STA 610L: Module 2.1

One way ANOVA (formulation and estimation)

Dr. Olanrewaju Michael Akande

STA 610L

Motivating Example: Cycling Safety

Dr. Ian Walker at University of Bath carried out a project to investigate how drivers overtake bicyclists.

His team modified a bicycle subtly to carry both a video system and an accurate ultrasonic distance sensor that could record the proximities of each passing vehicle.

The team then designed an experiment in which a cyclist (Dr. Walker) varied the distance he rode from the curb (the British spelling kerb is used in the dataset) from 0.25m to 1.25m in 0.25 m increments.

Motivating Example: Cycling Safety

We will consider the outcome of passing distance y_{ij}, which is the measured distance (in m) between the vehicle and the cyclist, as a function of the distance from the bike to the curb (indexed by j), as some cyclists have postulated that "the more room you take up, the more room they give you."

We'll use these data to test this "Theory of Big."

Our research question of interest is whether the distance from the bike to the curb is indeed related to the passing distance between the bike and a vehicle.

The data is in the PsychBikeData.RData file on Sakai.

STA 610L

EDA

```r
load("data/PsychBikeData.RData")
PsychBikeData$kerb <- as.factor(PsychBikeData$kerb)
dim(PsychBikeData)
```

```
## [1] 2355   11
```

```r
head(PsychBikeData)
```

```
## # A tibble: 6 x 11
##   vehicle colour `passing distan… street Time                hour               
##   <fct>   <fct>                <dbl> <fct>  <dttm>              <dttm>             
## 1 ordina… blue                 2.11  regul… 1904-01-01 16:30:00 1904-01-01 16:00:00
## 2 HGV     red                  0.998 regul… 1904-01-01 16:30:00 1904-01-01 16:00:00
## 3 minibus blue                 1.82  regul… 1904-01-01 16:30:00 1904-01-01 16:00:00
## 4 ordina… NA                   1.64  regul… 1904-01-01 16:30:00 1904-01-01 16:00:00
## 5 bus     other                1.54  regul… 1904-01-01 16:30:00 1904-01-01 16:00:00
## 6 ordina… silve…               1.51  regul… 1904-01-01 16:30:00 1904-01-01 16:00:00
## # … with 5 more variables: helmet <fct>, kerb <fct>, Bikelane <fct>,
## #   City <fct>, date <dttm>
```

STA 610L

EDA

```
str(PsychBikeData)
```

```
## tibble [2,355 × 11] (S3: tbl_df/tbl/data.frame)
##  $ vehicle         : Factor w/ 7 levels "ordinary","minibus",..: 1 5 2 1 4 1 2 1 4 7 ...
##  $ colour          : Factor w/ 8 levels "blue","red","silver/grey",..: 1 2 1 8 7 3 4 2 2 8 ...
##  $ passing distance: num [1:2355] 2.114 0.998 1.817 1.64 1.544 ...
##  $ street          : Factor w/ 6 levels "one way, one lane",..: 3 3 3 3 3 3 5 5 5 5 ...
##  $ Time            : POSIXct[1:2355], format: "1904-01-01 16:30:00" "1904-01-01 16:30:00" ...
##  $ hour            : POSIXct[1:2355], format: "1904-01-01 16:00:00" "1904-01-01 16:00:00" ...
##  $ helmet          : Factor w/ 2 levels "no","yes": 1 1 1 1 1 1 2 2 2 2 ...
##  $ kerb            : Factor w/ 5 levels "0.25","0.5","0.75",..: 2 2 2 2 2 2 4 4 4 4 ...
##  $ Bikelane        : Factor w/ 2 levels "no","yes": 1 1 1 1 1 1 1 1 1 1 ...
##  $ City            : Factor w/ 3 levels "Salisbury","Bristol",..: 1 1 1 1 1 1 1 1 1 1 ...
##  $ date            : POSIXct[1:2355], format: "2006-05-11" "2006-05-11" ...
```

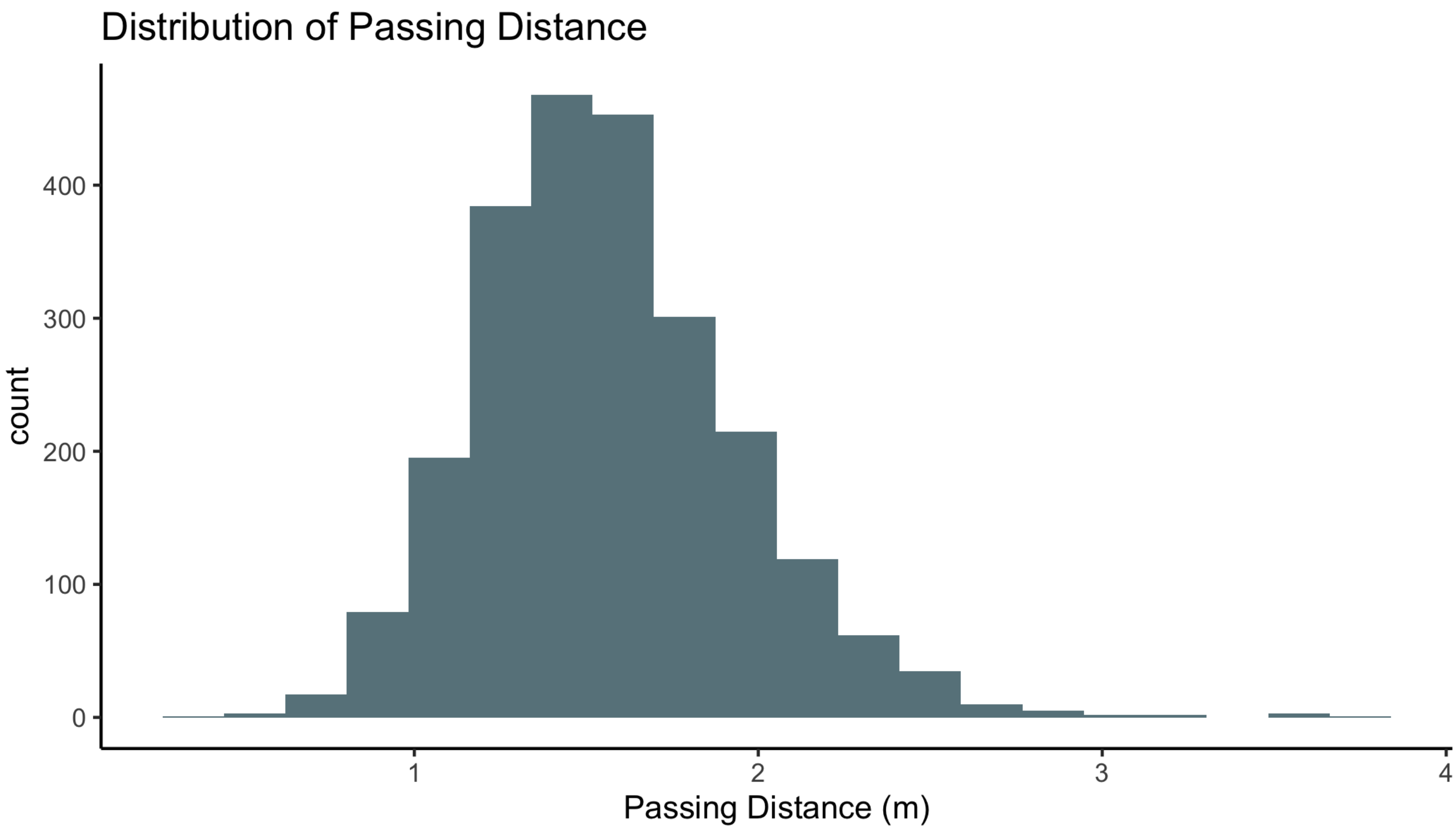
EDA

```
summary(PsychBikeData)
```

```
##                  vehicle              colour    passing distance
##  ordinary            :1708   blue       :636   Min.   :0.394   
##  minibus             : 293   silver/grey:531   1st Qu.:1.303   
##  SUV/pickup          : 143   red        :378   Median :1.529   
##  bus                 :  46   white      :333   Mean   :1.564   
##  HGV                 :  82   black      :262   3rd Qu.:1.790   
##  taxi                :  49   green      :149   Max.   :3.787   
##  powered two-wheeler:  34   (Other)    : 66                   
##                          street          Time                       
##  one way, one lane          :   9   Min.   :1904-01-01 07:46:00  
##  one way, 2 lanes           :  13   1st Qu.:1904-01-01 10:14:00  
##  regular urban street       : 655   Median :1904-01-01 12:13:00  
##  regular residential street:  39   Mean   :1904-01-01 12:40:09  
##  main road, regular         :1637   3rd Qu.:1904-01-01 15:30:00  
##  rural                      :   2   Max.   :1904-01-01 17:12:00  
##                                                                  
##       hour                         helmet       kerb     Bikelane  
##  Min.   :1904-01-01 07:00:00   no :1206   0.25:670   no :2305  
##  1st Qu.:1904-01-01 10:00:00   yes:1149   0.5 :545   yes:  50  
##  Median :1904-01-01 12:00:00              0.75:339              
##  Mean   :1904-01-01 12:05:38              1   :469              
##  3rd Qu.:1904-01-01 15:00:00              1.25:332              
##  Max.   :1904-01-01 17:00:00                                    
##                                                                  
##         City           date                       
##  Salisbury :1905   Min.   :2006-05-11 00:00:00  
##  Bristol   : 450   1st Qu.:2006-05-20 00:00:00  
##  Portsmouth:   0   Median :2006-05-27 00:00:00  
##                    Mean   :2006-05-27 12:08:15  
##                    3rd Qu.:2006-05-31 00:00:00  
##                    Max.   :2006-06-19 00:00:00  
## 
```

EDA

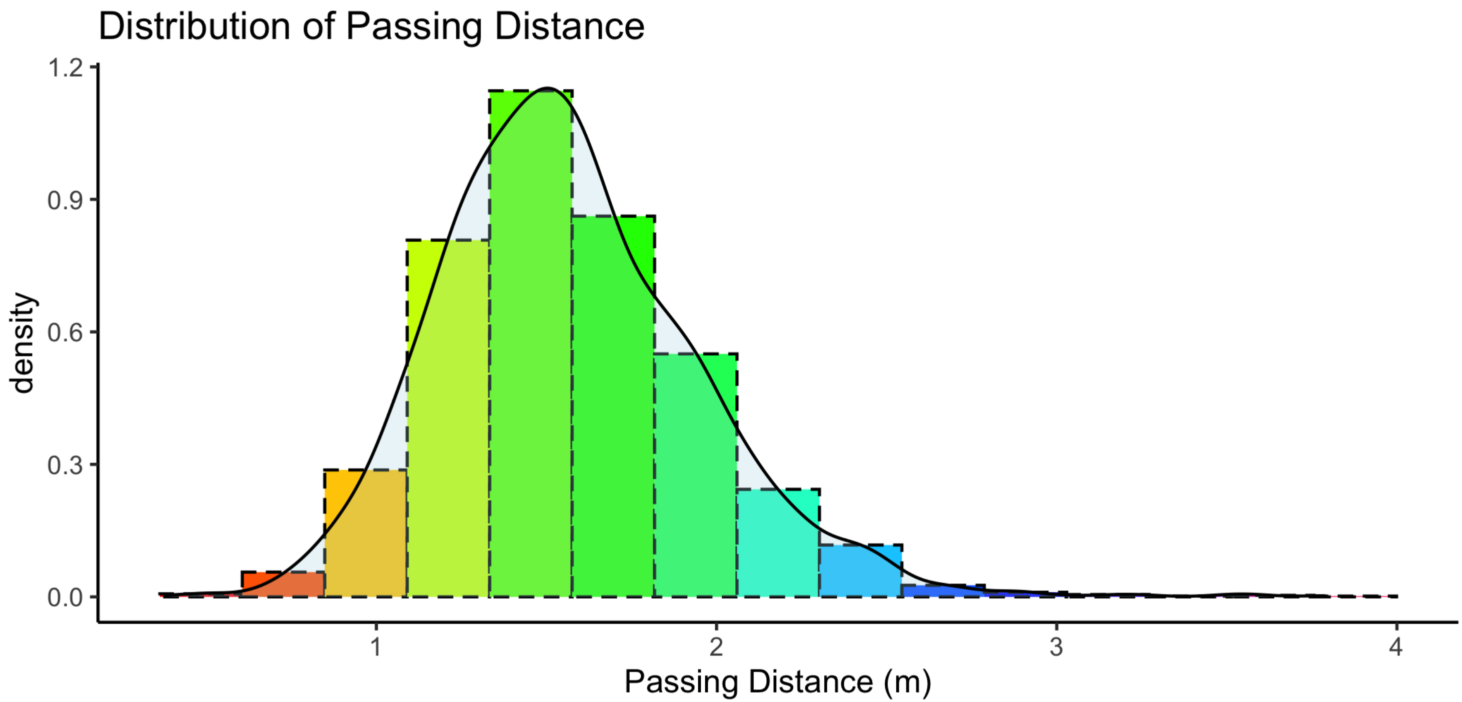
```
ggplot(PsychBikeData,aes(`passing distance`)) +
  geom_histogram(fill="lightblue4",bins=20) + theme(legend.position="none") +
  labs(title="Distribution of Passing Distance",x="Passing Distance (m)") +
  theme_classic()
```



STA 610L

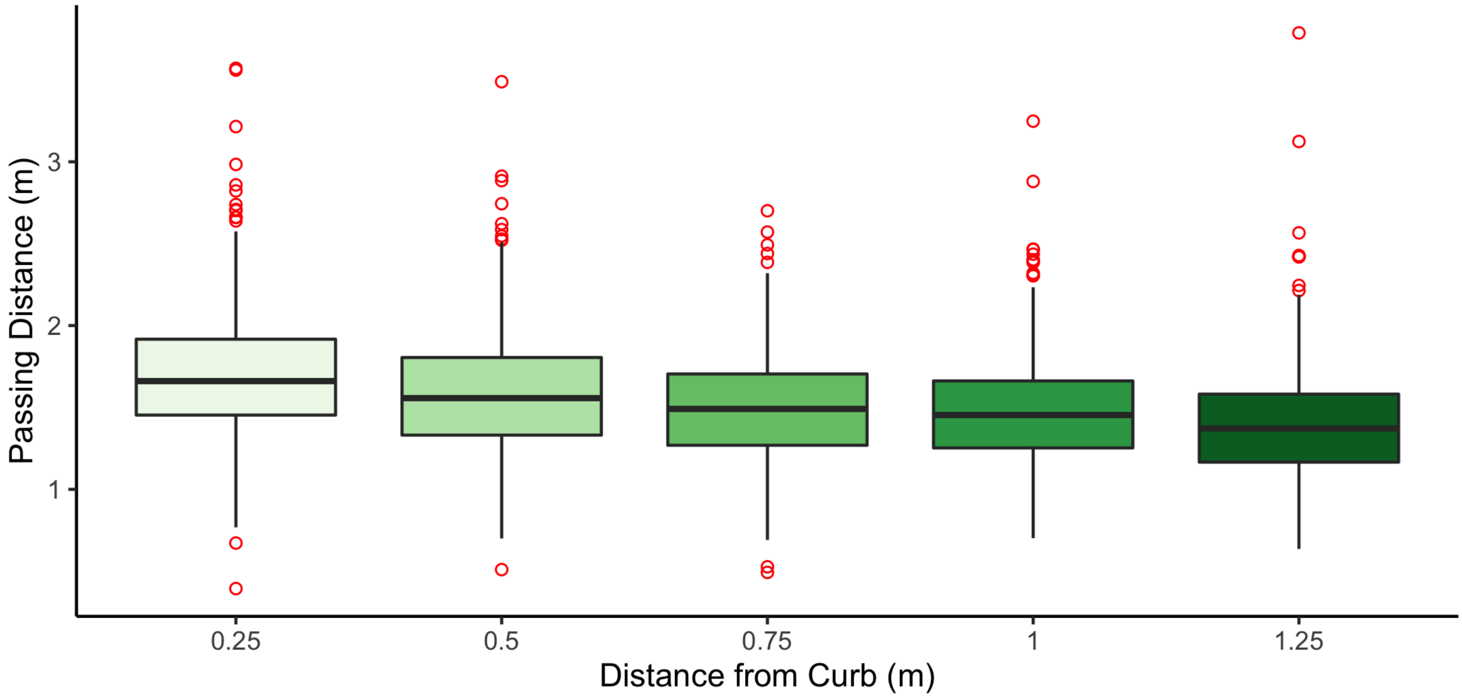
EDA

```
ggplot(PsychBikeData,aes(`passing distance`)) +
  geom_histogram(aes(y=..density..),color="black",linetype="dashed",
                fill=rainbow(15),bins=15) + theme(legend.position="none") +
  geom_density(alpha=.25, fill="lightblue") + scale_fill_brewer(palette="Blues") +
  labs(title="Distribution of Passing Distance",x="Passing Distance (m)") +
  theme_classic()
```



EDA

```
ggplot(PsychBikeData,aes(y=`passing distance`, x=kerb, fill=kerb)) +
  geom_boxplot(outlier.colour = "red", outlier.shape = 1) +
  scale_fill_brewer(palette="Greens") +
  labs(x="Distance from Curb (m)", y = "Passing Distance (m)") +
   theme_classic() + theme(legend.position="none")
```



Research question: is distance from curb related to passing distance?

EDA

```
table(PsychBikeData$kerb)
```

```
## 
## 0.25  0.5 0.75    1 1.25 
##  670  545  339  469  332
```

```
tapply(PsychBikeData$`passing distance`,PsychBikeData$kerb,mean)
```

```
##     0.25      0.5     0.75        1     1.25 
## 1.698054 1.590473 1.505519 1.490584 1.412813
```

STA 610L

ANOVA Model

Consider the model

$$
\begin{aligned}
y_{ij} &= \mu + \alpha_j + \varepsilon_{ij} && \text{(treatment effects model)} \\
&= \mu_j + \varepsilon_{ij} && \text{(treatment means model)}
\end{aligned}
$$

where $\mu_j = \mu + \alpha_j$.

These two equations are simply alternate parameterizations of the same model.

In each case, we estimate a separate mean passing distance $\mu_j = \mu + \alpha_j$ as a function of each of the 5 curb distances tested.

STA 610L

ANOVA Model

$$y_{ij} = \mu + \alpha_j + \varepsilon_{ij} = \mu_j + \varepsilon_{ij}$$

- μ: expected passing distance (grand mean).
- μ_j: expected passing distance for curb distance j, with $j = 1, \ldots, J = 5$.
- α_j: deviation between overall expected passing distance and expected passing distance for curb distance j.
- ε_{ij}: deviations of observed passing distance from curb-distance-specific expectations.
- In the standard ANOVA model $\sum_j \alpha_j = 0$ is assumed so that μ represents an overall mean across groups.
- Another coding scheme: set one $\alpha_j = 0$ so that μ represents the expected passing distance in that particular group, and remaining α_j measure differences from expected passing distance in that reference group.

STA 610L

ANOVA Model

We also assume that $\varepsilon_{ij} \overset{iid}{\sim} f(\varepsilon)$ with $\mathbb{E}(\varepsilon_{ij}) = 0$ within all groups j.

The expected passing distance for occasion i in with curb distance j is then

$$\begin{aligned}\mathbb{E}(y_{ij} \mid \mu, \alpha_1, \cdots, \alpha_J) &= \mathbb{E}(\mu + \alpha_j + \varepsilon_{ij} \mid \mu, \alpha_1, \cdots, \alpha_J) \\ &= \mu + \alpha_j \\ &= \mu_j\end{aligned}$$

If we assume $f(\varepsilon) = N\left(0, \sigma^2\right)$, then our model is $y_{ij} \sim N\left(\mu + \alpha_j, \sigma^2\right)$ or equivalently $y_{ij} \sim N\left(\mu_j, \sigma^2\right)$.

Parameter estimation

Let $\widehat{\mu} = (\widehat{\mu}_1, \cdots, \widehat{\mu}_J)$ be our estimates of the unknown parameters $\mu = (\mu_1, \cdots, \mu_J)$.

The residual for y_{ij} is the difference between the observed y_{ij} and our fitted value $\hat{y}_{ij}$ and is given by

$$\hat{\varepsilon}_{ij} = y_{ij} - \hat{y}_{ij} = y_{ij} - \widehat{\mu}_j.$$

The ordinary least squares (OLS) estimate of μ, $\widehat{\mu}_{OLS}$, is the value that minimizes the sum of squared residuals (sum of squared errors) given by

$$SSE(\mu) = \sum_j \sum_i (y_{ij} - \mu_j)^2.$$

OLS Estimates

You can show (homework!) that the OLS estimates are given by

- $(\widehat{\mu}_1, \cdots, \widehat{\mu}_J) = (\overline{y}_1, \cdots, \overline{y}_J)$, where $\overline{y}_j$ is the sample mean in group j.
- $\widehat{\mu} = \overline{y}$, where $\overline{y}$ is the grand mean over all observations.
- $\widehat{\mu} = \frac{1}{J}\sum_j \widehat{\mu}_j$ when the sample sizes in each group j, n_j, are equal for all groups.
- $\widehat{\alpha}_j = \widehat{\mu}_j - \widehat{\mu} = \overline{y}_j - \overline{y}$.

A helpful mnemonic may be the following "decomposition" of a single data point:

$$
\begin{aligned}
y_{ij} &= y_{ij} + \overline{y}_j - \overline{y}_j + \overline{y} - \overline{y} \\
&= \overline{y} + (\overline{y}_j - \overline{y}) + (y_{ij} - \overline{y}_j) \\
&= \widehat{\mu} \quad + \quad \widehat{\alpha}_j \quad + \quad \hat{\varepsilon}_{ij}
\end{aligned}
$$

STA 610L

Sums of Squares

Based on those ideas, let's decompose the variability of the data around the grand mean into variation within groups (error) and variation between or across groups (group effects).

For simplicity, suppose we have J groups with n_j observations in each group.

We break down the total variation of the data around the overall mean as follows:

$$SST = SSG + SSE,$$

where

- SST is the total sum of squared deviations around the overall mean,
- SSG is the portion of the total sum of squares due to group differences, and
- SSE is the portion of the total sum of squares due to differences between the individual observations and their own group means.

STA 610L

Sums of Squares

We define the sums of squares as follows:

- $\mathrm{SST} = \sum_{j=1}^{J} \sum_{i=1}^{n_j} \left(y_{ij} - \bar{y}\right)^2$

- $\mathrm{SSG} = \sum_{j=1}^{J} \sum_{i=1}^{n_j} \left(\bar{y}_j - \bar{y}\right)^2$

- $\mathrm{SSE} = \sum_{j=1}^{J} \sum_{i=1}^{n_j} \left(y_{ij} - \bar{y}_j\right)^2$

STA 610L

ANOVA Table

The main use of ANOVA is to evaluate the hypothesis that there are no differences across groups, e.g. $H_0 : \mu_j = \mu_{j'} \; \forall \; j \neq j'$ against the alternative that at least one mean is different.

Testing in ANOVA involves comparison of the mean squares for groups and the mean squares for error (we'll come back to why this is sensible) and can be summarized in the ANOVA table.

Let $N = \sum_j n_j$ be the overall sample size.

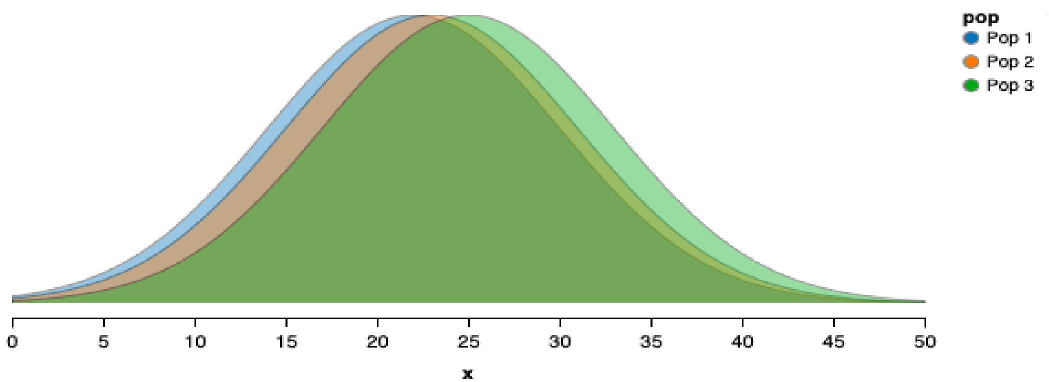
Source	DF	SS	MS	F	p-value
Groups	$J-1$	SSG	$MSG = \frac{SSG}{J-1}$	$\frac{MSG}{MSE}$	from $F_{J-1,N-J}$
Error	$N-J$	SSE	$MSE = \frac{SSE}{N-J}$		
Total	$N-1$	SST			

STA 610L

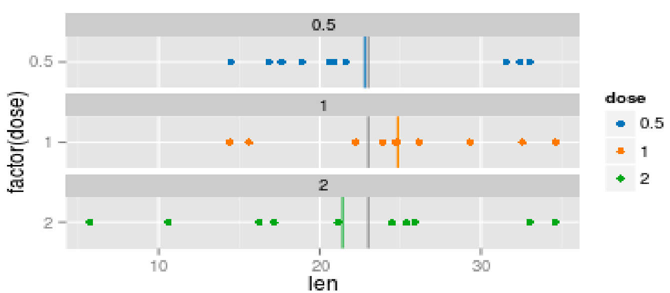
The variations in ANOVA

Using this Shiny app you can explore the roles of within-group and between-group variance in ANOVA.

Population distributions



Observed sample data



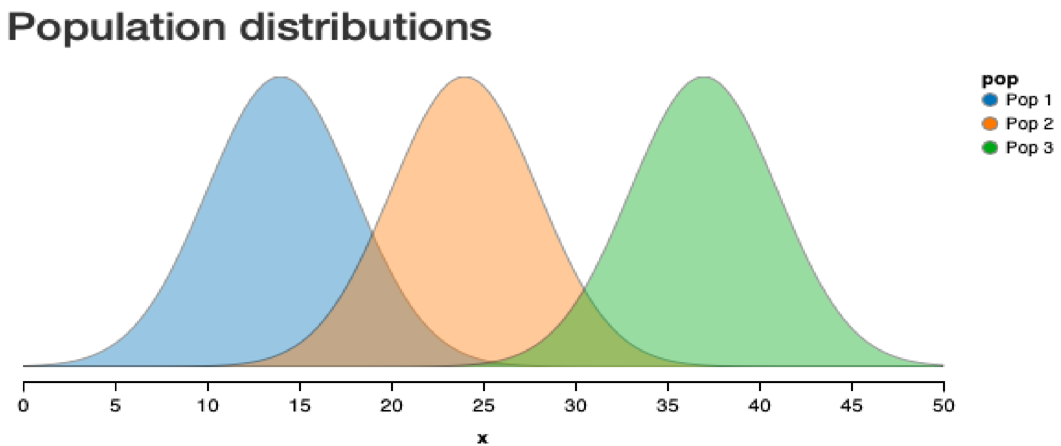
STA 610L

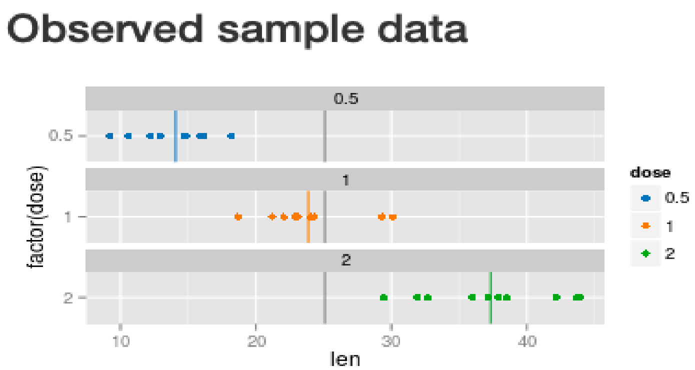
The variations in ANOVA

Here you see a situation with large within-group variance relative to the between-group variance (e.g., not much of a group effect relative to the variability within groups)

STA 610L

The variations in ANOVA





In this case, the means are further apart and the between-group variance is larger than in the prior figure, and differences among groups are more apparent.

STA 610L

MSE

The MSE can be written

$$
\begin{aligned}
MSE &= \frac{SSE}{\sum_j (n_j - 1)} \\
&= \frac{\sum_{j=1}^{J} \sum_{i=1}^{n_j} \left(y_{ij} - \bar{y}_j\right)^2}{\sum_j (n_j - 1)} \\
&= \frac{\sum_{i=1}^{n_1} \left(y_{i1} - \bar{y}_1\right)^2 + \cdots + \sum_{i=1}^{n_J} \left(y_{iJ} - \bar{y}_J\right)^2}{(n_1 - 1) + \cdots + (n_J - 1)} \\
&= \frac{(n_1 - 1)s_1^2 + \cdots + (n_J - 1)s_J^2}{(n_1 - 1) + \cdots + (n_J - 1)}
\end{aligned}
$$

STA 610L

MSE

In ANOVA, we typically assume independence of observations and equal variances within all the groups.

We see that the $MSE = \frac{(n_1-1)s_1^2+\cdots+(n_J-1)s_J^2}{(n_1-1)+\cdots+(n_J-1)}$ is a pooled estimate of the within-group sample variance, and you can show that $\mathbb{E}(\mathrm{MSE}) = \sigma^2$ if our assumption of equal variances holds.

Using algebra, you can show that $\mathbb{E}(MSG) = \sigma^2 + \frac{\sum n_j(\mu_j-\mu)^2}{J-1}$. Under the null hypothesis that all the group means are the same, this expectation reduces to σ^2.

Thus under H_0, $\mathbb{E}\left(F = \frac{MSG}{MSE}\right) = 1$, but if the group means are in fact different from each other, we expect $\mathrm{MSG} > \sigma^2$ and $F > 1$.

Under the assumption that $\varepsilon_{ij} \overset{iid}{\sim} N(0, \sigma^2)$, if H_0 is also true, then

$$F = \frac{MSG}{MSE} \sim F_{J-1,N-J}.$$

STA 610L

Back to Passing Bikes

```
aov.res=aov(`passing distance`~kerb,data=PsychBikeData)
summary(aov.res)
```

```
##               Df Sum Sq Mean Sq F value Pr(>F)
## kerb           4   23.7   5.925   43.18 <2e-16
## Residuals   2350  322.4   0.137
```

This F test indicates that it is very unlikely we would see differences in passing distance as large as we did under the null hypothesis that all groups have the same mean.

There is a difference in passing distance for at least one set of curb distances.

What's next?

Move on to the readings for the next module!

STA 610L