DA503 Applied Statistics

Lecture 09

Analysis of Variance (ANOVA)

Fall 2021 – H . Sait Ölmez

Variability in variance

- **F-distribution** is a continuous probability distribution that shows up frequently as the null distribution of a test-statistic, as in the analysis of variance (F-test)
- **F-test: Analysis of variance for two samples**
- We have methods for testing whether a difference in the means of samples is significant.
- What about testing for a difference in variance?
- **Problem**: The sodium contents (mg/125ml) of tomato sauces from two companies (1 and 2) are given below:

	Sodium content (mg/125dl)										Mean	Var
1	860	850	750	870	940	410	410	820	890	890	769	38254.4
2	540	640	600	640	300	610	430	280	300	610	495	23116.7

- We want to test the equality of two variances: $\sigma_1^2 = \sigma_2^2$?

Variability in variance – cont'd

- Here are the results from 2 samples:

  $n_1 = 10$ and $s_1^2 = 38254.4$

  $n_2 = 10$ and $s_2^2 = 23116.7$

- Variance from Company1 is higher than Company2. But remember, these are samples. So, there is going to be some sampling error.
- Is the difference in variability between (1) and (2) statistically significant, or is it due to random sampling error?
- For this analysis, we need a different technique. We're not comparing the sample variance to a hypothesized variance; we're comparing 2 sample variances with each other.

Variability in variance – cont'd

- The Null and Alternative hypotheses:

  H_0: $\sigma_1^2 = \sigma_2^2$

  H_A: $\sigma_1^2 > \sigma_2^2$

- The easiest way to compare the relative size of two measurements is by using a ratio:

$$F - ratio = F = \frac{S_1^2}{S_2^2}$$

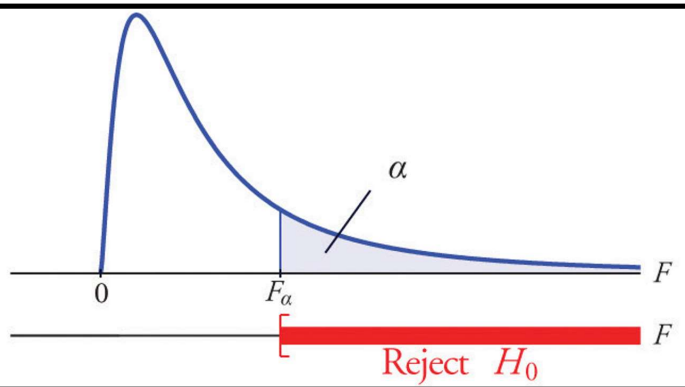
Larger variance
(n_1-1 degrees of freedom)

(a right-tailed test)

Smaller variance
(n_2-1 degrees of freedom)

- When independent random samples are taken from two normal populations with equal variances, the sampling distribution of the ratio of sample variances follows the F-distribution which is a distribution of ratios.

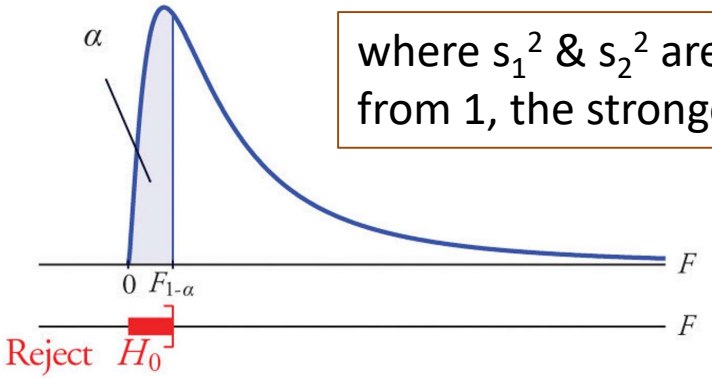
Variability in variance – cont'd



The test we'll end up using the most:

H_0: $\sigma_1^2 = \sigma_2^2$

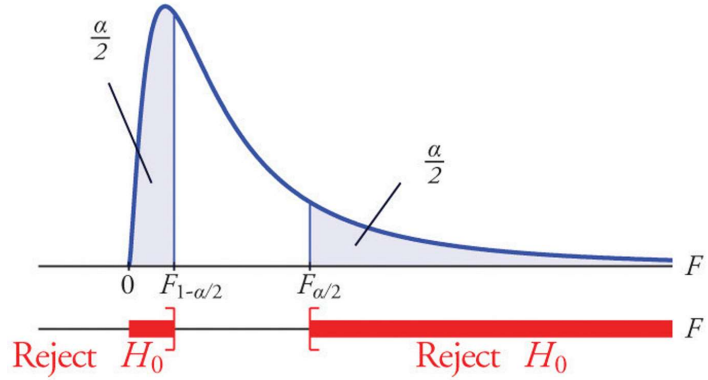
H_A: $\sigma_1^2 > \sigma_2^2$ for an upper one-tailed test



where s_1^2 & s_2^2 are the sample variances. The more this ratio deviates from 1, the stronger the evidence for unequal population variances.

H_0: $\sigma_1^2 = \sigma_2^2$

H_A: $\sigma_1^2 < \sigma_2^2$ for a lower one-tailed test



H_0: $\sigma_1^2 = \sigma_2^2$

H_A: $\sigma_1^2 \neq \sigma_2^2$ for a two-tailed test

Source: saylordotorg.github.io/text_introductory-statistics/s15-03-f-tests-for-equality-of-two-va.html

Variability in variance – cont'd

- For N (iid) observations from a $N(\mu, \sigma^2)$, we have – see Lec-07: $\frac{(N-1)s^2}{\sigma^2} \sim \chi^2_{N-1}$
- Sample variance has a distribution that has the same shape as a chi-square distribution with N−1 degrees of freedom:

$$s^2 \sim \frac{\sigma^2}{N-1}\chi^2_{N-1}$$

- Under the assumption of a normal distribution, we can see that the ratio of the two sample variances will have an **F distribution** multiplied by the ratio of the two population variances:

This gives rise to an extremely simple test for comparing two variances:
$H_0: \sigma_1^2 = \sigma_2^2$

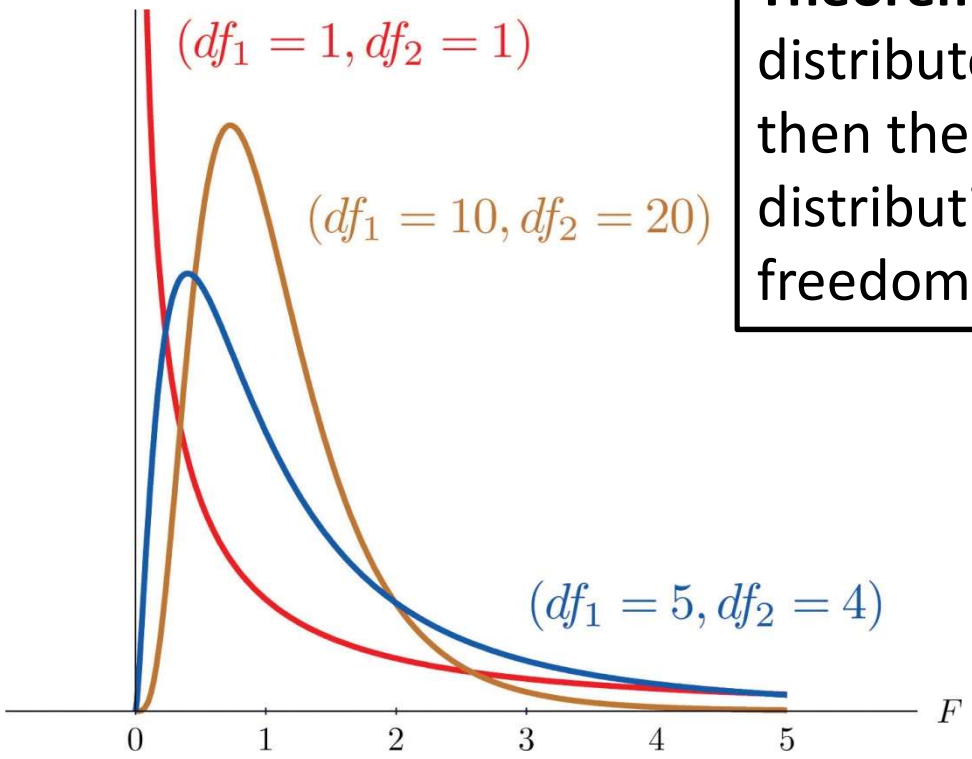
$$\frac{s_1^2}{s_2^2} \sim \frac{\frac{\sigma_1^2}{N_1-1}\chi^2_{N_1-1}}{\frac{\sigma_2^2}{N_2-1}\chi^2_{N_2-1}} = \frac{\sigma_1^2\chi^2_{N_1-1}/(N_1-1)}{\sigma_2^2\chi^2_{N_2-1}/(N_2-1)}$$

$$\frac{s_1^2}{s_2^2} \sim \frac{\sigma_1^2}{\sigma_2^2} F(df_1 = N_1-1,\ df_2 = N_2-1)$$

F-distribution

The F-distribution

- ANOVA uses F-tests to statistically test the equality of variances, named in honor of Ronald Fisher
- F-statistic is the ratio of two variances from samples
- Characterized by the following probability density function *f(F):*



Theorem: If χ_n^2 and χ_m^2 are independently distributed chi-squared random variables, then the random variable F(n,m) is the F distribution with n and m degrees of freedom, respectively:

$$F(n,m) = \frac{\chi_n^2 / n}{\chi_m^2 / m}$$

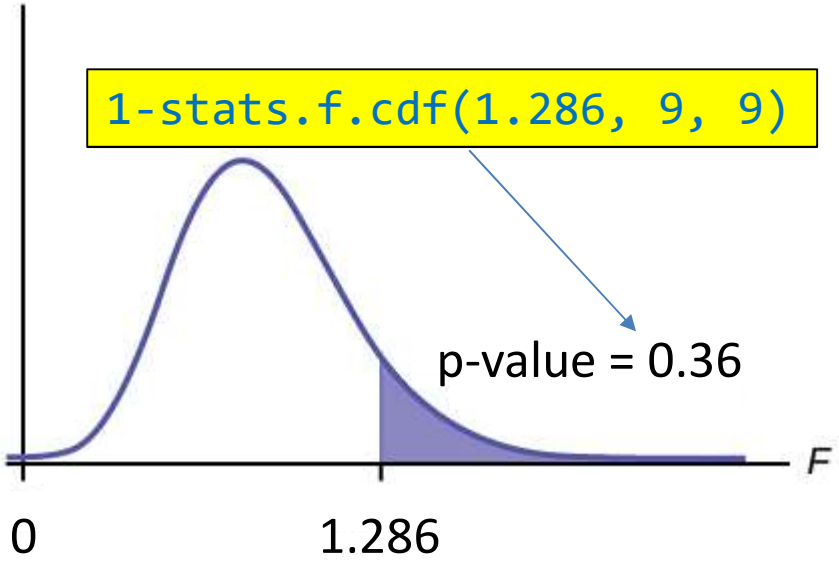
$$f(F) = cF^{(\upsilon_1/2)-1}\left(1+\frac{\upsilon_1 F}{\upsilon_2}\right)^{-(\upsilon_1+\upsilon_2)/2}$$

Variability in variance – cont'd

- Going back to tomato-sauce problem, the F-ratio is:

$$F = \frac{S_1^2}{S_2^2} = \frac{38254.4}{23116.7} = 1.286$$

- df1 = n_1 - 1 = 9 and df2 = n_2 – 1 = 9
- F-distribution for α = 0.05, df_1=9 and df_2=9



We can either determine $F_{critical} = F(\alpha, df_1, df_2)$ and compare this with 1.286

Or, compute the p-value directly for 1.286 : 0.36

- Since p > 0.05, we fail to reject H_0. There is **not significant evidence** at 0.05 level **that the variances are different**.

What is ANOVA?

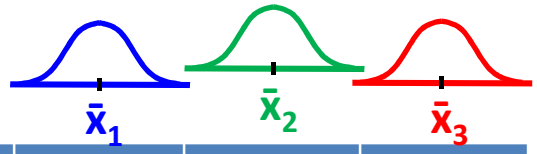
- **AN**alysis **O**f **VA**riance (ANOVA)
- To this point we've been comparing two populations
  - Independent samples t-test (random)
  - Matched-sample (paired) t-test
- This is limiting... What if we want to compare the means of more than two populations?
- What if we want to compare populations each containing several levels of subgroups?
- ANOVA can determine whether the means of three or more groups are different by using variances (somewhat interesting that it uses variances to determine if the means are different!)

Variables in ANOVA

- The value of the dependent variable (DV, **response** or target) depends on 1 or more explanatory variables (EV, **factors** or treatments) and on random effects
- Generally, we have a quantitative response variable as it relates to one or more categorical variables (want to study the effect of one or more ***qualitative*** variables on a ***quantitative*** outcome variable):
  - Which promotional campaigns lead to greatest sales during the Valentine's day?
    - DV: store income (continuous
    - EV: promotion type (nominal)
  - How does the number of days for exercising affect the amount of weight loss?
    - DV: Weight loss (continuous)
    - EV: Exercise frequency (on a scale of 1 to 5 days / week)

Example – cont'd

- Weight loss under 3 different diets: Are the differences in means of 3 diet types significant?
- H_0: No difference among groups



#	Diet1	Diet2	Diet3
1	1.0	1.1	1.2
2	1.2	1.4	1.3
3	1.8	1.9	1.7
4	2.4	2.3	2.5
5	3.6	3.5	3.7

$\bar{x}_1 = 2.00$ $\bar{x}_3 = 2.08$

$\bar{x}_2 = 2.04$

Overall mean for all 15 observations: $\bar{X}_{overall} = 2.04$

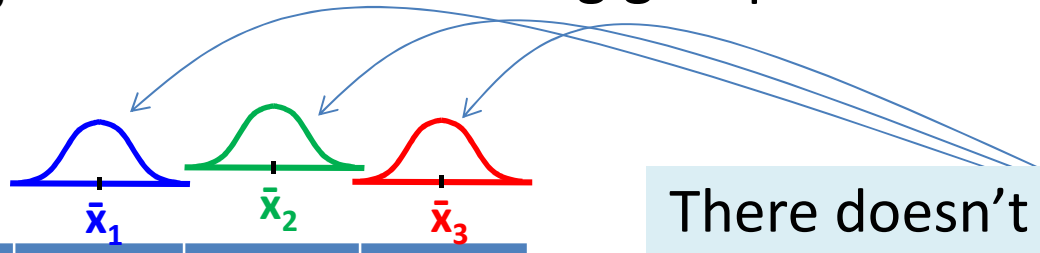
We have a considerable variation in each group. Different people have different reactions to different diet types?

We don't, however, seem to have a significant variation among different groups (factors) even though each group was subjected to a different type of diet.

Conclusion: It's the people that make the difference, not diet type.

Example – cont'd

- Consider the following case now:
- H_0: No difference among groups



#	Diet1	Diet2	Diet3
1	2.9	1.8	0.5
2	3.0	1.9	0.7
3	3.1	2.1	0.6
4	3.1	2.0	0.5
5	2.9	1.9	0.9

$\bar{x}_1 = 3.00$ $\bar{x}_3 = 0.64$

$\bar{x}_2 = 1.94$

Overall mean for all 15 observations: $\bar{X}_{overall} = 1.86$

There doesn't seem to be a considerable variance among people who followed the same type of diet.

Different diet types seem to have an effect on weight loss. Variation is considerable among the factors.

Conclusion: It's the diet type that makes the difference, not the people. So reject the Null.

Example – cont'd

- In this exercise, the variance among groups was obvious. In most cases, the variance between and within groups may not be so obvious.
- So, we want to be able to figure out **how much of total variance comes from**:
  - The **variance between** the groups
  - The **variance within** the groups

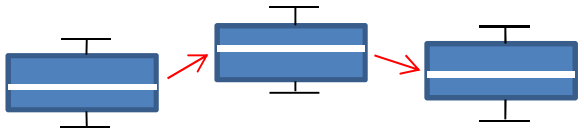
  Calculate the ratio: $F = \frac{Variance\ \mathbf{between}\ groups}{Variance\ \mathbf{within}\ groups}$

- **The larger the ratio, the more likely it is that groups have different means leading to rejection of H_0.**
- For the previous example: $F(.05, 2, 12) = 410.94$, $p = \mathbf{8.881e\text{-}12}$
- $p < .05$ => Reject H_0: At least one mean differs

More on the ratio

- We'll examine the variability between the sample means and within the sample for all groups (assuming same sample size)

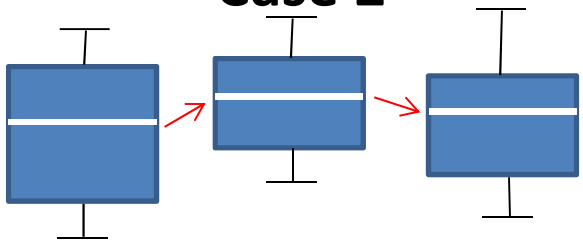
Case 1



$\bar{X}_1$=5 $\bar{X}_1$=7 $\bar{X}_1$=6

There is some variability **between** the sample means here.

Case 2



$\bar{X}_1$=5 $\bar{X}_1$=7 $\bar{X}_1$=6

Same variability **between** the sample means in comparison to case **1**.

Variability **within** samples in case **2** is much greater than the variability within samples in case **1**. But how will this affect the F-ratio?

- Computing the ratio of variability between the sample means and within the samples (assuming H_0: $\mu_1 = \mu_2 = \mu_3$), we have:

$$ratio = \frac{Between\ (B)}{Within\ (W)}$$

B	W	ratio	Verdict
LARGE	small	LARGE	**Reject H_0**
small	LARGE	small	Fail to reject H_0
Similar	Similar	~ 1	Fail to reject H_0

One-way ANOVA

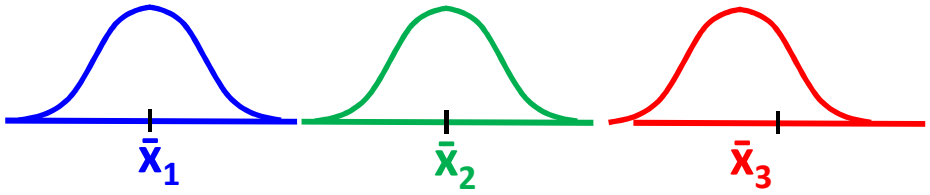
- **One-way** (single factor) **ANOVA**
  - Simplest case (outcome variable is what we're comparing)
  - An experiment that relates the values of a response variable to only **1 factor** is called a one-way (one-factor) ANOVA
  - ANOVA easily generalizes to more factors
  - It enables all groups to be compared with each other simultaneously rather than individually
  - ANOVA is based on the principle of **partitioning the total variance into components measuring the variance between the groups and within groups** (random effects – error)
- **Example**:
  - 500 patients tested on different diets (just **1 factor**: diet)

    Diet1 Diet2 Diet3 Diet4
  - Collision test scores on different car sizes (**1 factor**: car size)

    Compact Mid-size Full-size

Why not use multiple t-tests?

- **Question**: In the presence of 3 or more populations, why not use multiple t-tests (pairwise comparisons)?
  - Pairwise comparison means multiple t-tests all with $\alpha=0.05$ Type I error rate at 95% confidence
  - We have to compose 3 separate hypotheses:



$H_0: \bar{x}_1 = \bar{x}_2$ ($\alpha=0.05$) $\cap$ $H_0: \bar{x}_1 = \bar{x}_3$ ($\alpha=0.05$) $\cap$ $H_0: \bar{x}_2 = \bar{x}_3$ ($\alpha=0.05$)

  - If we do each test at level α, what is our chance of getting a rejection by at least one of these tests when in fact all 3 are equivalent?
  - If these 3 est are independent of each other, we have:

Why not use multiple t-tests?

$$\begin{aligned}\textbf{P(Overall Type-I error)} &= \textbf{P(reject at lesat one of these hypotheses)} \\ &= 1 - \textbf{P(fail to reject -FTR- all of these hypotheses)} \\ &= 1 - \textbf{P(FTR } H_{0,x_1,x_2} \cap \textbf{ FTR } H_{0,x_1,x_2} \cap \textbf{ FTR } H_{0,x_1,x_2}) \\ &= 1-(1-\alpha)^3 = 1-0.95^3 = 0.143\end{aligned}$$

- The error compounds with each t-test, and our significance level becomes 0.143. So, the chance of making a wrong decision at least once (the overall Type I error) is much higher than planned for in the individual tests.
- This is our overall α now (Type I error rate). This is the exact reason why multiple t-tests is not a solution to this problem unless a correction is used (such as Bonferroni correction).
- We'l revisit this concept when we get to the Post-hoc tests later in this chapter.

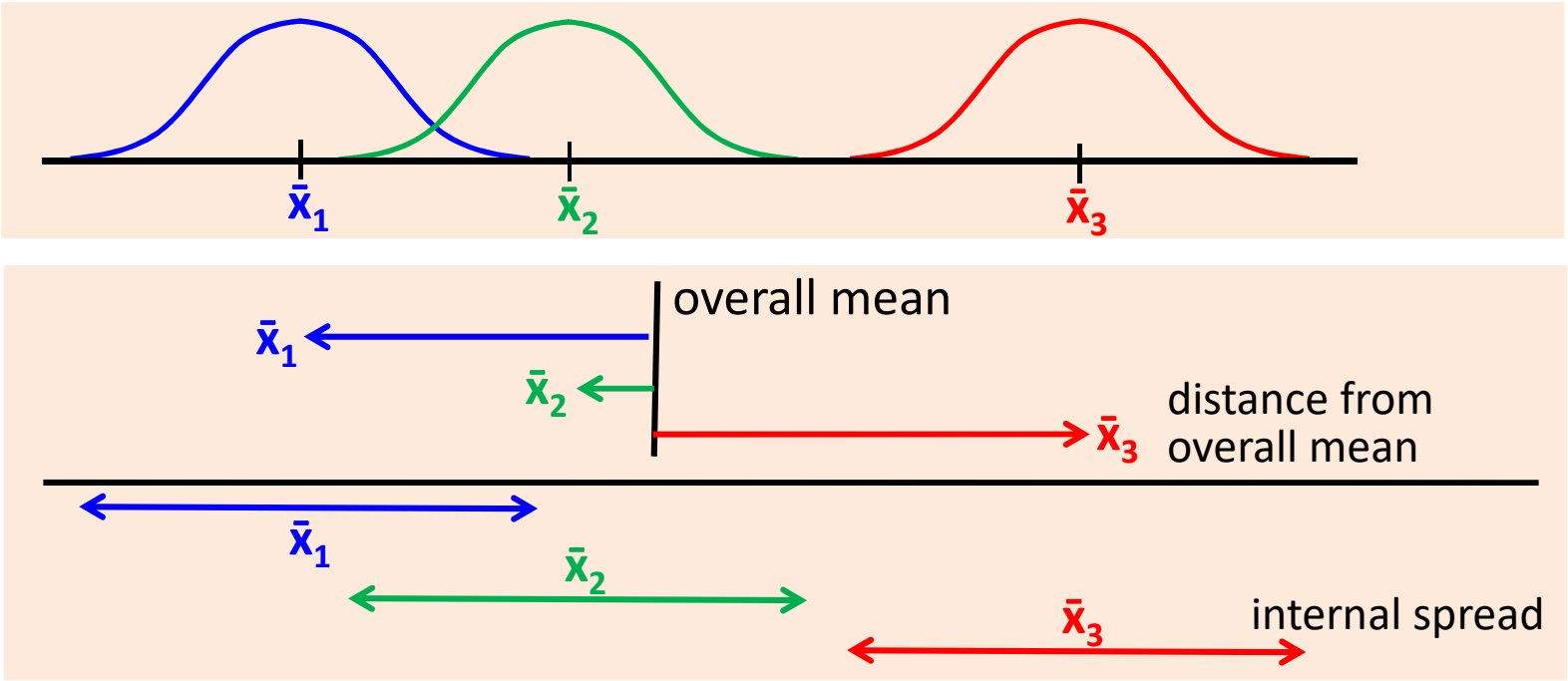
Basic ANOVA concepts – cont'd

- How does it work? ANOVA looks at a variability ratio

computed as: $$\text{Ratio} = \frac{\text{Var(BETWEEN the groups)}}{\text{Var(WITHIN the groups)}}$$

where TOTAL variance = Var(BETWEEN) + Var(WITHIN)

If the ratio >> 1 It's unlikely that samples come from a common population (reject H_0)

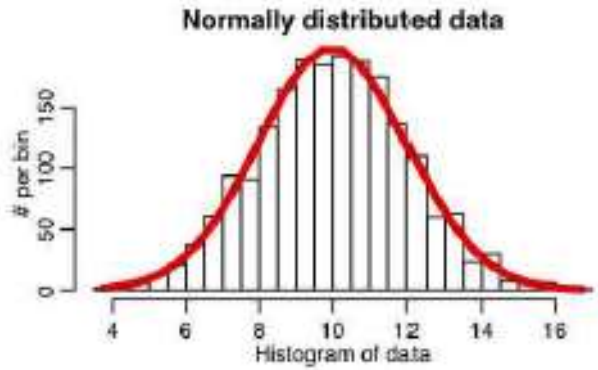


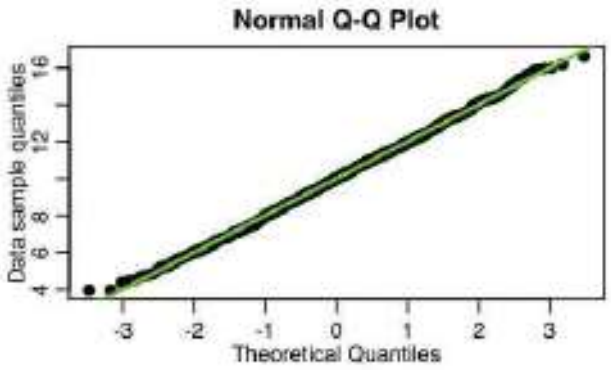
Basic ANOVA concepts – cont'd

- Underlying assumptions of ANOVA:
  1. Random samples and **independence** among data points.
  2. Your dependent variable (residuals) should be approximately **normally distributed** for each population. Otherwise, we may not be able to trust our *p*-values, which were built by assuming normality.
  3. The variance of your dependent variable (residuals) should be the the same among different groups (while the population means could be different from one group to next). This is called **Homoscedasticity**. This will impact the significance level, at least when sample sizes are not equal.
- Completely randomized design
  - If the experimental units (data points) are assigned to factor levels completely at random, then the experimental design is referred to as a completely randomized design.

Basic ANOVA concepts – cont'd

- You should test for the validity of these assumptions
- **Normality**
  - Can be tested by graphically (histograms, boxplots, QQ-plot)





  - Or by using statistical tests like
    - **Chi-square**, **Shapiro-Wilk**, **Jarque-Bera**, etc.
  - Small departures from normality are ok. The hypothesis tests discussed are only strongly affected if the data are VERY non-normal.
  - Non-normality can be fixed by increasing the sample size or by some data transformation.

Basic ANOVA concepts – cont'd

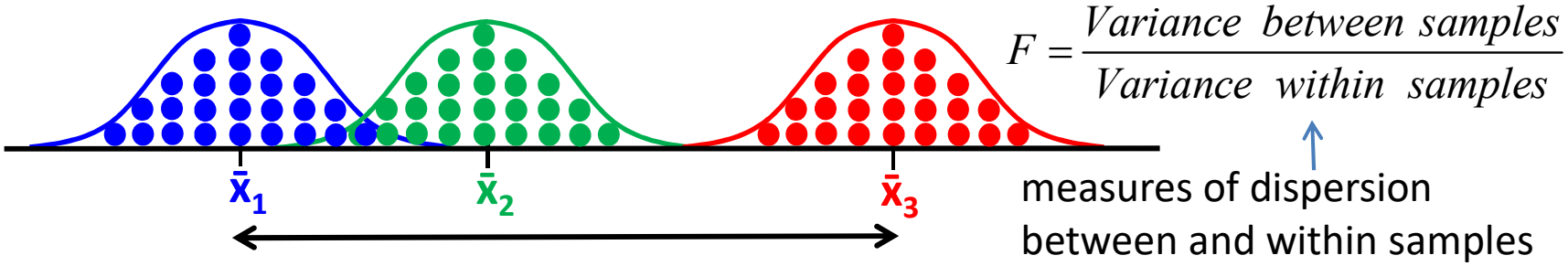
- **Homoscedasticity**
  - Can be tested by graphically (especially boxplots), or by using statistical tests like **Levene's test**, **Bartlett's test**, etc.
  - The ANOVA hypothesis tests are not strongly affected by small differences in variance between treatments, especially if the sample sizes for each treatment are equal. As a rule of thumb, the difference in variance is too great to trust our hypothesis test results if the following is true:

$$\frac{\textbf{largest group variance}}{\textbf{smallest group variance}} > 3$$

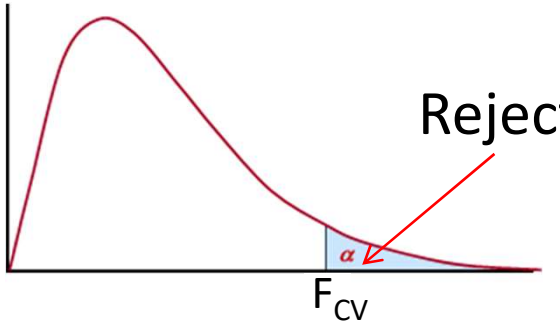
  - This gives us a simple check for constant error variance.
  - The ANOVA test is not too sensitive to inequality of variances if the sample sizes are equal.
- ANOVA is somewhat robust to slight deviations from normality and equal variance assumptions

Steps for Hypothesis testing in ANOVA

- **Step 1** Form the hypothesis:
  H_0: $\mu_1 = \mu_2 = \mu_3$ & H_A: At least one of the means is different
- **Step 2** Calculate the test-statistic F (a tedious process)



- **Step 3** Find the rejection region (F-distribution)

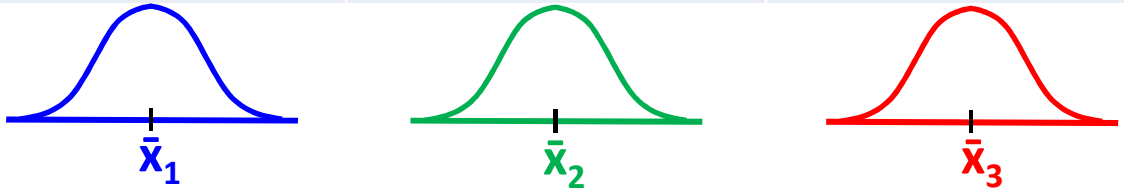


$F_{CV} = f(\alpha, df_1, df_2)$

- **Step 4** Decision => If test statistic $F > F_{CV}$, reject H_0
- **Step 5** Conclusion: The data provide sufficient evidence to conclude that at least one of the means is different.

Computing the F-statistic

obs#	Group1 (j=1)	Group2 (j=2)	Group3 (j=3)
i=1	x_{11}	x_{12}	x_{13}
i=2	x_{21}	x_{22}	x_{23}
...	...	...	...
i=n	... n_1	... n_2	... n_3



n_j : Size of the random sample selected from population "j"

$\bar{X}_j$: Mean of the group (sample) "j" => $\bar{X}_j = \sum_{i=1}^{n_j} \frac{x_{ij}}{n_j}$

$\bar{X}$: Overall (grand) mean => $\bar{X} = \sum_{j=1}^{c} \sum_{i=1}^{n_j} \frac{x_{ij}}{n} = \sum_{j=1}^{c} \frac{n_j \bar{X}_j}{n}$

Computing the F-statistic – cont'd

$$(x_{ij}-\bar{X})^2=(\bar{X}_j-\bar{X})^2+(x_{ij}-\bar{X}_j)^2+2(\bar{X}_j-\bar{X})(x_{ij}-\bar{X}_j)$$

$$\sum_{i=1}^{n_j}(x_{ij}-\bar{X}_j)=\sum_{i=1}^{n_j}x_{ij}-\sum_{i=1}^{n_j}\bar{X}_j$$

$$=n_j\bar{X}_j-\bar{X}_j\sum_{i=1}^{n_j}=0$$

SST (Sum of Squares Total)
amount of total variation among all observations in all samples

$$SST=\sum_{j=1}^{c}\sum_{i=1}^{n_j}(x_{ij}-\bar{X})^2$$

$$x_{ij}-\bar{X}=(\bar{X}_j-\bar{X})+(x_{ij}-\bar{X}_j)$$

c: number of groups (columns)
n_j : # of samples for each group
$\bar{X}$: Overall (grand) mean

SSB (**S**um of **S**quares **B**etween)
amount of variance between columns, hence also known as **SSC**

$$SSB\,(SSC)=\sum_{j=1}^{c}\sum_{i=1}^{n_j}(\bar{X}_j-\bar{X})^2$$

$$=\sum_{j=1}^{c}n_j(\bar{X}_j-\bar{X})^2$$

(weighted by the number of observations in sample ***j***)

SSW (**S**um of **S**quares **W**ithin)
amount of squared deviations from the sample means within each population

$$SSW\,(SSE)=\sum_{j=1}^{c}\sum_{i=1}^{n_j}(x_{ij}-\bar{X}_j)^2$$

(due to random errors, x_{ij} differs from $\bar{x}_j$, so this quantity is also called "Sum of Squares due to **Error**" and shown by **SSE**)

Overall (grand) mean :

$$\bar{X}=\sum_{j=1}^{c}\sum_{i=1}^{n_j}\frac{x_{ij}}{n}=\sum_{j=1}^{c}\frac{n_j\bar{X}_j}{n}$$

Degrees of freedom

- **Degrees of freedom between groups:** C groups, N instances

Group1	Group2	Group3
1	4	7
2	5	8
3	6	9
$\bar{x}_1$	$\bar{x}_2$	$\bar{x}_3$

$$Var_B = f(\bar{x}_j, \bar{X}) : \ (\mathbf{C\ groups},\ \bar{X}\ \mathbf{is\ known})$$

$$\Rightarrow df_B = C - 1$$

- **Degrees of freedom within groups:**

Group1	Group2	Group3
1	4	7
2	5	8
3	6	9
$\bar{x}_1$	$\bar{x}_2$	$\bar{x}_3$
Var: n_1-1	n_2-1	n_3-1

$$df\,(\mathbf{for\ Within}) = (n_1 - 1) + (n_2 - 1) + (n_3 - 1)$$

$$= n_1 + n_2 + n_3 - 3$$

# of all instances

# of groups

$$\Rightarrow df_W = N - C$$

- Degrees of freedom for "total"

$$Var_T = f(x_i, \bar{X}) : \ (i = 1,..,9\ \ \mathbf{N\ instances},\ \bar{X}\ \mathbf{is\ known})$$

$$\Rightarrow df_T = N - 1$$

Note that $df_T = df_B + df_W$

Computing the F-statistic – cont'd

- Sum of Squares Total: **SST = SSC + SSE**
- Calculating mean squares and degrees of freedom:
  - Before comparing SSC and SSE, we need to adjust the sums by dividing them by their associated degrees of freedom (df)

df for SSC	**c – 1**	Between groups: c group means computed, out of all c groups, c-1 are free (grand mean known)
df for SSE	**n – c**	Within groups (error): **c** groups, all with computed means => out of all **n** observations, only **n-c** are free (**c** group means are known)
df for SST	**n – 1**	c – 1 + n – c (n obs – grand mean = n-1)

Mean square between: $MSC = \frac{SSC}{c-1}$ Mean square within: $MSE = \frac{SSE}{n-c}$

- When F is large in a statistical sense, we would reject H_0

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

$H_0 : \mu_1 = \mu_2 = \ldots = \mu_c$

H_A: At least one of the means differs

Computing the F-statistic – cont'd

- All in a tidy ANOVA table:

Sources of variation	Sum of squares	Degrees of freedom	Mean square	F-statistic
Factors (between groups)	$SSC = \sum_{j=1}^{c} n_j (\bar{X}_j - \bar{X})^2$	c – 1	$MSC = \frac{SSC}{c-1}$	$F = \frac{MSC}{MSE}$
Error (within groups)	$SSE = \sum_{j=1}^{c} \sum_{i=1}^{n_j} (x_{ij} - \bar{X}_j)^2$	n – c	$MSE = \frac{SSE}{n-c}$	
Total	$SST = \sum_{j=1}^{c} \sum_{i=1}^{n_j} (x_{ij} - \bar{X})^2$	n – 1	$MST = \frac{SST}{n-1}$	

c: number of groups

n_j : number of samples for each group

$\bar{X}$: Overall (grand) mean

$$\bar{X} = \sum_{j=1}^{c} \frac{n_j \bar{X}_j}{n} = \sum_{j=1}^{c} \sum_{i=1}^{n_j} \frac{x_{ij}}{n}$$

Example

- Given the following data, any significant differences among the sample (group) means?

Group1	Group2	Group3
1	3	5
2	4	6
3	5	7

r = 3 rows)

c = 3 columns (groups)

n = r * c = 9

- **Grand** (overall) **mean** = (1+2+3+3+4+5+5+6+7) / 9 = **4**
- **T**otal **S**um of **S**quares (**SST** = Total variation):

SST = $(1-4)^2 + (2-4)^2 + (3-4)^2 +$
$(3-4)^2 + (4-4)^2 + (5-4)^2 +$
$(5-4)^2 + (6-4)^2 + (7-4)^2$

SST = **30**

Out of all 30 measurements, we have used the grand mean to compute the total variation. So we have 9-1= 8 independent measurements (only 8 of these provide information, df=8)

Example – cont'd

- Now we've computed the total variation, how much of this is coming from between and within groups?
- Variation from within groups (**SSW**): How much of SST is due to how far each of these data points from their central tendency (their respective means)?

Group1	Group2	Group3
1	3	5
2	4	6
3	5	7
$\bar{X}_1 = 2$	$\bar{X}_2 = 4$	$\bar{X}_3 = 6$

SSW = $(1-2)^2 + (2-2)^2 + (3-2)^2 +$
$(3-4)^2 + (4-4)^2 + (5-4)^2 +$
$(5-6)^2 + (6-6)^2 + (7-6)^2$

SSW = **6**

6 of the total variation (30) is coming from the variation within samples.

How many degrees of freedom (independent data points) do we have here? For each group, we have 3 data points and one group mean: 3 - 1 = 2 df per group.
For 3 groups, we have:
3 * 2 = 6 df, i.e., [**c * (r-1)**] df.

Alternatively, out of 9 points we have 3 group means, hence 9–3=6 df

Example – cont'd

- Variation from between groups (**SSB**): How much of the total variation is from between the sample means?

Group1	Group2	Group3
1	3	5
2	4	6
3	5	7
$\bar{X}_1 = 2$	$\bar{X}_2 = 4$	$\bar{X}_3 = 6$

For each data point, we compute the variation between the group mean that the data point belongs to and the grand mean.

Variability of the group means compared to the grand mean (the variability due to treatment)

$$SSB = n_1 * (2-4)^2 + n_2 * (4-4)^2 + n_3 * (6-4)^2$$

$$SSB = 24$$

$n_1 = n_2 = n_3 = 3$

Grand mean

How many df? If you know the sample means of 2 groups and the grand mean, you can always figure out the sample mean of the 3rd group. So, in general, in c groups, we have c -1 df. Here the df for SSB is 3 -1 = 2.

Example – cont'd

- So, we have:

	Variation	Degrees of freedom
SSB	24	c-1 = 2
SSW	6	c(r-1) = 6
SST	30	rc-1 = 8

$$F_{statistics} = \frac{\text{Variation between groups}}{\text{Variation within groups}} = \frac{\frac{SSB}{c-1}}{\frac{SSW}{c(r-1)}} = \frac{24/2}{6/6} = 12$$

$F_{critical} = F(\alpha, df_1, df_2) = F(0.05, 2, 6)$

- Compare ($F_{statistics} > F_{critical}$?) for statistical significance

Example

- In a study, the safety of compact cars, midsize cars, and full-size cars are examined. We collect a sample of three for each of the groups (cars types). Using the data provided below, test whether the mean pressure applied to the driver's head during a crash test is equal for each type of car based on a 95% confidence level (i.e., $\alpha = 5\%$)

Compact cars	Midsize cars	Full-size cars
665	447	495
707	417	427
673	543	507
. . .	. . .	. . .

  - H_0: $\mu_1 = \mu_2 = \mu_3$
  - H_A: At least one mean pressure is different
- We will compute the F-statistic as given in the last slide

Example – cont'd

Source	SS	df	MS	F	F-critical
Between	481693.33	2	240846.667	161.457	3.22
Within	62651.466	42	1491.701		
Total	544344.8	8			

$\bar{x}_{compact} = 678.60$
$\bar{x}_{midsize} = 475.27$
$\bar{x}_{full} = 445.93$

$\bar{X}_{overall} = 533.27$

$$df_{between} = c - 1 = 3 - 1 = 2$$
$$df_{within} = n - c = 45 - 3 = 42$$
$$df_{total} = n - 1 = 45 - 1 = 44$$

$$SSC = \sum_{j=1}^{c} n_j (\bar{x}_j - \bar{X})^2 = 481693.33$$

$$SST = \sum_{j=1}^{c} \sum_{i=1}^{n_j} (x_{ij} - \bar{X})^2 = 544344.8$$

$$SSE = SST - SSC = 544344.8 - 481693.33 = 62651.466$$

$$MSC = \frac{SSC}{c-1} = \frac{481693.33}{3-1} = 240846.667$$

$$MSE = \frac{SSE}{n-c} = \frac{62651.466}{45-3} = 1491.701$$

$$=> F = \frac{MSC}{MSE} = \frac{240846.667}{1491.701} = 161.4577$$

Example – cont'd

df	2	3
1	35.99	54.00
2	8.776	11.94
3	5.907	7.661
4	4.943	6.244
5	4.474	5.558
6	4.199	5.158

. . .

30	3.337	3.919
42 40	3.294	3.858
60	3.251	3.798
120	3.210	3.739
inf	3.170	3.682

- What is the critical value of F?
- Degrees of freedom for "between" and "within"
- $df_{between} = df_1 = 3-1=2$ / $df_{within} = df_2 = 45-3=42$
- For a 95% confidence level, we need $F^{CV}_{(\alpha=0.05,2,42)}$
- We read the critical value off the table for $\alpha=.05$:

Through linear interpolation: $F_{critical\_(0.05,2,42)} = 3.22$

- As the $F_{statistic}$ (161.457) > $F_{critical}$ (3.22), we reject H_0. So, we conclude that at least one of the mean head pressures is different from at least one other population mean (i.e., We're 95% confident that the mean head pressure is NOT statistically equal for compact, mid- and full-size cars.

Example – cont'd

- Working this out using Python:

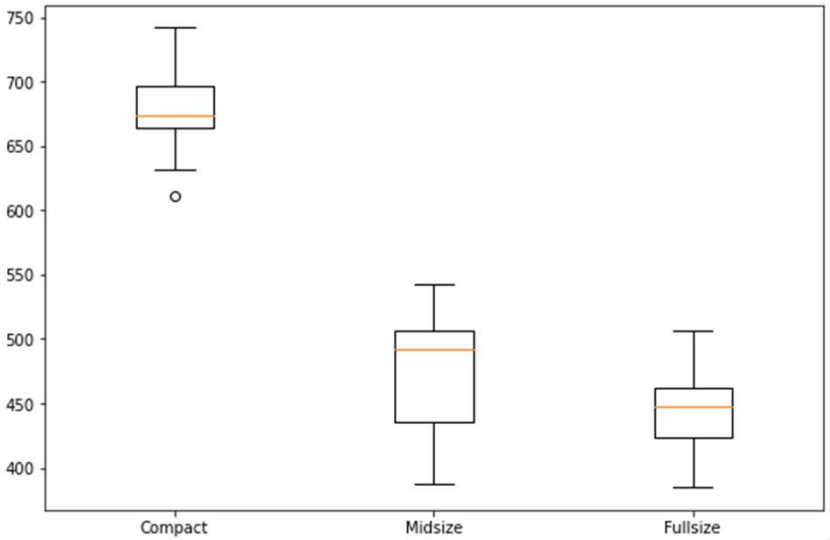
```
import numpy as np
import scipy.stats as stats
import matplotlib.pyplot as plt

compact  = [665,707,673,689,742,657,673,674,632,705,682,663,718,688,611]
midsize  = [447,417,543,503,500,510,425,492,467,412,387,501,522,480,523]
fullsize = [495,427,507,441,419,436,459,456,405,448,405,462,385,463,481]

cars = [compact,midsize,fullsize]
fig = plt.figure(1, figsize=(9, 6))
ax = fig.add_subplot(111)
bp = ax.boxplot(cars)
ax.set_xticklabels(['Compact', 'Midsize', 'Fullsize'])
plt.show()

from scipy import stats
F, p = stats.f_oneway(compact, midsize, fullsize)
print(F,p)
```

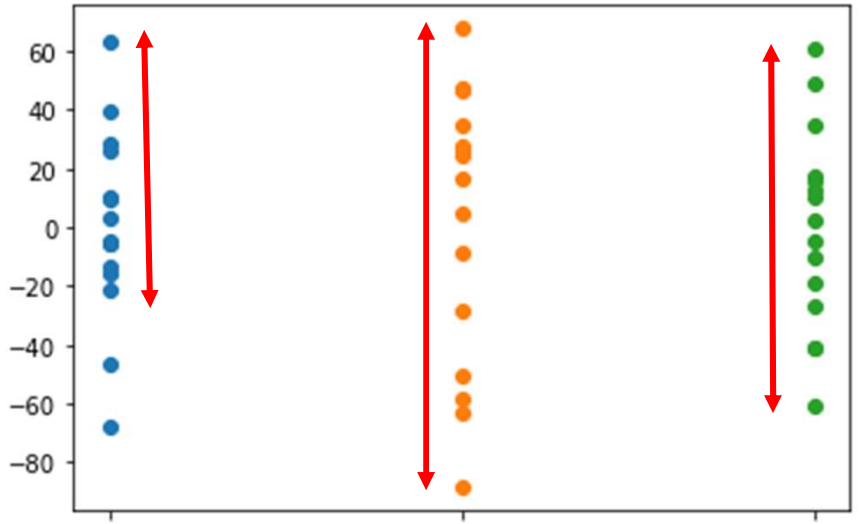
```
161.457    1.915139e-20
```

Python code

Example – cont'd

- Homoscedasticity (constant-variance) assumption?





compact – compact.mean = C-C_m M-M_m F-F_m

Example – cont'd

- Output of the ANOVA analysis:

```
161.457672712   1.9151396049e-20
```

F-statistic

p-value < 0.05, so reject H_0

- Getting the critical F value ($F^{CV}_{(\alpha=0.05,2,42)}$) in Python:

```
print('Critical F-value at %:',stats.f.ppf(0.95,2,42))
Critical F-value at 5%: 3.21994229
```

$df_{between} = df_1$ $df_{within} = df_2$

- Getting the p-value at F=161.457 ($\int_{161.5}^{\infty} f(F,\upsilon_1,\upsilon_2)dF$):

```
print('p-value at F=161.5 :', 1-stats.f.cdf(161.5,2,42))
p-value at F=161.5 : 1.11022302462e-16
```

- As the p-value is smaller than α=0.05, we reject H_0 and state that at least one of the means for the head pressure is significantly different. Which one?

Post Hoc Test

- ANOVA is an Omnibus test
  - It tests for an overall difference between groups: Is there a significant difference between group means? If so, it doesn't tell us exactly which means differ.
- So, we follow a significant ANOVA test with a post-hoc (meaning after) test. There are multiple ways we can conduct a **post-hoc test** (Bonferroni correction, Tukey's HSD test, etc.)
- Simplest is the **Bonferroni** (correction) **test** which is a series of t-tests performed on each pair of groups.
- Consider a case where you have 10 hypotheses to test, and a significance level of 0.05. What's the probability of observing at least one significant result just due to chance?

P(at least one significant result due to chance) = 1 − P(no significant results)

$= 1 - (1-0.05)^{10} \approx 0.40$

Post Hoc Test

- With 10 tests being considered, we have a 40% chance of observing at least one significant result, even if all the tests are not actually significant.
- So, we need to adjust **α** in some way, so that the probability of observing at least one significant result due to chance remains below our desired significance level (**α**).
  - For **k** groups, there are C(k,2) combinations of different pairs. We want to test them all with α_c where $\alpha_c = \alpha/C(k,2)$
  - If we have 5 groups, for example, we end up testing C(5,2)=10 pairs with a corrected level of significance α_c:

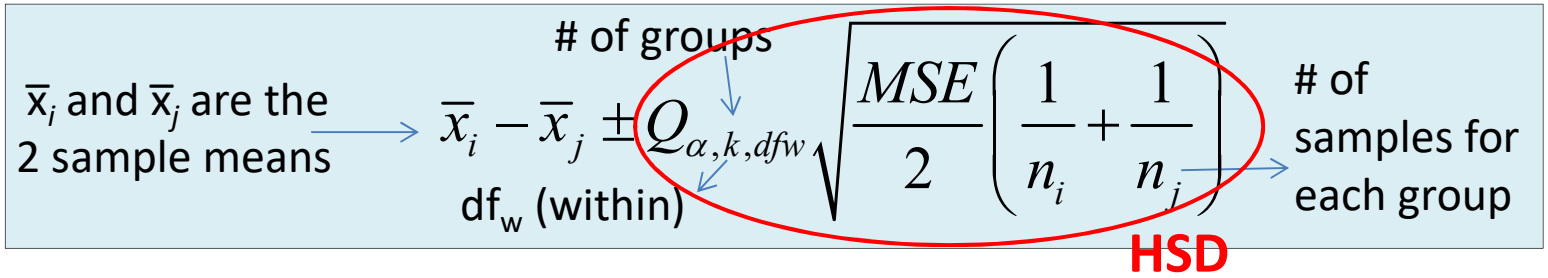
    For α=0.05 and 10 hypotheses, $\alpha_c = 0.05/10 = 0.005$

P(at least one significant result due to chance) = 1 − P(no significant results)
$= 1 - (1-0.005)^{10} \approx 0.049$

Slightly less than 0.05, so Bonferroni correction is a little conservative

Tukey's HSD test

- Tukey's HSD (**Honest Significant Difference**) test
- Which specific groups' means (compared with each other) are different? This is a test that compares all possible pairs of means (essentially a modified t-statistic that corrects for the compounding error in multiple comparisons).
- It calculates an "HSD" value based on the mean squared error, the sample size, and a value from Q-distribution (critical value of the studentized range for significance level, α):



- If the difference between two sample means is greater than HSD, then they are significantly different.
- We compute the CI similarly to the CI for the difference of two means but, using the Q distribution which avoids the problem of inflating alpha.

Tukey's HSD test – cont'd

- Going back to our pressure vs car type example:
- $F_{statistic} = 161.457 > F_{critical} = 3.22$ (where $F_{(0.05,\ 2,42)}=3.22$), so we rejected H_0.
- From the Q-table: $Q_{\alpha,k,dfw} = Q_{0.05,\ 3,\ 42} = 3.44$

$$HSD = Q_{0.05,3,42}\sqrt{\frac{MSE}{n}} = 3.44\sqrt{\frac{1491.7}{15}} = 34.3$$

equal sample sizes

- Min. difference between means must be 34.3 for significance.
- Now we'll arrange means in increasing order:

Fullsize	Midsize	Compact
445.93	475.27	678.6

→ Significantly different from both

< 34.3

- Underscore pairs that differ by less than w=34.3
- Pairs not underscored by same line are significantly different from one another.

Tukey's HSD test – cont'd

- Another way of looking at what we've found is the following:
- We're making pairwise comparisons. Having 3 types for the car sizes (Full, Mid and Compact) we need to look at the following pairs (F-M, F-C, M-C):
- HSD = 34.3 (margin of error for the confidence interval)

$\bar{x}_M - \bar{x}_F = 475.27 - 445.93 = 29.34 \Rightarrow 29.34 \pm 34.3$ ⇨ (-4.96 , 63.64)

$\bar{x}_C - \bar{x}_F = 678.6 - 445.93 = 232.67 \Rightarrow 232.67 \pm 34.3$ ⇨ (198.37 , 266.97)

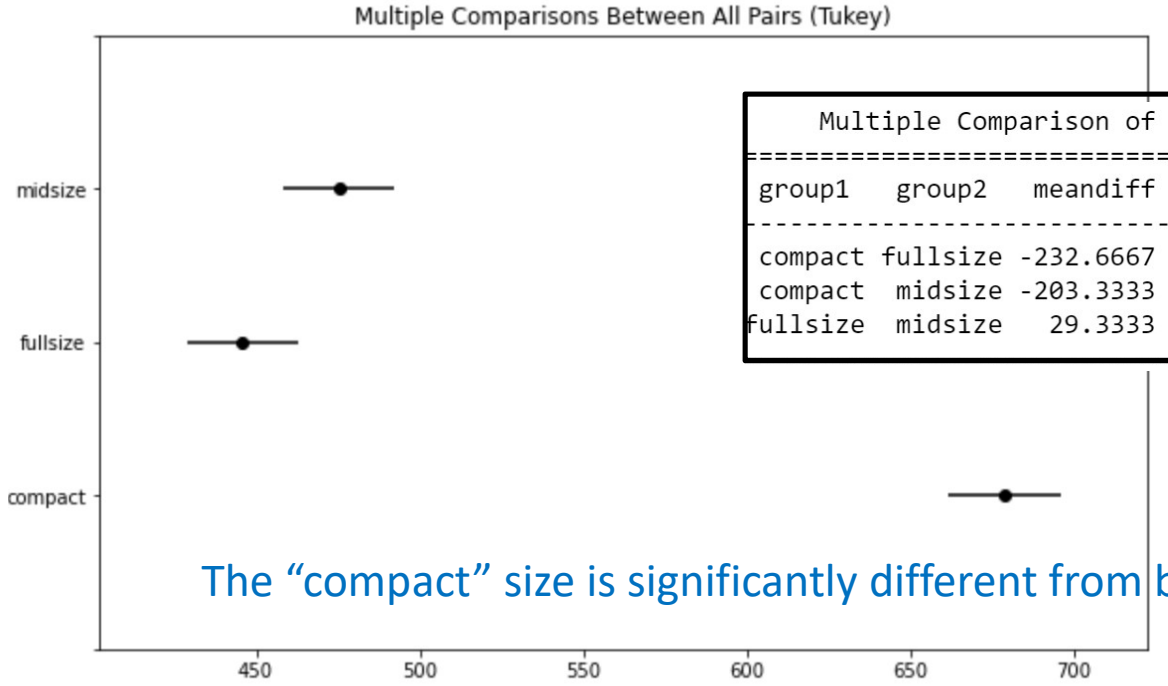
$\bar{x}_C - \bar{x}_M = 678.6 - 475.27 = 203.33 \Rightarrow 203.33 \pm 34.3$ ⇨ (169.03 , 237.63)

- Interpretation:
  - M-F: Inconclusive! Under 95% confidence, going from M to F, the pressure will possibly decrease or increase.
  - C-F: Significant difference. Going from F to C, the pressure will increase anywhere from 198.37 to 266.97
  - C-M: Significant difference.
- The "compact" size is significantly different from both "midsize" and "fullsize"

Tukey's HSD test – cont'd

- Running the Tukey's HSD test on Python we get:

```
from statsmodels.stats.multicomp import pairwise_tukeyhsd
from statsmodels.stats.multicomp import MultiComparison
pressure=np.array([665, 707, 673, . . ., 365, 463, 481])
cartype=np.array(['Compact',...,'Compact','Midsize',...,
                  'Midsize','Fullsize',...,'Fullsize'])
mc = MultiComparison(pressure,cartype)
result = mc.tukeyhsd()
print(result) ; result.plot_simultaneous()
```




```
    Multiple Comparison of Means - Tukey HSD, FWER=0.05
=====================================================================
group1   group2   meandiff p-adj    lower     upper   reject
---------------------------------------------------------------------
compact fullsize -232.6667  0.001 -266.9285 -198.4049   True
compact  midsize -203.3333  0.001 -237.5951 -169.0715   True
fullsize midsize   29.3333 0.1063   -4.9285   63.5951  False
```

The "compact" size is significantly different from both "midsize" and "fullsize"

Effect size for ANOVA

- The p-value ($1.9\ 10^{-20}$) we found for this problem indicated a significant difference? But what was the effect size?
- The effect size for a "between groups" ANOVA is given by a ratio called **eta squared** (***η²***):

Source	SS
Between	481693.33
Within (Error)	62651.47
Total	544344.8

$$\eta^2 = \frac{SS_{between}}{SS_{total}} = \frac{481693.33}{544344.8} = 0.885$$

A large effect (see below)

Interpretation: 89% of the total variance in our measurement is accounted by the size of the car.

$$\eta^2 = \frac{SS_{error}}{SS_{total}} = \frac{62651.47}{544344.8} = 0.115$$

This indicates that about one tenth of the variance is not accounted for at all.

- According to Cohen's (1988) guidelines:

η^2	Effect
0.01	Small
0.06	Medium
> 0.14	Large

Effect size for ANOVA

- η^2 is used specifically in ANOVA models. Each categorical effect in the model has its own η^2 , so it gives a specific measure of the effect of that categorical variable.
- η^2 has 2 shortcomings:
  1. As you add more variables, the proportion explained by any one variable will automatically decrease making it hard to compare the effect of a single variable in different cases. Partial η^2 solves this problem by making the comparison of the effect of the same variable in different cases with different covariates. In one-way ANOVA, η^2 = partial η^2.
  2. η^2 is a biased measure of population variance explained (an overestimation that gets smaller as sample size increases).
- ω^2 uses unbiased measures of the variance components and is always smaller than eta-squared.

Getting the ANOVA table using Python

- Using **statsmodels** to compute SS_{Total} and $SS_{Between}$:

```
import pandas as pd
raw = {'car': ['compact',  . . ., 'compact',
               'midsize',  . . ., 'midsize',
               'fullsize', . . ., 'fullsize'],
       'pressure': [665,707,673, . . ., 385, 463, 481]}
df = pd.DataFrame(raw, columns = ['car', 'pressure'])
```

```
import statsmodels.api as sm
from statsmodels.formula.api import ols

mod = ols('pressure ~ car', data=df).fit()

aov_t = sm.stats.anova_lm(mod, typ=2)
print(aov_t)
```

Use if there is no interaction effect

```
                sum_sq    df           F        PR(>F)
car       481693.333333   2.0  161.457673  1.915140e-20
Residual   62651.466667  42.0         NaN           NaN
```

```
aov_t['sum_sq'][0]/(aov_t['sum_sq'][0]+aov_t['sum_sq'][1])
>>> 0.88490481278
```

⇦ eta-square

Getting the ANOVA table using Python

- Using **pingouin** to compute SS_{Total} and $SS_{Between}$:

```
$ pip install pingouin
import pingouin as pg
```

```
pg.anova(dv='pressure', between='car', data=df, detailed=True)
```

	Source	SS	DF	MS	F	p-unc	np2
0	car	481693.333333	2	240846.666667	161.457673	1.915140e-20	0.884905
1	Within	62651.466667	42	1491.701587	NaN	NaN	NaN

```
pg.pairwise_tukey(data=df, dv='pressure', between='car')
```

	A	B	mean(A)	mean(B)	diff	se	T	p-tukey
0	compact	fullsize	678.600000	445.933333	232.666667	14.102962	16.497716	0.001000
1	compact	midsize	678.600000	475.266667	203.333333	14.102962	14.417775	0.001000
2	fullsize	midsize	445.933333	475.266667	-29.333333	14.102962	-2.079941	0.106306

```
from pingouin import welch_anova #for unequal variances
welch_anova(dv='  ... ', between='  ... ', data=df)
```

t-test vs ANOVA

- Independent samples t-tests (**equal variances**) are essentially a simplification of a one-way ANOVA for only two groups.
- In fact, if you run your t-test as an ANOVA, you'll get the same p-value. And the between-groups F statistic will be the square of the t statistic you got in your t-test.

```
import pandas as pd
raw = {'cat': ['a','a','a','a','a','a','a','a','a','a',
               'b','b','b','b','b','b','b','b','b','b'],
       'num': [860,850,750,870,940,410,410,820,890,890,
               540,640,600,640,300,610,430,280,300,610]}
df = pd.DataFrame(raw, columns = ['cat', 'num'])
```

```
import statsmodels.api as sm
from statsmodels.formula.api import ols

mod = ols('num ~ cat', data=df).fit()
aov_table = sm.stats.anova_lm(mod, typ=2)
print(aov_table)
```

```
            sum_sq    df          F    PR(>F)
cat       375380.0   1.0  12.233117  0.00257
Residual  552340.0  18.0        NaN      NaN
```

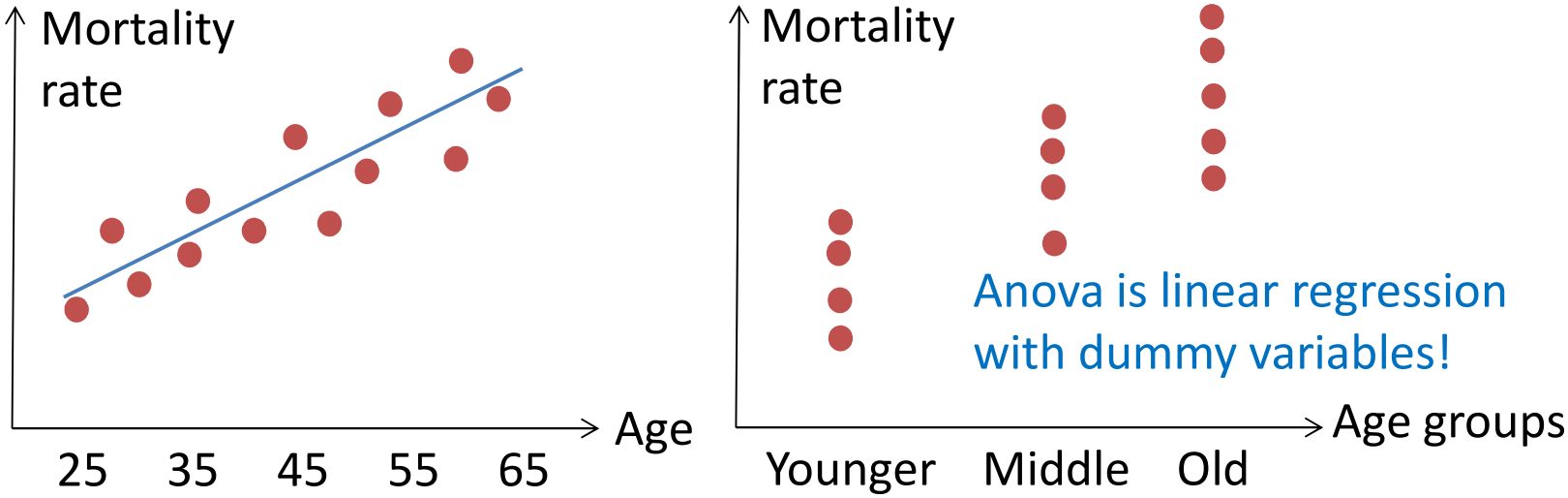
```
t, p = stats.ttest_ind(df[df.cat=='a'].num , df[df.cat=='b'].num)
print(t*t)
print(p)
```

```
12.233117282833039
0.002570460042718869
```

	cat	num
0	a	860
1	a	850
2	a	750
3	a	870
4	a	940
5	a	410
6	a	410
7	a	820
8	a	890
9	a	890
10	b	540
11	b	640
12	b	600
13	b	640
14	b	300
15	b	610
16	b	430
17	b	280
18	b	300
19	b	610

ANOVA vs Linear Regression

- Suppose we have data on mortality vs several age groups.



- Applying regression, we would reject H_0 that the slope=0 concluding age does affect the mortality rate.
- If the data were grouped crudely into 3 categories, the result would be the scatter shown on the right above.
- Same result with some loss of resolution in the age variable.

Non-parametric test for ANOVA

- **Kruskal-Wallis test**: A rank-based (rather than value) non-parametric test used when the assumptions are not met.
- Assumptions for ANOVA:
  - Random samples that are mutually independent
  - Groups have the same distribution (equal variance)
  - Measurement scale is at least ordinal for a continuous variable
- KW test is used as a test of stochastic dominance (testing whether samples originate from the same distribution). KW is a test on differences in ranks among the groups. If the samples are from an identical population, then the average rank should be about the same.
- If you can make the assumption of an identically shaped (and equal variance) distribution, the KW test can be used for testing the equality of population medians.

Non-parametric test for ANOVA

- If the data meet the requirements for a parametric test, it is better to use a one-way ANOVA because it is more powerful than the **KW test** which loses information when you substitute ranks for the original values.
- If the sample sizes are reasonably large and nearly equal, non-normality restriction can be relaxed, and the Welch-corrected one-way ANOVA can be used even for unequal variances.
- **KW test** using Scipy library:

```
from scipy.stats import kruskal
stat, p = kruskal(data1, data2, data3, …)
if p > alpha:
      print('Same distribution (fail to reject H0)')
else:
      print('Different distribution (reject H0)')
```

Two-way ANOVA

- A two-way ANOVA allows us to account for variation at the row level.

	Compact	Medium	Full
Japanese	…	…	…
European	…	…	…
US	…	…	…

- What impacts the variation in the pressure levels? Is it the size, or the origin of manufacturing, or both?
- We have 2 factors (size and origin), so this is a Two-way, or Two-factor ANOVA.
- Keep in mind that there may be no interactions between the two effects (only additive effects), or we may have interaction effects contributing to the target variable.