t tests

- useful for
  - comparing mean of 1 sample to some expected value
  - comparing means of 2 samples
- statistic “t”
  - sampling distribution is the t distribution
  - unimodal, symmetric about mean
  - shape governed by parameter degrees-of-freedom (df)
  - when NULL hypothesis is true, mean of t will be zero

R code

```r
# set seed
set.seed(555410)
mu <- 0 # population mean
n <- 20 # sample size
stddev <- 1 # population sd
R <- 10000

t.val <- rep(0,R)
p.val <- rep(0,R)
for(kk in 1:R){
  the.sample <- rnorm(n,mu,stddev)
  t.results <- t.test(the.sample)
  t.val[kk] <- t.results$statistic
  p.val[kk] <- t.results$p.value
}
```

• t-tests
• Effect Size
• Equivalence Tests
• Consequences of Low Power
### t test for single mean

**R code**

```r
set.seed(555410)
mu <- 3.5 # population mean
n <- 20 # sample size
stdev <- 1 # population sd
R <- 10000
t.val <- rep(0,R)
p.val <- rep(0,R)
for(kk in 1:R){
  the.sample <- rnorm(n,mu,stdev)
  t.val[kk] <- t.test(the.sample)
  p.val[kk] <- t.val$ statistic
}

# t test evaluates null hypothesis that mu=0

null hypothesis is false mu=0.5
```

### Comparing 2 means

- **Given** 2 populations of scores
  - means: \(^\mu_A\) and \(^\mu_B\)
  - variances: \(^\sigma^2_A\) and \(^\sigma^2_B\)
- **Distributions** of sample means:
  - means: \(^\bar{y}_A\) and \(^\bar{y}_B\)
  - variances: \(\frac{\sigma^2_A}{n_a} = \frac{\sigma^2_A}{n_1} = \frac{\sigma^2_B}{n_b}\)
- **Shape** normal (via Central Limit Theorem)
- **Distribution of difference** between sample means will be normal:
  - mean: \(^\mu_0 = \mu_A - \mu_B\)
  - variance: \(\frac{\sigma^2_A}{n_1} + \frac{\sigma^2_B}{n_2}\)

**R code**

```r
set.seed(555410)
m1 <- 100 # population 1 mean
m2 <- 100 # population 2 mean
n1 <- 25 # sample size
n2 <- 10 # sample size
stdev1 <- 10 # population 1 sd
stdev2 <- 10 # population 2 sd
R <- 10000
t.val <- rep(0,R)
p.val <- rep(0,R)
for(kk in 1:R){
  the.sample.1 <- rnorm(n1,m1, stdev1)
  the.sample.2 <- rnorm(n2,m2, stdev2)
  t.val[kk] <- t.test(the.sample.1, the.sample.2, var.equal=T)
  p.val[kk] <- t.val$p.value
}
```

### Simulation of 2-sample t-test

**R code**

```r
set.seed(555410)
m1 <- 100 # population 1 mean
m2 <- 100 # population 2 mean
n1 <- 25 # sample size
n2 <- 10 # sample size
stdev1 <- 10 # population 1 sd
stdev2 <- 10 # population 2 sd
R <- 10000
t.val <- rep(0,R)
p.val <- rep(0,R)
for(kk in 1:R){
  the.sample.1 <- rnorm(n1,m1, stdev1)
  the.sample.2 <- rnorm(n2,m2, stdev2)
  t.val[kk] <- t.test(the.sample.1, the.sample.2, var.equal=T)
  p.val[kk] <- t.val$p.value
}
```

**t statistic**

\[
t = \frac{(\bar{Y}_A - \bar{Y}_B) - (\mu_A - \mu_B)}{\hat{\sigma}_{\bar{Y}_A - \bar{Y}_B}}
\]

**Equation 1**

\[
\hat{\sigma}_{\bar{Y}_A - \bar{Y}_B}^2 = \frac{(n_A - 1)\sigma^2_A + (n_B - 1)\sigma^2_B}{n_A + n_B - 2} \left[ \frac{1}{n_A} + \frac{1}{n_B} \right]
\]

**Equation 2**

\[
\mu_0 = \mu_A - \mu_B
\]

**Equation 3**

\[
\sigma_0^2 = \frac{\sigma^2_A}{n_1} + \frac{\sigma^2_B}{n_2}
\]

**Equation 4**

**Above statement re: t assumes that means are distributed normally with equal variance**

**2-sample t-test H0 is true**

**2-sample t-test H0 is false**
Simulation of 2-sample t-test (equal variances)

R code
```r
set.seed(655410)
m1 <- 100 # population 1 mean
m2 <- 105 # population 2 mean
n <- 25 # sample size
sd1 <- 10 # population 1 sd
sd2 <- 10 # population 2 sd
R <- 10000
t.val <- rep(0,R)
p.val <- rep(0,R)
for(kk in 1:R){
  the.sample.1 <- rnorm(n,m1,sd1)
  the.sample.2 <- rnorm(n,m2,sd2)
  t.results <- t.test(the.sample.1,the.sample.2,var.equal=T)
  t.val[kk] <- t.results$statistic
  p.val[kk] <- t.results$p.value
}
```

R code
```r
stdev_diff <- sqrt((( ( (25-1)*100 + (25-1)*100 ) / (25+25-2) ) * ( (1/25)+(1/25) ) ) = 2.828
Expected Value of t = (100-105)/2.828 = -1.76 (actual mean = -1.81)
```

Simulation of 2-sample t-test (unequal variances)

R code
```r
set.seed(655410)
m1 <- 100 # population 1 mean
m2 <- 100 # population 2 mean
n <- 25 # sample size
sd1 <- 10 # population 1 sd
sd2 <- 3 # population 2 sd
R <- 10000
t.val <- rep(0,R)
p.val <- rep(0,R)
for(kk in 1:R){
  the.sample.1 <- rnorm(n,m1,sd1)
  the.sample.2 <- rnorm(n,m2,sd2)
  t.results <- t.test(the.sample.1,the.sample.2,var.equal=F)
  t.val[kk] <- t.results$statistic
  p.val[kk] <- t.results$p.value
}
```

one-sided t tests:
```r
> t.test(the.sample.1,the.sample.2, var.equal=F,alternative="less")

Welch Two Sample t-test
data:  the.sample.1 and the.sample.2
t = -1.919, df = 47.24, p-value = 0.0305
alternative hypothesis: true difference in means is less than 0
95 percent confidence interval:
-Inf -0.652423
sample estimates:
mean of x mean of y
98.3238  103.5093
```

100 x (1-alpha)% Confidence Interval

Probability of interval containing true population mean = 1-alpha
Probability of lower-bound being greater than true mean = alpha/2
Probability of upper-bound being less than true mean = alpha/2
Let alpha = original, 2-sided alpha

Probability of interval containing true population mean = (1-alpha) + (alpha/2)

Probability of lower-bound being greater than true mean = alpha/2

If 2-sided alpha was 0.05, then alpha/2 = 0.025. So this is 97.5% 1-sided CI.

To get 95% 1-sided Confidence Interval, our lower boundary would have to be moved.

Q: Should boundary be moved to lower or higher value?
R `t.test`

```r	n1 <- 100 # population 1 mean
tn2 <- 100 # population 2 mean
stdev1 <- 10 # population 1 sd
stdev2 <- 10 # population 2 sd
n <- 25 # sample size

the.sample.1 <- rnorm(n, mu1, stdev1)
the.sample.2 <- rnorm(n, mu2, stdev2)

two-sided t-test:
> t.test(the.sample.1, the.sample.2, var.equal=F)

Welch Two Sample t-test
data:  the.sample.1 and the.sample.2
t = -1.919, df = 47.24, p-value = 0.061
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
-10.620261   0.249236
sample estimates:
mean of x mean of y
98.3238  103.5093

one-sided t tests:
> t.test(the.sample.1, the.sample.2, var.equal=F, alternative="greater")

Welch Two Sample t-test
data:  the.sample.1 and the.sample.2
t = -1.919, df = 47.24, p-value = 0.0305
alternative hypothesis: true difference in means is less than 0
95 percent confidence interval:
-Inf -0.652423
sample estimates:
mean of x mean of y
98.3238  103.5093

> t.test(the.sample.1, the.sample.2, var.equal=F, alternative="less")

Welch Two Sample t-test
data:  the.sample.1 and the.sample.2
t = -1.919, df = 47.24, p-value = 0.969
alternative hypothesis: true difference in means is greater than 0
95 percent confidence interval:
-9.7186     Inf
sample estimates:
mean of x mean of y
98.3238  103.5093

one-sided t tests:
> t.test(the.sample.1, the.sample.2, var.equal=F, alternative="greater")

Welch Two Sample t-test
data:  the.sample.1 and the.sample.2
t = -1.919, df = 47.24, p-value = 0.0305
alternative hypothesis: true difference in means is less than 0
95 percent confidence interval:
-Inf -0.652423
sample estimates:
mean of x mean of y
98.3238  103.5093

Equivalence Tests

• reverse H0 & H1:
  - H0: there is an effect
  - H1: there is no effect

• true to reject H0 in favour of H1

• define Smallest Effect Sizes of Interest (SESOI)
  - test upper & lower bounds (SESOI) with 1-tailed t-tests
  - H0-u: observed effect unusually smaller than upper SESOI?
  - H0-l: observed effect unusually larger than lower SESOI?

• if both 1-tailed tests are significant, then we say observed effect is smaller than SESOI
  - e.g., two groups are “equivalent”

Effect Size

• what makes a p-value significant or non-significant?
  - alpha, sample size (power), effect size

• 2 classes of effect size
  - d - standardized differences (distances) btwn means
  - r - measures of association (variance accounted for)

Cohen’s $d_s$

• Used for between-subjects designs:

\[
d_s = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{(n_1 - 1)SD_1^2 + (n_2 - 1)SD_2^2}{(n_1 + n_2 - 2)}}}
\]

\[
d_s = t\sqrt{\frac{1}{n_1} + \frac{1}{n_2}} \approx 2 \times t/\sqrt{N}
\]
Hedges’ $g_s$

- unbiased estimate of population $d_s$

$$g_s = d_s \times \left(1 - \frac{3}{4(n_1 + n_2) - 9}\right)$$

Convert $d_s$ to $r$

$$r = \frac{d_s}{\sqrt{d_s^2 + (N^2 - 2N)/(n_1 n_2)}}$$

Cohen’s $d_z$

- used for within-subjects design (paired scores)
- $Md = \text{mean of difference scores}$
- $Di = \text{difference score i}$
- $N = \text{number of difference scores}$

$$d_z = \frac{Md}{\sqrt{(\sum(D_i - Md)^2/(N - 1)}}$$

$$d_z = \frac{t}{\sqrt{N}}$$
**eta-squared \( \eta^2 \)**
- measure of association between independent & dependent variables

\[
\eta^2 = \frac{SS_{\text{effect}}}{SS_{\text{total}}} \quad \text{proportion of variation in Y accounted for by group membership}
\]

\[
\eta_p^2 = \frac{SS_{\text{effect}}}{SS_{\text{effect}} + SS_{\text{error}}} \quad \text{for factorial experiments (more than 1 independent variable)}
\]

\[
\eta_p^2 = \frac{F \times df_{\text{effect}}}{F \times df_{\text{effect}} + df_{\text{error}}}
\]

**omega-squared**
- unbiased estimate of association in population

\[
\omega^2 = \frac{df_{\text{effect}} \times (MS_{\text{effect}} - MS_{\text{error}})}{SS_{\text{total}} + MS_{\text{error}}}
\]

\[
\omega_p^2 = \frac{df_{\text{effect}} \times (MS_{\text{effect}} - MS_{\text{error}})}{df_{\text{error}} \times MS_{\text{effect}} + (N - df_{\text{error}}) \times MS_{\text{error}}}
\]

---

**Why report effect size?**

- Why not just report p-values as index of “effect”?  
  - p-values depend on number of subjects/events/measures  
  - increased power leads to lower p values
- Would like to have measure doesn’t depend on experiment N  
  - or on particular aspects of experimental design
- Effect size measures try to do this  
  - and also give information about **magnitude** of effect

---

**Sample size affects precision of \( d_s \) estimation**

<table>
<thead>
<tr>
<th>Simulation Parameters</th>
</tr>
</thead>
<tbody>
<tr>
<td>8 (&lt;\ 5000 ) if 5K samples</td>
</tr>
<tr>
<td>n(&lt;\ 10 ) if mean of pop 1</td>
</tr>
<tr>
<td>m(&lt;\ 12 ) if mean of pop 2</td>
</tr>
<tr>
<td>s(&lt;\ 4 ) if sd of population</td>
</tr>
</tbody>
</table>

**Simulation 1:** n \(<\ 20 \) if sample size

**Simulation 2:** n \(<\ 70 \) if sample size

- effect size measures vary across samples  
  - variance of sampling distribution is larger for smaller samples
winner’s curse

- apply statistical threshold (e.g., p<.05) to effect size
  - what is median significant $d_s$?
- depends on sample size
  - for small samples, only big $d_s$ are significant
  - large samples more likely to get significant large & small $d_s$

if most published studies have low power, then published effect sizes are biased

- published effect sizes are too high

significant $d_s$ are more extreme with small samples

effect size measures vary across samples

variance of sampling distribution is larger for smaller samples