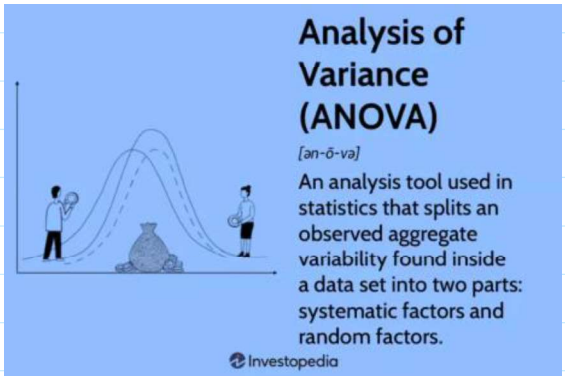
ANOVA & Chi square test

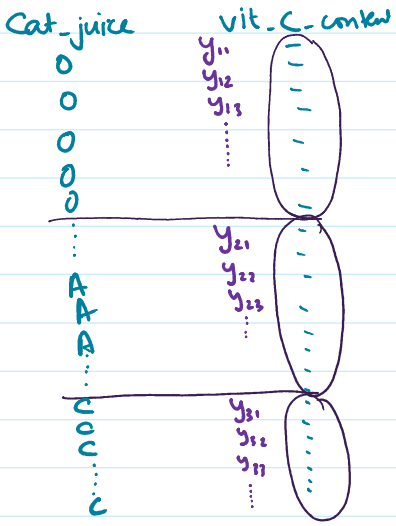
03 December 2022 07:10

ANOVA - Analysis of Variance : why is it needed

Numerical data which is categorized by categorical variables.

Orange Juice	2.8	3	2.95	3.10	2.78	
Apple Juice	1.97	2.16	2.31	1.86	2.11	Typhe
Carrot Juice	1.51	1.67	1.49	1.97	1.84	





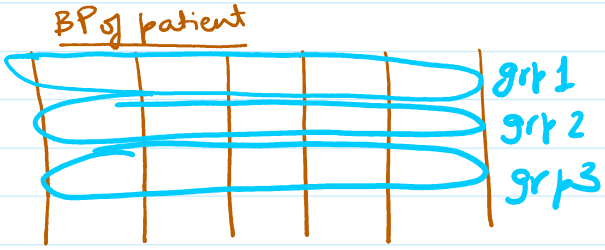
Problem statement 1: Does the average value of vit C change for different categories of juice?

Problem statement 2: Does my sales value differ significantly by region?

→ 16 stores countrywide: do all stores on an avg. sell uniformly

Real life examples: medical and vitC

→ Kidney disease of Diabetes
→ Kidney disease of Diabetes
→ No Kidney Disease



I want to check if BP values are affected by 3 groups.

Nike Bata Liberty Adidas

Agriculture example: 3 fertilizers. A, B, C.

farmers say that fert. A is the best and you being the govt. need to decide the best brand and give the brand an award.

→ Run an ANOVA test to verify weather or not the 3 brands are actually different or not.

One Way ANOVA - when t-test is not enough

The main goal of ANOVA is to test for multiple means.

like a t-test is used to compare means of 2 populations. An ANOVA test can do it for >2 categories / Populations

eg. for 6 categories I'll have to run t-test ${}^6C_2$ times

i.e. $\frac{6!}{2!4!} = \frac{6!}{4\times3\times4} = 15$ times.

Data design for one-way ANOVA

categories inside a single cat. variable

Class	Sample Observations				Total	Mean
1 (Apple Juice)	y_{11}	y_{12}	⋯	y_{1n_1}	$T_{1\cdot}$	$\bar{y}_{1\cdot}$
2 (Orange Juice)	y_{21}	y_{22}	⋯	y_{2n_2}	$T_{2\cdot}$	$\bar{y}_{2\cdot}$
⋮	⋮	⋮	⋮	⋮	⋮	⋮
i	y_{i1}	y_{i2}	⋯	y_{in_i}	$T_{i\cdot}$	$\bar{y}_{i\cdot}$
⋮	⋮	⋮	⋮	⋮	⋮	⋮
k (Carrot)	y_{k1}	y_{k2}	⋯	y_{kn_k}	$T_{k\cdot}$	$\bar{y}_{k\cdot}$

$\bar{y}_{Apple}$

$\bar{\bar{y}}$ or $\bar{y}_{\cdot\cdot}$ → grand mean

Ultimately I'm comparing $\bar{y}_{1\cdot}, \bar{y}_{2\cdot}, \ldots, \bar{y}_{k\cdot}$ to understand if $\mu_1, \mu_2, \ldots, \mu_k$ are different or not.

(multiple means for multiple categories / populations)

Variance overview

Variance is a measure of dispersion, i.e. how much does the data deviate from it's mean value.

Sample variance (s^2)

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

Population variance (σ^2)

$$\sigma^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2$$

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

$\bar{x}$ → sample mean
n → size of sample

$$\sigma^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2$$

μ → population mean
N → size of population

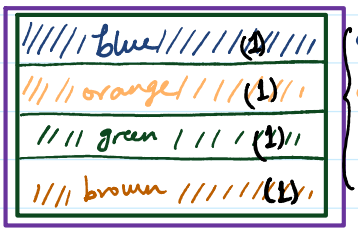
In a real life scenario we only have access to our sample so we know only the value of n, $\bar{x}$.

Basic principle of ANOVA

- The basic principle of ANOVA is to test for differences among the means of the populations by examining the amount of variation within each of these samples, relative to the amount of variation between the samples.
- Figure out how much of the total variance comes from:

1. The variance between the groups. ✓
2. The variance within the groups. ✓

Testing mean by variation



(1) Total variation ▭ can be split in 2 parts:
1) variation within groups.
i.e. variation of blue group +
variation of orange group +
variation of green group +
.. .. brown group.

2) Variation among the groups.
variation of (blue), (green), (orange), (brown)

Sum of squares and notation

y_{ij} → value in i^{th} group at j^{th} position

Total sum of squares = $\sum_i \sum_j (y_{ij} - \bar{\bar{y}})^2 = (y_{11} - \bar{\bar{y}})^2 + (y_{12} - \bar{\bar{y}})^2 + \dots + (y_{21} - \bar{\bar{y}})^2 + (y_{22} - \bar{\bar{y}})^2 + \dots$

TSS = sum of squares for all values of all groups.

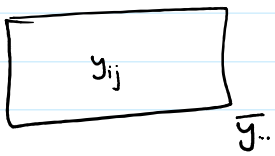
sum of squares (blue group) : $\sum_j (y_{blue\,j} - \bar{y}_{blue})^2$

$= (y_{b1} - \bar{y}_b)^2 + (y_{b2} - \bar{y}_b)^2 + \cdots + (y_{bn} - \bar{y}_b)^2$

ss. of all values in the blue group.

ss (brown group) : $\sum_j (y_{br\,j} - \bar{y}_{br})^2$

Splitting the sum of squares : TSS = SST + SSE or SSB + SSW



$y_{ij} - \bar{y}_{..} = y_{ij} - \bar{y}_{i.} + \bar{y}_{i.} - \bar{y}_{..}$

$y_{ij} - \bar{y}_{..} = (y_{ij} - \bar{y}_{i.}) + (\bar{y}_{i.} - \bar{y}_{..})$

$(y_{ij} - \bar{y}_{..})^2 = [(y_{ij} - \bar{y}_{i.}) + (\bar{y}_{i.} - \bar{y}_{..})]^2$

$(y_{ij} - \bar{y}_{..})^2 = (y_{ij} - \bar{y}_{i.})^2 + (y_{i.} - \bar{y}_{..})^2 + 2(y_{ij} - \bar{y}_{i.})(\bar{y}_{i.} - \bar{y}_{..})$

$\sum\sum (y_{ij} - \bar{y}_{..})^2 = \sum\sum (y_{ij} - \bar{y}_{i.})^2 + \sum\sum (\bar{y}_{i.} - \bar{y}_{..})^2$

$+ 2\sum\sum (y_{ij} - \bar{y}_{i.})(\bar{y}_{i.} - \bar{y}_{..}) = 0$

$\sum_i \sum_j (y_{ij} - \bar{y}_{i.})(\bar{y}_{i.} - \bar{y}_{..})$

$\sum_i [(y_{i1} - \bar{y}_{i.})(\bar{y}_{i.} - \bar{y}_{..}) + (y_{i2} - \bar{y}_{i.})(\bar{y}_{i.} - \bar{y}_{..}) + \cdots]$

$\sum_i (\bar{y}_{i.} - \bar{y}_{..}) [\quad] = [\quad] \sum_i (\bar{y}_{i.} - \bar{y}_{..})$ → 0

$$\sum\sum (y_{ij} - \bar{y}_{..})^2 = \sum_i \{\sum_j (y_{ij} - \bar{y}_{i.})^2\} + \sum\sum (\bar{y}_{i.} - \bar{y}_{..})^2$$

TSS = SSW + SSB

(1) (2)

* Total sum of Squares = Sum of sq. within groups. + Sum of sq. among groups.

Null hypothesis

$H_0 : \mu_1 = \mu_2 = \mu_3 = \cdots = \mu_K$

the population means for all the groups are the same.

H_1 : Atleast one group is significantly different from the others.

Assumptions

1. No. of data groups should be 3 or greater than 3. → o/w we can directly use t-test
2. Data collected should be equal for all the groups.
3. No outlier data should be included.
4. Data should be random and not biased.
5. The samples must be independent.
6. The data should be normally distributed.

The test statistic - F test

F - distribution and test of variances : Just like t-distribution is used to test means, F-distribution is used to test variances.

It tests if 2 variances are equal or not.

Calculate the ratio:

$$F = \frac{\text{between groups}}{\text{within groups}}$$

larger value of F ⇒ b/w groups variation is high → diff. pop^n means.

Smaller value of F ⇒ b/w groups variation is low → similar pop^n means.

The ANOVA table

Source of Variation	Sum of Squares (SS)	Degrees of freedom (df)	Mean Square (MS)	F
Factor (Between)	SS_{Factor} SSB	k-1	$MS_{Factor} = SS_{Factor}/k-1$	$F = MS_{Factor}/MS_{Error}$
Error (Within)	SS_{Error} SSW	n-k	$MS_{Error} = SS_{Error}/n-k$	
Total	SS_{Total} TSS	n-1		

now, when ANOVA is run in python / excel / or any other software, just like all hypothesis a p-value is generated.

→ If p-value < α (5%) → we reject H_0 and conclude : Pop^n means are different for atleast some groups

→ If p-value > α (5%) → we accept H_0 and conclude : Pop^n means are equal for all groups.

LSD introduction : Least Significant Difference

LSD technique is used after an ANOVA test says that pop^n mean are different. LSD technique helps us tell which groups are different.

2 way classification

TABLE 5-13 : TWO-WAY CLASSIFIED DATA

Treatments (Rations)	Varieties of Cows 1	2	...	j	...	h	Row Totals $= (\sum_j y_{ij})$	Row Means $= (\sum_j y_{ij})/h$
1	y_{11}	y_{12}	...	y_{1j}	...	y_{1h}	$T_{1\cdot}$	$\bar{y}_{1\cdot}$
2	y_{21}	y_{22}	...	y_{2j}	...	y_{2h}	$T_{2\cdot}$	$\bar{y}_{2\cdot}$
i	y_{i1}	y_{i2}	...	y_{ij}	...	y_{ih}	$T_{i\cdot}$	$\bar{y}_{i\cdot}$
k	y_{k1}	y_{k2}	...	y_{kj}	...	y_{kh}	$T_{k\cdot}$	$\bar{y}_{k\cdot}$
Column Totals	$T_{\cdot 1}$	$T_{\cdot 2}$	...	$T_{\cdot j}$	...	$T_{\cdot h}$	$G = \sum\sum y_{ij}$	
Column Means $= (\sum_i y_{ij})/k$	$\bar{y}_{\cdot 1}$	$\bar{y}_{\cdot 2}$	...	$\bar{y}_{\cdot j}$	...	$\bar{y}_{\cdot h}$		

Excel example

Chi Square tests (χ → 'chi' : greek letter)

test - statistic

$$x_c^2 = \frac{\Sigma (O_i - E_i)^2}{E_i}$$

The chi square statistic

$$\chi_c^2 = \sum \frac{(\text{observed} - \text{expected})^2}{\text{expected}}$$

Where

c = Degrees of freedom

test - statistic

$$x_c^2 = \frac{\Sigma\,(O_i - E_i)^2}{E_i}$$

The chi square statistic

$$\chi_c^2 = \sum \frac{(\text{observed} - \text{expected})^2}{\text{expected}}$$

Where

c = Degrees of freedom

O = Observed Value

E = Expected Value

Tests:

1. Goodness of fit
2. Independence

1. Independence ✓

The Chi-Square Test of Independence is a derivable (also known as inferential) statistical test which examines whether the two sets of variables are likely to be related with each other or not. This test is used when we have counts of values for two nominal or categorical variables and is considered as non-parametric test. A relatively large sample size and independence of observations are the required criteria for conducting this test.

For Example-

In a movie theatre, suppose we made a list of movie genres. Let us consider this as the first variable. The second variable is whether or not the people who came to watch those genres of movies have bought snacks at the theatre. Here the null hypothesis is that the genre of the film and whether people bought snacks or not are unrelatable. If this is true, the movie genres don't impact snack sales.

2. Goodness-Of-Fit ✓

In statistical hypothesis testing, the Chi-Square Goodness-of-Fit test determines whether a variable is likely to come from a given distribution or not. We must have a set of data values and the idea of the distribution of this data. We can use this test when we have value counts for categorical variables. This test demonstrates a way of deciding if the data values have a " good enough" fit for our idea or if it is a representative sample data of the entire population.

For Example-

Suppose we have bags of balls with five different colors in each bag. The given condition is that the bag should contain an equal number of balls of each color. The idea we would like to test here is that the proportions of the five colors of balls in each bag must be exact.

Goodness of fit: It's mainly used to test whether or not a sample comes from a given population or not.

eg: coin 1 — obs: H 48, T 52 — exp: 50, 50 — good fit for binomial (100, 0.5)

coin 2 — obs: 10, 90 — exp: 50, 50 — not a good fit for binomial (100, 0.5)

Uniform distribution H0: The given data follows a uniform distribution

obs: 1 2 3 4 5 6

exp: 1 2 3 4 5 6

obs:

eg. a dice

1	2	3	4	5	6
23	16	30	28	32	15

exp:

1	2	3	4	5	6
24	24	24	24	24	24

exp: $\frac{\text{sum}}{6} = \frac{144}{6} = 24$.

Degrees of freedom : n-1 [when using categorical data: n is no. of categories]

$\chi^2_5(0.05) = 11.07$

$\chi^2_{\text{(calculated)}} = \frac{(23-24)^2}{24} + \frac{(16-24)^2}{24} + \cdots + \frac{(15-24)^2}{24} = 10.916$

Excel example

Independence test - contingency table

var 1 → buying snacks → No / Yes

var 2 → movie genre → action / comedy / horror / thriller

	Yes	No
Act	20	21
Com	15	17
Hor	12	6
Thri	18	15

---

ice cream flavour & age-group are related

	5-10	10-15	15-20	20-25
Van				
Choc				
Straw				

Scientists want to test whether or not COVID-19 disease is affecting men & women equally.

var 1 → covid-19 + or −

var 2 → male / female.

observed

	male	female	total
+ve	25	27	52
−ve	108	96	204
total	133	123	256

Null hypothesis: the rows and columns are independent

eg. 1

Observed table:

	Republican	Democrat	Independent	Total
Male	100	70	30	200
Female	140	60	20	220
Total	240	130	50	~~440~~ 420

Expected values:

$$\text{Expected Value} = \frac{(\text{Row Total}) * (\text{Column Total})}{\text{Total Number Of Observations}}$$

EXPECTED	Republican	Democrat	Independent
Male	114.29	61.90	23.81
Female	125.71	68.10	26.19

$c = 2$
$\alpha = 5\%$

χ^2 calculated : 8.5027
χ^2_2 (tabulated) : 5.991

Since χ^2 calculated > χ^2 tabulated, calculated value lies in the rejection region hence we reject H_0 and conclude that the 2 variables are dependent.

eg.2

eg: Obs.

25	27	52
108	96	204
133	123	256

expected counts

$(52 \times 133)/256$	$(52 \times 123)/256$
$(204 \times 133)/256$	$(204 \times 123)/256$

O_i :

25	27
108	96

E_i :

27.01	24.98
105.98	98.01

$$\chi^2_c : \sum \frac{(O_i - E_i)^2}{E_i} = \frac{(25-27.01)^2}{27.01} + \frac{(27-24.98)^2}{24.98} + \frac{(108-105.98)^2}{105.98} + \frac{(96-98.01)^2}{98.01}$$

$= 0.1495 + 0.1633 + 0.038 + 0.0412$

$\chi^2_c = 0.392$

Since χ^2 value is low $\Rightarrow$ Observed values are close to

expected values ⇒ the actual split of data is very homogeneous and is as expected as if the variables were not effecting one another.

here $c = (2-1) \times (2-1) = 1$

Degrees of freedom = (rows -1) x (cols -1)

In χ^2_c ; $c \rightarrow$ degrees of freedom.

$c = (\text{rows} - 1) \times (\text{cols} - 1)$

Excel example

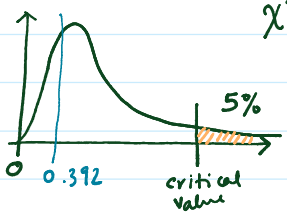
To test the null hypothesis, we see a tabulated χ^2_c value

eg. χ^2_1 (calculated) = 0.392

χ^2_1 (tabulated) (at $\alpha = 5\%$) =

χ^2_c (5%) we look in the table

rejection region



(3.841 for c = 1)

Since calculated χ^2_1 < tabulated χ^2_1 I accept the null hypothesis and state that my variables are independent of each other.