ANOVA

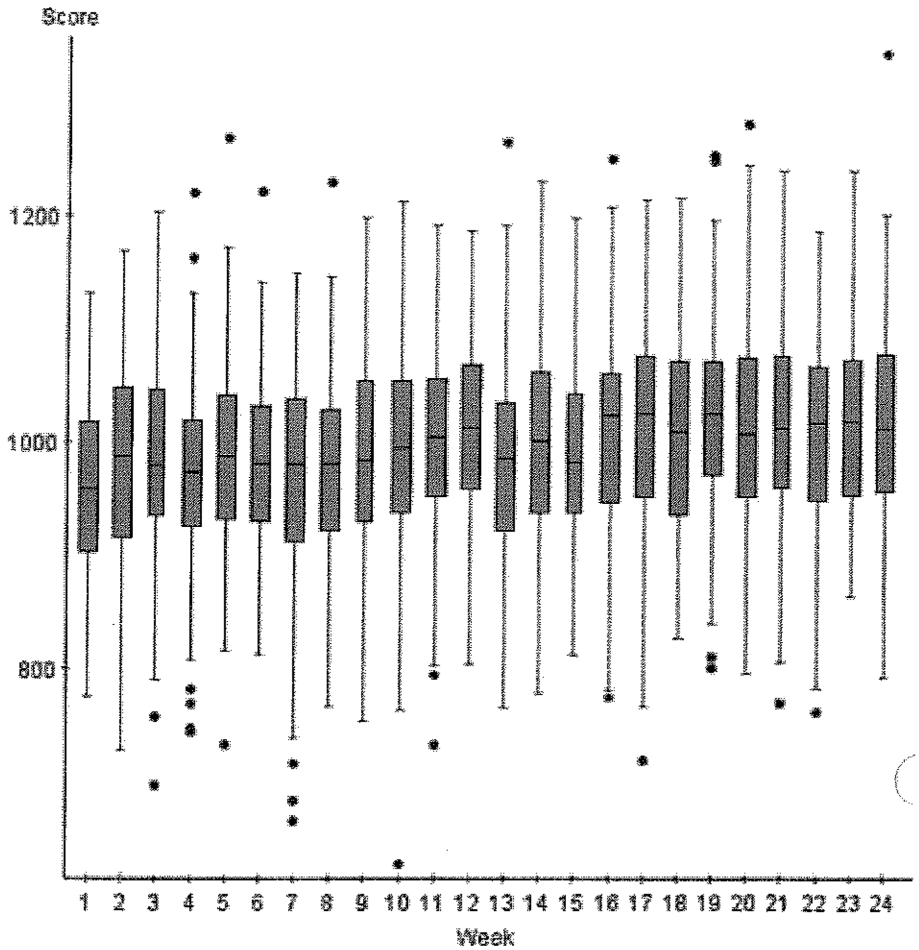
- The Tuesday Men's Handicap League at Blue Hens Lanes in Newark is essentially a randomized experiment.
- Each week, teams comprised of 5 men are randomly assigned to a pair of lanes and the bowlers bowl three games. The typical bowler in the league carries a 200 average.
- To keep things simpler, we will look at team scores each game rather than the 5 individual scores
- Bowlers frequently complain about everything. "***The lanes break down after the second game***", "***This pair of lanes sucks***", and "***This whole house is horrible this week***".
- As with regression, the variable of interest is called the response variable. Here, it is the team score for each game.
- We have three factors we will analyze:
  - Is there a game effect? In other words, do the means differ when looking at game 1 versus game 2 versus game 3.
  - Is there a lane effect? Are certain pairs of lanes better than others?
  - Is there a week effect? Is the oil so different one week that scores overall are worse? Are bowlers getting better as the season progresses?

- To statistically determine if these factors matter for scoring in the Tuesday league, we will run an ANOVA, or an Analysis of Variance.
- The wrong way to test all of these comparisons of means would be to run a ton of T-Tests. Testing all pairs (e.g. check if the mean score on lanes 1 and 2 differs from the mean score on lanes 3 and 4, ***And So On***) generates way too many Type I errors.
- Looking at the different pairs of lanes, we cannot expect the sample means to be exactly equal, even lanes don't matter!
- Whenever we test a factor with ANOVA, we start by assuming equality and we ask ourselves, "*Could the 16 pairs of lanes really have the same mean scoring and we just happened to get a difference like this because of natural sampling variability?*"

- Each week, there are 32 teams that bowl 3 games each. Each boxplot is a graph of 96 scores. **Is there a week effect?**

Summary statistics for Score:
Group by: Week

Week	Mean	Std. Dev.
1	956.8958	81.64074
2	984.73956	91.10887
3	980.09375	94.835045
4	972.9375	90.28293
5	985	87.166626
6	982.2292	78.13813
7	964.9583	100.241066
8	975.5	81.66582
9	986.26044	89.21874
10	989.5833	96.885895
11	1000.1042	85.45045
12	1014.7917	82.261444
13	984.42706	87.916794
14	1002.1667	94.48104
15	991.9375	87.653
16	1006.03125	92.18317
17	1013.28125	95.8627
18	1008.53125	89.5662
19	1018.9583	87.28585
20	1017.34375	91.5775
21	1011.9375	85.45135
22	1008.6458	87.308525
23	1020.84375	86.25487
24	1013.625	92.10329

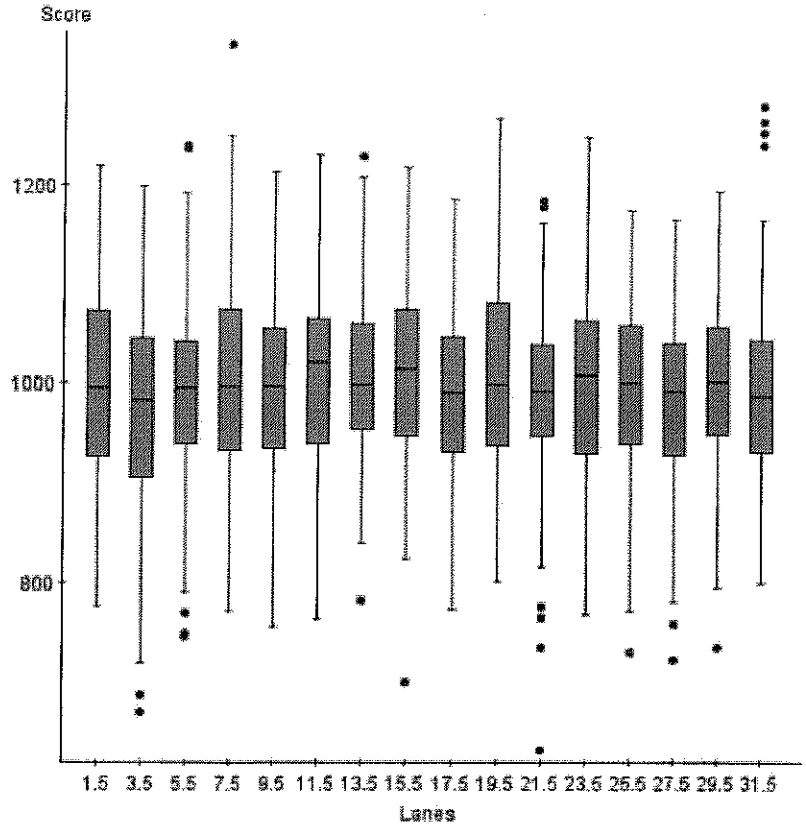


Do the means vary enough to convince us that every week was ~~not~~ created equally?

Is there a lane effect?

Summary statistics for Score:
Group by: Lanes

Lanes	Mean	Std. Dev.
1.5	996.8403	94.83631
3.5	970.4861	100.778435
5.5	990.3472	88.84349
7.5	1006.00696	97.40453
9.5	993.99304	92.414955
11.5	1002.9375	99.6508
13.5	1005.7778	81.115
15.5	1009.8542	89.4148
17.5	989.55554	83.11285
19.5	1005.7153	92.64087
21.5	987.7917	89.94038
23.5	999.30554	90.673416
25.5	997.3125	83.021225
27.5	981.2361	83.55175
29.5	998.7083	84.91179
31.5	991.3472	88.91367



Lanes 1 and 2 noted as 1.5, etc...

Are some pairs better than others?

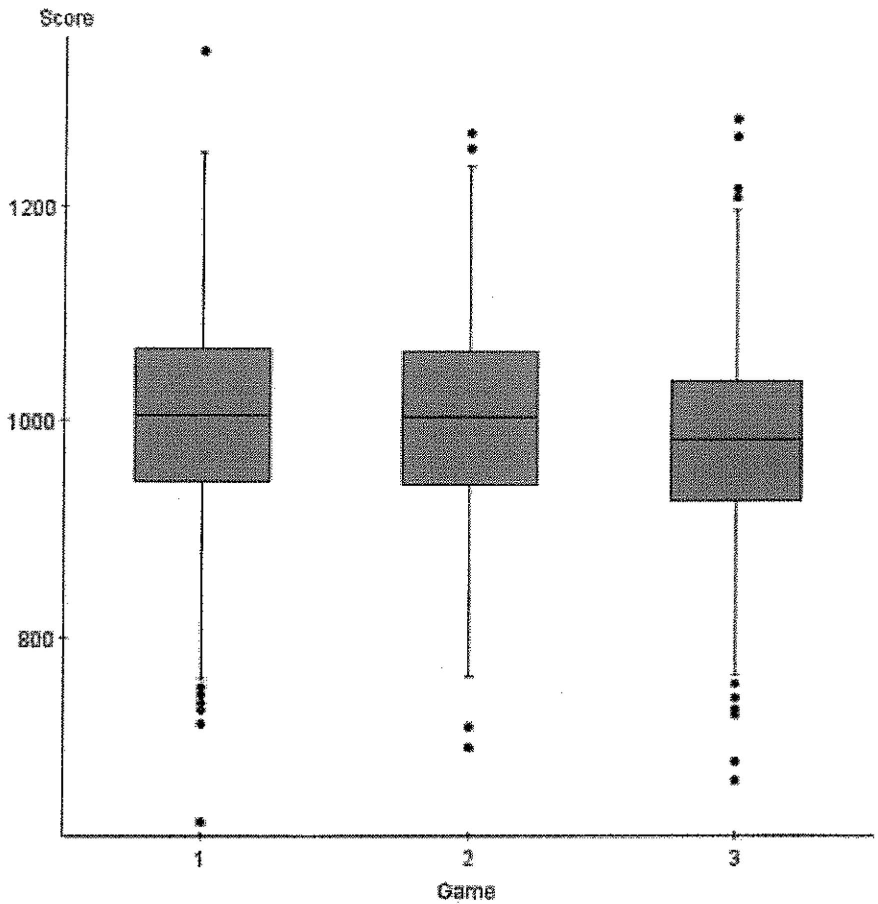
Worst is 3/4 at 970.5
Best is 15/16 at 1009.9

Is that statistically significant?

Is there a game effect?

Summary statistics for Score:
Group by: Game

Game	Mean	Std. Dev.
1	1005.0677	94.852844
2	1001.60547	86.78158
3	979.6797	87.720375



- We will test for a game effect first. Write down the appropriate hypotheses:

$H_0: \mu_1 = \mu_2 = \mu_3 = \mu$

H_A: At least one μ is different.

- To test these hypotheses, we need a new sampling distribution model called the F-Model, named for Sir Ronald Fisher.
- Since we have more than two groups, we cannot just look at the differences in means. If the null hypothesis were true, we'd still expect the Sample means to vary a bit. How much should they vary?

- If the null hypothesis were true, then each of the treatment means (Games 1, 2, 3) would be estimating the same underlying mean. For the game effect, we have 3 estimates. Treat these means as data points or observations and calculate the sample variance. We use this variance to assess how different the group means are from each other.
- If the group means are close, this variance will be Small. The more they differ, the larger it will be, indicating that the treatment (here, game 1, 2, 3) is actually meaningful.
- For the bowling games, we have:

Summary statistics for Score:
Group by: Game

Game	*n*	Mean
1	768	1005.0677
2	768	1001.60547
3	768	979.6797

Variance is 189.54616

- The sample variance (found using StatCrunch) for the 3 means is 189.54616
- We know the variance of a sample mean is σ^2/n, and with 768 observations in a group, that would be $\sigma^2/768$. The figure we just calculated, 189.54616 would estimate that quantity, and ~~the~~ to get back to the variance of the observations, multiply it by 768. We get 189.54616 × 768 = 145571.4509
- Is this large? This variance is called the Mean Square for Treatment, denoted MS_T.
- If MS_T is large, it means the treatment effect is large. We just need a something suitable to compare it to – and that something is called Mean Square for Error, MS_E.

- A suitable comparison of variability is the one due to the game-to-game differences in bowling scores. These surely will be different as bowlers have good games, average games and bad games. We need an independent estimate of σ^2. It cannot depend upon the null hypothesis being true (mean scores are equal across each game).
- What we do is this: Calculate a variance for each group, then we Pool them together. Because the sample sizes are equal, we can just average the variances.

Summary statistics for Score:
Group by: Game

Game	n	Mean	Variance	Std. Dev.
1	768	1005.0677	8997.062	94.852844
2	768	1001.60547	7531.0425	86.78158
3	768	979.6797	7694.8647	87.720375

$$MSE = \frac{8997.062 + 7531.0425 + 7694.8647}{3}$$

$$= 8074.323$$

- For the pooled variance, each variance is taken around its own treatment mean, so the pooled estimate does not depend on treatment means being equal
- Fixed The ~~estimate in which we use the three means as observations~~ do, the MS_T = 145571.66 does depend on the treatment means being equal. This number is much larger than the Mean Square Error.

It is 18 times larger than MSE

- We have two estimates of the underlying variance in bowling scores – one is based on the Differences between the group means. The other is based on the variation Within each group.
- If the null hypothesis is true, both MST and MSE estimate σ^2. Their ratio should be close to 1.
- If the null hypothesis is false, MST will be larger because the treatment means are different. Thus the ratio of MST / MSE will tend to be bigger than 1.

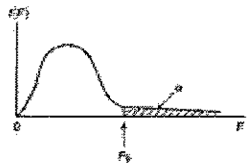
The *F*-Statistic

To test for a difference in more than two means, we calculate the *F*-statistic:

$$F = \frac{MS_T}{MS_E}$$

- The *F*-statistic is always Positive, and large values give Small *P*-values. TEST IS ALWAYS ONE-SIDED.
- Tables exist, but using them is a complete waste of time Technology gives an ANOVA table and all the relevant information to run the test.
- There are two separate degrees of freedom, one for the numerator and one for the denominator.
- We use N to denote the total number of cases.
- We use K to denote the number of groups (each with its own mean).
- The numerator, where we estimate the variance between group means has $k-1$ degrees of freedom (k things estimated).
- The denominator has the remaining $N-k$ degrees of freedom for the error.
- Since it all depends precisely on the two degrees of freedom, it is hard to tell what is a big *F*. Always use the *P*-value given in the computer output (shading right under the appropriate *F* model).

TABLE B-3
The *F* distribution (α = 0.10, 0.05, and 0.01)



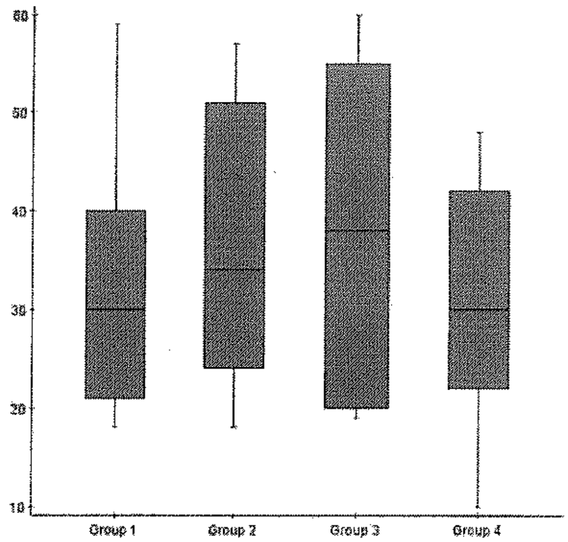
Given ν_1 and ν_2, the table gives the F_0 value with α of the area above it, that is, $P(F \geq F_0) = \alpha$

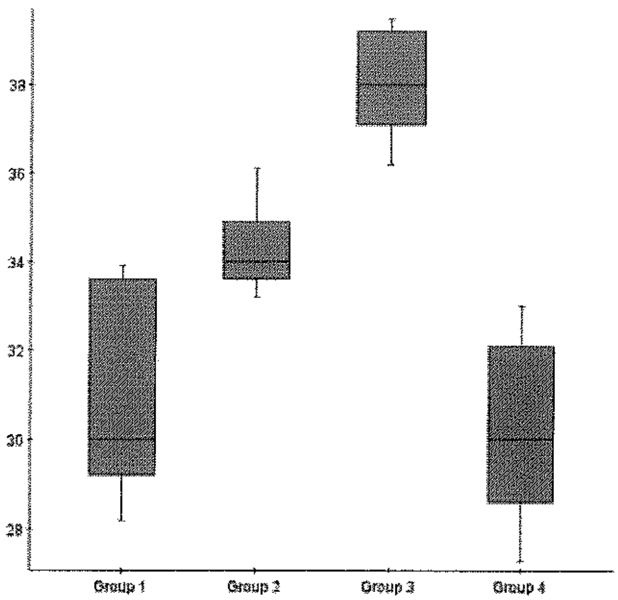
ν_2	α	1	2	3	4	5	6	7	8	9	10	11	12	14	15	19	20	24	30	50	100	500	∞
	.10	39.9	49.5	53.6	55.8	57.2	58.2	58.9	59.4	59.9	60.2	60.5	60.7	61.1	61.2	61.6	61.7	62.0	62.3	62.7	63.0	63.3	63.3
1	.05	161	200	216	225	230	234	237	239	241	242	243	244	245	246	248	248	249	250	252	253	254	254
	.10	8.53	9.00	9.16	9.24	9.29	9.33	9.35	9.37	9.38	9.39	9.40	9.41	9.42	9.42	9.44	9.44	9.45	9.46	9.47	9.48	9.49	9.49
2	.05	18.5	19.0	19.2	19.2	19.3	19.3	19.4	19.4	19.4	19.4	19.4	19.4	19.4	19.4	19.4	19.4	19.5	19.5	19.5	19.5	19.5	19.5
	.01	98.5	99.0	99.2	99.2	99.3	99.3	99.4	99.4	99.4	99.4	99.4	99.4	99.4	99.4	99.4	99.4	99.5	99.5	99.5	99.5	99.5	99.5
	.10	5.54	5.46	5.39	5.34	5.31	5.28	5.27	5.25	5.24	5.23	5.22	5.22	5.20	5.20	5.18	5.18	5.18	5.17	5.15	5.14	5.14	5.13
3	.05	10.1	9.55	9.28	9.12	9.01	8.94	8.89	8.85	8.81	8.79	8.76	8.74	8.71	8.70	8.67	8.66	8.64	8.62	8.58	8.55	8.53	8.53
	.01	34.1	30.8	29.5	28.7	28.2	27.9	27.7	27.5	27.3	27.2	27.1	27.1	26.9	26.9	26.7	26.7	26.6	26.5	26.4	26.2	26.1	26.1
	.10	4.54	4.32	4.19	4.11	4.05	4.01	3.98	3.95	3.94	3.92	3.91	3.90	3.88	3.87	3.86	3.84	3.83	3.82	3.80	3.78	3.76	3.76
4	.05	7.71	6.94	6.59	6.39	6.26	6.16	6.09	6.04	6.00	5.96	5.94	5.91	5.87	5.86	5.81	5.80	5.77	5.75	5.70	5.66	5.64	5.63
	.01	21.2	18.0	16.7	16.0	15.5	15.2	15.0	14.8	14.7	14.5	14.4	14.4	14.2	14.2	14.0	14.0	13.9	13.8	13.7	13.6	13.5	13.5
	.10	4.06	3.78	3.62	3.52	3.45	3.40	3.37	3.34	3.32	3.30	3.28	3.27	3.25	3.24	3.21	3.21	3.19	3.17	3.15	3.13	3.11	3.10
5	.05	6.61	5.79	5.41	5.19	5.05	4.95	4.88	4.82	4.77	4.74	4.71	4.68	4.64	4.62	4.57	4.56	4.53	4.50	4.44	4.41	4.37	4.36
	.01	16.26	13.27	12.06	11.39	10.97	10.67	10.46	10.29	10.16	10.05	9.96	9.89	9.77	9.72	9.58	9.55	9.47	9.38	9.24	9.13	9.04	9.02
	.10	3.78	3.46	3.29	3.18	3.11	3.05	3.01	2.98	2.96	2.94	2.92	2.90	2.88	2.87	2.84	2.84	2.82	2.80	2.77	2.75	2.73	2.72
6	.05	5.99	5.14	4.76	4.53	4.39	4.28	4.21	4.15	4.10	4.06	4.03	4.00	3.96	3.94	3.88	3.87	3.84	3.81	3.75	3.71	3.68	3.67
	.01	13.74	10.92	9.78	9.15	8.75	8.47	8.26	8.10	7.98	7.87	7.79	7.72	7.60	7.56	7.42	7.40	7.31	7.23	7.09	6.99	6.90	6.88

The degrees of freedom are ν_1 for the numerator and ν_2 for the denominator.

Understanding Why We Use Variances to Test Means

- In the four boxplots on the left, the medians are 30, 34, 38 and 30.
- Visually, there is no difference in means (centers / medians).
- There is too much variation within each group to sort of if there is a true difference in group means.





- In the four boxplots on the right, the medians are 30, 34, 38, and 30.
- Visually, there appears to be a difference in means (centers / medians).
- Because the variation within each group is small, we can see there is a difference between group means.
- The *F*-statistic sorts this out: Variation between group means in the numerator, variation within groups in the denominator (pooled together). When the numerator is larger, it indicates there is a difference between groups. Ingenious!

Assumptions and Conditions for ANOVA

The data collected must be generated with suitable randomization. The spread within each group must be roughly equal. The effects of outliers must be minimum or they should be removed / corrected. The boxplots should be fairly symmetric. Finally the residuals should be normal, but if all previous conditions are met, this will follow naturally.

Example: Back to bowling. There were 2304 team bowling scores through week 24 of the 2012 – 2013 season. Each of 32 teams bowls three games each week, and bowlers frequently complain that the lanes break down, killing game 3 scores. Run an ANOVA analysis to test for a difference in means if the factor is "***Game***".

Analysis of Variance results:

Responses stored in Score.

Factors stored in Game.

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

H_A: At least one μ_i is different

$N = 2304$

$k = 3$

Factor means

Game	n	Mean	Std. Dev.	Std. Error
1	768	1005.0677	94.852844	3.422707
2	768	1001.60547	86.78158	3.1314604
3	768	979.6797	87.720375	3.1653364

$F = \frac{MS_T}{MS_E} = \frac{145571.66}{8074.3228}$

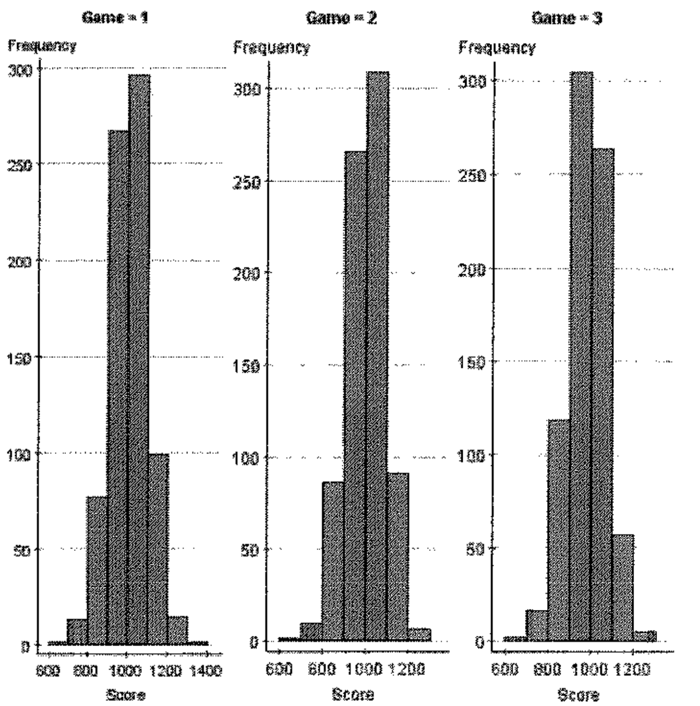
$= 18.028961$

ANOVA table

Source	df	SS	MS	F-Stat	P-value
Treatments	2	291143.3	145571.66	18.028961	<0.0001
Error	2301	1.8579018E7	8074.3228		
Total	2303	1.887016E7			

MS_T

MS_E



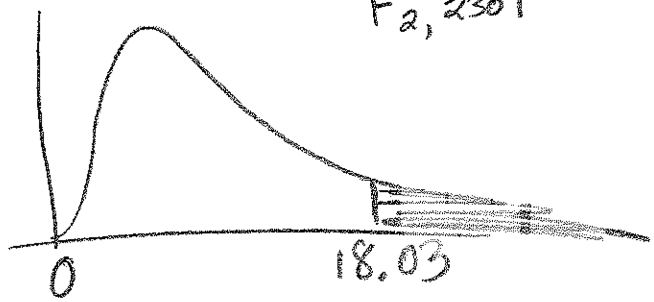
Numerator $df = k-1 = 3-1 = 2$

Denominator $df = N-K$
$= 2304-3$
$= 2301$

← Data look unimodal & symmetric

P-Value

$F_{2,2301}$



With P-Value < 0.0001,
Reject H_0 at any α level.

We have evidence the mean scores differ from game-to-game.

This test does not answer "Which games are lower/higher, 1, 2, or 3".

We explore that later!

Example: Continuing with the bowling example, run separate ANOVA analyses to test for a "*Lane*" effect and then a "*Week*" effect. Use StatCrunch.

TEST FOR "LANE" EFFECT:

$H_0: \mu_{1.5} = \mu_{3.5} = \cdots = \mu_{31.5}$

H_A: At least one is different

Conditions: Randomized data

* 16 Histograms all unimodal, roughly symmetric (lanes 3,4 a bit skewed)
* A few outliers on most pairs of lanes, but sample sizes very large so OK.
* Spread roughly equal in each group.

$N = 2304$, $K = 16$ groups

Df Numerator $= K-1 = 15$

Df Denominator $= N-K = 2288$

F-Statistic $= 1.86$

P-Value $= 0.0225$

There is some evidence of a "Lane" effect (not really strong evidence).

Lanes 3/4 @ 970.49

Lanes 15/16 @ 1009.85

Biggest Difference

TEST FOR A "WEEK" Effect

$H_0: \mu_{WEEK\ 1} = \cdots = \mu_{week\ 24}$

H_A: At least ONE WEEK IS DIFFERENT

24 HISTOGRAMS ALL LOOK UNIMODAL, SYMMETRIC, EQUAL SPREAD

F-Statistic = 4.04

P-Value < 0.0001

Reject H_0, conclude that at least one week has a different mean score

Which ones?

We can see a general increasing trend.

Run a "Means" Plot on SC.

Example: A student runs an experiment to study the effect of three different mufflers on gas mileage. He fits a system to his Scion TC to give the car exactly one gallon of gas. He tests each muffler 8 times, carefully recording the number of miles he goes before running out of gas. After analyzing the data, the *F*-ratio is 2.35 and the *P*-value is 0.1199.

a. What are the null and alternative hypotheses?

b. How many degrees of freedom does the treatment sum of squares have? How about the error sum of squares?

c. What can he conclude?

d. What else about the data would we like to know in order to check assumptions and conditions?

e. If our conclusion is wrong, what type of error have we made? What's it mean?

a.) $H_0: \mu_1 = \mu_2 = \mu_3$ (μ MPG FOR 3 Mufflers =)

H_A: At least one mean is different

b.) $N = 3 \times 8 = 24$

$K = 3$

SS_T has $3-1=2$ df

SS_E has $24-3=21$ df

c.) P-Value = 0.1199 is high, so no evidence the mean gas mileage differs based on muffler.

d.) Are boxplots symmetric, with equal spread & limited outlier effect?

e.) Type II. It means that in reality, muffler choice does have an effect on MPG (our data was unlucky)

The ANOVA Model

- We need to model the data that we collect from our designed experiment, so we start with the simplest model possible. This model says that the differences we observe are based on the differences in treatment means and that any variation around that mean is just random error:

$$y_{ij} = \mu_j + \varepsilon_{ij}$$

ith case in treatment j

jth treatment mean

error for the ith case in treatment j

- Recall the null hypothesis for ANOVA is that all means are equal:

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

- Think of our bowling example. There is an overall mean bowling score for the Tuesday League. The treatments (game 1, 2, or 3) add or subtract from this grand mean. We can rewrite our model then to showcase the j^{th} treatment effect:

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

Grand Mean

jth treatment effect

- Now, we could write the null hypothesis to emphasize the treatment effects instead of the means:

$$H_0: \tau_1 = \tau_2 = \cdots = \tau_K$$

(all treatment effects equal)

- So now there are three kinds of parameters – Overall Mean, Treatment Effects, and the Errors.

- To estimate the overall mean, we use the grand mean, or the mean of all our data: $\bar{\bar{Y}}$

- To estimate the mean of a particular treatment, we use the j^{th} treatment mean:

$$\bar{Y}_j$$

- To estimate the j^{th} treatment effect, subtract the grand mean from the particular treatment mean:

jth treatment effect is

$$\hat{\tau}_j = \overline{Y}_j - \overline{\overline{Y}}$$

- Each observation (each game bowled) has an error, ε_{ij}. Those are the residuals (just like for regression), and we can estimate those by taking the difference between the actual y value and the treatment mean:

True error is ε, estimated by actual e

So $e_{ij} = y_{ij} - \overline{Y}_i$

- Finally, each observation from the experiment (each game bowled), can be rewritten to look like the sum of three quantities:

$$\underbrace{y_{ij}}_{\text{Data Value}} = \underbrace{\overline{\overline{Y}}}_{\text{GRAND MEAN}} + \underbrace{(\overline{Y}_j - \overline{\overline{Y}})}_{\text{TREATMENT EFFECT}} + \underbrace{(y_{ij} - \overline{Y}_j)}_{\text{RANDOM ERROR}}$$

- This equation breaks down ANOVA into its skeleton. Each observation is split into its sources – the grand mean, the treatment effect, and random error.

- ANOVA tests if the treatment effect are large, compared to the random error.

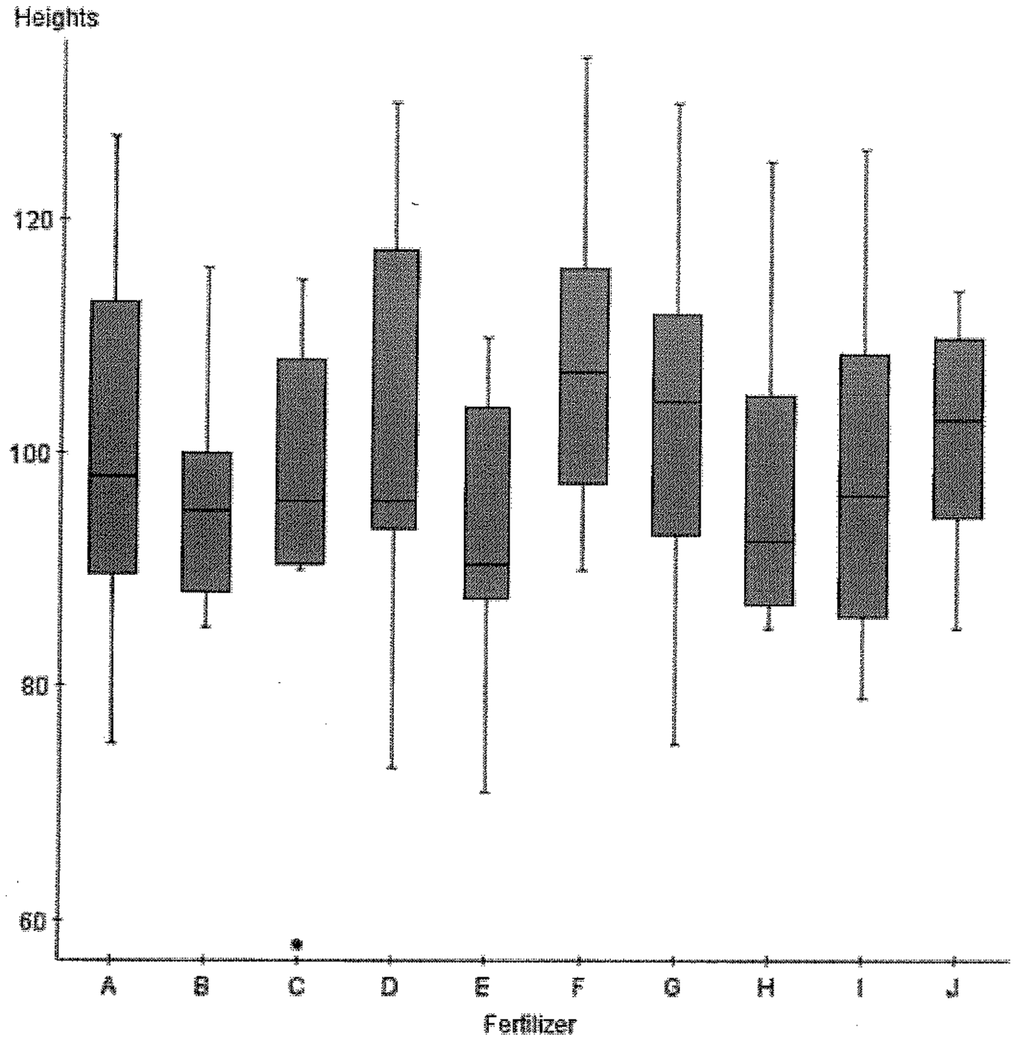
- The treatment effects are measured by MST and the error effects are measured by MSE.

- The ANOVA table gives sums of squares as well as mean squares for treatment and for errors. The mean squares are just the sums of squares divided by their degrees of freedom.

- If you want an estimate of the standard deviation, use the standard deviation of the errors. It is Pooled across all treatments. To compute it, simply do this:

$$S_p = \sqrt{MSE}$$

Example: A biologist investigated the effects of 10 different fertilizers on the growth of beans. Twelve beans each were placed in 10 different petri dishes, and the same amount of fertilizer was added to each dish. After a week, the heights of the 120 bean plants were measured in millimeters.



a. Do you have any concerns about the assumptions and conditions to run an ANOVA?

b. What are the hypotheses?

a.) Group C has a huge outlier and the variances within each group might be different.

b.) $H_0: \mu_1 = \mu_2 = \cdots = \mu_K$ OR

$\tau_1 = \tau_2 = \cdots = \tau_K$

HA: ~~At~~ least one different

Analysis of Variance results:
Responses stored in Heights.
Factors stored in Fertilizer.
Factor means

Fertilizer	n	Mean	Std. Dev.	Std. Error
A	12	99.833336	15.833413	4.5707126
B	12	95.25	8.9861	2.5940638
C	12	96.583336	14.853579	4.2878585
D	12	103.25	16.804356	4.8509994
E	12	93.416664	11.904608	3.4365644
F	12	108	14.154986	4.0861926
G	12	103.083336	16.384352	4.729755
H	12	97.416664	12.580348	3.6316335
I	12	98.5	15.974411	4.6114154
J	12	101.75	8.9861	2.5940638

ANOVA table

Source	df	SS	MS	F-Stat	P-value
Treatments	9	2073.7083	230.41203	1.1881874	0.3097
Error	110	21331.084	193.91895		
Total	119	23404.791			

c. What is the model we are trying to estimate with our collected data?

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

μ is overall mean height, τ_j is the effect of adding the jth fertilizer.

d. What can we conclude from the ANOVA table?

P-Value is high, so there is no evidence that any fertilizers are better or worse than the others.

Tukey's Honestly Significant Test (Tukey HSD)

- When running ANOVA, the F-test identifies that there is a difference in group means. It does not tell us which group means are significantly different.
- Tukey's test compares the means of every treatment to the means of every other treatment (all pairwise comparisons).
- It is based on a studentized range distribution, similar to the Student's t Model.
- The formula is beyond the scope of Stat II, but know this: When making multiple comparisons, the probability of making a Type I error increases. Tukey's test corrects for that. The intervals are essentially t-intervals with a corrected significance level.
- As a result of the correction to the experiment-wise error rate, the intervals will be wider than normal t-intervals. Thus, when you have an interval that does not include zero, it truly does indicate a significant difference in treatment means.

→ instead do regular t intervals

- If you did not do Tukey intervals, too many of them will conclude that the groups have different means, when in reality, they don't.

Example: Back to the fertilizer experiment. The researcher failed to reject the null hypothesis, so there was no significant difference in means. No need to test these since F-test failed.

As a result, with 10 treatments, there are 45 pairwise comparisons.

Every single one of the Tukey HSD intervals contains 0 inside, indicating there was no difference between fertilizer A and B, A and C, A and D, …, and H and I.

The Order of Operations

1. Run a designed experiment with multiple treatments and do an ANOVA.
2. If you reject the null hypothesis of equal means, then look to the Tukey HSD intervals to see which treatments have significantly different means.
3. It is possible to reject the F-test and then have no single pairing of treatments have different means.