

Chapter 15 - One-Way Analysis of Variance

COMPARING MANY POPULATION MEANS

- *Analysis of Variance* is often abbreviated as ANOVA.
- A one-way ANOVA considers $k > 2$ populations. The mean of the i^{th} population is μ_i . The variance of the i^{th} population is σ_i^2 .
- An ANOVA is used to compare the k population means, $\mu_1, \mu_2, \dots, \mu_k$.
- “One-way” means that the levels of a single factor define the populations being compared. In other words, the categories of a categorical variable define the populations.

Setting:

1. A SRS of size n_i has been chosen from population i for $i = 1, 2, \dots, k$.
2. Each SRS is independent of the others.
3. *Homogeneity of Variance*: $\sigma^2 = \sigma_1^2 = \sigma_2^2 = \dots = \sigma_k^2$.
4. Each population is normally distributed, so the distribution of the i^{th} population is $N(\mu_i, \sigma)$.

Note about Study Design:

- Completely Randomized Design Experiment
 - The k treatment groups from a CRD are considered to be independent samples from k populations.
 - In a CRD, if there is a difference among the k means, it is appropriate to claim that the factor (treatment) caused the difference in means.
 - If the groups of individuals in the CRD were chosen from a SRS, then conclusions about k means can be extended to the populations from which the individuals were drawn. However, if the individuals were not from a SRS, then conclusions to larger populations are dubious.
- Observational Study
 - The k samples to be compared in an observational study are considered independent if individuals in each sample were randomly chosen from each respective population.
 - Do not claim that the factor (explanatory variable) caused the difference in means.

The one-way ANOVA Model

$$X_{ij} = \mu_i + \epsilon_{ij} \quad \text{where } \epsilon_{ij} \sim N(0, \sigma)$$

- X_{ij} is the response of the j^{th} individual in the SRS from the i^{th} population.
- μ_i is mean of i^{th} population.
- ϵ_{ij} is the error term, $x_{ij} - \mu_i$, also called the *deviation* of x_{ij} from μ_i
- $\epsilon_{ij} \sim N(0, \sigma)$ is equivalent to the assumption that the i^{th} population is normal, $x_{ij} \sim N(\mu_i, \sigma)$.

The “Estimated” one-way ANOVA Model

$$X_{ij} = \bar{X}_i + e_{ij}$$

- \bar{X}_i is the sample mean of the SRS from the i^{th} population, an unbiased point estimator of μ_i .
- $e_{ij} = X_{ij} - \bar{X}_i$ is the residual, the deviation of X_{ij} from \bar{X}_i

HYPOTHESIS TEST TO COMPARE k POPULATION MEANS

The Overall Test:

The Idea: We will compare the variability **between** the sample means ($MSTr$) to the variability **within** each sample (MSE).

- If the variability between the \bar{x}_i 's is **large** relative to the variability within each sample ($MSTr \gg MSE$), then we will claim that there is a difference between the μ_i 's.
- If the variability among the \bar{x}_i 's is **not large** relative to the variability within each sample, then we will fail to claim that there is a difference between the μ_i 's.
- The statistic MSE is an unbiased estimator of the constant variance σ^2 .

1. Hypotheses:

$$H_0: \mu_1 = \mu_2 = \dots = \mu_k$$

$$H_a: \mu_i \neq \mu_j \text{ for some } i \text{ and } j \quad (\text{at least one of the } \mu_i \text{ are different than the others})$$

2. Check Assumptions:

- (a) Independent SRS's have been chosen, and so $.05N_i \geq n_i$ for each $i = 1, 2, \dots, k$.
- (b) Check normal probability plots to determine if the residuals are not normally distributed.
- (c) Assume that the constant variance assumption holds if $\frac{\text{largest } s}{\text{smallest } s} < 2$.

Perform Steps 3 and 4 assuming that H_0 is true!

3. Test Statistic:

$F = \frac{MSTr}{MSE}$, where $MSTr$ and MSE are from the following ANOVA table:

One-way ANOVA Table

Source	DF	Sum of Squares (SS)	Mean Squares (MS)	F	p-value
Treatments	$DF_{Tr}=k-1$	$SSTr = \sum_{i=1}^k n_i (\bar{x}_i - \bar{\bar{x}})^2$	$MSTr = \frac{SSTr}{DF_{Tr}}$	$F^* = \frac{MSTr}{MSE}$	$P(F > F^*)$
Error	$DFE=N-k$	$SSE = \sum_{i=1}^k (n_i - 1) s_i^2$	$MSE = \frac{SSE}{DFE}$		
Total	$DFTo=N-1$	$SSTo = \sum_{\text{all } x} (x_{ij} - \bar{\bar{x}})^2$			

where $N = \sum n_i$ is the *total sample size* and $\bar{\bar{x}} = \frac{\sum n_i \bar{x}_i}{\sum n_i}$ is the *grand mean*. Observe that $DFT = DFTr + DFE$ and $SST = SSTr + SSE$.

- When $H_0 : \mu_1 = \mu_2 = \dots = \mu_k$ is true, $\mu_{MSTr} = \mu_{MSE}$ and therefore $F^* = \frac{MSTr}{MSE} \approx 1$.
- When H_0 is false, $\mu_{MSTr} > \mu_{MSE}$ and therefore $F_{Tr} = \frac{MSTr}{MSE} \gg 1$.
- Large $F^* = \frac{MSTr}{MSE}$ values are strong evidence against H_0 and for H_a .

4. p-value: The p -value = $P(F > F^*)$.

- The test statistic has an F distribution, $F \sim F(DF_{Tr}, DFE)$, when H_0 is true.
- An F distribution $F(DF_{Tr}, DFE)$ has two parameters, the *numerator degrees of freedom* DF_{Tr} and the *denominator degrees of freedom* DFE .
- An F distribution is a beautiful right-skewed distribution. Probabilities can be found in Table 7 on pages 738-741 of your textbook, or by using R's `pf(x, df1=#, df2=#, lower.tail=FALSE)` function.

5. and 6. Make a Decision and give a Conclusion.

MULTIPLE COMPARISONS FOLLOW-UP TEST

If we reject H_0 in the overall test and conclude that at least one of the μ_i 's is different than the others, then we should ask “Which population means are different?”

Only do a follow-up test if you REJECT $H_0 : \mu_1 = \mu_2 = \dots = \mu_k!$

Tukey's Method

Tukey's Method calculates a family of CI's for all possible pairwise differences of the means $\mu_1, \mu_2, \dots, \mu_k$. The overall family-wise confidence level is held at some confidence level $C = 1 - \alpha$ (which means that the confidence level for each individual CI is more than C). For a given pair μ_i and μ_j (with $i \neq j$):

- The Tukey's confidence interval for $\mu_i - \mu_j$ is:

$$\bar{x}_i - \bar{x}_j \pm \frac{q_{1-\alpha, k, DFE}}{\sqrt{2}} \sqrt{MSE\left(\frac{1}{n_i} + \frac{1}{n_j}\right)}$$

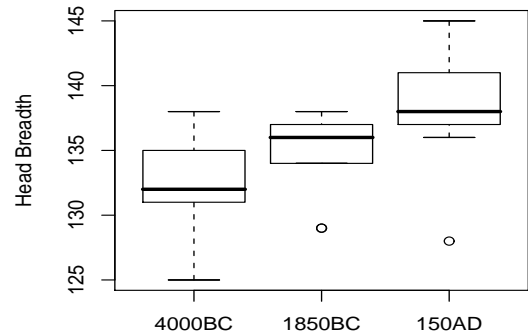
- The Tukey critical value $q_{1-\alpha, k, DFE}$ can be found:
 - In Table 8 on page 742 of your textbook. If you can not find the appropriate DFE in Table 8, ROUND DOWN to be conservative.
 - Using R's `qtukey(C, nmeans=k, df=DFE)`
- The Tukey's CI for $\mu_i - \mu_j$ can be used to test the hypothesis $H_0: \mu_i - \mu_j = 0$ versus $H_a: \mu_i - \mu_j \neq 0$.
 - If 0 is in the CI, then fail to reject H_0 .
 - If 0 is **not** in the CI, then reject H_0 .

EXAMPLE:

An archeologist is interested in studying skull breadths of humans from different epochs. Significant changes in head shape over time would suggest that interbreeding occurred with immigrant human populations. A sample of 27 head breadths were obtained by measuring skulls of Egyptian males from three different epochs: 4000BC, 1850BC, and 150AD. The data are from *Ancient Races of the Thebaid*, by Thomson and Randall-Maciver).

Display Your Data

```
> D = read.table("headbreadth.txt",header=TRUE)
> attach(D)
> Epoch = factor(as.character(Epoch),
  levels = c("4000BC","1850BC","150AD"))
> boxplot(HeadBreadth ~ Epoch,ylab="Head Breadth")
```



Fit the One-way ANOVA Model

```
> library(MASS)
> library(pastecs)
> tapply(HeadBreadth,Epoch,mean)
 4000BC  1850BC  150AD
132.6667 134.4444 138.1111

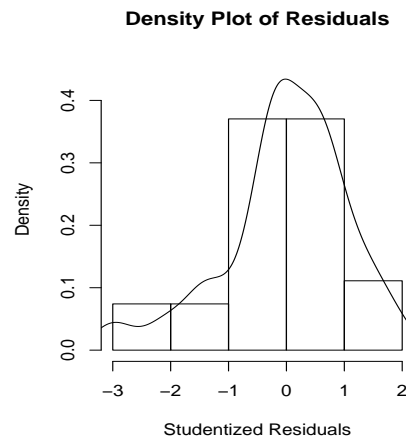
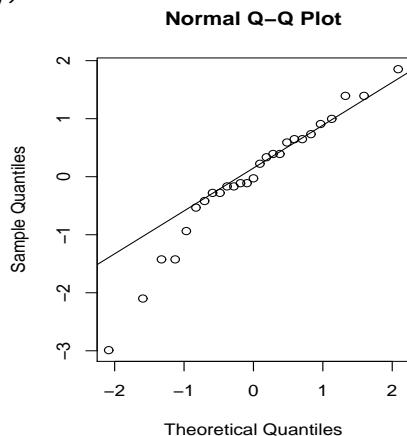
> hb.aov=aov(HeadBreadth ~ Epoch)
> summary(hb.aov) # Prints ANOVA table
```

Source	DF	SS	MS	F	p-value
Treatment	2	138.74	69.37	4.0497	0.03052
Error	24	411.11	17.13		
Total	26	549.85			

Check Your Assumptions

1. Independent Random Samples
2. Normal Distribution

```
> par(mfrow=c(1,2))
> qqnorm(studres(hb.aov))
> qqline(studres(hb.aov))
> hist(studres(hb.aov),freq=FALSE,ylim=c(0,0.45),main="Density Plot of Residuals",
  xlab="Studentized Residuals")
> lines(density(studres(hb.aov)))
> xy=qqnorm(studres(hb.aov))
> cor(xy$x,xy$y)
[1] 0.9722762
```



3. Check Constant Variance

```
> tapply(HeadBreadth, Epoch, sd)
 4000BC  1850BC  150AD
4.183300 3.358240 4.755114
```

$\frac{\text{largest } s}{\text{smallest } s} \approx \frac{4.76}{3.35} \approx 1.42 < 2$, so the constant variance assumption appears to hold.

1. Perform the **overall Hypothesis Test**

(a) Hypotheses:

(b) Test statistic value:

(c) Distribution of the test statistic:

(d) p-value:

(e) Decision at $\alpha = .05$:

(f) Conclusion:

2. Give an unbiased estimate of the constant variance σ^2 .

3. Perform **Tukey's Multiple Comparison Test**

```
> TukeyHSD(hb.aov, which="Epoch", conf.level=0.95)
```

Comparison	Estimate	Lower	Upper
$\mu_{1850BC} - \mu_{4000BC}$	1.7778	-3.0945	6.6501
$\mu_{150AD} - \mu_{4000BC}$	5.4444	0.5721	10.3168
$\mu_{150AD} - \mu_{1850BC}$	3.6667	-1.2057	8.5390

(a) **Conclusions**:

(b) Which epoch appears to have the largest mean head breadth? How much larger is the head breadth during this epoch?

Exercises

15.1 on p676: 1-15 odd
15.2 on p684: 19-23 odd

Reading

Sections 15.1-15.2. Unfortunately, we do not have time to cover ANOVA for RBD (15.3) and Two-way ANOVA (15.4), but you may want to read about these topics if they are applicable to your research.