Statistics for Astronomers: Lecture 16, 2020.12.09

Prof. Sundar Srinivasan

IRyA/UNAM

Review

Nonparametric hypothesis testing

Kolmogorov-Smirnov and Anderson-Darling tests. Wilcoxon-Mann-Whitney U-test.

Data visualisation

Box-and-whisker plots

Jupyter demos

1. Download this Jupyter notebook.
2. Navigate to Colaboratory.
3. Sign in
4. Click on "Upload" and upload the notebook you downloaded in step 1.

Histogram

Also non-parametric, generates piecewise constant estimator of underlying density distribution.

Histogram

Also non-parametric, generates piecewise constant estimator of underlying density distribution. Data of size N is placed into M bins of width h such that

$$\hat{f}(x) = \frac{1}{hN}\sum_{i=1}^{N}\sum_{b=1}^{M} \mathbb{I}\left(\frac{|x_i - x_b|}{h} \leq 1\right) \mathbb{I}\left(\frac{|x - x_b|}{h} \leq 1\right)$$

where x_i are the data values, x_b the location of the b^{th} bin, and $\mathbb{I}$ the Indicator Function.

Histogram

Also non-parametric, generates piecewise constant estimator of underlying density distribution. Data of size N is placed into M bins of width h such that

$$\hat{f}(x) = \frac{1}{hN}\sum_{i=1}^{N}\sum_{b=1}^{M} \mathbb{I}\left(\frac{|x_i - x_b|}{h} \leq 1\right) \mathbb{I}\left(\frac{|x - x_b|}{h} \leq 1\right)$$

where x_i are the data values, x_b the location of the b^{th} bin, and $\mathbb{I}$ the Indicator Function.

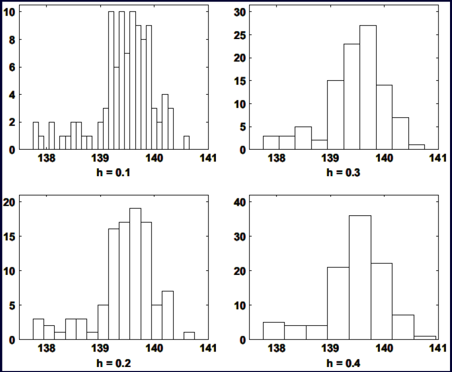
Advantages: easy and quick to compute, does well for large N.

Disadvantages:

Location information for data degraded (location for all points in a bin is now center of bin).

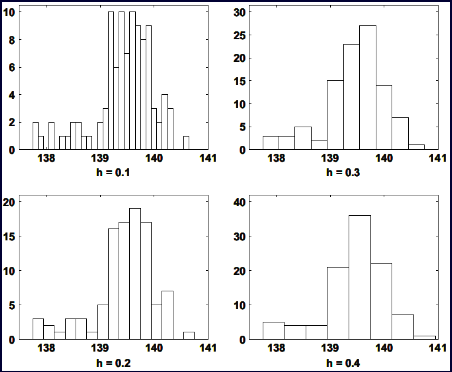
Shape highly sensitive on bin width and bin edges.

Histogram (contd.)

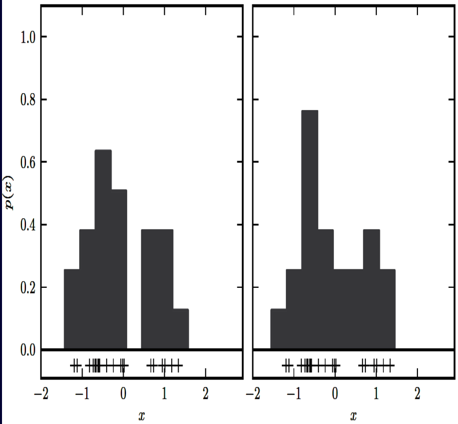


Effect of bin width.
Source: Applied Multivariate Statistical Analysis, Härdle & Simar

Histogram (contd.)



Effect of bin width.
Source: Applied Multivariate Statistical Analysis, Härdle & Simar



Effect of bin location.
Source: AstroML book

Optimal bin width for a histogram

Frequentist methods

Find width that optimises some function of (estimated density − true density).
Requires assumptions about true density.

Optimal bin width for a histogram

Frequentist methods

Find width that optimises some function of (estimated density − true density).
Requires assumptions about true density.

Scott's rule (Scott 1979): $h \approx \mathrm{IQR}\ N^{-1/3}$. Assumes normally-distributed data.

Freedman-Diaconis rule (Freedman & Diaconis 1981): $h = 2\ \mathrm{IQR}\ N^{-1/3}$.
Allows some departure from normality.

Optimal bin width for a histogram

Frequentist methods

Find width that optimises some function of (estimated density − true density).
Requires assumptions about true density.

Scott's rule (Scott 1979): $h \approx \mathrm{IQR}\ N^{-1/3}$. Assumes normally-distributed data.
Freedman-Diaconis rule (Freedman & Diaconis 1981): $h = 2\ \mathrm{IQR}\ N^{-1/3}$.
Allows some departure from normality.

Disadvantage of these methods: not sensitive to multimodal distributions.

Optimal bin width for a histogram

Frequentist methods

Find width that optimises some function of (estimated density − true density).
Requires assumptions about true density.

Scott's rule (Scott 1979): $h \approx \mathrm{IQR}\ N^{-1/3}$. Assumes normally-distributed data.
Freedman-Diaconis rule (Freedman & Diaconis 1981): $h = 2\ \mathrm{IQR}\ N^{-1/3}$.
Allows some departure from normality.

Disadvantage of these methods: not sensitive to multimodal distributions.

Bayesian methods:

No assumptions required about underlying distribution.
Can form data likelihood and assume appropriate priors for the problem.
Bayesian method allows computation of means and standard deviations of bin heights.
Good multimodal/unimodal distinction!

Optimal bin width for a histogram

Frequentist methods

Find width that optimises some function of (estimated density − true density).
Requires assumptions about true density.

Scott's rule (Scott 1979): $h \approx \mathrm{IQR}\ N^{-1/3}$. Assumes normally-distributed data.
Freedman-Diaconis rule (Freedman & Diaconis 1981): $h = 2\ \mathrm{IQR}\ N^{-1/3}$.
Allows some departure from normality.

Disadvantage of these methods: not sensitive to multimodal distributions.

Bayesian methods:

No assumptions required about underlying distribution.
Can form data likelihood and assume appropriate priors for the problem.
Bayesian method allows computation of means and standard deviations of bin heights.
Good multimodal/unimodal distinction!

Knuth (2006) used a multinomial likelihood and Jeffreys priors to find the optimal h.

Optimal bin width for a histogram

Frequentist methods

Find width that optimises some function of (estimated density − true density).
Requires assumptions about true density.

Scott's rule (Scott 1979): $h \approx \mathrm{IQR}\ N^{-1/3}$. Assumes normally-distributed data.
Freedman-Diaconis rule (Freedman & Diaconis 1981): $h = 2\ \mathrm{IQR}\ N^{-1/3}$.
Allows some departure from normality.

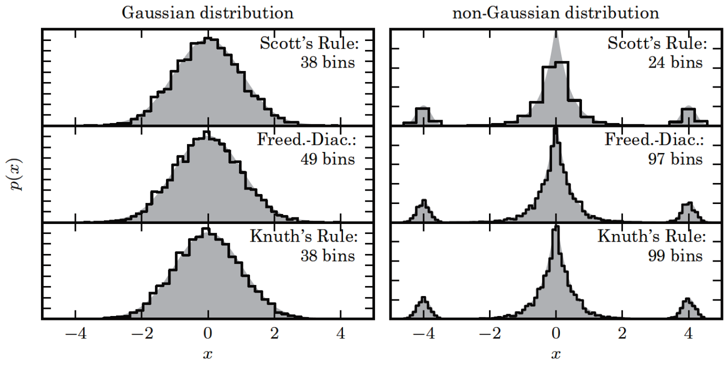
Disadvantage of these methods: not sensitive to multimodal distributions.

Bayesian methods:

No assumptions required about underlying distribution.
Can form data likelihood and assume appropriate priors for the problem.
Bayesian method allows computation of means and standard deviations of bin heights.
Good multimodal/unimodal distinction!

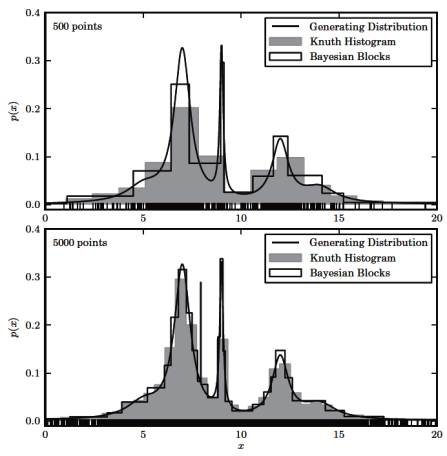
Knuth (2006) used a multinomial likelihood and Jeffreys priors to find the optimal h.
Bayesian Blocks (*e.g.*, Scargle et al. 2013, applied to time-series data):
designs a log-likelihood allowing for varying binsize.
The explanation by Jake VanderPlas here is worth a read! .

Comparison of optimal widths



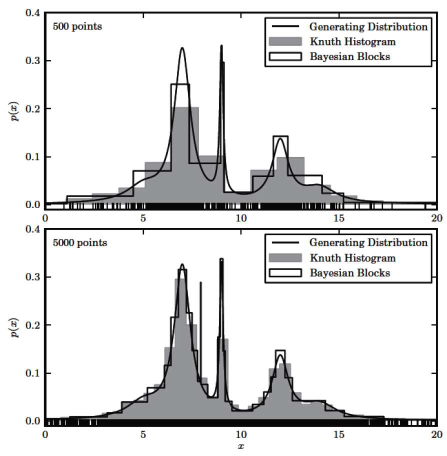
Source: AstroML book

Bayesian methods: constant vs. variable bin width



Source: AstroML book

Bayesian methods: constant vs. variable bin width



Source: AstroML book

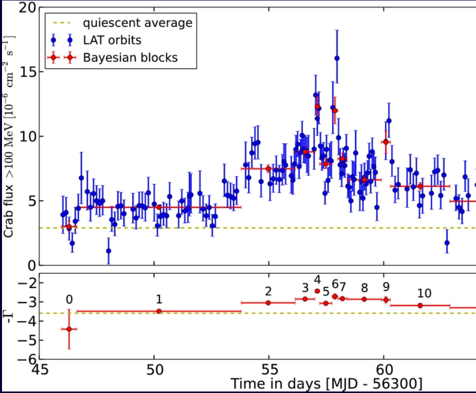


Fig. 2 from Mayer et al. 2013 ApJ Lett. 775, L37

Bayesian Blocks allow for more freedom in choice of bin width.

Kernel density estimate

Non-parametric density estimate. Recall the histogram estimator equation:

$$\hat{f}(x) = \frac{1}{hN}\sum_{i=1}^{N}\overbrace{\sum_{b=1}^{M}\mathbb{I}\left(\frac{|x_i - x_b|}{h} \leq 1\right)\mathbb{I}\left(\frac{|x - x_b|}{h} \leq 1\right)}^{K(u_i)}$$

Kernel density estimate

Non-parametric density estimate. Recall the histogram estimator equation:

$$\hat{f}(x) = \frac{1}{hN}\sum_{i=1}^{N}\overbrace{\sum_{b=1}^{M}\mathbb{I}\left(\frac{|x_i - x_b|}{h} \leq 1\right)\mathbb{I}\left(\frac{|x - x_b|}{h} \leq 1\right)}^{K(u_i)}$$

Generalisation: replace the inner sum with a function $K(u_i)$ of $u_i = \left(\frac{x - x_i}{h}\right)$.

The function $K(u)$ is called a kernel, with bandwidth h. It is evaluated at each data point x_i.

Kernel density estimate

Non-parametric density estimate. Recall the histogram estimator equation:

$$\hat{f}(x) = \frac{1}{hN}\sum_{i=1}^{N}\overbrace{\sum_{b=1}^{M}\mathbb{I}\left(\frac{|x_i - x_b|}{h} \leq 1\right)\mathbb{I}\left(\frac{|x - x_b|}{h} \leq 1\right)}^{K(u_i)}$$

Generalisation: replace the inner sum with a function $K(u_i)$ of $u_i = \left(\frac{x - x_i}{h}\right)$.

The function $K(u)$ is called a kernel, with bandwidth h. It is evaluated at each data point x_i.

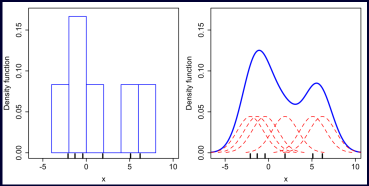
Histogram: each data value = delta function at its location.

KDE: "influence" of each data point "spread out" over "bin" of width h.

"Influence" = normalised function $K(u)$.

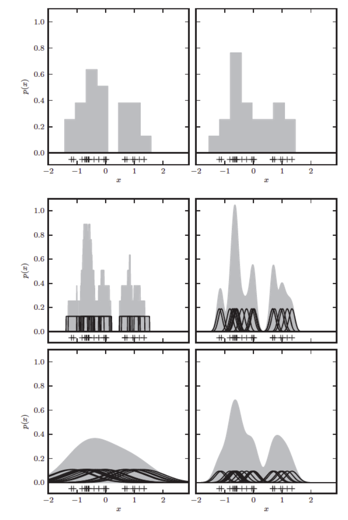
Bins of data points allowed to overlap.

Estimated density = sum of overlapping functions.



Credit: Wikipedia, User:Drleft, CC BY-SA 3.0

Histogram vs. KDE



AstroML book, Fig. 6.1

Top: Histograms with shifted bin centres

Centre left: adaptive histogram (one bin for each value, overlaps allowed).

Centre right, bottom: KDEs with Gaussian kernels of increasing bandwidth.

Small h: large variance. Large h: large bias (Bias-variance tradeoff).

KDE (contd.)

There are many functions used for $K(u)$.

Standard kernels: Gaussian, top hat, Epanechnikov (quadratic in u), exponential, linear, cosine.

Popular: Gaussian and Epanechnikov (minimises the mean square error).

For more, see here.

KDE (contd.)

There are many functions used for $K(u)$.

Standard kernels: Gaussian, top hat, Epanechnikov (quadratic in u), exponential, linear, cosine.

Popular: Gaussian and Epanechnikov (minimises the mean square error).

For more, see here.

Influence of $K(u)$ controlled by bandwidth h, which must be estimated.

Modern codes for computing KDEs have built-in options for this.

KDE (contd.)

There are many functions used for $K(u)$.

Standard kernels: Gaussian, top hat, Epanechnikov (quadratic in u), exponential, linear, cosine.

Popular: Gaussian and Epanechnikov (minimises the mean square error).

For more, see here.

Influence of $K(u)$ controlled by bandwidth h, which must be estimated.

Modern codes for computing KDEs have built-in options for this.

KDE implemented in `Python` packages such as `Scikit-learn`, `Scipy`, and `Statsmodels`.

KDE (contd.)

There are many functions used for $K(u)$.

Standard kernels: Gaussian, top hat, Epanechnikov (quadratic in u), exponential, linear, cosine.

Popular: Gaussian and Epanechnikov (minimises the mean square error).

For more, see here.

Influence of $K(u)$ controlled by bandwidth h, which must be estimated.

Modern codes for computing KDEs have built-in options for this.

KDE implemented in `Python` packages such as `Scikit-learn`, `Scipy`, and `Statsmodels`.

KDE can also be modified to handle measurement errors (see Section 6.1.2 in AstroML book).

KDE (contd.)

There are many functions used for $K(u)$.

Standard kernels: Gaussian, top hat, Epanechnikov (quadratic in u), exponential, linear, cosine.

Popular: Gaussian and Epanechnikov (minimises the mean square error).

For more, see here.

Influence of $K(u)$ controlled by bandwidth h, which must be estimated.

Modern codes for computing KDEs have built-in options for this.

KDE implemented in `Python` packages such as `Scikit-learn`, `Scipy`, and `Statsmodels`.

KDE can also be modified to handle measurement errors (see Section 6.1.2 in AstroML book).

	Bandwidth Selection	Available Kernels	Multi-dimension	Heterogeneous data	FFT-based computation	Tree-based computation
Scipy	Scott & Silverman	One (Gauss)	Yes	No	No	No
Statsmodels KDEUnivariate	Scott & Silverman	Seven	1D only	No	Yes	No
Statsmodels KDEMultivariate	normal reference cross-validation	Seven	Yes	Yes	No	No
Scikit-Learn	None built-in; Cross val. available	6 kernels x 12 metrics	Yes	No	No	Ball Tree or KD Tree

Summary table from Jake VanderPlas' blog.

Summary

1. If you're only interested in the general trend in your data, use box/violin plots. They'll also immediately identify outliers!
2. Histograms are fast but bad for various reasons – their shapes depend on bin size and bin location, and they degrade the information contained in the raw data.
3. There are ways to figure out the optimum bin size – both frequentist and Bayesian. The Bayesian versions are more sensitive to multimodal distributions, and allow for the computation of the optimum bin size without as few assumptions on the underlying distribution as possible.
4. The Bayesian Blocks method allows for variable bin size! It is especially applicable for small data sets.
5. If you're really interested in generating a function that mimics the true population distribution, use KDEs.