No Statistician Left Behind, or How (not) to Measure the Student Achievement Gap

Thomas Pietraho

Bowdoin College
Problem

I have always been bothered by the use of the following types of statistics to describe social problems:

- the difference in the percentage of students from group A and students from group B who go on to college,
- the difference in the percentage of students from group A and students from group B who graduate high school,
- the difference in the percentage of people from group A and people from group B who are proficient at some skill,
- the difference in the percentage of people from group A and people from group B who exceed some benchmark.
Mainly because they have the following “solutions:”

**Problem**
There is a significant difference in percentage of students from group A and students from group B who graduate high school.

**Solution**
Let everyone from group A and group B graduate. The difference in graduation rates is completely eliminated.

**Alternate Solution**
Fail every from group A and group B. No one graduates. The difference in graduation rates is completely eliminated.
Mainly because they have the following “solutions:”

**Problem**
There is a significant difference in percentage of students from group A and students from group B who graduate high school.

**Solution**
Let everyone from group A and group B graduate. The difference in graduation rates is completely eliminated.

**Alternate Solution**
Fail every from group A and group B. No one graduates. The difference in graduation rates is completely eliminated.
Sources

- The Lion’s Paw: an anonymous blog
The Texas Success Story

Table 1: White-Black achievement gap; percentage of white students minus percentage of black students passing state math exam. Grades 3 through 8 and 10.

<table>
<thead>
<tr>
<th>Year</th>
<th>Point Gap</th>
</tr>
</thead>
<tbody>
<tr>
<td>1994</td>
<td>35.2</td>
</tr>
<tr>
<td>1995</td>
<td>35.4</td>
</tr>
<tr>
<td>1996</td>
<td>30.0</td>
</tr>
<tr>
<td>1997</td>
<td>25.4</td>
</tr>
<tr>
<td>1998</td>
<td>21.4</td>
</tr>
<tr>
<td>1999</td>
<td>19.7</td>
</tr>
<tr>
<td>2000</td>
<td>16.6</td>
</tr>
<tr>
<td>2001</td>
<td>13.2</td>
</tr>
<tr>
<td>2002</td>
<td>10.0</td>
</tr>
</tbody>
</table>

Source: Texas Education Agency
"In my state, 73 percent of the white students passed the math test in 1994, while only 38 percent of African-American students passed it. So we made that the point of reference. We had people focused on the results for the first time – not process, but results. And because teachers rose to the challenge, because the problem became clear, that gap has now closed to 10 points. Because every child can learn, you’ve just got to focus the attention and the resources when necessary.”
Question

That’s all fine and good in practice, but how does it work in theory?
The Statistics of Exams

Exam distributions tend to be bell-shaped, or Gaussian.
The size of each bar represents the number of students within a score range.
Bell curves are smoothed-out versions of bar charts.
Observation

The area of a “slice” represents the percentage of students within a score range.
Observation

The area of a “slice” represents the percentage of students within a score range.
Observation

The area of the “slice” to the right of a passing score represents the percentage of students who passed the exam.
A Tale of Two Groups

Group A

Average test score: 60
A Tale of Two Groups

Group B

Average test score: 40
The usual way to measure the difference in performance between these groups is to measure the difference in the two averages, in this case 60 − 40, or a 20 point difference.
A Tale of Two Groups

Groups A and B

The differences between these two groups begin to disappear as the differences between the average scores decrease.
The differences between these two groups begin to disappear as the differences between the average scores decrease.
Comparing Populations, the Old Fashioned Way

**Problem:** Rarely do we know the theoretical distribution for a population exactly. Rather, we know details about a sample of the population:

![Histogram of Group A](attachment:image.png)

**Note:** Mean of sample can vary from theoretical.
**Problem:** Rarely do we know the theoretical distribution for a population exactly. Rather, we know details about a sample of the population:

![Group A](image)

**Note:** Mean of sample can vary from theoretical.
**Problem:** Rarely do we know the theoretical distribution for a population exactly. Rather, we know details about a sample of the population:

![Histogram of Group A](image)

**Note:** Mean of sample can vary from theoretical.

- Sometimes by quite a bit!
Comparing Populations, the Old Fashioned Way

**Problem:** Rarely do we know the theoretical distribution for a population exactly. Rather, we know details about a sample of the population:

![Histogram of Group A](image)

**Note:** Mean of sample can vary from theoretical.

**Natural Question:** After gathering data, “predict” the true mean of the underlying population.
**Problem:** Determine the difference between the means of two populations A: mean = 150, $\sigma_1 = 10$ and B: mean = 50, $\sigma_2 = 10$.

**Theory:**

![Groups A and B](image)
**Problem:** Determine the difference between the means of two populations A: mean = 150, $\sigma_1 = 10$ and B: mean = 50, $\sigma_2 = 10$.

**Practice:**

**Question:** How far off could we be? Quite a bit! (But this is not likely)

**Refined Question:** Predict difference with a measure of certainty.
Comparing Populations, the Old Fashioned Way

**Problem:** Determine the difference between the means of two populations

**Groups A and B**

**Difference of Means**

Difference: 93.97
Comparing Populations, the Old Fashioned Way

**Problem:** Determine the diff. between the means of two populations

**Groups A and B**

**Difference of Means**

Difference: 102.08
**Problem:** Determine the difference between the means of two populations.
**Problem:** Determine the difference between the means of two populations.

- **Groups A and B**
- **Difference of Means**
  - Difference: 99.08
**Problem:** Determine the diff. between the means of two populations
Comparing Populations, the Old Fashioned Way

**Problem:** Determine the difference between the means of two populations

![Groups A and B](image)

**Difference of Means**

Difference: 101.11
**Problem:** Determine the diff. between the means of two populations

Groups A and B

Difference of Means

Difference: 103.43
Problem: Determine the diff. between the means of two populations.

Groups A and B

Difference of Means

Difference: 98.22
**Problem:** Determine the diff. between the means of two populations

**Groups A and B**

**Difference of Means**

Difference: 99.16
Comparing Populations, the Old Fashioned Way

**Problem:** Determine the difference between the means of two populations.

**Observation:** In our random samples, the difference between the sample means has

- a bell-shaped distribution,
- and is
- centered at the correct theoretical value!
Comparing Populations, the Old Fashioned Way

**Problem:** Determine the diff. between the means of two populations

**Theorem**

The difference between the sample means has

- a Gaussian distribution,
- is centered at the correct theoretical value, and
- has standard deviation of

$$\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}.$$
Comparing Populations, the Old Fashioned Way

Fact: 95% of random samples will fall within two standard deviations of the true mean.

Restated: 95% of the time, the true difference of means will be within two standard deviations of the sample difference of means.

Example

Suppose that $\sigma_1$ and $\sigma_2$ are both 10 and the sample means are $m_1 = 147.2$ and $m_2 = 49.1$, and $n_1 = n_2 = 25$. Then

$$m_1 - m_2 = 98.1 \quad \text{and} \quad \sqrt{\frac{\sigma^2_1}{n_1} + \frac{\sigma^2_2}{n_2}} \approx 2.83$$

We can be 95% sure that the true difference of means is between $98.1 \pm 2 \times 2.83$, that is, 103.76 and 92.44.
Comparing Populations, the Old Fashioned Way

**Fact:** 95% of random samples will fall within two standard deviations of the true mean.

**Restated:** 95% of the time, the true difference of means will be within two standard deviations of the sample difference of means.

**Example**

Suppose that $\sigma_1$ and $\sigma_2$ are both 10 and the sample means are $m_1 = 147.2$ and $m_2 = 49.1$, and $n_1 = 1,000,000$ with $n_2 = 100,000$. Then

$$m_1 - m_2 = 98.1$$

and

$$\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} \approx 0.03$$

We can be 95% sure that the true difference of means is between $98.1 \pm 2 \times 0.03$, that is, 98.16 and 98.04.

**Note:** In large populations, with high probability, the difference of sample means will be very close to the true difference of means in the two populations.
Comparing Populations, the Old Fashioned Way

**Desiderata:** To compare two populations $A$ and $B$, it would be great to know

- the sample means $m_1$ and $m_2$ of populations $A$ and $B$,
- the standard deviations $\sigma_1$ and $\sigma_2$ of populations $A$ and $B$,
- all of the data, for instance, to test whether it is in fact Gaussian.

Armed with this, we could test the hypothesis of whether the difference between two populations is disappearing:

**Groups A and B**
Comparing Populations, the Old Fashioned Way

**Desiderata:** To compare two populations $A$ and $B$, it would be great to know

- the sample means $m_1$ and $m_2$ of populations $A$ and $B$,

- the standard deviations $\sigma_1$ and $\sigma_2$ of populations $A$ and $B$,

- all of the data, for instance, to test whether it is in fact Gaussian.

Armed with this, we could test the hypothesis of whether the diff. between two populations is disappearing:

*Groups A and B*
**Desiderata:** To compare two populations $A$ and $B$, it would be great to know

- the sample means $m_1$ and $m_2$ of populations $A$ and $B$,
- the standard deviations $\sigma_1$ and $\sigma_2$ of populations $A$ and $B$,
- all of the data, for instance, to test whether it is in fact Gaussian.

Armed with this, we could test the hypothesis of whether the difference between two populations is disappearing:
**Desiderata:** To compare two populations $A$ and $B$, it would be great to know

- the sample means $m_1$ and $m_2$ of populations $A$ and $B$,
- the standard deviations $\sigma_1$ and $\sigma_2$ of populations $A$ and $B$,
- all of the data, for instance, to test whether it is in fact Gaussian.

But this is not the way statistics is done in Texas.

Armed with this, we could test the hypothesis of whether the difference between two populations is disappearing:
Consider the same two imaginary populations, Group A and Group B.

- Set a passing score on the exam.
- Compute what percentage of students in Group A passed.
- Compute what percentage of students in Group B passed.
- Compute the difference between these percentages.

The difference in percentages of students passing the exam will be called the **achievement gap**.
Suppose that the passing grade was set at **40 points**

- 97 percent of Group A passes
- 50 percent of Group B passes

The achievement gap is: **47 points.**
Suppose that the passing grade was set at **45 points**

- 93 percent of Group A passes
- 31 percent of Group B passes

The achievement gap is:

**62 points.**
Already we’ve done something. We’ve changed the achievement gap from 47 to 62 points.

This was done without changing the populations at all, but merely by changing what we considered a passing score on a test they already took!
Suppose that the passing grade was set at **50 points**

- 84 percent of Group A passes
- 16 percent of Group B passes

The achievement gap is: **68 points.**
Suppose that the passing grade was set at 5 points

- 99 percent of Group A passes
- 99 percent of Group B passes

The achievement gap is:

0 points.
Suppose that the passing grade was set at 10 points

- 99 percent of Group A passes
- 99 percent of Group B passes

The achievement gap is:

0 points.
Suppose that the passing grade was set at 15 points

- 99 percent of Group A passes
- 99 percent of Group B passes

The achievement gap is: 0 points.
Suppose that the passing grade was set at 20 points

- 99 percent of Group A passes
- 97 percent of Group B passes

The achievement gap is:

2 points.
Suppose that the passing grade was set at 25 points

- 99 percent of Group A passes
- 93 percent of Group B passes

The achievement gap is:

6 points.
Suppose that the passing grade was set at 30 points

- 99 percent of Group A passes
- 84 percent of Group B passes

The achievement gap is: 15 points.
Suppose that the passing grade was set at 35 points

- 99 percent of Group A passes
- 69 percent of Group B passes

The achievement gap is:

29 points.
Suppose that the passing grade was set at 55 points

- 69 percent of Group A passes
- 6 percent of Group B passes

The achievement gap is: 62 points.
Suppose that the passing grade was set at **60 points**

- 50 percent of Group A passes
- 2 percent of Group B passes

The achievement gap is: **48 points.**
Suppose that the passing grade was set at 65 points

- 31 percent of Group A passes
- 0 percent of Group B passes

The achievement gap is:

30 points.
Suppose that the passing grade was set at **70 points**

- 16 percent of Group A passes
- 0 percent of Group B passes

The achievement gap is: **15 points.**
Suppose that the passing grade was set at 75 points

- 6 percent of Group A passes
- 0 percent of Group B passes

The achievement gap is:

6 points.
Suppose that the passing grade was set at 80 points

- 2 percent of Group A passes
- 0 percent of Group B passes

The achievement gap is: 2 points.
Suppose that the passing grade was set at 85 points

- 0 percent of Group A passes
- 0 percent of Group B passes

The achievement gap is:

0 points.
Conclusion: For Groups A and B, the achievement gap can be manipulated to be anywhere between 0 and 68 points. This can be accomplished merely by changing what we consider to be a passing score.
If the two groups became more similar, the achievement gap still could be manipulated, but the achievement gaps would trace out a different curve.
If the two groups became more similar, the achievement gap still could be manipulated, but the achievement gaps would trace out a different curve.
Evaluating the Texas Success Story

While we have shown that the achievement gap could be manipulated, this does not mean it was. We will pursue this further. A problem is the lack of data. Essentially, the only data available to us is

- the year of the exam,
- the achievement gap,
- the percent of white students who passed, and
- the percent of black students who passed.

Most importantly, we don’t have access to either what the passing score was each year, or what the mean scores were for the two populations.
It turns out that we do not need to know all this data to evaluate Texas’s results.

**Probability Density Function**

The “bell curve” describing the population with mean score $\mu$ and standard deviation $\sigma$ is

$$f(x|\mu, \sigma) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

Assume that

- $f(x|0, 1)$ describes population $A$, and
- $f(x|-D, r)$ describes population $B$. 
Then we can express the percentage of students who pass the exam if the passing score is $\Lambda$ as an integral:

$$\phi_A = \int_{\Pi}^{\infty} f(x | 0, 1) \, dx$$

$$\phi_B = \int_{\Pi}^{\infty} f(x | -D, r) \, dx$$

In terms of these integrals, the achievement gap is then:

$$\Gamma = \phi_A - \phi_B$$

Unfortunately, expression depends on $\Pi$, which we don’t know. However, we can eliminate it via some mathematical magic.
Evaluating the Texas Success Story

Since $\phi_A = \int_0^\infty f(x|0, 1) \, dx$ we can solve for $\Pi$:

$$\Pi = \sqrt{2} \operatorname{Erf}^{-1}(-2\phi_A)$$

Since $\phi_B = \int_0^\infty f(x|-D, r) \, dx$ we can again solve for $\Pi$:

$$\Pi = -D + r\sqrt{2} \operatorname{Erf}^{-1}(-2\phi_B)$$

From which we can write $\phi_B$ in terms of $\phi_A$ by setting the two expressions for $\Pi$ equal to each other:

$$\phi_B = -\frac{1}{2} \operatorname{Erf} \left( \frac{D + \sqrt{2} \operatorname{Erf}^{-1}(-2\phi_A)}{r\sqrt{2}} \right)$$
Evaluating the Texas Success Story

From which

\[ \Gamma = \phi_A - \phi_B = \phi_A + \frac{1}{2} \text{Erf} \left( \frac{D + \sqrt{2} \text{Erf}^{-1}(-2\phi_A)}{r\sqrt{2}} \right) \]

**Moral**

We do not need to know the passing score to predict the achievement gap. The achievement gap can be expressed in terms of only the percentage of students in group A who passed the exam.
Evaluating the Texas Success Story

After some mathematical manipulations, one can eliminate the dependence of our results on the passing score.

The achievement gap can be predicted just from the pass rate of one of the groups. The passing score itself is not necessary.
Evaluating the Texas Success Story

After some mathematical manipulations, one can eliminate the dependence of our results on the passing score.

The achievement gap can be predicted just from the pass rate of one of the groups. The passing score itself is not necessary.
Evaluating the Texas Success Story

After some mathematical manipulations, one can eliminate the dependence of our results on the passing score.

The achievement gap can be predicted just from the pass rate of one of the groups. The passing score itself is not necessary.
Evaluating the Texas Success Story

How would you know if the mean difference between the populations got smaller?

If the mean difference decreased from 20 to 10 points, this is what might be observed.

Punctuated improvement
Evaluating the Texas Success Story

How would you know if the mean difference between the populations got smaller?

Or perhaps this.

*Incremental improvement*
To evaluate the Texas Success story, we could compare Texas’s data with the Achievement gap graphs. But, to make such a graph, we need to know

- \( r \), the ratio of standard deviations,
- \( D \), the difference between mean scores.
Evaluating the Texas Success Story

**Ratio:** We can glean $r$ from other tests taken by Texas students for which the data is available:

- Take $r = 0.85$.

**Difference:** This is harder. We will use the value of $D$ which best fits the Texas data.

However, note that the curve should not fit the data perfectly, as presumably, this difference has been changing.
Evaluating the Texas Success Story

**Ratio:** We can glean $r$ from other tests taken by Texas students for which the data is available:
- Take $r = 0.85$.

**Difference:** This is harder. We will
- use the value of $D$ which best fits the Texas data.

However, note that the curve should not fit the data perfectly, as *presumably*, this difference has been changing.
Evaluating the Texas Success Story

When we compare the Texas data, we would expect to see a graph similar to this one if there was punctuated improvement.
Evaluating the Texas Success Story

When we compare the Texas data, we would expect to see a graph similar to this one if there was

- punctuated improvement
- incremental improvement.
The Texas Success Story

But this is what the data from Texas looks like:

**Achievement Gap**

<table>
<thead>
<tr>
<th>Year</th>
<th>Point Gap</th>
<th>W pass</th>
<th>B pass</th>
</tr>
</thead>
<tbody>
<tr>
<td>1994</td>
<td>35.2</td>
<td>73.3</td>
<td>38.1</td>
</tr>
<tr>
<td>1995</td>
<td>35.4</td>
<td>79.2</td>
<td>43.8</td>
</tr>
<tr>
<td>1996</td>
<td>30.0</td>
<td>85.0</td>
<td>55.0</td>
</tr>
<tr>
<td>1997</td>
<td>25.4</td>
<td>89.5</td>
<td>64.1</td>
</tr>
<tr>
<td>1998</td>
<td>21.4</td>
<td>91.1</td>
<td>70.5</td>
</tr>
<tr>
<td>1999</td>
<td>19.7</td>
<td>92.5</td>
<td>72.8</td>
</tr>
<tr>
<td>2000</td>
<td>16.6</td>
<td>93.6</td>
<td>77.0</td>
</tr>
<tr>
<td>2001</td>
<td>13.2</td>
<td>95.1</td>
<td>81.9</td>
</tr>
<tr>
<td>2002</td>
<td>10.0</td>
<td>96.5</td>
<td>86.5</td>
</tr>
</tbody>
</table>
The data points lie on the curve predicted by
- a mean difference $D$ of 0.97 standard deviation, and
- standard deviation ratio $r$ of 0.85,
which are consistent with other tests taken by Texas students.
The Texas Success Story

But this is what the data from Texas looks like:

Possible Conclusion: No *meaningful* gap reduction occurred in Texas from 1994 to 2002.
Conclusions

For the Honest
Statistics measuring the difference in the percentage of people from group A and people from group B who exceed some benchmark can be deceptive.

For the Dishonest
Statistics measuring the difference in the percentage of people from group A and people from group B who exceed some benchmark can be deceptive.
Conclusions

If you believe that the above is
- a mathematical curio, or
- statistical nuance,
then you are right. **However**, schools are rewarded and penalized on the basis of such statistics.

Unfortunately, this method of measuring population differences is:
- uninformative, and
- potentially deceptive.
**Columbus Dispatch, March 26th**

“The state has set a goal of cutting the gap in passing rates in half by 2017.”

**Detroit Free Press, June 29th**

“The State Board of Education and the Michigan Department of Education recently set closing the achievement gap – the difference in percentage of students passing the exam – as a key goal for the 2012-13 school year.”

Cringe...
Recent Headlines

**Columbus Dispatch, March 26th**

“The state has set a goal of cutting the gap in passing rates in half by 2017.”

**Detroit Free Press, June 29th**

“The State Board of Education and the Michigan Department of Education recently set closing the achievement gap – the difference in percentage of students passing the exam – as a key goal for the 2012-13 school year.”

Cringe...
**Recent Headlines**

**Columbus Dispatch, March 26th**

“The state has set a goal of cutting the gap in passing rates in half by 2017.”

**Detroit Free Press, June 29th**

“The State Board of Education and the Michigan Department of Education recently set closing the achievement gap – the difference in percentage of students passing the exam – as a key goal for the 2012-13 school year.”

Cringe...