§1 Random Variables

Note Title 14/10/2011

§1.1 Introduction

Example use: I've developed a new load-balancing algorithm for a web server. I want to test my algorithm, by means of simulation. My simulator needs a random number generator, to generate file sizes, request times etc. The performance of my algorithm will probably depend on the random number generator I use. How should I program this random number generator?

We would use random number generators in situations like this because the world is too complicated for us to model it in a Newtonian cause-and-effect system. Even though there might be deterministic explanations for every little variation in file size etc., it's too hard to take account of them all. Instead, we use random numbers to say "There is variability, and I can quantify the degree of variability, but I'm not going to look in excruciating detail for causes for every little variation."

Typically, we take real-world measurements, we look at the data, and we try to program a random number generator that produces output consistent with the data.

Perhaps the real-world measurements show a range of behaviours, e.g. web requests arrive close together at peak times, far apart at off-peak. How should I make my random number generator tunable, to capture this range?

Many standard random number generators come with tunable parameters. Typically, we look at the real-world measurements and try to estimate what values of the tuning parameters give the best fit.

Then, we can use the simulator to ask: how does the performance of my algorithm depend on these parameters? In some cases we don't even need to run the simulator — we can use maths to calculate the performance.

A <u>random sample</u> is a collection of numbers, i.e. a dataset, the output from a particular experiment or simulation.

e.g. "The first 7 requests at my webserver on 10 Oct were 3, 7, 2, 2, 108, 15, 1 kB big. Of these, only two were bigger than 10 kB. The average (mean) size was 19.7 kB."

A <u>random variable</u> is what is produced by a random number generator, i.e. it's a piece of code or a thought experiment. Typically it has tunable parameters.

e.g.

```
def rexp(r):
    return (-1.0/r) * math.log(random.random())
```

Let X be the output of rexp(3). Then, the probability that X is greater than 2 is e^{-6}, i.e. $\mathbb{P}(X>2) = e^{-6}$.

e.g. Let X be the output of rexp(λ). The expectation (mean value) of X is $\mathbb{E}(X) = \frac{1}{\lambda}$.

e.g. Let X be the outcome of throwing a fair six-sided dice. Then, $\mathbb{P}(X \text{ is even}) = \frac{1}{2}$, $\mathbb{P}(X \text{ is even} \mid X \leq 3) = \frac{1}{3}$.

NOTE: X doesn't "have" any particular value. X is a standin for "all the values you might get by calling the random number generator, weighted by their probabilites".

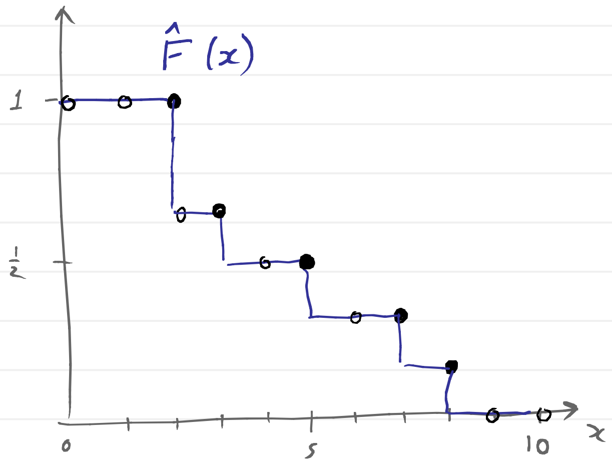
NOTE: There may be two different pieces of code that produce indistinguishable output. (Or course the output is random, so the two may be indistinguishable but not identical.) In this case we'd say that there are two different random number generators, but that the random variables that they output are the same.

§1.2 Visualizing random samples & variables

Example I take real-world measurements of some quantity, and get sample values 3, 8, 7, 2, 2, 5. A good way to illustrate this is by plotting the empirical tail distribution function

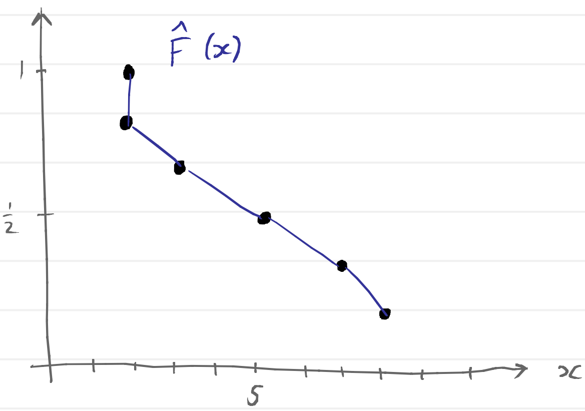
$$\hat{F}(x) = \frac{\text{number of readings that are } \geq x}{\text{total \# of readings}}.$$

x	# of readings that are $\geq x$	$\hat{F}(x)$
0	6	1
1	6	1
2	6	1
2.001	4	4/6
3	4	4/6
4	3	3/6
5	3	3/6
6	2	2/6
7	2	2/6
8	1	1/6
9	0	0



The only "interesting" points on this graph are the points at which it steps down. In practice, it's more convenient to only tabulate & plot those points. Also, if your plotting system doesn't do "step-style" curves, it's fine to simply connect the points. If you have a large enough sample (many more than 6 values) the difference will be negligible.

x	# of readings that are $\geq x$	$\hat{F}(x)$
2	6	1
2	5	5/6
3	4	4/6
5	3	3/6
7	2	2/6
8	1	1/6



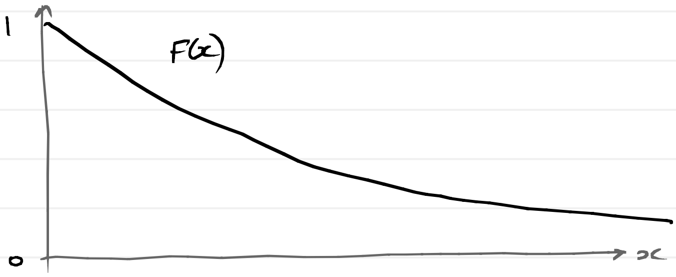
This row is a bit of a cheat. Really, it should be $\hat{F}(2) = 1$. But by using these values, (a) it's easier to generate the values in this table, and (b) I get a "step down" at 2, like the graph at the top of the page.

<u>Example</u> I have a random number generator

```
def rexp(r):
    return (-1.0/r) * math.log(random.random())
```

I have been told that, if X is the output of rexp(r), then
$$P(X \geq x) = e^{-rx}.$$

The function $F(x) = P(X \geq x)$ is called the <u>tail distribution function</u> of the random variable X.

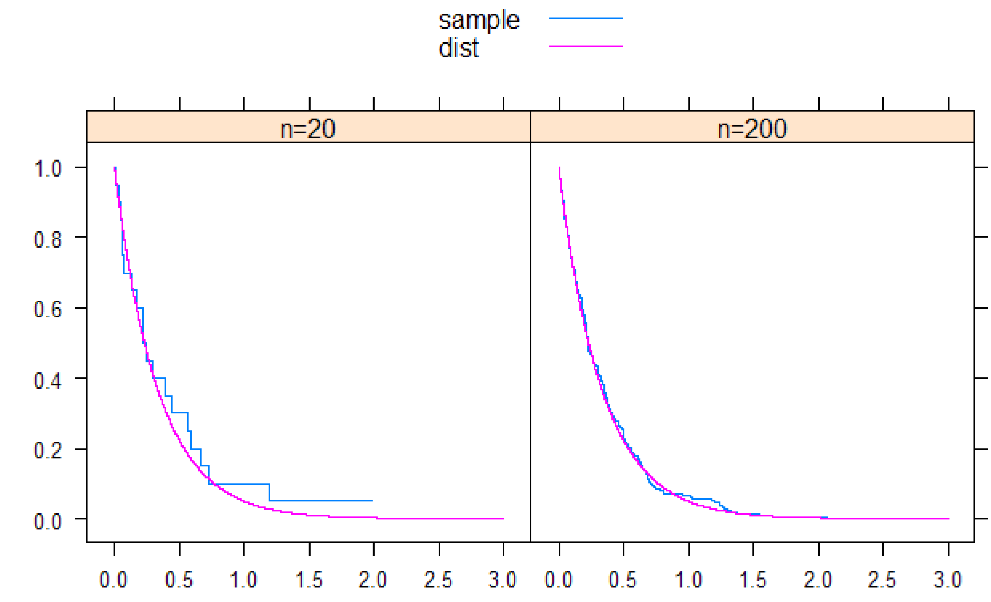


THE LAW OF LARGE NUMBERS.

This says: if you generate a big enough sample from a random number generator, then the sample's empirical distribution function and the RNG's distribution function should be very close. So, if you want to plot the distribution function for an RNG but don't know its formula, it doesn't matter — just generate a large sample from your RNG, and plot its empirical distribution.

Suppose I have a random number generator with tail distribution function $F(\cdot)$, and I generate a random sample of n values, and, for some arbitrary x, I count how many are $\geq x$. Call this N.

If n is large, then $N \approx n\,P(X \geq x) = n\,F(x)$
$$\Rightarrow \frac{N}{n} \approx F(x) \Rightarrow \hat{F}(x) \approx F(x).$$



CONTINUOUS RANDOM VARIABLES

If the random number generator gives floating point values (and all values are possible), it is called a <u>continuous random variable</u>, and the distribution function $F(\cdot)$ is continuous.

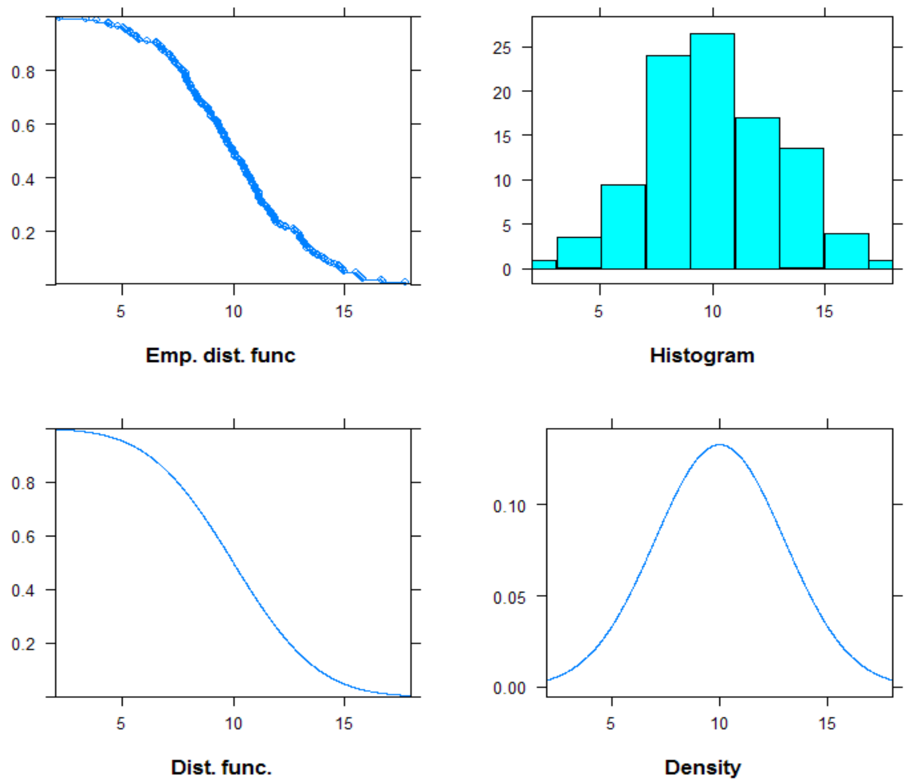
The <u>density</u> is $f(x) = -\frac{d}{dx} F(x)$. Equivalently, $F(x) = \int^{x} f(y)\,dy$.

For continuous random variables, the probability of any specific value is always 0, ie $\mathbb{P}(X = x) = 0$ for all x. This means

$$\mathbb{P}(X > x) = \mathbb{P}(X \geq x) - \mathbb{P}(X = x) = \mathbb{P}(X \geq x).$$

The law of large numbers tells us that if we generate a very large sample, then the sample's empirical distribution function will match up with the random variable's distribution function.

Furthermore, if we generate a very large sample and plot a histogram of it, and make the bin sizes small enough, then the histogram will match up with the random variable's density function.
(except that the vertical axis is different).



I don't like histograms. If you choose bin sizes wrong, you get misleading pictures. The empirical distribution function is nicer, and it fits more bits of information on the plot.

DISCRETE RANDOM VARIABLES

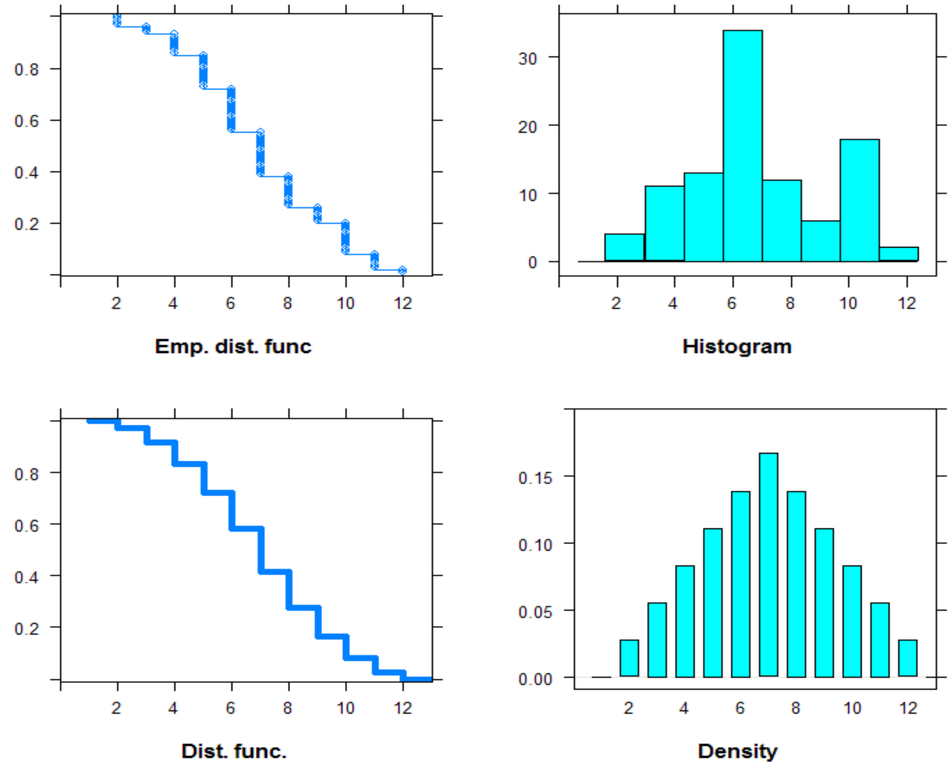
If the random number generator gives integer values, or values in some enumerable set (e.g. {♥, ♣, ♦, ♠}) then it is called a <u>discrete random variable</u>.

The <u>density</u> is $\Pi_x = \mathbb{P}(X = x)$. Sometimes called the <u>distribution</u>.
We can plot distribution function, histogram & density as before —
and the histogram won't be misleading if you have one bin per possible x-value.

e.g. I throw two dice, and sum up their values. Let X be the sum. Its distribution is

$$\Pi_2 = \frac{1}{36}, \quad \Pi_3 = \frac{2}{36}, \quad \Pi_4 = \frac{3}{36}, \quad \Pi_5 = \frac{4}{36}, \quad \Pi_6 = \frac{5}{36}, \quad \Pi_7 = \frac{6}{36},$$

$$\Pi_8 = \frac{5}{36}, \quad \Pi_9 = \frac{4}{36}, \quad \Pi_{10} = \frac{3}{36}, \quad \Pi_{11} = \frac{2}{36}, \quad \Pi_{12} = \frac{1}{36}.$$



This histogram is misleading because it doesn't have one bin per possible x-value.

§1.3 Generating Random Variables

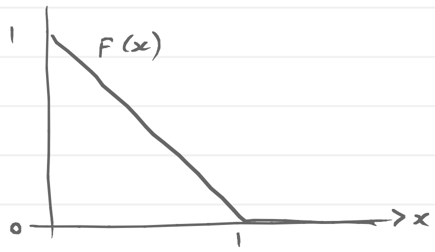
Exercise 1 Consider the standard Uniform $[0,1]$ random number generator, called random.random() in Python and runif(·) in R.
What is its distribution function?

$F(\frac{1}{2}) = \mathbb{P}(X \geq \frac{1}{2}) = \frac{1}{2}$, since it's just as likely to be $> \frac{1}{2}$ as $< \frac{1}{2}$.

$F(\frac{3}{4}) = \mathbb{P}(X \geq 3/4) = \frac{1}{4}$, since only a quarter of the range is $\geq 3/4$, and all parts of the range are equally likely.

$F(x) = \mathbb{P}(X \geq x) = 1-x$, we come to after some thought.

Also, $\mathbb{P}(X < x) = x$.



Technically speaking,

$$F(x) = \begin{cases} 1 & \text{if } x < 0 \\ 0 & \text{if } x > 1 \\ 1-x & \text{else} \end{cases}$$

But we don't normally bother with this pedantry.

Exercise 2 Let X be generated by the following random number generator

```
def myrng():
    Y = random.random()
    return Y*Y
```

What is its distribution function?

from the previous exercise.

$$F(x) = \mathbb{P}(X \geq x) = \mathbb{P}(Y*Y \geq x) = \mathbb{P}(Y \geq \sqrt{x}) = 1-\sqrt{x}.$$

THE INVERSION METHOD FOR GENERATING RANDOM VARIABLES

Given a distribution function $F(\cdot)$, how can we generate a random variable X with that distribution, ie with $\mathbb{P}(X \geqslant x) = F(x)$?

Often, we can achieve this by inverting the method used in Exercise 2.

Step 1. Generate $U \sim \text{Uniform}[0,1]$

Step 2. Find X such that $F(X) = U$.

Step 3. Use this X as the random variable we want.

Example Suppose we want to generate $X \sim \text{Exp}(\lambda)$, with $\mathbb{P}(X \geqslant x) = e^{-\lambda x}$.

1. Generate $U \sim \text{Uniform}[0,1]$
2. Solve $e^{-\lambda X} = U \Rightarrow -\lambda X = \log U \Rightarrow X = -\frac{1}{\lambda} \log U$.
3. Use $X = -\frac{1}{\lambda} \log U$ as our $\text{Exp}(\lambda)$ random variable.

In Python,

```
def myrng(λ):
    U = random.random()
    return (-1.0/λ) * math.log(U)
```

Example [non-examinable]

Why does this method work? Here is a derivation which shows you where the general method comes from,

Suppose we want $\mathbb{P}(X \geqslant x) = e^{-\lambda x}$.

First, invert the distribution function:

Let $y = e^{-\lambda x} \Rightarrow \log y = -\lambda x \Rightarrow x = -\frac{1}{\lambda} \log y$.

The basic equation we want can be rewritten in terms of y: we want

$\mathbb{P}(X \geqslant -\frac{1}{\lambda} \log y) = y. \Rightarrow \mathbb{P}(-\lambda X \leqslant \log y) = y \Rightarrow \mathbb{P}(e^{-\lambda X} \leqslant y) = y.$

We know that if $U \sim \text{Uniform}[0,1]$ then $\mathbb{P}(U \leqslant y) = \mathbb{P}(U < y) = y$.

So, to solve the equation we want, all we need is

$$e^{-\lambda X} = U \quad \Rightarrow \quad X = \frac{1}{\lambda} \log U.$$

SPECIAL CASE: THE BOX-MULLER METHOD (not examinable)

There are many cases where the distribution function can't be inverted. One example is the Normal distribution. For this case, there is a simple alternative.

Let U, V be independent Uniform$[0,1]$ random variables.

let $X = \sqrt{-2\log U}\cos(2\pi V)$

$Y = \sqrt{-2\log U}\sin(2\pi V)$

Then X, Y are independent Normal$(0,1)$ random variables.

DISCRETE RANDOM VARIABLES.

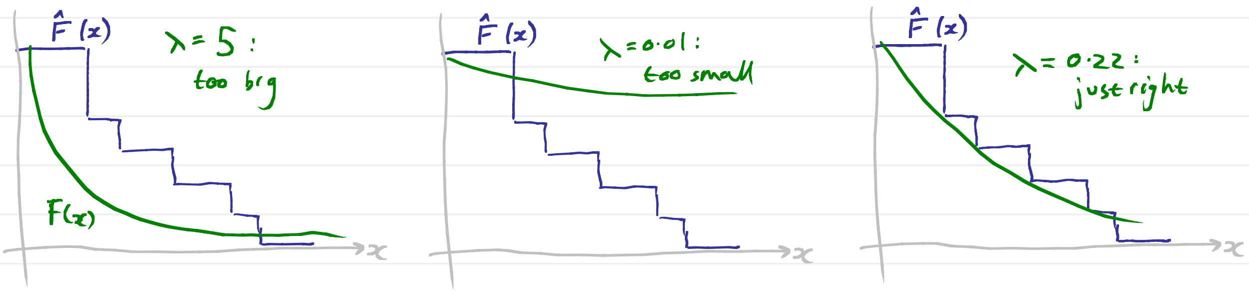
There are no good general-purpose methods for generating discrete random variables. It's generally up to you to invent a method for the distribution you need, or to look it up on wikipedia.

§1.4 Fitting Distributions

Suppose we have a collection of real-world measurements 3, 8, 7, 2, 2, 5.
Suppose I want a random number generator to mimic these outputs, and suppose I've settled on using an Exponential random variable, ie the random variable with distribution function $F(x) = e^{-\lambda x}$ and density function $f(x) = -\frac{d}{dx} F(x) = \lambda e^{-\lambda x}$.

This distribution depends on a parameter λ, and I need to pick a suitable λ.

I could plot the empirical distribution function $\hat{F}(x) = \frac{1}{6}$(# of measurements that are ≥ 6) theoretical distribution function $F(x) = \mathbb{P}(X \geq x) = e^{-\lambda x}$ for the Exponential distribution, and I could tweak λ until the two curves match up.



There is a systematic procedure for doing this:

1. Write down the density function $f_p(x)$ of the random variable whose parameters you want to fit. The density function depends on those parameters (there may be more than one). Here I've written p to denote "parameters to fit".

2. Write out $lik(p) = f_p(x_1) \times f_p(x_2) \times \cdots \times f_p(x_n) = \prod_{i=1}^{n} f_p(x_i)$.
where $x_1, \ldots, x_n$ are the n values in your sample.

3. Find the value of p that maximizes $lik(p)$, call it $\hat{p}$.
Often it's easier to maximize $\log(lik(p))$. This must give the same value of $\hat{p}$

4. This value $\hat{p}$ is the <u>maximum likelihood estimator</u> of p.
It gives the best-fitting distribution.

Example I observe bus inter-arrival times of 2 min, 10 min, 3 min, 8 min, 7 min.
I suspect these are independent $\sim \mathrm{Exp}(\lambda)$ random variables, and I want to estimate λ.

1. The density function of $\mathrm{Exp}(\lambda)$ is $f_\lambda(x) = \lambda e^{-\lambda x}$.

2. The likelihood function is

$$\mathrm{lik}(\lambda) = (\lambda e^{-\lambda \times 2})(\lambda e^{-\lambda \times 10})(\lambda e^{-\lambda \times 3})(\lambda e^{-\lambda \times 8})(\lambda e^{-\lambda \times 7})$$

$$= \lambda^5 e^{-30\lambda}$$

3. We'll choose λ to maximise $\log \mathrm{lik}(\lambda) = 5 \log \lambda - 30\lambda$:

$$\frac{d}{d\lambda} \log \mathrm{lik}(\lambda) = 0 \quad \Rightarrow \quad \frac{5}{\lambda} - 30 = 0 \quad \Rightarrow \quad \lambda = \frac{5}{30} = \frac{1}{6}.$$

4. Our maximum likelihood estimator is $\hat{\lambda} = \frac{1}{6}$

Note the units: λ must have units $[\frac{1}{\mathrm{min}}]$ since the density function is $\lambda e^{-\lambda x}$ and x is measured in [min] and you're only allowed to take exponentials of a unitless quantity. So really $\hat{\lambda} = \frac{1}{6} / \mathrm{min}$.

§1.5 Describing distributions

Here is some standard terminology for summarizing a distribution — describing the average and variability of a random variable.

The "average"

The expected or mean value of X is
$$\mathbb{E}X = \begin{cases} \sum_{\text{all values } x} x\,\mathbb{P}(X=x) & \text{when } X \text{ is discrete} \\ \int_{\text{entire range}} x\, f(x)\,dx & \text{when } X \text{ is continuous.} \end{cases}$$

The "typical value"

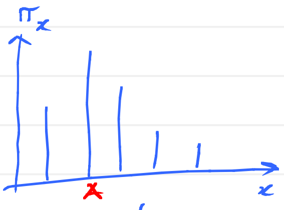
The median of X is x such that $\mathbb{P}(X \geq x) = \mathbb{P}(X \leq x) = \frac{1}{2}$.

NOTE. For discrete random variables, it may not be possible to find such an x exactly. In that case, use the closest x you can.

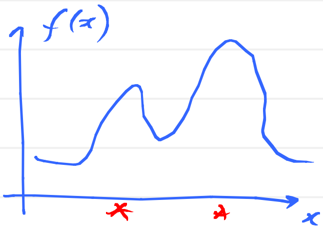
The "most likely value"

The mode of X is a local maximum on the density graph

There may be more than one mode



a unimodal rand.var.



a bimodal rand.var.

The "variability"

The variance of X is $\operatorname{Var}X = \mathbb{E}\left[(X-\mu)^2\right] = \begin{cases} \sum_{\text{all values } x} (x-\mu)^2\, \mathbb{P}(X=x) & \text{when } X \text{ is discrete} \\ \int_{\text{entire range}} (x-\mu)^2 f(x)\,dx & \text{when } X \text{ is continuous} \end{cases}$

where $\mu = \mathbb{E}X$.

The standard deviation of X is $\operatorname{sd}(X) = \sqrt{\operatorname{Var}X}$.

Likely ranges

The first quartile is a number x such that $\mathbb{P}(X \leq x) = 25\%$

median —— $\mathbb{P}(X \leq x) = 50\%$

third quartile —— $\mathbb{P}(X \leq x) = 75\%$

p-percentile —— $\mathbb{P}(X \leq x) = p$

The range $[x_1, x_2]$ is a 95% confidence interval if $\mathbb{P}(x_1 \leq X \leq x_2) = 95\%$.

Often we choose a two-sided confidence interval with $\mathbb{P}(X < x_1) = \mathbb{P}(X > x_2) = 2.5\%$.

In other contexts it may be useful to report a one-sided confidence interval — either an upper confidence interval $[x_1, \infty)$ or a lower confidence interval $(\infty, x_2]$

Confidence intervals are a way to express the variability of a random variable, rather like the standard deviation — but std.dev is often easier to calculate with.

We use the same words to describe a random sample. Use the same definitions as for a discrete random variable, but use "Fraction of the sample for which $X \leq x$" rather than "$\mathbb{P}(X \leq x)$", and "Fraction of sample for which $X = x$" rather than $\mathbb{P}(X=x)$.

§1.6 Independence

The concept of independent random variables is absolutely fundamental in modeling. Random variables X and Y are independent if knowing the value of one of them gives us no information about the value of the other. (Typically we assume that different users make independent requests, but that requests from a single user are not independent.)

Defn Two random variables X and Y are independent if

$$\mathbb{P}(X \geq x \text{ and } Y \geq y) = \mathbb{P}(X \geq x)\,\mathbb{P}(Y \geq y) \quad \text{for all } x \text{ and } y$$

Or, equivalently,

$$\mathbb{P}(X \geq x \mid Y \geq y) = \mathbb{P}(X \geq x). \quad \text{for all } x, y$$

Or, for discrete random variables, all we need is

$$\mathbb{P}(X = x \text{ and } Y = y) = \mathbb{P}(X=x)\,\mathbb{P}(Y=y). \quad \text{forall } x,y$$

or $\mathbb{P}(X = x \mid Y = y) = \mathbb{P}(X = x)$ for all x, y. (*)

e.g. I throw a fair die. Let Z be the result.

Let $X = Z$ mod 2 (ie $X=0$ if $Z = 2,4$ or 6; $X=1$ if $Z=1,3$ or 5)

let $Y = Z$ div 3 (ie $Y=0$ if $Z=1$ or 2; $Y=1$ if $Z=3,4$ or 5; $Y=2$ if $Z=6$)

Are X and Y independent?

Let's use the discrete equation (*).

We need to run through all values of x and y, and check that (*) holds.

- Try $x=0$, $y=0$.

  $\mathbb{P}(X=0 \text{ and } Y=0) = \mathbb{P}(Z=2,4 \text{ or } 6 \text{ and } Z = 1 \text{ or } 2) = \mathbb{P}(Z=2) = \frac{1}{6}.$

  $\mathbb{P}(X=0)\,\mathbb{P}(Y=0) = \mathbb{P}(Z=2,4 \text{ or } 6)\,\mathbb{P}(Z=1 \text{ or } 2) = \frac{1}{2}\times\frac{1}{3} = \frac{1}{6}.$

  So these

- Try $x=0$, $y=1$.

  $\mathbb{P}(X=0 \text{ and } Y=1) = \mathbb{P}(Z=2,4 \text{ or } 6 \text{ and } Z=3,4 \text{ or } 5) = \mathbb{P}(Z=4) = \frac{1}{6}$

  $\mathbb{P}(X=0)\,\mathbb{P}(Y=1) = \mathbb{P}(Z=2,4 \text{ or } 6)\,\mathbb{P}(Z=3,4 \text{ or } 5) = \frac{1}{2}\times\frac{1}{2} = \frac{1}{4}.$

  So these values of x and y fail the test.

Thus, X and Y are not independent.

Exercise. Let $X = (Z-1)$ mod 2 and $Y = (Z-1)$ div 2. Are X and Y independent?

Warning If X and Y are independent, it does NOT follow that

$\mathbb{P}(X = x \text{ or } Y = y) = \mathbb{P}(X=x) + \mathbb{P}(Y=y).$

§1.7 Working with random variables

For any random variable X,

$$\begin{aligned} \mathbb{E}(aX+b) &= a(\mathbb{E}X)+b && \text{for all constants } a \text{ and } b \\ \mathrm{Var}(aX+b) &= a^2 \,\mathrm{Var}\, X. && \text{"} \\ \mathrm{sd}(aX+b) &= a \,\mathrm{sd}(X) && \text{"} \end{aligned}$$

For any two random variables X and Y,

$$\mathbb{E}(X+Y) = (\mathbb{E}X) + (\mathbb{E}Y) \qquad (*)$$

For any two independent random variables X and Y,

$$\begin{aligned} \mathbb{E}(XY) &= (\mathbb{E}X)(\mathbb{E}Y) \\ \mathrm{Var}(X+Y) &= \mathrm{Var}\, X + \mathrm{Var}\, Y \\ \mathrm{sd}(X+Y) &= \sqrt{\mathrm{sd}(X)^2 + \mathrm{sd}(Y)^2} \end{aligned}$$

For two random variables X and Y which are not independent, we measure how related they are by the <u>covariance</u>

$$\mathrm{Cov}(X,Y) = \mathbb{E}\left[(X-\mathbb{E}X)(Y-\mathbb{E}Y)\right].$$

or by the correlation

$$\mathrm{Corr}(X,Y) = \frac{\mathrm{Cov}(X,Y)}{\mathrm{sd}(X)\,\mathrm{sd}(Y)}.$$

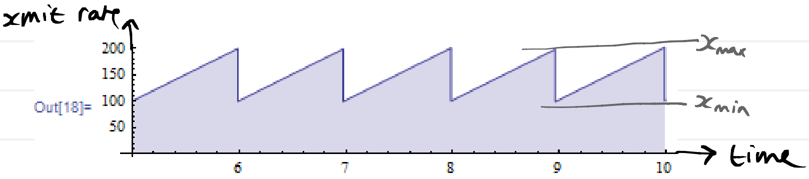
With a little algebra, we find

$$\begin{aligned} \mathrm{Var}(X+Y) &= \mathbb{E}\left[X+Y - \mathbb{E}(X+Y)\right]^2 \\ &= \mathbb{E}\left[X+Y-(\mathbb{E}X+\mathbb{E}Y)\right]^2 && \text{by } (*) \\ &= \mathbb{E}\left[(X-\mathbb{E}X)+(Y-\mathbb{E}Y)\right]^2 && \text{by rearranging} \\ &= \mathbb{E}(X-\mathbb{E}X)^2 + \mathbb{E}(Y-\mathbb{E}Y)^2 + 2\mathbb{E}(X-\mathbb{E}X)(Y-\mathbb{E}Y) && \text{by algebra} \\ &= \mathrm{Var}\, X + \mathrm{Var}\, Y + 2\,\mathrm{Cov}(X,Y). && \text{by definition of Var, Cov.} \end{aligned}$$

Exercise. Example sheet 2 has a variety of questions to build up your skill at working with random variables.

Example: Statistical multiplexing

For most traffic flows on the Internet, the rate at which data is transmitted is controlled by TCP. This algorithm steadily increases transmission rate (by 1pkt/sec every round-trip time) until it detects congestion in the form of a dropped packet, whereupon it cuts its transmission rate in half. This behaviour produces the characteristic "TCP sawtooth".



It's more convenient to think in terms of average xmit rate. Suppose that packet drops are periodic, hence xmit rate varies between x_{min} and x_{max}, and average xmit rate is $x = \frac{1}{2}(x_{min} + x_{max})$.

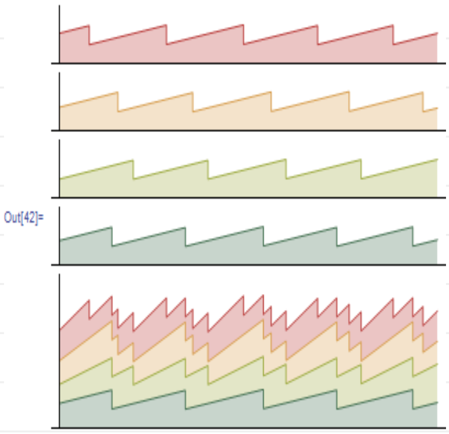
Also, we know from the "cut by half" rule that $x_{min} = \frac{1}{2}x_{max}$.

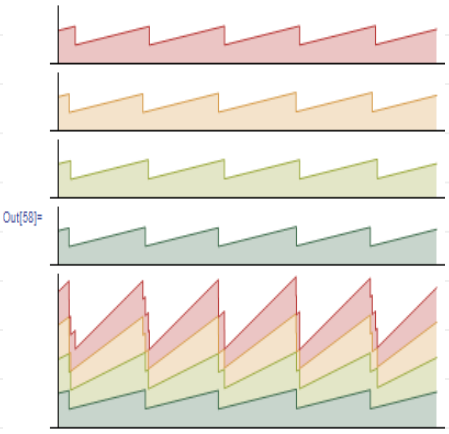
We can then solve:

$$x = \frac{1}{2}(x_{min} + x_{max}) = \frac{1}{2}\left(\frac{1}{2}x_{max} + x_{max}\right) = \frac{3}{4}x_{max} \Rightarrow x_{max} = \frac{4}{3}x$$

$$x_{min} = \frac{2}{3}x.$$

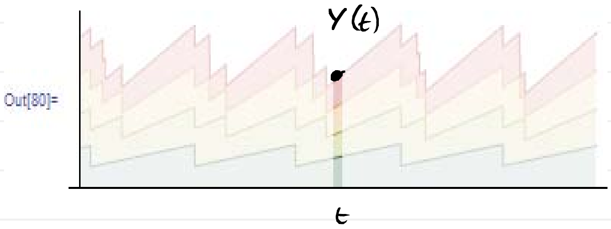
Suppose a network operator wants to lay in enough capacity to support 1000 users each running at 30 kB/sec. How much capacity is needed?





In the worst case, all the sawteeth might be aligned, giving peak rate of 1000 × $\frac{4}{3}$ × 30 kB/sec = 40 MB/sec. But intuitively we might guess that perfect alignment is unlikely, and that the peaks on one sawtooth cancel out those on another. If this happens, how much capacity is needed?

At an arbitrary instant in time t, let $X_1(t), \cdots, X_n(t)$ be the transmit rates of n flows, and let $Y(t) = X_1(t) + \cdots + X_n(t)$ be the total traffic.



We might have caught a flow at any point in its sawtooth, so each X_i might take any value in the range $[\frac{2}{3}x, \frac{4}{3}x]$ where $x = 30$ kB/s is the average rate we want to support. Furthermore, each value in this range is equally likely. The appropriate distribution is

$$X_i \sim \text{Uniform}\left(\tfrac{2}{3}x, \tfrac{4}{3}x\right).$$

$$\mathbb{E}X_i = \tfrac{1}{2}\left(\tfrac{2}{3}x + \tfrac{4}{3}x\right) = x$$

$$\text{Var}\, X_i = \frac{\left(\frac{4}{3}x - \frac{2}{3}x\right)^2}{12} = \frac{x^2}{27}$$

$$\text{sd}\, X_i = \sqrt{\text{Var}\, X_i} = \frac{x}{\sqrt{27}}.$$

If all the $X_i(t)$ are independent, then

$$\mathbb{E}Y = \mathbb{E}(X_1 + \cdots + X_n) = \mathbb{E}X_1 + \cdots + \mathbb{E}X_n = x + \cdots + x = nx$$

$$\text{Var}\, Y = \text{Var}(X_1 + \cdots + X_n) = \text{Var}\, X_1 + \cdots + \text{Var}\, X_n = \frac{x^2}{27} + \cdots + \frac{x^2}{27} = \frac{nx^2}{27}$$

$$\text{sd}(Y) = \sqrt{\text{Var}\, Y} = \sqrt{n}\,\frac{x}{\sqrt{27}}.$$

Recall, Y is a random variable, i.e. every time you sample it you get a different answer. We know from §1·8b that, roughly 95% of the time, an observation of Y will lie in the range

$$[\mathbb{E}Y - 2\,\text{sd}(Y),\ \mathbb{E}Y + 2\,\text{sd}(Y)].$$

In this case, 95% of the time, for $n = 1000$ and $x = 30$ kB/s, the total traffic rate Y lies in the range

$$\left[nx - 2\sqrt{n}\frac{x}{\sqrt{27}},\ nx + 2\sqrt{n}\frac{x}{\sqrt{27}}\right] = [29{\cdot}8, 30{\cdot}2] \text{ MB/s}.$$

This is much smaller than the worst-case value 40 MB/s.

Exercise. Suppose the capacity of the link is actually 32 MB/s.

- Explain why the overall fraction of traffic that is lost is $\frac{\mathbb{E}[(Y-32)^+]}{\mathbb{E}Y}$, and not $\mathbb{E}\left[\frac{(Y-32)^+}{Y}\right]$.

  Here, $(Y-32)^+$ is shorthand for $\max(Y-32, 0)$.

- calculate or compute the fraction of traffic that is lost.

  Hint: $\mathbb{E}[(Y-32)^+] = \int_{-\infty}^{\infty} (y-32)^+ f(y)\,dy$ where $f(y)$ is the density

$$= \int_{32}^{\infty} (y-32) f(y)\,dy$$

$$= \int_{32}^{\infty} y f(y)\,dy - 32\,\mathbb{P}(Y \geq 32).$$

You can find the integral either by numerical integration or by integration-by-parts. You can find $\mathbb{P}(Y \geq 32)$ either by a lookup table, a built-in function in a stats library, or numerical integration.

§ 1·8 Common Distributions [Not examinable]

There are a few distributions for random variables that crop up again and again in nature. See handout for details of these.

CONTINUOUS

Exponential distribution — used to model the time until an event happens for many natural processes
(Telnet session initiations, telephone call initiations, light bulb blows, radioactive nucleus decays)

Pareto distribution — time until an event happens in certain "cascade" processes, where one event can trigger others
(FTP transfer starts, landslide)
— size of a "cascade" event
(TCP flow size, insurance claim, terrorist attack)

Normal distribution — the size of a "natural" observation which doesn't deviate too much, (ie doesn't vary by many orders of magnitude) especially observations of the accumulation of many small factors
(height, weight, IQ)

DISCRETE

Geometric distribution — Like Exponential but in discrete time, i.e. the number of clock ticks until an event happens
(weeks until I win the lottery, packets sent until one is dropped)

Binomial / multinomial — The outcome of classifying n observations into categories
(e.g. number of heads in 100 tosses of a coin)
(e.g. survey 85 people, classify them by

	male	female
cute		
ugly		

)

Poisson — The number of events that occur in a fixed observation window, for well-behaved "natural" systems
(e.g. number of deaths by mule-kick each year in Napoleon's army, number of telnet sessions per hour)

Zipf — If e.g. city sizes have a Pareto distribution, and we pick a person at random, the rank of his/her city (1st, 2nd biggest, 3rd ...) has a Zipf distribution.

§ 1·8a : The Exponential Distribution

The random variable reference sheet tells us that the Exponential distribution is for a real-valued random variable taking values in $[0,\infty)$. It has one parameter, $\lambda > 0$. Its density is

$$f(x) = \lambda e^{-\lambda x}.$$

Random number generator :

```
def rexp(λ):
    return (-1.0/λ) * math.log(random.random()).
```

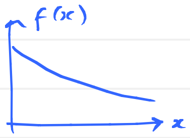
(see § 1·3)

Distribution function: $\mathbb{P}(X \geq x) = \int_x^\infty \lambda e^{-\lambda y}\, dy = e^{-\lambda x}$

Mean : $\mathbb{E}X = \int_0^\infty x \cdot \lambda e^{-\lambda x}\, dx = \left[-xe^{-\lambda x}\right]_0^\infty - \int_0^\infty -e^{-\lambda x}\, dx = \int_0^\infty e^{-\lambda x}\, dx = \left[-\frac{1}{\lambda}e^{-\lambda x}\right]_0^\infty = \frac{1}{\lambda}.$

Median : $\mathbb{P}(X \leq x) = \frac{1}{2} \Rightarrow 1 - e^{-\lambda x} = \frac{1}{2}$
$\Rightarrow e^{-\lambda x} = \frac{1}{2}$
$\Rightarrow -\lambda x = \log \frac{1}{2}$
$\Rightarrow x = \frac{1}{\lambda} \log 2.$

Mode :



mode is 0.

Variance : $\operatorname{Var} X = \mathbb{E}\left(X - \frac{1}{\lambda}\right)^2 = \int_0^\infty \left(x - \frac{1}{\lambda}\right)^2 \cdot \lambda e^{-\lambda x}\, dx = \cdots = \frac{1}{\lambda^2}$

Std. dev : $\operatorname{sd}(X) = \sqrt{\operatorname{Var} X} = \frac{1}{\lambda}.$

Two-sided 95% confidence interval :

$\mathbb{P}(X < x) = 0.025 \Rightarrow 1 - e^{-\lambda x} = 0.025 \Rightarrow e^{-\lambda x} = 0.975 \Rightarrow x = -\frac{1}{\lambda} \log 0.975$

$\mathbb{P}(X > x) = 0.025 \Rightarrow x = -\frac{1}{\lambda} \log 0.025.$

So $\left[-\frac{1}{\lambda} \log 0.975, -\frac{1}{\lambda} \log 0.025\right]$ is a 95% confidence interval

Example: let $X \sim \operatorname{Exp}(\lambda)$, and let $Y = aX$ for some constant $a > 0$. What is the distribution of Y?

We will work out the distribution function.

$\mathbb{P}(Y \geq y) = \mathbb{P}(aX \geq y)$

$= \mathbb{P}(X \geq y/a)$ — try to turn it into a statement about sthg we know

$= e^{-\lambda(y/a)}$ — we know the distribution function of X

$= e^{-(\lambda/a)y}$ — emphasize that y is the variable we're interested in

So $Y \sim \operatorname{Exp}(\lambda/a)$ — recognize the distribution function, and name it.

Exercise See examplesheet 2 questions 3, 4, 5 for further exercises in the same vein.

§1.9b The Normal Distribution

The Normal distribution is a very popular choice for data analysis because

- it's often a good fit for the aggregate of small quantities (eg individual flowrates)
- it is very simple to do algebra with it

ALGEBRA OF THE NORMAL DISTRIBUTION

If $X \sim \text{Normal}(\mu, \sigma^2)$ then

- density is $f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}(x-\mu)^2}$. There is no formula for the distribution function.
- the mean is $\mathbb{E}X = \mu$
- the variance is $\text{Var}\, X = \sigma^2$.
- $aX + b \sim \text{Normal}(a\mu + b, a^2\sigma^2)$ (*)
- $(X-\mu)/\sigma \sim \text{Normal}(0,1)$

If $X \sim \text{Normal}(\mu, \sigma^2)$ and $Y \sim \text{Normal}(\nu, \rho^2)$ and X and Y are independent, then

- $X + Y \sim \text{Normal}(\mu + \nu, \sigma^2 + \rho^2)$

To generate a $\text{Normal}(0,1)$ random variable, use the Box-Muller method (§1.3)
To generate a $\text{Normal}(\mu, \sigma^2)$ random variable, first generate $X \sim \text{Normal}(0,1$
then let $Y = \mu + \sigma X$: by (*), $Y \sim \text{Normal}(\mu, \sigma^2)$.

THE NORMAL DISTRIBUTION AS AN APPROXIMATION

Also referred to as "The central limit theorem".

If $X_1, X_2, \dots, X_n$ are independent random variables with the same distribution, from (nearly)* any distribution at all, and we let

$$Y = X_1 + \dots + X_n$$

$$\mu = \mathbb{E}Y = n\,\mathbb{E}X_1 , \quad \sigma^2 = \text{Var}\, Y = n \text{Var}\, X_1$$

then a good approximation is

$$Y \sim \text{Normal}(\mu, \sigma^2) \quad \text{ie} \quad Y \sim \text{Normal}(\mathbb{E}Y, \text{Var}\, Y)$$

The more conventional way to write this is

$$\frac{Y - n\mathbb{E}X_1}{\sqrt{n}\ \text{sd}(X_1)} \sim \text{Normal}(0,1).$$

In particular, one can prove that the distribution function of $\frac{Y - n\mathbb{E}X_1}{\sqrt{n}\ \text{sd}(X_1)}$ approaches that of $\text{Normal}(0,1)$ as n increases.

* assuming that μ and σ^2 are not infinite, plus some minor technical conditions

CONFIDENCE INTERVALS

A standard fact is

$$\mathbb{P}(\ -1.96 \le \text{Normal}(0,1) \le 1.96\ ) \approx 0.95$$

ie when we generate a Normal(0,1) random variable, we are 95% certain that the generated value lies in the range $[-1.96, 1.96]$

(You can use a computer to find the appropriate ranges for other levels of certainty).

If $Z \sim \text{Normal}(\mu, \sigma^2)$ then (from previous page) $Z = \mu + \sigma X$ where $X \sim N(0,1)$. And

$$\mathbb{P}(-1.96 \le X \le 1.96) = 0.95$$

$$\Rightarrow \mathbb{P}(\mu - 1.96\sigma \le \mu + \sigma X \le \mu + 1.96\sigma) = 0.95$$

$$\Rightarrow \mathbb{P}(\ \mu - 1.96\sigma \le Z \le \mu + 1.96\sigma) = 0.95$$

APPROXIMATE CONFIDENCE INTERVALS

We've seen:

* If Y is the sum of many small random variables, then we can approximate Y by $\text{Normal}(\ \mathbb{E}Y, \text{Var}\, Y)$.
* If $Z \sim \text{Normal}(\mu, \sigma^2)$ then a 95% confidence interval for Z is $[\mu - 1.96\sigma, \mu + 1.96\sigma]$.

Therefore, an approximate 95% confidence interval for Y is $[\ \mathbb{E}Y - 1.96\, sd(Y), \ \mathbb{E}Y + 1.96\, sd(Y)]$

e.g. ~~By~~ I throw a dice 100 times and compute the total score, Y. What is the typical range of values I get for Y?

Let X be the outcome of a single throw.

$$\mathbb{E}X = \tfrac{1}{6} \times 1 + \tfrac{1}{6} \times 2 + \cdots + \tfrac{1}{6} \times 6 = 7/2$$

$$\text{Var}\, X = \tfrac{1}{6} \times (1 - \tfrac{7}{2})^2 + \tfrac{1}{6} \times (2 - \tfrac{7}{2})^2 + \dots + \tfrac{1}{6} \times (6 - \tfrac{7}{2})^2 = \tfrac{35}{12}$$

By the Normal approximation, Y is approx. $\text{Normal}\left(100 \times \tfrac{7}{2}, \ 100 \times \tfrac{35}{12}\right)$.

Thus I am 95% confident that Y lies in the range $[317, 383]$

§1.8c The Pareto Distribution

The Pareto distribution has been found to arise in Internet traffic measurements. It causes particular problems for simulation and measurement.

The Random Variable Reference Sheet tells us that, if $X \sim \text{Pareto}(\alpha)$, then

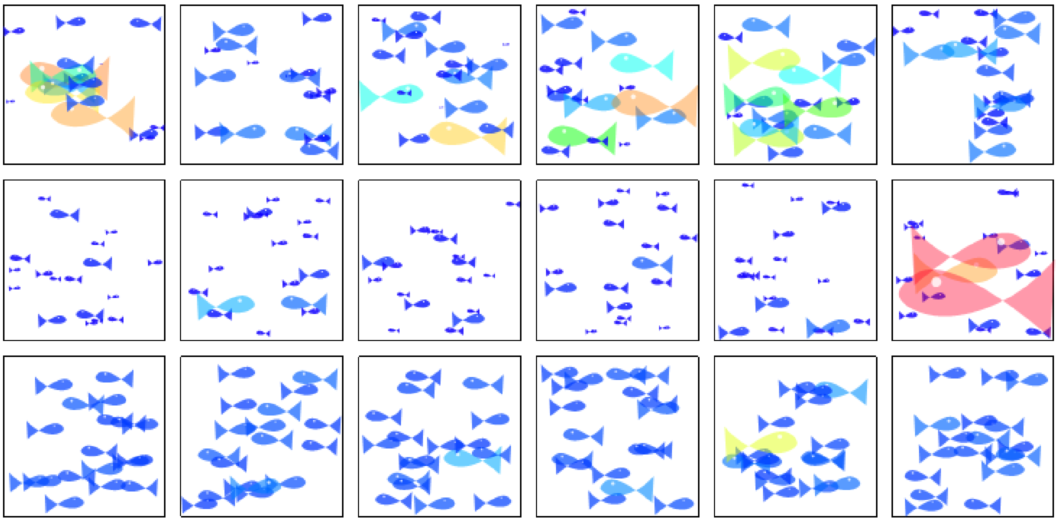
- X takes values in $[1,\infty)$
- X has density $f(x) = \alpha x^{-(\alpha+1)}$.

From this we can work out

- X has distribution function $\mathbb{P}(X \geq x) = x^{-\alpha}$
- $\mathbb{E}X = \begin{cases} \infty & \text{if } \alpha \leq 1 \\ \frac{\alpha}{\alpha-1} & \text{if } \alpha > 1 \end{cases}$
- $\text{Var}\, X = \begin{cases} \infty & \text{if } \alpha \leq 2 \\ \frac{\alpha}{(\alpha-1)^2(\alpha-2)} & \text{if } \alpha > 2. \end{cases}$
- To generate X using the inversion method (§1.4), generate $U \sim \text{Uniform}[0,1]$ and let $X = U^{-1/\alpha}$.

You may also come across more general versions of the Pareto distribution. For example, with density $f(x) = \alpha m^\alpha x^{-(\alpha+1)}$ and range $[m,\infty)$.

The Pareto distribution with $\alpha \leq 2$ tends to produce many small values ("mice") and very occassional huge values ("elephants"), and the elephants are so big that they make a significant contribution to the mean.



6 trials, each with 25 fish with size $\sim \text{Exp}(1)$

$\mathbb{E}\text{size} = 1$

size $\sim \frac{\alpha-1}{\alpha}$ Pareto(α) with $\alpha = 1.1$

$\mathbb{E}\text{size} = 1$

size $\sim \frac{\alpha-1}{\alpha}$ Pareto(α) with $\alpha = 5$

$\mathbb{E}\text{size} = 1$

This makes it hard to simulate — you need enough trials, and you need to run them long enough, to have a decent chance of catching the elephants.

§1·8d The Geometric distribution

Let $X \sim \text{Geom}(p)$, ie let X have a geometric distribution with parameter p. The handout tells us that the set of possible outcomes is $\{1, 2, \cdots\}$ which is countable — so it's a discrete random variable.

The density is $\mathbb{P}(X=r) = (1-p)^{r-1} p$.

The distribution function is $F(r) = \mathbb{P}(X \geq r) = \sum_{s=r}^{\infty} (1-p)^{s-1} p = (1-p)^{r-1}$.

Interpretation. I toss a biased coin (probability p of getting Heads) repeatedly. Let X be the number of tosses until my first Heads.

$\mathbb{P}(X=r) = \mathbb{P}($ first $r-1$ tosses were tails, then next is heads $) = (1-p)^{r-1} p$.

$\mathbb{P}(X \geq r) = \mathbb{P}($ first $r-1$ tosses were tails $) = (1-p)^{r-1}$.

Thus, $X \sim \text{Geom}(p)$.