PS303: Week 7

pp. 379-437

From categories to intervals and ratios

- We discussed strategies for describing and analyzing *categorical* data. Categories *cannot* be meaningfully sub-divided into smaller units as each level is distinct. Think of colors, species and religions.
  - The type of test we run depends on our data structure. If we have a *binary* outcome variable which can be conceptualized in terms of 'successes' and 'fails' (e.g., Heads vs Tails; Wins vs Losses), we can run *binomial* tests to estimate the likelihood of a 'success' and it's associated p-value.
  - When we have multiple (>2) levels of categorical data, we can estimate χ^2 statistics to determine whether observed category distributions are statistically significant in terms of being **different** (two-sided) or **greater/smaller** (one-sided) from an expected distribution (the null hypothesis, or H_O).
  - We generate decisions about retaining/rejecting H_O based on the p-value - e.g., is $p < .05$? The latter describes the probability of acquiring the observed data assuming the null hypothesis is true.

- Beyond discrete categories, you will encounter *continuous* data that can be conceptualized as intervals and ratios (intervals + absolute 0). Think of age, height, weight, distance, and brain activity - any parameter that can be sub-divided *ad infinitum* while retaining equal distance relative to adjacent values. For the remainder of the course, we will focus on describing, analyzing and interpreting continuous parameters using inferential tests common to the social sciences.

- We will focus on t-tests today, covering:
  - 1-sample vs 2-sample t-tests
  - One-sided ("greater", "less") vs two-sided ("two.sided") contrasts
  - Independent vs paired/repeated tests
  - Parametric vs non-parametric contrasts
  - Why Psychologists should use Welch's test (Delacre, Lakens & Leys, 2017)
  - Why Hedge's G should be reported with Welch's test (Delacre, Lakens, Ley, Liu & Leys, 2021).

1-sample t-tests

> *Use*: You have data for a single sample which you want to compare against a population parameter.

- Required parameters

- Sample mean ($\hat{X}$)

- Sample standard deviation (SD_s)

- Sample size (n)

- Population mean (μ)
  You can then compute the test statistic (t) as follows:

$$t = \frac{\hat{X} - \mu}{\frac{SD_s}{\sqrt{n}}}$$

Example: You collect the ages of 10 students in PS303 (n) in the classroom. You want to know if the mean age of your sample ($\hat{X}$) is *different* from the mean age of all 3$^{\text{rd}}$ year USP students, which is 22 years (μ).

The ages of the 10 students are 24, 25, 25, 27, 20, 20, 28, 21, 30, 20. The sample mean may then be calculated as

$$\frac{24 + 25 + 25 + 27 + 20 + 20 + 28 + 21 + 30 + 20}{10} = \frac{240}{10} = 24$$

.

We can compute the sample standard deviation as

$$SD_s = \sqrt{\frac{\sum (X_i - \hat{X})^2}{n - 1}}$$

. Plugging in our values gives us

$$SD_s = \sqrt{\frac{(24 - 24) + (25 - 24) + (25 - 24) + (27 - 24) + (20 - 24) + (20 - 24) + (28 - 24) + (21 - 24) + (30 - 24) + (20 - 2}{15 - 1}}$$

We can now compute the test statistic:

$$t = \frac{\hat{X} - \mu}{\frac{SD_s}{\sqrt{n}}} = \frac{24 - 22}{\frac{3.65}{\sqrt{15}}} = 1.73$$

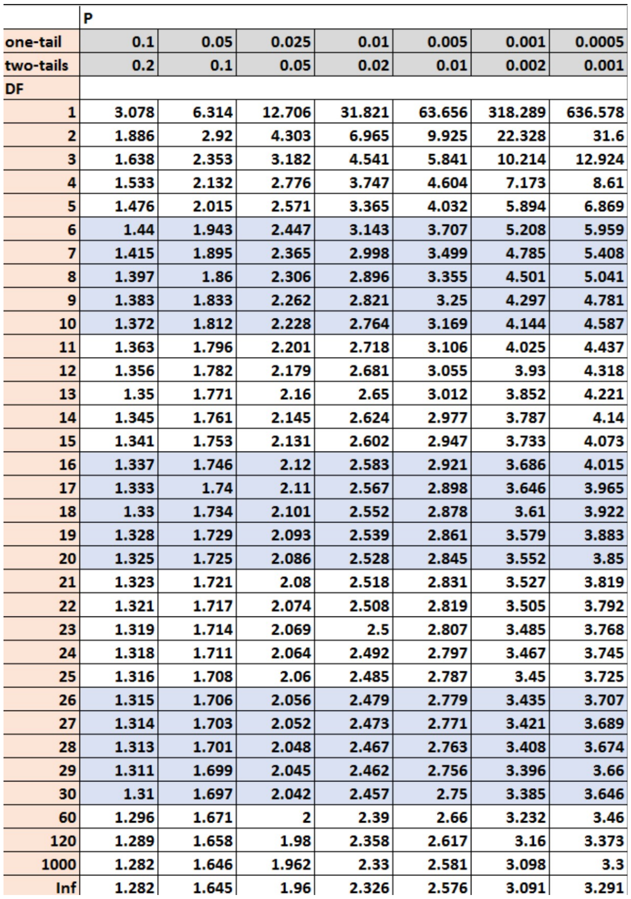
Is the outcome statistically significant at $p < .05$? We can look up the test statistic distribution to check…

We can speed up the process by running the entire test in *R*

```r
# store the sample's ages into a variable (called 'x'). Remember that the population parameter ('mu') i
x <- c(24,25,25,27,20,20,28,21,30,20)

# run the test!
t.test(x,mu=22,alternative = "two.sided")
```

```
##
##  One Sample t-test
##
## data:  x
## t = 1.7321, df = 9, p-value = 0.1173
## alternative hypothesis: true mean is not equal to 22
## 95 percent confidence interval:
##  21.38789 26.61211
## sample estimates:
## mean of x
##        24
```



	P						
one-tail	0.1	0.05	0.025	0.01	0.005	0.001	0.0005
two-tails	0.2	0.1	0.05	0.02	0.01	0.002	0.001
DF							
1	3.078	6.314	12.706	31.821	63.656	318.289	636.578
2	1.886	2.92	4.303	6.965	9.925	22.328	31.6
3	1.638	2.353	3.182	4.541	5.841	10.214	12.924
4	1.533	2.132	2.776	3.747	4.604	7.173	8.61
5	1.476	2.015	2.571	3.365	4.032	5.894	6.869
6	1.44	1.943	2.447	3.143	3.707	5.208	5.959
7	1.415	1.895	2.365	2.998	3.499	4.785	5.408
8	1.397	1.86	2.306	2.896	3.355	4.501	5.041
9	1.383	1.833	2.262	2.821	3.25	4.297	4.781
10	1.372	1.812	2.228	2.764	3.169	4.144	4.587
11	1.363	1.796	2.201	2.718	3.106	4.025	4.437
12	1.356	1.782	2.179	2.681	3.055	3.93	4.318
13	1.35	1.771	2.16	2.65	3.012	3.852	4.221
14	1.345	1.761	2.145	2.624	2.977	3.787	4.14
15	1.341	1.753	2.131	2.602	2.947	3.733	4.073
16	1.337	1.746	2.12	2.583	2.921	3.686	4.015
17	1.333	1.74	2.11	2.567	2.898	3.646	3.965
18	1.33	1.734	2.101	2.552	2.878	3.61	3.922
19	1.328	1.729	2.093	2.539	2.861	3.579	3.883
20	1.325	1.725	2.086	2.528	2.845	3.552	3.85
21	1.323	1.721	2.08	2.518	2.831	3.527	3.819
22	1.321	1.717	2.074	2.508	2.819	3.505	3.792
23	1.319	1.714	2.069	2.5	2.807	3.485	3.768
24	1.318	1.711	2.064	2.492	2.797	3.467	3.745
25	1.316	1.708	2.06	2.485	2.787	3.45	3.725
26	1.315	1.706	2.056	2.479	2.779	3.435	3.707
27	1.314	1.703	2.052	2.473	2.771	3.421	3.689
28	1.313	1.701	2.048	2.467	2.763	3.408	3.674
29	1.311	1.699	2.045	2.462	2.756	3.396	3.66
30	1.31	1.697	2.042	2.457	2.75	3.385	3.646
60	1.296	1.671	2	2.39	2.66	3.232	3.46
120	1.289	1.658	1.98	2.358	2.617	3.16	3.373
1000	1.282	1.646	1.962	2.33	2.581	3.098	3.3
Inf	1.282	1.645	1.96	2.326	2.576	3.091	3.291

Figure 1: Test statistic distribution

We can report the outcome as follows:

The mean age of 10 PS303 students ($\hat{X} = 24$) was not significantly different from the mean age of all 3$^{\text{rd}}$ year students ($\mu = 22$) following a one-sample t-test, $t(9) = 1.73$, $p = .117$, $d = 0.55$. The null hypothesis ($H_O : \hat{X} - \mu = 0$) was not rejected.

Q. What would you alter if you wanted to know whether the mean age of PS303 students was statistically *greater* than the (assumed) population mean?

2-sample t tests (Student's vs Welch)

> *Use*: You want to know whether the difference between two samples (of $n \geq 5$ per sample) varies significantly from the expected difference (knowledge of the population parameter is not necessary). The null hypothesis can be described as follows: $H_0 : \hat{X}_1 - \hat{X}_2 = 0$; that is, the mean difference between the two sample parameters is not statistically different from 0.

The conventional procedure for a 2-sample test involves Student's approach. The parameters for the equation are similar to those you have already encountered:

- $\hat{X}_1$ and $\hat{X}_2$ are mean parameters for the two groups being compared.
- SD_1 and SD_2 are the sample standard deviations for the two groups
- n_1 and n_2 are the sample sizes for each group

Student's 2-sample test statistic for *independent*(unpaired) samples can be estimated as follows:

$$t = \frac{\hat{X}_1 - \hat{X}_2}{\sqrt{\frac{(n_1-1)SD_1^2+(n_2-1)SD_2^2}{n_1+n_2-2} \times (\frac{1}{n_1} + \frac{1}{n_2})}}$$

where the degrees of freedom is estimated as $df = n_1 + n_2 - 2$.

Note the following features of Student's formulation:

- The error term (difference between an estimate and it's true value) is *pooled* across the two samples because (it is assumed) that both sample variances represent a common population, in which case variances should be *homogeneous*.

- Homogeneity of variance is vital for deriving an unbiased estimator but is not always the case when examining 'real-world' data due to (i) natural variability between groups (intellectual variability between men and women; non-random assignment to conditions) and (ii) the inclusion of the experimental treatment (e.g., placebo application).

- It is also assumed (for Student's t) that each sample contains an equal number of participants (is balanced), otherwise pooled variance will be biased towards the larger sample.

- Finally, both samples should be normally distributed to generate a reliable estimate.

In sum, the following assumptions should be met in order to reliably interpret a Student's t-test:

- Is the data normally distributed across samples?

- Are sample sizes balanced? Is the design parametric? (Not a concern for *paired* contrasts)

- Are the variances between the tested groups statistically equivalent?

Not meeting these assumptions renders the output of a Student's t-test biased. A lot of real-world data does not meet these assumptions, presenting unbalanced designs, non-normal data, and/or unequal variances. Although one could mathematically transform the data to try and meet these assumptions (e.g., by reducing spread) or run nonparametric tests (which look at ranked data and are more prone to Type-2 error), interpretive issues can be mitigated by using Welch's t-test by default..

$$t_{Welch} = \frac{\hat{X}_1 - \hat{X}_2}{\sqrt{\frac{SD_1^2}{n_1} + \frac{SD_2^2}{n_2}}}$$

with the degrees of freedom (from which the critical value is found) being

$$df_{Welch} = \frac{(\frac{SD_1^2}{n_1} + \frac{SD_2^2}{n_2})^2}{\frac{(\frac{SD_1^2}{n_1})^2}{n_1-1} + \frac{(\frac{SD_2^2}{n_2})^2}{n_2-1}}$$

Instead of pooling variance *across* samples (as was the case with Student's t), Welch's test divides the group mean difference with an *unpooled* error term, which takes different sample sizes and variances into account. As a result, the estimated p-value and test statistic is more robust to imbalanced sample sizes and unequal variance. However, the assumption of normality is still required for an unbiased Welch's estimate (although this is less important than for Student's t). For the remainder of this course, we will use Welch's t-test by default so as to avoid the 'conventional' two-step procedure required for Student's test statistic (test for variance homogeneity *and then* decide the test to run, which can lead to reduced Type-1 error).

Running 2-sample tests

> When we contrast two *independent* groups, we run an independent samples t-test. When we contrast the same group at two time points, we run a *paired* t-test.

- The arguments within the `t.test` function shown earlier allows running both types of contrasts.

Example of an independent t-test using Welch's method

You are interested in finding out whether the amount of time spent on social media influences depression rates. You recruit 40 students and ask them about their online habits. Using their responses, you split your students into a *High-Social-Media* group ($n_1 = 26$) and a *Low-Social-Media* group ($n_2 = 14$). Next, you have all students complete a depression inventory. Your raw data may look like the following (assume that higher scores correspond with greater depression):

- Scores for 26 participants in the *High-Social-Media* group: 7, 5, 6, 8, 10, 7, 6, 6, 6, 9, 10, 10, 7, 10, 6, 9, 9, 9, 5, 8, 6, 9, 6, 9, 6, 5.
- Scores for 14 participants in the *Low-Social-Media* group: 8, 3, 10, 5, 3, 9, 4, 7, 3, 2, 2, 10, 10, 9

Note that because our samples are unbalanced, we would use a Welch's test by default.

Let the null hypothesis be $H_O : \hat{X}_1 - \hat{X}_2 = 0$ in which case the alternative hypothesis would be $H_A : \hat{X}_1 - \hat{X}_2 \neq 0$ (two-sided).

```r
# Assign values to variables
high.soc <- c(7, 5, 6, 8, 10, 7, 6, 6, 6, 9, 10, 10, 7, 10, 6, 9, 9, 9, 5, 8, 6, 9, 6, 9, 6, 5)
low.soc  <- c(6, 10, 3, 6, 2, 8, 3, 10, 5, 3, 9, 4, 7, 3)

# Run the test
t.test(high.soc,low.soc,
       paired = F,    # This isn't necessary since the unequal sample sizes tells R to run Welch's test
       alternative = "two.sided") # Do you want a one- or two-sided test?
```

```
##
##  Welch Two Sample t-test
##
## data:  high.soc and low.soc
## t = 2.2157, df = 18.647, p-value = 0.03937
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  0.09847118 3.53889146
## sample estimates:
## mean of x mean of y
##  7.461538  5.642857
```

> Mean depression scores for the *High-Social-Media* group (M = 7.46; SD = 1.75) was statistically different relative to the mean depression scores for the *Low-Social-Media* group (M = 5.64; SD = 2.79) following a two-sample Welch's test, $t(18.6) = 2.22, p = .039$. The null hypothesis can be rejected.

Suppose you wanted to test a one-sided hypothesis: specifically, you suspect the *High-Social-Media* group may be more depressed relative to the *Low-Social-Media* group. In this case we can alter our null hypothesis to be $H_O : \hat{X}_1 - \hat{X}_2 \leq 0$. We can run the same test as before after changing the `alternative` argument.

```r
t.test(high.soc,low.soc,paired = F,alternative = "greater") # We want to know whether the FIRST variabl
```

```
##
##  Welch Two Sample t-test
##
## data:  high.soc and low.soc
## t = 2.2157, df = 18.647, p-value = 0.01968
## alternative hypothesis: true difference in means is greater than 0
## 95 percent confidence interval:
##  0.3979889       Inf
## sample estimates:
## mean of x mean of y
##  7.461538  5.642857
```

A one-sided test is less conservative than a two-sided test (it is 'easier' to reach significance) but it must be theoretically justified (e.g., what is the theoretical rationale for expecting *High-Social-Media* users to be more depressed?). In any case, we can report our outcomes as follows:

> Mean depression scores for the *High-Social-Media* group (M = 7.46; SD = 1.75) was statistically greater relative to the mean depression scores for the *Low-Social-Media* group (M = 5.64; SD = 2.79) following a one-sided two-sample Welch's test, $t(18.6) = 2.22, p = .02$. The null hypothesis can be rejected.

Paired *t*-test

> Use: Measuring the *same* sample at two time points (hence 'paired' or 'repeated').

For brevity, we can represent the difference score between the i^{th} estimate and the sample mean as delta, or $X_i - \hat{X} = X_\Delta$. The paired test statistic can be estimated as:

$$t_{paired} = \frac{\sum X_\Delta}{\sqrt{\frac{n(\sum X_\Delta^2)-(\sum X_\Delta)^2}{n-1}}}$$

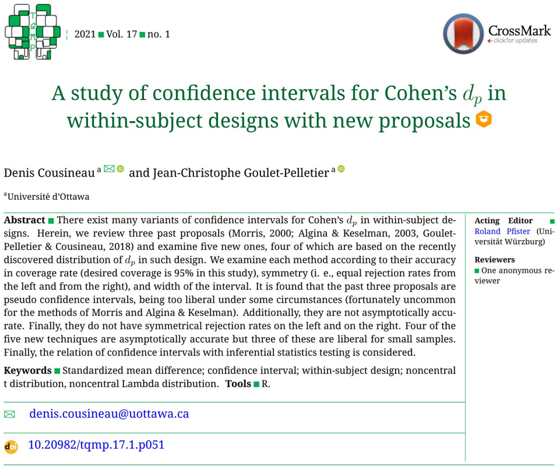
which can be simplified to

$$t_{paired} = \frac{\sum X_\Delta}{\frac{\sigma_\Delta}{\sqrt{n}}}$$

where σ_Δ is the standard deviation of X_Δ. Note that there is a single sample size parameter (n) because a paired test involves a single group of participants.

Similar to Student's and Welch's tests, the paired t-test assumes the data was sampled from a normal distribution. Contrary to independent tests, samples are not independent of each other as each subject is measured twice. Furthermore, confidence intervals are estimated using joint scores, *not* of the mean difference.

If you are interested in learning more about within-subject confidence intervals, have a look at the article by Cousineau and Pelletier (2021).



2021 ■ Vol. 17 ■ no. 1

CrossMark
click for updates

A study of confidence intervals for Cohen's d_p in within-subject designs with new proposals

Denis Cousineau[a] and Jean-Christophe Goulet-Pelletier[a]

[a]Université d'Ottawa

Abstract ■ There exist many variants of confidence intervals for Cohen's d_p in within-subject designs. Herein, we review three past proposals (Morris, 2000; Algina & Keselman, 2003, Goulet-Pelletier & Cousineau, 2018) and examine five new ones, four of which are based on the recently discovered distribution of d_p in such design. We examine each method according to their accuracy in coverage rate (desired coverage is 95% in this study), symmetry (i. e., equal rejection rates from the left and from the right), and width of the interval. It is found that the past three proposals are pseudo confidence intervals, being too liberal under some circumstances (fortunately uncommon for the methods of Morris and Algina & Keselman). Additionally, they are not asymptotically accurate. Finally, they do not have symmetrical rejection rates on the left and on the right. Four of the five new techniques are asymptotically accurate but three of these are liberal for small samples. Finally, the relation of confidence intervals with inferential statistics testing is considered.

Keywords ■ Standardized mean difference; confidence interval; within-subject design; noncentral t distribution, noncentral Lambda distribution. **Tools** ■ R.

Acting Editor ■ Roland Pfister (Universität Würzburg)

Reviewers ■ One anonymous reviewer

✉ denis.cousineau@uottawa.ca

doi 10.20982/tqmp.17.1.p051

Let's run through an example for a paired t-test.

Suppose we want to know whether engaging in solitary versus communal prayer influences subjective well-being. You recruit 20 individuals who you know pray on a regular basis. You ask half of them to pray in isolation for 30 minutes, then pray together for 30 minutes. You counter-balance this condition (the remaining half prays together for 30 minutes, than by themselves for 30 minutes). After each prayer session, you ask participants to complete a subjective well-being scale. You want to test the null hypothesis that there is no difference of praying in isolation versus communally in regards to subjective well-being.

Imagine we have collected the well-being scores for 20 participants after they prayed in **isolation** (3, 2, 3, 5, 2, 6, 2, 1, 4, 1, 1, 5, 5, 1, 5, 5, 4, 5, 2, 1) and after they prayed **communally** (5, 2, 1, 5, 4, 6, 6, 3, 1, 3, 8, 6, 7, 1, 6, 7, 1, 2, 4, 3). We can run a two-sided paired t-test to determine whether the mean difference in well-being scores across the two conditions is statistically different from the null.

```r
# Set up data
isolate  <- c(3, 2, 3, 5, 2, 6, 2, 1, 4, 1, 1, 5, 5, 1, 5, 5, 4, 5, 2, 1)
communal <- c(5, 2, 1, 5, 4, 6, 6, 3, 1, 3, 8, 6, 7, 1, 6, 7, 1, 2, 4, 3)

# Run the test (this time with 'paired = TRUE')
t.test(isolate, communal, paired=T, alternative="two.sided")
```

```
## 
##  Paired t-test
## 
## data:  isolate and communal
## t = -1.6446, df = 19, p-value = 0.1165
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -2.0453904  0.2453904
## sample estimates:
## mean of the differences 
##                    -0.9
```

> The subjective well-being of 20 students after praying in isolation (M = 3.15; SD =1.76) and communally (M = 4.05; SD =2.28) prayer was not significantly different, $t(19) = -1.64$, $p = .116$, $d = -0.37$. The null hypothesis can be retained.

Introducing effect sizes

- An effect size (ES) describes the magnitude of an observed effect and has primarily three purposes:

- They allow **interpretation** of data - researchers can assess whether a statistical effect may be practically 'meaningful'.
- One can **compare** between studies which incorporate effect sizes (e.g., assess reliability of similar procedures, run meta-analyses).

- ES's can be used for **inference** tests - you can check the viability of null hypotheses and generate confidence intervals to determine the precision of your estimate - *the narrower the interval, the more precise is your ES.*

- Not all ES's are the same - some are more biased than others, some are sensitive to sample size and become increasingly variable with larger samples (and/or sample inequality). Three properties of a good ES estimate are:

  - The ES is **unbiased** (the distribution of the ES is centered along the true population parameter).

  - The ES is **efficient** if it has less variance than it's competitors.

  - The ES is **consistent** if larger sample sizes lead to a convergence of the ES with the true population parameter.

A rough approximation of 'meaningfulness' is the magnitude of the effect size (d stands for 'standardized difference score').

- $d \leq 0.3$ - a small effect (not very meaningful)

- $.4 \leq d \leq .7$ - moderate effect

- $d \geq .8$ - a large effect (practically meaningful)

Effect size formulae (Cohen's *d*, Glass's Δ and Hedge's *g*)

- When contrasting across **large (*n*>20) and balanced samples**, the conventional effect size to report is Cohen's d for **samples**, which can be estimated as:

$$d_s^* = \frac{\hat{X}_1 - \hat{X}_2}{\sqrt{\frac{SD_1^2 + SD_2^2}{2}}}$$

Cohen's d-score tells us whether a statistically significant effect is practically meaningful. Note that Cohen's d_s^* assume samples have equal variance.

- *If*, homogeneity of variance can be assumed but samples are **unbalanced**, we may estimate can adjust the *pooled* standard deviation accordingly:

$$d_s = \frac{\hat{X}_1 - \hat{X}_2}{SD_{pooled}}$$

where

$$SD_{pooled} = \sqrt{\frac{SD_1^2(n_1 - 1) + SD_2^2(n_2 - 1)}{n_1 + n_2 - 2}}$$

. **Do not pool standard deviations** if the assumption of equal variances has not been met!

- Cohen's effect size is reliable when sample sizes are large (e.g., $n \geq 20$). For smaller samples, d_s^* can produce biased outcomes even if samples are otherwise balanced and there is homogeneity of variance.
- When variances are not homoscedastic (e.g., sample parameters have notably different standard deviations), an alternative approach for estimating effects is Glass's Δ.

- Instead of a pooled error term, the mean difference is divided by the standard deviation of a chosen group/condition (the 'control' group). So

$$\Delta_{Glass} = \frac{\hat{X}_{Treatment} - \hat{X}_{Control}}{SD_{Control}}$$

.

- In a paired design, the 'control' group is typically the pre-intervention condition. In an independent design, the 'control' group is typically the non-intervention condition.
- Because it is not always possible to designate a control group during a contrast (e.g., when comparing the effectiveness of two types of therapy on depression), Glass's Δ may not be applicable.

- It is also not recommended to use Glass's Δ when variances *are* homogeneous and/or there is no clearly specifiable 'control' group.

- When sample sizes are smaller than 20 and/or unbalanced, a recommended estimator of sample effect size is Hedge's g_s, which corrects for small samples.

$$g_s = d_s \times (1 - \frac{3}{4N - 9})$$

where N is the total sample size ($n_1 + n_2 = N$).

However, both Cohen's d and Hedge's g assume homogeneity of variance, which is not always the case across real-world scenarios.

- It has been recommended to use a variation of Hedge's formula (g_s^*) that does *not* involve pooling the standard deviation when comparing between independent samples for four reasons:

  - Does not rely on the homogeneity of variance assumption (as error is unpooled).

  - Ease of interpretation (roughly equivalent to Cohen's d_s).

  - ES variance remains consistent (even when normality is violated)/
  - Demonstrates convergence, even when variances are heterogeneous.

It is recommended to estimate Hedge's bias-corrected g_s^* score whenever running Welch's tests.

$$g_s^* = d_s^* \times \frac{N - 3}{N - 2.25} \times \sqrt{\frac{N - 2}{N}}$$

where d_s^* is based on Cohen's formulation described earlier.

- If you are interested in learning more about when and why other effect size measures may be used, you can go through the resources here

Table 1: Anxiety scores before and after intervention for 30 undergraduate students

Before intervention	After intervention
M = 11; SD = 2.5	M = 8.3; SD = 3.2

Note:
These are simulated scores that do not refer to any actual persons.

Activity

For each question, describe the steps you took to reach your outcome.

1. You have been provided IQ scores of 20 USP students. These are as follows:

   $$90, 93, 108, 94, 96, 101, 89, 95, 87, 100, 88, 107, 92, 105, 85, 110, 86, 106, 103, 104$$

   . You know that the average IQ for the population is 88. Is the mean IQ of the 20 USP students significantly *different* from the population mean?

2. You have been given the IQ scores of 20 FNU students, which are as follows:

   $$83, 80, 103, 98, 85, 100, 88, 90, 95, 104, 81, 96, 92, 93, 101, 97, 102, 86, 87, 84$$

   Is the mean difference in IQ between the 20 FNU students and 20 USP students statistically different from the null? If you find a statistically significant effect, report the effect size (Cohen's d_s).

3. You want to test whether a new psychotherapeutic procedure is effective in reducing anxiety across a sample of 30 University students. You record the anxiety scores of the 30 students before and after undergoing psychotherapy. The mean and standard deviations of the sample before and after treatment is provided near the top of this page.

Compute the effect size using *any one* of the three procedures discussed earlier (Cohen's d, Hedge's g, or Glass's Δ). Provide reasons for selecting the procedure that you did. Mention whether the difference is practically important based on the magnitude of the effect.

Submit your responses through the Moodle dropbox.