PSYCH 710

Review of Statistical Inference

part 2

Null Hypothesis Significance Testing

Null Hypothesis Significance Testing

- Create null (H0) & alternative (H1) hypotheses
  - mutually exclusive & exhaustive
- Determine if data are unusual assuming H0 is true
- If data are sufficiently unusual, then we reject H0
- If data are not sufficiently unusual, we do not reject H0
  - typically do not "accept H0"
    - the absence of evidence is not evidence of absence
- How do we determine if our data are "sufficiently unusual"?

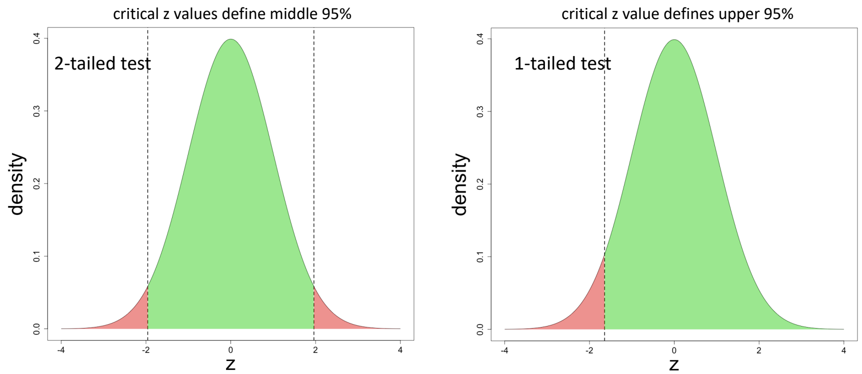
Null Hypothesis Significance Testing (for means)

- How different is observation from expected value when H0 is true?
- Express difference as a standardized distance
  - $z = \frac{\bar{Y} - \mu}{\sigma_{\bar{Y}}} = \frac{\bar{Y} - \mu}{\sigma / \sqrt{n}} \qquad t = \frac{\bar{Y} - \mu}{\hat{\sigma}_{\bar{Y}}} = \frac{\bar{Y} - \mu}{s / \sqrt{n}}$
- Assuming the means are distributed normally
  - z : distributed as standard normal variable
  - t : distributed as t statistic with appropriate degrees-of-freedom
- Calculate probability of getting our z or t (or one more extreme)
  - reject H0 if p value is below our "significance level" (i.e., alpha)

General Strategy

reject H0 if z falls outside critical values of z

$$z = \frac{\bar{Y} - \mu}{\sigma_{\bar{Y}}} = \frac{\bar{Y} - \mu}{\sigma / \sqrt{n}}$$

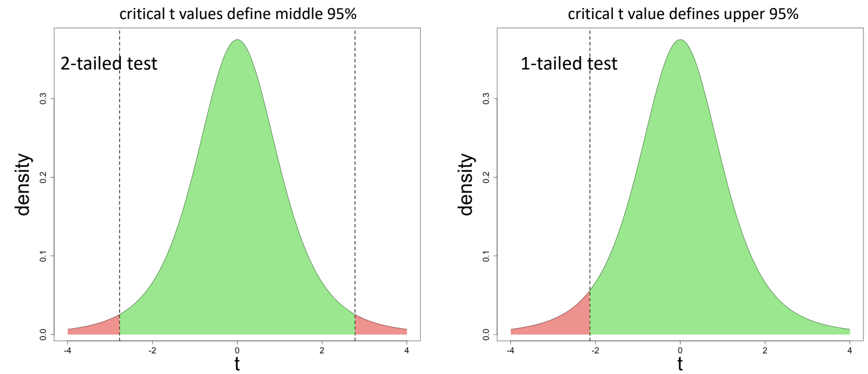


standard normal distribution

General Strategy

reject H0 if t falls outside critical values of t

$$t = \frac{\bar{Y} - \mu}{\hat{\sigma}_{\bar{Y}}} = \frac{\bar{Y} - \mu}{s / \sqrt{n}}$$



t distribution (df=4)

Possible Outcomes of Hypothesis Testing

Table 1: Possible outcomes of hypothesis testing.

decision	H0 is True	H0 is False
reject H0:	Type I ($p = \alpha$)	Correct ($p = 1 - \beta$ =power)
do not reject H0:	Correct ($p = 1 - \alpha$)	Type II error ($p = \beta$)

Type I Error: reject H0 when it is true (alpha)
Type II Error: fail to reject H0 when it is false (beta)
Power = Probability of rejecting false H0 (1-beta)

assumption of normal distribution

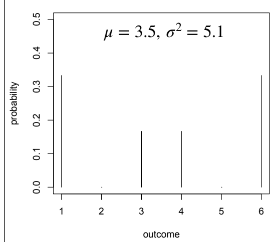
Central Limit Theorem

z & t tests for means

- tests assume that sample means are distributed normally
- if scores are distributed normally, then means are, too
- suppose <u>scores</u> are NOT distributed normally?
- **CENTRAL LIMIT THEOREM**:
  - **<u>irrespective</u>** of how the <u>scores</u> are distributed, the sample <u>means</u> will be distributed normally, provided that the sample size (n) is sufficiently large

Central Limit Theorem (Example)

video at https://www.simplypsychology.org/central-limit-theorem.html

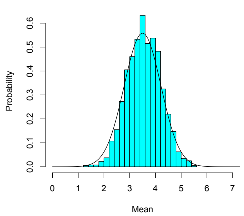


non-normal probability distribution of scores

```
pop.values <- c(1,1,3,4,6,6); # population
set.seed(7321083)
n <- 10 # sample size
B <- 2000 # number of iterations
samp.mean <- rep(0,B)
for(kk in 1:B){
 # randomly sample population of scores:
 cur.sample <- sample(pop.values,size=n,replace=T)
 # calculate and store mean:
 samp.mean[kk] <- mean(cur.sample)
 }
```

Central Limit Theorem (Example)

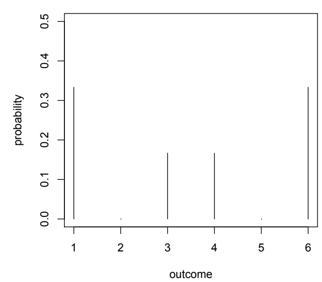
sample size = 10



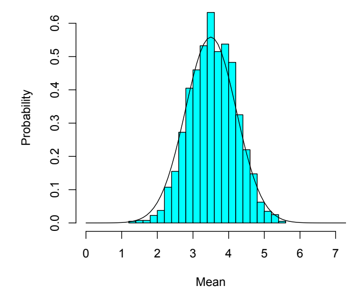
near-normal probability distribution of means

```
quartz()
# plot histogram
require(MASS) # load MASS package
truehist(samp.mean,xlab="Mean",
+          ylab="Probability",xlim=c(0,7))
mtext(text=bquote("Sample Size"~n==.(n)))
# draw predicted normal distribution:
mu <- mean(pop.values) # mean
sigma <- sd(pop.values)/sqrt(n) # SEM
xvals <- seq(0,8,.1)
yvals <- dnorm(xvals,mu,sigma) # normal density
lines(xvals,yvals,type="l") # draw in plot window
```

Central Limit Theorem (Example)



non-normal probability distribution of scores



near-normal probability distribution of means (n=10)

$N(\mu = 3.5,\ \sigma^2 = 5.1/10 = 0.51)$

2 independent samples

Comparing 2 independent means

- Given two independent sample means, $\bar{Y}_a$ & $\bar{Y}_b$
  - Question: are they "significantly different"?
- Define H0 & H1
  - H0: true population difference is zero, $\mu_a - \mu_b = 0$
  - H1: true population difference is not zero, $\mu_a - \mu_b \neq 0$
- Is the observed difference, $\bar{Y}_a - \bar{Y}_b = 0$, unusual when H0 is true?
- <u>Need to know the distribution</u> of $\bar{Y}_d = \bar{Y}_a - \bar{Y}_b$ when H0 is true

Comparing two independent means

- Given 2 populations of scores: means (μ_a & μ_b) variances: (σ_a^2 & σ_b^2)
- Distributions of sample <u>means</u>:
  - $N(\mu_a, \sigma_a^2/n), N(\mu_b, \sigma_b^2/n)$ (via Central Limit Theorem)
- Distribution of <u>difference</u> $\bar{Y}_d = (\bar{Y}_a - \bar{Y}_b)$:
  - mean: $\mu_d = \mu_a - \mu_b$
  - variance: $\dfrac{\sigma_a^2}{n} + \dfrac{\sigma_b^2}{n} - 2 \times \mathrm{COV(A, B)}$ [COV == covariance]
  - COV(A,B) is zero if A & B are independent, so $\sigma_d^2 = \sigma_{\bar{Y}_a}^2 + \sigma_{\bar{Y}_b}^2$
  - $N(\mu_d,\ \sigma_d^2)$ shape is normal (via Central Limit Theorem)

Comparing 2 independent means

- observed $\bar{Y}_d = (\bar{Y}_a - \bar{Y}_b)$ is a random sample from $N(\mu_d,\ \sigma_d^2)$
- is $\bar{Y}_d$ unusual assuming H0 is true?
- express $\bar{Y}_d$ as standardized distance from expected value
- $t = \dfrac{\bar{Y}_d - \mu_d}{\hat{\sigma}_d}$ follows t distribution with df = $n_1 + n_2 - 2$
  - df calculation assumes equal variance in two groups $s_a^2 = s_b^2$
  - when $s_a^2 \neq s_b^2$, t statistic follows t distribution with df < ($n_1 + n_2 - 2$)
- calculate probability of getting our t (or more extreme) when H0 is true
  - reject H0 if p < alpha

Effect of Spatial Uncertainty on Reaction Times

- measured simple reaction time for a spot of light
- stimulus presented at 1 of 4 locations
- each Ss detected spot in all 4 locations and RT was averaged across locations
- locations were either <u>blocked</u> or randomly <u>intermixed</u>
- 2 groups: Blocked vs Mixed was a between-subjects variable
  - RT blocked: $\bar{Y}_1 = 357, s_1 = 83$
  - RT mixed: $\bar{Y}_2 = 397, s_2 = 53$
  - Δ RT: $\bar{Y}_d = -40, s_d = 22.02$
- Assuming $\mu_d = 0$, is our observed $\bar{Y}_d = -45$ unusual?
- Use null hypothesis significance testing:
  - H0: $\mu_d = 0$, H1: $\mu_d \neq 0$, $t = -40/22.02 = -1.82$

Spot Detection Reaction Times

2-tailed test, var.equal = FALSE

```
> alpha <- .05
> t.test(x=blocked,y=mixed,
+        paired=FALSE,
+        var.equal=FALSE,
+        alternative="two.sided",
+        conf.level=1-alpha)

Welch Two Sample t-test
t = -1.8165, df = 32.286, p-value = 0.079
H1: true difference ≠ 0
95% CI:
 -84.84   4.84
sample estimates:
mean of x mean of y
      357       397
```

Notice df ≠ $n_1 + n_2 - 2$

Spot Detection Reaction Times

2-tailed test, var.equal = TRUE

```
> alpha <- .05
> t.test(x=blocked,y=mixed,
+        paired=FALSE,
+        var.equal=TRUE,
+        alternative="two.sided",
+        conf.level=1-alpha)

Two Sample t-test
t = -1.8165, df = 38, p-value = 0.077
H1: true difference ≠ 0
95% CI:
 -84.58   4.58
sample estimates:
mean of x mean of y
      357       397
```

Notice df = $n_1 + n_2 - 2$

Spot Detection Reaction Times

- We do not reject H0 $\mu_d = 0$
- However, we believe spatial uncertainty <u>increases</u> RT
  - $RT_{blocked} < RT_{mixed}$
- So a 1-tailed test is appropriate
  - H0: $\mu_d \geq 0$, H1: $\mu_d < 0$
- Reject H0 if $(\bar{Y}_{blocked} - \bar{Y}_{mixed})$ is unusually negative

Spot Detection Reaction Times

1-tailed test, var.equal = FALSE

```
> alpha <- .05
> t.test(x=blocked,y=mixed,
+        paired=FALSE,
+        alternative="less",
+        conf.level=1-alpha)

Welch Two Sample t-test
t = -1.82, df = 32.3, p-value = 0.039
H1: true difference < 0
95% CI:
    -Inf -2.7096
sample estimates:
mean of x mean of y
      357       397
```

- H0: $\mu_d \geq 0$, H1: $\mu_d < 0$
- alpha = 0.05
- reject H0

What affects our decision about H0?

Possible Outcomes of Hypothesis Testing

Table 1: Possible outcomes of hypothesis testing.

decision	H0 is True	H0 is False
reject H0:	Type I ($p = \alpha$)	Correct ($p = 1 - \beta$ =power)
do not reject H0:	Correct ($p = 1 - \alpha$)	Type II error ($p = \beta$)

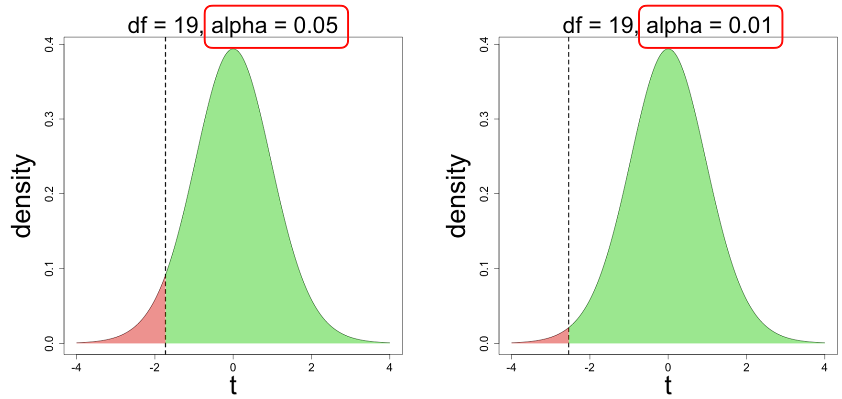
Type I Error: reject H0 when it is true (alpha)
Type II Error: fail to reject H0 when it is false (beta)
Power = Probability of rejecting false H0 (1-beta)

What factors determine the Type I error rate?

Table 1: Possible outcomes of hypothesis testing.

decision	H0 is True	H0 is False
reject H0:	Type I ($p = \alpha$)	Correct ($p = 1 - \beta$ =power)
do not reject H0:	Correct ($p = 1 - \alpha$)	Type II error ($p = \beta$)

Type I error rate is determined by alpha



t distribution (df=19)

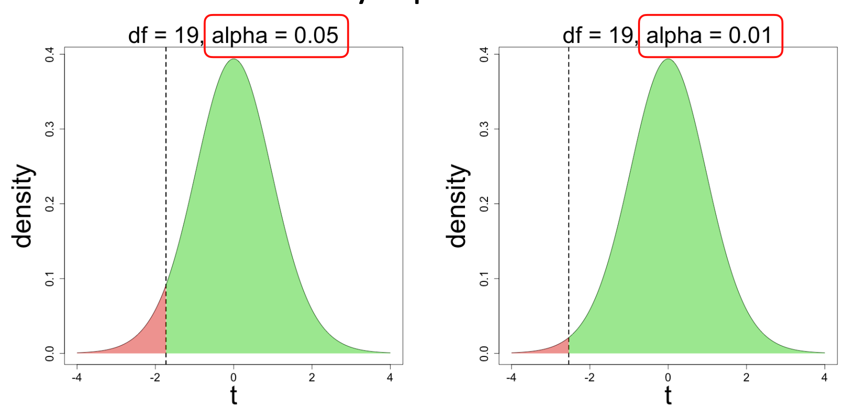
What factors determine the Power & Type II error rate?

Table 1: Possible outcomes of hypothesis testing.

decision	H0 is True	H0 is False
reject H0:	Type I ($p = \alpha$)	Correct ($p = 1 - \beta =$power)
do not reject H0:	Correct ($p = 1 - \alpha$)	Type II error ($p = \beta$)

What determines the Type II error rate?
What factors influence statistical power?

Power is influenced by alpha



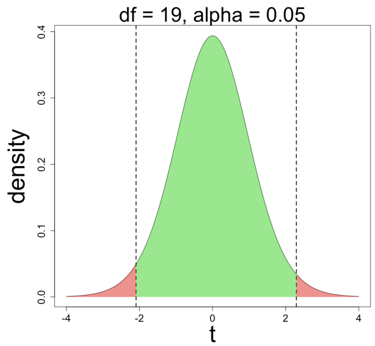
When H0 is false, red is “correct rejection” and green is Type II error
Lower alpha makes is harder to reject a false H0.

alpha & power

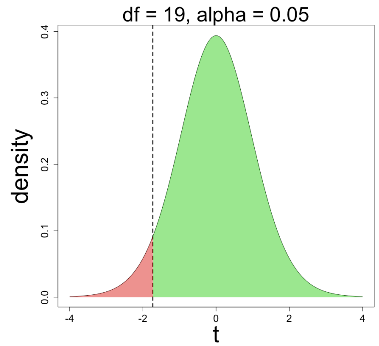
- using alpha of .001 instead of .05 reduces Type I error
- but also increases Type II error…
  - makes it harder to reject false H0
  - and therefore reduces power

Power is greater for 1-tailed tests

H0: $\mu_1 - \mu 2 = 0$ H1: $\mu_1 - \mu_2 \neq 0$



H0: $\mu_1 - \mu 2 \geq 0$ H1: $\mu_1 - \mu_2 < 0$



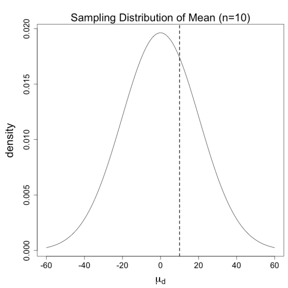
Easier to reject false H0 using focused, 1-tailed test

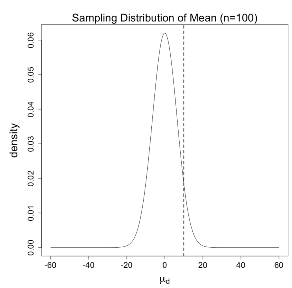
sample size

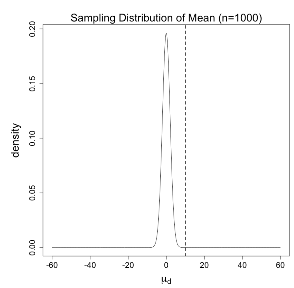
effect of sample size on power

$$\hat{\sigma}_{\bar{Y}_d} = \frac{s_d}{\sqrt{n}}$$

- increasing sample size decreases standard error of mean
- consequently, it becomes easier to reject a false H0 (i.e., increased power)







When H0 $\mu_d = 0$ is false, $\bar{Y}_d \neq 0$ becomes more unusual as n increases

Simple Reaction Times

- does visual processing speed differ across wavelength?
- measure simple reaction time for 2 wavelengths
- n=10; calculated RT difference for each S
  - RT w1: $\bar{Y}_1 = 357, s_1 = 83$
  - RT w2: $\bar{Y}_2 = 367, s_2 = 83$
  - Δ RT: $\bar{Y}_d = 10, s_d = 64.29$
- Assuming $\mu_d = 0$, is our observed $\bar{Y}_d = 10$ unusual?
- Use null hypothesis significance testing:
  - H0: $\mu_d = 0$, H1: $\mu_d \neq 0$

Simple Reaction Times

data frame organization

difference scores

```
> rt.df1
   sID  w1  w2     dRT
1   s1 291 304  12.450
2   s2 411 365 -46.205
3   s3 414 396 -17.631
4   s4 354 355   0.874
5   s5 242 393 150.332
6   s6 282 322  39.258
7   s7 288 250 -37.723
8   s8 450 531  81.118
9   s9 341 295 -46.437
10 s10 496 460 -36.036
```

Simple Reaction Times

difference scores

```
> t.test(rt.df1$dRT,mu=0,
         alternative="two.sided")

One Sample t-test
t = 0.5, df = 9, p-value = 0.6
H1: true mean ≠ 0
95% CI:
 -36  56
sample estimates:
mean of x
      10
```

paired samples

```
> t.test(rt.df1$w2,rt.df1$w1,mu=0,
+        paired=T,
+        alternative="two.sided")

Paired t-test
t = 0.5, df = 9, p-value = 0.6
H1: true mean difference ≠ 0
95% CI:
 -36  56
sample estimates:
mean difference
            10
```

Simple Reaction Times (n=100)

- repeat experiment with larger sample
- n=100; calculated RT difference for each S
  - RT w1: $\bar{Y}_1 = 357, s_1 = 83$
  - RT w2: $\bar{Y}_2 = 367, s_2 = 83$
  - Δ RT: $\bar{Y}_d = 10, s_d = 64.29$
- Assuming $\mu_d = 0$, is our observed $\bar{Y}$ unusual?
- Use null hypothesis significance testing:
  - H0: $\mu_d = 0$, H1: $\mu_d \neq 0$

same values as before

Simple Reaction Times

difference scores

n = 100

```
> t.test(rt.df1$dRT,mu=0,
+        alternative="two.sided")

One Sample t-test
t = 1.5554, df = 99, p-value = 0.123
H1: true mean ≠ 0
95% CI:
 -2.756833 22.756833
sample estimates:
mean of x
      10
```

n = 10

```
> t.test(rt.df1$dRT,mu=0,
         alternative="two.sided")

One Sample t-test
t = 0.5, df = 9, p-value = 0.6
H1: true mean ≠ 0
95% CI:
 -36  56
sample estimates:
mean of x
      10
```

Simple Reaction Times (n=1000)

- repeat experiment with very large sample
- n=1000; calculated RT difference for each S
  - RT w1: $\bar{Y}_1 = 357, s_1 = 83$
  - RT w2: $\bar{Y}_2 = 367, s_2 = 83$
  - Δ RT: $\bar{Y}_d = 10, s_d = 64.29$

same values as before

- Assuming $\mu_d = 0$, is our observed $\bar{Y}$ unusual?
- Use null hypothesis significance testing:
  - H0: $\mu_d = 0$, H1: $\mu_d \neq 0$

Simple Reaction Times

difference scores & paired samples

n = 1000

```
> t.test(rt.df3$dRT,mu=0,
+        alternative="two.sided")

One Sample t-test
t = 4.9187, df = 999, p-value < 0.001
H1: true mean ≠ 0
95% CI:
 6.01041 13.98959
sample estimates:
mean of x
      10
```

n = 10

```
> t.test(rt.df1$dRT,mu=0,
         alternative="two.sided")

One Sample t-test
t = 0.5, df = 9, p-value = 0.6
H1: true mean ≠ 0
95% CI:
 -36  56
sample estimates:
mean of x
      10
```

we now reject H0… increasing sample size increased our power

Sample Size & Statistical Power

Power = p(correctly rejecting a false H0) = 1 - (Type II Error Rate)

```
> power.t.test(n=10,
+        delta=10,
+        sd=64.29,
+        type="one.sample",
+        alternative="two.sided",
+        power=NULL)

1-sample power calculation
n = 10
delta = 10
sd = 64.29152
sig.level = 0.05
power = 0.06450793
alternative = two.sided
```

```
> power.t.test(n=100,
+        delta=10,
+        sd=64.29,
+        type="one.sample",
+        alternative="two.sided",
+        power=NULL)

1-sample power calculation
n = 100
delta = 10
sd = 64.29152
sig.level = 0.05
power = 0.337378
alternative = two.sided
```

```
> power.t.test(n=1000,
+        delta=10,
+        sd=64.29,
+        type="one.sample",
+        alternative="two.sided",
+        power=NULL)

1-sample power calculation
n = 1000
delta = 10
sd = 64.29152
sig.level = 0.05
power = 0.9984314
alternative = two.sided
```

effect size

Influence of effect size on statistical power

- big effects are easier to detect than small effects
- two-sample case:
  - H0: $\mu_1 - \mu_2 = 0$, H1: $\mu_1 - \mu_2 \neq 0$
  - easier to reject H0 when $(\mu_1 - \mu_2 \ll 0)$ or $(\mu_1 - \mu_2 \gg 0)$
- one-sample case:
  - H0: $\mu_d = 0$, H1: $\mu_d \neq 0$
  - easier to reject H0 when $(\mu_d \ll 0)$ or $(\mu_d \gg 0)$

Effect Size & Statistical Power

Easier to reject false H0 $\mu_d = 0$ when true $\mu_d \gg 0$

```
> power.t.test(n=20,
+         delta=20,
+         sd=100,
+         type="one.sample",
+         alternative="two.sided",
+         power=NULL)

1-sample power calculation
n = 20
delta = 20
sd = 100
sig.level = 0.05
power = 0.133
alternative = two.sided
```

```
> power.t.test(n=20,
+         delta=40,
+         sd=100,
+         type="one.sample",
+         alternative="two.sided",
+         power=NULL)

1-sample power calculation
n = 20
delta = 40
sd = 100
sig.level = 0.05
power = 0.397
alternative = two.sided
```

```
> power.t.test(n=sampleSize,
+         delta=80,
+         sd=100,
+         type="one.sample",
+         alternative="two.sided",
+         power=NULL)

1-sample power calculation
n = 20
delta = 80
sd = 100
sig.level = 0.05
power = 0.924
alternative = two.sided
```

Effect Size

Calculating and reporting effect sizes to facilitate cumulative science: a practical primer for t-tests and ANOVAs. Lakens D. Front Psychol. 2013 Nov 26;4:863. doi: 10.3389/fpsyg.2013.00863.

- 2 types of effect size measures:
  - d: standardized differences (distances) between means
  - r^2: measures of association
    - % variance accounted for by grouping variable
- there are MANY varieties of “d” and “r” measures:
  - we will consider just 1 variety of d here…
  - (we will consider more as we go through the term)
- ideally, measures should be invariant to sample size

Cohen’s d for RT (wavelength) study

cohens_d in effectsize library

$$\hat{d} = \frac{0 - \hat{\mu}_d}{\hat{\sigma}_d} = \frac{0 - \bar{Y}_d}{s_d}$$

```
> cohens_d(x=rt.df1$dRT)
Cohen's d |        95% CI
-------------------------
0.16      | [-0.47, 0.78]
```

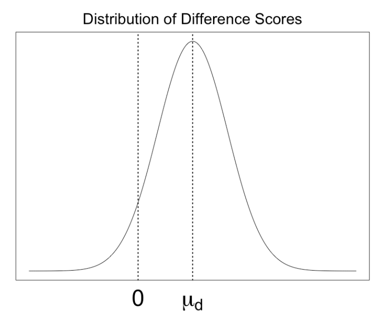
n = 10

```
> cohens_d(x=rt.df2$dRT)
Cohen's d |        95% CI
-------------------------
0.16      | [-0.04, 0.35]
```

n = 100

```
> cohens_d(x=rt.df3$dRT)
Cohen's d |       95% CI
-------------------------
0.16      | [0.09, 0.22]
```

n = 1000



factors affecting decision outcome

- Type I error: (alpha, significance level, critical p value)
- Power & Type II error:
  - alpha (Type I error)
  - general vs. focused statistical tests
    - 2-tailed vs 1-tailed t tests
  - sample size
  - effect size

equivalence tests

interpreting non-significant t tests

Interpreting non-significant t tests

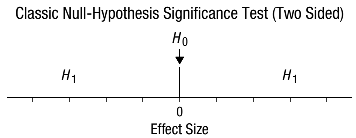
- An experiment compares drugs A & B
- Experimenter wants to know if 2 drugs yield same outcome
  - H0: $\mu_A - \mu_B = 0$ H1: $\mu_A - \mu_B \neq 0$
- Conduct a significance test that is not significant (i.e., p>0.05)
- Can we conclude that the two drugs are the same?

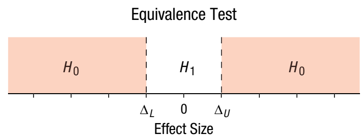
Interpreting Non-significant 2-sided Tests

- Can we conclude that the two drugs are the same?
- No. Why not?
- Failure to attain p<0.05 may be due to low power…
  - small sample size and/or noisy outcome measure
  - absence of evidence is not evidence of absence
- Only conclude we "do not reject H0"
- Can we make a stronger statement?
  - e.g., The two drugs have "equivalent" outcomes

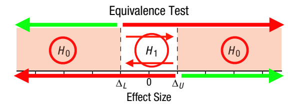
Equivalence Tests

- Standard NHST
  - H0: $(\mu_A - \mu_B) = 0$; H1: $\mu_A - \mu_B \neq 0$
- Equivalence tests reverse H0 & H1:
  - H0: there is a difference $(\mu_A - \mu_B) \neq 0$
  - H1: there is no difference $(\mu_A - \mu_B) = 0$
  - try to reject H0 with two 1-sided tests





Equivalence Tests



- Smallest Effect Size of Interest SESOI (Δ_L & Δ_U)
  - Δ_L & Δ_U are lower & upper bounds of equivalence region (i.e., $\mu_A \approx \mu_B$)
  - H0: $(\mu_A - \mu_B) \leq \Delta_L$ **OR** $(\mu_A - \mu_B) \geq \Delta_U$ [i.e., two means are not equivalent]
  - H1: $(\mu_A - \mu_B) > \Delta_L$ **AND** $(\mu_A - \mu_B) < \Delta_U$ [i.e., two means are equivalent]
- evaluate H0 with two 1-tailed t-tests:
  - H0$_L$: $(\mu_A - \mu_B) \leq \Delta_L$ H1$_L$: $(\mu_A - \mu_B) > \Delta_L$ [is difference > Δ_L?]
  - H0$_U$: $(\mu_A - \mu_B) \geq \Delta_U$ H1$_U$: $(\mu_A - \mu_B) < \Delta_U$ [is difference < Δ_U?]
- if both 1-tailed tests are significant, then
  - we accept H1$_L$: $(\mu_A - \mu_B) > \Delta_L$ **AND** H1$_U$: $(\mu_A - \mu_B) < \Delta_U$
  - difference is within ± SESOI
  - e.g., two groups are "equivalent"

Two One-Sided Test Method

TOST

TOST procedure

- Equivalence test hypotheses
  - define lower bound (LB) & upper bound (UB)
  - H0: $(\mu_d \leq \mathrm{LB})$ OR $(\mu_d \geq \mathrm{UB})$ [not equivalent]
  - H1: $(\mu_d > \mathrm{LB})$ AND $(\mu_d < \mathrm{UB})$ [equivalent]
- Evaluate H0 with two 1-tailed tests
  - lower bound: H0$_L$: $(\mu_d \leq \mathrm{LB})$ H1$_L$: $(\mu_d > \mathrm{LB})$
  - upper bound: H0$_U$: $(\mu_d \geq \mathrm{UB})$ H1$_U$: $(\mu_d < \mathrm{UB})$
- If both 1-tailed tests are significant $(p < \alpha)$
  - then reject H0 (not equivalent) $(p < \alpha)$

Equivalence Test Example

Two One-tailed Significance Tests (TOST)

```
> N <- 2500
> mu1 <- 100.5
> mu2 <- 100
> stdev <- 7
> set.seed(20912)
> x <- rnorm(N,mu1,stdev)
> y <- rnorm(N,mu2,stdev)

> # Is observed difference between
> # upper & lower bounds?
> UPPER.BOUND <- 1 # [ Δ_U ]
> LOWER.BOUND <- -1 # [ Δ_L ]
> # H0: (diff < -1) OR (diff > 1)
> # H1: (diff > -1) AND (diff < 1)
```

$HO_L : \mu_D \leq \Delta_L$

```
> # is mean > lower bound?
> t.test(x,y,mu=LOWER.BOUND,
+         alternative="greater")
t = 8.002, df = 4997, p-value = < .0001
H1L: true difference is > -1
95% CI: 0.2502728       Inf
sample estimates:
mean of x mean of y
 100.5765  100.0026
```

Equivalence Test Example

Two One-tailed Significance Tests (TOST)

```
> N <- 2500
> mu1 <- 100.5
> mu2 <- 100
> stdev <- 7
> set.seed(20912)
> x <- rnorm(N,mu1,stdev)
> y <- rnorm(N,mu2,stdev)

> # Is observed difference between
> # upper & lower bounds?
> UPPER.BOUND <- 1 # [ Δ_U ]
> LOWER.BOUND <- -1 # [ Δ_L ]
> # H0: (diff < -1) OR (diff > 1)
> # H1: (diff > -1) AND (diff < 1)
```

$HO_U : \mu_D \geq \Delta_U$

```
> # is mean < upper bound?
> t.test(x,y,mu=UPPER.BOUND,
+         alternative="less")
t = -2.167, df = 4997, p-value = 0.015
H1U: true difference < 1
95% CI:  -Inf 0.8974459
sample estimates:
mean of x mean of y
 100.5765  100.0026
```

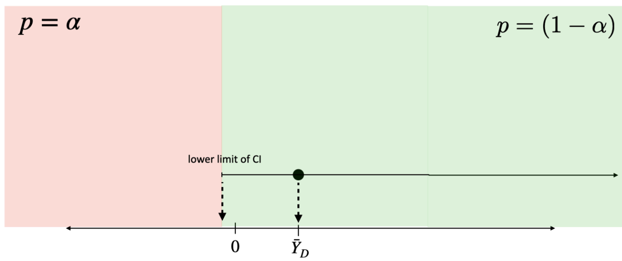
Two-sided Confidence Interval Method

Using 1 - (2 x alpha) two-sided CI

Equivalence Testing using 2-sided CI

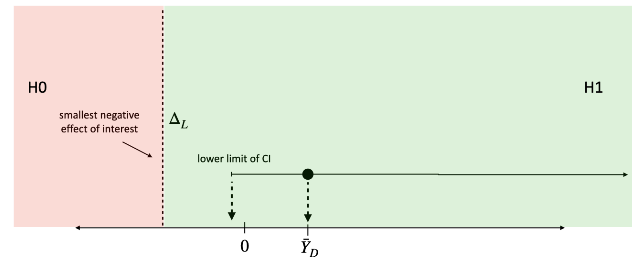
- Equivalence test using alpha = 0.05
  - H0: means are not equivalent
  - H1: means are equivalent
- Evaluate H0 using 2-sided t test
  - inspect 1 - (2 x 0.05) = 90% Confidence Interval
  - reject H0 (p < .05) if CI falls within equivalence zone

Testing Lower Bound of Equivalence Region
(is difference greater than Δ_L?)



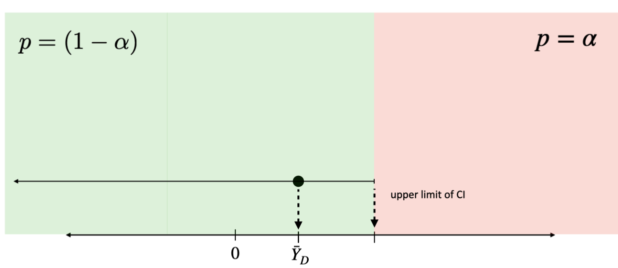
- Let alpha be set for a 1-sided test (alpha = 0.05)
- Probability of interval containing true population difference = 1-0.05 = 0.95
- Probability of true difference being less than lower limit of CI is 0.05

Testing Lower Bound of Equivalence Region
(is difference greater than Δ_L?)



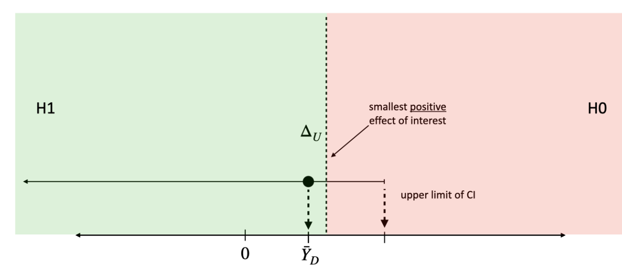
- Test of Lower Bound: $H0_L$: $(\mu_1 - \mu_2) \leq \Delta_L$ $H1_L$: $(\mu_1 - \mu_2) > \Delta_L$
- If CI does not include Δ_L, then reject $H0_L$ in favor of $H1_L$ (i.e., true difference between means is <u>greater than</u> Δ_L)
- In this case we reject $H0_L$ ($p < .05$)

Testing Upper-bound of Equivalence Region
(is difference less than Δ_U?)



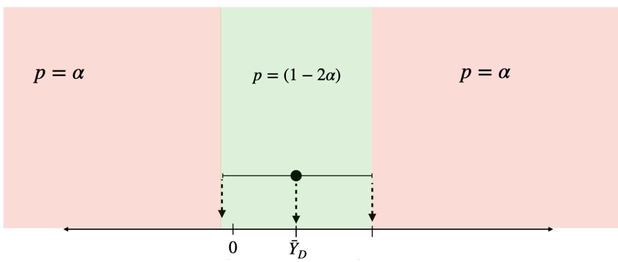
- Let alpha be set for a 1-sided test (alpha = 0.05)
- Probability of interval containing true population difference = 1-0.05 = 0.95
- Probability of true difference being greater than upper limit of CI is 0.05

Testing Upper-bound of Equivalence Region
(is difference less than Δ_U?)



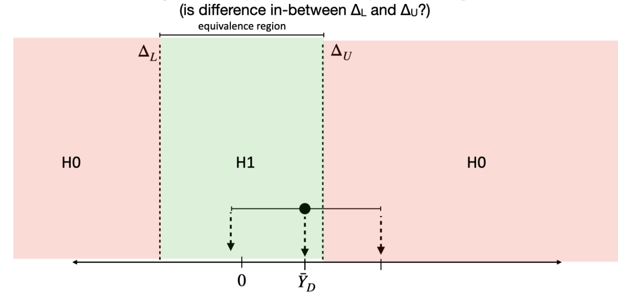
- Test of Upper Bound: $H0_U$: $(\mu_1 - \mu_2) \geq \Delta_U$ $H1_U$: $(\mu_1 - \mu_2) < \Delta_U$
- If CI does not include Δ_U then reject $H0_U$ in favor of $H1_U$ (i.e., true difference between means is <u>less than</u> Δ_U)
- In this case we do NOT reject $H0_U$

Testing Lower & Upper bounds of Equivalence Region
(is difference in-between Δ_L and Δ_U?)



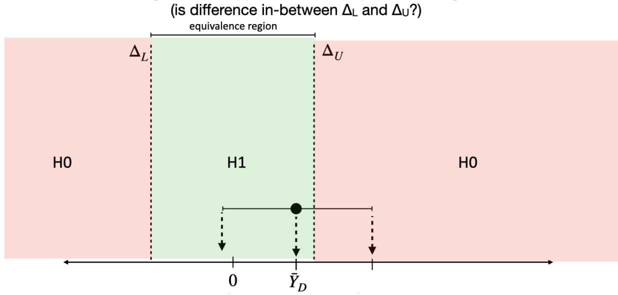
- Let alpha be set for a 2-sided test (alpha = 2 x 0.05)
- Probability of CI containing true population difference = (1 - 2 x alpha) = 0.90 (i.e., <u>probability of true difference lying below OR above 2-sided CI = 0.10</u>)

Testing Lower & Upper bounds of Equivalence Region
(is difference in-between Δ_L and Δ_U?)



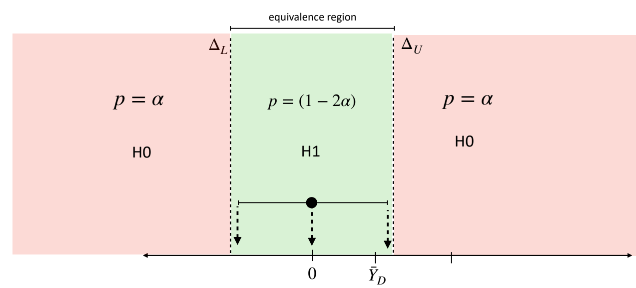
- Equivalence Test Hypothesis:
  - H0: $\{(\mu_A - \mu_B) \leq \Delta_L$ **<u>OR</u>** $(\mu_A - \mu_B) \geq \Delta_U\}$ [not equivalent]
  - H1: $\{(\mu_A - \mu_B) > \Delta_L$ **<u>AND</u>** $(\mu_A - \mu_B) < \Delta_U\}$ [equivalent]

Testing Lower & Upper bounds of Equivalence Region
(is difference in-between Δ_L and Δ_U?)



- If CI falls <u>within</u> the equivalence region, then $\bar{Y}_d$ is unusually small assuming H0 is true.
  - we reject H0 in favor of H1 — i.e., the two means are “equivalent” — p < .05
- In this case we do NOT reject H0 (i.e., that the two means are not equivalent)
  - because we cannot reject hypothesis that $(\mu_A - \mu_B) \geq \Delta_U$

Testing Lower & Upper bounds of Equivalence Region
(is difference in-between Δ_L and Δ_U?)



- Equivalence Test Hypothesis
  - H0: $\{(\mu_A - \mu_B) \leq \Delta_L$ **<u>OR</u>** $(\mu_A - \mu_B) \geq \Delta_U\}$ H1: $\{(\mu_A - \mu_B) > \Delta_L$ **<u>AND</u>** $(\mu_A - \mu_B) < \Delta_U\}$
- If 90% CI is within the equivalence region, then $\bar{Y}_d$ is unusually small assuming H0 is true.
  - we would reject H0 in favor of H1 — i.e., the two means are “equivalent” — p < .05
- In this case we DO reject H0 in favor of H1 [means are equivalent]

Equivalence Test Example

using two-sided 90% confidence interval

```
> UPPER.BOUND <- 1 [ Δ_U ]
> LOWER.BOUND <- -1 [ Δ_L ]
```

1-sided test results

```
upper bound: 95% CI = [-Inf, 0.897]
lower bound: 95% CI = [0.2502, Inf]
```

H0: $\{(\mu_x - \mu_y) \leq \Delta_L$ **OR** $(\mu_A - \mu_B) \geq \Delta_U\}$

H1: $\{(\mu_A - \mu_B) > \Delta_L$ **AND** $(\mu_A - \mu_B) < \Delta_U\}$

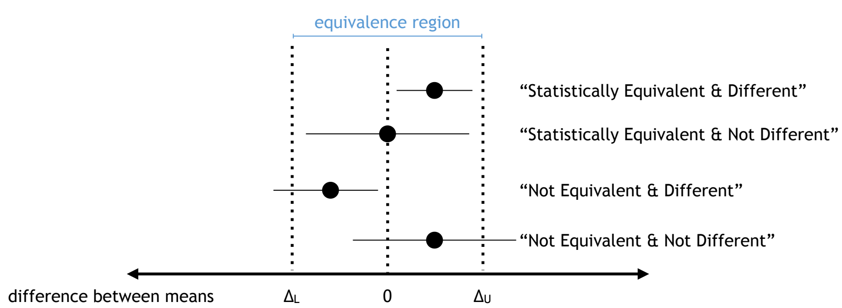
```
> alpha <- 0.05 # equivalence test alpha
> t.test(x,y,mu=0,
+        alternative="two.sided",
+        conf.level=(1-2*alpha) )
t = 2.9176, df = 4997, p-value = 0.003543
H1: true difference ≠ 0
90% CI: 0.2502728 0.8974459
sample estimates:
mean of x mean of y
 100.5765  100.0026

> (0.2502728 > LOWER.BOUND)
[1] TRUE
> (0.8974459 < UPPER.BOUND)
[1] TRUE

> # 90% CI falls within equivalence zone
> # reject H0 in favour of H1 (p < 0.05)
```

Null Hypothesis vs Equivalence Tests

Four possible outcomes when evaluating difference between 2 group means

		Equivalence Test	
		Equivalent	Not Equivalent
Null Hypothesis Significance Test	Not Different	+	?
	Different	?	+

NHST vs Equivalence Tests (4 outcomes)



Lakens, D. Equivalence tests: A practical primer for *t* tests, correlations, and meta-analyses. Social Psychological & Personality Science, 8(4), 355-362.

What do p values mean?

What does a significant p-value mean?

- A significant p-value indicates that the result is unusual
  - **assuming H0 is true** (and assumptions are correct)
- That is ALL that it means

decision	H0 is True	H0 is False
reject H0:	Type I ($p = \alpha$)	Correct ($p = 1 - \beta =$power)
do not reject H0:	Correct ($p = 1 - \alpha$)	Type II error ($p = \beta$)

What does a significant p-value mean?

- A significant p-value indicates that the result is unusual
  - **assuming H0 is true** (and assumptions are correct)
- That is ALL that it means
  - (1-p) is not equal to the probability of replicating result…
    - often, p(replication) << (1-p)

What does a significant p-value mean?

- A significant p-value indicates that the result is unusual
  - assuming H0 is true and assumptions are correct
- That is ALL it means
  - (1-p) is not equal to the probability of replicating result
  - p is not equal to the probability that H0 is TRUE…
    - p is not equal to the probability that the result is due to chance

alpha ≠ p(H0 is TRUE)

Table 1: Possible outcomes of hypothesis testing.

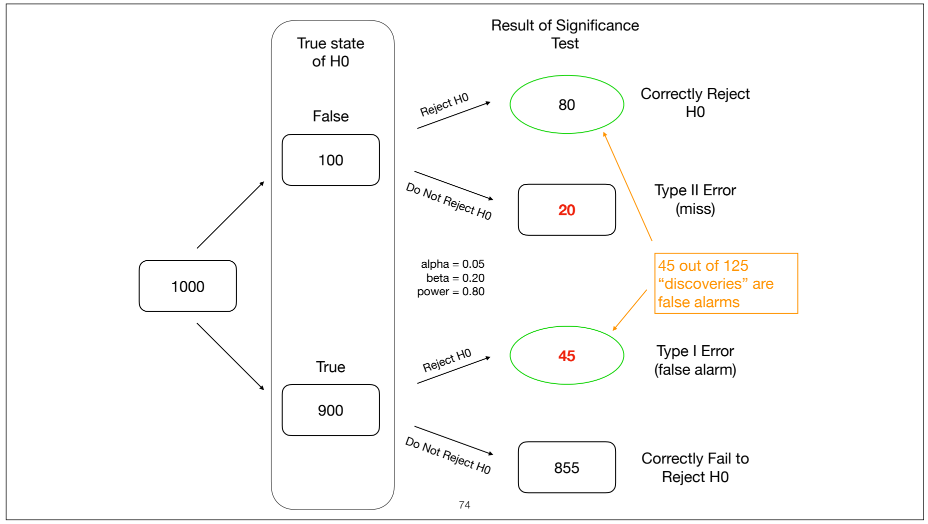
decision	H0 is True	H0 is False
reject H0:	Type I ($p = \alpha$)	Correct ($p = 1 - \beta =$power)
do not reject H0:	Correct ($p = 1 - \alpha$)	Type II error ($p = \beta$)

alpha = probability of making Type I error given that H0 is True
alpha ≠ probability that H0 is True

What does a significant p-value mean?

- A significant p-value indicates that the result is unusual
  - assuming H0 is true and assumptions are correct
- That is ALL it means
  - (1-p) is not equal to the probability of replicating result
  - p is not equal to the probability that H0 is TRUE…
    - p is not equal to the probability that the result is due to chance
  - p is not equal to the probability of making a false discovery…

73



What does a significant p-value mean?

- A significant p-value indicates that the result is unusual
  - assuming null hypothesis is true and assumptions are correct
- That is ALL it means
  - (1-p) is not equal to the probability of replicating the result
  - p is not equal to the probability that H0 is TRUE…
    - p is not equal to the probability that the result is due to chance
  - p is not equal to the probability of making a false discovery
  - p is not a measure of the strength of evidence in favour of H0
    - when H0 is true, all p values are EQUALLY likely (!)

75

What do p-values mean?

$P(\text{data}|H0) \neq P(H0|\text{data})$

- A p value is the probability of obtaining a result that is at least as extreme as observed result when H0 is true
  - it measures compatibility of our data with a specified model
- p values are statements about the data, not the hypotheses
- Used properly, p values control Type I error [N.B. this is good!]
  - When H0 is TRUE, in the long run Type I error rate equals alpha
  - But alpha does not equal the False Discovery Rate
    - FDR depends on alpha, statistical power, and p(H0 is True)

76

Lykken DT, Psychol Bulletin, 1968

"Statistical significance is perhaps the least important attribute of a good experiment; it is never a sufficient condition for claiming that a theory has been usefully corroborated, that a meaningful empirical fact has been established, or that an experimental report ought to be published."

Lykken DT, Psychol Bulletin, 1968

"Statistical significance is perhaps the least important attribute of a good experiment; it is never a sufficient condition for claiming that a theory has been usefully corroborated, that a meaningful empirical fact has been established, or that an experimental report ought to be published."

- Some important attributes are
  - Having a clear, logical framework for formulating the research question and deriving predictions
  - Using a good experimental design
  - Appropriate/interesting manipulations of relevant independent variables
  - Having a "good" sample of participants
  - Using sensitive and reliable dependent measures
  - and so on…

fin