Effect size and power analysis

Effect size

A measure of the **magnitude** (though usually not the sign) of the effect.

In a one-sample test, it might be defined in terms of the difference between the sample statistic (e.g., sample mean or median) and the null hypothesis parameter (e.g., the hypothesized population mean or median).

In a paired-sample test, it might be defined in terms of the sample mean (or median) of differences between the paired samples and a null hypothesis parameter (mean or median) of zero difference.

More generally: a measure of the impact of the independent variable (or predictor) on the dependent variable (or outcome)

Relation to other measures

Effect size is completely independent of sample size, and not directly related to p-values or statistical significance.
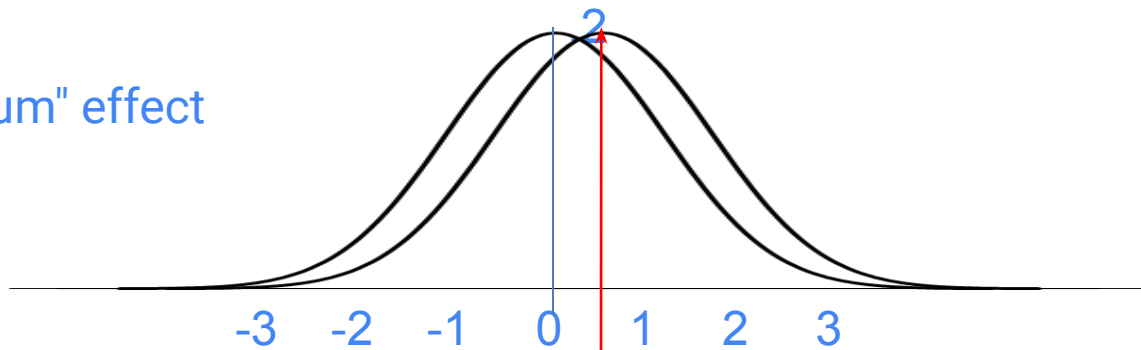
Thus both small and large effect sizes can give significant or non-significant results, depending on other factors.
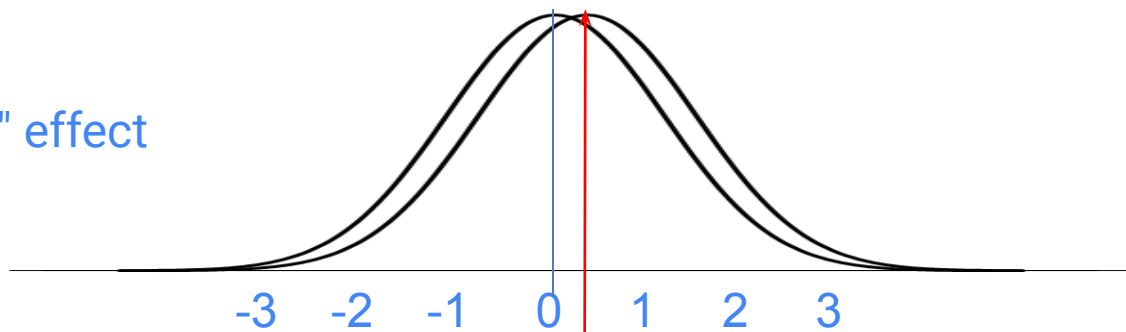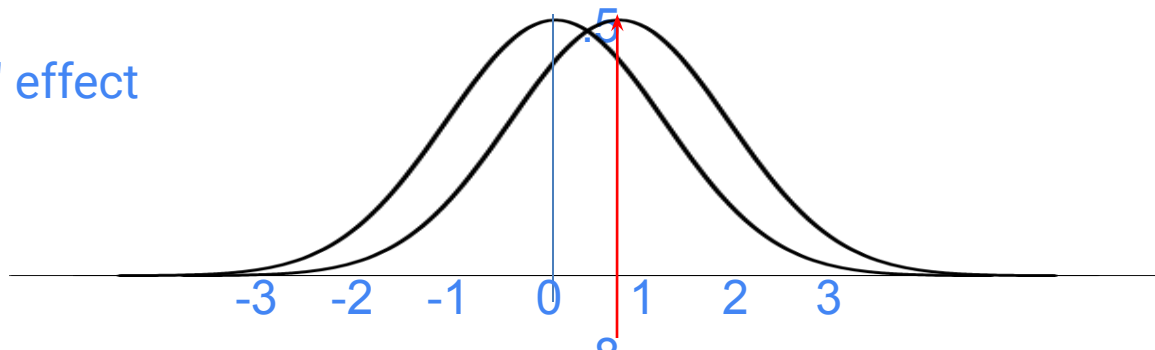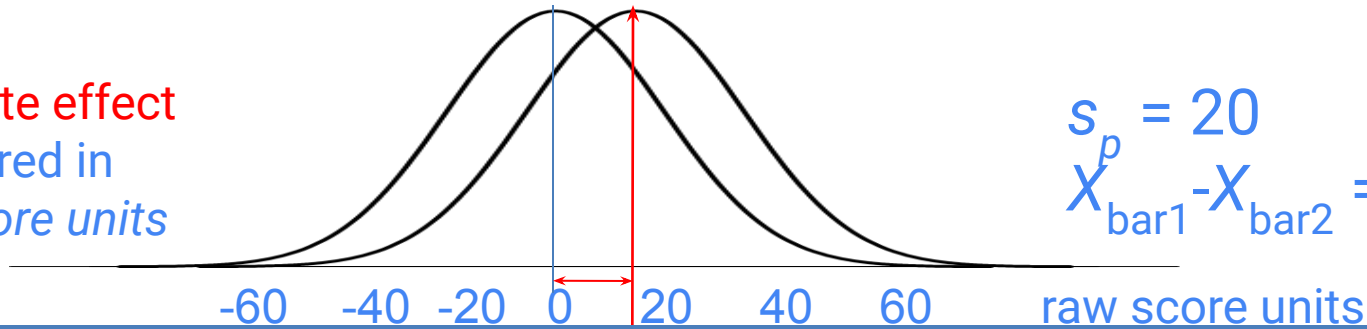
"small" effect $d = .2$

"medium" effect $d = .5$

"large" effect $d = .8$

Absolute effect measured in *raw score units*

$s_p = 20$

$X_{bar1} - X_{bar2} = 16$

raw score units

Relative effect measured in *std. dev. units*

16 / 20 = .80

standard deviation units
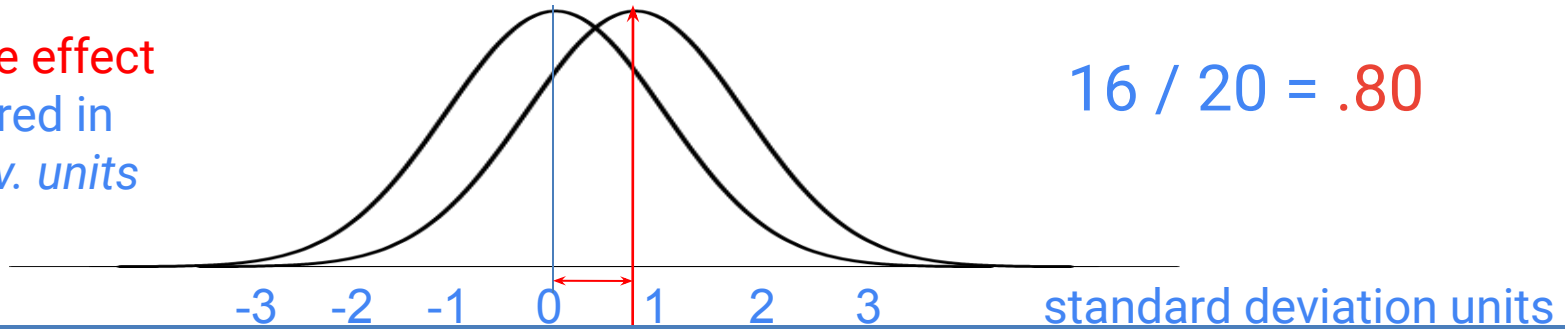
Test statistic (e.g., *t*) measured in *std. error units*
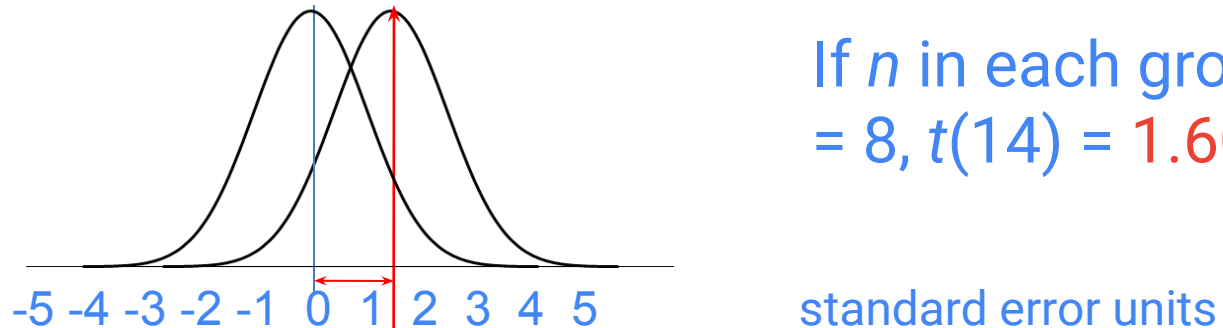
If *n* in each group = 8, $t(14) = 1.60$

standard error units

Sample size and effect size

- As effect size d (etc.) increases, the expected t (etc.) increases linearly
- As sample size n increases, the expected t increases as a function of $\sqrt{n / 2}$.

Power

- The degree to which we are able to detect a significant difference
- The probability that the null hypothesis will be rejected, given a particular sample size, effect size, and α
- The proportion of the alternative hypothesis distribution beyond the critical value for the given α

Power as a function of expected *t* (δ) and effect size
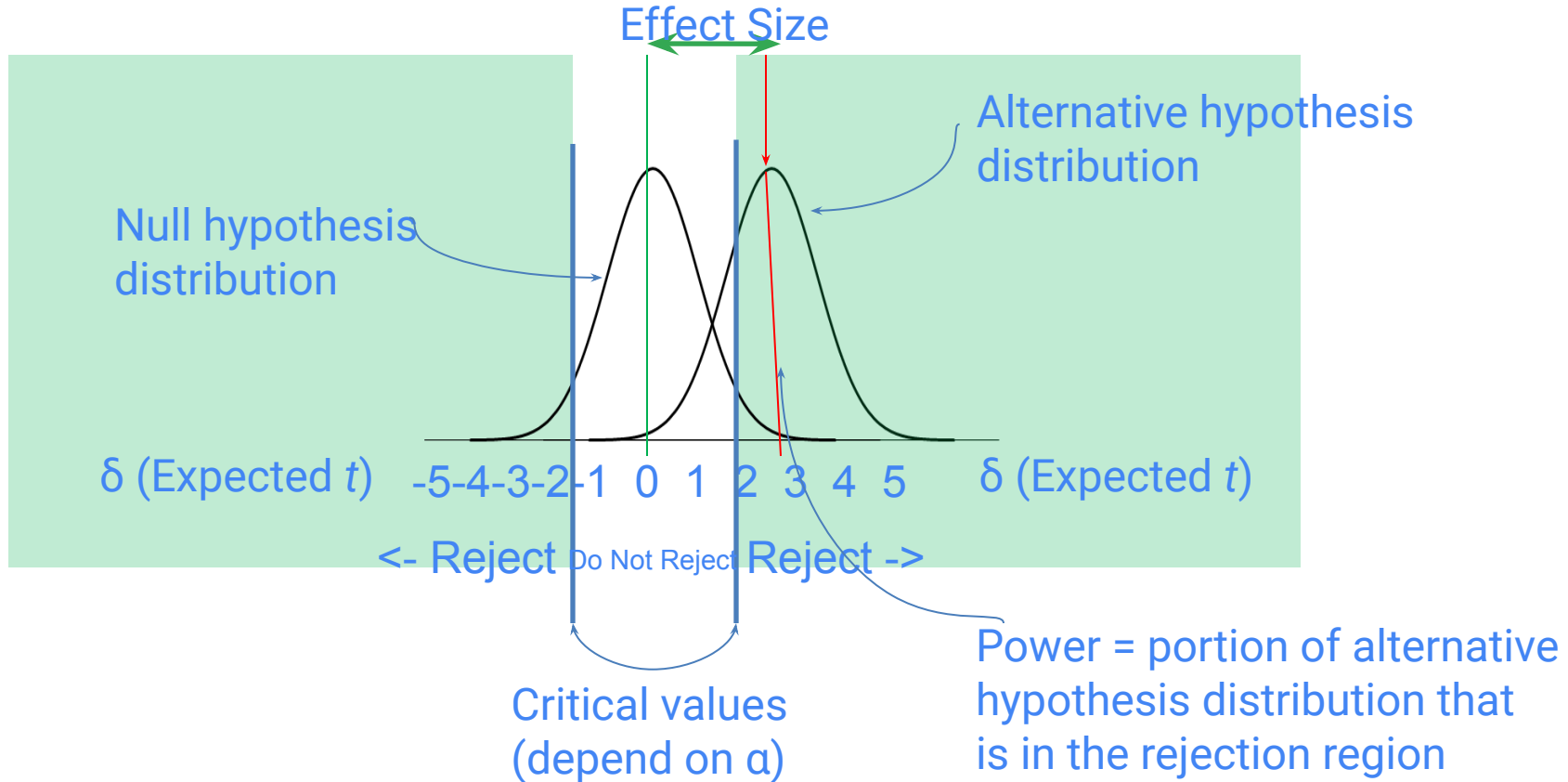
α = prop. of null hypothesis distribution that lies beyond critical value

Power = prop. of alternative hypothesis distribution that lies beyond critical value

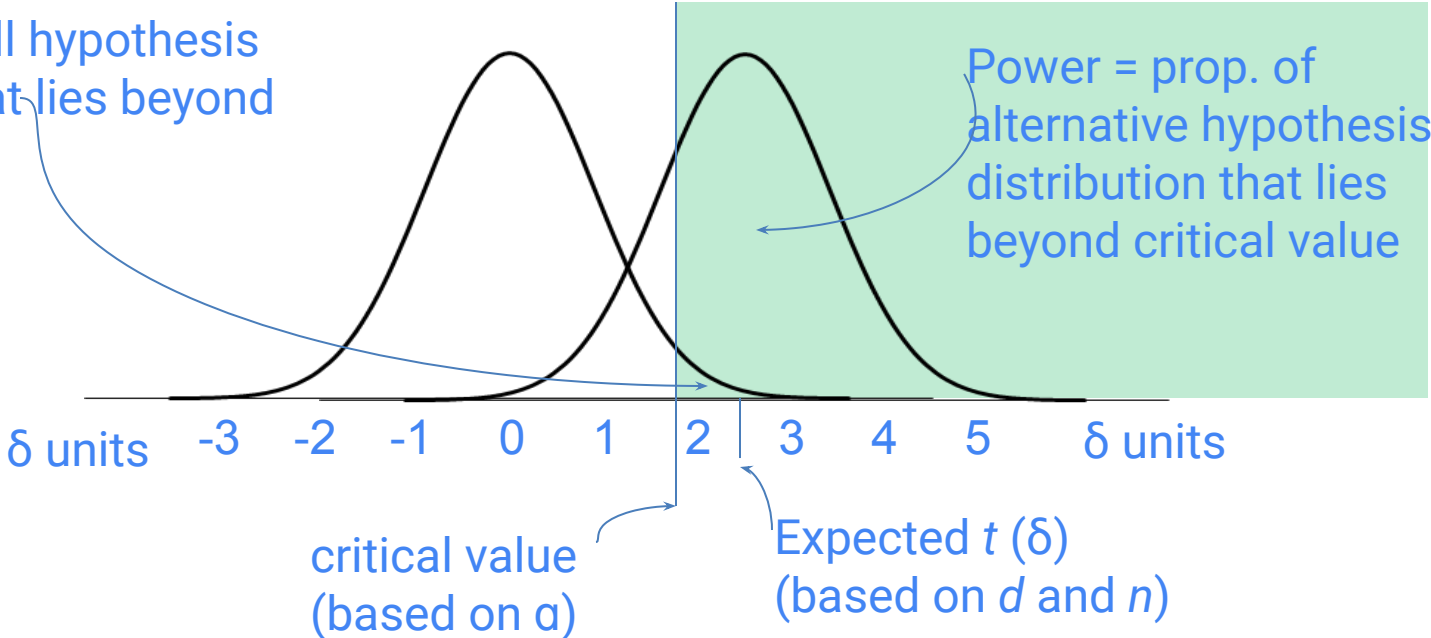δ units   -3   -2   -1   0   1   2   3   4   5   δ units

critical value (based on α)

Expected t (δ) (based on d and n)

Power as a function of δ and α

δ	.10	.05	.02	.01
1.00	.26	.17	.09	.06
1.10	.29	.20	.11	.07
1.20	.33	.22	.13	.08
1.30	.37	.26	.15	.10
1.40	.40	.29	.18	.12
1.50	.44	.32	.20	.14
1.60	.48	.36	.23	.17
1.70	.52	.40	.27	.19
1.80	.56	.44	.30	.22
1.90	.60	.48	.34	.25
2.00	.64	.52	.37	28
2.10	.68	.56	.41	.32
2.20	.71	.60	.45	.35
2.30	.74	.63	.49	.39
2.40	.78	.67	.53	.43
2.50	.80	.71	.57	.47
2.60	.83	.74	.61	.51
2.70	.85	.77	.65	.55
2.80	.88	.80	.68	.59
2.90	.90	.83	.72	.63
3.00	.91	.85	.75	.66
3.10	.93	.87	.78	.70
3.20	.94	.89	.81	.73
3.30	.95	.91	.84	.77
3.40	.96	.93	.86	.80
3.50	.97	.94	.88	.82
3.60	.98	.95	.90	.85
3.70	.98	.96	.92	.87
3.80	.98	.97	.93	.89
3.90	.99	.97	.94	.91
4.00	.99	.98	.95	.92

To increase power

- increase α (at the cost of Type I error),
- increase d,
- increase precision of measurement,
- decrease variability (e.g., by using a within-subjects design), or
- increase n.

Power analysis in R (`pwr` library)

`pwr.2p.test`	two proportions (equal n)
`pwr.2p2n.test`	two proportions (unequal n)
`pwr.anova.test`	balanced one-way ANOVA
`pwr.chisq.test`	Pearson chi-square test
`pwr.f2.test`	Generalized linear model
`pwr.p.test`	one proportion
`pwr.r.test`	correlation
`pwr.t.test`	t-test (one-sample, two-sample, paired)
`pwr.t2n.test`	t-test (two-sample, unequal n)

```
> library(pwr)
> pwr.t.test(n = 25, d = .5, sig.level = .05, type = "paired")

     Paired t test power calculation

              n = 25
              d = 0.5
      sig.level = 0.05
          power = 0.6697077
    alternative = two.sided

NOTE: n is number of *pairs*
```

```
> pwr.t.test(d = .5, sig.level = .05, power = .8,
             type = "paired")

        Paired t test power calculation

              n = 33.36713
              d = 0.5
      sig.level = 0.05
          power = 0.8
    alternative = two.sided

NOTE: n is number of *pairs*
```

Fisher's exact test

Testing for independence of categorical variables

Fisher's exact test tests for the independence of two categorical (usually boolean) variables.

The null hypothesis that the population conditional probabilities are equal. To compute the null hypothesis p-value, we compute the probability of each count table that has conditional probabilities *as extreme* or *more extreme* than those observed in the sample, then sum those probabilities.

This test makes obsolete the earlier *Pearson chi-square test*, one of the earliest null hypothesis tests, which is merely approximate (i.e., asymptotically correct), now that we have digital computers and can compute the factorial possibilities quickly...

The lady tasting tea

The "lady" (believed to be Muriel Bristol, a phycologist) claimed that she could tell whether milk or tea was added first.

> a = her guess as to whether milk or tea was added first
> b = whether milk or tea was added first

This is something of a class shibboleth in the British Isles:

- upper-class people usually put the tea in first
- lower-class people (and the Irish) usually put the milk in first

(cf. George Orwell's essay *A Nice Cup of Tea*.)

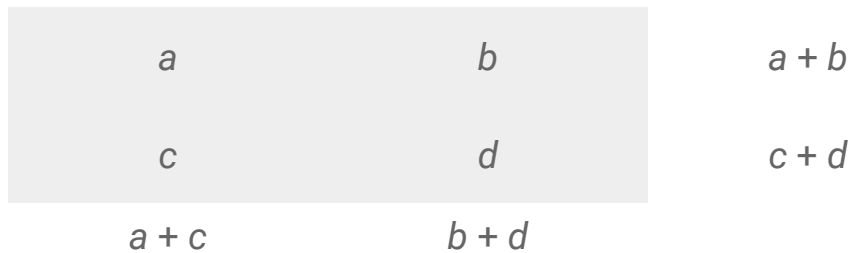# The lady tasting tea

Guess \ Reality	Milk first	Tea first
Milk first	7	2
Tea first	2	8

Ô1	Ô2	Ô3	Ô4	Ô5	Ô6	Ô7	Ô8	Ô9	Ô10
9 0	8 1	7 2	6 3	5 4	4 5	3 6	2 7	1 8	0 9
1 9	2 8	3 7	4 6	5 5	6 4	7 3	8 2	9 1	10 0

"this large or larger"

		$a + b$
a	b	$a + b$
c	d	$c + d$
$a + c$	$b + d$	

$$p = [(a + b)! \, (c + d) \, (a + c)! \, (b + d)!] \, / \, [(a + b + c + d)! \, a! \, b! \, c! \, d!]$$

```
> tab <- cbind(c(7, 2), c(2, 8))
> tab
     [,1]  [,2]
[1,]    7     2
[2,]    2     8
```

```
> fisher.test(tab)

    Fisher's Exact Test for Count Data

data:  tab
p-value = 0.02301
alternative hypothesis: true odds ratio is not equal to 1
95 percent confidence interval:
   1.116329 214.179160
sample estimates:
odds ratio
  11.64655
```