Research Methods II

MICHAEL BERNSTEIN
CS 376

Goal

Understand and use statistical techniques common to HCI research

Last time

- How to plan an evaluation
- What is a statistical test?
- Chi-square
- t-test
- Paired t-test
- Mann-Whitney U

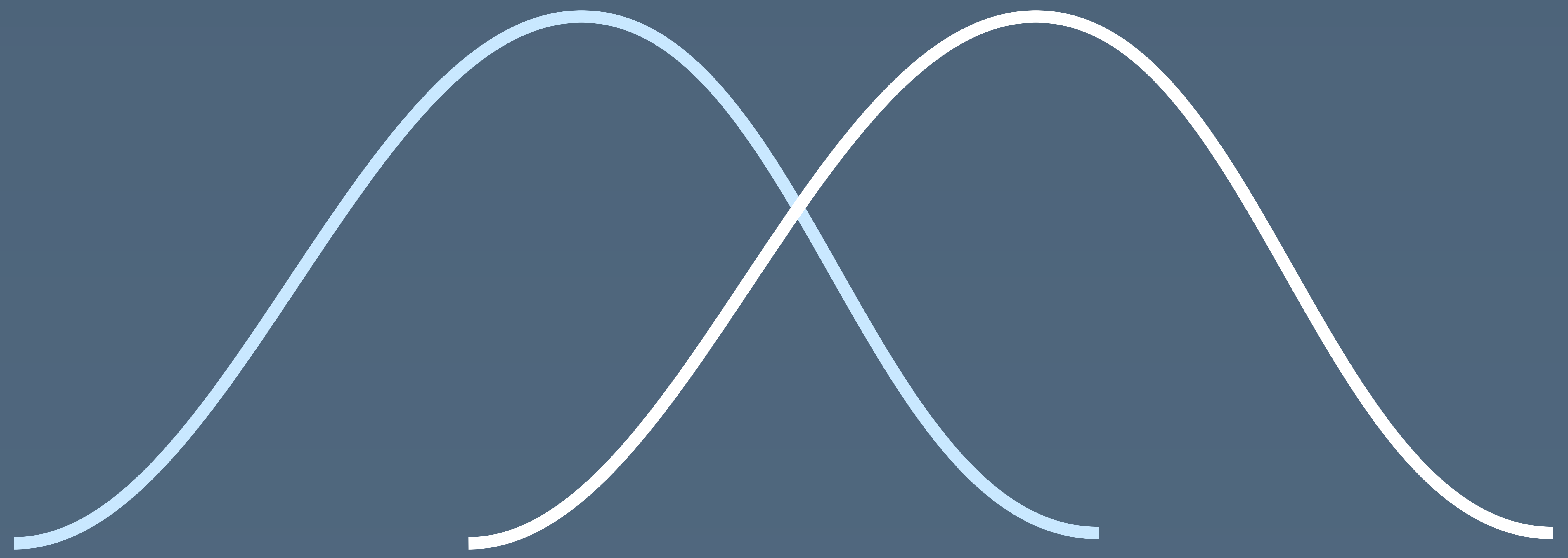
Today

- ANOVA
- Posthoc tests
- Two-way ANOVA
- Repeated measures ANOVA

ANOVA

t-test: compare two means

- "Do people fix more bugs with our IDE bug suggestion callouts?"



ANOVA: compare N means

- "Do people fix more bugs with our IDE bug suggestion callouts, with warnings, or with nothing?"

Cell means model

- Assume there are r factor levels
  e.g., laptop + tablet + phone: r=3
- Value of the jth observation for the ith factor level:

$$Y_{ij}$$

- e.g., $Y_{2,5}$ is the i=2nd condition and the j=5th user

Cell means model

- ANOVA characterizes each observation as a deviation from the mean of the factor level

Cell means model

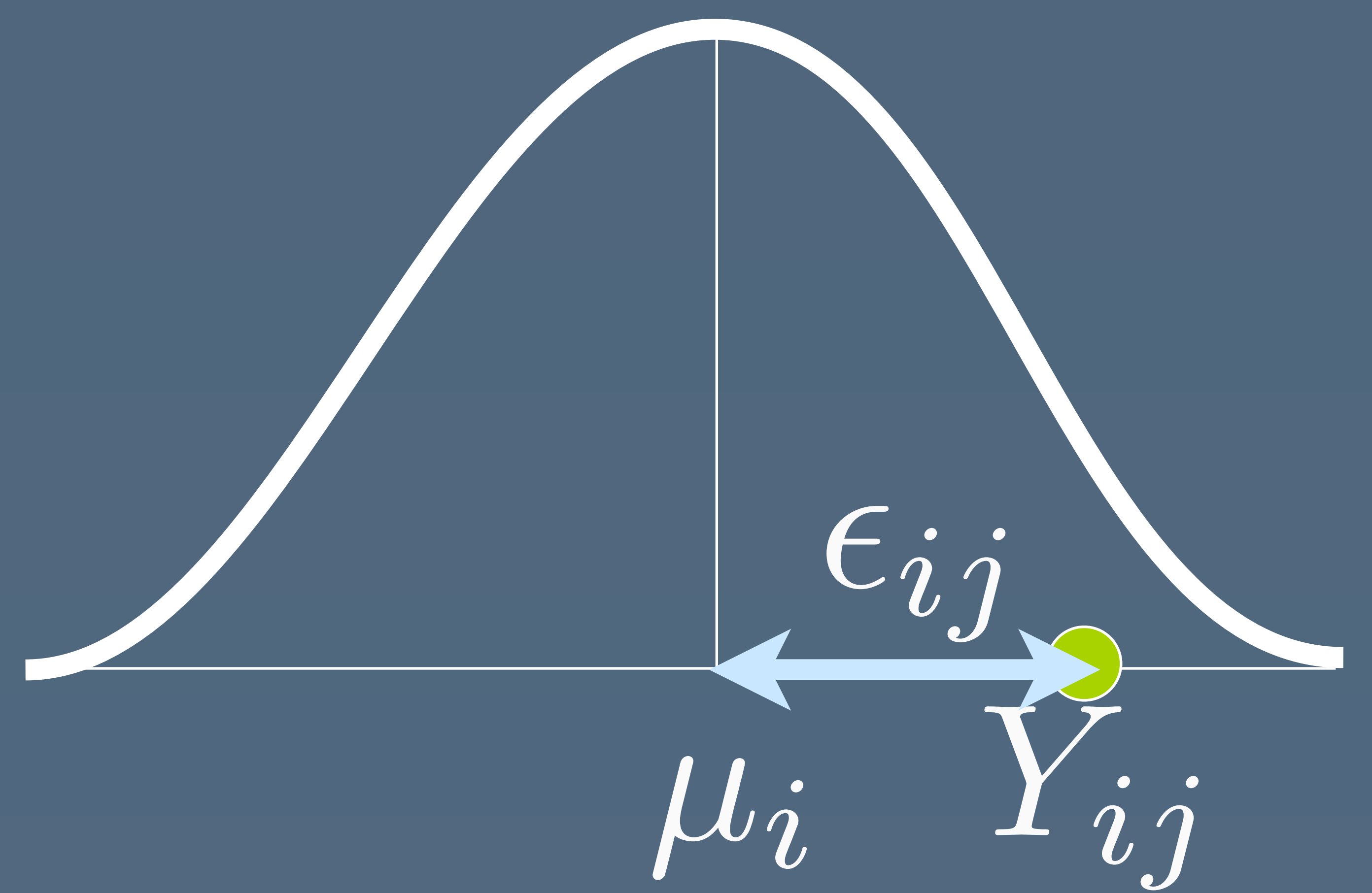
- Starter ANOVA model:

$$Y_{ij} = \mu_i + \epsilon_{ij}$$

mean for factor level i

error: difference between observed value and the mean

- Y_{ij} are independent $N(\mu_i, \sigma^2)$



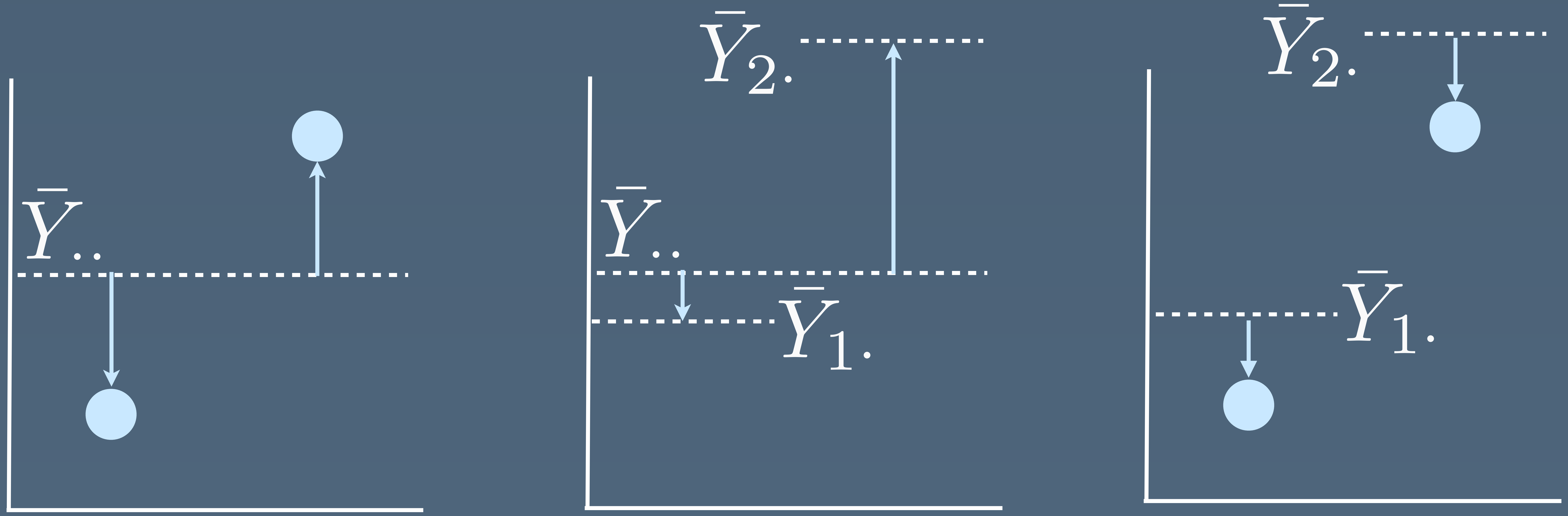
Partitioning the variance

- The total variability in Y is the difference between each observation Y_{ij} and the grand mean $\bar{Y}_{..}$

bar is the mean; dot is an aggregate over all observations, here both i and j

- Easier to understand if we separate it out via the factor level means

$$\underbrace{Y_{ij} - \bar{Y}_{..}}_{\text{total deviation from grand mean}} = \underbrace{\bar{Y}_{i.} - \bar{Y}_{..}}_{\text{deviation of factor mean from grand mean}} + \underbrace{Y_{ij} - \bar{Y}_{i.}}_{\text{deviation of response from factor mean}}$$



$$\underbrace{Y_{ij} - \bar{Y}_{..}}_{\text{total deviation from grand mean}} = \underbrace{\bar{Y}_{i.} - \bar{Y}_{..}}_{\text{deviation of factor mean from grand mean}} + \underbrace{Y_{ij} - \bar{Y}_{i.}}_{\text{deviation of response from factor mean}}$$

Partitioning the variance

- Total sum of squares SSTO:

$$SSTO = \sum_i \sum_j (Y_{ij} - \bar{Y}_{..})^2$$

Treatment sum of squares SSTR:

$$SSTR = \sum_i n_i(\bar{Y}_{i.} - \bar{Y}_{..})^2$$

- Error sum of squares SSE:

$$SSE = \sum_i \sum_j (Y_{ij} - \bar{Y}_{i.})^2$$

ANalysis Of VAriance (ANOVA)

- Provably true:

$$SSTO = SSTR + SSE$$

total variance | differences between factor level means | random variation around factor level means

- Degrees of freedom: how many values can vary? (Using n and r)

SSTO: n - 1

SSTR: r - 1

SSE: n - r

Studentizing the variance

- Divide each estimator by its degrees of freedom to produce a χ^2 random variable:
  - Treatment mean square is χ^2(r-1)

    $$MSTR = \frac{SSTR}{r-1}$$

  - Error mean square is χ^2(n-r)

    $$MSE = \frac{SSE}{n-r}$$

Turning variance into a statistic

- Null hypothesis: $\mu_1 = \mu_2 = \ldots = \mu_r$
- Alternate hypothesis: not all μ_i are equal
- Statistics magic: dividing two random variables distributed as χ^2 produces a random variable distributed as F
- $F^* = \frac{MSTR}{MSE}$ is $F(r-1, n-r)$

  Large MSTR relative to MSE suggests that the factor means explain most variance

Finally: run the test!

- How large is the value we constructed from the F distribution?
- Test if

$$F^* > F(1-\alpha; r-1, n-r)$$

```
> aov <- aov(value ~ group, data)
> summary(aov)
            Df Sum Sq Mean Sq F value  Pr(>F)    
group        2  22.75   11.38    12.1 0.00032 ***
Residuals   21  19.75    0.94                    
```

SSTR

SSE

3 factor levels

24 observations

SS

MS

F(2,21)

p < .001

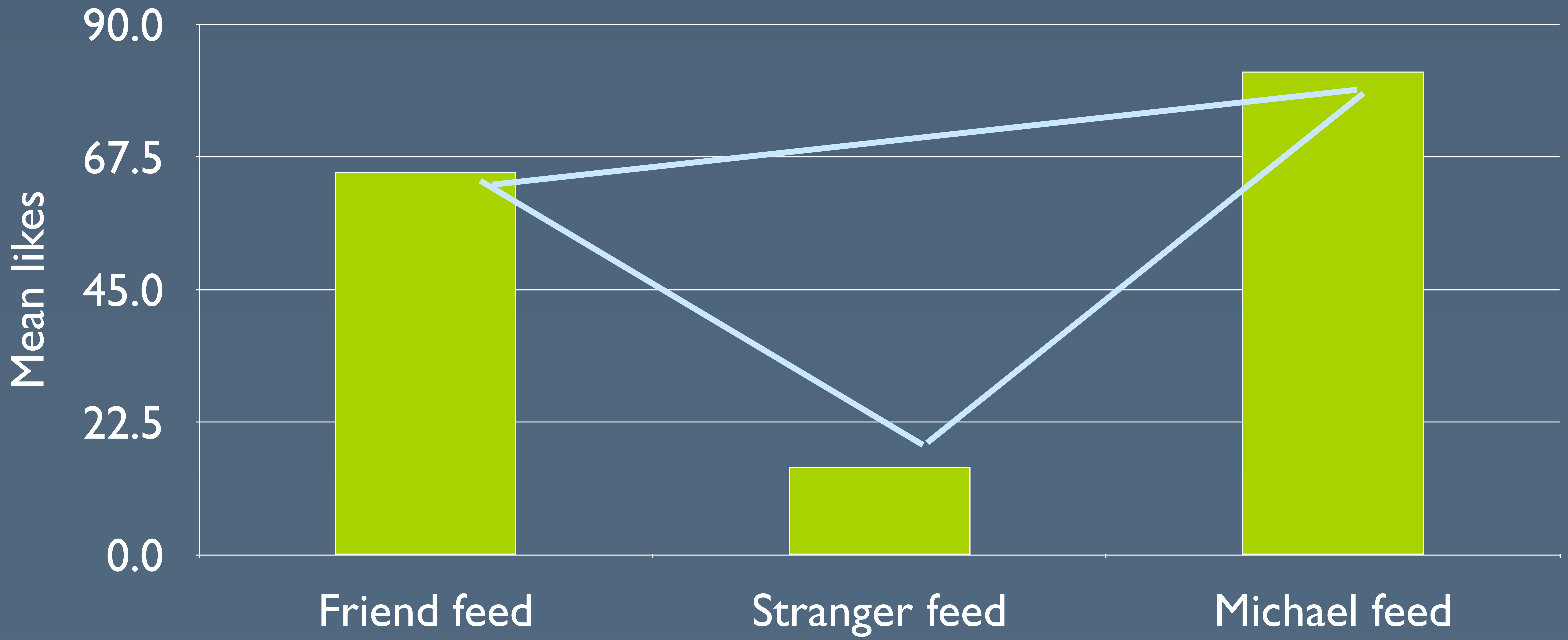
Posthoc tests

We’re done...or are we?

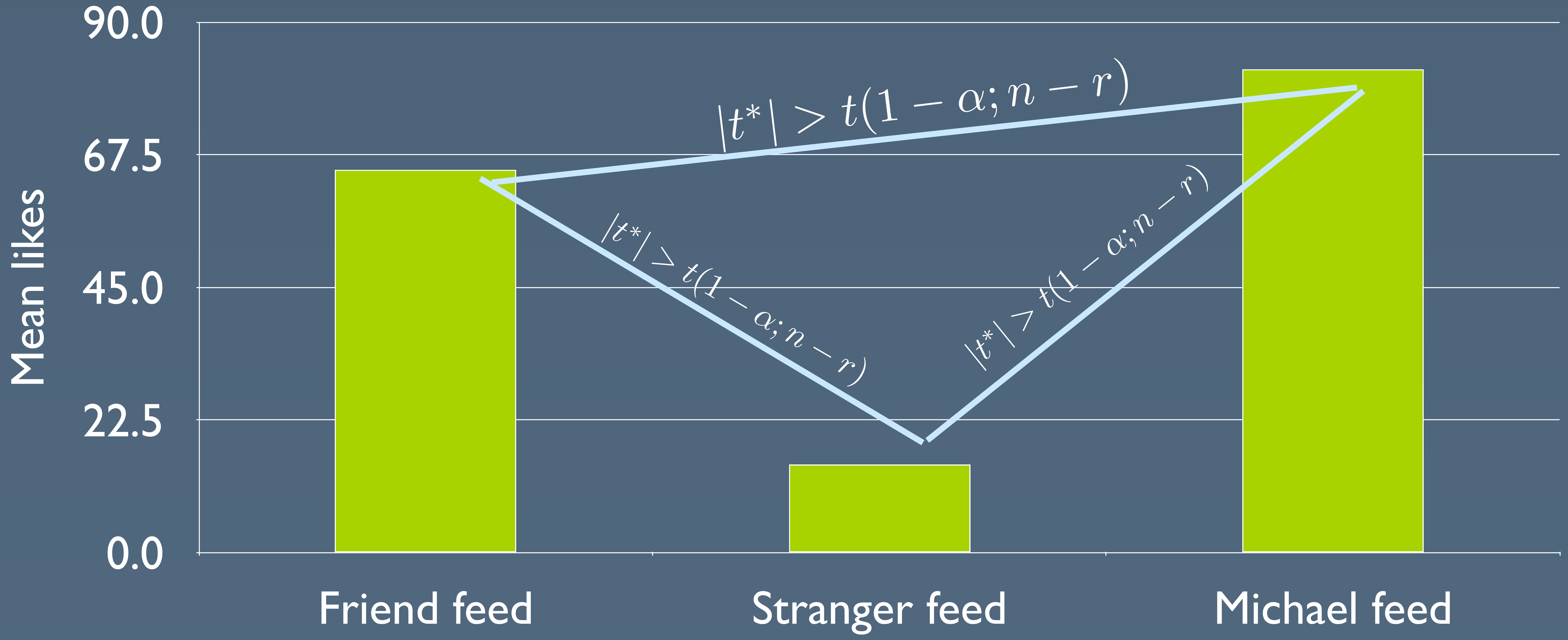
- Significant means “One of the μ_i are different.”
- That’s not very helpful: “There is some difference between populating the Facebook news feed with friends vs. strangers vs. only Michael’s status updates”

Estimating pairwise differences

- Which pairs of factor levels are different from each other?



Roughly: we do pairwise t-tests



But…familywise error!

- $\alpha = .05$ implies a .95 probability of being correct
- If we do m tests, the actual probability of being correct is now:

$$\alpha^m = .95 \cdot .95 \cdot .95 \cdot \ldots$$
$$< .95$$

Bonferroni correction

- Avoid familywise error by adjusting α to be more conservative
- Divide α by the number of comparisons you make
  - 4 tests at $\alpha = .05$ implies using $\alpha = .0125$
- Conservative but accurate method of compensating for multiple tests

Bonferroni correction

```
> pairwise.t.test(value, group, p.adj='bonferroni')

   Pairwise comparisons using t tests with pooled SD

data:  value and group

  A       B
B 0.02971 -
C 0.00023 0.15530

P value adjustment method: bonferroni
```

Reporting an ANOVA

- "A one-way ANOVA revealed a significant difference in the effect of news feed source on number of likes ($F(2, 21)=12.1, p<.001$)."

```
> aov <- aov(value ~ group, data)
> summary(aov)
            Df Sum Sq Mean Sq F value  Pr(>F)
group        2  22.75   11.38    12.1 0.00032 ***
Residuals   21  19.75    0.94
```

- "Posthoc tests using Bonferroni correction revealed that friend feed and Michael feed were significantly better than a stranger feed ($p<.05$), but the two were not significantly different from each other ($p=.32$)."

Two-way ANOVA

Crossed study designs

- Suppose you wanted to measure the impact of two factors on total likes on Facebook:
  - Strong ties vs. weak ties in your news feed
  - Presence of a reminder of the last time you liked each friend's content (e.g., "You last liked a story from John Hennessy in January")
- This is a 2 x 2 study: two factor levels for each factor {tie strength, reminder}

Basic two-factor ANOVA model

$$\mu_{ij} = \mu_{..} + \alpha_i + \beta_j$$

mean for ith level of 1st factor & jth level of 2nd factor

grand mean

difference between ith level of 1st factor and grand mean

difference between jth level of 2nd factor and grand mean

$$\mu_{ij} = \mu_{..} + \alpha_i + \beta_j$$

- Example: $\mu_{1,2}$
  - Mean user has 8 likes: $\mu_{..} = 8$
  - Mean user with strong ties (i=1) has 11 likes:
    $\alpha_1 = \mu_{i.} - \mu_{..} = 11 - 8 = 3$
  - Mean user with reminder has 7 likes:
    $\beta_2 = \mu_{.j} - \mu_{..} = 7 - 8 = -1$

Interaction effects

- Sometimes the basic model doesn't capture subtle interactions between factors
  - Data: People who see strong ties and have a reminder are especially active
  - Result: Grand mean 8, strong tie mean 11, reminder mean 7, but mean in this cell is 20

Two-factor ANOVA test

- Test for main effects and interaction

```
> anova(lm(time ~ device * technique))
Analysis of Variance Table

Response: time
                 Df Sum Sq Mean Sq  F value    Pr(>F)
device            1  981.0  981.02  94.5291 2.581e-12 ***
technique         2 3423.8 1711.90 164.9547 < 2.2e-16 ***
device:technique  2   75.3   37.65   3.6275   0.03522 *
Residuals        42  435.9   10.38
```

factor or interaction   SS   MS   F   p

- Main effects are significant, but interaction effect is also significant

Repeated measures ANOVA

Within-subjects studies

- Control for individual variation using the mean response for each participant
- Before: we found the mean effect of each treatment
- Now: we find the mean effect of each participant

Repeated measures in R

repeated measures
error term

effect of subtracting
out the participant
means

remaining
main effects

```
> aov <- aov(value ~ factor(group) +
+ Error(factor(participant)/factor(group)), repeatframe)
> summary(aov)

Error: factor(participant)
          Df Sum Sq Mean Sq F value Pr(>F)
Residuals  7  5.167  0.7381

Error: factor(participant):factor(group)
              Df Sum Sq Mean Sq F value  Pr(>F)
factor(group)  2  22.75  11.375   10.92 0.00139 **
Residuals     14  14.58   1.042
```

All together now

Always follow every step!

1. Visualize the data
2. Compute descriptive statistics (e.g., mean)
3. Remove outliers >2 standard deviations from the mean
4. Check for heteroskedasticity and non-normal data
   - Try log, square root, or reciprocal transform
   - ANOVA is robust against non-normal data, but not against heteroskedasticity
5. Run statistical test
6. Run any posthoc tests if necessary