Review of Introductory Statistics

Zack Treisman

Spring 2021

WESTERN
COLORADO UNIVERSITY
MATHEMATICS & COMPUTER
SCIENCE DEPARTMENT

Data

An **observation** is a single unit. A **variable** is a measurement made on that unit

- Record observations as **rows** and variables in **columns**.
- Variables can be **categorical** or **numerical**.
    - Categorical variables can be **binary** or not, **ordered** or not.
    - Numerical variables can be **discrete** or **continuous**.
- Dates, times and locations merit special consideration.
- Vocabulary is not universal: Factor, case, treatment . . .

```r
head(Sitka) # from package MASS
```

```
##   size Time tree treat
## 1 4.51  152    1 ozone
## 2 4.98  174    1 ozone
## 3 5.41  201    1 ozone
## 4 5.90  227    1 ozone
## 5 6.15  258    1 ozone
## 6 4.24  152    2 ozone
```

Distributions

The **distribution** of a variable is a measure of how often it takes each possible value.

- The distribution of a categorical variable is a list of the percentage of observations in each category.

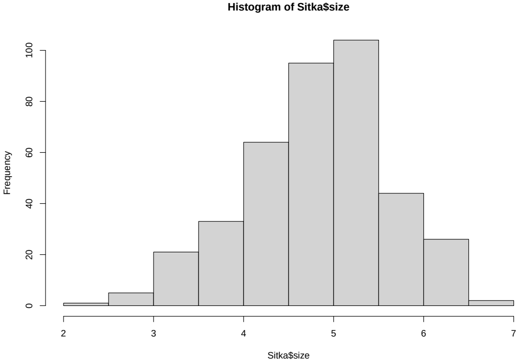
```r
table(Sitka$treat)/length(Sitka$treat)
```

```
##
##   control     ozone
## 0.3164557 0.6835443
```

Distributions

- Picture the distribution of a numerical variable with a histogram, boxplot or density estimate.

```
hist(Sitka$size)
```



- Shape: center, spread, skew, kurtosis

Numerical summaries

```r
summary(Sitka$size)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   2.230   4.345   4.900   4.841   5.400   6.630
```

```r
quantile(Sitka$size, c(0.025,0.975))
```

```
##   2.5%  97.5%
## 3.2370 6.2815
```

```r
sd(Sitka$size)
```

```
## [1] 0.7982084
```

```r
var(Sitka$size)
```

```
## [1] 0.6371367
```

```r
IQR(Sitka$size)
```

```
## [1] 1.055
```

Common distributions

- Normal distribution
  - The sum of many independent effects tends to be normal.
  - Formula that you never use: $N(\mu, \sigma^2)(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$
  - Observations on disparate scales can be standardized with z scores: $z = \frac{x-\mu}{\sigma}$
- Other distributions from Intro Stats
  - t - Like the normal distribution, but adjusted for small samples.
  - χ^2 - Sum of squared standard normals.
  - F - Similar to χ^2, used in ANOVA.
  - Binomial - How many successes in *n* trials?
  - Poisson - Count of discrete events in fixed time or space.
  - ...

Inference

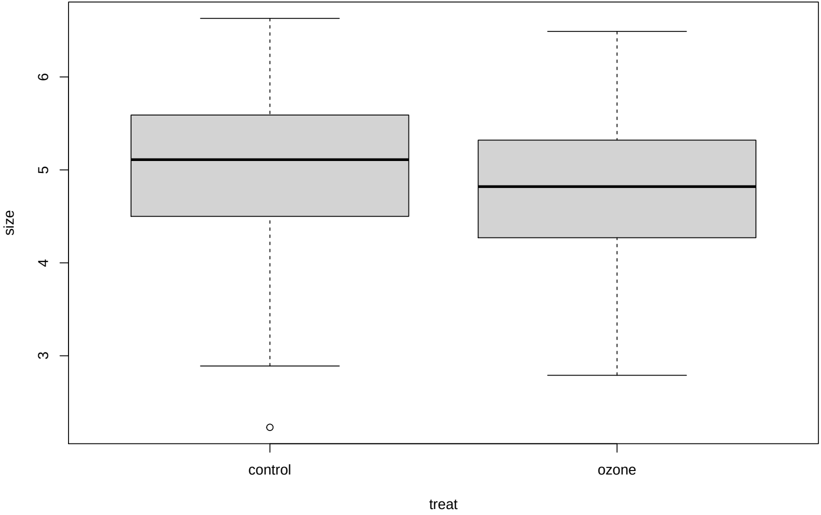
- Confidence intervals
- Hypothesis tests
  - p(robability)-values
  - Null and alternate hypotheses
  - t-tests, ANOVA, χ^2 tests

```
t.test(size~treat, data = Sitka)
```

```
##
##  Welch Two Sample t-test
##
## data:  size by treat
## t = 2.3163, df = 209.44, p-value = 0.02151
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  0.03144833 0.39086574
## sample estimates:
## mean in group control   mean in group ozone
##              4.985120              4.773963
```

Always plot your data!

```
boxplot(size~treat, data = Sitka)
```



Linear models

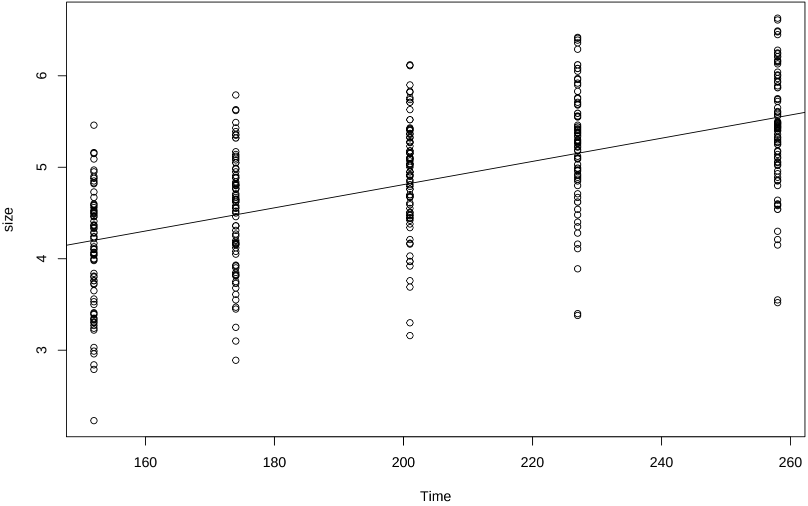
- Slope and intercept parameters
- Correlation
- Residuals
- Inference

```r
summary(lm(size~Time, data = Sitka))
```

```
##
## Call:
## lm(formula = size ~ Time, data = Sitka)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.02610 -0.37956  0.06948  0.41669  1.30948
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 2.2732443  0.1768643   12.85   <2e-16 ***
## Time        0.0126855  0.0008592   14.77   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.641 on 393 degrees of freedom
## Multiple R-squared:  0.3568, Adjusted R-squared:  0.3551
## F-statistic:   218 on 1 and 393 DF,  p-value: < 2.2e-16
```

Always plot your data!

```
plot(size~Time, data = Sitka)
abline(lm(size~Time, data=Sitka))
```


size
Time