PSY 503: Foundations of Statistics in Psychological Science

Comparing Two Means

Jason Geller, Ph.D. (he/him/his)

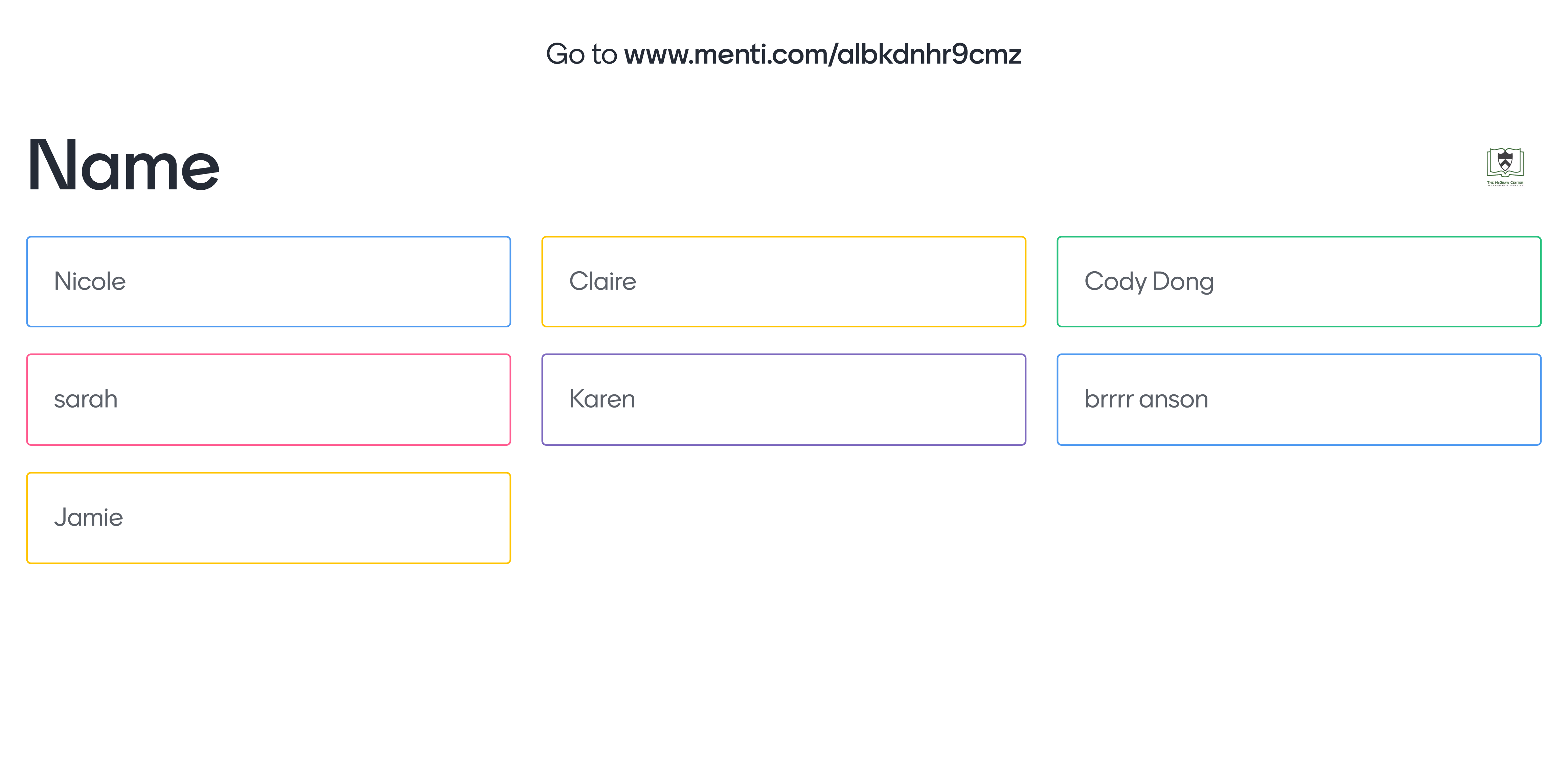
Princeton University

Last Updated: 2022-10-10

Housekeeping

- Problem Set 2 grades posted
- Problem Set 3 will be posted later today
- Data for the final project needs to be approved by October 31st

Knowledge Check



Go to **www.menti.com/albkdnhr9cmz**

Name

Nicole	Claire	Cody Dong
sarah	Karen	brrrr anson
Jamie		

Last Class

- NHST
    - We can only falsify a theory
        - p-value = likelihood of the observed data given the null is true
    - One- and two-sided hypotheses
    - One sample tests

Today

- Two sample *t*-tests
    - Independent
    - Dependent (paired)
- Non-parametric
- Multiple Comparisons

Experiments

- Simple experiments:
  - One IV that's binary with two options
  - One DV that's interval/ratio/continuous

- For example: manipulation of the independent variable involves having an experimental condition and a control
  - This situation can be analyzed with a *t*-test
  - We can also use *t*-tests to analyze any binary independent variable
  - The *t*-test is a simple regression model with one categorical predictor

Experiments

- Don't make a continuous variable categorical just so you can do a *t*-test
- People used to split variables into high versus low or simply split down the middle
    - You separate the people who are close together and lump them with people who are not really like them
    - Effect sizes get smaller
    - You will also decrease power and see Type II errors

Experiments

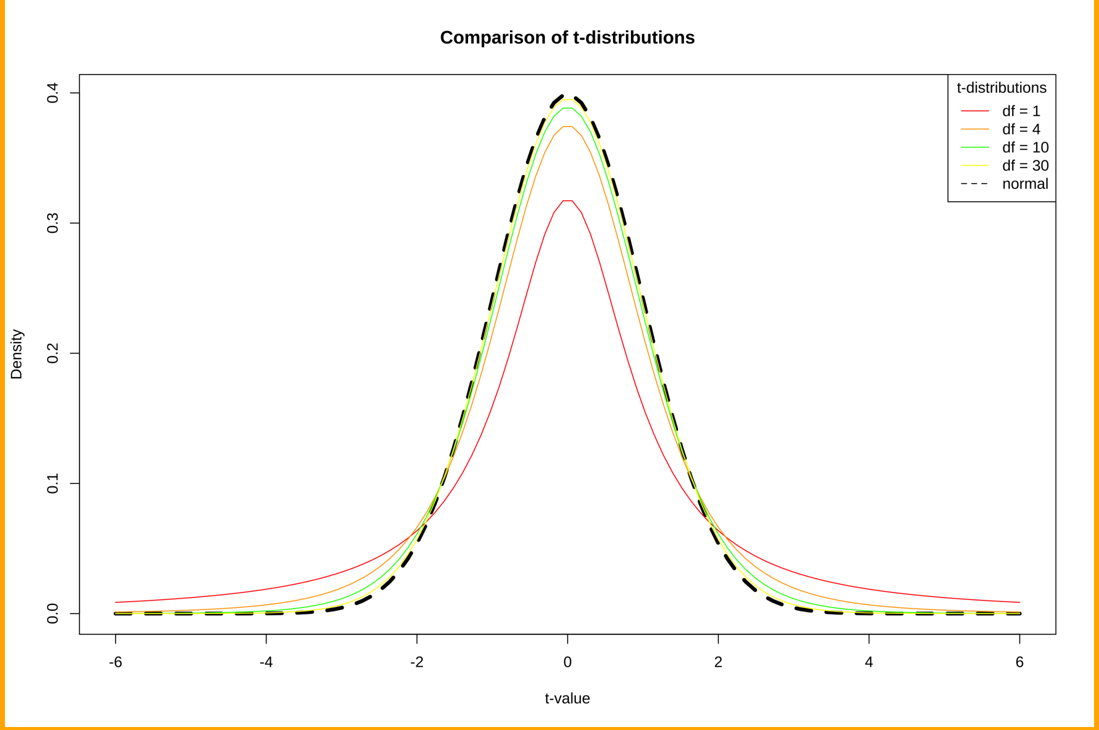
- Between subjects / Independent designs
  - Expose different groups to different experimental manipulations
- Repeated measures / within subjects / dependent designs
  - Take a single group of people and expose them to different experimental manipulations at different points in time

The *t*-test

- Independent *t*-test:
  - Compares two means based on independent data
  - Used when different participants were assigned to each condition of the study
- Dependent *t*-test:
  - Compares two means based on related data
  - Used when the same participants took part in both conditions of the study

t distribution

- William Gosset discovered it while working for Guinness
    - Often called student's *t* distribution
- Small samples: more conservative test
- *t*-distribution has fatter tails



Independent: Example

- Are invisible people mischievous?
- Manipulation
  - Placed participants in an enclosed community riddled with hidden cameras
  - 12 participants were given an invisibility cloak
  - 12 participants were not given an invisibility cloak
- Outcome measured how many mischievous acts participants performed in a week

```r
library(rio)
library(tidyverse)
library(easystats)
library(kableExtra)

longdata <- read_csv("https://raw.githubusercontent.com/doomlab/statsofdoom-files/master/graduate/R%20Flip/11_ttests/d

head(longdata)
```

```
## # A tibble: 6 × 2
##   Cloak    Mischief
##   <chr>       <dbl>
## 1 No Cloak        3
## 2 No Cloak        1
## 3 No Cloak        5
## 4 No Cloak        4
## 5 No Cloak        6
## 6 No Cloak        4
```

Independent: Understanding the NHST

- H_0: The no cloak and cloak groups would have the same mean

- H_1: The no cloak and cloak groups would have different means

```
M <- tapply(longdata$Mischief, longdata$Cloak, mean)
STDEV <- tapply(longdata$Mischief, longdata$Cloak, sd)
N <- tapply(longdata$Mischief, longdata$Cloak, length)
M;STDEV;N
```

```
##    Cloak No Cloak
##     5.00     3.75
```

```
##    Cloak No Cloak
## 1.651446 1.912875
```

```
##    Cloak No Cloak
##       12       12
```

Independent: Understanding the NHST

- Our means appear slightly different. What might have caused those differences?
  - Variance created by our manipulation: The cloak (systematic variance)
  - Variance created by unknown factors (unsystematic variance)

Independent: Understanding the NHST

- If the samples come from the same population, then we expect their means to be roughly equal
- Although it is possible for their means to differ by chance alone, here, we would expect large differences between sample means to occur very infrequently
- We compare the difference between the sample means that we collected to the difference between the sample means that we would expect to obtain if there were no effect (i.e. if the null hypothesis were true)

Independent: Understanding the NHST

- We use the standard error as a gauge of the variability between sample means
- If the difference between the samples we have collected is larger than what we would expect based on the standard error then we can assume one of two interpretations:
    - There is no effect and sample means in our population fluctuate a lot and we have, by chance, collected two samples that are atypical of the population from which they came *(Type 1 error)*
    - The two samples come from different populations but are typical of their respective parent population. In this scenario, the difference between samples represents a genuine difference between the samples *(and so the null hypothesis is incorrect)*

Independent: Understanding the NHST

- As the observed difference between the sample means gets larger, the more confident we become that the second explanation is correct (i.e., that the null hypothesis should be rejected)

- If the null hypothesis is incorrect, then we gain confidence that the two sample means differ because of the different experimental manipulation imposed on each sample

Independent: Formulas

$$t = \frac{(\overline{X}_1 - \overline{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\dfrac{s_1^2}{n_1} + \dfrac{s_2^2}{n_2}}}$$

$$t = \frac{\overline{X}_1 - \overline{X}_2}{\sqrt{\dfrac{s_1^2}{n_1} + \dfrac{s_2^2}{n_2}}}$$

$$s_p^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}$$

Independent: Data Screening

- Assumptions:
  - No missingness (NAs)
  - No outliers
  - Independence
  - Normality (each group should be approximately normally distributed)
  - Homogeneity: equal variances between groups

Independent: Data Screening

- Missingness

```
longdata %>%
  drop_na()
```

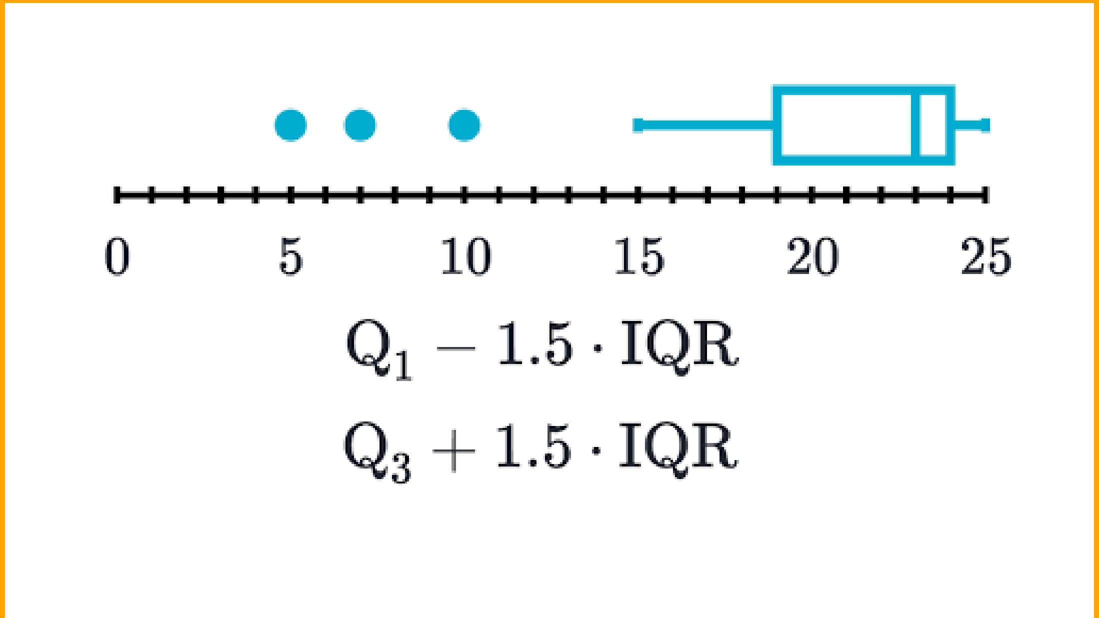
```
## # A tibble: 24 × 2
##    Cloak    Mischief
##    <chr>       <dbl>
##  1 No Cloak        3
##  2 No Cloak        1
##  3 No Cloak        5
##  4 No Cloak        4
##  5 No Cloak        6
##  6 No Cloak        4
##  7 No Cloak        6
##  8 No Cloak        2
##  9 No Cloak        0
## 10 No Cloak        5
## # … with 14 more rows
## # ℹ Use `print(n = ...)` to see more rows
```

Independent: Data Screening

- Outliers

```r
library(rstatix)

longdata %>%
  group_by(Cloak) %>%
  identify_outliers(Mischief)
```

```
## [1] Cloak       Mischief   is.outlier is.extreme
## <0 rows> (or 0-length row.names)
```



$$Q_1 - 1.5 \cdot \mathrm{IQR}$$

$$Q_3 + 1.5 \cdot \mathrm{IQR}$$

Independent: Data Screening

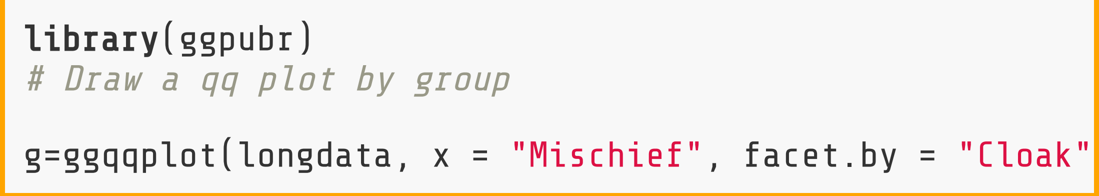
- Normality

```
longdata %>%
  group_by(Cloak) %>%
  shapiro_test(Mischief)
```

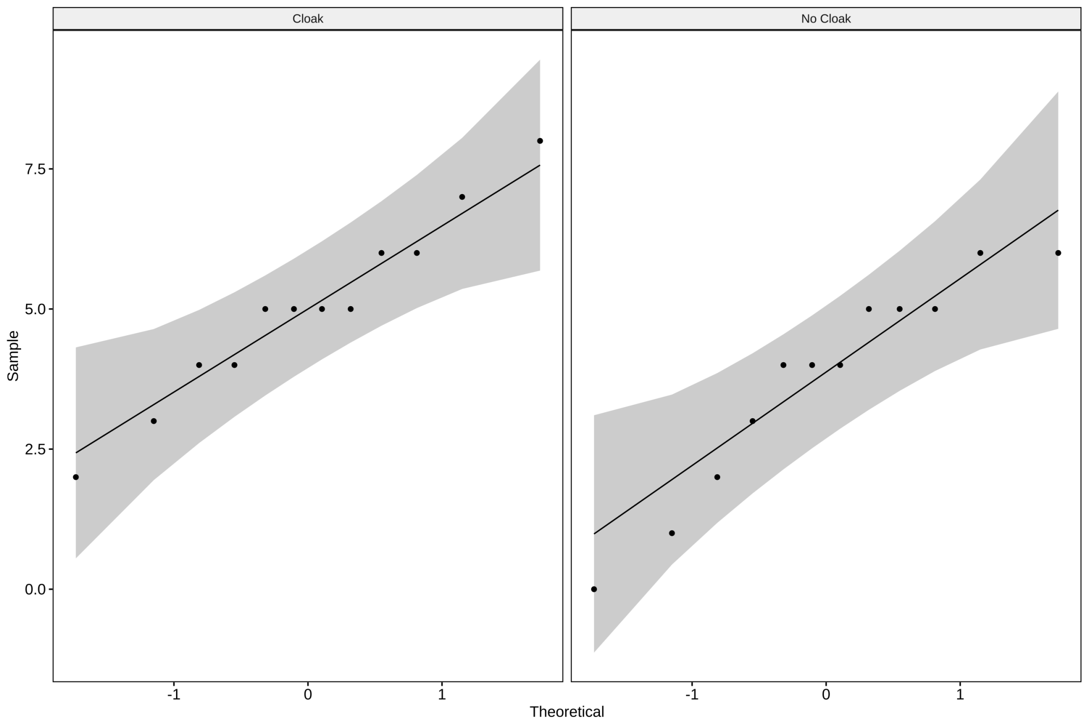
```
## # A tibble: 2 × 4
##   Cloak    variable statistic     p
##   <chr>    <chr>        <dbl> <dbl>
## 1 Cloak    Mischief     0.973 0.936
## 2 No Cloak Mischief     0.913 0.231
```

Independent: Data Screening

- Normality
  - qqplot



```
library(ggpubr)
# Draw a qq plot by group

g=ggqqplot(longdata, x = "Mischief", facet.by = "Cloak"
```



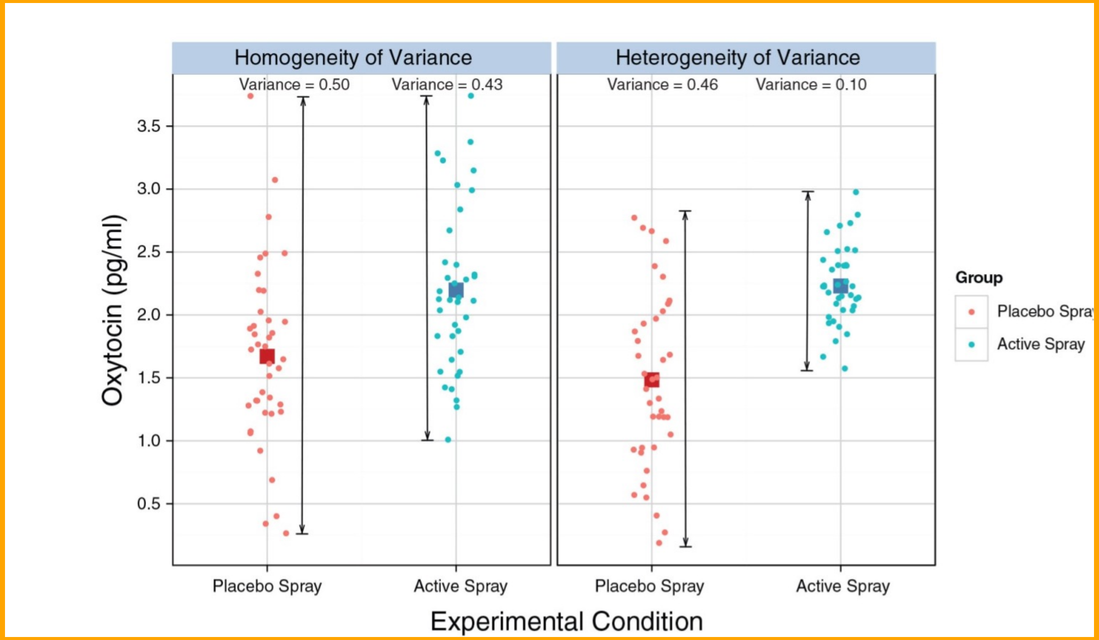
Independent: Homogeneity

- The most common problem is lack of homogeneity
    - where the group variance is not equal between groups

```
longdata %>%
  levene_test(Mischief~Cloak)
```

```
## # A tibble: 1 × 4
##     df1   df2 statistic     p
##   <int> <int>     <dbl> <dbl>
## 1     1    22     0.270 0.609
```

- if the *p*-value of the Levene's test is not-significant, we can use that as satisfying our variance assumption



2 Sample Welch’s *t*-test

- Welch’s *t*-test gives gives equivalent answer to traditional *t*-test when there is an equal sample size or variances, BUT can also handle unequal sample size and variance
    - http://daniellakens.blogspot.com/2015/01/always-use-welchs-t-test-instead-of.html
- Same t-statistic calculation and t-distribution, just have to apply correction for degrees of freedom (df)

2 Sample Welch's *t*-test

$$\mathrm{df} = \frac{\left(\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}\right)^2}{\frac{\left(\frac{\sigma_1^2}{n_1}\right)^2}{n_1 - 1} + \frac{\left(\frac{\sigma_2^2}{n_2}\right)^2}{n_2 - 1}}$$

$$A = \frac{s_1^2}{n1} \qquad \text{and} \qquad B = \frac{s_2^2}{n_2}$$

Independent: Analysis

```r
library(report)
d_ind <- t.test(Mischief ~ Cloak,
       data = longdata,
       var.equal = TRUE, #assume equal variances
       paired = FALSE) #independent

d_ind <- t.test(Mischief ~ Cloak,
       data = longdata,
       var.equal = FALSE, #assume unequal variances
       paired = FALSE) #independent
```

- No differences between groups was found: $t(22) = 1.71, p = .101$
- `Easystats` package in R can help write this up for you :)
- Effect sizes were labelled following Cohen's (1988) recommendations.

The Welch Two Sample t-test testing the difference of Mischief by Cloak (mean in group Cloak = 5.00, mean in group No Cloak = 3.75) suggests that the effect is positive, statistically not significant, and medium (difference = 1.25, 95% CI [-0.26, 2.76], t(21.54) = 1.71, p = 0.101; Cohen's d = 0.74, 95% CI [-0.14, 1.60])

```
report(d_ind)
```

The *t*-test as linear model

- Can be viewed through regression framework:

$$\text{Mischief} = \beta_0 + \beta_1(\text{Cloak}_{\text{No}}) + \epsilon$$

- Categorical variables are *dummy coded* or *treatment coded*
  - In R, levels of categorical variable transformed to 0 and 1
  - By default, 0 is attached to whatever variable comes first in alphabet
- β_1 = difference between the two groups

The *t*-test as linear model

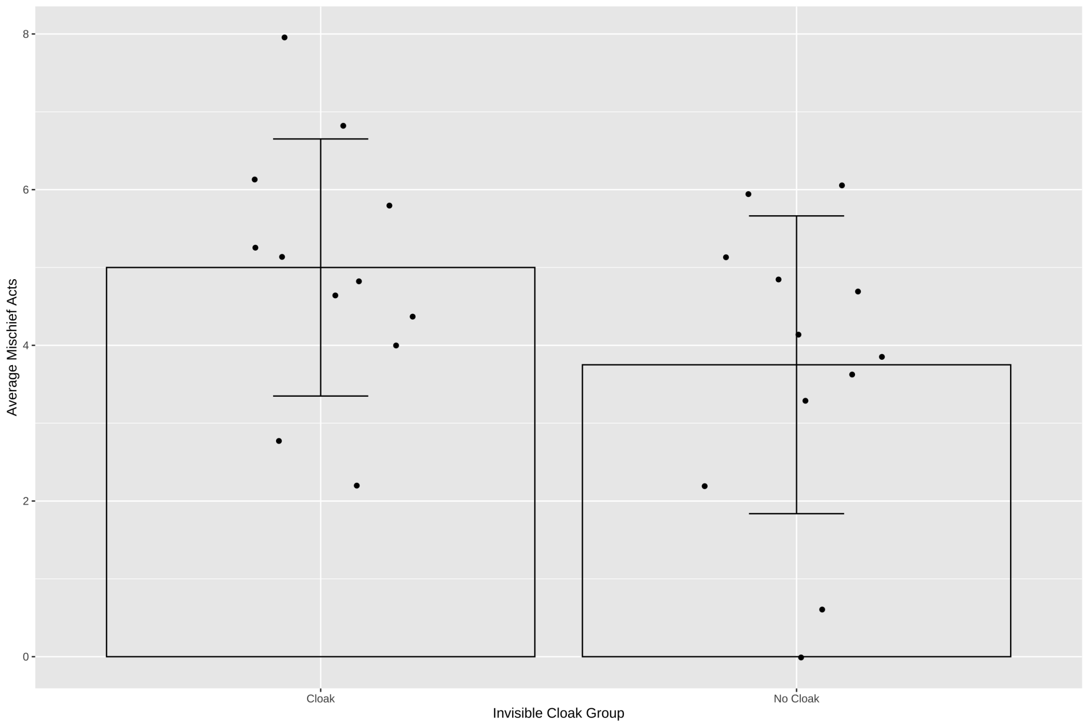
```
library(tidyverse)
library(broom)

d=lm(Mischief ~ Cloak,data = longdata)

broom::tidy(d)
```

```
## # A tibble: 2 × 5
##   term          estimate std.error statistic       p.value
##   <chr>            <dbl>     <dbl>     <dbl>         <dbl>
## 1 (Intercept)       5        0.516      9.69 0.00000000212
## 2 CloakNo Cloak    -1.25     0.730     -1.71 0.101
```

t-test: Visualization

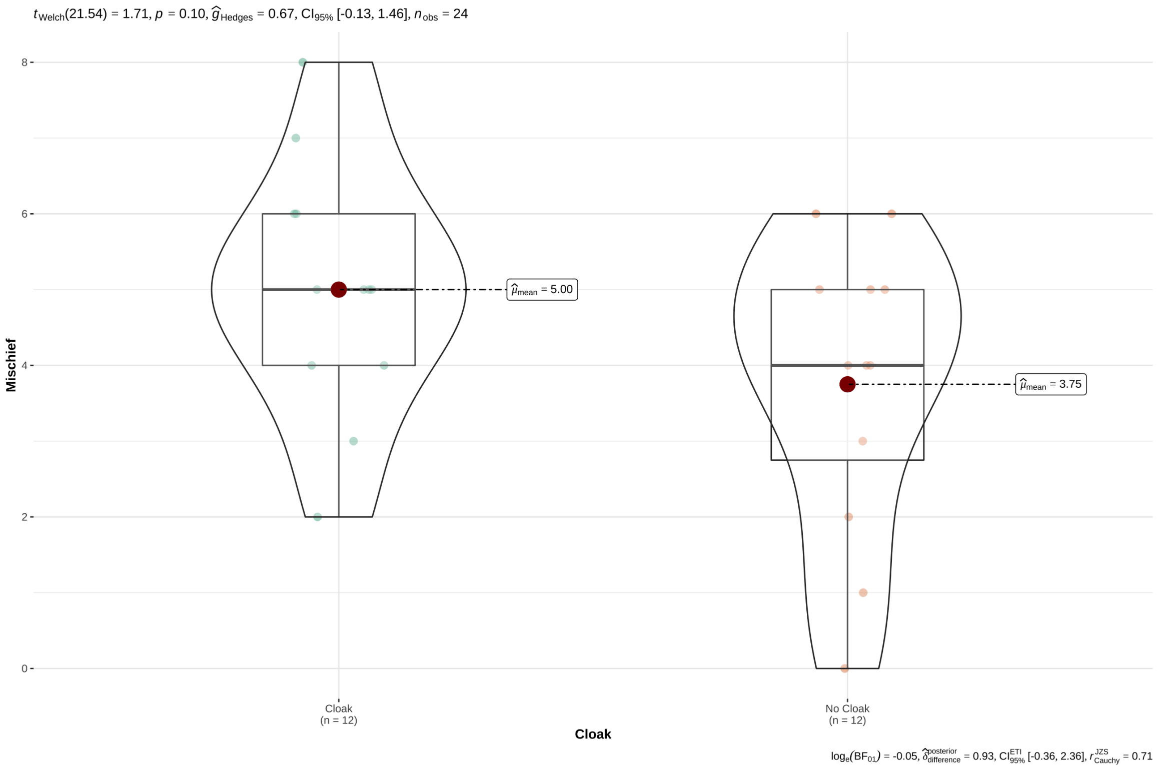
- Bar charts in ggplot2 with only one x variable (the different levels of your IV) and one y variable.

```r
library(ggplot2)
library(ggpubr)
df.summary <- longdata %>%
  group_by(Cloak) %>%
  summarise(
    sd = sd(Mischief, na.rm = TRUE),
    Mischief = mean(Mischief)
  )

d=ggplot(longdata, aes(Cloak, Mischief)) +
  geom_bar(stat = "identity", data = df.summary,
           fill = NA, color = "black") +
  geom_jitter( position = position_jitter(0.2),
               color = "black") +
  geom_errorbar(
    aes(ymin = Mischief-sd, ymax = Mischief+sd),
    data = df.summary, width = 0.2)  +
  xlab("Invisible Cloak Group") +
  ylab("Average Mischief Acts")
```



t-test: Visualization

```
ggstatsplot::ggbetweenstats(
  data = longdata,
  x = Cloak,
  y = Mischief)
```



2 Sample *t*-test Independant (Practice Example 1)

An educator believes that new directed reading activities in the classroom will help elementary school students improve some aspects of their reading ability. She arranges for a third grade class of 21 students to take part in these activities for an 8-week period. A control classroom of 23 third graders follows the same curriculum without the activities. At the end of 8 weeks, all students are given a Degree of Reading Power (DRP) test, which measures the aspects of reading ability that the treatment is designed to improve.

```
treatment=c(24,43,58,71,43,49,61,44,67,49,53,56,59,52,62,54,57,33,46,43,57)

control=c(42,43,55,26,62,37,33,41,19,54,20,85,46,10,17,60,53,42,37,42,55,28,48)
```

NHST Steps

1. State hypotheses

2. Check assumptions

3. Run `t.test`

4. Decision/conclusion

5. Visualize

2 Sample Welch's *t*-test (Practice Example 2)

A math test was given to 300 17 year old students in 1978 and again to another 17 year old students in 1992.

Group 1: X1 = 300.4, S1 = 34.9, n = 300 Group 2: X2 = 306.7, S2 = 30.1, n = 350

- Use α = 0.01

NHST Steps

1. State hypotheses
2. Simulate data (`rnorm`)
3. Check assumptions
4. Calculate t and DF correction
5. Run `t.test`
6. Decision/conclusion

Calculating Scores

- Group 1: X1 = 300.4, S1 = 34.9, n = 300
- Group 2: X2 = 306.7, S2 = 30.1, n = 350

```
n1=300
n2=350

t.stat=(300.4-306.7)/sqrt(34.9^2/300+30.1^2/350)

#df correction Welsch

#A=s1^2/n1
#B=s2^2/n2

A=34.9^2/300
B=30.1^2/350

df=(A+B)^2/(A^2/(n1-1)+B^2/(n2-1))

df
```

```
## [1] 594.7025
```

R Calculation

```
t4 =((306.7    - 300.4)-(0-0))/(34.9^2 / 300 + 30.1^2 / 350)^(1/2)

v4 =(34.9 ^2 / 300 + 30.1^2 / 350)^2 / (34.9^4 / (300^2 * (300 - 1)) + 30)

alpha4 =0.01

tcrit4 = qt(alpha4/2, v4)

pval4 =2 *pt(-abs(t4),v4)

abs(t4)
```

```
## [1] 2.443286
```

In R: Welch's *t*-test

```r
group1 <- rnorm(300,mean =300.4,sd=34.9)
group2 <- rnorm(350,mean =306.7,sd=30.1)

c=t.test(group1, group2,alternative ="two.sided")
```

Effect sizes were labelled following Cohen's (1988) recommendations.

The Welch Two Sample t-test testing the difference between group1 and group2 (mean of x = 303.72, mean of y = 304.89) suggests that the effect is negative, statistically not significant, and very small (difference = -1.17, 95% CI [-6.48, 4.14], t(566.72) = -0.43, p = 0.665; Cohen's d = -0.03, 95% CI [-0.19, 0.12])

Dependent (paired *t*-test)

Dependent: Example

- Are invisible people mischievous?
- Manipulation
    - Placed participants in an enclosed community riddled with hidden cameras
    - For first week participants normal behavior was observed
    - For the second week, participants were given an invisibility cloak
- Outcome: We measured how many mischievous acts participants performed in week 1 and week 2
- Note: Same data, but instead the study is dependent. Let's see what happens to our *t*-test

Dependent: Understanding the NHST

$$t = \frac{\bar{D} - \mu_D}{S_D/\sqrt{N}}$$

$$S_D = \sqrt{\frac{(d_1 - \overline{d})^2 + (d_2 - \overline{d})^2 + \cdots + (d_n - \overline{d})^2}{n - 1}}$$

- We are going to use the standard error of the differences rather than standard error
- The standard error of the differences is calculated by subtracting the two sets of scores and calculating standard deviation on that difference score

Dependent: Data Screening

- The data screening can be treated in the same fashion
  - Normality
  - Missingness
  - Outliers
- However, homogeneity between groups is not examined, because you do not have separate groups!
- The variance is calculated on one difference score, so there is not a homogeneity concern

Dependent: Analysis

- The cloak and no cloak conditions were different: $t(11) = 3.80, p = .003$

- Why is this result different than independent t?

```r
d_pair <-t.test(Mischief ~ Cloak,
       data = longdata,
       var.equal = TRUE, #ignored in dependent t
       paired = TRUE) #dependent t
```

Dependent: Reporting

- Effect sizes were labelled following Cohen's (1988) recommendations.

The Paired t-test testing the difference of Mischief by Cloak (mean difference = 1.25) suggests that the effect is positive, statistically significant, and large (difference = 1.25, 95% CI [0.53, 1.97], t(11) = 3.80, p = 0.003; Cohen's d = 1.15, 95% CI [0.37, 1.89])

Dependent GLM

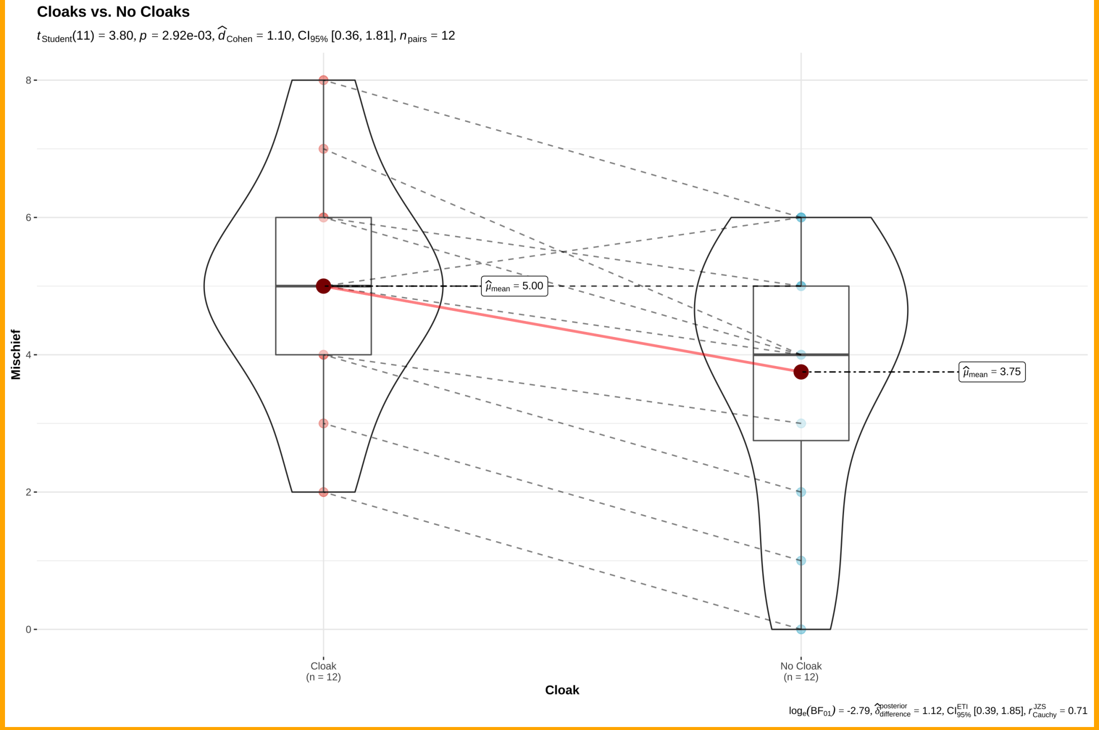
- Use the lm to test for significance
- Calculate the t value
    - Why is there a discrepancy?

```r
library(lme4)
library(sjPlot)

longdata$id<-rep(1:12, length(longdata))

d_reg<-lme4::lmer(Mischief~Cloak + (1|id),  data=longdata)
```

	Dependent variable				*Dependent variable*			
Predictors	Estimates	CI	p	df	Estimates	CI	p	df
(Intercept)	5.00	3.89 – 6.11	<0.001	13.45				
Cloak [No Cloak]	-1.25	-1.97 – -0.53	0.003	11.00				
Mischief					1.25	0.53 – 1.97	0.003	11.00
Random Effects								
σ^2	0.65							
τ_{00}	2.55 $_{\text{id}}$							
ICC	0.80							

t-test: Visualization

```
library(ggstatsplot)
## parametric t-test
p1 <- ggwithinstats(
  data = longdata,
  x = Cloak,
  y = Mischief,
  type = "p",
  effsize.type = "d",
  conf.level = 0.95,
  title = "Cloaks vs. No Cloaks",
  package = "ggsci",
  palette = "nrc_npg"
)
```

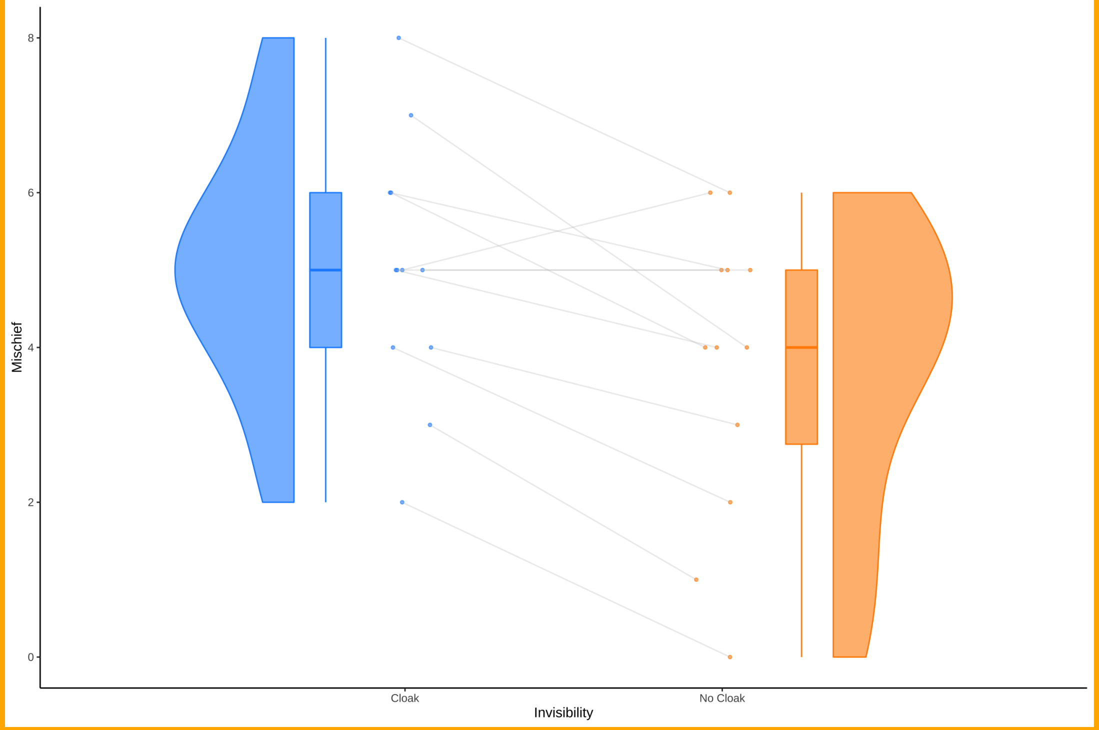

Cloaks vs. No Cloaks

$t_{\mathrm{Student}}(11) = 3.80$, $p = 2.92\mathrm{e}{-03}$, $\widehat{d}_{\mathrm{Cohen}} = 1.10$, $\mathrm{CI}_{95\%}$ [0.36, 1.81], $n_{\mathrm{pairs}} = 12$

$\log_e(\mathrm{BF}_{01}) = -2.79$, $\widehat{\delta}^{\mathrm{posterior}}_{\mathrm{difference}} = 1.12$, $\mathrm{CI}^{\mathrm{ETI}}_{95\%}$ [0.39, 1.85], $r^{\mathrm{JZS}}_{\mathrm{Cauchy}} = 0.71$

t-test: Visualization

```r
library(raincloudplots)

wide <- longdata %>%
    pivot_wider(names_from = "Cloak", values_from ="Mis

df_1x1 <- data_1x1(
  array_1 = wide$Cloak,
  array_2 = wide$`No Cloak`)

raincloud_2 <- raincloud_1x1_repmes(
  data = df_1x1,
  colors = (c('dodgerblue', 'darkorange')),
  fills = (c('dodgerblue', 'darkorange')),
  line_color = 'gray',
  line_alpha = .3,
  size = 1,
  alpha = .6,
  align_clouds = FALSE) +

scale_x_continuous(breaks=c(1,2), labels=c("Cloak", "No
  xlab("Invisibility") +
  ylab("Mischief") +
  theme_classic()

raincloud_2
```



NHST Steps

1. State hypotheses

2. Check assumptions

3. run `t.test`

4. Decision/conclusion

5. Visualize the data

Paired *t* in R

```r
data5diff=data5a-data5b
t5 =( mean(data5diff) - 0) / (sd(data5diff) / (length(data5diff))^(1/2.))
tcrit5 = qt(0.05/2, length(data5diff)-1)
pval5 =2 *pt(-abs(t5),length(data5diff)-1)
abs(t5)
```

```
## [1] 1.714286
```

Paired *t* in R

```
t.test(data5a, data5b,paired =TRUE,alternative ="two.sided")
```

```
##
##      Paired t-test
##
## data:  data5a and data5b
## t = 1.7143, df = 9, p-value = 0.1206
## alternative hypothesis: true mean difference is not equal to 0
## 95 percent confidence interval:
##  -0.5113467  3.7113467
## sample estimates:
## mean difference
##             1.6
```

Paired *t* (Practice Problem 2)

We know the weight of 10 mice before and after a treatment

```
before <-c(200.1, 190.9, 192.7, 213, 241.4, 196.9, 172.2, 185.5, 205.2, 193.7)
# Weight of the mice after treatment
after <-c(392.9, 393.2, 345.1, 393, 434, 427.9, 422, 383.9, 392.3, 352.2)
```

We want to know, if there is any significant difference in the mean weights after treatment? Assume α .05

NHST Steps

1. State hypotheses

2. Check assumptions

3. run `t.test`

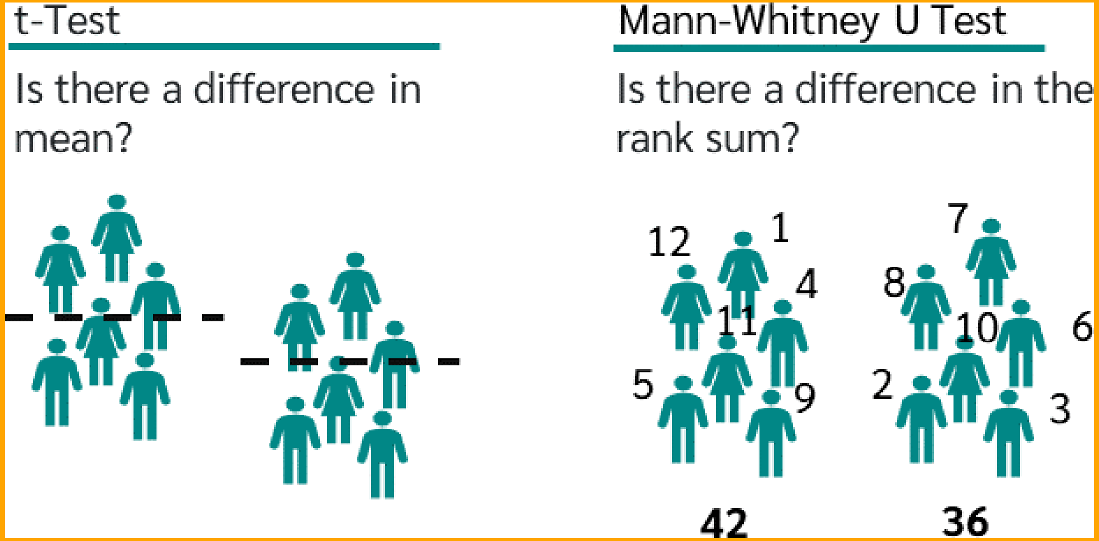
4. Decision/conclusion

5. Visualize the data

Non-parametric

- Sometimes data is non-normal (skewed, bimodal, etc.), or ordinal, so what do we do?
    - Use Shapiro-Wilk normality test
    - Can transform data (e.g., log, sqrt, etc.), but these also make assumptions
    - Robust methods
- Mann-Witney U test (indep)
- Wilcoxon (paired)
    - Rank method based on calculating all possibilities to form distribution

Mann-Witney U Example



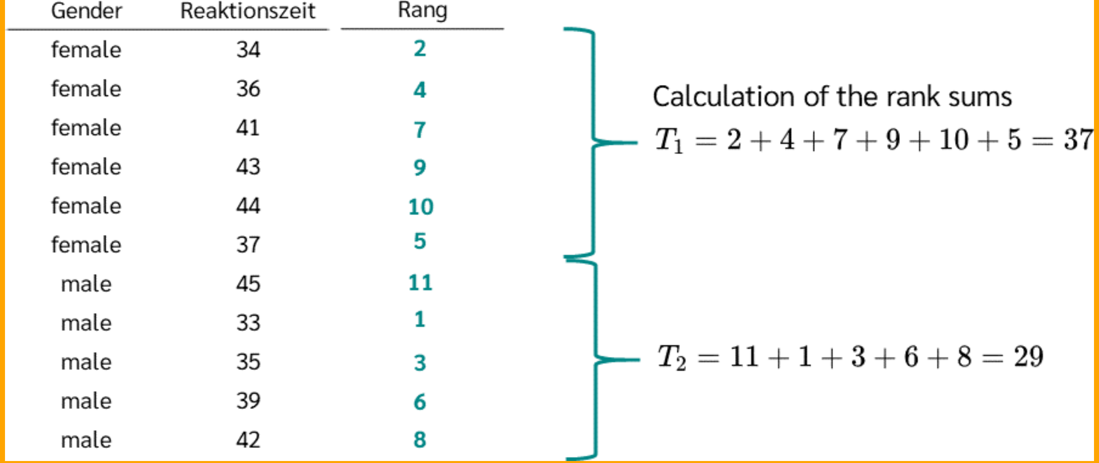
t-Test
Is there a difference in mean?
Mann-Whitney U Test
Is there a difference in the rank sum?
12
1
4
11
5
9
7
8
10
6
2
3
42
36

$$U_1 = n_1 n_2 + \frac{n_1(n_1+1)}{2} - R_1$$

$$U_2 = n_1 n_2 + \frac{n_2(n_2+1)}{2} - R_2$$

$$U = min(U_1, U_2)$$

- If U < Ucrit we reject the null (opposite from *t*-test; always reject if t > tcrit)

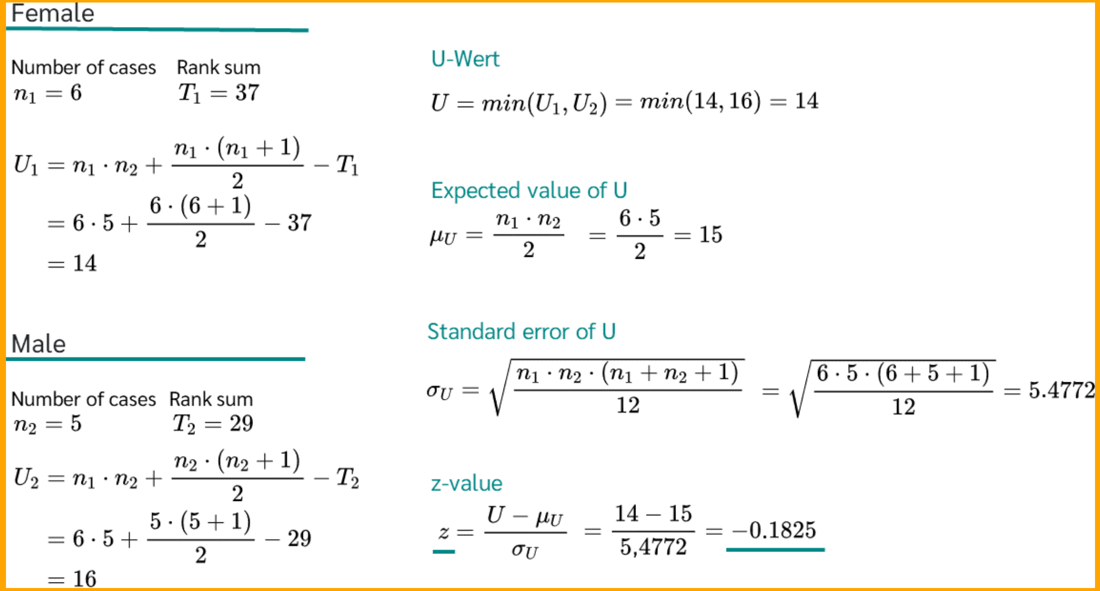


Gender	Reaktionszeit	Rang
female	34	2
female	36	4
female	41	7
female	43	9
female	44	10
female	37	5
male	45	11
male	33	1
male	35	3
male	39	6
male	42	8

Calculation of the rank sums

$$T_1 = 2 + 4 + 7 + 9 + 10 + 5 = 37$$

$$T_2 = 11 + 1 + 3 + 6 + 8 = 29$$



Female

Number of cases $n_1 = 6$

Rank sum $T_1 = 37$

$$U_1 = n_1 \cdot n_2 + \frac{n_1 \cdot (n_1+1)}{2} - T_1$$
$$= 6 \cdot 5 + \frac{6 \cdot (6+1)}{2} - 37$$
$$= 14$$

Male

Number of cases $n_2 = 5$

Rank sum $T_2 = 29$

$$U_2 = n_1 \cdot n_2 + \frac{n_2 \cdot (n_2+1)}{2} - T_2$$
$$= 6 \cdot 5 + \frac{5 \cdot (5+1)}{2} - 29$$
$$= 16$$

U-Wert

$$U = min(U_1, U_2) = min(14, 16) = 14$$

Expected value of U

$$\mu_U = \frac{n_1 \cdot n_2}{2} = \frac{6 \cdot 5}{2} = 15$$

Standard error of U

$$\sigma_U = \sqrt{\frac{n_1 \cdot n_2 \cdot (n_1 + n_2 + 1)}{12}} = \sqrt{\frac{6 \cdot 5 \cdot (6+5+1)}{12}} = 5.4772$$

z-value

$$z = \frac{U - \mu_U}{\sigma_U} = \frac{14 - 15}{5{,}4772} = -0.1825$$

```r
female = c(34,36, 41, 43, 44, 37)
male = c(45, 33, 35, 39, 42)

wilcox.test(male, female, paired = FALSE)
```

```
##
##     Wilcoxon rank sum exact test
##
## data:  male and female
## W = 14, p-value = 0.9307
## alternative hypothesis: true location shift is not equal to 0
```

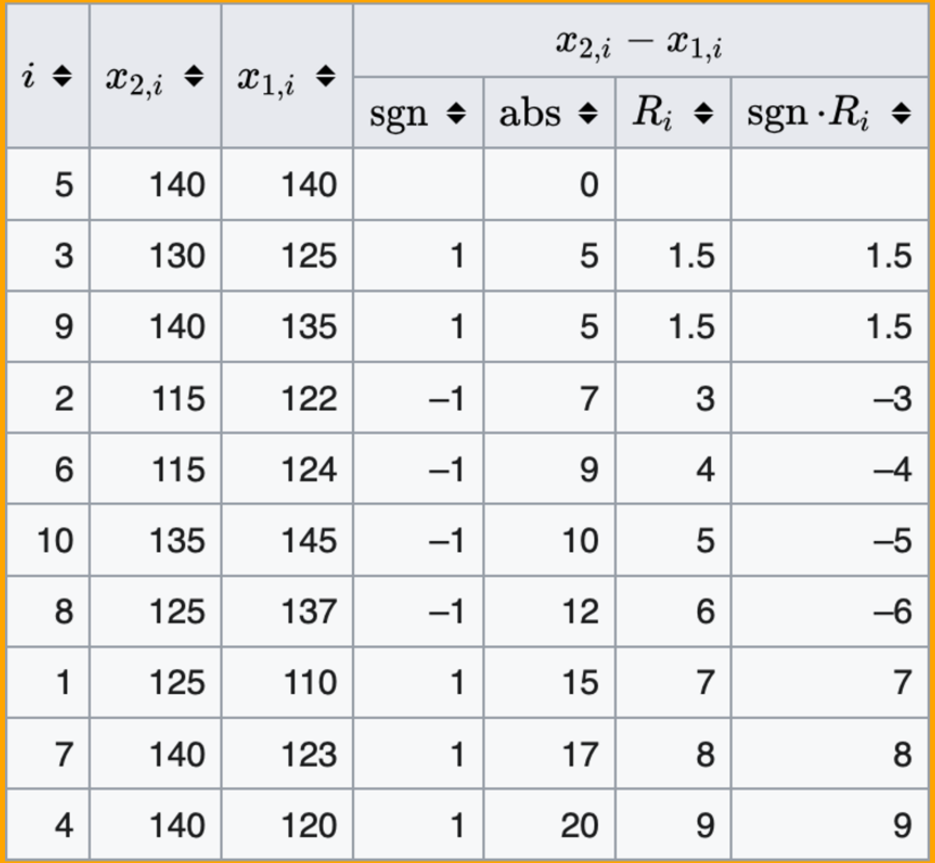
Wilcoxon Signed-Rank (Paired)

$$W = \sum_{i=1}^{N_r} sgn(x_2 - x_1, i)R_i$$

- Where sgn is an indicator variable with if is negative and if is positive
- R = Rank
- W is then the sum of the positive signed ranks
- Exclude pairs where difference equals zero, Nr is the reduced sample size
- If W < Wcrit we reject the null (opposite from *t*-test; always reject if t > tcrit)

Wilcoxon Signed-Rank (Paired)

```
G1 = c(125,115,130,140,140,115,140,125,140,135)
G2 = c(110,122,125,120,140,124,123,137,135,145)
```



i ⬍	$x_{2,i}$ ⬍	$x_{1,i}$ ⬍	$x_{2,i} - x_{1,i}$			
			sgn ⬍	abs ⬍	R_i ⬍	$\text{sgn} \cdot R_i$ ⬍
5	140	140		0		
3	130	125	1	5	1.5	1.5
9	140	135	1	5	1.5	1.5
2	115	122	–1	7	3	–3
6	115	124	–1	9	4	–4
10	135	145	–1	10	5	–5
8	125	137	–1	12	6	–6
1	125	110	1	15	7	7
7	140	123	1	17	8	8
4	140	120	1	20	9	9

Paired, Wilcoxon signed-rank: R

- V = the sum of (W+) ranks

```
G1 = c(125,115,130,140,140,115,140,125,140,135)
G2 = c(110,122,125,120,140,124,123,137,135,145)
wilcox.test(G1, G2,paired =TRUE,alternative ="two.sided")
```

```
##
##      Wilcoxon signed rank test with continuity correction
##
## data:  G1 and G2
## V = 27, p-value = 0.6353
## alternative hypothesis: true location shift is not equal to 0
```

Multiple Comparisons

Multiple Comparisons

- We want our tests to find true positives and true negatives
- Multiple comparisons
    - Type I error (false positive)
    - α-inflation
    - Testing each new pairwise comparison is costly

Bonferroni

- simplest

$$\alpha/m$$

- m = number of comparisons
- Controls for false positives (Type I errors)
- Overly conservative
  - Leads to false negatives (Type II errors)

```r
pvals = c(0.01,0.02,0.04)

p.adjust(pvals,method ="bonferroni", n = length(pvals))
```

```
## [1] 0.03 0.06 0.12
```

Holm-Bonferroni

- Strikes a balance between Type I and Type II errors

1. Sort p-values from smallest to largest

2. Test whether $p < \frac{\alpha}{m+1-k}$

   - m = number of comparisons
   - k = rank

- If so, reject and move to the next

- Typically you report the adjusted p-value. Just multiply your p-value by the adjusted alpha’s denominator

```
pvals = c(0.01,0.02,0.04)
```

```r
pvals = c(0.01,0.02,0.04)

p.adjust(pvals,method ="holm",n = length(pvals))
```

```
## [1] 0.03 0.04 0.04
```

Many Multiple Comparison Corrections

- Tukey - all possible comparisons: TukeyHSD()
- Scheffe
- Dunnett
- Fisher’s LSD (least significant difference)
- Newman-Keuls
- Find what your field does and, more importantly, justify your decisions

Summary

- In this lecture, you've learned:
  - All things *t*-tests
  - The logic of *t*-tests
  - Independent and dependent *t*-tests

Coming Up

- Effect size and power
- Regression