whether or not a linear relationship exists between two variables, X and Y

Correlation

- Recall that correlation tells us the degree to which two random variables are (linearly) associated or related
- We denote the true population correlation of X and Y as ρ ("rho")
- We estimate ρ with Pearson's coefficient of correlation, r (Chapter 3)

$$r = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\left[\sum_{i=1}^n (x_i - \bar{x})^2 \right] \left[\sum_{i=1}^n (y_i - \bar{y})^2 \right]}} = \frac{1}{(n-1)} \sum_{i=1}^n \left(\frac{x_i - \bar{x}}{s_x} \right) \left(\frac{y_i - \bar{y}}{s_y} \right)$$

- Note that $-1 \leq r \leq 1$

Correlation: Example

- Setup: Suppose we examine $n = 7$ subjects for which we have age and weight measurements
- We want to determine whether a significant linear relationship exists between age (X) and weight (Y)

Correlation: Inference

- Question: Is the population correlation, ρ , between two variables, X and Y , different from 0?
- In other words, we're investigating whether a linear relationship exists between X and Y (age and weight)
- Use the sample correlation r for our statistic in hypothesis testing

Correlation: Inference

- Hypotheses: $H_0 : \rho = 0$ vs. $H_1 : \rho \neq 0$
- If pairs (x_i, y_i) come from normally distributed X and Y , then if we standardize r , we get a statistic that has a t distribution with $n - 2$ degrees of freedom

- Test statistic: $t = \frac{r - \rho}{SE(r)}$, where $SE(r) = \sqrt{\frac{1 - r^2}{n - 2}}$

- We thus get $t = \frac{r - \rho}{SE(r)} = r \sqrt{\frac{n - 2}{1 - r^2}}$

Correlation: Inference Example

- Returning to setup: Suppose we examine $n = 7$ subjects for which we have age and weight measurements
- We want to determine whether a significant linear relationship exists between age (X) and weight (Y)
 - $H_0 : \rho = 0$ vs. $H_1 : \rho \neq 0$
- We know that the correlation between weight and age for this sample is $r = 0.865$
- Test this null hypothesis at the $\alpha = 0.05$ significance level

Correlation: Inference Example

- $H_0 : \rho = 0$ vs. $H_1 : \rho \neq 0$, test at $\alpha = 0.05$
- The correlation between weight and age for this sample is $r = 0.865$
- Test statistic:
- p-value:
- Conclusion:

Correlation: Inference Example

- $H_0 : \rho = 0$ vs. $H_1 : \rho \neq 0$, test at $\alpha = 0.05$
- The correlation between weight and age for this sample is $r = 0.865$
- Test statistic:

$$t = r \sqrt{\frac{n-2}{1-r^2}} = 0.865 \cdot \sqrt{\frac{7-2}{1-0.865^2}} = 3.856$$

- p-value:

$$p = 2 \cdot \Pr(T \geq 3.856) = 2 * (1 - \text{pt}(3.856, \text{df}=5)) = 0.0119$$

- Conclusion: Since the p-value is less than $\alpha = 0.05$, we reject the null hypothesis and conclude that a linear relationship does exist between age and weight

Correlation: Inference in R

- We can also calculate directly in R using the `cor.test()` function

```
> x<-c(220,215,179,145,145,177,136)
> y<-c(68,58,43,37,20,58,36)
> cor.test(x,y)
```

Pearson's product-moment correlation

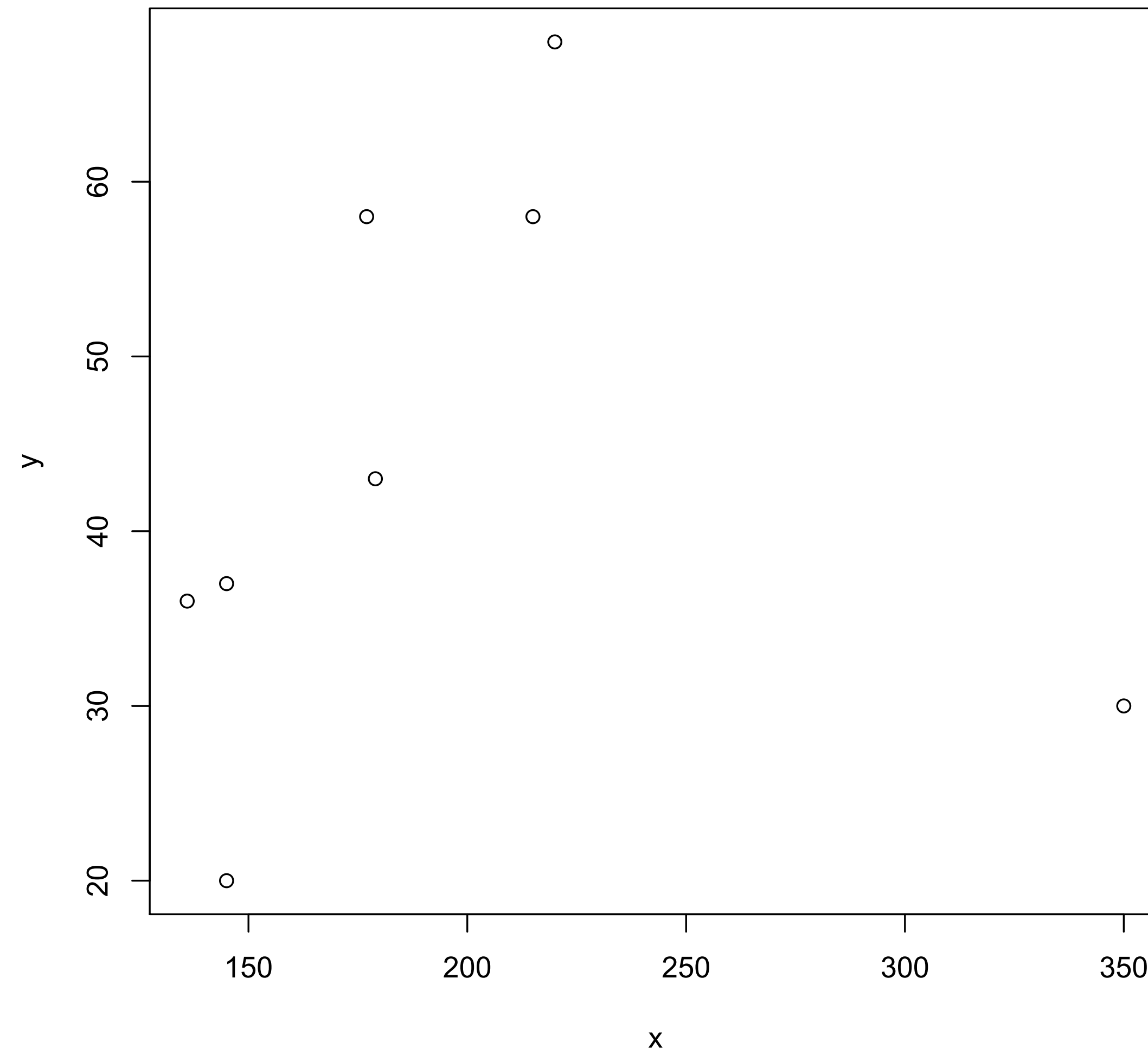
```
data:  x and y
t = 3.856, df = 5, p-value = 0.01193
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.3213733 0.9798243
sample estimates:
      cor
0.8650727
```

Correlation Limitations

- Only describes linear relationships
 - We could be missing a nonlinear relationship if we don't examine the scatterplot
- Hypothesis testing only works for the null hypothesis $H_0 : \rho = 0$
 - For any $\rho \neq 0$, normality assumptions are not met and our hypothesis testing procedures are invalid
- Correlation can be very sensitive to outliers and can thus give misleading results when outliers are present

Correlation Limitations

- Suppose we have another subject who is 30 years old with a weight of 350 pounds



Correlation Limitations

```
> x<-c(220,215,179,145,145,177,136, 350)
> y<-c(68,58,43,37,20,58,36, 30)
> plot(x,y)
> cor(x,y)
[1] 0.06260467
```

Spearman's Rank Correlation Coefficient

- Need a more robust measure that isn't as sensitive to outliers
- Instead of using the actual observations, we rank the data and then use the *ranks* as our data
 - If multiple values are the same, assign the average rank
- Rank all the x values, and call these ranks x_r
- Rank all the y values, and call these ranks y_r
- Compute the Pearson's correlation coefficient for this ranked data (x_r, y_r) instead of the actual data
- This is *Spearman's correlation coefficient*

Spearman's Rank Correlation Coefficient

- Calculate Spearman's correlation coefficient as follows:

$$r_s = \frac{\sum_{i=1}^n (x_{ri} - \bar{x}_r)(y_{ri} - \bar{y}_r)}{\sqrt{\left[\sum_{i=1}^n (x_{ri} - \bar{x}_r)^2 \right] \left[\sum_{i=1}^n (y_{ri} - \bar{y}_r)^2 \right]}}$$

- Note that $-1 \leq r_s \leq 1$
- Same interpretations of association between X and Y

Spearman's Rank Correlation Coefficient

- Consider the following age and weight measurements for the 7 subjects

Patient	Weight	Rank	Age	Rank
1	220		68	
2	215		58	
3	179		43	
4	145		37	
5	145		20	
6	177		58	
7	136		36	

Spearman's Rank Correlation Coefficient

- Consider the following age and weight measurements for the 7 subjects

Patient	Weight	Rank	Age	Rank
1	220	7	68	7
2	215	6	58	5.5
3	179	5	43	4
4	145	2.5	37	3
5	145	2.5	20	1
6	177	4	58	5.5
7	136	1	36	2

Spearman's Rank Correlation Coefficient

- Using this information, we can calculate r_s as follows:
 - $\bar{x}_r =$
 - $\bar{y}_r =$
 - $r_s =$
- Recall that Pearson's correlation coefficient was given as 0.865. How does Spearman's rank correlation coefficient compare?

Spearman's Rank Correlation Coefficient

- Using this information, we can calculate r_s as follows:

- $\bar{x}_r = 4$

- $\bar{y}_r = 4$

- $$r_s = \frac{\sum_{i=1}^n (x_{ri} - 4)(y_{ri} - 4)}{\sqrt{\left[\sum_{i=1}^n (x_{ri} - 4)^2 \right] \left[\sum_{i=1}^n (y_{ri} - 4)^2 \right]}} = 0.8727$$

- Recall that Pearson's correlation coefficient was given as 0.865. How does Spearman's rank correlation coefficient compare? Matches closely because we had normally distributed data and no obvious outliers

Spearman's Rank Correlation Coefficient: R

- In R:

```
> x<-c(220,215,179,145,145,177,136)
> y<-c(68,58,43,37,20,58,36)
> cor(x,y, method="spearman")
[1] 0.8727273
>
> x1 <- rank(x)
> y1 <- rank(y)
> x1; y1
[1] 7.0 6.0 5.0 2.5 2.5 4.0 1.0
[1] 7.0 5.5 4.0 3.0 1.0 5.5 2.0
> cor(x1,y1)
[1] 0.8727273
```

Spearman's Rank Correlation Coefficient: Interpretation

- Spearman's rank correlation coefficient is a measure of concordance of ranks for the outcomes x and y
- If measurements of X and Y are ranked in the same order for each variable, then $r_s = 1$
- If measurements of X and Y are ranked in the reverse order from each other, then $r_s = -1$
- If there is no linear correspondence between the ranks, then $r_s = 0$

Spearman's Rank Correlation Coefficient: Outliers

- What is the Spearman rank correlation coefficient when we have an outlier in our data?

```
> x<-c(220,215,179,145,145,177,136,350)
> y<-c(68,58,43,37,20,58,36,30)
> cor(x,y)
[1] 0.06260467
> cor(x,y, method="spearman")
[1] 0.373494
```

Spearman's Rank Correlation Coefficient: Inference

- We can similarly perform hypothesis tests for ρ based on r_s
- If the sample size is large enough (generally $n \geq 10$) and if we can assume that pairs of ranks (x_{ri}, y_{ri}) are chosen randomly, then we can test the null hypothesis $H_0 : \rho = 0$ vs. the alternative hypothesis $H_1 : \rho \neq 0$
- Use a similar test statistic: $t_s = r_s \cdot \sqrt{\frac{n-2}{1-r_s^2}}$
- Compare t_s to a t distribution with $n - 2$ degrees of freedom

Spearman's Rank Correlation Coefficient: Inference

- Although we only have $n = 7$, let's test $H_0 : \rho = 0$ vs. $H_1 : \rho \neq 0$ at $\alpha = 0.05$
- Calculating our t-statistic:
- Calculating the p-value:
- Conclusion:

Spearman's Rank Correlation Coefficient: Inference

- Although we only have $n = 7$, let's test $H_0 : \rho = 0$ vs. $H_1 : \rho \neq 0$ at $\alpha = 0.05$
- Calculating our t-statistic:

$$t_s = r_s \cdot \sqrt{\frac{n-2}{1-r_s^2}} = 0.8727 \cdot \sqrt{\frac{7-2}{1-0.8727^2}} = 3.997$$

- Calculating the p-value:

$$p = 2 \Pr(T > 3.997) = 2 * (1 - \text{pt}(3.997, \text{df}=5)) = 0.01035$$

- Conclusion:

$$p < \alpha = 0.05, \text{ so we reject } H_0 \text{ and conclude nonzero correlation}$$

Spearman's Rank Correlation Coefficient: R

```
> x<-c(220,215,179,145,145,177,136)
> y<-c(68,58,43,37,20,58,36)
> cor.test(x,y,method="spearman",exact=FALSE)
```

Spearman's rank correlation rho

```
data:  x and y
S = 7.1273, p-value = 0.01035
alternative hypothesis: true rho is not equal to 0
sample estimates:
      rho
0.8727273
```