Week 3: More Random Variables

Bayesian Statistics for Machine Learning

Dr Daniel Worrall

AMLab, University of Amsterdam

September 13, 2019

What is estimation?

Last week we considered sampling, estimation, and expectations.

$$\begin{tikzcd}
p(x) \arrow[r, "\text{expectation}"] \arrow[d, "\text{sampling}"'] & \{\mu, \sigma^2, ...\} \\
\{x_1, x_2, ..., x_N\} \arrow[r, "\text{estimation}"] & \{\hat{\mu}, \hat{\sigma}^2, ...\} \arrow[u, "\lim_{n\to\infty}"']
\end{tikzcd}$$

This week we consider we consider an easier way to compute expectation, multiple random variables, and transformations of random variables.

I: Sums of random variables

Moment Generating Functions

Moment Generating Functions

Computing moments is bothersome. The *moment generating function* (MGF)[1], is an elegant method to find moments with somewhat less bother.

The MGF $M_x(t)$ is defined as

$$\boxed{M_x(t) = \mathbb{E}\left[e^{tx}\right] \qquad \forall t \text{ where } M_x(t) \geq 0}$$

Now look at the n^{th} derivative at $t = 0$:

$$\frac{\mathrm{d}^n}{\mathrm{d}t^n} M_x(t)\bigg|_{t=0} = \frac{\mathrm{d}^n}{\mathrm{d}t^n}\mathbb{E}\left[e^{tx}\right]\bigg|_{t=0} = \mathbb{E}\left[\frac{\mathrm{d}^n}{\mathrm{d}t^n} e^{tx}\right]\bigg|_{t=0} = \mathbb{E}\left[x^n e^{tx}\right]\big|_{t=0} = \mathbb{E}\left[x^n\right]$$

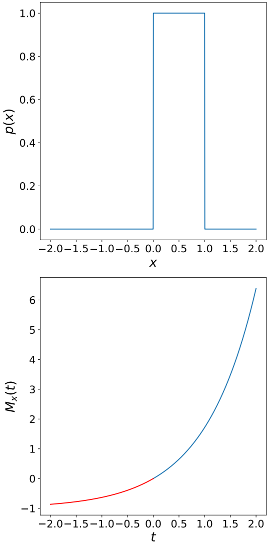
The n^{th} derivative is exactly the n^{th} moment!

[1] The MGF has a more sophisticated cousin, the *characteristic function*, prefered in practice.

Moment Generating Functions

e.g. What is the MGF of the uniform distribution between a and b?

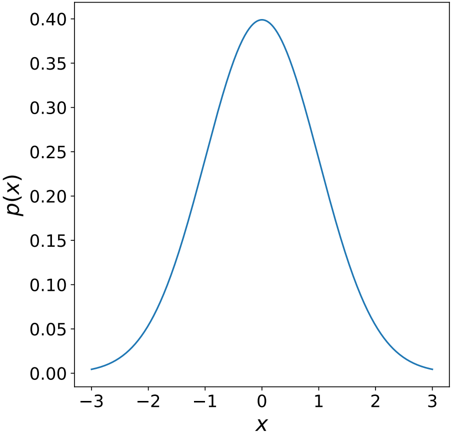
$$
\begin{aligned}
M_x(t) &= \mathbb{E}\left[e^{tx}\right] \\
&= \int_{-\infty}^{\infty} e^{tx} \frac{\mathbb{I}[x \in [a,b]]}{b-a} \,\mathrm{d}x \\
&= \frac{1}{b-a} \int_a^b e^{tx} \,\mathrm{d}x \\
&= \left[\frac{e^{tx}}{t(b-a)}\right]_a^b \\
&= \frac{e^{tb} - e^{ta}}{t(b-a)}
\end{aligned}
$$

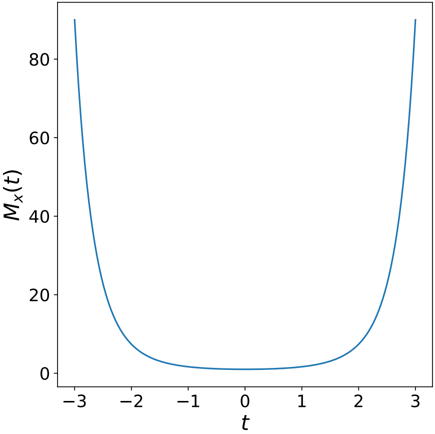


Moment Generating Functions

e.g. The MGF of the Gaussian is $M_x(t) = e^{\mu t + \frac{1}{2}\sigma^2 t^2}$. What is its mean?

$$\frac{\mathrm{d}}{\mathrm{d}t} M_x(t)\bigg|_{t=0} = (\mu + \sigma^2 t) e^{\mu t + \frac{1}{2}\sigma^2 t^2}\bigg|_{t=0} = \mu$$





Moment Generating Functions

e.g. The Gamma distribution and its MGF are

$$p(x; k, \theta) = \frac{1}{\Gamma(k)\theta^k} x^{k-1} e^{-\frac{x}{\theta}} \mathbb{I}[x \geq 0], \qquad M_x(t) = (1 - \theta t)^{-k} \mathbb{I}\left[t < \theta^{-1}\right]$$

What is the n^{th} moment of the Gamma distribution?

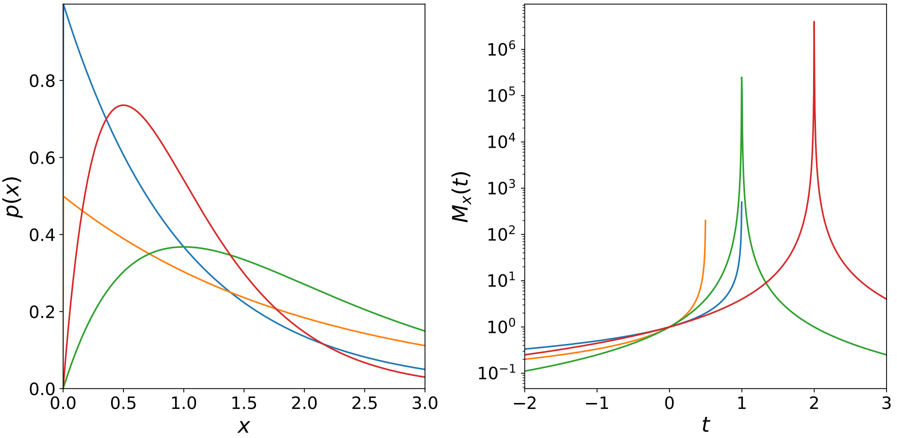
$$\frac{\mathrm{d}}{\mathrm{d}t}(1 - \theta t)^{-k} = -k(1 - \theta t)^{-(k+1)} \cdot (-\theta)$$

$$\frac{\mathrm{d}^2}{\mathrm{d}t^2}(1 - \theta t)^{-k} = +k(k+1)(1 - \theta t)^{-(k+2)} \cdot (-\theta)^2$$

$$\implies \frac{\mathrm{d}^n}{\mathrm{d}t^n}(1 - \theta t)^{-k} = (1 - \theta t)^{-(k+n)} \cdot \theta^n \prod_{i=0}^{n-1} (k + i)$$

$$\implies \mathbb{E}[x^n] = \theta^n \prod_{i=0}^{n-1} (k + i)$$

Moment Generating Functions



Moment Generating Functions: Properties

Properties

- What is $M_x(0)$?

$$M_x(0) = \mathbb{E}[x^0] = \mathbb{E}[1] = 1$$

- Matched MGFs implies matched CDFs, thus matched distributions

$$M_x(t) = M_y(t) \iff F_x(x) = F_y(y)$$

- Scaling $y = ax$

$$M_y(t) = M_x(at)$$

Moment Generating Functions: Properties

Sums of independent random variables. If $z = x + y$ then $M_z(t) = M_x(t)M_y(t)$

$$
\begin{aligned}
M_z(t) = \mathbb{E}_z[e^{tz}] = \mathbb{E}_x\left[\mathbb{E}_y[e^{t(x+y)}]\right] = \mathbb{E}_x\left[\mathbb{E}_y[e^{tx+ty}]\right] \\
= \mathbb{E}_x\left[\mathbb{E}_y[e^{tx}e^{ty}]\right] = \mathbb{E}_x[e^{tx}]\mathbb{E}_y[e^{ty}] = M_x(t)M_y(t)
\end{aligned}
$$

So if $z = \sum_{n=1}^{N} x_i$ then

$$
M_z(t) = M_{x_1}(t)M_{x_2}(t)\cdots M_{x_N}(t) = \prod_{n=1}^{N} M_x(t)
$$

And the mean of z is thus

$$
\begin{aligned}
\frac{\mathrm{d}}{\mathrm{d}t}M_z(t)\bigg|_{t=0} &= \frac{\mathrm{d}}{\mathrm{d}t}M_{x_1}(t)M_{x_2}(t)\cdots M_{x_N}(t)\bigg|_{t=0} \\
&= M'_{x_1}(t)M_{x_2}(t)\cdots M_{x_N}(t) + M_{x_1}(t)M'_{x_2}(t)\cdots M_{x_N}(t)...\big|_{t=0} \\
&= M'_{x_1}(0) + M'_{x_2}(t) + ... + M'_{x_N}(0) \\
&= \sum_{n=1}^{N} \mathbb{E}[x_i]
\end{aligned}
$$

Moments of summed random variables

It is not too difficult to see that

$$\mathbb{E}\left[\sum_{i=1}^{N} x_n\right] = \sum_{i=1}^{N} \mathbb{E}_{x_n}\left[x_n\right]$$

for *independent* random variables.

The well known Bienaymé formula[2] is a little trickier to show

$$\mathbb{V}\left[\sum_{i=1}^{N} x_n\right] = \sum_{i=1}^{N} \mathbb{V}_{x_n}\left[x_n\right]$$

for *independent* random variables.

[2] Discovered in 1853

Sums of random variables

e.g. What is the variance of the mean estimator $\frac{1}{N}\sum_{i=1}^{N} x_i$ for large N, if $\mathbb{E}[x] = 0$ and $\mathbb{V}[x] = \sigma$?

The variance of the sums of iid RVs is the sum of their variances. So

$$\mathbb{V}\left[\sum_{i=1}^{N} x_i\right] \overset{\text{Bienaymé}}{=} \sum_{i=1}^{N} \mathbb{V}[x_i] \overset{\text{iid}}{=} \sum_{i=1}^{N} \mathbb{V}[x] = N\sigma^2$$

$$\mathbb{V}\left[\frac{1}{N}\sum_{i=1}^{N} x_i\right] = \frac{1}{N^2} N\sigma^2 = \frac{1}{N}\sigma^2$$

Notice that as $N \to \infty$ the variance goes to zero.

II: Transforming random variables

Statistical independence

If the random variable X is statistically independent of Y then we saw last week that

$$p(x, y) = p(x|y)p(y) = p(x)p(y).$$

For N variables $X_1, X_2, ..., X_N$, this would become

$$\boxed{p(x_1, ..., x_N) = p(x_1)p(x_2) \cdots p(x_N) = \prod_{i=1}^{N} p(x_i)}$$

We call such a joint, *factorizable*.

Bivariate Gaussian

Let's say x and y are iid RVs from standard Gaussians, so

$$
\begin{aligned}
x &\sim \frac{1}{\sqrt{2\pi}} \exp\left\{-\frac{x^2}{2}\right\} \\
y &\sim \frac{1}{\sqrt{2\pi}} \exp\left\{-\frac{y^2}{2}\right\}
\end{aligned}
$$

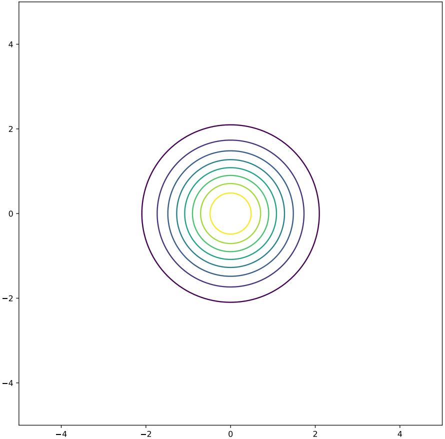
Their joint probability is written

$$
\begin{aligned}
p(x, y) &= p(x)p(y) \\
&= \frac{1}{\sqrt{2\pi}} \exp\left\{-\frac{x^2}{2}\right\} \cdot \frac{1}{\sqrt{2\pi}} \exp\left\{-\frac{y^2}{2}\right\} \\
&= \frac{1}{2\pi} \exp\left\{-\frac{x^2 + y^2}{2}\right\}
\end{aligned}
$$

Bivariate Gaussian

The level sets of this distribution are circles since if

$$\frac{1}{2\pi} \exp\left\{-\frac{x^2+y^2}{2}\right\} = \text{const} \implies x^2+y^2 = \text{const}$$



Transformation of random variables

e.g. Arrows are shot at a target. If the x-position and y-position of the arrows are distributed about the centre with zero mean and variance σ^2, what is the distribution of radii from the centre?

$$p(x,y) = p(x)p(y) = \mathcal{N}(x|0,\sigma^2)\mathcal{N}(y|0,\sigma^2) = \frac{1}{2\pi\sigma^2}\exp\left\{-\frac{x^2+y^2}{2\sigma^2}\right\}$$

In polar coordinates

$$x = r\cos\theta \qquad y = r\sin\theta \qquad p(r,\theta) = \frac{1}{2\pi\sigma^2}\exp\left\{\frac{-r^2}{2\sigma^2}\right\} \qquad \mathrm{d}x\mathrm{d}y = r\mathrm{d}r\mathrm{d}\theta$$

So

$$P(R<r) = \int_0^r \int_{\theta=-\pi}^{\theta=\pi} p(r',\theta)r'\mathrm{d}\theta\mathrm{d}r' = \int_0^r \int_{\theta=-\pi}^{\theta=\pi} \frac{1}{2\pi\sigma^2}\exp\left\{\frac{-r'^2}{2\sigma^2}\right\} r'\mathrm{d}\theta\mathrm{d}r'$$

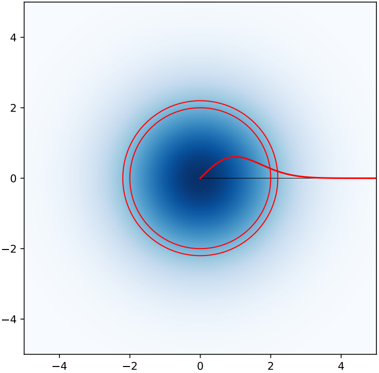
$$= \int_0^r \frac{\cancel{2\pi}\cdot r'}{\cancel{2\pi}\sigma^2}\exp\left\{\frac{-r'^2}{2\sigma^2}\right\}\mathrm{d}r = \left[-\exp\left\{\frac{-r^2}{2\sigma^2}\right\}\mathrm{d}r\right]_0^r = 1 - \exp\left\{\frac{-r^2}{2\sigma^2}\right\}$$

Transformation of random variables

Given $P(R < r) = 1 - \exp\left\{\frac{-r^2}{2\sigma^2}\right\}$, we can figure out the PDF using

$$
\begin{aligned}
p(r) &= \frac{\mathrm{d}}{\mathrm{d}r} P(R < r) \\
&= \frac{\mathrm{d}}{\mathrm{d}r}\left(1 - \exp\left\{\frac{-r^2}{2\sigma^2}\right\}\right) \\
&= \frac{r}{\sigma^2} \exp\left\{\frac{-r^2}{2\sigma^2}\right\}
\end{aligned}
$$

The radial PDF is *Rayleigh distributed*.



Transformation of random variables

In general, if $x \sim p(x)$ and we are given $x = f(y)$, what is $p(y)$?

We know that probability of x in some region $\mathcal{X}$ should equal probability of y in the transformed region $\mathcal{X} = f(\mathcal{Y})$.

$$\int_{\mathcal{X}} p_X(x)\,\mathrm{d}x = \int_{\mathcal{Y}} p_Y(y)\,\mathrm{d}y$$

So if we shrink the region on integration to an infinitesimal slither

$$|p_X(x)\mathrm{d}x| = |p_Y(y)\mathrm{d}y| \quad \Longrightarrow \quad \boxed{p_Y(y) = p_X(x)\left|\frac{\mathrm{d}x}{\mathrm{d}y}\right|}$$

This only works for f invertible!

Useful in e.g. random number generation, computer simulations, Monte Carlo integrals, generating images (GLOW by Durk Kingma)

Transformation of random variables

e.g. Find the distribution of $y = ax + b$ if $p(x) = \mathcal{N}(x|\mu, \sigma^2)$.

$$\begin{aligned}
p(y) &= p(x) \left| \frac{\mathrm{d}x}{\mathrm{d}y} \right| \\
&= \frac{1}{\sigma\sqrt{2\pi}} \exp\left\{ -\frac{(x-\mu)^2}{2\sigma^2} \right\} \frac{\mathrm{d}}{\mathrm{d}y} \frac{y-b}{a} \\
&= \frac{1}{a\sigma\sqrt{2\pi}} \exp\left\{ -\frac{(\frac{y-b}{a} - \mu)^2}{2\sigma^2} \right\} \\
&= \frac{1}{a\sigma\sqrt{2\pi}} \exp\left\{ -\frac{(y-b-a\mu)^2}{2a^2\sigma^2} \right\} \\
&= \mathcal{N}(y|a\mu + b, a^2\sigma^2)
\end{aligned}$$

This matches what we know that $\mathbb{E}[ax+b] = a\mathbb{E}[x] + b$ and $\mathbb{V}[ax+b] = a^2\mathbb{V}[x]$.

Transformation of random variables

e.g.
Find the distribution of $x = -y$ where $x \sim \mathbb{I}[x \in [0, 1]]$.

$$p_Y(y) = p_X(x) \left| \frac{\mathrm{d}x}{\mathrm{d}y} \right|$$

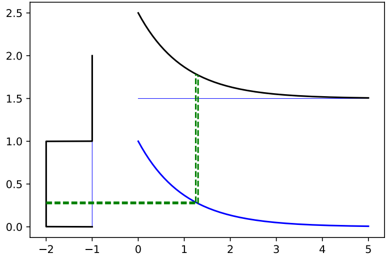
$$\begin{aligned} &= p_X(x) \left| \frac{\mathrm{d}}{\mathrm{d}y}(-y) \right| \\ &= p_X(-y) \left| -1 \right| \\ &= \mathbb{I}[-y \in [0, 1]] \\ &= \mathbb{I}[y \in [-1, 0]] \end{aligned}$$

Transformation of random variables

e.g.

Find the distribution of $x = e^{-y}$ where $x \sim \mathbb{I}[x \in [0,1]]$.

$$
\begin{aligned}
p(y) &= p(x) \left| \frac{\mathrm{d}x}{\mathrm{d}y} \right| \\
&= p(x) \left| \frac{\mathrm{d}}{\mathrm{d}y} e^{-y} \right| \\
&= p(x) \left| -e^{-y} \right| \\
&= \underbrace{p(x)}_{=1} e^{-y} \\
&= \mathsf{Exp}(y; 1)
\end{aligned}
$$



Transformation of random variables

What if we now have the opposite process? I give you $p(x)$ and $p(y)$ and you have to tell me $x = f(y)$? Recall

$$\int_{\mathcal{X}} p_X(x)\,\mathrm{d}x = \int_{\mathcal{Y}} p_Y(y)\,\mathrm{d}y$$

Let's choose the regions of integration such that $\int_{\mathcal{X}} p_X(x)\,\mathrm{d}x$ is a CDF

$$F_X(x) = F_Y(y).$$

The transformation is now

$$\boxed{x = F_X^{-1}\left(F_Y(y)\right).}$$

Transformation of random variables

e.g. If $p(y) = \mathbb{I}(y \in [0,1])$ and $p(x) = \frac{1}{2}\mathbb{I}[x \in [0,2]]$, what is $x = f(y)$?

Matching the CDFs

$$F_X(x) = F_Y(y)$$

$$\begin{aligned}\int_{-\infty}^{x} \frac{1}{2}\mathbb{I}[x' \in [0,2]]\,\mathrm{d}x' &= \int_{-\infty}^{y} \mathbb{I}(y \in [0,1])\mathrm{d}y' \\ \int_{0}^{x} \frac{1}{2}\,\mathrm{d}x' &= \int_{0}^{y} \mathrm{d}y' \\ \frac{x}{2} &= y \\ x &= 2y\end{aligned}$$

Transformation of random variables

e.g. If $p(y) = \mathbb{I}(y \in [0,1])$ and $p(x) = \frac{x}{\sigma^2} e^{-\frac{x^2}{2\sigma^2}}$, what is $x = f(y)$?

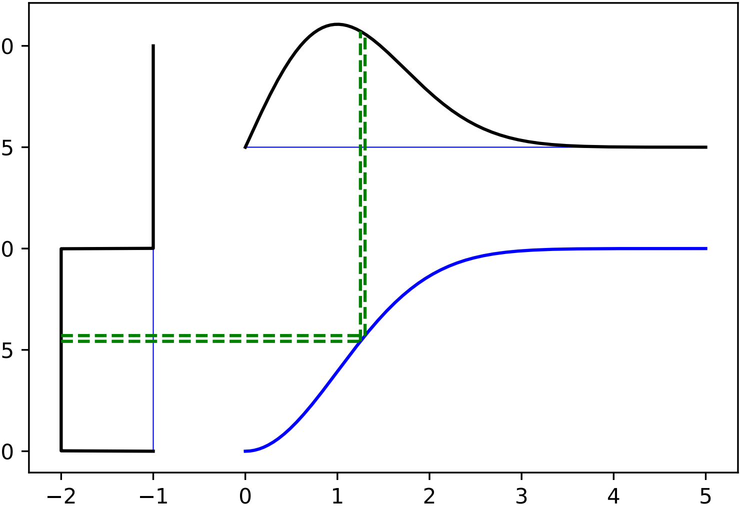
Matching the CDFs

$$F_X(x) = F_Y(y)$$

$$\begin{aligned}
\int_0^x \frac{x'}{\sigma^2} e^{-\frac{x'^2}{2\sigma^2}} \, \mathrm{d}x' &= \int_0^y \mathrm{d}y' \\
1 - e^{-\frac{x^2}{2\sigma^2}} &= y \\
-\frac{x^2}{2\sigma^2} &= \ln(1-y) \\
x^2 &= -2\sigma^2 \ln(1-y) \\
x &= \sqrt{-2\sigma^2 \ln(1-y)}
\end{aligned}$$

Transformation between Uniform and Rayleigh

$\sigma = 1$



Transformation of random variables

If $p(x)$ is a zero mean Gaussian of width σ and y is uniform, what is $y = f(x)$?

$$F_X(x) = F_Y(y) \implies y = \int_{-\infty}^{x} \mathcal{N}(x'|0, \sigma^2)\, \mathrm{d}x' = \Phi(x)$$

so

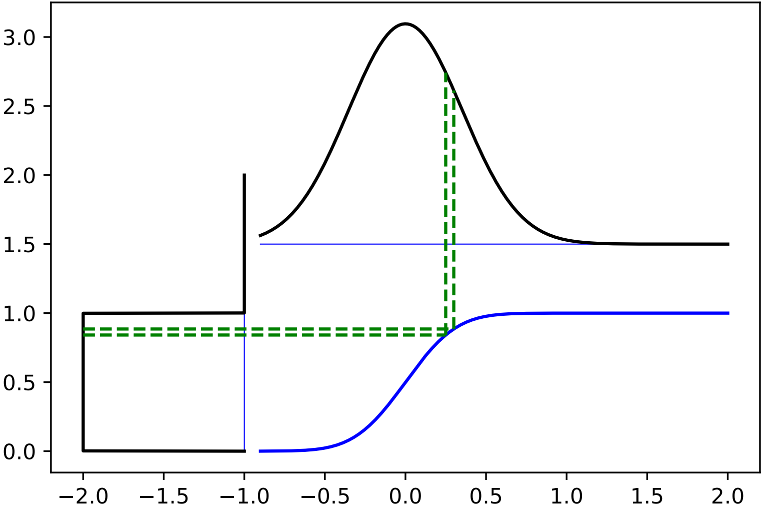
$$x = \Phi^{-1}(y)$$

This requires that we have access to the inverse of the CDF of the Gaussian. As we know, this is intractable, but we can use good numerical approximations.

This is the typical way to generate Gaussian random numbers on your computer.

Transformation between Uniform and Gaussian

$\sigma = 0.25$



Transformation of random variables

e.g. The Gompertz distribution is $p(x) = b\eta e^{\eta + bx - \eta e^{bx}}$, for $b, \eta > 0, x \geq 0$ with

$$F(x) = 1 - e^{-\eta(e^{bx}-1)}.$$

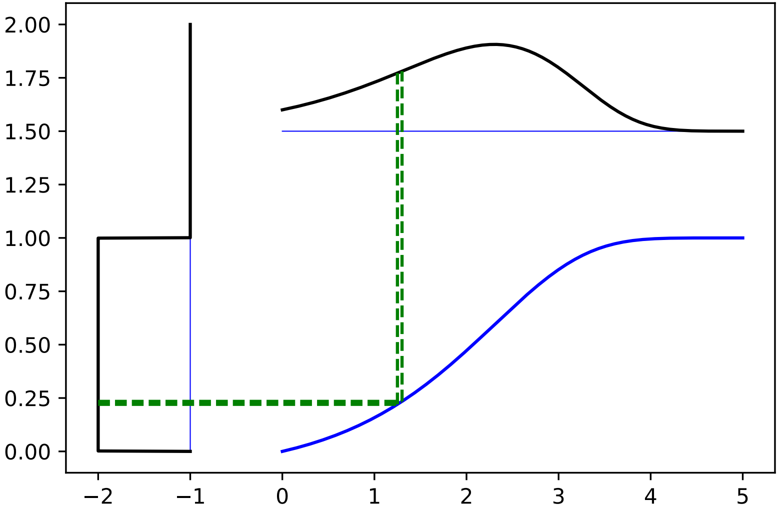
If $p(y) = \mathbb{I}[y \in [0,1]]$, what f satisfies $x = f(y)$?

$$F(x) = F(y)$$

$$\begin{aligned} 1 - e^{-\eta(e^{bx}-1)} &= y \\ -\eta(e^{bx} - 1) &= \log(1-y) \\ e^{bx} &= 1 - \frac{1}{\eta}\log(1-y) \\ x &= \frac{1}{b}\log\left(1 - \frac{1}{\eta}\log(1-y)\right) \end{aligned}$$

Transformation between Uniform and Gompertz

$\eta = 0.1, b = 1$



III: Miscellanea

Statistical Independence

When we have a collection of n Bernoulli trials? How do we model that? Say I have n *identical* coins and I flip each one once. The outcome $\{\texttt{head}, \texttt{tail}\}$ of each coin is *independent* of every other. We formalise this scenario with the concept of *independent and identically distributed* or *iid*.

Statistical Independence The joint probability of n statistically independent outcomes is the product of their marginals

$$p(x_1, x_2, ..., x_n) = p(x_1)p(x_2)...p(x_n) = \prod_{i=1}^{n} p(x_i)$$

e.g. A pharmacologist is developing a new strain of antibiotics on the superbug MRSA. She prepares 25 petri dishes with bacteria and drops a little of the new antibiotics on each. $X \sim \text{Ber}(0.99)$ is the RV whether the antibiotics kills the bacteria in a dish. The probability the bacteria is killed in every dish is

$$\prod_{i=1}^{25} 0.99 = \underbrace{0.99 \cdot 0.99 \cdot ... \cdot 0.99}_{\text{25 times}} = 0.99^{25} = 0.7778 \text{ (4 s.f.)}$$

Combinatorics

What if we had asked the question "What is the probability that in exactly one dish the bacteria is not killed"? Let's write $\mathbf{X} = X_1 X_2 X_3 ...$ and denote success as 1 and failure as 0 then

$$
\begin{aligned}
P(\mathbf{X} = \texttt{01111...}) &= 0.01 \cdot 0.99 \cdot 0.99 \cdot 0.99 \cdot 0.99 ... = 0.01 \cdot 0.99^{24} \\
P(\mathbf{X} = \texttt{10111...}) &= 0.99 \cdot 0.01 \cdot 0.99 \cdot 0.99 \cdot 0.99 ... = 0.01 \cdot 0.99^{24} \\
P(\mathbf{X} = \texttt{11011...}) &= 0.99 \cdot 0.99 \cdot 0.01 \cdot 0.99 \cdot 0.99 ... = 0.01 \cdot 0.99^{24} \\
P(\mathbf{X} = \texttt{11101...}) &= 0.99 \cdot 0.99 \cdot 0.99 \cdot 0.01 \cdot 0.99 ... = 0.01 \cdot 0.99^{24} \\
P(\mathbf{X} = \texttt{11110...}) &= 0.99 \cdot 0.99 \cdot 0.99 \cdot 0.99 \cdot 0.01 ... = 0.01 \cdot 0.99^{24} \\
&\vdots
\end{aligned}
$$

If $K = \sum_{i=1}^{n} X_i$, then

$$
\begin{aligned}
P(K = 24) &= \underbrace{0.01 \cdot 0.99^{24} + 0.01 \cdot 0.99^{24} + ... + 0.01 \cdot 0.99^{24}}_{\text{25 times}} \\
&= 25 \cdot 0.01 \cdot 0.99^{0.24} = 0.1964 \text{ (4 s.f.)}
\end{aligned}
$$

Combinatorics

What is $P(K = 23)$?

There are 25 ways for the first dish to fail, and 24 ways for the second dish to fail, so there are $25 \cdot 24$ ways to choose two failed dishes. BUT order does not matter so we have $25 \cdot 24/2$ ways.

The binomial coefficient

$$^{n}C_k = \frac{n!}{k!(n-k)!}$$

where the ! symbol is called a factorial, meaning $n! = n \cdot (n-1) \cdot (n-2) \cdot ... \cdot 2 \cdot 1$.

The factorial is ridiculously fast growing function

1!=1, 2!=2, 3!=6, 4!=24, 5!=120

6!=720, 7!=5040, 8!=40320, 9!=362880, 10!=3628800

By the way, we define $0! = 1$.

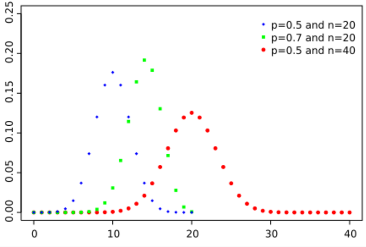
Binomial distribution

The Binomial distribution describes the number of *successes* k out of n independent and identically distributed Bernoulli trials of probability p

$$k \sim \mathrm{Bin}(k; n, p),$$

where $n = 1, 2, 3, ...,$ and $0 \leq p \leq 1.$ We have

$$P(k) = {}^{n}C_k p^k (1-p)^{n-k}$$



Binomial distribution

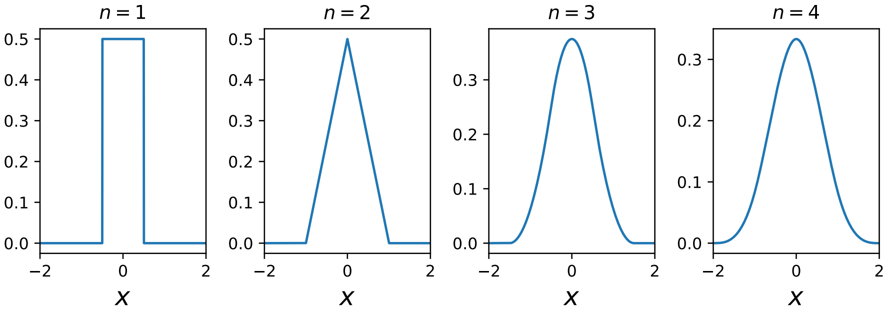
e.g. This has mean

$$
\begin{aligned}
\mathbb{E}[k] &= \sum_{k=0}^{n} k \, \mathsf{Bin}(k; n, p) = \sum_{k=0}^{n} k \frac{n!}{k!(n-k)!} p^{k} (1-p)^{n-k} \\
&= \sum_{k=1}^{n} \frac{n!}{(k-1)!(n-k)!} p^{k} (1-p)^{n-k} \\
&= np \sum_{k=1}^{n} \frac{(n-1)!}{(k-1)!(n-k)!} p^{k-1} (1-p)^{n-k} \\
&= np \sum_{k'=0}^{n'} \frac{n'!}{k'!(n'-k')!} p^{k'} (1-p)^{n'-k'} = np.
\end{aligned}
$$

Well that was fiddly! Next we consider a simpler way to compute the mean of such a distribution.

Sum of random variables

n is the number of uniform random variables in the sum



Do you notice something about this?

The distributions start to look like Gaussians with our already-known fact

$$\mu = \sum_{i=1}^{N} \mu_i, \qquad \sigma^2 = \sum_{i=1}^{N} \sigma_i^2.$$

The Central Limit Theorem

If X_i are iid RVs with shared mean 0 and variance σ^2, then the quantity

$$S_N = \frac{1}{\sqrt{N}} \sum_{i=1}^{N} x_i$$

is Gaussian distributed[3] in the limit of $N \to \infty$ i.e.

$$\lim_{N \to \infty} S_N \sim \mathcal{N}(\cdot | 0, \sigma^2).$$

The proof is beyond the scope of this course.

Why did we divide by $\sqrt{N}$ and not N? Hint: Think about the variance $\frac{x}{\sqrt{N}}$.

[3] Note this only applies to random variables which have finite moments (which is most of them). An example of a distribution with undefined moments is the Cauchy distribution $p(x) = (\pi(x^2+1))^{-1}$. It has infinite mean and variance! The sum of two Cauchy random variables is itself Cauchy distributed.