Computational Physics Section 1: Visualization of Data

Prof. Lisa W. Koerner

University of Houston

Department of Physics

Visualization

- Reading: Chapter 3 Graphics and Visualization
- Technical:
  - I will use matplotlib, a plotting library for Python
  - There are other options in Python for making plots – you are welcome to use whatever you want
- Start with two commons ways of visualizing data
  - Graphs: You have two variables (x,y) where one is functionally dependent on the other. Plotting y vs x on a graph to show that functional dependence, and perhaps compare to a model if x and y are measurements
  - Histograms: You make multiple measurements of a random variable and want to see the distribution of the variable

Example: Making a graph

Example: Plotting a function

Example: Multiple lines on the same plot

Example: Math on Numbers, Arrays, and Lists

Exercise 1

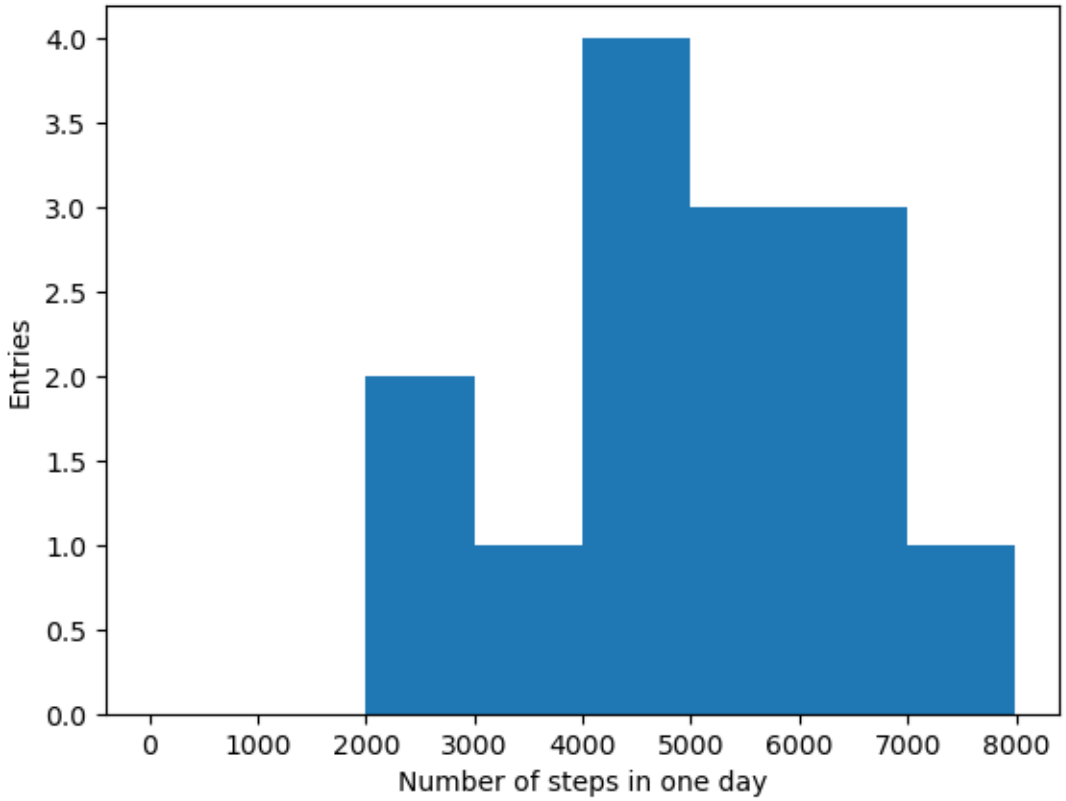
Histograms: An Example

- Start with a variable that can be measured multiple times
  - Example: number of steps I take per day
  - For a 2-week period, here are the number of steps per day: 2725, 6908, 2294, 4967, 5999, 6513, 4549, 5745, 4726, 3144, 7274, 6109, 4786, 5354
  - (Don't judge me – this was during 2020 quarantine!)
- Next, we want to "bin" the data, i.e. a sequence of consecutive non-overlapping intervals that span the whole range of values
  - Let's use these ranges: 0-999, 1000-1999,2000-2999,3000-3999,4000-4999,5000-5999,6000-6999,7000-7999
  - That's 8 bins with a range of 0 to 7999 steps with a bin size of 1000 steps

Histograms: An Example

- Data: 2725, 6908, 2294, 4967, 5999, 6513, 4549, 5745, 4726, 3144, 7274, 6109, 4786, 5354
- Now we count how many time one of the values shows up in each bin:
  - 0-999: 0 entries
  - 1000-1999: 0 entries
  - 2000-2999: 2 entries
  - 3000-3999: 1 entry
  - 4000-4999: 4 entries
  - 5000-5999: 3 entries
  - 6000-6999: 3 entries
  - 7000-7999: 1 entry
  - Sum: 2+1+4+3+3+1 = 14 (as it should)

Histograms: An Example

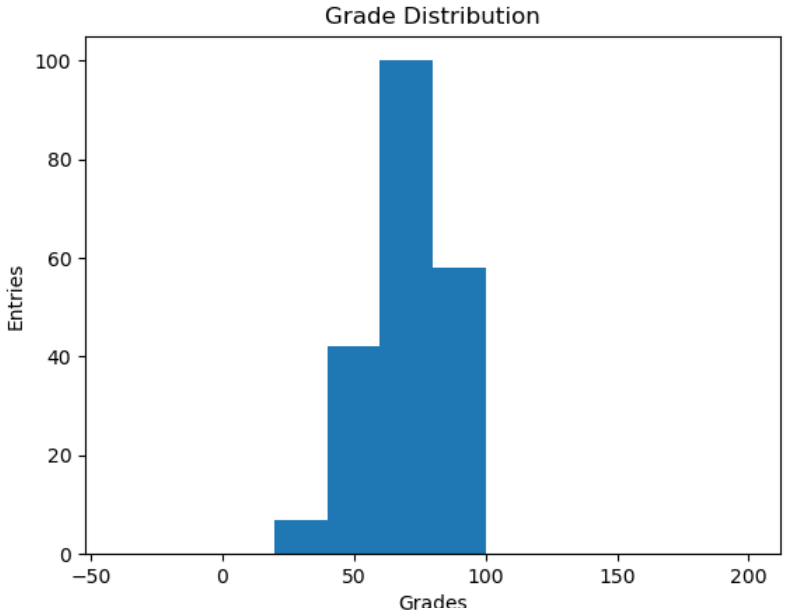


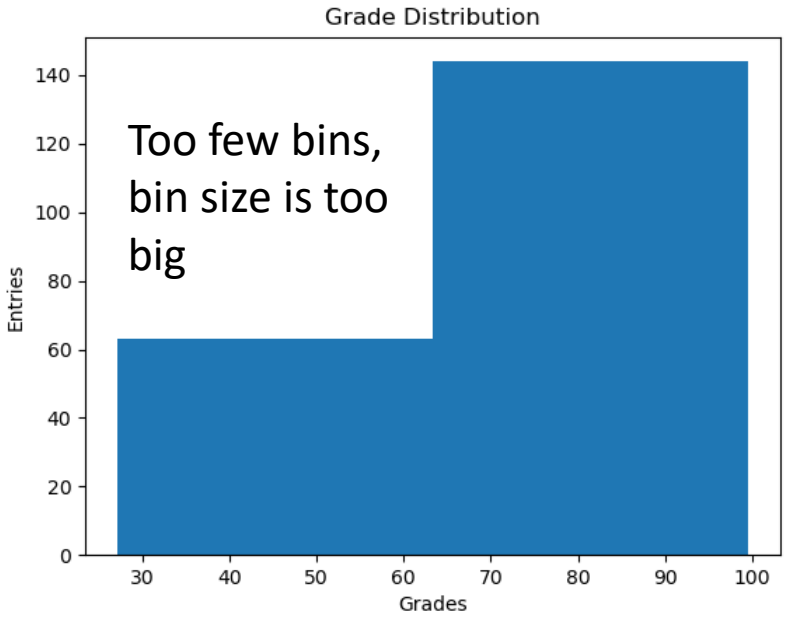
Example: Make a Histogram

Example: Histogram with Specified Binning

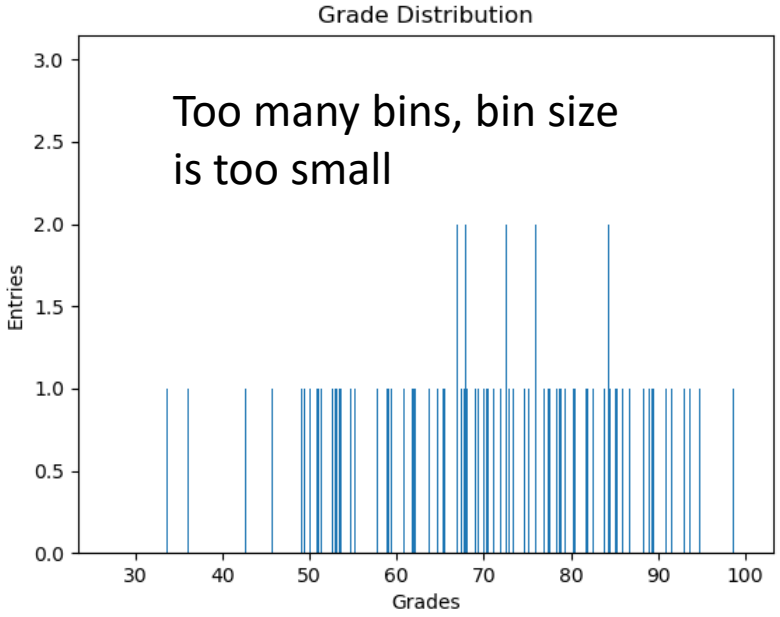
Histogram Binning

- Choose a range and bin size/number of bins that make sense for your data
  - For a given range, the bin size or width (distance between neighboring bins) and the number of bins are inversely correlated
- Range
  - Consider the possible range of your data values
    - i.e. for grades: It doesn't make sense to have bins below 0 or above 100
- Bin size
  - Consider the resolution of your data: the grades are specified to 2 decimal places, so it doesn't make sense to have a bin size smaller than 0.01 (but even that is too small)
  - Related to instrument accuracy/measurement error as well: if you are measuring a length with a measurement error of +-1 cm, you shouldn't have a bin size smaller than this



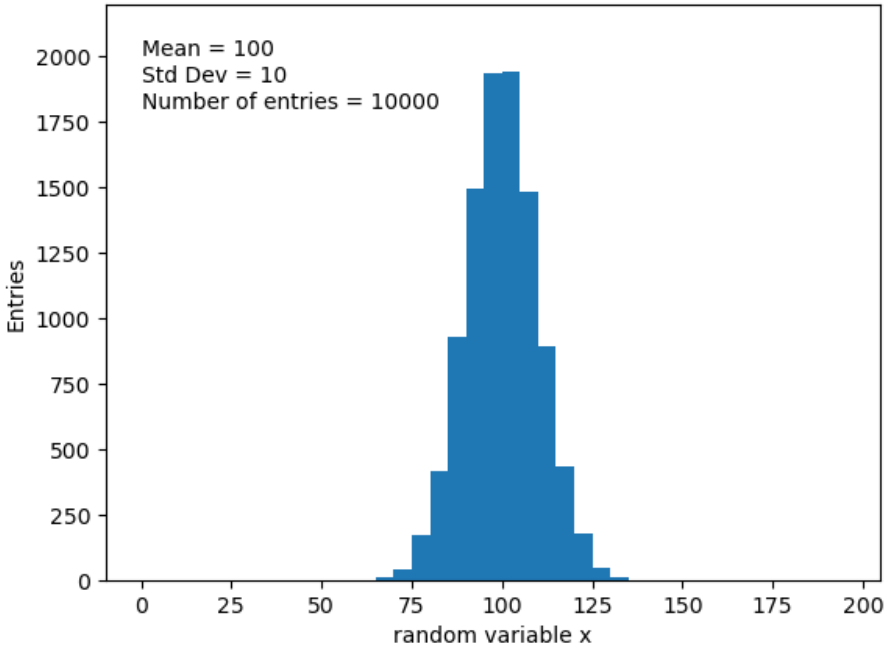


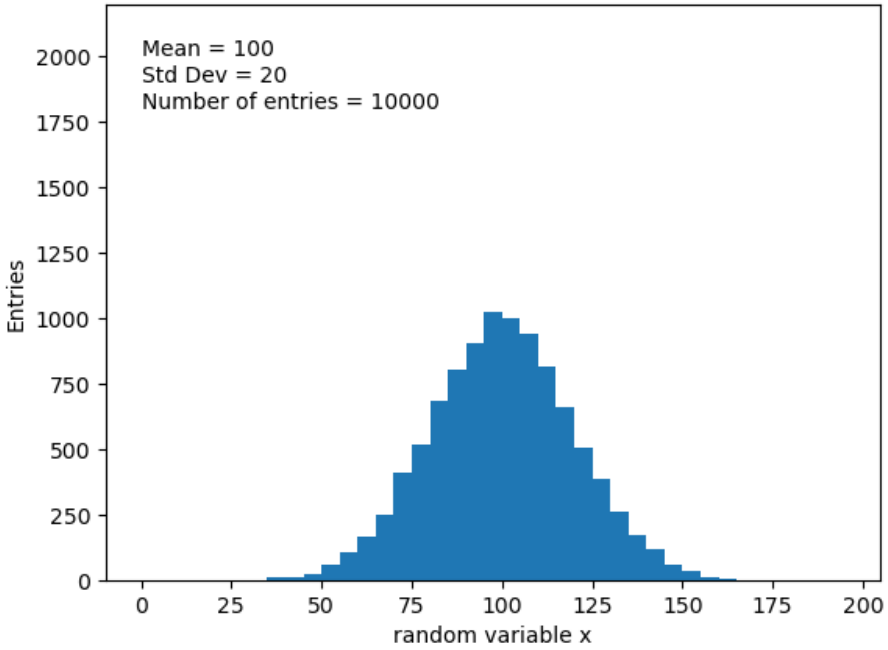
Inappropriate range



Some Statistics

- Mean (average): a measure of the central or typical value for a set of values
  - $\mu = \frac{1}{n}\sum_{i=1}^{n} x_i$
- Standard deviation: a measure of the amount of variation in a set of values
  - $\sigma = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \mu)^2}$


Mean = 100
Std Dev = 10
Number of entries = 10000
Entries
random variable x
0
250
500
750
1000
1250
1500
1750
2000
0
25
50
75
100
125
150
175
200


Mean = 100
Std Dev = 20
Number of entries = 10000
Entries
random variable x
0
250
500
750
1000
1250
1500
1750
2000
0
25
50
75
100
125
150
175
200

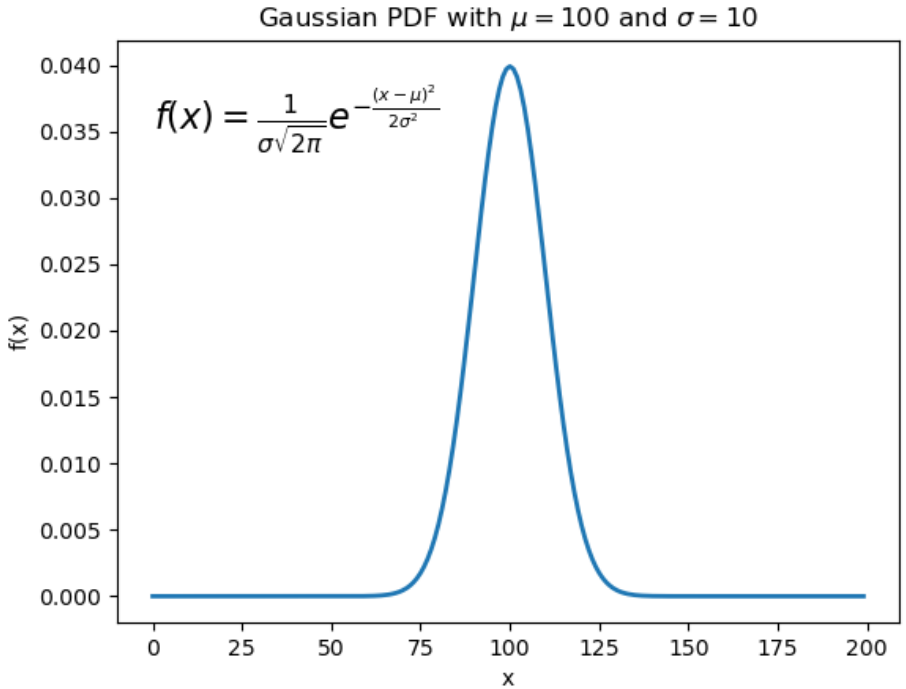
Example: Mean and Standard Deviation

Exercise 2

Histograms and PDFs

- A histogram gives us an idea of what the underlying probability distribution is for a variable, i.e. the relative probability for the variable to take different values
- For example, if a variable is “normally distributed,” we expect the histogram to have the shape of a Gaussian probability density function (PDF):

The integral of this function is 1 (the integral of a PDF is always 1). We’ll revisit this integral when we talk about numerical integrals.



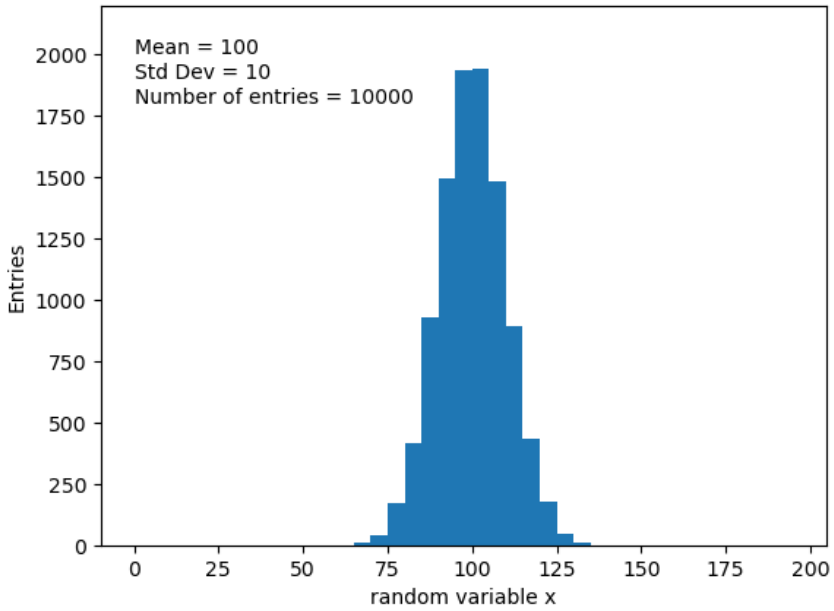
μ = mean
σ = standard deviation

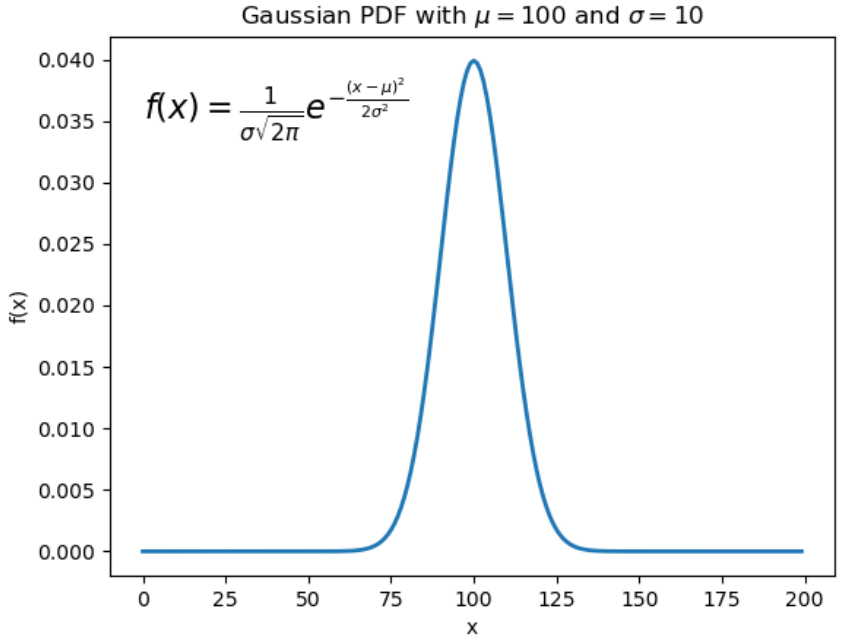
Probability Density Functions

- Probability density function (PDF), f(x): f gives the probability per unit x that the random variable will take value x.
  - So if x is a variable measured in inches (like height), the PDF will have units of 1/inches
- Density: Like mass density has units of mass per volume, probability density has units of probability per something
- Why "per unit x"? For a continuous variable, the variable can have an infinite number of values, so the probability for any one specific value is zero. BUT the probability of getting a value between x and x+ Δ can be quantified. The PDF is defined as that probability over the interval length, in the limit that the interval length goes to zero:
  - $f(x) = \lim_{\Delta \to 0} \frac{P(x<value<x+\Delta)}{\Delta}$
- In the Gaussian pdf, the σ in the denominator gives it units of x

Histograms and PDFs

- So, how do we compare a histogram to a PDF?
  - Our histogram has a Gaussian shape, but the scales are way different



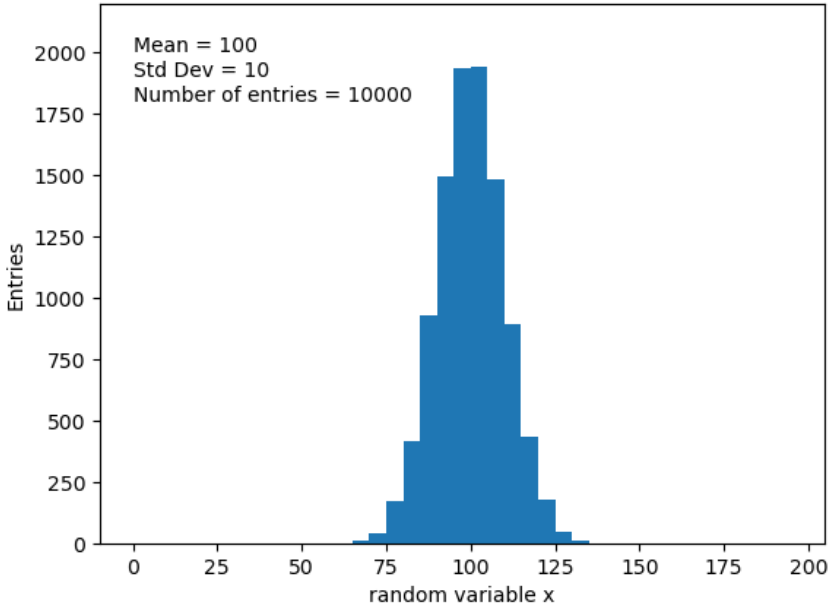


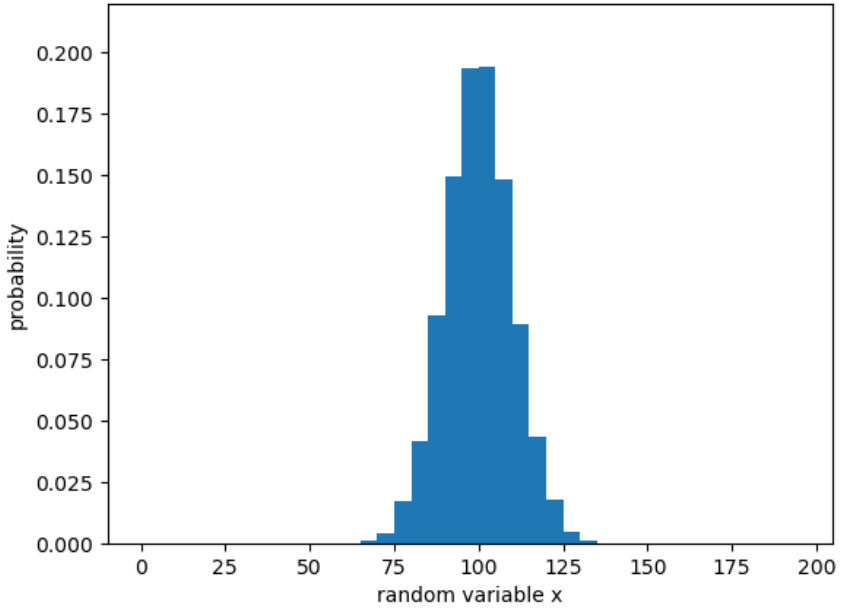
Histograms and PDFs

- Our histogram shows number of entries per bin. To get it to show a probability, we need to divide by the total number of entries in the histogram.
- For example, in the x=85-90 bin of the histogram on the previous slide, there are 931 entries in that bin. There are 10,000 entries total
  - The probability of a getting a value in the x=85-90 bin is 931/10,000 = 0.0931
- So we need to divide the bin contents (number of entries in each bin) by the total number of entries
- Dividing by the total number of entries is equivalent to filling each bin with a weight of 1/10000 instead of 1
  - You can do this with a “weights” option in the hist() function

Histograms and PDFs

Histogram → probability



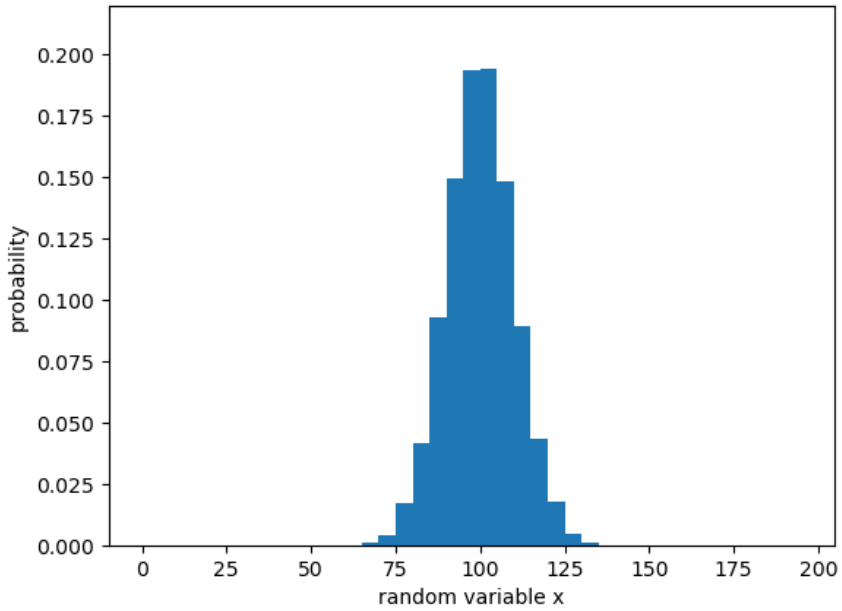


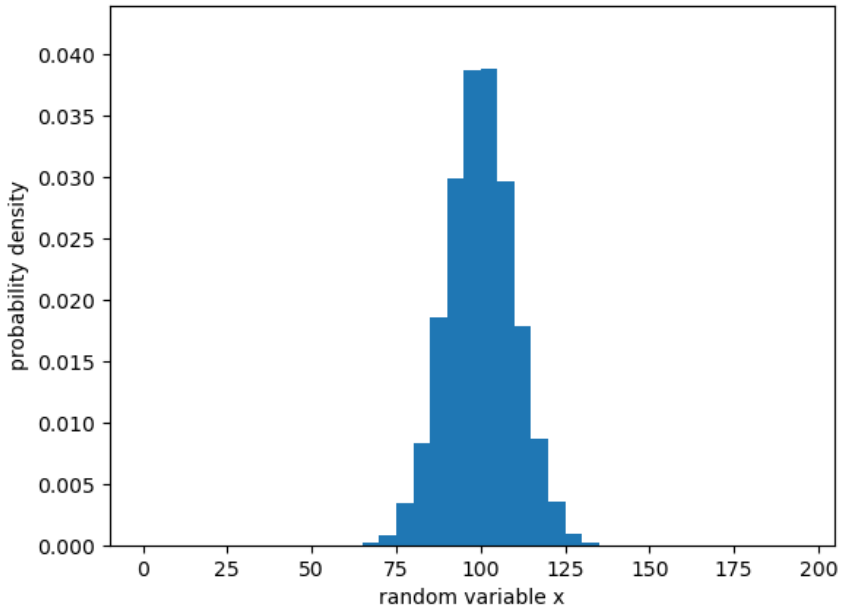
Same shape, we've
just scaled it down by
a factor of 1/10000.

Histograms and PDFs

- Now we've got the probability for each bin
- To get the probability density, we divide the probability by the interval length, which is the bin width (5 in this case)

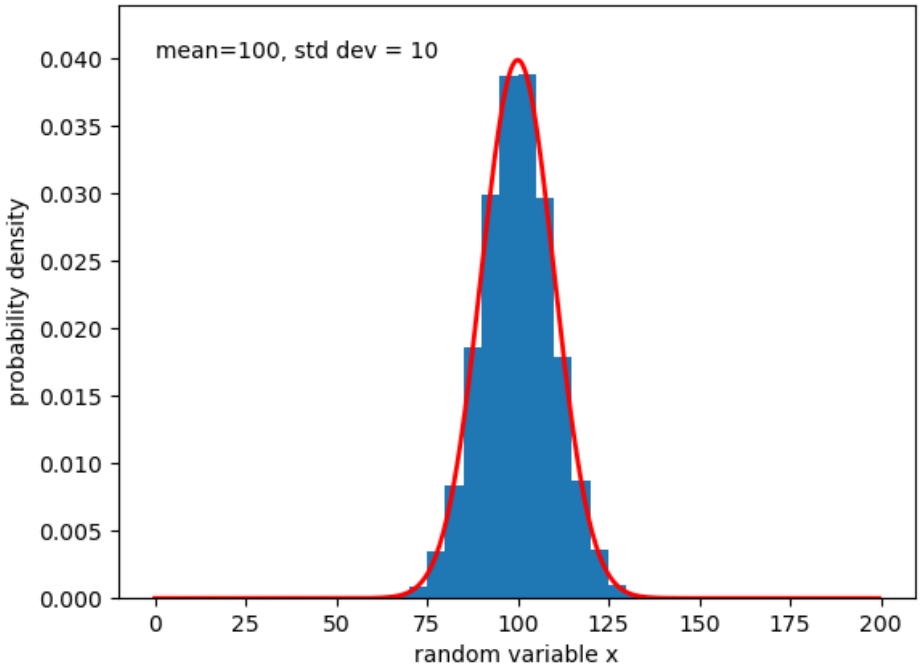
Probability → probability density





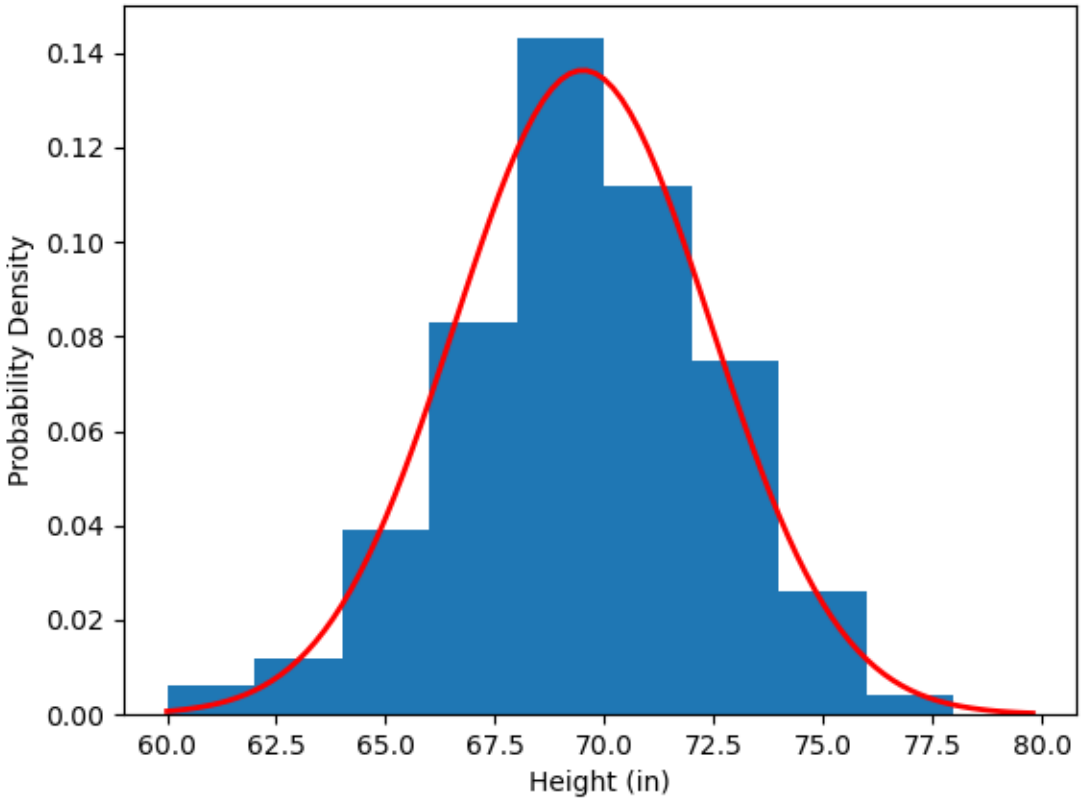
Histograms and PDFs

- Now that we've turned our histogram into a probability density, we can compare it directly to a Gaussian function with the same mean and standard deviation



Example: Probability Density

Example: Probability Density



The normalization looks right, but it's not a perfect match…

That's OK! Even if a variable is truly Gaussian, with low statistics (small sample size), it's not expected to look perfectly Gaussian.

We will revisit this when we talk about random numbers later in the semester.

Other PDFs

- Not all random variables are Gaussian
  - Our example was height, which is described well by a Gaussian
- For example…
  - Landau distribution: energy loss of a charged particle passing through matter
  - Poisson distribution: probability of a given number of events occurring in a fixed interval of time if the events occur with a known constant rate and are independent (photon counting)

Mean, Median, Mode

- Different measures of “central tendency”
  - Median is the middle value of a set of ordered numbers
  - Mode is the most frequent number in the set (location of the peak)

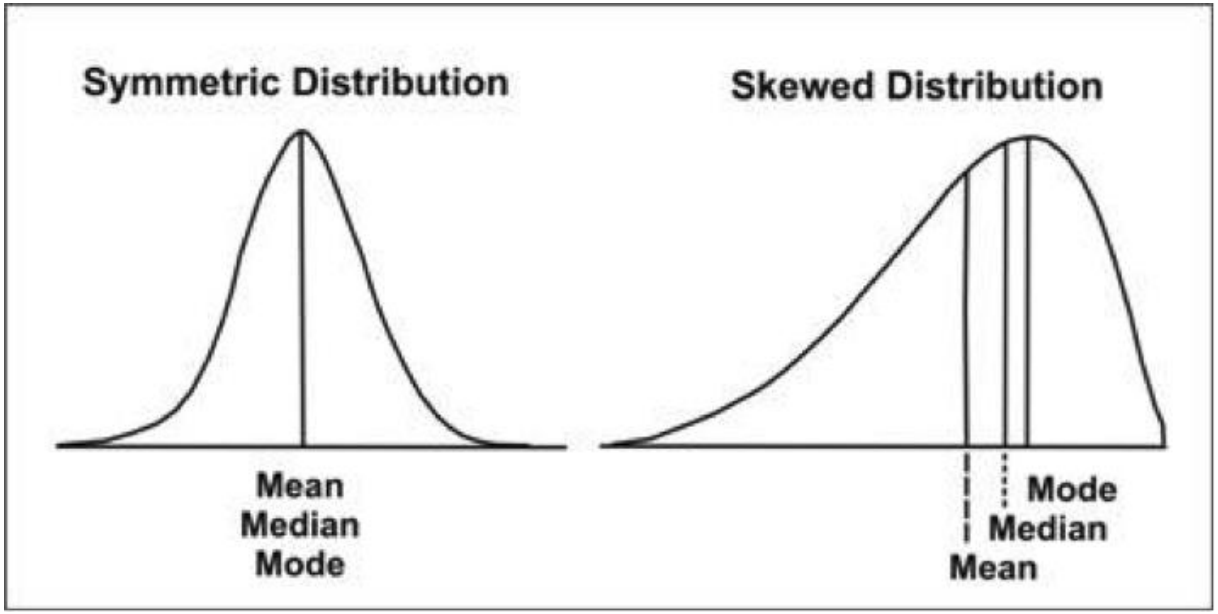


Figure from: https://soc.utah.edu/sociology3112/normal-distribution.php

You can calculate the mean and standard deviation for any set of numbers, but if the underlying PDF isn’t Gaussian, that information (mean and std. dev.) isn’t sufficient to describe the distribution

In Physics we like to assume things are Gaussian, but it’s not always true (“Assume a spherical cow…”)

Scatterplots and 2D Histograms

- What if you want to look at the distribution of two related variables?
- Scatterplot: let one variable be the x value and the other value be the y value; put a point on the plot for each pair of values
  - Allows you to see correlations
- 2D histogram: divide the x and y axis into ranges. For each bin (rectangular region defined by the intersection of an x interval and a y interval), count the number of entries (data points) in that bin. Use a color scale to represent the number of entries
  - When a scatterplot gets too dense, a 2D histogram may better allow you to compare different regions
  - Also use if you are doing a 2D binned fit of data (we'll talk more about this later)
- Back to the grade distribution example: let's consider the semester grade and the final exam grade for each student

Example: Make a 2D Histogram

Example: Plot in 3D

Summary

- Brief introduction to drawing graphs and histograms in Python
- Some best practices
  - Labels
  - Appropriate binning choices
- Discussion of histograms and probability density functions