Cancer data visualization and analysis

Jeff Oliver

04 June, 2019

A two-hour workshop for participants in STEP-UP summer program on cancer prevention and control. There are 16 students, mostly upper-division undergraduates.

Learning objectives

1. Examine data to ensure quality
2. Develop hypotheses to explain trends
3. Visualize comparisons between two groups
4. Visualize relationships between two variables
5. Test relationships between variables
6. Understand the difference between correlation and causation

[DESCRIPTION OR MOTIVATION; 2-4 sentences that would be used for an announcement]

Getting started

- Start RStudio
- Make `data` and `output` directories
- Download data (come from https://statecancerprofiles.cancer.gov/)
    - These data are lung cancer data incidence rates for each state, along with some demographic data

```
download.file(url = "tinyurl.com/cancer-data-csv", destfile = "data/cancer-data.csv")
```

- Look at data

```
cancer.data <- read.csv(file = "data/cancer-data.csv")
head(cancer.data)
```

```
##        state male.lung female.lung income poverty uninsured unemployed
## 1    Alabama      89.0        51.6  44758    14.0      13.8        8.3
## 2     Alaska      65.3        50.1  74444     7.0      17.5        7.8
## 3    Arizona      54.7        45.0  51340    12.9      13.6        8.0
## 4   Arkansas      98.7        61.6  42336    13.8      11.6        6.9
## 5 California      49.2        39.0  63783    11.8      10.5        8.7
## 6   Colorado      46.9        40.7  62520     8.1      10.2        6.0
##   lang.isolation
## 1            1.2
## 2            2.2
## 3            4.5
## 4            1.6
## 5            9.4
## 6            3.0
```

```r
summary(cancer.data)
```

```
##         state      male.lung       female.lung        income
##  Alabama   : 1   Min.   : 32.30   Min.   :23.70   Min.   :40528
##  Alaska    : 1   1st Qu.: 63.17   1st Qu.:50.52   1st Qu.:49037
##  Arizona   : 1   Median : 69.85   Median :53.45   Median :54384
##  Arkansas  : 1   Mean   : 72.29   Mean   :53.47   Mean   :56031
##  California: 1   3rd Qu.: 82.85   3rd Qu.:58.65   3rd Qu.:62519
##  Colorado  : 1   Max.   :112.80   Max.   :79.00   Max.   :76067
##  (Other)   :45   NA's   :1        NA's   :1
##       poverty         uninsured       unemployed    lang.isolation
##  Min.   : 5.30   Min.   : 3.7   Min.   :2.800   Min.   :0.300
##  1st Qu.: 8.10   1st Qu.: 7.7   1st Qu.:5.750   1st Qu.:1.500
##  Median :10.20   Median :11.2   Median :7.100   Median :2.400
##  Mean   :10.37   Mean   :11.1   Mean   :6.859   Mean   :3.031
##  3rd Qu.:12.60   3rd Qu.:13.7   3rd Qu.:8.050   3rd Qu.:4.200
##  Max.   :17.40   Max.   :22.3   Max.   :9.600   Max.   :9.400
##
```

- Describe the data

Exercise

- Get in groups; take five minutes to come up with a hypothesis you can test with these data

---

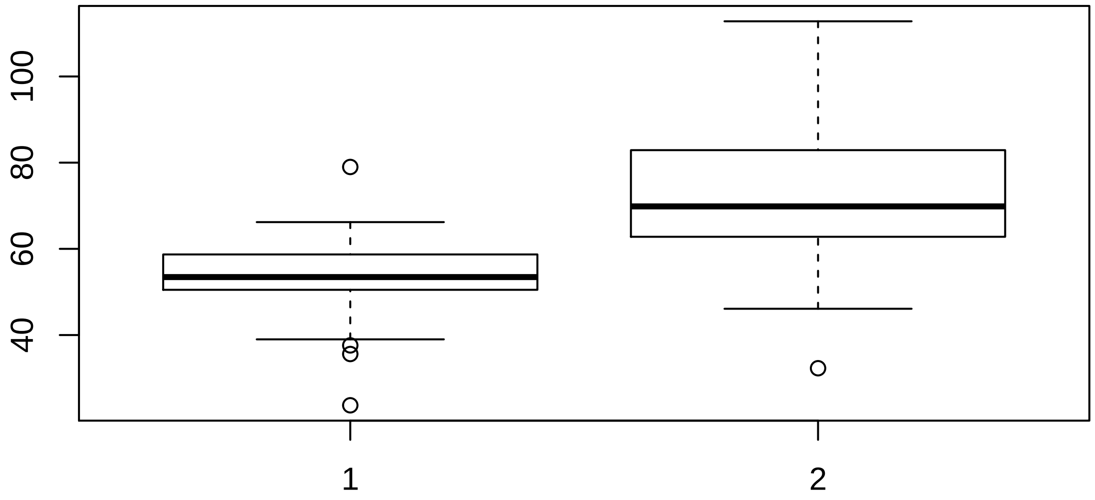
Comparing groups

New script (reproducibility)

```r
# Compare cancer incidence between sexes
# Jeff Oliver
# jcoliver@email.arizona.edu
# 2019-06-05
```

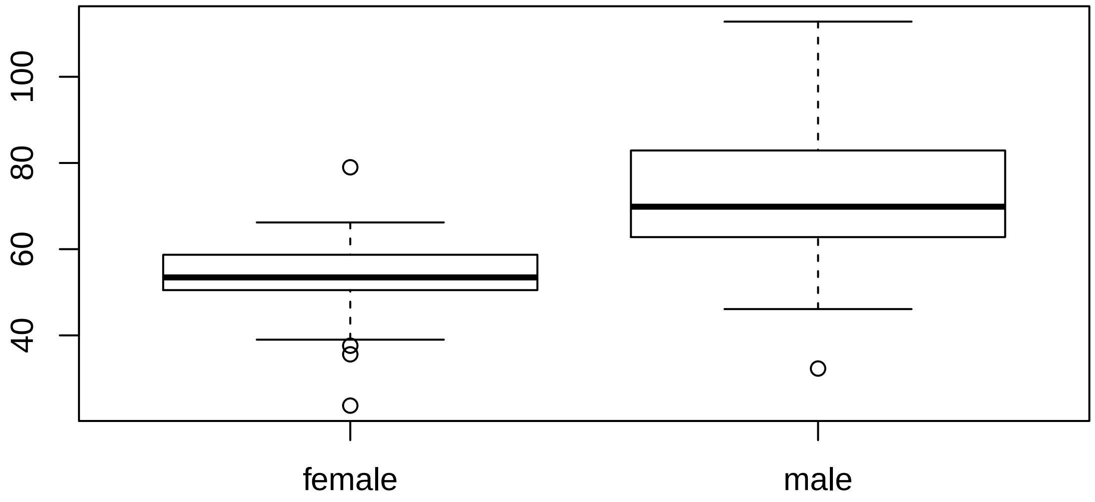
Read in data & visualize male v. female incidence

```r
cancer.data <- read.csv(file = "data/cancer-data.csv")
boxplot(cancer.data$female.lung, cancer.data$male.lung)
```



But the x-axis doesn't tell us what "1" and "2" correspond to.

```
boxplot(list(female = cancer.data$female.lung, male = cancer.data$male.lung))
```



Run t-test

```
t.test(x = cancer.data$female.lung, y = cancer.data$male.lung)
```

```
##
##  Welch Two Sample t-test
##
## data:  cancer.data$female.lung and cancer.data$male.lung
## t = -7.3113, df = 77.884, p-value = 2.017e-10
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -23.95238 -13.69962
## sample estimates:
## mean of x mean of y
##    53.466    72.292
```

Exercise

- Get in groups; take five minutes to draw how you might show relationship between variables; ideally comes from one of the hypotheses students generated

---

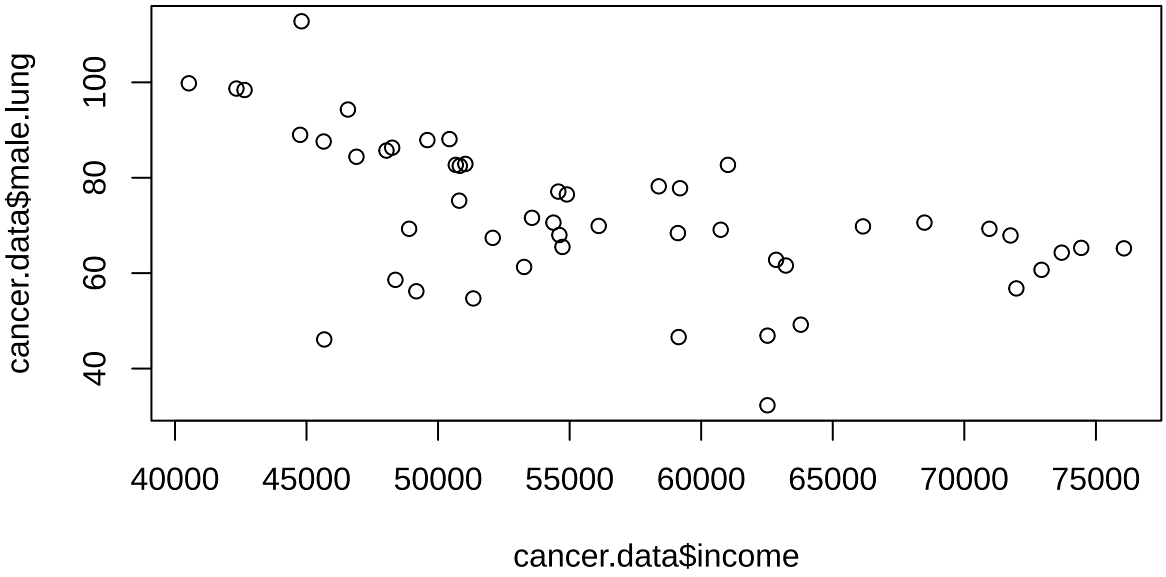
Testing relationships

New script

```
# Compare cancer incidence between sexes
# Jeff Oliver
# jcoliver@email.arizona.edu
# 2019-06-05
```

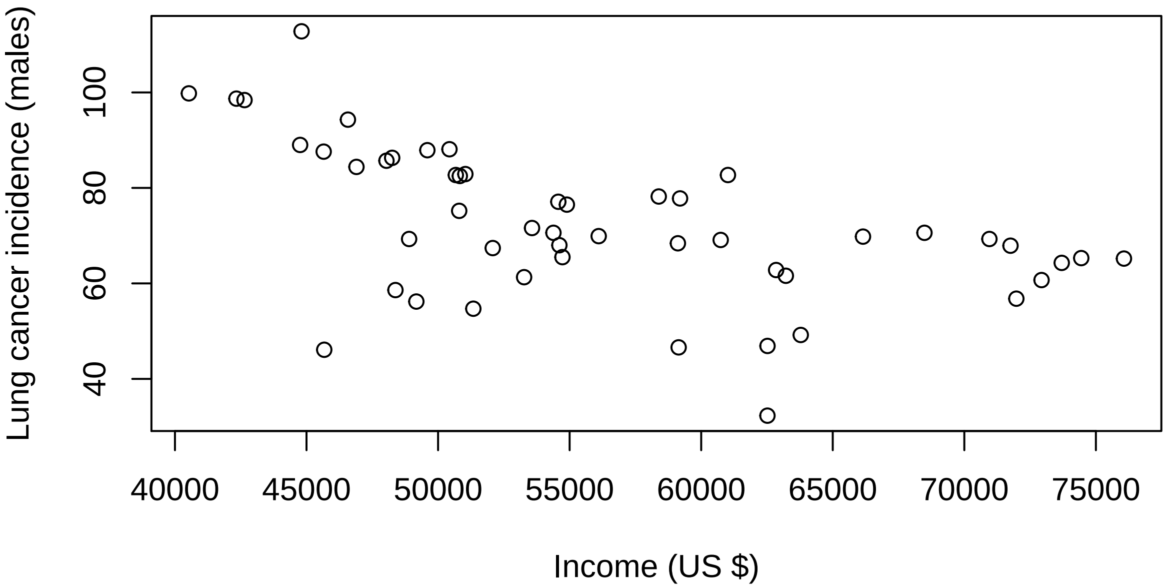
Plot data & describe hypothesis

```
cancer.data <- read.csv(file = "data/cancer-data.csv")
plot(x = cancer.data$income, y = cancer.data$male.lung)
```



Clean up the axis labels

```
plot(x = cancer.data$income,
     y = cancer.data$male.lung,
     xlab = "Income (US $)",
     ylab = "Lung cancer incidence (males)")
```



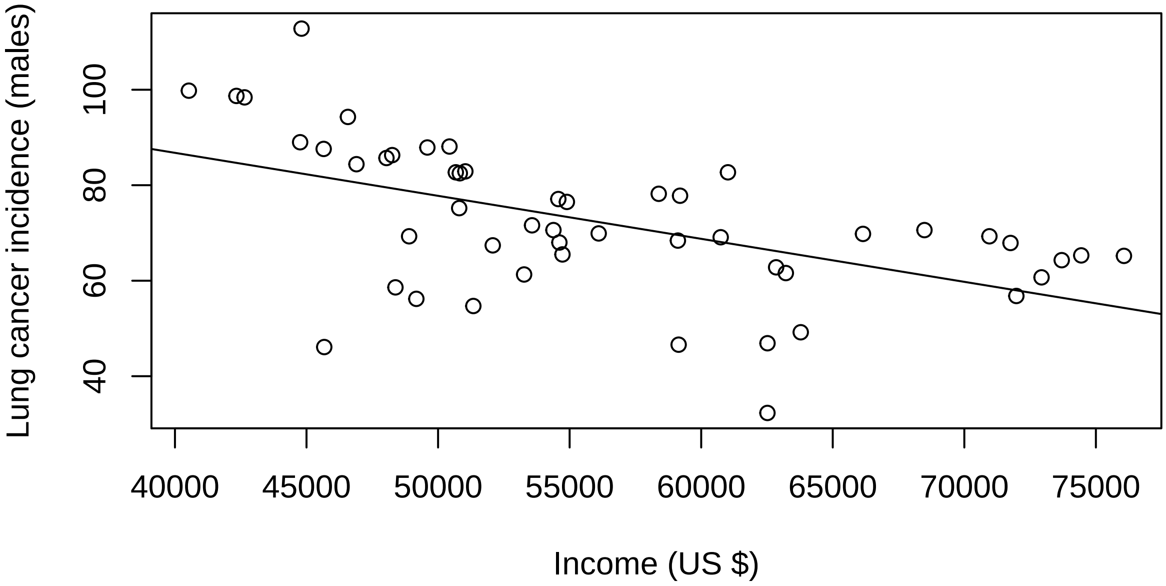
Test relationship

```
income.lm <- lm(cancer.data$male.lung ~ cancer.data$income)
summary(income.lm)
```

```
##
## Call:
## lm(formula = cancer.data$male.lung ~ cancer.data$income)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -35.573  -5.287   3.705   9.180  30.350
##
## Coefficients:
##                      Estimate Std. Error t value Pr(>|t|)
## (Intercept)         1.228e+02  1.150e+01  10.677 2.84e-14 ***
## cancer.data$income -9.007e-04  2.022e-04  -4.453 5.04e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 13.44 on 48 degrees of freedom
##   (1 observation deleted due to missingness)
## Multiple R-squared:  0.2924, Adjusted R-squared:  0.2776
## F-statistic: 19.83 on 1 and 48 DF,  p-value: 5.041e-05
```

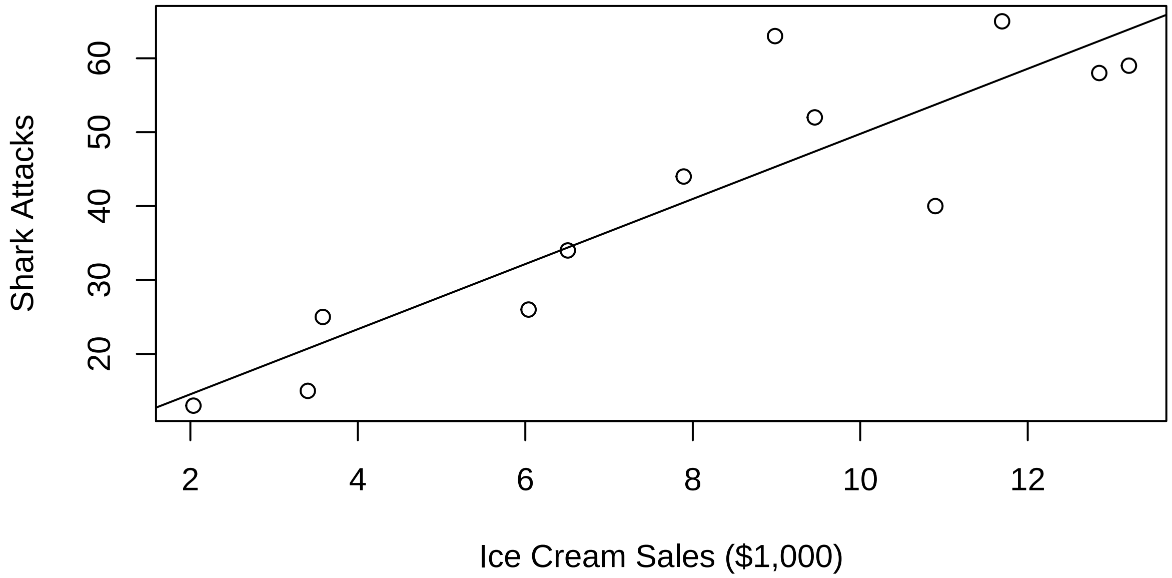
Update plot

```
plot(x = cancer.data$income,
     y = cancer.data$male.lung,
     xlab = "Income (US $)",
     ylab = "Lung cancer incidence (males)")
abline(income.lm)
```



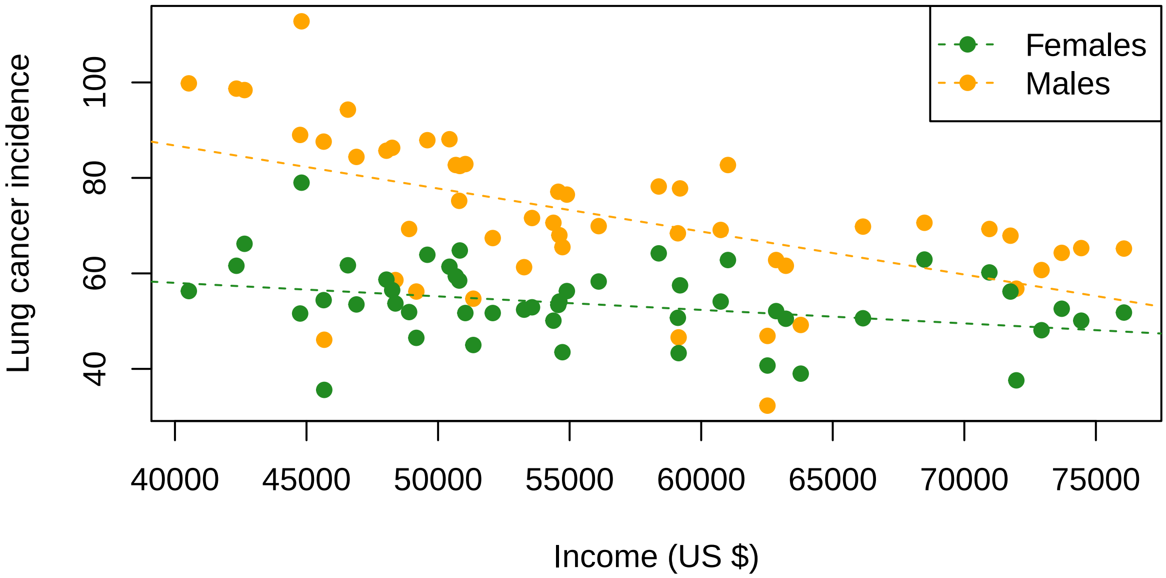
Correlation vs. Causation

Discuss:



Some additional plotting options

```r
plot(x = cancer.data$income,
     y = cancer.data$male.lung,
     xlab = "Income (US $)",
     ylab = "Lung cancer incidence",
     pch = 19,
     col = "orange")
points(x = cancer.data$income,
       y = cancer.data$female.lung,
       pch = 19,
       col = "forestgreen")
abline(income.lm, col = "orange", lty = 2)
abline(lm(female.lung ~ income, data = cancer.data), col = "forestgreen", lty = 2)
legend("topright",
       legend = c("Females", "Males"),
       col = c("forestgreen", "orange"),
       pch = 19,
       lty = 2)
```



Additional resources

- resource one
- resource two
- A PDF version of this lesson

Questions? e-mail me at jcoliver@email.arizona.edu.