Homework 2

Aniket Kesari

October 5, 2017

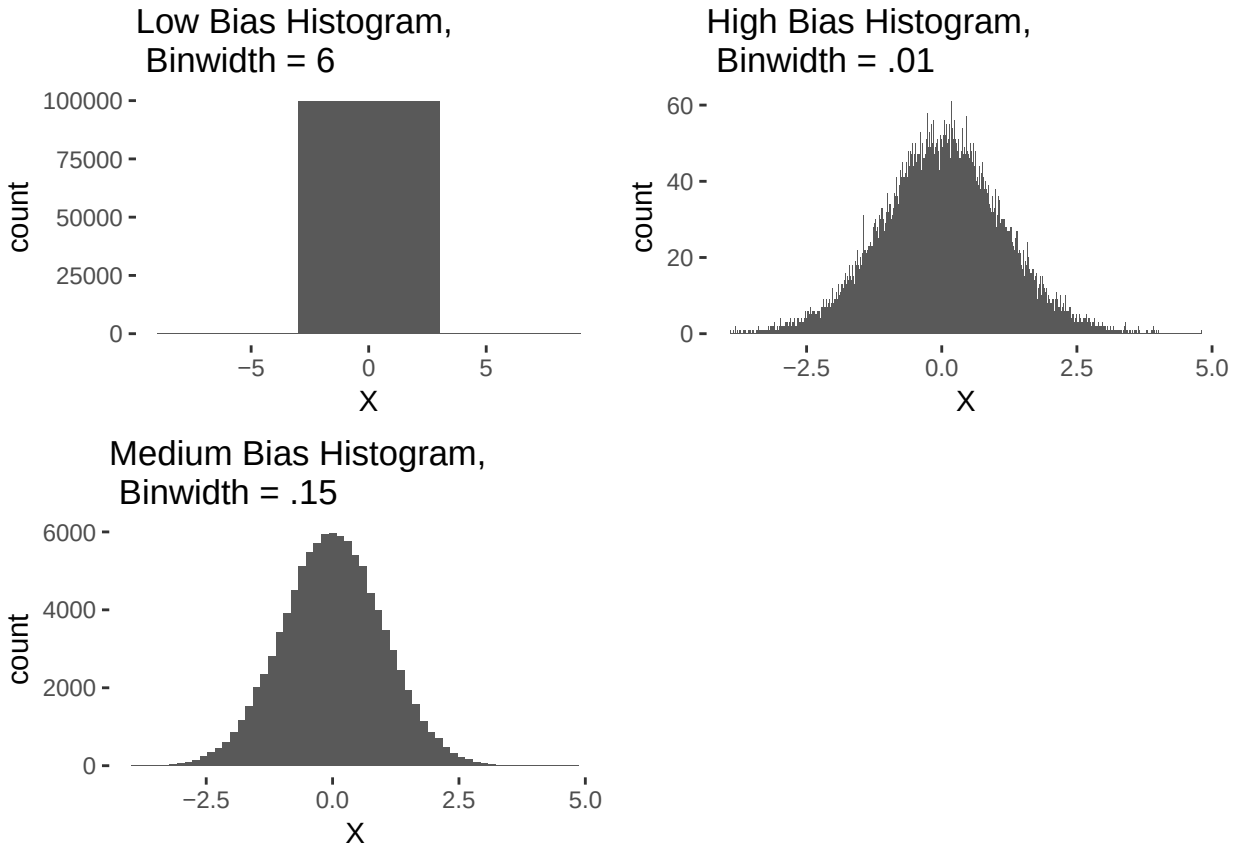
Kernel Density Estimation

Bias-Variance Trade-off in Histograms

Simply put, the bias-variance tradeoff is the balancing act between two sources of error. Bias refers to how well the model fits the data, while variance refers to how well the model maps the spread of the data to prediction. Reducing bias increases the variance of an estimate, and vice versa. A low-bias, high-variance estimate will explain the observed data well but will likely model every irregularity and make bad predictions (overmodels the "noise"). A high-bias, low-variance model will fail to detect patterns arising from the variation in data spread (undermodels the "signal"). The key is to optimize bias and variance so that the model is neither over- nor under-fit.

I illustrate this concept by experimenting with histograms and kernel density plots. The main objective here is to illustrate how the bias-variance tradeoff relates to estimating the underlying probability density function. Both histograms and kernel density plots are used for this purpose, and it is an important part of visualizing and inferring from the data.

First, with histograms I illustrate how large bins generate high bias, low variance estimates and conversely smaller bins generate low bias, high variance estimates. I start by using R's pseudo-random number generator (rnorm) to create a dataframe with 100,000 observations that are distributed approximately normally. I show an extremely large binwidth that captures basically all of the data, an extremely small binwidth that does not have enough data within each bin to properly visualize the data, and an appropriate binwidth that approximates ~30 observations per bin.

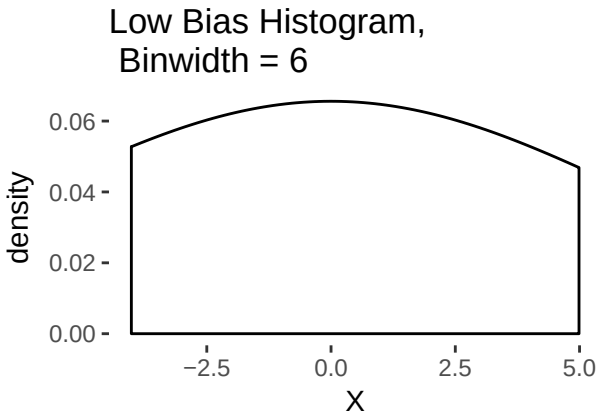


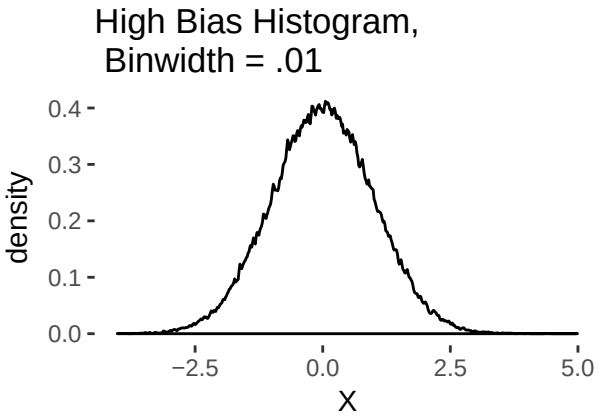
Examining these plots, it is clear that neither a no-bias nor no-variance estimate does a good job approximating the underlying distribution. The low-bias plot appears captures all of the data in one bin of size 6 which making it appear as one large rectangle. This shape does not give any indiciation as to the actual shape of the data. The high bias, low-variance estimate with a binwidth of .001, and does a bad job of approximating the data near the center of the distribution because there are not enough datapoints to fill each bin. This causes the visualization to underestimate how much data should be at the center. Extending to other statistical applications, this is bad because a model that does not approximate central tendency will probably do worse when extrapolating to the population or making predictions.

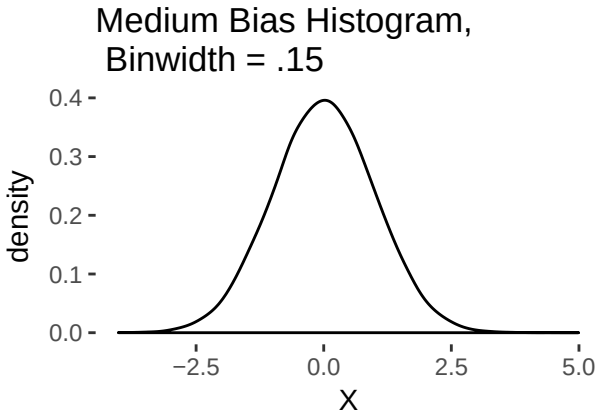
Meanwhile, the third plot on the bottom optimizes between bias and variance by using a binwidth of .15, which allows for about 30 observations per bin (there are 100,000 points in the overall dataframe). The sample size is still too small to create a totally smooth curve, but it does a much better job approximating a normal shape. This plot illustrates nicely the balance between capturing enough points to lower bias, while not lowing bias so much that it creates a high variance estimate that misses important signals.

Bias-Variance Tradeoff in Kernel Density Estimates

Next, I illustrate the same concept with kernel density plots.







Using the same binwidths as the histograms, I show that the density plots also highlight the bias-variance tradeoff. The large binwidth captures all of the data and does not provide any indication of the shape. In this case, the small binwidth plot (high bias) does a much better job approximating normality, but still does a poor job with smoothing. Again, the binwidth of .15 seems to provide a happy medium by visualizing the smooth, normal shape that one would expect.

Bias Term of the Mean Squared Error

The bias term of the Mean Squared Error is:

Bias $= E[\hat{f}(x)] - f(x)$

Essentially, this expression says that the bias is measured as the difference between the expected values of the outcome variable and the observed values.

Does Bias Change with n? With h? Does this make sense?

Bias does not change with n, but it does change with h. This makes sense given my previous answers. Although variance gets smaller as sample size increases (because there is more data with which to estimate the overall population), there is no mathematical reason as to why bias would decrease solely because of increased sample size. However, adjusting the smoothing parameter, h, will change the bias term. This is because the smoothing parameter, h, penalizes x_i by limiting its shape, thus creating a gap between each x_i and x. A large smoothing parameter will alow x_i to get close to x making for low bias, while a small h will restrict how close they can be, thus creating more bias.

Multidimensional Scaling

Show if D is the matrix of Euclidean interpoint distances for a con???guration...

Using the definition of H, we can rewrite B to:

$B = A - \frac{1}{n}A11' - \frac{1}{n}11'A + \frac{1}{n^2}11'A11'$

So, we can rewrite: $b_{rs} = a_{rs} - \bar{a_s} - \bar{a_r} + \bar{a}$

A cell in B can be expressed as:

$b_{rs} = a_{rs} - \bar{a}_r - \bar{a}_s + \bar{a}$

We can then substitute a_{rs} for b_{rs} and get to:

$b_{rs} = -\frac{1}{2}(z_r - z_s)'(z_r - z_s) + \frac{1}{2n}\sum_{s=1}^{n}(z_r - z_s)'(z_r - z_s) + \frac{1}{2n}\sum_{r=1}^{n}(z_r - z_s)'(z_r - z_s) - \frac{1}{2n^2}\sum_{r=1}^{n}\sum_{s=1}^{n}(z_r - z_s)'(z_r - z_s) = -\frac{1}{2}(z_r - z_s)'(z_r - z_s) + \frac{1}{2n}(nz_r'z_r - 2nz_r'\bar{z} + \sum_{s=1}^{n} z_s'z_s) + \frac{1}{2n}(nz_s'z_s - 2nz_s'\bar{z} + \sum_{r=1}^{n} z_r'z_r) - \frac{1}{2n^2}(2n\sum_{s=1}^{n} z_s'z_s - 22n^2\bar{z}'\bar{z}) = \frac{1}{2}(z_s'z_r + z_r'z_s) - z_r'\bar{z} - z_s'\bar{z} + \bar{z}'\bar{z} = (z_r - \bar{z})'(z_s - \bar{z})$

To show that B is positive semi-definite, we need to show that all of its eigenvalues are nonnegative.

First we can rewrite B as: $B = Z_s'Z_s$

This implies that: $x'Bx = (Z_sx)'(Z_sx) \geq 0$

Let B. b p.s.d. of rank p with positive eigenvalues...

Show that the points with coordinates...

B can be written as:

$B = XX'$

We know from the result above that this can be rewritten as:

$x_r - x_s)'(x_r - x_s) = x_r'x_r - 2x_r'x_s + x_s'x_s = b_{rr} - 2b_{rs} + b_{ss} = a_{rr} - 2a_{rs} + a_{ss} = -2a_{rs} = d_{rs}^2$ (from the definition of A)

This shows that D is the interpoint matrix.

Further show that the center of gravity $\bar{x} = 0$ and B represents the inner product matrix

Be can be multiplied by a 1 eigenvector, and HAH can be rewritten as:

$B1 = HAH1 = 0$

Assuming the 1 eigenvector has an eigenvalue of 0, this implies that:

$B1 = HAH1 = X'1 = 0$

This means that every coordinate in X sums to 0, implying that the center of gravity is also there.