Agenda

1. Parameters and statistics
2. Hypothesis testing
3. Partitioning variability
4. Confidence and prediction intervals

Parameters and statistics Let's take another moment to review the relationship between the population and the samples we take.

Hypothesis testing In this class, the hypothesis we are most often testing is about the slope coefficient(s).

$$H_0 : \beta_1 = 0$$
$$H_A : \beta_1 \neq 0$$

Why do we write the hypothesis test about the parameter?

To test the hypothsis, we will use the following test statistic:

$$t = \frac{\hat{\beta}_1}{SE_{\hat{\beta}_1}}$$

Why is the test statistic including the estimate of the parameter?

Similarly, we could also compute a confidence interval for the slope

$$\hat{\beta}_1 \pm t^* \cdot SE_{\hat{\beta}_1}$$

An example Let's consider the cereal example again

```
require(mosaic)
require(Stat2Data)
data(Cereal)
m1 <- lm(Calories~Sugar, data=Cereal)
summary(m1)

##
## Call:
## lm(formula = Calories ~ Sugar, data = Cereal)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -37.428  -9.832   0.245   8.909  40.322
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  87.4277     5.1627  16.935   <2e-16 ***
## Sugar         2.4808     0.7074   3.507   0.0013 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 19.27 on 34 degrees of freedom
## Multiple R-squared:  0.2656,Adjusted R-squared:  0.244
## F-statistic:  12.3 on 1 and 34 DF,  p-value: 0.001296
```

Partitioning variability When we do Analysis of Variance (ANOVA) we are partitioning the variability.

Looking at the relationship between the y_is, the $\hat{y}$s, and $\bar{y}$, we can see that

$$(y_i - \bar{y}) = (\hat{y}_i - \bar{y}) + (y_i - \hat{y}_i)$$

We can use the relationship between those quantities to gain some intuition for this:

$$\begin{aligned} \sum_{i=1}^{n}(y_i - \bar{y})^2 &= \sum_{i=1}^{n}(\hat{y}_i - \bar{y}) + \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \\ SST &= SSM + SSE \end{aligned}$$

Example Again, let's think about the cereal example

```
anova(m1)

## Analysis of Variance Table
##
## Response: Calories
##           Df  Sum Sq Mean Sq F value   Pr(>F)
## Sugar      1  4567.2  4567.2  12.299 0.001296 **
## Residuals 34 12626.2   371.4
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Confidence and prediction intervals This is one of those places where statisticians shouldn't have been put in charge of the terms. We're going to talk about two different varieties of ”confidence intervals,” but we're going to call one a confidence interval and the other a prediction interval.

- Confidence intervals are about means
- Prediction intervals are about individuals

The difference is in the computation of the standard error

$$\begin{aligned} SE_{\hat{\mu}} &= \hat{\sigma_{\epsilon}}\sqrt{\frac{1}{n} + \frac{(x * -\bar{x})^2}{\sum(x-\bar{x})^2}} \\ SE_{\hat{y}} &= \hat{\sigma_{\epsilon}}\sqrt{1 + \frac{1}{n} + \frac{(x * -\bar{x})^2}{\sum(x-\bar{x})^2}} \end{aligned}$$

Example If we wanted to create a confidence interval about cereals with 6 grams of sugar, and a prediction interval about a particular cereal with 6 grams of sugar, which interval would be wider?