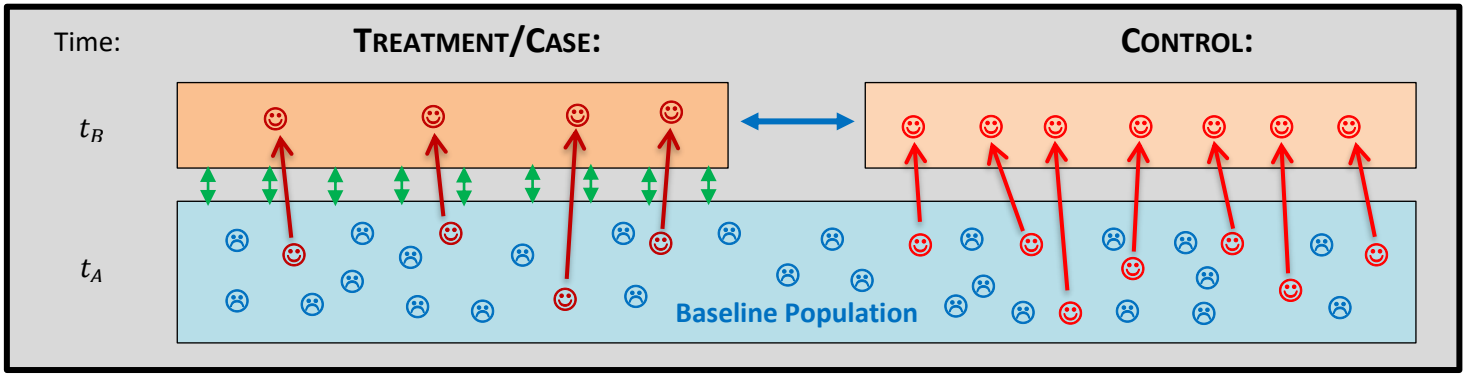
Comparisons of Means

Introduction

- Objective: Test whether the ***population*** means μ_A and μ_B of either dependent or independent samples ***differ*** by a hypothesized quantity D_0.
  Frequently, the difference is ***assumed zero*** under the null hypothesis $D_0 = 0$.
- The null and the alterative hypotheses are:

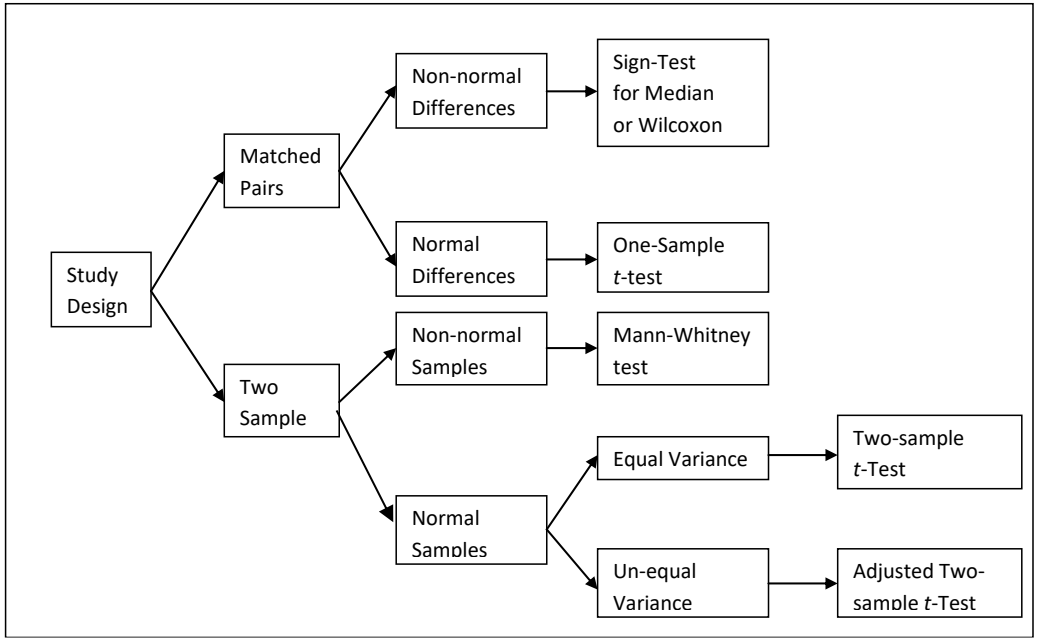
	$D_0 = 0$		$D_0 \neq 0$	
	H_0	H_1	H_0	H_1
two sided	$H_0: \mu_A = \mu_B$	$H_1: \mu_A \neq \mu_B$	$H_0: \mu_A - \mu_B = D_0$	$H_1: \mu_A - \mu_B \neq D_0$
one-sided	$H_0: \mu_A \leq \mu_B$	$H_1: \mu_A > \mu_B$	$H_0: \mu_A - \mu_B \leq D_0$	$H_1: \mu_A - \mu_B > D_0$
	$H_0: \mu_A \geq \mu_B$	$H_1: \mu_A < \mu_B$	$H_0: \mu_A - \mu_B \geq D_0$	$H_1: \mu_A - \mu_B < D_0$

- ***Study designs*** to compare means:



- In the ***related/matched sample design*** the means within cases before (time t_A) and after the treatment (time t_B) are compared (green arrows).
- In the ***independent sample design*** the cases, which received treatment, are compared (blue arrow) against the controls, which have not received the treatment, at time t_B.
- In either case, sampling is performed from the ***baseline population*** which has not been exposed to a treatment.

Two-Sample Test Decision Rules



Two Related Sample

- Two measurements may be related either because
  - they are performed ***repeatedly*** at the same object before and after the treatment or
  - cases are ***matched*** to untreated members (controls), which are closely related with the cases.
    For example, treated and untreated siblings, cases and a member of their neighborhood, spouses.
- In either case, one can expect a high degree of ***correlation*** between both measurements.
  $\Rightarrow$ This implies for the underlying data structure that each observation has two measurements and the treated and untreated number of measurements are identically.

The test statistic

- Gain Score: The two measurements at t_A and t_B for the i^{th} observations can be expressed by the gain score D_i as difference $D_i = X_{iA} - X_{iB}$.
- The test statistic becomes:

$$\overline{D} = \frac{\sum_{i=1}^{n} D_i}{n}.$$

- The standard error of the average gain $\overline{D}$ is $s_D/\sqrt{n}$ with the estimator of the variance being $s_D^2 = \frac{\sum_{i=1}^{n}(D_i - \overline{D})^2}{n-1}$.

- The test statistic becomes $t = \frac{\bar{D} - D_0}{s_D/\sqrt{n}}$ and for $D_0 = 0$ it becomes $t = \frac{\bar{D}}{s_D/\sqrt{n}}$.
- Assuming the sample measurements come for a population that satisfies the null hypothesis the test statistic will follow a t-distribution with $n - 1$ degrees of freedom.
- The test is equivalent to the one-sample test of the mean $H_0: \mu_D = D_0$.

DEFINITION: SAMPLING DISTRIBUTION OF PAIRED-OBSERVATION MEAN $\bar{D}$

Assume X_1 and X_2 are normal with a difference in means $\mu_1 - \mu_2 = D_0$. Given a random sample of n paired observations, the following has an approximate t-distribution:

$$T = \frac{\bar{D} - D_0}{S_d/\sqrt{n}} \qquad (10\text{-}10)$$

with $n - 1$ degrees of freedom.

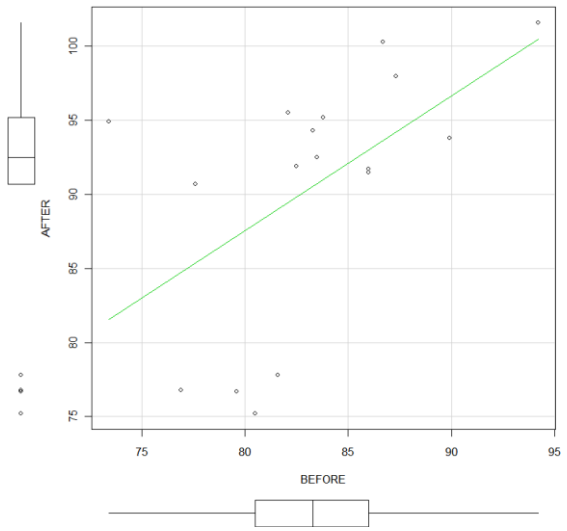
Advantages and disadvantages of using related samples:

- ***Variability among*** the sample objects becomes irrelevant because it cancels out using the differences at t_A and t_B, therefore just leaving the effect size:
  Let $X_{iA} = c_i + x_{iA}$ and $X_{iB} = c_i + x_{iB}$ then the gain becomes
  $$D_i = (c_i + x_{iA}) - (c_i + x_{iB}) = x_{iA} - x_{iB}$$
- ***External factors*** c_i are controlled because they are constant within each measurement object and therefore cancel out (this is underlying concept of panel data analysis).

- ***Less*** objects/individuals are needed to be recruited into the sample because we repeat measurements at the same object/individual.
  That is, we achieve higher power of rejecting H_0 when, in fact, it is false.
- There may be a carry-over effect that arises when the measurement at t_A influences the outcome at t_B.
- Potential loss of observations because a second measurement cannot be performed on them.

Example: Everitt's anorexia data

Investigate the data structure of **DEPSAMPLE** and **INDEPSAMPLE**




```
Call:
lm(formula = AFTER ~ BEFORE, data = DepSample)

Residuals:
    Min      1Q  Median      3Q     Max
-12.812  -2.759   1.760   4.187  13.343

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  14.8198    30.6500   0.484   0.6357
BEFORE        0.9092     0.3676   2.473   0.0258 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1
' ' 1

Residual standard error: 7.377 on 15 degrees of freedom
Multiple R-squared: 0.2897,   Adjusted R-squared: 0.2423
F-statistic: 6.117 on 1 and 15 DF,  p-value: 0.02584
```

Note: The 4 observations in the lower left corner may be special cases, because these low weight girls lose weight even after treatment.

Perform the *t*-test (important: order of both variables):

```
> t.test(DepSample$AFTER, DepSample$BEFORE, alternative='two.sided',
+   conf.level=.95, paired=TRUE)

Paired t-test

data:  DepSample$AFTER and DepSample$BEFORE
t = 4.1849, df = 16, p-value = 0.0007003
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
  3.58470 10.94471
sample estimates:
mean of the differences
               7.264706
```

Two Independent Samples

- In the two-independent sample design, we distinguish between a ***case group*** that has been ***exposed*** to a particular treatment and the ***control group*** remains ***unexposed*** to the treatment. We want the ***test for the effect of the treatment*** on the case group compare to the control group.
- The members of the case and the control groups were ***randomly sampled*** from a common population and ***randomly assigned*** to either group.
- Both sample groups can be of different sizes.

The test statistic

- The key problem of developing the test statistics lies in the ***evaluation*** of its ***standard error***.

- While both samples, under the null hypothesis, come from the same population and, therefore, should have the identical variances,
  - the ***different sample*** sizes of the cases and controls complicate the estimation of the variance from the sample observations
  - the ***treatment*** of the cases compared to the controls may also change the variance.
- Theoretically, for subtraction or addition of independent means their joint variance is equal to their sums:

$$\sigma^2_{\bar{X}_A \pm \bar{X}_B} = \sigma^2_{\bar{X}_A} + \sigma^2_{\bar{X}_B} = \frac{\sigma_A^2}{n_A} + \frac{\sigma_B^2}{n_B}.$$

- Therefore, using the sample estimates s_A^2 and s_B^2, the general structure of the test statistic becomes

$$t = \frac{(\bar{X}_A - \bar{X}_B) - E[(\bar{X}_A - \bar{X}_B)]}{\sqrt{Var[(\bar{X}_A - \bar{X}_B)]}} = \frac{(\bar{X}_A - \bar{X}_B) - (\mu_A - \mu_B)}{\sqrt{\frac{s_A^2}{n_A} + \frac{s_B^2}{n_B}}} = \frac{\bar{X}_A - \bar{X}_B}{\sqrt{\frac{s_A^2}{n_A} + \frac{s_B^2}{n_B}}}$$

  where under the null hypothesis $\mu_A - \mu_B = 0$.
- The test statistic t follows a t-distribution with $n_A + n_B - 2$ degrees of freedom.
  Two degrees of freedom are lost because the estimated sample variances consume each one degree of freedom.

Pooling variances

- The expression for the standard error of the difference of means $s^2_{\bar{X}_A \pm \bar{X}_B} = \frac{s_A^2}{n_A} + \frac{s_B^2}{n_B}$ ***weights each sample variance equally***, which is misleading if the samples sizes differ substantially, e.g., $n_A \neq n_B$.
- A better estimate is the pooled variance estimator which employs differential weights:

$$s^2_{pooled} = \frac{\sum_{i=1}^{n_A}(X_{iA} - \bar{X}_A)^2 + \sum_{j=1}^{n_B}\left(X_{jB} - \bar{X}_B\right)^2}{n_A + n_B - 2}$$
$$= \frac{(n_A - 1) \cdot s_A^2 + (n_B - 1) \cdot s_B^2}{n_A + n_B - 2}.$$

- The test statistics becomes

$$t = \frac{\bar{X}_A - \bar{X}_B}{\sqrt{\frac{s^2_{pooled}}{(n_A + n_B)}}}.$$

DEFINITION: SAMPLING DISTRIBUTION OF $\bar{X}_1 - \bar{X}_2$, EQUAL POPULATION VARIANCES

Assume X_1 and X_2 are normal with a difference in means $\mu_1 - \mu_2 = D_0$. If the variance σ^2 is the same for both populations, then the following has a t-distribution:

$$T = \frac{\bar{X}_1 - \bar{X}_2 - D_0}{\hat{\sigma}_{\bar{X}_1 - \bar{X}_2}} = \frac{\bar{X}_1 - \bar{X}_2 - D_0}{S_p\sqrt{1/n_1 + 1/n_2}} \qquad (10\text{-}3)$$

with degrees of freedom

$$\mathrm{df} = n_1 + n_2 - 2$$

Heterogeneity of Variances

- We can usually assume that the treatment just shifts the mean level without affecting the variance.
  Therefore, the test statistic t follows a t-distribution with $n_A + n_B - 2$ degrees of freedom.
- However, if this assumption is incorrect difficulties arise (see Wikipedia for the Behrens-Fisher Problem).
- Several ***conservative adjustments*** are proposed in the literature.
  Conservative means: the actual probability of rejecting a true null hypothesis incorrectly by error is smaller than the nominal level α.

That implies that we are on the safe side and if the null hypothesis is rejected, the error probability is smaller than α.

- Two proposed adjustments are used for the degrees of freedom of the t-test
  1. the smaller of the two numbers $(n_A - 1)$ or $(n_B - 1)$, or
  2. $df \approx \frac{\left(S_A^2/n_A + S_B^2/n_B\right)^2}{\left(S_A^2/n_A\right)^2/(n_A-1) + \left(S_B^2/n_B\right)^2/(n_B-1)}$.

DEFINITION: SAMPLING DISTRIBUTION OF $\bar{X}_1 - \bar{X}_2$, POPULATION VARIANCES UNEQUAL

Assume X_1 and X_2 are normal with a difference in means $\mu_1 - \mu_2 = D_0$ and variances $\sigma_1^2 \neq \sigma_2^2$. Then the following has an approximate t-distribution:

$$T = \frac{\bar{X}_1 - \bar{X}_2 - D_0}{\hat{\sigma}_{\bar{X}_1 - \bar{X}_1}} = \frac{\bar{X}_1 - \bar{X}_2 - D_0}{\sqrt{S_1^2/n_1 + S_2^2/n_2}} \tag{10-4}$$

with degrees of freedom given by

$$\text{df} = \frac{(S_1^2/n_1 + S_2^2/n_2)^2}{(S_1^2/n_1)^2/(n_1 - 1) + (S_2^2/n_2)^2/(n_2 - 1)} \tag{10-5}$$

Alternatively, the (approximate) degrees of freedom can be found from

$$\text{df} = \min(n_1 - 1, n_2 - 1)$$

Impact of the Deviation from the Underlying Assumptions

- See the script **TTESTSIMULATION.R** for an investigation on how violations of the standard assumptions [a] normality and [b] equality of variances affect the significance levels of the *t*-test.

Test for the Equality of Variances / Homoscedasticity

- The specification of the two-sample difference-of-means test depends on $\sigma_1^2 = \sigma_2^2$ or $\sigma_1^2 \neq \sigma_2^2$.
- The *two-sided* hypothesis is formulated in terms of *ratios*, i.e., $H_0 : \frac{\sigma_1^2}{\sigma_2^2} = 1$ against $H_A : \frac{\sigma_1^2}{\sigma_2^2} \neq 1$
- Distribution under the null hypothesis H_0

DEFINITION: SAMPLING DISTRIBUTION OF THE RATIO OF VARIANCES

Assume X_1 and X_2 are both normally distributed, with variances σ_1^2 and σ_2^2. Given independent random samples of size n_1 and n_2, then the statistic

$$F = \frac{S_1^2/\sigma_1^2}{S_2^2/\sigma_2^2}$$

will follow an F distribution with $n_1 - 1$ and $n_2 - 1$ degrees of freedom.

- Because under the null hypothesis H_0 the population variance are identically, the test statistic becomes $F = \frac{S_1^2}{S_2^2}$.

Analysis of the Everitt's anorexia example

- Comparison of the sample variance:

```
> var.test(GAIN ~ Treatment, alternative='two.sided', conf.level=.95,
+   data=IndepSample)

F test to compare two variances

data:  GAIN by Treatment
F = 0.8027, num df = 16, denom df = 25, p-value = 0.6587
alternative hypothesis: true ratio of variances is not equal to 1
95 percent confidence interval:
 0.3367083 2.0981634
sample estimates:
ratio of variances
         0.8027132
```

  We cannot reject $H_0: \frac{\sigma_A^2}{\sigma_B^2} = 1$ and therefore tentatively assume equal variances.

- The t-test assuming identical variances (see option **`var.equal=TRUE`**)[1] of a two-sided null hypothesis $H_0: \mu_{case} = \mu_{control}$ is accomplished by:

```
> t.test(GAIN~Treatment, alternative='two.sided', conf.level=.95,
+   var.equal=TRUE, data=IndepSample)

Two Sample t-test

data:  GAIN by Treatment
t = 3.2227, df = 41, p-value = 0.002491
```

[1] Use **`var.equal=FALSE`** if the *F*-test indicates heteroscedasicity.

```
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:   2.880164 12.549248
sample estimates:
   mean in group Case mean in group Control
             7.264706              -0.450000
```

- Alternatively we could have achieved the same results with a regression model that uses **TREATMENT** as a dummy variable.

```
Call: lm(formula = GAIN ~ Treatment, data = IndepSample)

Coefficients:
                    Estimate Std. Error t value Pr(>|t|)
(Intercept)            7.265      1.861   3.903 0.000347 ***
Treatment[T.Control]  -7.715      2.394  -3.223 0.002491 **

Residual standard error: 7.675 on 41 degrees of freedom
Multiple R-squared: 0.2021,   Adjusted R-squared: 0.1827
F-statistic: 10.39 on 1 and 41 DF,  p-value: 0.002491
```

- This model implies for
  - cases: $GAIN = 7.265 + \underbrace{0}_{Treatment=0} \cdot -7.715 = 7.265$
  - controls: $GAIN = 7.265 + \underbrace{1}_{Treatment=1} \cdot -7.715 = -0.45$