Still Crazy After All These Years: Complexity, Principles, and Practice in Multivariate Statistics

James H. Steiger

Department of Psychology and Human Development
Vanderbilt University

Still Crazy After All These Years: Complexity, Principles, and Practice in Multivariate Statistics

Introduction

In our introductory statistics education, we were taught a number of key principles to guide our research.

But principles that are obvious in one context may be obscured in another.

And so, year after year, we keep doing the same thing.

In this talk, I'll briefly consider 3 examples. I'm preparing a manuscript with several more examples and I treat all of them in much more detail.

Replicability, Representativeness, Visualization

Multivariate Nonnormal Random Number Generation

Many studies of robustness have used the technique due to Fleishman(1978) and Vale and Maurelli (1983) to simulate data.

By 2010, there were dozens of citations to the Fleishman and Vale-Maurelli articles.

Fleishman(1978).

A power transformation of a standardized normal variable Z.

$$Y = b_0 + b_1 Z + b_2 Z^2 + b_3 Z^3 \tag{1}$$

Manipulated skewness and kurtosis independently while maintaining zero mean and unit variance.

Vale and Maurelli(1983).

Extended the method to allow the generation of sets of standardized variables with desired marginal skewness and kurtosis, *and* specified intercorrelations.

Replicability, Representativeness, Visualization

Multivariate Nonnormal Random Number Generation

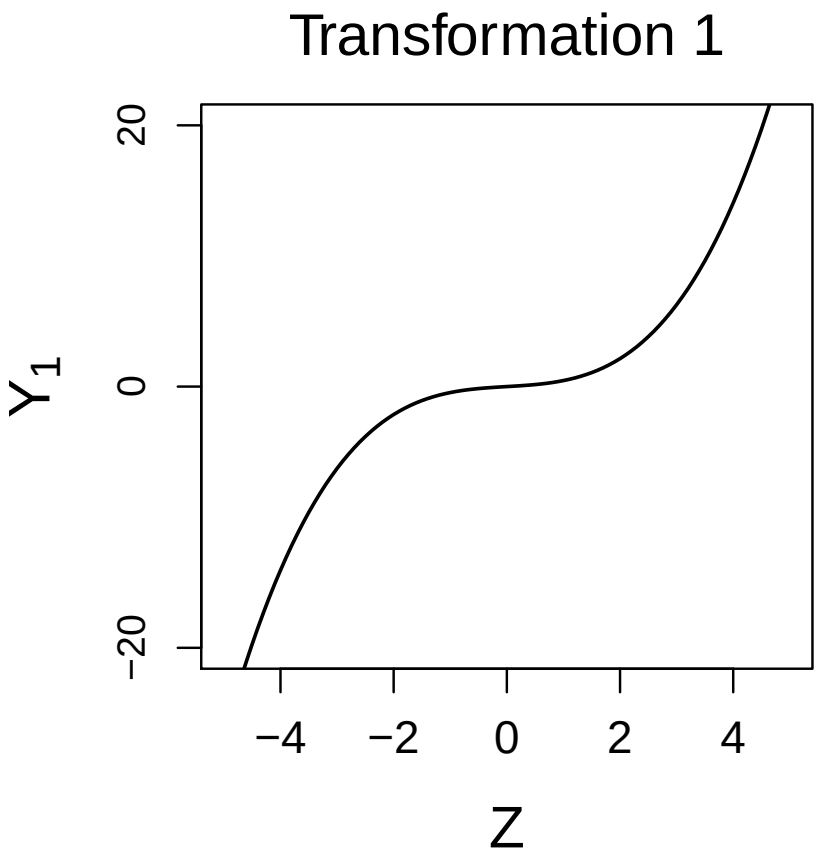
Fleishman had given only one set of weights for any skewness-kurtosis combination.

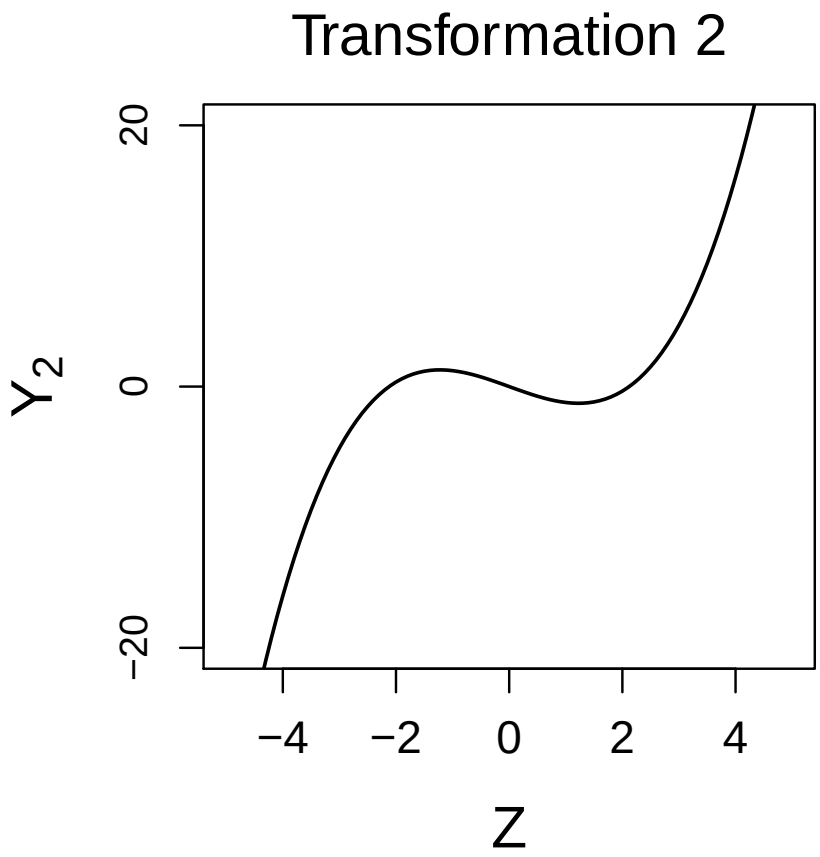
Around 2010, a student, Miriam Kraatz, and I noticed that the solution for polynomial weights in the Fleishman transformation is not unique.

On the next slide are plots of two polynomial transformations of Z that both yield skewness of $\gamma_1 = 0$ and kurtosis of $\gamma_2 = 25$.

Replicability, Representativeness, Visualization

Multivariate Nonnormal Random Number Generation

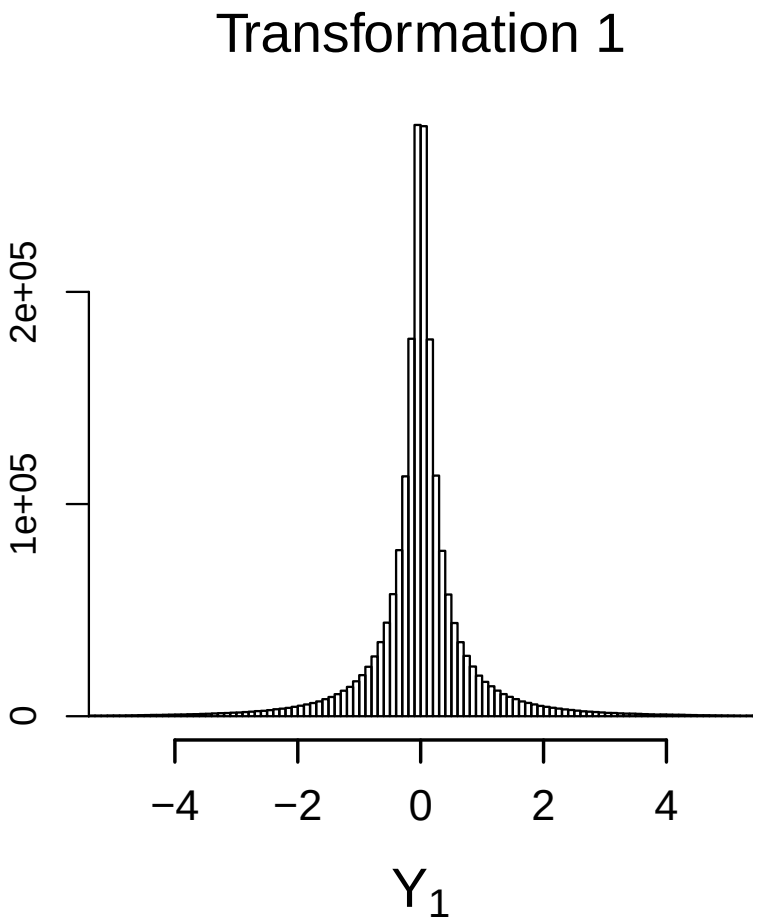


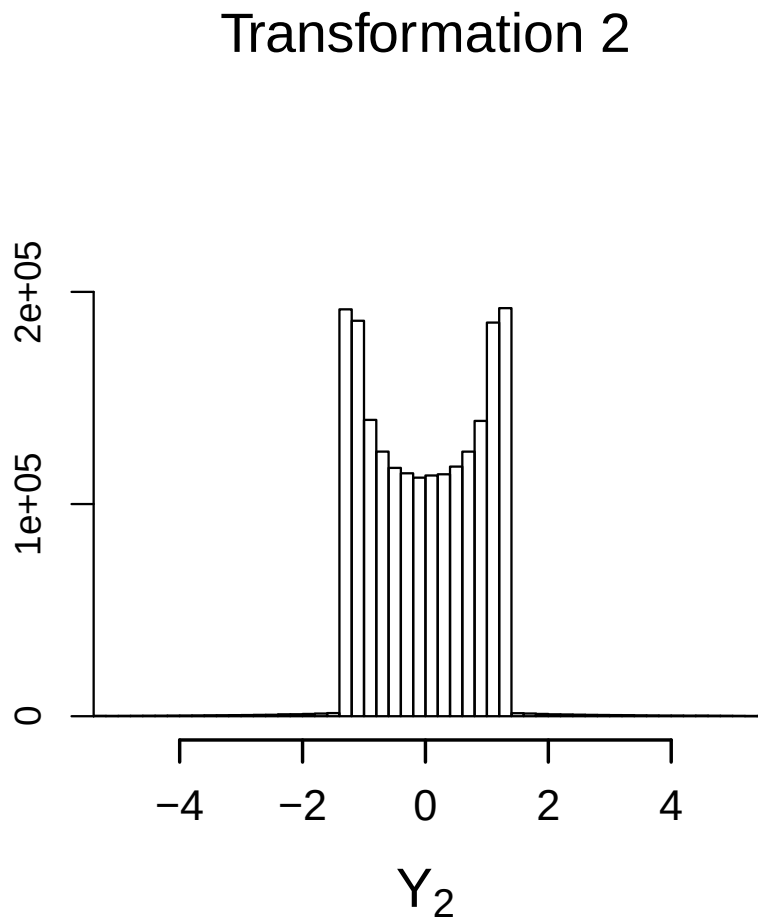


Replicability, Representativeness, Visualization

Multivariate Nonnormal Random Number Generation

This apparently subtle difference has a big effect on the distribution of the two transformed variables.



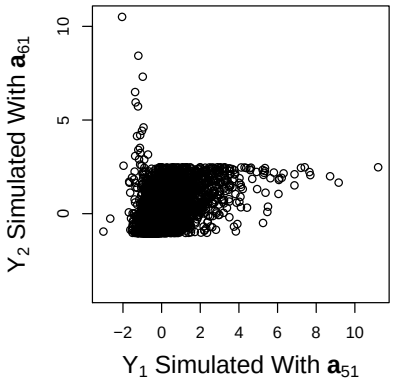


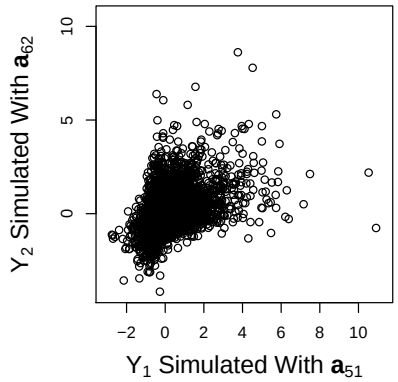
Replicability, Representativeness, Visualization

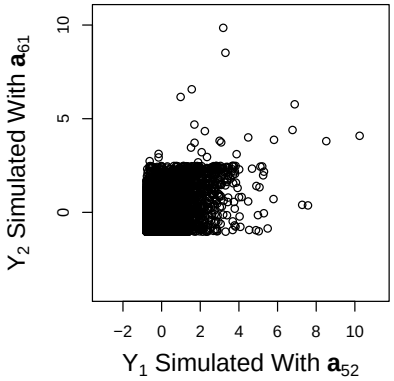
Multivariate Nonnormal Random Number Generation

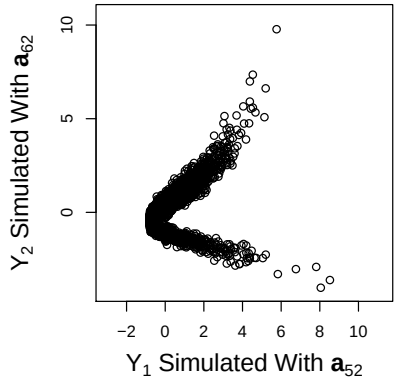
The difference is even more obvious when a bivariate distribution is produced.

Only one of the distributions looks remotely like anything I've encountered. How about you?









Replicability, Representativeness, Visualization

Multivariate Nonnormal Random Number Generation

Which of several possible transforms is produced may be an accident of fate, depending on how the software is written, or whether the coefficients were taken directly from Fleishman's article.

Replicability, Representativeness, Visualization

Multivariate Nonnormal Random Number Generation

In her dissertation research, Kraatz found, not surprisingly, that some statistical methods perform rather differently with different versions of "the" transform.

This means that replicability and clarity are in doubt for articles using these methods.

Moreover, some simulated data produced by the method of Vale and Maurelli (1983) aren't representative of what we normally see in the real world.

Replicability, Representativeness, Visualization

Multivariate Nonnormal Random Number Generation

Having opened Pandora's box, Kraatz proceeded to investigate further.

She found several published papers that claimed to simulate combinations of skewness and kurtosis that are not possible — because there are joint bounds on skewness and kurtosis for any distribution, and the claimed values violated these bounds.

Apparently, software solving for Fleishman coefficients "converged" to a solution that is outside the permissible parameter space.

Reviewers and editors had failed to detect this.

Replicability, Representativeness, Visualization

Multivariate Nonnormal Random Number Generation

Think of how the quality of the overall research effort might have been positively impacted had authors been encouraged to

1. Report the weights used to generate skewness and kurtosis combinations.
2. Show visualizations of data generated by conditions in their studies.
3. Discuss representativeness of the data.

The Strange Case of the Fixed Decision Criterion

A Fixed Cutoff Test for Sample Mean Differences?

Suppose you are talking your undergraduate class through hypothesis testing, with a very basic example comparing means from two independent groups with equal sample size and known σ.

The null hypothsesis is that $\mu_1 - \mu_2 = 0$.

A bright and enthusiastic student jumps the gun and suggests a "new cutoff strategy" to compare the means.

The Strange Case of the Fixed Decision Criterion

A Fixed Cutoff Test for Sample Mean Differences?

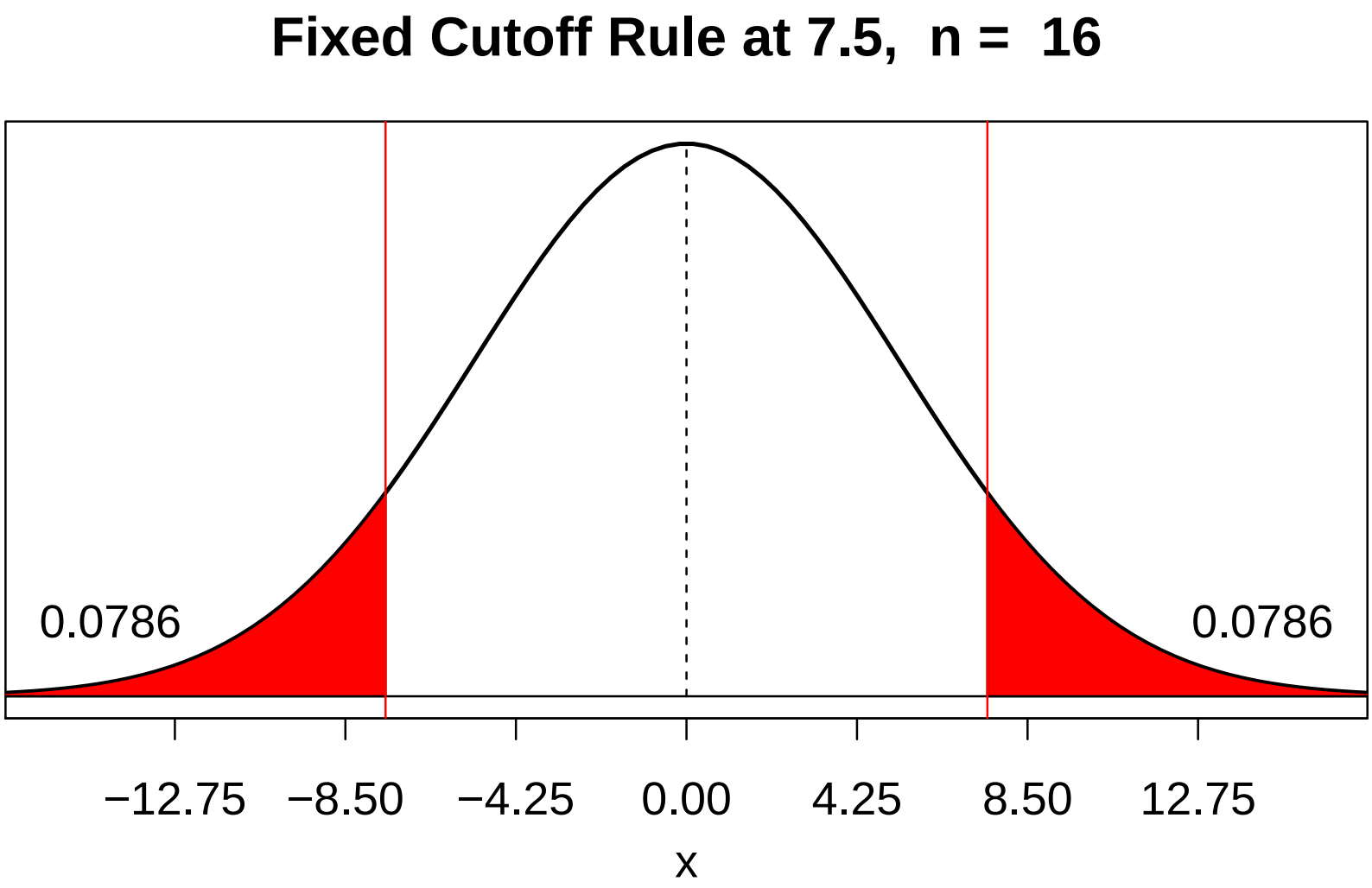
The New Cutoff Strategy. Reject the null hypothesis that $\mu_1 - \mu_2 = 0$ if and only if $|D| = |\overline{X}_{\bullet 1} - \overline{X}_{\bullet 2}| > c$, where $c = 0.5\sigma$.

Anyone proposing such a rule in the simple context of tests on means would quickly be reminded that a fixed cutoff for the sample estimate D doesn't work consistently across sample sizes, because the sampling variability of D varies.

Here is a demonstration of the fixed cutoff for testing $\mu_1 - \mu_2 = 0$ when $\sigma = 15$, $c = 7.5$, and the null hypothesis is true.

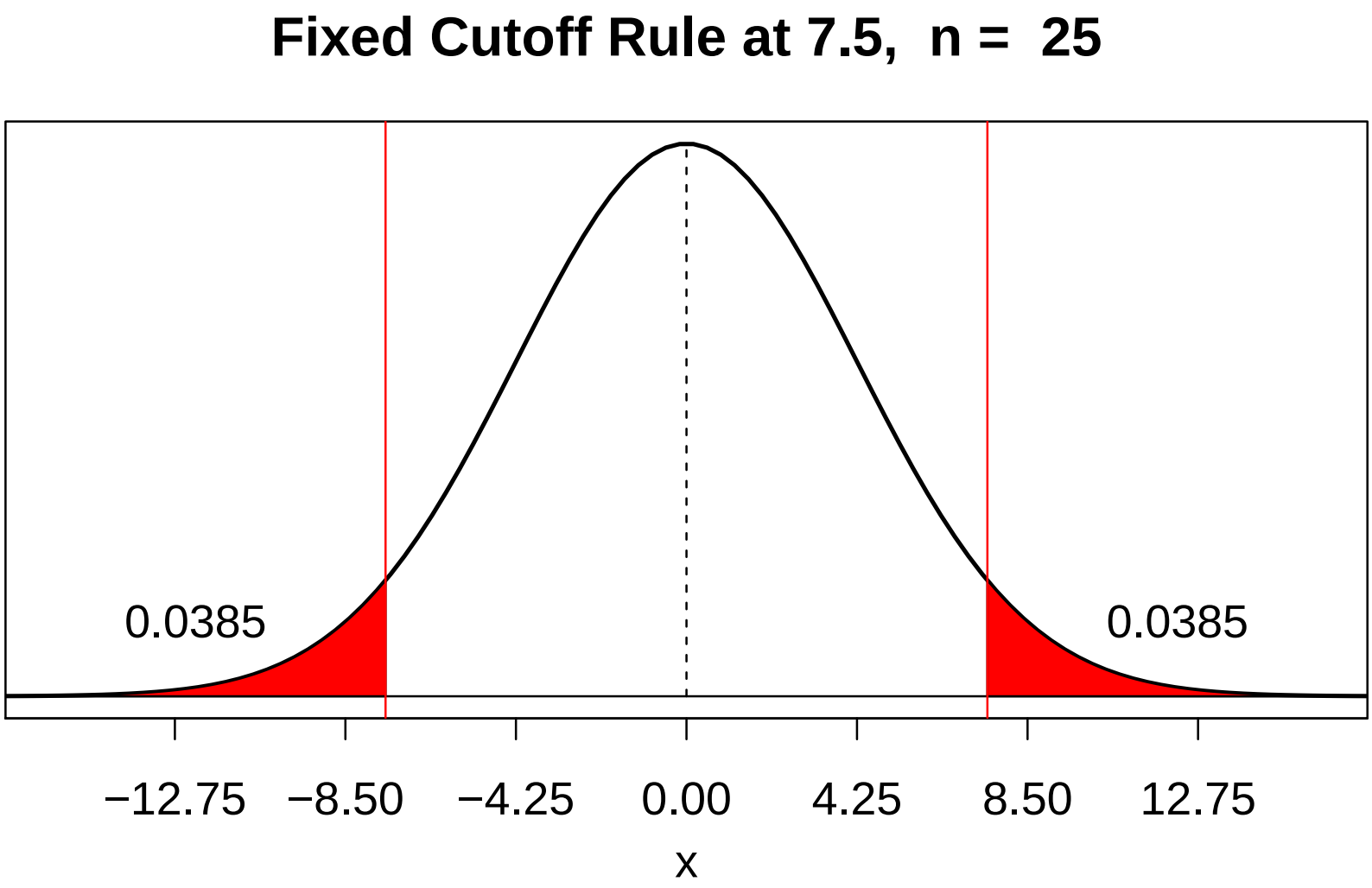
The Strange Case of the Fixed Decision Criterion

A Fixed Cutoff Test for Sample Mean Differences?



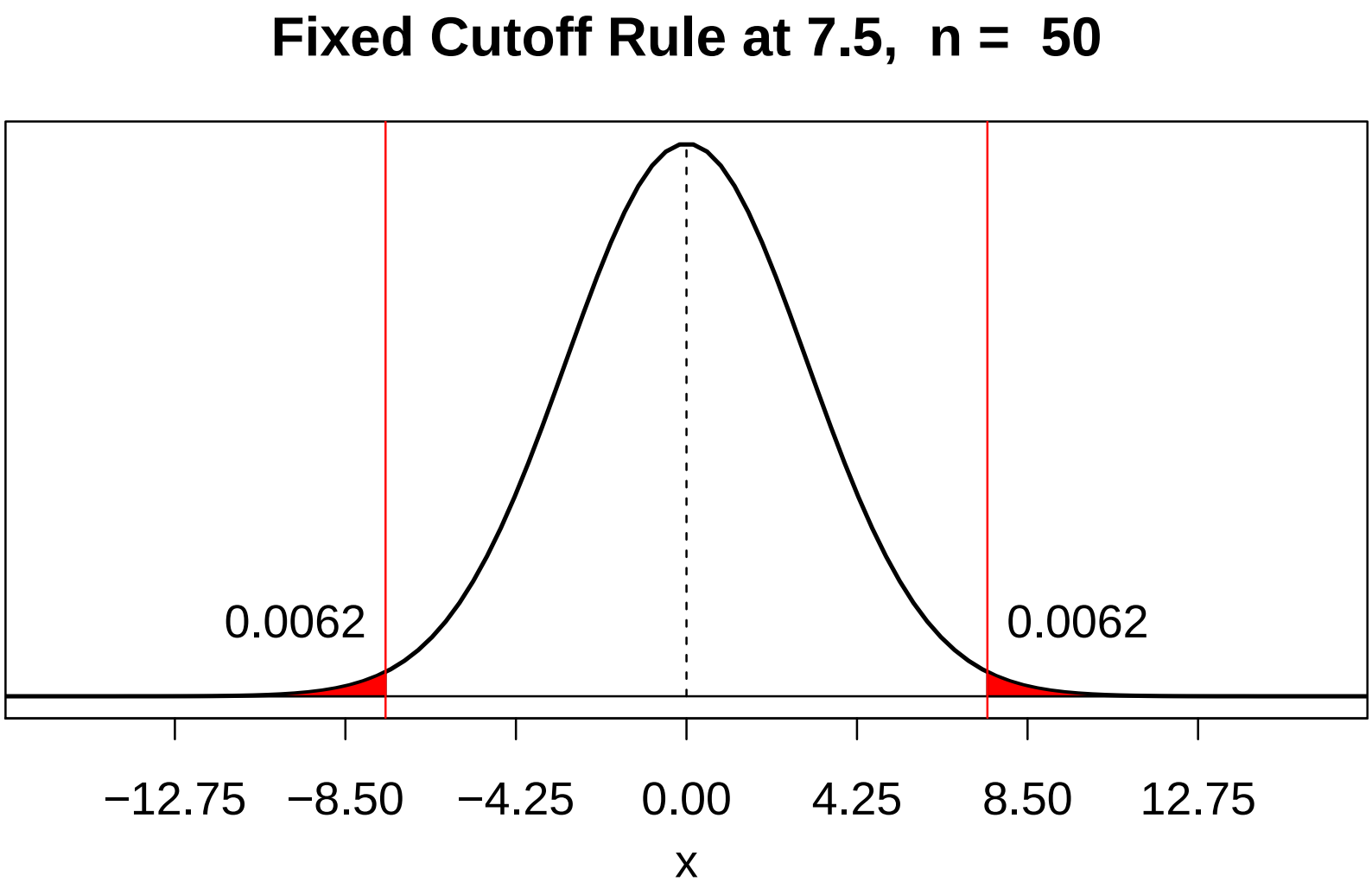
The Strange Case of the Fixed Decision Criterion

A Fixed Cutoff Test for Sample Mean Differences?



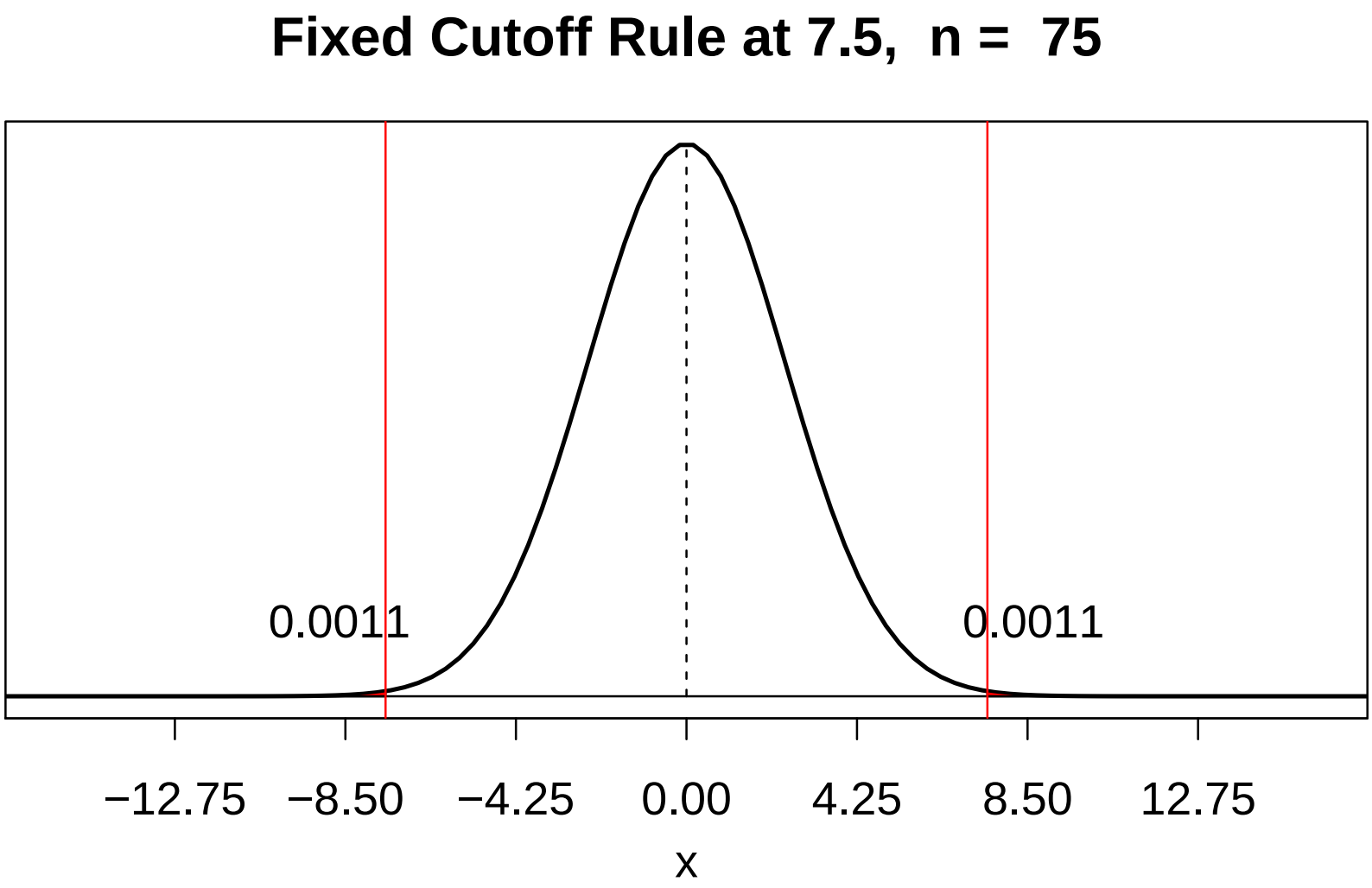
The Strange Case of the Fixed Decision Criterion

A Fixed Cutoff Test for Sample Mean Differences?



The Strange Case of the Fixed Decision Criterion

A Fixed Cutoff Test for Sample Mean Differences?



The Strange Case of the Fixed Decision Criterion

A Fixed Cutoff Test for Sample Mean Differences?

If you expect the fixed cutoff rule to replicate a legitimate hypothesis testing approach, you will be disappointed.

You find that the "half sigma test" rejects too often with small samples and hardly ever with large samples.

The Strange Case of the Fixed Decision Criterion

A Fixed Cutoff for Sample Fit Indices SEM

Although the fixed cutoff fallacy is obvious in the context of tests on means, the identical fallacy has persisted in covariance structure modeling for decades in various forms.

For example, a number of studies examined the efficacy of a cutoff value of .05 of the *sample* RMSEA as a device for determining when fit is not perfect—apparently oblivious to the fact that such a cutoff value cannot be found.

Perhaps lost in the shuffle was the fact that the entire concept of fit indices evolved from the consideration that model fit is almost never perfect!

The Strange Case of the Fixed Decision Criterion

A Fixed Cutoff for Sample Fit Indices SEM

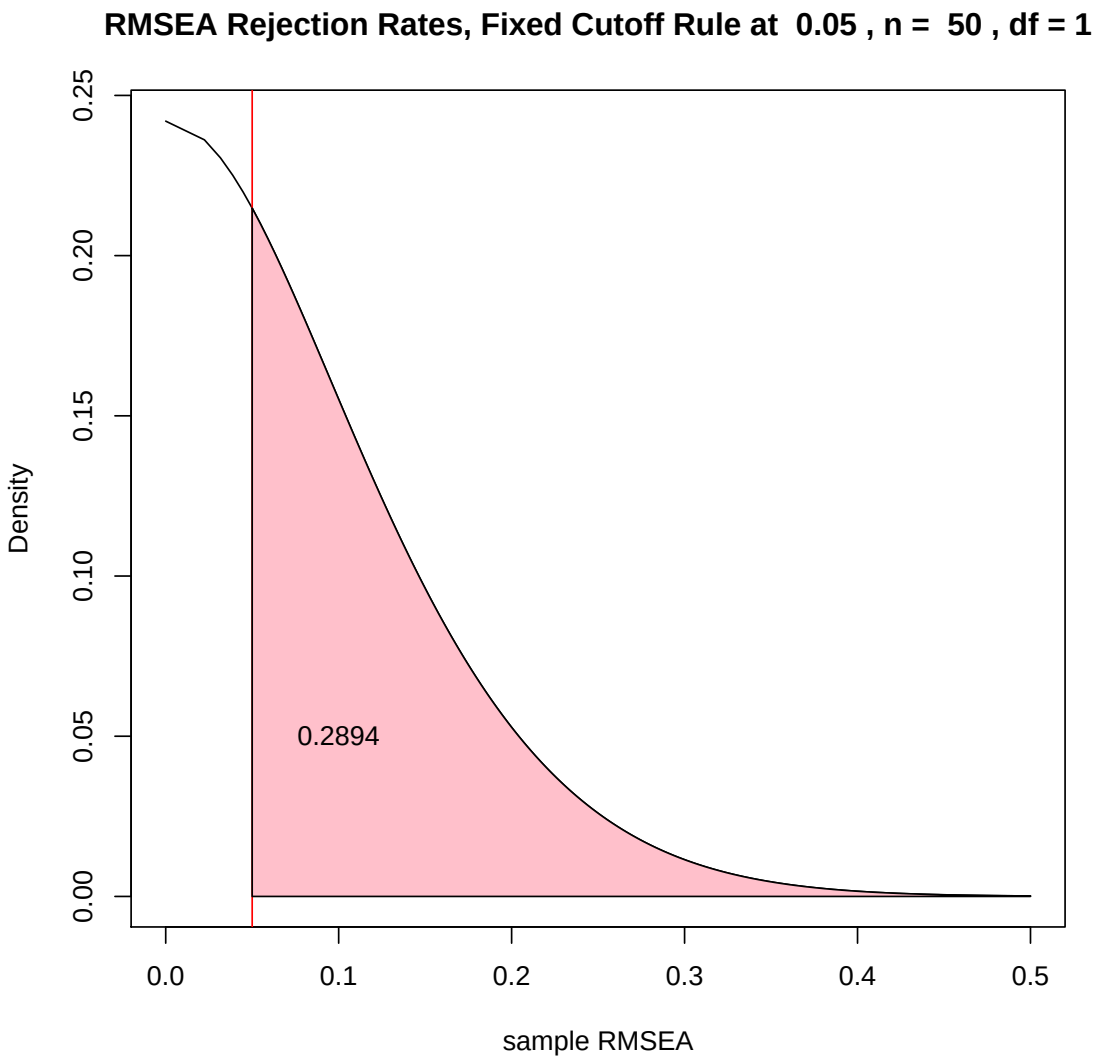
Imagine that someone proposed a sample RMSEA cutoff value of 0.05 as a criterion for rejecting the hypothesis of perfect model fit.

On the next few slides, we examine approximately what would happen with different degrees of freedom and different sample sizes.

We start with a single df model and vary the sample size.

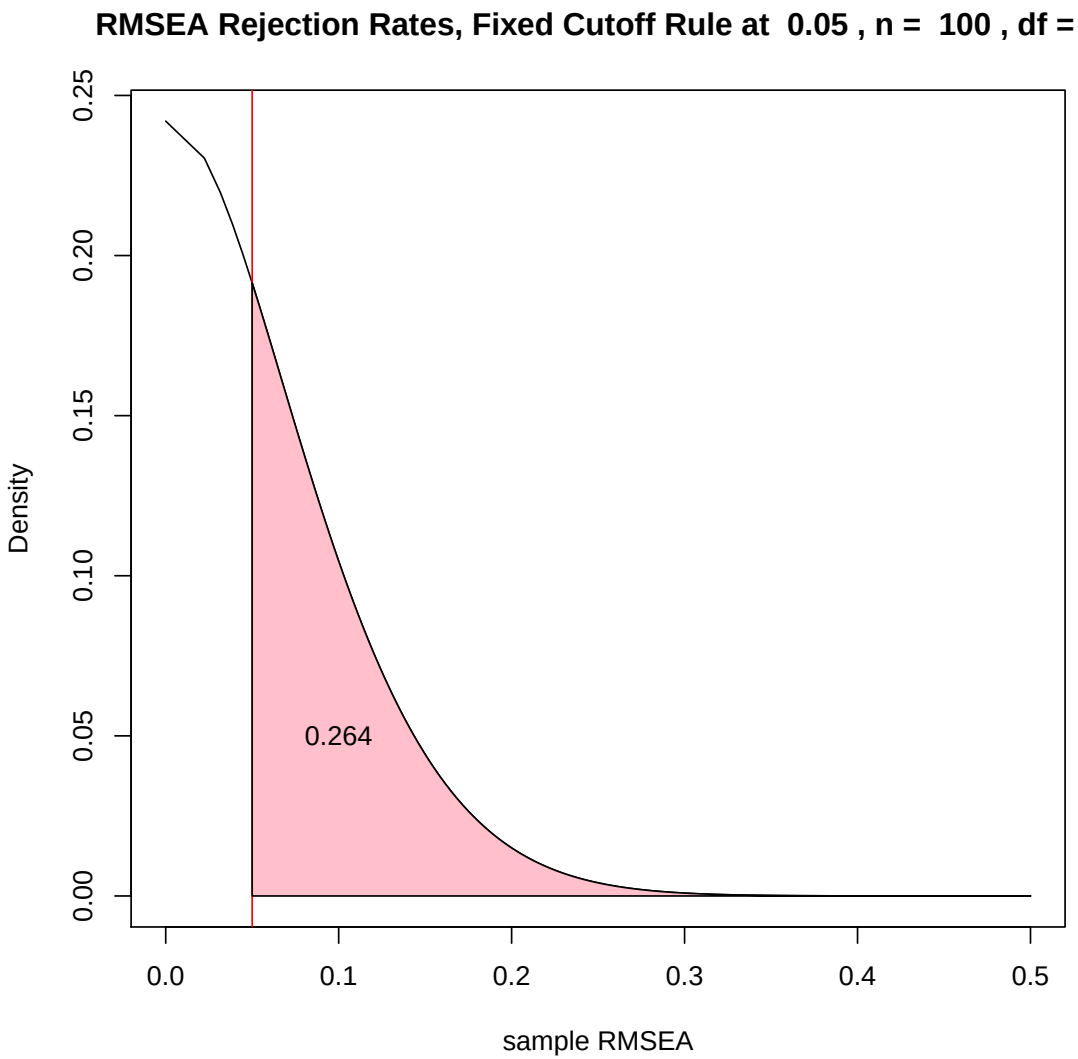
The Strange Case of the Fixed Decision Criterion

A Fixed Cutoff for Sample Fit Indices SEM



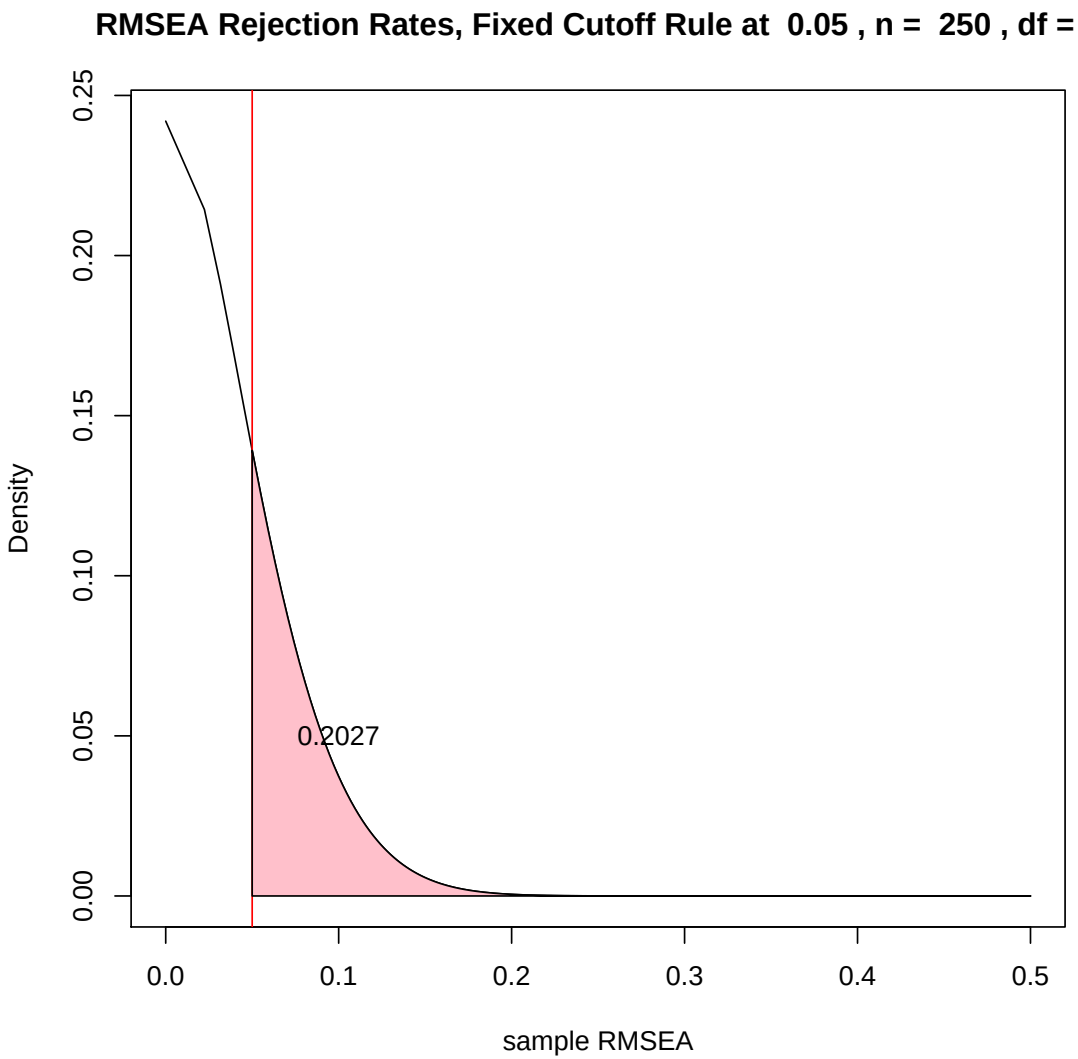
The Strange Case of the Fixed Decision Criterion

A Fixed Cutoff for Sample Fit Indices SEM



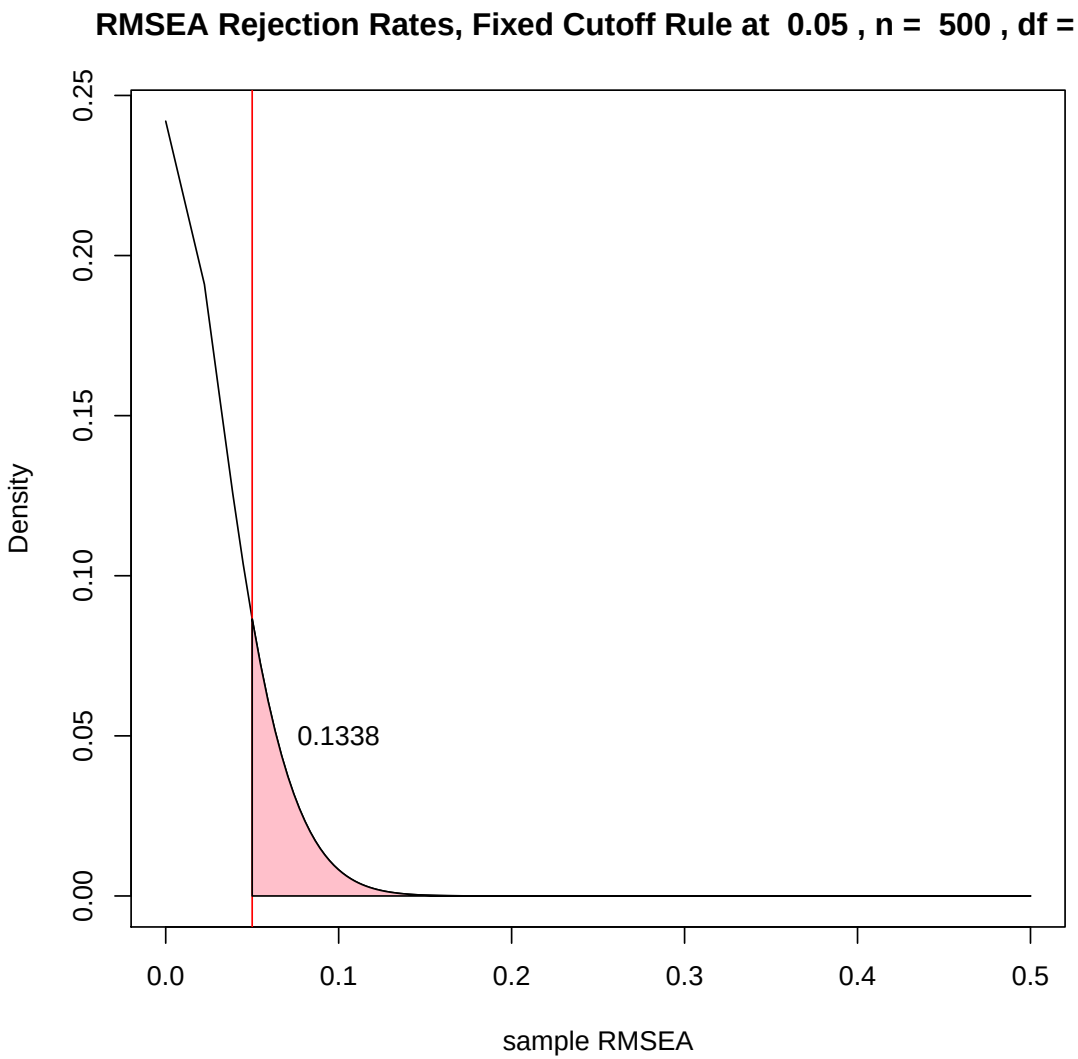
The Strange Case of the Fixed Decision Criterion

A Fixed Cutoff for Sample Fit Indices SEM



The Strange Case of the Fixed Decision Criterion

A Fixed Cutoff for Sample Fit Indices SEM



The Strange Case of the Fixed Decision Criterion

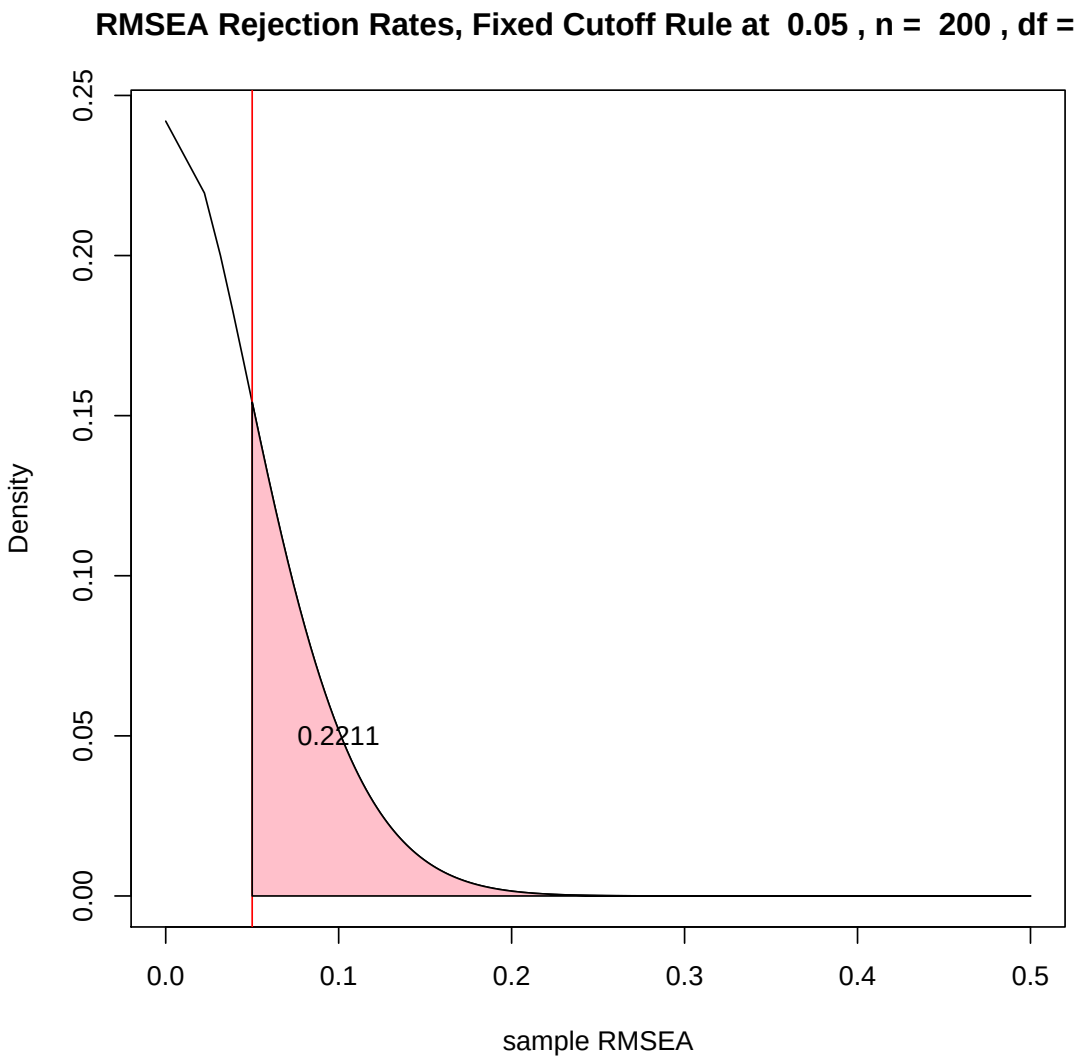
A Fixed Cutoff for Sample Fit Indices SEM

As n increases, the variability of the RMSEA decreases and the rejection rate goes from very high to very low.

Next, we examine the effect of degrees of freedom.

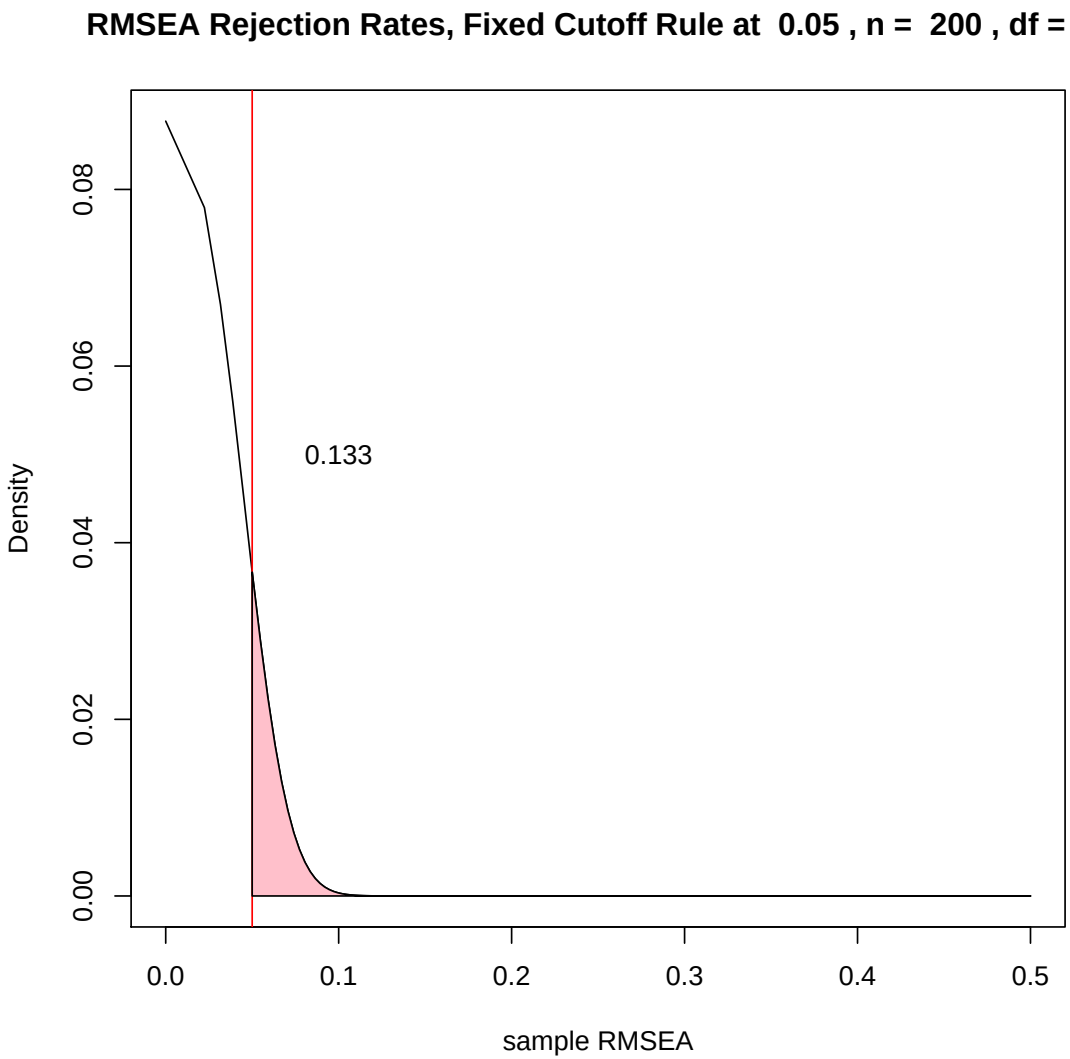
The Strange Case of the Fixed Decision Criterion

A Fixed Cutoff for Sample Fit Indices SEM



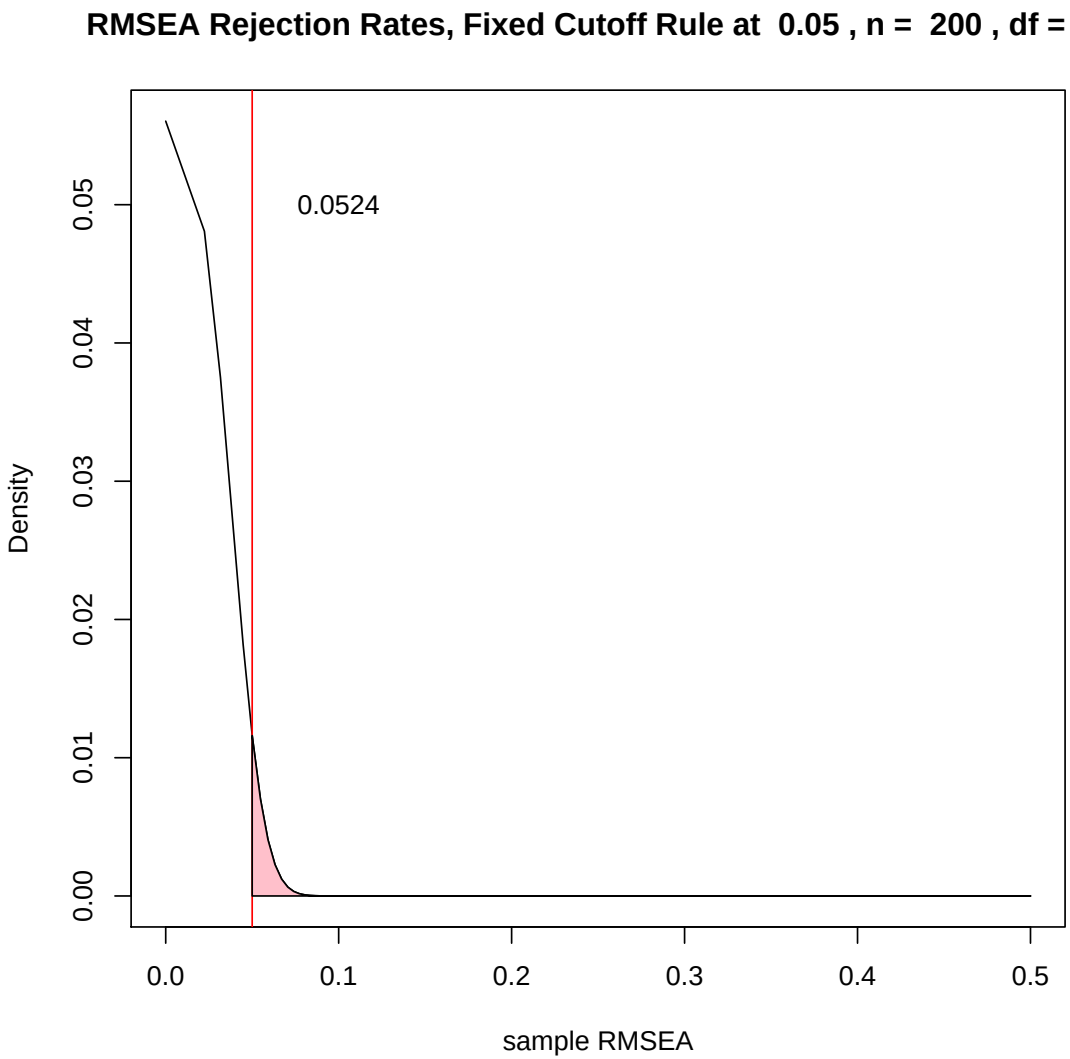
The Strange Case of the Fixed Decision Criterion

A Fixed Cutoff for Sample Fit Indices SEM



The Strange Case of the Fixed Decision Criterion

A Fixed Cutoff for Sample Fit Indices SEM



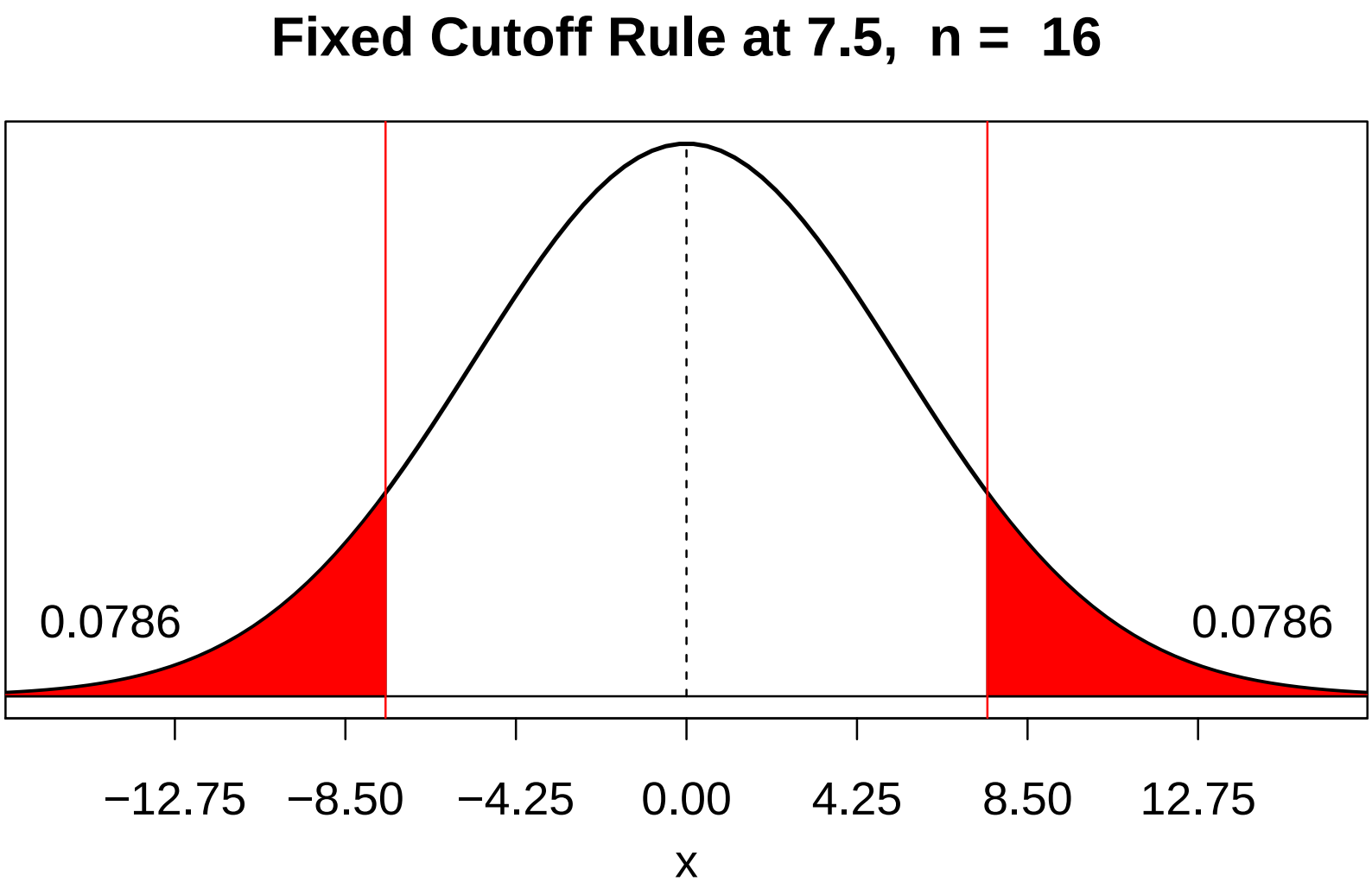
The Strange Case of the Fixed Decision Criterion

A Fixed Cutoff for Sample Fit Indices SEM

Let's play it again. First, tests on means with a fixed cutoff.

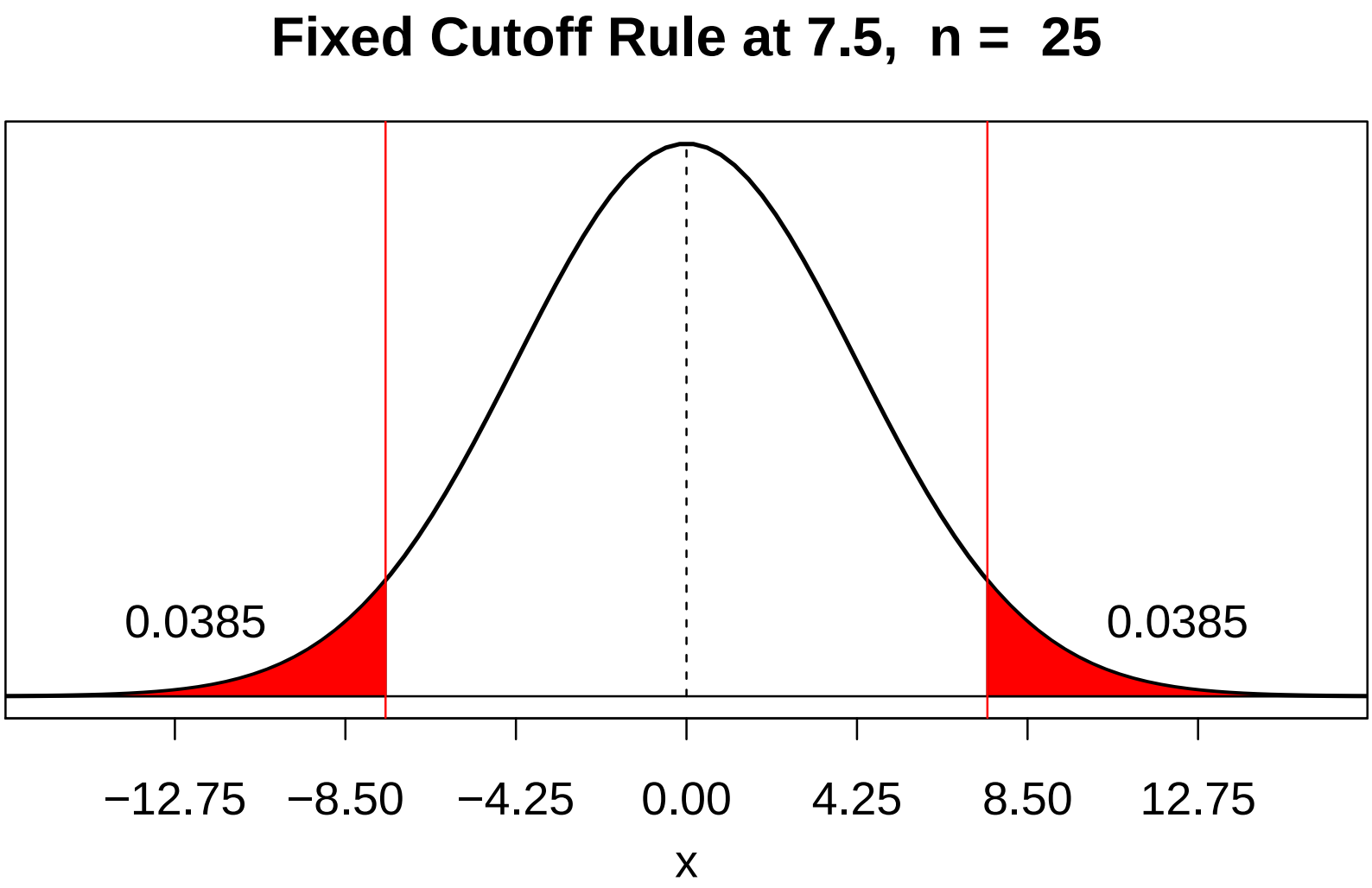
The Strange Case of the Fixed Decision Criterion

A Fixed Cutoff for Sample Fit Indices SEM



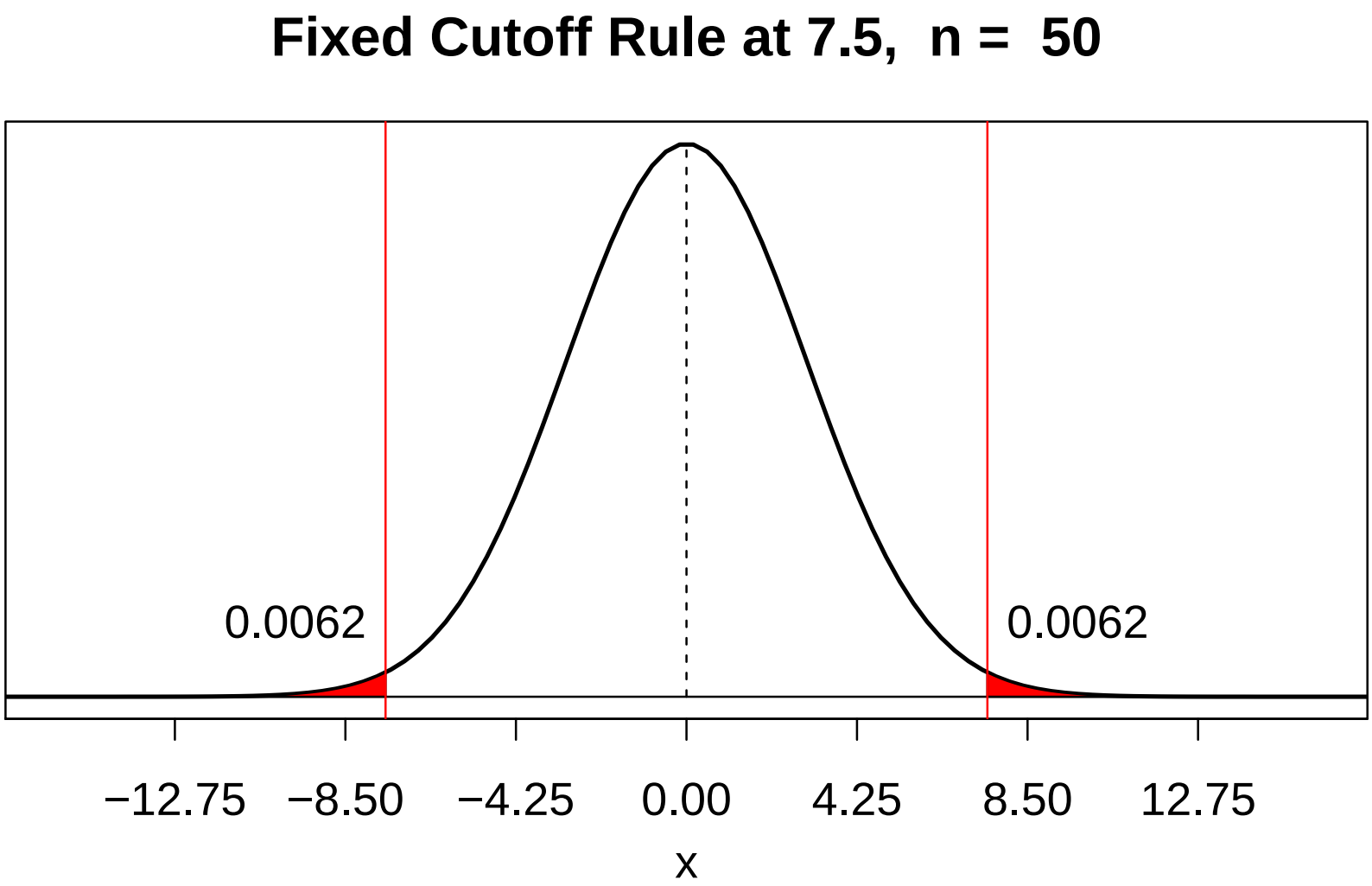
The Strange Case of the Fixed Decision Criterion

A Fixed Cutoff for Sample Fit Indices SEM



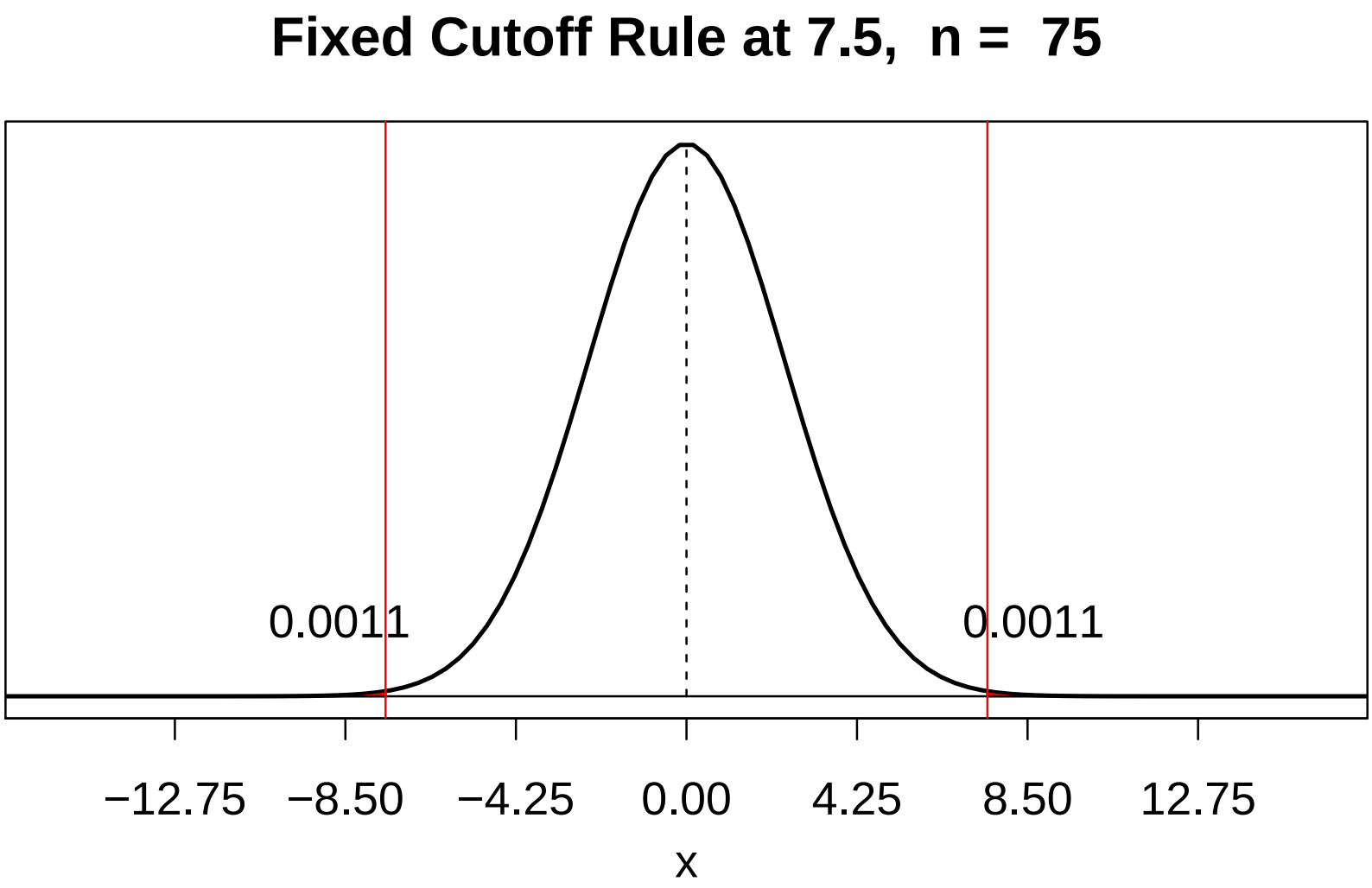
The Strange Case of the Fixed Decision Criterion

A Fixed Cutoff for Sample Fit Indices SEM



The Strange Case of the Fixed Decision Criterion

A Fixed Cutoff for Sample Fit Indices SEM



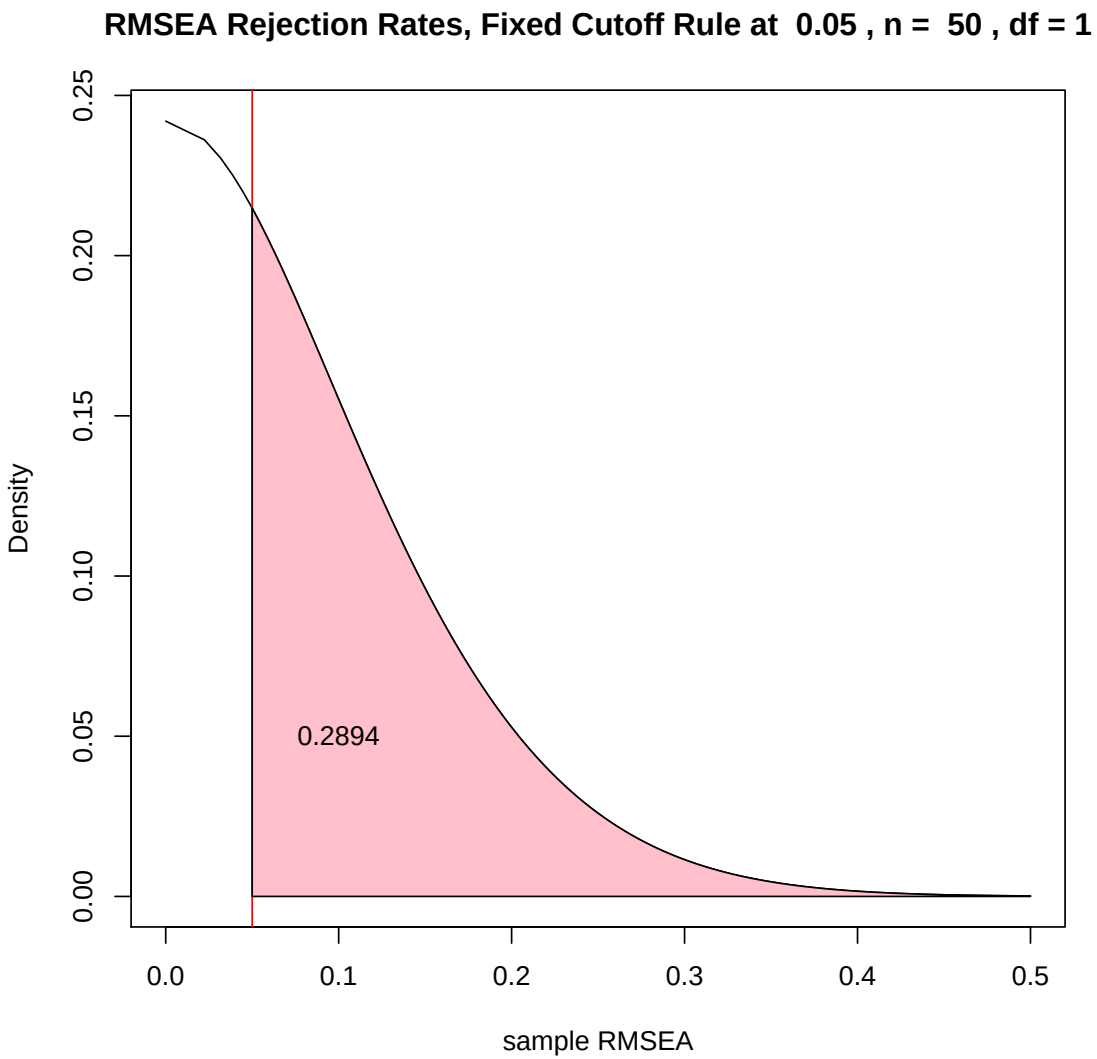
The Strange Case of the Fixed Decision Criterion

A Fixed Cutoff for Sample Fit Indices SEM

Next, tests of perfect fit using a fixed RMSEA cutoff.

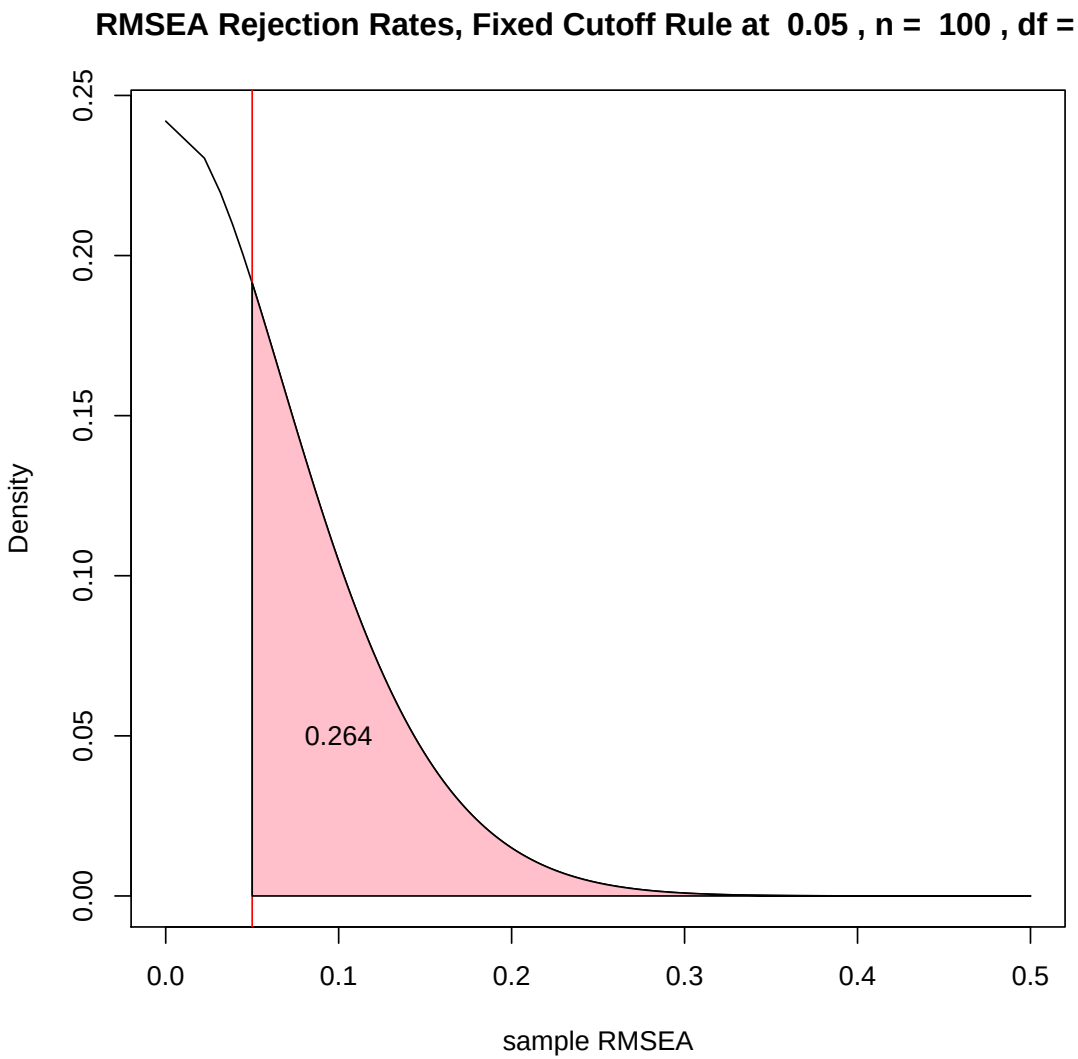
The Strange Case of the Fixed Decision Criterion

A Fixed Cutoff for Sample Fit Indices SEM



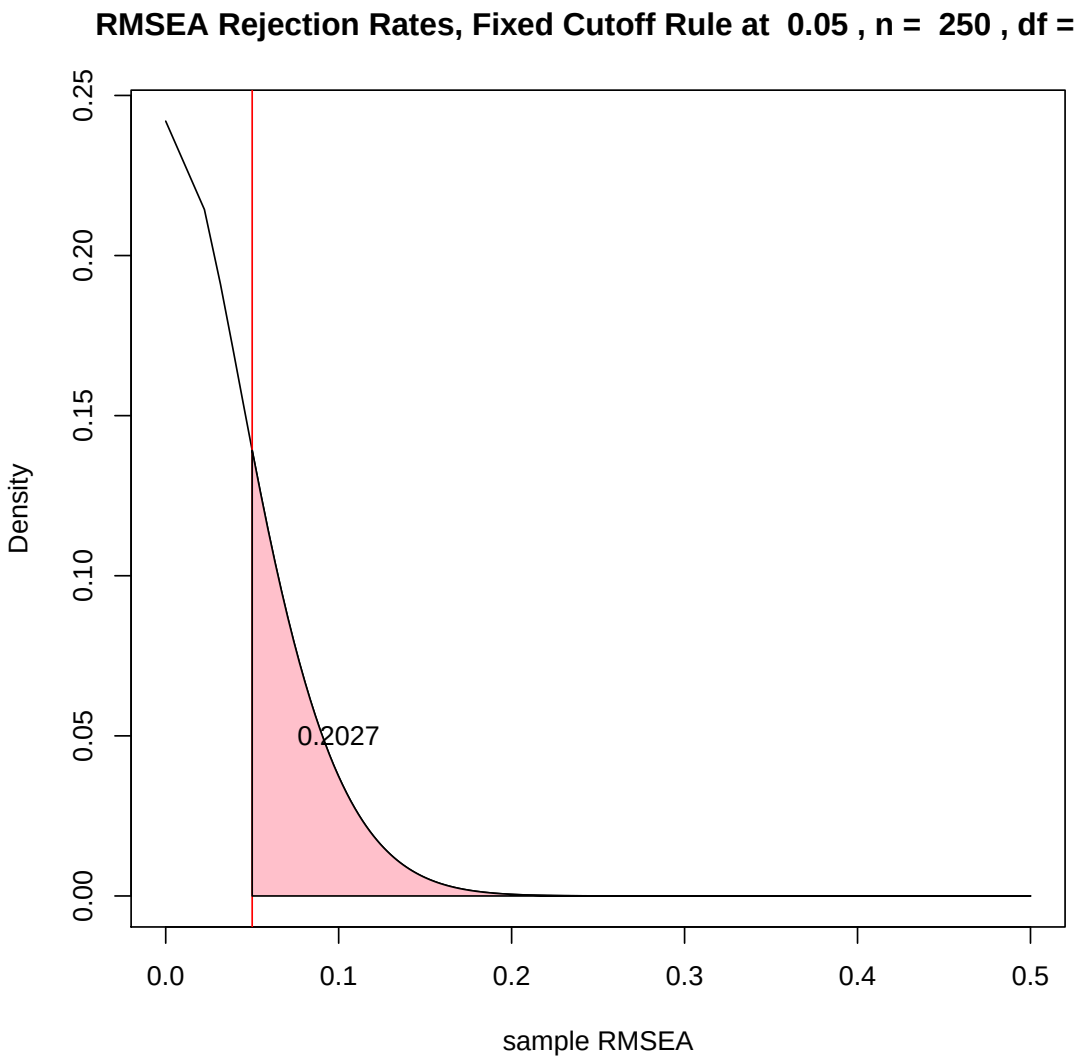
The Strange Case of the Fixed Decision Criterion

A Fixed Cutoff for Sample Fit Indices SEM



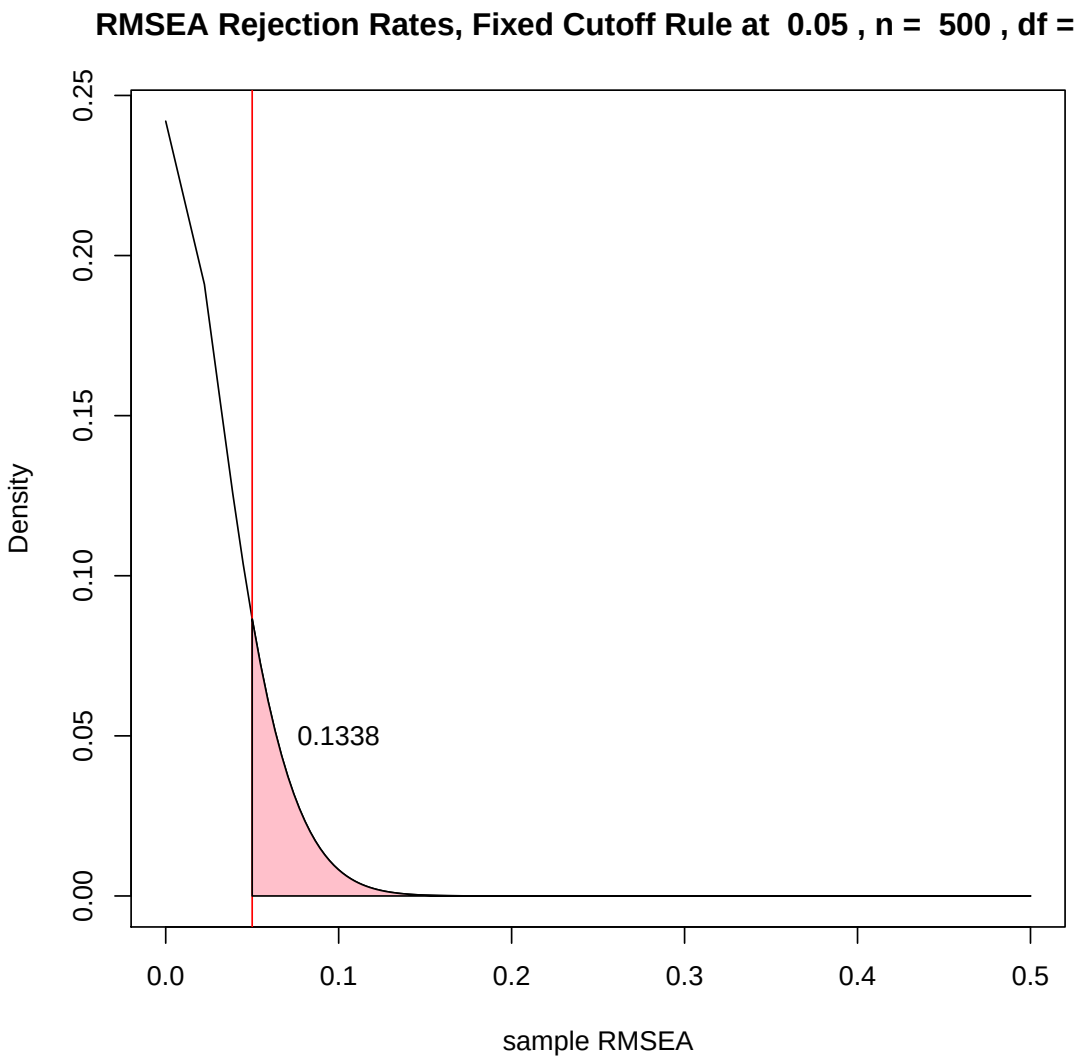
The Strange Case of the Fixed Decision Criterion

A Fixed Cutoff for Sample Fit Indices SEM



The Strange Case of the Fixed Decision Criterion

A Fixed Cutoff for Sample Fit Indices SEM



The Strange Case of the Fixed Decision Criterion

A Fixed Cutoff for Sample Fit Indices SEM

Does everyone get the picture?

One thing that emerges from the RMSEA plots is that things get worse as degrees of freedom get smaller.

This is because, as MacCallum, Browne, and Sugawara (1996) have pointed out, precision of estimation of population fit suffers when degrees of freedom are small.

Without an index of precision of estimation, a point estimate can be misleading.

Before I continue, let me ask you: Would you expect a fixed cutoff criterion to work consistently well with *any* point estimate of *any* SEM fit coefficient?

The Strange Case of the Fixed Decision Criterion

A Fixed Cutoff for Sample Fit Indices SEM

Several studies examined the fixed cutoff approach as if it was a viable method for model fitting.

Others, including papers by Lance, Marsh, and Kenny, have criticized the fixed cutoff approach in vague terms that fail to get to the heart of the matter.

This lack of precision often leads to statements and conclusions that are at best seriously misleading.

The Strange Case of the Fixed Decision Criterion

A Fixed Cutoff for Sample Fit Indices SEM

An example: a recently published article by Kenny, Kaniskan, and McCoach(2014).

The title of the article is "The Performance of the RMSEA in Models with Small Degrees of Freedom."

The title alone manages to be misleading in 3 different respects.

1. The article does not discuss the performance of the RMSEA as correctly used. It actually discusses the performance of the fixed cutoff rule. Steiger and Lind (1980) never mentioned either a point estimate or a fixed cutoff rule.
2. The title obscures the fact that the fixed cutoff rule doesn't work for degrees of freedom in general, not just small df.
3. The title deflects from the fact that other fit indices don't work with fixed cutoff rules either. The article mentions this only vaguely in a footnote.

The Strange Case of the Fixed Decision Criterion

A Fixed Cutoff for Sample Fit Indices SEM

Things go downhill from there, although, not surprisingly, I certainly agree that fixed cutoffs don't work well with small df.

Two Kinds of Standardization

Introduction

Standardization is an important tool in statistics. It can help us see more clearly what our data mean. But there are at least two different kinds of standardization, which I call *stochastically-motivated* and *measurement-motivated*.

Two Kinds of Standardization

Measurement-Motivated Standardization

A standardization is *Measurement-Motivated* if it removes the metric from data to allow it to be more consistently interpretable.

A classic example is in the context of the 2-sample t-statistic. We are interested in the standardized effect size

$$\delta = \frac{\mu_1 - \mu_2}{\sigma} \tag{2}$$

and we compute a t statistic as

$$t = \sqrt{n/2}\frac{\overline{X}_{\bullet 1} - \overline{X}_{\bullet 2}}{S} \tag{3}$$

Two Kinds of Standardization

Stochastically-Motivated Standardization

A standardization is *Stochastically-Motivated* if it leads to a distributional result for the hypothesis test.

It also happens to be the case that

$$\lambda = \sqrt{n/2}\ \delta = \sqrt{n/2}\frac{\mu_1 - \mu_2}{\sigma}$$

provides the formal noncentrality parameter for the t distribution that describes our t statistic. Hence, standardizing the mean difference by dividing by σ (and S) in this case is both measurement-motivated and stochastically-motivated.

Two Kinds of Standardization

Stochastically-Motivated Standardization

What makes this correspondence particularly useful is that we can compute a confidence interval on the noncentrality parameter of any of the classic noncentral distributions (t,F,χ^2), as Rachel Fouladi and I described in a tutorial article on "Noncentrality Interval Estimation" in 1997.

We get a CI for $\lambda = \sqrt{n/2}\ \delta$, divide the endpoints by $\sqrt{n/2}$, and we get back a CI on the quantity we are interested in.

Two Kinds of Standardization

The Correlated Sample t-Test

Consider the paired sample t test applied to two repeated measures. We again wish to test whether $\mu_1 - \mu_2 = 0$. In this case, the two kinds of standardization might be different. Suppose you decide on substantive grounds that what you are interested in is, once again,

$$\delta = \frac{\mu_1 - \mu_2}{\sigma} \tag{4}$$

Two Kinds of Standardization

The Correlated Sample t-Test

To get a test statistic, you need to divide the sample mean difference by s_D, the standard deviation of the difference scores. But in this case, the noncentrality parameter for the non-null distribution of the t statistic is (assuming equal variances)

$$\lambda = \sqrt{n/2}\frac{\mu_1 - \mu_2}{\sigma_D} \tag{5}$$

$$= \sqrt{n/2}\frac{\mu_1 - \mu_2}{\sqrt{\sigma^2 - \sigma_{12}}} \tag{6}$$

$$= \sqrt{n/2}\frac{\mu_1 - \mu_2}{\sigma\sqrt{1-\rho}} \tag{7}$$

$$= \frac{1}{\sqrt{1-\rho}}\sqrt{n/2}\ \delta \tag{8}$$

Two Kinds of Standardization

The Correlated Sample t-Test

In other words, in order to characterize the distribution correctly, by dividing the sample mean difference by s_D, we get a tractable distribution, but the quantity we are interested in is "polluted by covariance." In this case, measurement-motivated and stochastically-motivated standardization differ.

A discussion of this issue in the context of t-tests can be found in Steiger(1999).

Two Kinds of Standardization

Standardization in SEM

A virtually identical problem occurs in the context of SEM, in which, as discussed in Steiger(2000), the noncentrality parameter's standardization is clearly stochastically-motivated.

Consider the extremely simple SEM model that $\rho = \rho_0$. The asymptotic χ^2 statistic for this hypothesis is

$$\chi^2 = nF = n\frac{(r-\rho_0)^2}{(1-\rho_0^2)^2}$$

The noncentrality parameter is approximately

$$\lambda = nF^* = n\frac{(\rho-\rho_0)^2}{(1-\rho_0^2)^2}$$

Two Kinds of Standardization

Standardization in SEM

And, since there is only 1 df, the population RMSEA is approximately

$$\frac{|\rho - \rho_0|}{1 - \rho_0^2}$$

So the RMSEA will match $\rho - \rho_0$ well only for lower values of ρ_0.

Two Kinds of Standardization

Standardization in SEM

Note, however, that the "best" characterization of effect size in the test of $\rho = \rho_0$ is open to debate. Is it $\rho - \rho_0$? Is it (assuming positive correlations) $\rho^2 - \rho_0^2$?

These are dilemmas.

Two Kinds of Standardization

Standardization in SEM

Clearly, fixed *population* cutoffs for the RMSEA need to be taken with a grain of salt. But their use as a point of calculation has yielded some valuable insights. Can we do better? In simple cases, certainly.

In the previous example, one obvious solution is to create a confidence interval on $r - \rho_0$, the simplest special case of the RMSR. More complex situations are definitely challenging.

Conclusions

We can find many examples in the published literature in which principles that seem obvious in simple contexts slip by us.

Can we do better? I hope so.

I'm not being rhetorical here, or holier than thou. I'm one of the offenders.

Hindsight can be 20/20. One thing is sure – it is hard to solve a problem you don't see.

I hope to be back next year with a solution to at least one problem.