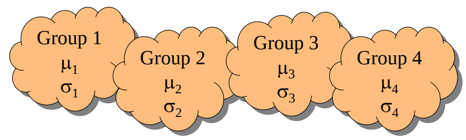
One-way ANOVA

First textbook presentation in 1925.

What Is ANOVA?

ANOVA = ANalysis Of VAriance

ANOVA compares the *means* of several groups.
The groups are sometimes called "treatments"

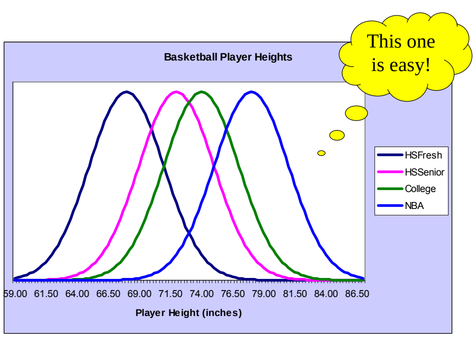


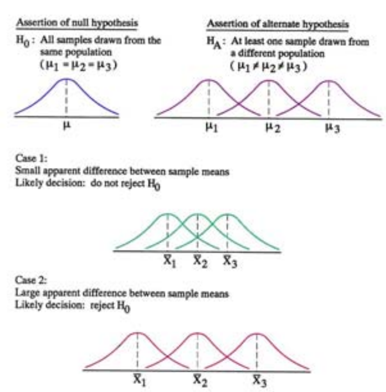
- A One-Way ANOVA and a Two-Way ANOVA were talking shop one day.
- The One-Way said, "I sure do envy the interaction you have with your variables."
- The Two-Way frowned and replied,
- "Yah, but the minute it diminishes to any significant extent they really become independent and go their own separate ways."

One way ANOVA (the F ratio test)

- it is the standard parametric test of difference between 3 or more samples,
- it is the parametric equivalent of the Kruskal-Wallis test
- $H_0 = \mu_1 = \mu_2 = \mu_3 = \mu_k$ the population means are all equal
- so that if **just one** is different, we will reject H_0

Comparing Groups





Null and Alternate Hypotheses in Analysis of Variance (ANOVA)

Why not multiple t-tests?

- If you have 5 groups you end up with 10 t-tests, too difficult to evaluate
- The greater the number of tests you make, the more likely you commit a type-I error (that is, you reject H_0 when you should accept it)
- There are methods to do pair wise tests that we'll discuss later

Some Terminology

- The ANOVA model specifies a single *dependent variable* (continuous)
- There will be one or more *explanatory factors* (categorical)
- Each factor has several levels (or groups) called *treatments*.

ANOVA reveals whether or not the mean depends on the treatment group from which an observation is taken.

- we expect that the sample means will be different, question is, are they significantly different from each other?
- H_0: differences are not significant - differences between sample means have been generated in a random sampling process
- H_1: differences are significant - sample means likely to have come from different population means

Assumptions

- 1) data must be at the interval/ratio level
- 2) sample data must be drawn from a normally distributed population
- 3) sample data must be drawn from independent random samples

---

- 4) populations have the same/similar variances - homoscedasticity assumption from which the samples are drawn
  - The dependent variable should have the same variance in each category of the independent variable.
    - The reason for this assumption is that the denominator of the F-ratio is the within-group mean square, which is the average of group variances taking group sizes into account. When groups differ widely in variances, this average is a poor summary measure. However, ANOVA is robust for small and even moderate departures from homogeneity of variance

One-Factor ANOVA

Model Form

$$X_{ij} = \mu + \tau_j + \varepsilon_{ij}$$

Hypotheses

H_0: $\tau_j = 0$ (no treatment effect exists)

H_1: $\tau_j \neq 0$ (treatment effect exists)

Definitions

X_{ij} = observation i in group j

μ = common mean

τ_j = effect due to treatment group j

ε_{ij} = random error

Note If H_0 is true, the model collapses to $X_{ij} = \mu + \varepsilon_{ij}$ which says that each observed data value is just the mean perturbed by some random error.

One Factor: DVD Price

General Form:	X = f(T)
Specific Form:	$X_{ij} = \mu + \tau_j + \varepsilon_{ij}$
Verbal Form:	Price = f(Store Type)

Music Store	Bookstore	Discount Store
18.95	14.95	11.50
14.95	15.95	12.50
15.95	21.95	9.50
11.00	13.75	11.75
17.00		13.75
14.50		
13.00		

This example shows that group sizes can be unequal.

X = price of a recent DVD (continuous variable)
T = store type (discrete - 3 treatment levels)

- ANOVA examines differences in means by looking at estimates of variances
- essentially ANOVA poses the following partition of a data value
- observation=overall mean + deviation of group mean from overall mean + deviation of observation from group mean
- the overall mean is a constant to all observations

- deviation of the group mean from the overall mean is taken to represent the effect of belonging to a particular group
- deviation from the group mean is taken to represent the effect of all other variables other than the group variable

Case example

Lower Variation, Higher Quality

Ford and Mazda were producing similar transmissions that were supposed to be made with the same specifications. But the American-made transmissions required more warranty repairs than the Japanese-made transmissions. When investigators inspected samples of the Japanese transmission gearboxes, they first thought that their measuring instruments were defective

because they weren't detecting any variability among the Mazda transmission gearboxes. They realized that although the American transmissions were within the specifications, the Mazda transmissions were not only within the specifications, but consistently close to the desired value. By reducing variability among transmission gearboxes, Mazda reduced the costs of inspection, scrap, rework, and warranty repair.

Example 1

- consider the numbers below as constituting **One** data set

	observations				$\bar{x}$
Sample 1	1		7	3	3.7
Sample 2	3	3	4		3.3
Sample 3	3	7	1	2	3.3

- Much of the variation is **within** each row, it is difficult to tell if the means are significantly different
- as stated before, the variability in all observations can be divided into 3 parts
- 1) variations due to differences within rows
- 2) variations due to differences between rows
- 3) variations due to sampling errors

Example 2

much of the variation is **between** each row, it is easy to tell if the means are significantly different [note: 1 way ANOVA does not need equal numbers of observations in each row]

	observations				$\bar{x}$
Sample 1	1	1	1	2	1.25
Sample 2	3	3	4		3.33
Sample 3	7	7	8	7	7.25

- in example 1: between row variation is small compared to row variation, therefore F will be small
- large values of F indicate difference
- conclude there is no significant difference between means
- in example 2: between row variation is large compared to within row variation, therefore, F will be large
- conclude there is a significant difference

Sample 1 vs Sample 2

	obs				$\bar{x}$
S_1	1		7	3	3.7
S_2	3	3	4		3.3
S_3	3	7	1	2	3.3

	obs				$\bar{x}$
S_1	1	1	1	2	1.25
S_2	3	3	4		3.33
S_3	7	7	8	7	7.25

- the first step in ANOVA is to make two estimates of the variance of the hypothesized common population
  - 1) the within samples variance estimate
  - 2) the between samples variance estimate

within sample variance estimate

$$\sigma_w^2 = \frac{\sum_{i=1}^{k}\sum_{j=1}^{n}(x-\bar{x})^2}{N-k}$$

k - is the number of sample
N - total number of individuals in all samples
X bar - is the mean

Between sample variance

$$\sigma_b^2 = \frac{\sum_{i=1}^{k} n(\bar{x}-\bar{x}_G)^2}{k-1}$$

$\bar{x}_G$ is the grand mean

F ratio

- having calculated 2 estimates of the population variance how probable is it that 2 values are estimates of the same population variance
- to answer this we use the statistic known as the F ratio

$$F\ ratio = \frac{between\ row\ variation}{within\ row\ variation}$$

$$F\ ratio = \frac{estimate\ of\ \ variance\ between\ samples}{estimate\ of\ \ variance\ within\ samples}$$

- significance: critical values are available from tables
- df are calculated as
- 1) for the between sample variance estimate they are the number of sample means minus 1 (k-1)
- 2) for the within sample variance they are the total number of individuals in the data minus the number of samples (N-k)
- since the calculations are somewhat complicated it should be done in a table

Example Test winning times for the men's Olympic 100 meter dash over several time periods

	winning time (in seconds)				X_k
1900-1912	10.8	11	10.8	10.8	10.85
1920-1932	10.8	10.6	10.8	10.3	10.625
1936-1956	10.3	10.3	10.4	10.5	10.375

- source: Chatterjee, S. and Chatterjee, S. (1982) 'New lamps for old: an exploratory analysis of running times in Olympic Games',
- Applied statistics, 31, 14-22.

Hypotheses

- H_0: There is no significant difference in winning times. The difference in means have been generated in a random sampling process
- H_1: There are significant differences in winning times. Given observed differences in sample means, it is likely they have been drawn from different populations.
- Confidence at p=0.01, 99% confident from different population

1900-1912	1920-1932	1936-1956	
10.8	10.8	10.3	
11	10.6	10.3	
10.8	10.8	10.4	
10.8	10.3	10.5	X_G=0.62
∑x=43.4	∑x=42.5	∑x=41.5	
n=4	n=4	n=4	
$\bar{X}$ =10.85	$\bar{X}$=10.625	$\bar{X}$=10.375	

1900-1912	1920-1932	1936-1956	1900-1912	1920-1932	1936-1956
$(x-\bar{x})$	$(x-\bar{x})^2$	$(x-\bar{x})$	$(x-\bar{x})^2$	$(x-\bar{x})$	$(x-\bar{x})^2$
-.05	.175	-.075	.0025	.0306	.0056
.15	-.025	-.075	.0225	.0006	.0056
-.05	.175	.025	.0025	.0306	.0006
-.05	-.325	.125	.0025	.1056	.0156
$\Sigma(x-\bar{x})^2$=.03		$\sum(x-\bar{x})^2$ =.1674		$\sum(x-\bar{x})^2$=.0274	

$$\sigma_w^2 = \frac{\sum_{i=1}^{k}\sum_{j=1}^{n}(x-\bar{x})^2}{N-k} = \frac{0.2248}{12-3} = 0.025$$

calculation of between samples variance estimate

$$\sigma_B^2 = \frac{\sum n(\bar{x}-\bar{x}_G)^2}{k-1} = \frac{0.2116+0+0.24}{3-1} = 0.2258$$

1900-1912	X=10.85	n=4	$n(X-X_G)$	$4(10.85-10.62)^2$	4(.0529)	.2116
1920-1932	X=10.625	n=4	$n(X-X_G)$	$4(10.625-10.62)^2$	4(.0000)	0
1936-1956	X=10.375	n=4	$n(X-X_G)$	$4(10.375-10.62)^2$	4(.0600)	.24

	variance estimate	df
between samples	.2258	2 (k-1)
within samples	.025	9 (N-k)

$$F\ ratio = \frac{estimate\ of\ \ variance\ between\ samples}{estimate\ of\ \ variance\ within\ samples} = \frac{0.2258}{0.25} = 9.032$$

critical value=8.02, page 277 in text, α=.01
therefore we reject the null hypothesis

- one problem in using these formulas, however, is that if the means are approximated, the multiple subtractions compound rounding errors. To get around this, the formulas can rewritten as:

$$T = \sum_{i=1}^{r}\sum_{j=1}^{k} x_{ij}$$

$$SST = (\sum_{i=1}^{r}\sum_{j=1}^{k} x_{ij}^2) - \frac{1}{n}T^2$$

N=total number of observations
T is the total sum of observations

- It can be shown that:

$$\sum_{i=1}^{k}\sum_{j=1}^{n_i}(x_{ij}-\bar{\bar{x}})^2 = \sum_{i=1}^{k} n_i(\bar{x}_i-\bar{\bar{x}})^2 + \sum_{i=1}^{k}\sum_{j=1}^{n_i}(x_{ij}-\bar{x}_i)^2$$

Total sum of squares =
Between row sum of squares
+ within row sum of squares

- SST = SSR + SSE

1900-1912	X²	1920-1932	X²	1936-1956	X²	
10.8	116.64	10.8	116.64	10.3	106.09	
11	121	10.6	112.36	10.3	106.09	
10.8	116.64	10.8	116.64	10.4	108.16	
10.8	116.64	10.3	106.09	10.5	110.25	
Totals	470.92		451.73		430.59	1353.24

- Where:

$$SST = \sum (x_{ij}^2) - \frac{1}{N}T^2 \qquad SSR = \sum \frac{T_i^2}{n_i} - \frac{1}{N}T^2$$

T = sum of all observations

T_i is the sum of all the observations in a row

$$SSR = [\sum_{i=1}^{k} \frac{T_i^2}{n_i}] - \frac{1}{N}T^2 = [\frac{43.4^2}{4} + \frac{42.5^2}{4} + \frac{41.5^2}{4}] - [\frac{1}{12}(127.4^2] = 0.45$$

SSE=SST-SSR

$$SST = 1353.64 - \frac{127.4^2}{12} = 0.68$$

SSE=.68 - .45 = 0.23

$$MSR = \frac{SSR}{k-1} = \frac{0.45}{3-1} = 0.225$$

$$MSE = \frac{SSE}{N-k} = \frac{0.23}{12-3} = 0.0255$$

$$F = \frac{MSR}{MSE} = \frac{0.225}{0.0255} = 8.82$$

$$df_1 = k-1 = 2 \quad df_2 = N-k = 12-3 = 9$$

Source of variation	df	sum of squares	mean square	F-statistic
between rows/groups /samples	k-1 (2)	SSR (0.45)	SSR/k-1 (0.225)	MSR/MSE (8.82)
within rows/groups /samples	N-k (9)	SSE (0.23)	SSE/N-k (0.0255)	
Total	N-1	0.68		

- F distribution is One tail only (no such thing as a two tailed test), it doesn't make sense to specify direction with k samples
- computed value of F (8.83) > critical value
- df_1=2, df_2=9, α=0.05, critical value=4.26
- df_1=2, df_2=9, α=0.01, critical value=8.02
- We can be > 99% certain the differences between the time periods are significant and that the observations in each time period are drawn from distributions with different population means

F Distribution (α = 0.05 in the right tail)

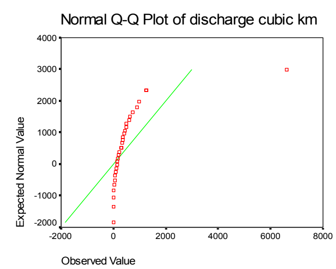
Numerator degrees of freedom

Denominator degrees of freedom	1	2	3	4	5	6	7	8	9
1	161.45	199.50	215.71	224.58	230.16	233.99	236.77	238.88	240.54
2	18.513	19.000	19.164	19.247	19.296	19.330	19.353	19.371	19.385
3	10.128	9.5521	9.2766	9.1172	9.0135	8.9406	8.8867	8.8452	8.8123
4	7.7086	6.9443	6.5914	6.3882	6.2561	6.1631	6.0942	6.0410	5.9988
5	6.6079	5.7861	5.4095	5.1922	5.0503	4.9503	4.8759	4.8183	4.7725
6	5.9874	5.1433	4.7571	4.5337	4.3874	4.2839	4.2067	4.1468	4.0990
7	5.5914	4.7374	4.3468	4.1203	3.9715	3.8660	3.7870	3.7257	3.6767
8	5.3177	4.4590	4.0662	3.8379	3.6875	3.5806	3.5005	3.4381	3.3881
9	5.1174	$\boxed{4.2565}$	3.8625	3.6331	3.4817	3.3738	3.2927	3.2296	3.1789
10	4.9646	4.1028	3.7083	3.4780	3.3258	3.2172	3.1355	3.0717	3.0204
11	4.8443	3.9823	3.5874	3.3567	3.2039	3.0946	3.0123	2.9480	2.8962
12	4.7472	3.8853	3.4903	3.2592	3.1059	2.9961	2.9134	2.8486	2.7964
13	4.6672	3.8056	3.4105	3.1791	3.0254	2.9153	2.8321	2.7669	2.7144
14	4.6001	3.7389	3.3439	3.1122	2.9582	2.8477	2.7642	2.6987	2.6458
15	4.5431	3.6823	3.2874	3.0556	2.9013	2.7905	2.7066	2.6408	2.5876
16	4.4940	3.6337	3.2389	3.0069	2.8524	2.7413	2.6572	2.5911	2.5377
17	4.4513	3.5915	3.1968	2.9647	2.8100	2.6987	2.6143	2.5480	2.4943
18	4.4139	3.5546	3.1599	2.9277	2.7729	2.6613	2.5767	2.5102	2.4563
19	4.3807	3.5219	3.1274	2.8951	2.7401	2.6283	2.5435	2.4768	2.4227
20	4.3512	3.4928	3.0984	2.8661	2.7109	2.5990	2.5140	2.4471	2.3928
21	4.3248	3.4668	3.0725	2.8401	2.6848	2.5727	2.4876	2.4205	2.3660
22	4.3009	3.4434	3.0491	2.8167	2.6613	2.5491	2.4638	2.3965	2.3419
23	4.2793	3.4221	3.0280	2.7955	2.6400	2.5277	2.4422	2.3748	2.3201
24	4.2597	3.4028	3.0088	2.7763	2.6207	2.5082	2.4226	2.3551	2.3002
25	4.2417	3.3852	2.9912	2.7587	2.6030	2.4904	2.4047	2.3371	2.2821
26	4.2252	3.3690	2.9752	2.7426	2.5868	2.4741	2.3883	2.3205	2.2655
27	4.2100	3.3541	2.9604	2.7278	2.5719	2.4591	2.3732	2.3053	2.2501
27	4.1960	3.3404	2.9467	2.7141	2.5581	2.4453	2.3593	2.2913	2.2360
29	4.1830	3.3277	2.9340	2.7014	2.5454	2.4324	2.3463	2.2783	2.2229
30	4.1709	3.3158	2.9223	2.6896	2.5336	2.4205	2.3343	2.2662	2.2107
40	4.0847	3.2317	2.8387	2.6060	2.4495	2.3359	2.2490	2.1802	2.1240
60	4.0012	3.1504	2.7581	2.5252	2.3683	2.2541	2.1665	2.0970	2.0401
120	3.9201	3.0718	2.6802	2.4472	2.2899	2.1750	0.2087	2.0164	1.9588
∞	3.8415	2.9957	2.6049	2.3719	2.2141	2.0986	2.0096	1.9384	1.8799

Violations

- Lack of independence
  - Example is time series
- Outliers
  - Outliers tend to increase the estimate of sample variance, thus decreasing the calculated F statistic for the ANOVA and lowering the chance of rejecting the null hypothesis

- Nonnormality
  - Do a histogram to check
  - If sample size is small it may be difficult to detect
  - If the samples are not seriously imbalanced in size skewness won't have much impact
  - Do a normal Q-Q plot or normal quantile-quantile plot, it's a plot of the ordered data values (as Y) against the associated quantiles of the normal distribution (as X)



Pairwise comparisons

- There are procedures available for doing comparisons between the means of the classes in the anova
- Some of those available in SPSS
  - Scheffe's test
  - Bonferrani's test
  - Tukey-Kramer
    - → best for ANOVA of unequal sample sizes
  - Tukey test → best for balanced designs