STAT021 Statistical Methods II

Lecture 6 One-way ANOVA Table

Lu Chen
Swarthmore College
9/20/2018

Outline

- Review: one-way ANOVA model
- ASSESS model
  - F distribution
  - ANOVA F test
  - One-way ANOVA table
- ASSESS error
  - Check assumptions $\epsilon \overset{iid}{\sim} N(0, \sigma)$
- Deal with violation in assumptions
  - Re-CHOOSE, re-FIT and re-ASSESS

Review - One-way ANOVA Model

One-Way Analysis of Variance Model

The **ANOVA model** for a quantitative response variable and one categorical explanatory variable with K values is

$$\begin{array}{ccccccc} \text{Data} & = & \text{Grand Mean} & + & \text{Group Effect} & + & \text{Error} \\ Y & = & \mu & + & \alpha_k & + & \epsilon \end{array}$$

where k refers to the specific category of the explanatory variable and $k = 1, 2, \cdots, K$, and $\epsilon \overset{iid}{\sim} N(0, \sigma)$.

The null and alternative hypotheses for the ANOVA model are

$H_0 : \alpha_1 = \alpha_2 = \cdots = \alpha_K = 0$;

H_a: at least one $\alpha_k \neq 0$.

Review - ASSESS model: Triple decomposition

$$\text{Data} = \text{Grand Mean} + \text{Group Effect} + \text{Error}$$

$$Y = \mu + \alpha_k + \epsilon$$

$$y = \bar{y} + \bar{y}_k - \bar{y} + y - \bar{y}_k$$

Data

$$y - \bar{y} = \bar{y}_k - \bar{y} + y - \bar{y}_k$$

Sum of squares

$$SSTotal = SSGroup + SSE$$

$$\sum(y - \bar{y})^2 = \sum(\bar{y}_k - \bar{y})^2 + \sum(y - \bar{y}_k)^2$$

Total variability in data (null model residuals)

Variability explained by the ANOVA model

Variability left in the ANOVA model residuals

Degrees of freedom

$$df_{Total} = df_{Group} + df_{Error}$$

$$n - 1 = K - 1 + n - K$$

Review - ASSESS model: Mean square

- Average variability: Mean square

$$\text{Mean Square} = \frac{\text{Sum of Squares}}{\text{Degree of Freedom}}$$

$$MSGroup = \frac{SSGroup}{df_{Group}} = \frac{\sum(\bar{y}_k - \bar{y})^2}{K-1}$$

$$MSE = \frac{SSE}{df_{Error}} = \frac{\sum(y - \bar{y}_k)^2}{n-K}$$

- MSE, **mean square error**, is the estimate of σ^2 (or $\sqrt{MSE}$ is the estimate of σ)
- We denote the estimate of σ as $\hat{\sigma}$.

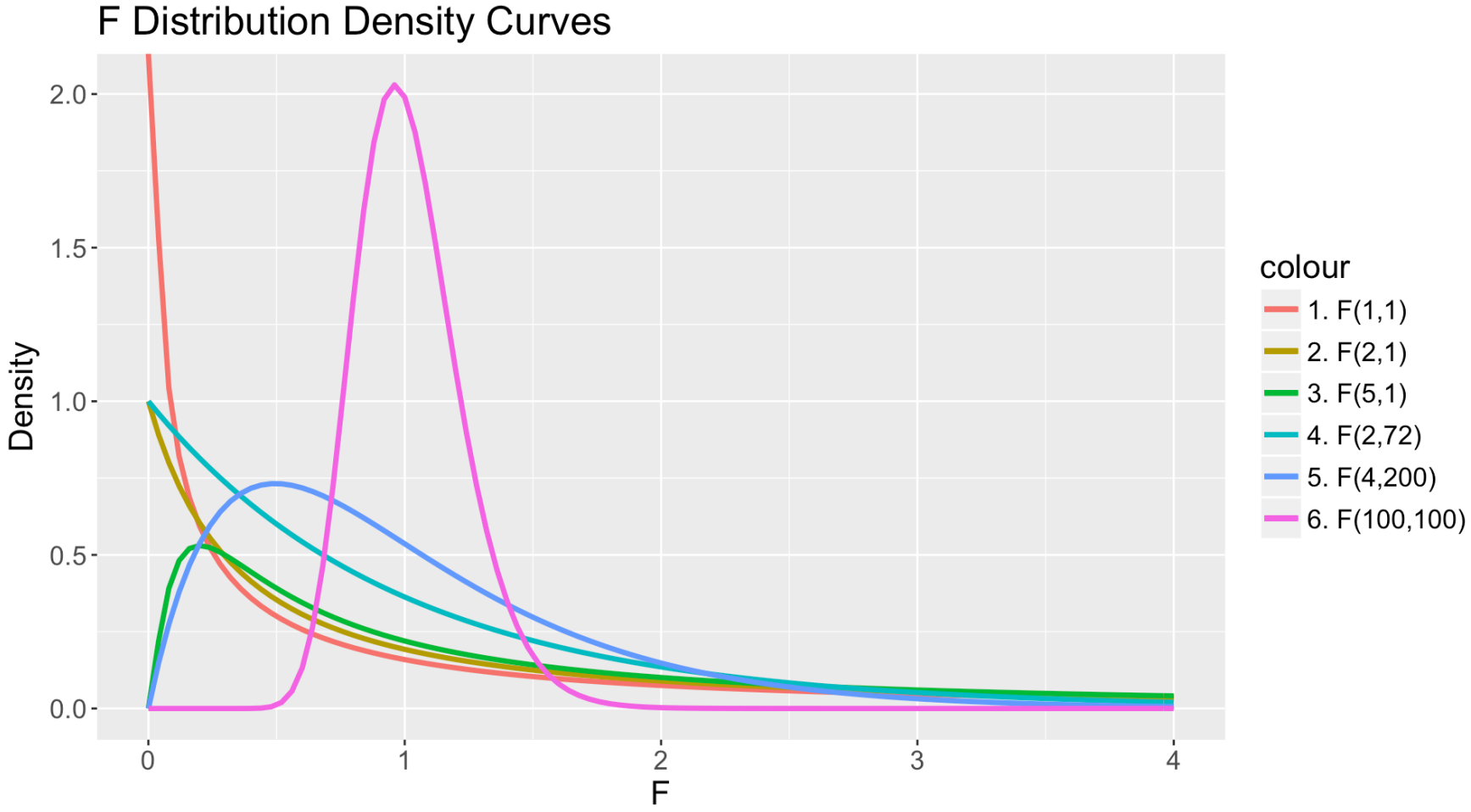
ASSESS model: F statistic

The ANOVA method compares the **average variability explained by the model** to the **average variability left in the residuals** using the F **statistic**

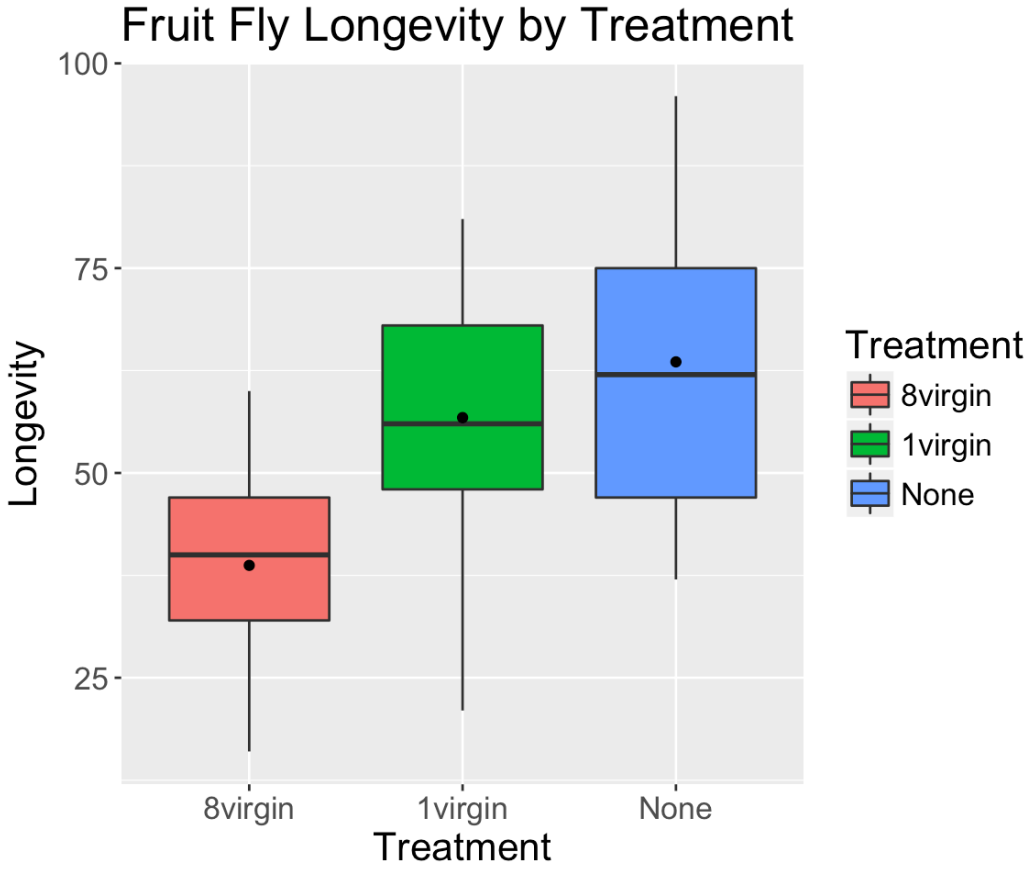
$$F = \frac{MSGroup}{MSE} \sim F(K-1, n-K)$$

- $F > 0$, the value of F is always positive.
- F distribution has two parameters; both are called **degree of freedom**, one from $MSGroup$ and the other from MSE.
- F is large, the ANOVA model is better than the null model. We use the ANOVA model to describe the data.
- F is small, the ANOVA model is NOT better than the null model. We use the simpler null model to describe the data.

ASSESS model: *F* distribution



ASSESS model: ANOVA *F* test



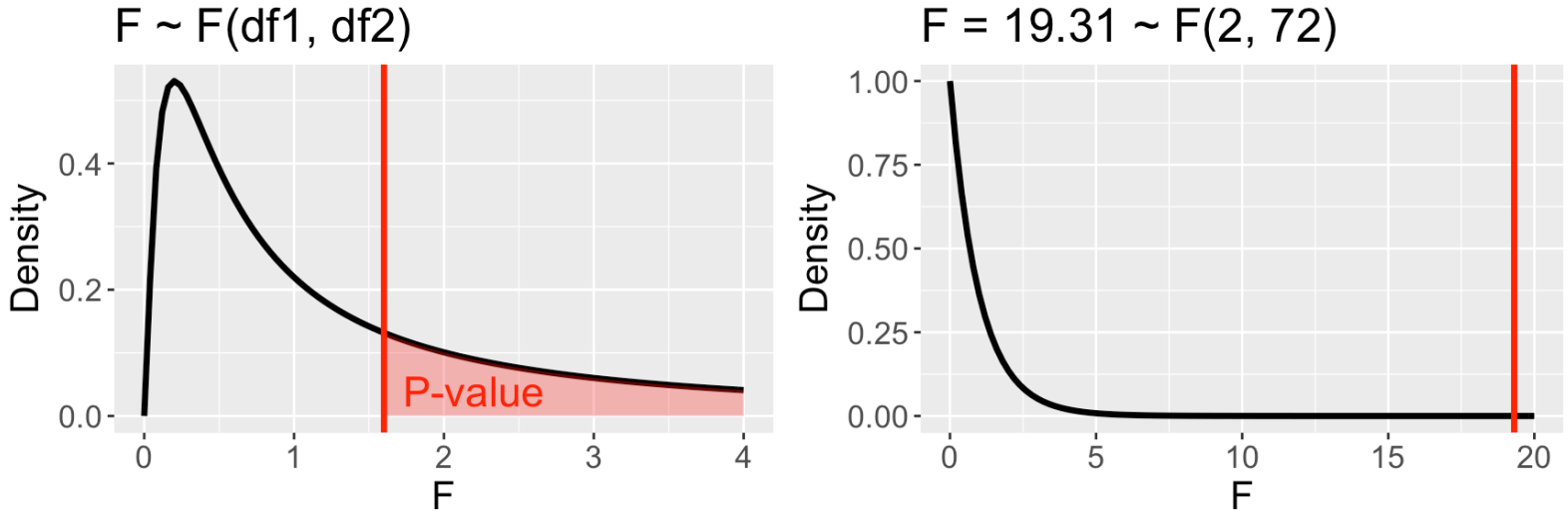
- For the fruit fly example, 75 male fruit flies were assigned to each of the *8virgin*, *1virgin* and *None* group. Their *Longevity* was compared.
- What is the distribution of the F statistic?
- $F = \frac{MSGroups}{MSE} \sim F(K-1, n-K)$

  $K - 1 = 3 - 1 = 2$

  $n - K = 75 - 3 = 72$

  $\Rightarrow F = 19.31 \sim F(2, 72)$

ASSESS model: ANOVA F test




```
1-pf(19.31, df1=2, df2=72)
```

```
## [1] 1.931783e-07
```

- The P-value of an F test is computed as the probability that an $F(df1, df2)$ variable is greater than a certain F value: $P\text{-value} = P(F_{df1,df2} > F)$
- For the fruit fly example, $P\text{-value} = P(F_{2,72} > 19.31) = 1.93 \times 10^{-7}$.

ASSESS model: One-way ANOVA table

The **null and alternative hypotheses** of the ANOVA model are

- $H_0 : \alpha_0 = \alpha_1 = \cdots = \alpha_K = 0$
- H_a : at least one $\alpha_k \neq 0$

and the **ANOVA table** is

	Degree of Freedom	Sum of Squares	Mean Square	F statistic	P-value
Model	$K-1$	SSG	MSG	$F = \frac{MSG}{MSE}$	$P(F_{K-1,n-K} > F)$
Error	$n-K$	SSE	MSE		
Total	$n-1$	SST			

If the proper conditions hold, the P-value is calculated using the upper tail of an F distribution with $K-1$ and $n-K$ degrees of freedom.

ASSESS model: One-way ANOVA table in R

```
flymodel <- aov(Longevity ~ Treatment, data=fly) # analysis of variance model
summary(flymodel) # get the ANOVA table
```

```
##             Df Sum Sq Mean Sq F value   Pr(>F)
## Treatment    2   8239    4120   19.31 1.93e-07 ***
## Residuals   72  15360     213
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

- $df_{Group} = 2, df_{Error} = 72$
- $SSGroup = 8239, SSE = 15350$
- $MSGroup = 4120 = 8239/2, MSE = 213 = 15360/72, \hat{\sigma} = \sqrt{MSE} = 14.6$
- $F = 4120/213 = 19.31, P = 1.93 \times 10^{-7} < 0.05$
- We reject H_0 that all the groups have the same population mean and conclude that at least one of the *8virgin*, *1virgin* and *None* group has a significantly different mean.

ASSESS error: Check model assumptions

ANOVA model assumptions: $\epsilon \overset{iid}{\sim} N(0, \sigma)$

1. Zero mean: mean of ϵ is 0.

- Mean of residuals for each group and overall are always zero.

2. Equal variance: $\text{Var}(\epsilon) = \sigma^2$ is the same for all groups.

- Plot residuals vs. fitted (predicted) values.
- Compute s_{max}/s_{min} and compare the ratio to 2.

3. Normal distribution: $\epsilon \sim N(0, \sigma)$.

- Normal Q-Q plot

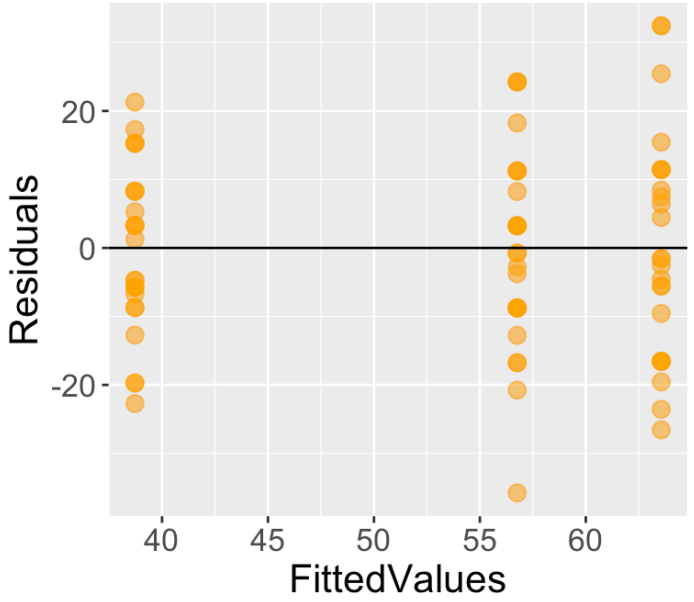
4. Independence: errors are independent of each other.

- Consider how data were collected.
- If the observations are not independent, ANOVA model is not applicable.

ASSESS error: Check model assumptions

Assumption: Equal variance

```
Assess <- data.frame(FittedValues=flymodel$fitted.values,
                     Residuals=flymodel$residuals) # prepare the data
ggplot(data=Assess, aes(x=FittedValues, y=Residuals))+
  geom_point(size=3, color="orange", alpha=0.6)+ # plot points
  geom_hline(yintercept=0) # add y=0 line
```

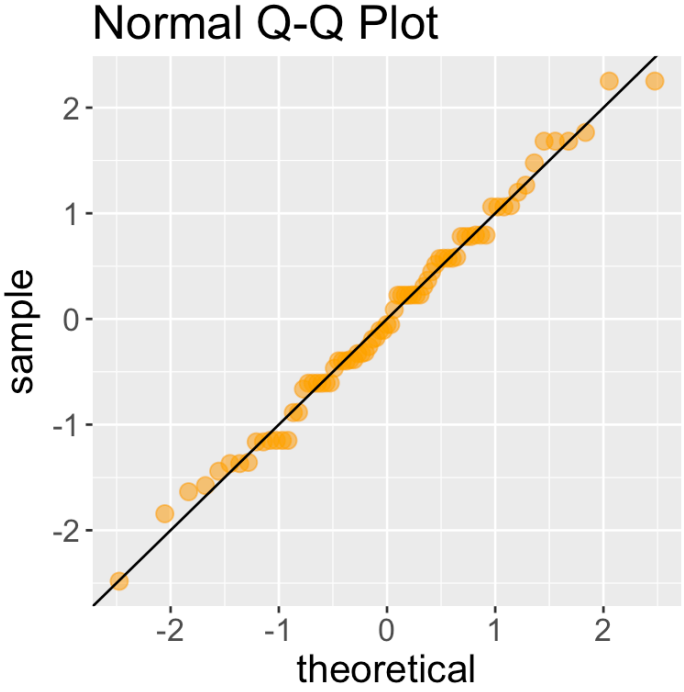


- Fitted values: predicted y, group means. The spread of the points about the $y = 0$ line should be roughly the same for all groups.
- Ratio of the largest and the smallest group SD $s_{max}/s_{min} \leq 2$. It can be somewhat larger than 2 if number of groups is large and sample size of each group is small. For the fruit fly example, $s_1 = 12.1$, $s_2 = 14.9$ and $s_3 = 16.5 \Rightarrow s_{max}/s_{min} = 16.5/12.1 = 1.4$

ASSESS error: Check model assumptions

Assumption: Normal distribution

```
ggplot(data=Assess, aes(sample = scale(Residuals)))+ # scale(): standardizing
  stat_qq(size=3, color="orange", alpha=0.6)+ # Q-Q plot
  geom_abline(intercept=0, slope=1)+ # add y=x line
  ggtitle("Normal Q-Q Plot")
```



- All points lie very closely to the $y = x$ line. The Normal assumption is satisfied.

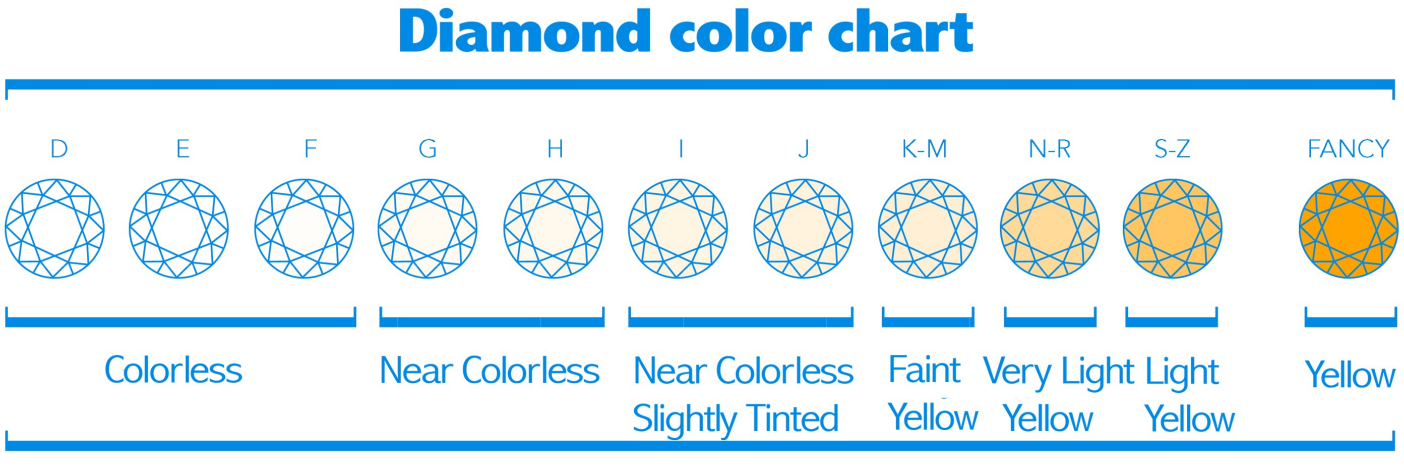
What if the equal variance and/or Normal distribution assumption are **violated**?

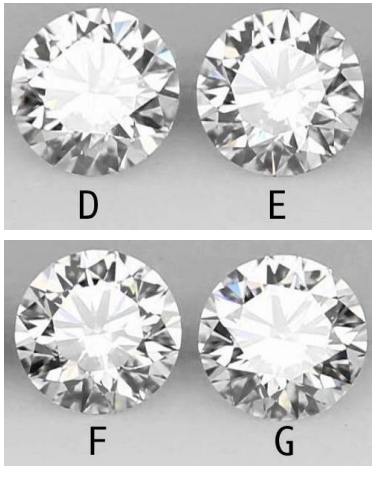
1. Transform the data to be Normally distributed
2. Use another distribution to describe the error
3. Non-parametric methods

   ...

Data example: Diamond carats and colors

Diamonds have several different characteristics that people consider before buying them. Most think about the number of carats, color, cut, clarity in a particular diamond, and probably also the price. A prospective buyer who is interested in diamonds with more carats might want to know **if a particular color of diamond is associated with more or fewer carats**. This dataset contains diamonds with color D, E, F, G, and number of carats from 0.31 to 3.35.

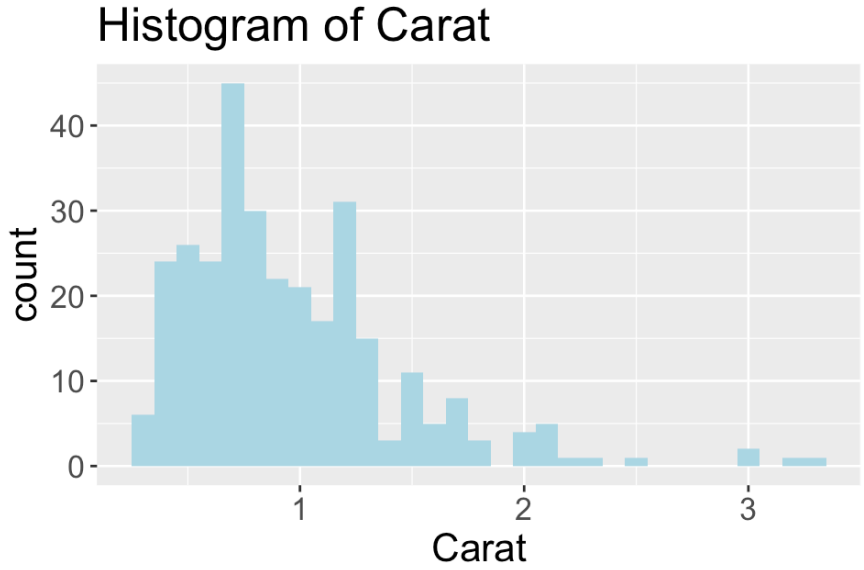


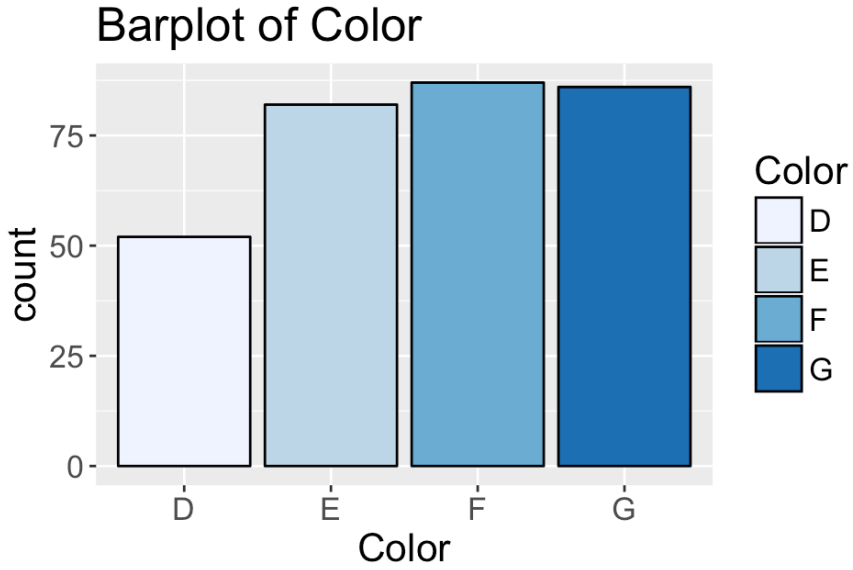


CHOOSE

Exploratory data analysis

- Response variable: *Carat*, quantitative.
  - Mean 0.97, SD 0.49; Sample size 307.
- Explanatory variable: *Color*, categorical.
  - D: 52; E: 82; F: 87; G: 86.



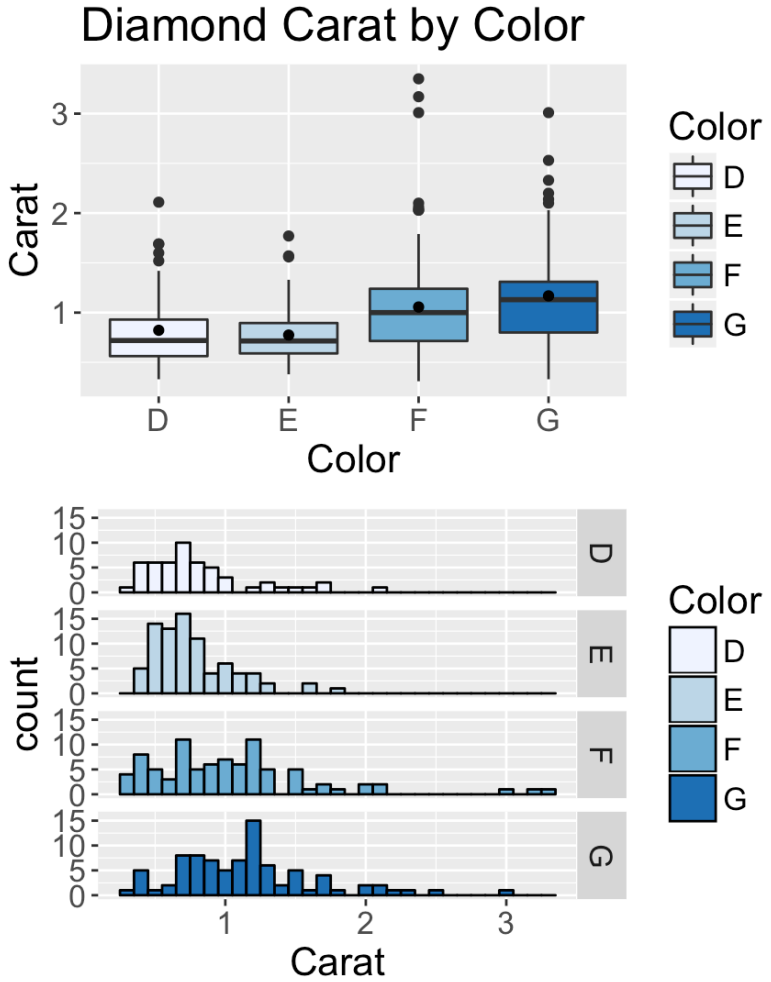


CHOOSE

Exploratory data analysis

- Response variable: *Carat*, quantitative.
  - Mean 0.97, SD 0.49; Sample size 307.
- Explanatory variable: *Color*, categorical.
  - D: 52; E: 82; F: 87; G: 86.

Color	Size	Mean	SD
D	$n_1 = 52$	$\bar{y}_1 = 0.82$	$s_1 = 0.39$
E	$n_2 = 82$	$\bar{y}_2 = 0.77$	$s_2 = 0.29$
F	$n_3 = 87$	$\bar{y}_3 = 1.06$	$s_3 = 0.59$
G	$n_4 = 86$	$\bar{y}_4 = 1.17$	$s_4 = 0.50$
All	$n = 307$	$\bar{y} = 0.97$	$s = 0.49$



CHOOSE and FIT: ANOVA model

Model: $y = \mu + \alpha_k + \epsilon$, where $k = 1, 2, 3, 4$ and $\epsilon \overset{iid}{\sim} N(0, \sigma)$.

- Null hypothesis: Number of diamond carats is the same for different colors
  $H_0 : \alpha_1 = \alpha_2 = \alpha_3 = \alpha_4 = 0$
- Alternative hypothesis: At least one color has different number of carats
  H_a: at least one $\alpha_k \neq 0$

FIT

- Parameters: $\mu, \alpha_1, \alpha_2, \alpha_3, \alpha_4, \sigma$. It is equivalent to estimating $\mu_1, \mu_2, \mu_3, \mu_4$ and σ.
- $\bar{y}_1, \bar{y}_2, \bar{y}_3$ and $\bar{y}_4$ are calculated in the exploratory data analysis. σ needs to be estimated by R.

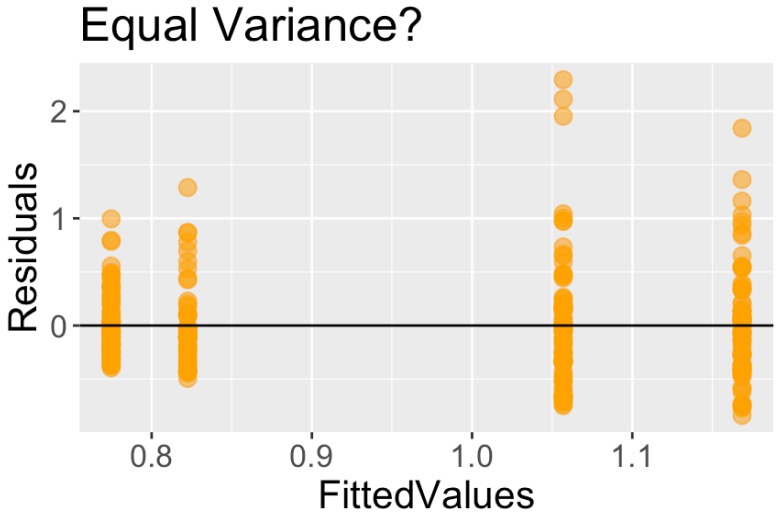
ASSESS model

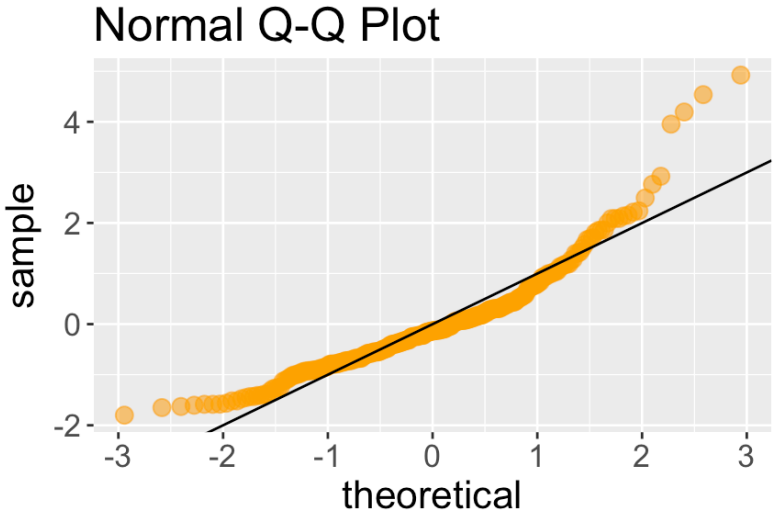
```
diamodel <- aov(Carat ~ Color, data=Diamonds)
summary(diamodel)
```

```
##              Df Sum Sq Mean Sq F value  Pr(>F)
## Color         3   8.30   2.767   12.63 8.4e-08 ***
## Residuals   303  66.36   0.219
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

- $\hat{\sigma} = \sqrt{MSE} = \sqrt{0.219} = 0.468$.
- Before we draw our final conclusion based on the F statistic and P-value, we should first assess the error term assumptions. Check the residuals and see whether they have
  - Zero mean (always true)
    Equal variance
    Normal distribution
    Independence (yes based on the data collecting process)

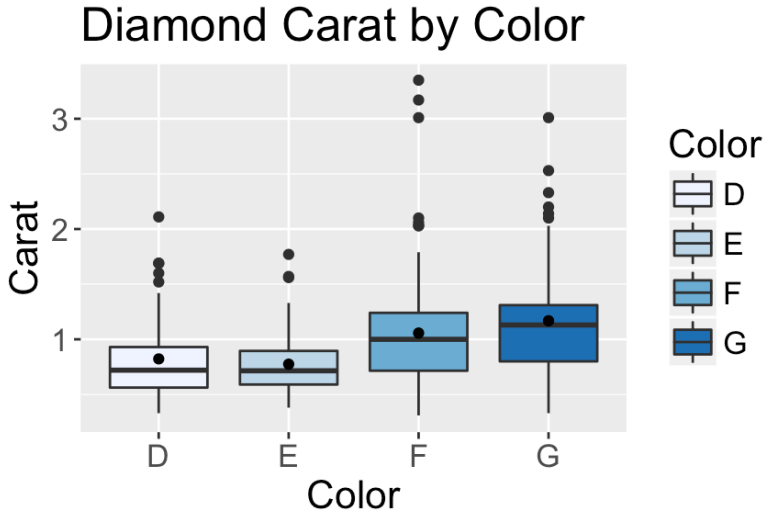
ASSESS error

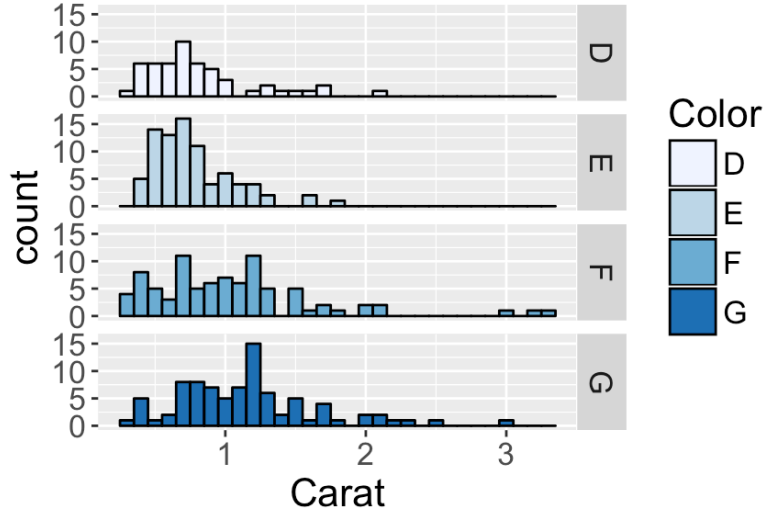




- **Equal variance**: the residuals are not evenly distributed around the $y = 0$ line. The spread is not consistent across the groups, either.
  $s_{max}/s_{min} = 0.59/0.29 = 2.1 > 2$
- **Normal distribution**: the points in the Q-Q plot have a curved trend and some lie far away from the $y = x$ line.
- Both the equal variance and Normal distribution assumptions are **violated**.

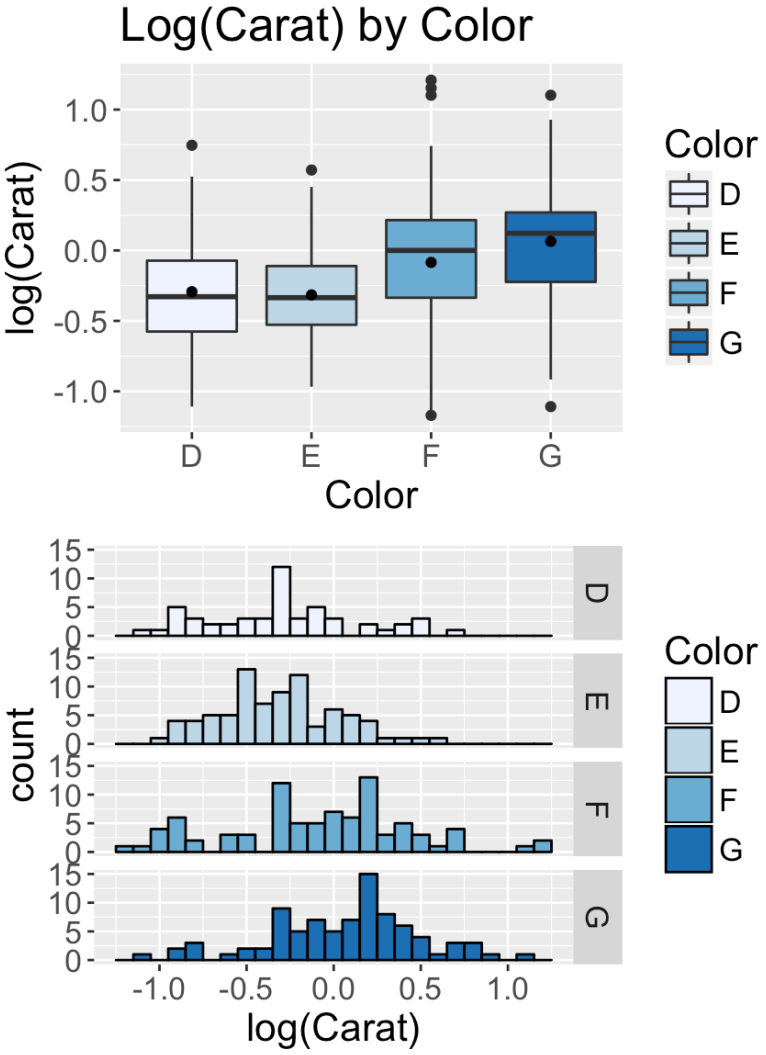
Re-CHOOSE





- The distribution of *Carat* is highly right skewed with quite some outliers in every *Color* group.
- When distribution is right skewed, we can try the **natural logarithm tranformation**.
- In the new model, Y is no longer *Carat* but $log(Carat)$.

Re-CHOOSE: After tranformationrmation



- Response variable: *log(Carat)*, quantitative.
- Explanatory variable: *Color*, categorical.

Color	Size	Mean	SD
D	$n_1 = 52$	$\bar{y}_1 = -0.29$	$s_1 = 0.44$
E	$n_2 = 82$	$\bar{y}_2 = -0.32$	$s_2 = 0.34$
F	$n_3 = 87$	$\bar{y}_3 = -0.08$	$s_3 = 0.54$
G	$n_4 = 86$	$\bar{y}_4 = 0.06$	$s_4 = 0.44$
All	$n = 307$	$\bar{y} = -0.14$	$s = 0.47$

Note: the y values could be negative because of the log() function.

Re-CHOOSE and Re-FIT: After tranformation

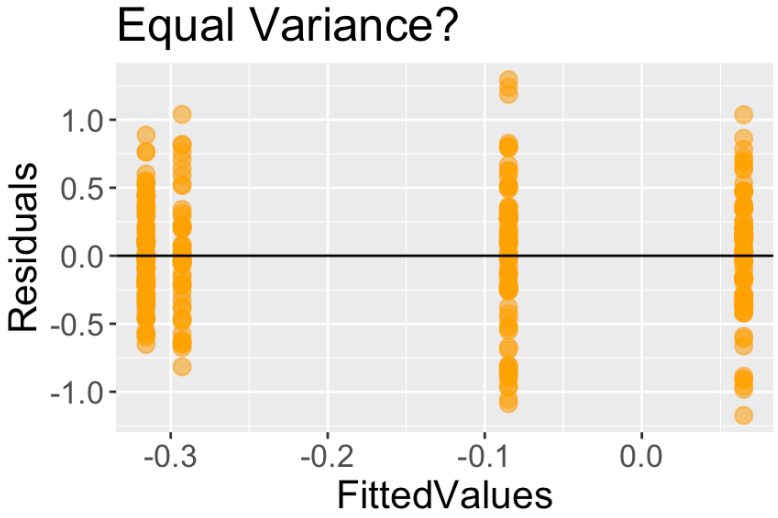
Model: $Y = \mu + \alpha_k + \epsilon$, where $Y = log(Carat)$, $k = 1, 2, 3, 4$ and $\epsilon \overset{iid}{\sim} N(0, \sigma)$.

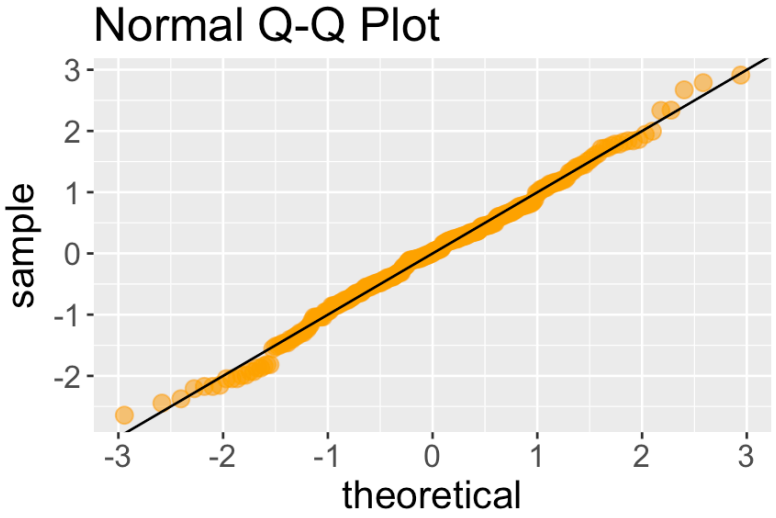
- Null hypothesis: Log number of carats is the same for different colors
  $H_0 : \alpha_1 = \alpha_2 = \alpha_3 = \alpha_4 = 0$
- Alternative hypothesis: At least one color has different log number of carats
  H_a: at least one $\alpha_k \neq 0$

FIT

- $\bar{y}_1, \bar{y}_2, \bar{y}_3$ and $\bar{y}_4$ are calculated in the exploratory data analysis. σ needs to be estimated by R.

Re-ASSESS error: After tranformation





- **Equal variance**: the residuals are evenly distributed around the $y = 0$ line. The spread is roughly the same across the groups.
  $s_{max}/s_{min} = 0.54/0.34 = 1.6 < 2$
- **Normal distribution**: points in the Q-Q plot lie very close to the $y = x$ line.
- Both the equal variance and Normal distribution assumptions are now **satisfied**.

Re-ASSESS model: After tranformation

```
summary(diamodel)
```

```
##              Df Sum Sq Mean Sq F value  Pr(>F)
## Color         3   8.30   2.767   12.63 8.4e-08 ***
## Residuals   303  66.36   0.219
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
summary(diamodel2)
```

```
##              Df Sum Sq Mean Sq F value   Pr(>F)
## Color         3   7.62  2.5392   12.74 7.28e-08 ***
## Residuals   303  60.38  0.1993
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

- The conclusion based on the model after transformation is more reliable because all model assumption are now satisfied.
- We reject H_0 that all group means are the same.

USE

- The mean of the natural log of the number of carats is significantly different across different colors (D, E, F, G) of the diamonds.
- There is statistically significant association between the color and number of carats of diamonds.

Note:

- This is an observational study but not an experiment. Therefore NO causal (but only association) relationship between the two variables can be inferred.
- If the diamonds were randomly chosen from the population, this conclusion of significant differences found in the ANOVA F-test can be extended to population data.

Summary

- ASSESS model
  - F distribution $F(df1, df2)$
  - ANOVA F test
  - One-way ANOVA table
- ASSESS error
  - Check assumptions $\epsilon \overset{iid}{\sim} N(0, \sigma)$
- Deal with violation in assumptions
  - Re-CHOOSE, re-FIT and re-ASSESS
  - *Natural logarithm transformation of right skewed data.*

R codes

```
# Slide 16 bar plot
ggplot(data=Diamonds, aes(Color))+
  geom_bar(aes(fill=Color),color="black")+ # bar plot
  scale_fill_brewer()+ # default: light blue to dark blue
  ggtitle("Barplot of Color")

# Slide 22 boxplot and histograms after tranformation
ggplot(Diamonds, aes(x=Color, y=log(Carat)))+
  geom_boxplot(aes(fill=Color))+
  scale_fill_brewer()+
  stat_summary(fun.y=mean, geom="point")+
  ggtitle("Log(Carat) by Color")

ggplot(Diamonds, aes(log(Carat)))+
  geom_histogram(binwidth=0.1, aes(fill=Color), color="black")+
  scale_fill_brewer()+
  facet_grid(Color ~ .)
```

R codes

```
# Slide 24 residuals vs. fitted values plot and Normal Q-Q plot after tranformation.
# Similar codes on Slide 13 and 14.
diamodel2 <- aov(log(Carat) ~ Color, data=Diamonds)
Assess <- data.frame(FittedValues=diamodel2$fitted.values,
                     Residuals=diamodel2$residuals)

ggplot(data=Assess, aes(x=FittedValues, y=Residuals))+
  geom_point(size=3, color="orange", alpha=0.6)+
  geom_hline(yintercept=0)+
  ggtitle("Equal Variance?")

ggplot(data=Assess, aes(sample = scale(Residuals)))+
  stat_qq(size=3, color="orange", alpha=0.6)+
  geom_abline(intercept=0, slope=1)+
  ggtitle("Normal Q-Q Plot")
```