Comparing means of 3 or more groups (ANOVA)

Single Factor ANOVA

Example: The theory we want to test is one studied by 7 years ago by a researcher visiting the department. The theory is: "*For STAT 511 students, there is an association between smoking and GPR*." To test this theory, the researcher used data from an anonymous survey taken of STAT 302 students in Fall 02. Assume that this data set is equivalent to a random sample. Each student in the survey was asked which category they fell into: *never smoked, former smoker, currently smokes.* Each student was also asked their GPR.

- Two variables are of interests Explanatory: Smoking status - categorical
  Response: GPR – numerical
  We are interested in the relationship between the explanatory and response variable.

- How many groups are we comparing?

Can we just use Independent Samples t test 3 times to test the following alternative hypotheses?

1. H_A: $\mu_{\text{Ex-Smoker}} \neq \mu_{\text{Never Smoked}}$ 2. H_A: $\mu_{\text{Smoker}} \neq \mu_{\text{Never Smoked}}$ and 3. H_A: $\mu_{\text{Ex-Smoker}} \neq \mu_{\text{Smoker}}$

ANSWER: Because if H_0 is true for each of the 3 different hypotheses given above (that is if there is no association between smoking and GPR) then we'd have 3 chances to make a type I error.

ANOVA *Test used when the response is numerical and the explanatory variable is categorical taking 3 or more values. This explanatory variable divides the population into groups.*

Notations and Assumptions:

There are T different groups corresponding to T different,
$X_{1,1},\dots,X_{1,n_1} \sim N(\mu_{1,}\sigma^2)$ are the n_1 samples of the first group;
$X_{2,1},\dots,X_{2,n_2} \sim N(\mu_{2,}\sigma^2)$ are the n_2 samples of the second group;
….
$X_{T,1},\dots,X_{T,n_T} \sim N(\mu_T,\sigma^2)$ are the n_T samples of the T-th group.
(The above statement contains two critical assumption: *NORMAL* and *EQUAL VARIANCE*)

Denote:

$$\bar{X}_{\cdot\cdot} = \frac{\sum_{i=1}^{T}\sum_{j=1}^{n_i} X_{i,j}}{\sum n_i} = \text{Grand Mean}$$

$$\bar{X}_{i\cdot} = \frac{\sum_{j=1}^{n_i} X_{i,j}}{n_i} = \text{Group Sample Mean}$$

Hypothesis Test:
H_0: there is no relationship between smoke status and GPR; or
H_0: group index has nothing to do with the mean of GPR; or

H_0: $\mu_1 = \ldots = \mu_T = \text{constant}$.

H_1: smoke status does influence the group mean of GPR; or
H_1: There are at least two groups that their group mean are different; or
H_1: $\mu_1 = \ldots = \mu_T$ doesn't hold

Note: Graphical approach, such as side-by-side boxplot can visual display the comparison. But usually, side-by-side boxplot can only detect practical significance, but not statistical significance.

Intuition of ANOVA:Sum of Square (variability) decomposition

$$\sum_{i=1}^{T} \sum_{j=1}^{n_i} (X_{i,j} - \bar{X}_{..})^2 = \sum_{i=1}^{T} \sum_{j=1}^{n_i} (X_{i,j} - \bar{X}_{i\cdot})^2 + \sum_{i=1}^{T} \sum_{j=1}^{n_i} (\bar{X}_{i\cdot} - \bar{X}_{..})^2$$
$$= \sum_{i=1}^{T} \sum_{j=1}^{n_i} (X_{i,j} - \bar{X}_{i\cdot})^2 + \sum_{i=1}^{T} n_i (\bar{X}_{i\cdot} - \bar{X}_{..})^2$$

Total sum of square (SST) = treatment sum of square (SSTr) + error sum of squre (SSE)
Total variation in data = variation within each group + variation among groups

The significance of group difference depends on the *comparison of between-group-variation vs within-group-variation.*

Test statistic: $F = \dfrac{\text{SSTrt}/(T-1)}{\text{SSE}/(N-T)} = \dfrac{\text{MSTrt}}{\text{MSE}}$, where $N = \sum n_i$.

In order to determine the rejection region, we need to under the sampling distribution of F-statistic.

Note: under equal sample size situation, MSE = average of sample group variance

Proposition

- Under both Null and Alternative hypothesis, $\text{SSE}/\sigma^2 \sim \chi^2_{N-T}, E(\text{MSE}) = \sigma^2$
- Under Null hypothesis, $SSE/\sigma^2 \sim \chi^2_{N-T}, E(MSE) = \sigma^2$
- Under Alternative hypothesis, $E(MSE) > \sigma^2$
- Under Null hypothesis, F follows a F distribution with df T-1 and N-T.

The above proposition implies 1) MSTrt is an unbiased estimator for variance; 2) F is approximately 1 under Null, but tends to be larger when alternative is true.

Rejection region: $F > F_{\alpha, T-1, N-T}$

ANOVA table:

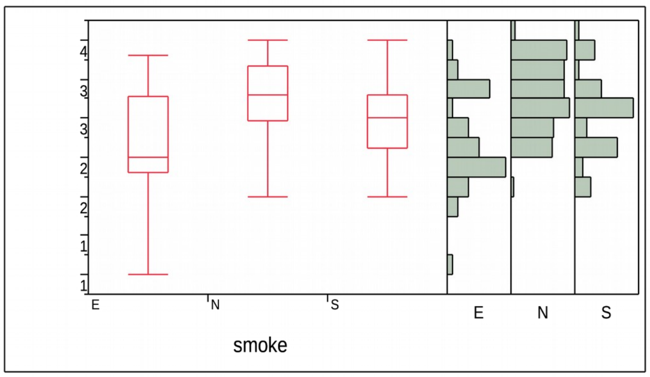
Source	DF	Sum of Squares	Mean Square	F Ratio	Prob > F
Treatment	T-1	SSTrt	MSTr=SSTr/(T-1)	F=MSTr/MSE	<.0001
Error	N-T	SSE	MSE=SSE/(N-T)		
Total	N-1	SST			

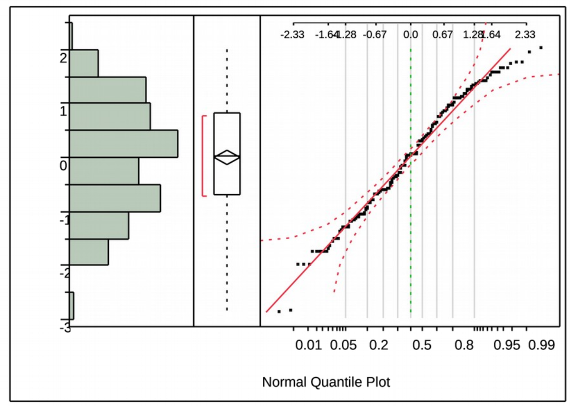
Smoke status example

Descriptive Statistics is given:

One Variable Summary	**Ex-smoker**	**Never**	**Smoker**
Mean	2.64	3.27	2.99
Std. Dev.	0.62	0.44	0.50
Median	2.50	3.30	3.00
Minimum	1.00	2.00	2.00
Maximum	3.80	4.00	4.00
Count	29	145	49
1st Quartile	2.30	2.97	2.65
3rd Quartile	3.25	3.66	3.30

One way Analysis of GPR By smoking habit





Above right is the normal quantile plot of residuals. This is used to determine if the normality condition is satisfied. The box plot to the right is the one we check for outliers.

What to check before using an ANOVA to test the hypotheses

1. We have a normal distribution requirement for the residuals (difference between the data values and the group mean). In other words, the QQ plot for all data value $X_{i,j} - \bar{X}_{i\cdot}$)

2. We require that all groups have close to the same standard deviation. Usually, we want the largest sample SD can be no greater than twice the size of the smallest sample SD.

3.**ANOVA table**

Analysis of Variance

Source	**DF**	**Sum of Squares**	**Mean Square**	**F Ratio**	**Prob > F**
smoke	2	12.31	6.156	26.566	<.0001
Error	231	53.53	0.232		
C. Total	233	65.84			

4. Conclusion:

The data demonstrates statistical evidence that for there is association between smoking and GPR.

```
R code
y<-c(rnorm(10),rnorm(12, mean=1),rnorm(9))
x<-factor(c(rep(1,10),rep(2,12),rep(3,9)))

boxplot(y~x)
qqnorm(y-c (rep(mean(y[1:10]),10),rep(mean(y[11:22]),12),rep(mean(y[23:31]),9)))
qqline(y-c (rep(mean(y[1:10]),10),rep(mean(y[11:22]),12),rep(mean(y[23:31]),9)))

a1<-aov(y~x)
summary(a1)

TukeyHSD(a1, conf.level=0.95)
```

Multiple Comparison of treatments

If the ANOVA F-test is not significant, then we claims there is no difference between groups. This ends the analysis.
If the test is significant, we claim there is difference among T groups. Then a natural follow-up question: where is the difference? More specifically, among **all T(T-1)/2 possible pairwise comparison**, which are significantly different, which are not?

Remember, we can not perform several 2-sample t- or z-tests, since it fails to control the probability of type I error.

Simultaneous confidence interval

Given d parameters $\theta_1,\ldots,\theta_d$, a simultaneous confidence set is a random set $C(X)\in R^d$ which based on random sample X, such that

$$P((\theta_1,\ldots,\theta_d)\in C(X))=1-\alpha$$

If C(X) is a hyper-rectangle such as $C(X)=[L_1(X),U_1(X)]\times\ldots\times[L_d(X),U_d(X)]$, then this leads to the simultaneous confidence interval:

$$P(\theta_1\in[L_1(X),U_1(X)],\text{ and } \ldots \text{ and } \theta_d\in[L_d(X),U_d(X)])=1-\alpha .$$

Note: Regular Marginal confidence interval concerns probability of event $\{\theta_i\in[L_i(X),U_i(X)]\}$, while simultaneous interval concerns the probability of the union of these events.

Tukey pairwise interval provides simultaneous confidence interval for ***all* pairwise group mean difference**,

$$P\left(\mu_i-\mu_j\in\bar{X}_{i\cdot}-\bar{X}_{j\cdot}\pm\frac{Q_{\alpha,T,N-T}}{\sqrt{2}}\sqrt{MSE\left(\frac{1}{n_i}+\frac{1}{n_j}\right)}\text{ for all } 1\leqslant i<j\leqslant T\right)=1-\alpha,$$

where $Q_{\alpha,T,N-T}$ is the critical value obtained for studentized range distribution. This Q value can be obtained from Tukey table, or R function qtukey(1-alpha, T, N-T).

Identifying significant pairs of difference:

Groups i and j have significant different group population mean

$$\leftrightarrow \quad 0 \notin \bar{X}_{i\cdot} - \bar{X}_{j\cdot} \pm \frac{Q_{\alpha,T,N-T}}{\sqrt{2}} \sqrt{MSE\left(\frac{1}{n_i} + \frac{1}{n_j}\right)}$$

$$\leftrightarrow \quad |\bar{X}_{i\cdot} - \bar{X}_{j\cdot}| > \frac{Q_{\alpha,T,N-T}}{\sqrt{2}} \sqrt{MSE\left(\frac{1}{n_i} + \frac{1}{n_j}\right)}$$

The quantity $\frac{Q_{\alpha,T,N-T}}{\sqrt{2}} \sqrt{MSE\left(\frac{1}{n_i} + \frac{1}{n_j}\right)}$ is called the critical difference.

In summary: we only need to compare group sample mean difference against critical difference. If the former is larger, then this pairwise comparison is significant, otherwise not.

Example:

A possible contradictory comparison result of 3-group ANOVA:

No difference between mean of group 1 and group 2. (Fail to reject the null that group 1 = group 2)
No difference between mean of group 2 and group 3. (Fail to reject the null that group 2 = group 3)
Significant difference between mean of group 2 and group 3. (Reject the null that group 1 = group 3)

If all n_i are the same, the critical difference is a constant across all (i,j). This critical difference is

$$Q_{\alpha,T,N-T} \sqrt{\frac{MSE}{n}} \quad .$$

Presentation of the Tukey comparison result

Marginal confidence interval for linear function of group mean $\sum c_i \mu_i$

The point estimator of $\sum c_i \mu_i$ is $\sum c_i \bar{X}_{i\cdot}$, whose sampling distribution is $N\left(\sum c_i \mu_i, \sigma^2 \sum \frac{c_i^2}{n_i}\right)$

Thus we can derive that $\frac{\sum c_i \bar{X}_{i\cdot} - \sum c_i \mu_i}{\sqrt{\sigma^2 \sum \frac{c_i^2}{n_i}}} \sim N(0,1)$ and $\frac{\sum c_i \bar{X}_{i\cdot} - \sum c_i \mu_i}{\sqrt{\text{MSE} \sum \frac{c_i^2}{n_i}}} \sim t_{N-T}$. Note that the pooled estimator of variance is used, this estimator combines the whole data information and provides higher df.

This lead to confidence interval

$$\sum c_i \bar{X} \pm t_{\alpha/2, N-T} \sqrt{\text{MSE} \sum \frac{c_i^2}{n_i}}$$

More on Single Factor ANOVA

Power and sample size

Type II error is $\beta = P(F > F_{\alpha, T-1, N-T} | \text{Alternative hypothesis is true})$
To figure out the above quantity, one need to know what is the sampling distribution of F under alternative hypothesis.
It turns out that F has a so-called noncentral F distribution, which is positively related to $\sum_i n_i (\mu_i - \bar{\mu})^2 / \sigma^2$ where $\bar{\mu} = \sum_i \mu_i / T$.

It is impossible to conduct hand calculation for power and sample size.

R code:

```
mu<-c(0, 0,  -1, 1)
#sample size calculation#
power.anova.test(group = 4, between.var=var(mu), within var = 1, sig.level=0.05, power = 0.8)
#power  calculation#
power.anova.test(group = 4, n=8, between.var=var(mu), within var = 1, sig.level=0.05)
```

In the R function argument, group is the number of group, n is the number of observation in each group, between.var = $\sum_i (\mu_i - \bar{\mu})^2 / (T-1)$, and within.var = σ^2

Relationship between F- and t-test

If T=2, both F-test and 2-sided pooled t-test can be used to test the difference of 2 means of 2 population that shares the same normal variance. *Are they equivalent?*

*Answer:*Yes, exactly. One can prove that F statistic = (t-statistics)^2, and $t^2_{\alpha/2, df} = F_{\alpha, 1, df}$.

Data transformation

Data tranformation is used, when the equal variance or normality assumption is not met (e.g. your boxplots shows big variance difference across groups.)

A particular popular transformation is the **Boxcox transformation**, that is

$$f(X_{i,j},\lambda)=X_{i,j}^{\lambda}, \text{ and } f(X_{i,j},0)=\log(X_{i,j})\, if\, \lambda=0.$$

We find the best lambda, such that after the transformation, the data look like normal and equal variance.

```
R code
y<-c(rnorm(10, mean=3, sd=1),rnorm(12, mean=6,sd=2),rnorm(9, mean=9, sd=3));
x<-factor(c(rep(1,10),rep(2,12),rep(3,9)))

boxplot(y~x)

A<-boxcox(y~x)
lambda<-A$x[which.max(A$y)]

z<-y^lambda
if(abs(lambda)<0.1 ) z<-log(y)

boxplot(z~x)
```