Oneway ANOVA: follow-up tests & statistical power

Week 7

Last week

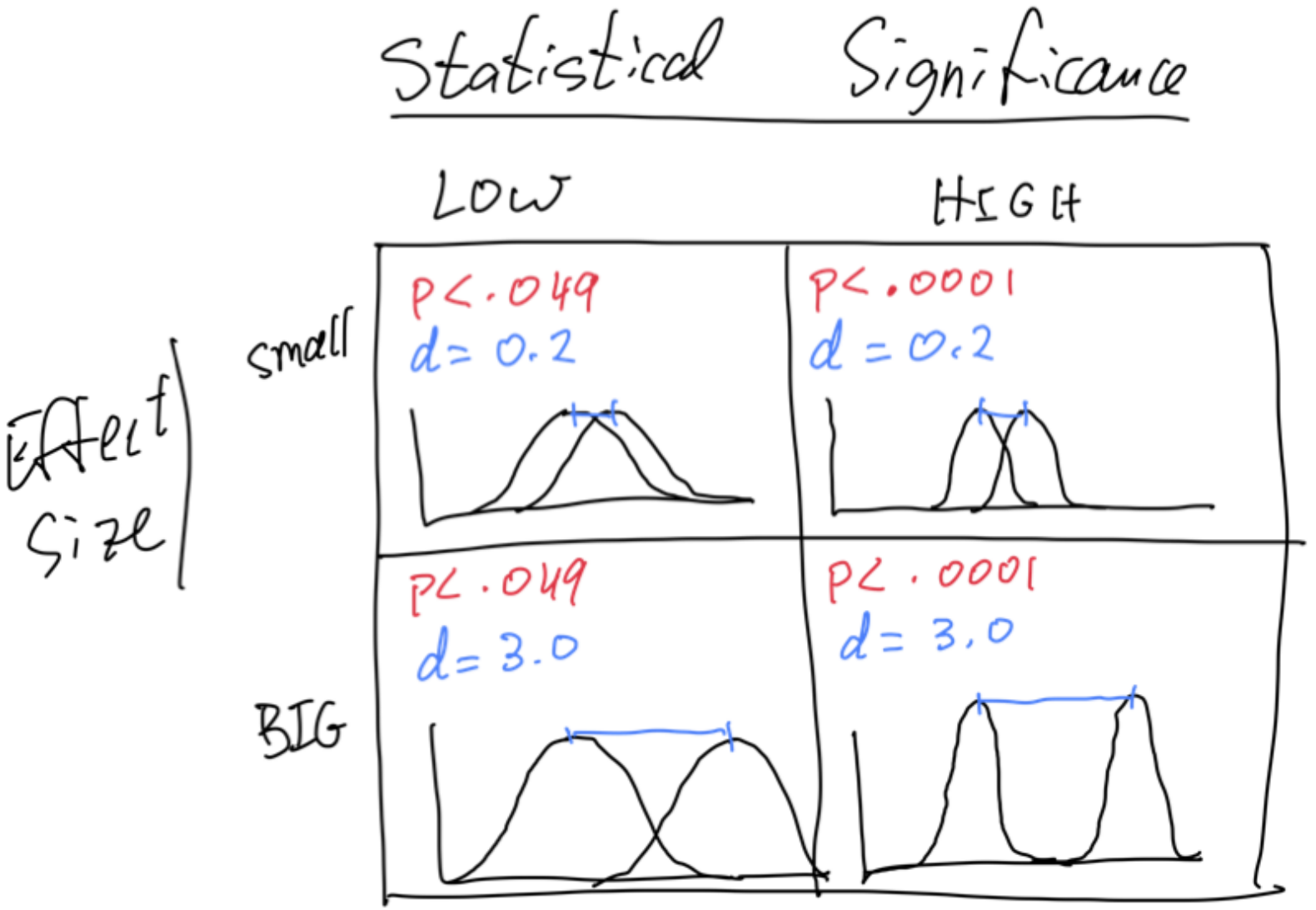
Statistical Significance vs Effect Size

- **statistical significance** is given by the p-value
  - probability of observing a difference as large as the one we observed, under the null hypothesis H_0 (*difference between sample means due only to random sampling from same population*)
  - it is a metric of our confidence in the null hypothesis

Statistical Significance vs Effect Size

- **effect size** gives us a measure of the size of the difference between groups
  - how much larger is the mean of one group compared to the mean of another group, compared to the variability within each group?
  - it is a metric of the strength of the effect

Statistical Significance vs Effect Size



Effect Size: η^2

- η^2 (eta-squared) is one measure of effect size
- $\eta^2 = \frac{SS_{between}}{SS_{total}}$
- η^2 is a proportion of the total variance that a factor explains
  - ranges between 0 and 1 (just like R^2 in regression)
- η^2 is a measure of the strength of the effect
- `eta_quared()` function in the `effectsize` package

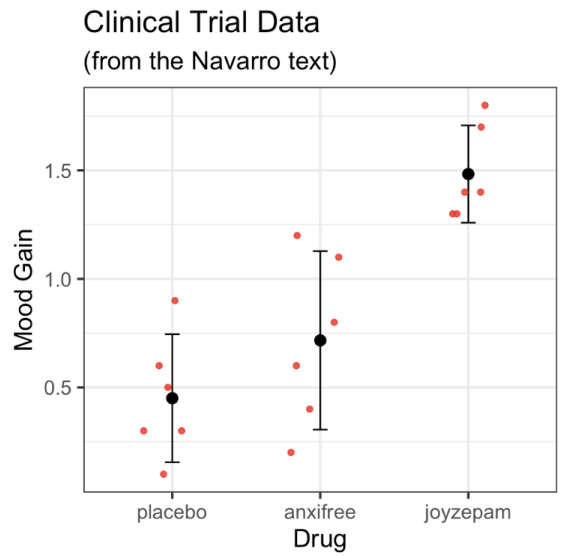
Effect Size: ω^2 & Cohen's d

- some researchers prefer ω^2 (omega-squared)
  - it can be less biased than η^2 for small samples
  - in R: `library(effectsize)` and `omegaSquared()` function
- Cohen's d is another measure of effect size
  - it is the ratio of difference in means to the standard deviation of the groups
  - in R: library(`effectsize`) and `cohens_d()` function

ANOVA: Omnibus F-test

```
1 my.anova <- aov( mood.gain ~ drug, data = clin.trial )
2 summary(my.anova)
```

```
            Df Sum Sq Mean Sq F value   Pr(>F)
drug         2  3.453  1.7267   18.61 8.65e-05 ***
Residuals   15  1.392  0.0928
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```



- which means are different?
- we can conduct “post-hoc” tests
- like a series of pairwise t-tests
  - (with some slight changes)

ANOVA: Follow-up tests

Your competitor's drug (Anxifree) is no better than a placebo (i.e., $\mu_A = \mu_P$)

Your drug (Joyzepam) is no better than a placebo (i.e., $\mu_J = \mu_P$)

Anxifree and Joyzepam are equally effective (i.e., $\mu_J = \mu_A$)

possibility:	is $\mu_P = \mu_A$?	is $\mu_P = \mu_J$?	is $\mu_A = \mu_J$?	which hypothesis?
1	✓	✓	✓	null
2	✓	✓		alternative
3	✓		✓	alternative
4	✓			alternative
5		✓	✓	alternative
6		✓		alternative
7			✓	alternative
8				alternative

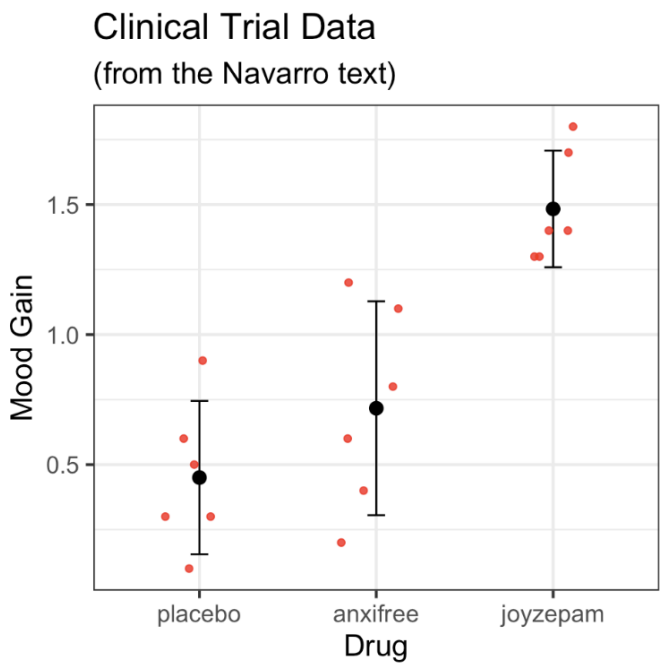
ANOVA: Follow-up tests: `pairwise.t.test()`

- `pairwise.t.test()` is a function in base R
- feed it your DV and your IV

```
1 library(tidyverse)
2 load(url("https://www.gribblelab.org/2812/data/clinicaltrial.Rdat
3 clin.trial <- tibble(clin.trial)
```

```
1 pairwise.t.test( x = clin.trial$mood.gain,
2                  g = clin.trial$drug,
3                  p.adjust.method = "none") # we will change this
```

```
    Pairwise comparisons using t tests with pooled SD

data:  clin.trial$mood.gain and clin.trial$drug

         placebo anxifree
anxifree 0.15021 -
joyzepam 3e-05   0.00056

P value adjustment method: none
```



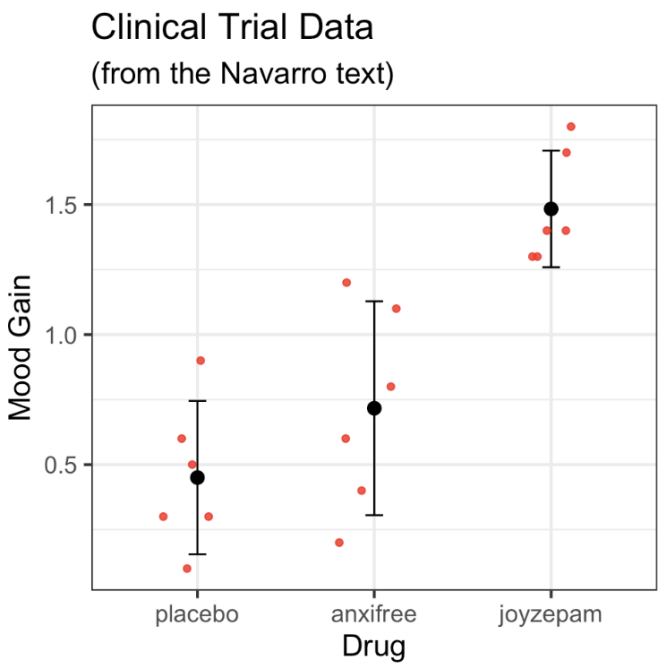
ANOVA: Follow-up tests: `posthocPairwiseT()`

- `posthocPairwiseT()` is a function from the `lsr` package
- feed it your anova model object

```
1  library(lsr) # you will have to install.packages("lsr") once
2  my.anova <- aov(mood.gain ~ drug, data = clin.trial)
```

```
1  posthocPairwiseT(my.anova, p.adjust.method = "none") # we will ch
```

```
    Pairwise comparisons using t tests with pooled SD

data:  mood.gain and drug

          placebo anxifree
anxifree  0.15021 -
joyzepam  3e-05   0.00056

P value adjustment method: none
```



The Multiple Comparisons Problem

- "post-hoc" tests are like going fishing for differences
- we are testing many hypotheses on the same dataset
- Type-I errors accumulate
- actual Type-I error rate is inflated above $\alpha = .05$
- called "Familywise Error Rate"
- but before we get into this—let's review Type-I and Type-II errors

review: Type-I Errors

- a Type-I error is when you reject the null hypothesis when it is actually true
- you conclude there is a difference when there is actually no difference

	H_0 rejected	Fail to reject H_0
H_0 false	Correct	Type II error
H_0 true	Type I error	correct

Alpha (α) = Prob (Type I error)

Beta (β) = Prob (Type II error)

Power = $1 - \beta$

review: Type-I Error rate

- how do you set the Type-I error rate?
  - your **threshold for rejecting the null hypothesis**
- is called α and is typically set to 0.05
  - you are willing to make a Type-I error 5% of the time when the null hypothesis is true

Type-I Error rate: simulations

- make null hypothesis true
- take 3 samples of size 10 from **one** normal distribution
- all three samples come from same population (same mean)
- conduct a one-way ANOVA
- if p-value < 0.05, then reject the null hypothesis
- repeat this 10,000 times and count how many times we reject the null hypothesis
  - (these are Type-I errors)

Type-I Error rate: simulations

```r
set.seed(2812)
N <- 10                    # sample size
nsims <- 10000             # number of simulated experiments
p.values <- rep(0, nsims)   # to store our p-values
F.values <- rep(0, nsims)   # to store our F values
decisions <- rep("", nsims) # to store our decisions
for (i in seq(nsims)) {
  y1 <- rnorm(n=N, mean=0, sd=1) # sample group 1
  y2 <- rnorm(n=N, mean=0, sd=1) # sample group 2
  y3 <- rnorm(n=N, mean=0, sd=1) # sample group 3
  y <- c(y1,y2,y3)
  g <- factor(c(rep("g1",N), rep("g2",N), rep("g3",N))) # construct our group factor
  my.df <- tibble(y=y, g=g)                # construct our data tibble
  my.anova <- aov(y ~ g, data=my.df)       # conduct ANOVA
  my.anova.summary <- summary(my.anova)    # get ANOVA summary
  p <- my.anova.summary[[1]]$`Pr(>F)`[1]  # extract p-value of omnibus F-test
  F <- my.anova.summary[[1]]$`F value`[1] # extract F-value of omnibus F-test
  if (p < .05) {                           # make our decision
    decisions[i] = "reject H0"             # reject the null hypothesis
  } else {
    decisions[i] = "accept H0"             # or accept the null hypothesis
  }
  p.values[i] <- p                         # store the p-value
  F.values[i] <- F                         # store the F-value
}
my.sims <- tibble(p.values, F.values, decisions)
```

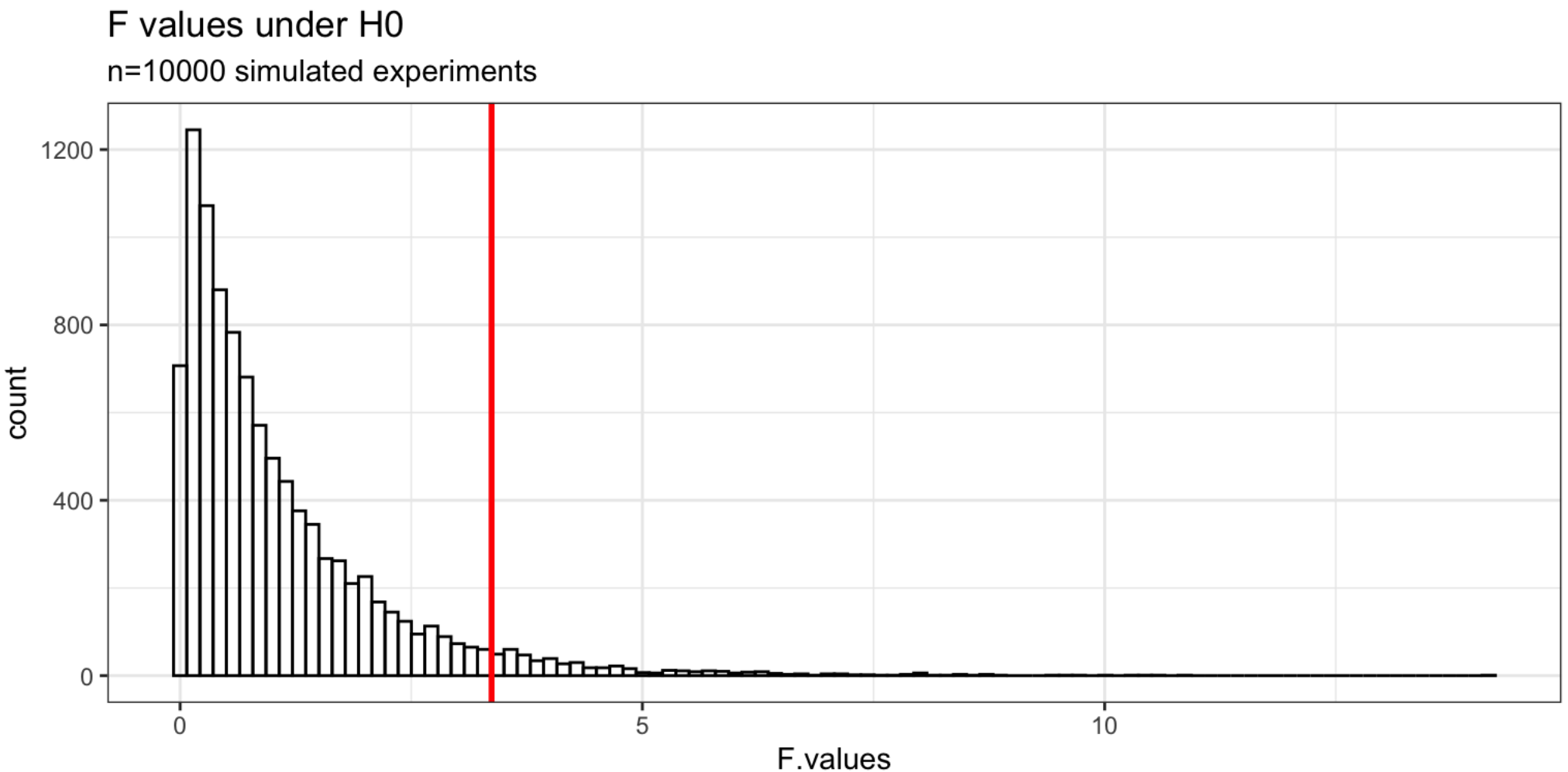
Type-I Error rate: simulations

```
1 my.sims
```

```
# A tibble: 10,000 × 3
   p.values F.values decisions
      <dbl>    <dbl> <chr>
 1    0.668   0.409  accept H0
 2    0.878   0.130  accept H0
 3    0.950   0.0514 accept H0
 4    0.813   0.209  accept H0
 5    0.430   0.870  accept H0
 6    0.777   0.255  accept H0
 7    0.631   0.469  accept H0
 8    0.719   0.334  accept H0
 9    0.101   2.50   accept H0
10    0.827   0.191  accept H0
# ℹ 9,990 more rows
```

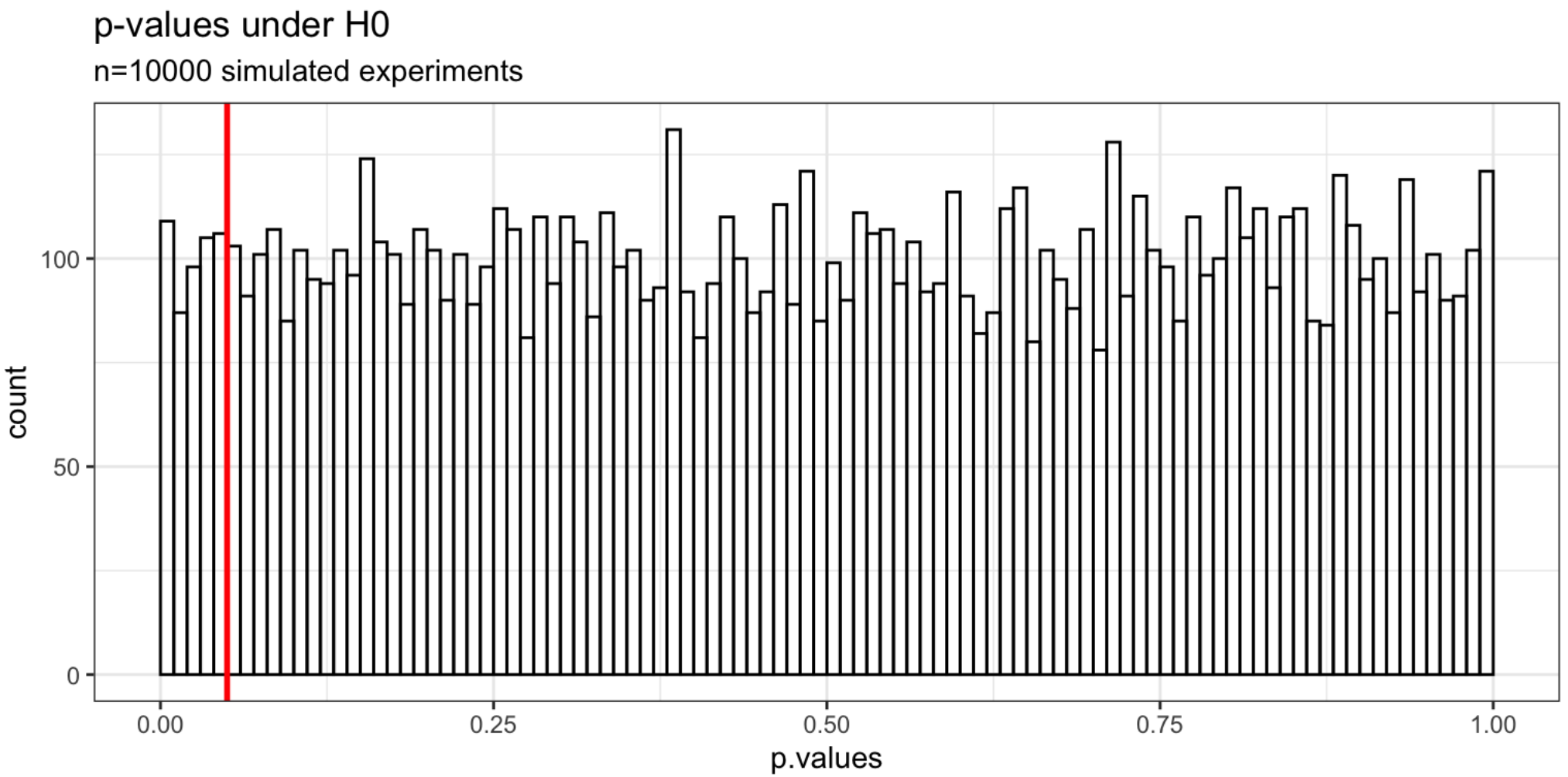
Type-I Error rate: simulations

▶ Code



Type-I Error rate: simulations

▶ Code



Type-I Error rate: simulations

```r
my.sims %>% group_by(decisions) %>%
    summarise(n = n(), percent = n/nsims)
```

```
# A tibble: 2 × 3
  decisions     n percent
  <chr>     <int>   <dbl>
1 accept H0  9495  0.950
2 reject H0   505  0.0505
```

- **decision** based on $\alpha = .05$, so we made 5% Type-I errors
- if decision based on $\alpha = .01$, we make 1% Type-I errors
- if decision based on $\alpha = .001$, we make 0.1% Type-I errors
- **you get to choose** your Type-I error rate

Type-I Error rate

- **important**: we never know whether H_0 is true or H_1 is true
- we compute p under the assumption that H_0 is true
- decision to reject or accept H_0 based on comparing p to α

Type-I Error rate

- H_0 is actually true: then α determines our Type-I error rate
- H_0 is actually false: rejecting H_0 is the **correct decision**; no Type-I error

	H_0 rejected	Fail to reject H_0
H_0 false	Correct	Type II error
H_0 true	Type I error	correct

Alpha (α) = Prob (Type I error)

Beta (β) = Prob (Type II error)

Power = $1 - \beta$

Type-II Error rate

- a Type-II error is when you fail to reject the null hypothesis even though it is actually false—the alternate hypothesis is true
- you conclude there is not a difference when there is actually is one

	H_0 rejected	Fail to reject H_0
H_0 false	Correct	Type II error
H_0 true	Type I error	correct

Alpha (α) = Prob (Type I error)

Beta (β) = Prob (Type II error)

Power = $1 - \beta$

Type-II Error rate

- β is the probability of making a Type-II error
- the *power* of your statistical test is defined as $1 - \beta$
- **statistical power** is the probability of correctly rejecting the null hypothesis when it is false
- the probability that you will correctly detect a difference when there is one

Determinants of statistical power

- effect size
  - how big is the difference between the groups relative to within-group variance?
- sample size
  - bigger sample size: more power to detect a difference when there is one
- α
  - smaller α: less power to detect a difference when there is one

Computing power in R

- `power.anova.test()` will compute power for an ANOVA
- `power.t.test()` will compute power for a t-test
- **important**: power calculations are always based on the assumption that H_1 is true
  - (groups came from different populations with different means)

Power example: ANOVA (H_1 true)

- we have 3 groups; we want to detect a difference between groups with a power of 0.8
- **assume H_1 is true**: we expect within-groups variance to be 4 times as big as the between-groups variance

```
power.anova.test(groups = 3, between.var = 1, within.var = 4, power = .80)
```

```
     Balanced one-way analysis of variance power calculation

         groups = 3
              n = 20.30205
    between.var = 1
     within.var = 4
      sig.level = 0.05
          power = 0.8

NOTE: n is number in each group
```

- `power.anova.test()` tells us we would need n=20 subjects per group to detect a difference between groups with a power of 0.8

Power example: ANOVA (H_1 true)

```
power.anova.test(groups = 3, between.var = 1, within.var = 4, power = .80)
```

```
     Balanced one-way analysis of variance power calculation

         groups = 3
              n = 20.30205
    between.var = 1
     within.var = 4
      sig.level = 0.05
          power = 0.8

NOTE: n is number in each group
```

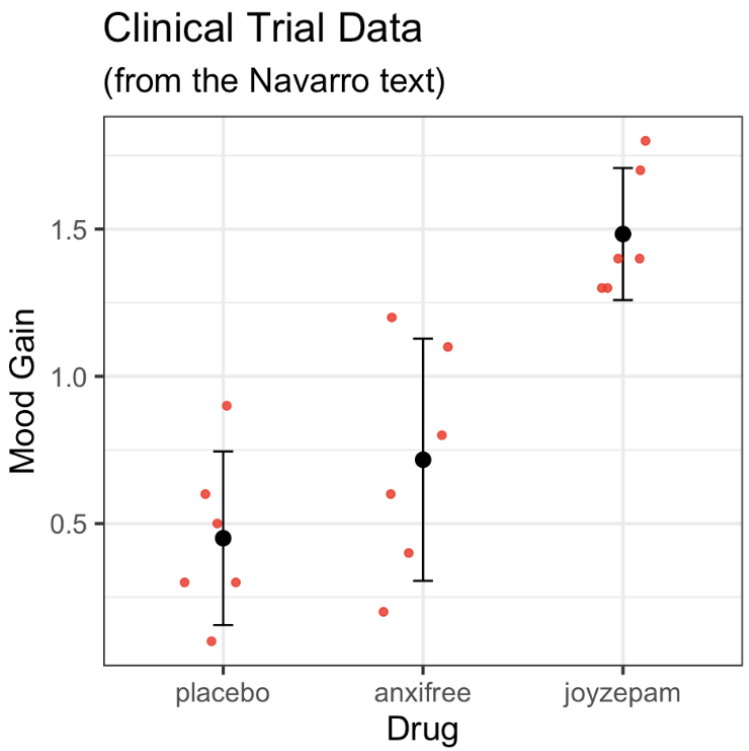
- **if there actually is a difference in the population** and we run this experiment 100 times, 80 times out of 100 we will correctly reject the null hypothesis
- 20 times out of 100 we will make a Type-II error and conclude there is not a difference
- play with parameters to see how sample size n depends on power and effect size

Power example: ANOVA (H_0 true)

- if there actually is **NOT** a difference in the population:
- and if we adopt $\alpha = 0.05$:
- if we run this experiment 100 times, 95 times out of 100 we will correctly fail to reject the null hypothesis
- 5 times out of 100 we will make a Type-I error and conclude there is a difference even though there isn't
- this is because our decision to reject or fail to reject H_0 is based on comparing p to $\alpha = 0.05$

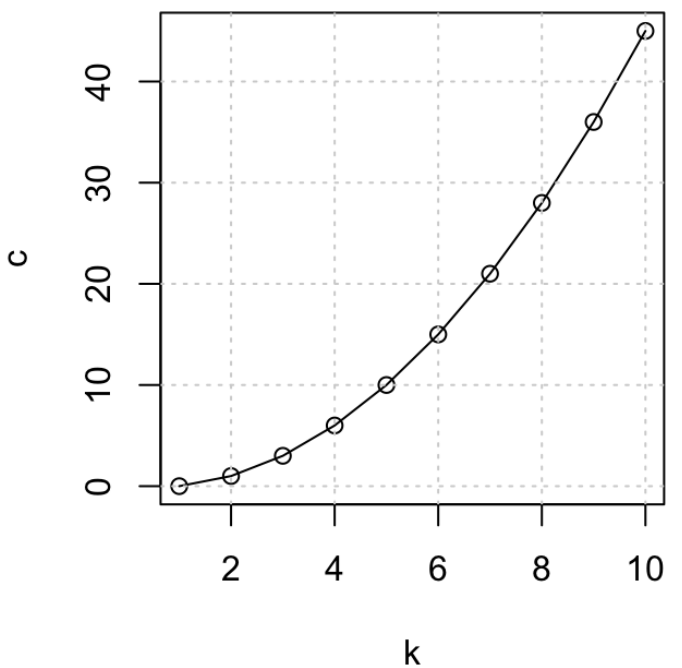
Multiple Comparisons

- how many possible pairwise tests?
  - placebo vs anxifree
  - placebo vs joyzepam
  - anxifree vs joyzepam
- For K groups the number of possible pairwise comparisons c is:
  - $c = K(K-1)/2$



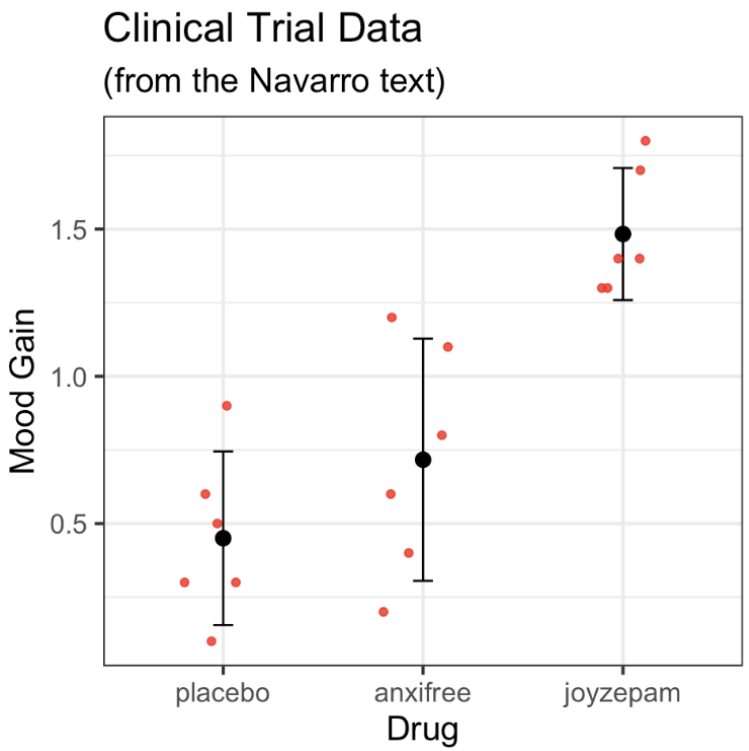
Multiple Comparisons

- For K groups the number of possible pairwise comaprisons c is:
  - $c = K(K-1)/2$
- for 3 groups: 3*(3-1)/2 = 3
- for 4 groups: 4*(4-1)/2 = 6
- for 6 groups: 6*(6-1)/2 = 15
- for 10 groups: 10*(10-1)/2 = 45 !



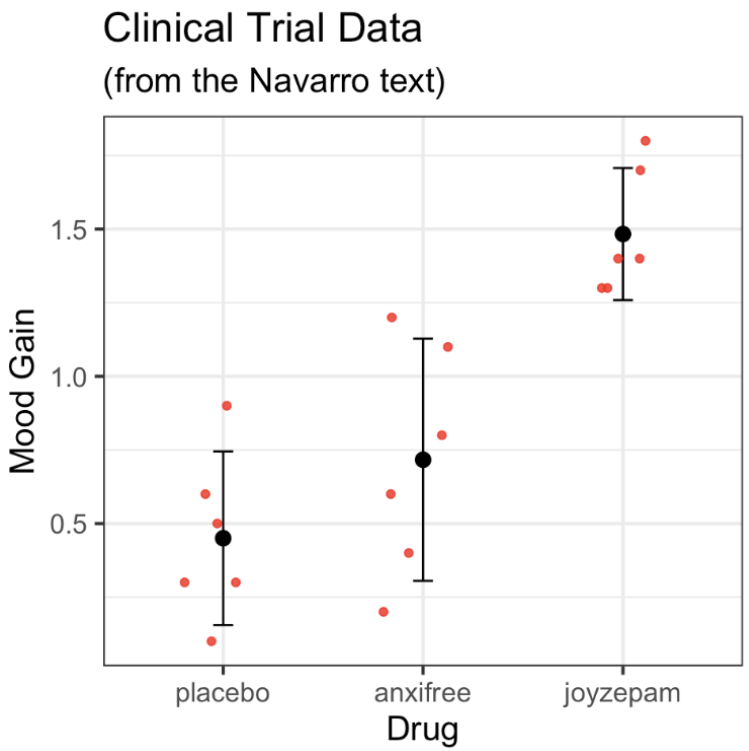
Multiple Comparisons

- 3 post-hoc tests:
  - placebo vs anxifree
  - placebo vs joyzepam
  - anxifree vs joyzepam
- each test has a p-value
- we make a decision about each test based on $\alpha = .05$
- Q: is our overal (**family-wise**) Type-I error rate still 5 %?



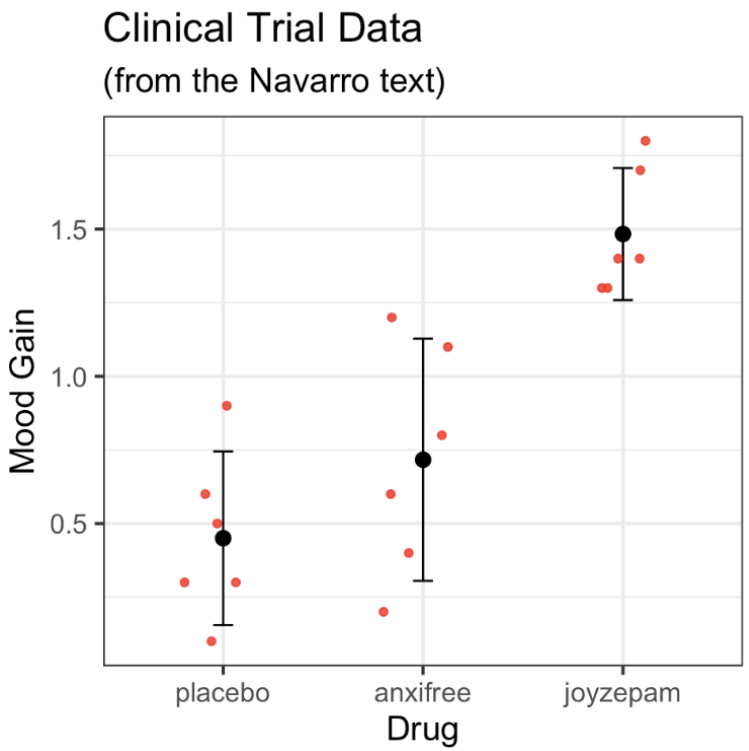
Multiple Comparisons

- is family-wise Type-I error still 5 %?
- Answer: No!
  - Type-I error rate is **inflated**
- Type-I error rate on the whole **family of tests** is way above 5 %
- Pr(Type-I error) = $1 - (1 - \alpha)^C$
- C is the number of post-hoc tests



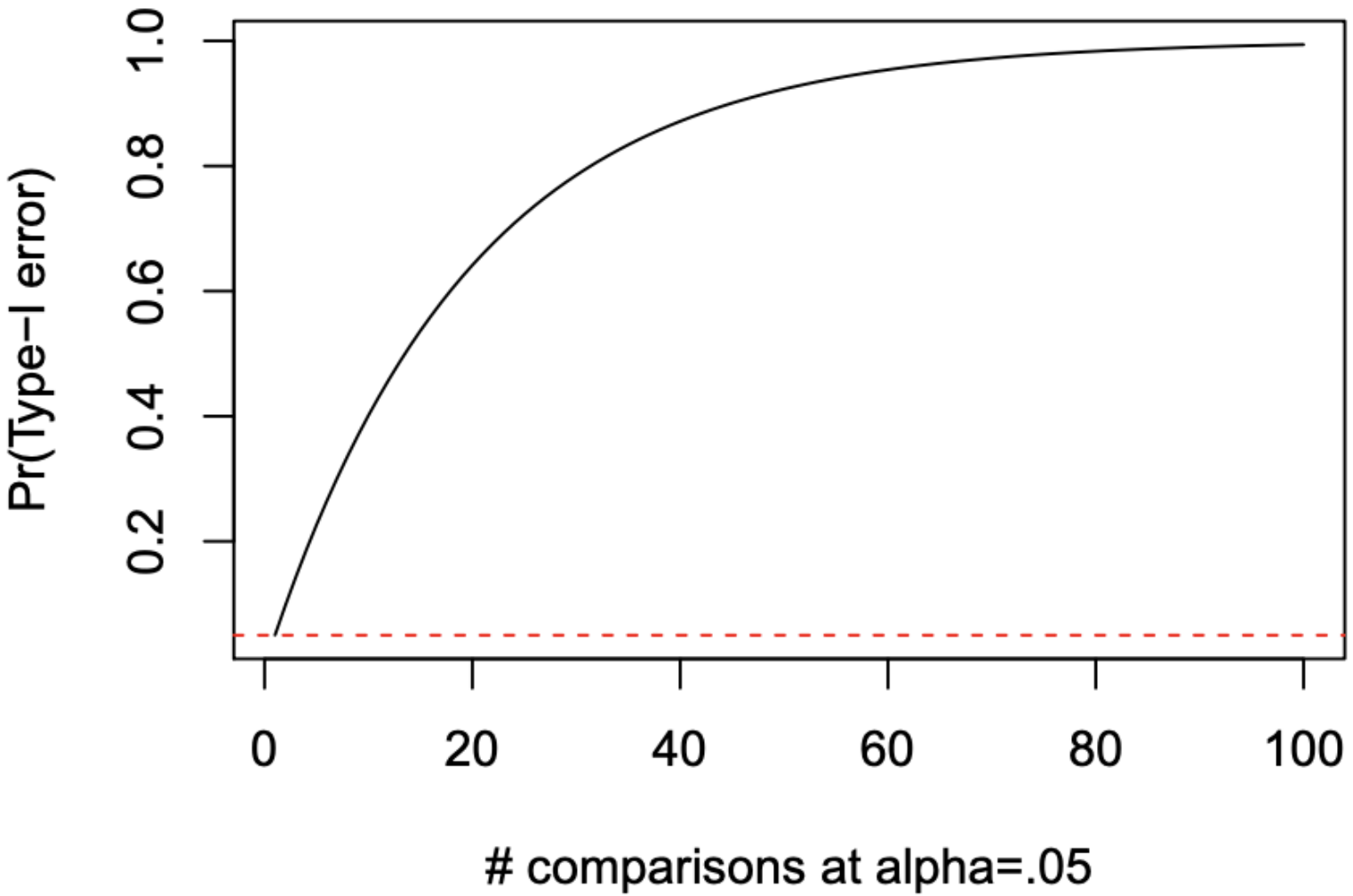
Multiple Comparisons

- Pr(Type-I error) = $1 - (1 - \alpha)^C$
- when $C = 3$
  - we get **14.3% family-wise Type-I error rate**



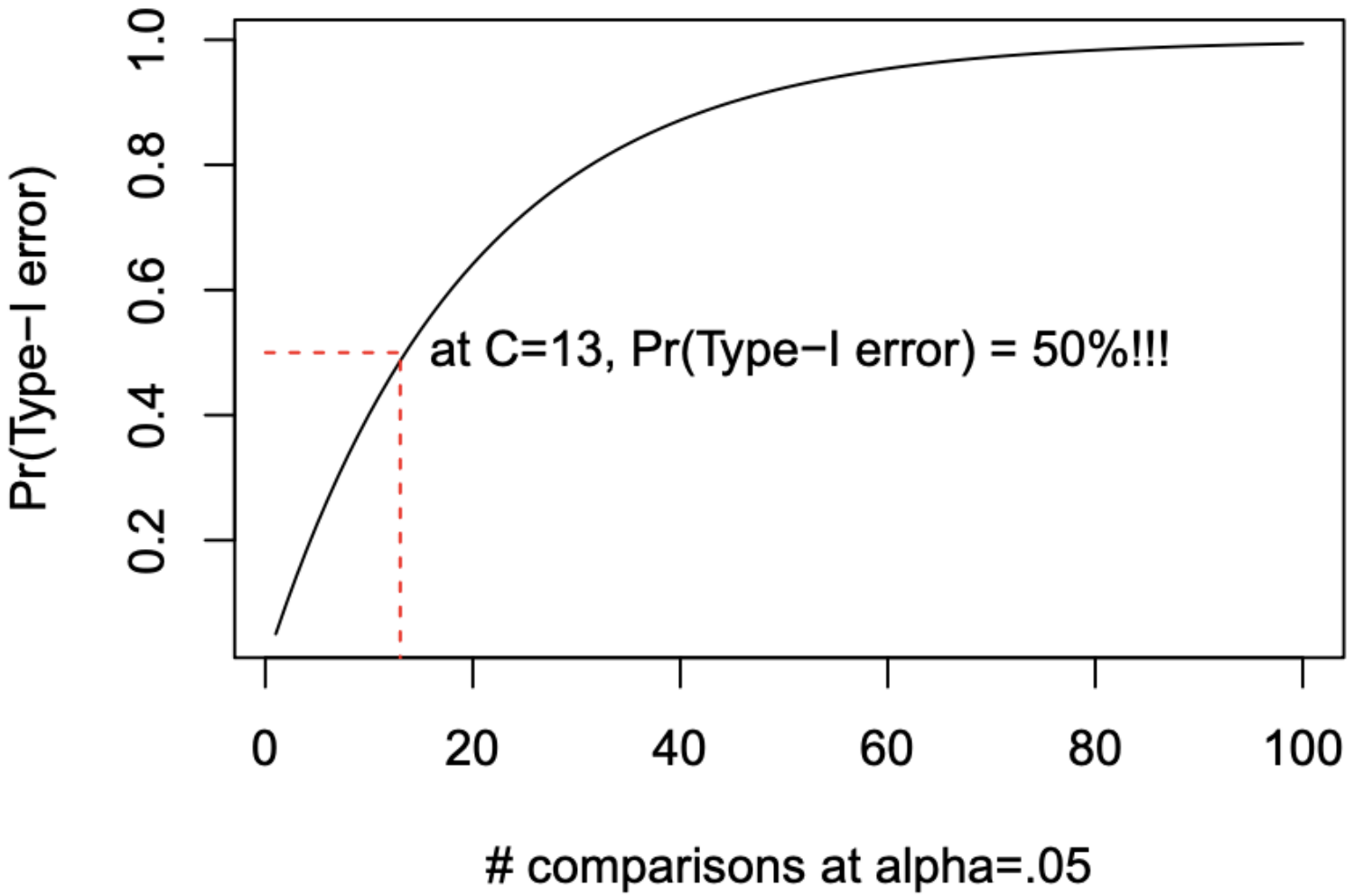
Multiple Comparisons

- Pr(Type-I error) = $1 - (1 - \alpha)^C$



Multiple Comparisons

- Pr(Type-I error) = $1 - (1 - \alpha)^C$



Why does Type-I error rate inflate?

- when we do post-hoc tests, we are **increasing the number of tests** performed on the same dataset
- the more tests we do, the more likely to make a Type-I error
- especially if we look at the data first and test the differences that **look largest**
- we end up **chasing the big differences**
- big differences (even if they are due to chance) are more likely to be identified as statistically significant

Correcting for Type-I error inflation

- there are several ways to correct for Type-I error inflation
- we will look at three of them:
  1. Bonferroni
  2. Tukey
  3. Bonferroni-Holm

Bonferroni Correction

- corrects for Type-I error inflation
- $p_{\text{corrected}}$ becomes $p * C$ for each test
- if $C = 3$ then $p_{\text{corrected}} = 0.05 * 3 = 0.15$
- for each test, reject H_0 only if $p_{\text{corrected}} < 0.05$
- simple, but overly conservative when doing many tests

Bonferroni Correction in R

```
1 posthocPairwiseT( my.anova, p.adjust.method = "bonferroni")
```

```
    Pairwise comparisons using t tests with pooled SD

data:  mood.gain and drug

          placebo anxifree
anxifree  0.4506  -
joyzepam  9.1e-05 0.0017

P value adjustment method: bonferroni
```

Tukey Correction

- allows for **all pairwise** comparisons among groups even when there are many
- still maintains overall $\alpha = 0.05$
- less conservative than Bonferroni when doing a large number of tests
- corrects the p-value with a formula based on number of groups
- reject H_0 when $p_{\text{corrected}} < 0.05$

Tukey Correction in R

```
1 TukeyHSD(my.anova)
```

```
  Tukey multiple comparisons of means
    95% family-wise confidence level

Fit: aov(formula = mood.gain ~ drug, data = clin.trial)

$drug
                        diff        lwr       upr     p adj
anxifree-placebo   0.2666667 -0.1901184 0.7234518 0.3115006
joyzepam-placebo   1.0333333  0.5765482 1.4901184 0.0000854
joyzepam-anxifree 0.7666667  0.3098816 1.2234518 0.0015284
```

Bonferroni-Holm Correction

- a good general purpose correction
  - less conservative than Bonferroni & Tukey
  - still allows for any number of pairwise comparisons
- sorts tests by uncorrected p-value
- then corrects p-values more for most significant tests and less for less significant tests
- reject H_0 when $p_{\text{corrected}} < 0.05$

Bonferroni-Holm Correction

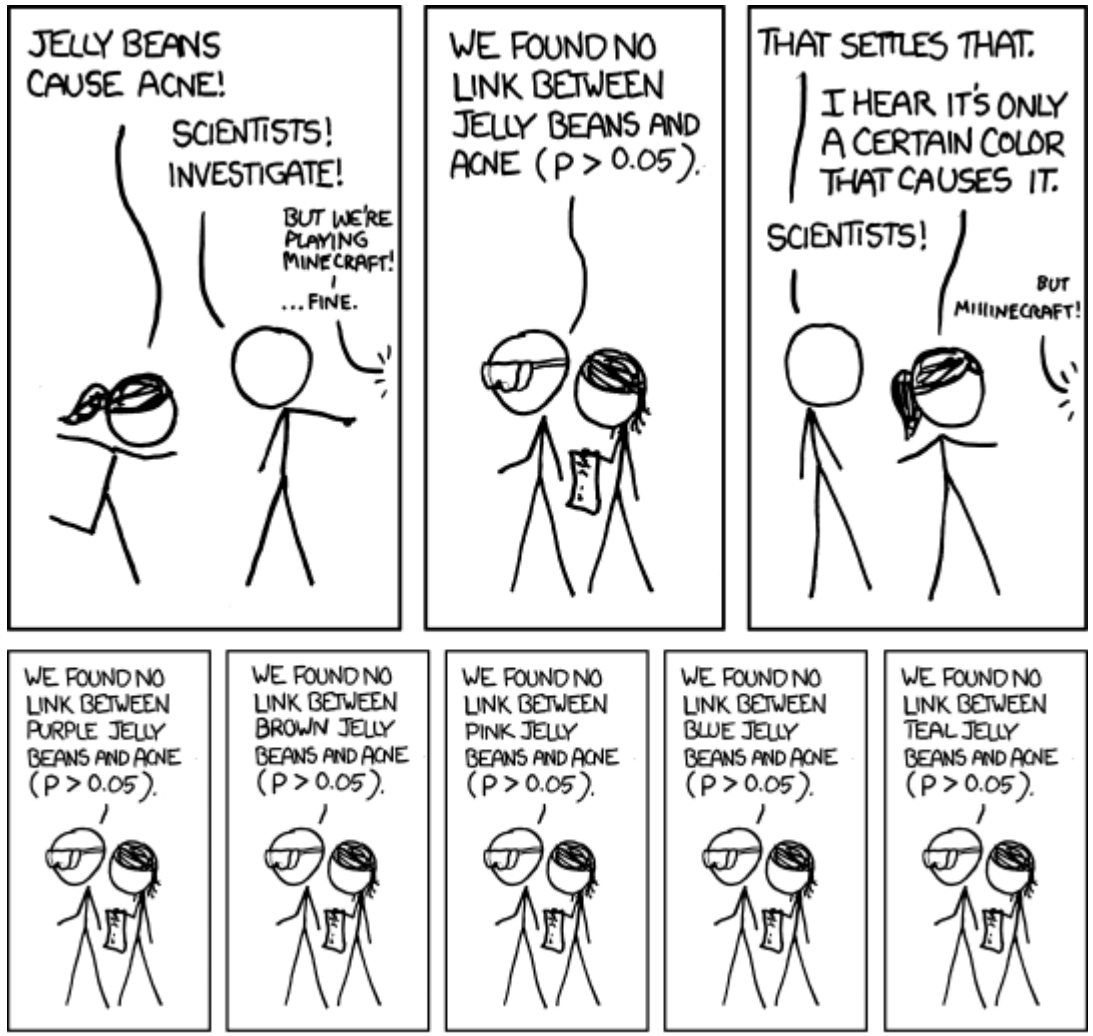
Bonferroni-Holm Correction in R

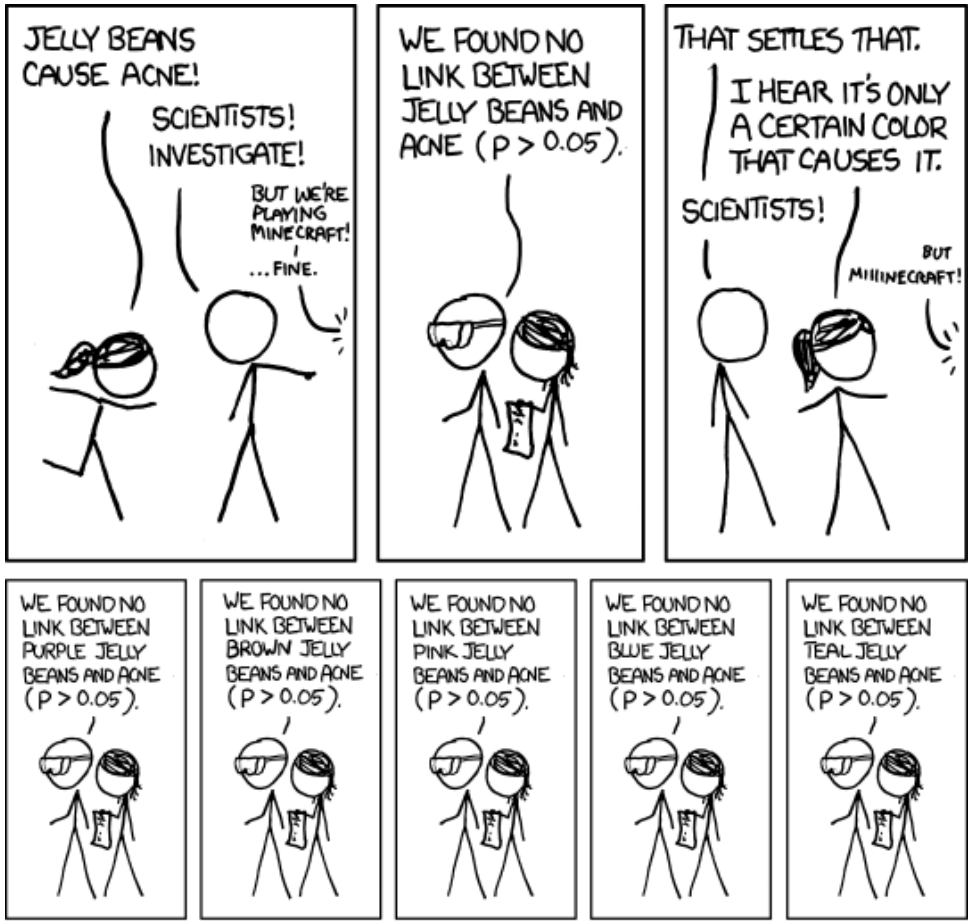
```
posthocPairwiseT( my.anova, p.adjust.method = "holm")
```

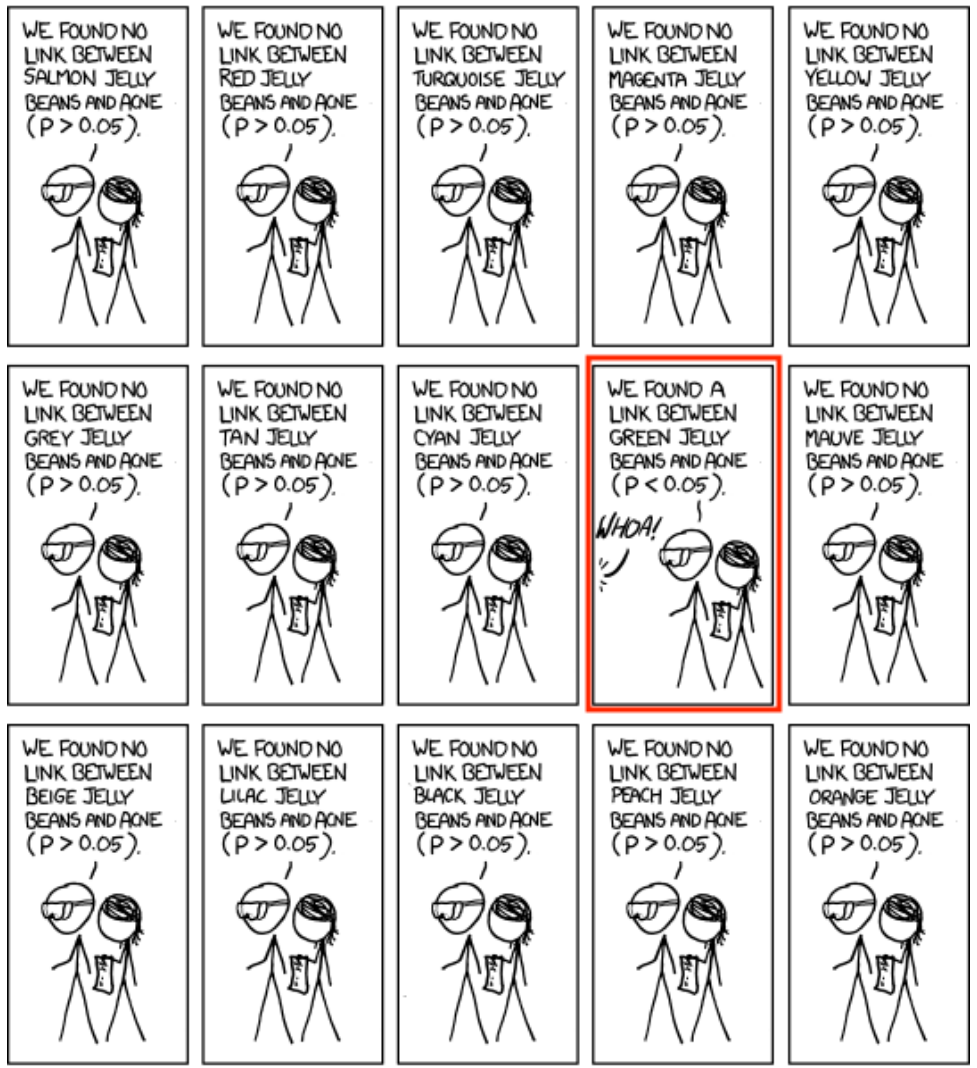
```
    Pairwise comparisons using t tests with pooled SD

data:  mood.gain and drug

         placebo anxifree
anxifree 0.1502  -
joyzepam 9.1e-05 0.0011

P value adjustment method: holm
```

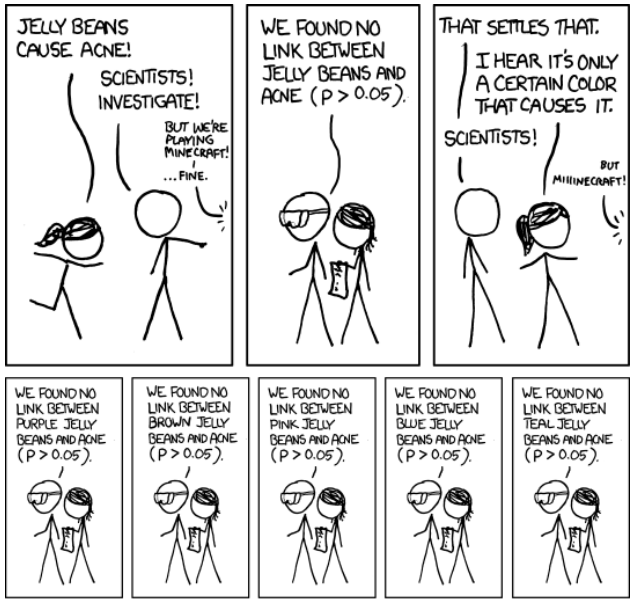
Unbalanced Designs

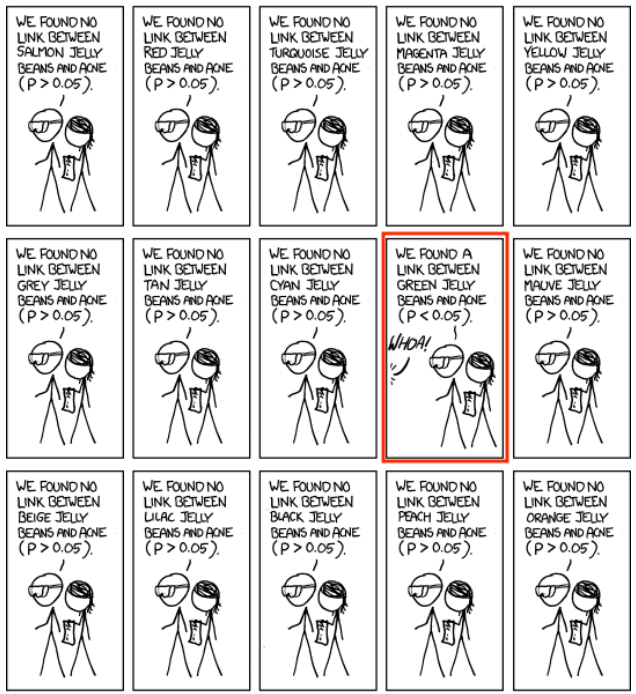
- when we have unequal sample sizes in the ANOVA groups
- can be a problem for the ANOVA test
  - think about why? (e.g. homogeneity of variances)
- calculations to correct for unequal sample sizes are complicated and tedious
- sometimes reasonable people even disagree on the best approach
- try hard not to have unequal sample sizes in your ANOVA groups!!
- we will return to this later in the course when we talk about Factorial ANOVA
- as we saw in a previous lecture, for oneway ANOVA you can use Welch's `oneway.test()` if you are concerned about homogeneity of variance due to unequal sample sizes
  (see Navarro, Ch. 14.8 for details)

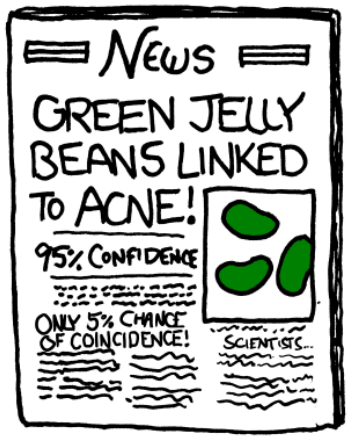

JELLY BEANS CAUSE ACNE!
SCIENTISTS! INVESTIGATE!
BUT WE'RE PLAYING MINECRAFT!
...FINE.
WE FOUND NO LINK BETWEEN JELLY BEANS AND ACNE (P > 0.05).
THAT SETTLES THAT.
I HEAR IT'S ONLY A CERTAIN COLOR THAT CAUSES IT.
SCIENTISTS!
BUT MIIINECRAFT!
WE FOUND NO LINK BETWEEN PURPLE JELLY BEANS AND ACNE (P > 0.05).
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).


JELLY BEANS CAUSE ACNE!
SCIENTISTS! INVESTIGATE!
BUT WE'RE PLAYING MINECRAFT!
... FINE.
WE FOUND NO LINK BETWEEN JELLY BEANS AND ACNE (P > 0.05).
THAT SETTLES THAT.
I HEAR IT'S ONLY A CERTAIN COLOR THAT CAUSES IT.
SCIENTISTS!
BUT MIIIINECRAFT!
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).
WE FOUND NO LINK BETWEEN BEIGE JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN LILAC JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN BLACK JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN PEACH JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN ORANGE JELLY BEANS AND ACNE (P > 0.05).







https://xkcd.com/882/