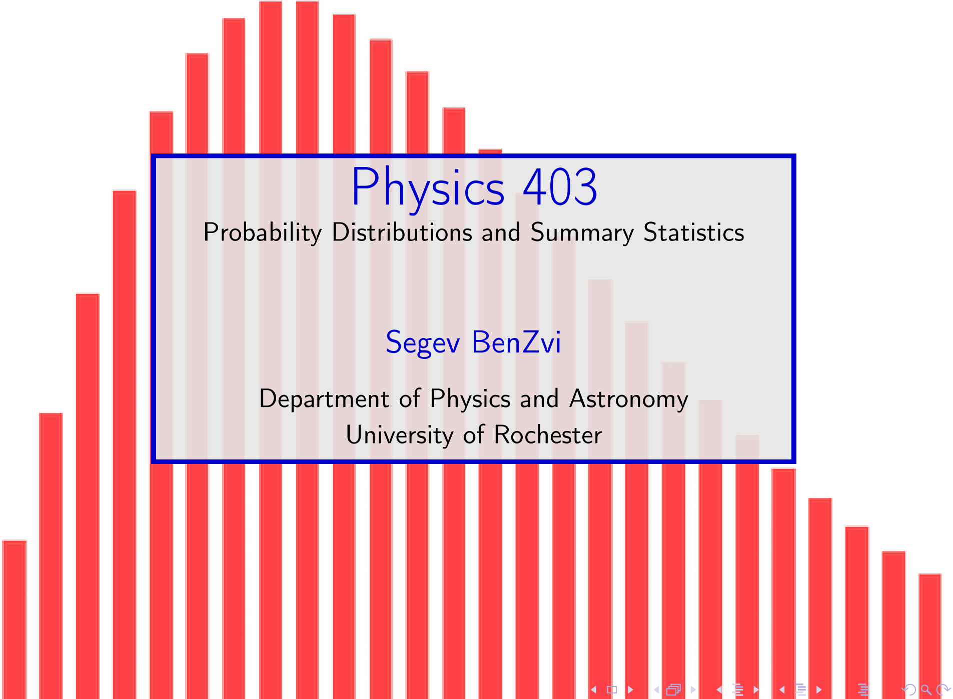


Physics 403

Probability Distributions and Summary Statistics

Segev BenZvi

Department of Physics and Astronomy
University of Rochester

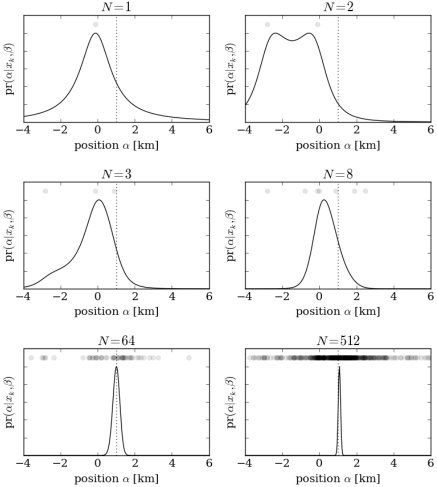
Table of Contents

1. Review of Last Class
   - Basic Rules of Probability
   - Marginalization
   - The Probability Density Function

2. Summary Statistics
   - Mean (Expectation Value)
   - Variance and Covariance
   - High-Order Moments
   - Ordered Rank Statistics
   - The Cumulative Distribution Function
   - Histograms

Last Time

Basics of Probabilistic Reasoning



- Degrees of plausibility are represented by real numbers.
- As data supporting a hypothesis accumulate, its plausibility increases continuously and monotonically.
- If there are two different ways to use the same information, both methods should give the same conclusion.
- All probability is conditional on some assumption.

Last Time

Introduction to Probability

- **Sum Rule**

$$P(A|I) + P(\overline{A}|I) = 1$$
$$\sum P(H_i|I) = 1 \quad \text{for exclusive } H_i$$

- **Product Rule (Joint Probability)**

$$P(A, B|I) = P(A|B, I)P(B|I)$$

- **Bayes' Theorem**

$$P(A|B, I) = \frac{P(B|A, I)P(A|I)}{P(B|I)}$$

- **Law of Total Probability**

$$P(A|I) = \sum_i P(A, B_i|I) = \sum_i P(A|B_i, I)P(B_i|I)$$

Marginalization

Discrete "Events"

Given a set of mutually exclusive possibilities Y_k, we can estimate the probablity of some event X as

$$P(X|I) = \sum_k P(X, Y_k|I), \qquad \text{where } \sum_k P(Y_k|X, I) = 1$$

Example

Suppose there are 5 presidential candidates in an election, which we represent by Y_k with $k = 1, \ldots, 5$. Then the probability that the unemployment rate will go down next year (X) irrespective of who wins the election is given by

$$P(X|I) = \sum_{k=1}^{5} P(X, Y_k|I)$$

Marginalization

Continuum Limit

Suppose we don't have a set of discrete events or hypotheses to test, but an arbitrarily large set of propositions in a range of values? In this case, we go to the $M \to \infty$ limit:

$$P(X|I) = \int_{-\infty}^{\infty} p(X, Y|I)dY, \text{ where}$$

$$p(X, Y|I) = \lim_{\delta y \to 0} \frac{P(X, y \leq Y < y + \delta y|I)}{\delta y}$$

is called the *probability density function* (PDF) of X and $Y \in [y, y + \delta y]$.

Example

We want to calculate the mass of a particle like the Higgs. We consider a parameter space where m_H may take on any continuous value inside a physically motivated range.

The Probability Density Function

- The PDF is a probability per unit volume (hence *density*).
- The quantity we want is a probability. To get it we calculate volume integrals of the PDF.
- Obviously, it doesn't have to be a joint distribution. The 1D case:

$$P(a \leq X < b|I) = \int_a^b p(x|I)dx$$

- The PDF must be normalized since the values of x are mutually exclusive:

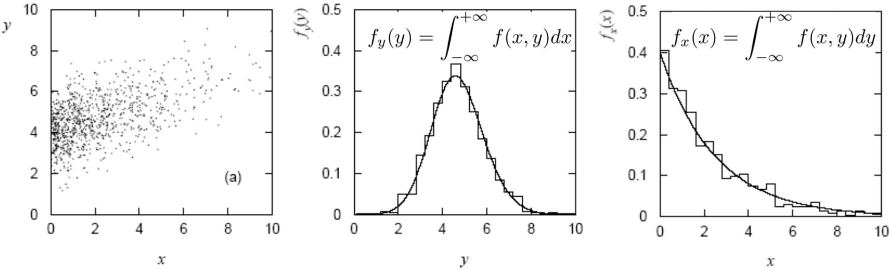
$$\int_{-\infty}^{\infty} p(x|I)dx = 1$$

- The PDF contains all the information we need to make probabilistic inferences about a parameter, event, or a hypothesis. Its maximum gives the most probable value of a parameter.

Comment: Marginalization vs. Projection

Marginalization eliminates an unwanted parameter from a joint PDF:

$$p(x|I) = \int p(x,y|I)\ dy \qquad \text{(marginal PDF)}$$



This is not the same as projection, in which you calculate the PDF of x for some fixed y (see [1]), giving you a conditional PDF:

$$p(x|y,I) = \frac{p(x,y|I)}{\int p(x,y|I)\ dx} = \frac{p(y|x,I)p(x|I)}{p(y|I)}$$

Table of Contents

1. Review of Last Class
   - Basic Rules of Probability
   - Marginalization
   - The Probability Density Function

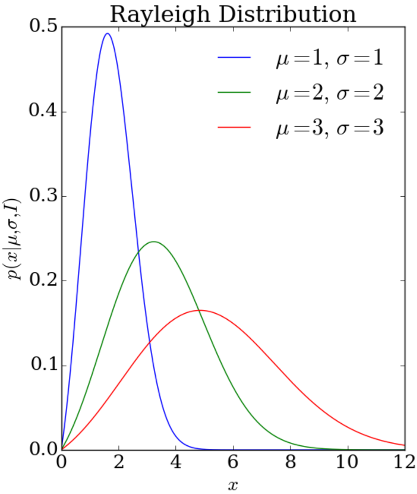
2. Summary Statistics
   - Mean (Expectation Value)
   - Variance and Covariance
   - High-Order Moments
   - Ordered Rank Statistics
   - The Cumulative Distribution Function
   - Histograms

Summary Statistics

Often we don't have access to the full PDF. Or we do, but we wish to summarize it in just a few numbers:

- **Mean**: "location"
- **Variance**: "width" or "spread"
- **Mode**: most probable value
- **Median**: central value
- **Percentiles**: rank/scoring
- Skew: asymmetry of PDF
- Kurtosis: "peakedness"

Can you think of a case where these might not be sufficient?



Expectation Value

The Mean of a Distribution

- In terms of a PDF the expectation value or mean of a distribution is given by

$$\mu = \langle x \rangle = \int x \ p(x|I) dx$$

- Other notations: $\mathrm{E}\left(x\right)$ and $\bar{x}$. Read the latter as "x-bar" instead of "not-x." It isn't logical negation.
- Typical usage: μ, $\langle x \rangle$, and $\mathrm{E}\left(x\right)$ refer to the expectation value of a PDF, while $\bar{x}$ refers to the mean of a set of measurements $\{x_i\}$:

$$\bar{x} = \frac{1}{N} \sum_{i-1}^{N} x_i$$

- Weighted mean: if not all data should contribute equally to the sum,

$$\bar{x} = \frac{\sum_{i-1}^{N} w_i x_i}{\sum_{i-1}^{N} w_i}$$

Special Case

Cauchy/Lorentzian/Breit-Wigner Distribution

- The Cauchy distribution is defined by the PDF

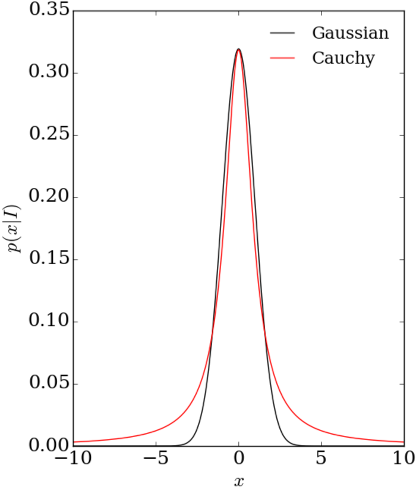
$$p(x|x_0,\Gamma) = \frac{1}{2\pi}\frac{\Gamma}{(x-x_0)^2+(\Gamma/2)^2}$$

- If you try to calculate

$$\langle x\rangle = \int_{-\infty}^{\infty} x\ p(x|x_0,\gamma)\ dx$$

you will find that it diverges!

- This function describes spectral lines and resonances, so we do come across it.



Variance

The Width of a Distribution

- In terms of a PDF the variance of a distribution is

$$\sigma_x^2 = \mathrm{V}\left(x\right) = \langle (x-\mu)^2 \rangle = \int (x-\mu)^2 \; p(x|I)dx$$

- Note how variance is defined in terms of the mean μ; it measures the spread of squared deviations of x about μ. This is more obvious if you remember the definition of variance for a data set $\{x_i\}$:

$$\hat{\mathrm{V}}(x) = \frac{1}{N}\sum_{i=1}^{N}(x_i-\mu)^2$$

- The square root of the variance, called the standard deviation or RMS error σ_x, is a measure of the width of the PDF in the same units as x.

Calculating Variance

Known and Unknown Mean

- Note that the calculation of the variance of a data set will differ if the mean is known vs. calculated from the data.

Known Mean

$$\hat{\mathrm{V}}(x) = \frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2$$

Unknown Mean

$$\hat{\mathrm{V}}(x) = \frac{1}{N-1}\sum_{i=1}^{N}(x_i - \bar{x})^2$$

- If we compute $\bar{x}$ from the data but use the formula on the left, our *estimate* of the variance of the PDF will be too small (biased).
- Underestimating $V(x)$, in this or any other way, can result in serious mistakes. For example, for small N you could underestimate the probability of observing a particular x_i.

Calculating Variance

"Online" Formula

- Suppose you have a detector that is measuring events x_i in real time. How do you calculate $\mathrm{V}\left(x\right)$ as the data are recorded?
- If you use the formula

$$\hat{\mathrm{V}}(x) = \frac{1}{N-1}\sum_{i=1}^{N}(x_i - \bar{x})^2$$

  then you need to estimate $\bar{x}$ and then recalculate all of the deviations from $\bar{x}$, requiring a second pass through the data. Inefficient!
- But, if you realize that

$$\mathrm{V}\left(x\right) = \langle (x-\mu)^2 \rangle = \langle x^2 \rangle - \langle x \rangle^2 = \overline{x^2} - \bar{x}^2$$

  then you can write an algorithm that computes both the mean and variance on the fly. You will do this in your next problem set.

Covariance

- The covariance of two quantities x and y is given by

$$\begin{aligned}\sigma_{xy}^2 = \operatorname{cov}(x,y) &= \langle (x-\mu_x)(y-\mu_y) \rangle \\ &= \iint (x-\mu_x)(y-\mu_y)\ p(x,y|I)\ dx\ dy\end{aligned}$$

- As with variance, there is a nice simplification of covariance that makes calculations easy:

$$\operatorname{cov}(x,y) = \langle xy \rangle - \langle x \rangle \langle y \rangle$$

  Clearly, $\sigma_{xx}^2 = \operatorname{cov}(x,x) = \mathrm{V}(x) = \sigma_x^2$.

- Often (but not so much in physics) people use a dimensionless version of covariance called the correlation coefficient,

$$\rho = \frac{\operatorname{cov}(x,y)}{\sqrt{\mathrm{V}(x)\,\mathrm{V}(y)}} = \frac{\sigma_{xy}^2}{\sigma_x \sigma_y}$$

Covariance

Independent x and y

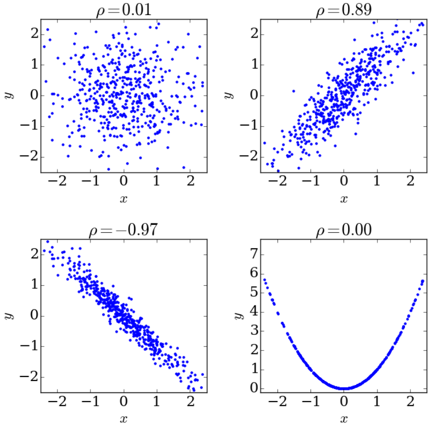
Example

If x and y are independent, what is their covariance?

$$
\begin{aligned}
\operatorname{cov}(x, y) &= \langle xy \rangle - \langle x \rangle \langle y \rangle, \text{ but} \\
\langle xy \rangle &= \iint xy\ p(x, y|I)\ dx\ dy \\
&= \iint xy\ p(x|I) p(y|I)\ dx\ dy \\
&= \int x\ p(x|I)\ dx \int y\ p(y|I)\ dy \\
&= \langle x \rangle \langle y \rangle
\end{aligned}
$$

So clearly $\operatorname{cov}(x, y) = 0$ if x and y are independent.

Examples of Covariance and Correlation



- Correlations work as you expect; they can be positive, negative, or zero.
- Note: x, y independent will have $\text{cov}\,(x, y) = 0$.
- Note: $\text{cov}\,(x, y) = 0$ does not imply that x, y are independent.
- Get comfortable with the concept of covariance. It is central to fitting and parameter estimation.

Higher-Order Summary Statistics

- The mean ("central value") is the first moment of a PDF and the variance ("spread") is the second moment.
- The third moment ("asymmetry") is called the skew, and it is defined as

$$\begin{aligned}\mathrm{skew}\,(x) = \gamma_x &= \int (x-\mu_x)^3\ p(x|I)\ dx \\ &= \frac{1}{N\sigma_x^3}\sum_{i=1}^{N}(x_i-\bar{x})^3\end{aligned}$$

- The fourth moment is called the kurtosis.
- You could keep going like this, but eventually it becomes easier to just characterize your distribution with the full PDF or at least a compressed representation like a histogram.

The Median

- The median is defined as the value in a PDF or a data set where 50% of the data are expected to be above or below the value.
- For an ordered data set x_i of length N,

$$\text{median}\,(x) = x_{0.5} = \begin{cases} x_{(N+1)/2} & N \text{ is odd} \\ (x_{N/2} + x_{N/2+1})/2 & N \text{ is even} \end{cases}$$

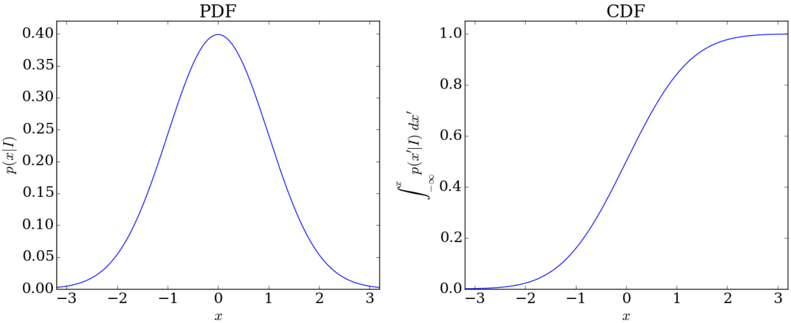
- For a PDF of x, the median is given by the value $x_{0.5}$ which satisfies the condition

$$P(x \leq x_{0.5}|I) = \int_{-\infty}^{x_{0.5}} p(x|I)\ dx = 0.5$$

- This is literally the definition above expressed in terms of the cumulative distribution $P(x \leq x_{0.5}|I)$.

The Cumulative Distribution Function

- The cumulative distribution function, or CDF, of x is the probability of observing a value at or below some x. It is the integral of the PDF.



- For a normalized one-dimensional PDF, the CDF will go to zero as $x \to -\infty$ and one as $x \to +\infty$.

In-Class Exercise

The Discrete CDF

Example

You flip a fair coin twice and let X be the number of heads. What are the possible outcomes and their probabilities? What is the CDF?

In-Class Exercise

The Discrete CDF

Example

You flip a fair coin twice and let X be the number of heads. What are the possible outcomes and their probabilities? What is the CDF?

$P(X = 0|I) = 1/4$, $P(X = 1|I) = 1/2$, $P(X = 2|I) = 1/2$

In-Class Exercise

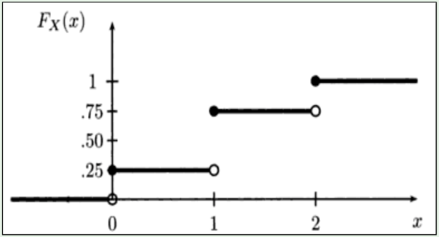
The Discrete CDF

Example

You flip a fair coin twice and let X be the number of heads. What are the possible outcomes and their probabilities? What is the CDF?

$P(X = 0|I) = 1/4$, $P(X = 1|I) = 1/2$, $P(X = 2|I) = 1/2$

$$F_X(x) = \begin{cases} 0, & x < 0 \\ 1/4, & 0 \leq x < 1 \\ 3/4, & 1 \leq x < 2 \\ 1 & x \geq 2 \end{cases}$$



Rank Statistics

Quantiles and Data Scoring

- Let's extend the definition of the median. We define the quantile x_α as the value which satisfies the definition

$$P(x \leq x_\alpha|I) = \int_{-\infty}^{x_\alpha} p(x|I)\ dx = \alpha$$

Example

The 25th percentile of a distribution $x_{0.25}$ satisfies

$$P(x \leq x_{0.25}|I) = \int_{-\infty}^{x_{0.25}} p(x|I)\ dx = 0.25$$

- Quantiles are tail statistics; they tell us how probable it is to find x in one of the tails of the PDF $p(x|I)$. These are used all the time for *scoring*.

Why Use the Median?

- Aside from scoring data like exams, when is the median ever useful?
- It is a measure of centrality that is less sensitive to the tails of of a PDF than other measures like the mean.

Example

Let $\{x_i\} = 1, 2, 1, 1, 1, 2, 3, 1, 1000$. The mean and median are given by

$$\bar{x} \approx 112.4$$
$$\text{median}\,(x) = 1$$

- The mean in the example is sensitive to an outlier far from the main cluster of values, while the median is not. It is said to be “robust” against outliers.
- **Question**: how should we define an outlier?

Aside: Outliers in Physics

- You're doing a measurement. *Given an accepted model of the data* – the so-called null hypothesis – you observe something very unlikely.
- Is this a good thing or a bad thing? Could be either.
- "Bad:" the null hypothesis is correct, and you don't understand something in your data. Maybe there is an unknown systematic effect or poor calibration in your instruments.
- "Good:" you did everything right, and the null hypothesis does not describe the data well. Congratulations, you made a discovery!

Example

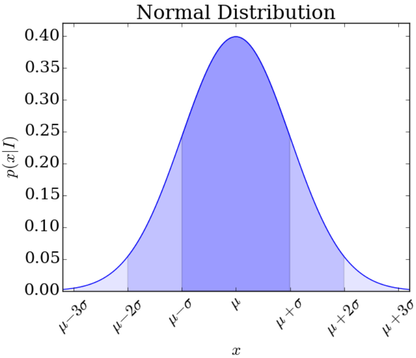
In 2011, the OPERA Collaboration measured the time of flight of a ν_μ beam from CERN and found $(v_\nu - c)/c = (2.48 \pm 0.28 \pm 0.30) \times 10^{-5}$ [2]. *Given the null hypothesis* $(v_\nu - c)/c \leq 0$, this was a significant outlier. Not the good kind [3].

Aside: Decision Making in Physics

The 68-95-99 Rule

In physics we tend to express rare events in terms of the tails of the Gaussian PDF

$$p(x|I) = \frac{1}{\sqrt{2\pi\sigma}} \exp\left\{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right\}$$



The "68-95-99" quantile rule:

- 68.27% of the data are within 1σ of the mean.
- 95.45% of the data are within 2σ of the mean.
- 99.73% of the data are within 3σ of the mean.

Aside: Decision Making in Physics

The 5σ Rule

The "sigma" nomenclature is a nice shorthand for quantiles. For example, "3σ" is physicist-speak for something outside the central 99% of a distribution (or upper/lower 99th percentile). So even when your PDF isn't Gaussian, everyone knows that "3σ" means the 99.7*th* percentile.

Example

The 5σ **Rule**: the gold standard for a discovery in HEP is a 5σ deviation of data from the null hypothesis. For an upper-tail test, this corresponds to

$$P(x \leq \mu + 5\sigma | I) = \int_{-\infty}^{\mu+5\sigma} \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right\}$$
$$\approx 3 \cdot 10^{-7}$$

Why so strict? Why not use 1%, like in medical trials? We'll come back to this later in the course. You may find the answer... disturbing.

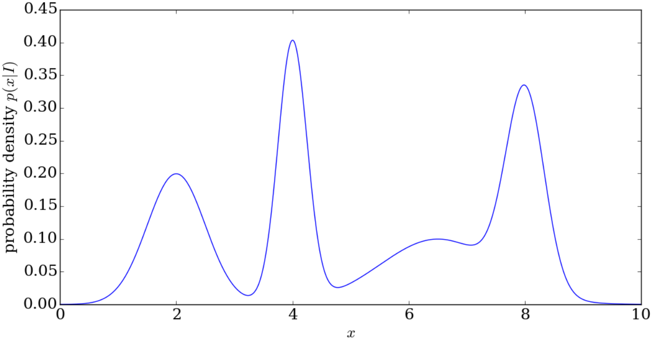
The Mode

- The most probable value in a distribution (or most common value in a data set), called the mode, is given by the maximum of the PDF.
- The mode is a **location parameter** like the mean. Unlike the mean, it does not account for the skewness of the PDF, so the mean may perform better for asymmetric distributions.
- However, when we do parameter estimation, we are most interested in the maximum (the mode) of the PDF and the shape of the distribution around the maximum.
- All the information you need for parameter estimation is in the PDF. Summary statistics are nice, but they can mislead you.

Breakdown of Summary Statistics

Multimodal Distributions

Where would the mean be in this distribution? What is the variance?

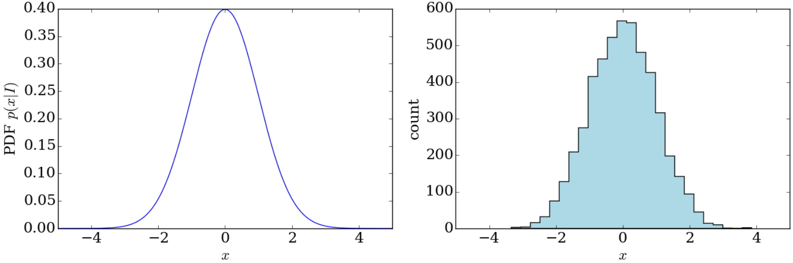


Would any or all of the moments of the PDF that we defined today be sufficient to describe this?

Binning of Data

Data Compression with Histograms

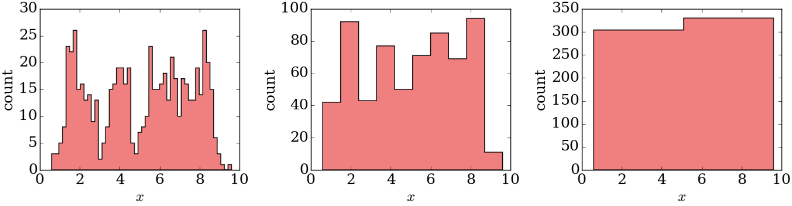
- Often you will want to bin your data, or you will be given binned data.
- A histogram is a division of N data points into m subintervals or bins of width Δx_i. A value x is sorted into bin i if $x \in [x_i, x_i + \Delta x_i]$.



- Normalization: $N = \sum_i n_i \cdot \Delta x_i$, with n_i the count in bin i.
- Note: data can also be weighted when filling the histogram.

Data Compression with Histograms

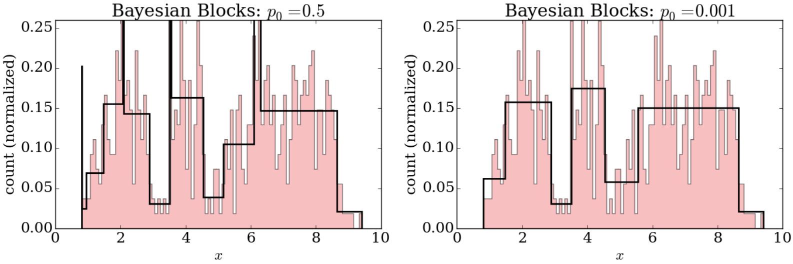
- Histograms are a great way to summarize a large data set, but never forget that they are a compression technique. When you bin data you are throwing away information.



- Ideally: bin edges are chosen such that the PDF changes very little across the width of the bin.
- Typically the bin widths are set to the same value Δx, but it's better to have equal counts per bin.

Automatic Binning Schemes

- There is a large literature on optimally binning a given data set. One scheme now common in astronomy is called Bayesian Blocks [4].
- Idea: iteratively sort through the data for "changepoints" that indicate whether or not a bin should be split.



- You can be more or less aggressive about splitting bins by tuning the false positive rate p_0 of accidentally splitting a bin.

Some Warnings About Binning

- You have seen that with a poor choice of binning you can effectively wipe out features in your data.
- You can bin more finely, though eventually you'll reach a point where every bin contains only 0 or 1 counts. So much for compression.
- Another issue: because you're binning some random x, the counts in each bin are themselves random numbers with some uncertainty.
- Most binned statistics, like the χ^2 test we'll talk about later in the course, assume the uncertainty on the counts in each bin is Gaussian.
- However, if the counts in a bin are low (< 10) then the distribution will actually be Poisson, violating the conditions of your χ^2 test.
- Possible consequence: you misinterpret your χ^2 statistic and publish a false discovery.

Summary

- The probability density function (PDF) is the probability per unit volume of one or more parameters in a parameter space.
- The PDF contains all the information you need to know about a parameter.
- Most often we are interested in the most probably location of a parameter and its distribution about this point.
- There are various summary statistics we can use to capture the essence of a distribution but there are pathological cases which you encounter frequently in research.
- Binning data is an effective way of summarizing it in m values (counts). Due to the freedom you have in choosing bins, you have to be careful not to throw away too much information.

Further Reading I

[1] Glen Cowan. *Statistical Data Analysis*. New York: Oxford University Press, 1998.

[2] T. Adam et al. "Measurement of the neutrino velocity with the OPERA detector in the CNGS beam". In: *JHEP* 1210 (2012), p. 093. arXiv: 1109.4897 [hep-ex].

[3] M. Antonello et al. "Measurement of the neutrino velocity with the ICARUS detector at the CNGS beam". In: *Phys.Lett.* B713 (2012), pp. 17–22. arXiv: 1203.3433 [hep-ex].

[4] J.D. Scargle et al. In: *Astrophys.J.* 764 (2013), p. 167.