Effect size vs. Test statistic

JvD

2023-09-20

Test Statistic

A test statistic quantifies the departure of the data vs. a hypothesized value. In the case of a difference in means, the test statistic (t) looks at the difference between the observed difference in means ($\bar{X_A} - \bar{X_B}$), and the hypothesized difference in means ($H_0 : \bar{\mu_A} - \bar{\mu_B} = 0$). This difference is then standardized by the pooled standard error SE_p:

$$SE_p = \sqrt{\frac{S_p^2}{n_A} + \frac{S_p^2}{n_B}},$$

with the pooled variance S_p^2 calculated as:

$$S_p^2 = \frac{(n_A - 1)s_A^2 + (n_B - 1)s_B^2}{n_A + n_B - 2}.$$

The full formula for t is as follows, when we fill in 0 for the hypothesized difference between means:

$$t_{n_A+n_B-2} = \frac{(\bar{X_A} - \bar{X_B}) - 0}{SE_p}.$$

The standard error of t is affected by the sample size ($n_A + n_B$: the greater the sample size, the smaller the standard error, which then leads to a higher t-value. So, in order to reach an extreme t-value (i.e., further away from 0), we need to either observe a large mean difference, have a large sample size (or both). Consider we observe a mean difference in IQ-scores of -5 (pooled standard deviation = 20), we set $\alpha = 0.05$ and conduct a two-sided test. Whether this observed difference will lead to rejection of the null hypothesis depends on the sample size that was used. First we use ($n_A = n_B = 5$):

```
observedDifference <- -5
nA <- nB <- 5
df <- nA + nB - 2
sdA <- 20
sdB <- 20

observedPooledSD <- sqrt(((nA - 1) * sdA^2 + (nA - 1) * sdB^2) / (nA + nB - 2))
pooledSE <- sqrt(observedPooledSD / nA + observedPooledSD / nB)
tStat <- observedDifference / pooledSE

tStat
```

```
## [1] -1.767767
```

```
# Convert the t-statistic to a p-value by calculating area under the curve of t-distribution:
# We multiply this area by 2 because we are doing a 2-sided test
pt(tStat, df = nA + nB - 2, lower.tail = TRUE) * 2
```

```
## [1] 0.1150771
```

```r
# Alternatively, we look at the critical value corresponding to our alpha (0.05)
criticalValue <- qt(0.025, nA + nB - 2)
criticalValue
```

```
## [1] -2.306004
```

```r
# Then, we look whether our observed t-value is more extreme than this critical value:
tStat < criticalValue # If true, we achieve significance
```

```
## [1] FALSE
```

If we now do exactly the same, but only for an increased sample size ($n_A = n_B = 100$):

```r
observedDifference <- -5
nA <- nB <- 100
df <- nA + nB - 2
sdA <- 20
sdB <- 20

observedPooledSD <- sqrt(((nA - 1) * sdA^2 + (nA - 1) * sdB^2) / (nA + nB - 2))
pooledSE <- sqrt(observedPooledSD / nA + observedPooledSD / nB)
tStat <- observedDifference / pooledSE

tStat
```

```
## [1] -7.905694
```

```r
# Convert the t-statistic to a p-value by calculating area under the curve of t-distribution:
# We multiply this area by 2 because we are doing a 2-sided test
pt(tStat, df = nA + nB - 2, lower.tail = TRUE) * 2
```

```
## [1] 1.825085e-13
```

```r
# Alternatively, we look at the critical value corresponding to our alpha (0.05)
criticalValue <- qt(0.025, nA + nB - 2)
criticalValue
```

```
## [1] -1.972017
```

```r
# Then, we look whether our observed t-value is more extreme than this critical value:
tStat < criticalValue # If true, we achieve significance
```

```
## [1] TRUE
```

So what we see now is that if we increase the sample size, while keeping everything else constant, the t-value increases quite a bit. This is not necessarily a bad thing, because we also grow more certain as we collect a greater sample size. Achieving significance more easily with a larger sample size reflects this increase in certainty. However, achieving a significant difference does *not* mean having found a large effect size.

Effect Size

To keep the focus solely on the size of the effect that was found, we can instead look at *effect sizes*. For a difference in means, we can look at d or r. These effect sizes are standardized, meaning that the original measurement scale is irrelevant (this makes it easy to compare effect sizes across studies, for instance). Furthermore, these effect sizes are not influenced by the sample size:

```r
observedDifference <- -5
nA <- nB <- 5
df <- nA + nB - 2
sdA <- 20
```

```r
sdB <- 20

observedPooledSD <- sqrt(((nA - 1) * sdA^2 + (nA - 1) * sdB^2) / (nA + nB - 2))
pooledSE <- sqrt(observedPooledSD / nA + observedPooledSD / nB)
tStat <- observedDifference / pooledSE

# Convert the t-statistic to effect size d:
d <- 2*tStat / sqrt(df)
d
```

```
## [1] -1.25
```

```r
# Or we can compute r:
r <- sqrt(tStat^2/(tStat^2 + df))
r
```

```
## [1] 0.5299989
```

And repeated with a higher sample size:

```r
observedDifference <- -5
nA <- nB <- 100
df <- nA + nB - 2
sdA <- 20
sdB <- 20

observedPooledSD <- sqrt(((nA - 1) * sdA^2 + (nA - 1) * sdB^2) / (nA + nB - 2))
pooledSE <- sqrt(observedPooledSD / nA + observedPooledSD / nB)
tStat <- observedDifference / pooledSE

tStat
```

```
## [1] -7.905694
```

```r
# Convert the t-statistic to effect size d:
d <- 2*tStat / sqrt(df)
d
```

```
## [1] -1.123666
```

```r
# Or we can compute r:
r <- sqrt(tStat^2/(tStat^2 + df))
r
```

```
## [1] 0.4898196
```

While the effect sizes are still somewhat affected by the sample size, it is much less affected by sample size than the t- and p-value. We therefore generally also include an effect size when reporting on a study, preferably also including a confidence interval to express our uncertainty about the value of the effect size (just as we would do with a mean difference, or t-statistic).

Practical Significance

To illustrate why both effect size and test statistic are important, consider the following table:

	Big effect size	Small effect size
$p < \alpha$	Evidence for difference, and of practical importance	Evidence for difference, but might not be interesting
$p > \alpha$	no effect observed	no effect observed

While the test statistic quantifies evidence against the null hypothesis (and informs our decision to reject or not reject it), it does not guarantee a practically significant effect. Similarly, finding a large effect does not guarantee a significant effect.

Imagine you are researching some new experimental treatment, and find it significantly improves chances of survival, it can be the case that it only increases chances of survival by 0.01%. When this treatment is extremely expensive to implement at a hospital, the hospital might then opt to not implement this treatment because it is not worth it. Of course, whether an effect is practically significant is a lot more debatable than whether an effect is statistically significant, since the latter is simply a function of what α was used, while the former involves all sorts of things that might be hard to quantify (and statisticians do not like things that cannot be quantified...).