New(er) Quantitative Methods for Documenting Health Disparities

James Jaccard
New York University

Health Disparities

Health disparities are widely studied in social work and the health sciences

Health disparities focus on group differences in the incidence, prevalence, and/or burden of adverse health conditions

Commonly studied group differences include (but are not limited to):

> Ethnicity, place of residence (rural areas), gender, age (the elderly, children, adolescents), persons with disabilities, social class, sexual orientation, and many, many more

Health Disparities

The range of outcomes studied covers a wide spectrum including both continuous and dichotomous outcomes. Some examples are:

Depression
Anxiety
Suicide ideation
Suicide
Educational attainment
Income
Access to services
Unintended pregnancies

HIV and STDs
Alcohol use
Drug use
Violence
Injuries
Tobacco use
Exercise and nutrition
Incarceration

Health Disparities

Traditional approaches to documenting health disparities are to compare groups….

For continuous outcomes: on mean outcome values using independent groups t tests, ANOVA, or multiple regression (with dummy variables)

For dichotomous outcomes: on percentages using z tests, or logistic/probit regression (with dummy variables) or the modified linear probability model

Health Disparities

I am going to focus on other analytic methods for documenting group differences on health outcomes

For continuous outcomes, we consider a method called quantile regression

We will learn the basics of quantile regression and how it can (a) reveal disparities that otherwise might be overlooked, and (b) for known disparities, provide greater insight into the nature of those disparities

Health Disparities

Next, we will consider the analysis of health disparities using analytic approaches that reject null hypothesis testing frameworks (and p values) as a way of asserting disparities

This approach applies to all types of outcomes (continuous, dichotomous, etc.) and evolved from studies many years ago on generic versus brand name drugs, where interest was in declaring generic drugs as being equally effective as brand name drugs

The approach is known as *equivalence testing* and *non-inferiority testing*. It has major implications for the study and documentation of health disparities

A Quick Review of Traditional Regression

Review of Traditional Regression Methods

When we apply regression analysis to compare groups on mean values, we regress the outcome onto one or more dummy variables to document group differences

$$\text{Age First Sex} = a + b\,D_B$$

D_B is a dummy variable where Blacks = 1 and Whites = 0

Review of Traditional Regression Methods

$$\text{Age First Sex} = 16 + 1.0\, D_B$$

The intercept is the predicted mean age when $D_B = 0$, i.e., the mean age of first intercourse for Whites is 16

The regression coefficient is the mean Y difference between the group scored 1 on D_B and the group scored 0, i.e., it is the mean age for Blacks minus the mean age for Whites

The significance test for *b* is a test of the mean difference

(The mean age of first sex for Blacks above is 17)

Review of Traditional Regression Methods

$$\text{Age First Sex} = a + b_1 D_B + b_2 \text{ Income}$$

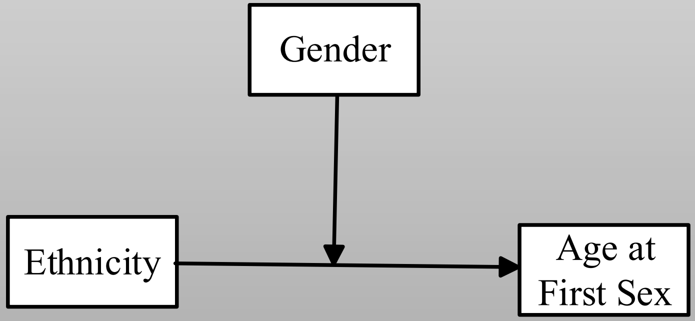
We can include in the analysis additional predictors that can serve as covariates. In this case, I decide to control for parental income

a is the predicted mean on Y when all predictors = 0, i.e., for White youth whose parents had no income

If b_1 is, say 0, then there is no difference in the mean age at first intercourse for Whites and Blacks when income is held constant

Review of Traditional Regression Methods

We also can include interaction terms for moderator analysis. I am interested in whether ethnic differences in age at first intercourse vary depending on gender:



We model this using product terms:

$$\textbf{Age First Sex} = \mathbf{a + b_1\, D_B + b_2\, D_F + b_3\, (D_B)(D_F)}$$

Review of Traditional Regression Methods

Suppose the means are as follows:

	Females	Males
Blacks	14	13
Whites	15	15

I model this using product terms

Review of Traditional Regression Methods

$$\text{Age First Sex} = a + b_1 D_B + b_2 D_F + b_3 (D_B)(D_F)$$

	Females	Males
Blacks	14	13
Whites	15	15

***b*₁** is the ethnic difference for males = 13 – 15 = -2. The significance test for *b*₁ is the test of this difference

Note that the ethnic difference for females is 14 – 15 = -1.

***b*₃** is the difference between the two ethnic differences, (-2) – (-1) = -1. It tests the interaction effect

Review of Traditional Regression Methods

In sum, we use multiple regression to study mean differences between groups

We can include dummy variables or continuous predictors

We can include covariates, as appropriate

We can model statistical interactions

Quantile Regression

Describing Distributions

A common method for describing distributions is to report the mean and standard deviation of it. The former is an index of central tendency and the latter is an index of variability

When we document health disparities on continuous outcomes, we almost always compare groups on means

However, there are other facets of a distribution that we can compare groups on

Describing Distributions

A quantile is what many of us learned as a percentile

Informally speaking, a quantile is a score in a distribution that a specified percentage of scores are less than or equal to

If I tell you a GRE score of 161 (using the new scoring methods) defines the 80th percentile, this means that 80% of individuals scored 161 or less

If I tell you a GRE score of 153 defines the 50th percentile, this means that 50% of individuals scored 153 or less

Describing Distributions

For reasons I will not go into, statisticians hate the use of the term *percentile* and use the term *quantile* in its place

And, the percentile associated with a score is expressed as a proportion (or probability) instead of a percent

The 0.80 quantile (q = 0.80) for the GRE is a score of 161

The 0.50 quantile (q = 0.50) for the GRE is a score of 153

Describing Distributions

A widely used quantile is the 0.50 quantile because it is the median of a distribution (half the scores are below it and half are above it)

A property of the median is that it is outlier resistant, so it is often used to describe the central tendency of variables that have outliers, such as income

The 0.50 quantile ($q = 0.50$) for income was $45,000

The median income was $45,000

Describing Distributions

The median of

20,000
23,000
25,000
27,000
100,000

is 25,000. Even if the last score was 1,000,000, the median would be 25,000

Means (and standard deviations) are outlier sensitive. Quantiles are not, which is a desirable property of them.

Comparing Groups on Quantiles

We can compare groups on quantiles. If we compare them on the 0.50 quantile, we are comparing their medians

Suppose we compare males and females on their median depression scores on the CES-D (a well known scale that ranges from 0 to 60, where scores of 16 or greater are assumed to be clinically significant)

Males: 9.0

Females: 12.0

This tells us what is going on in the middle of the distribution, i.e., we are comparing the central tendency of the two groups

Comparing Groups on Quantiles

But what about at the lower and upper ends of the distribution? What if we compare the 0.25 quantile for males and females and find

Males: 5.0

Females: 5.0

We see at the lower end of the depression distribution, there is no difference between males and females: 25% of males have scores less than or equal to 5 and 25% of females also have scores less than or equal to 5.

Comparing Groups on Quantiles

Suppose at q = 0.90, we find the following quantiles

Males: 19.0

Females: 25.0

Only 10% of males have depression scores above 19, whereas 10% of females have scores above 25. This difference of 6 units in the quantile is more pronounced than the difference at the median (which was 3)

Comparing Groups on Quantiles

As students of health disparities are becoming sensitized to more exaggerated or less exaggerated differences in different portions of the distribution of an outcome, they are expanding their focus to the analysis of quantiles

Numerous studies that have not observed health disparities when focusing on means have documented them at different parts of the distribution using quantiles

This has led to the use of a method called quantile regression to analyze such differences.

Comparing Groups on Quantiles

Actually, there are two approaches to comparing groups on distributions that are used and both offer somewhat different perspectives on data.

One approach analyzes *quantiles* (which we are going to explore) in the form of quantile regression and the other, which is more common, analyzes *breakpoints*.

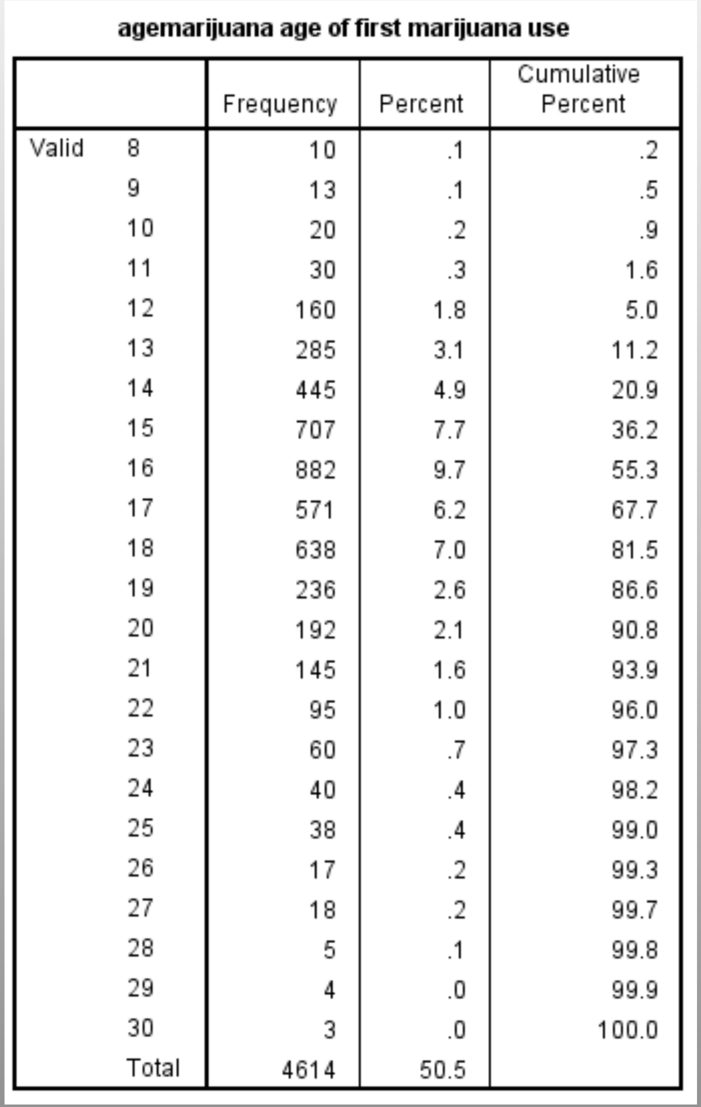
We can appreciate the difference in the two approaches best by examining cumulative density functions of scores for two groups.

Comparing Groups on Quantiles

We are familiar with an informal type of a cumulative density function when we do one way frequency distributions in SPPS.

Note the column of cumulative percents to the right

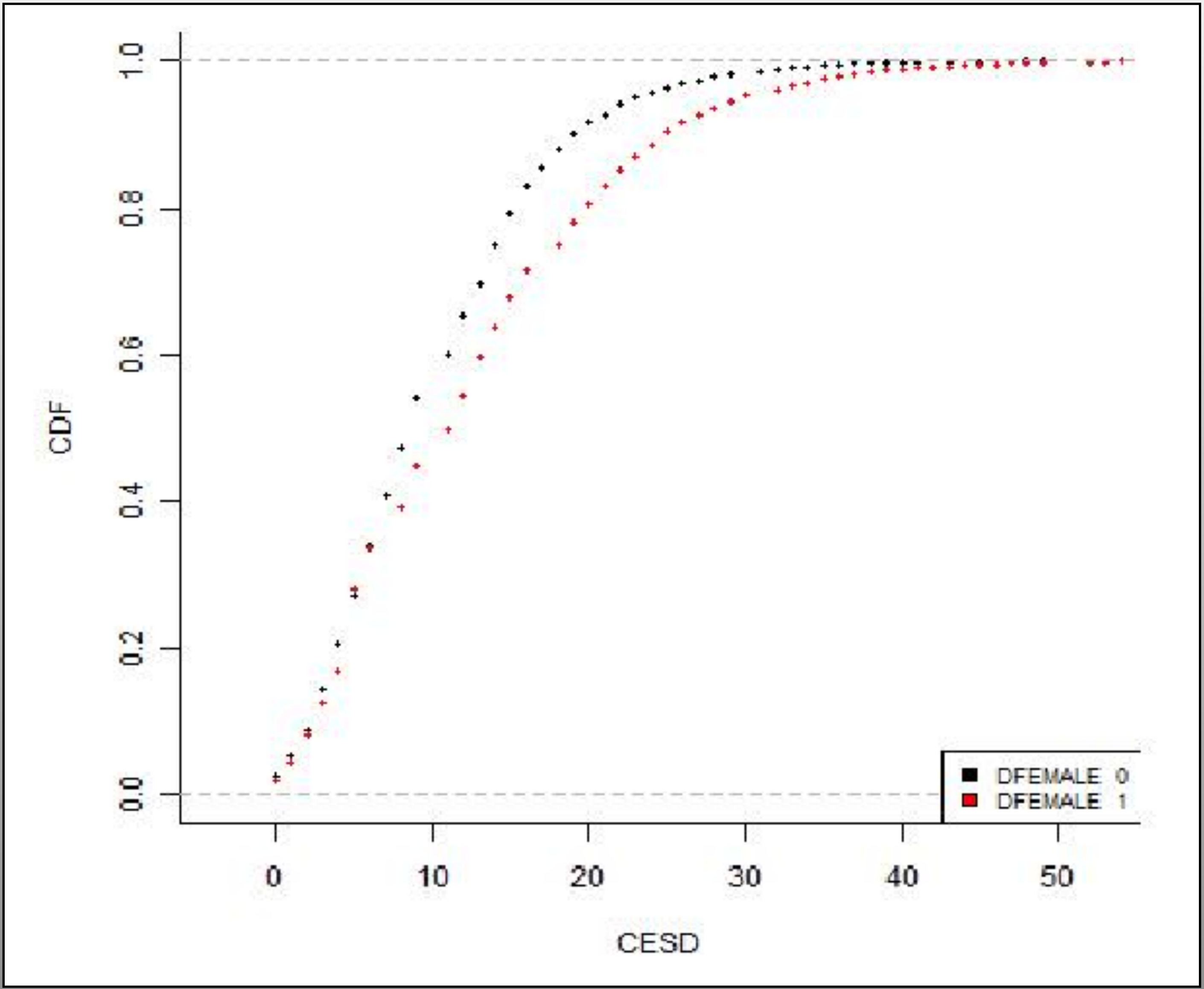
On the next slide is a plot of the CDF of depression scores (CESD) for males and females



agemarijuana age of first marijuana use

		Frequency	Percent	Cumulative Percent
Valid	8	10	.1	.2
	9	13	.1	.5
	10	20	.2	.9
	11	30	.3	1.6
	12	160	1.8	5.0
	13	285	3.1	11.2
	14	445	4.9	20.9
	15	707	7.7	36.2
	16	882	9.7	55.3
	17	571	6.2	67.7
	18	638	7.0	81.5
	19	236	2.6	86.6
	20	192	2.1	90.8
	21	145	1.6	93.9
	22	95	1.0	96.0
	23	60	.7	97.3
	24	40	.4	98.2
	25	38	.4	99.0
	26	17	.2	99.3
	27	18	.2	99.7
	28	5	.1	99.8
	29	4	.0	99.9
	30	3	.0	100.0
	Total	4614	50.5	

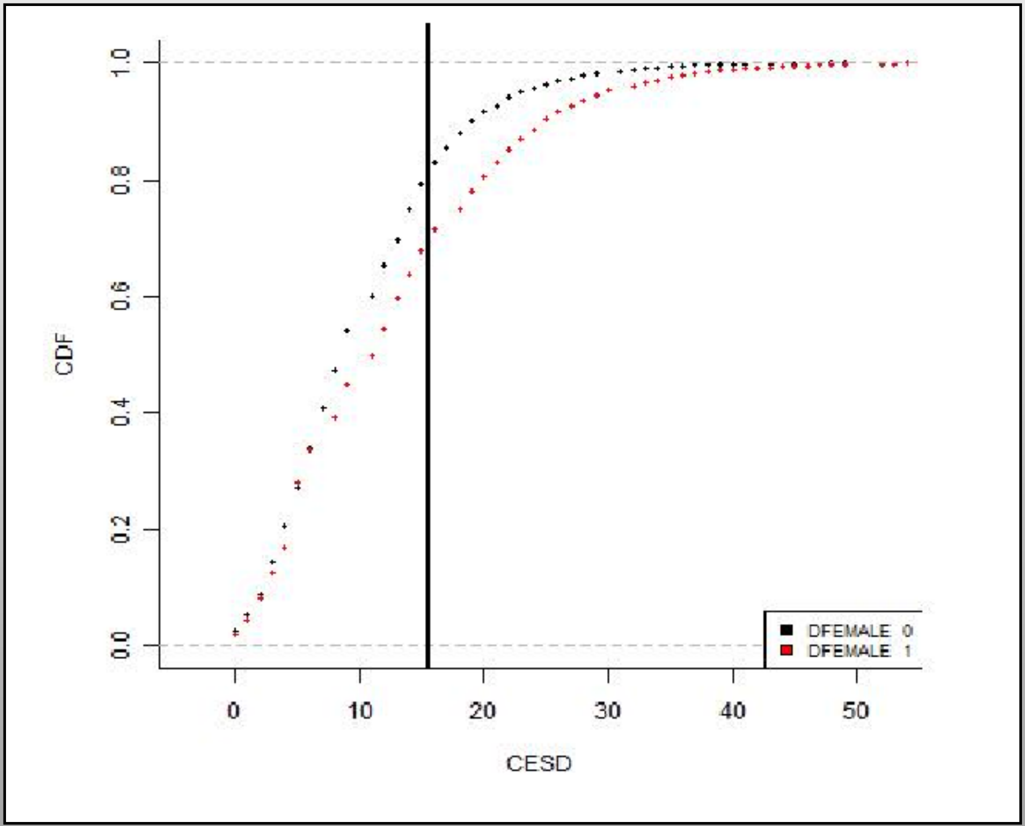
CDF Plot



Breakpoint Analysis

This defines a CES-D score of interest (e.g. 16) and uses the CDF to compare the % of cases above and below the score (the breakpoint) for the two groups

20% of males have a score above 16, with 80% below 16. 32% of females have a score above 16, with 68% below 16



Breakpoint Analysis

In breakpoint analysis, the CES-D is dichotomized into a 0-1 metric (0 = below the breakpoint, 1 = above the breakpoint) and logistic/probit regression or a modified linear probability model is used to explore its relationship to other variables:

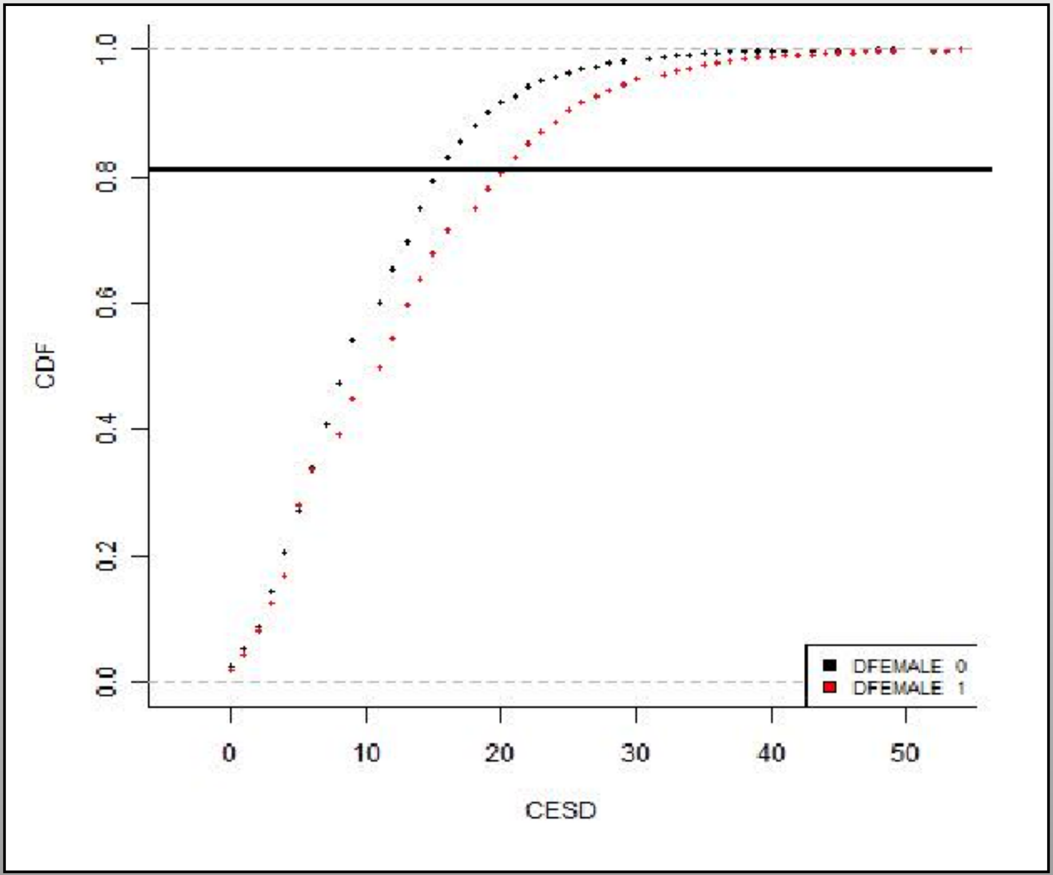
$$\text{Dichot CES-D} = a + b_1 D_F$$

Choice of breakpoints is theoretically guided

Quantile Analysis

This defines a quantile of interest (e.g. q=0.80) and uses the CDF to identify the score that maps onto to that quantile for each group

q = 0.80 quantile for males is 15 and for females it is 20. 20% of males are above 15, whereas 20% of females are above 20.



Breakpoint and Quantile Analysis

The quantiles are then related to other variables using quantile regression methods:

$$q = 0.80 \text{ for CES-D} = a + b_1 D_F$$

Both approaches have their virtues and researchers often analyze matters from both perspectives, as appropriate

Implementing Quantile Regression

Go over computer program and output

Show program to plot CDFs

We can basically use quantile regression just as we would traditional linear regression to explore a wide range of predictors of health outcomes in ways different from OLS.

Quantile Regression Methods

We can include dummy variables or continuous predictors

We can include covariates, as appropriate

We can model statistical interactions

There are many technical issues that need to be considered, especially for longitudinal data

SEM Model

Can use in SEM frameworks by adopting a limited information approach to the linear equations suggested by a path model

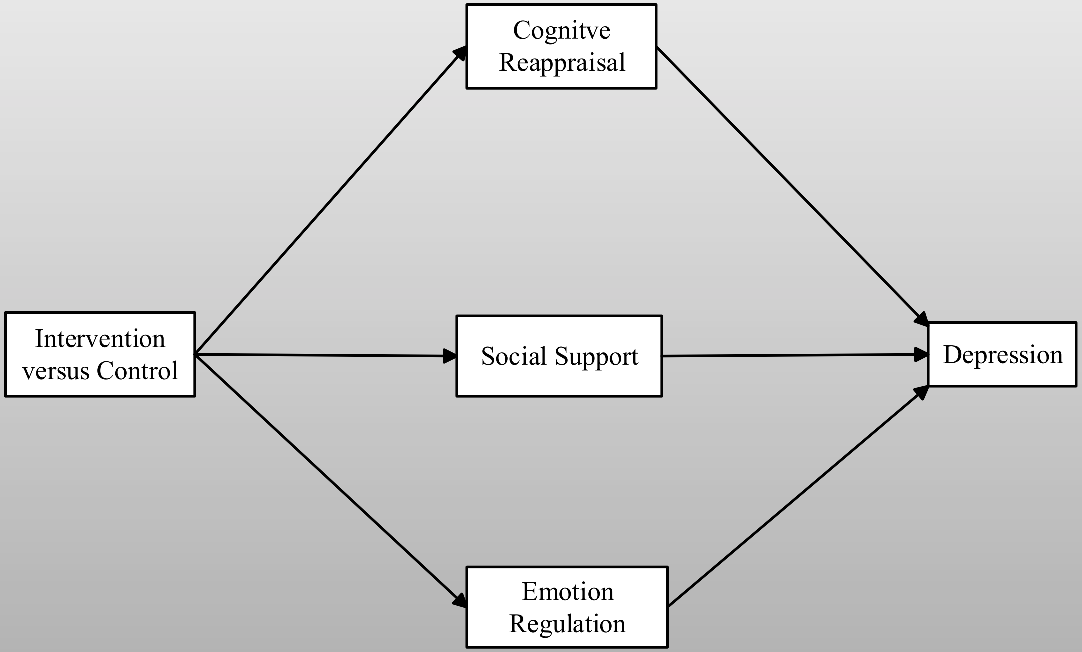
Consider a randomized explanatory design where an intervention (versus control) is designed to improve three distinct factors associated with depression

Cognitive reappraisals

Use of social support

Affect regulation

Application to SEM



This model implies four linear equations. Can work with each using standard or quantile regression.

Some References

Hao, L. and Naiman, D. (2007). Quantile regression. Newbury Park: Sage

Koenker, R. (2005). Quantile regression. New York, NY: Cambridge University Press

Gebregziabher, M., Lynch, C., Mueller, M. et al. (2011). Using quantile regression to investigate racial disparities in medication non-adherence. BMC Medical Research Methodology, 11, 88-95.

Juarez, D, Tan, C., Davis, J. et al. (2014). Using quantile regression to assess disparities in medication non-adherence. American Journal of Health, 38, 53-62

Health Disparities and Equivalence Testing

Equivalence Testing

Equivalence testing evolved from scenarios where the FDA wanted to compare the effectiveness of generic drugs to brand name drugs to determine their equivalence

In traditional null hypothesis testing frameworks, we formulate a null and alternative hypothesis to test with respect to some effectiveness outcome:

$$H_0: \mu_{BN} - \mu_G = 0$$

$$H_1: \mu_{BN} - \mu_G \neq 0$$

Equivalence Testing

We conduct a study on the outcome and determine if the sample means are different for the two groups (generic vs. brand name). They almost always are, so we are interested in whether the difference can be attributed to sampling error.

We calculate a p value for our sample difference which is the probability that our sample mean difference would occur given H_0 is true.

If the p value < 0.05, we reject the null hypothesis of no population difference in means

If the p value > 0.05, we fail to reject the null hypothesis and conclude the sample result could be sampling error

Equivalence Testing

$$H_0: \mu_{BN} - \mu_G = 0$$

$$H_1: \mu_{BN} - \mu_G \neq 0$$

Note that we do not *accept* the null hypothesis. Rather we *fail to reject it*.

We can never accept the null hypothesis because it refers to a specific population value rather than a range of values. It is virtually impossible, given sampling error, to say that a population mean or mean difference is exactly equal to a specified value

Equivalence Testing

This is the dilemma faced by the FDA. They want to say that generic drugs and brand name drugs are equivalent; but they can never accept the null hypothesis

Also, if a study is conducted with small N, it will have low power and lead to “acceptance” of the null hypothesis. FDA wants strong tests of equivalence, not weak tests based on small N

From this dilemma, equivalence testing evolved

Coefficient of Functional Equivalence

Suppose we are comparing the mean annual starting salaries of male and female assistant professors of social work in the U.S. as a first step to determine if there is gender bias in salaries

Suppose the true population mean for females is $50,000 and for males it is $50,001

The null hypothesis that the population mean difference is 0 is false.

If we sample 100 male and 100 female assistant professors and perform a t test on sample means, and if we fail to reject the null hypothesis, we will have committed a Type II error

Coefficient of Functional Equivalence

But do we really care if we make such an error given the population mean difference is so small?

It can be argued that a $1 population difference in annual income is so small that it simply does not matter, i.e., that the two groups are *functionally equivalent* on annual mean salaries

What if the population difference is $10? How about $100. How about $1,000. How about $10,000?

At some point, the difference becomes non-trivial and meaningful. What is that point?

Coefficient of Functional Equivalence

The point that separates a trivial from a meaningful difference is called a *coefficient of function equivalence* (CFE).

When we evaluate mean differences (or percent differences) we need to specify a CFE to work with.

This is, in some respects, an evaluation of effect size. We need to state a standard for declaring an effect size (mean difference) meaningful or trivial

Surprisingly, social scientists have been lax at addressing this very fundamental question

Coefficient of Functional Equivalence

A common approach is to defer to Cohen's (1988) classic standards as expressed in the form of Cohen's d.

d is the population mean difference between two groups divided by the (pooled) population standard deviation.

According to Cohen, a *small effect size* is when the mean difference equals 1/5 of a standard deviation (d = 0.20)

Mean for group 1: 102   Mean for group 2: 100   SD = 10

d = (102 – 100) / 10 = 0.20

Coefficient of Functional Equivalence

A *small effect size* is when the mean difference equals 1/5 of a standard deviation (d = 0.20)

A *medium effect size* is when the mean difference equals 1/2 of a standard deviation (d = 0.50)

A *large effect size* is when the mean difference equals 4/5 of a standard deviation (d = 0.80)

Essentially, the standard deviation becomes the “standard” against which we compare a mean difference to determine its meaningfulness

Coefficient of Functional Equivalence

A d of 0.50 translates into a percent of variance accounted for of abut 6%. So if a factor accounts for less than 6% of the variance in an outcome, it might be deemed “small” and insignificant

But why 6%? How did Cohen come up with his standards?

Cohen’s standards are widely used in power analysis and in evaluations of mean differences reported in studies. Just how did he come up with these?

Coefficient of Functional Equivalence

It turns out, Cohen (1988) states they are arbitrary. To quote him:

"*Characterizing effects as small, medium or large is an operation fraught with many dangers because such "definitions are arbitrary*" (p. 12)

"*The terms are relative…to the specific content and research method being employed in an investigation*"… [and his criteria are "*recommended for use only when no better basis for estimating the effect size index is available*" (p. 25)

Coefficient of Functional Equivalence

Cohen also said:

"*…these proposed conventions were set forth… with much diffidence, qualifications, and invitations not to employ them if possible,*" noting that "*the values [have]… no more reliable basis than my own intuition*" (p. 532)

And that the criteria…

"*were needed in a research climate characterized by a neglect of attention to issues of magnitude*"

Coefficient of Functional Equivalence

Other researchers agree with Cohen

Glass, McGaw, and Smith (1981, p. 104) state "*there is no wisdom whatsoever in attempting to associate regions of the effect-size metric with descriptive adjectives such as 'small,' 'moderate,' 'large,' and the like*.

Lenth (2009) refers to Cohen's effect size criteria of "small", "medium", and "large" as T-shirt effect sizes that lead power analyses to arrive at the same required sample size no matter what the characteristics of the outcome or the setting.

Coefficient of Functional Equivalence

Examples of the failure of Cohen’s standards include….

1987 study of aspirin, that was halted for ethnical reasons despite an effect size of less than .001 percent explained variance

If a company pays all its new employees virtually the same amount, the SD for salaries might be quite low, say $10. If the mean for males is $50,010 and for females it is $50,000, Cohen’s d is 1.0 and the difference is declared as being “large”

We need to get serious about effect size evaluation and not seek simplistic rules that allow us not to think them through

Coefficient of Functional Equivalence

Factors to consider include…

How *likely* it is to affect the overall quality of life of individuals

The *degree of impact* (severity) it has on people

***How many* individuals are affected**

The *sustainability* or *reversibility* of the effect over time

The *vulnerability* (ability to "defend oneself" against the negative event) or entitlement (helping the rich get richer at the expense of the poor) of the affected individuals

The *costs* (broadly defined) of addressing the event

Meaningfulness Intervals

Suppose, to continue to illustrate our logic model, that we set a CFE for a gender salary difference at $600. This is $50 a month (note: $50 a month is more important for populations that are poor than those that are wealthy)

If the population mean difference is outside -600 to +600, then we will deem the difference to be meaningful. Otherwise, we will declare the groups as being functionally equivalent in terms of their salary

We refer to the interval defined by the CFE to specify a meaningful effect as a *meaningfulness interval*

Margins of Error

When we calculate a sample estimate of a parameter/difference, we can state a margin of error for that estimate

We see this practice for political polls reported in the media (the percents have a margin of error of plus or minus 5%)

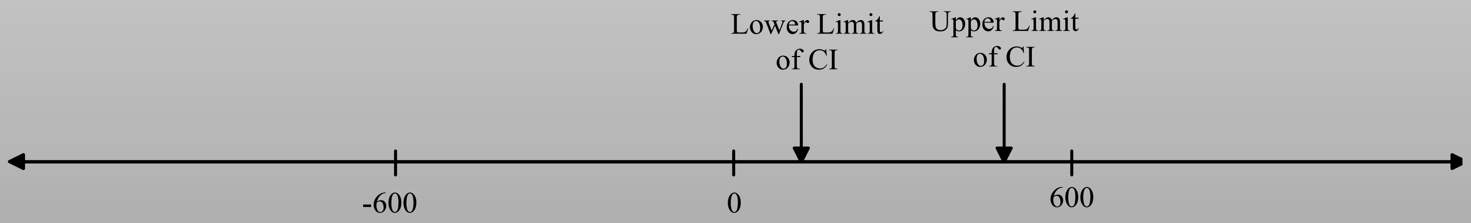
We use confidence intervals (CIs, or credible intervals) to form margins of error

If a mean difference in annual salary is $300, with a 95% confidence interval of 100 to 500, then the estimate of $300 has a margin of error (MOE) of plus or minus $200

Decision Rules

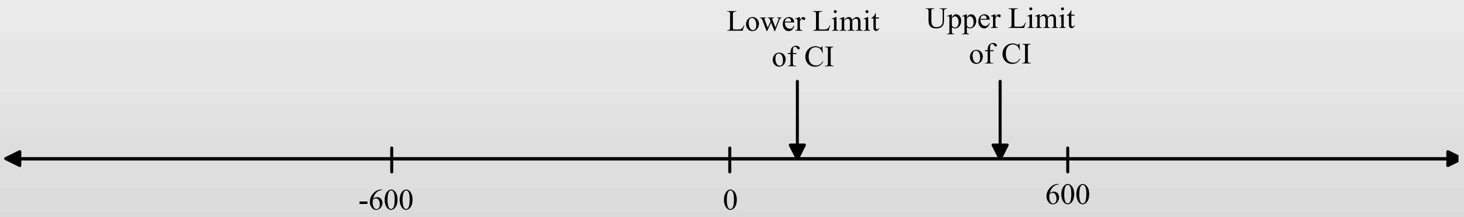
Suppose we conduct a study and find a sample estimate for gender differences in salaries is $300 with a 95% CI of $100 to $500. Note that this is completely contained in our meaningfulness interval

Here is a graphical depiction, known as an *equivalence diagram*:



Note that because 0 is not in the CI, the result is statistically significant ($p < 0.05$) in traditional hypothesis testing

Decision Rules



Because the CI is completely contained within the meaningfulness interval, we are confident the true population mean is within the -600 to +600 standard. We declare the groups “functionally equivalent.”

This is the basic logic model of equivalence testing

Equivalence Testing

Here are the basic steps we followed:

Specify a CFE that distinguishes a trivial from a meaningful effect

Translate the CFE into a meaningfulness interval

Collect data and calculate the mean or percentage difference and its associated 95% CI

Determine if the CI is completely contained within the meaningfulness interval. If it is, declare functional equivalence. If it is not, do not declare such equivalence

Equivalence Testing

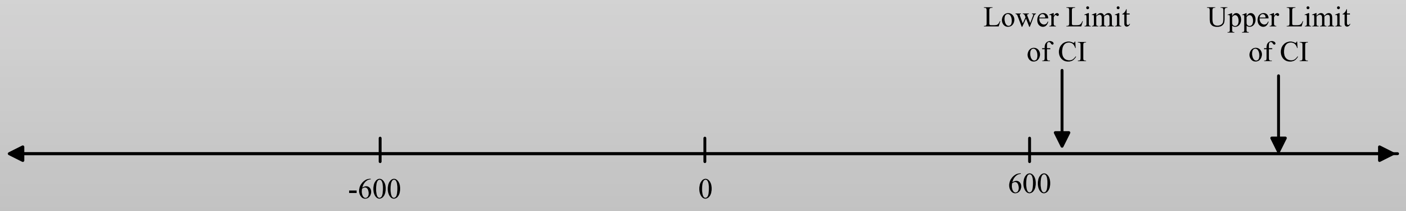
There are numerous issues that arise in applying this framework and I will identify these shortly

The framework can be applied to health disparities research to decide if there are meaningful group differences on a health outcome or if groups are functionally equivalent on those outcomes

Research in health disparities that is using this approach are sometime making different conclusions about the presence of disparities as compared to studies that rely just on statistical significance (p values)

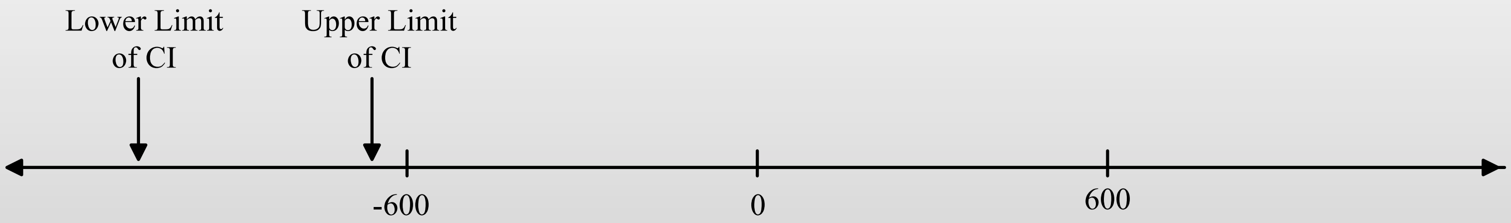
More Possible Results

Here are some other results that can happen in the equivalence testing framework

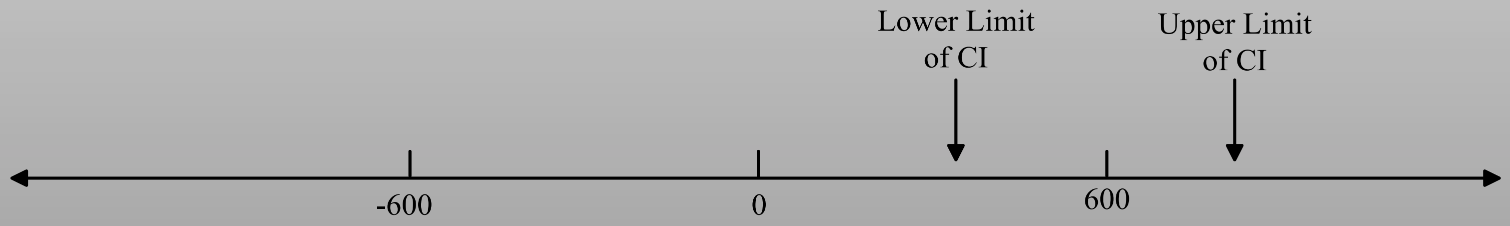


In this case, we confidently conclude that there is a meaningful difference between the groups

More Possible Results

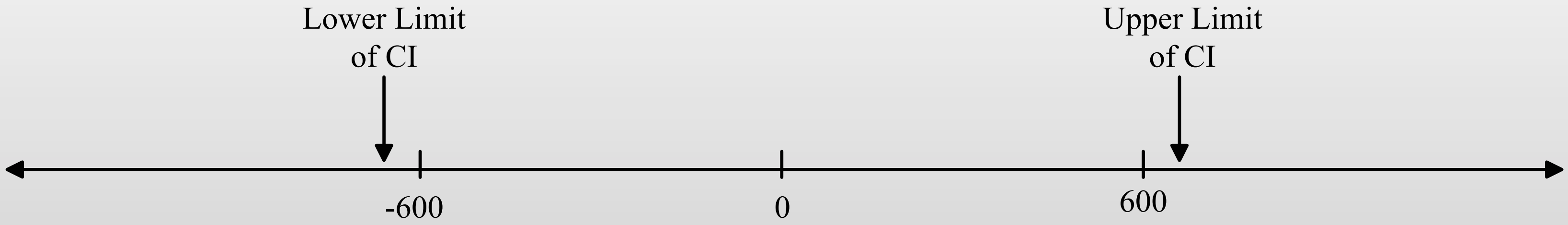


Here, we also confidently conclude that there is a meaningful difference between the groups



Here, we can not conclude one way or the other. We need to suspend judgment

More Possible Results



Here, the CI is so wide that we can't say anything. In essence, we need to re-run the study to get the width of the CI more narrow by having a larger sample size

This latter point is important because if you conduct a study with small N, you are going to end up with a wide CI, making it harder to declare functional equivalence (or to conclude anything, for that matter)

Equivalence Testing and Null Hypothesis Testing

Note that in this approach, we never said a word about p values. Significance testing, as we know it, is irrelevant

In null hypothesis testing, one implicitly sets a CFE. The CFE one uses is 0! Always!

$$H_0: \mu_F - \mu_M = 0$$

$$H_1: \mu_F - \mu_M \neq 0$$

Destructive Dichotomous Thinking

Null hypothesis testing has become perverted into rigid dichotomous thinking with respect to p values. A p value of 0.04999 represents a meaningful effect, but one of 0.05001 does not

In my work that applies equivalence testing to health disparities, I seek to minimize such thinking

A CFE has somewhat of the same qualities as a p value of 0.05, although it at least can vary from study to study and outcome to outcome

Destructive Dichotomous Thinking

When specifying a CFE, there will be ranges of values that everyone agrees represents trivial effects (such as a $1 difference in annual income)

There also will be ranges of values that everyone agrees represents meaningful effects (such as a $10,000 difference in annual income)

There also will be a “gray area” which is a range of CFEs values that people might disagree on. I try to identify this latitude of values and then explore conclusions using different values from that latitude

Additional Notes

Unfortunately, the FDA in using this framework has fallen into the trap, like Cohen’s standards, of using a simple a priori standard for defining a CFE

A generic drug must be at least 80% as effective as a brand name drug to be declared “functionally equivalent.” It is easy to find problems with this standard (but the FDA recognizes this)

In social science research that uses this equivalence testing, there remains a tendency to rely on arbitrary standards for defining a CFE, often relative to a SD. We just seem to be looking for ways not to have to think about difficult issues.

Additional Notes

I am pursing research to try to make metrics of many of our scales more meaningful and not so arbitrary.

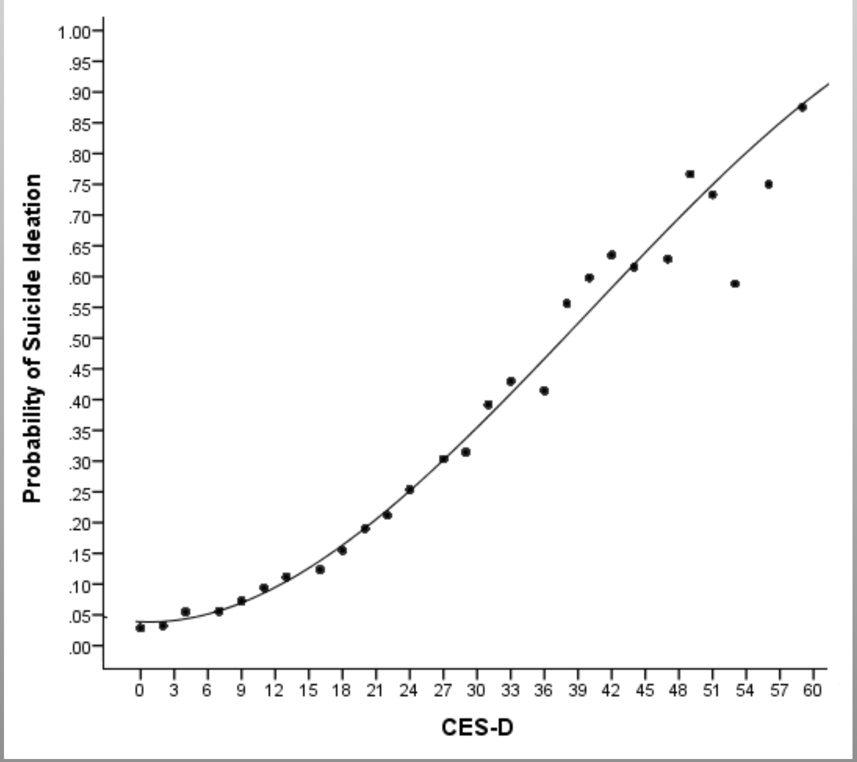
Consider the CES-D scale of depression that ranges from 0 to 60. People have said that a score of 16 is “clinically meaningful” but when I examined the empirical bases for this standard, there was none. This is based on “clinical judgment.”

What is the meaning of a CES-D score of 14 and what does it mean when we change means from 14 to 10? Or from 20 to 16? Or from 21 to 18?

(Too often, we just revert to Cohen’s standards to deal with this)

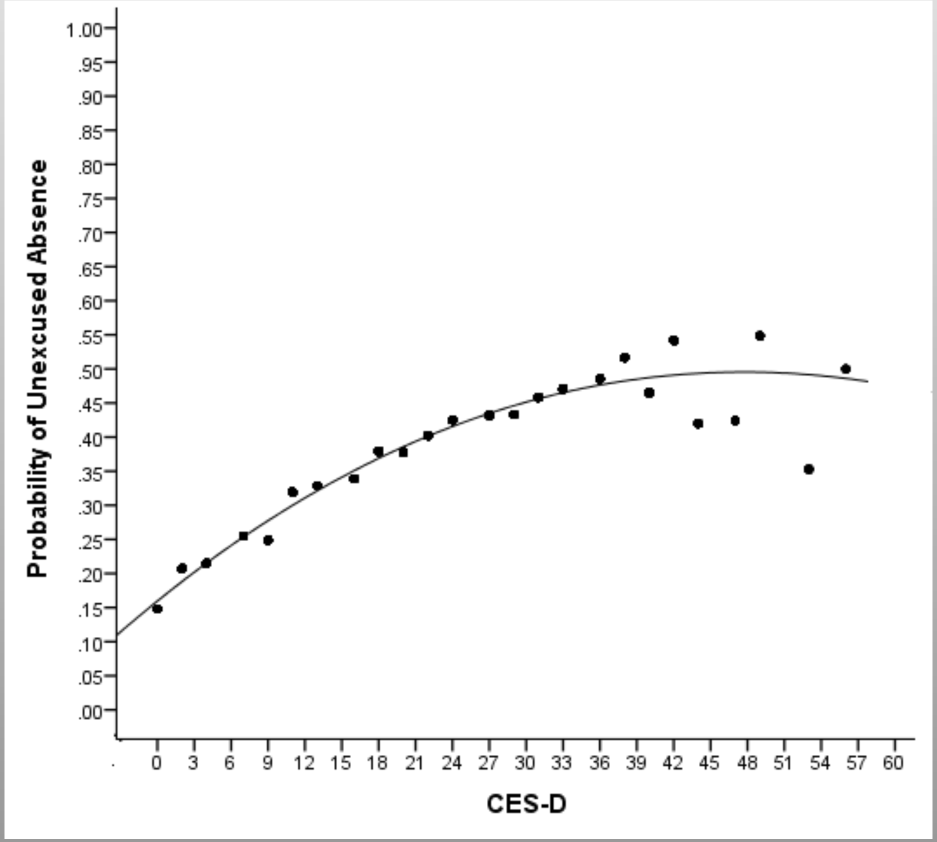
Additional Notes

I my research, I am relating each scale point on the CES-D to meaningful benchmarks in large national samples. Here is a plot of the CES-D and suicide ideation for a nationally representative sample of 20,000 adolescents



Additional Notes

Here is a plot for the probability of an unexcused absence from school in the past 6 months:



Additional Notes

I am relating scores on the CES-D metric to 20 different benchmarks to try to give the scale more meaning and for us to more fully understand the implications in changes in values on it

Concluding Comments

Concluding Comments

Heath disparities are a major concern for all of us

One approach we can profitably use in studying and documenting health disparities is to move away from over-reliance on mean values and central tendencies. Quantile regression is a useful tool in this regard

Another approach we can profitably use is to move away from reliance on p values and to incorporate some of the perspectives offered by equivalence testing