ANOVA: Assumptions and miscellany

Research Methods for Human Inquiry
Andrew Perfors

This lecture:

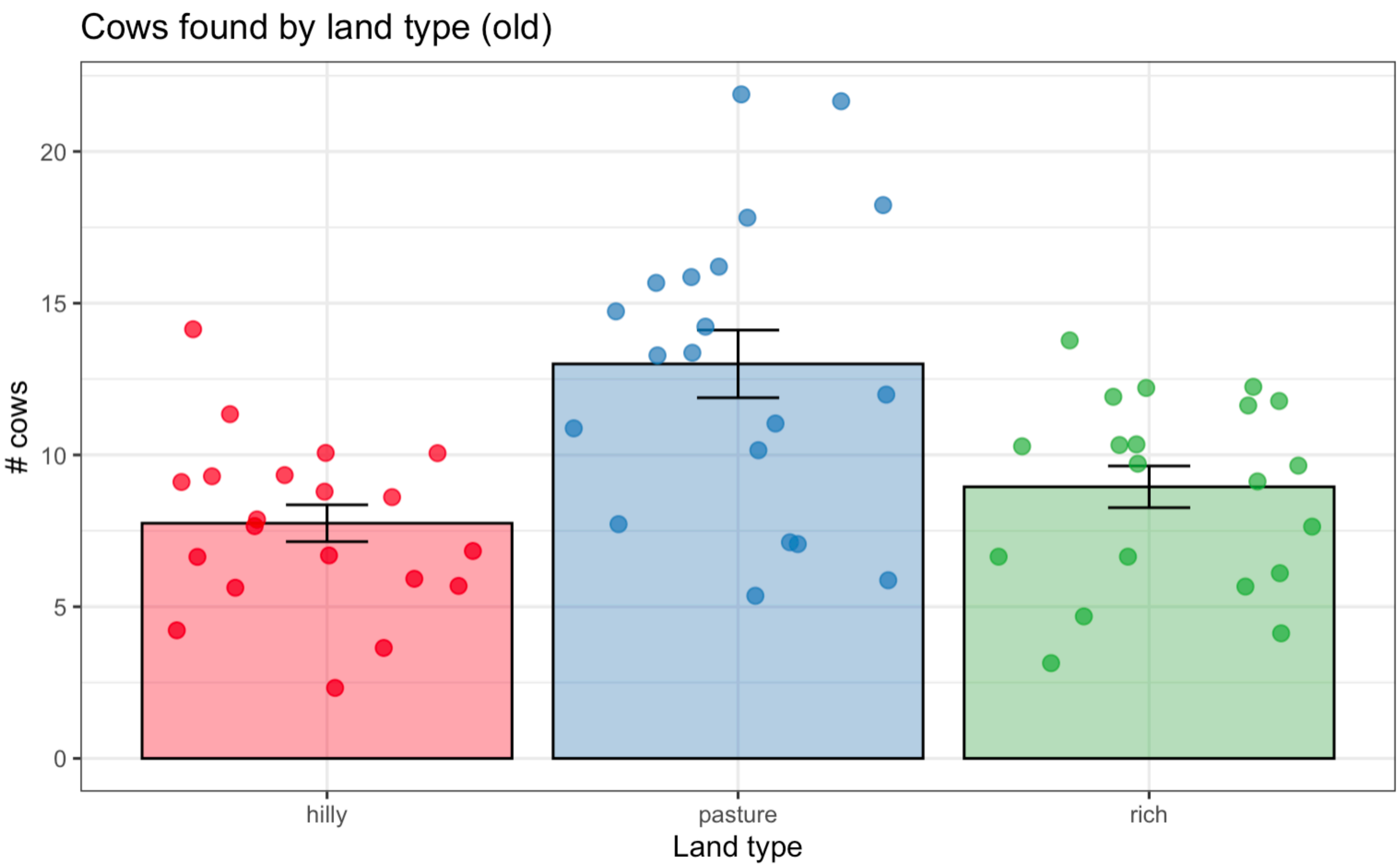
- Post-hoc tests
- What assumptions does ANOVA rely on?
- How do we check these assumptions?
- (How do we fix it if the assumptions are wrong)

ANOVA is unsatisfying

- The problem...
  - Our ANOVA tells us that 15 years go the # of cows was significantly different on different types of land
  - It doesn't tell us which types of land (if any) have significantly more cows

- Usually we do want to know which groups are different to one another, and which ones aren't

Solution #1: Draw a picture



Solution #2: Run lots of t-tests...

- We need one t-test for every pair of groups
  - pasture v rich
  - pasture v hilly
  - rich v hilly

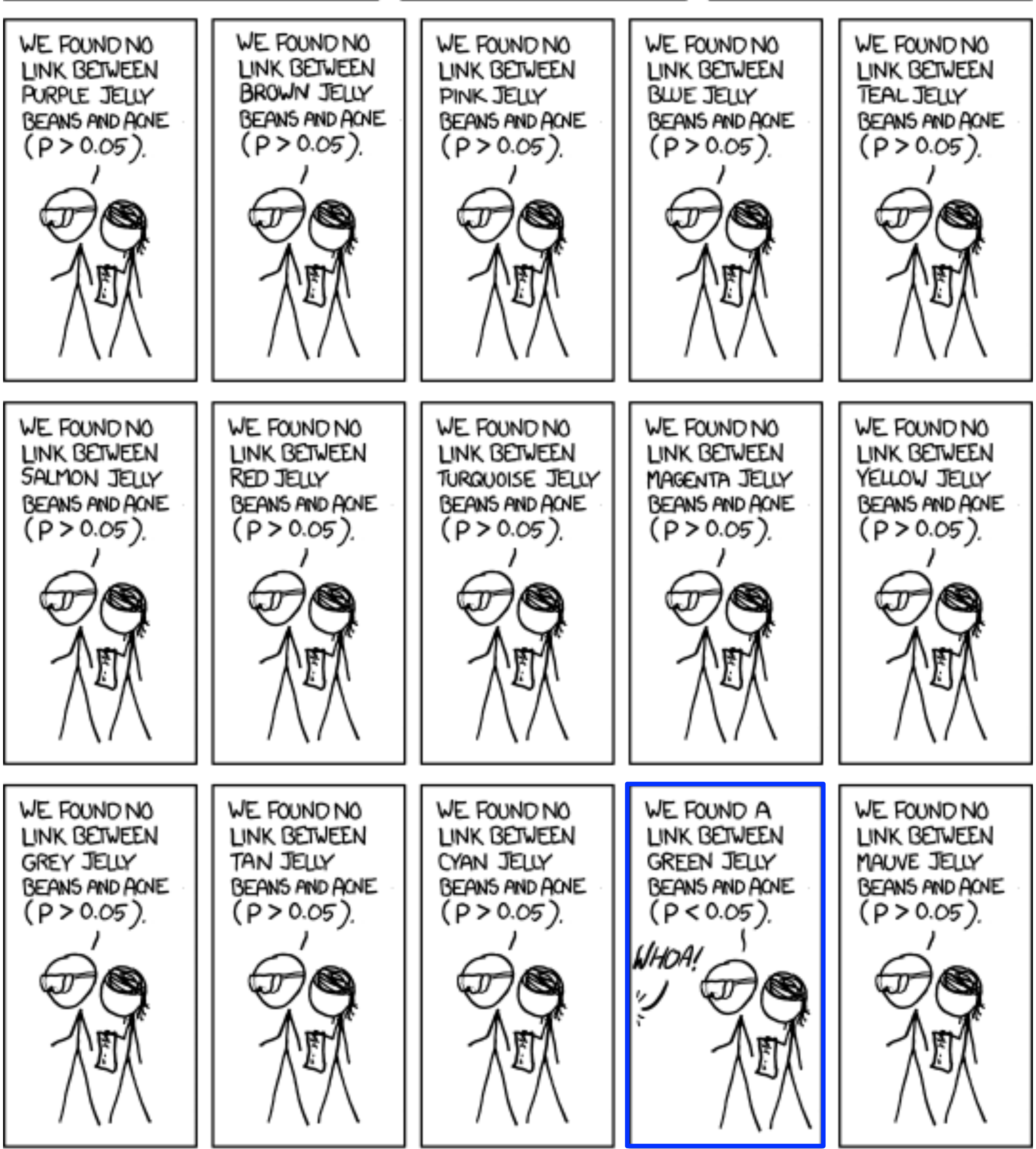
- Not too bad if there are only 3 groups, but what if you have more than that...

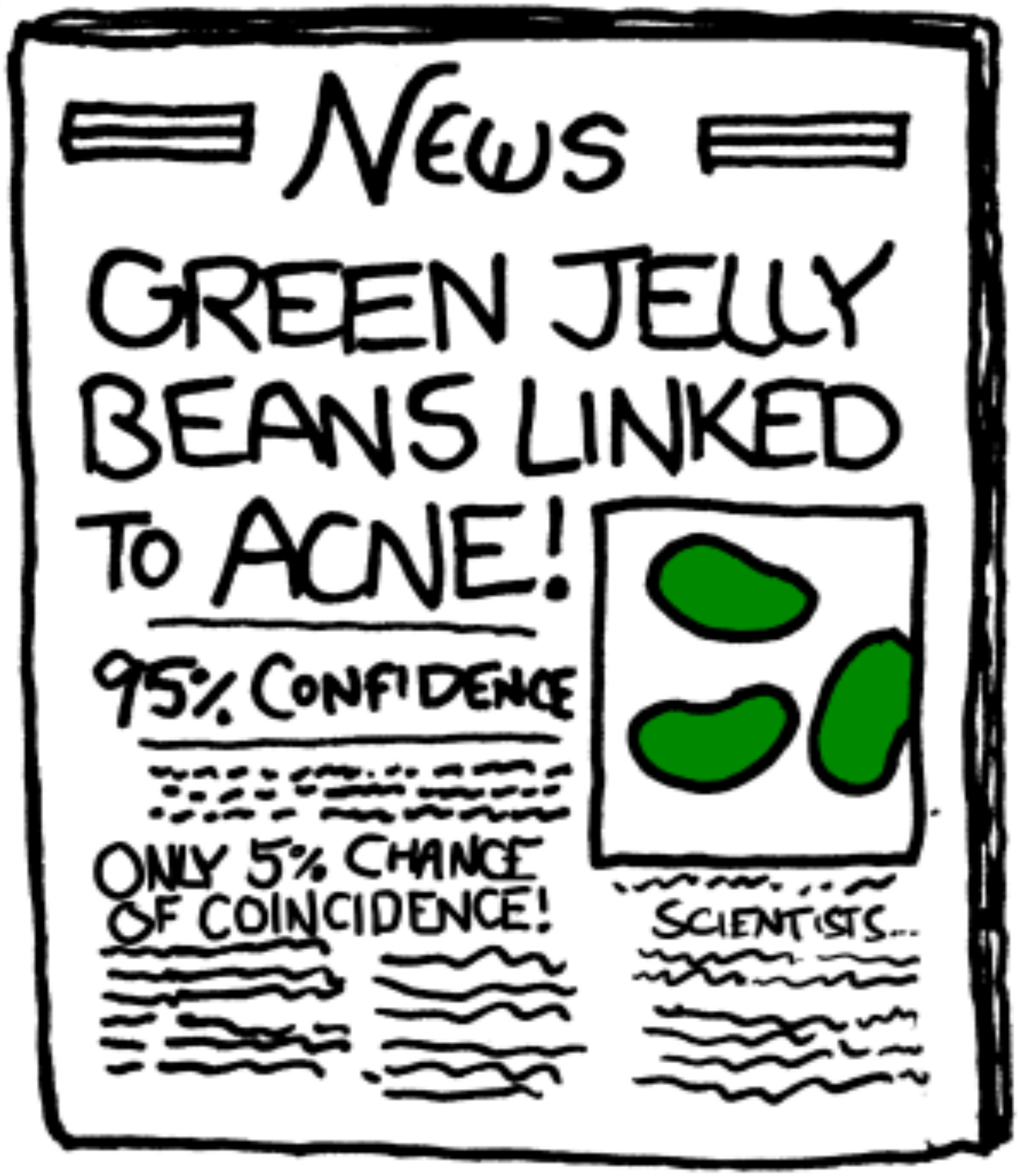
Solution #2: Run lots of t-tests...

- Suppose we have 6 groups: pasture, hilly, rich, forest, urban, desert
- We need 15 tests...
  - (pasture v hilly), (pasture v urban),
  - (pasture v rich), (pasture v desert),
  - (pasture v forest), (hilly v urban),
  - (hilly v rich), (hilly v desert),
  - (hilly v forest), (urban v rich),
  - (urban v desert), (urban v forest),
  - (rich v desert), (rich v forest),
  - (desert v forest)
- The number of tests needed rapidly gets large!

Does this matter?

- Yes, it does.
  - Remember the central goal of hypothesis testing is to control the Type 1 error rate at 5% (for instance)
  - Each individual t-test has a 5% Type I error rate.
  - If you're running lots of t-tests, then the probability of getting *at least one* Type I error is now much larger than 5%.
  - Or, to explain this by way of an XKCD comic...


WE FOUND NO LINK BETWEEN PURPLE JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN BROWN JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN PINK JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN BLUE JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN TEAL JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN SALMON JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN RED JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN TURQUOISE JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN MAGENTA JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN YELLOW JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN GREY JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN TAN JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN CYAN JELLY BEANS AND ACNE (P > 0.05).
WE FOUND A LINK BETWEEN GREEN JELLY BEANS AND ACNE (P < 0.05).
WHOA!
WE FOUND NO LINK BETWEEN MAUVE JELLY BEANS AND ACNE (P > 0.05).


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

Not exactly convincing, is it?

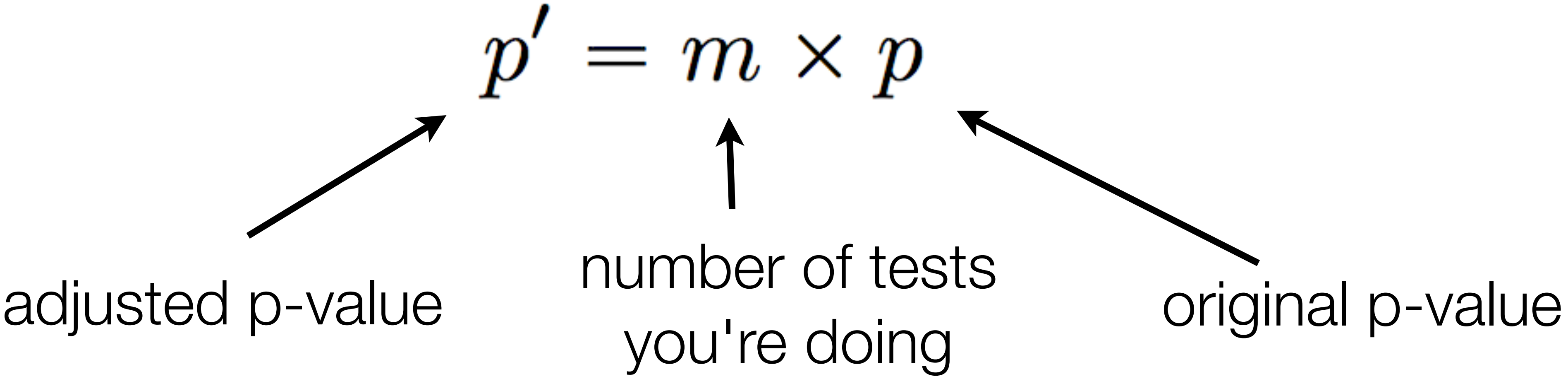
Corrections for multiple comparisons

- The <u>family wise</u> Type I error rate is the probability of obtaining at least one Type I error across multiple tests

- If you want to keep the family-wise error rate at 5%, then you need to "adjust" the raw p-values using some method

The Bonferroni correction

- The simplest way to do this is to multiply all your original p-values by the number of tests...



$$p' = m \times p$$

adjusted p-value

number of tests you're doing

original p-value

- This works just fine, but it's very conservative, meaning that you lose a *lot* of power relative to more sophisticated methods.

The Holm correction

- Superior to Bonferroni: same Type I error risk, but lower Type II error risk
- Sorts the p-values from lowest to highest, and adjust each one as follows:

Lowest one: $p' = m * p$
2nd-lowest one: $p' = (m-1) * p$
3rd-lowest one: $p' = (m-2) * p$
...
Highest one: $p' = p$

- Stops once it gets to one that it can't reject, and retains all of the ones that have higher p-values

How to do it in R

- Use the `posthocPairwiseT()` function [`lsr` package]
- Two arguments:
  - `x`: the aov object
  - `p.adjust.method`: text indicating what correction to use (e.g., "none", "bonferroni", "holm"). The default is the Holm.

(There are also other functions, like `PostHocTest()` in `DescTools`, that will do this. I’m teaching the `lsr` one because it’s most straightforward (and nicely does Holm) but it is not as general — these other functions have many other kinds of correction as well)

No corrections

```
> posthocPairwiseT( x = cows1waynewModel, p.adjust.method = "none" )

   Pairwise comparisons using t tests with pooled SD

data:  cows and type

        hilly   pasture
pasture 3.9e-05 -
rich    0.3126  0.0011

P value adjustment method: none
```

Bonferroni correction

```
> posthocPairwiseT( x = cows1waynewModel, p.adjust.method = “bonferroni" )

    Pairwise comparisons using t tests with pooled SD

data:   cows and type

        hilly   pasture
pasture 0.00012 -
rich    0.93773 0.00330

P value adjustment method: bonferroni
```

Holm correction

```
> posthocPairwiseT( x = cows1waynewModel, p.adjust.method = “holm” )

    Pairwise comparisons using t tests with pooled SD

data:   cows and type

        hilly   pasture
pasture 0.00012 -
rich    0.31258 0.00220

P value adjustment method: holm
```

Note: when you report these (including for this subject), it’s often sufficient to only include p-values and direction (which you get from the figure) plus the correction method. If you need to also report t and df, you will need to calculate them using `t.test()` on different subsets of the data

I suggest using the Holm method if you can. If not, Bonferroni. Most important is that you make some correction!

Some thoughts and terminology

- Terminology...
  - A **post hoc test** refers to a test that you conduct after you've done your ANOVA, and for which you don't have any particular hypotheses... e.g., pairwise t-tests run with no particular plan in mind
  - A **multiple test correction** is a method used to control your overall (e.g. family wise) Type 1 error rate.... e.g., Bonferroni, Holm.

- When you're running post hoc tests, you usually need to apply a multiple test correction

Some thoughts and terminology

- Real life is messy.
  - The XCKD "jelly beans" example feels like definitely needed a multiple test correction.
  - But what if I run a study for which I have 3 specific hypotheses that I want to test, all of which are motivated by a theoretical idea, and all of which I wrote down before running the study? Do I need to make a correction here??? General consensus is "no".

- This is an example of a **planned comparison**...
  - it was always my intention ONLY to look at a few specific cases, so I don't need to make corrections
  - The problem is that in practice, few people actually stick to the plan... so I'm often suspicious. This is where pre-registration is often very helpful.

Assumptions of ANOVA

Assumptions

- Residuals are normally distributed
  - Check using Shapiro-Wilk test
  - If violated, use the Kruskal-Wallis test

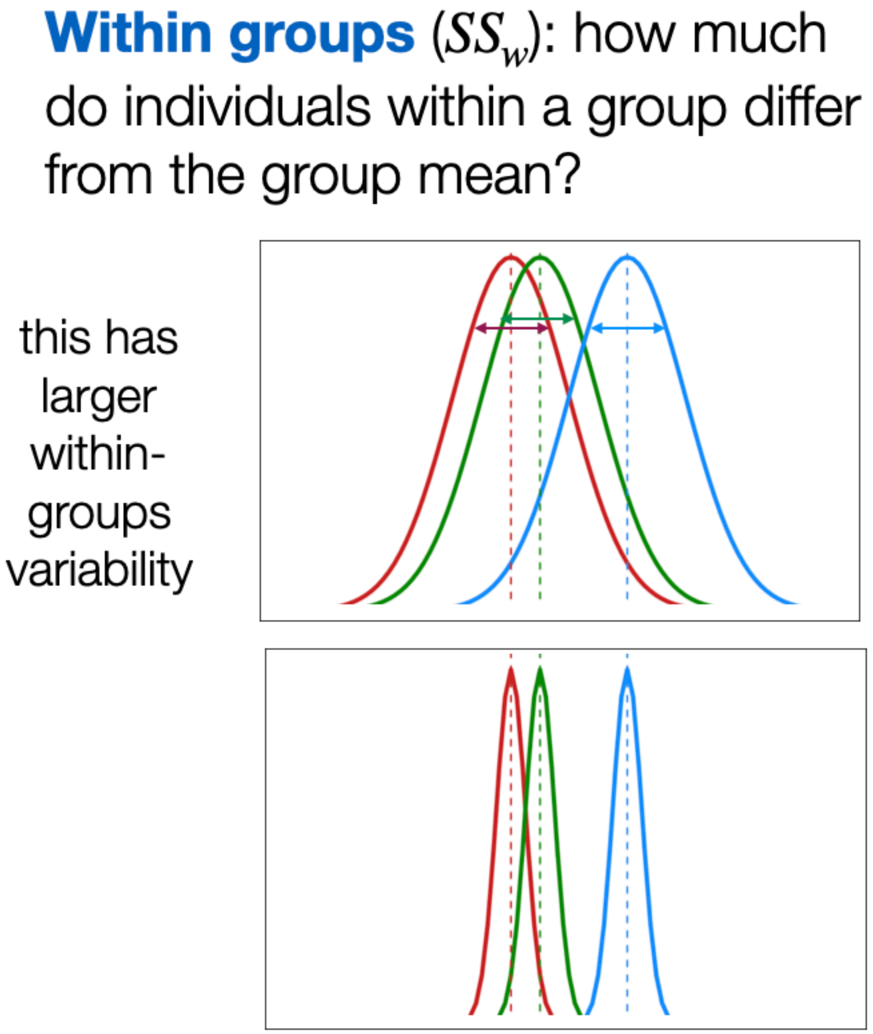
- Homogeneity of variance across all groups
  - Check using Levene's test
  - If violated, use the Welch one way ANOVA

- Independence (we'll talk about this later)

If both are violated, do the Kruskal-Wallis as it doesn't assume homogeneity of variance either!

1. Residuals are normally distributed

Remember that “residuals” is a word for the within-group variance



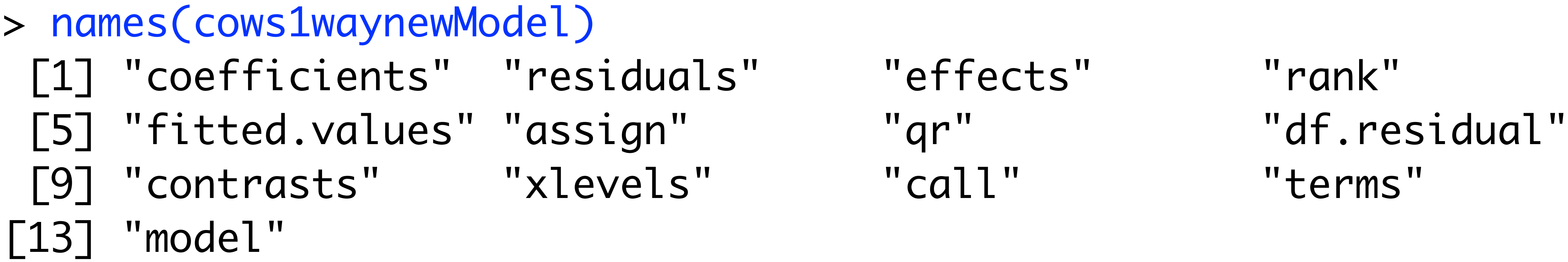
Within groups (SS_w): how much do individuals within a group differ from the group mean?

this has larger within-groups variability

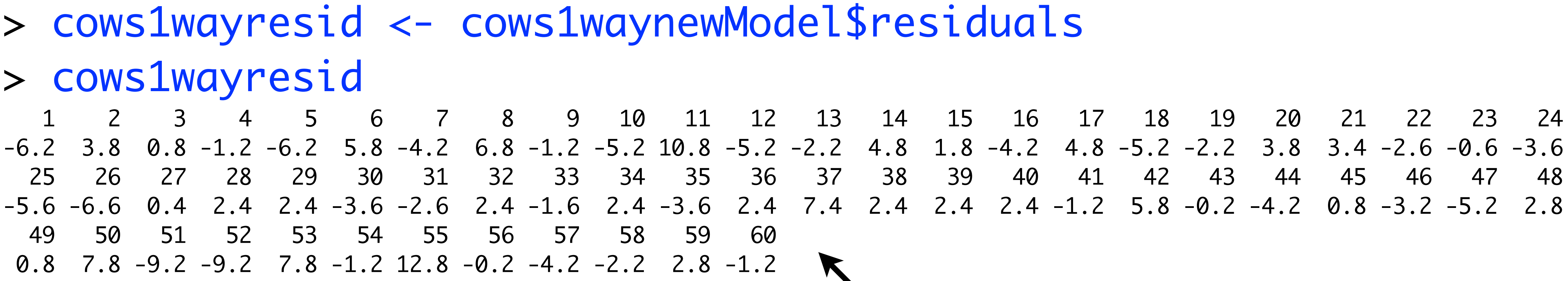
The maths assumes that this variance is normally distributed

1. Residuals are normally distributed

To check this assumption, first we have to get the residuals. We use the fact that the aov object contains a lot of information in it



```
> names(cows1waynewModel)
 [1] "coefficients"  "residuals"     "effects"       "rank"
 [5] "fitted.values" "assign"        "qr"            "df.residual"
 [9] "contrasts"     "xlevels"       "call"          "terms"
[13] "model"
```



```
> cows1wayresid <- cows1waynewModel$residuals
> cows1wayresid
   1    2    3    4    5    6    7    8    9   10   11   12   13   14   15   16   17   18   19   20   21   22   23   24
-6.2  3.8  0.8 -1.2 -6.2  5.8 -4.2  6.8 -1.2 -5.2 10.8 -5.2 -2.2  4.8  1.8 -4.2  4.8 -5.2 -2.2  3.8  3.4 -2.6 -0.6 -3.6
  25   26   27   28   29   30   31   32   33   34   35   36   37   38   39   40   41   42   43   44   45   46   47   48
-5.6 -6.6  0.4  2.4  2.4 -3.6 -2.6  2.4 -1.6  2.4 -3.6  2.4  7.4  2.4  2.4  2.4 -1.2  5.8 -0.2 -4.2  0.8 -3.2 -5.2  2.8
  49   50   51   52   53   54   55   56   57   58   59   60
 0.8  7.8 -9.2 -9.2  7.8 -1.2 12.8 -0.2 -4.2 -2.2  2.8 -1.2
```

For each of the 60 datapoints, it contains the variance that wasn't accounted for by the group

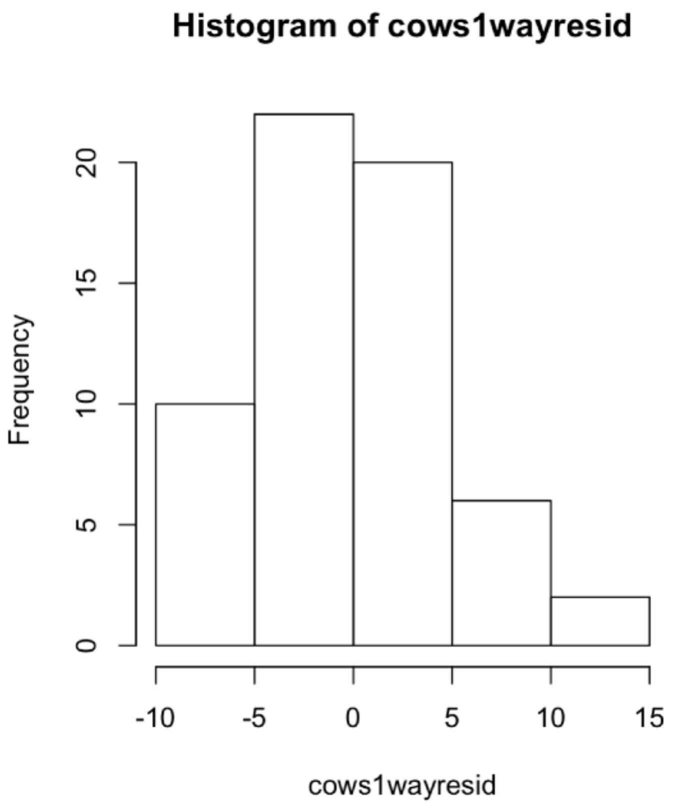
1. Residuals are normally distributed

We want to check if these are normal, so we can use the same techniques we already have!

```
> shapiro.test(x=cows1wayresid)

	Shapiro-Wilk normality test

data:  cows1wayresid
W = 0.98008, p-value = 0.4317
```



P-value is not significant, which suggests that the residuals are normally distributed

But what if the residuals *aren’t* normally distributed?

We use a test called the Kruskal-Wallis test, which is very similar in the basic idea as the Wilcoxon test: it rank orders the data and conducts the analysis on the ranks

```
> kruskal.test(cows ~ type, data=d_old)

	Kruskal-Wallis rank sum test

data:  cows by type
Kruskal-Wallis chi-squared = 13.653, df = 2, p-value = 0.001085
```

But what if the residuals *aren't* normally distributed?

Instead of using η^2 for effect size, we need to calculate the non-parametric equivalent

```
> library(rstatix)
> kruskal_effsize(cows ~ type, data=d_old)

  .y.       n effsize method  magnitude
* <chr> <int>   <dbl> <chr>   <ord>
1 cows     60   0.204 eta2[H] large
```

Interpreted similarly to η^2, as % of variance accounted for. So this suggests that land type accounts for 20.4% of the variance in # of cows

Assumptions

- Residuals are normally distributed
  - Check using Shapiro-Wilk test
  - If violated, use the Kruskal-Wallis test

- Homogeneity of variance across all groups
  - Check using Levene's test
  - If violated, use the Welch one way ANOVA

- Independence (we’ll talk about this later)

If both are violated, do the Kruskal-Wallis as it doesn’t assume homogeneity of variance either!

The Levene test

Used to check whether the different groups have the same standard deviation (i.e., whether variance is homogeneous).

```
> library(car)
> leveneTest(cows ~ type, data=d_old)

Levene's Test for Homogeneity of Variance (center = median)
```

```
      Df F value  Pr(>F)
group  2  4.1716 0.02038 *
      57
```

This output is just an abbreviated ANOVA table... a significant result means that the groups have unequal variance, and therefore your assumptions are violated

What do we do if the variance isn’t homogeneous?

There’s an analogue of the Welch t-test called the Welch one-way ANOVA that doesn’t assume homogenous variance

```
> oneway.test(cows ~ type, data = d_old)

    One-way analysis of means (not assuming equal variances)

data:  cows and type
F = 8.4063, num df = 2.000, denom df = 36.359, p-value = 0.000998
```

Exercises are in w8day2exercises.Rmd