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.

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 posterior and as a presentation tool.

Also useful in some simulation and MCMC algorithms.

Parametric Solution:

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

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

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

Danger: Relying on an incorrect model $f_{Y|\underline{\theta}}$ can lead to serious inferential 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!

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

how to estimate f w/ data

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 $\hat{f}(x) = \frac{1}{h}\left\{\hat{F}_n(b_{j+1}) - \hat{F}_n(b_j)\right\}$

$= \frac{1}{h}\left\{\frac{\#\{Y_i \le b_{j+1}\} - \#\{Y_i \le b_j\}}{n}\right\}$ where $(b_j, b_{j+1}]$ defines the boundaries of the jth bin.

equivalently, $\hat{f}(x) = \frac{n_j}{n \cdot h}$ where n_j = # observations in jth bin

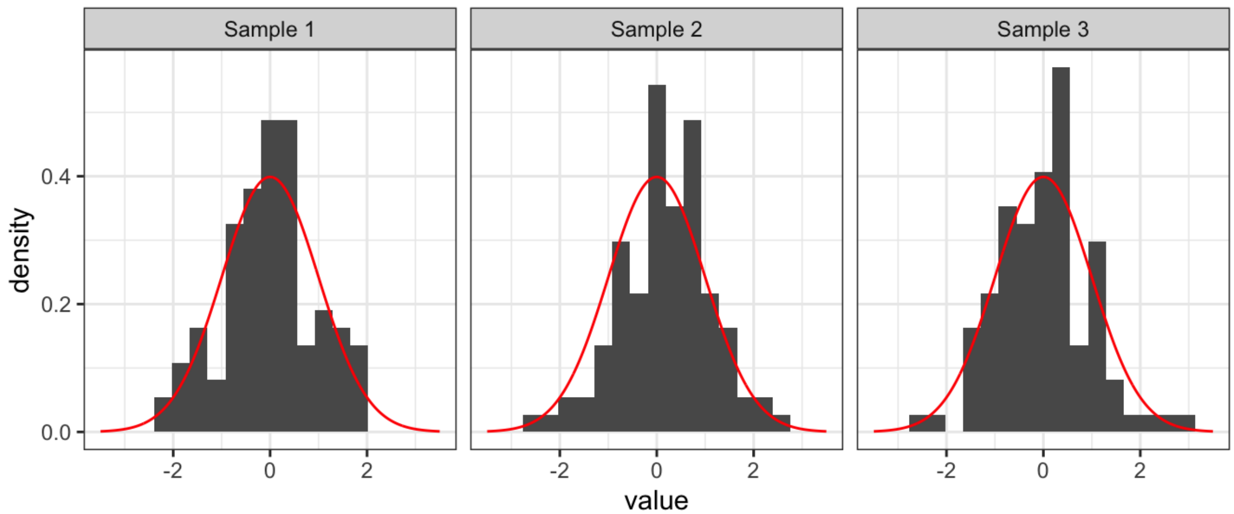
$h = b_{j+1} - b_j$ (length of bin).

bin width choice h is crucial to construction of histograms.

1.2 Bin Width

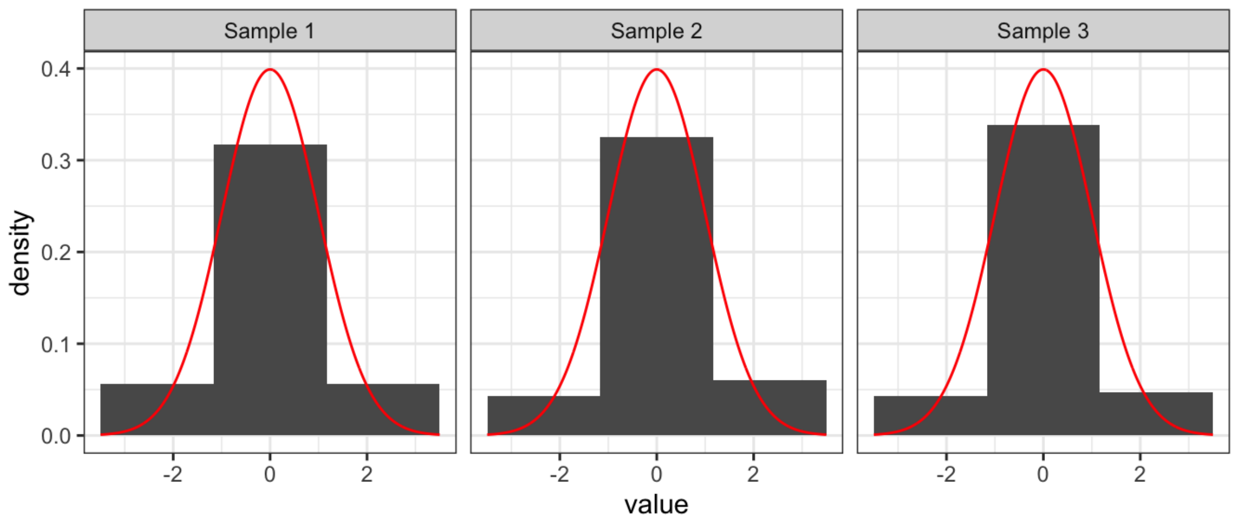
3 random samples of size 100 taken from $N(0,1)$.

20 bins



high variability.

4 bins.



high bias!

Top row: under smoothing. Histograms vary greatly between samples $\Rightarrow$ high variability, low bias.

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

$\Rightarrow$ low variability, high bias.

1.3 Measures of Performance

Squared Error

$$SE_h\left(\hat{f}(y)\right) = \left[\hat{f}_h(y) - f(y)\right]^2$$

local: at a point y

depends on realization $Y_1, \dots, Y_n$ (through $\hat{f}$).

Mean Squared Error

$$MSE\left(\hat{f}(y)\right) = E_f\left[\hat{f}(y) - f(y)\right]^2 = Var\left(\hat{f}(y)\right) + \left[bias\left(\hat{f}(y)\right)\right]^2$$

local at y

but now describes a property of the dsn (mean) of error.

Integrated Squared Error

$$ISE = \int_{-\infty}^{\infty}\left[\hat{f}(u) - f(u)\right]^2 du$$

No longer local

depends on realization.

Mean Integrated Squared Error

$$MISE = \int_{-\infty}^{\infty} MSE(\hat{f}(u))\,du = \int_{-\infty}^{\infty} Var(\hat{f}(u))\,du + \int_{-\infty}^{\infty}\left[bias\left(\hat{f}(y)\right)\right]^2 du.$$

Not local

describes property of dsn of error.

this is what we are estimating!

Of course these are all theoretical because we have to know f to calculate

↳ useful for discussing properties of $\hat{f}$

1.4 Optimal Binwidth

$\hat{f}(y) = \frac{n_j}{n \cdot h}$ for $y \in (b_j, b_{j+1}]$ and $h = b_{j+1} - b_j$. (n_j = count of points in $(b_j, b_{j+1}]$)

We will investigate bias and variance of $\hat{f}$ pointwise, because
$\text{MSE}(y) = (\text{bias}(\hat{f}(y)))^2 + \text{Var}\hat{f}(y)$.

$n_j \sim \text{Binomial}(n, p_j)$, where $p_j = P(b_j < Y \le b_{j+1}) = \int_{b_j}^{b_{j+1}} f(y)\,dy$ (if density exists)

$\Rightarrow E[\hat{f}(y)] = \frac{n \cdot p_j}{n \cdot h} = \frac{p_j}{h} \Rightarrow \text{bias}(\hat{f}(y)) = \frac{p_j}{h} - f(y)$

$\text{Var}[\hat{f}(y)] = \frac{1}{n^2 h^2} n p_j (1-p_j) = \frac{1}{nh^2} p_j(1-p_j)$

Assumption: Let's suppose $f(y)$ is Lipschitz continuous over the interval $B_j = (b_j, b_{j+1}]$, i.e. $\exists$ a constant γ_j s.t. $|f(x) - f(y)| < \gamma_j |x-y|$ $\forall x, y \in B_j$.

Then by MVT, $p_j = \int_{B_j} f(y)\,dy = h f(\xi_j)$ for some $\xi_j \in B_j$.

$\Rightarrow \text{Var}[\hat{f}(y)] = \frac{p_j(1-p_j)}{nh^2} \overset{p_j \in [0,1]}{\le} \frac{p_j}{nh^2} = \frac{\boxed{f(\xi_j)}}{nh}$ for some $\xi_j \in B_j$. as $h \to 0$, increases; as $n \to \infty$, decreases.

and $|\text{Bias}\, \hat{f}(y)| = \left| \frac{p_j}{h} - f(y) \right| = |f(\xi_j) - f(y)| \le \gamma_j |\xi_j - y| \le \gamma_j h$ (decreases as $h \to 0$, unaffected by n.)

So if f is Lipschitz continuous, $\text{MSE}(\hat{f}(y)) = [\text{bias}\,\hat{f}(y)]^2 + \text{Var}\,\hat{f}(y) \le \gamma_j^2 h^2 + \frac{f(\xi_j)}{nh} \equiv M$.

$\Rightarrow$ If as $n \to \infty$, $h \to 0$ and $nh \to \infty$ then $\hat{f}(y)$ is mean square consistent ($\lim_{n\to\infty} \text{MSE}\,\hat{f}(y) = 0$).

optimal bin width based on MSE

$\frac{\partial M}{\partial h} = -\frac{f(\xi_j)}{nh^2} + 2\gamma_j^2 h \overset{\text{set}}{=} 0$

$2\gamma_j^2 h^3 = \frac{f(\xi_j)}{n} \Rightarrow h = \left[\frac{f(\xi_j)}{2\gamma_j^2 n}\right]^{1/3} \Rightarrow$ optimal bin width decreases at a rate proportional to $n^{-1/3}$.

optimal $\text{MSE}[\hat{f}(x)] = \frac{f(\xi_j)}{n \underbrace{[\alpha n^{-1/3}]}_{\text{optimal } h}} + \gamma_j^2 (\alpha n^{-1/3})^2 = K n^{-2/3}$. MSE is not rate n^{-1} (parametric estimation), $\underbrace{\text{but instead } n^{-2/3}}_{\text{cost of being non parametric.}}$

Global histogram error: Consider integrated bias & variance separately.

$IV = \int_{-\infty}^{\infty} \text{Var}\,\hat{f}(y)\,dy = \sum_j \int_{B_j} \text{Var}\,\hat{f}(y)\,dy = \sum_j \int_{B_j} \frac{p_j(1-p_j)}{nh^2}\,dy = \sum_j \frac{p_j(1-p_j)}{nh} = \frac{1}{nh}\left[\underbrace{\sum_j p_j}_{=1} - \sum_j p_j^2\right]$

$\sum_j p_j^2 \overset{MVT}{=} \sum_j f^2(\xi_j) h^2 = h \sum f^2(\xi_j) h = h\left[\int f^2(y)\,dy + o(1)\right]$

So, $IV = \frac{1}{nh}\left(1 - h\int f^2(y)\,dy + h\,o(1)\right) = \frac{1}{nh} - \frac{R(f)}{n} + o(n^{-1})$ where $R(f) := \int f^2(y)\,dy$ "roughness"

Consider a typical bin $B_0 = [0, h]$.

The bin probability $p_0 = \int_0^h f(t)\,dt = \int_0^h \left[f(x) + (t-x)f'(x) + \frac{(t-x)^2}{2} f''(x) + \dots\right] dt$

$= \left[t f(x) + \frac{(t-x)^2}{2} f'(x) + \frac{(t-x)^3}{2\cdot 3} f''(x) + \dots\right]_0^h$

$= h f(x) + \left[\frac{(h-x)^2}{2} - \frac{x^2}{2}\right] f'(x) + O(h^3) = hf(x) + h\left(\frac{h}{2} - x\right) f'(x) + O(h^3)$.

$\Rightarrow$ bias at a point $x \in B_0$ is $\text{Bias}(\hat{f}(x)) = \frac{p_0}{h} - f(x) = \left(\frac{h}{2} - x\right) f'(x) + O(h^2)$.

Integrated squared bias for bin B_0

$ISB_0 \approx \int_{B_0} \left(\frac{h}{2} - x\right)^2 f'(x)^2\,dx \underset{\text{generalized MVT}}{=} f'(\eta_0)^2 \int_0^h \left(\frac{h}{2} - x\right)^2 dx = \frac{h^3}{12}[f'(\eta_0)]^2$

Total Integrated squared bias:

$$ISB \approx \frac{h^3}{12}\sum_{\text{all } j}[f'(\eta_j)]^2 = \frac{h^2}{12}\sum_{\text{all } j}[f'(\eta_j)]^2 h$$

$$= \frac{h^2}{12}\left[\int [f'(x)]^2 dx + o(1)\right]$$

$$= \frac{h^2}{12}\underbrace{\int [f'(x)]^2 dx}_{R(f')} + o(h^2)$$

$$\Rightarrow MISE = IV + ISB = \frac{1}{nh} - \frac{R(f)}{n} + O(n^{-1}) + \frac{1}{12}h^2 R(f') + O(h^2)$$

$$= \underbrace{\frac{1}{nh} + \frac{h^2 R(f')}{12}}_{AMISE} + o(n^{-1}) + o(h^2)$$

"asymptotic"

narrower bins give an estimator that is less biased but more variable. As $h \to 0$, $\hat{f} \to$ set of spikes at each observation (0 bias).

The minimizer of AMISE is $h_0 = \left[\frac{6}{R(f')}\right]^{1/3} n^{-1/3}$

and minimum AMISE is $AMISE_0 = \left[\frac{9R(f')}{16}\right]^{1/3} n^{-2/3}$.

The roughness of the underlying density, as measured by $R(f')$ determines the optimal level of smoothing and the accuracy of the histogram estimate.

Densities w/ few bumps (smaller $R(f')$) and require wider bins

Bumpy densities (larger $R(f')$) require smaller bins.

We cannot find the optimal binwidth without known the density f itself.

this is what we are estimating!!

Simple (plug-in) approach: Assume f is a $N(\mu, \sigma^2)$, then

$$h_0 = 3.491\, \sigma\, n^{-1/3}$$

could use sample st. dev or interquartile range to estimate

For non-normal data, multiple modes inflate $\hat{\sigma}^2$ $\Rightarrow$ Gaussian based histogram will be oversmoothed.

No theoretical justification, just something we can do and often passes the "eye test".

"cross-validation"

Data driven approach:

$$\begin{aligned} ISE &= \int [f(u) - \hat{f}(u)]^2 \, du \\ &= R(f) + \underbrace{R(\hat{f})}_{\text{computed in closed form}} - 2\int \hat{f}(u) f(u) \, du. \end{aligned}$$

irrelevant for choosing h

this is not wrt to data sample we have

$$-2\int \hat{f}(u) f(u) \, du = \underbrace{-2\, E[\hat{f}(U)]}, \quad U \sim f$$

2 Frequency Polygon

The histogram is simple, useful and piecewise constant.

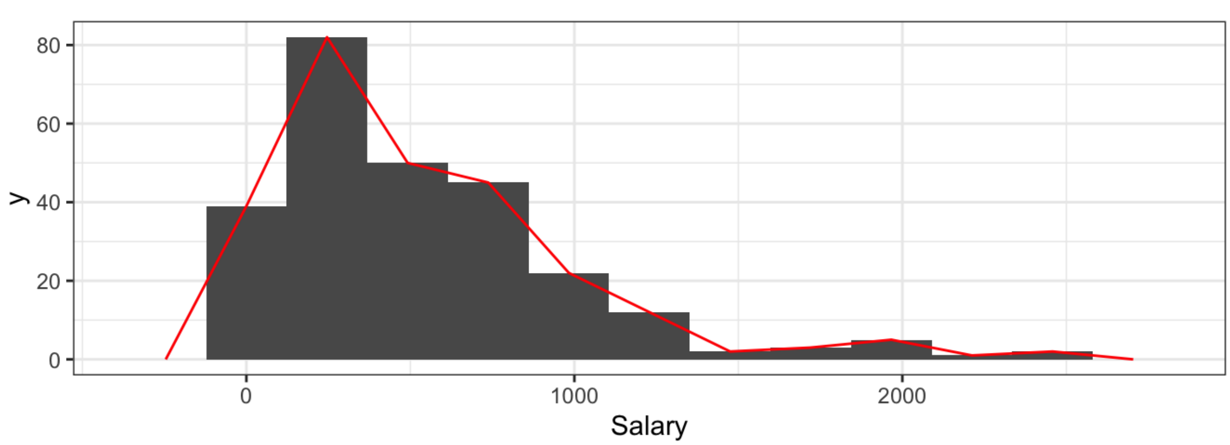
```
library(ISLR)

# optimal h based on normal method
h_0 <- 3.491 * sd(Hitters$Salary, na.rm = TRUE) *
        sum(!is.na(Hitters$Salary))^(-1/3)

## original histogram with optimal h
ggplot(Hitters) +
  geom_histogram(aes(Salary), binwidth = h_0) -> p

## get values to build freq polygon
vals <- ggplot_build(p)$data[[1]]
poly_dat <- data.frame(x = c(vals$x[1] - h_0,
                             vals$x, vals$x[nrow(vals)] + h_0),
                       y = c(0, vals$y, 0))

## plot freq polygon
p + geom_line(aes(x, y), data = poly_dat, colour = "red")
```



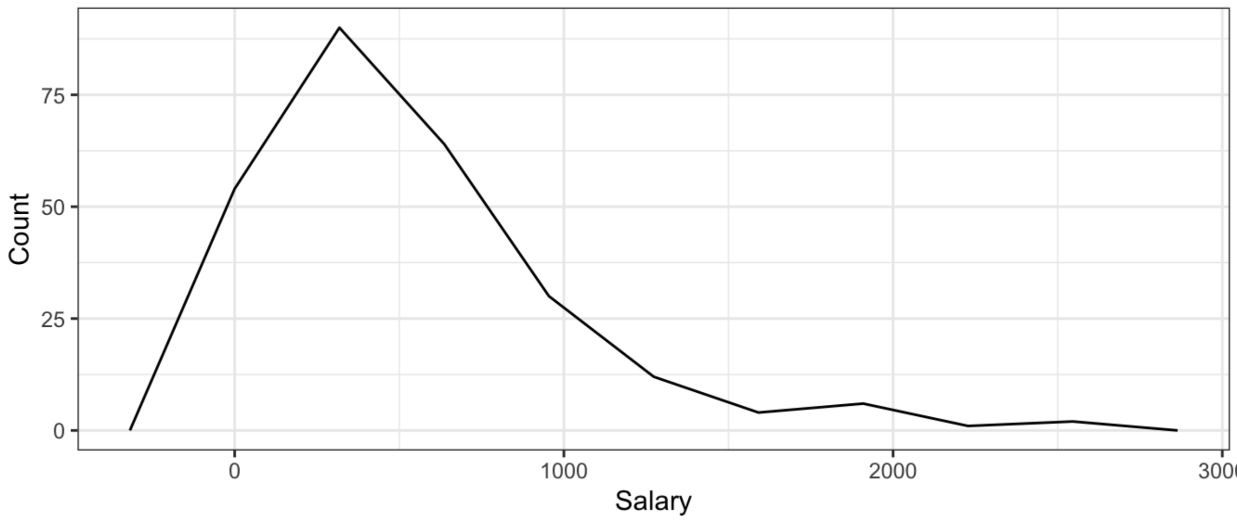
Let $b_1, \ldots, b_{K+1}$ represent bin edges of bins with width h and $n_1, \ldots, n_K$ be the number of observations falling into the bins. Let $c_0, \ldots, c_{k+1}$ be the midpoints of the bin interval.

The frequency polygon is defined as

MISE

AMISE

Gaussian rule for binwidth


Count
75
50
25
0
0
1000
2000
3000
Salary

In practice, a simple way to construct locally varying binwidth histograms is by transforming the data to a different scale and then smoothing the transformed data. The final estimate is formed by simply transforming the constructed bin edges $\{b_j\}$ back to the original scale.