# 2.1 Histograms

Saturday, February 5, 2022 11:18 AM

https://bookdown.org/egarpor/NP-UC3M/kde-i-hist.html

## 2.1 Histograms

Histograms

- estimate density $f$ from window of $[x_0, x_0+h)$

$x \in [x_0, x_0+h)$:

$$f(x_0) = F'(x_0) = \lim_{h\to 0^+} \frac{F(x_0+h) - F(x_0)}{h}$$

$$= \lim_{h\to 0^+} \frac{\mathbb{P}[x_0 < X < x_0+h]}{h}$$

- Bins:
  - $B_k := [t_k, t_{k+1})$ : $t_k = t_0 + hk$, $k \in \mathbb{Z}$
  - samp pts
  - bandwidth $h$
- hist defined as:

$$\hat{f}_H(x; t_0, h) := \frac{1}{nh} \sum_{i=1}^{n} 1_{\{X_i \in B_k : x \in B_k\}}$$

$$= \frac{v_k}{nh} \qquad v_k := \#\text{ obs in bin } B_k$$

- analysis of $\hat{f}_H(x; t_0, h)$ as RV:

$v_k \sim B(n, p_k)$ $\qquad p_k := \mathbb{P}[X \in B_k] = \int_{B_k} f(t)\,dt$

$n$: # trials, $p_k$: prob(succ)

→ if $f$ cont, then by MVT:

$p_k = h f(\xi_{k,h})$ (so $p_k = \frac{v_k}{n}$)

in $\xi_{k,h} \in (t_k, t_{k+1})$

$$\mathbb{E}[\hat{f}_H(x; t_0, h)] = \frac{np_k}{nh} = f(\xi_{k,h})$$

→ $\mathbb{E}[\hat{f}_H] = \frac{1}{nh}\mathbb{E}[v_k] = \frac{np_k}{nh}$ b/c mean of binomial

$$\mathrm{Var}[\hat{f}_H(x; t_0, h)] = \frac{np_k(1-p_k)}{n^2h^2} = \frac{f(\xi_{k,h})(1 - hf(\xi_{k,h}))}{nh}$$

→ $\mathrm{Var}[\hat{f}_H] = \frac{1}{n^2h^2}\mathrm{Var}[v_k] = \frac{1}{n^2h^2}(np(1-p))$

Observations:

1) if $h \to 0$, then $\xi_{k,h} \to x \Rightarrow f(\xi_{k,h}) \to f(x)$

Note

↳ $x \in [t_k, t_{k+1}]$, $k$ ↓ as $h \to 0$

& interval collapses to $x$

↳ unbiased estimator of $f(x)$ asymptotically when $h \to 0$ (small bins)

2) if $h \to 0$, ↑var

↳ $nh \to \infty$ to decrease

(more points $n$)

3) $f(\xi_{k,h})(1 - hf(\xi_{k,h})) \to f(x)$ as $h \to 0$

↳ more variability where higher density

↳ more variability where higher density

- $t_0$ matters!

## Moving Histogram "Naive Density Estimator"

- Goal: aggregate $X_1, \dots, X_n$ in intervals $(x-h, x+h)$

$$f(x) = F'(x) = \lim_{h \to 0^+} \frac{F(x+h) - F(x-h)}{2h}$$

$$= \lim_{h \to 0^+} \frac{\mathbb{P}[x-h < X < x+h]}{2h}$$

→ symmetric derivative

↳ intervals based on eval point $x$ and are centered around it
⇒ directly estimates $f(x)$ w/o proxy $f(x_0)$

- Here, we have:

$$\hat{f}_n(x;h) := \frac{1}{2nh} \sum_{i=1}^{n} 1_{\{x-h < X_i < x+h\}}$$

- Analysis

$$\sum 1 \sim B(n, p_{x,h})$$

$$p_{x,h} := \mathbb{P}[x-h < X < x+h] = F(x+h) - F(x-h)$$

$$\mathbb{E}[\hat{f}_n(x;h)] = \frac{F(x+h) - F(x-h)}{2h}$$

↳ $1/2nh\,(np)$

$$\mathbb{V}ar[\hat{f}_n(x;h)] = \frac{F(x+h) - F(x-h)}{4nh^2} - \frac{[F(x+h) - F(x-h)]^2}{4nh^2}$$

↳ $Var(X) = E[(X - E(X))^2]$ + mechanics

Observations:

1) if $h \to 0$, then
   1.a) $\mathbb{E}[\hat{f}_n(x;h)] \to f(x)$
   1.b) $\mathbb{V}ar[\hat{f}_n(x;h)] \approx \frac{f(x)}{2nh} - \frac{f(x)^2}{n} \to \infty$
2) if $h \to \infty$
   2.a) $\mathbb{E}[\hat{f}_n(x;h)] \to 0$
   2.b) $\mathbb{V}ar[\hat{f}_n(x;h)] \to 0$
3) $var \to 0$ if $nh \to \infty$

Consider – should we weight points closer to $x$ more heavily?