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

t test for single mean

R code
```r
set.seed(555410)
mu <- 0 # 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.results <- t.test(the.sample)
  t.val[kk] <- t.results$statistic
  p.val[kk] <- t.results$p.value
}
```

Statistical Inference Part 2
• t-tests
• Effect Size
• Equivalence Tests
• Consequences of Low Power

```latex
\frac{t}{s/\sqrt{n}} \text{ or } t = \frac{Y - \mu}{s/\sqrt{n}}
\]
Comparing 2 means

- Given 2 populations of scores
  - means: \( \mu_A, \mu_B \)
  - variances: \( \sigma^2_A, \sigma^2_B \)
- Distributions of sample means:
  - means: \( \bar{\mu}_A, \bar{\mu}_B \)
  - variances: \( \sigma^2_{\bar{\mu}_A}, \sigma^2_{\bar{\mu}_B} \)
- shape: normal (via Central Limit Theorem)
- Distribution of difference between sample means will be normal:
  - mean: \( \bar{\mu}_A - \bar{\mu}_B \)
  - variance: \( \sigma^2_{\bar{\mu}_A - \bar{\mu}_B} = \sigma^2_{\bar{\mu}_A} + \sigma^2_{\bar{\mu}_B} \)

R code

```r
set.seed(555418)
mu1 <- 100 # population 1 mean
mu2 <- 100 # population 2 mean
n1 <- 20 # sample size
s1 <- 10 # population 1 sd
sd1 <- 10 # population 2 sd
R <- 10000

# Set the null hypothesis
h0 <- c(mu1 - mu2)

# Null hypothesis is accepted if t-statistic falls below 95th percentile
# and can be written as:


# Calculate the t-statistic

df <- n1 + n2 - 2

t.val <- t(test(x = sample.A, y = sample.B, alternative = "two.sided", var.equal = TRUE))

# Calculate the p-value
p.val <- t.results$p.value

# Set the null hypothesis
h0 <- c(mu1 - mu2)

# Null hypothesis is accepted if t-statistic falls below 95th percentile
# and can be written as:


# Calculate the t-statistic

df <- n1 + n2 - 2

t.val <- t(test(x = sample.A, y = sample.B, alternative = "two.sided", var.equal = TRUE))

# Calculate the p-value
p.val <- t.results$p.value
```

Simulation of 2-sample t-test

R code

```r
set.seed(555418)
mu1 <- 100 # population 1 mean
mu2 <- 100 # population 2 mean
n1 <- 20 # sample size
s1 <- 10 # population 1 sd
sd1 <- 10 # population 2 sd
R <- 10000

# Calculate the t-statistic

df <- n1 + n2 - 2

t.val <- t(test(x = sample.A, y = sample.B, alternative = "two.sided", var.equal = TRUE))

# Calculate the p-value
p.val <- t(results$p.value)
```

Above statement re: t assumes that means are distributed normally with equal variance.
Simulation of 2-sample t-test (equal variances)

```r
code
set.seed(655410)
mu1 <- 100 # population 1 mean
mu2 <- 105 # population 2 mean
n <- 25 # 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(n,mu1,stdev1)
  the.sample.2 <- rnorm(n,mu2,stdev2)
  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
}
```

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:

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

$$d_s \approx t \sqrt{\frac{1}{n_1} + \frac{1}{n_2}} \approx 2 \times t / \sqrt{N}$$
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)
  - $Md$ = mean of difference scores
  - $Di$ = difference score $i$
  - $N$ = 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

$$\eta^2 = \frac{SS_{effect}}{SS_{total}}$$

proportion of variation in Y accounted for by group membership

$$\eta_p^2 = \frac{SS_{effect}}{SS_{effect} + SS_{error}}$$

for factorial experiments (more than 1 independent variable)

$$\eta_p^2 = \frac{F \times df_{effect}}{F \times df_{effect} + df_{error}}$$

omega-squared

- unbiased estimate of association in population

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

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

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

- 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 $d_s$

**effect size inflation (winner's curse)**

- effect size measures vary across samples
- variance of sampling distribution is larger for smaller samples
- significant $d_s$ are more extreme with small samples

**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
  - it is not equal to the probability that H0 is true
  - it is not equal to the probability of replicating the result...

**probability of replicating p=.05 effect**

- observed effect significant $p<0.05$
- expect 50% of replications to yield smaller effect
- critical value of $t(8)=1.86$

- $n=9$, $\mu=1.86\times\text{SEM}$
What does a p-value mean?

• A non-significant p-value indicates that the result is NOT unusual assuming null hypothesis is true.
• It does NOT mean that the null hypothesis is true
  - e.g., a non-significant 2-sample t-test does NOT mean that the populations means are equal
• You do not accept the null hypothesis, you simply fail to reject it.
• To “accept” the null hypothesis, perform an equivalence test

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”

Spurious findings?

• Over time:
  - there has been a dramatic increase in the ease with which we can analyze a single data set in multiple ways
  - also, the effects/phenomena of interest typically become increasingly smaller & more subtle as a research area matures
  - but, sample sizes tend not to change very much
  - nor have the unwritten rules about deciding if a paper is publishable
• these trends suggest many statistically significant findings are spurious

Power (Type I & Type II Errors)

<table>
<thead>
<tr>
<th>decision</th>
<th>H0 is True</th>
<th>H0 is False</th>
</tr>
</thead>
<tbody>
<tr>
<td>reject H0</td>
<td>Type I (p = α)</td>
<td>Correct (p = 1 − β = power)</td>
</tr>
<tr>
<td>do not reject H0</td>
<td>Correct (p = 1 − α)</td>
<td>Type II error (p = β)</td>
</tr>
</tbody>
</table>

Table 1: Possible outcomes of hypothesis testing.

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)
Positive Predictive Value (PPV)

- Given that we’ve rejected H0 in favour of H1
  - what is probability that our conclusion is correct?

<table>
<thead>
<tr>
<th>decision</th>
<th>H0 is True</th>
<th>H0 is False</th>
</tr>
</thead>
<tbody>
<tr>
<td>reject H0</td>
<td>Type I (p = α)</td>
<td>Correct (p = 1 - β = power)</td>
</tr>
<tr>
<td>do not reject H0</td>
<td>Correct (p = 1 - α)</td>
<td>Type II error (p = β)</td>
</tr>
</tbody>
</table>

\[
PPV = \frac{p_A}{(p_A + p_B)}
\]

Factors that lower PPV

- small studies & small effects (low power)
- testing more hypotheses (lowering R)
- greater flexibility in design and analysis (increasing bias)
- more independent research teams studying phenomenon
  - due to selective reporting of “positive” results

Ioannidis 2005

**Table 1. Research Findings and True Relationships**

<table>
<thead>
<tr>
<th>Research Finding</th>
<th>True Relationship</th>
<th>Yes</th>
<th>No</th>
<th>Total</th>
</tr>
</thead>
</table>
| Yes              | c1                    | c1 - α | c0 | c1 /
|                  | false              | 1     | c0 | c1 |
| No               | c2                    | c2 - α | c0 | c2 |
|                  | true                | 1     | c0 | c2 |
||| total                  | c1 + c2              | c1 + c2 |

R is ratio of number of “true effects” divided number of “no effects”
\(c\) is number of effects begin examined
\(\alpha\) is Type I error rate
\(1 - \beta\) is power (1 minus Type II error rate)

\[
PPV = \frac{1 - \beta}{R(1 - \beta) + \alpha}
\]

A statistically significant finding has a greater than 50% chance of being true/correct if:
\[R(1 - \beta) > \alpha\]