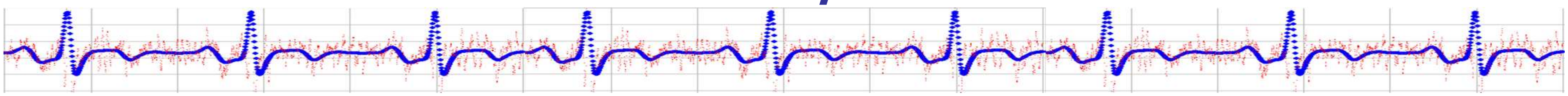
Empirical Research Methods in Information Science

IS 4800 / CS6350



Lecture 21

Outline

- Power
- One-way ANOVA

- Work in teams for T3 – Experimental!

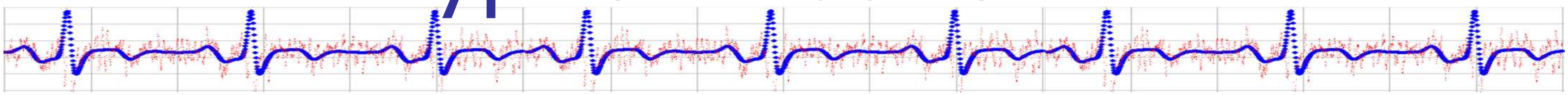
Power

- The "power" of a statistical test is its ability to detect differences in data that are inconsistent with the null hypothesis.
  - p(rejecting H0|H1)
  - Aka – the ability to find a significant result, if your hypotheses are actually true.
- What is it called when this fails (i.e., accepting H0 when H1 is true)?
- Why is this a bad situation?

Effect size

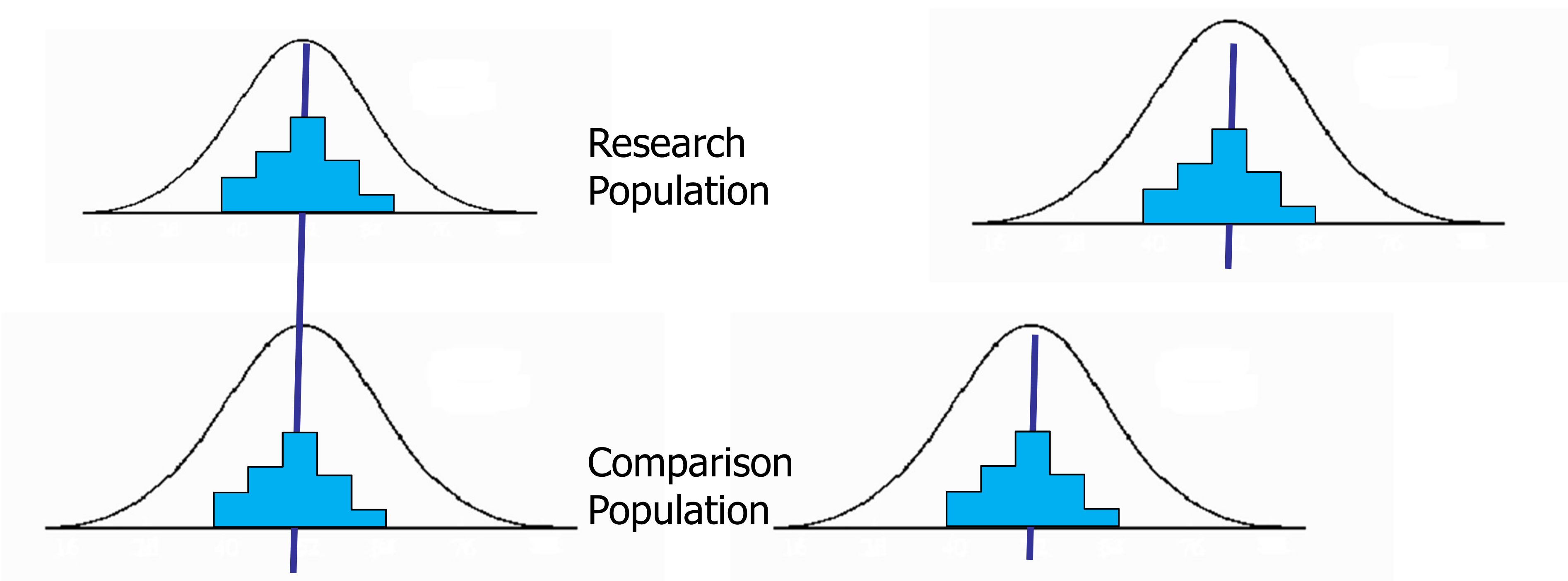
- The *amount* of measured difference between study conditions
- The greater the effect size, the easier it is to show there is a significant difference in your study (i.e., the greater the power)
- Effect size formula is different for each hypothesis test procedure
- Tabulated standard values for “small”, “medium”, and “large” effect sizes
- Only talk about effect size IF significance is established – but then DO present it in your results

The typical situation

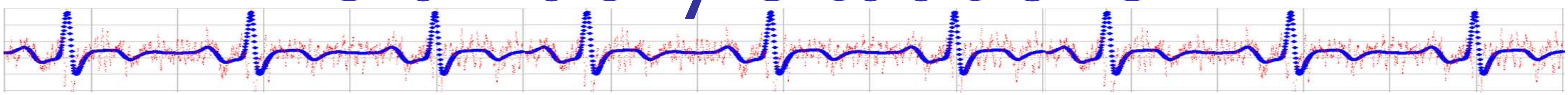


H0 Actually True

H1 Actually True

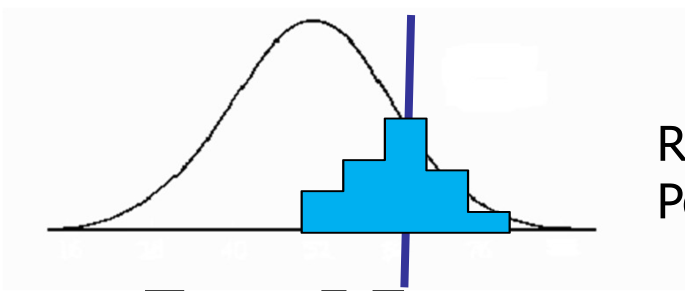


The unlucky situations

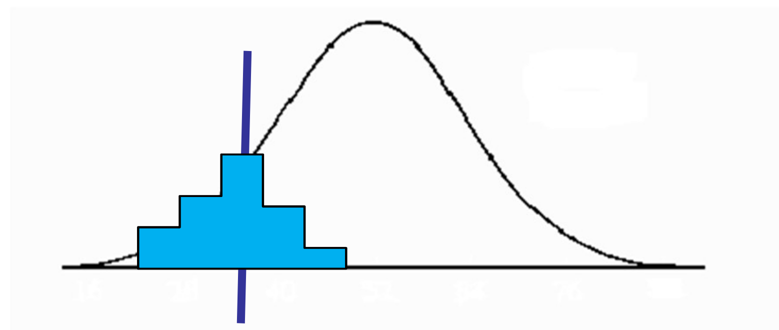


H0 Actually True

H1 Actually True

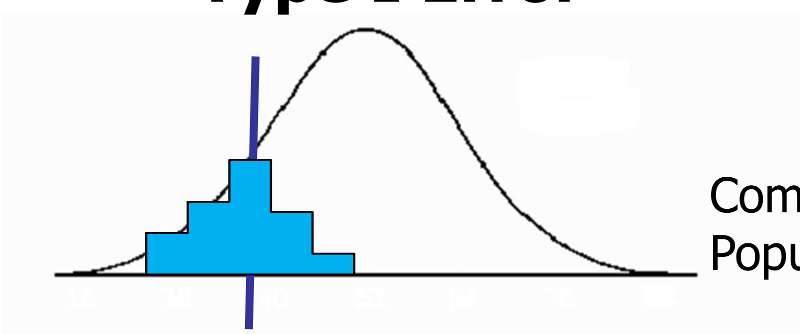


Research
Population

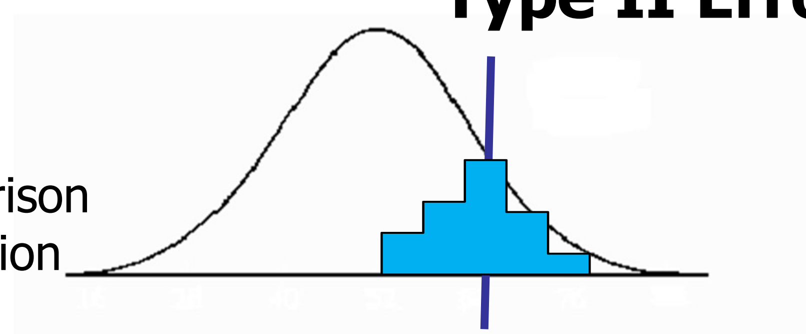


Type I Error

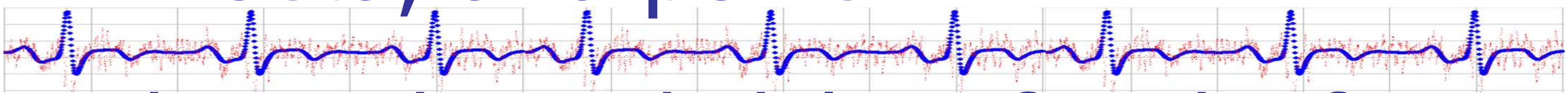
Type II Error



Comparison
Population



Relationship between alpha, beta, and power



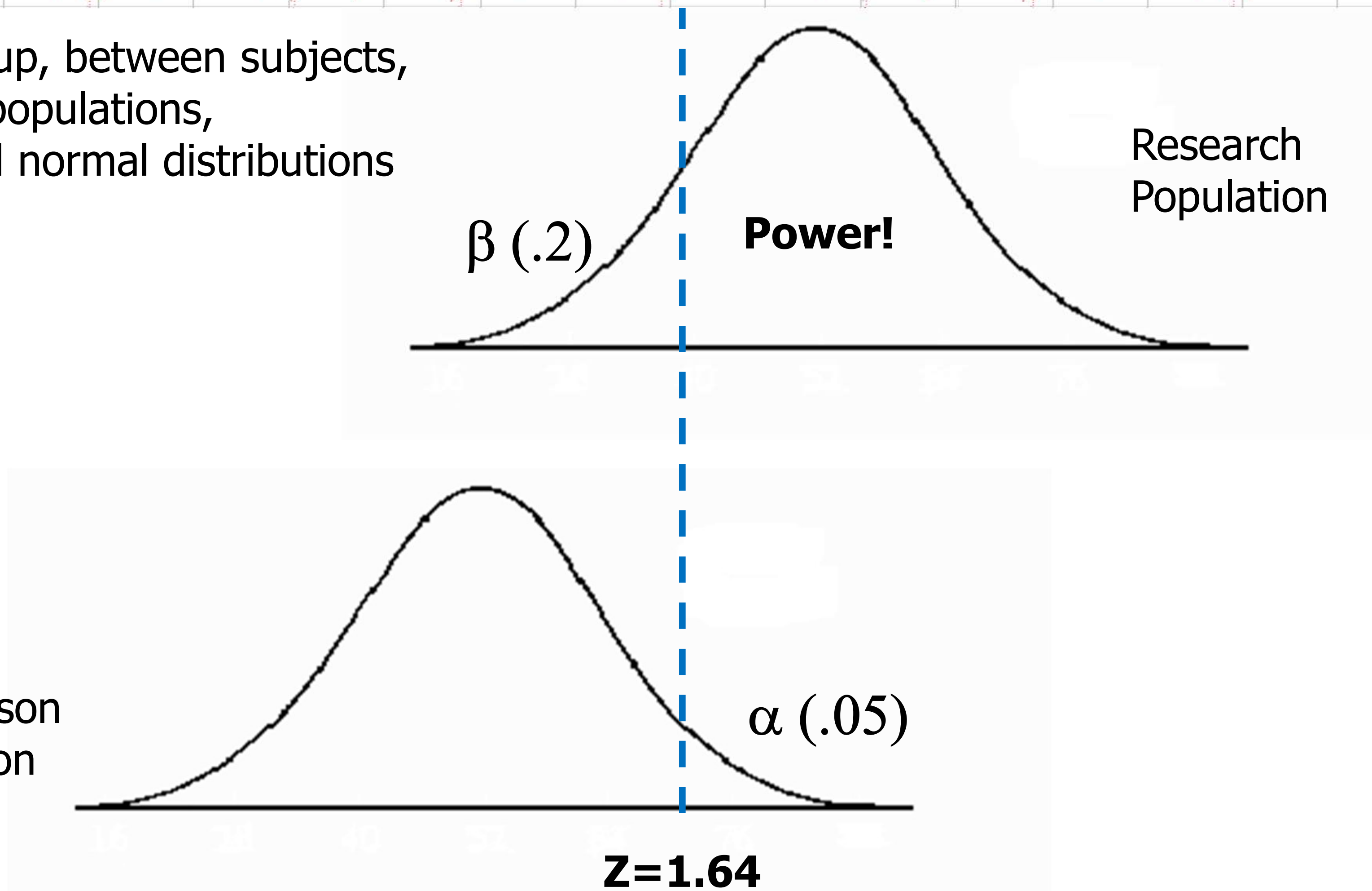
What is the probability of each of these situations occurring?

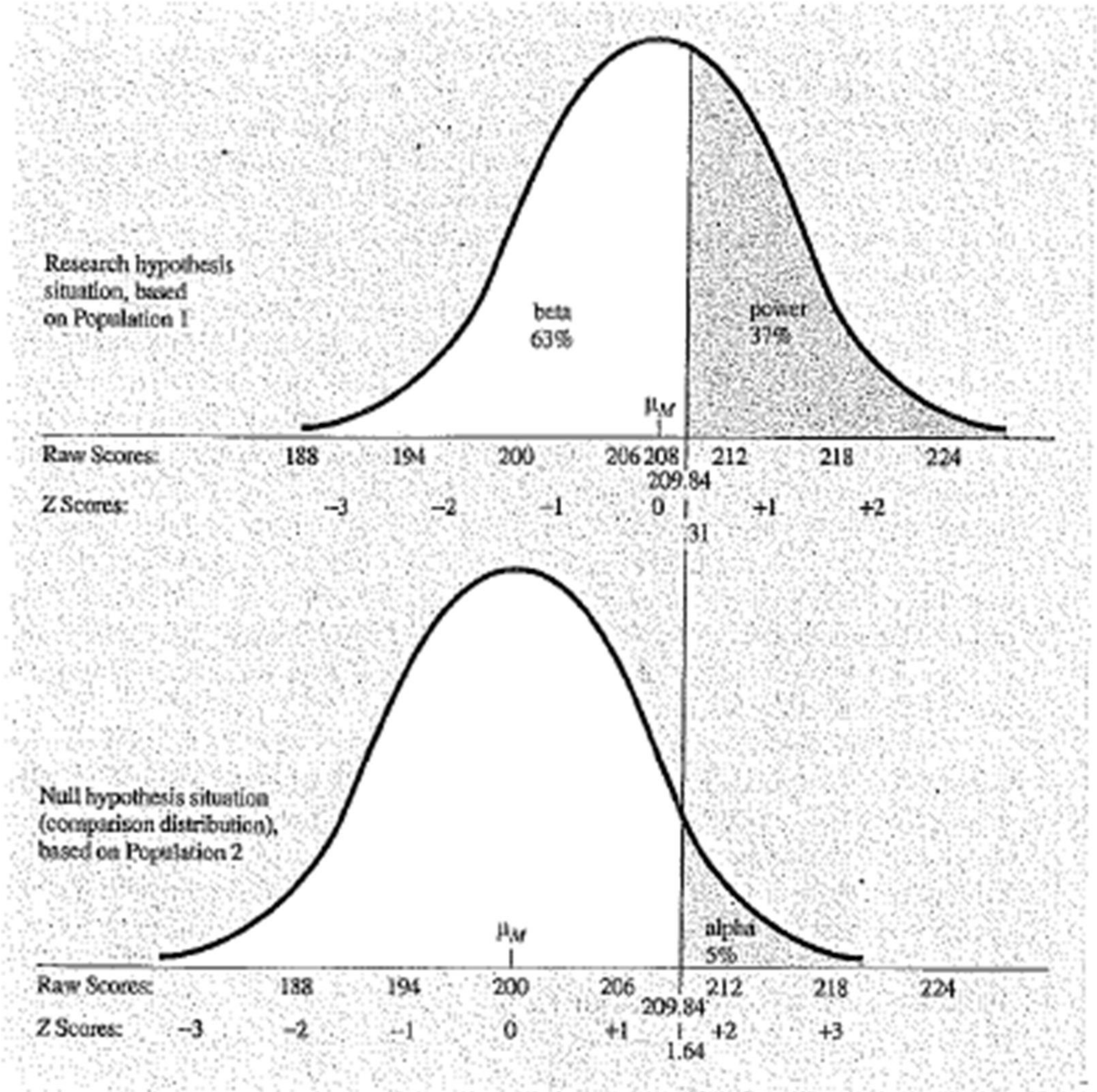
"The Truth"

	H1 True	**H1 False**
Decide to Reject H0 & accept H1	Correct p = power	Type I err $p = \alpha$
Do not Reject H0 & do not accept H1	Type II err $p = \beta$	Correct $p = 1-\alpha$

Relationship between power and effect size

Two group, between subjects,
normal populations,
standard normal distributions

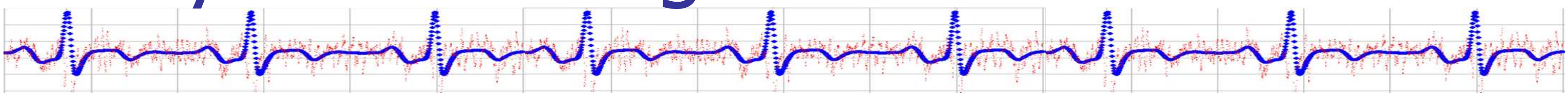



Research hypothesis situation, based on Population 1
beta 63%
power 37%
μ_M
Raw Scores: 188 194 200 206 208 212 218 224
209.84
Z Scores: −3 −2 −1 0 +1 +2
.31
Null hypothesis situation (comparison distribution), based on Population 2
μ_M
alpha 5%
Raw Scores: 188 194 200 206 212 218 224
209.84
Z Scores: −3 −2 −1 0 +1 +2 +3
1.64

Power Analysis

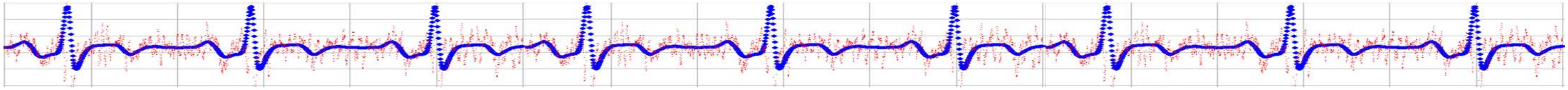
- Should determine number of subjects you need ahead of time by doing a ‘power analysis’
- Standard procedure (part of your study plan):
  - Determine statistic you will use
  - Fix alpha and beta (1-power) (and number of tails if appropriate)
  - Estimate expected effect size from prior studies
  - Then: Determine number of subjects you need
- Note: Power
  - Increases with effect size
  - Increases with sample size
  - Decreases with decreasing alpha

Power analyses are different depending on the statistical test you are using…



t-test for independent means

Effect Size



$$d = \frac{(\mu_1 - \mu_2)}{\sigma}$$

Parameters for population of individuals.
(so, use SD-pooled for t-test of indep means)

Cohen:
d~0.2 small
d~0.5 medium
d~0.8 large

Power table

TABLE 8–4 Approximate Power for Studies Using the *t* Test for Independent Means Testing Hypotheses at the .05 Significance Level

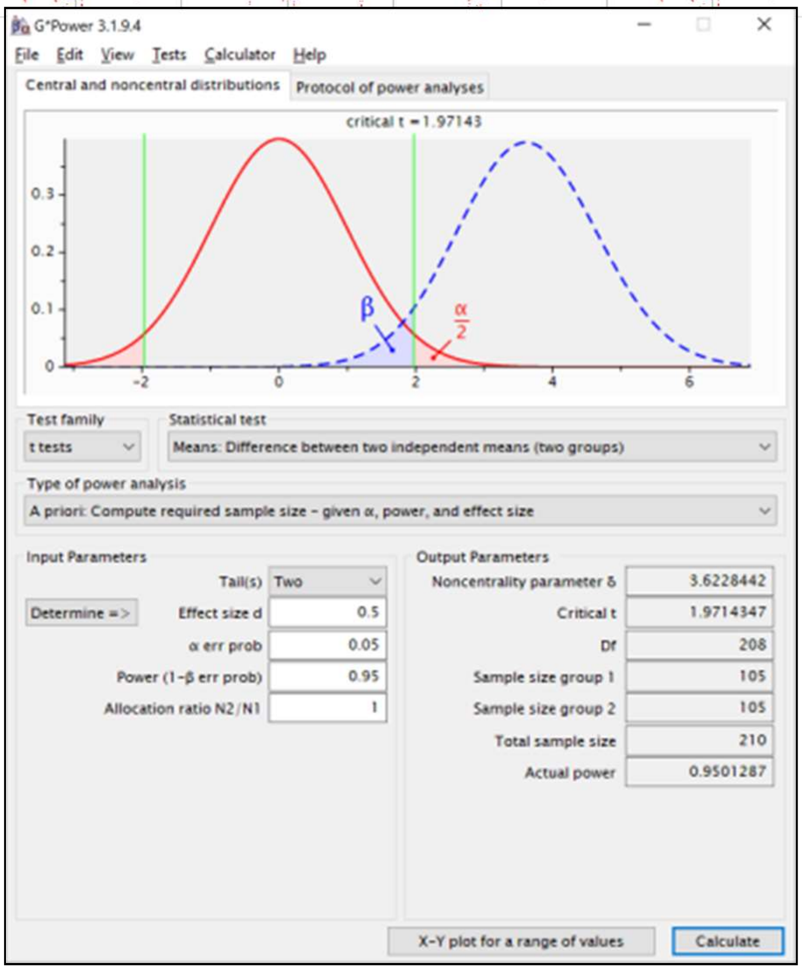
Number of Participants in Each Group	Effect Size Small (.20)	Medium (.50)	Large (.80)
One-tailed test			
10	.11	.29	.53
20	.15	.46	.80
30	.19	.61	.92
40	.22	.72	.97
50	.26	.80	.99
100	.41	.97	*
Two-tailed test			
10	.07	.18	.39
20	.09	.33	.69
30	.12	.47	.86
40	.14	.60	.94
50	.17	.70	.98
100	.29	.94	*

More Useful and Concise
(for practical purposes use a power calculator)

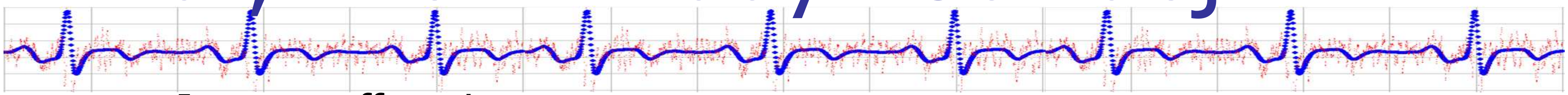
TABLE 8–5 Approximate Number of Participants Needed in Each Group (Assuming Equal Sample Sizes) for 80% Power for the *t* Test for Independent Means, Testing Hypotheses at the .05 Significance Level

	Effect Size		
	Small (.20)	Medium (.50)	Large (.80)
One-tailed	310	50	20
Two-tailed	393	64	26

G*Power


G*Power 3.1.9.4
File Edit View Tests Calculator Help
Central and noncentral distributions
Protocol of power analyses
critical t = 1.97143
0.3
0.2
0.1
0
-2
0
2
4
6
β
α
2
Test family
t tests
Statistical test
Means: Difference between two independent means (two groups)
Type of power analysis
A priori: Compute required sample size – given α, power, and effect size
Input Parameters
Tail(s) Two
Determine =>
Effect size d 0.5
α err prob 0.05
Power (1-β err prob) 0.95
Allocation ratio N2/N1 1
Output Parameters
Noncentrality parameter δ 3.6228442
Critical t 1.9714347
Df 208
Sample size group 1 105
Sample size group 2 105
Total sample size 210
Actual power 0.9501287
X-Y plot for a range of values
Calculate

But, I can’t study 786 subjects!



- Increase effect size
  - Increase difference in population means (change manipulation)
  - Decrease population variance (better measures, control more extraneous vars)
  - Redesign study to collect many trials of measures per subject
- Relax criteria for Type I error
  - Increase α threshold
  - Change from Two-tailed => one-tailed test
  - *Decreases credibility of your findings*
- Decrease power
  - *Decreases likelihood of getting a significant result*
- Use a different statistic
  - *If possible, maybe consult a statistician*

- Practically
  - usually, redesign experiment so that we have increased effect size or better measures for decreased variance
  - OR, call it a “pilot study”

Interpreting results: Significance & effect size

- Significance
  - Just indicates that it is likely there is a non-zero difference between populations
  - Says nothing about how big the difference is
- Effect Size
  - Only meaningful if result is significant
  - Indicates how big the difference is (usually normalized to number of std-deviations)

Interpreting results: Significance & effect size

- Significant & small effect => ?
  - Real difference, but slight.
  - Probably not of practical importance.
- Significant & large effect => ?
  - Real difference, likely meaningful.
- Significant & small sample => ?
  - Significant & possibly important.
- Non-significant & small sample => ?
  - Inconclusive
- Non-significant & large sample => ?
  - Evidence there really is no difference

Power & effect size for correlation

- Effect size = |r|
- Power, see table 11-7, pg 465 Aron
  - Usually, given
    - Expected effect size
    - Test criteria
      - Desired significance level (usually 0.05)
      - Desired power (usually 0.8)
      - Directionality of test

Table 11-7 Approximate Power of Studies Using the Correlation Coefficient (r) for Testing Hypotheses at the .05 Level of Significance

		Effect Size		
		Small ($r = .10$)	Medium ($r = .30$)	Large ($r = .50$)
Two-tailed				
Total N:	10	.06	.13	.33
	20	.07	.25	.64
	30	.08	.37	.83
	40	.09	.48	.92
	50	.11	.57	.97
	100	.17	.86	*
One-tailed				
Total N:	10	.08	.22	.46
	20	.11	.37	.75
	30	.13	.50	.90
	40	.15	.60	.96
	50	.17	.69	.98
	100	.26	.92	*

*Power is nearly 1.

Table 11-8, Aron

Approximate number of participants needed for 80% power for a study using the correlation coefficient (r) for testing a hypothesis at the .05 significance level

Effect size		
Small (r=0.1)	**Medium (r=0.3)**	**Large (r=0.5)**
783	**85**	**28**

Effect size & power for X^2 test for independence

- Completely different formulas than for Pearson r or t-test.
- Dependent on df.
- For 2x2, effect size = "phi"

$$\sqrt{\frac{x^2}{N}}$$

Effect Size & Power for X^2

Table 13-10 Approximate Total Number of Participants Needed for 80% Power for the Chi-Square Test for Independence for Testing Hypotheses at the .05 Significance Level

	Effect Size		
Total *df*	Small	Medium	Large
1	785	87	26
2	964	107	39
3	1,090	121	44
4	1,194	133	48

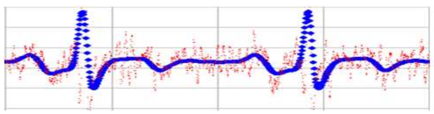


Table 13-9 Approximate Power for the Chi-Square Test for Independence for Testing Hypotheses at the .05 Significance Level

		Effect Size		
Total *df*	Total *N*	Small	Medium	Large
1	25	.08	.32	.70
	50	.11	.56	.94
	100	.17	.85	*
	200	.29	.99	*
2	25	.07	.25	.60
	50	.09	.46	.90
	100	.13	.77	*
	200	.23	.97	*
3	25	.07	.21	.54
	50	.08	.40	.86
	100	.12	.71	.99
	200	.19	.96	*
4	25	.06	.19	.50
	50	.08	.36	.82
	100	.11	.66	.99
	200	.17	.94	*

*Nearly 1.

Computing effect size

- Some authors do not include means & stddevs (per group) in their article…
- R package ‘compute.es’ contains a variety of methods for computing effect size given other info (e.g., t score, N1, N2)

- Morale: Always include means & stddevs
- Better: Report effect sizes yourself!

T3 planning

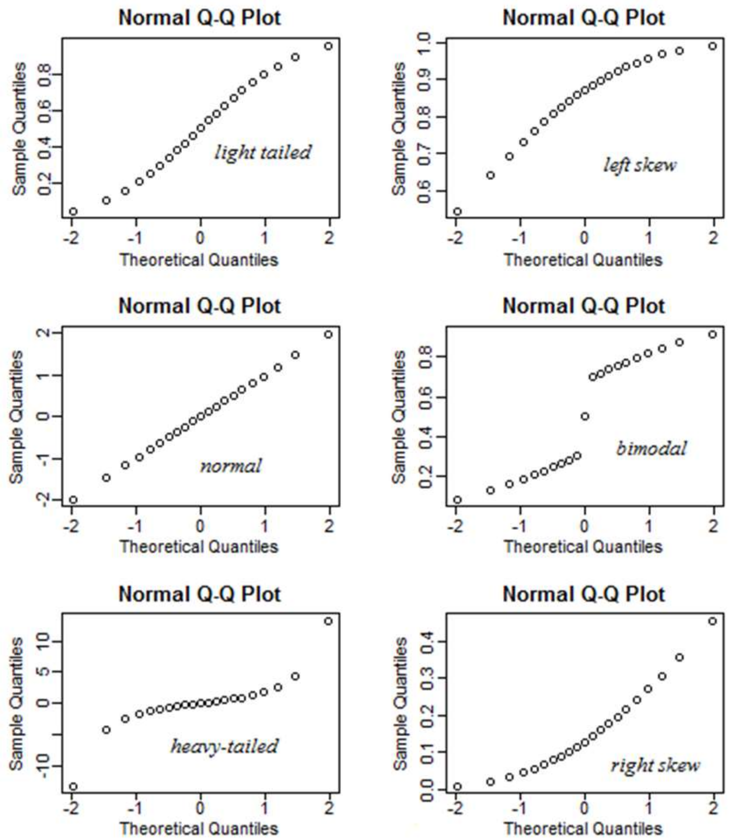
T1	T2	T3
Justin	Justin	Justin
Travis	Binh	Kenneth
Kenneth	Ian	Atamai
	Jake	
Zach		Travis
Bin	Travis	Zach
Ian	Jonathan T.	Eli
	Hao	Wilson
Eli		
Wilson	Kenneth	Binh
Jake	Eli	Erica Y.
	Atamai	Hao
Jonathan T.		
Atamai	Zach	Ian
Erica Y.	Wilson	Jake
Hao	Erica Y.	Jonathan T

Are my data normal?

- Eyeballing histogram is a crude measure
- Inspect Q-Q plot (quantile-quantile)
  - Compare shapes of distributions by plotting quantiles against each other
- Run statistical test

Python Guide + https://machinelearningmastery.com/a-gentle-introduction-to-normality-tests-in-python/

Q-Q plot



http://seankross.com/2016/02/29/A-Q-Q-Plot-Dissection-Kit.html

Types of Study Designs

- Qualitative
  - Ethnography
- Quantitative
  - Descriptive
  - Correlational
  - Demonstrative
  - Experimental
    - Between-subjects
      - Single factor, two-level
    - Within-subjects
      - Single factor, two-level

Factor = IV

Levels = different values of the factor

1-factor, N-level, between-subjects (N>2) Experimental Design

- Trivial generalization of two-level between-subjects design
- Randomize uniformly across the treatment levels
  - Random number generator
  - Blocked randomization still works
  - Baseline analysis generalizes to N
- Everything else is the same as 2 level

Accompanying Statistics

- Experimental
  - Between-subjects
    - Single factor, N-level (for N>2)
      - One-way Analysis of Variance (ANOVA)
    - Two factor, two-level (or more!)
      - Factorial Analysis of Variance
      - AKA N-way Analysis of Variance (for N IVs)
      - AKA N-factor ANOVA
  - Within-subjects (for N>2 treatments)
    - Repeated-measures ANOVA (not discussed)
      - AKA Within-subjects ANOVA

Basic Logic of ANOVA

- Null hypothesis
  - Means of all groups are equal.
  - H0: $\mu_1 = \mu_2 = \mu_3 \ldots = \mu_n$

Analyze this using variance!

- Test: do the means differ more than expected given the null hypothesis?

- Terminology
  - Group = Condition = Cell = treatment

ANOVA: Single factor, N-level (for N>2)

- The *Analysis of Variance* is used when you have more than two groups in an experiment
  - The *F-ratio* is the statistic computed in an Analysis of Variance and is compared to critical values of *F*
  - A significant overall *F* may require further planned or unplanned (*post hoc*) follow-up analyses
  - The analysis of variance may be used with unequal sample size (weighted or unweighted means analysis)

1-factor, 2 level?

- Could use ANOVA, but t-test between independent means simpler and gives same answer

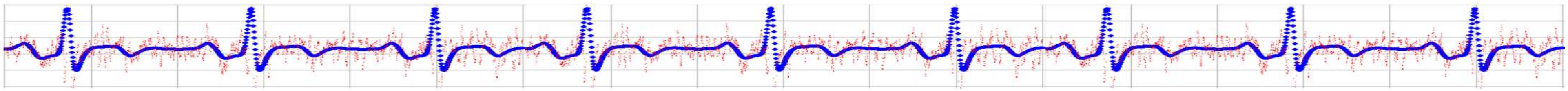
Pop. variance from variation within samples

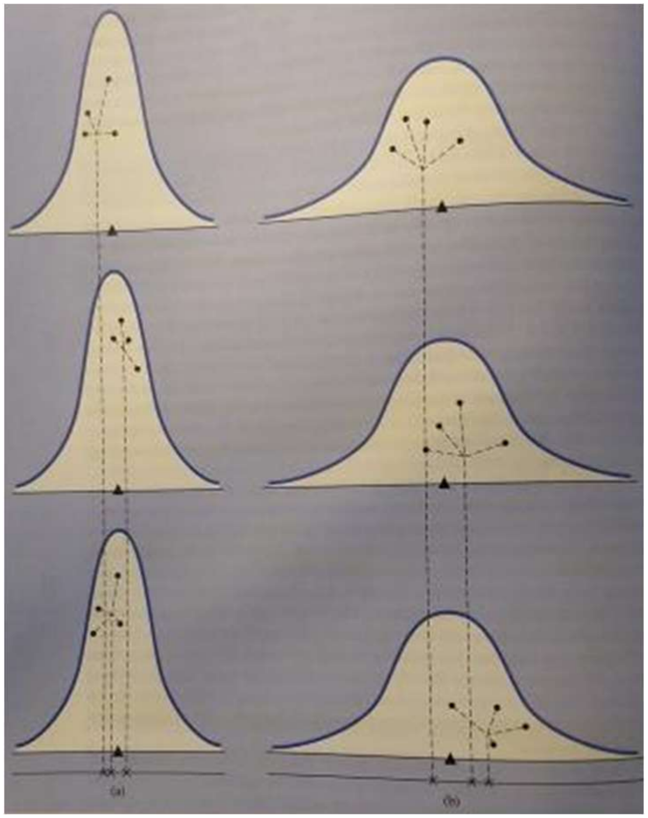
- As with t test
  - Don't know true population variances
  - Estimate from samples
  - Assume populations have same variance
- Average estimates of each sample into a within-groups estimate of pop. variance

How far apart means are doesn't matter. Focus only on variation inside each population. Thus, not affected by whether null hypothesis is true.

Pop. variance from variation between means of samples

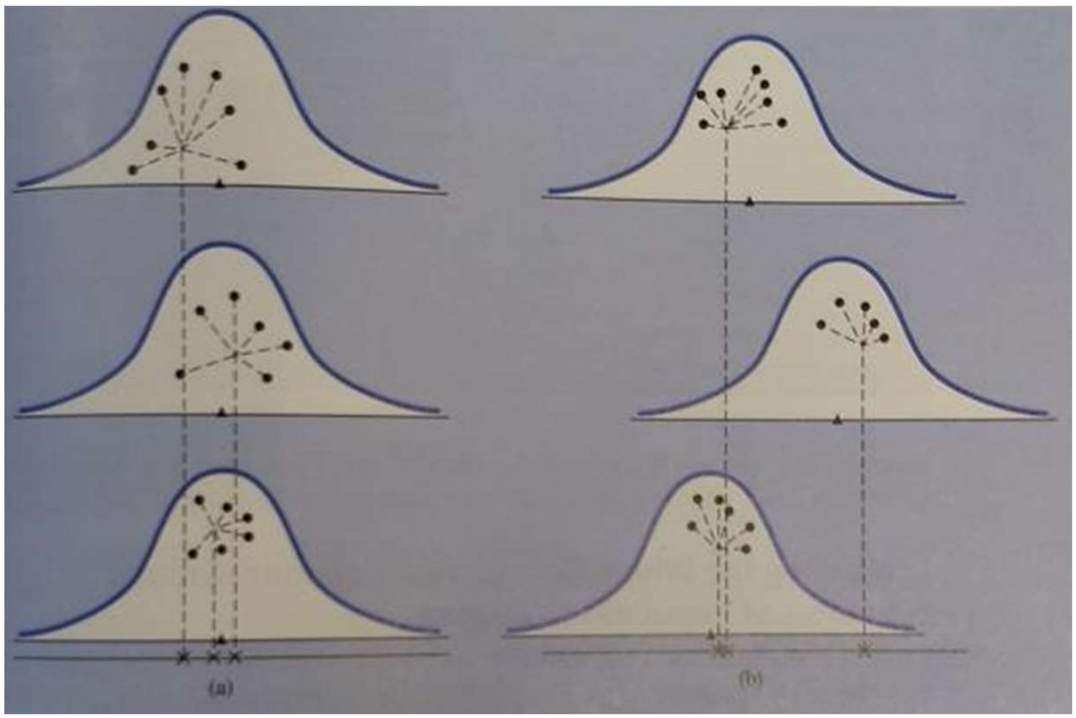
- The more variance there is within several identical populations, the more variance there will be among the means of samples when you take a random sample from each population





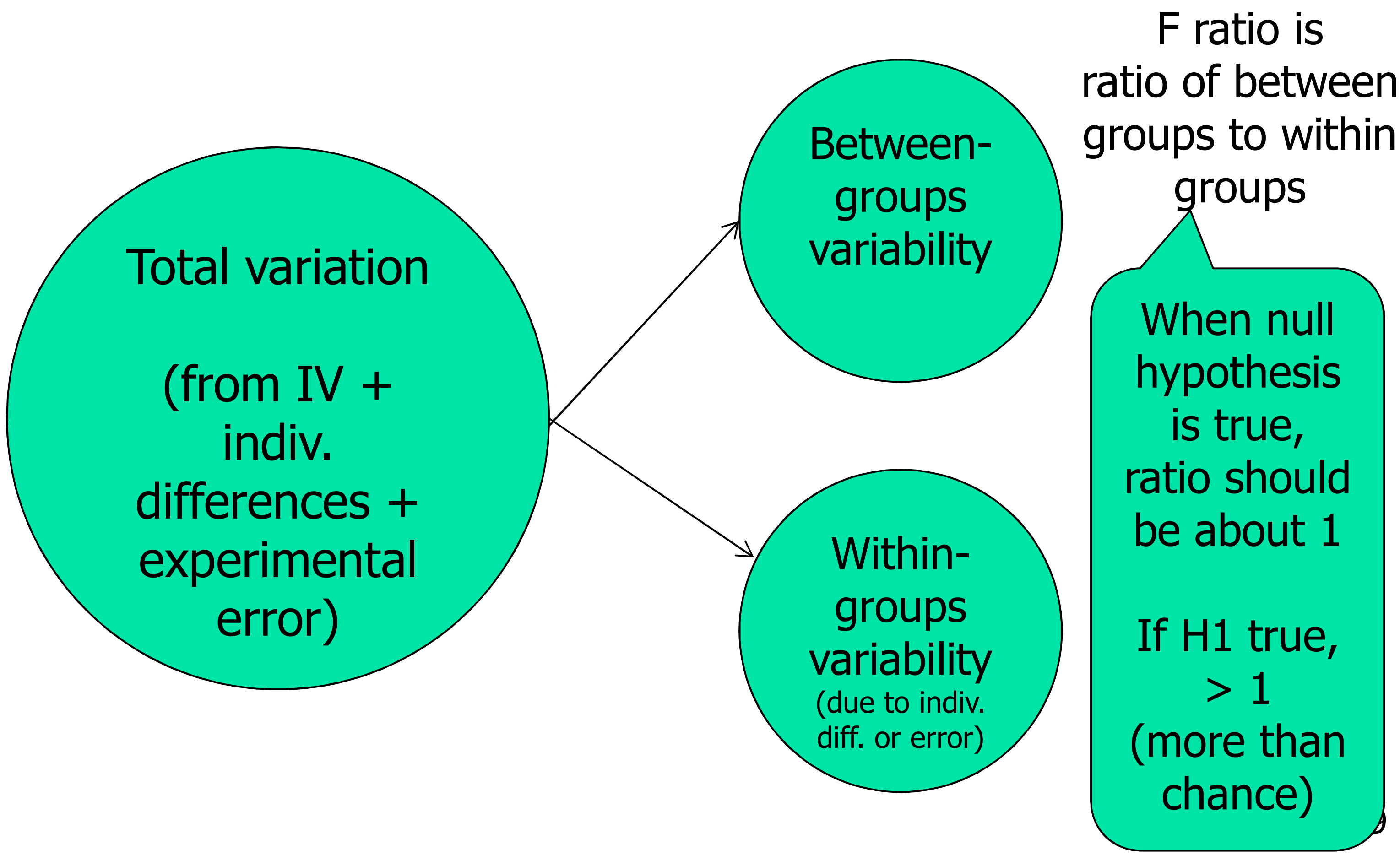
- Means of pop. the same, but means of samples are not
- Samples means from populations that have small variance have less variance amont them

Implication: Estimate variance in each pop from variation in means of samples



- Spread (right) due to differences in population means

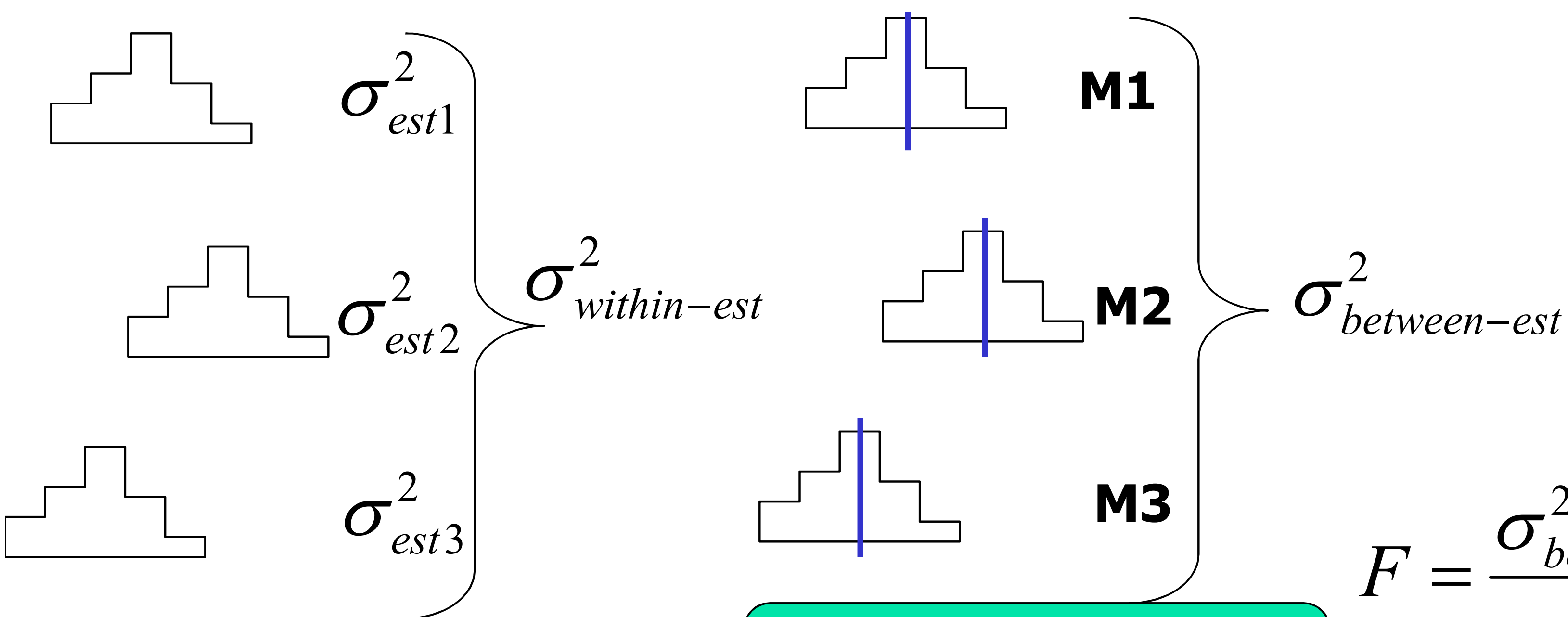
ANOVA – F ratio (F for Sir Ronald Fisher)



One-Way ANOVA – Assuming Null Hypothesis is True…

Within-Group Estimate Of Population Variance

Between-Group Estimate Of Population Variance



$$F = \frac{\sigma^2_{between-est}}{\sigma^2_{within-est}}$$

Somewhat analogous to a signal-to-noise ratio

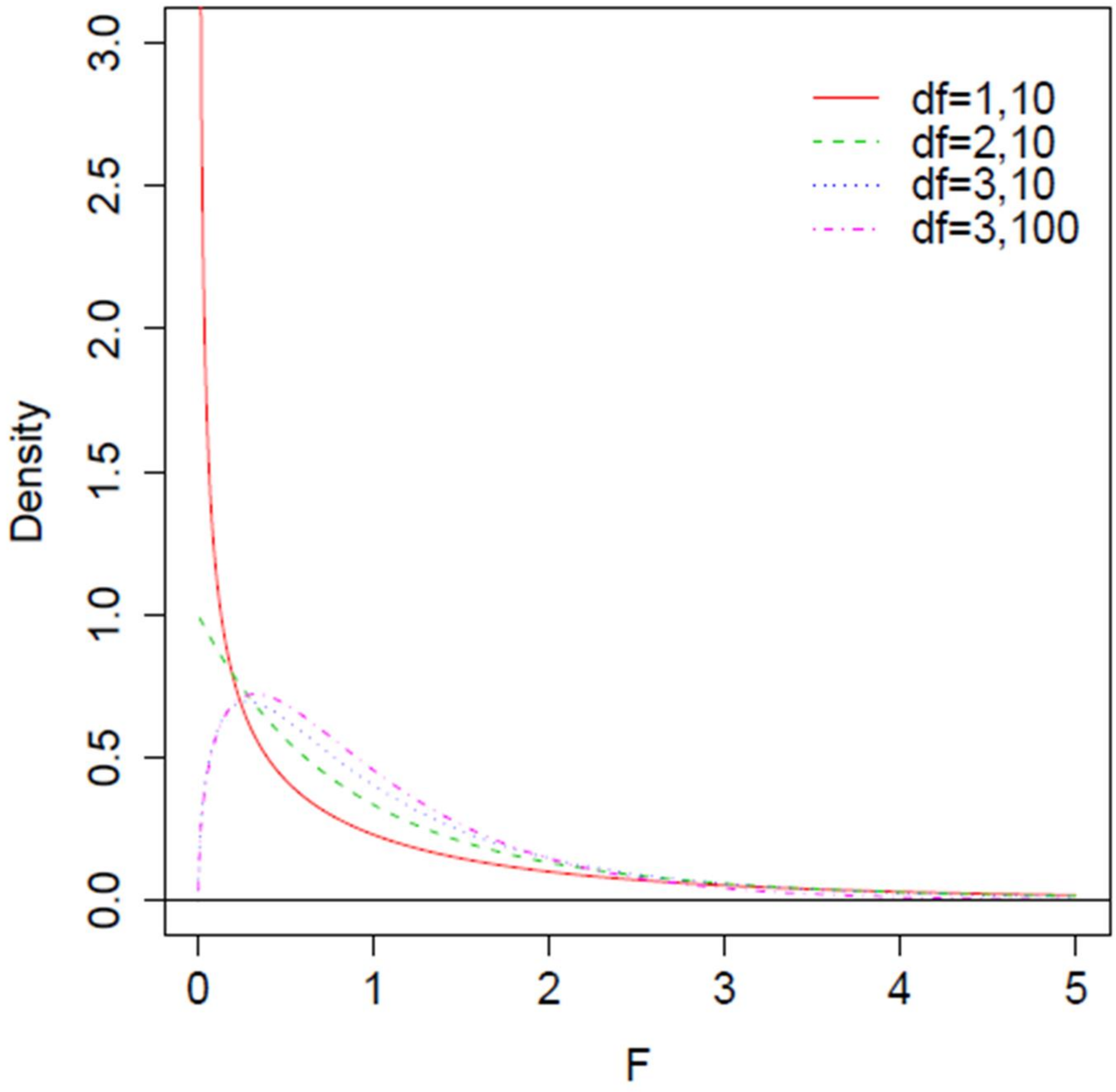
Degrees of freedom

- F(between-df,within-df)

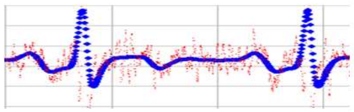
- beween-df = num groups - 1
- within-df = sum df for each group

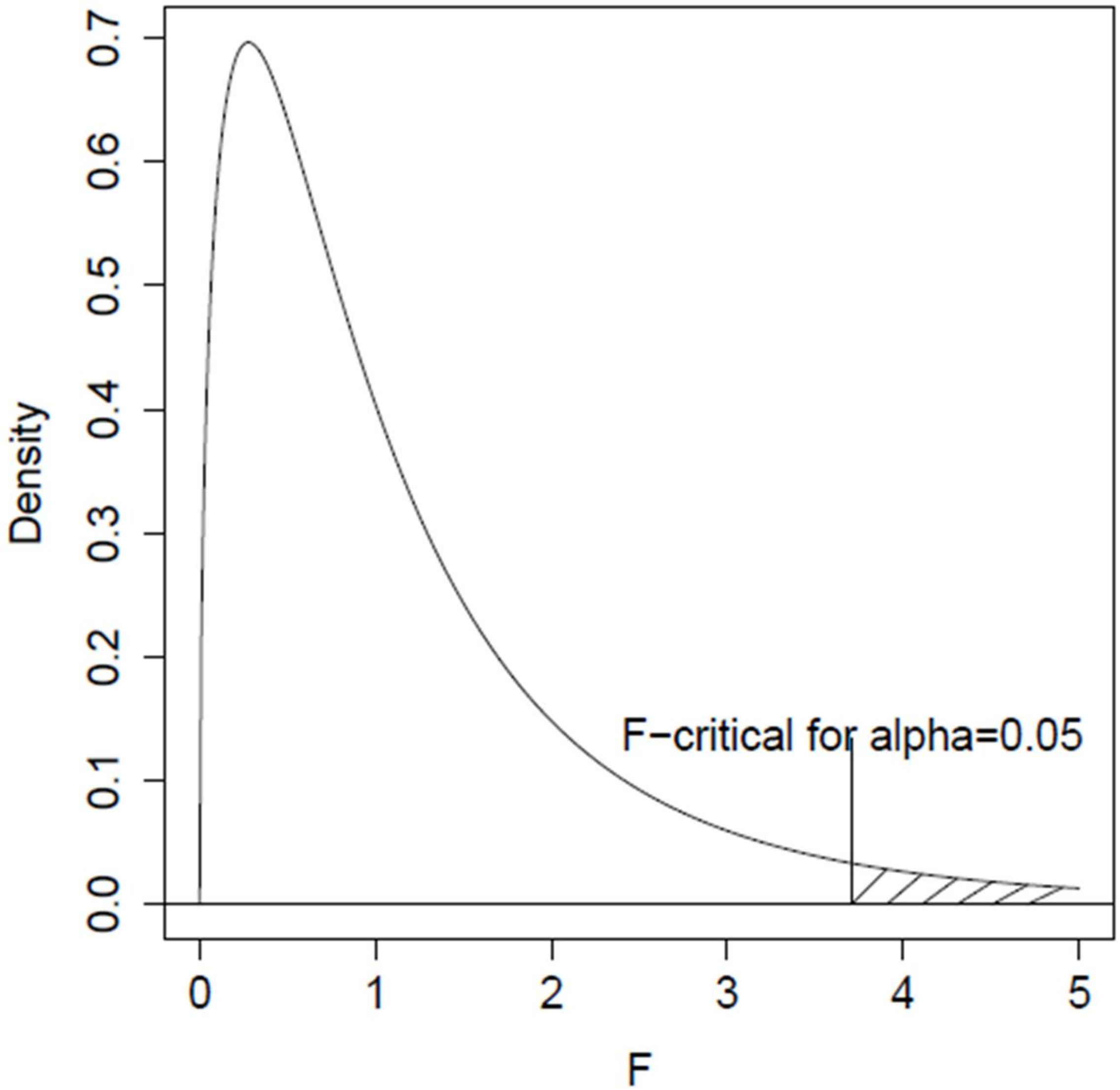
- Each group df = N_{group}-1
  - So, within-df = total N – num groups

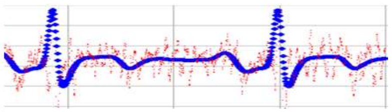
Sample F Distributions



Sample critical value for F(3,10)







Interpreting F ratio

- Significant F ration:
  - At least some of the differences among means probably not caused by chance but by variations in IV

  - DOES NOT tell you where! Do planned or unplanned test between means:
    - Planned (specific, pre-experimental hypotheses)
    - Unplanned (post hoc comparisons)

Planned contrasts

- Can use pairwise F tests or *t* tests
- Two types of error to consider:
  - Per-comparison error (alpha for each comparison)
  - Familywise error (takes into account probability of error given repeated tests
    $\alpha_{FW} = 1 - (1 - \alpha)^c$
    c is the number of comparisons
    (With c =4, α=.05, 3+ times chance to get at least one significant result)
- Correction example:
  - Bonferroni procedure (Dunn's test)
    (divide alpha by number of tests)

Post hoc analysis

- Bonerroni often no longer practical (adjusted alpha too small, power for any comparison too low)
- There are many post hoc tests (B&A 452)
  - Most obvious: Fisher’s Least Significant Difference (LSD)
    - Same as t-tests on every pair of treatments
    - Has inflated Type I error due to multiple tests
  - Many others: Sheffe,, Tukey, Dunnett etc.

Post hoc analysis

- Bonerroni often no longer practical (adjusted alpha too small, power for any comparison too low)
- There are many post hoc tests (B&A 452)
  - Most obvious: Fisher’s Least Significant Difference (LSD)
    - Same as t-tests on every pair of treatments
    - Has inflated Type I error due to multiple tests
  - Many others: Sheffe,, Tukey, Dunnett etc.

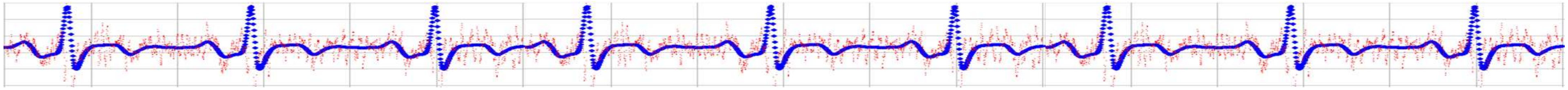
Example post hoc test

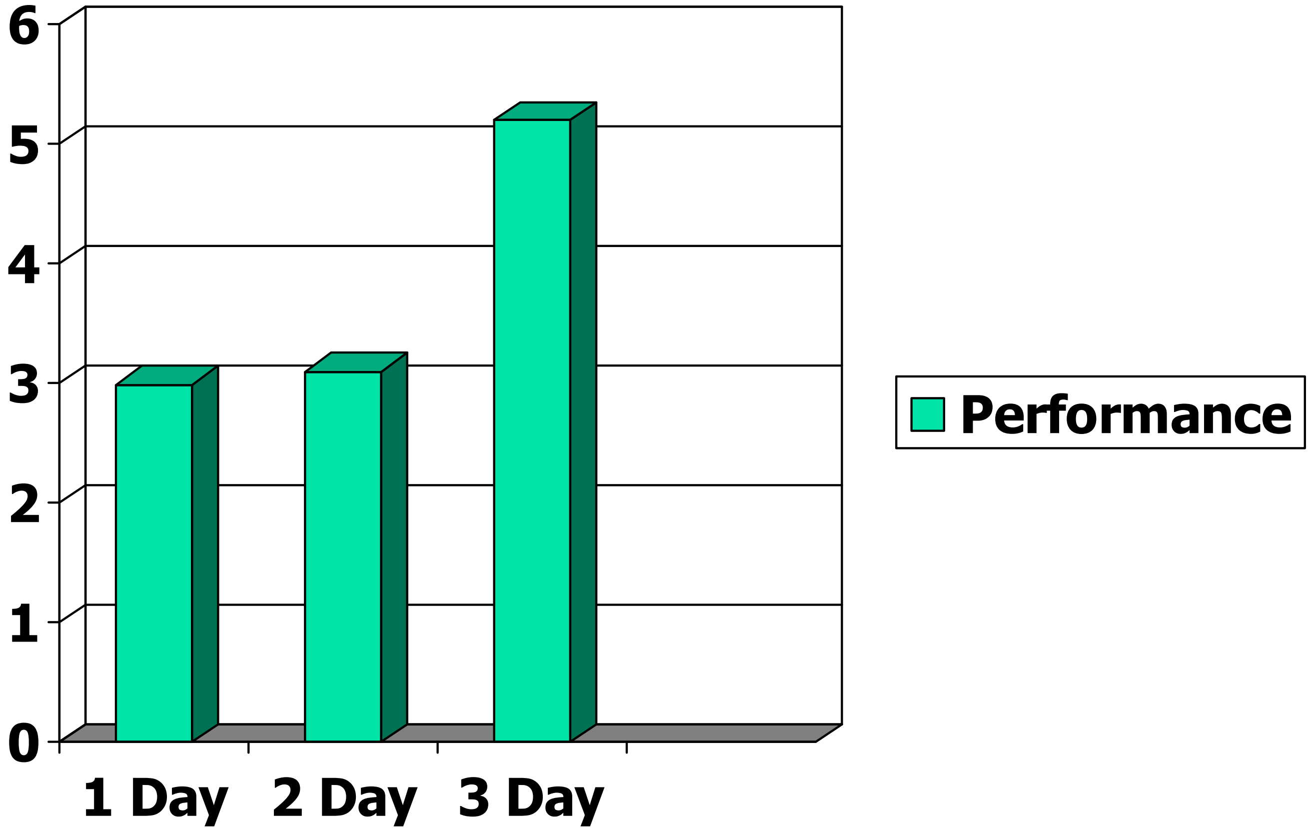
- Scheffe
  - Figure F for comparison in usual way
  - Divide F by the overall study's $df_{Between}$ (number of groups – 1)
  - Compare this smaller F to the overall study's F cutoff

One-way ANOVA in R

SID	TrainingDays	Performance
1	1	4.0
2	2	3.0
3	3	6.0
4	1	3.5
5	2	4.5
6	3	6.5
7	1	2.5

Data





One-way ANOVA in R

```
> one$TrainingDays <- factor(one$TrainingDays)
> res <- aov(one$Performance ~ one$TrainingDays)
> summary(res)
                 Df Sum Sq Mean Sq F value   Pr(>F)
one$TrainingDays  2 24.812  12.406  9.4417 0.001188 **
Residuals        21 27.594   1.314
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’
0.1 ‘ ’ 1
```

F(2,21)=9.44, p<.05

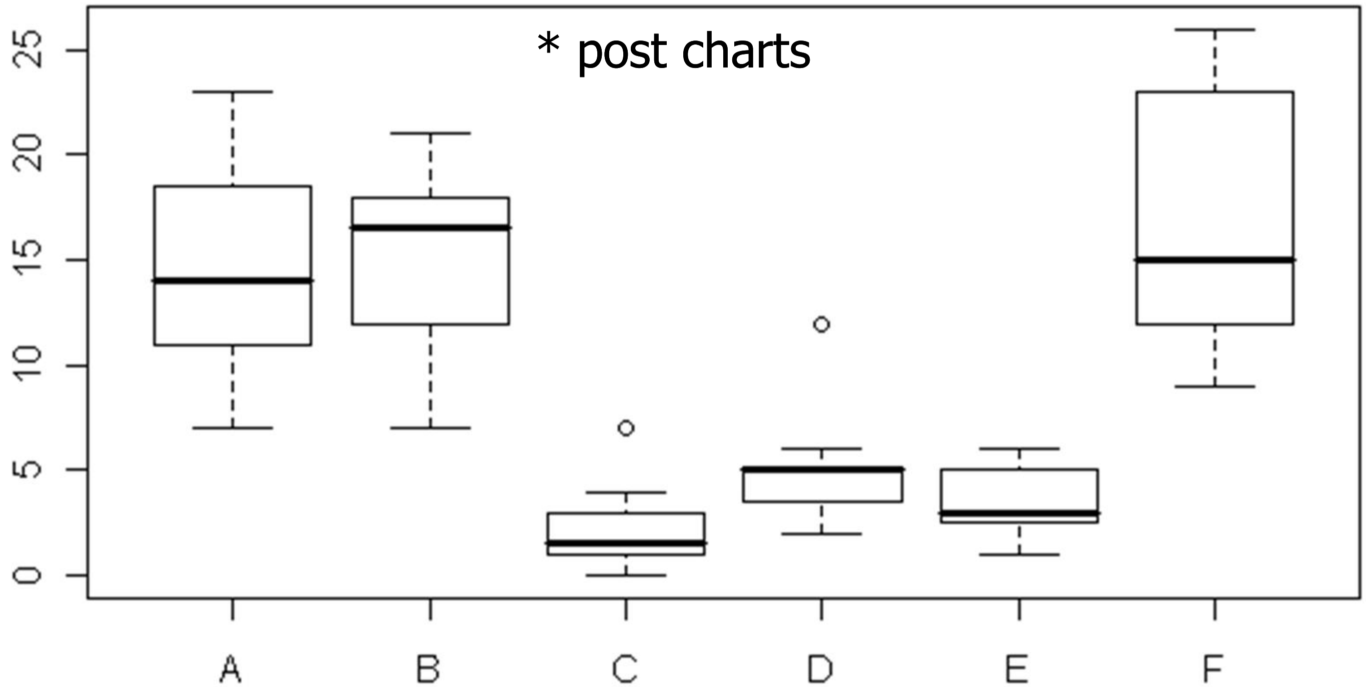
One-way ANOVA in R

```
#'d' is dataframe
#'d$Performance' is DV
#'d$TrainingDays' is factor (IV)
> oneway.test(d$Performance ~ d$TrainingDays,
                var.equal=TRUE)

      One-way analysis of means

data:  d$Performance and d$TrainingDays
F = 9.4417, num df = 2, denom df = 21, p-value =
0.001188
```

$F(2,21)=9.442, p<.05$

Visualizing results

- boxplot(DV ~ IV)



LSD aka unadjusted t-tests

```
> pairwise.t.test(DV, IVfactor,
          p.adjust="none", pool.sd = T)
Pairwise comparisons using t tests with pooled SD data:
DV and IVfactor
```

	Compact	Other	Pickup
Other	0.50197	-	-
Pickup	0.32786	0.72507	-
Sports	5.9e-05	0.00019	0.00064

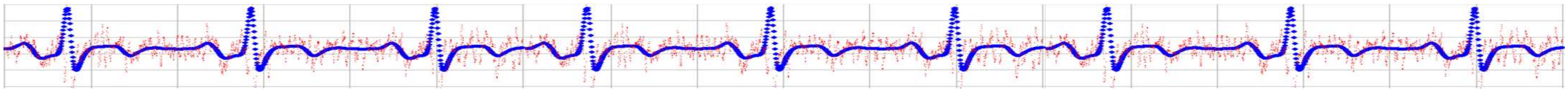
```
P value adjustment method: none
```

Note: p.adjust can also be "holm", "hochberg", "hommel", "bonferroni", "BH", "BY”

Post-hoc tests in R
Tukey HSD ("Honest Sig Diffs")

```
> res <- aov(one$Performance ~ one$TrainingDays)
> TukeyHSD(res)
  Tukey multiple comparisons of means
    95% family-wise confidence level

Fit: aov(formula = one$Performance ~ one$TrainingDays)

$`one$TrainingDays`
      diff        lwr      upr     p adj
2-1 0.0625 -1.3821563 1.507156 0.9934676
3-1 2.1875  0.7428437 3.632156 0.0027729
3-2 2.1250  0.6803437 3.569656 0.0035777
```

Publication format



The overall ANOVA was significant, $F(2,21)=9.44$, $p<.05$, indicating significant differences among the three study treatments.

Between df (numGroups – 1)

Within df (TotalN-numGroups)

Tukey HSD post-hoc tests (at .05 significance) indicated significant differences between 3-day training and the other conditions, but not between 1-day and 2-day training.

Another example

- "The means for the CRCR and NI groups were 8.0, 4.0, and 5.0, respectively. These were significantly different, $F(2,12) = 4.07$, $p<.05$. We also carried out two planned contrasts: The CR versus the NI condition, $F(1,12)=4.22$, $p<.10$;and the CrimR versus the CR condition, $F(1,12)=7.50$,$p<.05$. Although the first contrast approached significance, after a Bonferroni correction (for two planned contrasts), it does not even reach the .10 level."

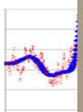


Table 9-11 Love Subscale Means for the Three Attachment Types (Newspaper Sample)

Scale Name	Avoidant	Anxious/ Ambivalent	Secure	*F*(2, 571)
Happiness	3.19_a	3.31_a	3.51_b	14.21***
Friendship	3.18_a	3.19_a	3.50_b	22.96***
Trust	3.11_a	3.13_a	3.43_b	16.21***
Fear of closeness	2.30_a	2.15_a	1.88_b	22.65***
Acceptance	2.86_a	3.03_b	3.01_b	4.66**
Emotional extremes	2.75_a	3.05_b	2.36_c	27.54***
Jealousy	2.57_a	2.88_b	2.17_c	43.91***
Obsessive preoccupation	3.01_a	3.29_b	3.01_a	9.47***
Sexual attraction	3.27_a	3.43_b	3.27_a	4.08*
Desire for union	2.81_a	3.25_b	2.69_a	22.67***
Desire for reciprocation	3.24_a	3.55_b	3.22_a	14.90***
Love at first sight	2.91_a	3.17_b	2.97_a	6.00**

Note: Within each row, means with different subscripts differ at the .05 level of significance according to a Scheffé test.*$p < .05$; **$p < .01$; ***$p < .001$.

Source: Hazan, C., & Shaver, P. (1987). Romantic love conceptualized as an attachment process. *Journal of Personality and Social Psychology, 52*, 511–524. Published by the American Psychological Association. Reprinted with permission.