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 $\hat{\underline{\theta}}$ are found (e.g. MLE, EM, 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 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 local 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 $F(y)$ 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

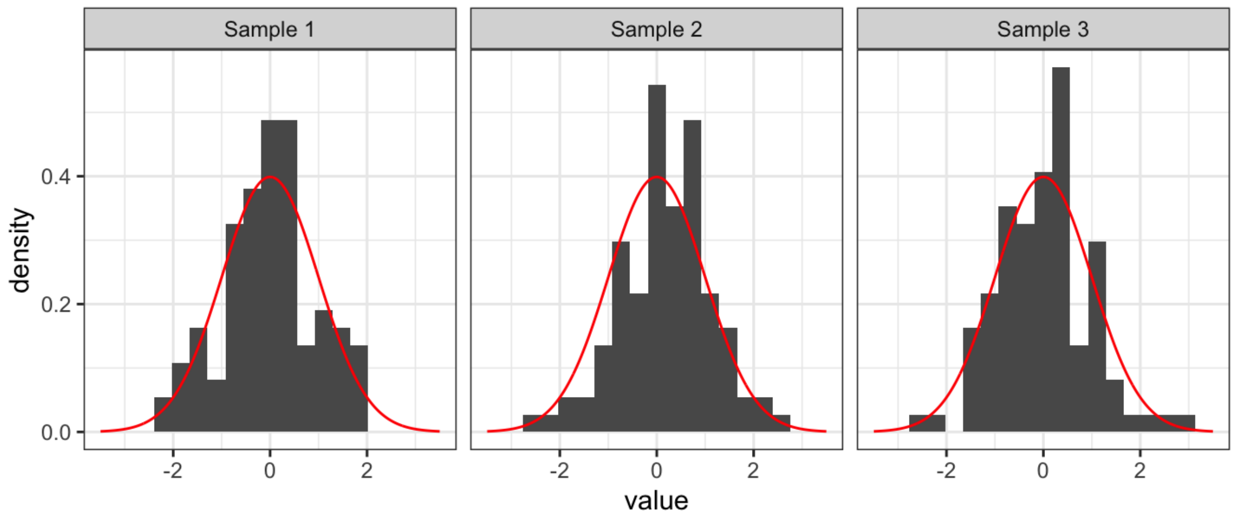
$h = b_{j+1} - b_j$ (length of bin).

1.2 Bin Width

bin width choice h is crucial to construction of histograms.

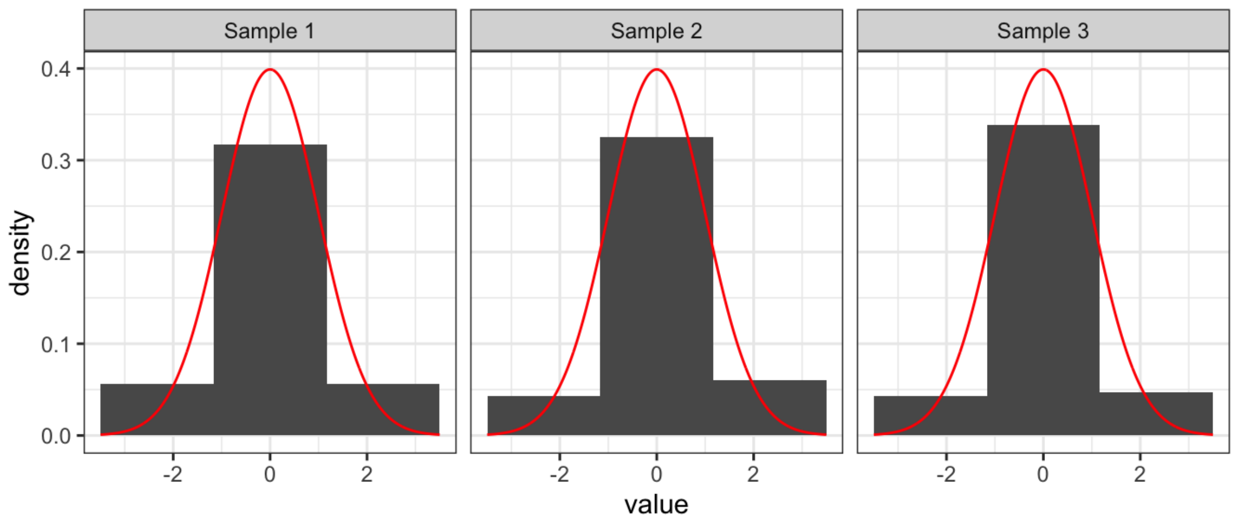
3 random samples of size 100 taken from $N(0,1)$.

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

n_j = count of points in $(b_j, b_{j+1}]$

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

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 $\xi_j \in B_j$.

$p_j \in [0,1]$

$\Rightarrow \text{Var}[\hat f(y)] = \frac{p_j(1-p_j)}{nh^2} \le \frac{p_j}{nh^2} = \frac{\boxed{f(\xi_j)}}{nh}$ 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

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

$$2\gamma_j^2 h^3 = \frac{f(\xi_j)}{n} \Rightarrow h = \left[\frac{f(\xi_j)}{2\gamma_j^2 n}\right]^{1/3}$$

$\Rightarrow$ 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), but instead $\underbrace{n^{-2/3}}_{\text{cost of being non-parametric.}}$

Global histogram error: Consider integrated bias & variance separately.

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

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

"roughness"

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 $R(f) := \int f^2(y)\,dy$

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

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

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

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

$\Rightarrow$ 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

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

Total Integrated squared bias:

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

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

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

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

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

"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:

$$ISE_h = \int [f(u) - \hat{f}_h(u)]^2 du$$

$$= \underset{\text{irrelevant for choosing } h}{R(f)} + \underbrace{R(\hat{f})}_{\text{computed in closed form}} - 2\int \hat{f}(u) f(u)\, du.$$

$$-2\int \hat{f}(u) f(u)\, du = \underbrace{-2\,E[\hat{f}(U)]}_{\text{this is not wrt to data sample we have}},\ U \sim f$$

$\Rightarrow$ one idea is to estimate d ISE so that we can find an h that would minimize it

Dropping the i^{th} case from $\{Y_1, \dots, Y_n\}$ leads to an estimate $\hat{f}_{-i}(Y_i)$. ← histogram of $n-1$ points, leaving out Y_i

$\Rightarrow$ we can obtain

$$CV(h) = -\frac{2}{n}\sum_{i=1}^{n} \hat{f}_{-i}(Y_i)$$

evaluate this for many values of h, choose the h that minimize $CV(h)$.

Why use ISE, not MISE? We don't know f and we are trying to come up with something we can actually do w/ what we know (practical)

Does this work?

Unforturnately often leads to undersmoothing (high variance) in histogram.

2 Frequency Polygon

connect midpoint of histogram w/ straight lines.

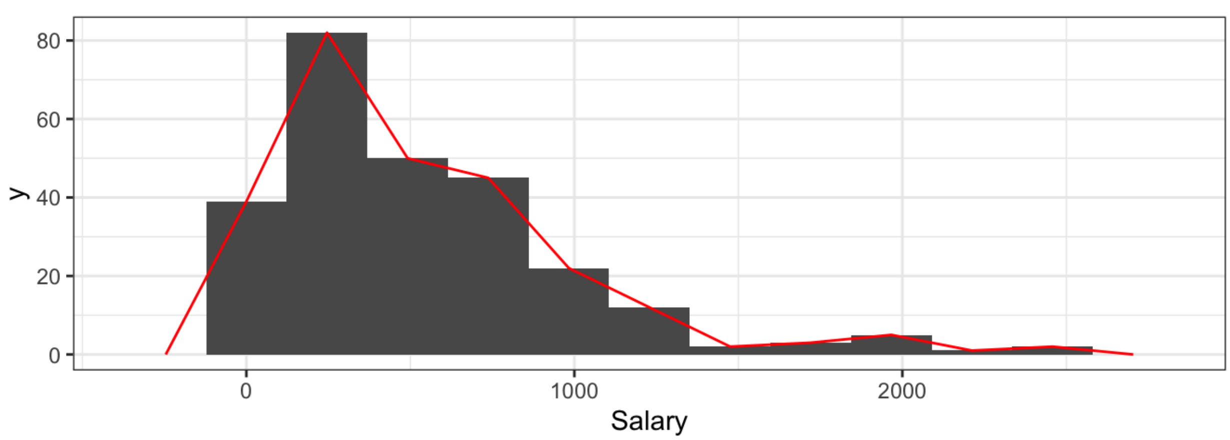
The histogram is simple, useful and piecewise constant.

For continuous RV, need something smoother.

```
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")
```

optimal binwidth for histogram w/ plug-in method.

pull out bin information



Major League Baseball salaries (in 1000s of dollars) in 1986 and 1987.

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.

$$c_j = \frac{(b_j + b_{j+1})}{2} \qquad j = 1, \ldots, K$$

$$c_0 = b_1 - \frac{h}{2}, \quad c_{K+1} = b_{K+1} + \frac{h}{2}$$

The frequency polygon is defined as

$$\hat{f}(x) = \frac{1}{nh^2}\left[n_j c_{j+1} - n_{j+1} c_j + (n_{j+1} - n_j)\, x\right] \text{ for } x \in [c_j, c_{j+1}].$$

If f'' is absolutely continuous, and $R(f), R(f'), R(f''), R(f''')$ all finite.

$$\text{MISE} = \underbrace{\frac{2}{3nh}}_{\text{as } h\downarrow\text{, this increases}}^{\text{var}} + \underbrace{\frac{49h^4 R(f'')}{2880}}_{\text{as } h\downarrow\text{, this decreases.}}^{\text{bias}^2} + O(n^{-1}) + O(h^6).$$

compare to MISE for histogram: $\frac{1}{nh} + \frac{h^2 R(f')}{12} + O(n^{-1}) + O(h^3)$.

$$\text{AMISE} \quad \frac{2}{3nh} + \frac{49h^4 R(f'')}{2880}$$

minimization of AMISE results in $h_0 = 2\left[\frac{15}{49R(f'')}\right]^{1/5} n^{-1/5}$

results in minimal AMISE: $\text{AMISE}_0 = \frac{5}{12}\left[\frac{49R(f'')}{15}\right]^{1/5} \underbrace{n^{-4/5}}$ improved convergence rate (vs. histogram $n^{-2/3}$)

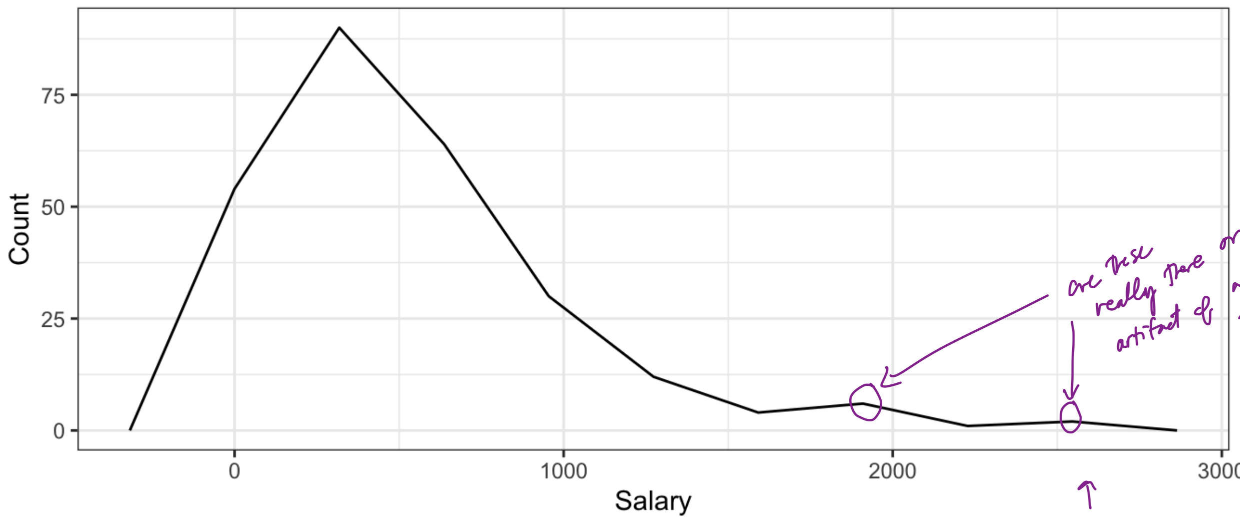
Optimal binwidth for freq polygon will be asymptotically larger than histogram.

Gaussian rule for binwidth

Again we don't know $R(f'') \Rightarrow$ Assume f is Gaussian and plug in:

$$h_0 = 2.15\,\sigma\, n^{-1/5}$$

mode of salary at ≈ 400k



are these really there or artifact of noise in data?

right skew

Let's return to the histogram (local error) for a moment:

$$MSE(\hat{f}(x)) = Var(\hat{f}(x)) + Bias(\hat{f}(x))^2$$

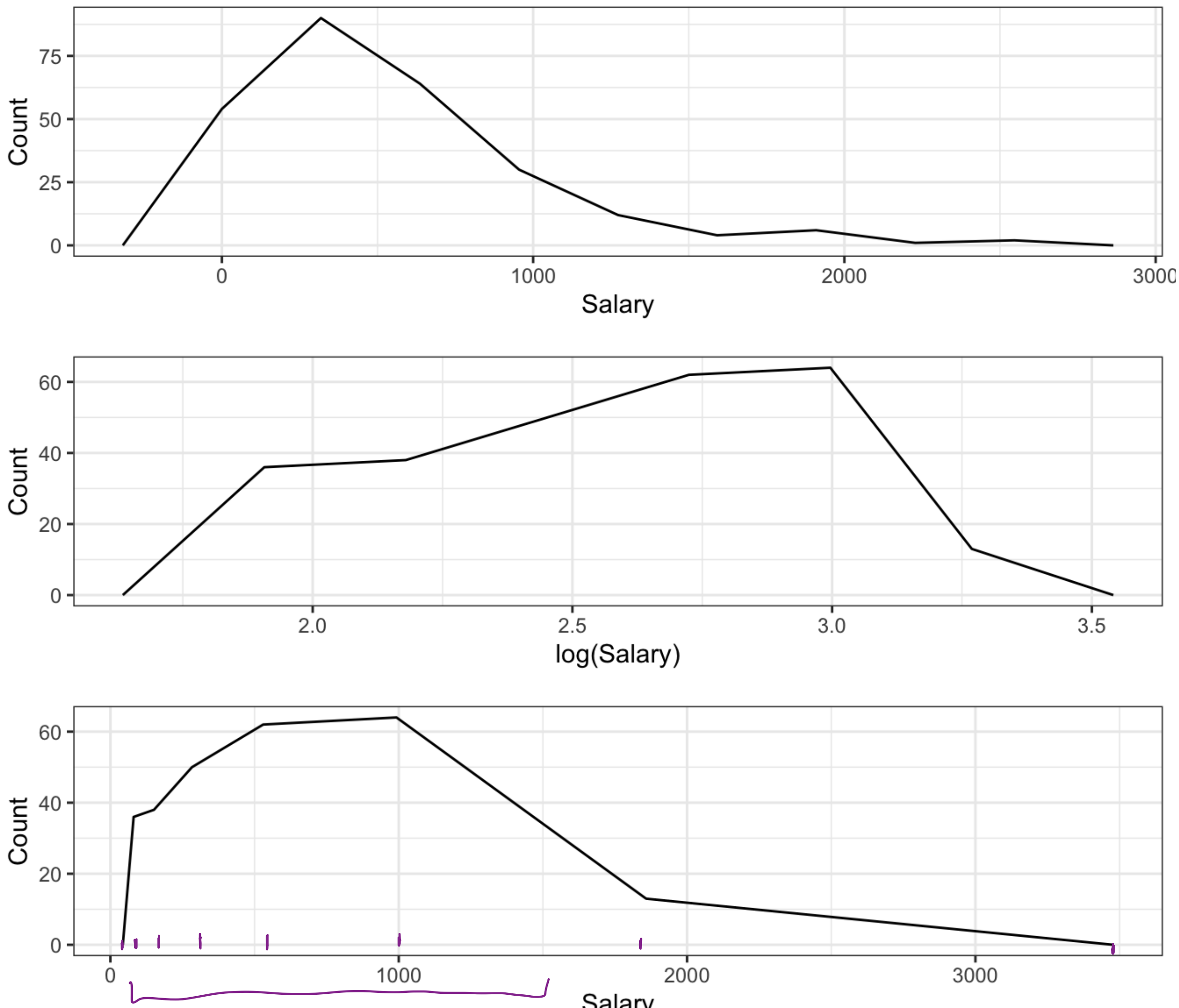
$$= \frac{f(x)}{nh} + \frac{f'(x)^2}{4}\left[h - 2(x - b_j)\right]^2 + o(n^{-1}) + O(h^3) \quad \text{for } x \in (b_j, b_{j+1}].$$

binwidth should be larger in regions w/ more data to minimize 1st term. (variance)
binwidth should be inversely related to $|f'(x)|$ to minimize 2nd term (bias).

$\Rightarrow$ a histogram (or freq. polygon) w/ locally varying bin width could be more accurate than fixed binwidth.

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. how about $\hat{f}(x)$?

Again, Salary information for MLB 1986 & 1987



Fixed Bin

Fixed bin on $\log_{10}$ scale

more detailed information where we have more data.

after back transforming, have varying bin widths.