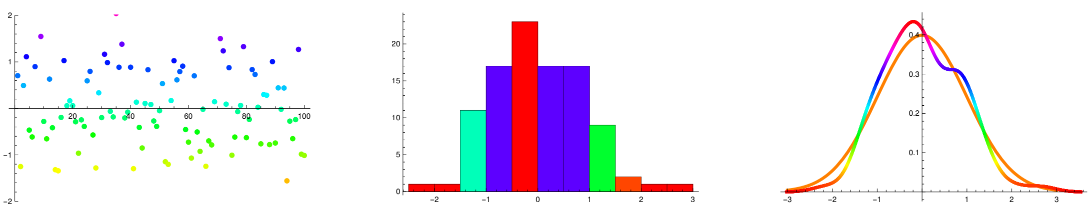
INTRODUCTION TO CALCULUS

MATH 1A

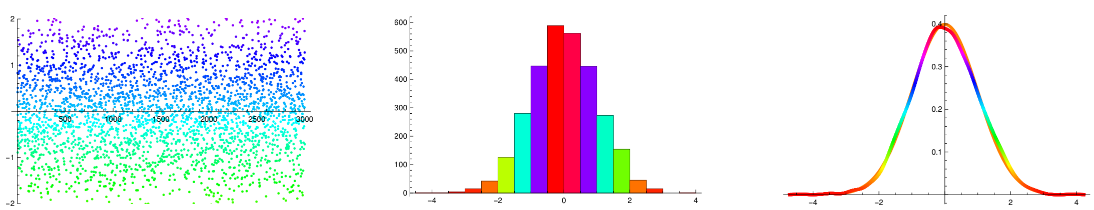
Unit 32: Statistics

32.1. Statistics describes data using functions called **probability distributions**. The topic allows us to review some integration. To see how one gets from data to functions, lets look at the following 100 data points. If we count, how many data values fall into a specific interval, we get a **histogram**. Smoothing this histogram and scaling so that the total integral is 1 produces a **probability distribution function**. This allows us to describe data, discrete sets of points with functions.



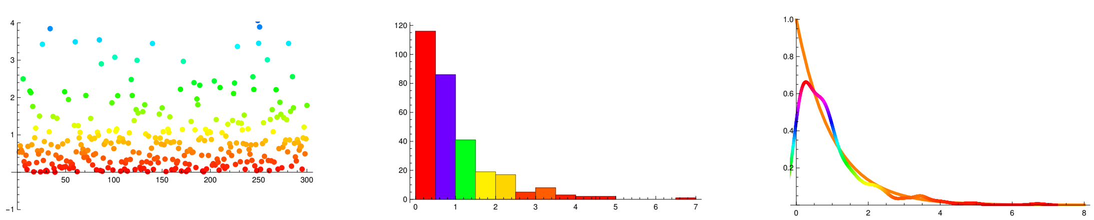
100 data points, the histogram and a smooth interpolation PDF.

32.2. If we take 3200 points, the data distribution has a bell curve shape. In the last picture, we also included the graph of $f(x) = e^{-x^2/2}/\sqrt{2\pi}$.

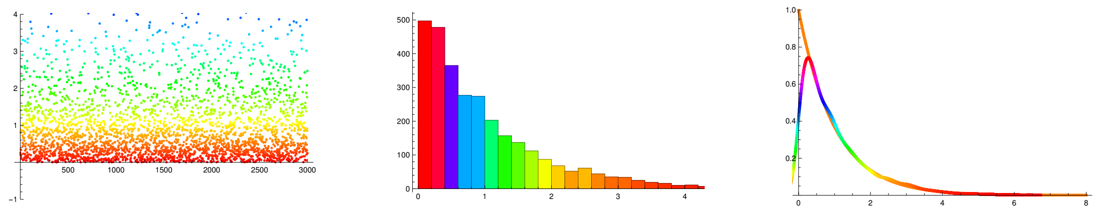


3200 data points, the histogram and a smooth interpolation PDF.

32.3. There are some data of “waiting times. These data are positive. We then draw the histogram and a smooth interpolation function. Lets do it again first for 320 data points and then for 3200 data points.



320 data points, the histogram and a smooth interpolation PDF



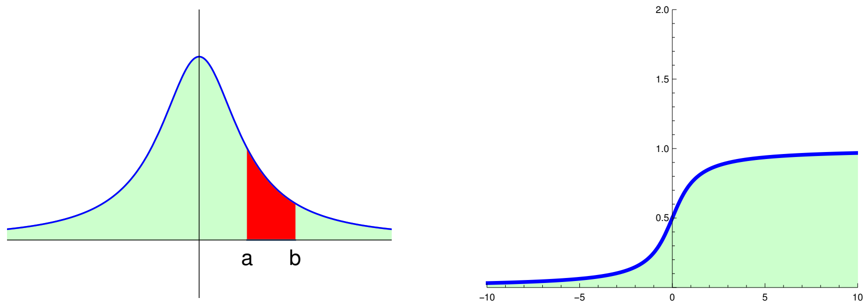
3200 other data points, the histogram and a smooth interpolation PDF

Definition: A **probability density function** is a piecewise non-negative continuous function f with the property $\int_{-\infty}^{\infty} f(x)\, dx = 1$. The anti derivative $F(x) = \int_{-\infty}^{x} f(t)\, dt$ is called the **cumulative distribution function**.

The cumulative distribution function is increasing from 0 to 1 as its derivative is non-negative.

Definition: **moments** of the PDF are integrals of the form

$$M_n = \int_{-\infty}^{\infty} x^n f(x)\, dx$$



PDF and CDF

32.4. For $n = 0$, we know $M_0 = 1$. The first moment M_1 is called the **expectation** or **average**:

Definition: The **expectation** of probability density function f is

$$m = \int_{-\infty}^{\infty} x f(x)\, dx \; .$$

32.5. The second moment allows us to get the **variance**:

Definition: The **variance** of probability density function f is

$$\mathrm{Var}(f) = \int_{-\infty}^{\infty} x^2 f(x)\, dx - m^2 \; ,$$

where m is the expectation. We can write $\mathrm{Var}(f) = M_2 - M_1^2$.

Definition: The integral $\mu_k = \int_{-\infty}^{\infty} (x-m)^k f(x)\, dx$ is called the k'th **central moment**. We have $\mu_2 = \mathrm{Var}$.

Definition: The square root σ of the variance is called the **standard deviation** It is the expected deviation from the mean.

32.6. From it, one can get the **normalized central moment** $C_k = \mu_k/\sigma^k$ which is $C_k = \int_{-\infty}^{\infty} (\frac{(x-m)}{\sigma})^k f(x)\,dx$. Computing moments, central moments and normalized central moments leads often to "integration by parts" problems:

The expectation of the geometric distribution $f(x) = e^{-x}$

$$\int x e^{-x}\,dx = 1 \,.$$

The variance of the geometric distribution $f(x) = e^{-x}$ is 1 and the standard deviation 1 too. To see this, let us compute

$$\int_0^{\infty} x^2 e^{-x}\,dx \,.$$

x^2	e^{-x}	
$2x$	$-e^{-x}$	$\oplus$
2	e^{-x}	$\ominus$
0	e^{-x}	$\oplus$

You have already computed the expectation of the standard Normal distribution $f(x) = (2\pi)^{-1/2} e^{-x^2/2}$

$$\frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} x e^{-x^2/2}\,dx = 0 \,.$$

The variance of the standard Normal distribution $f(x)$ is $\frac{1}{\sqrt{2\pi}}$ times

$$\int_{-\infty}^{\infty} x^2 e^{-x^2/2}\,dx \,.$$

We can compute this integral by partial integration too but we have to split it as $u = x$ and $v = xe^{-x^2/2}$.

$$-xe^{-x^2/2}|_{-\infty}^{\infty} + \int_{-\infty}^{\infty} e^{-x^2/2}\,dx = \sqrt{2\pi} \,.$$

The variance therefore is $\boxed{1}$.

32.7. Example: The distribution on $[-1,1]$ with function $(1/\pi)(1-x^2)^{-1/2}$ there and 0 everywhere else is called the **arcsin-distribution**. What is the cumulative distribution function? What is the mean m? What is the standard deviation σ? As we will see in class, $m = 0, \sigma = 1/\sqrt{2}$.

Homework

Problem 32.1: The function $f(x) = \cos(x)/2$ on $[-\pi/2, \pi/2]$ is a probability density function. Its mean is 0. Find its variance $\int_{-\pi/2}^{\pi/2} x^2 \cos(x)\, dx/2$.

Solution:
It is piecewise continuous, non-negative and the integral from $[-\pi/2, \pi/2]$ is equal to 1. The mean is 0 so that the variance is $M_2 = \int_{-\pi/2}^{\pi/2} \cos(x)x^2\ dx/2 = (\pi^2 - 8)/4$. This is 0.51...

Problem 32.2: The **uniform distribution on** $[0, 1]$ is a distribution with probability density function is $f(x) = 1$ for $0 \leq x \leq 1$ and 0 elsewhere. Compute: a) the n'th moment M_n, b) the variance $\mathrm{Var}[f] = M_2 - M_1^2$, and c) the standard deviation $\sigma = \sqrt{\mathrm{Var}[f]}$.

Solution:
a) Just integrate $x^n/(5-1)$ from a to b $(5^{1+n} - 1)/(4(1+n))$.
b) $4/3$ and $\sqrt{4/3}$.

Problem 32.3: We have seen in lecture 27 that the **sigmoid function** $F(x) = (\tanh(\frac{x}{2}) + 1)/2$ has the derivative $f(x) = \frac{1}{4\cosh^2(\frac{x}{2})}$. It is called the **logistic distribution**. You have already checked that $\int_{-\infty}^{\infty} f(x)\ dx = 1$. Use both a computer algebra system as well as an AI tool to compute.
a) The variance $\int_{-\infty}^{\infty} x^2 f(x)\ dx$ and the entropy $-\int_{-\infty}^{\infty} f(x) \log(f(x))\ dx$

1

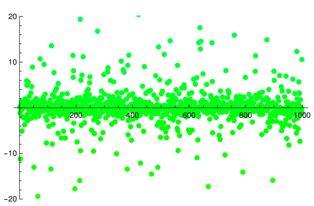
Solution:
a) $\pi^2/3$
b)

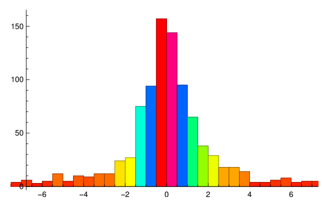
[1] In both cases, give the entry you entered into the computer algebra system like wolfram alpha and the prompt you used to extract the number from a tool like chat gpt 4.

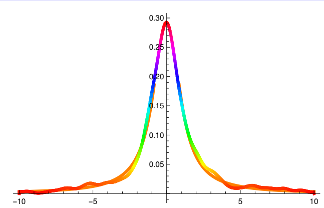
Problem 32.4: a) Verify again that the **Cauchy distribution** with PDF $f(x) = \frac{(1/\pi)}{x^2+1}$ has the CDF $F(y) = 1/2 + \arctan(y)/\pi$.
b) What can you say about the variance of this distribution?

Solution:
It is continuous, non-negative and the total integral is 1 as the anti-derivative is $\arctan(x)/\pi + 1/2$.

Problem 32.5: If we take random numbers x_k in $[0,1]$ then $\tan(\pi x_k)$ are Cauchy distributed. a) What is the probability that such a random number hits $[-1,1]$? (Hint: it is $F(b) - F(a)$ where F is the function in 31.4).
b) Use a calculator (or your favorite program) to compute at least 20 Cauchy distributed random numbers. How many hit the interval $[-1,1]$? How many would you have expected to hit? (The picture below does that for a few hundred data). In Mathematica these numbers can be accessed by Tan[Pi $*$ Random[]].







Solution:
The probability is $1/4 + 1/4 = 1/2$. We would expect 10 to hit the interval $[-1,1]$. More explanation: The probability that $y = \tan(\pi x)$ is in $[a,b]$ is the same than that $x = \arctan(y)/\pi$ is in $[\arctan(a)/\pi, \arctan(b)/\pi]$. But since the distribution of x is uniform in $[0,1]$ this probability is $\arctan(b)/\pi - \arctan(a)/\pi$. That is the same than $F(b) - F(a)$ for the CDF F of the Cauchy distribution f given in 4). Here is a program: X := Tan[Pi Random[]]; Total[Table[If[Abs[X] ¡ 1, 1, 0], 1000]]

Oliver Knill, knill@math.harvard.edu, Math 1a, Spring, 2024