Research Methods II

MICHAEL BERNSTEIN

SPRING 2013

cs376.stanford.edu

Goal:

Understand and use statistical techniques common to HCI research

Last time

- Model of statistical tests
- t-test
- Linear regression
  - General form $Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \ldots + \beta_m X_{im} + \epsilon_i$
  - Significance test

    $$\left| \frac{b_1}{s\{b_1\}} \right| > t(1 - \alpha/2; n - 2)$$

  - $R^2 = 1 - \frac{SSE}{SSTO}$

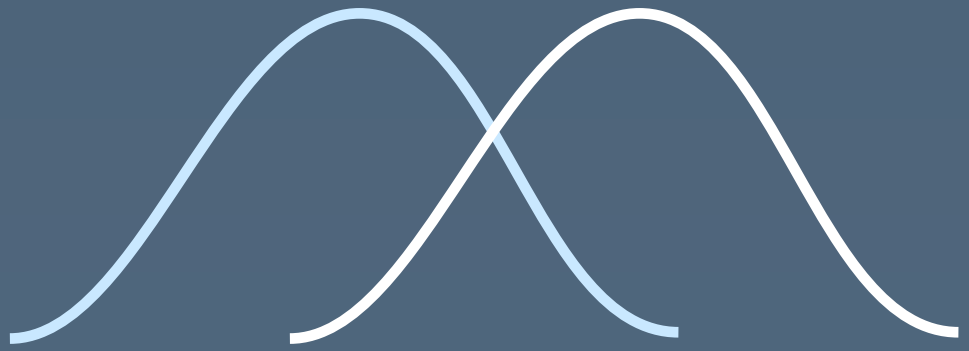
Today

- ANOVA
- Posthoc tests
- Two-way ANOVA
- Repeated measures ANOVA
- Intro to nonparametric tests

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 means r=3
- Value of the j*th* observation for the *i*th factor level:

$$Y_{ij}$$

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

Cell means model

- Useful framework:
  ANOVA characterizes each observation as a deviation from the mean of the factor level
  - Or later: the mean of the subject, or the mean of the control variable...

Cell means model

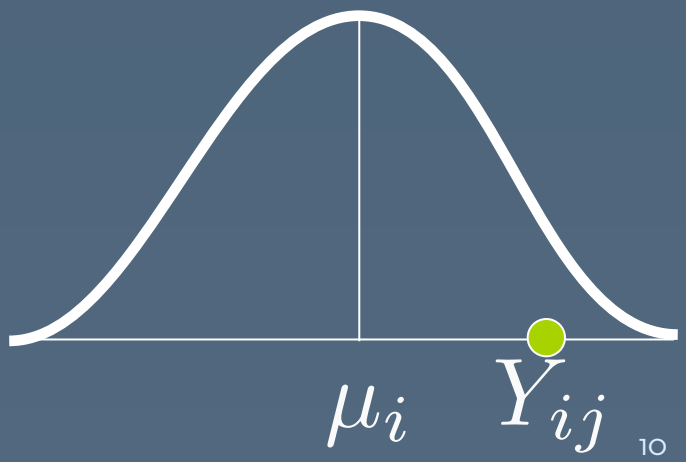
- Starter ANOVA model:

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

mean response for factor level *i*

error: difference between observed value and the mean

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



Cell means model

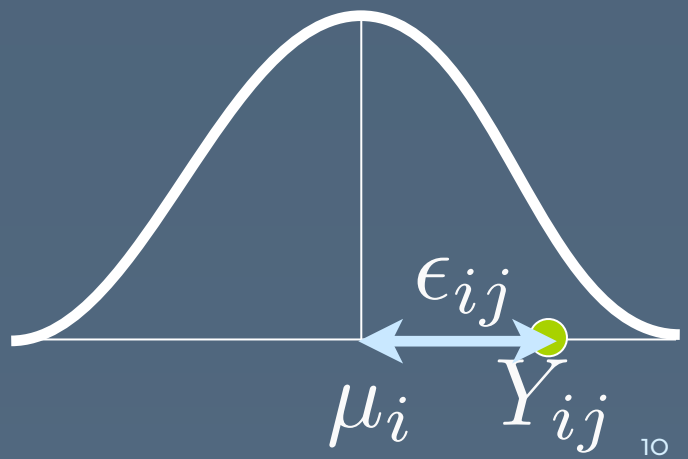
- Starter ANOVA model:

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

mean response for factor level *i*

error: difference between observed value and the mean

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



Partitioning the variance

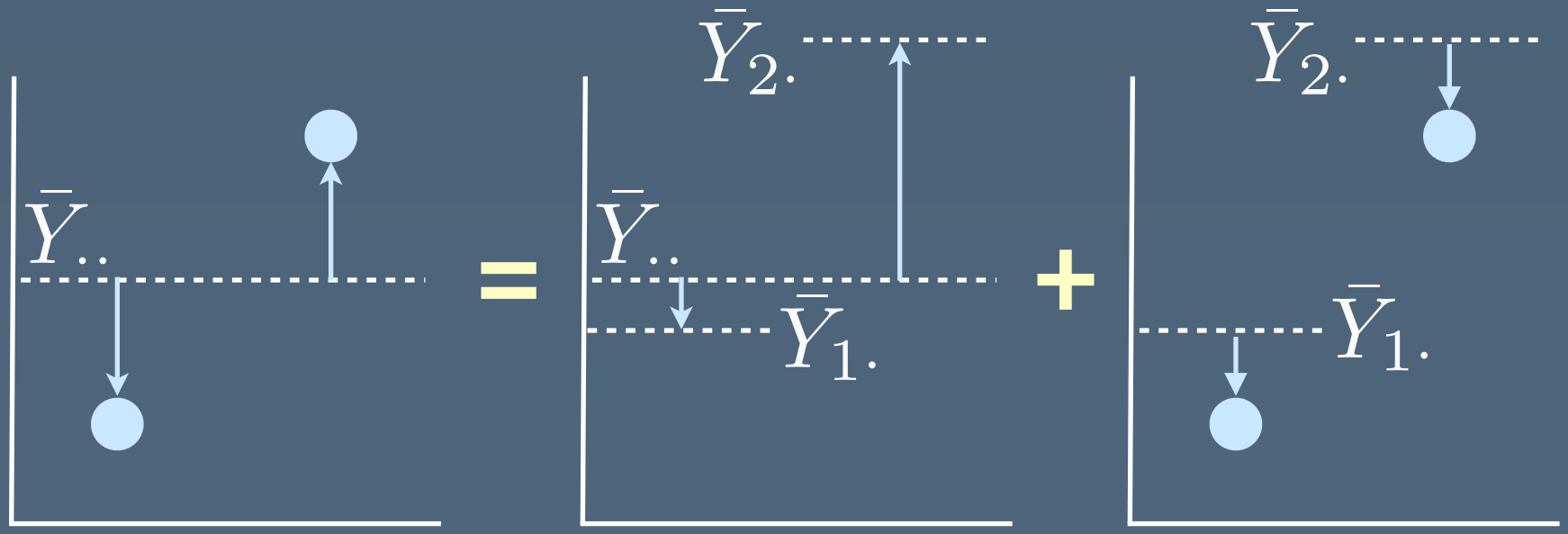
- The total variability in Y is the difference between each observation Y_{ij} and the grand mean $\bar{Y}_{..}$

bar is 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

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

Partitioning the variance



$$Y_{ij} - \bar{Y}_{..} = \bar{Y}_{i.} - \bar{Y}_{..} + Y_{ij} - \bar{Y}_{i.}$$

total deviation **treatment deviation** **error deviation**

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)

- Like in regression:

$$SSTO = SSTR + SSE$$

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*)

ANalysis Of VAriance (ANOVA)

- Like in regression:

$$SSTO = SSTR + SSE$$

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

ANalysis Of VAriance (ANOVA)

- Like in regression:

$$SSTO = SSTR + SSE$$

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

ANalysis Of VAriance (ANOVA)

- Like in regression:

$$SSTO = SSTR + SSE$$

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
- 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

SS MS F(2,21) p < .001

3 factor levels
24 observations

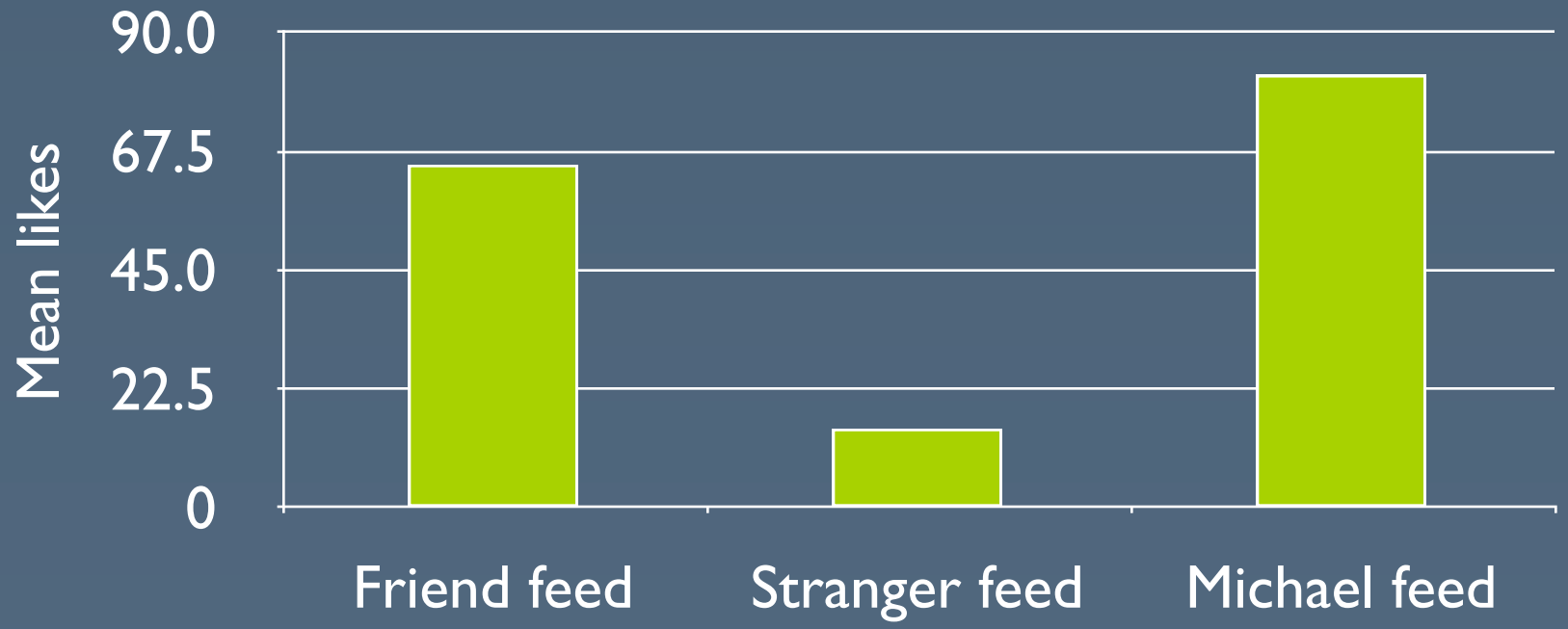
Posthoc tests

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"

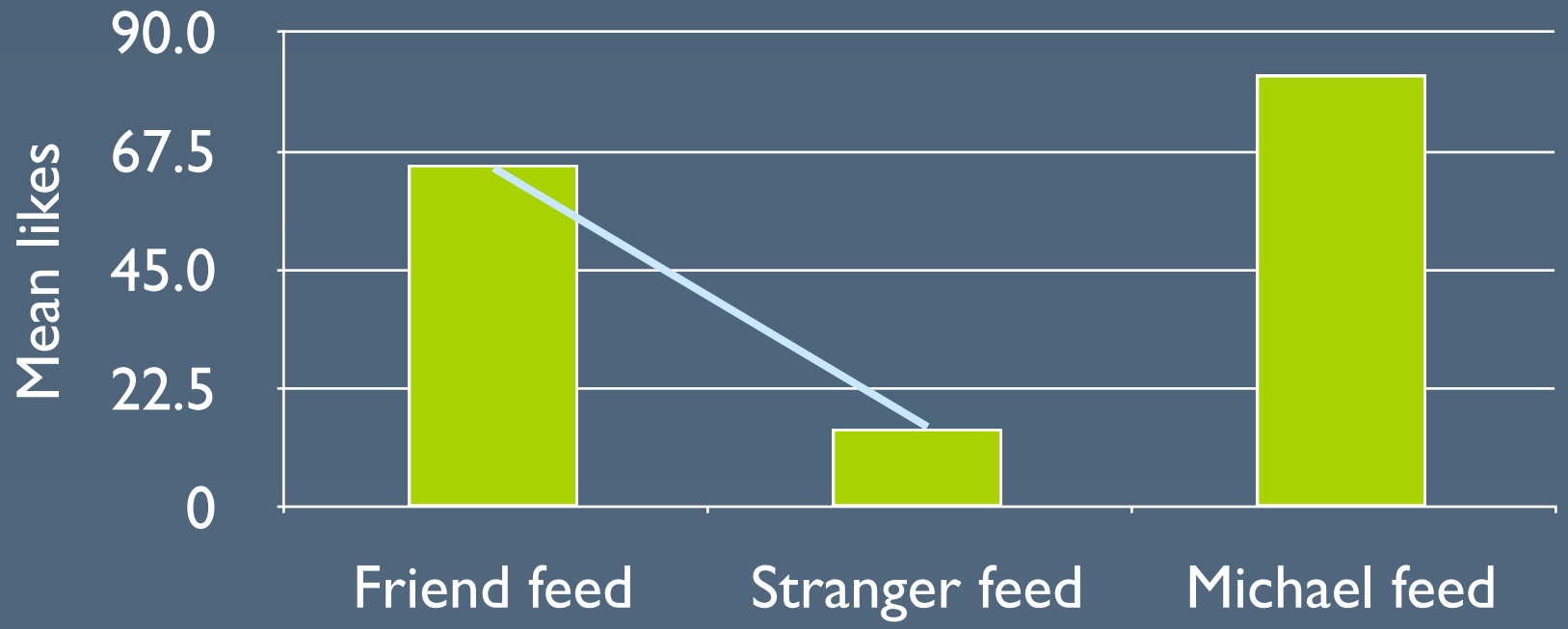
Estimating pairwise differences

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



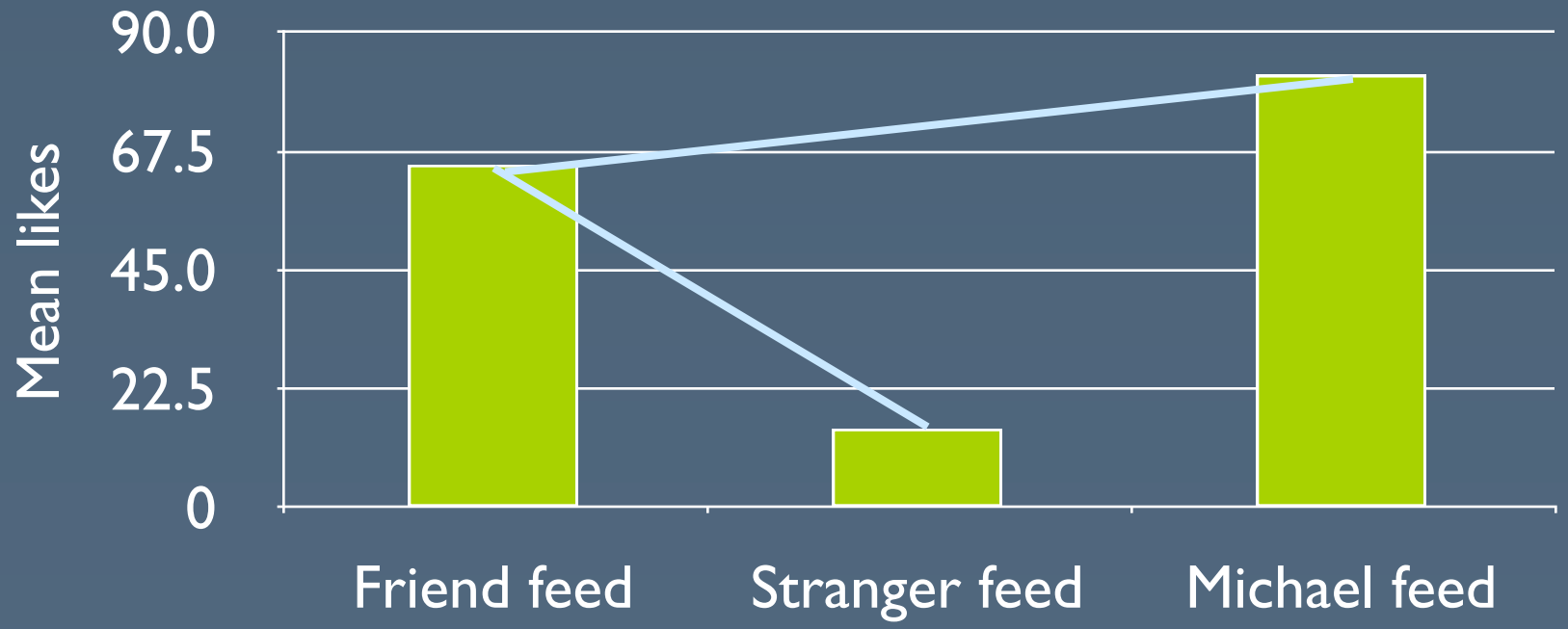
Estimating pairwise differences

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



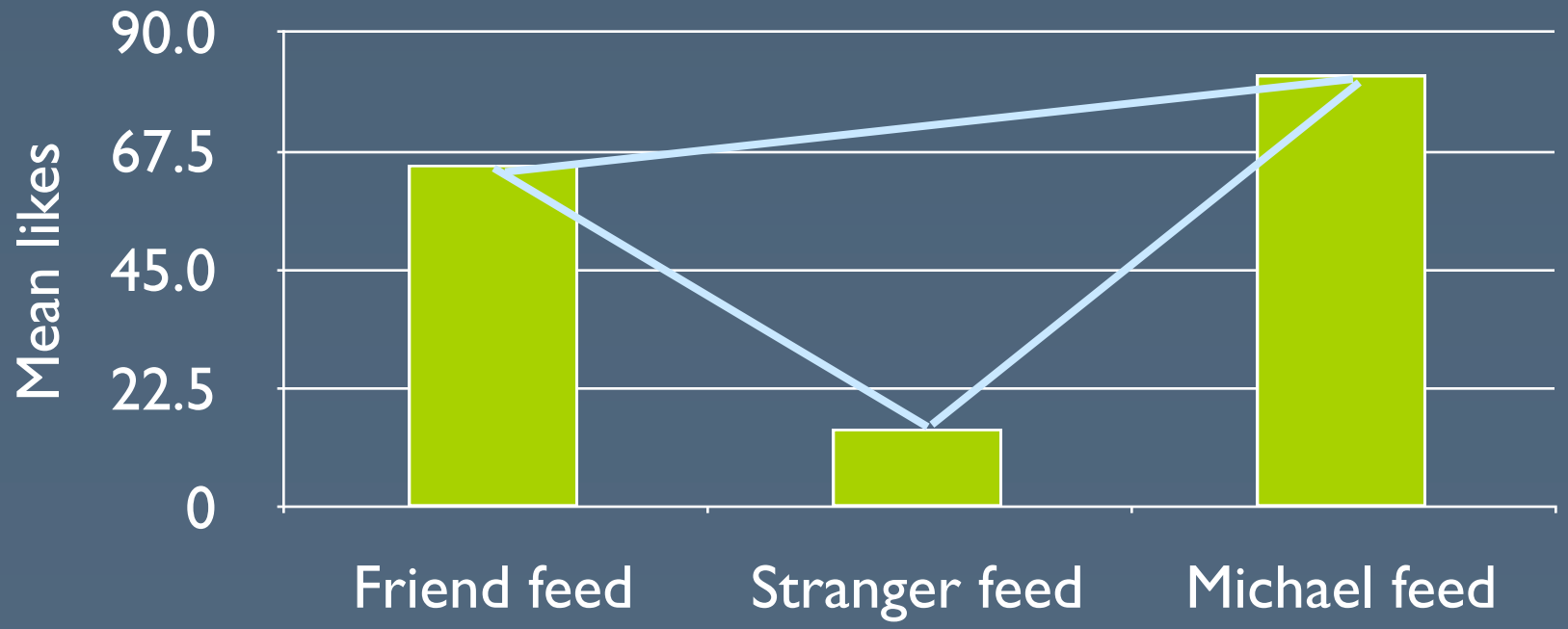
Estimating pairwise differences

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



Estimating pairwise differences

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



Comparing two means

- $D = \mu_i - \mu_j$ is a normal variable (sum of two normal variables)
- We studentize the statistic:

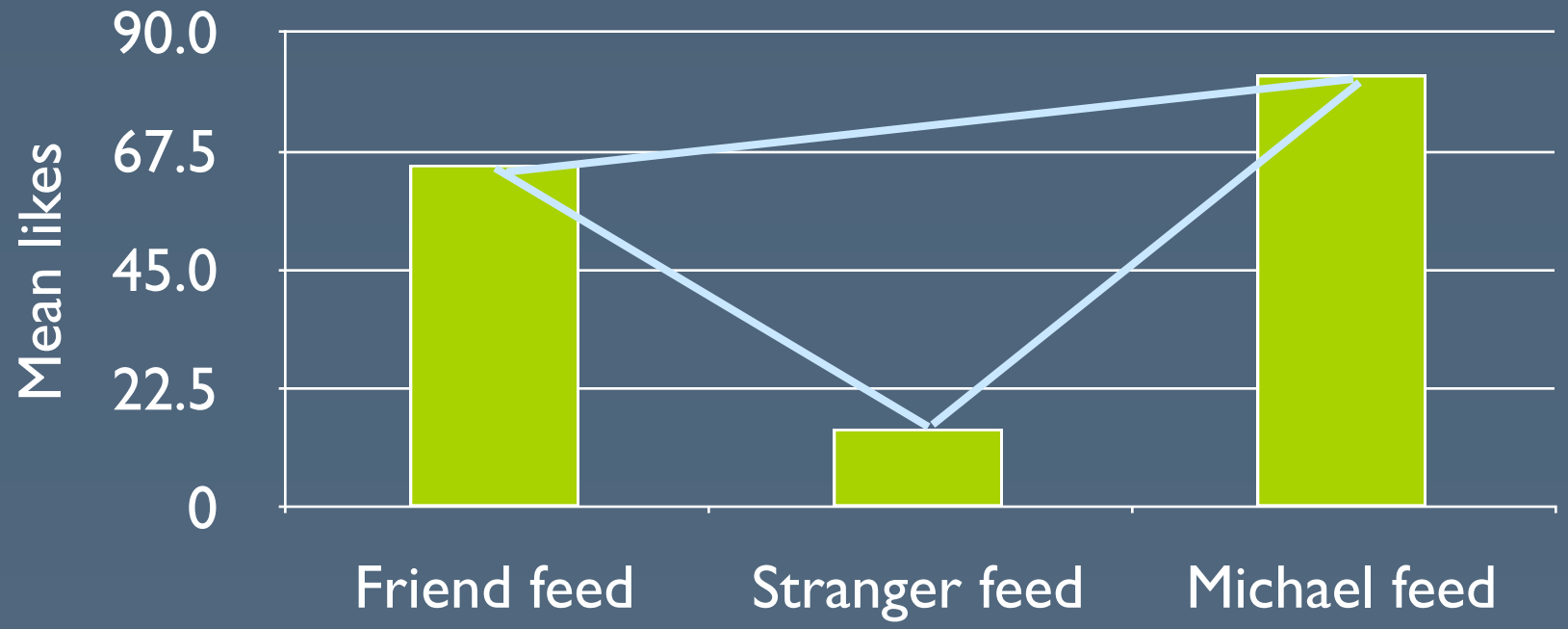
$$t^* = \frac{D}{s\{D\}}$$

- Now test:

$$|t^*| > t(1 - \alpha; n - r)$$

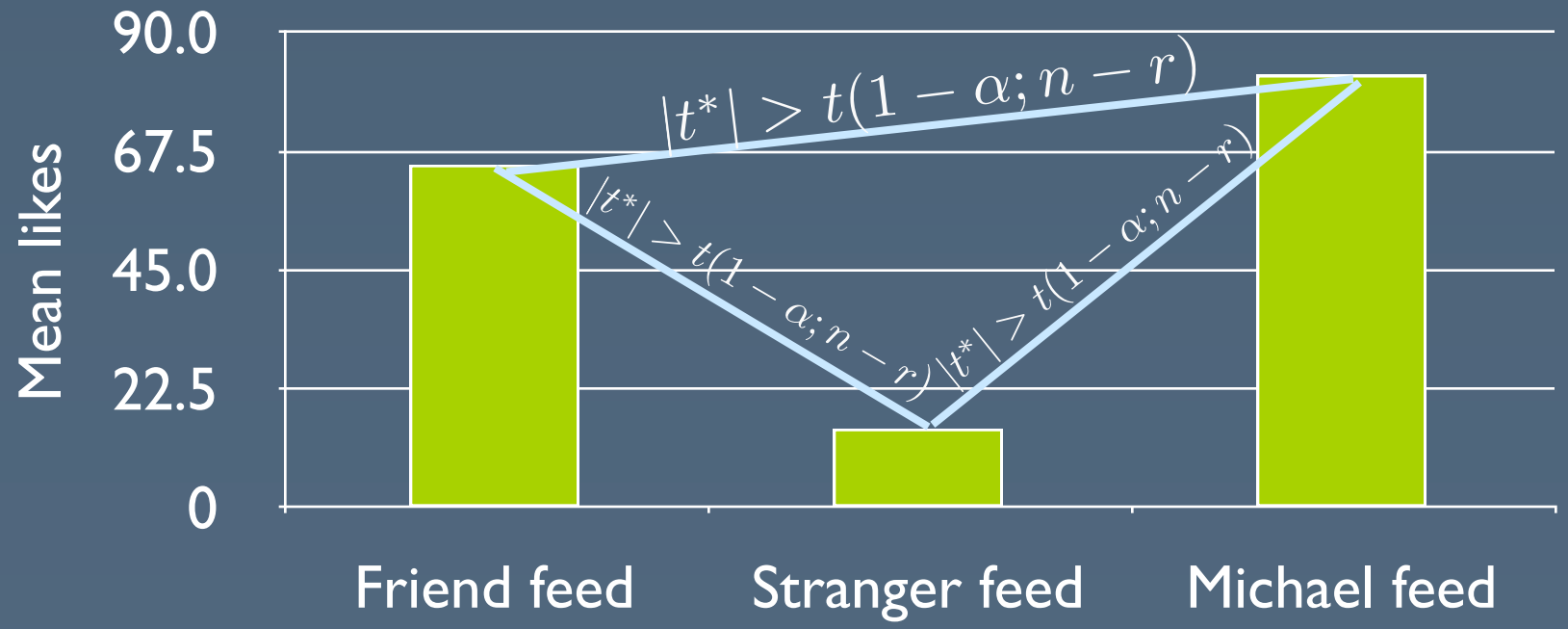
Estimating pairwise differences

- Testing each pair for significance



Estimating pairwise differences

- Testing each pair for significance



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
```

Tukey test

- Less conservative than Bonferroni
- Compares all pairs of factor level means

```
> TukeyHSD(aov)
  Tukey multiple comparisons of means
    95% family-wise confidence level

Fit: aov(formula = value ~ group, data = data)

$group
     diff         lwr      upr     p adj
B-A 1.375  0.1527988 2.597201 0.0257122
C-A 2.375  1.1527988 3.597201 0.0002167
C-B 1.000 -0.2222012 2.222201 0.1222307
```

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 a Tukey pairwise comparison 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 *i*th level of 1st factor *j*th level of 2nd factor

grand mean

difference between *i*th level of 1st factor and grand mean

difference between *j*th level of 2nd factor and grand mean

- 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

Interaction effects

- Interaction terms $(\alpha\beta)_{ij}$ capture the effect of a single constellation of the factor levels

$$(\alpha\beta)_{ij} = \mu_{ij} - (\mu_{..} + \alpha_i + \beta_j)$$
$$= 20 - (8 + 3 + -1) = 10$$

- Intuitively: how far off is the cell from the mean you would predict?

Full ANOVA model

$$Y_{ijk} = \mu_{..} + \alpha_i + \beta_j + (\alpha\beta)_{ij} + \epsilon_{ijk}$$

- We now have more sums of squares:
  - *SSTO*
  - *SSTR = SSA + SSB + SSAB*
  - *SSE*

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

Significant interaction?

- Significant interactions mean that you can't just report the main effects — the story is more complicated
- Inspect to figure it out:

	Pen	Touch
Technique A	15.3	21.1
Technique B	23.9	33.1
Technique C	32.9	44.9

Significant interaction?

- Significant interactions mean that you can't just report the main effects — the story is more complicated
- Inspect to figure it out:

	Pen	Touch
Technique A	15.3	21.1
Technique B	23.9	33.1
Technique C	32.9	44.9

The slower techniques (B, C) harm Touch more than Pen

Repeated measures ANOVA

Within-subjects studies

- Control for individual variation using the mean response for each participant
- Recall: single-factor ANOVA, simpler form:

$$Y_{ij} = \mu_{\cdot} + \tau_i + \epsilon_{ij}$$

grand mean

treatment effect of *i*th factor level

Within-subjects studies

- Control for individual variation using the mean response for each participant
- Recall: single-factor ANOVA, simpler form:

$$Y_{ij} = \mu_{\cdot} + \tau_i + \epsilon_{ij}$$

grand mean — **treatment effect of *i*th factor level**

- Shiny new repeated measures model:

$$Y_{ij} = \mu_{\cdot\cdot} + \rho_i + \tau_j + \epsilon_{ij}$$

mean effect of *i*th participant

Running a repeated measures ANOVA

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
```

More complex ANOVA variants

- In general, adding more complexity to the ANOVA model just means adding terms to:

$$Y_{ij} = \mu_{\cdot} + \tau_i + \epsilon_{ij}$$

- Examples:
  - controlling for participant (repeated measures)
  - controlling for another variable
  - nested designs
- It's called the General Linear Model (GLM)

Nonparametric tests

What if data is non-normal or not really continuous?

- Example: people rate an interface as “hate”, “ok”, or “love”
- Most tests we’ve discussed have been **parametric**: they assume normal distributions
- We need to consider outcomes that can’t be normally distributed, requiring **nonparametric** tests

Chi-square test

Things you know already

- We count 29 hates, 12 oks, and 19 loves.
- Are there significantly different numbers of votes for each category?

Ranks

- Many nonparametric tests require **ordinal** data so they can rank the results:
  “hate”, “love”, “love”, “meh”, “hate”
  to
  “hate”, “hate”, “meh”, “love”, “love”
- Transforming into ranks allows rough equivalents of many parametric tests

Equivalent nonparametric tests

Parametric	Nonparametric
Unpaired t-test	Mann-Whitney U
Paired t-test	Wilcoxon matched pairs
ANOVA	Kruskal-Wallis
Repeated measures ANOVA	Friedman test

Statistical pipeline

Always follow every step!

1. Visualize the data
2. Compute descriptive statistics (e.g., mean)
3. Remove outliers >2 s.d. from 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