Lecture 19: ANOVA Part I

Chapter 5.5

Discussion of Quiz

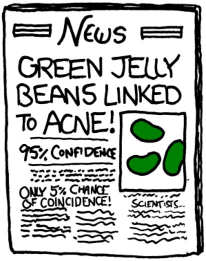
Question 1: Why did $\frac{1}{20}$ studies yield a positive/significant result i.e. that there is a link between jelly beans and acne?

Not that the p-value is 0.05, rather that $\alpha = 0.05$:

- (pre-specified) significance level AKA
- type I error rate AKA
- false positive rate

i.e. we expect 1 out of 20 results to be significant even if there is no effect by chance alone.

Publication Bias


NEWS
GREEN JELLY BEANS LINKED TO ACNE!
95% CONFIDENCE
ONLY 5% CHANCE OF COINCIDENCE!
SCIENTISTS...

Publication Bias

Publication bias: people only highlight significant/positive results. Wikipedia: "occurs when the publication of research results depends on their nature and direction."

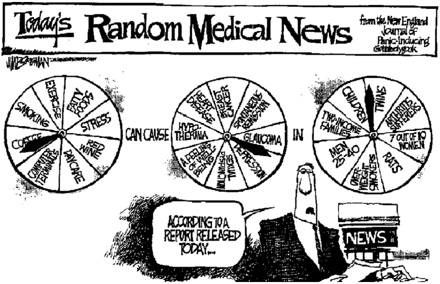
To counter this, some medical journals like

- New England Journal of Medicine
- The Lancet
- Journal of the American Medical Association

require registration of a trial before it starts so that unfavorable results are not withheld from publication.

Journal of Negative Results: `http://www.jnrbm.com/`

Publication Bias



REPRINTED WITH SPECIAL PERMISSION OF NORTH AMERICAN SYNDICATE

Discussion of Quiz

Question 2: Say a very successful entrepreneur named Jamie puts out an autobiography called “How to win at life.” In it, Jamie details a plan to "win" at the various dimensions of life. Jamie states "I followed these steps, and look at me now! You should do the same!" Critique this statement keeping the comic in mind.

There might have been 9999 people who did the same things but perhaps aren’t as successful. Those people generally don’t get book deals so we don’t know about them.

Interpreting Confidence Intervals

Page 180: They use the term "we are 95% confident that the population parameter is between . . . ".

- This is shorthand for "if we repeat this procedure 100 times, then we expect 95 . . . "
- and not "there is a 95% probability the CI contains the population parameter."
- Think as "using a procedure that is 95% reliable, this interval is where we think the population parameter is."

Analysis of Variance (ANOVA)

A farmer wants to compare the performance of 4 fertilizers in terms of tomato yield. We have $k = 4$ groups AKA levels of a factor.

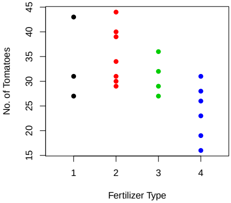
- They assign n_i plants to each of the $k = 4$ fertilizers:

n_1	n_2	n_3	n_4	total n
3	7	4	6	20

- They count the number of tomatoes on each plant

Tomato Fertilizer

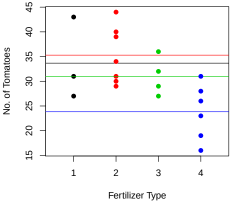
They compare the performance in terms of # of tomatoes yielded.



Tomato Fertilizer

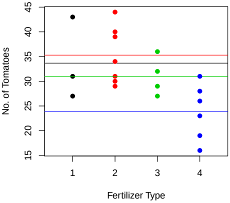
They compare the performance in terms of # of tomatoes yielded.

Plot the sample mean of each level.



Tomato Fertilizer

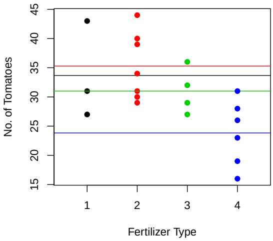
They compare the performance in terms of # of tomatoes yielded. Plot the sample mean of each level. Question: are the mean tomato yields different?



Analysis of Variance

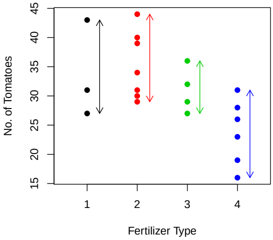
Tomato Fertilizer Example

Numerator: the between-group variation refers to the variability between the levels (the 4 horizontal lines):



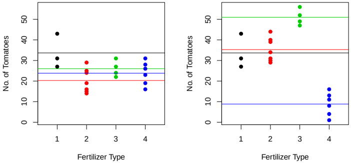
Tomato Fertilizer Example

Denominator: the within-group variation refers to the variability within each level (the 4 vertical arrows):



Tomato Fertilizer Example

Now compare the following two plots. Which has “more different” means?

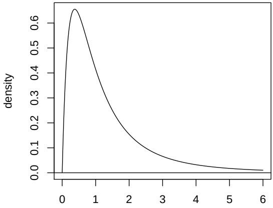


Tomato Fertilizer Example

F Distributions

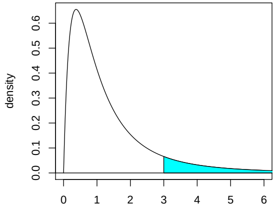
F Distributions

For $df_1 = 4$ and $df_2 = 6$, the F distribution looks like:



F Distributions

p-values are computed where "more extreme" means larger. Say the $F = 3$, the p-value is the area to the right of 3.



Conducting An F-Test

The results are typically summarized in an ANOVA table:

Source of Variation	df	SS	MS	F	p-value
Between groups	$k-1$	$SSTr$	$MSTr = \frac{SSTr}{k-1}$	$\frac{MSTr}{MSE}$	p
Within groups	$n-k$	SSE	$MSE = \frac{SSE}{n-k}$		
Total	$n-1$	SST			

Conditions

1. The observations have to be independent. 10% rule.
2. Trade off of n and normality of observations within each group.
3. Each of the groups has constant variance $\sigma_1^2 = \ldots = \sigma_k^2 = \sigma^2$. Check via:
    - boxplots
    - comparing the sample standard deviations $s_1, \ldots, s_k$