STAT021 Statistical Methods II

Lecture 5 One-way ANOVA Model

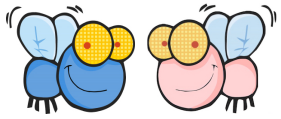
Lu Chen
Swarthmore College
9/18/2018

Review: statistical modeling

- Variability
- How to quantify variability
- Standard deviation (SD)
  - Sample standard deviation
  - Degree of freedom
- Variance
- Sampling variability of statistics
  - Definition
  - Standard error (SE)
  - Example
  - Sample size

Outline

One-way ANOVA

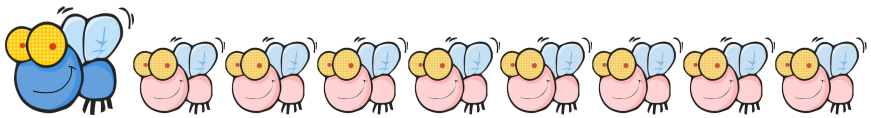


- Data example: Fruit flies
- CHOOSE
  - Exploratory data analysis
  - Null and alternative model
- FIT
  - Parameter estimation
- ASSESS model
  - Triple decomposition
    - Sum of squares and degrees of freedom
  - Mean square
  - F statistic

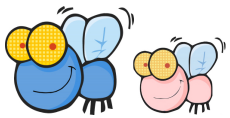
Example: Fruit flies

Increased reproduction leads to shorter life spans for female fruit flies. But the question remained **whether an increase in sexual activity would also reduce the life spans of male fruit flies**. The researchers designed an experiment to answer this question. They randomly assigned 75 male fruit flies to one of the following three groups:

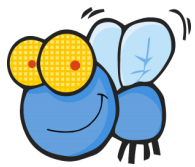
- **8 virgins**: Each male fruit fly was assigned to live with 8 virgin female fruit flies.



- **1 virgin**: Each male fruit fly was assigned to live with 1 virgin female fruit fly.



- **None**: Each male fruit fly lived alone.



Example: Fruit flies

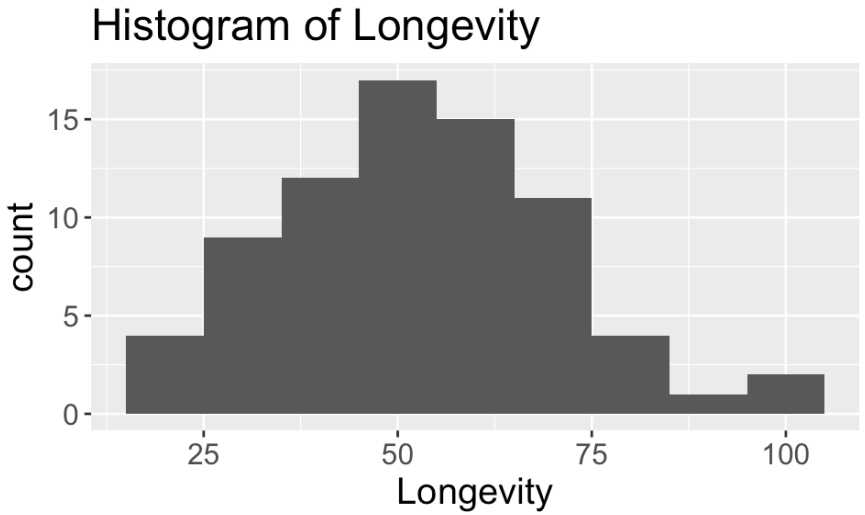
Questions of interest

- Is there statistically significant difference in life span (*Longevity*) of fruit flies among the three groups (*Treatment*)?
  - t tests are for two-group problems only
  - One-way ANOVA compares more than two population means
- If yes, which group(s) is(are) statistically significantly different?
  - Multiple pairwise comparisons

CHOOSE: exploratory data analysis

- **Response variable**: *Longevity*, quantitative; Mean: 53.0 days; SD: 17.9 days.
- **Explanatory variable**: *Treatment*, categorical; $n_1 = n_2 = n_3 = 25$

```
# install.packages("ggplot2") # install the package
library(ggplot2) # make the pacakge available
theme_update(text=element_text(size=15)) # set font size
ggplot(data=fly, aes(Longevity))+ # specify dataset and variable
  geom_histogram(binwidth=10)+ # plot histogram
  ggtitle("Histogram of Longevity") # main title
```



CHOOSE: exploratory data analysis

Relationship between *Longevity* and *Treatment*

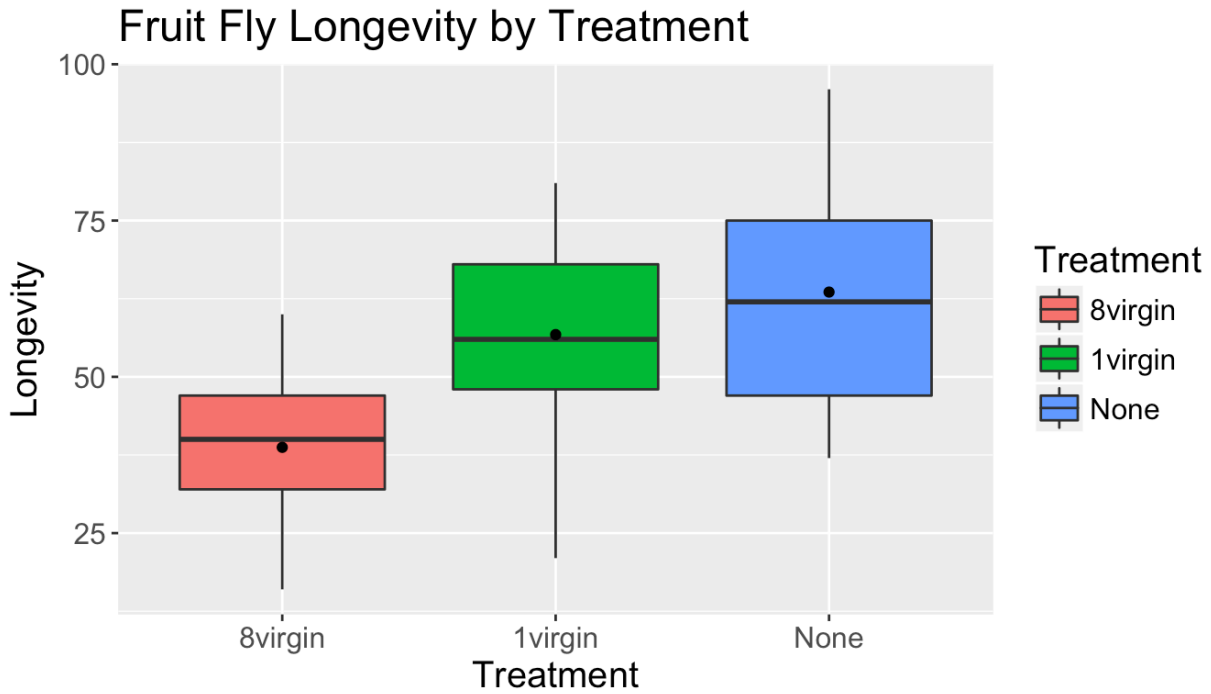
```
# define my own function to calcuate mean, sd and n at once
mysummary <- function(x){ c(mean=mean(x), sd=sd(x), n=length(x))}
# apply the function on Longevity by Treatment
aggregate(Longevity ~ Treatment, data=fly, FUN=mysummary)
```

```
##   Treatment Longevity.mean Longevity.sd Longevity.n
## 1   8virgin       38.72000     12.10207    25.00000
## 2   1virgin       56.76000     14.92838    25.00000
## 3      None       63.56000     16.45215    25.00000
```

Treatment	Size	Mean	SD
8virgin	$n_1 = 25$	$\bar{y}_1 = 38.7$	$s_1 = 12.1$
1virgin	$n_2 = 25$	$\bar{y}_2 = 56.8$	$s_2 = 14.9$
None	$n_3 = 25$	$\bar{y}_3 = 63.6$	$s_3 = 16.5$
Overall	$n = 75$	$\bar{y} = 53.0$	$s = 17.9$

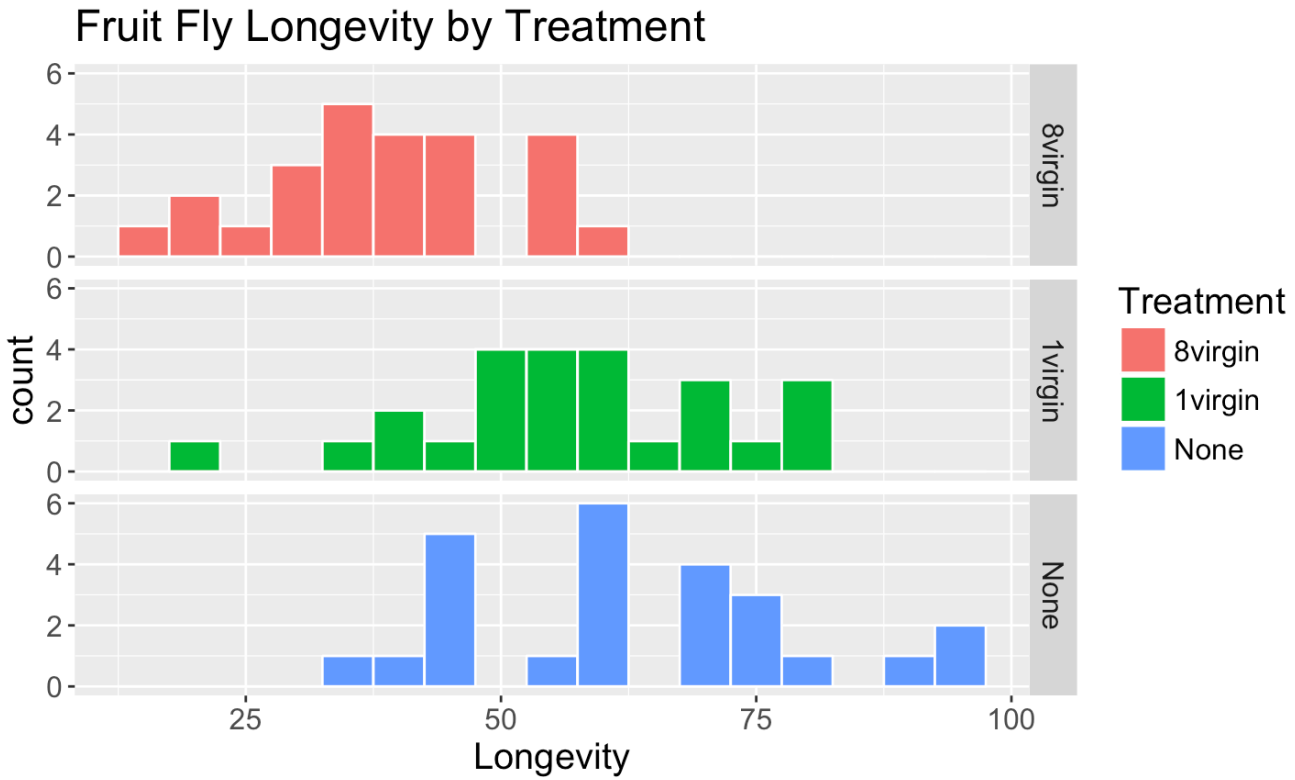
CHOOSE: exploratory data analysis

```
ggplot(fly, aes(x=Treatment, y=Longevity))+
  geom_boxplot(aes(fill=Treatment))+ # color by Treatment
  stat_summary(fun.y=mean, geom="point")+ # add the points of mean
  ggtitle("Fruit Fly Longevity by Treatment")
```



CHOOSE: exploratory data analysis

```
ggplot(fly, aes(Longevity))+
  geom_histogram(binwidth=5, aes(fill=Treatment), color="white")+
  facet_grid(Treatment ~ .)+ # plot histogram by Treatment
  ggtitle("Fruit Fly Longevity by Treatment")
```



- Based on exploratory data analysis, it seems that the three groups do have different means.
- Let's verify it using one-way ANOVA
- Denote *Longevity* as Y

CHOOSE: the ANOVA model

ANOVA model: each group has a unique population mean.

$$
\begin{array}{rcccll}
\text{Data} & = & \text{Model} & + & \text{Error} & \\
Y & = & \mu_1 & + & \epsilon & \text{for the 8virgin group} \\
Y & = & \mu_2 & + & \epsilon & \text{for the 1virgin group} \\
Y & = & \mu_3 & + & \epsilon & \text{for the None group}
\end{array}
$$

or in general

$$
\begin{array}{rcccll}
\text{Data} & = & \text{Model} & + & \text{Error} & \\
Y & = & \mu_k & + & \epsilon & \text{for } k = 1, 2, 3
\end{array}
$$

- μ_k is the **group mean** of the kth group, where $k = 1, 2, 3$.
- $\epsilon \sim N(0, \sigma)$.

CHOOSE: the ANOVA model

ANOVA model: each group has a unique population mean.

$$\begin{array}{lcccl} \text{Data} & = & \text{Model} & + & \text{Error} \\ Y & = & \mu_k & + & \epsilon \\ Y & = & \mu + \alpha_k & + & \epsilon \qquad \text{for } k = 1, 2, 3 \end{array}$$

- μ_k: **group mean** of the kth group. $\mu_k = \mu + \alpha_k$.
  - μ: **grand mean** of Y.
  - α_k: **group effect** of the kth group.

We are interested in whether

- $\mu_1 = \mu_2 = \mu_3 = \mu$, which is equivalent to
- $\alpha_1 = \alpha_2 = \alpha_3 = 0$.

CHOOSE: the ANOVA model

ANOVA model: each group has a unique population mean.

$$\begin{aligned} \text{Data} &= \text{Model} + \text{Error} \\ Y &= \mu_k + \epsilon \\ Y &= \mu + \alpha_k + \epsilon \qquad \text{for } k = 1, 2, 3 \text{ and } \epsilon \overset{iid}{\sim} N(0, \sigma) \end{aligned}$$

Model assumptions for the error term ϵ:

- Zero mean: mean of ϵ is 0.
- Equal variance: $\text{Var}(\epsilon) = \sigma^2$ is the same for all groups.
- Normal distribution: $\epsilon \sim N(0, \sigma)$.
- Independence: errors are independent of each other $\epsilon \overset{iid}{\sim} N(0, \sigma)$
  - $\overset{iid}{\sim}$ `\overset{iid}{\sim}`: independently and identically distributed.

CHOOSE: the ANOVA model

We will compare the **ANOVA model** (each group has a unique population mean)

$$\begin{aligned} \text{Data} &= \text{Model} + \text{Error} \\ Y &= \mu_k + \epsilon \\ Y &= \mu + \alpha_k + \epsilon \qquad \text{for } k = 1, 2, 3 \text{ and } \epsilon \overset{iid}{\sim} N(0, \sigma) \end{aligned}$$

to the **simpler/null model** (every group has the same population mean)

$$\begin{aligned} \text{Data} &= \text{Model} + \text{Error} \\ Y &= \mu + \epsilon \qquad \text{for } \epsilon \overset{iid}{\sim} N(0, \sigma) \end{aligned}$$

Null and alternative hypotheses of one-way ANOVA:

- $H_0 : \alpha_1 = \alpha_2 = \alpha_3 = 0.$
- H_a: at least one $\alpha \neq 0.$

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 $\alpha_k \neq 0$.

FIT: parameter estimation

Null model:

$$Y = \mu + \epsilon \quad \text{for } \epsilon \overset{iid}{\sim} N(0, \sigma)$$

- Parameters: μ and σ.

ANOVA model:

$$Y = \mu_k + \epsilon$$

$$Y = \mu + \alpha_k + \epsilon \quad \text{for } k = 1, 2, 3 \text{ and } \epsilon \overset{iid}{\sim} N(0, \sigma)$$

- Parameters: μ, α_k and σ.
  - Note: μ_k is also a parameter. Estimating μ and α_k is equivalent to estimating μ_k.

FIT: parameter estimation

$$\begin{array}{rccccl} & \text{Data} & = & \text{Model} & + & \text{Error} \\ \text{Population:} & Y & = & \mu_k & + & \epsilon \quad \text{for } \epsilon \overset{iid}{\sim} N(0, \sigma) \\ \text{Sample:} & y & = & \bar{y}_k & + & e \end{array}$$

- y: observed values of Y (*Longevity*)
- $\bar{y}_k$: sample mean of Y (*Longevity*) for the kth group
- **Residual** $e = y - \bar{y}_k$; standard deviation of e is the estimate of σ.

Null model:

$$\begin{array}{rcccl} Y & = & \mu & + & \epsilon \quad \text{for } \epsilon \overset{iid}{\sim} N(0, \sigma) \\ y & = & \bar{y} & + & e \end{array}$$

- Residual $e = y - \bar{y}$; standard deviation of e is the estimate of σ.

FIT: parameter estimation

Null model:

$$
\begin{aligned}
Y &= \mu + \epsilon && \text{for } \epsilon \overset{iid}{\sim} N(0, \sigma) \\
y &= \bar{y} + e \\
y &= \bar{y} + y - \bar{y} \\
61 &= 53.0 + 61 - 53.0 \\
53 &= 53.0 + 53 - 53.0 \\
33 &= 53.0 + 33 - 53.0 \\
&\ldots
\end{aligned}
$$

- $\bar{y} = 53.0$
- SD of e = SD of $y - \bar{y} = 17.9$
- Here we consider $\bar{y}$ as the **predicted value of y**, which is usually denoted as $\hat{y}$. So in the null model, $\hat{y} = \bar{y}$.

FIT: parameter estimation

ANOVA model:

$$
\begin{aligned}
Y &= \mu_k + \epsilon \qquad \text{for } k = 1, 2, \cdots, K \text{ and } \epsilon \overset{iid}{\sim} N(0, \sigma) \\
Y &= \mu + \alpha_k + \epsilon \\
y &= \bar{y}_k + e \\
y &= \bar{y} + \bar{y}_k - \bar{y} + y - \bar{y}_k
\end{aligned}
$$

- The estimate of μ is sample mean $\bar{y}$
- The estimate of α_k is sample group effect $\bar{y}_k - \bar{y}$
- $e = y - \bar{y}_k$
- In the ANOVA model, $\hat{y} = \bar{y}_k$.

FIT: parameter estimation

ANOVA model:

$$
\begin{aligned}
Y &= \mu_k + \epsilon \qquad \text{for } k = 1, 2, \cdots, K \text{ and } \epsilon \overset{iid}{\sim} N(0, \sigma) \\
Y &= \mu + \alpha_k + \epsilon \\
y &= \bar{y}_k + e \\
y &= \bar{y} + \bar{y}_k - \bar{y} + y - \bar{y}_k
\end{aligned}
$$

- Group mean: $\bar{y}_1 = 63.6, \bar{y}_2 = 56.8, \bar{y}_3 = 38.7$
- Grand mean: $\bar{y} = 53.0$
- Group effect: $\bar{y}_1 - \bar{y} = 10.6, \bar{y}_2 - \bar{y} = 13.8, \bar{y}_3 - \bar{y} = -14.3$
- SD of e = SD of $y - \bar{y}_k = 14.6$

FIT: parameter estimation

		Null model $Y = \mu + \epsilon$		ANOVA model $Y = \mu + \alpha_k + \epsilon$	
Treatment	**Observed** y	**Predicted** $\hat{y} = \bar{y}$	**Residual** $e = y - \bar{y}$	**Predicted** $\hat{y} = \bar{y}_k$	**Residual** $e = y - \bar{y}_k$
8virgin	33	53.0	-20.0	38.7	-5.7
8virgin	26	53.0	-27.0	38.7	-12.7
1virgin	53	53.0	0.0	56.8	-3.8
1virgin	60	53.0	7.0	56.8	3.2
None	61	53.0	8.0	63.6	-2.6
None	71	53.0	18.0	63.6	7.4
...	...	...	...	...	...

- In general, the ANOVA model has smaller residuals (in absolute value) than the null model.

ASSESS model: Triple decomposition

Data	=	Grand Mean	+	Group Effect	+	Error
Y	=	μ	+	α_k	+	ϵ
y	=	$\bar{y}$	+	$\bar{y}_k - \bar{y}$	+	$y - \bar{y}_k$
33	=	53.0	+	−14.3	+	−5.7
26	=	53.0	+	−14.3	+	−12.7
53	=	53.0	+	3.8	+	−3.8
60	=	53.0	+	3.8	+	3.2
61	=	53.0	+	10.6	+	−2.6
71	=	53.0	+	10.6	+	7.4
...		↑		↑		↑
		No variability		Variability **between** groups		Variability **within** groups

ASSESS model: Triple decomposition

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

$$\text{Total sum of squares} = \text{Group sum of squares} + \text{Error sum of squares}$$

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

$$SSTotal = SSGroup + SSE$$

$$SST = SSG + SSE$$

↑ Total variability in data

Variability of the null model residuals

↑ Variability from the group effects

Variability explained by the ANOVA model

↑ Variability in the residuals

Variability left in the ANOVA model residuals

ASSESS model: Triple decomposition

- Thus, the ANOVA method analyzes the variability in the data and measures how it can be explained by the explanatory variable.
- Specifically, the ANOVA method compares the average variability explained by the model to the average variability left in the residuals.
  - If the ratio is large, the ANOVA model explained a lot variability that was left in the null model residuals. ANOVA model is better than the null model.
  - If the ratio is small, the ANOVA model still cannot explain the variability that was left in the null model residuals. ANOVA model is no better than the null model.
- How to quantify the *average* variability?
- It involves the concept of **degree of freedom**: the number of values in the final calculation of a statistic that are free to vary.

ASSESS model: Triple decomposition

Data

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

Sum of squares

$$\begin{aligned} SSTotal &= SSGroup + SSE \\ \sum(y - \bar{y})^2 &= \sum(\bar{y}_k - \bar{y})^2 + \sum(y - \bar{y}_k)^2 \end{aligned}$$

Degrees of freedom

$$\begin{aligned} df_{Total} &= df_{Group} + df_{Error} \\ \#\text{ of } y - 1 &= \#\text{ of groups} - 1 + \#\text{ of } y - \#\text{ of groups} \\ n - 1 &= K - 1 + n - K \\ 75 - 1 &= 3 - 1 + 75 - 3 \end{aligned}$$

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 $\boldsymbol{F}$ **statistic**

$$F = \frac{MSGroups}{MSE}$$

- F is large: the ANOVA model is better than the null model. The groups have significantly different means.
- F is small: the ANOVA model is NOT better than the null model. The group means are not that different.
- For our fruit fly example, $F = 19.31$. Do you consider it to be large enough for us to say the ANOVA model is better than the null?

Summary

- Data example: Fruit flies
- CHOOSE
  - Exploratory data analysis
  - Null and alternative model $Y = \mu + \epsilon$ vs. $Y = \mu + \alpha_k + \epsilon$
- FIT
  - Parameter estimation
- ASSESS model
  - Triple decomposition
    - Sum of squares and degrees of freedom
  - Mean square
  - F statistic