Stat 414 — Day 1
Models

Example 1: Predicting Airfares

Airfares from San Luis Obispo to a "random" sample of 12 major U.S. cities as found March 31, 2017 on Travelocity.com for travel on May 8-May 12, 2017 are found in airfare.txt.

(a) Load the data

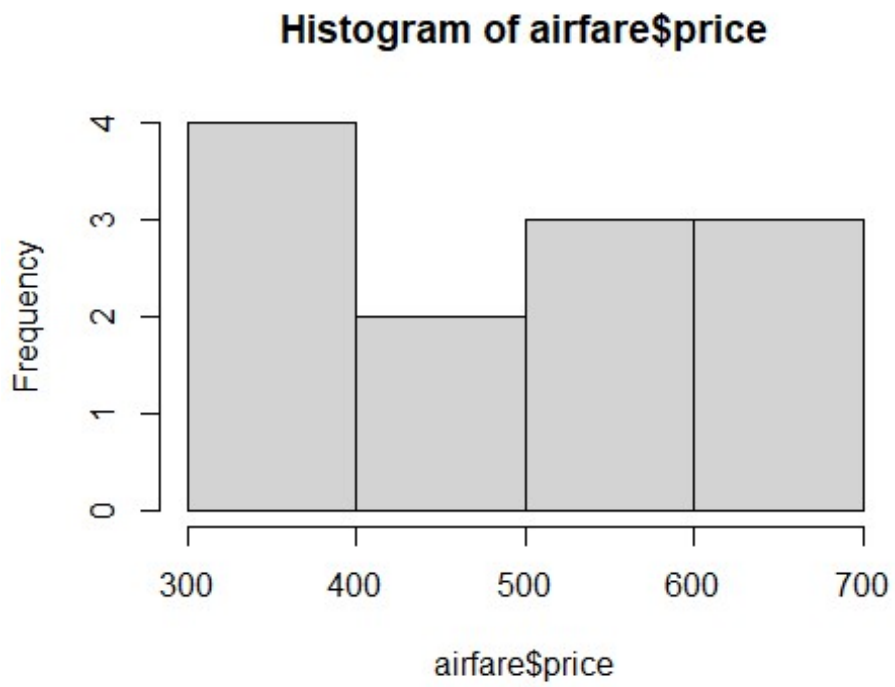
```
airfare <- read.table("https://www.rossmanchance.com/stat414/data/airfare.txt", hea
der = TRUE)
head(airfare)
  price distance                    city
1 631.8     2604    JacksonvilleFlorida
2 338.6      850       SaltLakeCityUtah
3 627.9     2590 CharlotteNorthCarolina
4 352.6      673          TucsonArizona
5 699.8     2370     JacksonMississippi
6 470.7     1990        StLouisMissouri
```

(b) What are the observational units and response variable?

The observational units are the destination cities and the response variable is the price (quantitative, dollars).

(c) Examine descriptive statistics and a graph

```
#I like to use some functions I made
#load in these functions
load(url("https://www.rossmanchance.com/iscam3/ISCAM.RData"))
iscamsummary(airfare$price)
missing       n    Min       Q1  Median      Q3     Max    Mean      SD    skew
      0      12 321.100 350.625 509.150 584.348 699.800 490.836 132.502   0.049
hist(airfare$price)
```



Histogram of airfare prices (12 cities), slightly skewed right

Is the distribution of prices approximately normal? How are you deciding? Do we care?

The distribution of airfare prices for these 12 cities is slightly skewed to the right. We probably don't care. If a distribution was strongly skewed right, and normality was a requirement of some later inference procedure, we could consider a data transformation.

Definitions

- The *population* is the entire collection of observational units we are interested in
- A *parameter* is a numerical summary of the population
- The "sample" is the collection of observational units which we gather data for
  - A *statistic* is a numerical summary of the sample

(d) Define in words the population mean you would like to estimate.

The average aifrare to all US cities or to destination cities above a certain population size?

(e) What one number would you use to estimate that population mean?

The sample mean would be an unbiased estimator of the population mean.

(f) What one number would you use to estimate the price of an individual flight?

The sample mean would also be a good way to estimate any individual city.

(g) What makes your number in (f) 'best'?

Overall, it's the 'closest' to all of the data values, at least as we use the sum of squared errors as our metric. In other words, the sample mean would minimize the sum of the squared deviations between the observed data values and the one number.

(h) How accurate would you say your estimate in (f) is?

Could report that minimized sum of squared errors. Or divide that value by n-1 and that's the sample variance. Or take the square root of that value and that's the sample standard deviation.

(i) How accurate would you say your estimate in (e) is?

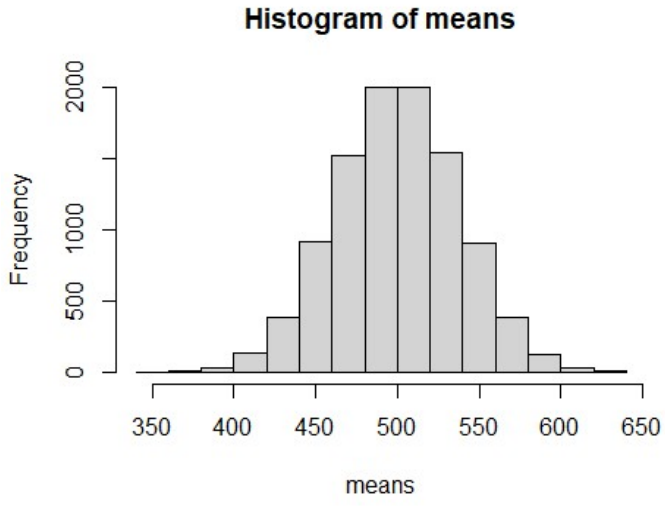
Now we need to consider how much sample to sample variation there is in the sample mean and how far we can expect the sample means, on average, to fall from the population mean we are trying to estimate.

Demo:

Let's assume airfare prices follow a normal distribution with mean 500 dollars and standard deviation 250 dollars. How much *sampling variation* is there in the sample mean?

```
means <- 0  #initialize the vector where we will story our results
for (i in 1:10000){
  sim_airfares <- rnorm(n = 12, mean = 500, sd = 132)
  means[i] = mean(sim_airfares)
}

iscamsummary(means)
missing       n     Min       Q1  Median      Q3     Max    Mean      SD    skew
      0   10000 351.633 473.947 500.041 525.570 637.045 499.980  37.763  -0.006
hist(means)
```



Histogram of sample means

(j) Does the distribution of sample means behave as you would have predicted?

Using rules for random variables, we expect distribution of sample means to have a mean close to the poulation mean (500) with standard deviation $SD(\bar{x}) = \sigma/\sqrt{n} = 132/\sqrt{12} \approx 38.1$. Because the samples came from a normal distribution, we expected the distribution of sample means to also follow a normal distribution.

Fitting a model

Let's fit a *model* to find the best predictor of airfare. Our model has the form $y_i = \beta_0 + \epsilon_i$. Here β_0 is the *parameter* we are trying to estimate and ϵ_i is *random*, so follows a probability distribution. This means y_i is also random and also follows a probability distribution.

(k) What are the expected value and variance of the random variable Y?

$E(Y_i) = E(\beta_0 + \epsilon_i) = E(\beta_0) + E(\epsilon_i)$. We typically assume $\epsilon_i \sim N(0, \sigma^2)$ so this reduces to β_0.
$V(Y_i) = V(\beta_0 + \epsilon_i) = Var(\epsilon_i) = \sigma^2$

(l) What is the *correlation* between two different airfares?

$Cov(Y_i, Y_j) = 0$ assuming the observations are independent.

Fit the least squares regression model:

```
m0 <- lm(price ~ 1, data = airfare)
summary(m0)

Call:
lm(formula = price ~ 1, data = airfare)

Residuals:
    Min      1Q  Median      3Q     Max
-169.74 -140.21   18.31   93.51  208.96

Coefficients:
            Estimate Std. Error t value     Pr(>|t|)
(Intercept)   490.84      38.25   12.83 0.0000000582 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 132.5 on 11 degrees of freedom
```

(m) What do you notice about the intercept and about the *standard error of the intercept*? What do you notice about the *residual standard error*?
The intercept matches the sample mean airfare and the standard error of the intercept matches the predicted standard deviation $\sigma/\sqrt{n} = 132/\sqrt{12} \approx 38.1$ though may not be exact.

Fitting a model using distance

Examine a scatterplot using distance to the city to predict the price

```
m1 <- lm(price ~ distance, data = airfare)
summary(m1)

Call:
lm(formula = price ~ distance, data = airfare)

Residuals:
    Min      1Q  Median      3Q     Max
-138.21  -55.91   22.13   45.06  147.19

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 214.99440   69.81197    3.08   0.0116 *
distance      0.14246    0.03384    4.21   0.0018 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 83.46 on 10 degrees of freedom
Multiple R-squared:  0.6393,    Adjusted R-squared:  0.6032
F-statistic: 17.72 on 1 and 10 DF,  p-value: 0.0018
```

(n) What do you notice about the residual standard error? Which model is more 'accurate'? What do you know about the *Multiple R-squared*? How do we interpret this value?

When we bring in more information, the typical (overall/average) prediction gets smaller. R^2 measures the proportion of variation (sum of squared residuals) explained, e.g.,

$$1 - \frac{\sum(y_i - \hat{y}_i)^2}{\sum(y_i - \bar{y})^2}$$

Example 2: Demonstration

Follow this link for the Two Quantitative Variables applet

I want to show you another simulation, but first I need to create a population to sample from.

- Check the **Design Population** box.
- Use pull-down menu to select **Observed x**
- Press the **Create Population** button

The applet will create a population that matches our sample's characteristics but assumes a population slope of zero.

- Check the **Show Sampling Options** box.
- Specify **1000** samples
- Specify a sample size of **12**
- Use the pull-down menu to select the **slope** as the statistic

(a) Describe the resulting *sampling distribution of slopes*.

Should be approximately normal with mean around zero and standard deivation around

(b) How does the standard deviation of the distribution you created compare to the *standard error of the slope* in the regression output?

Should be pretty similar.

Definition

$SE(\hat{\beta})$ estimates the sample to sample variation in the coefficient (e.g., intercept, slope). While it is related to SD(Y) and SD(X), it's important to be clear what variation is being measured.