Nonparametric Function Estimation
Chapter 3 Stat 550[1]

David W Scott[2]

Rice University

September 12
Fall 2023
Rice University

[1] A course based upon the 2nd edition of *Multivariate Density Estimation; Theory, Practice, and Visualization*, *John Wiley & Sons, 2015*

[2] www.stat.rice.edu/∼scottdw/

Chapter III: Histograms: Theory and Practice

- histogram most intuitive; first thing taught in intro stat course; how taught?
- non-density forms: bin counts or as a stem-and-leaf plot
- distinction between histogram as a density estimator and as a data presentation device?
- **Sturges' Rule** for Choosing the Number of Bins
  - Histogram with k bins, labelled $0, 1, \ldots, k-1$:

$$B(k-1, p = \tfrac{1}{2}) \approx N\left(\frac{k-1}{2}, \frac{k-1}{4}\right) \sim \binom{k-1}{x} 0.5^x (1-0.5)^{k-1-x}$$

  - Sturges took the binomial coefficients as the bin counts:

$$n = \sum_{x=0}^{k-1} \binom{k-1}{x} = 2^{k-1} \quad \Rightarrow \quad k = 1 + \log_2 n$$

  - Check boundary condition: $n = 1$ implies $k = 1$.

Example with 21 Bins

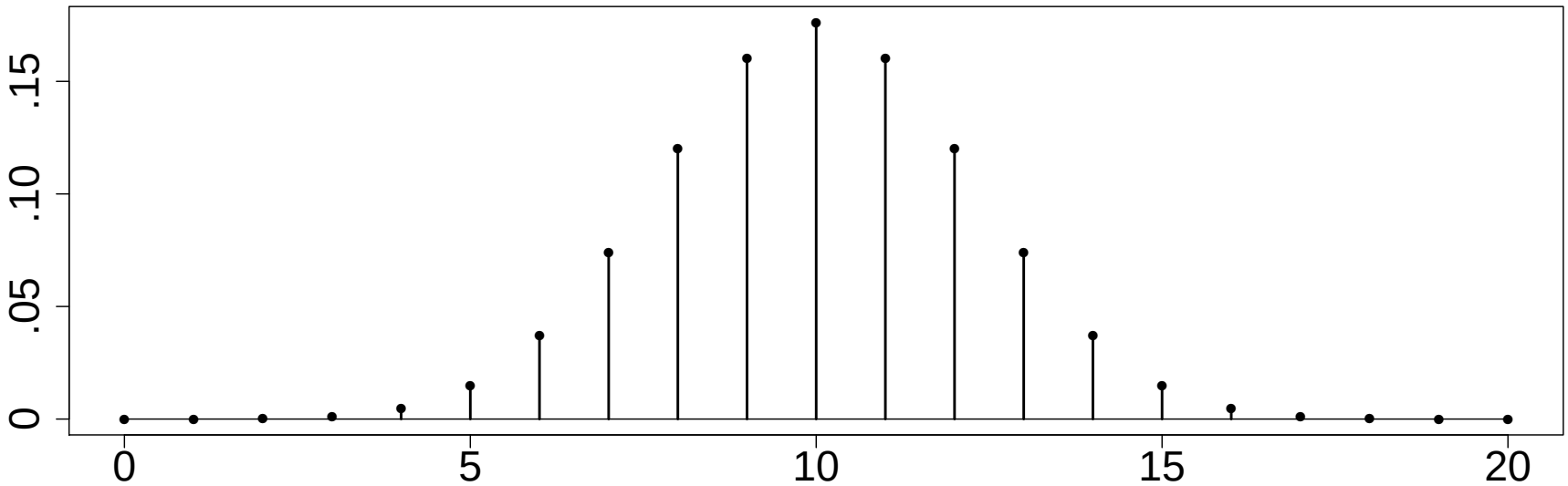


Figure: Binomial pdf with $p = 0.5$ used by Sturges to determine the number of histogram bins.

$$k - 1 = 20 \quad \Longrightarrow \quad n = 2^{21} \approx 2 \text{ million}$$

Question: only 21 bins?

The L_2 Theory of Univariate Histograms

Pointwise Mean Squared Error and Consistency

- construct an equally spaced mesh with 0 as one mesh point and let h denote the bin width

$$\{t_k = kh, -\infty < k < \infty\}$$

- define the kth bin, B_k as the interval $[t_k, t_{k+1})$

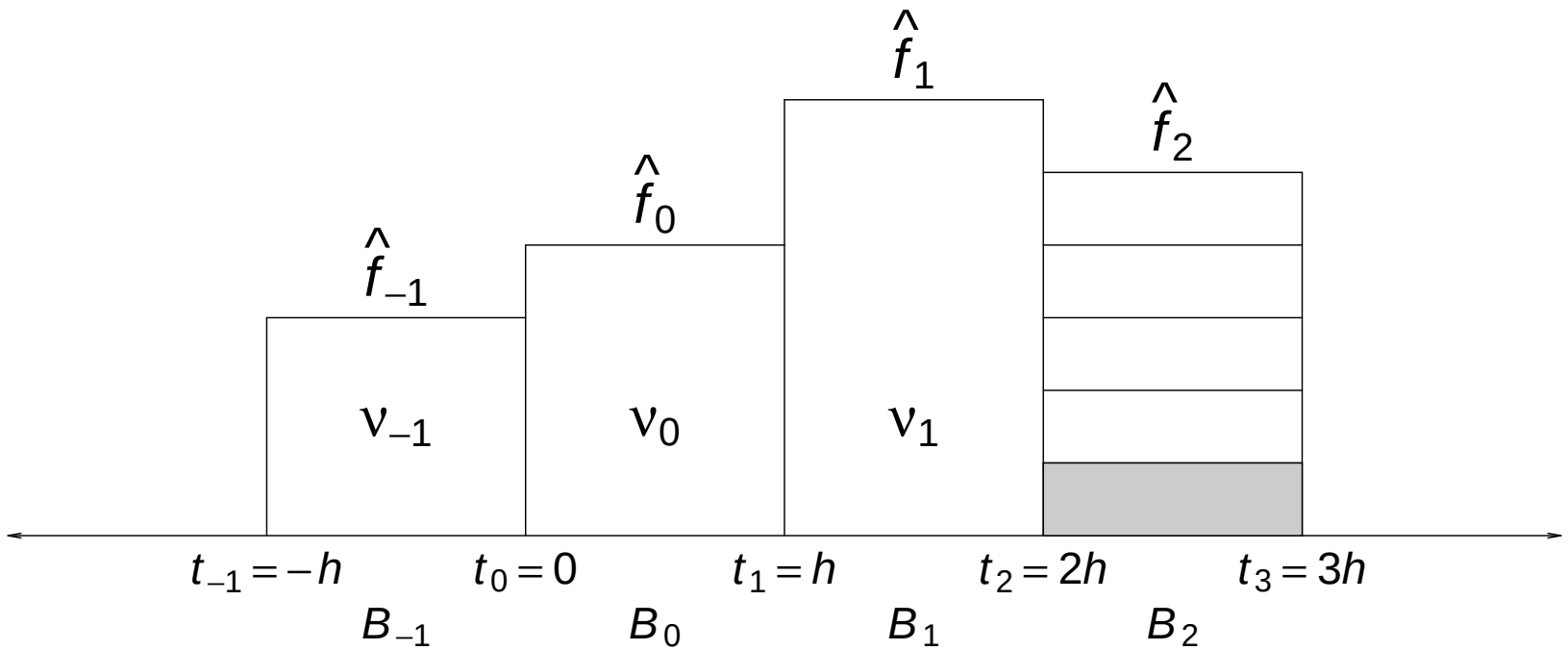


Figure: Notation for construction of an equally-spaced histogram.

- Define the density form of the histogram as

$$\hat{f}(x) = \frac{\nu_k}{nh} \quad \text{for } x \in B_k$$

- ν_k is the bin count; clearly, $\sum_k \nu_k = n$
- can you verify that

$$\int_{x=-\infty}^{\infty} \hat{f}(x)\, dx = 1\,?$$

- break up the integral bin-by-bin:

$$\begin{aligned}\int_{x=-\infty}^{\infty} \hat{f}(x)\, dx &= \sum_{k=-\infty}^{\infty} \int_{x \in B_k} \hat{f}(x)\, dx \\ &= \sum_{k=-\infty}^{\infty} \frac{\nu_k}{nh} \times h = \frac{n}{n} = 1\end{aligned}$$

- The bin count $\nu_k \sim B(n, p_k)$, where $p_k = \int_{B_k} f(t)\,dt$
- For a fixed point $x \in B_k$, what is the mean squared error of $\hat{f}(x)$? Recall $MSE = Var + Bias^2$.

$$Var\,\hat{f}(x) = \frac{Var\,\nu_k}{(nh)^2} = \frac{np_k(1-p_k)}{n^2h^2} = \frac{p_k(1-p_k)}{nh^2}$$

$$Bias\,\hat{f}(x) = E\,\hat{f}(x) - f(x) = \frac{E\,\nu_k}{nh} - f(x) = \frac{p_k}{h} - f(x)$$

- Let's try to be exact as long as possible. By the Mean Value Theorem,

$$p_k = \int_{B_k} f(t)\,dt = h\,f(\xi_k) \qquad \text{for some} \quad \xi_k \in B_k$$

- Also assume $f(x)$ is Lipschitz continuous (weaker than assuming the derivative $f'(x)$ exists):

$$|f(y) - f(x)| < \gamma_k\,|y - x| \quad \text{for all } x, y \in B_k$$

- Then we have the following results for the variance and bias:

$$Var\,\hat{f}(x) = \frac{p_k(1-p_k)}{nh^2} \leq \frac{p_k}{nh^2} = \frac{h\,f(\xi_k)}{nh^2} = \frac{f(\xi_k)}{nh}$$

$$Bias\,\hat{f}(x) = \frac{p_k}{h} - f(x) = \frac{h\,f(\xi_k)}{h} - f(x) = f(\xi_k) - f(x)$$

$$\therefore \quad \left|Bias\,\hat{f}(x)\right| = \left|f(\xi_k) - f(x)\right| \leq \gamma_k \left|\xi_k - x\right| \leq \gamma_k\, h$$

- Combining, we have

$$MSE\,\hat{f}(x) \leq \frac{f(\xi_k)}{nh} + \gamma_k^2\, h^2 \qquad (var + bias^2)$$

- Observations and conclusions:
  - histogram consistent if, as $n \to \infty$, $h \to 0$ and $nh \to \infty$
  - In fact, $h^* = O(n^{-1/3})$ which results in $MSE^* = O(n^{-2/3})$.
  - So the convergence is **slower** than $O(n^{-1})$ for parameters.
  - Very different than the logarithmic rate suggested by Sturges' rule.

Understanding the Noise in a Histogram

- The noise inherent in the histogram varies directly with the square root of its height, since $\text{var}\{\hat{f}(x)\} \approx f(x)/(nh)$
- There is a variance-stabilizing transformation for Poisson data, namely, the square root function.
- A little work gives us the result

$$\sqrt{\text{var}\ \sqrt{\hat{f}(x)}} \approx \frac{1}{2\sqrt{f(x)}}\sqrt{\frac{f(x)}{nh}} = \frac{1}{2\sqrt{nh}}$$

  which does not depend on the unknown density function value $f(x)$
- Tukey advocated plotting the histogram on a square root scale, which he called the **rootgram**.
- True in $\Re^d$ as well. But the contours of $\hat{f}(x)$ and $\sqrt{\hat{f}(x)}$ are identical! Conclusion?

Global L_2 Histogram Error

- Want to use the decomposition $IMSE = IV + ISB$ for this purpose (integrated variance and the integrated squared bias)

$$
\begin{aligned}
IV &= \int_{-\infty}^{\infty} Var\, \hat{f}(x)\, dx \\
&= \sum_{k=-\infty}^{\infty} \int_{B_k} Var\, \hat{f}(x)\, dx \\
&= \sum_{k=-\infty}^{\infty} \frac{p_k(1-p_k)}{nh^2} \times h \\
&= \frac{1}{nh} \sum_{k=-\infty}^{\infty} \left[p_k - p_k^2 \right] \\
&= \frac{1}{nh} - \frac{1}{n} \int_{-\infty}^{\infty} f(t)^2 dt + \cdots
\end{aligned}
$$

Bias Calculations

Here is a picture of how the bias behaves in the limit:

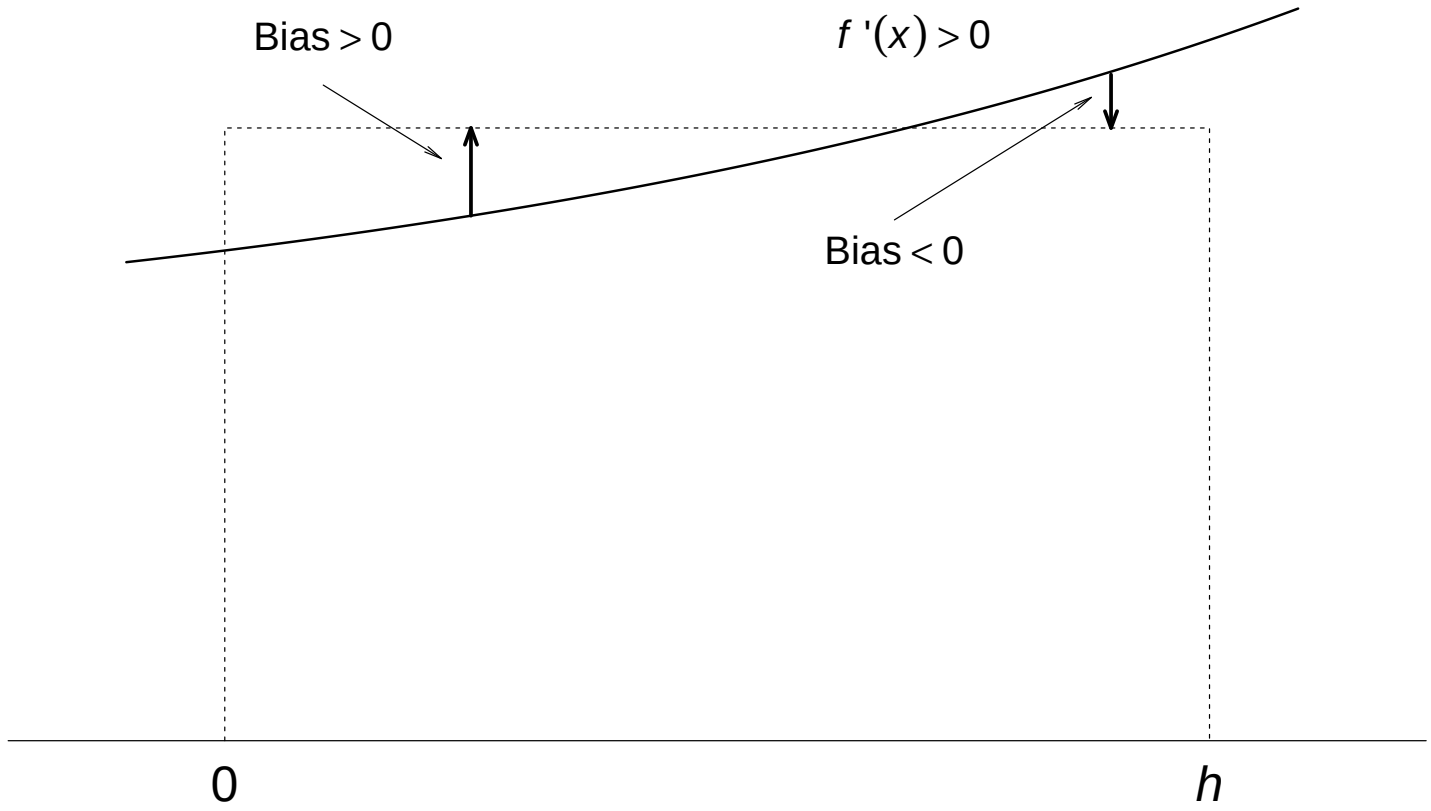


Figure: Bias of histogram estimator in a typical bin.

Taylor Series approximation for the bin probability p_0

The bin probability, p_0, may be approximated by

$$p_0 = \int_0^h f(t)\,dt = \int_0^h \left[f(x) + (t-x)f'(x) + \frac{1}{2}(t-x)^2 f''(x) + \cdots \right] dt$$
$$= hf(x) + h\left(\frac{h}{2} - x\right) f'(x) + O(h^3)$$

It then follows that

$$\text{Bias } \hat{f}(x) = \frac{p_0}{h} - f(x) = \left(\frac{h}{2} - x\right) f'(x) + O(h^2)$$

Note that the bias is greater where the slope of $f(x)$ is greater (see figure again). The integrated squared bias over the bin B_0 is given by: (here we use the Generalized MVT for some $\eta_0 \in B_0$)

$$\int_{B_0} \left(\frac{h}{2} - x\right)^2 f'(x)^2\,dx = f'(\eta_0)^2 \int_0^h \left(\frac{h}{2} - x\right)^2 dx = \frac{h^3}{12} f'(\eta_0)^2$$

Accumulated the Integrated Squared Bias bin-by-bin

- The previous result for bin B_0 easily generalizes to bin B_k:

$$
\begin{aligned}
ISB &= \sum_{k=-\infty}^{\infty} ISB_k \\
&= \sum_{k=-\infty}^{\infty} \frac{h^3}{12} f'(\eta_k)^2 \qquad \text{for some } \eta_k \in B_k \\
&= \frac{h^2}{12} \sum_{k=-\infty}^{\infty} f'(\eta_k)^2 \times h \qquad \text{(a Riemannian sum)} \\
&= \frac{h^2}{12} \int_{-\infty}^{\infty} f'(x)^2 \, dx + o(h^2) \\
&= \frac{1}{12} h^2 R(f') + o(h^2) \,,
\end{aligned}
$$

- Introducing a new notation for the "roughness" of a function:

$$
R(\phi) = \int_{-\infty}^{\infty} \phi(t)^2 \, dt
$$

$AMISE(h)$ Result

- Combining the asymptotic approximations to the IV and ISB, we have

$$AMISE(h) = \frac{1}{nh} + \frac{1}{12}h^2 R(f') \qquad \text{hence,}$$
$$h^* = [6/R(f')]^{1/3} n^{-1/3}$$
$$AMISE^* = (3/4)^{2/3} R(f')^{1/3} n^{-2/3}$$

- A practical data-based rule (Normal Reference Rule):

$$f \sim N(\mu, \sigma^2) \quad \Longrightarrow \quad R(f') = \frac{1}{4\sqrt{\pi}\sigma^3}$$

- Plugging into h^* gives us **Scott's Rule**:

$$h^* = \left(24\sqrt{\pi}\sigma^3\right)^{1/3} n^{-1/3}$$
$$\approx 3.5\,\hat{\sigma}\, n^{-1/3}$$

Literature Review

- These results first appeared in
  - Scott, D.W. (1979) "On Optimal and Data-Based Histograms," *Biometrika*, 66:605–610
  - Freedman, D. and Diaconis, P. (1981). "On the Histogram as a Density Estimator: L_2 Theory," *Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete*, 57:453–476
- Freedman and Diaconis proposed an alternative data-based rule:

$$h_{FD} = 2\,(IQR)\,n^{-1/3}$$

  where IQR is the interquartile range, a more robust measure of scale
- For normal data, the Freedman-Diaconis rule is about 23% narrower than Scott's rule
- Soon we will look at other estimates of the unknown $R(f')$

Comparison of Bandwidth Rules

Table: Comparison of Number of Bins from Three Normal Reference Rules

n	Sturges' Rule	Scott's Rule	F-D Rule
50	5.6	6.3	8.5
100	7.6	8.0	10.8
500	10.0	13.6	18.3
1,000	11.0	17.2	23.2
5,000	13.3	29.4	39.6
10,000	14.3	37.0	49.9
100,000	17.6	79.8	107.6

Clearly, Sturges' Rule is too conservative, oversmoothing the histogram and losing information with bins that are too wide.

Asymptotic MISE Curves for *Beta*(5, 5) Density

The *Beta*(5, 5) is close to normal, but with finite support.

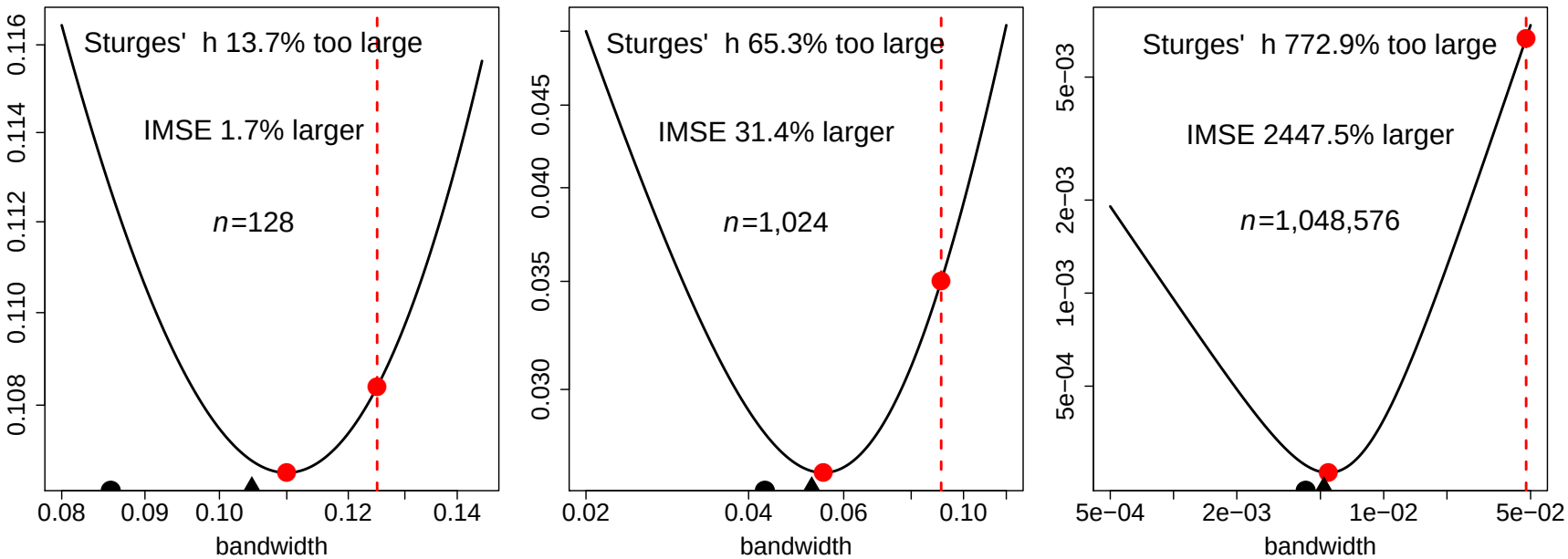


Figure: AMISE versus bandwidth for a $B(5, 5)$ density. The best and Sturges' bandwidths are indicated by points on the curves. The Freedman-Diaconis and Scott reference bandwidths are shown as semicircular and triangular points, respectively, along the x-axis.

Equivalent Sample Sizes for Normal Data

- Here we compare the sample size required to achieve a certain *AMISE* by the histogram and several parametric estimators.
- From our earlier results, it follows that if $f = N(0, 1)$, then the optimal AMISE of the histogram is

$$\text{AMISE}^* = [9/(64\sqrt{\pi})]^{1/3} n^{-2/3} \approx 0.4297 n^{-2/3}.$$

- Then we find

Table: Equivalent Sample Sizes for Several Normal Density Estimators

	Estimator			
AMISE	$N(\bar{x}, s^2)$	$N(\bar{x}, 1)$	$N(0, s^2)$	Histogram
0.002468	100	57	43	2,297
0.000247	1,000	571	429	72,634

- Steep penalty for such generality?

Sensitivity of MISE to Bin Width

- How close to h^* can be get, and how close should we be?
- Suppose we use a bin width, h, which is multiple of h^*. How is the error *AMISE* affected?

$$\frac{\text{AMISE}(ch^*)}{\text{AMISE}(h^*)} = \frac{2+c^3}{3c} \tag{1}$$

- This author's rule-of-thumb is to be within 10-15% of h^*

Table: Sensitivity of AMISE to Error in Bin Width Choice $h = ch^*$

$(d=1, r=0)$	$p=1$	$p=2$	$p=4$
c	$(c^3+2)/(3c)$	$(c^5+4)/(5c)$	$(c^9+8)/(9c)$
1/2	1.42	1.61	1.78
3/4	1.08	1.13	1.20
1	1	1	1
4/3	1.09	1.23	1.78
2	1.67	3.60	28.89

Simulation evaluation of ch^* for $N(0,1)$ data ($n = 1000$)

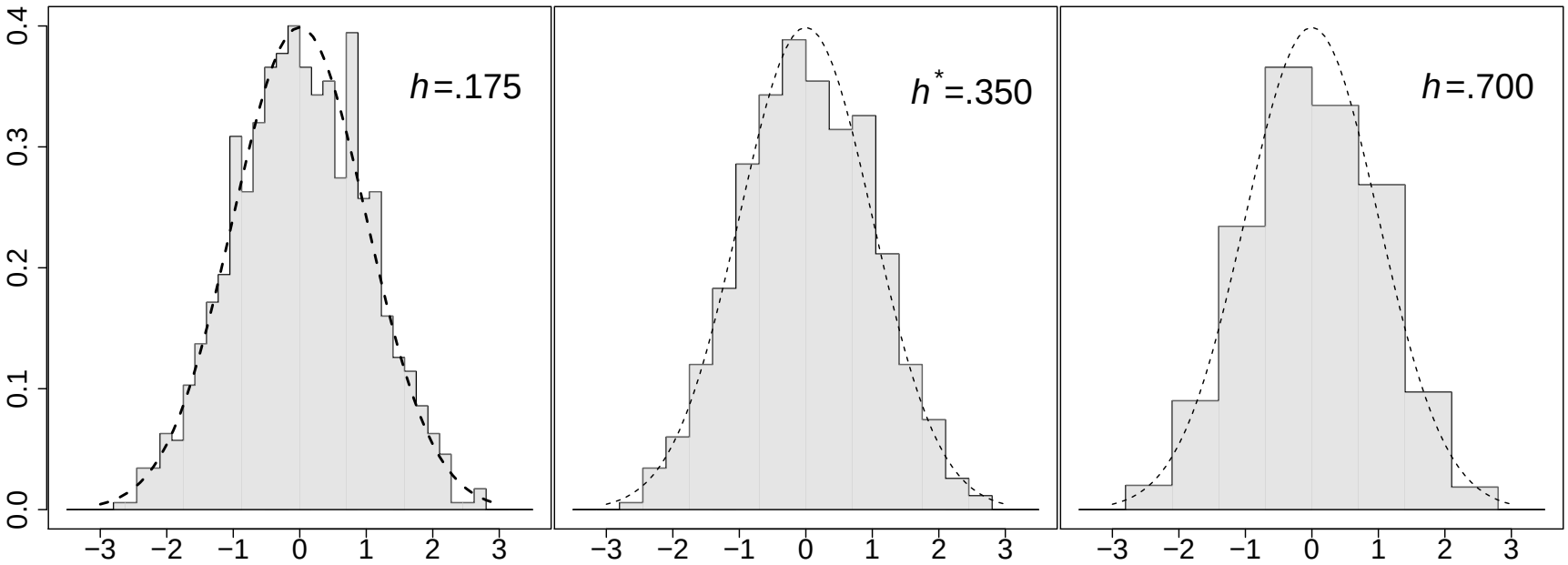


Figure: Three histograms of 1,000 normal observations with bin widths $h = (\frac{1}{2}h^*, h^*, 2h^*)$.

Simulation evaluation of ch^* for $N(0,1)$ data ($n = 10^6$)

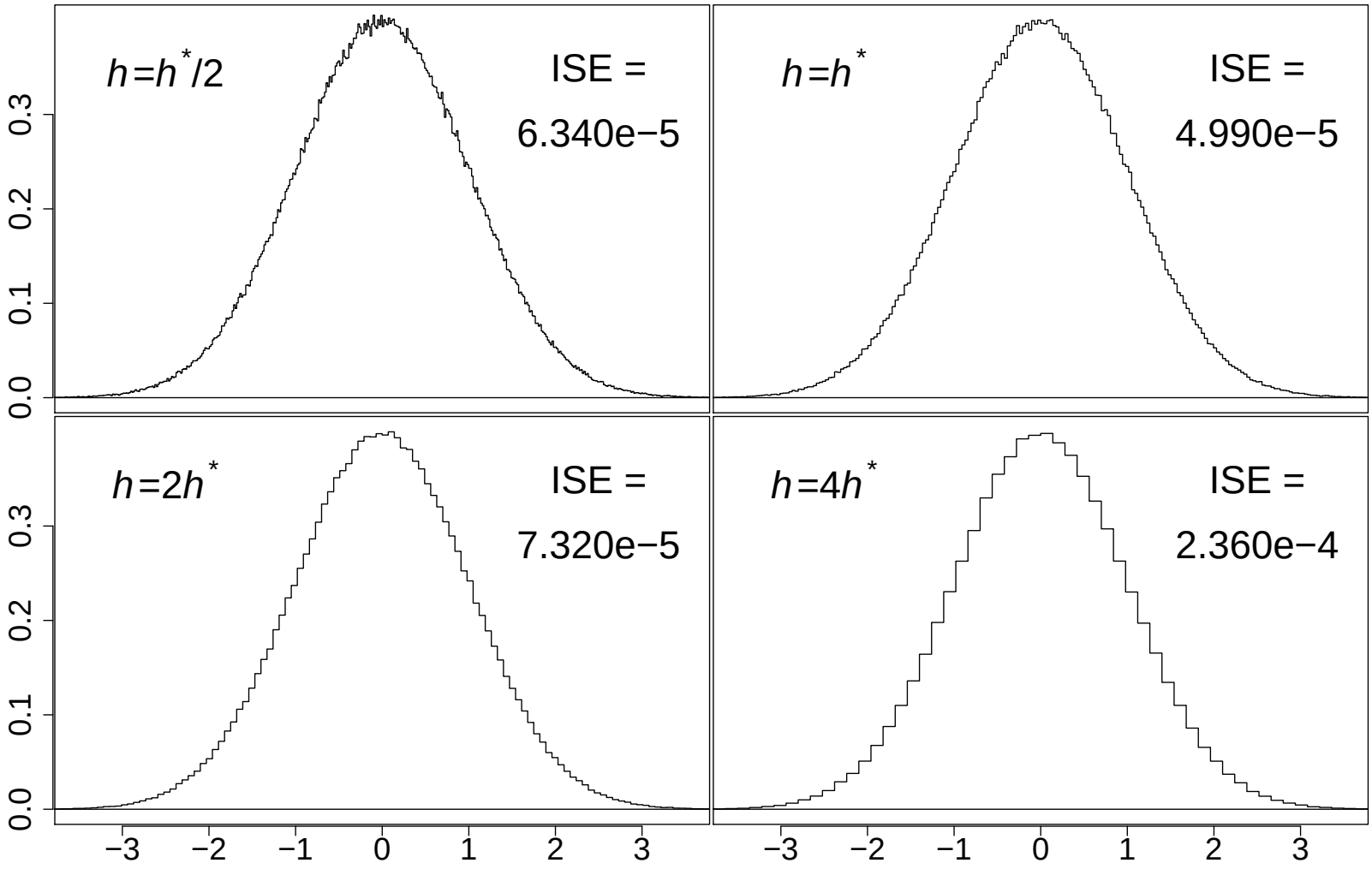


Figure: Exact ISE of four histograms of a million normal points with bin widths $h = (h^*/2, h^*, 2h^*, 4h^*)$, where $h^* = 0.035$.

Exact *MISE* versus Asymptotic *MISE*

- Generalize the definition of the histogram mesh to allow the bin width of B_k to be h_k. The it is straightforward to show we have the following (exact) expressions:

$$\text{IV} = \frac{1}{n}\sum_k \frac{p_k(1-p_k)}{h_k} \quad \text{and}$$

$$\text{ISB} = R(f) - \sum_k \frac{p_k^2}{h_k}$$

- For the special case where $h_k = h$, the (exact)

$$MISE(h, t_0, n) = \frac{1}{nh} - \frac{n+1}{nh}\sum_k p_k^2 + R(f)$$

AMISE versus *MISE* for $N(0, 1)$ Data

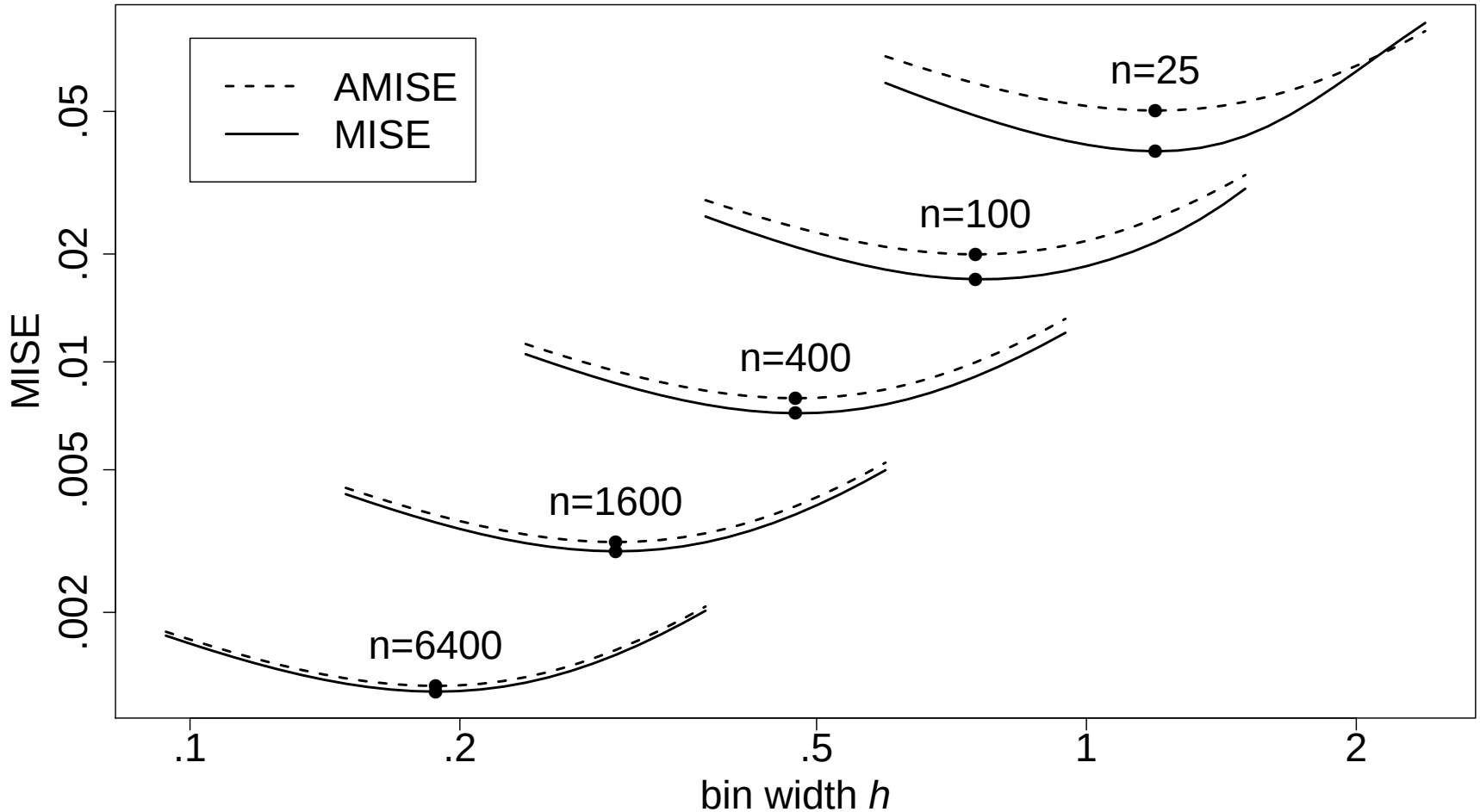


Figure: AMISE and exact MISE for the $N(0, 1)$ density.

Decomposition of *MISE* into *IV* and *ISB* for $N(0,1)$ Data

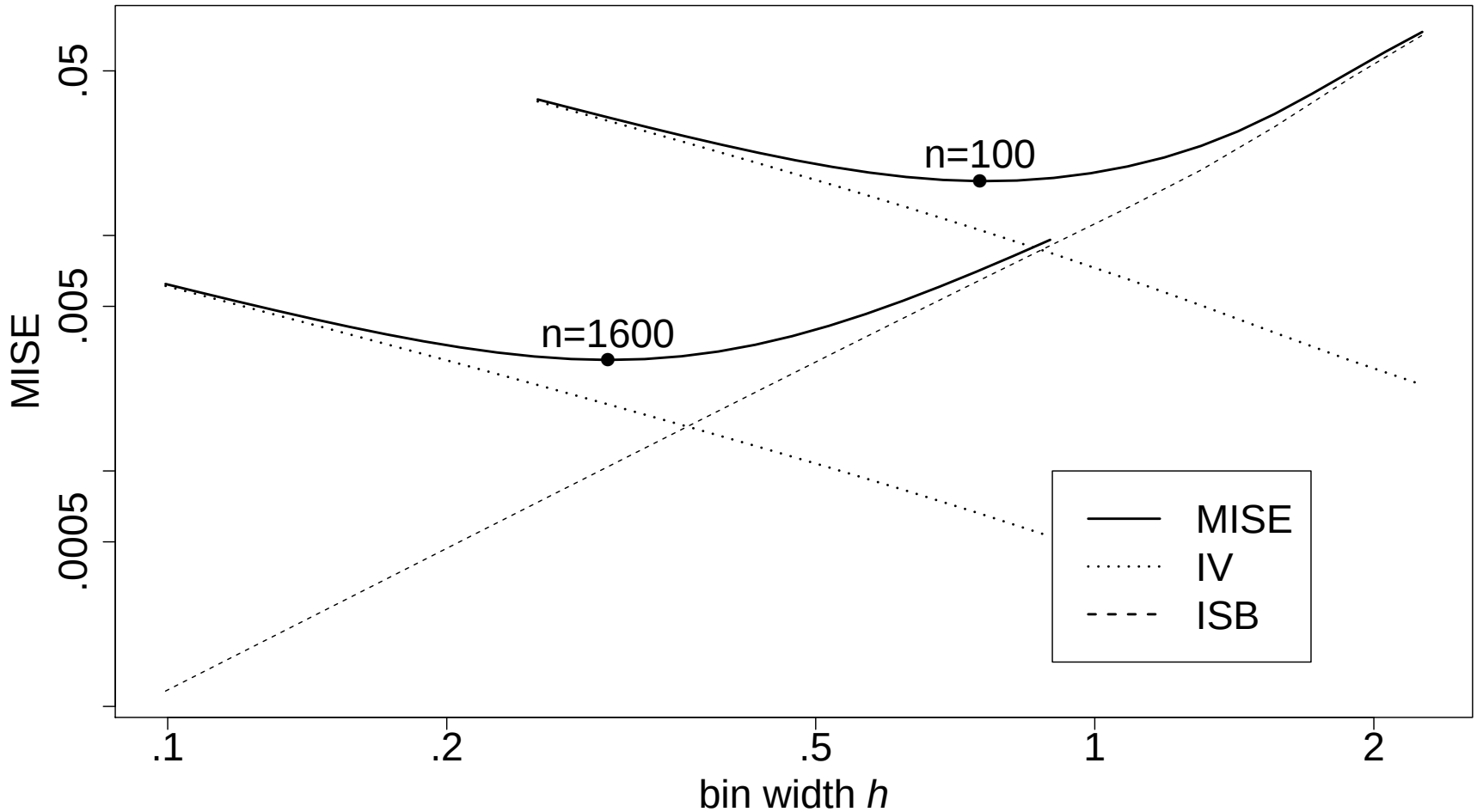


Figure: Integrated squared-bias/variance decomposition of MISE for the $N(0,1)$ density.

How Good Is the Sensitivity Curve for $h = ch^*$?

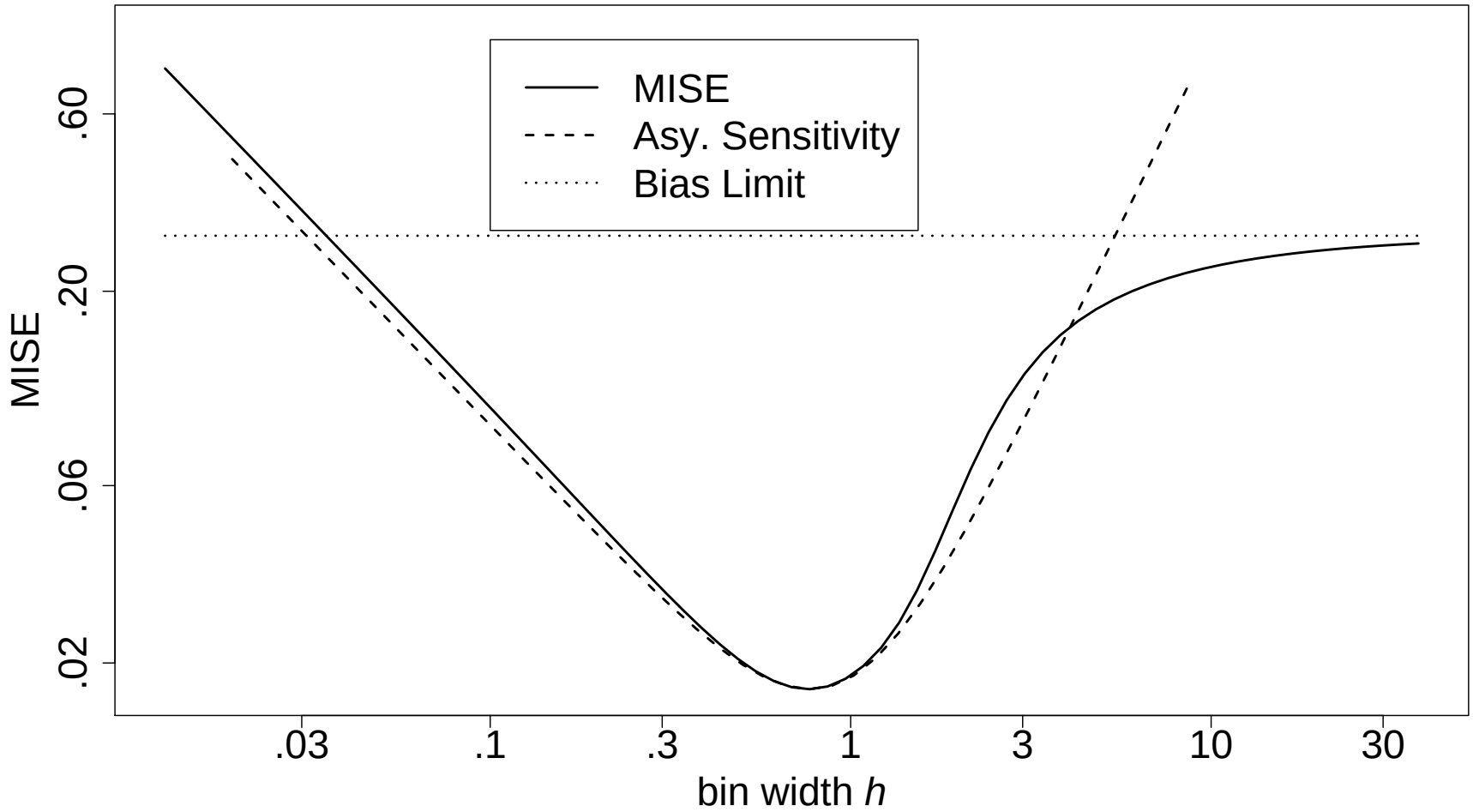


Figure: Complete MISE curve for the $N(0, 1)$ density when $n = 100$. The asymptotic sensitivity relationship holds over a wide range of bin widths.

Similar Results for the Lognormal Density?

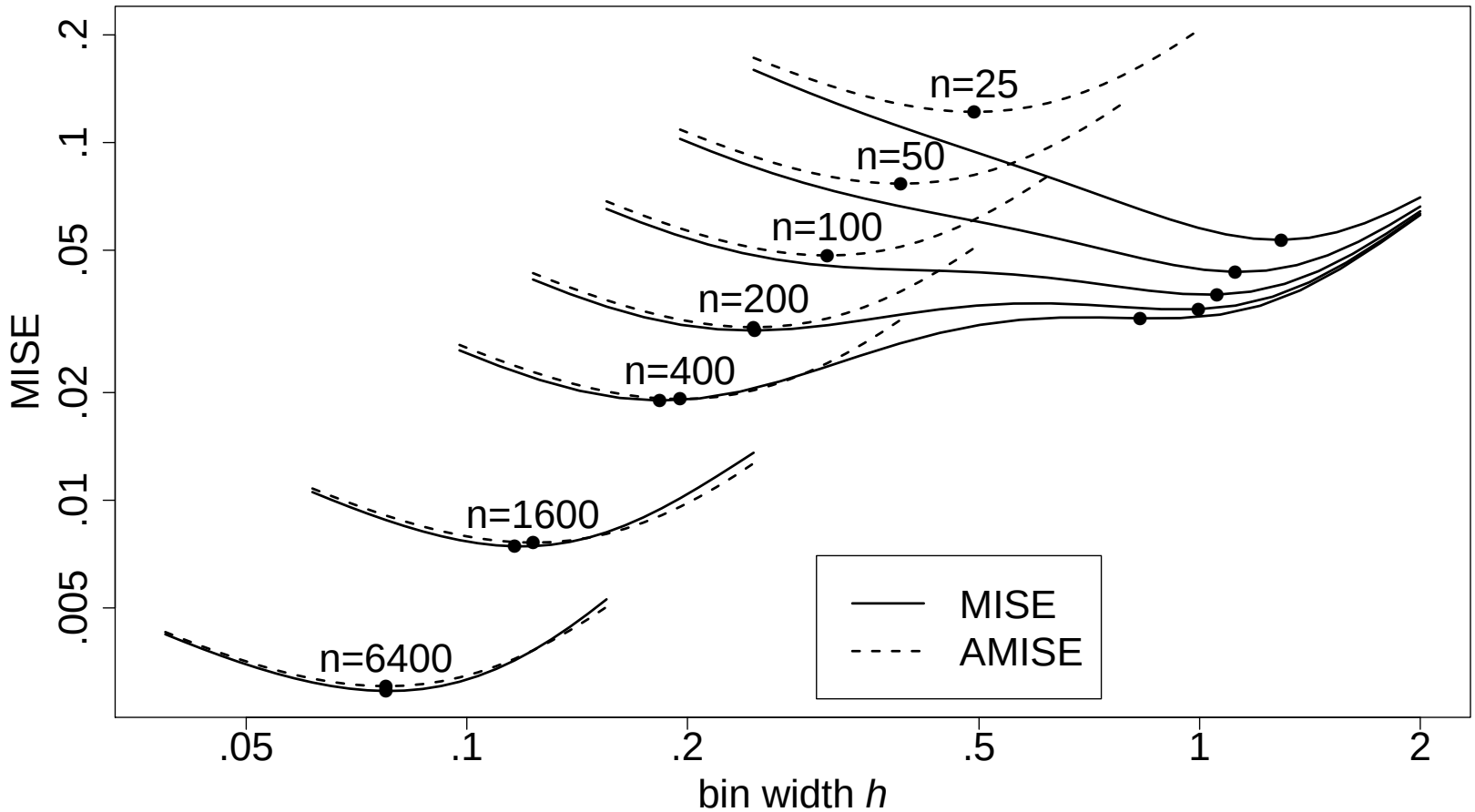


Figure: AMISE and exact MISE for the lognormal density.

Thoughts on the Lognormal Density

- The roughness of a density is summarized in $R(f')$
- The lognormal density is infinitely differentiable, but is very rough statistically!
- In fact, 90% of the roughness, $R(f') = 3e^{9/4}/(8\sqrt{\pi})$, comes from the small interval [0, 0.27], even though the mode is located at $x = 0.368$ and the $99th$ percentile is $x = 10.2$.
- This suggests we should try to use an adaptive mesh; that is, use h_k rather than a fixed mesh where all $h_k = h$. But hard to do in practice.
- For small samples, the histogram cannot track the rapid rise near $x = 0$
- Useful to think of sample sizes as being one of
  - inadequate
  - transitional
  - sufficient

Influence of Bin Edge Location on MISE

- Our Taylor Series analysis of the *AMISE* revealed that the choice of the bin origin, t_0, is asymptotically negligible
- A discontinuity in the density at a boundary
  - causes no problem if place t_0 at the (known) boundary
  - can reduce the convergence rate to $O(n^{-1/2})$ if unknown

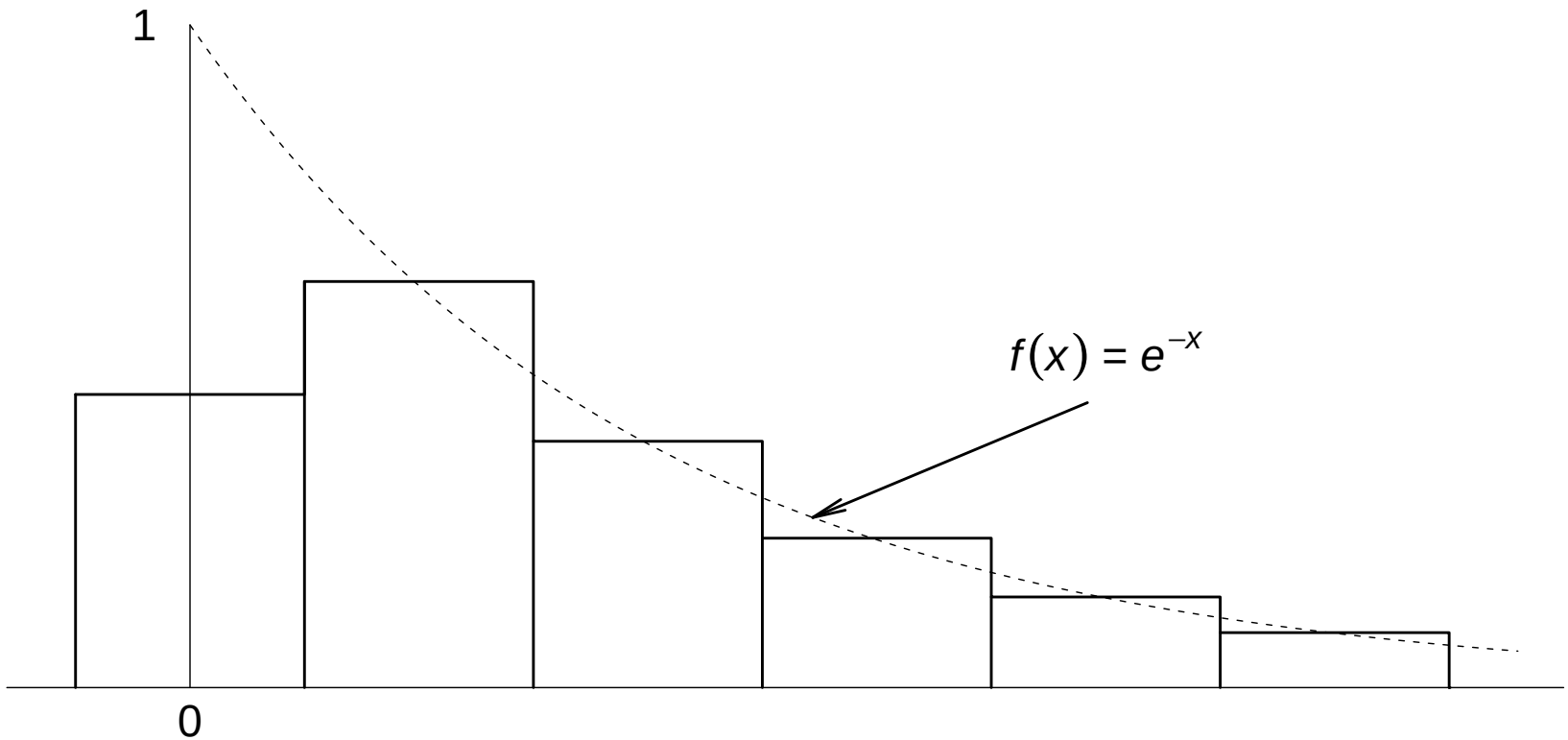


Table: Potential Impact on AMISE of Lack of Knowledge of Boundary Discontinuities for Negative Exponential Data

	$X > 0$ Known		$X > 0$ Unknown		
n	h^*	AMISE*	h^*	AMISE*	Error Ratio
10	1.063	0.14116	0.3162	0.31623	2.24
100	0.493	0.03041	0.1	0.1	3.29
1,000	0.229	0.00655	0.0316	0.03162	4.83
10,000	0.106	0.00141	0.01	0.01	7.09
100,000	0.049	0.00030	0.0032	0.00316	10.54

Optimally Adaptive Histogram Meshes

Bounds on MISE Improvement for Adaptive Histograms

- Try to find h^* for each point x by considering the asymptotically adaptive pointwise histogram MSE (AAMSE):

$$AAMSE(x) \approx \frac{f(x)}{nh} + \frac{1}{12} h^2 f'(x)^2$$

- Therefore, we obtain the following results:

$$h^*(x) = \left[\frac{6f(x)}{nf'(x)^2} \right]^{1/3} \quad \Rightarrow \quad \text{AAMSE}^*(x) = \left[\frac{3f(x)f'(x)}{4n} \right]^{2/3}.$$

- This leads to the best adaptive *AMISE*

$$AAMISE^* = (3/4)^{2/3} \left(\int_{-\infty}^{\infty} [f'(x)f(x)]^{2/3} dx \right) n^{-2/3}$$

How Much Better is *AAMISE** versus *AMISE**?

Table: Reduced AMISE Using an Optimally Adaptive Histogram Mesh

Density	$\int (f^2 f'^2)^{1/3} \div [\int f'^2]^{1/3}$
$N(0,1)$	$0.4648/0.5205 = 89.3\%$
$3/4\ (1-x^2)_+$	$0.8292/1.1447 = 72.4\%$
$15/16\ (1-x^2)_+^2$	$2.1105/2.8231 = 74.8\%$
$315/256\ (1-x^2)_+^4$	$1.4197/1.6393 = 86.6\%$
Cauchy	$0.3612/0.4303 = 84.0\%$
Lognormal	$0.6948/1.2615 = 55.1\%$

Notes: Given how hard it is likely to be in practice to actually construct an adaptive mesh, it is reassuring to see that the (potential) improvement is marginal for $N(\mu, \sigma^2)$ data. Jensen's inequality assures the ratio in the table is always less than 1.

Some Optimal Meshes

These were found by numerical minization of the exact *MISE* formulae.

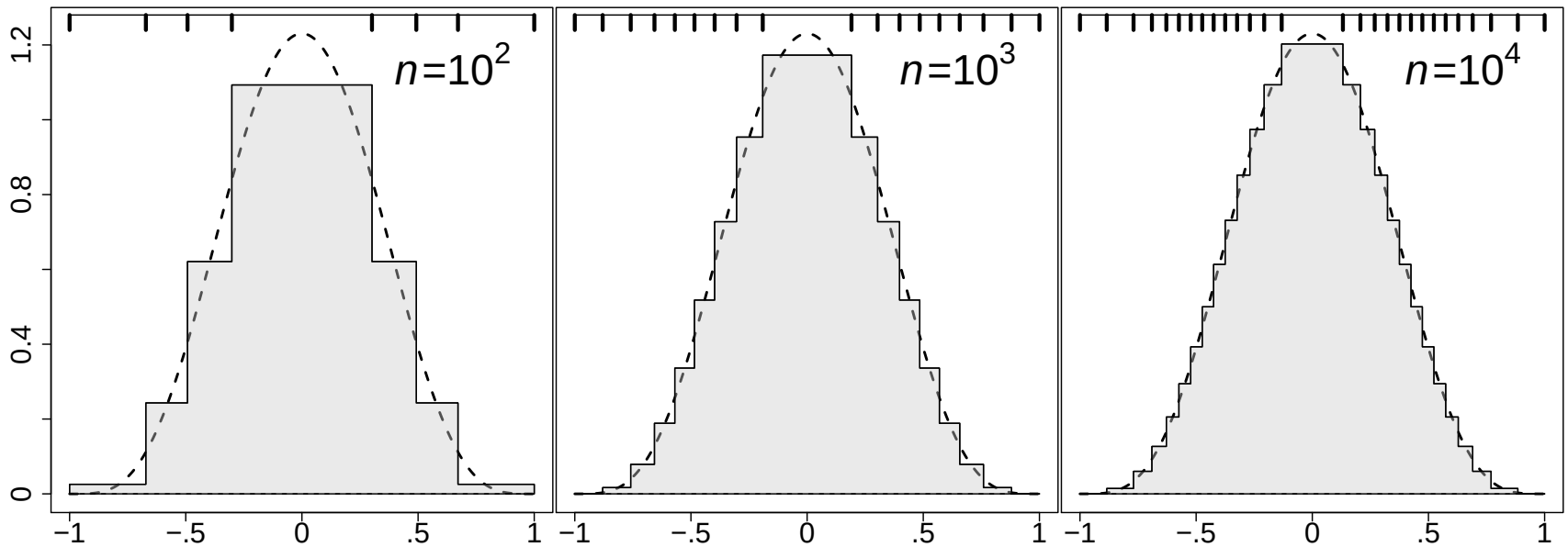


Figure: Representation of optimal adaptive meshes for the transformed Beta(5,5) density, which equals $315/256\ (1-x^2)_+^4$. The optimal adaptive mesh is also indicated by tick marks above each graph.

An Intuitively Appealing Adaptive Mesh: Percentile Meshes or Adaptive Histograms with Equal Bin Counts

This can be modeled by picking the mesh to satisfy

$$t_k = F_X^{-1}\left(\frac{k}{m}\right), \quad k = 0, \ldots, m\,.$$

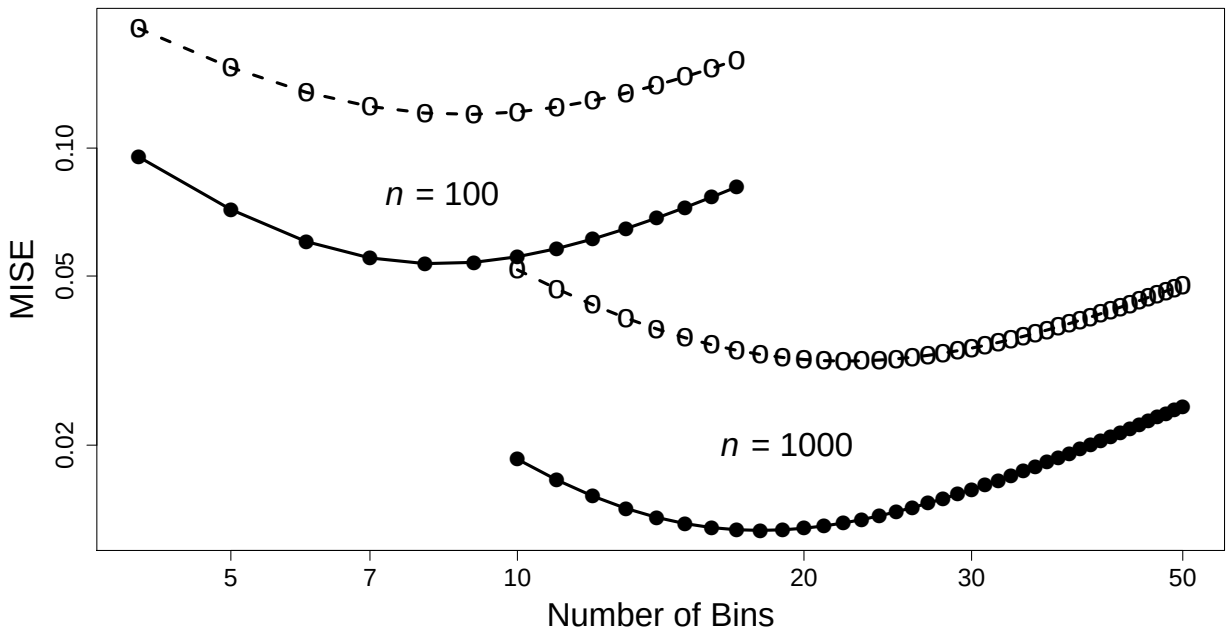


Figure: Exact MISE for fixed (●●●●●) and percentile (ooooo) meshes with k bins over $(-1, 1)$ for the transformed Beta(5,5) density with $n = 100$ and 1,000.

Using Adaptive Meshes vs. Transformation

- Since skewed data have greater roughness (and hence approximation error), it is easy to consider transformations of the raw data before constructing the histogram
- For example, Economists usually transform household income, x, to $\log(1+x)$. This is a nice choice, as $x = 0$ is 0 on both scales.
- Also used to justify percentage raises, rather than absolute raises?
- Tukey's transformation ladder:

$$\ldots, x^{-2}, x^{-1}, x^{-1/2}, x^{-1/4}, \log(x), x^{1/4}, x^{1/2}, x, x^2, x^3, \ldots$$

- Box-Cox family is similar, but continuous in λ for $x > 0$:

$$x^{(\lambda)} = \begin{cases} \frac{x^\lambda - 1}{\lambda} & \lambda \neq 0 \\ \log x & \lambda = 0\,. \end{cases}$$

More Practical Data-Based Bin Width Rules: Oversmoothed Bin Widths

- Since the density $f(x)$ can be very rough, there is no **lower bound** on h^*
- However, it turns out there are useful **upper bounds** on h^*
- Recall that $h^* = \left(\frac{6}{R(f')}\right)^{1/3} n^{-1/3}$
- George Terrell proposed the following variational problem:

$$\min_f \int_{-\infty}^{\infty} f'(x)^2 dx \quad \text{s/t} \quad \text{support of } f = [-0.5, 0.5]$$

- The solution is

$$f_1(x) = \frac{3}{2}(1 - 4x^2) I_{[-.5,.5]}(x) = \frac{3}{2}(1 - 4x^2)_+$$

- $f_1(x)$ is the **smoothest density** with fixed support ("oversmoothed" in George's jargon)

Oversmoothed Bin Width Rule

- Note that $R(f_1') = 12$.
- For a (known) support interval (a, b), we have the inequality

$$R(f') \geq 12/(b-a)^3$$

- Therefore, we have the useful result that

$$h^* = \left(\frac{6}{nR(f')}\right)^{1/3} \leq \left(\frac{6(b-a)^3}{n \cdot 12}\right)^{1/3} = \frac{b-a}{\sqrt[3]{2n}} \equiv h_{\mathrm{OS}}$$

- Rearranging, gives us a **lower bound** on the number of bins:

$$\boxed{\text{number of bins} = \frac{b-a}{h^*} \geq \frac{b-a}{h_{\mathrm{OS}}} = \sqrt[3]{2n}}$$

- Compare to Sturges' Rule: number of bins $= 1 + \log_2(n)$

Examples: Buffalo snowfall ($n = 63$, $\sqrt[3]{126} = 5.01$) and LRL ($n = 25{,}752$, $\sqrt[3]{51504} = 37.2$) data

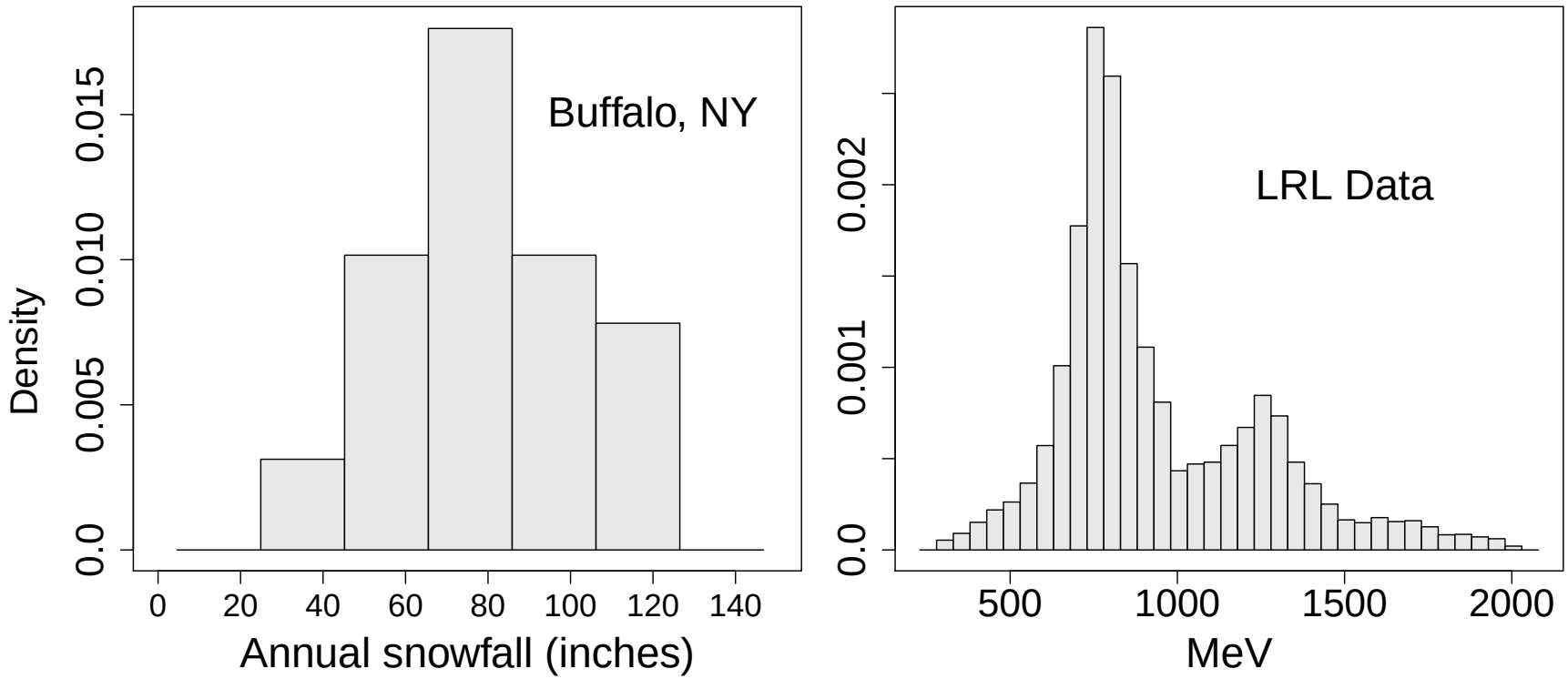


Figure: Oversmoothed histograms of the Buffalo snowfall and LRL data.

Note that the true "optimal" histogram must be rougher. Why?

Another Oversmoothed Rule Based Upon Variance

The variational problem

$$\min_f \int_{-\infty}^{\infty} f'(x)^2 dx \quad \text{s/t} \quad \text{variance of } f = \sigma^2$$

has the solution

$$f_2(x) = \frac{15}{16\sqrt{7}\sigma}\left(1 - \frac{x^2}{7\sigma^2}\right)^2 I_{[-\sqrt{7}\sigma,\sqrt{7}\sigma]}(x).$$

for which $R(f_2') = 15\sqrt{7}/(343\sigma^3)$.
Hence, we have the result

$$\boxed{h^* \le \left(\frac{6}{nR(f_2')}\right)^{1/3} = \left(\frac{686\sigma^3}{5\sqrt{7}n}\right)^{1/3} \approx 3.729\,\sigma\, n^{-1/3} \equiv h_{\mathrm{OS}}\,.}$$

Note: Recall the constant is 3.5 for $N(\mu, \sigma^2)$ data.

Yet Another Oversmoothed Rule Based Upon IQR

- Consider the variation problem where the constraint is the robust scale measure provided by the interquartile range (IQR).
- This leads to the oversmoothed rule

$$\boxed{h^* \leq 2.603(\text{IQR})n^{-1/3} \equiv h_{\text{OS}}}$$

- Recall that the Freedman-Diaconis rule used the constant 2.
- In practice, very good estimates of σ and IQR are available, as well as the interval of support (a, b).
- Can compute all three oversmoothed rules and use the smallest.

Examples

- For $N(\mu, \sigma^2)$ data,

$$h^* = 3.5\,\sigma\,n^{-1/3} < 3.729\,\sigma\,n^{-1/3} = h_{\mathrm{OS}} \quad (= \text{upper bound})$$

  which is only off by 6.4%.
- However, for lognormal data, $\sigma^2 = e(e-1)$; hence,

$$h^* = 1.44 n^{-1/3} < 3.729 \times 2.161 \times n^{-1/3} = 8.059 n^{-1/3}$$

  which is only 18% of h_{OS}. Not a "tight" approximation.
- For standard Cauchy data, $R(f') = 1/(4\pi)$ and the IQR covers $(-1, 1)$; hence,

$$h^* = 4.225 n^{-1/3} < 2.603 \times 2 \times n^{-1/3} = 5.206 n^{-1/3}$$

  which is useful formula for a density whose variance is undefined.

Bin Width Selection: Biased & Unbiased Cross-Validation

- the goal now is to produce bin widths $\hat{h}_{CV}$ that are close to h^* **for finite samples**
- we know that $h^* \in (0, h_{OS})$, which is better than $(0, \infty)$
- biased cross-validation tries to estimate $R(f')$ with the data
- unbiased cross-validation tries to estimate $ISE(h)$ itself
- References:
  - BCV: Scott and Terrell (1987) JASA
  - UCV: Rudemo (1982) and Bowman (1984)

Biased Cross-Validation (or Plug-In)

- The only unknown quantity in the AMISE is $R(f')$; use the histogram itself to estimate the derivative by finite differences:

$$\hat{f}'(t_k) = \frac{\left[\frac{\nu_{k+1}}{nh} - \frac{\nu_k}{nh}\right]}{h}$$

- This leads us to

$$\hat{R}_1 = \sum\nolimits_k [\hat{f}'(t_k)]^2 \cdot h = \frac{1}{n^2 h^3} \sum\nolimits_k (\nu_{k+1} - \nu_k)^2$$

- Similar Taylor Series arguments lead to the result

$$\mathrm{E}[\hat{R}_1] = R(f') + 2/(nh^3) + O(h)\,.$$

- The term $2/(nh^3)$ does not vanish; indeed, with optimal smoothing, it converges to $R(f')/3$ by our theorem
- So biased upwards by a factor of a third

Biased Cross-Validation (continued)

- An asymptotically unbiased estimate of $R(f')$ is simply

$$\hat{R}_h(f') = \frac{1}{n^2 h^3} \sum\nolimits_k (\nu_{k+1} - \nu_k)^2 - \frac{2}{nh^3}$$

- Plugging this into the $AMISE(h)$ expression

$$AMISE(h) = \frac{1}{nh} + \frac{1}{12} h^2 R(f')$$

gives us

$$\boxed{BCV(h) = \frac{5}{6nh} + \frac{1}{12n^2 h} \sum\nolimits_k (\nu_{k+1} - \nu_k)^2}$$

- Try different histograms giving different bin counts $\{\nu_k\}$ and find minimum, subject to the constraint that $\hat{h}_{BCV} \leq h_{OS}$

Unbiased Cross-Validation

- Expand the *ISE* and consider each integral separately

$$\begin{aligned} ISE(h) &= \int \left[\hat{f}(x) - f(x)\right]^2 dx \\ &= R(\hat{f}) - 2\int \hat{f}(x) f(x)\, dx + R(f) \\ &= R(\hat{f}) - 2E\,\hat{f}(X) + constant \end{aligned}$$

- Using a clever unbiased estimator for the second term gives us

$$UCV(h) = R(\hat{f}) - \frac{2}{n}\sum_{i=1}^{n} \hat{f}_{-i}(x_i)$$

where $\hat{f}_{-i}(x_i)$ is the histogram constructed without the data point x_i and then evaluated in the bin where x_i falls

$$\boxed{UCV(h) = \frac{2}{(n-1)h} - \frac{n+1}{n^2(n-1)h}\sum_{k} \nu_k^2}$$

JAVA Histogram Applet: Scott and Lane

- Long term NSF grant with David Lane, Dept of Psychology
- on-line statistics book
- case studies
- demonstrations
- click here: `http://www.davidmlane.com`

Examples: Vertical Scale Carefully Matched

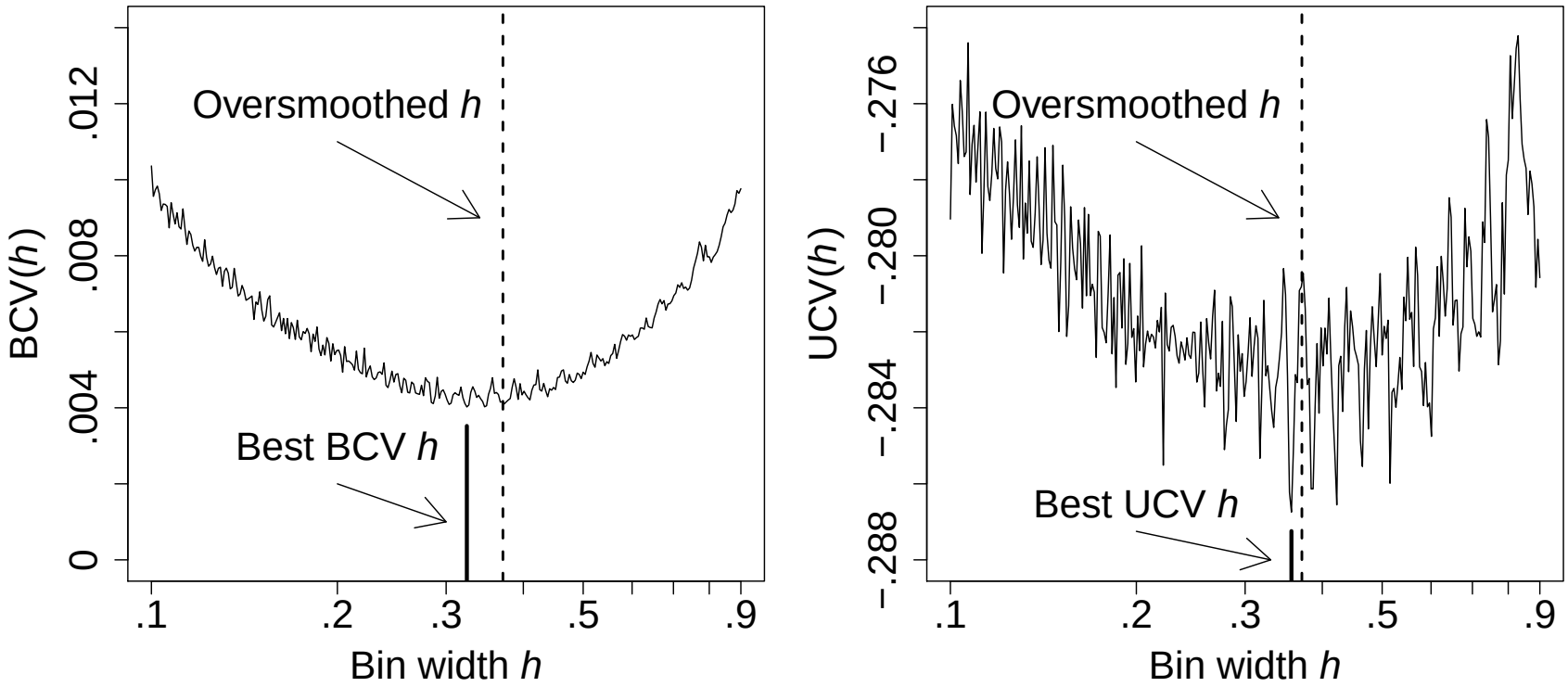


Figure: BCV and UCV functions for a $N(0,1)$ sample of 1,000 points.

Example: German Household Income Data

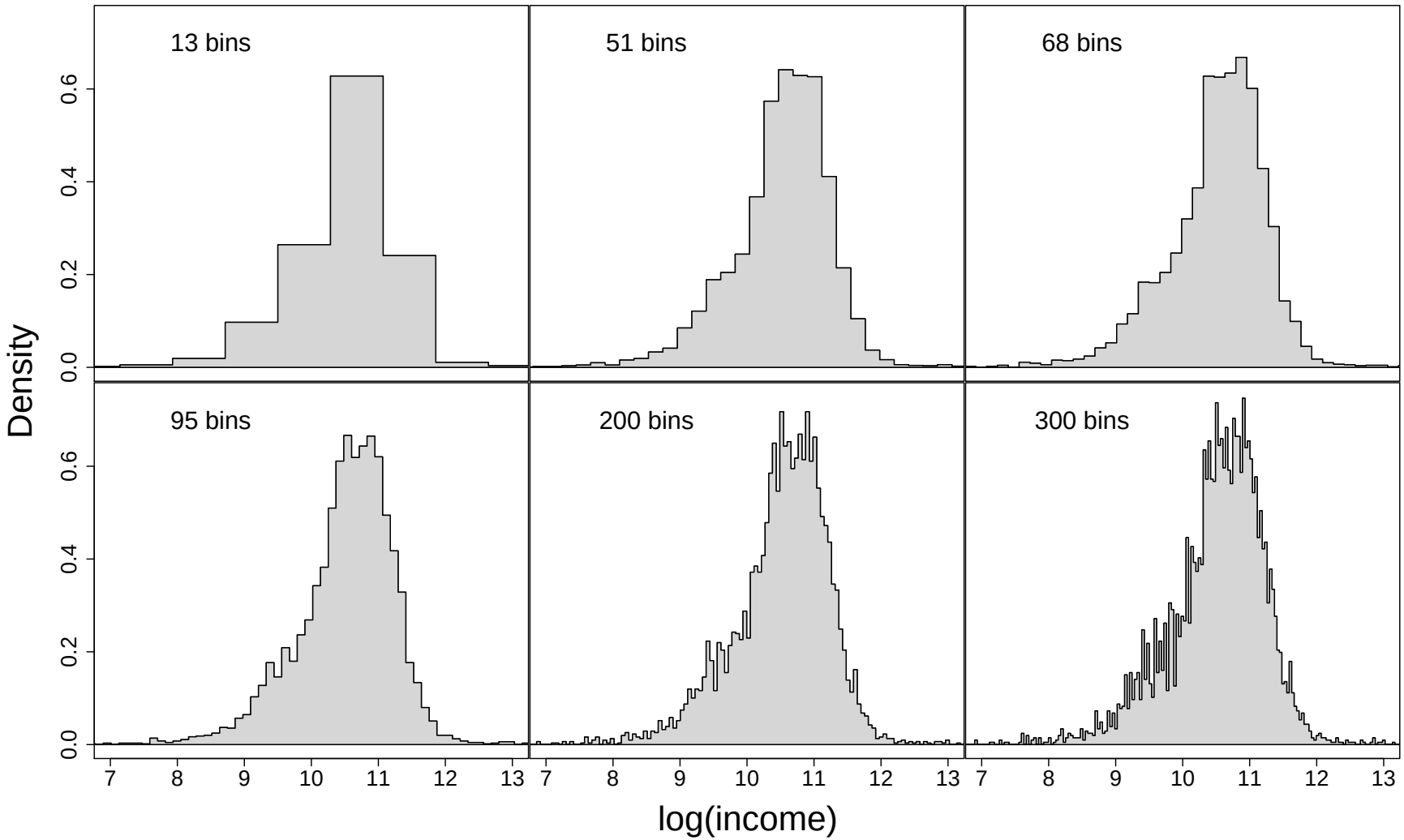


Figure: Six histograms of the 1983 German income sample of 5,625 households (Sturges 13 bins; BCV 95 bins; UCV 95-200 bins)

Example: German Household Income Data (continued)

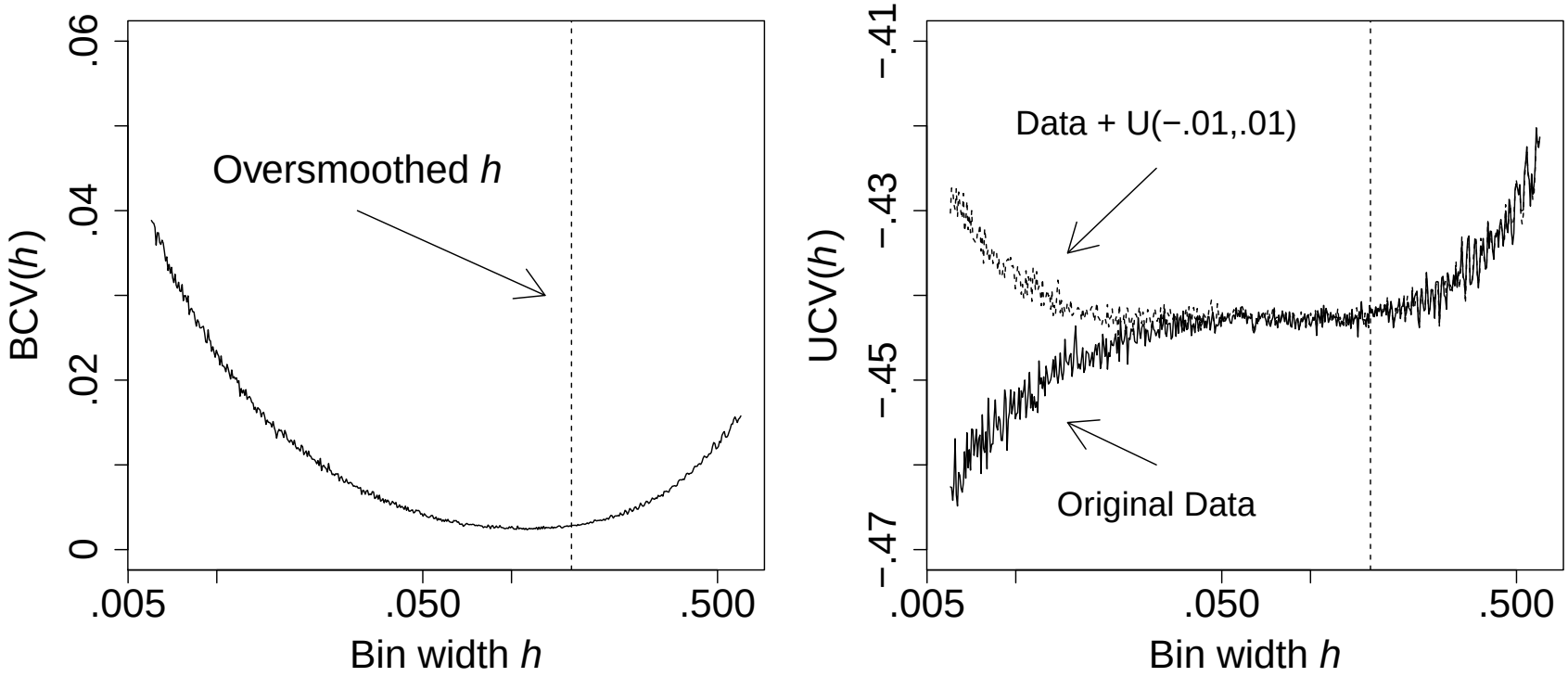


Figure: BCV and UCV functions of the 1983 German income sample of 5,625 households. A second UCV curve is shown for the blurred logarithmic data. (Note: UCV will return $h = 0$ if thinks discrete data.)

Remarks on *BCV* and *UCV*

- Both $BCV(h)$ and $UCV(h)$ converge to 0 as $h \to \infty$
  - Since $UCV(h)$ omitted the (unknown) constant $R(f)$, this is unbiased and why the minimizer occurs at negative values
  - For $BCV(h)$, the value 0 would be the global minimum, so must pay attention to the constraint $\hat{h} \leq h_{OS}$
- Scott and Terrell (1987) showed the *UCV* is much noiser than *BCV*; however, $UCV(h)$ is unbiased for all values of h, whereas $BCV(h)$ is accurate only in a range around h^*
- $BCV(h)$ often has no local minimum; so choose one of the other rules
- Check the histogram demo at **davidmlane.com**
- *BCV* targets the *AMISE* criterion, while *UCV* goes directly after *ISE* or *MISE*
- However, slow convergence: $\sigma_{h_{\mathrm{CV}}}/h_{\mathrm{CV}} = O(n^{-1/6})$
- Hall, Marron,... showed can construct a better (point) estimator of $R(f')$ with higher rate of convergence

L_2 Theory for Multivariate Histograms

- The derivation of the MISE for the multivariate histogram is only slightly more complicated than in the univariate case.
- Having gone through some of the details and issues in the univariate case, we will not dwell on those details from now on (see the book)
- Consider a regular partition of size $h_1 \times h_2 \times \cdots \times h_d$

$$\hat{f}(\mathbf{x}) = \frac{\nu_k}{n h_1 h_2 \cdots h_d} \quad \text{for } \mathbf{x} \in B_k$$

- Similar Binomial approximations lead to

$$AMISE(\mathbf{h}) = \text{AIV} + \text{AISB} = \frac{1}{n h_1 h_2 \cdots h_d} + \frac{1}{12} \sum_{i=1}^{d} h_i^2 R(f_i).$$

Best Bin Widths for Multivariate Histogram

- $R(f_i)$ is the multivariate integral of the square of the partial derivative

$$h_k^* = R(f_k)^{-1/2} \left(6 \prod_{i=1}^{d} R(f_i)^{1/2} \right)^{1/(2+d)} n^{-1/(2+d)},$$

$$AMISE^* = \frac{d+2}{2} 6^{-d/(2+d)} \left(\prod_{i=1}^{d} R(f_i) \right)^{1/(2+d)} n^{-2/(2+d)}.$$

- Note the bin widths get wider as the dimension d increases
- But the rate of convergence of the *AMISE* slows as d increases
- Aside: Surprising fact is that any orthogonal rotation of the multivariate bins does not change the total *AMISE*

EXAMPLE: Multivariate Normal Case

- Suppose that $X \sim N(\mu, \Sigma), \quad \Sigma = \text{Diag}(\sigma_1^2, \sigma_2^2, \ldots, \sigma_d^2)$.
- Let $c_d = \sigma_1 \sigma_2 \cdots \sigma_d$; then

$$R(f_i) = (2^{d+1} \pi^{d/2} \sigma_i^2 c_d)^{-1}$$

and

$$h_k^* = 2 \cdot 3^{1/(2+d)} \pi^{d/(4+2d)} \sigma_k n^{-1/(2+d)}$$

$$AMISE^* = (2+d)\, 2^{-(1+d)} 3^{-d/(2+d)} \pi^{-d^2/(4+2d)} c_d^{-1} n^{-2/(2+d)}$$

- Note: The constant in h_k^* varies from 3.4908 in one dimension to the limiting value of $2\sqrt{\pi} = 3.5449$ as $d \to \infty$

normal reference rule :	$h_k^* \approx 3.5\, \sigma_k\, n^{-1/(2+d)}$

Curse of Dimensionality

- Bellman first coined the phrase "curse of dimensionality" to describe the exponential growth in combinatorial optimization as the dimension increases.
- In our context, it is the number of bins that grows exponentially as the dimension increases (and most will be empty)

Table: Example of Asymptotically Optimal Bin Widths and Errors for $f = N(\mathbf{0}_d, I_d)$

Dimension d	h_d^*	AMISE_d^*
1	$3.491n^{-1/3}$	$0.430n^{-2/3}$
2	$3.504n^{-1/4}$	$0.163n^{-2/4}$
3	$3.512n^{-1/5}$	$0.058n^{-2/5}$
4	$3.518n^{-1/6}$	$0.020n^{-2/6}$

- The convergence rates of the parametric estimates are $O(n^{-1})$

How to Compare *AMISE* Across Dimesions

- Epanechnikov created a dimensionless quantity

$$\epsilon_d \equiv \frac{\text{MISE}}{R(f)} \quad \left\{ \approx \frac{2+d}{2} 3^{\frac{-d}{2+d}} \pi^{\frac{d}{2+d}} n^{-\frac{2}{2+d}} \quad \text{when } f = N(\mathbf{0}_d, I_d) \right\}$$

Table: Equivalent Sample Sizes Across Dimensions for the Multivariate Normal Density, Based on Epanechnikov's Criterion

d	Equivalent Sample Sizes (Read Down Each Column)		
1	10	100	1,000
2	39	838	18,053
3	172	7,967	369,806
4	838	83,776	8,377,580
5	4,446	957,834	206,359,075

- Rather pessemistic

Another Comparison of *AMISE* Across Dimesions

- Count the bins in a region of interest
- For normal data, this would be a sphere with radius r_d containing 99% of the probability mass; recall

$$\text{Prob}\left(\sum_{i=1}^{d} Z_i^2 \leq r_d^2\right) = 0.99 \quad \Rightarrow \quad r_d = \sqrt{\chi^2_{.99}(d)}$$

Table: Approximate Number of Bins in the Region of Interest for a Multivariate Histogram of 1,000 Normal Points

d	h_d^*	r_d	Number of Bins
1	0.35	2.57	15
2	0.62	3.03	75
3	0.88	3.37	235
4	1.11	3.64	573
5	1.30	3.88	1,254

A Special Case: $d = 2$ with Nonzero Correlation

- Have ignored the effect of correlation on h^*
- With $f(x_1, x_2) = N(\mu_1, \mu_2, \sigma_1^2, \sigma_2^2, \rho)$

$$R(f_1) = \left[8\pi(1-\rho^2)^{3/2}\sigma_1^3\sigma_2\right]^{-1}$$
$$R(f_2) = \left[8\pi(1-\rho^2)^{3/2}\sigma_1\sigma_2^3\right]^{-1}$$

- Now can minimize the $AMISE(h_1, h_2)$ and get

$$h_i^* = 3.504\,\sigma_i\,(1-\rho^2)^{3/8}\,n^{-1/4}$$
$$AMISE^* = \frac{0.122}{\sigma_1\sigma_2}(1-\rho^2)^{-3/4}\,n^{-1/2}$$

- As expected, as the correlation increases, the bin widths decrease to try to follow the steep sides of the bivariate normal density; but the $AMISE$ blows up as $\rho \to \pm 1$. Full rank necessary!

Interesting Problem: Optimal Regular Bivariate Meshes

- There are 3 regular polygons one might use as bins for a bivariate histogram. Is one better than the others? Why?
- squares, equilateral triangles, or hexagons

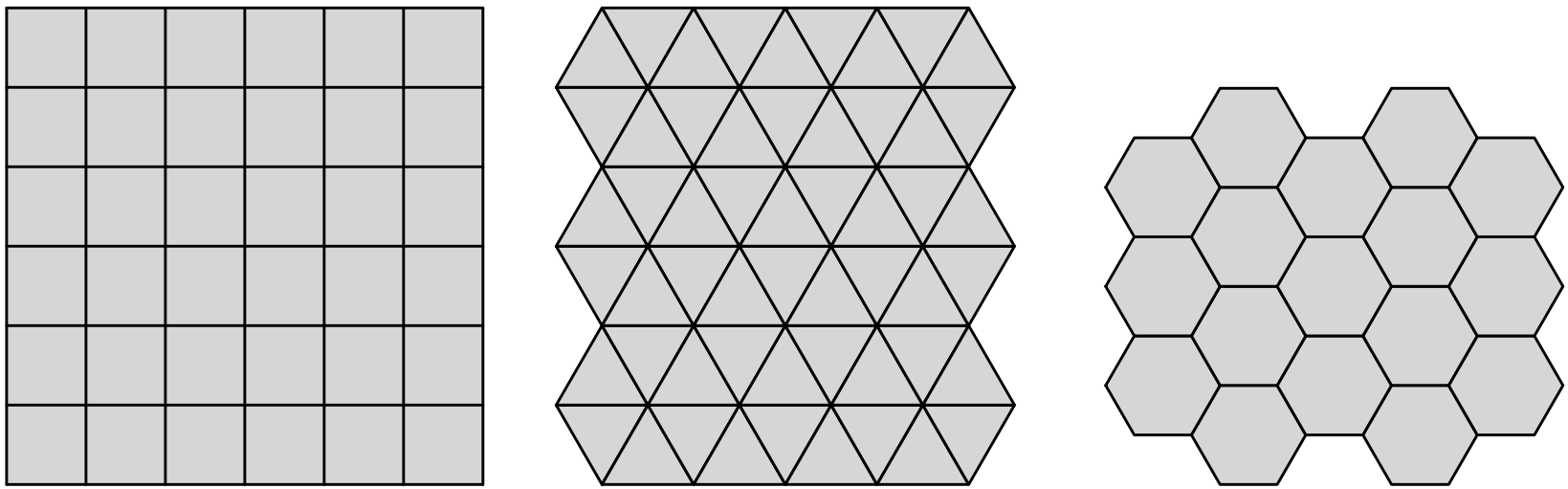


Figure: The 3 possible regular polygon meshes for bivariate histograms.

Optimal Regular Bivariate Meshes (Results)

- Scott (1988) showed that if parameterize the bins so that have same area, h^2, then

$$\text{AMISE}(h) = \frac{1}{nh^2} + ch^2\left[R(f_x) + R(f_y)\right]$$

where

$$c = \left[\frac{1}{12}, \frac{1}{6\sqrt{3}}, \frac{5}{36\sqrt{3}}\right] = \left[\frac{1}{12}, \frac{1}{10.39}, \frac{1}{12.47}\right]$$

- Conclusions:
  - triangles are the worst
  - hexagons are best, but only a bit better than squares
  - circles would be optimal, but cannot tile $\Re^2$
  - note data can always to scaled so that rectangular bins same as squares

Optimal Regular Bivariate Meshes (Uses)

- Dan Carr (GMU) has advocated using histogram glyphs as an alternative to bivariate scatter diagrams for massive datasets.
- He observed that it his idea did not work well using square/rectangular bins, and that hexagonal bins broke the visual North/South East/West vice

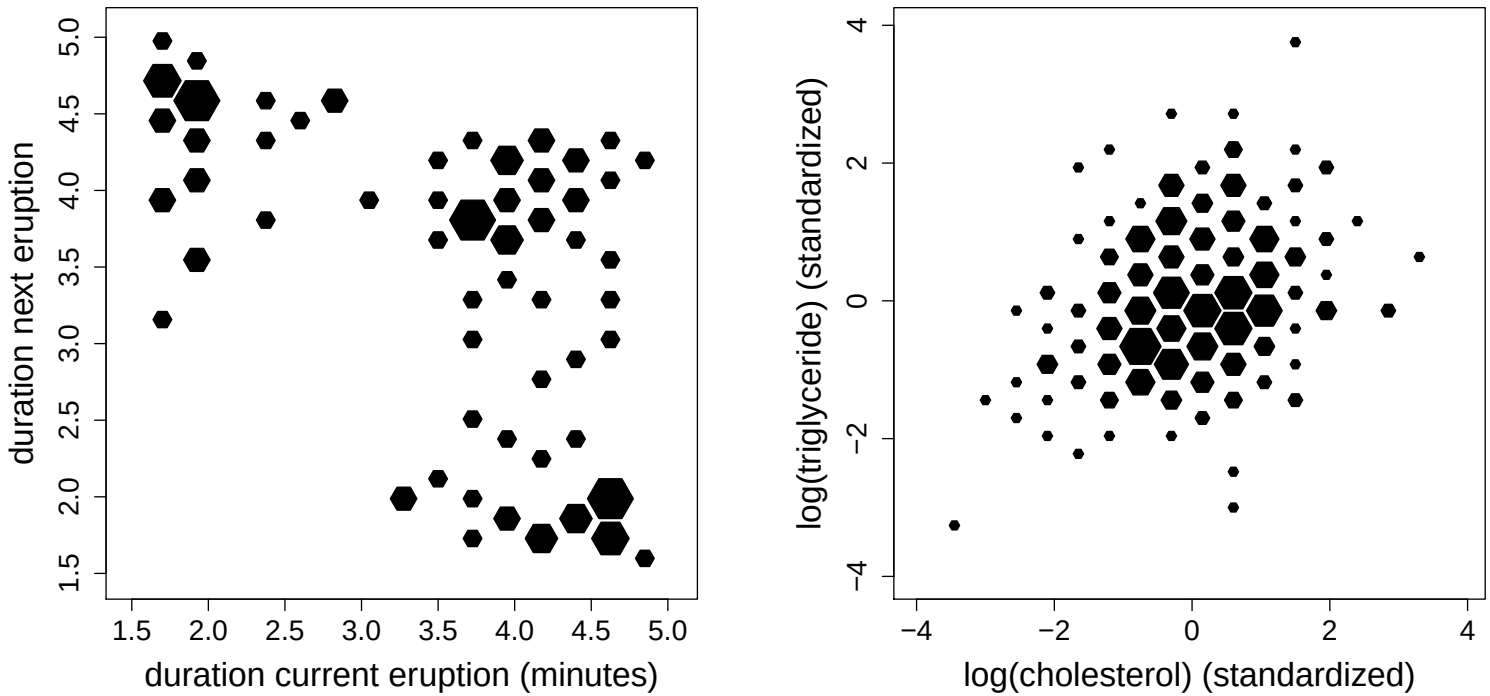


Figure: Hexagon glyphs for the lagged Old Faithful duration data and for the lipid dataset for 320 males with heart disease.

Applications: Modes and Bumps in a Histogram

- Many of our case studies display multimodal densities, an "unexpected" feature of the data
- I.J. Good wrote an influential JASA paper on using nonparametric density estimation to identify and evaluation modes and bumps in his maximum penalized likelihood estimate. In one dimension $\Re^1$,
  - a **mode** is a set (a collection of points) where $f'(x) = 0$ and $f''(x) < 0$
  - a **bump** is a set (a collection of disjoint intervals) where $f''(x) < 0$
- Thus bump-hunting is more general than estimating the mode
- A bump does not necessarily contain a mode, although a mode is always located in a bump.

Bump-hunting Examples

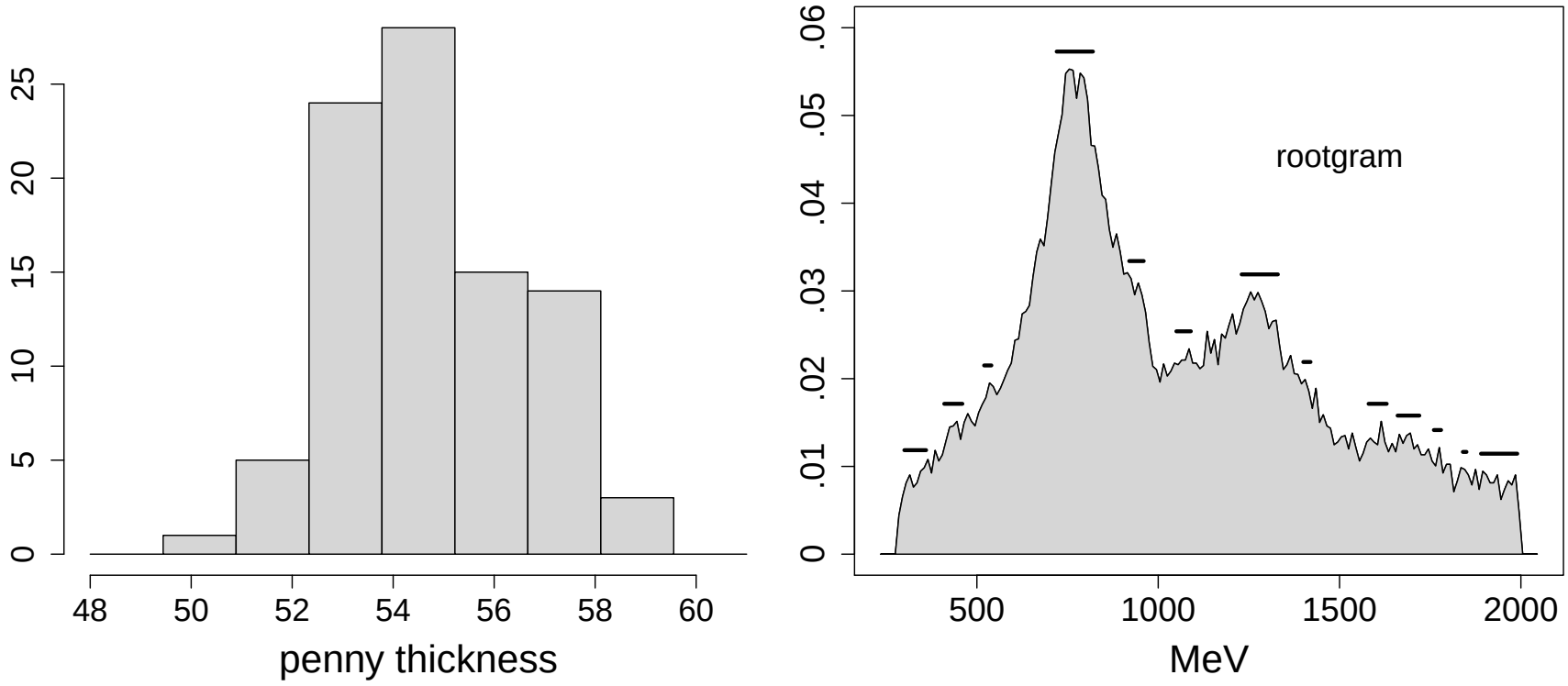


Figure: Histogram of U.S. penny thickness with one mode and a bump. In the right frame, the histogram of the LRL dataset with $h = 10$ MeV is plotted on a square root scale; the 13 bumps found by Good and Gaskins are indicated by the line segments above the histogram.

Aside: Bump-hunting using mixture models

- Bumps may be thought of as clusters
- The mode in a bump may be thought of as its exemplar
- In practical bump-hunting situations, the density is often modeled as a mixture of several component densities

$$f(x) = \sum_{i=1}^{q} w_i \, \phi(x | \mu_i, \sigma_i^2) \qquad \text{where} \sum_{i=1}^{q} w_i = 1$$

- Interestingly, although there may be q clusters/components, there may be only one mode, or anything in-between
- Thus, the mixture problem is even more general than bump-hunting, since a normal mixture density can reveal more clusters than bumps/modes
- However, unless widely separated components, the estimates of the parameters $\{q, w_i, \mu_i, \sigma_i^2\}$ are highly unstable

Rough and Ready Tests for Modes and Bumps

- natural to examine plots of (standardized) first and second differences for evidence of modes and bumps, respectively

$$\frac{\nu_{k+1} - \nu_k}{\sqrt{\nu_{k+1} + \nu_k}} \quad \text{and} \quad \frac{\nu_{k+1} - 2\nu_k + \nu_{k-1}}{\sqrt{\nu_{k+1} + 4\nu_k + \nu_{k-1}}}$$

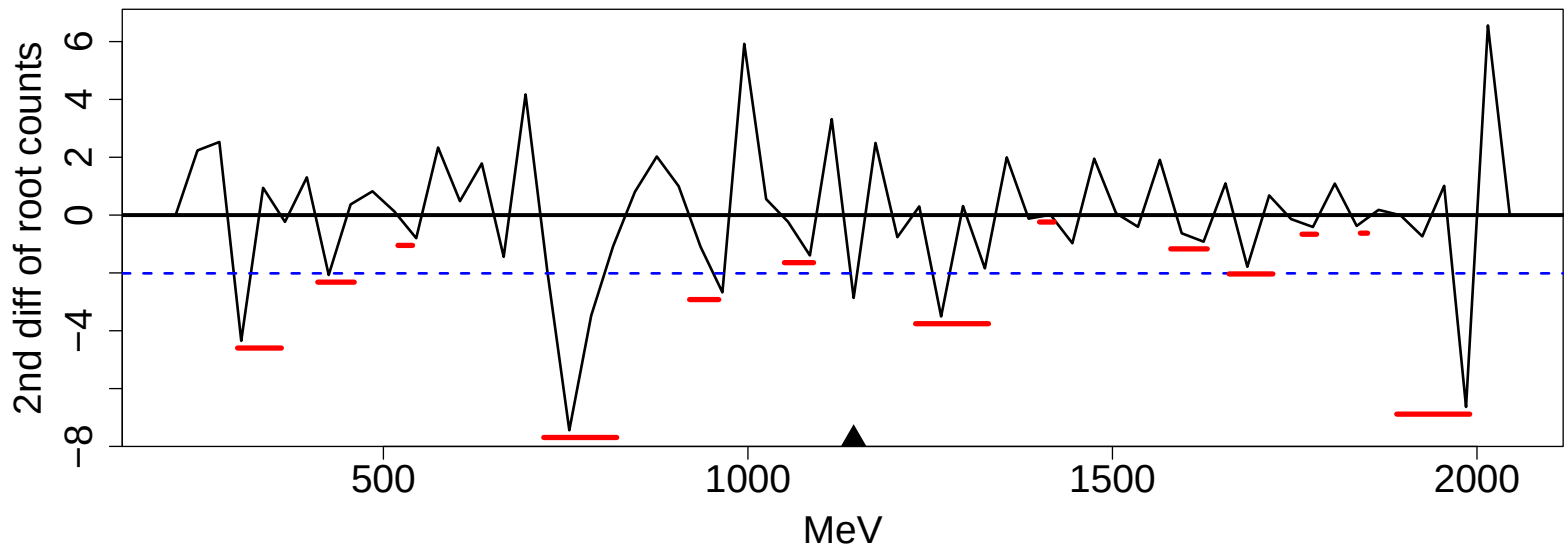
- since the bin counts can be approximately modeled as independent Poisson random variables. For example, $\text{var}\,(\nu_{k+1} - \nu_k) \approx \nu_{k+1} + \nu_k$.
- Alternatively, we may focus on finite differences of the rootgram, using the fact that the square root is the variance stabilizing transformation for Poisson random variables, with variance of $1/4$:

$$\sqrt{\nu_{k+1}} - \sqrt{\nu_k} \quad \text{and} \quad \sqrt{\nu_{k+1}} - 2\sqrt{\nu_k} + \sqrt{\nu_{k-1}}$$

which have approximate variances $1/2 = \frac{1}{4} + \frac{1}{4}$ and $3/2 = \frac{1}{4} + (-2)^2\frac{1}{4} + \frac{1}{4}$, respectively

Good's LRL Data and 13 Bumps Re-Visited

- Consider a plot of the second differences of the root bin counts of the LRL histogram with $h = 30$ MeV. Red lines indicated the 13 bumps found by Good and Gaskins. The dashed line indicates the approximate 5% cutoff level for a bump to be significant.



- Under the null hypothesis that there is no bump, i.e., the second difference is nonnegative, the one-sided 5% test level will be at -1.645$\sqrt{3/2}$ =-2.015. (Missed bump at 1145 MeV?)

Properties of Histogram "Modes:" Some Theory

- Suppose, without loss of generality, that the true density f has a mode at 0
- Construct a histogram where bin $B_0 = [-h/2, h/2)$
- The bin count $\nu_0 \sim B(n, p_0) \approx P(\lambda_0)$, which is the Poisson density with $\lambda_0 = np_0$
- Asymptotically, the adjacent bin counts, $(\nu_{-k}, \ldots, \nu_0, \ldots, \nu_k)$, are independent and normally distributed with $\nu_i \approx N(\lambda_i, \lambda_i)$
- **Question:** What is the probability that the histogram will have a **sample mode** in bin B_0?

Properties of Histogram "Modes:" (continued)

- Conditioning on the observed bin count in B_0, we obtain

$$
\begin{aligned}
&Prob\left(\nu_0 = \arg\max_{|j|\le k} \nu_j\right) \\
&\quad = \sum_{x_0} \Pr\left(\nu_0 = \arg\max_{|j|\le k} \nu_j \middle| \nu_0 = x_0\right) f_{\nu_0}(x_0) \\
&\quad = \sum_{x_0} \Pr(\nu_j < x_0; |j| \le k, j \neq 0) f_{\nu_0}(x_0) \\
&\quad \approx \int_{x_0} \prod_{\substack{j=-k \\ j\neq 0}}^{k} \Phi\left(\frac{x_0 - \lambda_j}{\sqrt{\lambda_j}}\right) \phi\left(\frac{x_0 - \lambda_0}{\sqrt{\lambda_0}}\right) \frac{1}{\sqrt{\lambda_0}}\, dx_0
\end{aligned}
$$

- Skipping a number of Taylor Series approximations to bin probabilities, we obtain

$$\Pr\left(\nu_0 = \arg\max_{|j|\le k} \nu_j\right) \approx \int_y \prod_{\substack{j=-k \\ j\neq 0}}^{k} \Phi\left(\frac{\lambda_0 - \lambda_j + y\sqrt{\lambda_0}}{\sqrt{\lambda_j}}\right) \phi(y)\, dy$$

$$\approx \int_y \prod_{\substack{j=-k \\ j\neq 0}}^{k} \Phi\left(y - \frac{j^2 h^{5/2}\sqrt{n}}{2}\frac{f''(0)}{\sqrt{f(0)}} + \cdots\right) \phi(y)\, dy$$

- In the case of an optimal histogram, $h = cn^{-1/3}$, so that

$$\lim_{n\to\infty} [h^{5/2}\sqrt{n}] = O(n^{-1/3}) \to 0 \qquad \text{hence,}$$

$$\lim_{n\to\infty} \Pr\left(\nu_0 = \arg\max_{|j|\le k} \nu_j\right) = \int_y \Phi(y)^{2k}\phi(y)dy = \frac{1}{2k+1} \quad (!!)$$

- Tentative conclusion: The "optimal" *AMISE* histogram is **too flat** near the mode!! Bins are too narrow.

Simulation Confirmation of This Result

Here is a simulation of optimally smoothed normal data with $n = 1{,}000{,}000$, the average number of sample modes in the histogram was 20! Now admittedly, most of these "modes" were just small aberrations, but the result is rather unexpected.

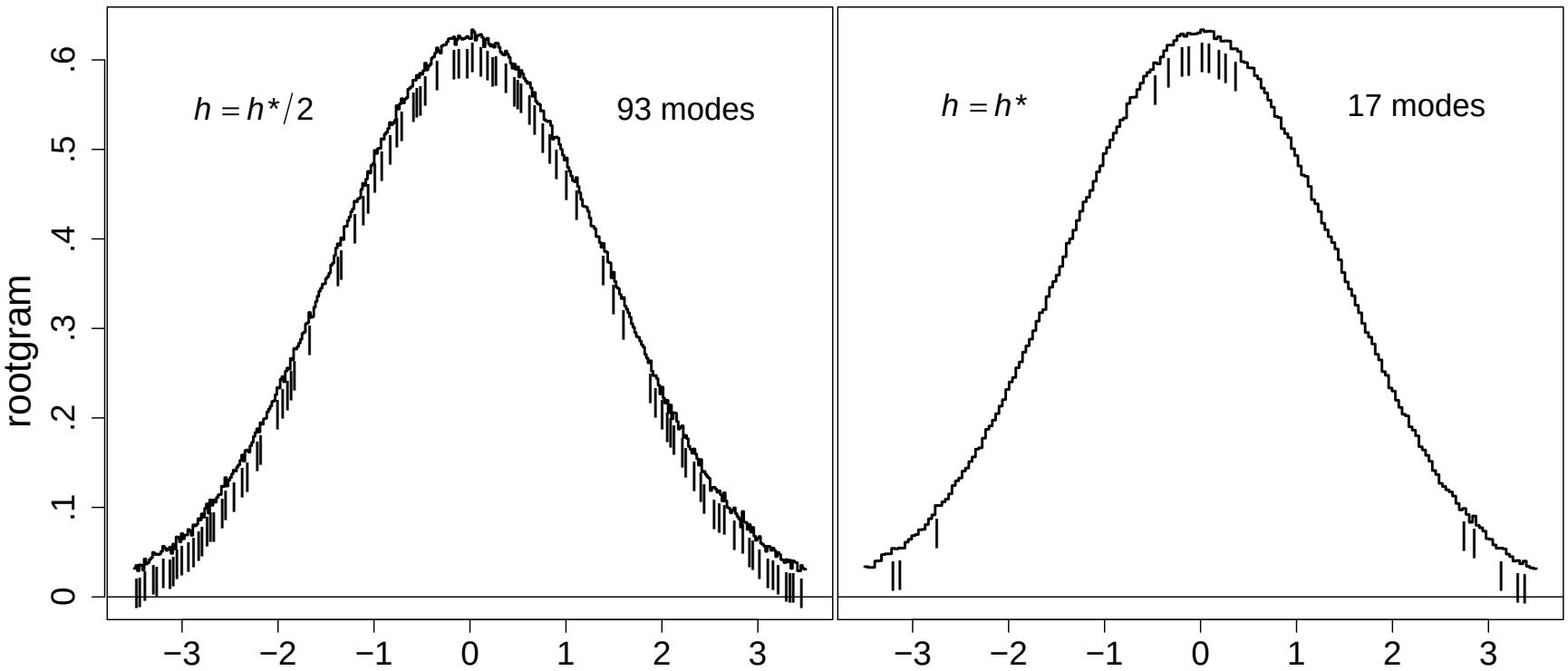


Figure: Rootgram of 2 histograms of a million normal points $(h^*/2, h^*)$

Properties of Histogram "Modes:" (continued)

- In the previous slides, we assumed there was a mode at $x = 0$
- Now we assume that there is not a mode at $x = 0$
- A similar analysis shows the probability that bin B_0 is a mode (which it is not!) converges to a fixed nonzero probability as $n \to \infty$.
- The probability is smaller the larger the magnitude of $f'(0)$.
- Note the simulation showed false modes only near the mode and in the tails where $\phi'(x) \approx 0$, and not where the slope is large
- The bottom line is that histograms which are "optimal" with respect to MISE may not be "optimal" for other purposes.

Optimal Histogram Bandwidths for Modes

- What is the bandwidth that provides better estimates of the derivative of the unknown density?
- Want the finite difference derivative estimate to be close to the true $f'(x)$:

$$\hat{f}'(x) = \frac{\hat{f}_0 - \hat{f}_{-1}}{h} = \frac{\nu_0 - \nu_{-1}}{nh^2} \qquad -h/2 < x < h/2\,.$$

- Finding the *AMISE* of $\hat{f}'(x)$, the final result is

$$\text{AMISE}_{\hat{f}'}(h) = \frac{2}{nh^3} + \frac{1}{12}h^2 R(f''); \quad \text{therefore,}$$
$$h^* = 6^{2/5}\, R(f'')^{-1/5}\, n^{-1/5}$$
$$\text{AMISE}^* = 5 \cdot 6^{-6/5} R(f'')^{3/5} n^{-2/5}.$$

- If $f \sim N(\mu, \sigma^2)$, then $h^* \approx 2.8\,\sigma\, n^{-1/5}$ (not $n^{-1/3}$)
- Notice the result uses $R(f'')$ rather than $R(f')$; intuition?

Simulations of Normal reference rule for histogram derivative: Start by using h^* for the density

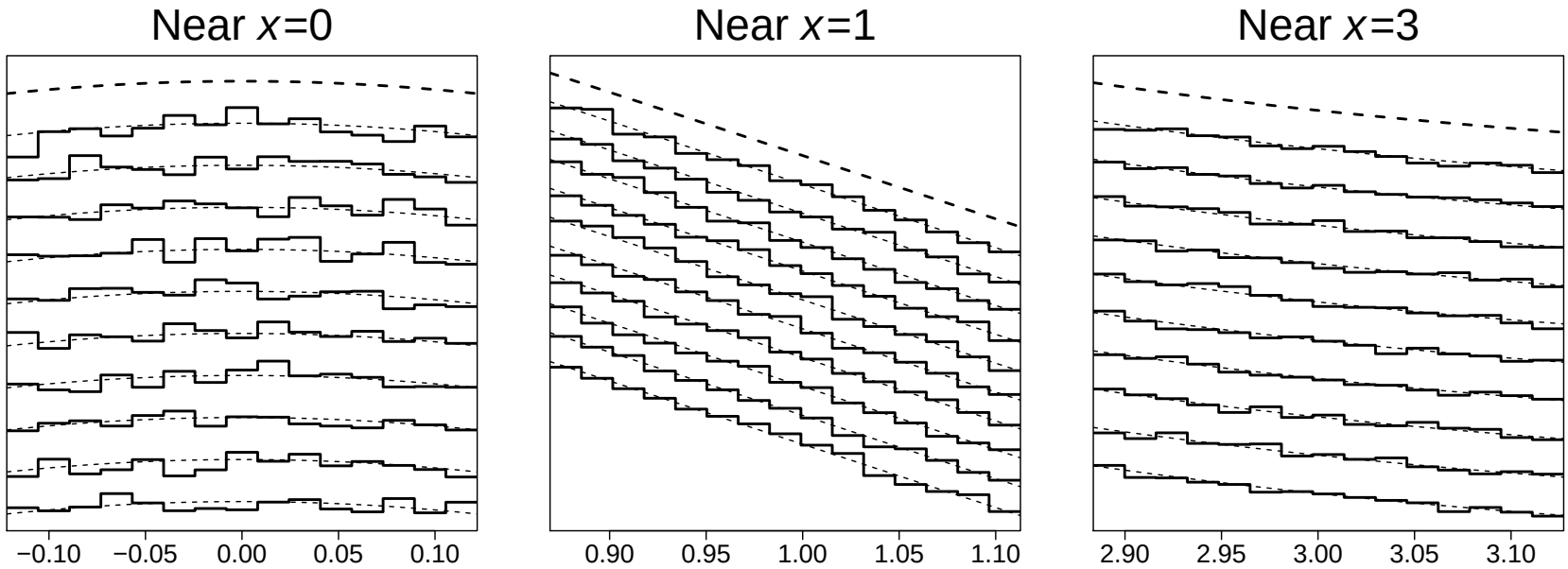


Figure: For 10 $N(0,1)$ samples of size $n = 10^7$, (vertically shifted) blowups of portions of the ten histograms using $h^* = 0.01625$ in the vicinity of $x = 0$, 1, and 3. Each histogram snippet shows the bin of interest and 7 bins on either side.

Simulations of Normal reference rule for histogram derivative: Next use h^* for the derivative

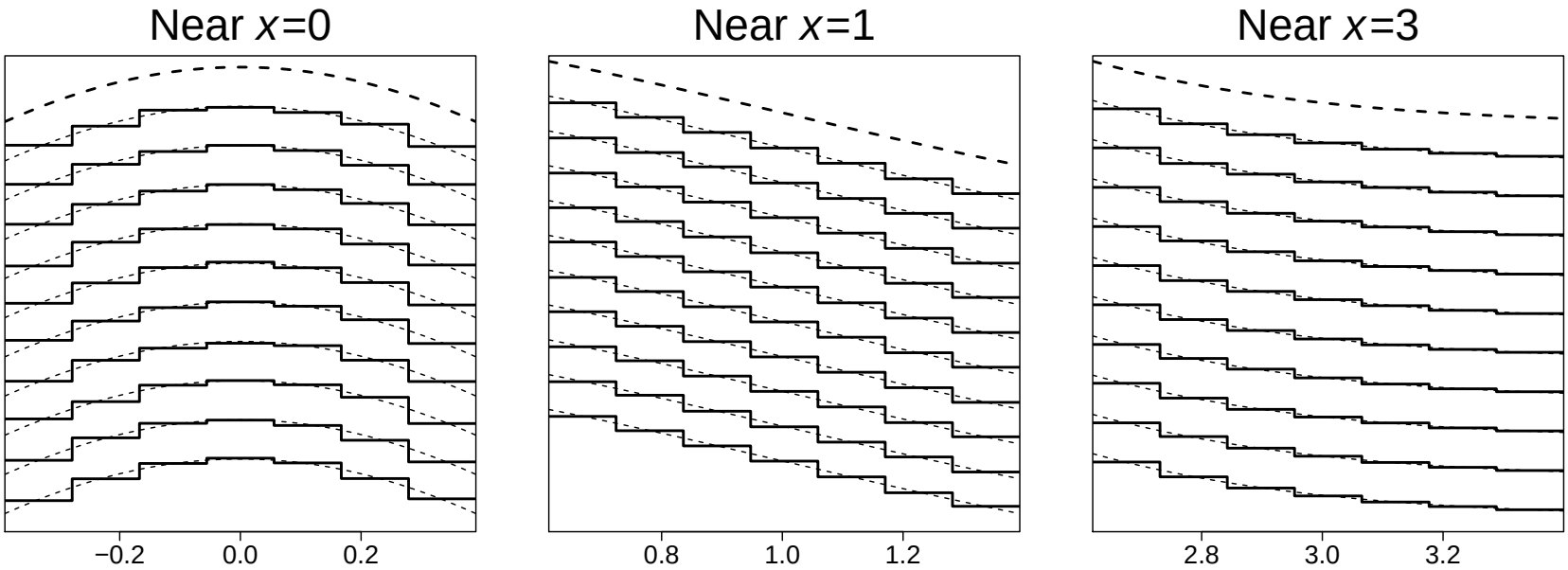


Figure: For the same 10 $N(0,1)$ samples of size $n = 10^7$, (vertically shifted) blowups of portions of the ten histograms using $h^* = 0.1115$ in the vicinity of $x = 0$, 1, and 3. Each histogram snippet shows the bin of interest and 3 bins on either side.

Simulations of Normal reference rule for estimating $\phi'(x)$

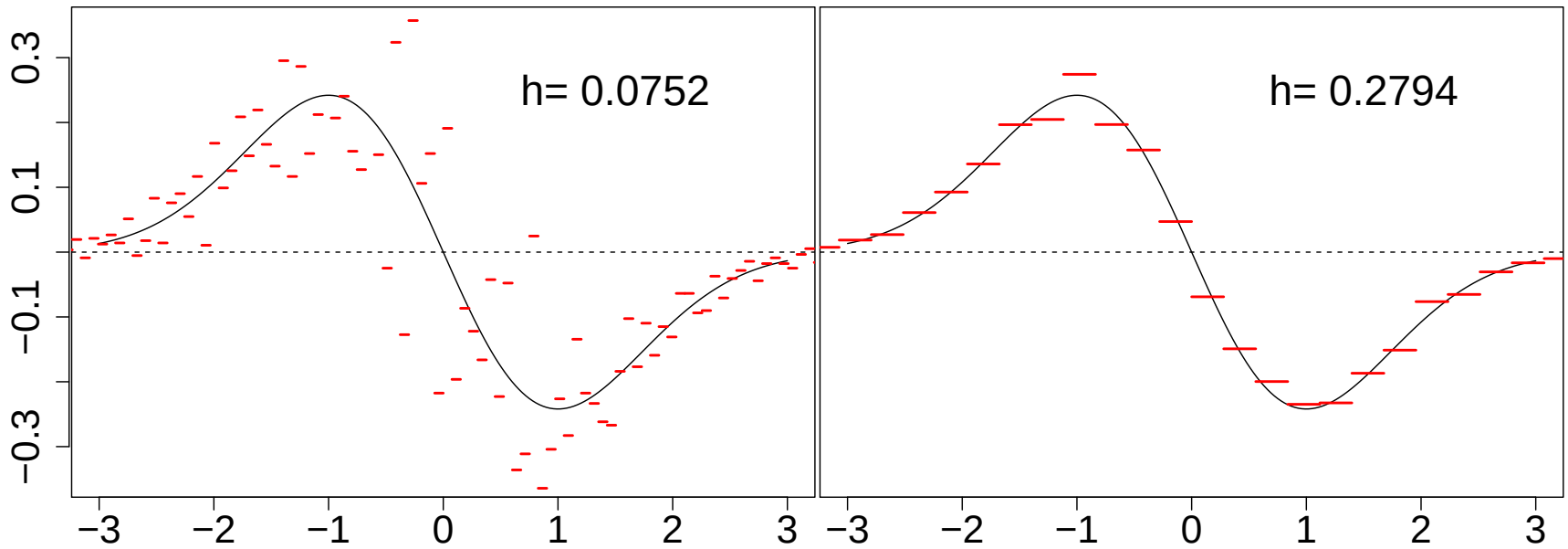


Figure: For a standard normal sample $n = 10^5$ points, comparison of the histogram "derivative" estimates using the optimal density and derivative bandwidths.

A Useful Bimodal Mixture Density to Understand

- A useful density to understand estimation difficulties

$$f_M(x) = \frac{3}{4}\phi(x|0,1) + \frac{1}{4}\phi(x|3,\tfrac{1}{3^2})$$

- Designed so that $f(0) \approx f(3)$, but $f''(0) \neq f''(3)$

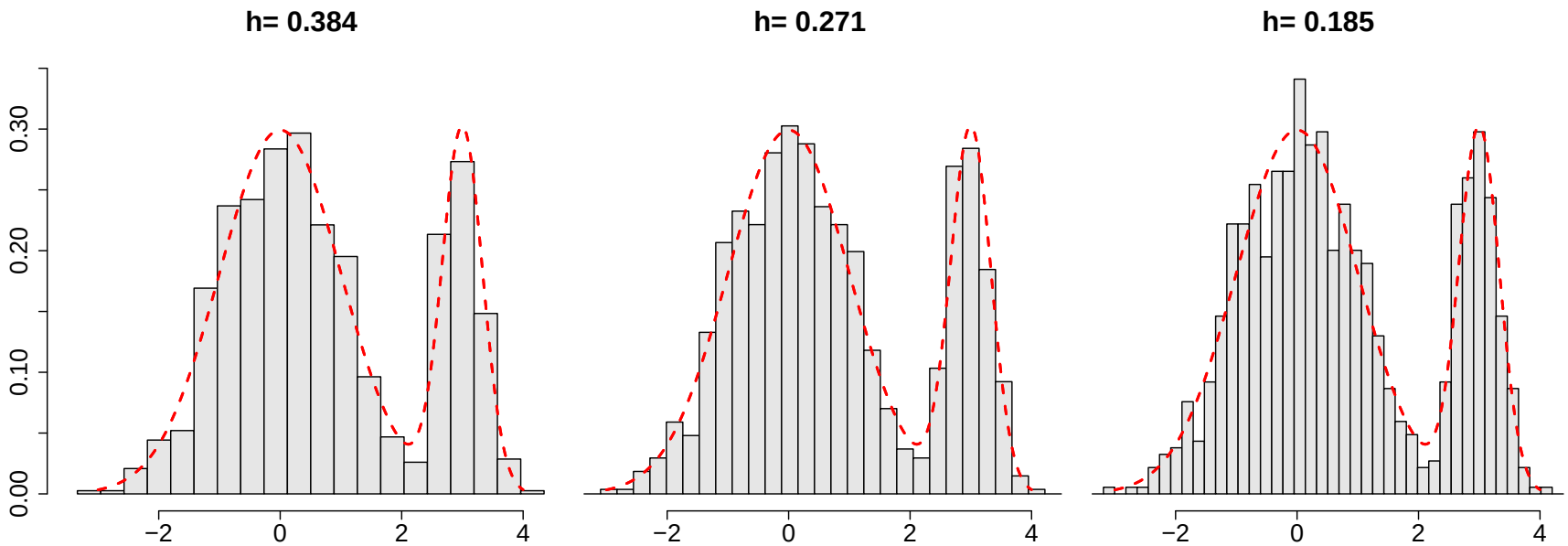


Figure: Three histograms of 1,000 points from the two-component mixture. The bandwidths (from left to right) are optimal for the left component, the mixture, and the right component, respectively.

Wrapping Up: Other Error Criteria: L_1, L_4, L_6, L_8, and L_∞

Table: Optimal Bandwidths for $N(0,1)$ Data with Different L_p Criteria

Error Criterion	Optimal Bin Width	Expected Error
L_1 (upper bound)	$2.72n^{-1/3}$	$1.6258n^{-1/3}$
L_1 (numerical)	$3.37n^{-1/3}$	$1.1896n^{-1/3}$
L_2	$3.49n^{-1/3}$	$(0.6555n^{-1/3})^2$
L_4	$3.78n^{-1/3}$	$(0.6031n^{-1/3})^4$
L_6	$4.00n^{-1/3}$	$(0.6432n^{-1/3})^6$
L_8	$4.18n^{-1/3}$	$(0.6903n^{-1/3})^8$

- Increasing the order p gives more weight to the high-density regions, and less in the tails. Thus, wider bins.
- Again, any fixed-bandwidth criterion will pay most attention to regions where the density is rough; that region is not necessarily in the tails.

Concluding Thoughts about Histograms

- The histogram is the most commonly used nonparametric probability density estimator.
- The histogram turns out not to be the most statistical efficient estimator.
- However, the histogram is the most computationally efficient estimator.
- If analyzing massive datasets, the loss of statistical efficiency may not be important.
- That said, there are better estimators for data in dimensions $\Re^1 \longrightarrow \Re^5$ that we will study next.
- Lots of room to teach histograms in introductory courses in an improved manner.