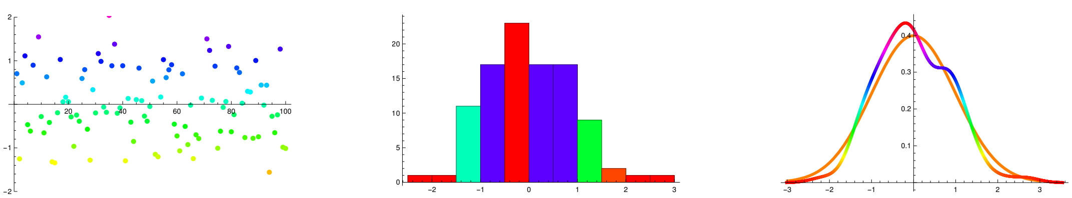
INTRODUCTION TO CALCULUS

MATH 1A

Unit 30: Statistics

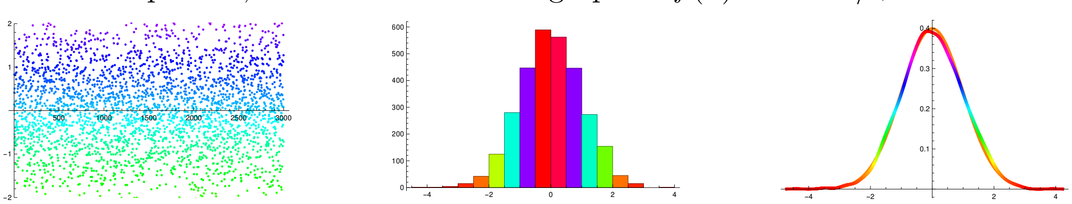
Functions

30.1. Statistics describes the distribution of data using 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.



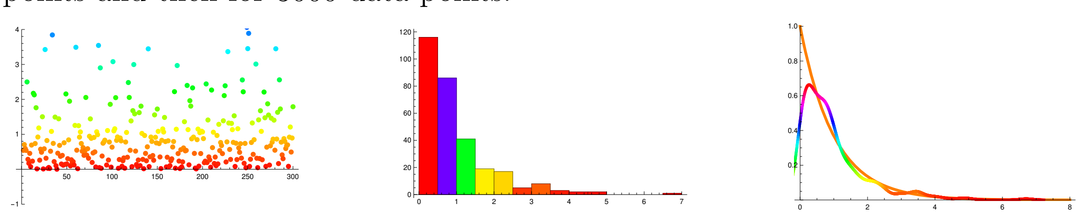
100 data points, the histogram and a smooth interpolation PDF.

30.2. Let us now look at 3000 points. The data distribution has a bell curve shape. In the last picture, we also included the graph of $f(x) = e^{-x^2/2}/\sqrt{2\pi}$.

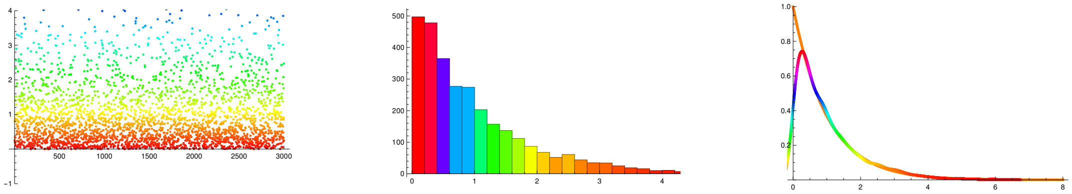


3000 data points, the histogram and a smooth interpolation PDF.

30.3. 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 300 data points and then for 3000 data points.



300 data points, the histogram and a smooth interpolation PDF



3000 other data points, the histogram and a smooth interpolation PDF

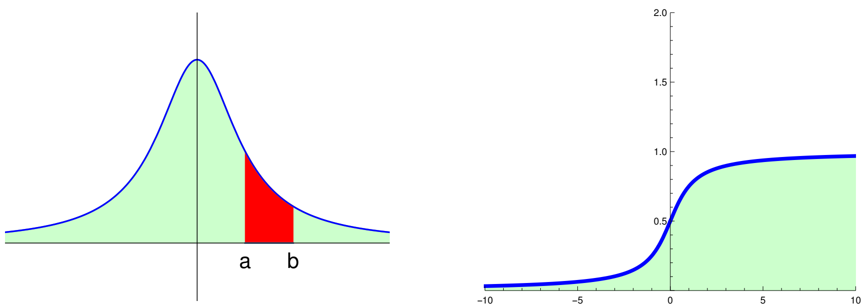
Integration

30.4. We have already defined the probability density function f called PDF and its anti-derivative $F(x)$, the cumulative distribution function CDF.

Definition: Recall that a probability density function is a piecewise continuous function f satisfying $\int_{-\infty}^{\infty} f(x)\,dx = 1$ and which is ≥ 0 everywhere.

Definition: Of great interest are **moments** of the PDF. These are integrals of the form

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



PDF and CDF

30.5. For $n = 0$, we know the answer is always $M_0 = 1$. The first moment M_1 is the expectation or average:

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

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

30.6. The second moment allows us to get the **variance** which is of great importance:

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: More generally, one can look at $\mu_k = \int_{-\infty}^{\infty} (x-m)^k f(x)\, dx$ which is called the k'th **central moment**. We have $\mu_2 = \text{Var}$.

Definition: The square root σ of the variance is called the **standard deviation**.

30.7. The standard deviation tells us what deviation we expect from the mean. 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:

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

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

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

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

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

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

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

Example: 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 $\boxed{1}$.

30.8. The next example is for trig substitution:

Example: The distribution supported 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? We will compute this in class. The answers are $m = 0, \sigma = 1/\sqrt{2}$.

Homework

Problem 30.1: The function $f(x) = \cos(x)/2$ on $[-\pi/2, \pi/2]$ is a probability density function. Its mean is 0. Find its variance $\int_{-\pi/2}^{\pi/2} x^2 \cos(x)\, dx$.

Problem 30.2: The **uniform distribution on** $[a, b]$ is a distribution with probability density function is $f(x) = 1/(b-a)$ for $a \leq x \leq b$ and 0 elsewhere. Let $a = 1$ and $b = 5$;
a) Find the n'th moment $M_n = \int_{-\infty}^{\infty} x^n f(x)\, dx$ in general.
b) Now compute the variance $\text{Var}[f] = M_2 - M_1^2$ and the standard deviation $\sigma = \sqrt{\text{Var}[f]}$.

Problem 30.3: Define $f(x)$ to be 0 for $x < 0$ and for $x > 0$ to be

$$f(x) = \frac{1}{\log(2)} \frac{e^{-x}}{1 + e^{-x}} .$$

a) Find the CDF F and verify that f is a probability density function. (Note that $F(x) = 0$ for $x \leq 0$ and especially $F(0) = 0$.)
b) Use a computer to numerically compute the expectation $m = M_1$.
c) Use a computer to compute the second moment M_2.
d) What is the standard deviation $\sigma = \sqrt{M_2 - M_1^2}$ of the distribution?
P.S. Your computer algebra system might tell you $M_1 = \zeta(2)/2, M_2 = 3\zeta(3)/2$. In general $M_n = \zeta(1+n)n!(2^n - 1)/2^n$ for the **zeta function**.

Problem 30.4: 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?

Problem 30.5: Let us quickly verify that if we take random numbers x in $[0, 1]$ then the data $\tan(x)$ are Cauchy distributed: just check that the probability that $y = \tan(\pi x)$ is in $[a, b]$ is $F(b) - F(a)$ for the Cauchy CDF appearing in the last problem. Now, use a calculator and compute 10 random Cauchy distributed numbers. In Mathematica such numbers can be accessed by Tan[Pi ∗ Random[]].