Discussion of Optimal bandwidth choice for robust bias-corrected inference in regression discontinuity designs

Bas Machielsen (discussant)

2025-06-18

Introduction

Introduction

- The regression discontinuity (RD) design is widely used in program evaluation, causal inference, and treatment effect settings.
- In the past, empirical work in RD designs often employs a mean square error (MSE) optimal bandwidth for local polynomial estimation of and inference on treatment effects.
- This MSE-optimal bandwidth choice yields an MSE-optimal RD point estimator, but is by construction invalid for inference.
- Robust bias-corrected (RBC) inference methods provide a natural solution to this problem: RBC confidence intervals and related inference procedures remain valid even when the MSE-optimal bandwidth is used.
- This paper: the authors establish Coverage Error-optimal RBC confidence interval estimators
  - Analogously, minimizing the error in rejection probability of the associated hypothesis testing procedures for RD treatment effects

RBC Optimal Inference vs. CE Optimal Inference

- Robust bias-corrected (RBC) inference methods are valid when using the MSE-optimal bandwidth
  - Paper shows that they yield suboptimal confidence intervals in terms of coverage error.
  - Establishes coverage error expansions for RBC confidence interval estimators and uses these results to propose new inference-optimal bandwidth choices for forming these intervals.
- Authors derive a CE-optimal bandwidth choice designed to minimize coverage error of the interval estimator, which is a fundamentally different goal than minimizing the MSE of the point estimator.
  - The MSE- and CE-optimal bandwidths are complementary, as both can be used in empirical work to construct, respectively, optimal point estimators and optimal inference procedures for RD treatment effects.

Set-Up

- Estimand:

$$\tau_\nu = \tau_\nu(c) = \frac{\partial^\nu}{\partial x^\nu} E[Y_i(1) - Y_i(0) | X_i = x]\Big|_{x=c}$$

- A generalized set of treatment effects (regression discontinuity corresponding to $v = 0$, kink to $v = 1$, etc.)
- Restrict attention to $v = 0$ case in presentation.

Set-Up [2]

- Point Estimate:

The idea is to first choose a neighbourhood around the cutoff c via a positive bandwidth choice h, and then employ (local) weighted polynomial regression using only observations with score X_i lying within the selected neighbourhood.

- Estimator that follows from this idea:

$$\hat{\tau}(h) = \nu! e' \hat{\beta}_{+,p}(h) - \nu! e' \hat{\beta}_{-,p}, \qquad v = 0, 1, \ldots, p$$

where (e.g.):

$$\beta\hat{}_{-,p} = \text{argmin}_{\beta \in R^{p+1}} \sum_{i=1}^{N} 1(c > X_i)(Y_i - r_p(X_i - c)'\beta)^2 K_h(X_i - c)$$

with $r_p(x) = (1, x, \ldots, x^p)$.

Optimal Bandwidth

- With this estimator, the MSE can be approximated as follows:

$$E\left[(\hat{\tau}_\nu(h) - \tau_\nu)^2 \mid X_1, \ldots, X_n\right] \approx_p h^{2p+2-2\nu}\mathcal{B}^2 + \frac{\mathcal{V}}{nh^{1+2\nu}}$$

- Which by the first order condition yields:

$$h_{\mathrm{MSE}} = \left[\frac{(1+2\nu)\mathcal{V}}{2(1+p-\nu)\mathcal{B}^2}\right]^{\frac{1}{2p+3}} n^{-\frac{1}{2p+3}}$$

Plug-in Estimators

- The authors also derived closed form expressions of $\mathcal{V}$ and $\mathcal{B}$
- Since we don't know them, rely on plug-in estimators to pick the h_{MSE}
  - The resulting estimator is "feasible" since a preliminary bandwidth, e.g. Imbens and Kalyanaraman (2012), is used to arrive at a plug-in estimate of $\mathcal{B}$ and $\mathcal{V}$ respectively.

Robust Bias-corrected Inference

- The infeasible estimator $\hat{\tau}_\nu(h_{\mathsf{MSE}})$ and its data-driven counterpart $\hat{\tau}_\nu(\hat{h}_{\mathsf{MSE}})$ are MSE-optimal point estimators of τ_ν in large samples.
- These point estimators are used not only to construct the "best guess" of the unknown RD treatment effect τ_ν, but also to conduct statistical inference, in particular for forming confidence intervals for τ_ν.
- The standard approach employs a Wald test statistic under the null hypothesis, and inverts it to form the confidence intervals. Specifically, for some choice of bandwidth h, the t-test statistic takes the form

$$T(h) = \frac{\hat{\tau}_\nu(h) - \tau_\nu}{\sqrt{\hat{V}(h)/(nh^{1+2\nu})}},$$

where it is assumed that $T(h) \sim \mathcal{N}(0,1)$, at least in large samples.

Confidence Interval Based on RBC Estimate

- Hence the corresponding confidence interval estimator for τ_ν is

$$I_{\mathrm{US}}(h) = \left[\hat{\tau}_\nu(h) - z_{1-\alpha/2} \cdot \sqrt{\frac{\hat{V}(h)}{nh^{1+2\nu}}},\ \hat{\tau}_\nu(h) + z_{1-\alpha/2} \cdot \sqrt{\frac{\hat{V}(h)}{nh^{1+2\nu}}}\right],$$

where z_α denotes the (100α)-percentile of the standard normal distribution.

The Problem This Paper Solves

- The confidence interval $I_{\mathrm{US}}(h)$ will only have correct asymptotic coverage, in the sense of $P[\tau_\nu \in I_{\mathrm{us}}(h)] = 1 - \alpha + o(1)$, if h obeys $nh^{2p+3} \to 0$, that is, the bandwidth is "small enough".
  - In particular, the MSE-optimal bandwidth is too large: the CI does not approach $(1 - \alpha)$, rendering inference and confidence intervals based on the naive t-test invalid.

Why is the naive t test invalid?

In a nutshell:

- **MSE-Optimal Bandwidth (h_{MSE})**: This bandwidth is chosen to make the **squared bias** of the same order of magnitude as the **variance**. It balances them perfectly to minimize the total error of the *point estimate*.
- **Valid Naive Inference**: This requires the t-statistic $T(h)$ to be centered around zero. For this to happen, the **bias** must be of a *smaller order of magnitude* than the **standard error**.
- h_{MSE} creates a bias that is "too big" relative to the standard error, so the t-statistic is not centered at zero, and the confidence interval is systematically off-center.

Asymptotic distribution of the raw estimator $\hat{\tau}_\nu(h)$.

Standard results in local polynomial regression (which form the basis for the MSE formula) show that the estimator $\hat{\tau}_\nu(h)$ is asymptotically normal, but with a bias term. For a large sample size n, its distribution is:

$$\hat{\tau}_\nu(h) \quad \approx \quad \mathcal{N}\left(\tau_\nu + \underbrace{h^{p+1-\nu}\mathcal{B}}_{\text{Bias}}, \quad \underbrace{\frac{\mathcal{V}}{nh^{1+2\nu}}}_{\text{Variance}}\right)$$

Where `B` and `V` are the bias and variance constants from the MSE formula.

Asymptotic Distribution of t Stat

The naive t-statistic is defined as:

$$T(h) = \frac{\hat{\tau}_\nu(h) - \tau_\nu}{\sqrt{\hat{V}(h)/(nh^{1+2\nu})}}$$

Let's analyze the numerator and denominator:

- The denominator is the estimated **standard error**. Since $\hat{V}(h)$ is a consistent estimator for `V`, the denominator converges to $\sqrt{\mathcal{V}/(nh^{1+2\nu})}$.
- The numerator is the centered estimator: $\hat{\tau}_\nu(h) - \tau_\nu$. From Step 1, this part is distributed as $\mathcal{N}(\text{Bias}, \text{Variance})$.

Distribution of t stat.

Now, let's divide the numerator by the (true) standard error to see the distribution of the t-statistic:

$$T(h) \approx \frac{\mathcal{N}\left(h^{p+1-\nu}\mathcal{B}, \frac{\mathcal{V}}{nh^{1+2\nu}}\right)}{\sqrt{\frac{\mathcal{V}}{nh^{1+2\nu}}}} = \mathcal{N}\left(\frac{h^{p+1-\nu}\mathcal{B}}{\sqrt{\frac{\mathcal{V}}{nh^{1+2\nu}}}}, \quad 1\right)$$

- Simplify the mean term:

$$\text{Mean of } T(h) = \frac{h^{p+1-\nu}\mathcal{B} \cdot \sqrt{n}h^{(1+2\nu)/2}}{\sqrt{\mathcal{V}}} = \frac{\mathcal{B}}{\sqrt{\mathcal{V}}}\sqrt{nh^{2p+3}}$$

So, the asymptotic distribution of the naive t-statistic is:

$$T(h) \xrightarrow{d} \mathcal{N}\left(\frac{\mathcal{B}}{\sqrt{\mathcal{V}}}\sqrt{nh^{2p+3}}, \quad 1\right)$$

Convergence result

- Check the condition $nh^{2p+3} \to 0$.

From the result before, $T(h)$ will only converge to a *standard* normal $N(0,1)$ if its mean converges to zero. This happens if and only if:

$$nh^{2p+3} \to 0$$

This is the condition for a "small enough" bandwidth, often called an **"undersmoothed" bandwidth**.

Plug-in h_{MSE}

Now we plug in the MSE-optimal bandwidth `h_MSE`. The MSE-optimal bandwidth is defined by its rate:

$$h_{\text{MSE}} = C \cdot n^{-1/(2p+3)}$$

where `C` is the constant $[(1+2\nu)V/(2(1+p-\nu)B^2)]^(1/(2p+3))$.

Let's substitute this specific h into the term nh^{2p+3}:

$$\begin{aligned}
n(h_{\text{MSE}})^{2p+3} &= n\left(C \cdot n^{-1/(2p+3)}\right)^{2p+3} \\
&= n \cdot C^{2p+3} \cdot \left(n^{-1/(2p+3)}\right)^{2p+3} \\
&= n \cdot C^{2p+3} \cdot n^{-1} \\
&= C^{2p+3}
\end{aligned}$$

Conclusion

This is a **non-zero constant**. It does **not** converge to 0. Because $n(h_{MSE})^{2p+3}$ converges to a non-zero constant, the mean of the t-statistic `T(h_MSE)` also converges to a non-zero constant:

$$T(h_{\text{MSE}}) \xrightarrow{d} \mathcal{N}(\mu, 1) \quad \text{where } \mu = \frac{\mathcal{B}}{\sqrt{\mathcal{V}}}\sqrt{C^{2p+3}} \neq 0$$

The confidence interval $I_{US}(h_{MSE})$ is constructed assuming the test statistic follows a $N(0,1)$ distribution. But in reality, it follows a normal distribution shifted by μ.

Therefore, the probability of coverage is:

$$P[\tau_\nu \in I_{\text{US}}(h_{\text{MSE}})] = P[-z_{1-\alpha/2} \leq T(h_{\text{MSE}}) \leq z_{1-\alpha/2}] \to P[-z_{1-\alpha/2} \leq \mathcal{N}(\mu, 1) \leq z_{1-\alpha/2}]$$

This probability is the area under a normal curve centered at μ. Because μ is not zero, this area is **not** equal to $1-\alpha$. The inference is invalid.

Solution

- Bias correction is an alternative to undersmoothing. Calonico et al. (2014) introduced a robust bias-correction method:

$$T_{\text{RBC}}(h) = \frac{\hat{\tau}_{\nu,\text{BC}}(h) - \tau_\nu}{\sqrt{\hat{V}_{\text{BC}}(h)/(nh^{1+2\nu})}}, \quad \text{where} \quad \hat{\tau}_{\nu,\text{BC}}(h) = \hat{\tau}_\nu(h) - h^{1+p-\nu}\hat{B}(b),$$

and

$$I_{\text{RBC}}(h) = \left[\hat{\tau}_{\nu,\text{BC}}(h) - z_{1-\alpha/2} \cdot \sqrt{\frac{\hat{V}_{\text{BC}}(h)}{nh^{1+2\nu}}},\ \hat{\tau}_{\nu,\text{BC}}(h) + z_{1-\alpha/2} \cdot \sqrt{\frac{\hat{V}_{\text{BC}}(h)}{nh^{1+2\nu}}}\right],$$

For inference, a key feature is that $\hat{V}_{\text{BC}}(h)$ is an estimator of the variance of $\hat{\tau}_{\nu,\text{BC}}(h)$, not of the variance of $\hat{\tau}_\nu(h)$.

Stepwise

- The Main Task (using bandwidth h):
  - Running a local polynomial regression of order p using a bandwidth h. This gives us our initial, biased estimate $\hat{\tau}_\nu(h)$.
- The Source of the Bias:
  - The bias in $\hat{\tau}_\nu(h)$ comes from the fact that our polynomial of order p fails to perfectly capture the true shape of the CE-function. The dominant part of this approximation error is related to the next highest-order term we omitted, which is the $(p+1)$-th derivative. The bias constant B is a function of this $(p+1)$-th derivative.
- Estimating the Bias (using bandwidth b):
  - To correct for the bias, we can't use the true bias B because it depends on an unknown derivative. We must estimate it. The text explicitly states that estimating $\hat{B}$ requires a local polynomial regression of order $p+1$.
  - We need to use a regression of order $p+1$ because that's the order required to estimate the $(p+1)$-th derivative.
  - Like any local polynomial regression, this procedure requires its own bandwidth. We call this the auxiliary bandwidth, b.

Three claims

Paper shows that:

- $I_{RBC}(h)$ has asymptotic coverage error $(P(\tau \in CI(X)) - (1-\alpha))$that is no larger than $I_{US}(h)$, and is strictly smaller in most practically relevant cases, even when the corresponding best possible bandwidth is used to construct each confidence interval.
- Employing the MSE-optimal bandwidth h_{MSE} to construct $I_{RBC}(h)$ is valid but suboptimal in terms of coverage error.
- Paper drives new optimal bandwidth choices that minimize the coverage error of the RBC confidence intervals.

Three Proof Sketches: Claim 1

Claim: $I_{RBC}(h)$ has an asymptotic coverage error that is no larger than that of $I_{US}(h)$, and is strictly smaller in most cases, even when using the best possible bandwidth for each.

Proof Sketch: The proof rests on comparing the rates at which the coverage errors of the two intervals vanish as the sample size n grows. A faster rate means a better interval. The key tool is the Edgeworth expansion of the coverage error from the paper's Theorem 3.1.

- Coverage Error for Undersmoothing (I_{US}): The naive t-statistic $T(h)$ has a non-zero asymptotic mean (bias) unless h is chosen to be "small enough." The coverage error of $I_{US}(h)$ is dominated by two terms: one from variance and one from this uncorrected bias.

$$\text{Coverage Error}(I_{\text{US}}(h)) \approx \underbrace{C_1 \cdot n^{-1}h^{-1}}_{\text{Variance Term}} + \underbrace{C_2 \cdot (nh^{p+1-\nu})^2}_{\text{Squared Bias Term}} + \dots$$

- To get the best possible US interval, we must choose h to balance these two terms. This leads to the optimal undersmoothing bandwidth, h_{US}, which has a rate of $h_{US} \propto n^{-1/(p+2)}$. Plugging this optimal rate back into the error formula shows that the fastest possible coverage error for I_{US} vanishes at a rate of $O(n^{-(p+1)/(p+2)})$.

Proof Sketch: Claim 1 - Why is that the Coverage Error?

Given the earlier derived expression for the shift of our t-statistic from zero:

$$\mu_T \approx \frac{\mathcal{B}}{\sqrt{\mathcal{V}}} \cdot \sqrt{nh^{2p+3}}$$

A confidence interval $[-z, z]$ for a standard normal variable covers $1-\alpha$ of the probability mass. If our variable is actually a normal variable with mean μ_T instead of 0, the coverage probability is $P(-z \leq \mathcal{N}(\mu_T, 1) \leq z)$.

Using a Taylor expansion of this probability around $\mu_T = 0$, (▸ Derivation), we find that the first-order effect of the shift cancels out due to symmetry, and the dominant error term is proportional to the **square of the mean shift**:

$$\text{Coverage Error from Bias} \approx \text{Constant} \times (\mu_T)^2$$

Now, substitute our expression for μ_T:

$$\text{Error from Bias} \approx \text{Constant} \times \left(\frac{\mathcal{B}}{\sqrt{\mathcal{V}}} \cdot \sqrt{nh^{2p+3}} \right)^2 = \underbrace{\left(\frac{\mathcal{B}^2}{\mathcal{V}} \cdot \text{Const} \right)}_{C_2} \cdot \left(nh^{2p+3} \right)$$

Claim 1: Origin of the Variance Term: $C_1 \cdot n^{-1}h^{-1}$

This term captures all other deviations from a perfect $N(0,1)$ distribution. It is not about the *center* of the distribution, but about its *shape* and *randomness*.

- The numerator has its own **kurtosis** (i.e., fatter or thinner tails than a normal distribution).
- The denominator $\hat{V}(h)$ is a **random variable** itself. It has its own variance. Dividing by a random variable, instead of a constant, induces extra variability and non-normality in the ratio.

Claim 1: Origin of the Variance Term: $C_1 \cdot n^{-1}h^{-1}$ [2]

- Formal analysis using Edgeworth expansions, but use heuristic: **the effective sample size**.
- The estimator $\hat{\tau}_\nu(h)$ is a local average. The number of data points it effectively uses is roughly proportional to nh.
  - Higher-order approximation errors (related to skewness, kurtosis, etc.) decrease as the sample size increases.
  - The leading error term in an Edgeworth expansion for a symmetric statistic (after accounting for bias) is typically of the order 1/(sample size).
  - In our case, the relevant sample size is nh. Therefore, the leading error term that captures all these shape distortions is proportional to:

$$\text{Error from Shape/Variance} \approx \frac{C_1}{nh} = C_1 \cdot n^{-1}h^{-1}$$

This term represents how much the coverage probability is distorted because our t-statistic isn't a perfect bell curve, primarily due to the randomness in the standard error estimate and the kurtosis of the point estimate.

Proof Sketch: Claim 1 [2]

- **Coverage Error for RBC ($I_R BC$)**: By construction, the RBC procedure removes the first-order bias term C_2. Its coverage error (similar argument) is therefore dominated by a variance term and a higher-order bias term.

$$\text{Coverage Error}(I_{\text{RBC}}(h)) \approx \underbrace{C_1' \cdot n^{-1} h^{-1}}_{\text{Variance Term}} + \underbrace{C_2' \cdot (nh^{p+2-\nu})^2}_{\text{Higher-Order Bias}} + \dots$$

- To get the best possible RBC interval, we choose h to balance these different terms.
  - This leads to the new CE-optimal bandwidth, h_{RBC}, which has a rate of $h_{RBC} \propto n^{-1/(p+3)}$. Plugging this optimal rate back in shows that the fastest possible coverage error for I_{RBC} vanishes at a rate of $O(n^{-(p+2)/(p+3)})$.

Proof Sketch: Claim 1 [3]

- Compare the two best-case error rates: $n^{-(p+1)/(p+2)}$ for US versus $n^{-(p+2)/(p+3)}$ for RBC.
  - For any $p \geq 0$, the exponent for RBC is larger in magnitude (e.g., for $p = 1$, we compare $n^{-2/3}$ vs $n^{-3/4}$).
  - This means the coverage error for the best RBC interval vanishes strictly faster than the coverage error for the best US interval.
  - This establishes the "strictly smaller" claim.
  - The "no larger than" part holds even under minimal smoothness assumptions where the first-order bias is still removed by RBC, making it at least as good as US.

Proof Sketch: Claim 2 [1]

Claim: Using the MSE-optimal bandwidth h_{MSE} to construct $I_{RBC}(h)$ is valid for inference, but the resulting interval is suboptimal in terms of coverage error.

Proof Sketch:

- Validity:
  - The RBC t-statistic is $T_{RBC}(h) = (\hat{\tau}_\nu(h) - \text{BiasEstimate} - \tau_\nu)/SE$.
  - The entire point of the BiasEstimate is to cancel the asymptotic bias of $\hat{\tau}_\nu(h)$.
  - This cancellation works regardless of the specific (valid) bandwidth h used.
  - Therefore, even when $h = h_{MSE}$, the numerator of $T_{RBC}(h_{MSE})$ is asymptotically centered at zero, and the whole statistic converges to a standard $N(0,1)$.
- This means hypothesis tests have the correct size and confidence intervals have the correct coverage in the limit, so inference is valid.

Proof Sketch: Claim 2 [2]

- Suboptimality comes from a mismatch in optimization goals.
  - h_{MSE} is derived by minimizing the Mean Squared Error of the point estimate: $MSE(h) = \text{Bias}^2 + \text{Variance}$. This involves balancing a variance term with the first-order bias term, which yields the rate $h_{MSE} \propto n^{-1/(2p+3)}$.
  - The Coverage Error of the RBC interval, as shown in sketch (a), is determined by a balance between the variance term and a higher-order bias term. Minimizing this error yields the optimal rate $h_{RBC} \propto n^{-1/(p+3)}$.
  - Since $1/(2p+3) \neq 1/(p+3)$ (for $p \geq 1$), the h_{MSE} rate is not the rate that optimally minimizes the coverage error of the RBC interval.
  - Plugging h_{MSE} into the RBC coverage error formula creates an imbalance between the error components. This leads to a coverage error that vanishes more slowly than the error achieved by using the purpose-built h_{RBC}. Therefore, while valid, it's not the best one can do.

Proof Sketch: Claim 3 [1]

Claim: We can derive a new optimal bandwidth h_{RBC} that minimizes the RBC coverage error, and this choice has positive consequences for interval length.

Proof Sketch (Theorem 3.2): The optimal bandwidth h_{RBC} is found by explicitly minimizing the Edgeworth expansion of the RBC coverage error with respect to h. This involves taking the derivative of the coverage error formula (from sketch 1) with respect to h, setting it to zero, and solving.

- This defines both the optimal rate $h_{RBC} \propto n^{-1/(p+3)}$ and the associated constant $\mathcal{H}$. In practice, this requires estimating the unknown quantities in the error formula and numerically solving for the optimal bandwidth (a "direct plug-in" approach).
- Consequences for Interval Length: The length of a confidence interval is directly proportional to its standard error. A larger bandwidth h leads to a shorter interval.
- We found the optimal rates: $h_{RBC} \propto n^{-1/(p+3)}$ and $h_{US} \propto n^{-1/(p+2)}$.
- Since $1/(p+3) < 1/(p+2)$, the h_{RBC} bandwidth vanishes more slowly, meaning for any large n, h_{RBC} will be larger than h_{US}.
- Because I_{RBC} optimally employs a larger bandwidth, its standard error is smaller, and its length is therefore asymptotically shorter than the best possible undersmoothed interval I_{US}.

The End

Appendix

- In an ideal world, our t-statistic T follows a standard normal distribution, $T \; N(0,1)$. A two-sided $(1-\alpha)$ confidence interval corresponds to finding the region $[-z_c, z_c]$ where $P(-z_c \leq T \leq z_c) = 1-\alpha$. Here, z_c is the critical value (e.g., 1.96 for $\alpha = 0.05$).
  - In the undersmoothing case with a "too large" bandwidth, our t-statistic $T(h)$ is *not* centered at zero. It is approximately normal, but with a non-zero mean μ_T.

$$T(h) \approx N(\mu_T, 1)$$

- We still use the same interval $[-z_c, z_c]$ because we are *assuming* it's a standard normal. The actual coverage probability is now $P(-z_c \leq N(\mu_T, 1) \leq z_c)$.
- We want to understand the difference between the actual and nominal coverage, which is the coverage error:

$$\text{Coverage Error} = P(-z_c \leq N(\mu_T, 1) \leq z_c) - (1-\alpha)$$

- We will analyze this by defining a function for the coverage probability and expanding it with a Taylor series.

The Taylor Expansion Derivation

Let $g(\mu)$ be the probability that a normal random variable with mean μ and variance 1 falls within the interval $[-z_c, z_c]$.

$$g(\mu) = P(-z_c \leq N(\mu, 1) \leq z_c)$$

Let $\Phi(x)$ be the CDF and $\phi(x)$ be the PDF of a standard normal $N(0, 1)$. We can write $g(\mu)$ using the standard normal CDF:

$$g(\mu) = \Phi(z_c - \mu) - \Phi(-z_c - \mu)$$

Our goal is to approximate $g(\mu_T)$ when μ_T is small. We use a second-order Taylor expansion of $g(\mu)$ around the point $\mu = 0$:

$$g(\mu) \approx g(0) + g'(0)\mu + \frac{g''(0)}{2}\mu^2$$

The coverage error is $g(\mu) - g(0)$. So, we need to find the first and second derivatives of $g(\mu)$ and evaluate them at $\mu = 0$.

Step 1: Calculate the First Derivative $g'(\mu)$

We use the chain rule and the fact that $d/dx\Phi(x) = \phi(x)$:

$$\begin{aligned} g'(\mu) &= \frac{d}{d\mu}\left[\Phi(z_c - \mu) - \Phi(-z_c - \mu)\right] \\ &= \phi(z_c - \mu) \cdot (-1) - \phi(-z_c - \mu) \cdot (-1) \\ &= -\phi(z_c - \mu) + \phi(-z_c - \mu) \end{aligned}$$

Now, evaluate at $\mu = 0$:

$$g'(0) = -\phi(z_c) + \phi(-z_c)$$

A key property of the standard normal PDF is that it is symmetric: $\phi(x) = \phi(-x)$. Therefore:

$$g'(0) = -\phi(z_c) + \phi(z_c) = 0$$

This is a crucial result. It means that for very small shifts in the mean, the coverage error is *not* proportional to μ. The leading error term must come from the second derivative.

Step 2: Calculate the Second Derivative $g''(\mu)$

We differentiate $g'(\mu)$ with respect to μ. We need the rule for the derivative of the PDF: $d/dx\phi(x) = -x\phi(x)$.

$$\begin{aligned}
g''(\mu) &= \frac{d}{d\mu}\left[-\phi(z_c - \mu) + \phi(-z_c - \mu)\right] \\
&= -\left(\phi'(z_c - \mu)\cdot(-1)\right) + \left(\phi'(-z_c - \mu)\cdot(-1)\right) \\
&= \phi'(z_c - \mu) - \phi'(-z_c - \mu) \\
&= (-(z_c - \mu)\phi(z_c - \mu)) - (-(-z_c - \mu)\phi(-z_c - \mu)) \\
&= -(z_c - \mu)\phi(z_c - \mu) - (z_c + \mu)\phi(-z_c - \mu)
\end{aligned}$$

Now, evaluate at $\mu = 0$:

$$g''(0) = -(z_c)\phi(z_c) - (z_c)\phi(-z_c)$$

Using the symmetry $\phi(z_c) = \phi(-z_c)$ again:

$$g''(0) = -z_c\phi(z_c) - z_c\phi(z_c) = -2z_c\phi(z_c)$$

This is a non-zero negative constant.

Step 3: Assemble the Result

Now we plug our derivatives back into the Taylor expansion for the coverage error, $g(\mu) - g(0)$:

$$\begin{aligned} \text{Coverage Error} &\approx g'(0)\mu + \frac{g''(0)}{2}\mu^2 \\ &\approx (0)\mu + \frac{-2z_c\phi(z_c)}{2}\mu^2 \\ &= -z_c\phi(z_c)\mu^2 \end{aligned}$$

Finally, we substitute our specific mean shift, $\mu = \mu_T$:

$$\text{Coverage Error from Bias} \approx \underbrace{(-z_c\phi(z_c))}_{\text{A negative constant}} \cdot (\mu_T)^2$$

This explicitly shows that the dominant component of the coverage error arising from a bias in the test statistic is proportional to the **square of that bias**. This is the mathematical origin of the "Squared Bias Term" in the proof sketch.

Appendix II: Edgeworth Expansions (Example)

- Imagine you have a statistic, like the sample mean. The Central Limit Theorem (CLT) tells us that for a large sample size, the distribution of this statistic is *approximately* a normal distribution.
- An **Edgeworth expansion** is a way to make that approximation much more precise.
  - It takes the normal distribution from the CLT as a starting point and then adds a series of correction terms to it.
  - These correction terms account for the ways the true distribution deviates from a perfect normal curve, primarily due to **skewness** and **kurtosis** in the original data.
- Think of it like a Taylor series for probability distributions.
  - A Taylor series approximates a complex function with a simple polynomial.
  - The more terms you add, the better the approximation.
  - An Edgeworth expansion approximates a complex probability distribution with a simple normal distribution plus polynomial correction terms.
  - The more terms you add, the more accurately you capture the true shape of the distribution.

Formal Definition

Let Z_n be a standardized statistic based on a sample of size n. For example, $Z_n = \sqrt{n}(\bar{X}_n - \mu)/\sigma$. Let $F_n(z)$ be the true cumulative distribution function (CDF) of Z_n, so $F_n(z) = P(Z_n \leq z)$. The Central Limit Theorem states:

$$F_n(z) \approx \Phi(z)$$

where $\Phi(z)$ is the CDF of a standard normal distribution. This approximation has an error that shrinks as n gets larger.

An **Edgeworth expansion** provides an asymptotic expansion for $F_n(z)$ in powers of $n^{-1/2}$. It gives a more detailed approximation:

$$F_n(z) \approx \Phi(z) - \phi(z)\left[\frac{p_1(z)}{\sqrt{n}} + \frac{p_2(z)}{n} + \frac{p_3(z)}{n^{3/2}} + \dots\right]$$

Where:

- $\Phi(z)$ is the standard normal CDF (the CLT part).
- $\phi(z)$ is the standard normal probability density function (PDF).
- $p_1(z), p_2(z), \dots$ are polynomials in z. The crucial part is that the **coefficients of these polynomials depend on the moments (like skewness and kurtosis) of the underlying data distribution**.

The first correction term, governed by $p_1(z)$, accounts for skewness. The second term, governed by $p_2(z)$, accounts for kurtosis and other higher-order features.

Simple Example: The Sample Mean

Let's look at the first-order Edgeworth expansion for the standardized sample mean $Z_n = \sqrt{n}(\bar{X}_n - \mu)/\sigma$.

The first polynomial, $p_1(z)$, is given by:

$$p_1(z) = \frac{\gamma}{6}(z^2 - 1)$$

where $\gamma = E[(X - \mu)^3]/\sigma^3$ is the **skewness** of the original data distribution.

Plugging this in, the first-order Edgeworth expansion is:

$$P(Z_n \leq z) \approx \Phi(z) - \phi(z)\frac{\gamma}{6\sqrt{n}}(z^2 - 1)$$

What does this tell us?

1. **The Role of Skewness:** The first and most important correction to the normal approximation comes from the skewness (γ) of the original data. If the original data is perfectly symmetric (like a t-distribution or the normal distribution itself), then $\gamma = 0$, and this entire first correction term vanishes. In that case, the normal approximation is much more accurate.
2. **The Role of Sample Size:** The correction term is divided by $\sqrt{n}$. This shows that as the sample size n gets very large, the correction term goes to zero, and we are left with the simple CLT result, as expected.
3. **Improving Approximations:** If you have data from a skewed distribution, and a finite sample size, this formula will give you a much more accurate p-value or critical value than relying on the simple normal approximation alone.

In the context of the Calonico, Cattaneo, and Farrell paper, they use Edgeworth expansions on their complex RD t-statistic to get a highly accurate formula for its true distribution. This allows them to precisely characterize the coverage error of confidence intervals and then find the bandwidth that minimizes this error.