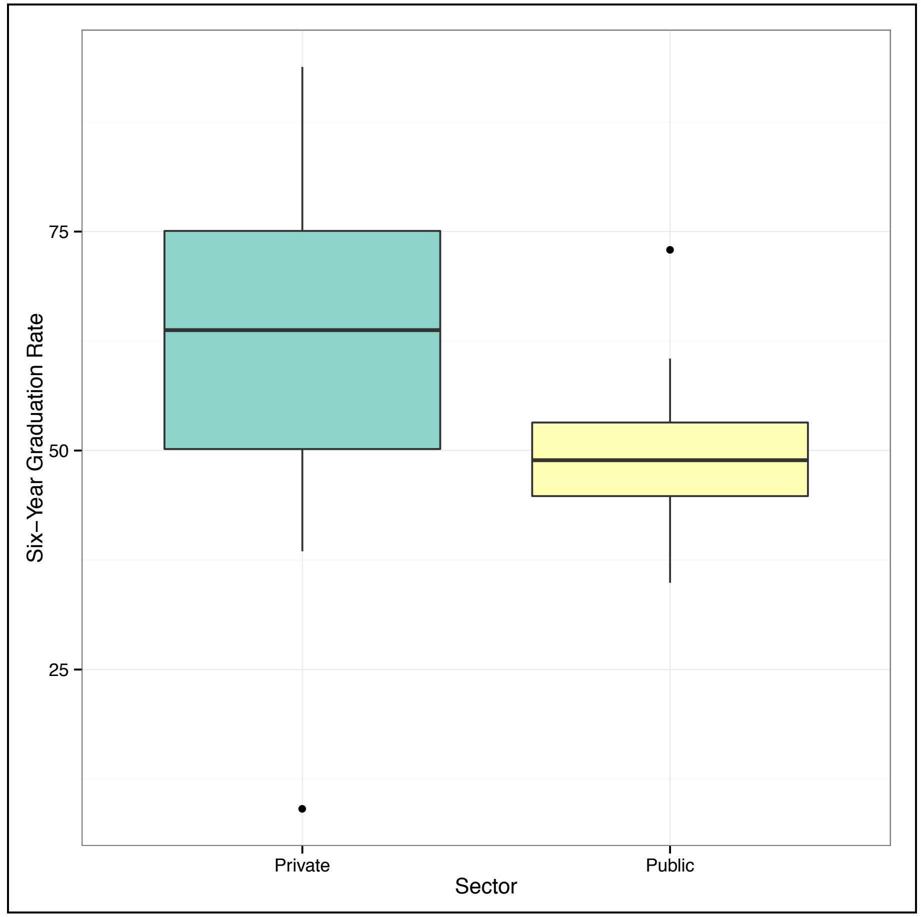
Effect Size

Graduation data for graduating class of 2011 from collegeresults.org

- Data have been cleaned and only include institutions that reported a graduation rate
- The outcome (`gradRate`) is the proportion of first-time, full-time, bachelor's or equivalent degree-seeking students who graduate within 6 years (freshmen class of 2006).

http://collegecompletion.chronicle.com

Use the `MN-Colleges-2012.csv` data to examine whether there are mean differences in the graduation rates between public and private schools




```
group: Private not-for-profit
  vars  n mean    sd median trimmed   mad min  max range skew kurtosis   se
1    1 24 63.1 19.08  63.75   64.17 19.27 9.1 93.8  84.7 -0.7     0.56 3.89
-------------------------------------------------------------------------------
-
group: Public
  vars  n  mean    sd median trimmed  mad  min  max range skew kurtosis   se
1    1 11 50.35 10.15   48.9   49.56 6.97 34.9 72.9    38 0.67    -0.12 3.06
```

```
# Fit the statistical model using the lm() function
> lm.1 = lm(gradRate ~ public, data = mn)

# Examine the results using the summary() function
> summary(lm.1)

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   63.104      3.446  18.313   <2e-16 ***
public       -12.759      6.147  -2.076   0.0458 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 16.88 on 33 degrees of freedom
Multiple R-squared:  0.1155,   Adjusted R-squared:  0.08868
F-statistic: 4.309 on 1 and 33 DF,  p-value: 0.04579
```

To examine whether there is a mean difference in graduation rates between public ($M = 50.4$, $SD = 10.2$) and private ($M = 63.1$, $SD = 19.1$) colleges and universities, the six-year graduation rate for $n = 35$ colleges and universities (located in Minnesota) was regressed on a dummy-coded predictor for sector (private universities = 0; public universities = 1). The results of this analysis were statistically reliable, $F(1, 33) = 4.31$, $p = 0.046$, indicating that there is likely a difference in the mean graduation rates. Differences in sector accounted for 11.55% of the variation in graduation rates.

Questions of Interest

According to Kirk (2001), questions regarding group differences are often three-fold:

(a) Is an observed effect real or should it be attributed to chance?
(b) If the effect is real, how large is it?
(c) Is the effect large enough to be useful?

ANOVA only provides an answer to the first!

Kirk, R. E. (2001). Promoting good statistical practices: Some suggestions. *Educational and Psychological Measurement*, *61*(2), 213–218.

Effect Size

Effect size is a term used to describe a family of indices that characterize the *extent to which sample results diverge from the expectations specified in the null hypothesis*.

Effect size can provide an answer to the second two questions.

Effect size can also provide a method by which to compare the results between different studies.

The *APA Publication Manual* (American Psychological Association, 2009, p. 25) suggests that for a reader to fully understand the importance of any finding, it is almost **always necessary to include some index of effect size or strength of relationship** in the Results section.

It also suggests that failure to report effect sizes may be found by editors to be one of the **defects in the design and reporting of results**.

Types of Effect Size

There are many different types of effect size that can be reported. Two common categories of effect size are:

- **Distance measures** of effect, and
- **Variance accounted for measures** of effect

When the analysis focuses on only **two groups**, it is typical to report a *distance* measure of effect.

When the analysis focuses on **more than two groups**, it is typical to report a *variance accounted for* measure of effect.

Consider what we **know** about the average graduation rates after conducting an regression analysis:

- Based on the *F*-test, there is statistical evidence of an **average difference** in graduation rates between the population of public and private colleges and universities.
- The negative regression coefficient for public suggests that public colleges and universities have a **lower** graduation rate than private colleges and universities
- The magnitude of the coefficient suggests that public colleges and universities have a graduation rate that is, on average, **12.76% lower** than private colleges and universities

If we would have carried out an ANOVA rather than a regression, we would only know there was a likely difference...not how nor to what magnitude the difference was.

Sample evidence (e.g., the regression coefficients and sample means) can provide some insight into how the two groups differ

- Public schools in Minnesota have an average graduation rate of 50.4%
- Private schools in Minnesota have an average graduation rate of 63.1%

Direction of the effect

Private schools have a higher graduation rate, on average, than public schools.

Magnitude: of the effect

The difference is, on average, 12.7%.

Interpretation

Private schools have a graduation rate that is 12.7% higher than public schools, on average.

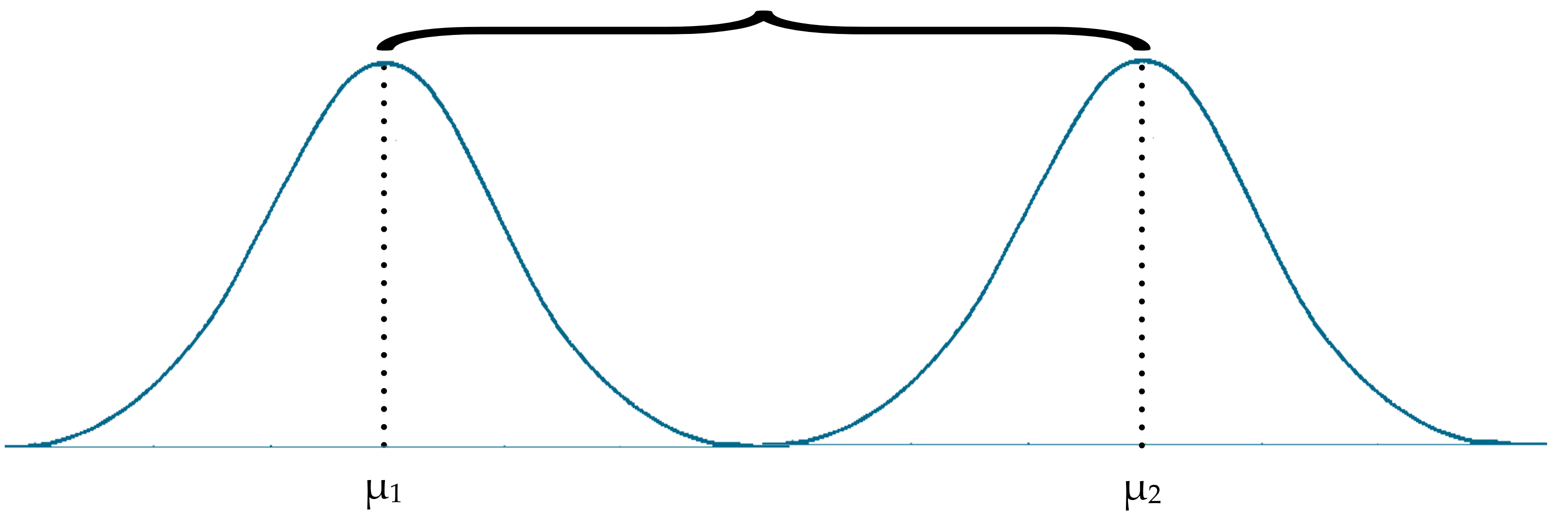
Based on the assumptions of the model,

- The conditional population distributions are both normally distributed.
- The conditional population distributions have the same variance.

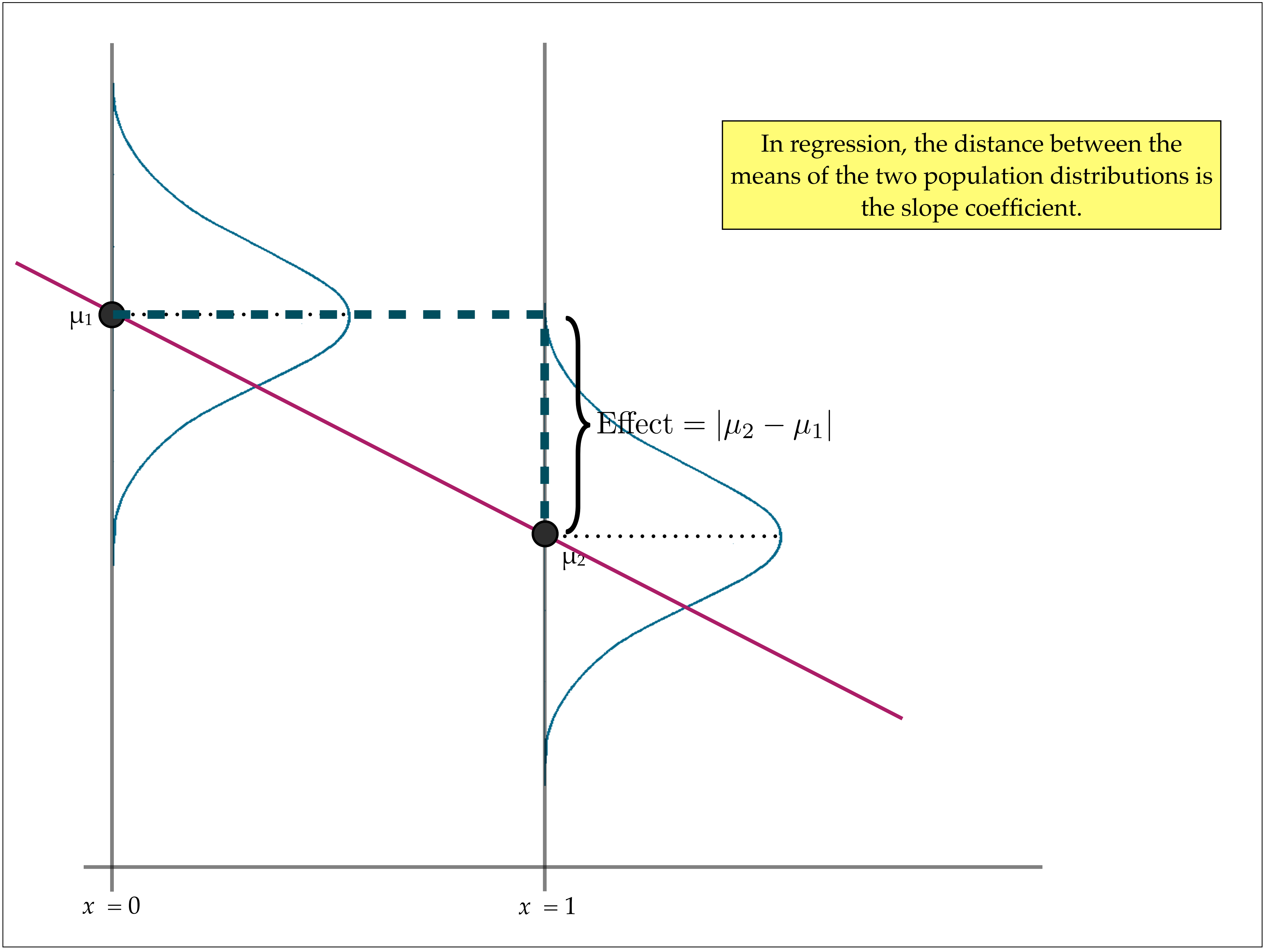
Based on the rejection of the null hypothesis,

- It is likely that the conditional population distributions do not have the same mean.

The distance between the means of the two population distributions is called the distance measure of effect.



$$\text{Effect} = |\mu_2 - \mu_1|$$

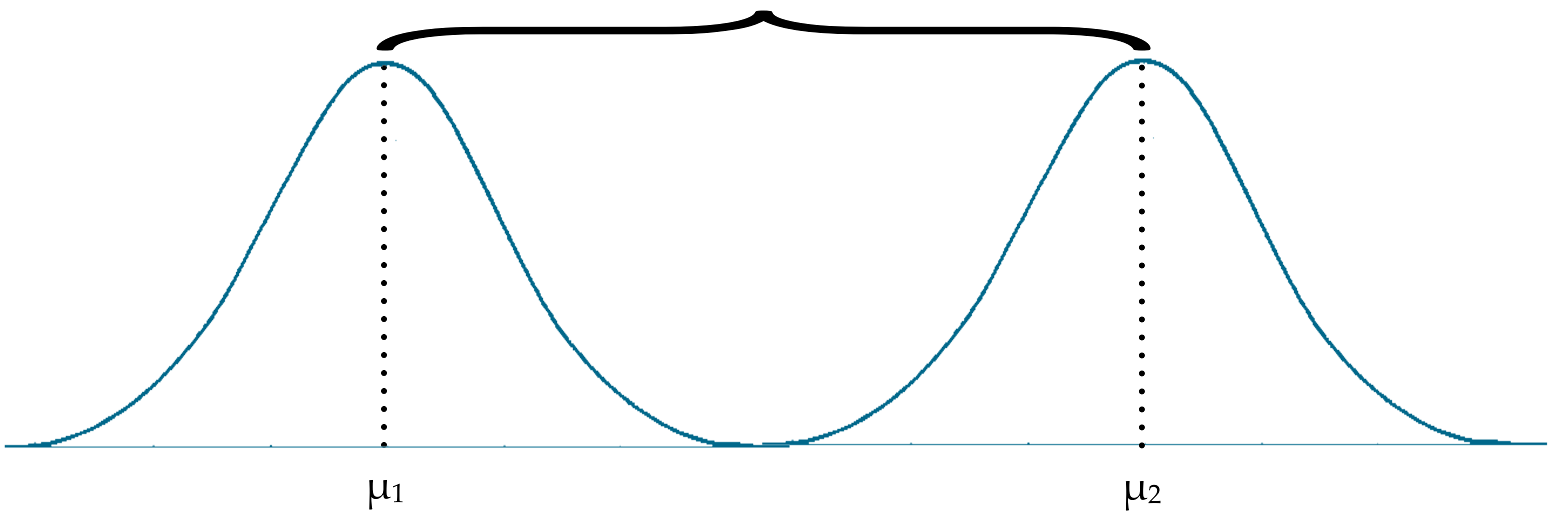

In regression, the distance between the means of the two population distributions is the slope coefficient.
μ1
μ2
Effect $= |\mu_2 - \mu_1|$
$x = 0$
$x = 1$

The **distance measure of the effect** between two populations is estimated using the absolute value of the *sample mean difference*.

$$
\begin{aligned}
\widehat{\text{Effect}} &= \left|\bar{Y}_{\text{Private}} - \bar{Y}_{\text{Public}}\right| \\
&= |63.1 - 50.4| \\
&= |12.7| \\
&= 12.7
\end{aligned}
$$

This is called a **point estimate** of the population effect.

Based on the sample data, our best guess for this distance is 12.7%.



Based on the sample data, we believe that μ_1 is approximately 50.4%

Based on the sample data, we believe that μ_2 is approximately 63.1%

Sometimes the **raw metric** that the response variable is measured in is not easily interpretable...this will imply that the estimated effect will also be not easily interpretable.

For example,

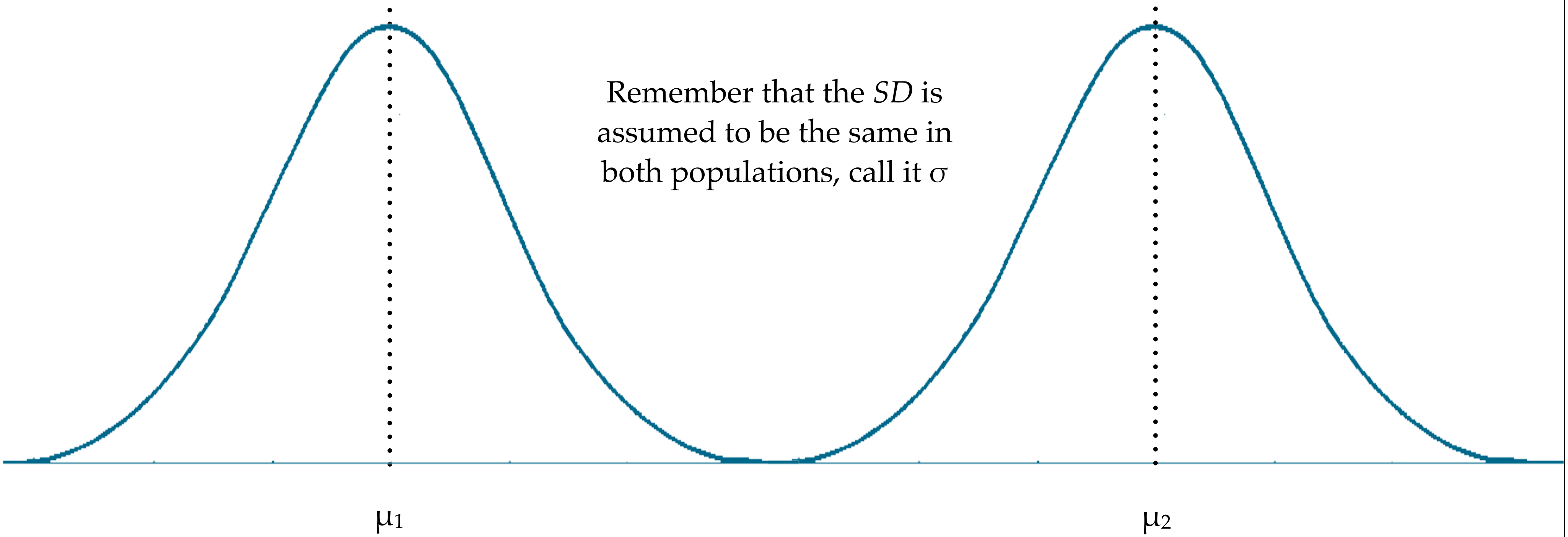
- Consider comparing the average MCA-III scores between the St. Cloud Public School District (M = 852.5) and the Richfield Public School District (M = *845.7*).

The *estimated effect* is an average difference of 6.8 points.

How do we interpret this difference? Is it small? moderate? large?

Many researchers choose to report a *standardized effect* rather than a raw effect

To standardize a value, you divide that value by the standard deviation of the distribution the value is a part of.



To standardize μ_1 we find:

$$\frac{\mu_1}{\sigma}$$

To standardize μ_2 we find:

$$\frac{\mu_2}{\sigma}$$

The distance measure of effect is then:

$$\frac{\mu_2}{\sigma} - \frac{\mu_1}{\sigma}$$

Since the denominators are equal: $\dfrac{\mu_2 - \mu_1}{\sigma}$

Since we are measuring distance, it is easier to use the absolute value in the numerator.

$$\frac{|\mu_2 - \mu_1|}{\sigma}$$

This standardized measure of the effect is referred to as Cohen's *d* (named after Jacob Cohen who invented it).

Even though the metric of graduation rate is easily interpretable in its raw form, we will find the estimate of Cohen's d (for fun!)

We need three values:

- An *estimate* of the mean for population 1
- An *estimate* of the mean for population 2
- An *estimate* of the population standard deviation (which is assumed to be the same in both populations)

The first two estimates we can easily obtain from the sample data

$$\bar{Y}_{\text{Private}} = 63.1$$

$$\bar{Y}_{\text{Public}} = 50.4$$

To paraphrase *Naughty By Nature*, the last estimate, while "that's not that simple".

We have two estimates for the *SD* of the population, one from each sample

$$\hat{\sigma}_{\text{Private}} = 19.1$$

$$\hat{\sigma}_{\text{Public}} = 10.2$$

Option 1: Simple Average

Compute the simple mean of the two estimates and use that as the estimate for the standard deviation of the population. This is what Cohen proposed originally.

$$\hat{\sigma} = \frac{19.1 + 10.2}{2} = 14.7$$

The estimate for Cohen's d would then be

$$\hat{d} = \frac{|63.1 - 50.4|}{14.7} = 0.86$$

Option 2: Weighted Average

A second option is to compute a weighted average based on the sample sizes. This provides a better estimate when the sample sizes are unequal. This solution was proposed by Hedges and is called Hedges' *g*.

$$\hat{\sigma} = \frac{19.1(24) + 10.2(11)}{35} = 16.30$$

The estimate for Cohen's *d* would then be

$$\hat{d} = \frac{|63.1 - 50.4|}{16.30} = 0.78$$

Option 3: Root Mean Square Error

A third option is to use the RMSE from either the summary() or anova() output.

```
Residual standard error: 16.88 on 33 degrees of freedom
Multiple R-squared:  0.1155,   Adjusted R-squared:  0.08868
F-statistic: 4.309 on 1 and 33 DF,  p-value: 0.04579
```

$$\hat{\sigma} = 16.88$$

The estimate for Cohen's *d* would then be

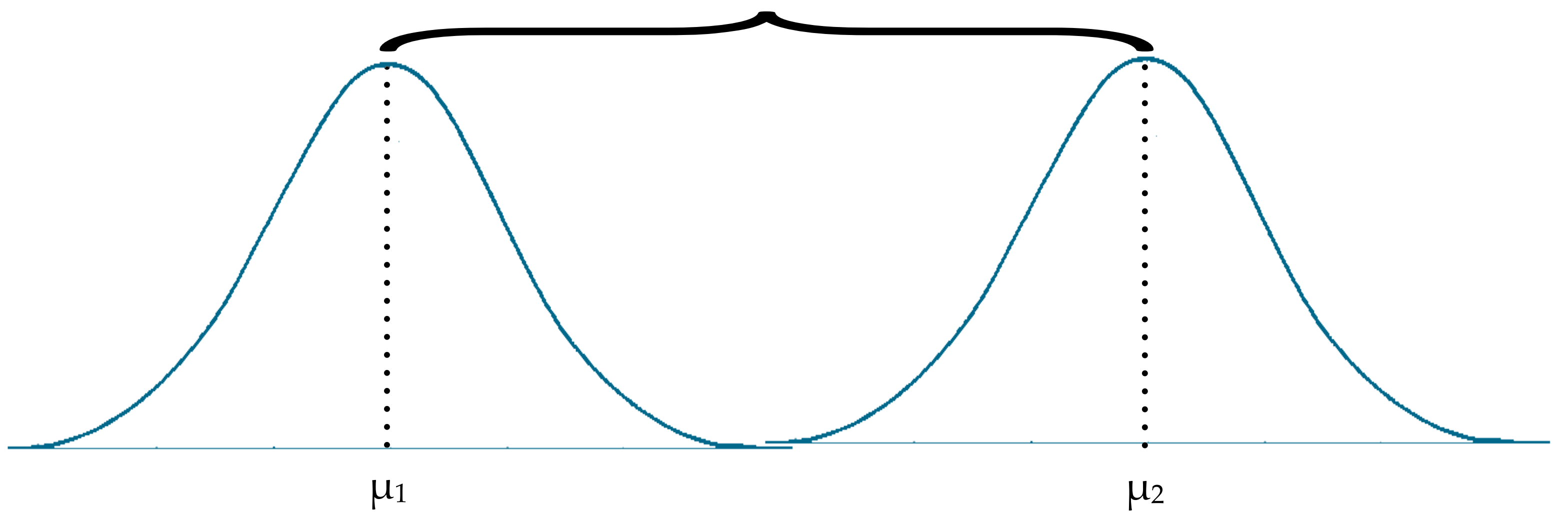
$$\hat{d} = \frac{|63.1 - 50.4|}{16.88} = 0.76$$

In this example, both estimates for the population standard deviation are similar and therefore produce estimates for Cohen's *d* that are relatively close to one another.

How do we interpret a Cohen's *d* estimate of 0.76?

Recall that standardizing a metric changes the metric from the raw metric to the **standard deviation metric**. Rather than measuring how different the graduation rates were in the raw percentages (%) we are now measuring how different they are in standard deviation units.

Based on the sample data, our best guess for this distance is 0.76 standard deviations



The average graduation rate for private schools is 0.76 standard deviations higher than the average graduation rate for public schools.

Computing Cohen's *d* Using R

We will use the `smd()` function from the **MBESS** library to compute the estimate of the standardized effect.

```
# Load MBESS library (you may need to install it first)
> library(MBESS)

# Point estimate for the standardized effect
> smd(
    Mean.1 = 63.1,
    s.1 = 19.1,
    n.1 = 24,
    Mean.2 = 50.4,
    s.2 = 10.2,
    n.2 = 11
    )

[1] 0.7512442
```

This is based on Option #2

Interpreting the Magnitude of *d*

Cohen provided "rules of thumb" to help applied researchers interpret the magnitude of standardized difference

- $d = 0.2$ corresponds to a small effect size
- $d = 0.5$ corresponds to a medium effect size
- $d = 0.8$ corresponds to a large effect size

The terms "small," "medium," and "large" are relative, not only to each other, but to the area of behavioral science or even more particularly to specific content and research method being employed in any given investigation (Cohen, 1965, p. 23).

Cohen, J. (1965). *Statistical power analysis for the behavioral sciences*. New York: Academic Press.

Let's go back to the distance measure of effect in the raw metric.

We conclude that the estimated graduation rate for private colleges and universities in Minnesota is, on average, 12.7% higher than that for public colleges and universities.

This statement is, in itself, a **model or hypothesis** about the true population mean difference in graduation rates between public and private universities.

It can be expressed mathematically as,

$$\mu_{\text{Private}} - \mu_{\text{Public}} = 12.7$$

It is stressed that this model is only a hypothesis—the true population difference cannot be *determined* based on sample data.

Any model or hypothesis, however, can be empirically examined using the sample data.

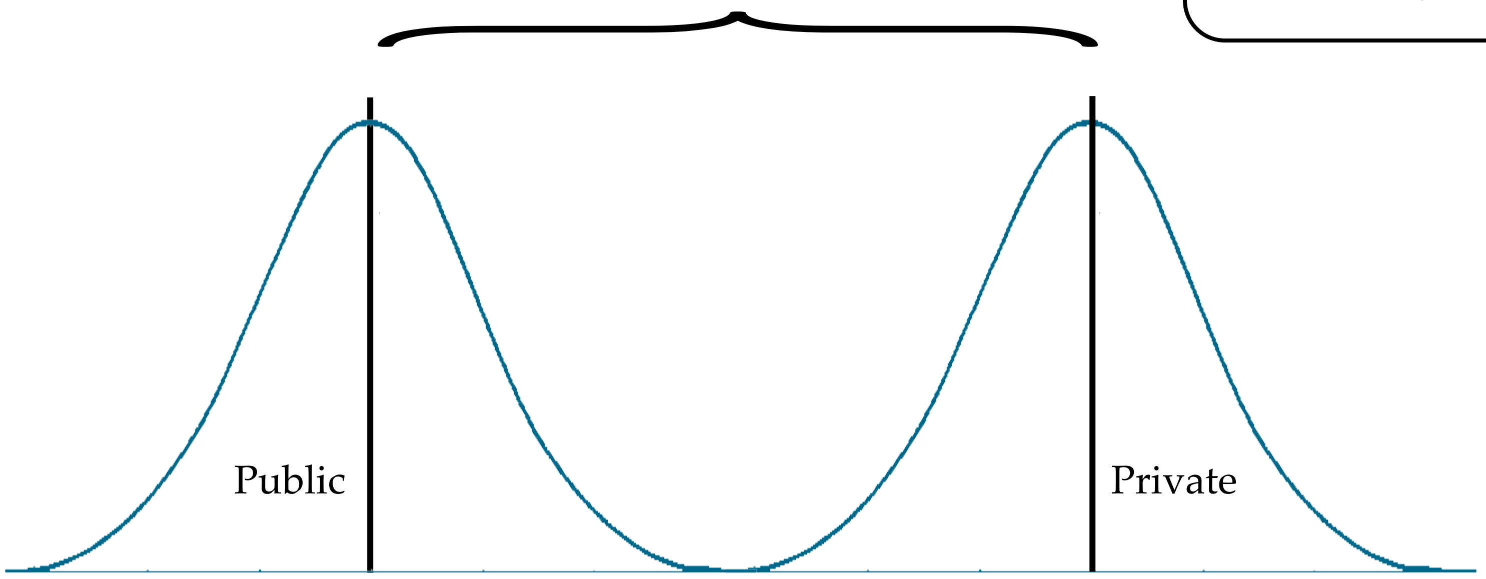
For example, we earlier tested a very specific model (the null model)

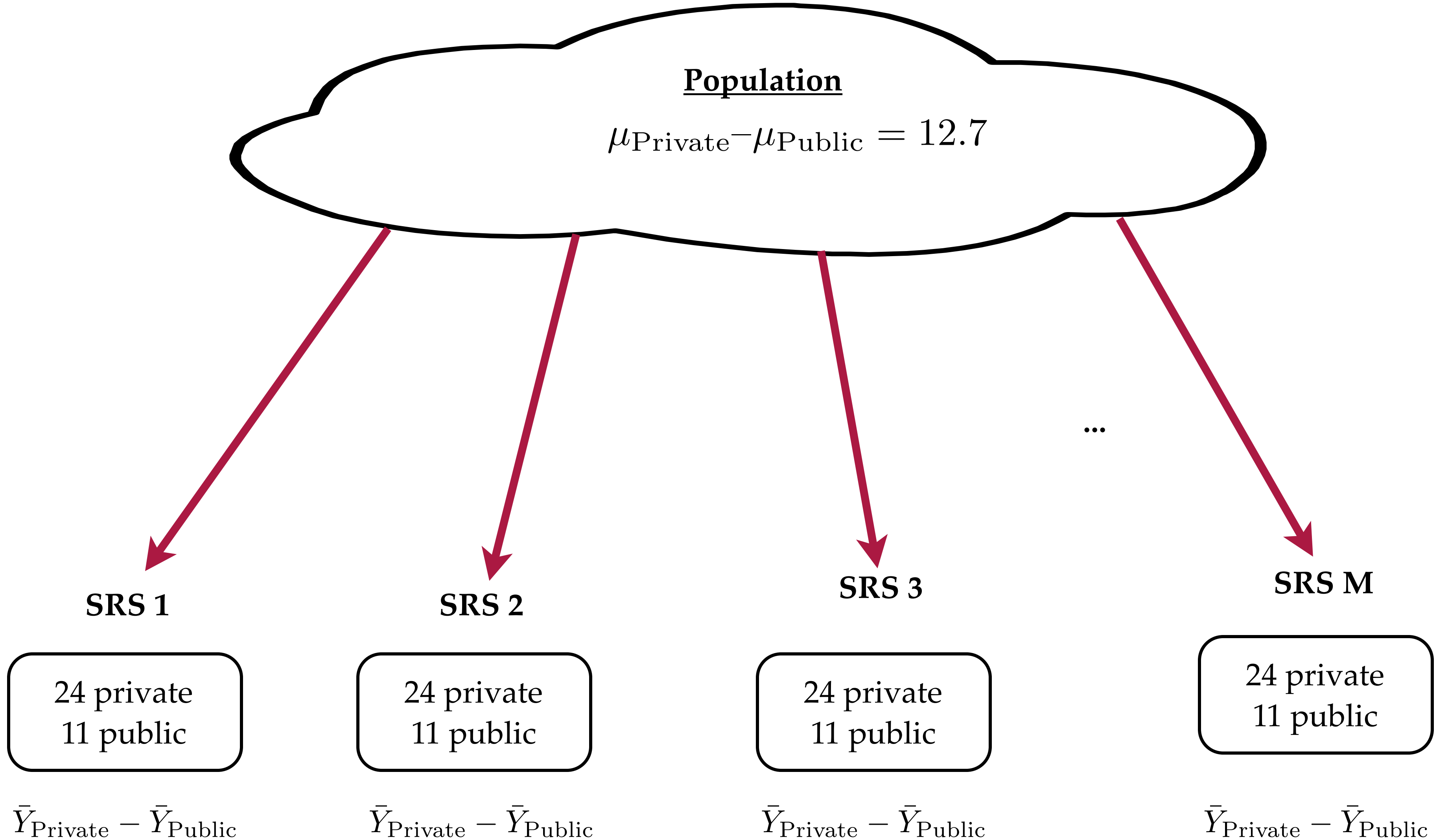
$$H_0 : \mu_{\text{Private}} - \mu_{\text{Public}} = 0$$

Let's assume that the first model, which has a parameter value of 12.7, is the true population model.

$$\mu_{\text{Private}} - \mu_{\text{Public}} = 12.7$$

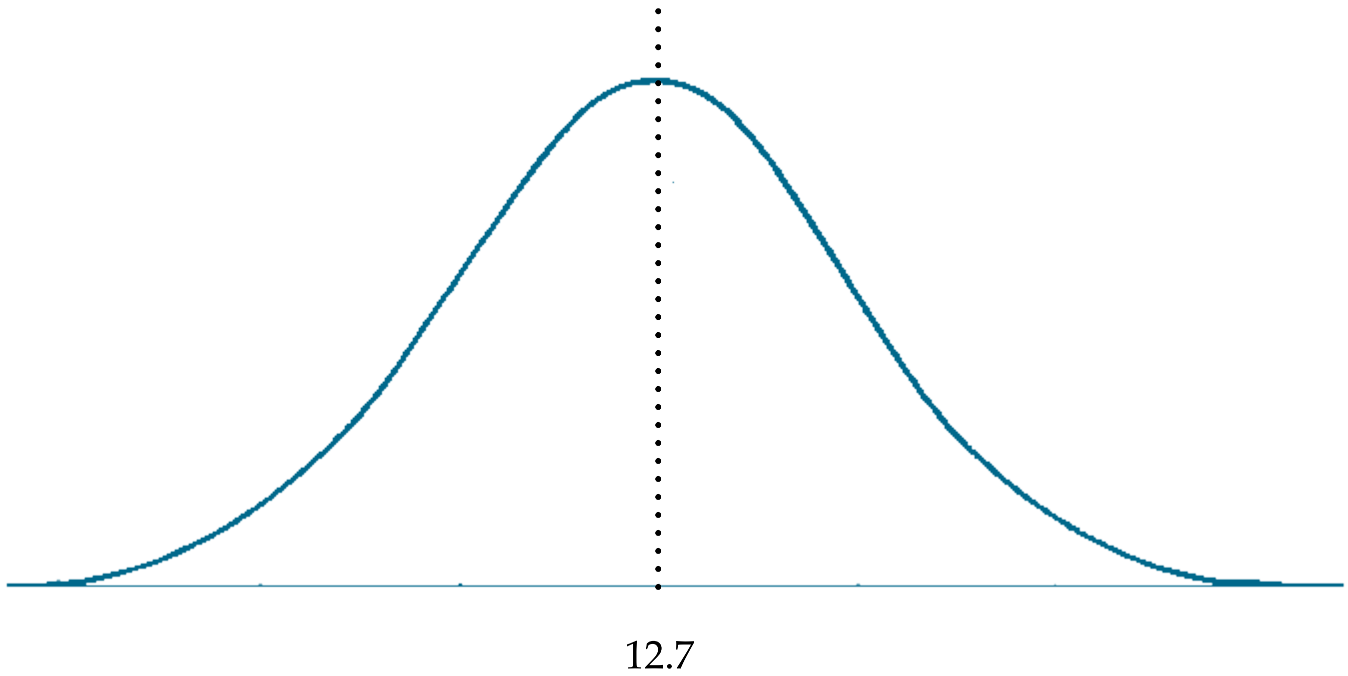
Note that this model, or hypothesis, is directional.



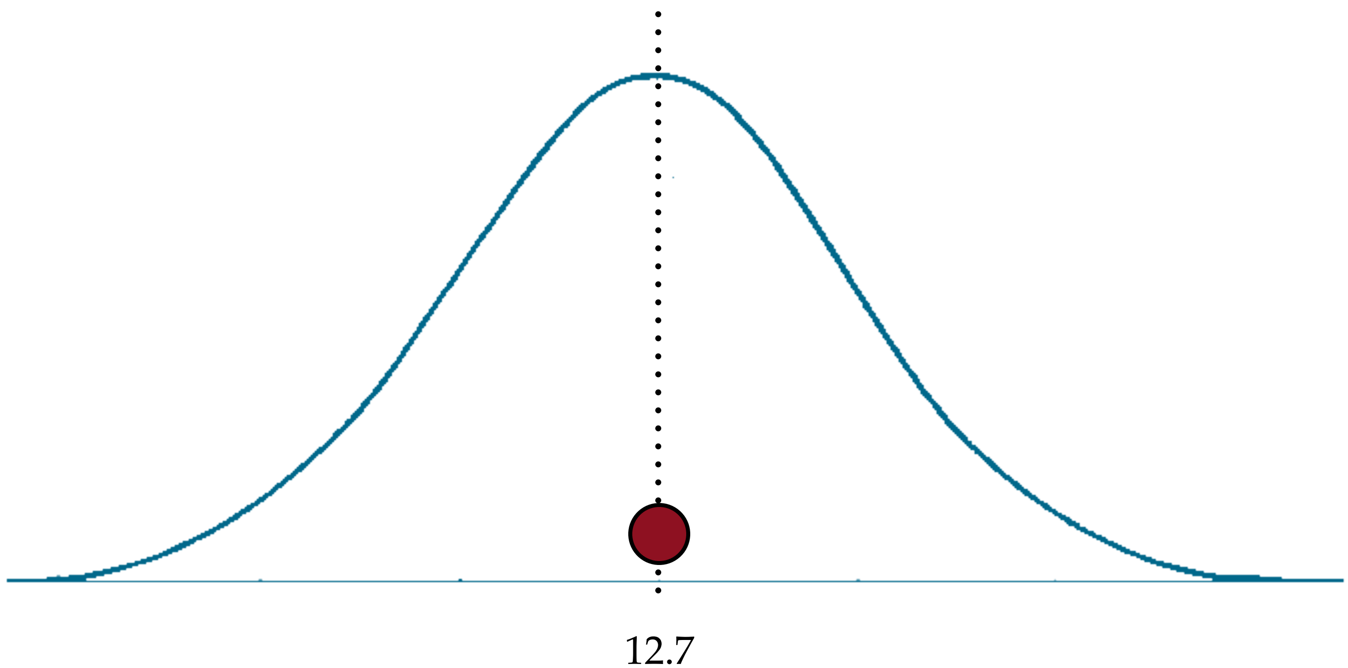


Now let us hypothetically draw many, many random samples the same size and make up of our observed data from this model.

If we plot all of these mean differences we get the following distribution:



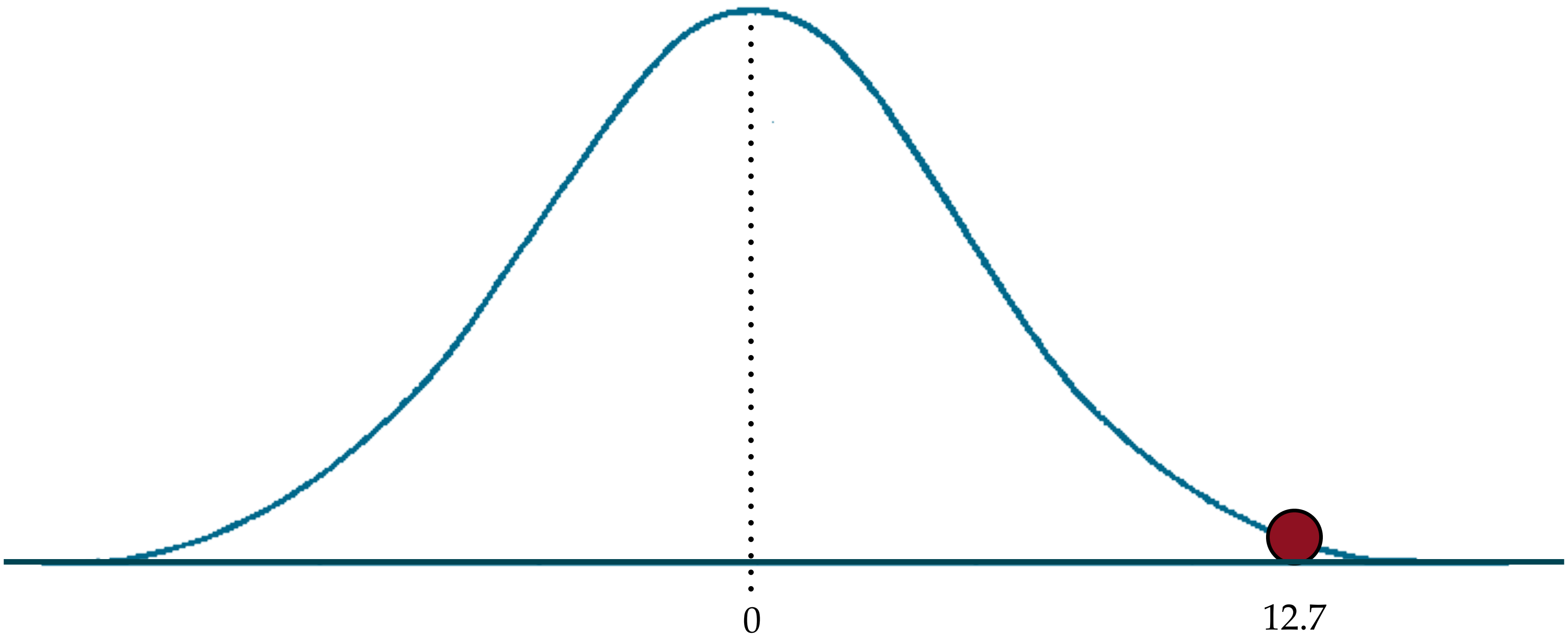
Remember that the observed mean difference in our original sample was also 12.7, it is marked in the distribution with a point.



Based on where the observed value lies in this distribution, does the assumed model have a high or low likelihood of reproducing the observed result?

Compare that to the distribution when we sample from a population where the null model is true.

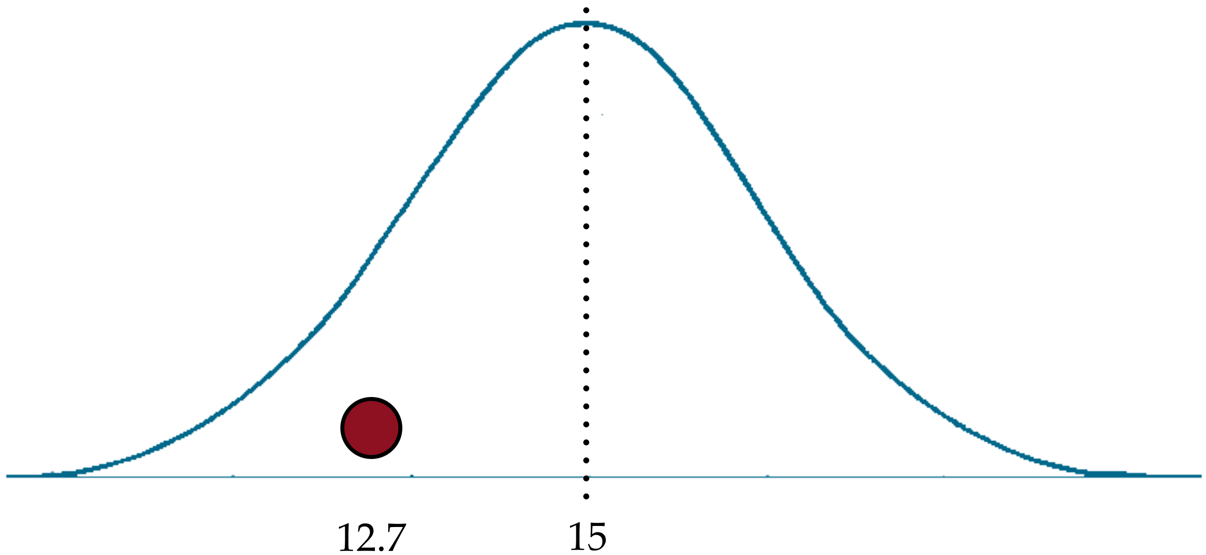
$$\mu_{\text{Private}} - \mu_{\text{Public}} = 0$$



Based on where the observed value lies in this distribution, does the null model have a high or low likelihood of reproducing the observed result?

What about a population model where the assumed mean difference is 22?

$$\mu_{\text{Private}} - \mu_{\text{Public}} = 15$$



Based on where the observed value lies in this distribution, does this model have a high or low likelihood of reproducing the observed result?

The null model, which has a **parameter value of 0**, reproduces the observed mean difference of 12.7 with a low likelihood. The *p*-value produced when we test this model is 0.0458.

The model which has a **parameter value of 12.7**, reproduces the observed mean difference of 12.7 with a high likelihood. The *p*-value produced when we test this model is 0.50.

The third model, which has a **parameter value of 15**, reproduces the observed mean difference of 12.7 with a fairly high likelihood. The *p*-value produced when we test this model is 0.36.

These examples should illustrate the following:

- Different models (or hypotheses) for the population mean difference will generate different distributions of the mean differences.
- Some models will generate sampled data that reasonably reproduce the observed result. These models can be considered as plausible.
- Other models will **not** generate sampled data that seem to reasonably reproduce the observed result. These models should probably be considered as implausible.

Might there be other models that would also reproduce the observed mean difference of 12.7 with reasonably high likelihood?

Confidence Intervals

We can identify the parameters of the models that would reproduce the data with a high likelihood by producing a confidence interval.

```
# Confidence interval for the raw effect
> confint(lm.1)

                2.5 %     97.5 %
(Intercept)  56.09345 70.1148840
public      -25.26419 -0.2532387
```

Look at the line showing the CI for public.

The 95% CI for the effect is [–25.3, –0.3]

The values in this interval represent parameter values for the model that produce the observed result with high likelihood.

The observed result was a mean difference of 12.7.

The observed result should be exactly in the middle of the CI.
(That is the value of the parameter that will reproduce the data with the highest likelihood.)

The exact middle of our CI is –12.7.

In order to interpret the interval based on how we have been writing the models/hypotheses,

$$\mu_{\text{Private}} - \mu_{\text{Public}}$$

flip the signs on the CI (and reverse the limits).

The new CI will be [0.3, 25.3].

With 95% confidence, we believe that the true population effect is that private schools have a higher graduation rate than public schools in Minnesota and that difference could be as low as 0.3% or as high as 25.3%.

Note that we can use the CI to also examine other models or hypotheses regarding the size of the effect.

Model: $\mu_{Private} - \mu_{Public} = 30$

The CI of [0.3, 25.3] represent the likely parameter values.
The value 30 is not a parameter that will likely reproduce the observed result of 12.7.
Thus the model is **not** a good candidate model.

We would reject the model.

Confidence Interval for the Standardized Effect

What about an interval estimate for the *standardized* effect?

To compute the CI for the standardized effect, we will use the `ci.smd()` function from the **MBESS** library.

```
# Confidence interval for the standardized effect
> ci.smd(smd = 0.7512442, n.1 = 24, n.2 = 11)

$Lower.Conf.Limit.smd
[1] 0.009779527

$smd
[1] 0.7512442

$Upper.Conf.Limit.smd
[1] 1.48203
```

The CI for Cohen's d is [0.01, 1.48].

With 95% confidence, we believe that the true population effect is that private schools in Minnesota have a higher graduation rate than public schools, on average and that difference could be as little as 0.01 of a standard deviation or as much as 1.48 of a standard deviation.