Leibniz Universität Hannover

Mensch-Computer-Interaktion 2

Data Analysis

Human-Computer Interaction Group

Prof. Dr. Michael Rohs
michael.rohs@hci.uni-hannover.de

Lectures

Session	Date	Topic	
1	6.4.	Introduction	
2	13.4.	Interaction elements	GUI toolkits, interaction techniques
3	20.4.	Event handling	
4	27.4.	Scene graphs	
5	4.5.	Interaction techniques	
	11.5.	no class (CHI)	
	18.5.	no class (spring break)	
6	25.5.	Experiments	design and analysis of experiments
7	1.6.	Data Analysis	
8	8.6.	Data Analysis	
9	15.6.	Visualization	
10	22.6.	Visualization	current topics beyond-desktop UIs
11	29.6.	Modeling interaction	
12	6.7.	Computer vision for interaction	
13	13.7.	Computer vision for interaction	

Klausur:
28.7.2016
8-11 Uhr
HG E214

Review

- Properties: Nominal scale, ordinal scale, interval scale, ratio scale?
- Why does correlation not imply causality? Counterexample?
- Internal validity? External validity?
- Explain: factor, independent variable, dependent variable, level, condition, trial
- Control variable? Random variable? Confounding variable?
- What is a "good" task?
- Why written instructions?
- What is counterbalancing?
- How to construct a balanced Latin square for n = 4?

HYPOTHESIS TESTING

Null Hypothesis Significance Testing (NHST)

- NHST: The use of statistical procedures to answer research questions
- Typical research question (generic):

  Is the time to complete a task less
  using Method A than using Method B?

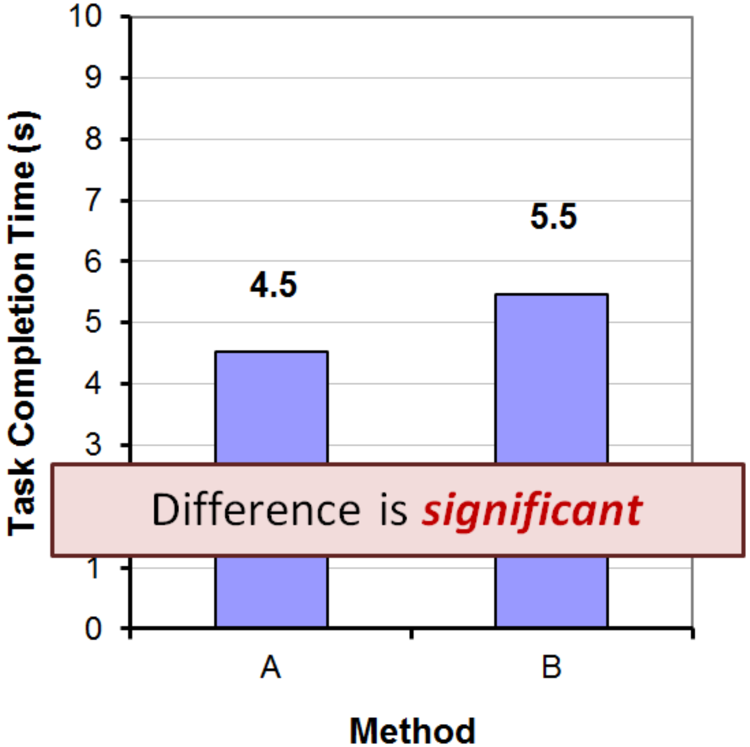
- For hypothesis testing, research questions are statements:

  There is no difference in the mean time to complete a task
  using Method A vs. Method B.

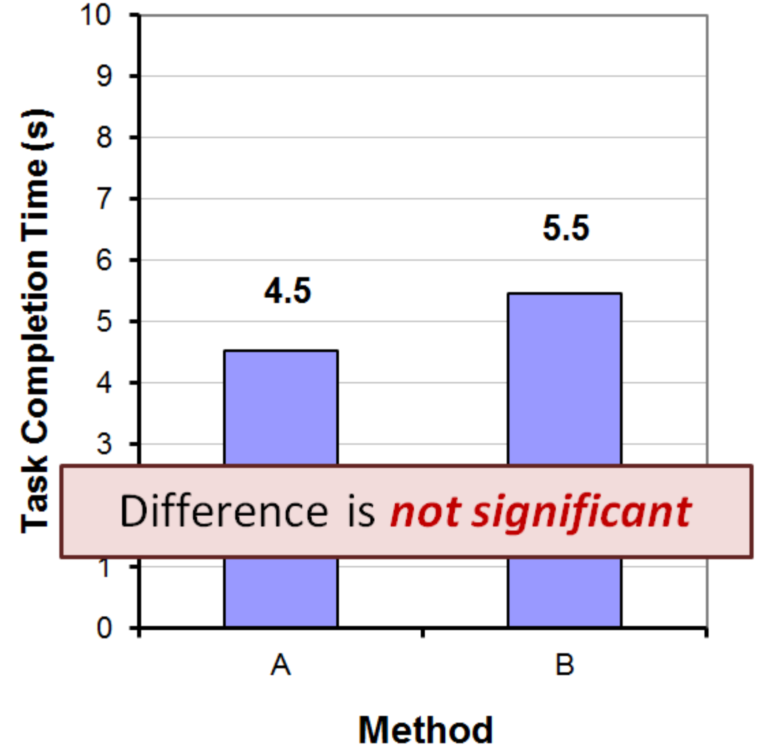
- This is the null hypothesis (assumption of “no difference”)
- Statistical procedures seek to reject or accept the null hypothesis

MacKenzie: Human-Computer Interaction - An Empirical Research Perspective.

Example #1



Example #2



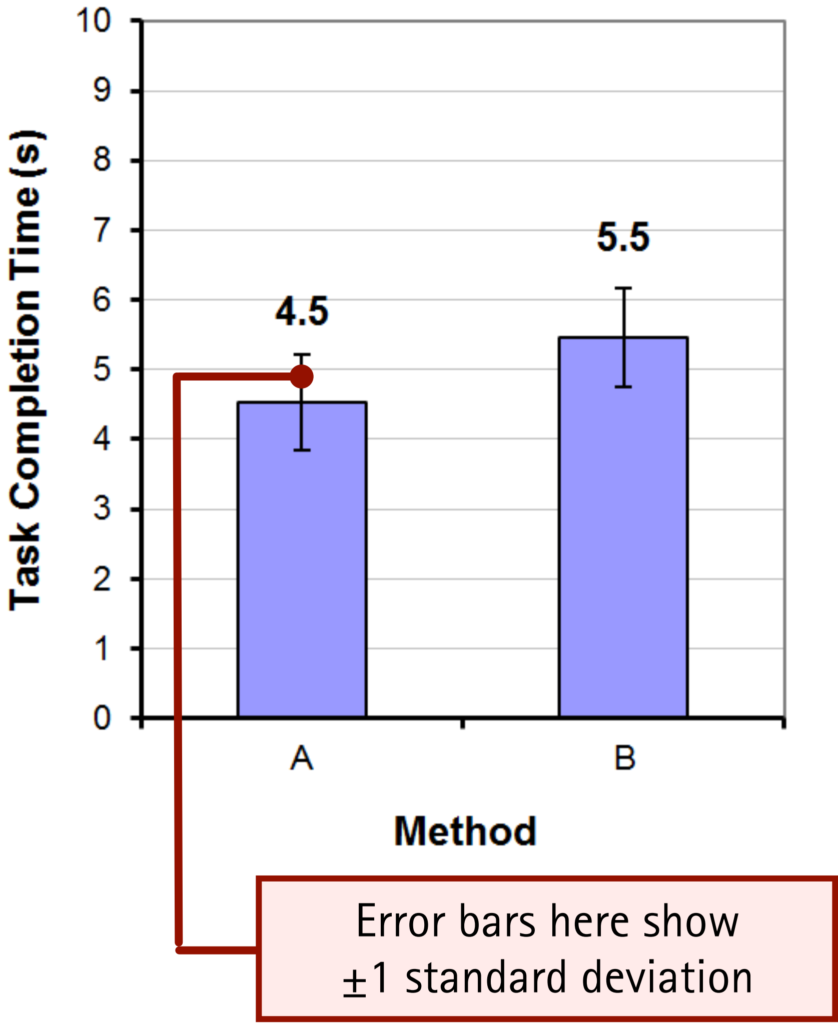
"Significant" implies that in all likelihood the difference observed is due to the test conditions (Method A vs. Method B).

"Not significant" implies that we cannot tell whether the difference observed is due to Method A vs. B or due to chance.

MacKenzie: Human-Computer Interaction - An Empirical Research Perspective.

Example #1 – Details

Note: Within-subjects design

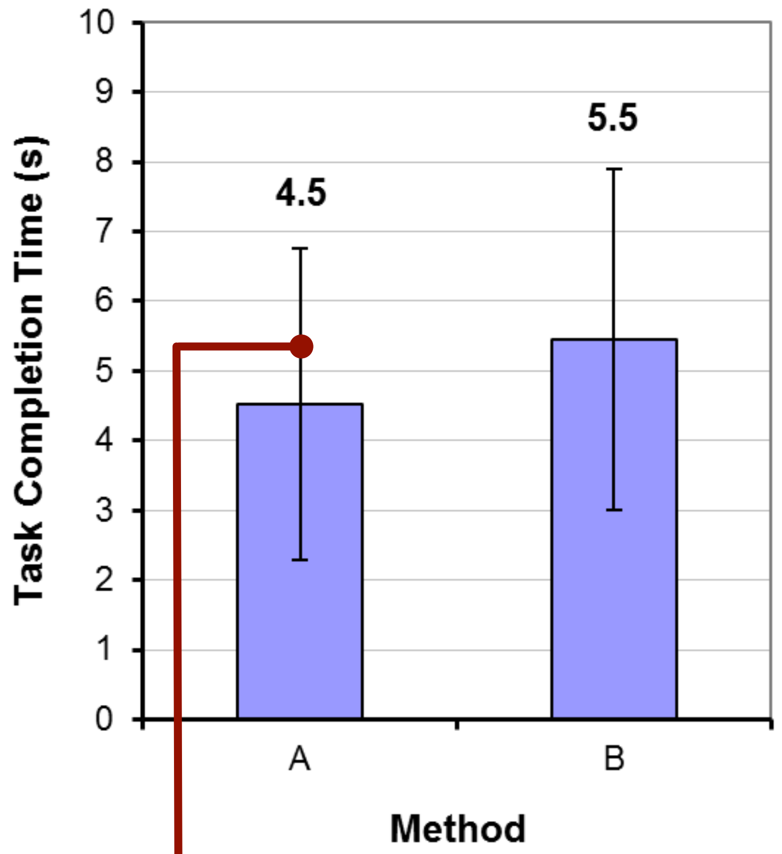


Participant	Method A	Method B
1	5.3	5.7
2	3.6	4.8
3	5.2	5.1
4	3.6	4.5
5	4.6	6.0
6	4.1	6.8
7	4.0	6.0
8	4.8	4.6
9	5.2	5.5
10	5.1	5.6
Mean	4.5	5.5
SD	0.68	0.72

MacKenzie: Human-Computer Interaction - An Empirical Research Perspective.

Example #2 – Details

Note: Within-subjects design

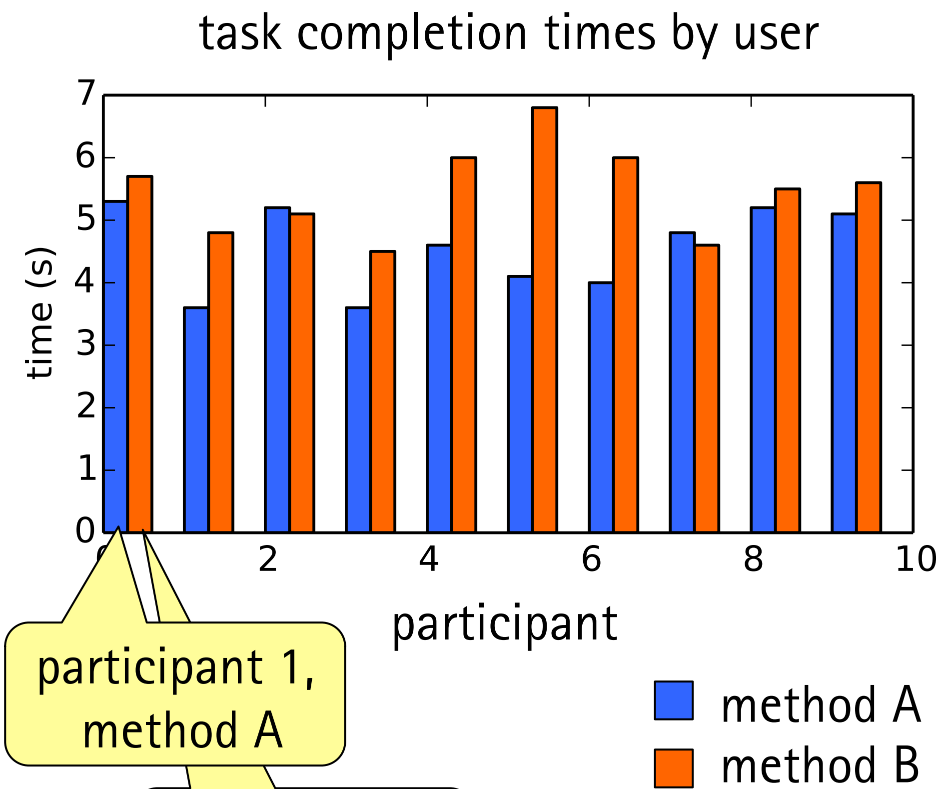


Error bars here show
±1 standard deviation

Participant	Method	
	A	B
1	2.4	6.9
2	2.7	7.2
3	3.4	2.6
4	6.1	1.8
5	6.4	7.8
6	5.4	9.2
7	7.9	4.4
8	1.2	6.6
9	3.0	4.8
10	6.6	3.1
Mean	4.5	5.5
SD	2.23	2.45

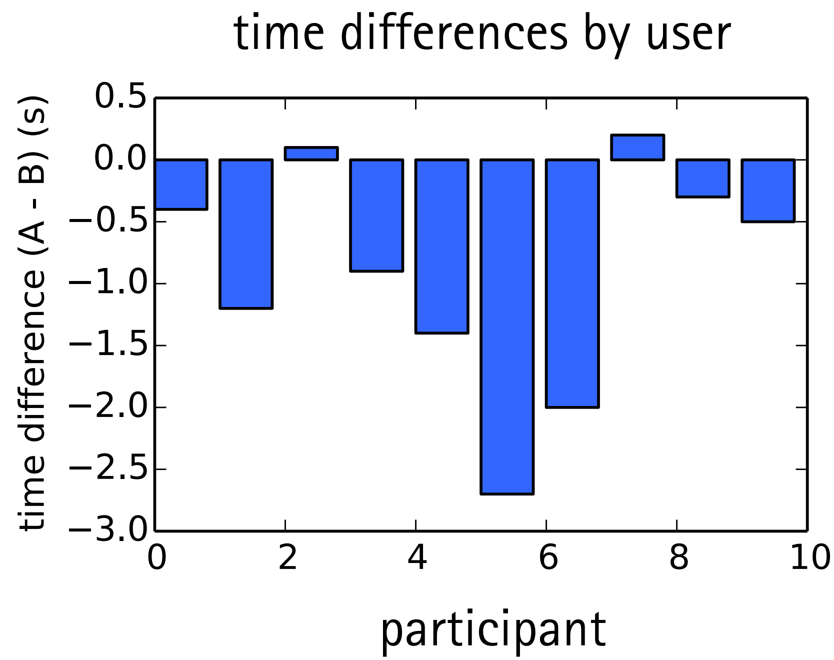
MacKenzie: Human-Computer Interaction – An Empirical Research Perspective.

Example #1: Comparison of Time Differences



participant 1, method A

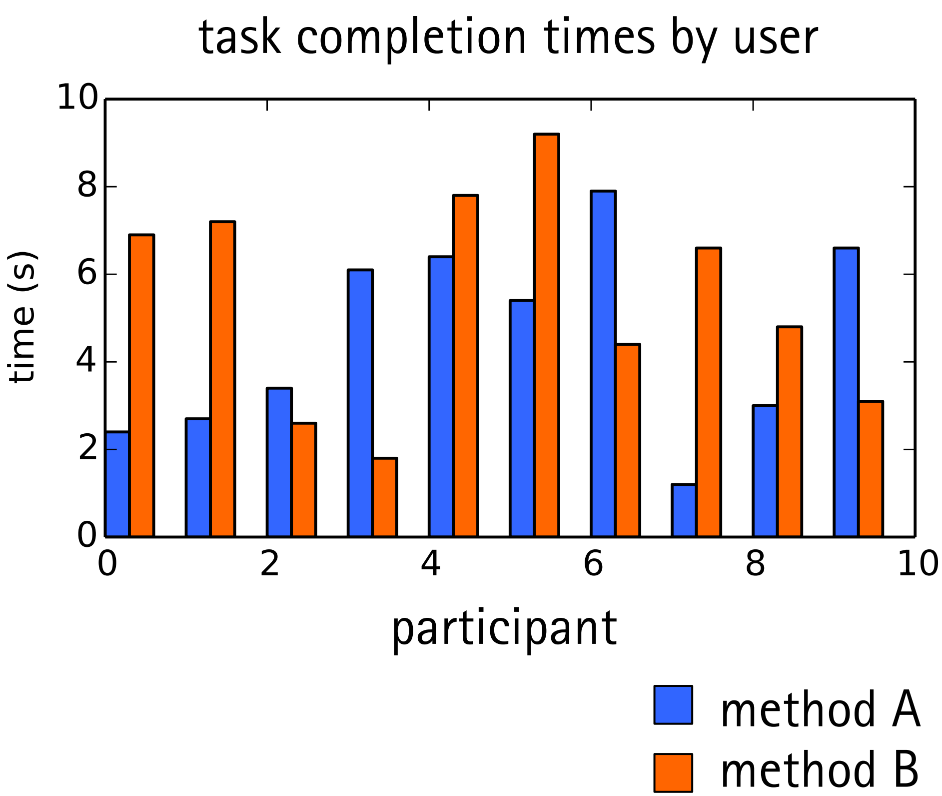
participant 1, method B

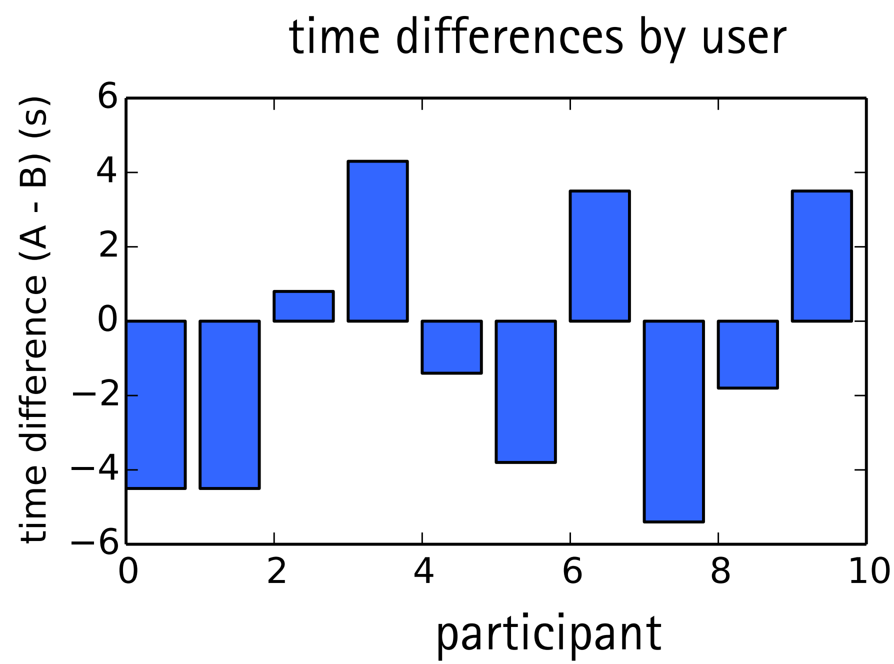


on average A - B = -0.9 s

8 of 10 participants are faster with method A than with method B

Example #2: Comparison of Time Differences



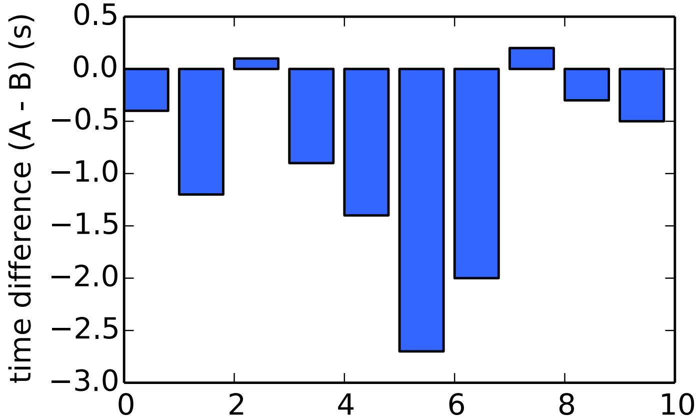


on average A - B = -0.9 s

6 of 10 participants are faster with method A than with method B

Confidence that "Method" has a Systematic Effect?

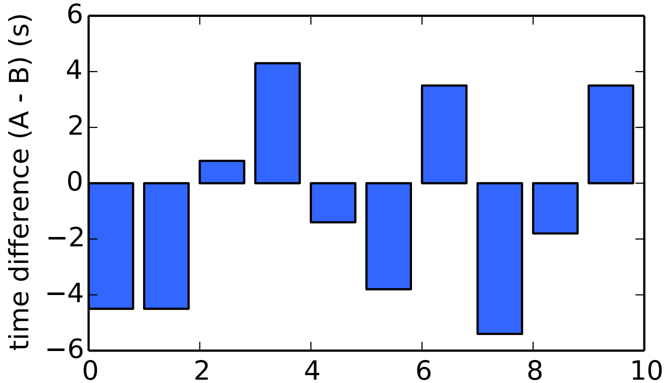
Example 1:



time differences by user

"looks like there is an effect of method (A/B)"

Example 2:

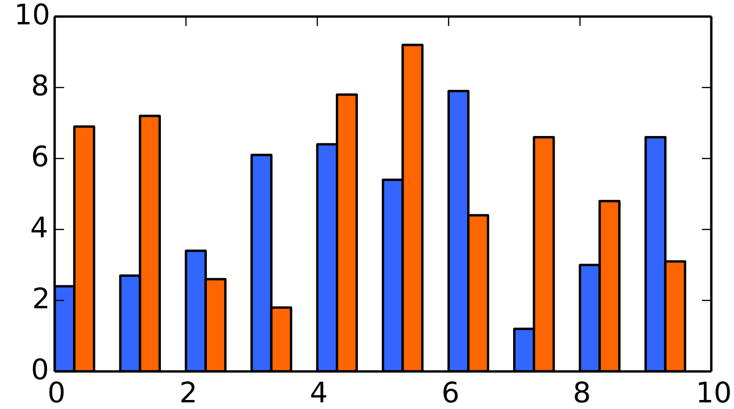


time differences by user

"no clear effect of method (A/B)"

Trick: How likely are the observed results under the assumption that "method" has no effect?

Null Hypothesis: Assume "Method" has No Effect



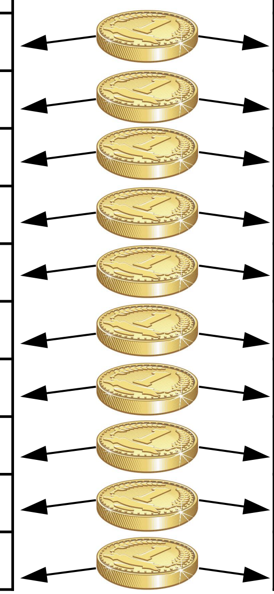
- Assumption: Factor "method" (with levels A and B) has no effect
  - If so, then methods A and B are in fact the same condition
  - If so, then whether A or B is shorter only depends on a participant's variability in execution speed
  - If so, then for each participant the shorter execution time could equally likely have occurred with A as with B (→ coin flip)
  - If so, then each direction (A < B, B < A) is equally likely; and for 10 users there are $2^{10} = 1024$ equally likely possibilities
  - If so, then it is very unlikely that for all users the shorter execution time occurs with method A ($p = 1/2^{10} = 1/1024$)
- (other factors controlled or randomly distributed, no order effects)

Probability of Obtained Result under Null Hypothesis

- Assumption: Method has no effect, rows are independent
- For participant 1, values A=5.3 and B= 5.7 as likely as A=5.7 and B=5.3
- Generate all possible $2^{10} = 1024$ orders (for 10 participants)
  - Each order equally likely
  - Unlikely that all small values with A
- Compute mean difference for each
- Compute fraction of means that is as extreme or more extreme as the obtained mean difference (-0.9)

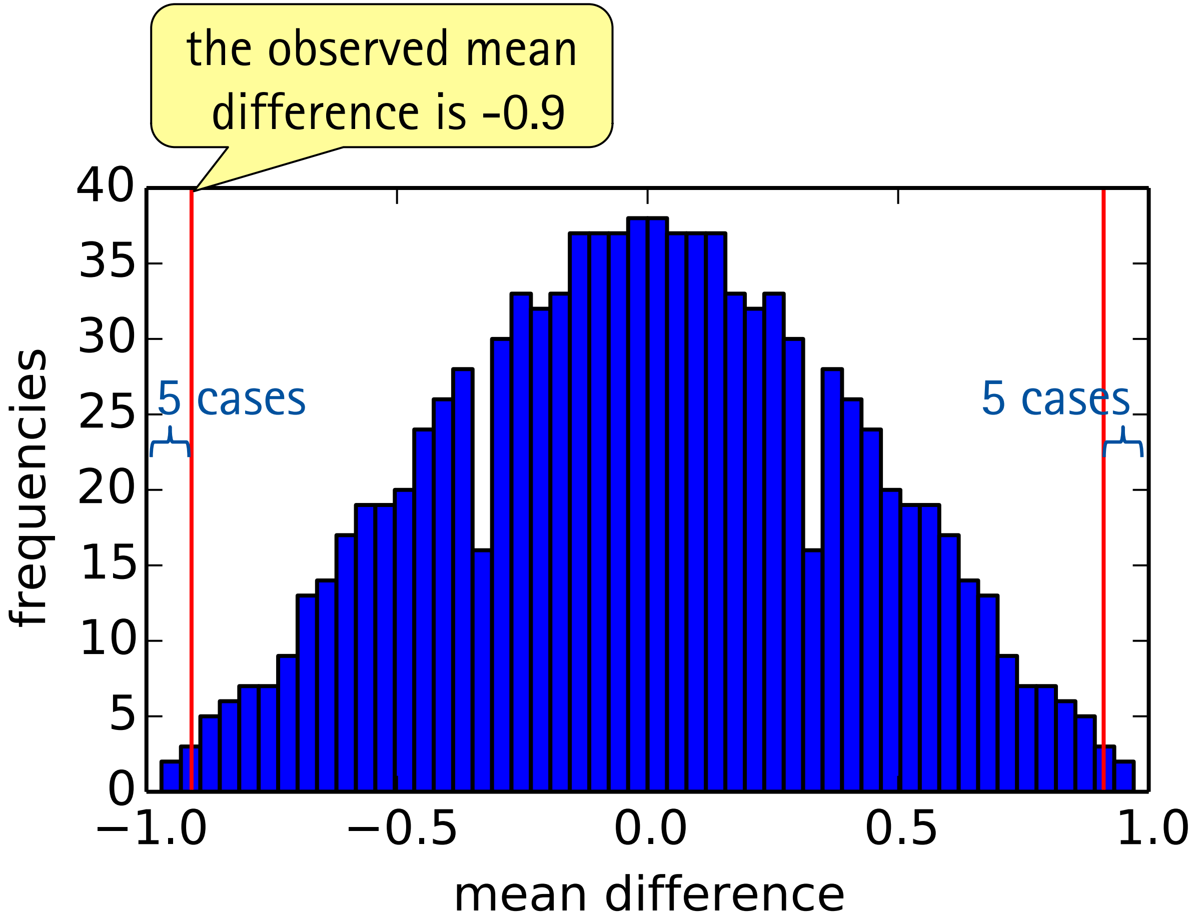
Example 1:

P	A	B
1	5.3	5.7
2	3.6	4.8
3	5.2	5.1
4	3.6	4.5
5	4.6	6.0
6	4.1	6.8
7	4.0	6.0
8	4.8	4.6
9	5.2	5.5
10	5.1	5.6



A	B
5.3	5.7
4.8	3.6
5.1	5.2
4.5	3.6
6.0	4.6
6.8	4.1
6.0	4.0
4.6	4.8
5.5	5.2
5.6	5.1

Probability of Obtained Result under Null Hypothesis

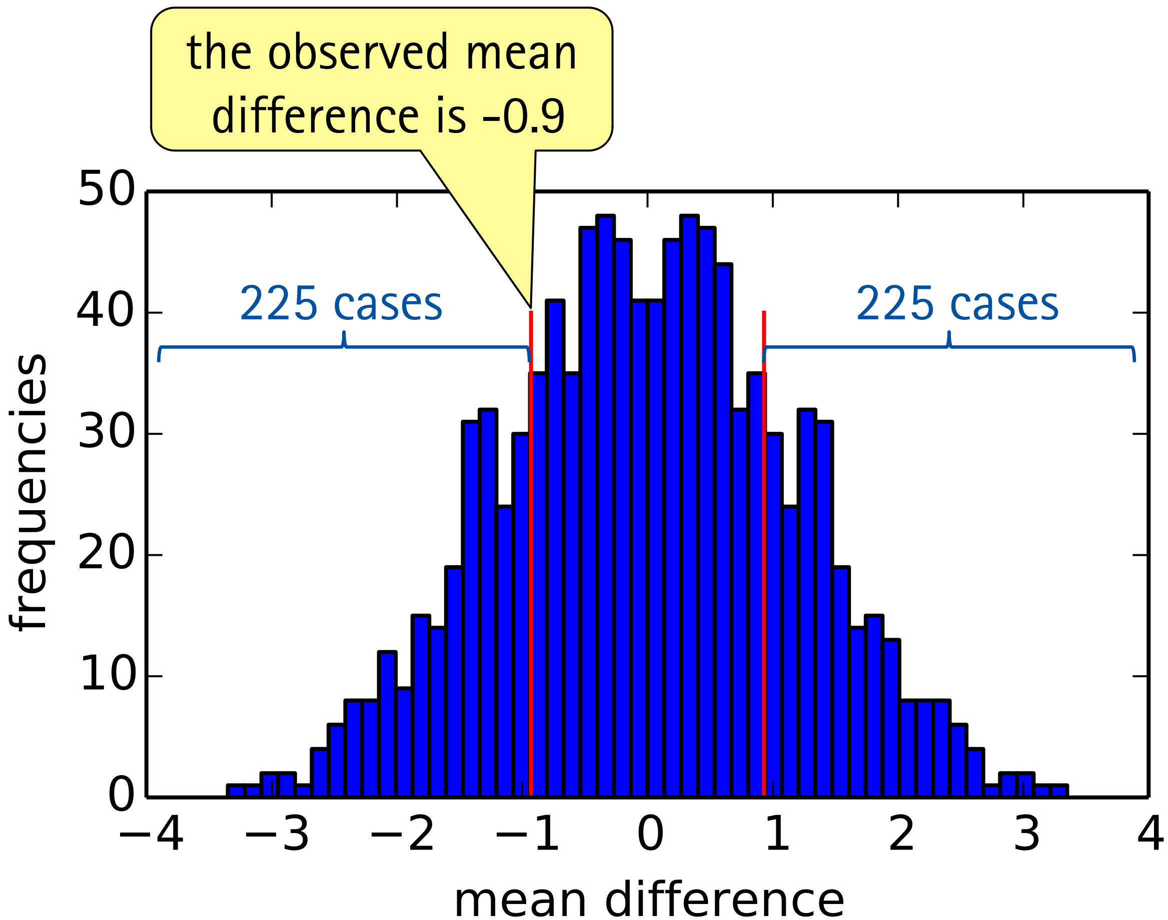


Example 1:

P	A	B	diff
1	5.3	5.7	-0.4
2	3.6	4.8	-1.2
3	5.2	5.1	0.1
4	3.6	4.5	-0.9
5	4.6	6.0	-1.4
6	4.1	6.8	-2.7
7	4.0	6.0	-2.0
8	4.8	4.6	0.2
9	5.2	5.5	-0.3
10	5.1	5.6	-0.5
		mean diff.	**-0.9**

$p = 10 / 2^{10} = 0.0098$ → unlikely that null hypothesis is true

Probability of Obtained Result under Null Hypothesis



Example 2:

P	A	B	diff
1	2.4	6.9	-4.5
2	2.7	7.2	-4.5
3	3.4	2.6	0.8
4	6.1	1.8	4.3
5	6.4	7.8	-1.4
6	5.4	9.2	-3.8
7	7.9	4.4	3.5
8	1.2	6.6	-5.4
9	3.0	4.8	-1.8
10	6.6	3.1	3.5
		mean diff.	**-0.9**

$p = 450 / 2^{10} = 0.4395 \rightarrow$ no evidence that null hypothesis is false

Randomization Tests

- If n is large, too much effort to enumerate all possibilities
- Alternative: Randomly pick large number (e.g., 10 000) of the possible assignments
  - Each possibility should have equal chance of being selected

Randomization Tests with Matched Samples

- Data discussed above are “matched samples”
  - Each participant generates two data points (one for each condition)
- Data from a single participant are not statistically independent
  - Slow participant is likely slow for both methods
  - Fast participant is likely fast for both methods
- Each participant generates a “matched pair“
  - (time method A, time for method B)
- Occurs with within-subject experiments
- For test, compute difference within each pair
  - If no effect, expect 0.0 difference on average
  - Differences are statistically independent

Randomization Test with Matched Samples

```
data = {(t11, t12), (t21, t22), ..., (tn1, tn2)} // matched pairs of n participants
dataMeanDiff = abs(mean(t11 - t12, t21 - t22, ..., tn1 - tn2))
r = 10000                  // repetitions
M = zeros(r)               // mean differences of each trial
for j = 1..r:              // r repetitions
    D = zeros(n)           // difference vector for jth trial
    for i = 1..n:          // n participants
        (t1, t2) = datai   // data pair of participant i
        Di = if (coin flip is head) t1 - t2 else t2 - t1
    Mj = mean(D)
p = (count(Mi ≤ -dataMeanDiff) + count(Mi ≥ dataMeanDiff)) / r // extreme values
if p ≤ 0.05 then "significant" else "not significant"
```

Randomization Tests with Independent Samples

- Occurs with between-subject experiments
  - Each participant randomly assigned to one group
- Each participant provides a single data point
  - Either time for method A or time for method B
- Data points are assumed to be statistically independent
- Null hypothesis: Method has no effect, i.e.,
  the two groups represent the same condition
- Generate many random reassignments of users to groups
- Compute test statistic for each random reassignment
  - Example: Absolute difference between group means
- Compute fraction of reassignments that generate an effect
  as strong or stronger as the observed one

Example: Between-Subjects Designs

- Research question:

  Do left-handed users and right-handed users differ in the time to complete an interaction task?

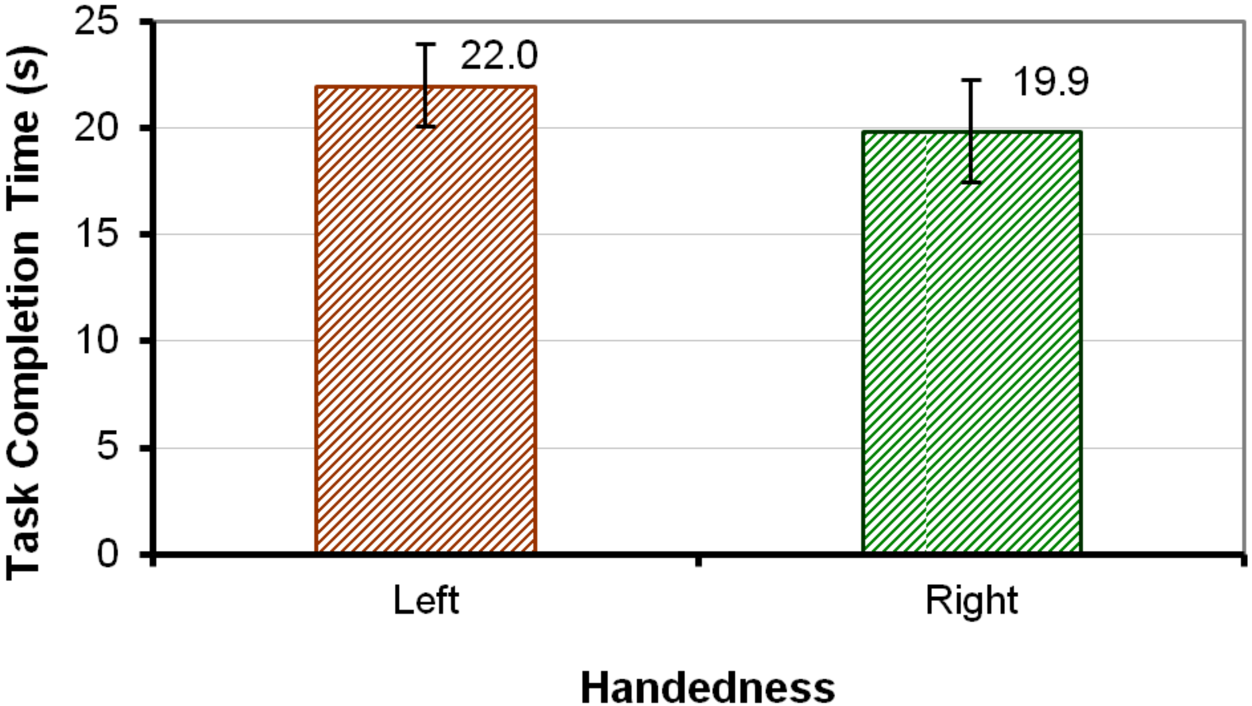
- The independent variable (handedness) must be assigned between-subjects
- There is one data point per participant
- The data points are independent

Participant	Task Completion Time (s)	Handedness
1	23	L
2	19	L
3	22	L
4	21	L
5	23	L
6	20	L
7	25	L
8	23	L
9	17	R
10	19	R
11	16	R
12	21	R
13	23	R
14	20	R
15	22	R
16	21	R
Mean	20.9	
SD	2.38	

MacKenzie: Human-Computer Interaction - An Empirical Research Perspective.

Summary Data and Chart

Handedness	Task Completion Time (s)	
	Mean	*SD*
Left	22.0	1.93
Right	19.9	2.42

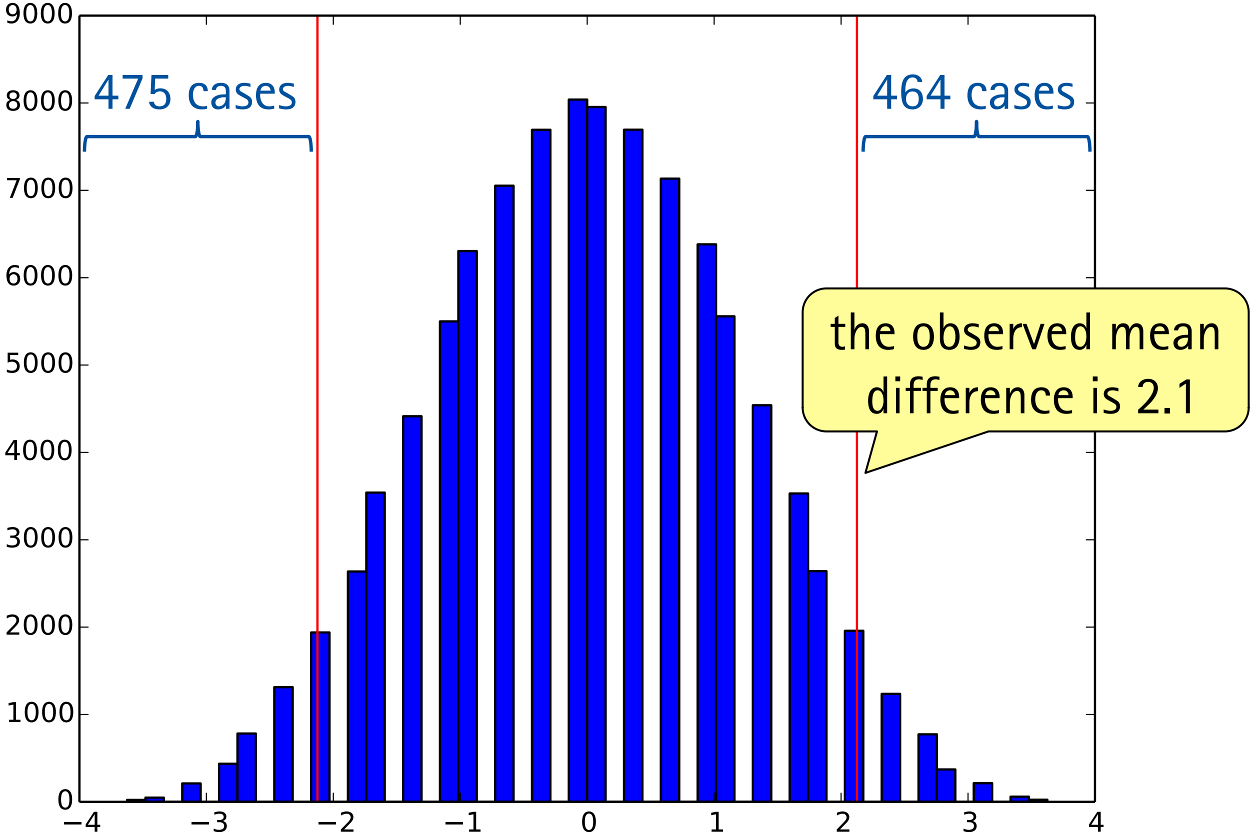


MacKenzie: Human-Computer Interaction - An Empirical Research Perspective.

Randomization Test with Independent Samples

```
A = {t11, t12, ..., t1n}               // goup 1, participants 1..n
B = {t2,(n+1), t2,(n+2), ..., t2,2n} // goup 2, participants n+1..2n
meanDiff = abs(mean(A) - mean(B)) // observed difference between groups
C = stack(A, B)    // combine A and B, length(C) = 2n
r = 10000          // repetitions
M = zeros(r)       // mean differences of each trial
for j = 1..r:      // r repetitions
    permuted = shuffle(C)          // randomly shuffle all data
    group1 = permuted[1..n]        // randomized group 1
    group2 = permuted[n+1..2n]     // randomized group 2
    Mj = mean(group1) - mean(group2) // difference between randomized groups
p = (count(Mi ≤ -meanDiff) + count(Mi ≥ meanDiff)) / r // extreme values
if p ≤ 0.05 then "significant" else "not significant"
```

Between-Subjects Designs



Participant	Task Completion Time (s)	Handedness
1	23	L
2	19	L
3	22	L
4	21	L
5	23	L
6	20	L
7	25	L
8	23	L
9	17	R
10	19	R
11	16	R
12	21	R
13	23	R
14	20	R
15	22	R
16	21	R
Mean	20.9	
SD	2.38	

p = 939 / 10000 = 0.0939 → no evidence that null hypothesis is false

Parametric and Non-Parametric Statistical Tests

- Non-parametric tests
  - Randomization tests are non-parametric
  - Data are not assumed to come from a particular statistical distribution
- Parametric tests
  - Data are assumed to come from a particular statistical distribution (for example the normal distribution)
- Selecting a test: Match the type of test with the experimental design and measurement scale of the data

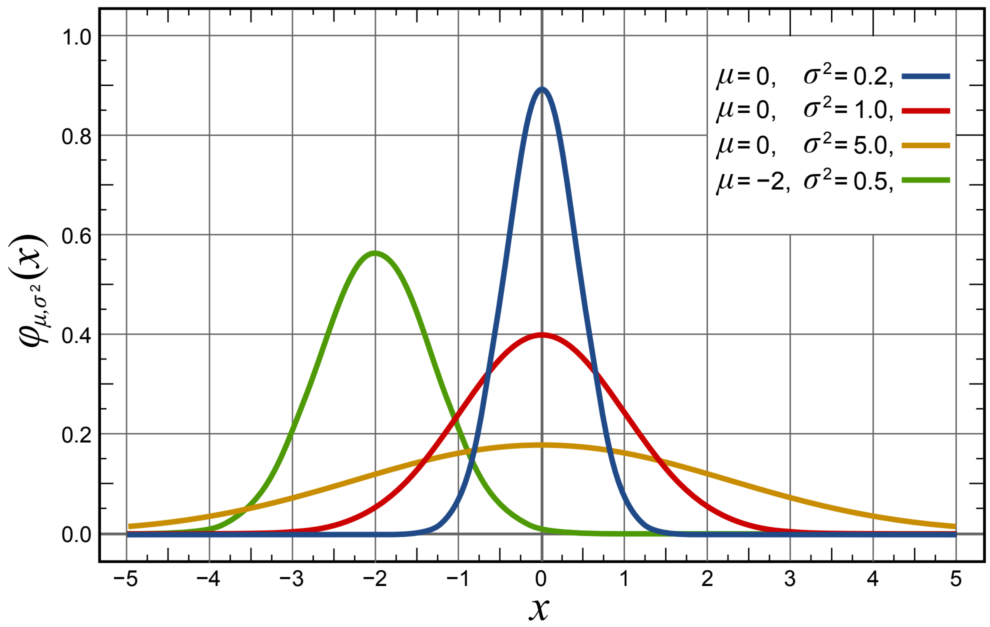
MacKenzie: Human-Computer Interaction - An Empirical Research Perspective.

Reminder: Normal Distribution

Some tests assume that DV is (roughly) normally distributed (under the null hypothesis)

$$DV = N\left(\mu, \sigma^2\right)$$

- μ = population mean
- σ = population standard deviation



Normal distribution:

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left[-\frac{(x-\mu)^2}{2\sigma^2}\right]$$

Measurement Scales and Statistics

Scale	Relations	Statistics	Tests
Nominal	• Equivalence	• Mode • Frequency	Non-parametric tests
Ordinal	• Equivalence • Order	• Median • Percentile	
Interval	• Equivalence • Order • Ratio of intervals	• Mean • Standard deviation	Parametric tests and non-parametric tests
Ratio	• Equivalence • Order • Ratio of intervals • Ratio of values	• Geometric mean • Coefficient of variation	

MacKenzie: Human-Computer Interaction - An Empirical Research Perspective.

Measurement Scales and Statistical Tests

Kind of Test	Statistical Tests
Non-parametric	• Chi-square
	• Mann-Whitney U (2 groups, between) • Wilcoxon Signed-Rank (2 groups, within) • Kruskal-Wallis (3+ groups, between) • Friedman (3+ groups, within)
Parametric	• t-test (2 groups, independent samples, matched samples) • ANOVA (2+ groups, one-way between, one-way within, two-way within, mixed factors, etc.)

MacKenzie: Human-Computer Interaction - An Empirical Research Perspective.

Tests Discussed

- Parametric
  - t-test
    - Comparison of two groups
  - Analysis of variance (ANOVA)
    - Most common statistical procedure in HCI research
    - Used for ratio data and interval data
- Non-parametric
  - Randomization tests
  - Chi-square test
    - Used for nominal data
  - Mann-Whitney U, Wilcoxon Signed-Rank, Kruskal-Wallis, and Friedman tests
    - Used for ordinal data

MacKenzie: Human-Computer Interaction - An Empirical Research Perspective.

ANALYSIS OF VARIANCE (ANOVA)

Analysis of Variance (ANOVA)

- Most widely used statistical test for hypothesis testing in factorial experiments
- Goal: Determine if an independent variable has a significant effect on a dependent variable
  - Remember, an independent variable has at least two levels (values, settings, test conditions)
- Goal (put another way): Determine if the test conditions yield different outcomes on the dependent variable
  - E.g., one of the test conditions is faster/slower than the other

MacKenzie: Human-Computer Interaction - An Empirical Research Perspective.

Why Analyze the Variance?

- We are interested in differences between means
  - Is the time to complete a task less using Method A than using Method B?
- Seems odd that we analyse the variance...

MacKenzie: Human-Computer Interaction - An Empirical Research Perspective.

Standard Deviation, Variance

- Standard deviation is a measure of variability about the mean
- Variance is the squared standard deviation
- Standard deviation of the entire population (of n data points)

$$\sigma = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(data_i - \overline{data}\right)^2}$$

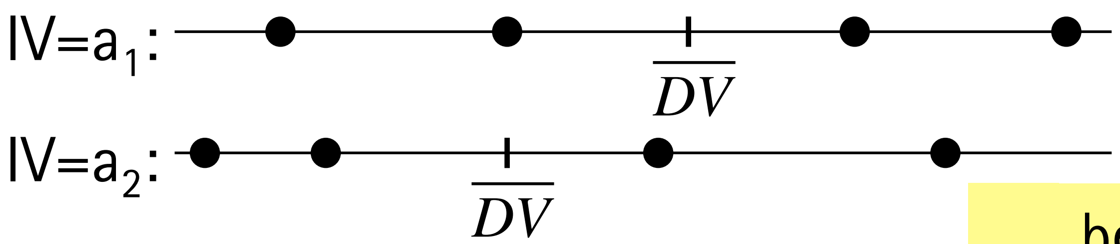
- Standard deviation of a sample (of a larger population)

$$s = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}\left(data_i - \overline{data}\right)^2}$$

  - Dividing by n would introduce a bias

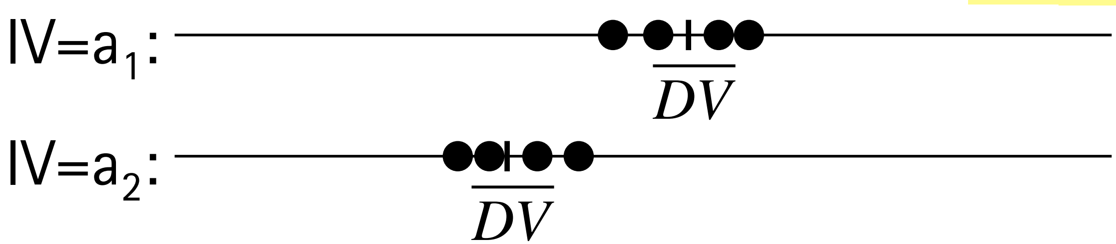
Basic Idea of Analysis of Variance

- Given: Independent variable IV with levels a_1 and a_2
- Continuous dependent variable DV is measured
- Idea: Look at variability within groups and across group means
  - More confident if small variability within group and large variability across means



situation 1: small confidence in a systematic effect

$$F = \frac{\text{between-group variability}}{\text{within-group variability}}$$

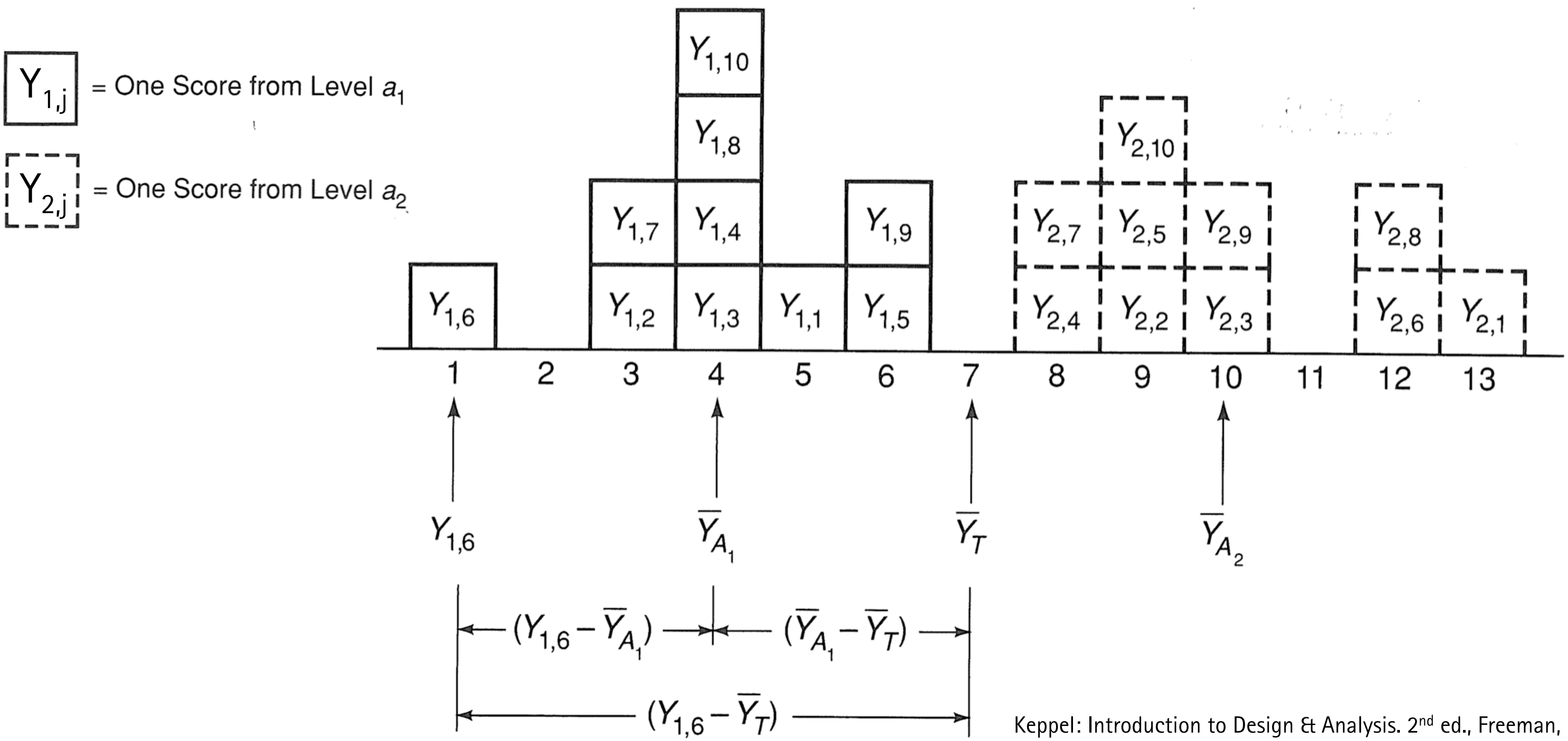


situation 2: larger confidence in a systematic effect

Partitioning Sums of Squares

total deviation = within-group deviation + between-group deviation

$$Y_{i,j} - \overline{Y}_T = (Y_{i,j} - \overline{Y}_{Ai}) + (\overline{Y}_{Ai} - \overline{Y}_T)$$



Keppel: Introduction to Design & Analysis. 2nd ed., Freeman, 1992.

Partitioning Sums of Squares

total deviation = within-group deviation + between-group deviation

$$Y_{i,j} - \overline{Y}_T \quad = \quad (Y_{i,j} - \overline{Y}_{Ai}) \quad + \quad (\overline{Y}_{Ai} - \overline{Y}_T)$$

this is also true for the sum of the squared deviations (no proof here)

$$\sum_{i=1}^{k}\sum_{j=1}^{n}\left(Y_{i,j} - \overline{Y}_T\right)^2 = \sum_{i=1}^{k}\sum_{j=1}^{n}\left(Y_{i,j} - \overline{Y}_{A_i}\right)^2 + \sum_{i=1}^{k}\left(\overline{Y}_{A_i} - \overline{Y}_T\right)^2$$

$$SS_{total} = SS_{within\text{-}groups} + SS_{between\text{-}groups}$$

$$SS_{total} = SS_{error} + SS_{condition}$$

The sum of squared total deviations equals the sum of the squared within-group deviations (due to error sources) and the sum of the squared between-group deviations (due to conditions, if there is an effect)

Keppel: Introduction to Design & Analysis. 2nd ed., Freeman, 1992.

Degrees of Freedom (df)

- df = the number of independent sources of variation
- Sums of squares are divided by df to obtain a variance
  - Average sum of squares (mean squares)
  - MS = SS / df
- $df_{total} = n * k - 1$

$$SS_{total} = \sum_{i=1}^{k} \sum_{j=1}^{n} \left(Y_{i,j} - \bar{Y}_T \right)^2$$

k = number of conditions

n = number of observations in each condition

- $df_{between-groups} = k - 1$

$$SS_{between-groups} = \sum_{i=1}^{k} \left(\bar{Y}_{A_i} - \bar{Y}_T \right)^2$$

- $df_{within-groups} = n * k - k$

$$SS_{within-groups} = \sum_{i=1}^{k} \sum_{j=1}^{n} \left(Y_{i,j} - \bar{Y}_{A_i} \right)^2$$

- $df_{total} = df_{within-groups} + df_{between-groups}$

F-Ratio (F-Statistic)

- Variability due to experimental condition
  - $MS_{condition} = MS_{between\text{-}groups} = SS_{between\text{-}groups} / df_{between\text{-}groups}$
- Variability due to error sources
  - $MS_{error} = MS_{within\text{-}groups} = SS_{within\text{-}groups} / df_{within\text{-}groups}$

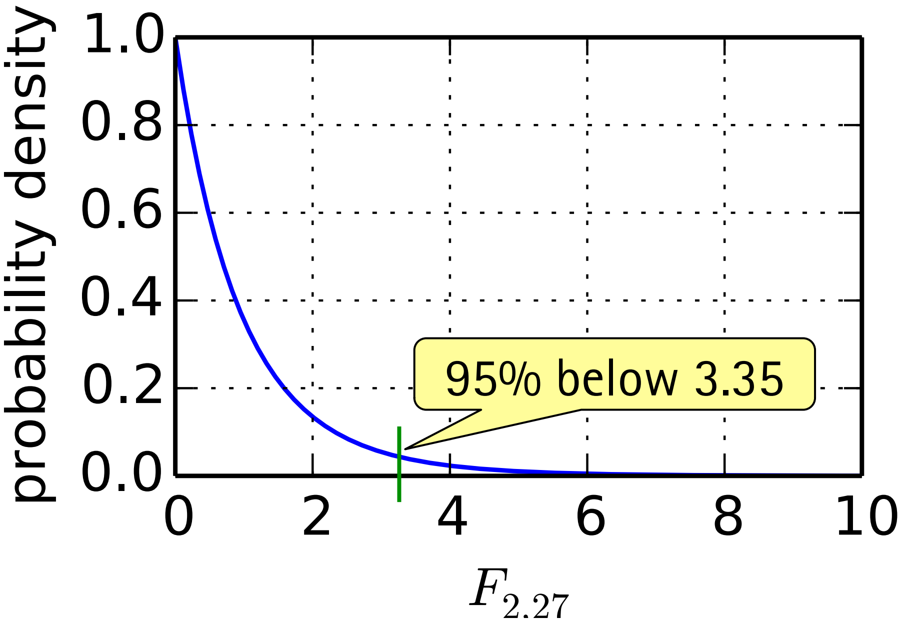
- F-statistic
  - $F = MS_{condition} / MS_{error}$

$$F = \frac{\text{between-group variability}}{\text{within-group variability}}$$

F-Distribution

$$F = \frac{\text{between-group variability}}{\text{within-group variability}}$$

- F-ratio follows F-distribution with k-1, nk-k degrees of freedom, if null hypothesis is true and data is normally distributed
  - F(k-1, nk-k)
  - k groups, n values per group
- Probability density graph:
  - k = 3 groups
  - n = 10 values per group
  - df_{bg} = k - 1 = 2
  - df_{wg} = nk - k = 27



Between-Subjects ANOVA

ss = sum of squares
df = degrees of freedom
MS = mean square
F = Fisher-ratio
p = probability that F is larger than the given value

data	condition 1	condition 2	condition 3		
participants 1, 2, 3	9	7	5		
participants 4, 5, 6	3	3	0		
participants 7, 8, 9	3	2	4		
mean for condition	5	4	3	4	grand mean
squared differences from grand mean	25	9	1		
	1	1	16		
	1	4	0		
total sum of squares (ssTotal)	58				ssTotal = ssCondition + ssError
squared differences from condition means	16	9	4		
	4	1	9		
	4	4	1		
sum of squares from condition means	52				ssError

	ss	df	MS	F	p
ssCondition	6	2	3.000	0.346	0.721
ssError	52	6	8.667		
ssTotal	58	8			

Within-Subjects ANOVA

data	condition 1	condition 2	condition 3	mean for participant	
participant 1	9	7	5	7	
participant 2	3	3	0	2	
participant 3	3	2	4	3	
mean for condition	5	4	3	4	grand mean

ss = sum of squares
df = degrees of freedom
MS = mean square
F = Fisher-ratio
p = probability that F is larger than the given value

squared differences from grand mean	25	9	1
	1	1	16
	1	4	0
total sum of squares (ssTotal)	58		
squared differences from condition means	16	9	4
	4	1	9
	4	4	1
sum of squares from condition means	52		
squared differences from participant means	4	0	4
	1	1	4
	0	1	1
sum of squares from participant means	16		

ssTotal = ssCondition + ssParticipant + ssError

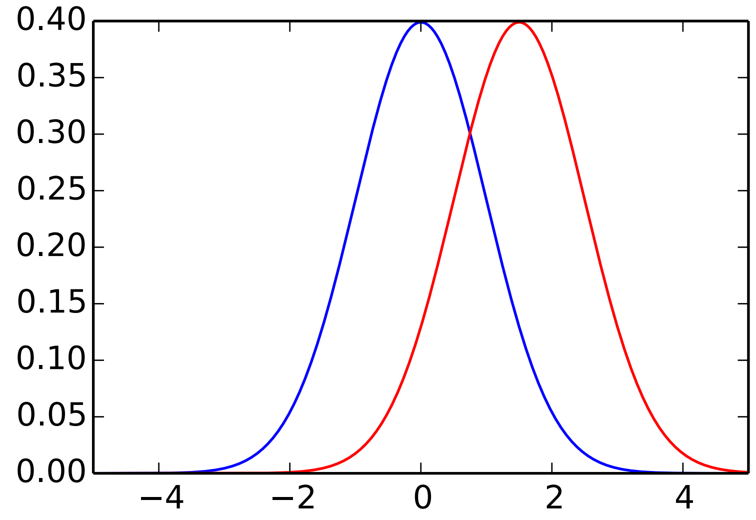
sum of squares from condition means: ssParticipant + ssError

sum of squares from participant means: ssCondition + ssError

	ss	df	MS	F	p
ssCondition = ssT – (ssP + ssE)	6	2	3.000	1.200	0.391
ssParticipant = ssT – (ssC + ssE)	42	2	21.000		
ssError = ssT – ssC – ssP	10	4	2.500		
ssTotal	58	8			

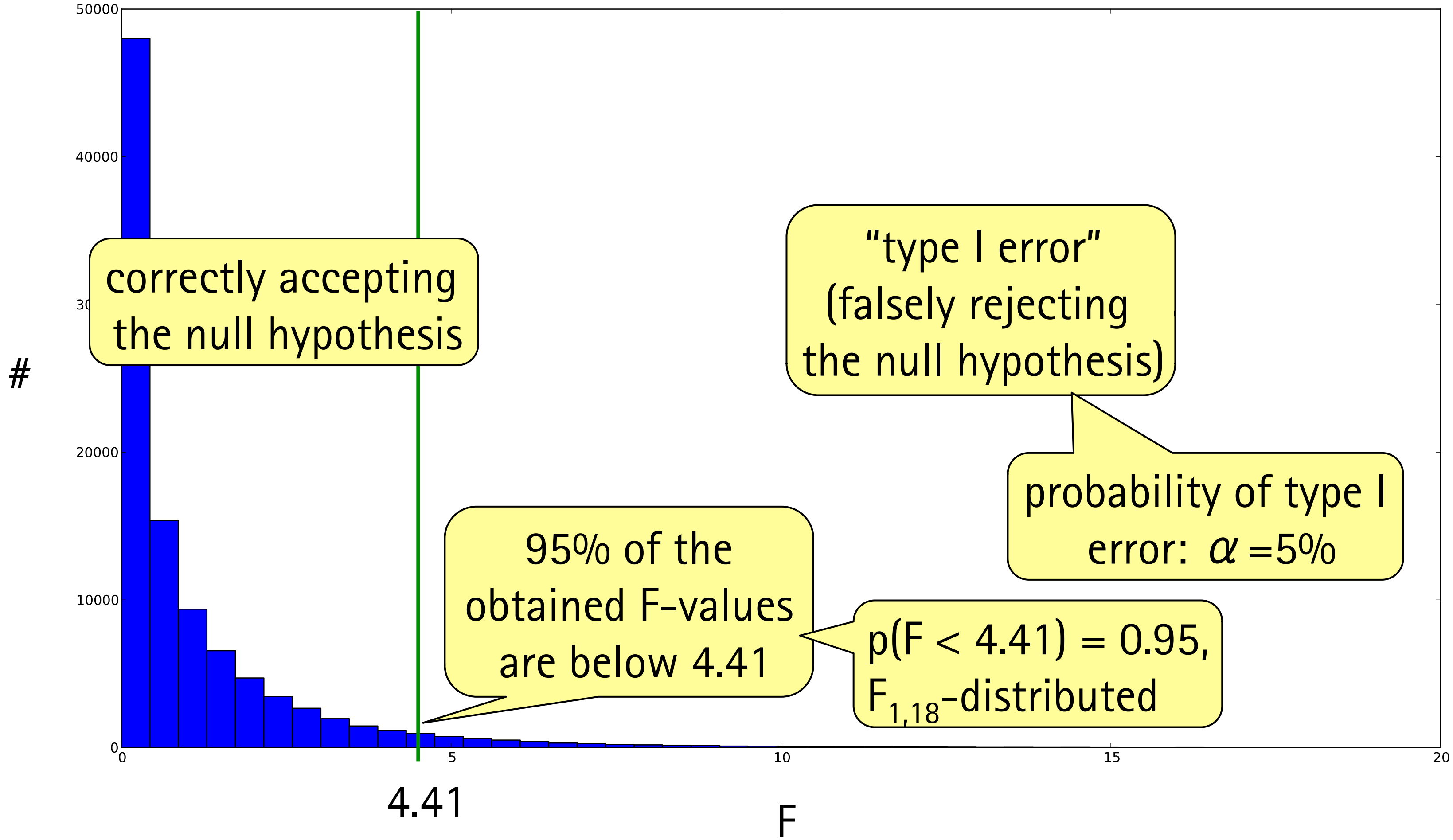
Experiment Simulations, F-Histograms

- Simulate 100k experiments
- Generate data for 2 groups with 10 data points each
- Data points are independent
- Data points are from normal distributions with σ=1.0
- Group 1: μ_1=0.0 (fixed mean)
- Group 2: μ_2=0.0, 0.5, 1.0, 1.5

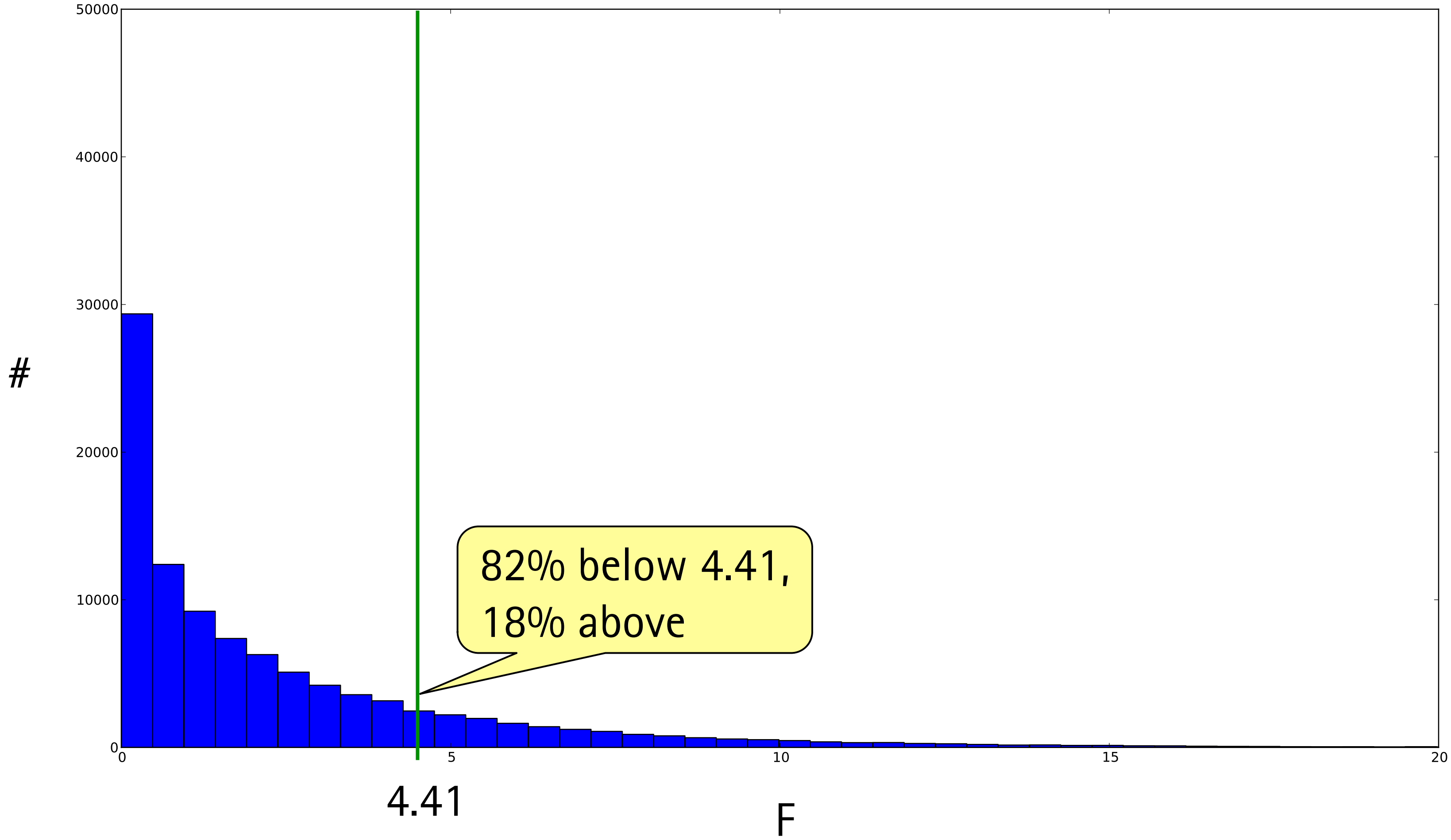


- Produce histograms of F-statistic
- Normalize histograms, produce cumulative histograms

F-Histogram
(k = 2, n = 10, σ=1, μ_1=0, μ_2=0)

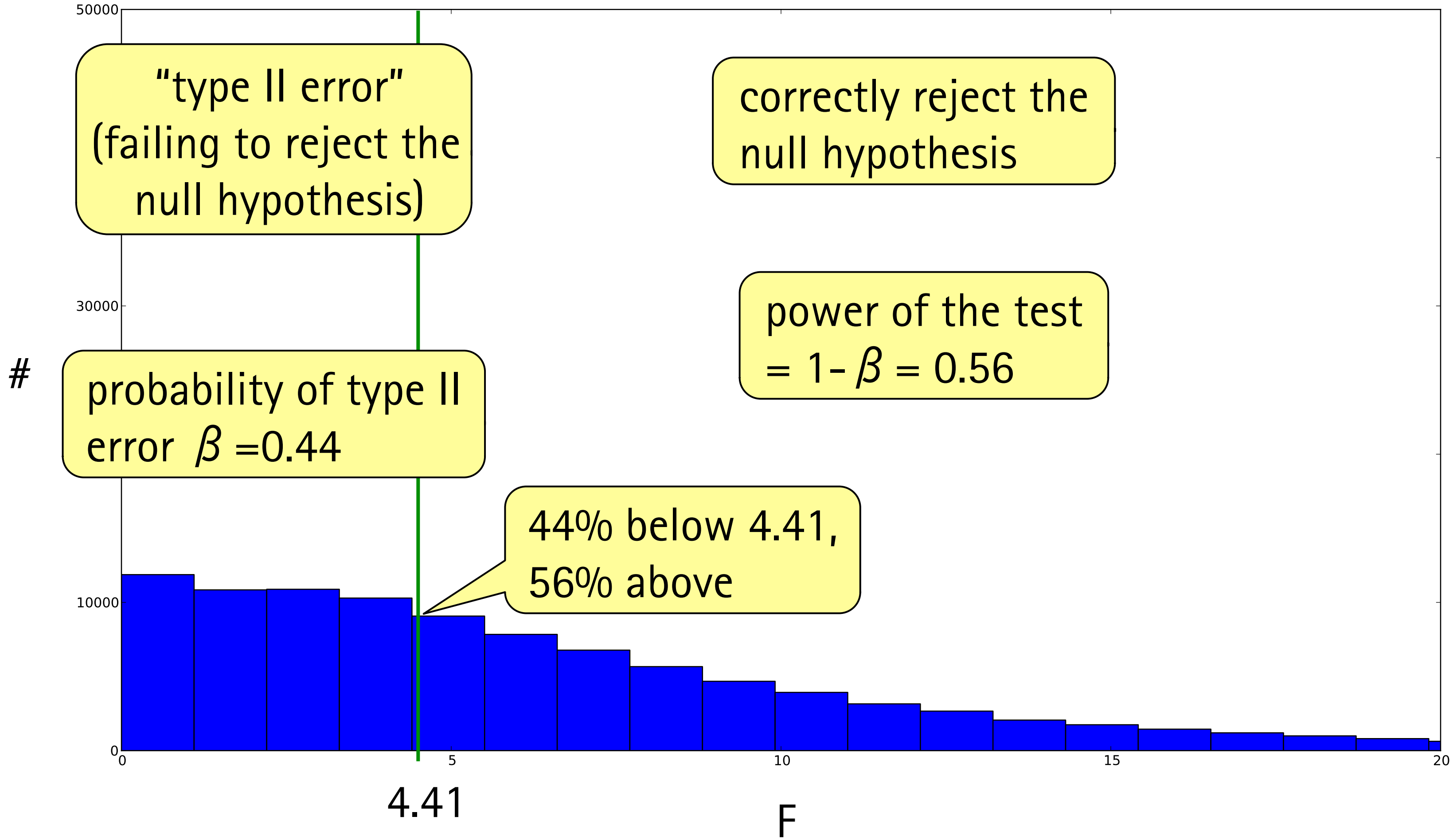


F-Histogram
(k = 2, n = 10, σ=1, μ_1=0, μ_2=0.5)

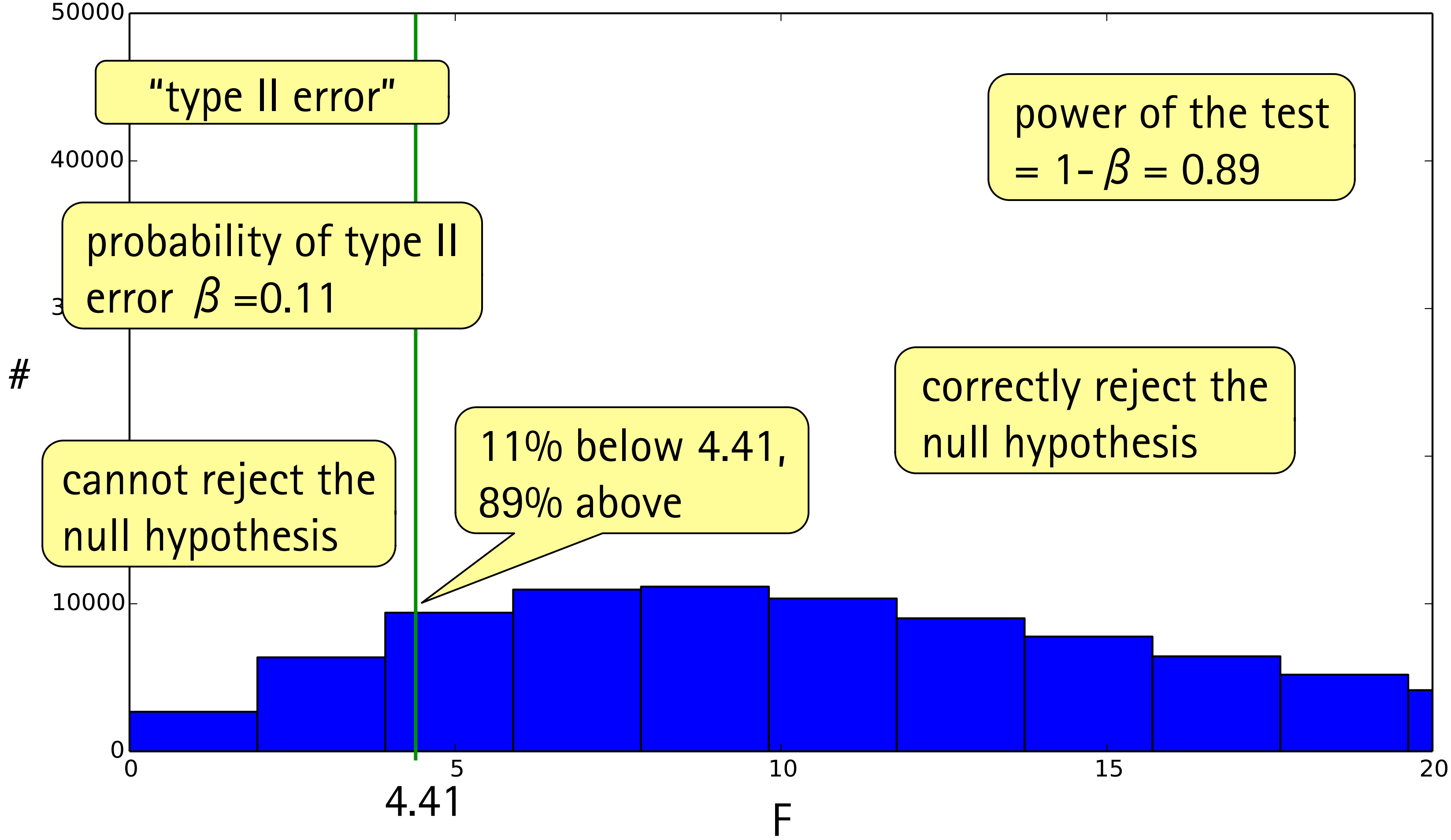


F-Histogram

(k = 2, n = 10, σ=1, μ_1=0, μ_2=1)



F-Histogram
(k = 2, n = 10, σ=1, μ_1=0, μ_2=1.5)



Numpy Script to Simulate Experiments (1/3)

```
import numpy as np # import numpy
import matplotlib.pyplot as pp # import pyplot for graphical output

sigma = 1.0 # both distributions have the same standard deviation 1.0
mu1 = 0.0 # mean of group 1
mu2 = 1.0 # mean of group 2
groups = 2 # 2 groups
groupSize = 10 # 10 data points in each group
experiments = 100000 # a huge number of experiments
results = np.zeros(experiments) # this array will hold the results
```

Numpy Script to Simulate Experiments (2/3)

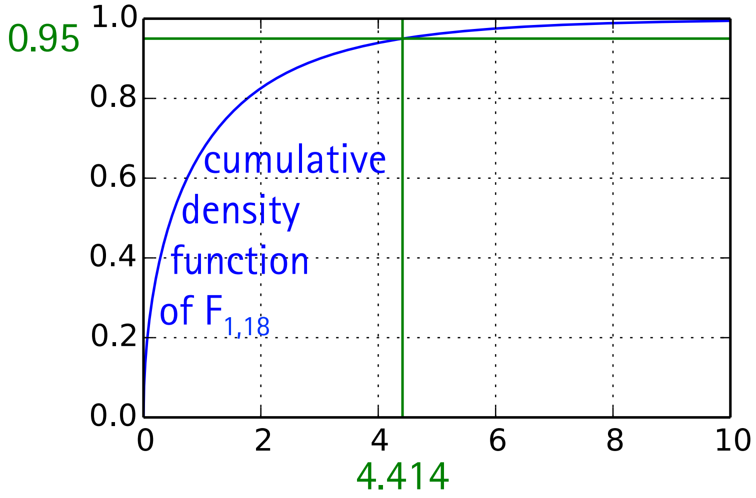
```
for i in xrange(experiments): # simulate experiments
    data1 = sigma * np.random.randn(groupSize, 1) + mu1; # normally distributed
    data2 = sigma * np.random.randn(groupSize, 1) + mu2; # normally distributed
    data = np.hstack((data1, data2)) # horizontally stack data columns
    grandMean = np.mean(data) # overall mean
    groupMeans = np.mean(data, 0) # mean for each group

    bwGroupVar = np.sum(groupSize * (grandMean - groupMeans) ** 2) / (groups - 1)
    wiGroupVar = np.sum((data - groupMeans) ** 2) / (groups * groupSize - groups)
    F = bwGroupVar / wiGroupVar
    results[i] = F
```

Numpy Script to Simulate Experiments (3/3)

```
# histogram output
pp.figure() # a new figure window
# a normalized, cumulative histogram with 100 bins
pp.hist(results, bins = 100, normed = True, cumulative = True)
pp.hold(True) # keep histogram when adding line
pp.plot([0, 20], [0.95, 0.95]) # add horizontal line
pp.xlim(0, 20) # limit x-axis from 0 to 20
pp.show() # actually show the result
```

SciPy F-Statistic: scipy.stats.f

- Import module
  - import scipy.stats as st

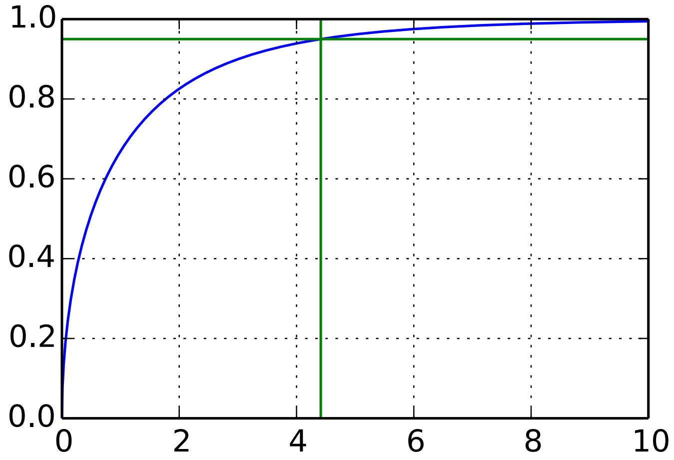


- Probability density function
  - pdf
- Cumulative density function
  - print st.f.cdf(4.414, 1, 18) # output: 0.950002951227
  - 95% of the distribution of $F_{1,18}$ is below 4.414
- Percent point function (inverse of cdf)
  - print st.f.ppf(0.95, 1, 18) # output: 4.41387341917
  - 4.414 is the point at which 95% of the cdf is reached

SciPy F-Statistic: scipy.stats.f

- Script to generate the figure

```
import numpy as np
import matplotlib.pyplot as pp
import scipy.stats as st
dfn, dfd = 1, 18
x = np.arange(0, 10, 0.01)
y = st.f.cdf(x, dfn, dfd)
pp.figure(figsize=(3,2))
pp.plot(x, y, 'b')
pp.grid()
p95 = st.f.ppf(0.95, dfn, dfd)
pp.plot([p95, p95], [0, 1], 'g')
pp.plot([0, 10], [0.95, 0.95], 'g')
```



Summary Analysis of Variance

- Measurements of DV are random samples of populations
- Null hypothesis: all measurements are from one population
  - H_0: $\mu_1 = \mu_2$ (population means are equal)
- Alternative hypothesis: not all means are equal
  - Many possibilities, difficult to analyze → focus on H_0
- The larger F, the more likely a systematic effect is present

$$F = \frac{\text{between-group variability}}{\text{within-group variability}}$$

  - The larger F, the smaller the likelihood of H_0
  - If probability of H_0 is low enough (typically $\alpha = 5\%$): reject H_0 → accept alternative hypothesis
  - However, this is not a proof!

Errors in Hypothesis Testing

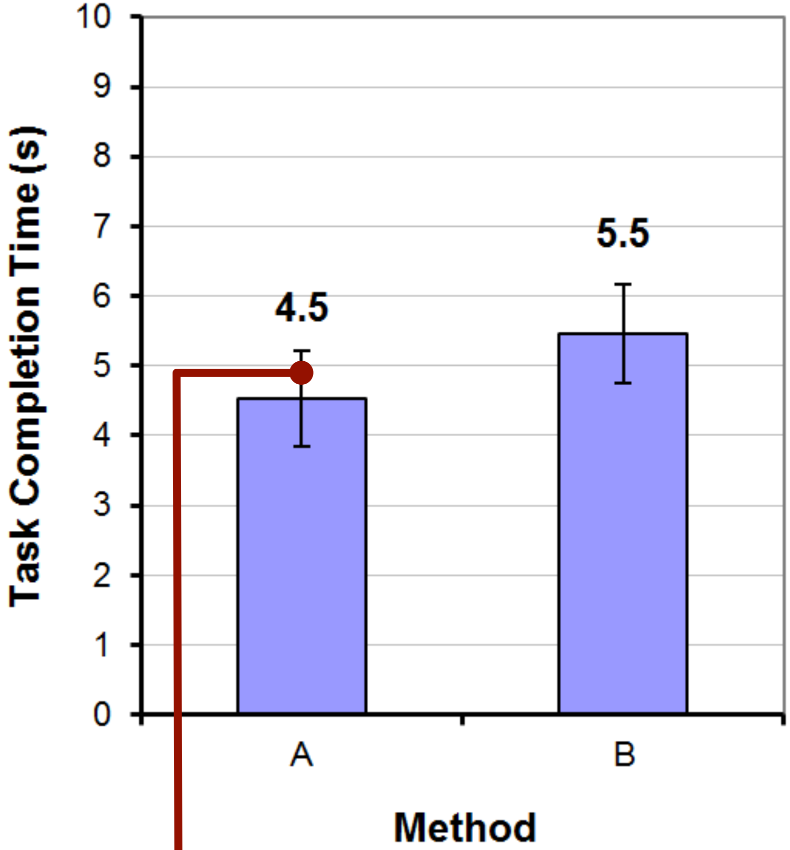
- Null hypothesis is true (there is no effect)
  - Experiment yields $p \leq 0.05$, null hypothesis is rejected
    → type I error (falsely rejecting the null hypothesis) , probability α
  - Experiment yields $p > 0.05$
    → correctly accepting the null hypothesis
- Null hypothesis is false (there is an effect)
  - Experiment yields $p \leq 0.05$, null hypothesis is rejected
    → correctly rejecting the null hypothesis
  - Experiment yields $p > 0.05$
    → type II error (failing to reject the null hypothesis), probability β
  - Power: Probability of finding an effect if there is one ($1-\beta$)

Sensitivity and Specificity

		condition (reality)		
	there is an effect	condition positive	condition negative	there is no effect
test result	test positive	true positive	false positive (type I error)	positive predictive value = $\frac{\text{true positive}}{\text{test positive}}$
	test negative	false negative (type II error)	true negative	negative predictive value = $\frac{\text{true negative}}{\text{test negative}}$
		sensitivity = $\frac{\text{true positive}}{\text{condition positive}}$	specificity = $\frac{\text{true negative}}{\text{condition negative}}$	

Example #1 – Details

Note: Within-subjects design



Error bars here show
±1 standard deviation

Participant	Method	
	A	B
1	5.3	5.7
2	3.6	4.8
3	5.2	5.1
4	3.6	4.5
5	4.6	6.0
6	4.1	6.8
7	4.0	6.0
8	4.8	4.6
9	5.2	5.5
10	5.1	5.6
Mean	4.5	5.5
SD	0.68	0.72

MacKenzie: Human-Computer Interaction - An Empirical Research Perspective.

Example #1 – ANOVA with Statistics Software

ANOVA Table for Task Completion Time (s)

	DF	Sum of Squares	Mean Square	F-Value	P-Value	Lambda	Power
Subject	9	5.080	.564				
Method	1	4.232	4.232	9.796	.0121	9.796	.804
Method * Subject	9	3.888	.432				

Probability of obtaining the observed data if the null hypothesis is true

- If the data points of both groups come from the same normal distribution, then the F-ratio follows the F-distribution
- $p(F_{1,9} < 9.796) = 0.9879$
- $p(F_{1,9} \geq 9.796) = 1 - p(F_{1,9} < 9.796) = 0.0121$
- Reported as: $F_{1,9} = 9.80$, $p < .05$

(permutation test yielded p = 0.0098)

MacKenzie: Human-Computer Interaction - An Empirical Research Perspective.

How to Report an ANOVA Result

The mean task completion time for Method A was 4.5 s. This was 20.1% less than the mean of 5.5 s observed for Method B. The difference was statistically significant ($F_{1,9} = 9.80$, $p < .05$).

- Report means
  - 4.5 s vs. 5.5 s
- Report effect sizes (as ratios or differences)
  - 20.1% (ratio of "improvement")
- State results of ANOVA
  - $F_{1,9} = 9.80$, $p < .05$

MacKenzie: Human-Computer Interaction - An Empirical Research Perspective.

How to Report an ANOVA Result

> The mean task completion time for Method A was 4.5 s. This was 20.1% less than the mean of 5.5 s observed for Method B. The difference was statistically significant ($F_{1,9} = 9.80$, $p < .05$).

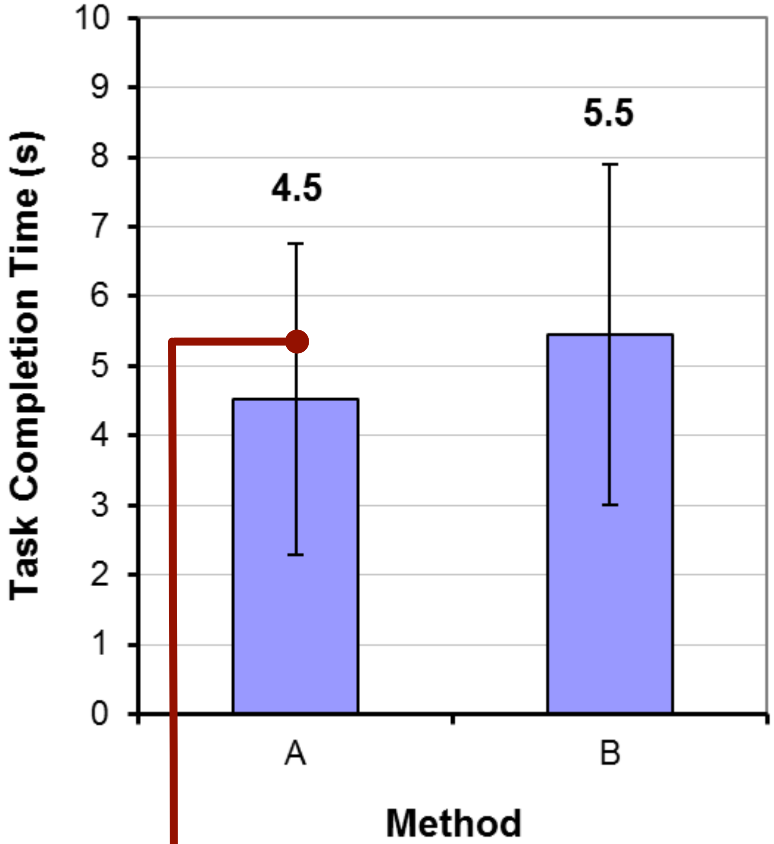
- The actual results are the observations and measurements
  - So always report them!
- Statistical tests just have a supporting role
  - They allow estimating confidences in the conclusions to be drawn

MacKenzie: Human-Computer Interaction - An Empirical Research Perspective.

Effect Size

- Describes magnitude of the association between IV and DV
  - Helps judging the practical relevance of a (stat. sig.) difference
- Unstandardized effect size ← typically used in HCI
  - Absolute difference between means
    - E.g., 1.0 s
  - Relative difference between means
    - E.g., 20.1%
  - Does not consider variability within groups
- Standardized effect size ← less often used in HCI
  - Size of effect relative to variability in the sample
  - Example: Cohen's d is absolute difference of means
    divided by standard deviation in the sample

Example #2 – Details



Error bars here show ±1 standard deviation

Note: Within-subjects design

Participant	Method	
	A	B
1	2.4	6.9
2	2.7	7.2
3	3.4	2.6
4	6.1	1.8
5	6.4	7.8
6	5.4	9.2
7	7.9	4.4
8	1.2	6.6
9	3.0	4.8
10	6.6	3.1
Mean	4.5	5.5
SD	2.23	2.45

MacKenzie: Human-Computer Interaction - An Empirical Research Perspective.

Example #2 – ANOVA

ANOVA Table for Task Completion Time (s)

	DF	Sum of Squares	Mean Square	F-Value	P-Value	Lambda	Power
Subject	9	37.372	4.152				
Method	1	4.324	4.324	.626	.4491	.626	.107
Method * Subject	9	62.140	6.904				

(permutation test yielded p = 0.4395)

Probability of obtaining the observed data if the null hypothesis is true

Reported as...

$F_{1,9} = 0.626$, ns

Note: For non-significant effects, use "ns" if $F < 1.0$, or "$p > .05$" if $F > 1.0$.

MacKenzie: Human-Computer Interaction - An Empirical Research Perspective.

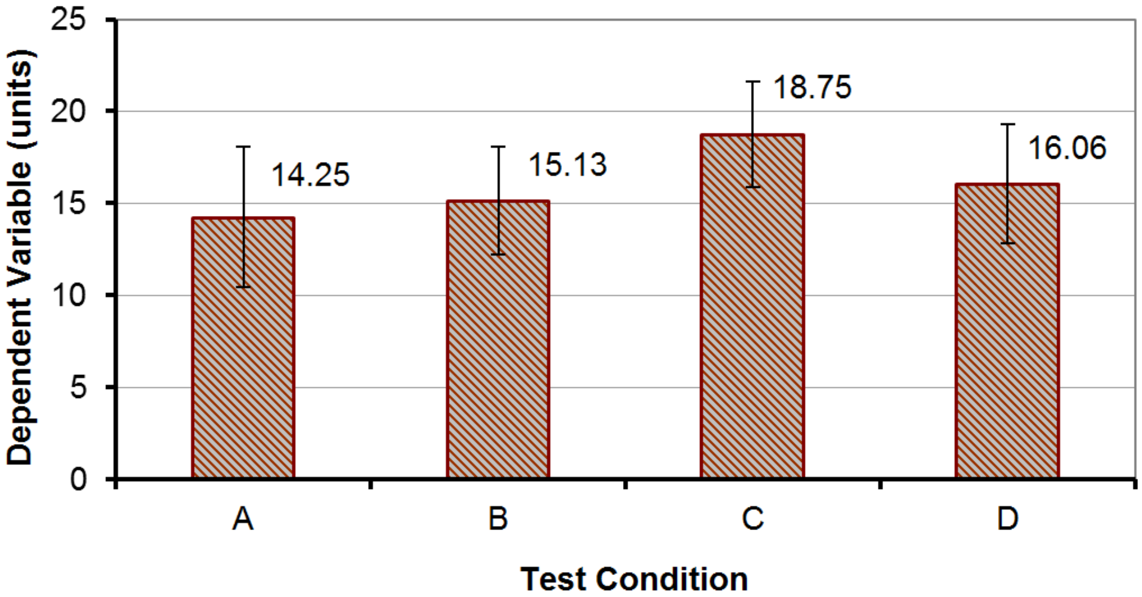
Example #2 – Reporting

The mean task completion times were 4.5 s for Method A and 5.5 s for Method B. As there was substantial variation in the observations across participants, the difference was not statistically significant as revealed in an analysis of variance ($F_{1,9}$ = 0.626, ns).

MacKenzie: Human-Computer Interaction - An Empirical Research Perspective.

More Than Two Test Conditions

Participant	Test Condition			
	A	B	C	D
1	11	11	21	16
2	18	11	22	15
3	17	10	18	13
4	19	15	21	20
5	13	17	23	10
6	10	15	15	20
7	14	14	15	13
8	13	14	19	18
9	19	18	16	12
10	10	17	21	18
11	10	19	22	13
12	16	14	18	20
13	10	20	17	19
14	10	13	21	18
15	20	17	14	18
16	18	17	17	14
Mean	14.25	15.13	18.75	16.06
SD	3.84	2.94	2.89	3.23



MacKenzie: Human-Computer Interaction - An Empirical Research Perspective.

ANOVA (Single Factor, Within Subjects)

ANOVA Table for Dependent Variable (units)

	DF	Sum of Squares	Mean Square	F-Value	P-Value	Lambda	Power
Subject	15	81.109	5.407				
Test Condition	3	182.172	60.724	4.954	.0047	14.862	.896
Test Condition * Subject	45	551.578	12.257				

- Single-factor, within-subjects design
- There was a significant effect of test condition on the dependent variable ($F_{3,45} = 4.95$, $p < .005$)
- Degrees of freedom (k=4 conditions, n=16 participants)
  - If k is the number of test conditions and n is the number of participants:
  - Participant dfs → n - 1
  - Condition dfs → k – 1
  - Error dfs → (n – 1)(k – 1)

MacKenzie: Human-Computer Interaction - An Empirical Research Perspective.

Post Hoc Comparisons Tests

- A significant F-test means that at least two test conditions differed significantly
- Does not indicate which test conditions differed significantly from one another
- To determine which pairs differ significantly, post hoc comparisons are used
- Typically: t-tests, adjusted for multiple comparisons
  - Adjustment: Avoid inflation of type I errors
- Examples:
  - Bonferroni/Dunn, Fisher PLSD, Dunnett, Tukey/Kramer, Games/Howell, Student-Newman-Keuls, orthogonal contrasts, Scheffé

MacKenzie: Human-Computer Interaction - An Empirical Research Perspective.

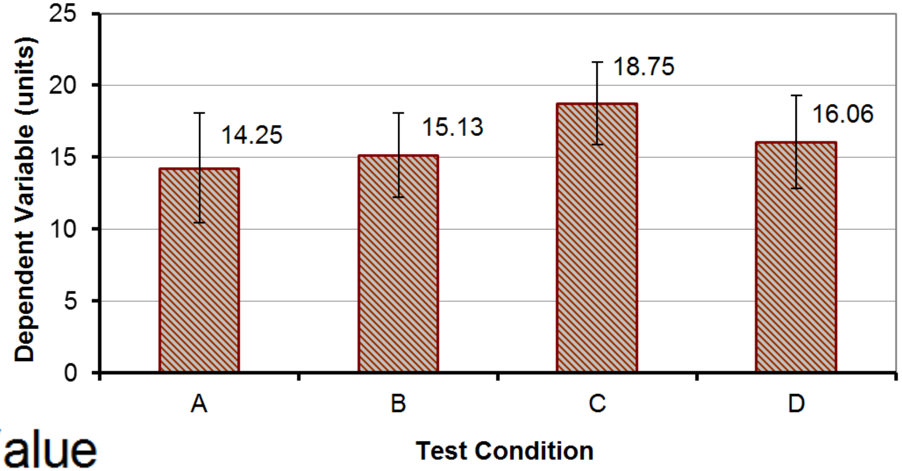
Type I Error Inflation under Multiple Comparisons

- Type I error: Falsely rejecting the null hypothesis, corresponds to α-level, typically $\alpha = 5\%$
- A single test: Type I error rate $\alpha^* = \alpha$
- Two (independent) tests: Type I error rate $\alpha^* = 1 - (1 - \alpha)^2$
  - First test no type I error ($p = 1 - \alpha$) and second test no type I error ($p = 1 - \alpha$)
  - Both tests no error (if independent) $(1 - \alpha)^2$
  - Combined error rate: $\alpha^* = 1 - (1 - \alpha)^2$
- k (independent) tests: $\alpha^* = 1 - (1 - \alpha)^k$
- If tests not independent: $\alpha^* = k\alpha$
- Modify α such that $\alpha^* \leq 5\%$
- Bonferroni correction: $\alpha = \alpha^*/k$ (conservative)

k	$1 - (1 - \alpha)^k$	$k\alpha$
1	5.0%	5.0%
2	9.8%	10.0%
3	14.3%	15.0%
4	18.5%	20.0%
5	22.6%	25.0%
6	26.5%	30.0%

Post Hoc Comparisons

Simple Bonferroni correction can be too conservative (inflating false negatives)



	Mean Diff.	Crit. Diff.	P-Value	
A, B	-.875	3.302	.9003	
A, C	-4.500	3.302	.0032	S
A, D	-1.813	3.302	.4822	
B, C	-3.625	3.302	.0256	S
B, D	-.938	3.302	.8806	
C, D	2.688	3.302	.1520	

significance level: 5%

⇨ Test conditions A:C and B:C differ significantly ($\alpha \leq 0.05$)

MacKenzie: Human-Computer Interaction - An Empirical Research Perspective.