t-tests

Testing system A vs system B

AB t-tests

Today's goal:

Teach you about the t-test, the test used to measure the difference between two conditions

Outline:

- The independent t-test (for between-subjects studies)
- The dependent t-test (for within-subjects studies)

Independent t-test

for between-subjects studies

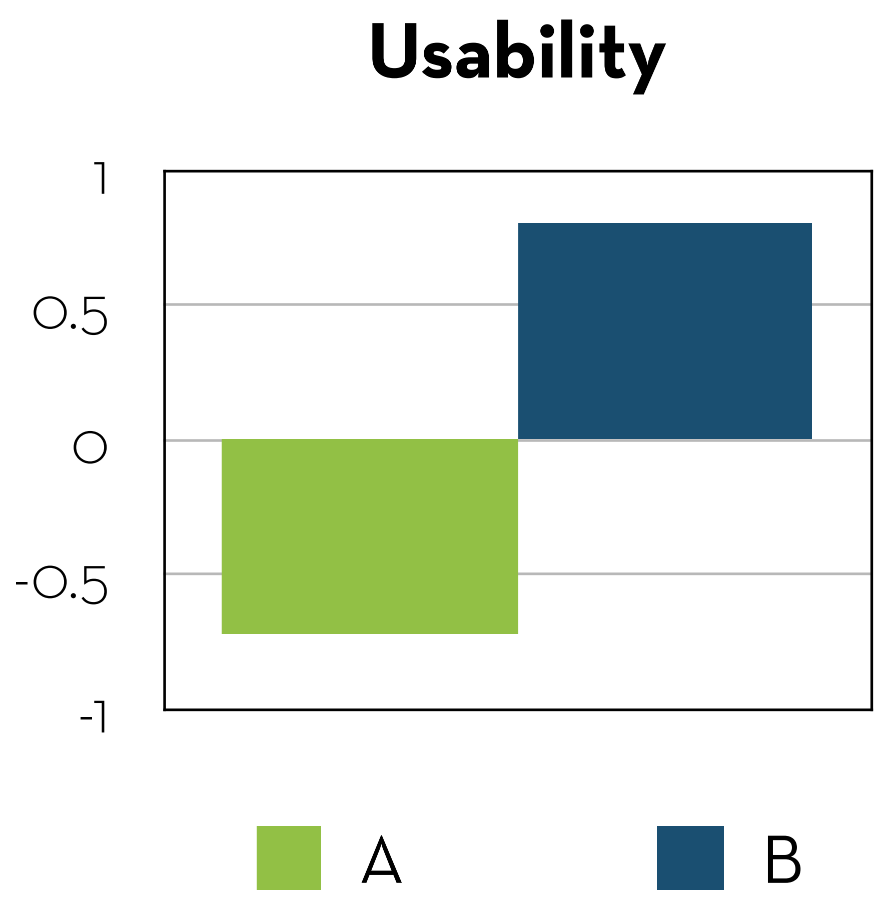
Independent t-test

Difference between two systems:

Do these two UIs (A and B) lead to a different level of usability?

Differences between two groups of people:

Do men (A) and women (B) perceive different levels of usability?



AB Independent t-test

Usability for users of system A:

3, 2, 3, 4, 1

Usability for users of system B:

5, 4, 5, 4, 5

Which system is more usable?

Is this difference significant?

AB Independent t-test

Calculate the means. Do they differ a little or a lot?

Given no effect, we expect the means to be roughly equal.

- May differ by chance, but no large differences expected
- Null hypothesis (H0): Ma = Mb

Compare the found difference to the standard error of the difference

- If the SE is small, we expect small differences under H0
- If it is large, large differences are more likely

Independent t-test

If the difference is larger than expected based on the SE:

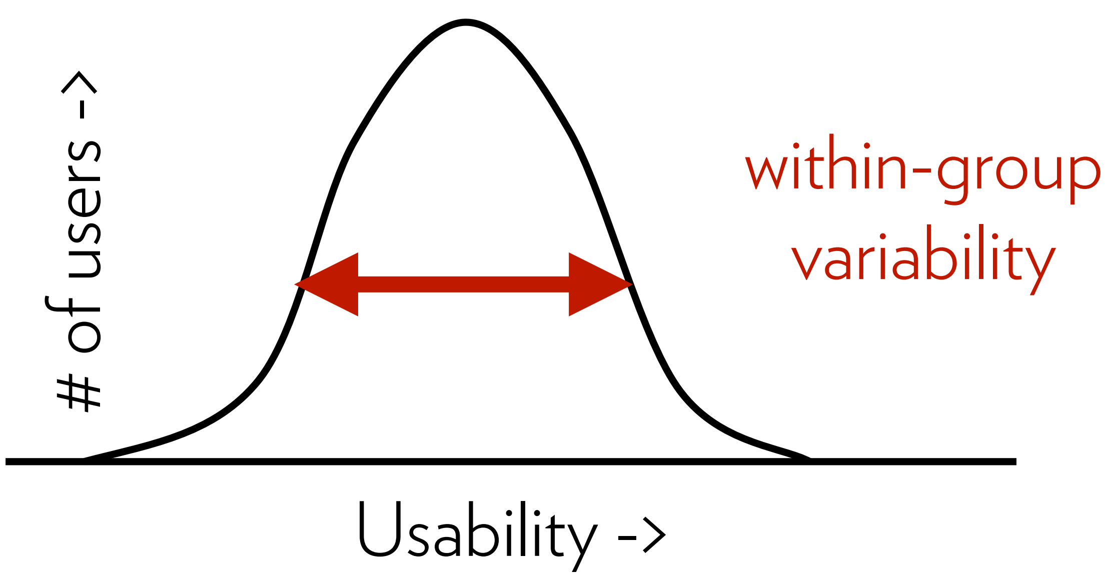
- We may still have found a difference by chance (no real effect), or...
- There is a real difference in means (H0 is incorrect).

The larger the difference, the more confident we are that H0 is incorrect. Then, H1 is supported

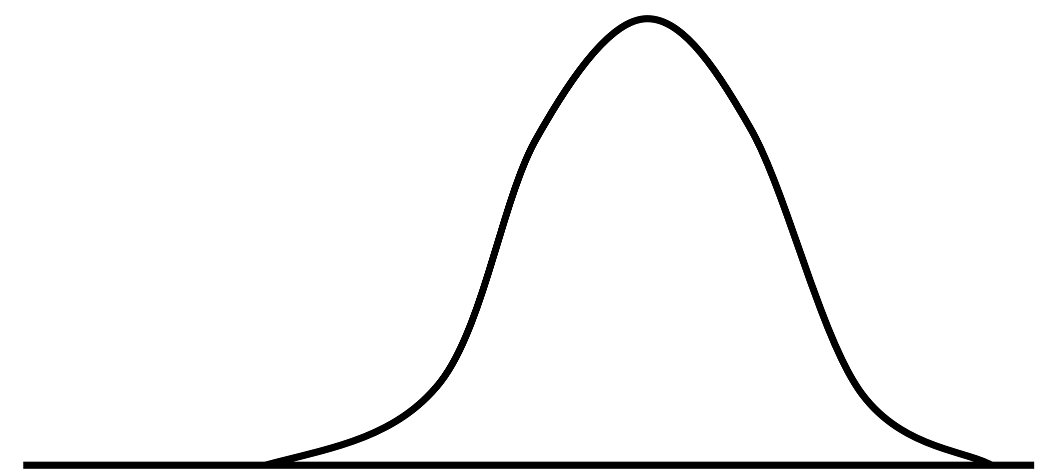
But never **proven**, because the first option may still apply!

t-test concept

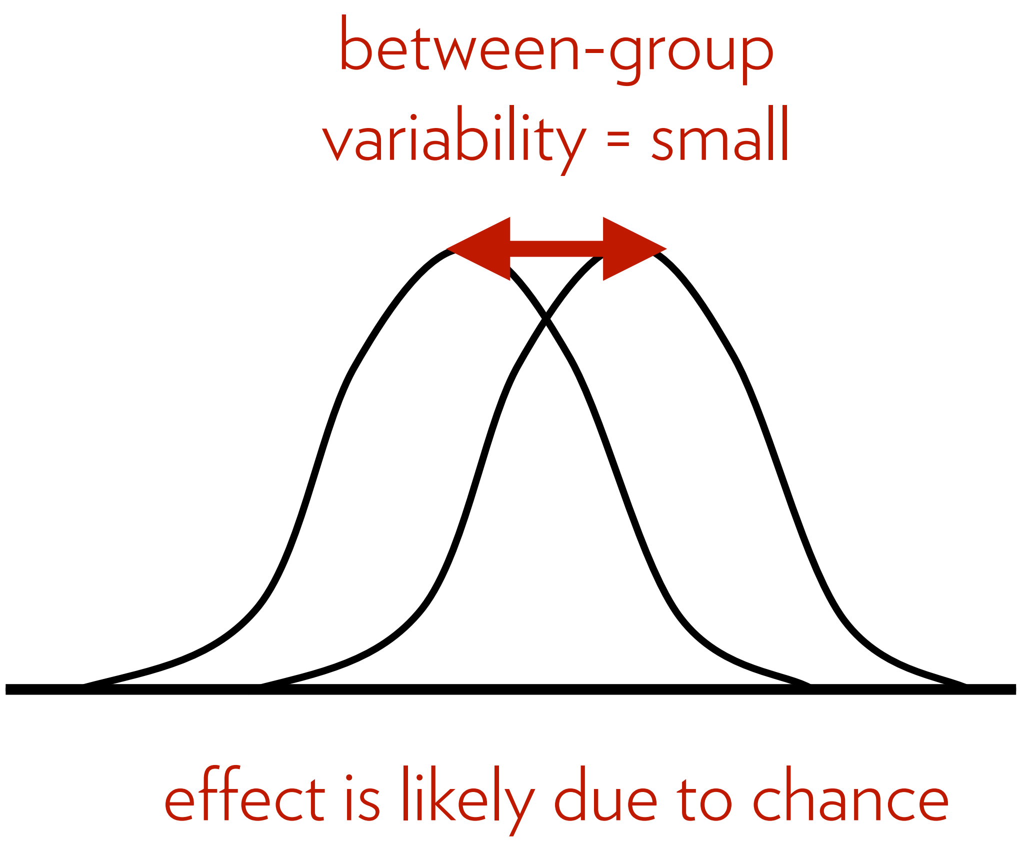
Usability for users of system A:



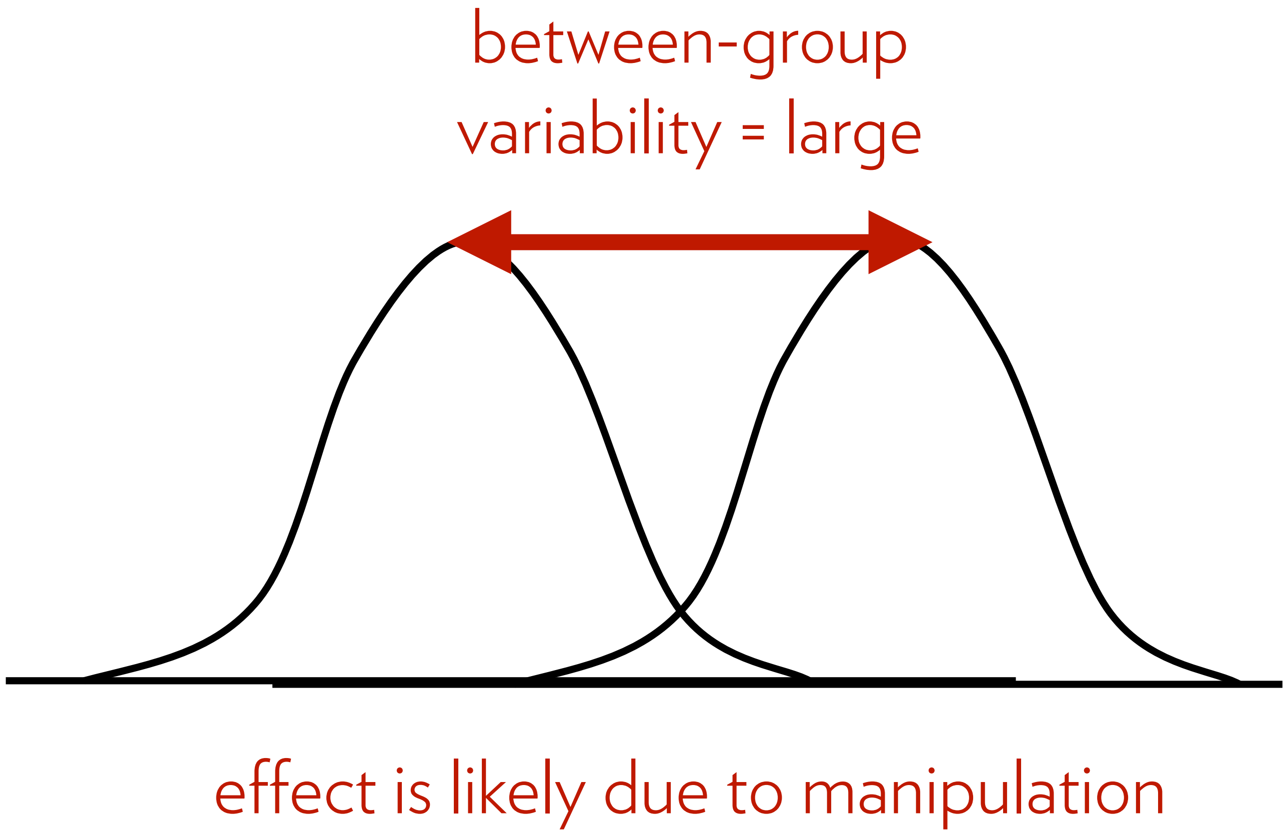
Usability for users of system B:



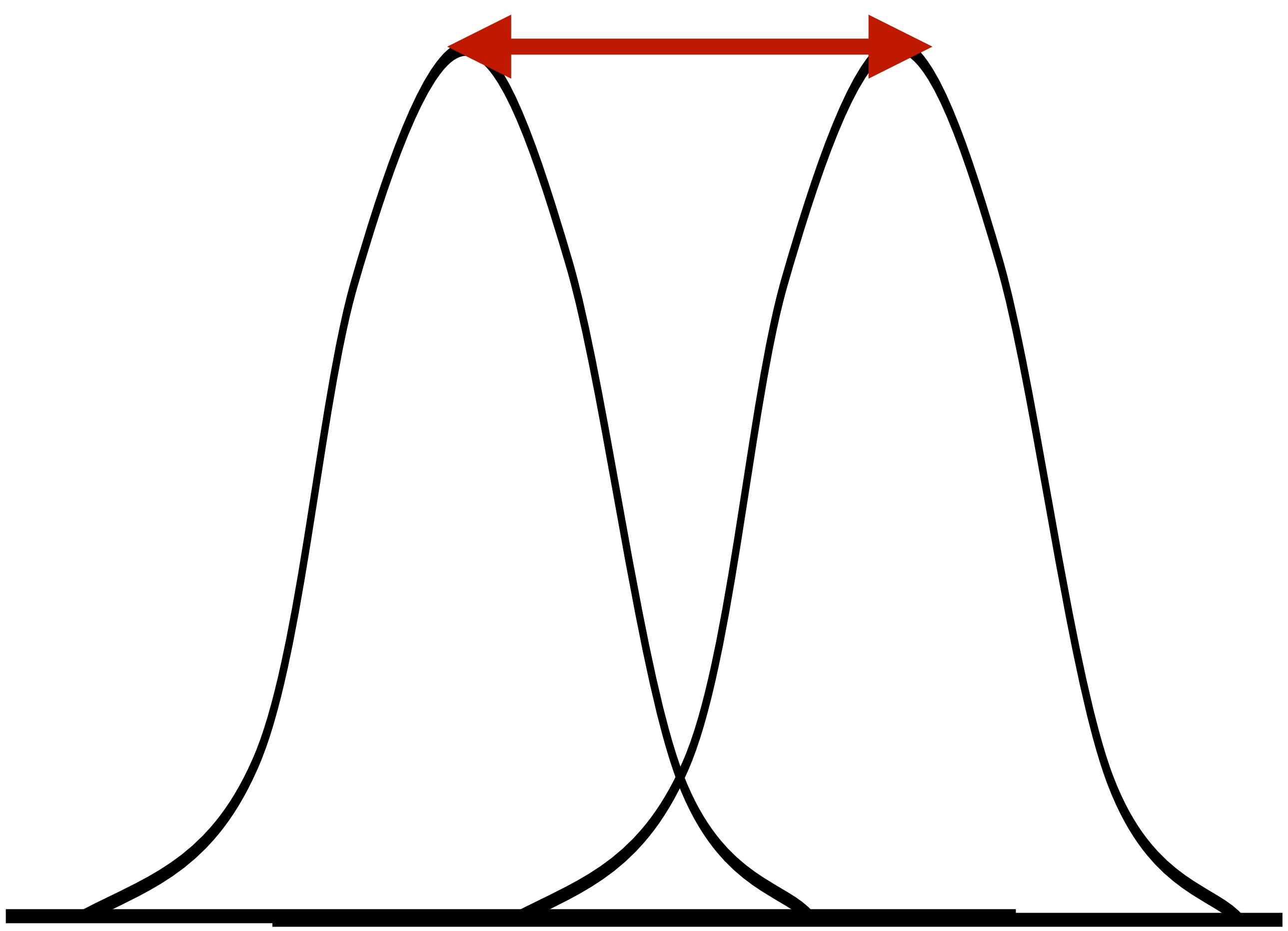
AB t-test concept



t-test concept

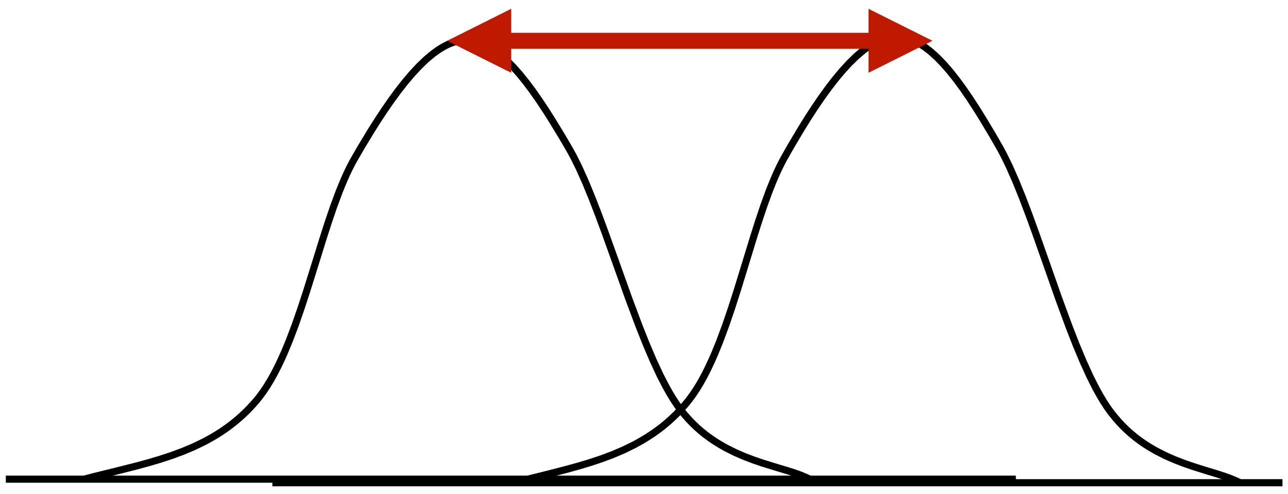


t-test concept

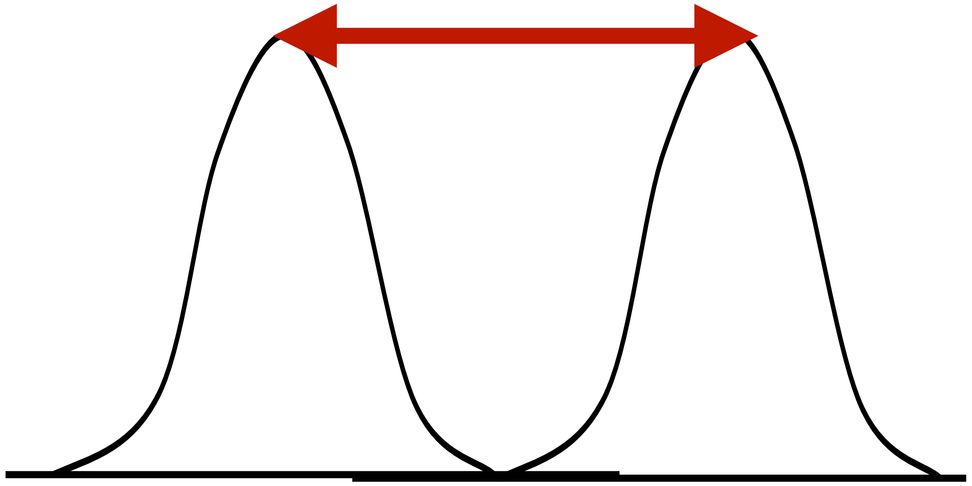


more data = stronger test

t-test concept



t-test concept



lower variability = stronger test

AB t-test formula

t-test: compare the difference in means (the variation explained by the model model) with the standard error of that difference (the residual variation)

$t = (Ma - Mb)/SE_{Ma\text{-}Mb}$

Ma and Mb each have their own SE, but what is the SE of the difference?

The variance of a difference between two independent variables is equal to the sum of their variances!

(and variance = SE^2)

AB Calculating the SE

SE of mean A = $s_a/\sqrt{}N_a$, so the variance of mean 1 = s^2_a/N_a

SE of mean B = $s_b/\sqrt{}N_b$, so the variance of mean 2 = s^2_b/N_b

Sum: $s^2_a/N_a + s^2_b/N_b$

Translate back to SE: $\sqrt{}(s^2_a/N_a + s^2_a/N_a)$

t-test formula

t-test: compare the difference in means (M) with the standard error ($\sqrt{(s^2a/Na+s^2b/Nb)}$)

$t = (Ma - Mb)/\sqrt{(s^2a/Na+s^2b/Nb)}$

this test has Na + Nb – 2 degrees of freedom

For our example:

$Ma = 2.6$, $s^2a = 1.3$, $Na = 5$

$Mb = 4.6$, $s^2b = 0.3$, $Nb = 5$

$t = 3.53$, $p = 0.01317$

It is all the same!

Regression: $Y = a + bX + e$

T-test: let's say you test system A versus B

Your X is a dummy variable:

- $X = 0$ for system A, and 1 for system B
- For system A: $Y = a + b*0 = a$
- For system B: $Y = a + b*1 = a + b$

Parameter b tests the difference between system A and B!

Let's do it in R:

Dataset "SpiderLong.dat" -> set name to spiderLong

- Effect exposure to a real spider vs. a picture on anxiety

Variables:

- Group: whether participants saw a Picture or a Real Spider
- Anxiety: anxiety level

AB Plotting

Bar chart with error bars:

ggplot(spiderLong,aes(Group,Anxiety))
+stat_summary(fun.y=mean, geom="bar", fill="white",
color="black") + stat_summary(fun.data=mean_cl_normal,
geom="errorbar", width=0.2)

Boxplot:

ggplot(spiderLong,aes(Group,Anxiety))+geom_boxplot()

AB Descriptives

Descriptives per group:

- by(spiderLong$Anxiety, spiderLong$Group, stat.desc, basic = F, norm = T)
- looks pretty normal!

The t-test

difT <- t.test(Anxiety ~ Group, data = spiderLong)

difT

```
	Welch Two Sample t-test

data:  Anxiety by Group
t = -1.6813, df = 21.385, p-value = 0.1072
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -15.648641   1.648641
sample estimates:
    mean in group Picture mean in group Real Spider 
                       40                        47 
```

AB As a regression

difR <- lm(Anxiety ~ Group, data = spiderLong)

summary(difR)

```
Coefficients:
                 Estimate Std. Error t value Pr(>|t|)    
(Intercept)        40.000      2.944  13.587 3.53e-12 ***
GroupReal Spider    7.000      4.163   1.681    0.107    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 10.2 on 22 degrees of freedom
Multiple R-squared:  0.1139,	Adjusted R-squared:  0.07359 
F-statistic: 2.827 on 1 and 22 DF,  p-value: 0.1068
```

AB Assumptions

Normal distribution

Interval level data

Independence

Heteroscedasticity is okay!

The two groups can have different variances, because we conduct "Welch's t-test"

AB Robust methods

What if the data is not normal? —> Robust methods!

Note: these have been updated since Field's book came out

Wide format no longer needed!

We can run yuen (in WRS2):

```
yuen(Anxiety ~ Group, data = spiderLong)
```

Or with less trimming (default is 20%):

```
yuen(Anxiety ~ Group, data = spiderLong, tr = 0.1)
```

AB Robust methods

For bootstrapping we can run yuenbt:

yuenbt(Anxiety ~ Group, data = spiderLong, nboot = 2000)

Or using M-estimators (no trimming needed):

pb2gen(Anxiety ~ Group, data = spiderLong, boot = 2000)

In sum, all of them seem to suggest that there is no significant difference!

AB Effect size

$r = \sqrt{(t^2 / (t^2 + df))}$

In R:

```
t <- difT$statistic[[1]]
df <- difT$parameter[[1]]
r <- sqrt(t^2/(t^2+df))
round(r, 3)
```

r = .342, a medium effect, even though it is not significant!

AB Effect size

Cohen's d = $(M_a - M_b) / sd_{Ma\text{-}Mb}$

In R:

- load "psych" package
- `cohen.d(Anxiety~Group, data=spiderLong)`

d = .72 (also gives you r!)

Reporting

On average, participants experienced greater anxiety from real spiders (M = 47.00, SE = 3.18) than from pictures of spiders (M = 40.00, SE = 2.68). This difference was not significant $t(21.39) = -1.68$, $p = .107$; however, it did represent a medium-sized effect $r = .342$.

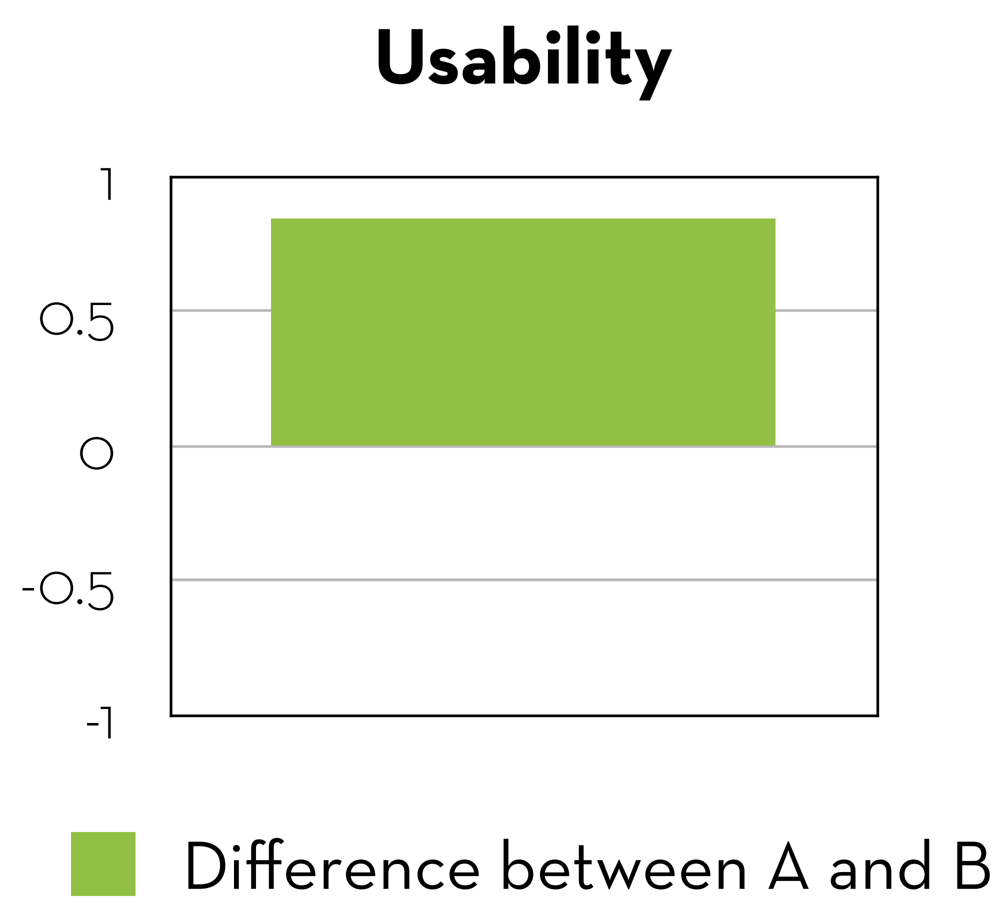
Dependent t-test

for within-subjects studies

Dependent t-test

Difference between two systems, tested by the same user

Differences in user evaluation of Facebook vs. Google Plus



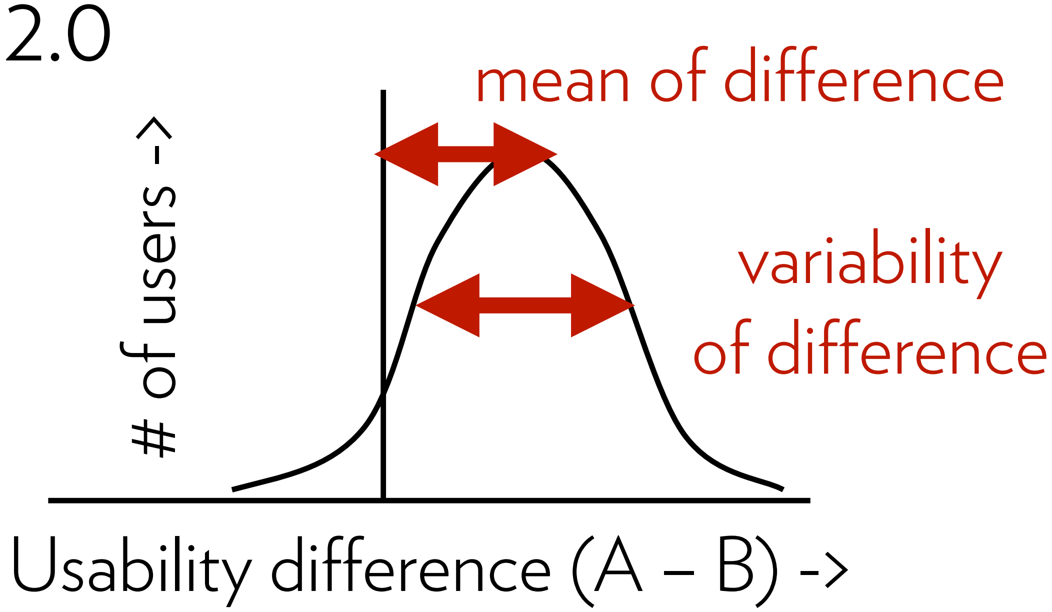
AB 1-sample t-test

Participant uses system A —> usability evaluation: 4.0

Participant uses system B —> usability evaluation: 2.0

Calculate the difference: 2.0

Tabulate all differences:



AB T-test example

	u1	u2	u3	u4	u5
A	3	2	3	4	1
B	5	4	5	4	5
Diff	2	2	2	0	4

T-test: compare the difference (D) with SE_D ($S_D/\sqrt{N}$)

$$t = D/(S_D/\sqrt{N})$$

For our example:

$D = 2.0$, $S_D = 1.41$, $N = 5$

$t = 3.16$, $p = 0.034$

Let's do it in R:

Dataset "SpiderWide.dat" -> set name to spiderWide

Effect exposure to a real spider vs. a picture on anxiety, but now tested within subjects

Variables:

picture: anxiety when seeing a picture of a spider

real: anxiety when seeing a real spider

AB

Plotting

Stack the data:

```
spiderStack <- stack(spiderWide)
names(spiderStack) <- c(“Anxiety”,”Group”)
```

Bar chart with error bars:

```
ggplot(spiderStack,aes(Group,Anxiety))
+stat_summary(fun.y=mean, geom=“bar”, fill=“white”,
color=“black”) + stat_summary(fun.data=mean_cl_normal,
geom=“errorbar”, width=0.2)
```

Plotting

Huh? Same as spiderLong?

- Wasn't within-subjects supposed to be better?

Problem: error contains between-subjects differences

- Solution: remove those!

Plotting

How?

Subtract participant mean, add grand mean:

```
spiderAdjusted <- spiderWide

spiderAdjusted$picture <- spiderWide$picture -
(spiderWide$picture + spiderWide$real)/2 +
mean((spiderWide$picture + spiderWide$real)/2)

spiderAdjusted$real <- spiderWide$real -
(spiderWide$picture + spiderWide$real)/2 +
mean((spiderWide$picture + spiderWide$real)/2)
```

AB Plotting

Stack the data:

```
spiderStack <- stack(spiderAdjusted)
names(spiderStack) <- c(“Anxiety”,”Group”)
```

Bar chart with error bars:

```
ggplot(spiderStack,aes(Group,Anxiety))
+stat_summary(fun.y=mean, geom=“bar”, fill=“white”,
color=“black”) + stat_summary(fun.data=mean_cl_normal,
geom=“errorbar”, width=0.2)
```

That looks better!

Descriptives

Descriptives per group:

```
stat.desc(spiderWide, basic = F, norm = T)
```

Better: descriptives of the difference

```
stat.desc(spiderWide$real-spiderWide$picture, basic = F, norm = T)
```

looks pretty normal!

The t-test

dif <- t.test(spiderWide$real, spiderWide$picture, paired=T)

dif

```
        Paired t-test

data:  spiderWide$real and spiderWide$picture
t = 2.4725, df = 11, p-value = 0.03098
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
  0.7687815 13.2312185
sample estimates:
mean of the differences
                      7
```

AB Robust methods

What if the data is not normal? —> Robust methods!

Need to load the "source" version of WRS2:

```
install.packages("WRS2", type="source")
```

Let's start with yuend:

```
WRS2::yuend(spiderWide$real, spiderWide$picture)
```

Robust methods

For bootstrapping we can run ydbt:

```
WRS2::ydbt(spiderWide$real, spiderWide$picture, nboot = 2000)
```

Or using M-estimators (no trimming needed):

```
WRS2::bootdpci(spiderWide$real, spiderWide$picture, est=tmean, nboot = 2000)
```

In sum, the robust methods seem to disagree…

Effect size

Same as before: $r = \sqrt{(t^2 / (t^2 + df))}$

In R:

```
t <- dif$statistic[[1]]
df <- dif$parameter[[1]]
r <- sqrt(t^2/(t^2+df))
round(r, 3)
```

r = .598, a large effect

AB Effect size

Cohen's $d_z = M_{difference} / sd_{difference}$

In R:

```
mean(spiderWide$real–spiderWide$picture)/
sd(spiderWide$real–spiderWide$picture)
```

$d_z = .714$

Reporting

On average, participants experienced significantly greater anxiety from real spiders (M = 47.00, SE = 3.18) than from pictures of spiders (M = 40.00, SE = 2.68), $t(11)$ = 2.47, p = .030, r = .598.

“It is the mark of a truly intelligent person to be moved by statistics.”

T H A N K S !

George Bernard Shaw