# Lecture 19: ANOVA Part I

Chapter 5.5

# Previously: Conditions for Using t Distribution

We use the $t$ distribution when you have

- $n$ is small. E.g. 3, 5, 10, 15.
- Independence: $n \leq 10\%$ rule
- Observations come from a nearly normal distribution:
  - Look at a histogram of the data (difficult when $n$ is small)
  - Consider whether any previous experiences alert us that the data may be normal

# New Topic: Analysis of Variance (ANOVA)

A farmer has the choice of four tomato fertilizers and wants to compare their performance in terms of crop yield.

We have $k = 4$ groups AKA levels of a factor: the 4 types of fertilizer. Say we:

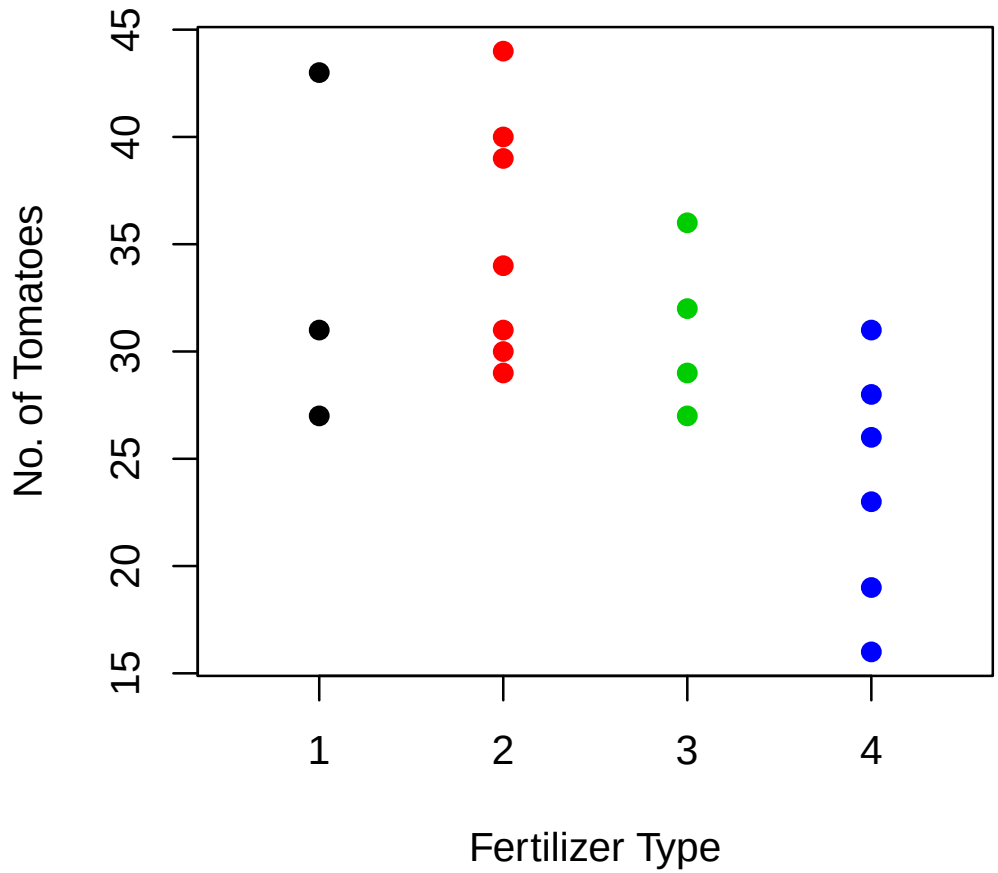
- assign $n_i$ plants to each of the $k = 4$ fertilizers as follows:

| $n_1$ | $n_2$ | $n_3$ | $n_4$ | total $n$ |
|---|---|---|---|---|
| 3 | 7 | 4 | 6 | 20 |

- we evaluate the number of tomatoes on each plant

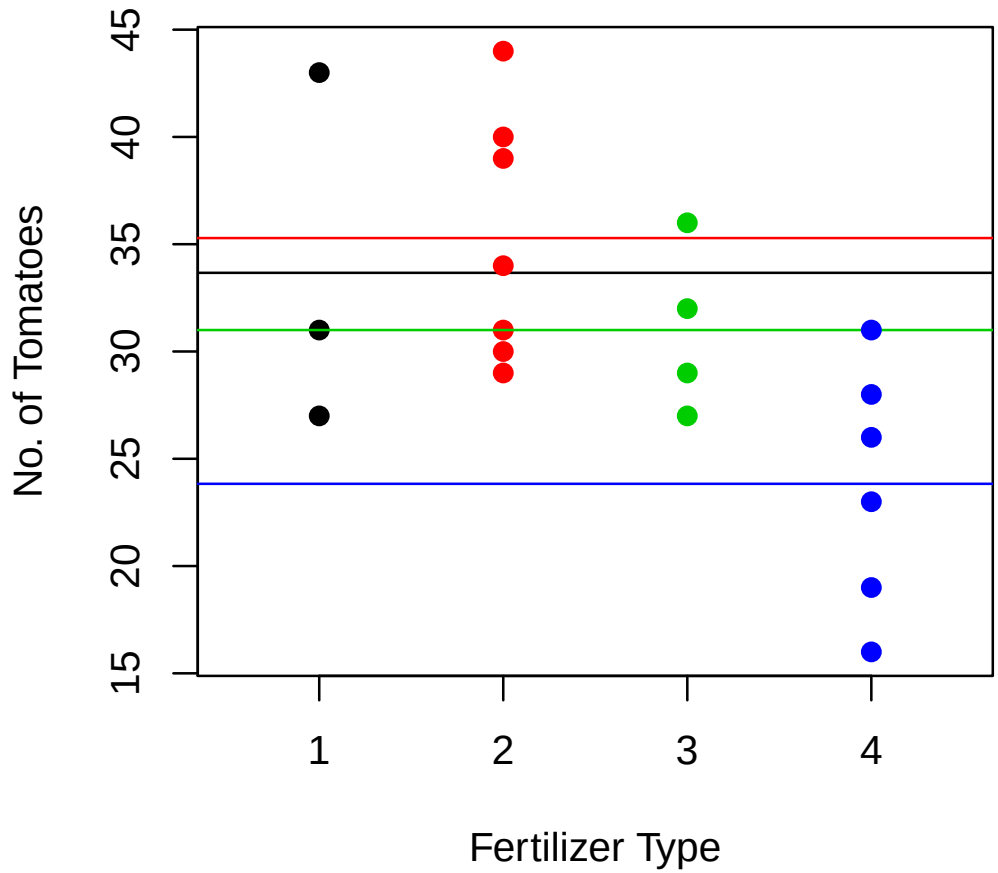
# Tomato Fertilizer

We observe the following, where each point is one tomato plant.



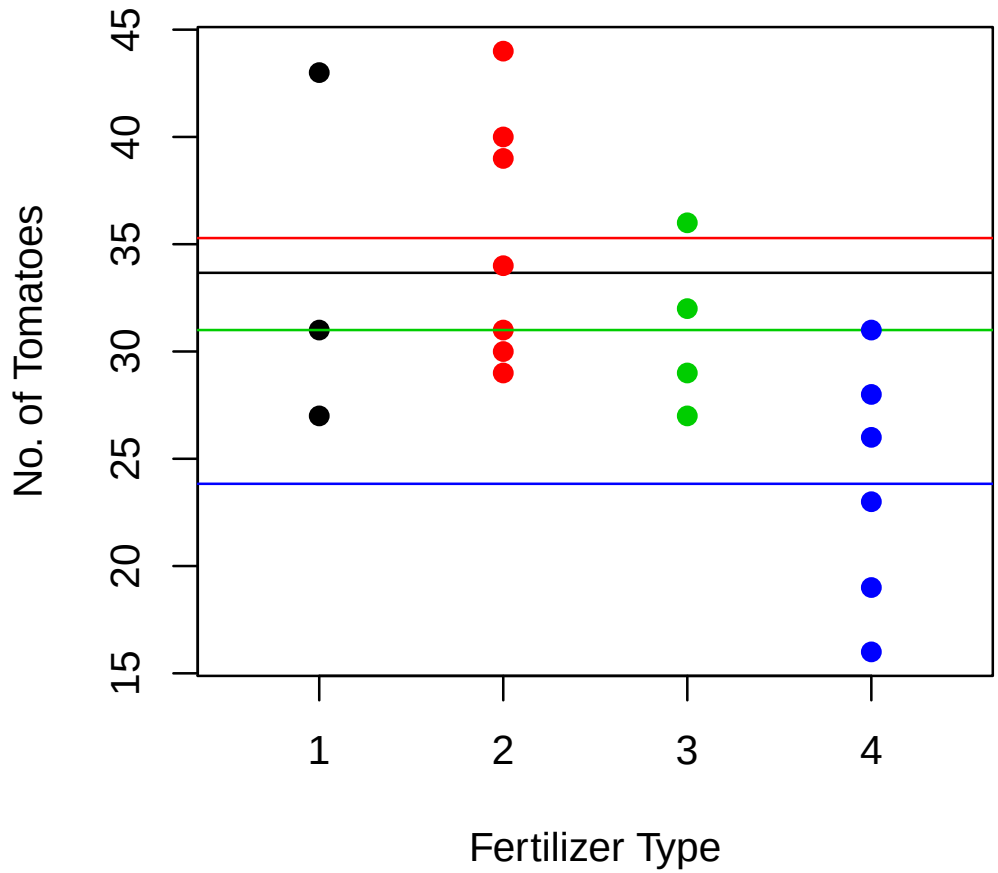
# Tomato Fertilizer

We observe the following, where each point is one tomato plant. Plot the sample mean of each level.



# Tomato Fertilizer

We observe the following, where each point is one tomato plant. Plot the sample mean of each level. Question: are the mean tomato yields different?



# Analysis of Variance

Say we have $k$ groups and want to compare the $k$ means:

$$\mu_1, \mu_2, \ldots, \mu_k$$

We could do $\binom{k}{2}$ individual two-sample tests. Ex. for groups 1 & 2:

$$\begin{aligned} H_0: &\quad \mu_1 = \mu_2 \\ \text{vs. } H_a: &\quad \mu_1 \neq \mu_2 \end{aligned}$$

i.e. no difference in means

# Analysis of Variance

Or, rather than conducting all $\binom{k}{2}$ tests, we do a single overall test via Analysis of Variance ANOVA:

The hypothesis test is:

$$\begin{aligned} H_0 &: \quad \mu_1 = \mu_2 = \ldots = \mu_k \\ \text{vs. } H_a &: \quad \text{at least one of the } \mu_i\text{'s are different} \end{aligned}$$

# How ANOVA Tests Work

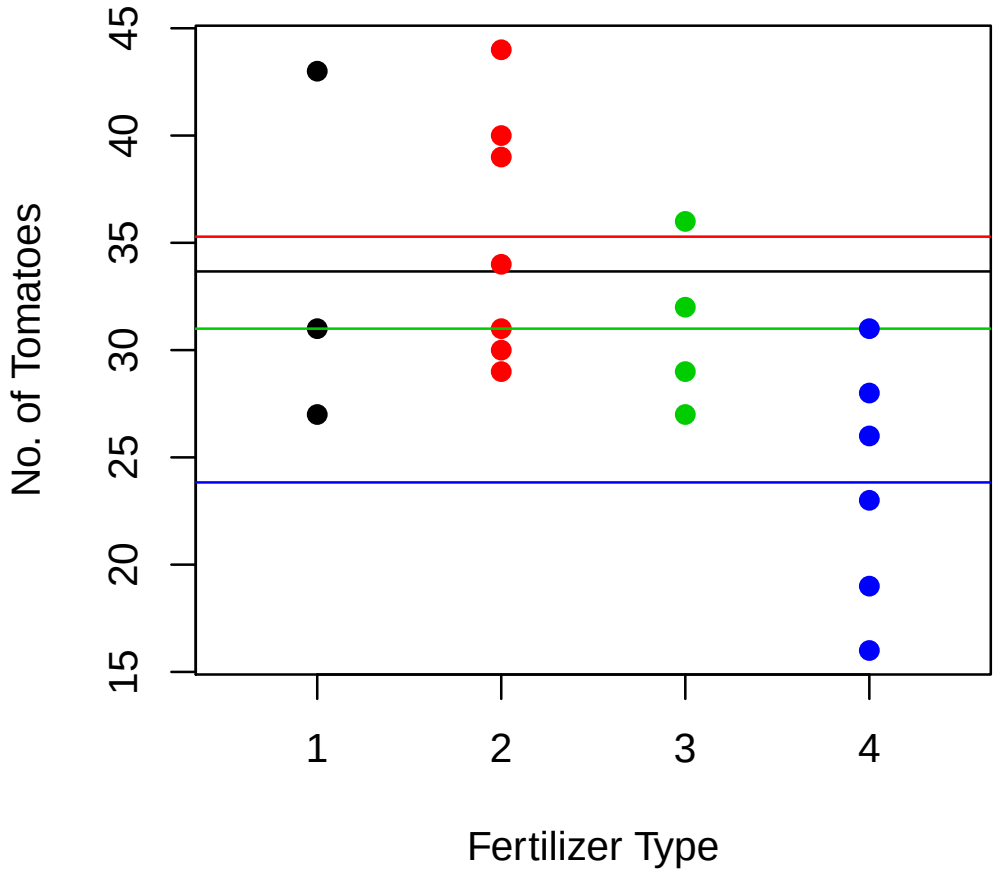
ANOVA asks: where is the overall variability of the data originating from?

The test statistic used to compute a $p$-value is now the F-statistic:

$$F = \frac{\text{measure of between-group variability}}{\text{measure of within-group variability}}$$

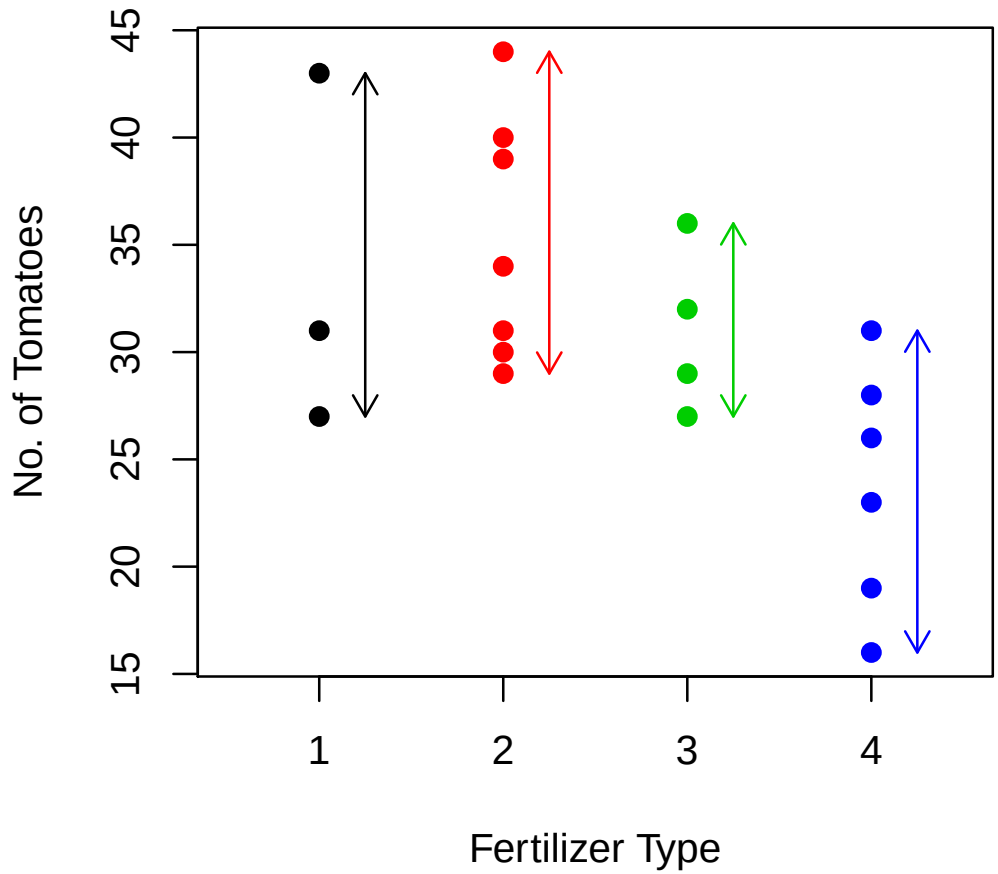
# Tomato Fertilizer Example

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



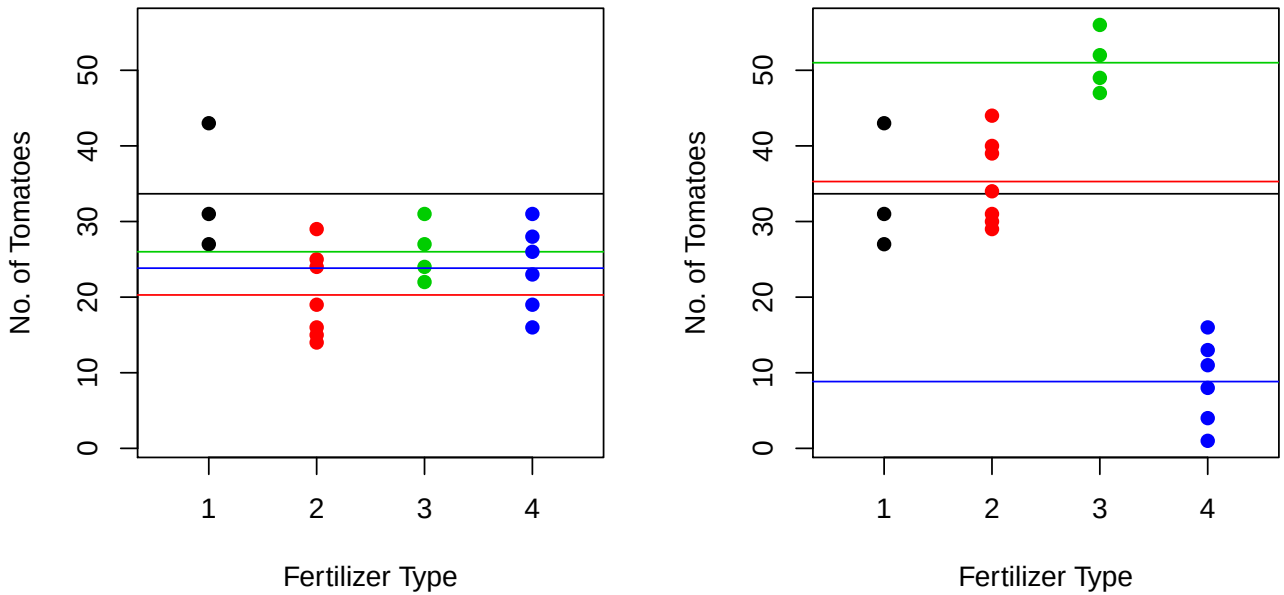
# Tomato Fertilizer Example

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



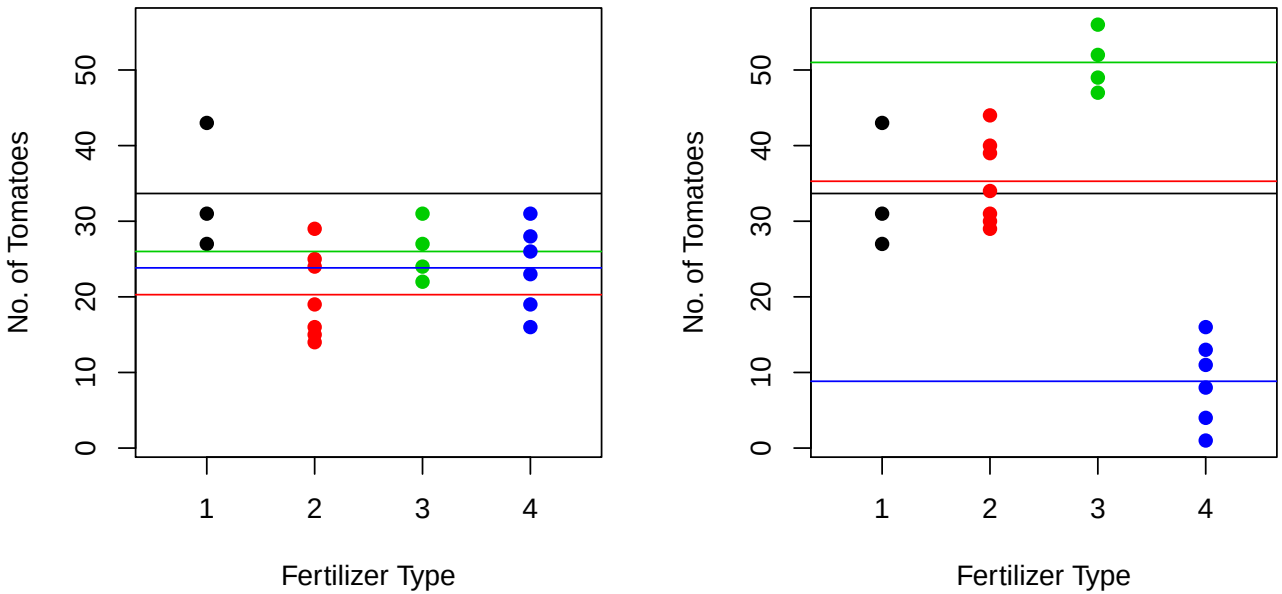
# Tomato Fertilizer Example

Now compare the following two plots:




- They have the same within-group variability. Call this value $W$
- The right plot has higher between group variability b/c the 4 means are more different. Call these values $B_{left}$ and $B_{right}$ with $B_{left} < B_{right}$

# Tomato Fertilizer Example




$$\text{Recall } F = \frac{\text{measure of between-group variability}}{\text{measure of within-group variability}}$$

$$\begin{aligned} \text{Since} \quad \frac{B_{left}}{W} &< \frac{B_{right}}{W} \\ \text{thus} \quad F_{left} &< F_{right} \end{aligned}$$

# $F$ Distributions

Assuming $H_0$ is true (that $\mu_1 = \mu_2 = \ldots = \mu_k$), the $F$-statistic

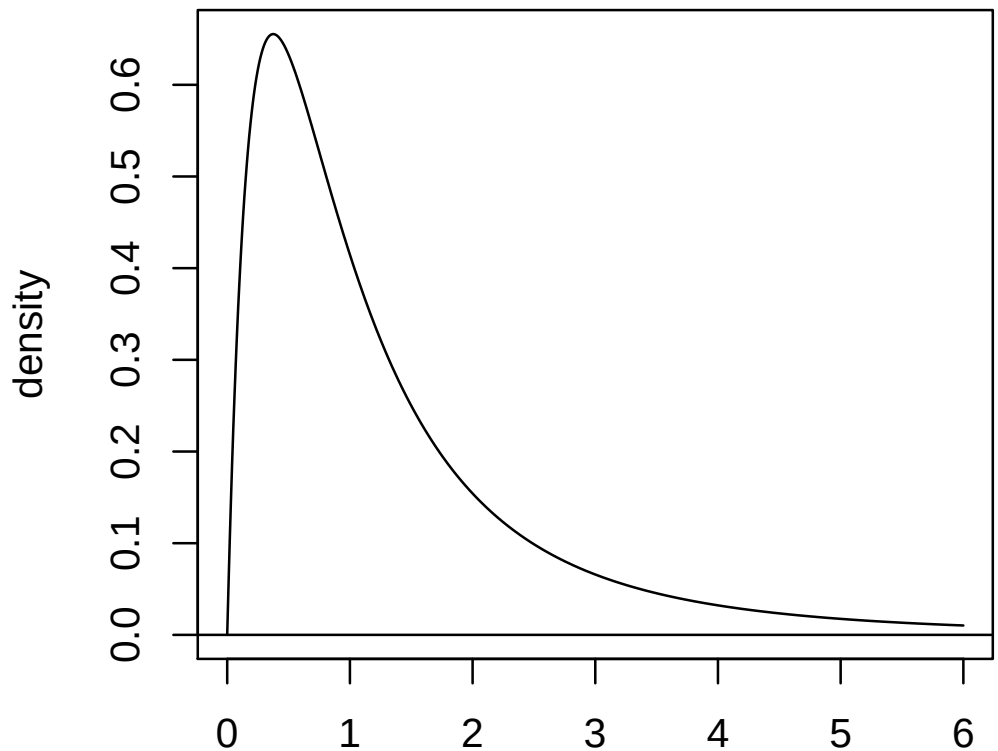
$$F = \frac{\text{measure of between-group variability}}{\text{measure of within-group variability}}$$

follows the $F$ distribution with degrees of freedom $df_1 = k - 1$ and $df_2 = n - k$ where

- $n$ is the total number of observations
- $k$ is the number of groups

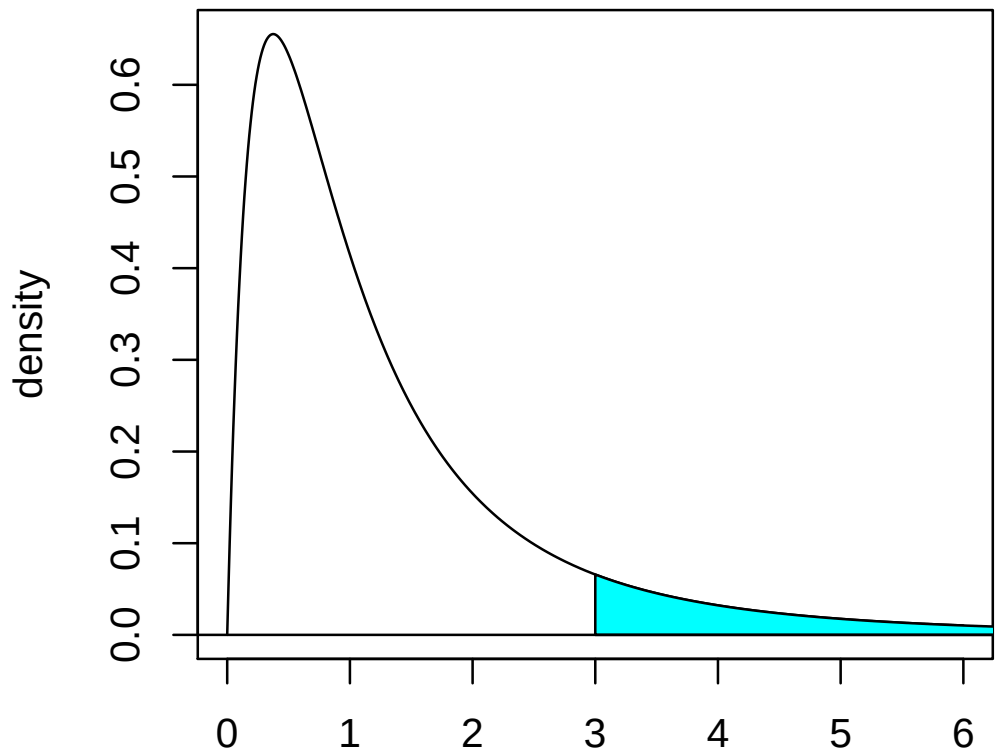
# $F$ Distributions

Much like the $t$ distribution with degrees of freedom $df$, the parameters for the $F$ distribution are $df_1 = k - 1$ and $df_2 = n - k$. Example with $df_1 = 4$ and $df_2 = 6$:



# $F$ Distributions

$p$-values are computed as before where "more extreme" means larger. You can compute these in `R` using `pf(F,df1,df2)`. Say the $F$-statistic is equal to 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. If the sample sizes are small within each group, normality of the data is important. If not small, we can be lax about this.
3. Each of the groups has constant variance $\sigma_1^2 = \ldots = \sigma_k^2 = \sigma^2$. We can check this with boxplots and by comparing the sample standard deviations $s_1, \ldots, s_k$

# Discussion of Yesterday's Quiz

Question 1: Why did 1 out of the 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 $\alpha = 0.05$ (significance level AKA type I error rate AKA false positive rate)
i.e. we expect 1 out of 20 results to be significant

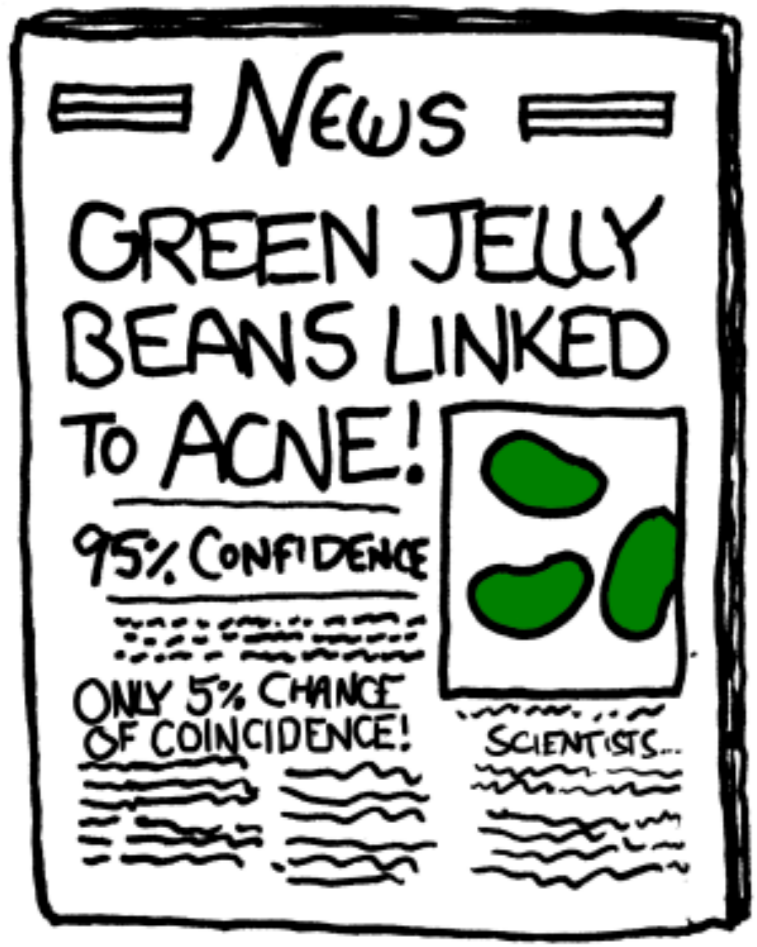
# Publication Bias

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

To counter this, some prominent medical journals (incl. the New England Journal of Medicine, The Lancet, Annals of Internal Medicine, and JAMA) require registration of a trial before it starts so that unfavorable results are not withheld from publication.

http://www.jnrbm.com/content/10/1/6

# Publication Bias



# Publication Bias




From: Sterne JA, Davey Smith G (2001) Sifting the evidence - What's wrong with significance tests. BMJ 322: 226231.

# Multiple Testing

A related issue is the statistical concept of multiple testing.

Say the null is true (i.e. nothing is going on). If you repeat the experiment more and more times, you're bound to get a significant result eventually just by chance alone.

# Multiple Testing

You conduct $n = 20$ tests of different jelly beans colors at the $\alpha = 0.05$ significance level. We expect 5% of them to be significant by chance alone.

If the tests are independent, the number of significant tests is Binomial($n = 20, p = \alpha = 0.05$).

We saw in Chapter 3

$$\mu = np = 20 \times 0.05 = 1 \text{ test}$$

# Multiple Testing

What is

$$\begin{aligned} P(\text{at least one sig result in 20}) &= 1 - P(\text{no sig results in 20}) \\ &= 1 - \binom{20}{0}\alpha^0(1-\alpha)^{20} \\ &= 1 - (1-0.05)^{20} \\ &\approx 0.64 \end{aligned}$$

We have 64% chance of observing at least one significant result, even if "nothing is going on."

Why? Because we did so many tests! That is a huge chance of a false positive!

# Multiple Testing

What do people do? Make the $\alpha$ stricter! i.e.

- make the $\alpha$ smaller
- so we have less chance the p-value is smaller than $\alpha$
- so we have less chance of incorrectly rejecting the null when it is true

Bonferroni correction: If you are conducting $n$ tests, use $\alpha^* = \frac{\alpha}{n}$

# Multiple Testing

So in our case, use $\alpha^* = \frac{0.05}{20} = 0.0025$. Now

$$\begin{aligned} P(\text{at least one sig result in 20}) &= 1 - P(\text{no sig results in 20}) \\ &= 1 - \binom{20}{0}\alpha^{*0}(1-\alpha^*)^{20} \\ &= 1 - (1 - 0.0025)^{20} \\ &\approx 0.0488 \end{aligned}$$

Closer to true desired $\alpha = 0.05$.

Note: the correction is conservative in that the overall type I error rate is $0.0488 < 0.05$.

# What $\alpha$ To Set for Single Tests

About using $\alpha = 0.05$ as your significance level. Before using it, at least put some thought into the balance between:

- Type I errors. If type I errors matter more, set $\alpha$ small.
- Type II errors. If type II errors matter more, set $\alpha$ high.