Density Estimation

Goal: We are interested in estimation of a density function f using observations of random variables $Y_1, \ldots, Y_n$ sampled independently from f.

we will focus on univariate density estimation, but multivariate also exist.

In EDA, estimate of density can be used to assess multimodality, skew, tail behavior, etc.

Useful for summarizing posteriors in Bayesian analyses and as a presentation tool.

Useful for understand sampling dsn of statistics (i.e. in bootstrap).

Parametric Solution:

Begin by assuming a parametric model $Y_1, \ldots, Y_n \overset{iid}{\sim} f_{Y|\underline{\theta}}$

Parameter estimates $\hat{\underline{\theta}}$ are found (e.g. MLE, MoM, Bayesian)

The resulting density estimate at y is $f_{Y|\underline{\theta}}(y \mid \hat{\underline{\theta}})$.

Danger: Relying on an incorrect model $f_{Y|\underline{\theta}}$ can lead to serious errors, regardless of estimation strategy.

We will focus on **nonparametric** approaches to density estimation.

assume very little about the form of f.

predominantly use <u>local</u> information to estimate f at a point y.

1 Histograms

→ piecewise constant density estimator.

One familiar density estimator is a histogram. Histograms are produced automatically by most software packages and are used so routinely to visualize densities that we rarely talk about their underlying complexity.

We will remedy this today!

1.1 Motivation

Recall the definition of a density function

$$f(y) \equiv \frac{d}{dy}F(y) \equiv \lim_{h\to 0}\frac{F(y+h)-F(y-h)}{2h} = \lim_{h\to 0}\frac{F(y+h)-F(y)}{h},$$

where $F(x)$ is the cdf of the random variable Y.

Now, let $Y_1, \ldots, Y_n$ be a random sample of size n from the density f.

Empirical cdf $\hat{F}_n(y) = \dfrac{\sum_{i=1}^{n} \mathbb{I}(Y_i \le y)}{n} = \dfrac{\#\{Y_i \le y\}}{n}$

A natural finite-sample analog of $f(y)$ is to divide the support of Y into a set of K equi-sized bins with small width h and replace $F(x)$ with the empirical cdf.

This leads to $\hat{f}(x) = \dfrac{1}{h}\left\{\dfrac{\#\{Y_i \le b_{j+1}\} - \#\{Y_i \le b_j\}}{n}\right\}$

$= \dfrac{1}{h}\left\{\hat{F}_n(b_{j+1}) - \hat{F}(b_j)\right\}$ where $(b_j, b_{j+1}]$ defines the boundaries of the j^{th} bin.

equivalently $\hat{f}(x) = \dfrac{n_j}{n \cdot h}$ where $n_j = \#$ observations in j^{th} bin

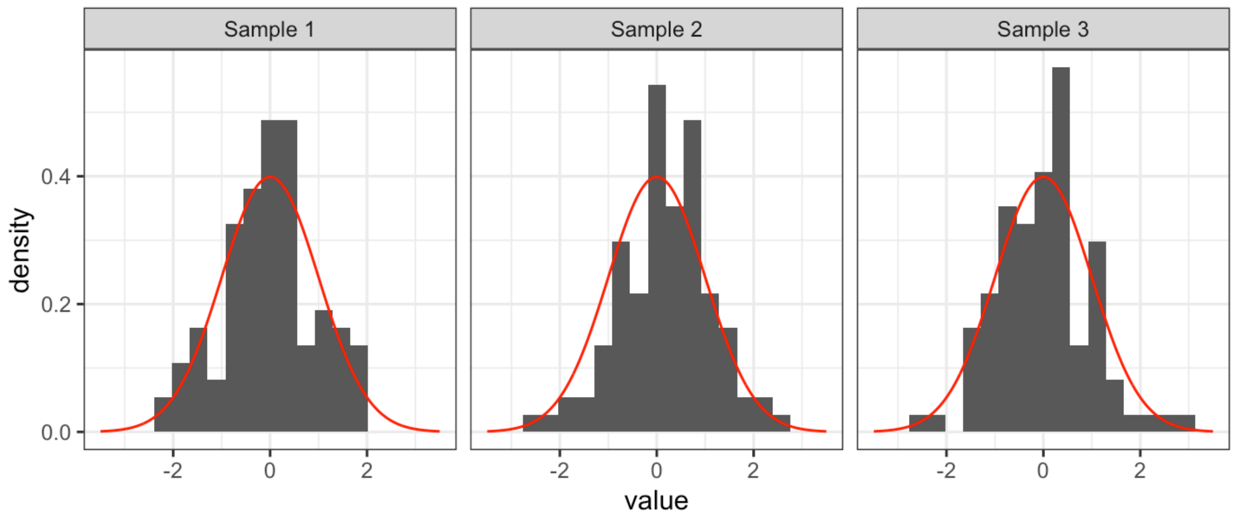
$h = b_{j+1} - b_j$ (width of the bin).

1.2 Bin Width

— bin width choice h is crucial in construction of histograms.

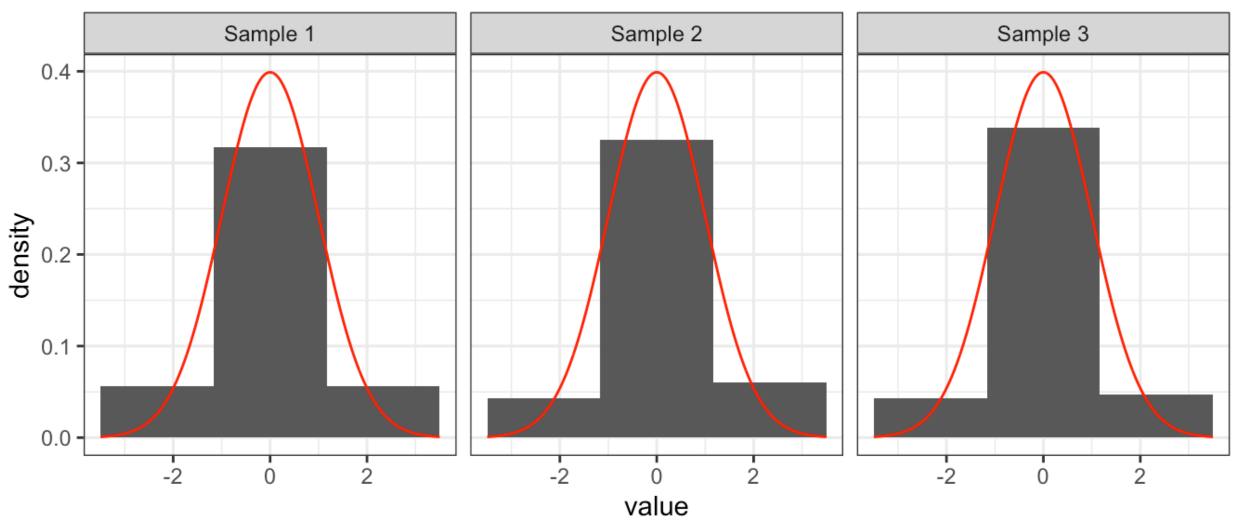
3 sets of simple random samples of size 100 taken from N(0,1).

20 bins.



high variability.

4 bins.



high bias!

Top row: undersmoothing. Histograms vary greatly between samples ⇒ we are seeing small biases and high variance in the estimator

↳ the histogram is the statistic!

Bottom row: oversmoothing. Histograms are stable, but don't follow the density very well

⇒ low variance but high bias.