Curtin College

DIPLOMA OF INFORMATION TECHNOLOGY

RNI1006 REGRESSION AND NONPARAMETRIC INFERENCE

Curtin College

in association with

Curtin University

Acknowledgement

We respectfully acknowledge the Elders and custodians of the Whadjuk Nyungar nation, past and present, their descendants and kin. Curtin College Bentley Campus enjoys the privilege of being located in Whadjuk / Nyungar Boodjar (country) on the site where the Derbal Yerrigan (Swan River) and the Djarlgarra (Canning River) meet. The area is of great cultural significance and sustains the life and well being of the traditional custodians past and present.

Aims of this lecture

1. One-factor analysis of variance (ANOVA) – Parametric

(Moore et al 2021 Chapter 12)

1.1 The One Way ANOVA model (model; parameters; hypothesis testing; assumptions)

1.2 The ANOVA Table (F test; partition of total variations)

2. **Kruskal-Wallis rank test for differences in *c* (or *K*) medians – Nonparametric**

(Moore et al 2021 Chapter 15)

Aim 1 ANOVA
General Experimental Setting

- Investigator controls one or more independent variables
  - Called treatment variables or factors
  - Each treatment factor contains two or more levels (or categories/classifications)
- Observe effects on dependent variable
  - Response to levels of independent variable
- Experimental design: the plan used to test hypothesis

Completely Randomized Design

- Experimental units (subjects) are assigned randomly to treatments
  - Subjects are assumed homogeneous
- Only one factor or independent variable
  - With two or more treatment levels
- Analyzed by
  - One-factor analysis of variance (one-way ANOVA)

	Factor (Training Method)		
Factor Levels (Treatments)			
Randomly Assigned Units			
Dependent Variable (Response)	21 hrs	17 hrs	31 hrs
	27 hrs	25 hrs	28 hrs
	29 hrs	20 hrs	22 hrs

The Idea of ANOVA

- When comparing different populations or treatments, the data are subject to **sampling variability.** We can pose the question for inference in terms of the *mean* response.
- The two-sample *t* procedures compared the means of two populations or the mean responses to two treatments in an experiment.
- We now extend those methods to problems involving more than two populations using **analysis of variance (ANOVA).**

One-way ANOVA is used for situations in which there is only one factor or only one way to classify the populations of interest.

Example: To compare tread lifetimes of five brands of tyre, we use one-way ANOVA with tyre brand as our factor.

Two-way ANOVA is used to analyze the effect of two factors.

Example: The tire researcher also may want to consider the temperature of the environment as a second factor.

Note: We are comparing ***means***, even though the procedure is called analysis of *variance.*

Comparing Means 1

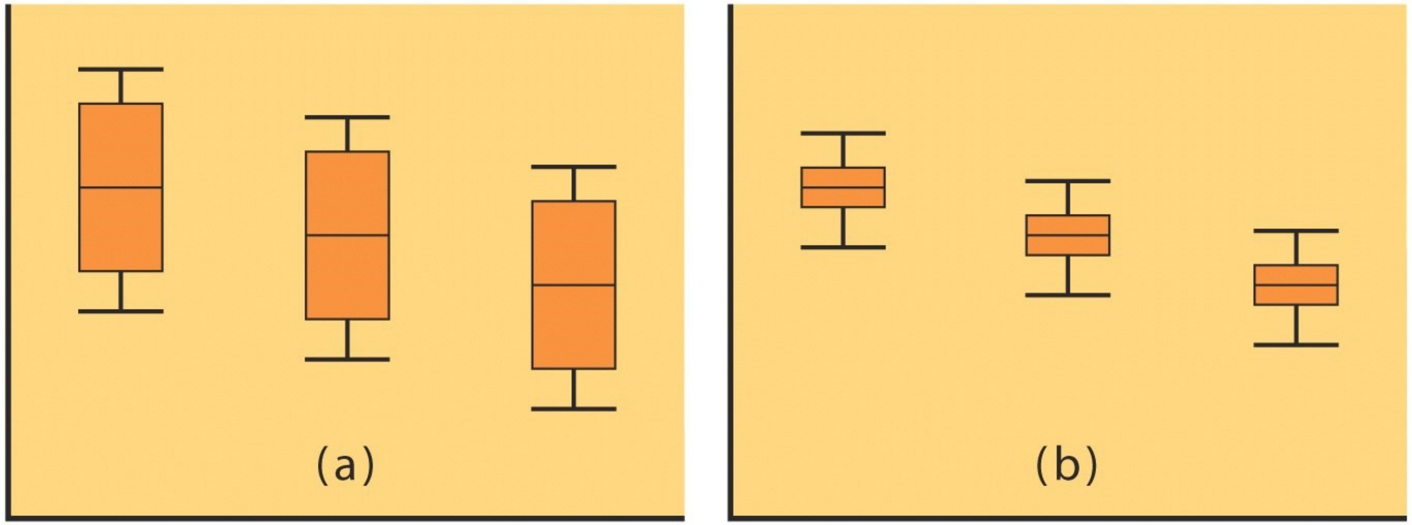
The details of ANOVA are a bit daunting. The main idea is that, when we ask if a set of means gives evidence for differences among the population means, what matters is not how far apart the sample means are but how far apart they are *relative to the variability of individual observations*.

The Analysis of Variance Idea

Analysis of variance compares the variation due to specific sources with the variation among individuals who should be similar. In particular, ANOVA tests whether several populations have the same means by comparing *how far apart the sample means are* with *how much variation there is within a sample*.

The ANOVA question: Do all groups have the same population mean? The purpose of ANOVA is to assess whether the observed differences among sample means are *statistically significant*.

Comparing Means 2



- ✓ The sample means for the three samples are the same for each set.
- ✓ The variation *among sample means* for (a) is identical to that for (b).
- ✓ The variation *among the individuals within* each of the three samples is much less for (b).

CONCLUSION: The samples in (b) contain a larger amount of variation among the sample means *relative to* the amount of variation within the samples, so ANOVA will find *more significant differences among the means in (b)*

- assuming equal sample sizes here for (a) and (b).
- **Note:** Larger samples will find more significant differences.

Example 1: Facebook Profiles
An Overview of ANOVA

Number of Friends	*n*	Mean	*s*
102	24	3.82	1.00
302	33	4.88	0.85
502	26	4.56	1.07
702	30	4.41	1.43
902	21	3.99	1.02

Offline, having more friends is associated with higher ratings of likability and trustworthiness.

134 students were randomly assigned to observe one of five Facebook profiles, which were identical except for the number of friends.

Response variable: attractiveness score

Groups: Which profile was viewed: 102 friends, 302 friends, etc. K=5

Comparing Several Means

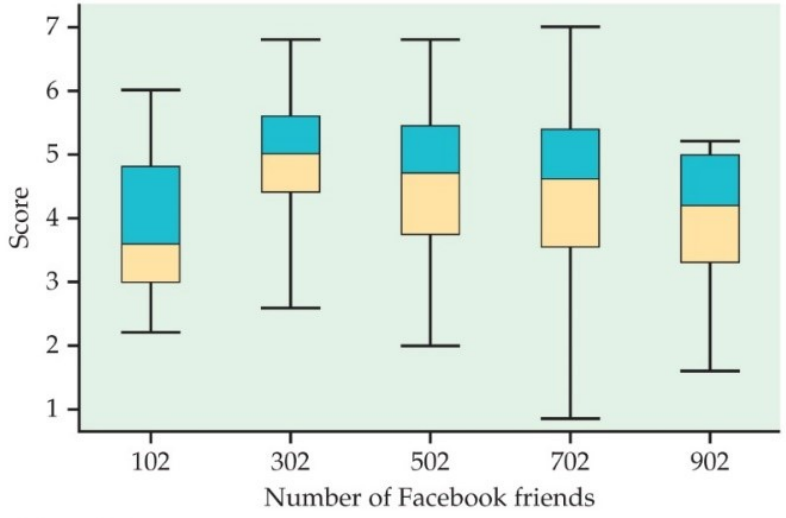


Figure 12.4
Moore/McCabe/Craig, *Introduction to the Practice of Statistics*, 9e, © 2017 W. H. Freeman and Company

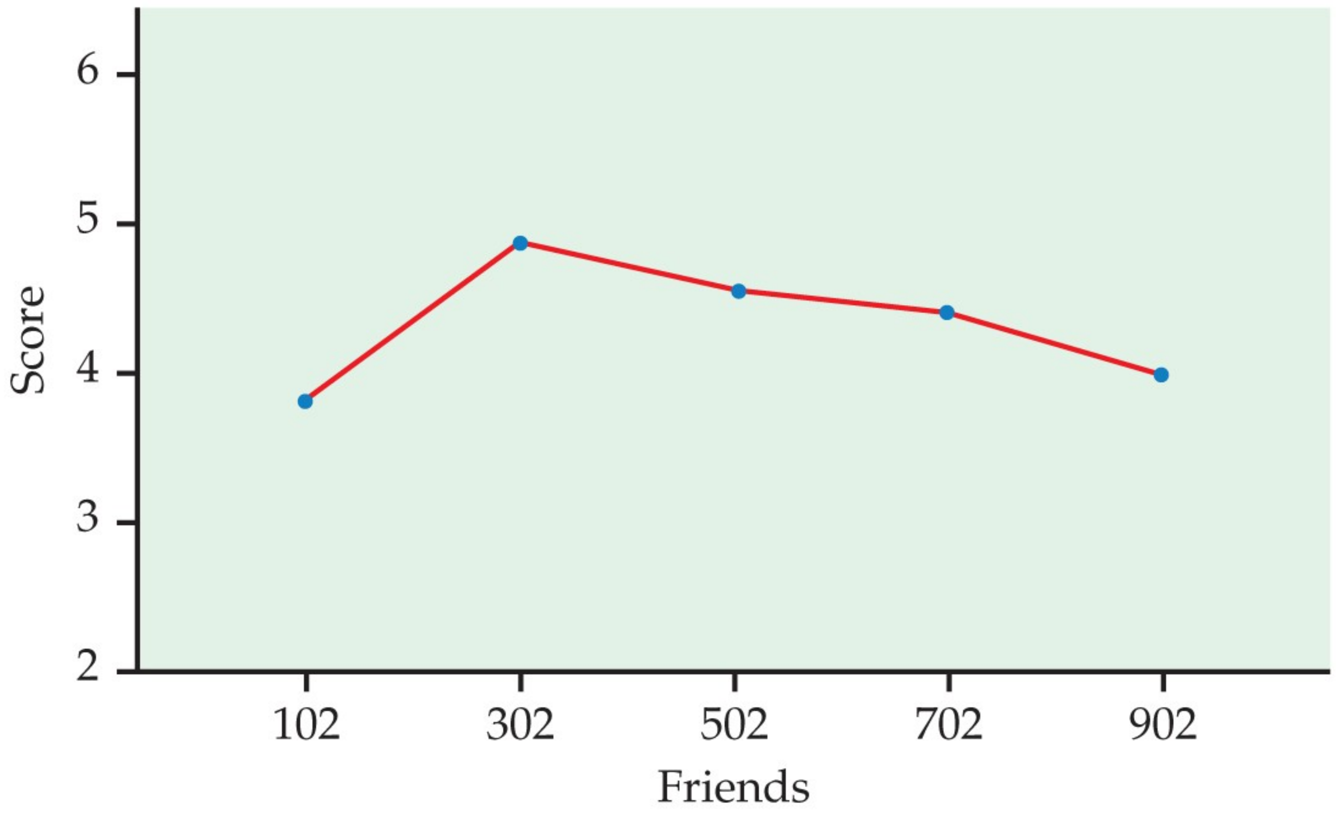


Figure 12.5
Moore/McCabe/Craig, *Introduction to the Practice of Statistics*, 9e, © 2017 W. H. Freeman and Company

- ✓ Are the corresponding population means different or can we conclude that the variability we see is random chance?
- ✓ Do the means follow a curvilinear relationship?
- ✓ Which means differ?

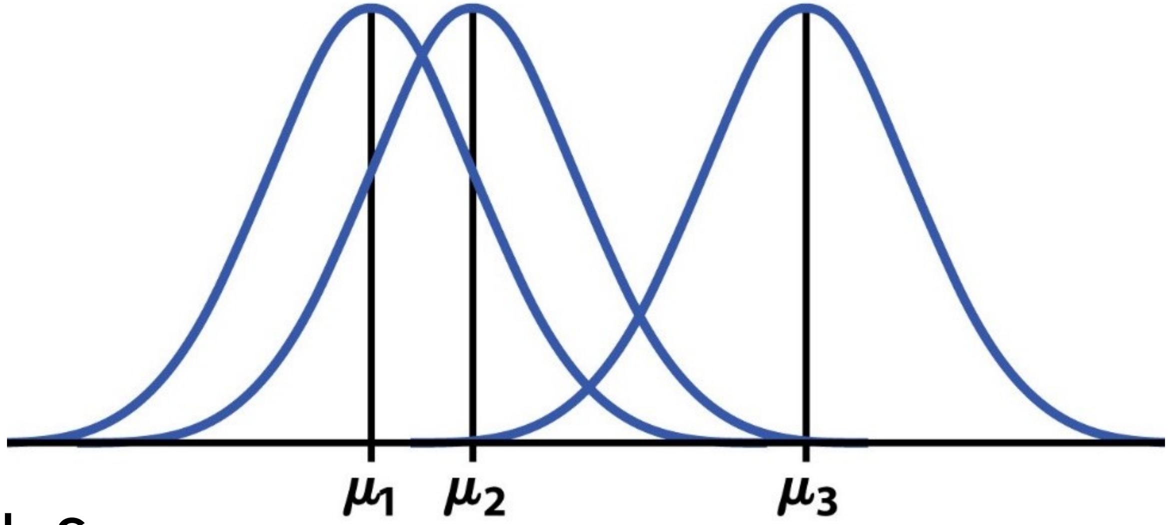
Aim 1.1 The One-Way ANOVA Model

- Random sampling always produces chance variations. Any "factor effect" would thus show up in our data as the factor-driven differences plus chance variations ("error"):

Data = factor effect + error

or **Data = FIT + RESIDUAL**

- The one-way ANOVA model analyzes data x_{ij} where chance variations are Normally distributed $N(0,\sigma)$:



$$x_{ij} = \mu_i + \varepsilon_{ij}$$

for $i = 1,\ldots,K$ and $j = 1,\ldots,n_i$.

- The ε_{ij} are assumed to be from a $N(0,\sigma)$ distribution. The **parameters of the model** are the population means $\mu_1,\mu_2,\ldots,\mu_K$ and the common standard deviation σ.

Estimates of the Population Parameters

- The unknown parameters in the model are the K population means μ_i and the common population standard deviation σ.
- To estimate μ_i, we use the sample mean:

$$\bar{x}_i = \frac{1}{n_i}\sum_{j=1}^{n_i} x_{ij}$$

- To estimate σ, we use the pooled standard deviation:

$$s_p = \sqrt{\frac{(n_1-1)s_1^2 + (n_2-1)s_2^2 + \cdots + (n_K-1)s_K^2}{(n_1-1) + (n_2-1) + \cdots + (n_K-1)}}$$

Testing Hypotheses in One-Way ANOVA

- We want to test the null hypothesis that there are *no differences* among the means of the populations:

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

- The basic conditions for inference are that we have a random sample from each population and that each population is Normally distributed.

- The alternative hypothesis is that there is *some difference*. That is, not all means are equal. This hypothesis is not one-sided or two-sided. It is "many-sided."

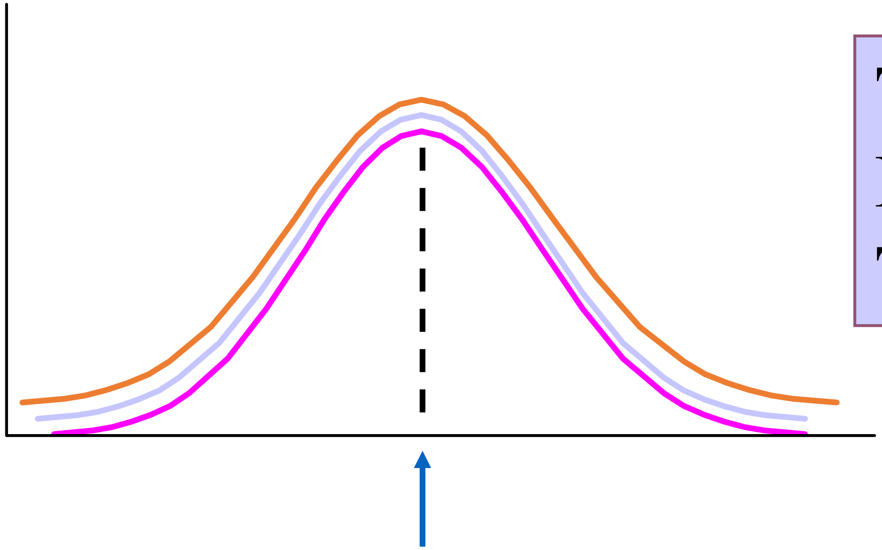
$$H_1\text{: not all of the means are equal}$$

- This test is called the **analysis of variance *F* test (ANOVA).**

One-Way ANOVA : Under Ho
No Treatment Effect

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

$$H_1: \text{ Not all } \mu_i \text{ are the same}$$



The Null Hypothesis is True

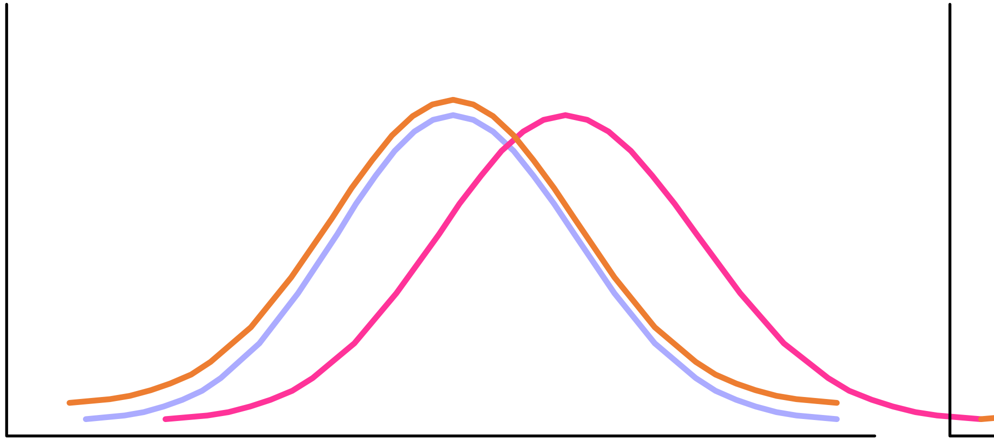
$$\mu_1 = \mu_2 = \mu_3$$

One-Way ANOVA : Under Ha (Treatment Effect Present)

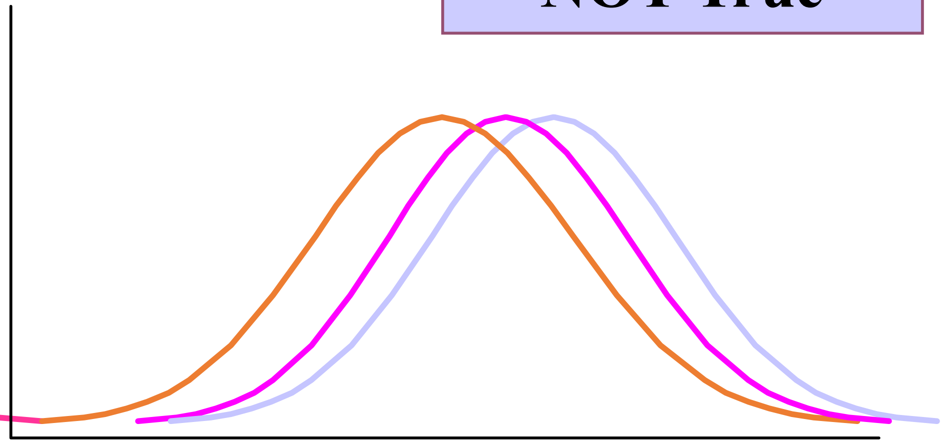
$$H_0: \mu_1 = \mu_2 = \cdots = \mu_K$$

$$H_1: \text{ Not all } \mu_i \text{ are the same}$$

The Null Hypothesis is NOT True



$\mu_1 = \mu_2 \neq \mu_3$



$\mu_1 \neq \mu_2 \neq \mu_3$

Aim 1.2 One-Way ANOVA: Partition of Total Variation

Total Variation SST

= **Variation Due to Treatment/Group SSG** + **Variation Due to Random Sampling SSE**

Commonly referred to as:

- Sum of Squares Among
- Sum of Squares Between
- Sum of Squares Model
- Sum of Squares Explained
- Sum of Squares Treatment
- Among Groups Variation

Commonly referred to as:

- Sum of Squares Within
- Sum of Squares Error
- Sum of Squares Unexplained
- Within Groups Variation

Total Variation, SST=Sum of Squares Total

$$SST = \sum_{i=1}^{K}\sum_{j=1}^{n_i}(X_{ij} - \bar{\bar{X}})^2$$

$$SST = (X_{11} - \bar{\bar{X}})^2 + \cdots + (X_{1n_1} - \bar{\bar{X}})^2 + \ldots + (X_{K1} - \bar{\bar{X}})^2 + \ldots + (X_{Kn_K} - \bar{\bar{X}})^2$$

where

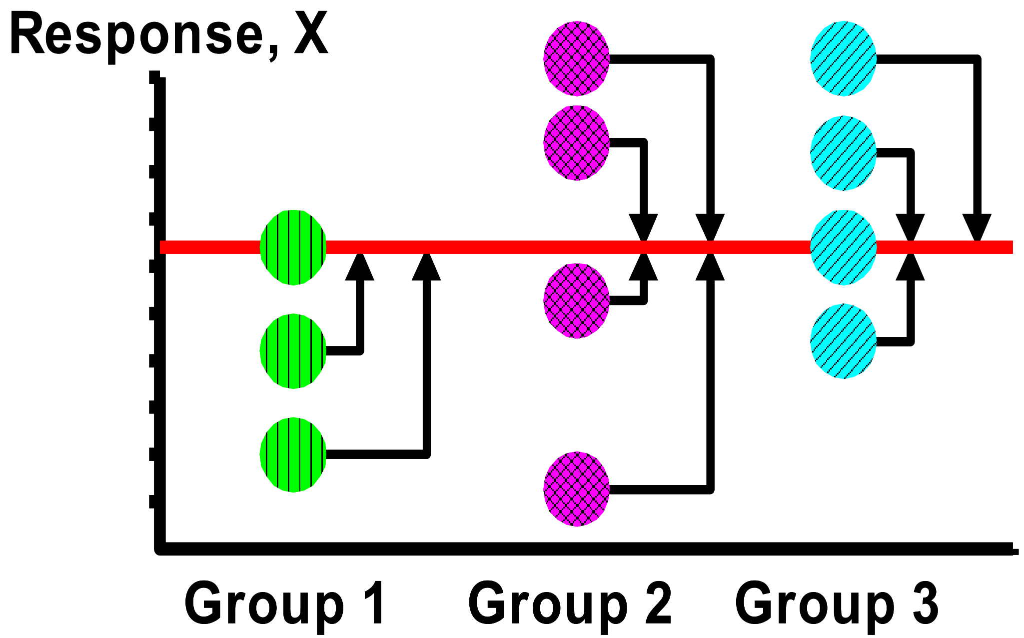
X_{ij} : the j-th observation in group-i

n_i: the number of observations in group-i

N: the total number of observation in all groups

K: number of groups

$\bar{\bar{X}} = \frac{\sum_{i=1}^{K}\sum_{j=1}^{n_i} X_{ij}}{n}$, the overall mean or grand mean



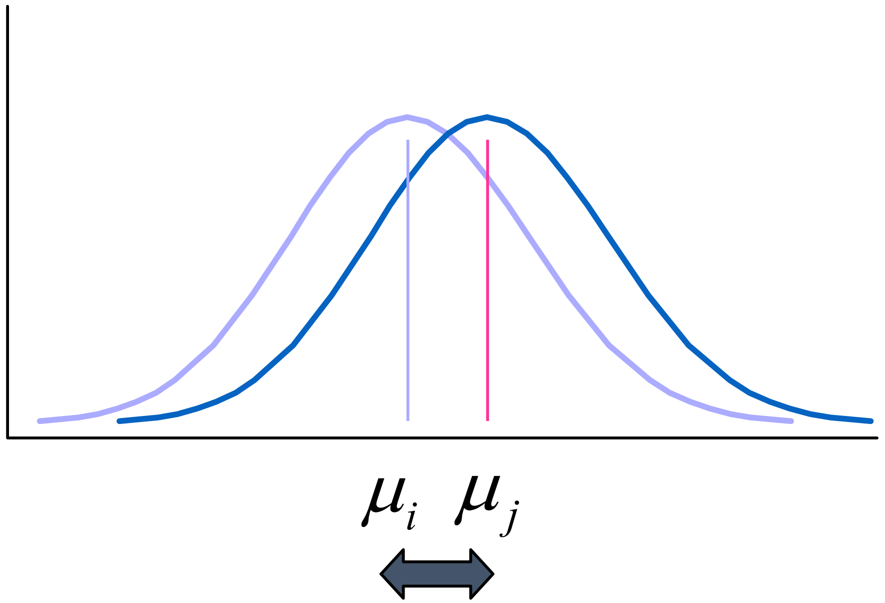
Among-Group Variation (SSG)

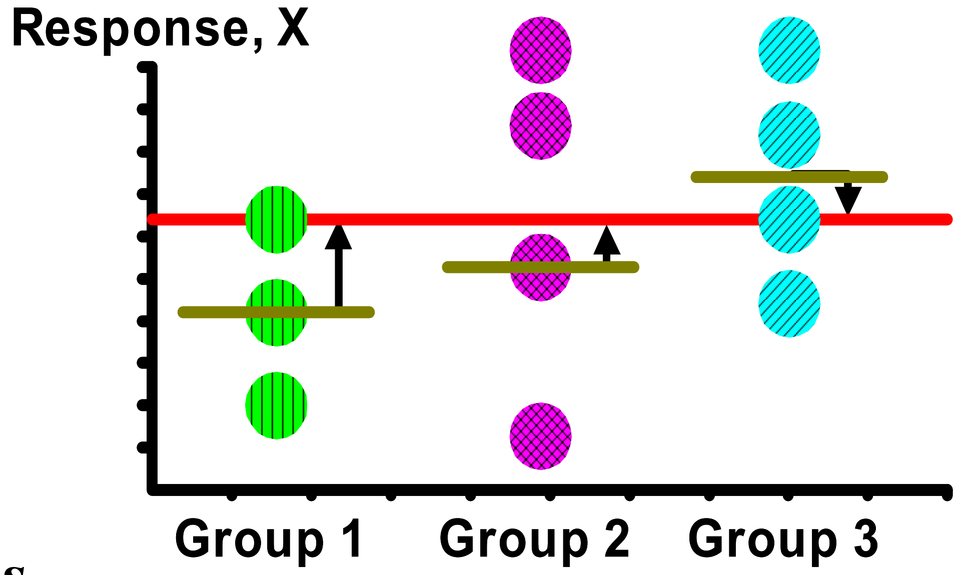
$$SSG = \sum_{i=1}^{K} n_i(\bar{X}_i - \bar{\bar{X}})^2$$

$\bar{X}_i$ *: The sample mean of group-i*

$\bar{\bar{X}}$: The overall or grand mean

$$SSG = n_1(\bar{X}_1 - \bar{\bar{X}})^2 + \ldots + n_K(\bar{X}_K - \bar{\bar{X}})^2$$





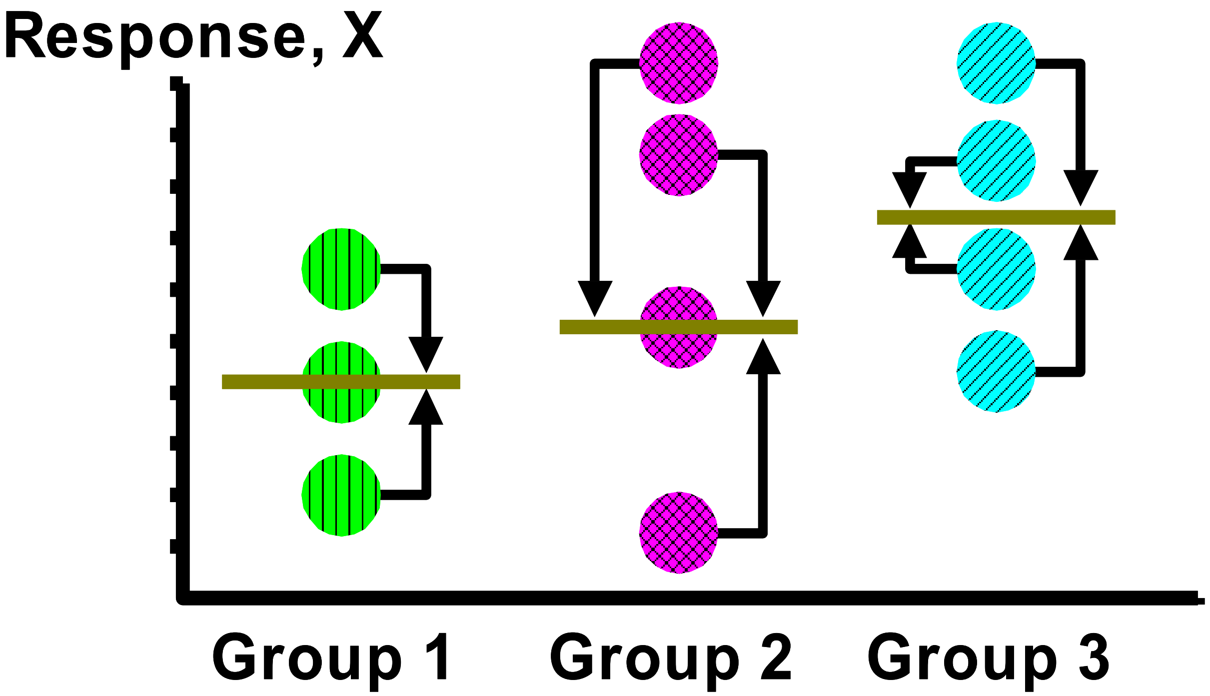
Variation Due to Differences Among Groups.

Within-Group Variation (SSE)

$\bar{X}_i$: *The sample mean of group-i*

$\bar{\bar{X}}$: The overall or grand mean

$$SSE = \sum_{i=1}^{K}\sum_{j=1}^{n_i}(X_{ij} - \bar{X}_i\ )^2$$



$SSE =$
$(X_{11} - \bar{X}_1\ )^2 + \ldots + (X_{1n_1} - \bar{X}_1\ )^2 + \ldots +$
$(X_{K1} - \bar{X}_K\ )^2 + \ldots + (X_{Kn_K} - \bar{X}_K\ )^2$

Summing the variation within each group and then adding over all groups.

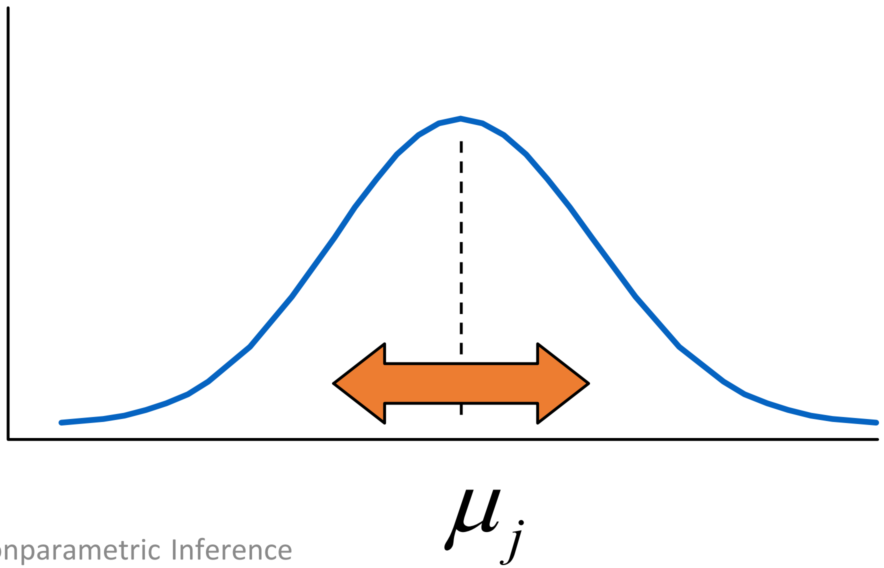
$$MSE = \frac{SSE}{N-K}$$

Within-Group Variation *(continued)*

$$\text{M}SE = \frac{SSE}{N-K}$$
$$= \frac{(n_1-1)S_1^2+(n_2-1)S_2^2+\cdots+(n_K-1)S_K^2}{(n_1-1)+(n_2-1)+\cdots+(n_K-1)}$$

For K=2, this is the pooled-variance in the t-Test.

- ***If more than two groups, use F Test.***
- ***For two groups, use t-Test. F Test more limited.***



The ANOVA *F* Test

$$F = \frac{\text{variation among the sample means}}{\text{variation among individuals in the same sample}}$$

The measures of variation in the numerator and denominator are ***mean squares:***

- Numerator: **Mean Square for Groups** (MSG)

$$MSG = \frac{n_1(\bar{x}_1 - \bar{x})^2 + n_2(\bar{x}_2 - \bar{x})^2 + \cdots + n_K(\bar{x}_K - \bar{x})^2}{K - 1}$$

- Denominator: **Mean Square for Error** (MSE)

$$MSE = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2 + \cdots + (n_K - 1)s_K^2}{N - K}$$

  - MSE is also called the **pooled sample variance**, written as s_p^2 (s_p is the ***pooled standard deviation***).
  - s_p^2 estimates the common variance σ^2.

F Distributions

The *F* distributions are a family of right-skewed distributions that take only values greater than 0. A specific *F* distribution is determined by the degrees of freedom of the numerator and denominator of the *F* statistic.

When describing an *F* distribution, always give the numerator degrees of freedom first. Our brief notation will be *F*(df1, df2) with df1 degrees of freedom in the numerator and df2 degrees of freedom in the denominator.

Degrees of Freedom for the *F* Test

We want to compare the means of K populations. We have an SRS of size n_i from the i^{th} population, so that the total number of observations in all samples combined is

$$N = n_1 + n_2 + \cdots + n_K$$

If the null hypothesis that all population means are equal is true, the ANOVA *F* statistic has the *F* distribution with $K - 1$ degrees of freedom in the numerator and $N - K$ degrees of freedom in the denominator.

The ANOVA Table

Source of variation	Sum of squares SS	DF	Mean square MS	*F*	*P*-value	*F* crit
Among or between "groups"	$\sum n_i(\bar{x}_i - \bar{x})^2$	*K* -1	MSG= SSG/DFG	MSG/MSE	Tail area above *F*	Value of *F* for α
Within groups or "error"	$\sum (n_i - 1)\, s_i^2$	*N* - *K*	MSE= SSE/DFE			
Total	SST=SSG+SSE $\sum \left(x_{ij} - \bar{x}\right)^2$	*N* – 1				

R^2 = SSG/SST
Coefficient of determination

$\sqrt{\text{MSE}} = s_p$
Pooled standard deviation

The sums of squares represent different sources of variation in the data:

SST = SSG + SSE

The degrees of freedom mirror the sums of squares:

DFT = DFG + DFE

Data ("Total") = Factor effect ("Groups") + Error ("Error")

Conditions for ANOVA

Like all inference procedures, ANOVA is valid only in some circumstances. The conditions under which we can use ANOVA are:

Conditions for ANOVA Inference

- We have ***K*** **independent SRSs**, one from each population. We measure the same response variable for each sample.
- The *i*th population has a **Normal distribution** with unknown mean μ_i.
- All the populations have the **same standard deviation** σ, whose value is unknown. Use Levene Test to check the equal variances or use Welch ANOVA test for unequal variances instead of F test.

Checking Standard Deviations in ANOVA

- The results of the ANOVA *F* test are approximately correct when the largest sample standard deviation is no more than twice as large as the smallest sample standard deviation.

6 Steps ANOVA Hypothesis Testing

STEP 1 Hypotheses Testing

$$H_0: \mu_1 = \mu_2 = \cdots = \mu_K \qquad H_1: \text{ Not all } \mu_i \text{ are the same}$$

STEP 2 Test Statistic Fo= MSG/MSE (see ANOVA Table)

STEP 3 The sampling distribution Fo distributed as F with df= (K-1, N-K)

STEP 4 p-value = Prob(F(K-1, N-K) > Fo)

STEPS 5 and 6. Decision and Conclusion

ANOVA *F* Test: Example 1 Facebook (continued)

IBM SPSS Statistics Viewer

ANOVA

Score

	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	19.890	4	4.973	4.142	.003
Within Groups	154.867	129	1.201		
Total	174.757	133			

Figure 12.8

STEP 1 Hypotheses Testing, i= 1,2,…, K; K=5; N=134

$H_0\colon \mu_1 = \mu_2 = \cdots = \mu_K$ $\quad H_1\colon$ Not all μ_i are the same

STEP 2 Test Statistic

Fo= MSG/MSE= 4.973/1.201= 4.142 (see ANOVA Table)

STEP 3 The sampling distribution Fo ~ F(4, 129)

Df1=K-1= 5-1=4 ; Df2= N-K = 134-5=129

STEP 4 p-value = (F(4, 129) > 4.142)= 0.003

STEP 5 Decision. As the p-value=0.3% < 5%, then we reject Ho.

STEP 6 Conclusion. There is significant evidence that attractiveness ratings depend on the number of friends reported on a Facebook profile.

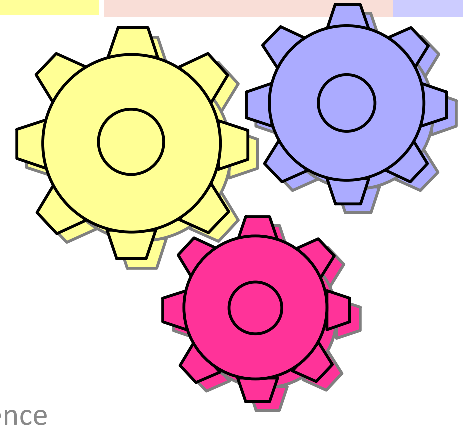
One-Way ANOVA *F* Test: Example 2

- As production manager, you want to see if three filling machines have different mean filling times.

- You assign 15 similarly trained and experienced workers, five per machine, to the machines.

- At the .05 significance level, is there a difference in mean filling times?

Machine1	Machine2	Machine3
25.40	23.40	20.00
26.31	21.80	22.20
24.10	23.50	19.75
23.74	22.75	20.60
25.10	21.60	20.40

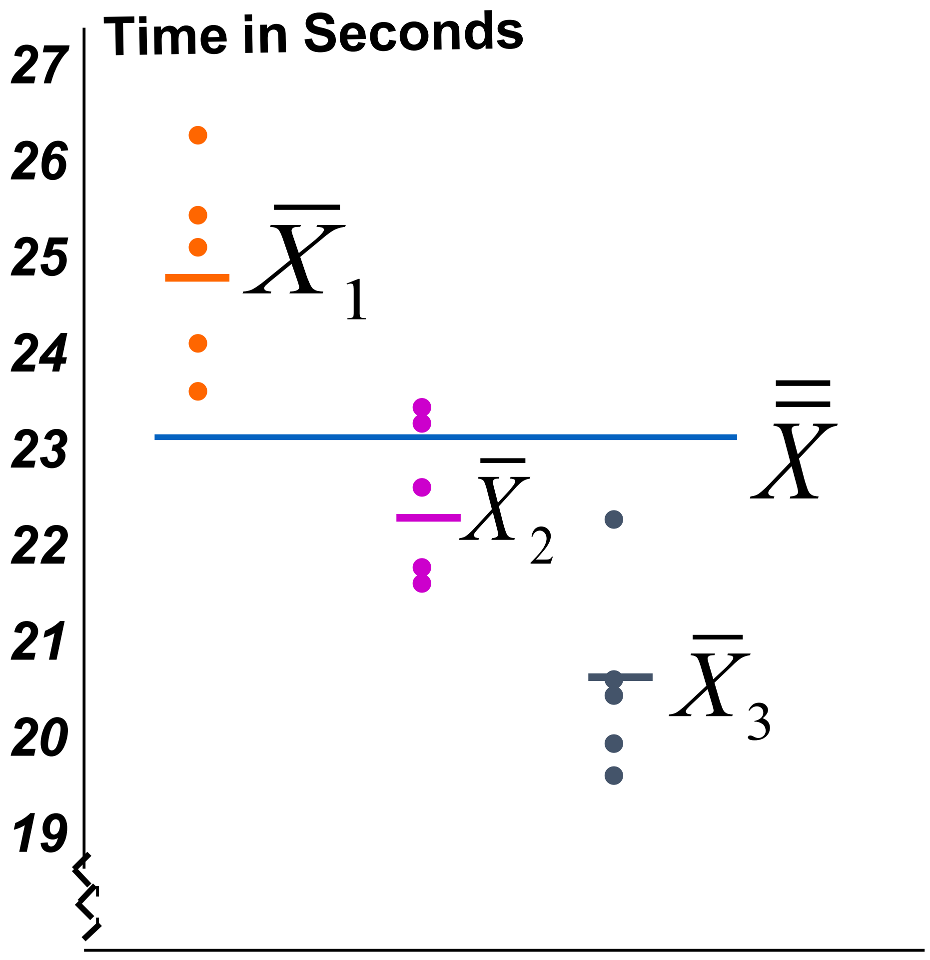


One-Way ANOVA Example 2: Scatter Diagram

Machine1	Machine2	Machine3
25.40	23.40	20.00
26.31	21.80	22.20
24.10	23.50	19.75
23.74	22.75	20.60
25.10	21.60	20.40

$\bar{X}_1 = 24.93 \qquad \bar{X}_2 = 22.61$

$\bar{X}_3 = 20.59 \qquad \bar{\bar{X}} = 22.71$



One-Factor ANOVA Example 2 Computations

Machine1	Machine2	Machine3
25.40	23.40	20.00
26.31	21.80	22.20
24.10	23.50	19.75
23.74	22.75	20.60
25.10	21.60	20.40

$\bar{X}_1 = 24.93$

$\bar{X}_2 = 22.61$

$\bar{X}_3 = 20.59$

$\bar{\bar{X}} = 22.71$

$n_i = 5, i = 1,2,3$

K=3

N=15

$SSG = \sum_{i=1}^{K} n_i(\bar{X}_i - \bar{\bar{X}})^2$=5[$(24.93 - 22.71)^2$+$(22.61 - 22.71)^2$+$(20.59 - 22.71)^2$]=47.164

$$SSE = (X_{11} - \bar{X}_1)^2 + \ldots + (X_{1n_1} - \bar{X}_1)^2 + \ldots + (X_{K1} - \bar{X}_K)^2 + \ldots + (X_{Kn_K} - \bar{X}_K)^2$$
$$= (25.40 - 24.93)^2 + (25.10 - 24.93)^2 + \ldots + (20.00 - 20.59)^2 + (20.40 - 20.59)^2 = 11.0532$$

MSG= $\frac{SSG}{K-1}$=47.164/2=23.5820; MSE $= \frac{SSE}{n-K}$=11.0532/12=0.9211

ANOVA Table: Example 2

Using R

```
> mach.aov <- aov(time ~ machines)
> summary.aov(mach.aov)
            Df  Sum Sq   Mean Sq   F value   Pr(>F)
machines     2  47.16    23.582    25.6      4.68e-05 ***
Residuals   12  11.05     0.921
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Source of Variation	Degrees of Freedom	Sum of Squares	Mean Squares (Variance)	F Statistic
Among (Factor)	3-1=2	47.1640	23.5820	MSA/MSW =25.60
Within (Error)	15-3=12	11.0532	.9211	
Total	15-1=14	58.2172		

One-Factor ANOVA Example 2 Solution

STEP 1 H_0: $\mu_1 = \mu_2 = \mu_3$ *K=3*

H_1: Not All Equal

K= 3 N=15 df_1= K-1=2 df_2 = N-K=12

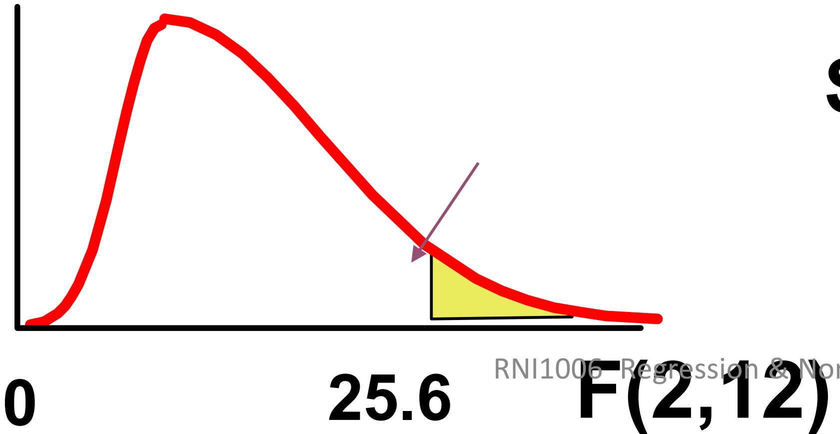
Step 3 The sampling distribution of the test statistic F is F with df=(K-1, N-K)= (2, 12)

STEP 4 p-value

=P (F(2,12)>25.6)

=pf(25.6,2,12,lower.tail=FALSE)= 4.685814e-05

(using R)



STEP 2 Test Statistic:

$$F = \frac{MSG}{MSE} = \frac{23.5820}{.9211} = 25.6$$

STEP 5 Decision: As the p-value 0.0000469 is very small,

Reject at $\alpha = 0.05$

STEP 6 Conclusion:

There is very strong evidence that at least one μ_i differs from the rest.

Aim 2 The Kruskal-Wallis Test

The Kruskal-Wallis test is a rank test that can replace the ANOVA F test. The idea of the Kruskal-Wallis rank test is to rank all the responses from all groups together and then apply one-way ANOVA to the ranks rather than to the original observations.

The Kruskal-Wallis Test

Draw independent SRSs of sizes $n_1, n_2, \ldots, n_K$ from K populations. There are N observations in all. Rank all N observations and let R_i be the sum of the ranks for the i^{th} sample. The **Kruskal-Wallis statistic** is

$$H = \frac{12}{N(N+1)} \sum \frac{R_i^2}{n_1} - 3(N+1)$$

When the sample sizes are large and all populations have the same continuous distribution, H has approximately the chi-square distribution with K–1 degrees of freedom. The **Kruskal-Wallis test** rejects the null hypothesis that all populations have the same distributions when H is large.

Kruskal-Wallis Rank Test for *K* Medians

- Extension of Wilcoxon rank-sum test
  - Tests the equality of more than 2 ($K>2$) population medians
- Distribution-free test procedure
- Used to analyze completely randomized experimental designs
- Use χ^2 distribution to approximate if each sample group size $n_i > 5$
  - $df = K - 1$

Kruskal-Wallis Rank Test

- **Assumptions**
  - Independent random samples are drawn
  - Continuous dependent variable
  - Data may be ranked both within and among samples
  - Populations have same variability
  - Populations have same shape
- **Robust with regard to the last two conditions**
  - Use F test in completely randomized designs and when the more stringent assumptions hold

Kruskal-Wallis test

1. Draw independent SRSs of sizes $n_1, n_2, \dots, n_K$ from K populations.
2. There are N observations in all
3. **Rank all $\boldsymbol{N}$ observations and let $\boldsymbol{R_i}$ be the sum of the ranks for the $\boldsymbol{i^{th}}$ sample**
4. The Kruskal-Wallis statistic is
$$H = \frac{12}{N(N+1)} \sum_{i=1}^{K} \frac{R_i^2}{n_i} - 3(N+1)$$
5. When the sample sizes n_i are moderately large and all K populations have the same continuous distribution*,
$$H \sim \chi^2_{K-1}$$
6. When H is large, we reject the null hypothesis that all populations have the same distribution

Kruskal-Wallis test

- Kruskal-Wallis test can replace ANOVA F-test
  - Can relax assumption about Normality
  - Independent random sampling still important
- Hypotheses tested are either:

  H_0: response has same distribution in all groups
  H_A: response is systematically different in some groups than in others

  OR, if distributions have same shape,

  $H_0: \tilde{\mu}_1 = \cdots \tilde{\mu}_K$
  H_A: not all medians are the same
- Basic idea is to rank all the responses from all groups together and then apply one-way ANOVA to ranks.
- Kruskal-Wallis test statistic is essentially SSG for ranks

6 Steps Hypothesis Testing: Kruskal-Wallis Test

STEP 1 Hypotheses Testing

H_0: response has same distribution in all groups

H_A: response is systematically different in some groups than in others

OR if distributions have same shape,

$H_0: \tilde{\mu}_1 = \cdots \tilde{\mu}_K$ H_A: not all medians are the same

STEP 2 Test Statistic $H = \frac{12}{N(N+1)} \sum_{i=1}^{K} \frac{R_i^2}{n_i} - 3(N+1)$

STEP 3 The sampling distribution $H \sim \chi^2_{K-1}$

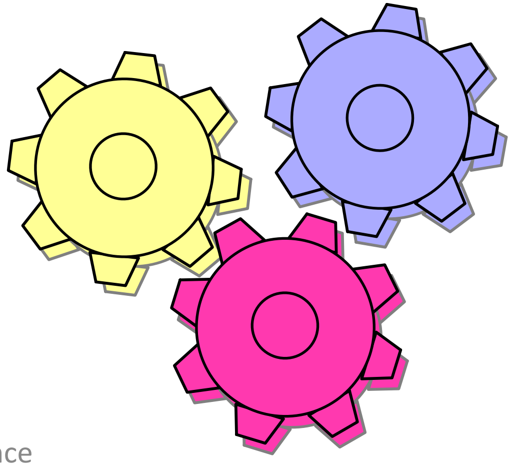
STEP 4 p-value = Prob(χ^2_{K-1}> H)

STEPS 5 and 6. Decision and Conclusion

Kruskal-Wallis Rank Test: Example 3 (was Example 2)

As production manager, you want to see if three filling machines have different median filling times. You assign 15 similarly trained & experienced workers, five per machine, to the machines. At the .05 significance level, is there a difference in median filling times?

Machine1	Machine2	Machine3
25.40	23.40	20.00
26.31	21.80	22.20
24.10	23.50	19.75
23.74	22.75	20.60
25.10	21.60	20.40



Example 3 Solution: Step 1 Obtaining a Ranking

Raw Data

Machine1	Machine2	Machine3
25.40	23.40	20.00
26.31	21.80	22.20
24.10	23.50	19.75
23.74	22.75	20.60
25.10	21.60	20.40

Ranks

Machine1	Machine2	Machine3
14	9	2
15	6	7
12	10	1
11	8	4
13	5	3
65	**38**	**17**

Example 3 Solution: Step 2 Test Statistic Computation

$$H = \frac{12}{N(N+1)} \sum_{i=1}^{K} \frac{R_i^2}{n_i} - 3(N+1)$$

$$= \frac{12}{15(15+1)} \left(\frac{65^2}{5} + \frac{38^2}{5} + \frac{17^2}{5}\right) - 3(15+1)$$

$$= 11.58$$

Kruskal-Wallis Test Example 3: 6 Steps Hypothesis Testing

STEP 1

H0: $\eta_1 = \eta_2 = \eta_3$

H1: Not all equal

$\alpha = .05$

$df = K - 1 = 3 - 1 = 2$

STEP 2 Test Statistic:

$$H = 11.58$$

STEP 3 Sampling

Distribution $H \sim \chi^2_{df=2}$

STEP 4 p-value

P-value = P($\chi^2_{df=2}$ > 11.58) (using R)

=pchisq(11.58, 2, lower.tail=FALSE)=0.0031

STEP 5 Decision: $\alpha = .05$

As p-value =0.0031 < 0.05, reject Ho.

STEP 6 Conclusion:

There is evidence that population medians are not all equal.

Kruskal-Wallis Test: Example 4

Effect of weed density on yield of corn

- A researcher planted corn at the same rate in 16 plots, and then randomly assigned the plots to four groups based on weed density
  - 0, 1, 3, and 9 weed plants/meter of corn
- Is the yield of corn affected by the density of weed plants?

Kruskal-Wallis Test: Example 4

Weeds per Meter	Corn Yield	Weeds per Meter	Corn Yield	Weeds per Meter	Corn Yield	Weeds per Meter	Corn Yield
0	166.7	1	166.2	3	158.6	9	162.8
0	172.2	1	157.3	3	176.4	9	142.4
0	165.0	1	166.7	3	153.1	9	162.7
0	176.9	1	161.1	3	156.0	9	162.4

Weeds	*n*	Median	Mean	Std. Dev.
0	4	169.45	170.200	5.422
1	4	163.65	162.825	4.469
3	4	157.30	161.025	10.493
9	4	162.55	157.575	10.118

Kruskal-Wallis Test: Example 4

- Yields do go down as more weeds are present
- Some outliers present – difference between medians and means
- Sample standard deviations do not quite satisfy rule-of-thumb for constant population standard deviations
- Might prefer to use nonparametric test rather than ANOVA
- Because of small sample size, hypotheses tested are
- **STEP 1 State the hypotheses**

  H_0: yields have the same distribution in all groups
  H_A: yields are systematically higher in some groups than others

Kruskal-Wallis Test: Example 4

Yield								
Rank								

Yield								
Rank								

Weeds	Ranks				Sum of Ranks
0	10	12.5	14	16	52.5
1	4	6	11	12.5	33.5
3	2	3	5	15	25.0
9	1	7	8	9	25.0

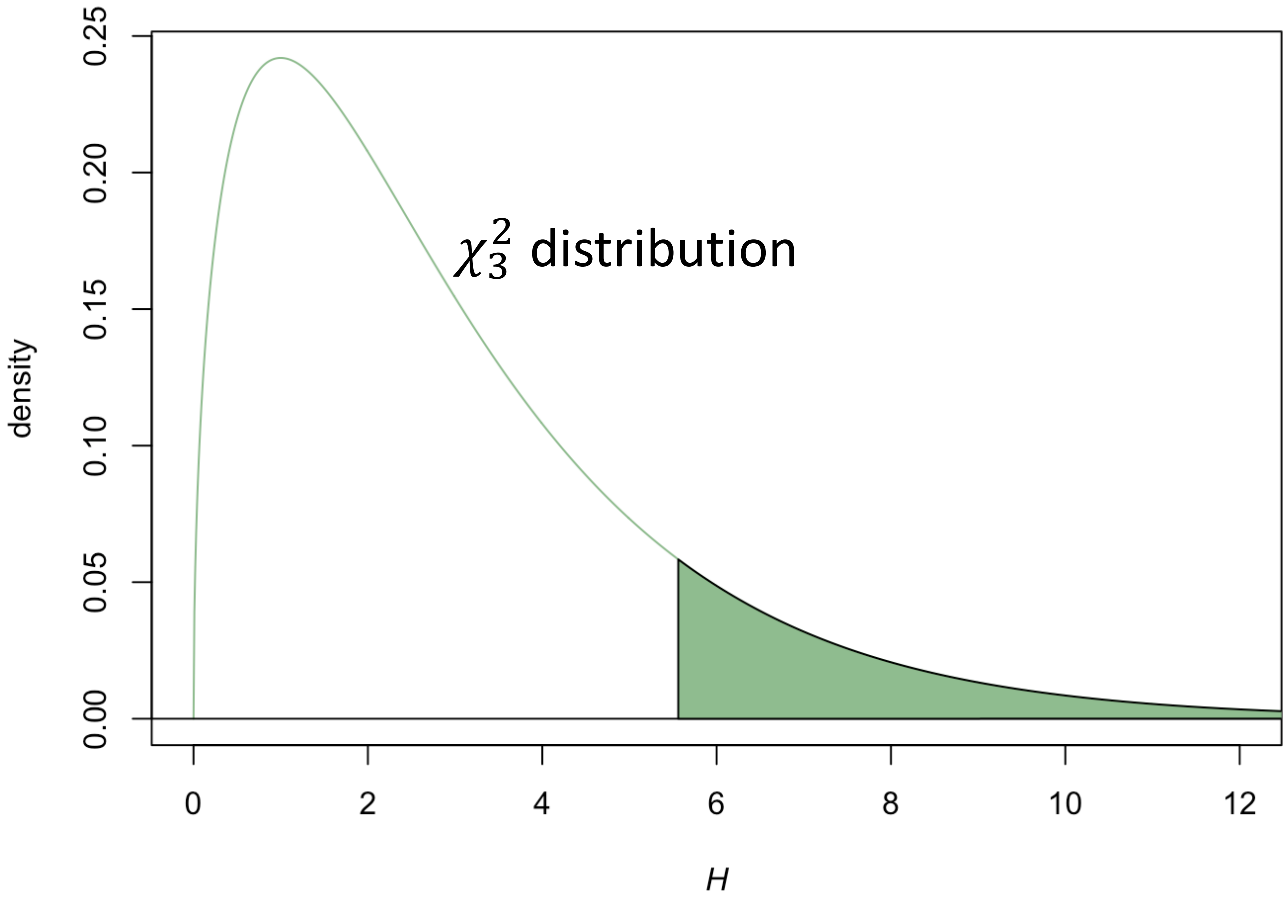
Kruskal-Wallis Test: Example 4

- **Step 2 Test statistic**

$$\mathrm{H} = \frac{12}{16(16+1)}\left(\frac{52.5^2}{4}+\frac{33.5^2}{4}+\frac{25.5^2}{4}+\frac{25^2}{4}\right)-3(16+1)$$
$$= \frac{12}{272}(1282.125)-51$$
$$= 5.56 \sim\chi_3^2$$

- **STEP 3**. The sampling distribution $\mathrm{H} \sim\chi_3^2$
- **STEP 4.** The p-value. Use *R* to get p-values

Kruskal-Wallis Test: Example 4



- **STEP 5**

$p(H \geq 5.56) = 0.135$ so we have insufficient evidence to reject null hypothesis

- **STEP 6**

This small experiment suggests that more weeds decrease yield, but the evidence is not convincing

- Not unexpected!