Unit 5: Probability Density Estimation

Chapter 12 in "Statistical Computing with R"

Anna Ly

Department of Mathematical and Computational Sciences
University of Toronto Mississauga

January 19, 2026

Overview

Introduction... To Histograms!

Introduction

- Density estimation is a collection of methods for constructing an estimate of a probability density function of an observed sample of data.
- A histogram is a type of density estimator.
- There are many methods for estimation as outlined in the textbook; due to the interest of time, we will cover three basic univariate methods.
- These are not required for the term tests and examination, but there will be a quiz dedicated to this topic on the high-level topics (the discussion). The formulas will be on the formula sheet in case you're paranoid.

The Histogram

- The probability histogram is the most widely used density estimate in descriptive statistics. (After all, who taught you anything else?)
- We ought to talk about these questions more:
  - How to determine the number of bins?
  - What should the boundaries and width of class intervals be?
  - How do we handle unequal class interval widths (should we even have them)?
- Fortunately, can R automatically make this choice for you! But if you're picky, these results can be undesirable, and you can fine-tune this!
- Data is usually contaminated by noise; a narrow bin may undersmooth the data, presenting too many details, but wider bins may oversmooth the data, obscuring important features.
- We will discuss rules to suggest an optimal choice of bin width.

The Histogram

Example

Suppose $X_1 \sim N(-2, 1)$, $X_2 \sim N(2, 1)$, and $X_1 \perp\!\!\!\perp X_2$. Simulating the following using R:

$$F_X(x) := 0.5F_{X_1}(x_1) + 0.5F_{X_2}(x_2)$$

Create three histograms; the first one shall have 10 bins, the second 25, and the third 50. What patterns do you notice?

Solution.

The Histogram

Histogram Density Estimate

Given class intervals of equal width h, the histogram density estimate based on a sample size n is:

$$\hat{f}(x) = \frac{\nu_k}{nh}, \quad t_k \leq x < t_{k+1},$$

Where $\nu_k$is the number of sample points in the class interval $[t_k, t_{k+1})$.

Sturges' Rule

Sturges' Rule

- Sturges' rule tends to oversmooth the data, and is the default in many statistical packages (including R).
- This choice of class interval is designed for data sampled from symmetric, unimodal populations, but is not a good choice for skewed distributions or distributions with more than one mode.
- Sturges' rule is based on the implicit assumption that the sampled population is normally distributed.
- In this case, it is natural to choose a family of discrete distributions that converge in distribution to normal as the number of classes (and sample size n) tend to infinity.
- The most obvious candidate is the binomial distribution with probability of success $p = 1/2$.

Sturges' Rule

For example, if the sample size is $n = 64$, one could select seven class intervals such that the frequency histogram corresponding to a $Binomial(6, 1/2)$ sample has expected class frequencies:

$$\binom{6}{0}, \binom{6}{1}, \binom{6}{2}, \ldots, \binom{6}{6} = 1, 6, 15, 20, 15, 6, 1$$

These sum to 64.

Sturges' Rule

- Now consider sample sizes $n = 2^k$, $k = 1, 2, \ldots$ If k is large (or n is large), the distribution of $Binomial(n, 1/2)$ is approximately $N(\mu = n/2, \sigma^2 = n/4)$.
- Here, $k = \log_2 n$ and we have $k + 1$ bins with expected class frequencies:

$$\log_2 \binom{n}{j}, \quad j \in \{0, 1, \ldots, k\}.$$

Sturges' Rule

Let R represent the sample range. The optimal **width** of class intervals is given by:

$$\frac{R}{1 + \log_2(n)}$$

Struges' Rule

Example

Generate a random sample from the standard Gaussian distribution. Compute the number of breaks manually in R using Struges' rule, and then compare this to the default settings of hist().

Solution.

Scott's Normal Reference Rule

Scott's Normal Reference Rule

- One approach to select an optimal smoothing parameter for density estimation is to minimize the squared error in the estimate. (Similar to using MSE as a criterion to compare two estimators...)
- Consider the integrated squared error (ISE) which is the L_2 norm:

$$ISE(\hat{f}(x)) = \int (\hat{f}(x) - f(x))^2 dx$$

- Practically, people use the mean integrated squared error (MISE):

$$MISE = \mathbb{E}[ISE]$$

Scott's Normal Reference Rule

MISE = IMSE

Show that:

$$MISE = \mathbb{E}[ISE] = \int MSE[\hat{f}(x)]dx := IMSE$$

Solution.

Scott's Normal Reference Rule

After some painful computation, Scott showed that (don't worry you don't need to prove this):

$$MISE = \frac{1}{nh} + \frac{h^2}{12} \int f'(x)^2 dx + O\left(\frac{1}{n} + h^3\right)$$

They solved for h ignoring the last term; intuitively, the bin width should not get too huge, so $O\left(\frac{1}{n} + h^3\right)$ shouldn't be too large. Let's solve for an optimal value for h:

Scott's Normal Reference Rule

Scott's Normal Reference Rule

Assume the model is Gaussian with variance σ^2, the specified bin width is:

$$\hat{h} = 2 \times 3^{1/3} \times \pi^{1/6} \hat{\sigma} n^{-1/3} \approx 3.49 \hat{\sigma} n^{-1/3}$$

Scott's Normal Reference Rule

Example

Generate a random sample from the standard Gaussian distribution. Compute the number of breaks manually in R using Scott's normal reference rule, and then compare this to setting hist($\cdot$, breaks $=$ "Scott").

Solution.

Freedman-Diaconis Rule

Freedman-Diaconis Rule

- Remember that Scott's method claims a reasonable choice for bin width is:

$$\hat{h} = 3.49\hat{\sigma}n^{-1/3}$$

- However, $\hat{\sigma}$ is only a reasonable choice of the data somewhat follows a symmetric, Gaussian distribution. It may not be that good for skewed data.
- Naturally when doing skewed data analysis it's better to use IQR as a heuristic of spread.
- Freedman-Diaconis Rule argued that the following provides reasonable results, tested on some simulations:

Freedman-Diaconis Rule

The specified bin width is:

$$\hat{h} = 2(IQR)n^{-1/3}$$

Freedman-Diaconis Rule

Example

Generate a random sample from the standard Gaussian distribution. Compute the number of breaks manually in R using the Freedman-Diaconis rule, and then compare this to setting hist($\cdot$, breaks = "FD").

Solution.

What's Testable?

Testable Material

Basically, you need to understand the following concepts for the quiz:

- What is the most common probability density estimator?
- Which method(s) are designed for symmetric distributions?
- Which method(s) make sense for skewed data?
- Which method was inspired by the Binomial distribution?
- Based on the code, what method was used? (If you understand my lecture code, you're good to go. Note that my code differs a lot from the textbook!)