8.044 Lecture Notes
Chapter 2: Probability for 8.044

Lecturer: McGreevy

Contents

Reading: Notes by Prof. Greytak

How we count in 8.044: first one random variable, then two random variables, then 10^{24} random variables.

2.1 One random variable

A **random variable** (RV) is a quantity whose value can't be predicted (with certainty) given what we know. We have limited knowledge, in the form of a probability distribution.

Two basic sources of uncertainty (and hence RVs) in physics:

1. quantum mechanics (8.04)

   Even when the state of the system is completely specified, some measurement outcomes can't be predicted with certainty. Hence they are RVs.

2. ignorance (8.044)

   This happens when we have insufficient information to fully specify the state of the system, *e.g.* because there are just too many bits for us to keep track of. Suppose we know P, V, T, E, S of a cylinder of gas – this determines the macrostate. $\vec{x}, \vec{p}$ of the 10^{24} molecules are not specified by this. So we will use RVs to describe the microstate.

Types of RVs:

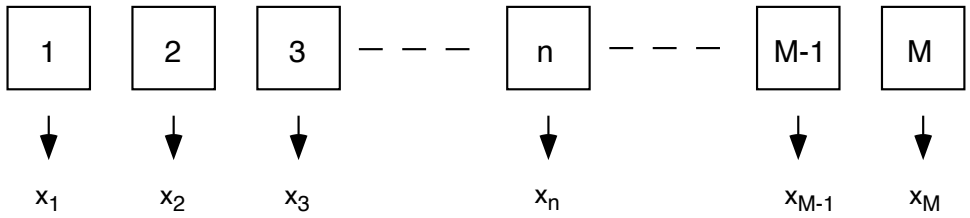
- continuous: (*e.g.* position or velocity of gas atoms) real numbers.
- discrete: (*e.g.* number of gas atoms in the room) integers.
- mixed: both continuous and discrete components (*e.g.* energy spectrum of H atom has discrete boundstate energies below a continuum)

Probability theory can be based on an ensemble of similarly prepared systems.

e.g. many copies of a cylinder of gas, all with the same P, V, T, E, S.
all *determined* (*i.e.* macroscopic) are the same.

Imagine M copies of the system, with $M \gg 1$. Let x be an RV.

(*e.g.* suppose we are interested in the air in the room. Imagine M copies of the room, fill each with gas of $N_2, O_2, ...$ and one Xe atom. $x =$ position of the Xe atom.)



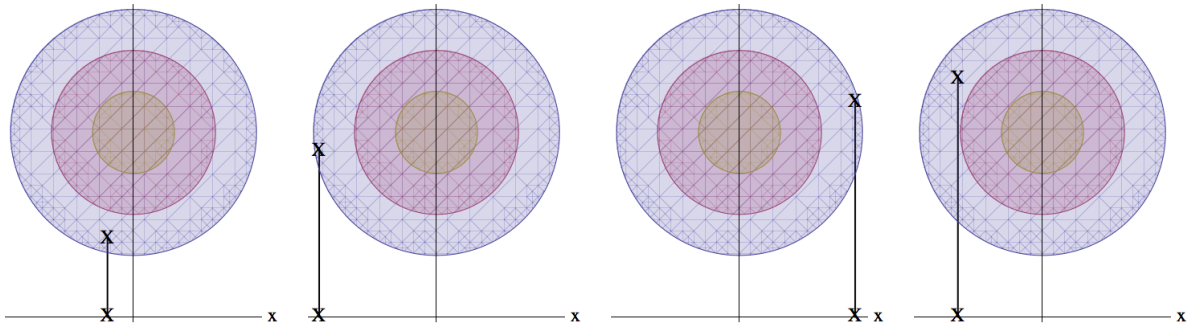
Make a histogram of how many of the rooms have each value of x, or rather have their value of x in some interval of possible values. We define the probability density for x to be

$$p_x(\zeta) = \lim_{d\zeta \to 0, M \to \infty} \frac{\#\text{ of systems with } x \in (\zeta, \zeta + d\zeta)}{M d\zeta},$$

that is, it's the fraction of systems with x in a certain bin, as we take the bin size to zero, and the number of copies of the system in the ensemble to infinity. ζ ("zeta") is a dummy variable with units of x. Once we get used to this, we'll occasionally write the less precise expression $p(x)$ for $p_x(\zeta)$.

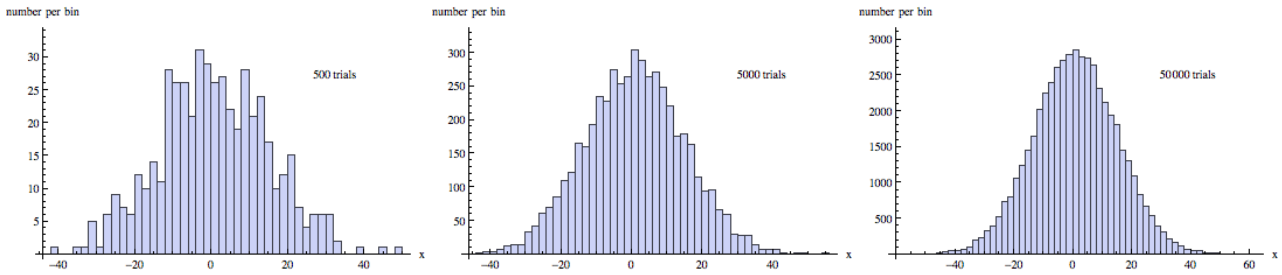
[End of Lecture 3.]

A simple example of assembling a probability distrbution: shoot arrows at a target.



x is horizontal displacement from bullseye (in some units, say millimeters).

Make a histogram of values of x: divide up the possible locations into bins of size $d\zeta$, and count how many times the arrow has an x-position in each bin. Shoot M arrows.



In these plots, I am just showing raw numbers of hits in each bin on the vertical axis, and I am fixing the bin size. In the three plots I've taken $M = 500, 5000, 50000$.

Claim: [histogram] $\overset{M\to\infty, d\zeta\to 0}{\longrightarrow}$ [smooth distribution]

Properties of p_x:

- Prob(x between ζ and $\zeta + d\zeta$) $= p_x(\zeta)d\zeta$. (So the probability of hitting any particular point exactly is zero if p is smooth. But see below about delta functions.)
- p is real
- $p_x(\zeta) \geq 0$.
- Prob($x \in (a, b)$) $= \int_a^b p_x(\zeta)d\zeta$
- The fact that the arrow always ends up *somewhere* means that the distribution is *normalized*:
  $\int_{-\infty}^{\infty} p_x(\zeta)d\zeta = 1$.
- the units of p_x are $\frac{1}{\text{units of } x}$. (*e.g.* 1 is dimensionless.)

The probability density is a way of packaging what we *do* know about a RV. Another convenient package for the same info is:

Cumulative probability

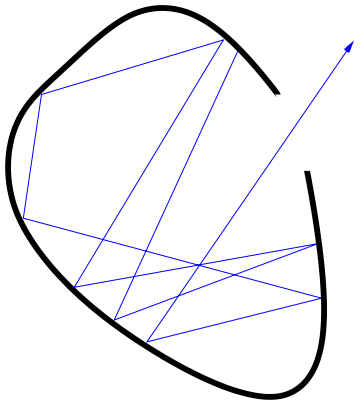
$$P(\zeta) \equiv Prob(x < \zeta) = \int_{-\infty}^{\zeta} d\zeta' p_x(\zeta').$$

(Note the serif on the capital P.) To get back the probability density:

$$p_x(\zeta) = \frac{d}{d\zeta} P(\zeta).$$

The question we have to answer in this chapter of the course is: "Given p or P, what can we learn?" How to determine p is the subject of 8.04 and the whole rest of 8.044. So in the next few lectures, we'll have to cope with Probability Densities from Outer Space, *i.e.* I'm not going to explain where the come from.

Example of a discrete density:



Consider an atom bounces around in a cavity with a hole.

Prob(atom escapes after a given bounce) $= a$
for some small constant a.

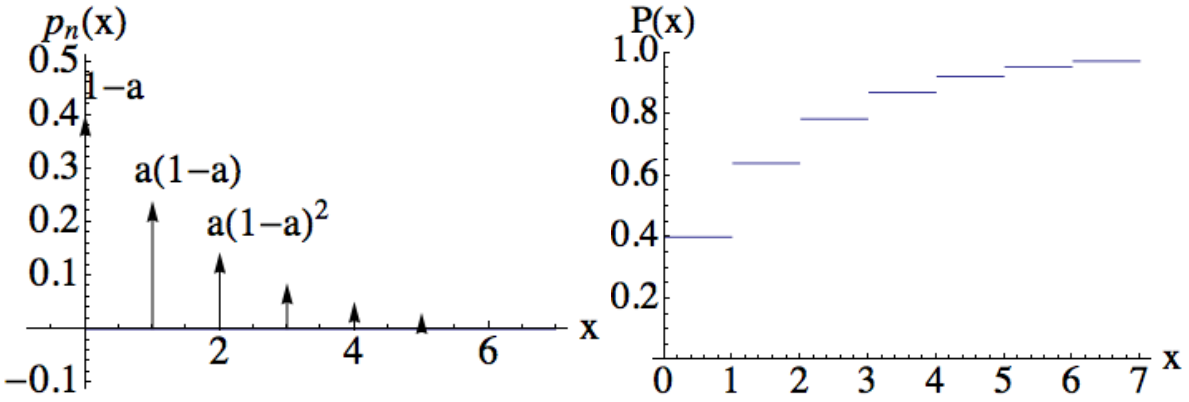
$p(n) \equiv$ Prob(atom escapes after n bounces)
is a probability, not a density. $n = 0, 1, 2...$

$p(n) =$ $\underbrace{(1-a)^n}_{\text{prob that it failed to escape } n \text{ times}}$ $\times$ $\underbrace{a}_{\text{and then escaped on the } n\text{th}}$

Repackage:

$$p_n(\zeta) = \sum_{n=0}^{\infty} p(n)\delta(\zeta - n) = \sum_{n=0}^{\infty} (1-a)^n a \delta(\zeta - n).$$

($\delta(x)$ here is the Dirac delta function, which has the property that it vanishes for any $x \neq 0$ but integrates to 1 around any neighborhood of $x = 0$.) The point of this interlude is that all manipulations can be done in the same formalism for discrete and continuum cases, we don't need to write separate equations for the discrete case.



(Note convention for sketching δ-functions.)

P always asymptotes to 1, since $\int_{ALL} p = 1$.

Check normalization: $\sum_{n=0}^{\infty} a(1-a)^n = a\frac{1}{1-(1-a)}$. You should get used to checking the normalization whenever anyone gives you an alleged probability distribution. Trust no one!

This is called a 'geometric' or 'Bose-Einstein' distribution.

What to do with a probability density? $p(x)$ is a lot of information. Here's how to extract the bits which are usually most useful.

Averages

$$\text{Mean:} \qquad \langle x \rangle \equiv \int_{-\infty}^{\infty} x p(x) dx$$

$$\text{Mean square:} \qquad \langle x^2 \rangle \equiv \int_{-\infty}^{\infty} x^2 p(x) dx$$

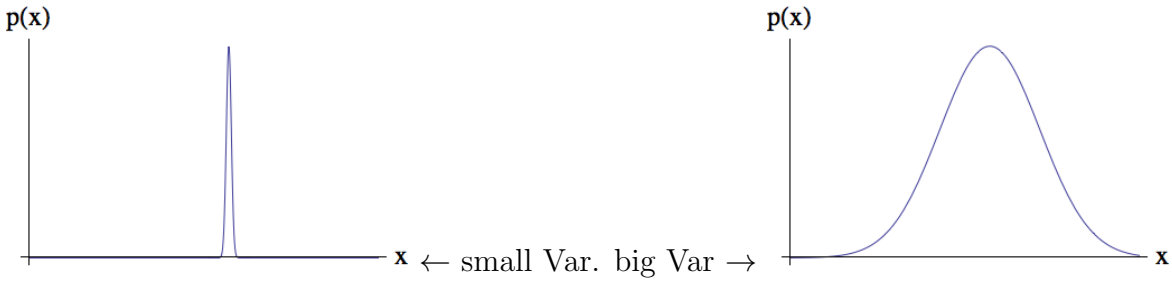
$$n\text{th Moment:} \qquad \langle x^n \rangle \equiv \int_{-\infty}^{\infty} x^n p(x) dx$$

$$\langle f(x) \rangle \equiv \int_{-\infty}^{\infty} f(\zeta) p_x(\zeta) d\zeta = \int_{-\infty}^{\infty} f(x) p(x) dx$$

Meaning: make a histogram of $f(x)$ instead of a histogram of x. What is the mean of this histogram?

$$\text{Variance:} \qquad \text{Var}(x) \equiv \langle (x - \langle x \rangle)^2 \rangle$$

Variance is an answer to the question "How wide is the histogram for x?" A quadratic measure of the fluctuations of x about its mean.



$$\text{Standard deviation} \equiv \text{stdev} \equiv \sqrt{\text{Variance}}$$

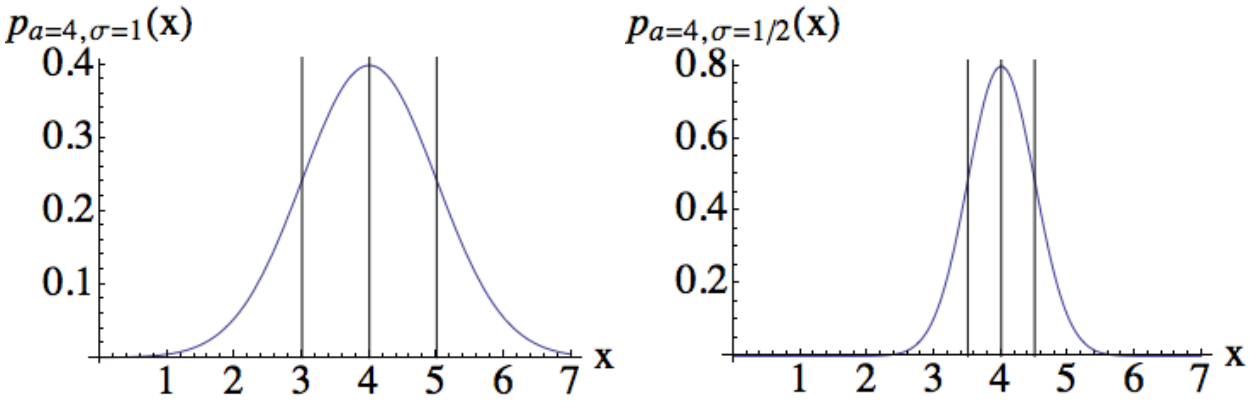
$$\begin{aligned} \text{Note:} \qquad \text{Var}(x) \equiv \langle (x - \langle x \rangle)^2 \rangle &= \int_{-\infty}^{\infty} dx p(x) (x - \langle x \rangle)^2 \\ &= \int_{-\infty}^{\infty} dx p(x) x^2 - 2\langle x \rangle \int_{-\infty}^{\infty} dx p(x) x + \langle x \rangle^2 \int_{-\infty}^{\infty} dx p(x) \\ &= \langle x^2 \rangle - 2\langle x \rangle^2 + \langle x \rangle^2 \end{aligned}$$

$$\boxed{\text{Var}(x) = \langle x^2 \rangle - \langle x \rangle^2}$$

Example: the Gaussian density (ubiquitous, and provably so)

$$p(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{(x-a)^2}{2\sigma^2}\right]$$

Specified by two parameters a, σ.



The plots shows the gaussian distribution for $a = 4, \sigma = 1$ and $a = 4, \sigma = 1/2$. The lines appear at $x = a - \sigma, a, a + \sigma$. Visibly, a determines the center of the peak, and σ determines its width.

Find the mean, variance, standard deviation for this distribution.

$$\begin{aligned}
\langle x \rangle &= \frac{1}{\sqrt{2\pi\sigma^2}} \int_{-\infty}^{\infty} dx\ x \exp\left[-\frac{(x-a)^2}{2\sigma^2}\right] \\
&\overset{\eta \equiv x-a}{=} a \frac{1}{\sqrt{2\pi\sigma^2}} \underbrace{\int_{-\infty}^{\infty} d\eta\ \exp\left[-\frac{\eta^2}{2\sigma^2}\right]}_{\sqrt{2\pi\sigma}.\ \text{Normalized!}} + \frac{1}{\sqrt{2\pi\sigma^2}} \underbrace{\int_{-\infty}^{\infty} d\eta\ \eta\ \exp\left[-\frac{\eta^2}{2\sigma^2}\right]}_{\text{odd integral=0}} \\
&= a.
\end{aligned}$$

$$\begin{aligned}
\langle x^2 \rangle &= \int_{-\infty}^{\infty} dx\ x^2 p(x) \\
&\overset{\eta \equiv x-a}{=} \frac{1}{\sqrt{2\pi\sigma^2}} \int_{-\infty}^{\infty} d\eta \left(\underbrace{\eta^2}_{\text{see 2.1.2 below}} + \underbrace{2\eta a}_{\text{odd integral}} + \underbrace{a^2}_{\langle 1 \rangle} \right) \exp\left[-\frac{\eta^2}{2\sigma^2}\right] \\
&= \sigma^2 + a^2.
\end{aligned}$$

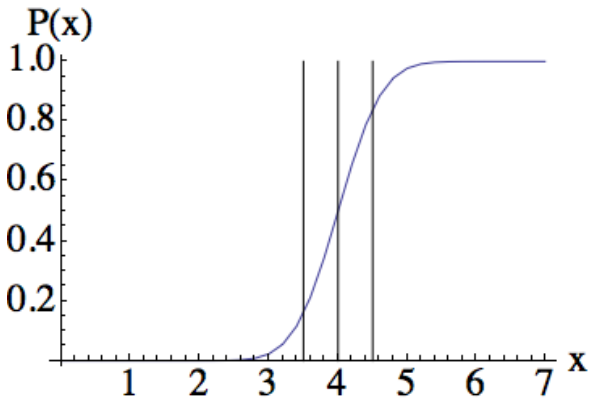
Gaussian density, cont'd

Please see Prof. Greytak's probability notes (page 13), and the subsections below, and your recitation instructors, for more on how to do the integrals. The conclusion here is that for the gaussian distribution,

$$\mathrm{Var}(x) = \sigma^2, \qquad \mathrm{stdev}(x) = \sigma.$$

Cumulative probability:

$$P(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \int_{-\infty}^{x} d\zeta \; e^{-\frac{(\zeta-a)^2}{2\sigma^2}}.$$



This is an example to get us comfortable with the notation, and it's an example which arises over and over. Soon: why it arises over and over (Central Limit Theorem).

2.1.1 A trick for doing the gaussian integral by squaring it

Consider the distribution

$$P(x_1, x_2) = C^2 e^{-\frac{x_1^2}{2s^2}} e^{-\frac{x_2^2}{2s^2}}$$

describing two statistically independent random variables on the real line, each of which is governed by the Gaussian distribution. What is C?

$$1 = C^2 \int dx_1 dx_2 e^{-\frac{x_1^2}{2s^2}} e^{-\frac{x_2^2}{2s^2}} = C^2 \int r dr d\theta e^{-\frac{r^2}{2s^2}} = 2\pi C^2 \int_0^\infty r dr e^{-\frac{r^2}{2s^2}}$$

This is the square of a single gaussian integral. Let $u = \frac{r^2}{2s^2}$, $\quad r dr = s^2 du$:

$$1 = 2\pi C^2 s^2 \int_0^\infty du e^{-u} = 2\pi C^2 s^2 \left(-e^{-u}\right) |_0^\infty = 2\pi C^2 s^2$$

2.1.2 A trick for evaluating moments of the gaussian distribution

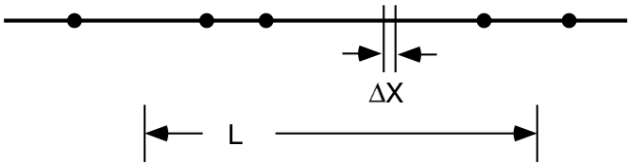
The key formula to use is: for any $\mathcal{X}$,

$$\int_{-\infty}^{\infty} d\eta \ \eta^2 e^{-\mathcal{X}\eta^2} = -\frac{\partial}{\partial \mathcal{X}} \int_{-\infty}^{\infty} d\eta \ e^{-\mathcal{X}\eta^2}.$$

Another example: Poisson density

Imagine a situation where events occur over and over at random points on the real line. This real line could describe *e.g.* the random times at which a geiger counter clicks from a radioactive source, or random positions of galaxies. More precisely, suppose the events satisfy the conditions:

1. In the limit $dx \to 0$, the probability that one and only one event occurs between x and $x + dx$ is $r dx$ with r independent of x. (r is a *rate*.)
2. The probability of an event occurring in some interval dx is independent of what happens in other intervals. (Roughly: the clicks of the counter don't care about each other. **Not:** each radioactive decay triggers an avalanche...)



Under these circumstances, in a *finite* interval of length L (*i.e.* not an infinitesimal interval like dx),

$$\text{Prob}(n \text{ events occur}) \equiv p(n) = \frac{1}{n!}(rL)^n e^{-rL} \qquad \text{(Poisson density)}$$

Notice that the product

$$rL = \text{rate} \times \text{sample size}$$

is dimensionless, as it must be if we are going to exponentiate it. We have pulled this expression out of nowhere. A proof that this expression follows from the starting assumptions 1,2 appears in §2 of Greytak notes. A slightly different explanation of its origin appears at the end of these notes in subsection 2.5. Like the Gaussian, this distribution also makes many appearances, such as:

- radioactive decay (then x is time)
- locations of impurities
- typos in a set of lecture notes
- deaths by horse of Prussian cavalry officers
- more examples on pset 3

Here: analyze consequences.

Poisson density, cont'd

Rewrite as a continuous density:

$$p(y) = \sum_{n=0}^{\infty} p(n)\delta(y-n).$$

To do: check normalization, compute mean and variance.

$$\begin{aligned}
\text{Normalize:} \qquad 1 &\stackrel{!}{=} \int_{-\infty}^{\infty} dy p(y) = \int_{-\infty}^{\infty} dy \sum_{n=0}^{\infty} p(n)\delta(y-n) \\
&= \sum_{n=0}^{\infty} p(n) \underbrace{\left(\int_{-\infty}^{\infty} dy \delta(y-n) \right)}_{=1} \\
&= e^{-rL} \sum_{n=0}^{\infty} \frac{1}{n!}(rL)^n = e^{-rL} e^{rL} = 1.
\end{aligned}$$

$$\langle n \rangle = \int_{-\infty}^{\infty} dy p(y) y = \int_{-\infty}^{\infty} dy \sum_{n=0}^{\infty} p(n)\delta(y-n)y = \sum_{n=0}^{\infty} np(n)$$

$$= e^{-rL} \underbrace{\sum_{n=0}}_{n=0 \text{ doesn't contribute}}^{\infty} \frac{n}{n!}(rL)^n = e^{-rL} rL \underbrace{\sum_{n=1}^{\infty} \frac{1}{(n-1)!}(rL)^{n-1}}_{e^{rL}} = rL.$$

$$\text{or:} \qquad \frac{\partial}{\partial r}\left[e^{rL} = \sum_{n=0}^{\infty} \frac{1}{n!}(rL)^n \right]$$

$$\implies \qquad Le^{rL} = \underbrace{\sum_{n=0}^{\infty} \frac{n}{n!}(rL)^n}_{\text{what we want}} \frac{1}{r}.$$

Poisson density, cont'd

$$\langle n^2 \rangle = ... = \sum_{n=0}^{\infty} n^2 p(n) = ... = (rL)(rL+1).$$

Differentiate e^{rL} twice. See Greytak notes for more of this.

$$\Longrightarrow \qquad \mathrm{Var}(n) = \langle n^2 \rangle - \langle n \rangle^2 = rL.$$
$$rL = \langle n \rangle.$$

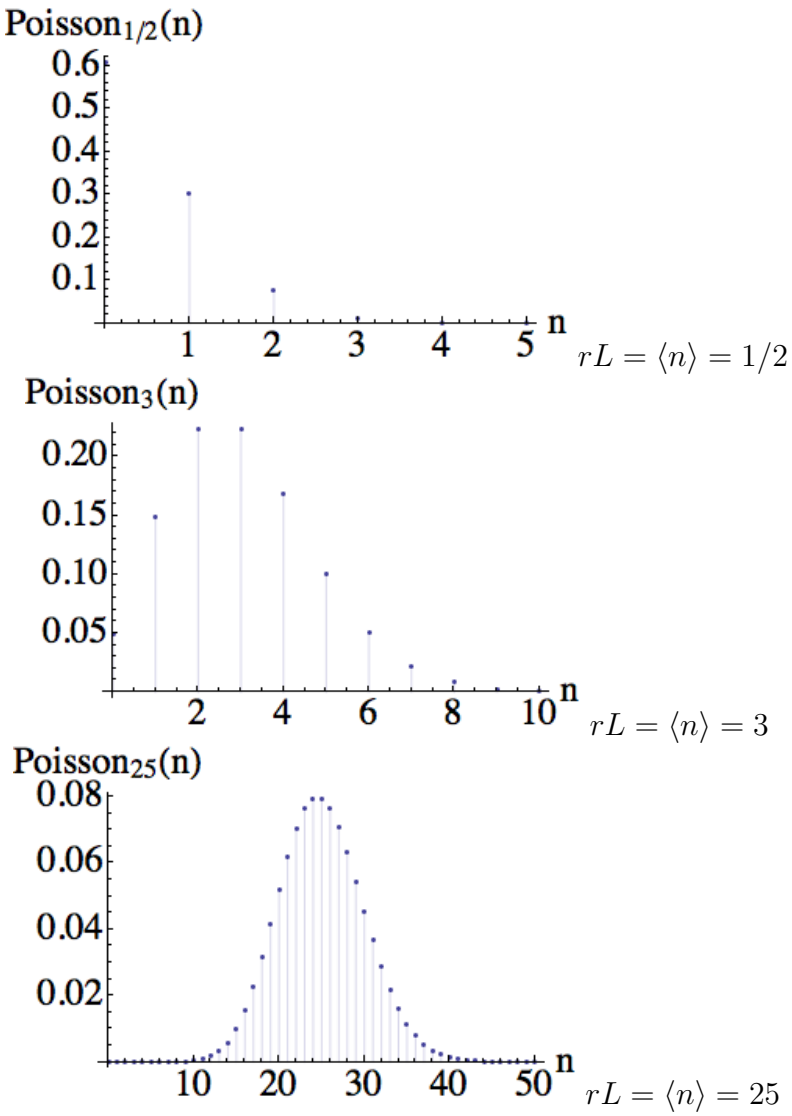
The Poisson density has the same mean and variance. The std deviation is $\sqrt{rL}$.

Note that only the (dimensionless) combination rL appears in the distribution itself.

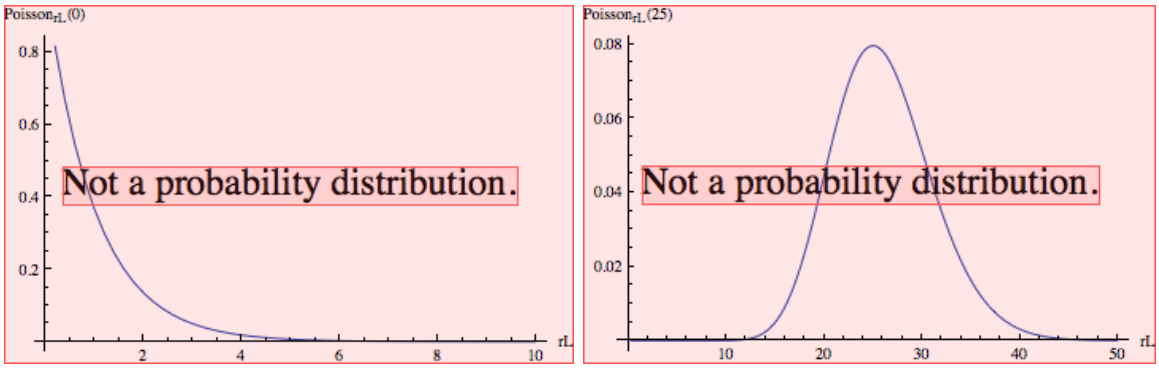
Rewrite:

$$p(n) = \frac{1}{n!} \langle n \rangle^n e^{-\langle n \rangle}.$$

Getting used to this abuse of notation is useful: n is the random variable.
$\langle n \rangle$ is a *number*.



As we increase rL the (envelope of the) distribution is visibly turning into a gaussian with $\langle n \rangle = rL$, $\mathrm{Var}(n) = rL$, stdev $= \sigma = \sqrt{rL}$. (This is an example of a more general phenomenon which we'll discuss soon, in 2.3.)



2.2 Two random variables

The extension of the previous discussion to encode partial information about *two* degrees of freedom rather than one is pretty straightforward.

$$d\zeta d\eta \underbrace{p_{x,y}}_{\downarrow}(\zeta,\eta) \qquad \equiv \qquad \mathrm{Prob}\big(\zeta \le x < \zeta + d\zeta \text{ and } \eta \le y < \eta + d\eta\big)$$

"joint probability density"

Think of x, y as coordinates on a map; $p(x,y)$ is a mountain range of probability.

$$\underbrace{P(\zeta,\eta)}_{\downarrow} \qquad\qquad \equiv \mathrm{Prob}\big(x \le \zeta \text{ and } y \le \eta\big)$$

$$\text{"joint cumulative probability"} \quad = \int_{-\infty}^{\zeta} d\zeta' \int_{-\infty}^{\eta} d\eta' \; p_{x,y}(\zeta',\eta')$$

$$p_{x,y}(\zeta,\eta) = \partial_\zeta \partial_\eta P(\zeta,\eta).$$

$$\langle f(x,y)\rangle = \int_{-\infty}^{\infty} d\zeta \int_{-\infty}^{\infty} d\eta \; f(\zeta,\eta)\; p_{x,y}(\zeta,\eta).$$

As usual, the fact that some outcome must occur requires us to normalize the distribution as

$$1 = \int_{\text{all possibilities}} p = \langle 1\rangle = \int_{-\infty}^{\infty} d\zeta \int_{-\infty}^{\infty} d\eta \; p_{x,y}(\zeta,\eta).$$

[End of Lecture 4.]

Reduction to a single variable

A new twist with two variables is the following new thing to do: There are several different ways to get a distribution for a single random varaible from a joint distribution.

If we have no knowledge of y, we can get a probability for x by integrating over all y against the joint distribution:

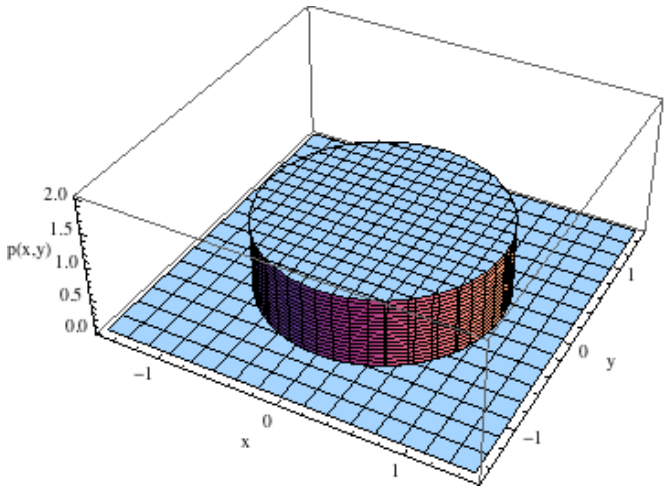
$$p_x(\zeta) = \int_{-\infty}^{\infty} d\eta \; p_{x,y}(\zeta,\eta)$$

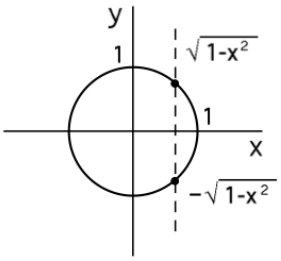
This operation is usefully called "squashing the mountain of probability".

(Similarly, if we have no knowledge of x, we can integrate over all possible values of x to get a probability distribution for y: $p_y(\eta) = \int_{-\infty}^{\infty} d\zeta \; p_{x,y}(\zeta,\eta)$.)

Example: "hockey puck" distribution

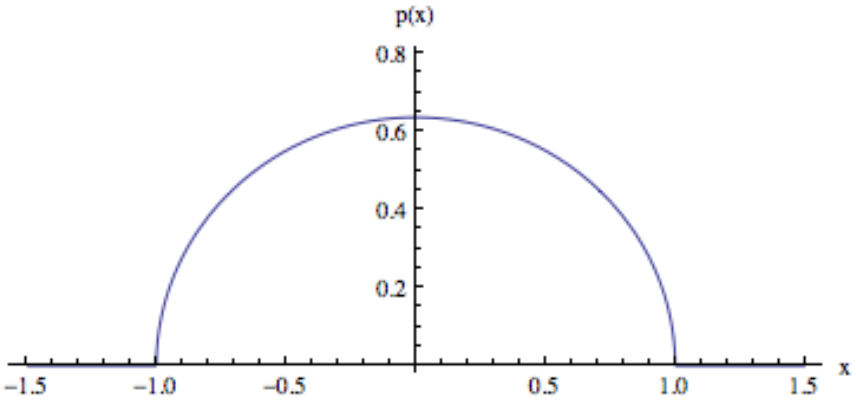
$$p(x,y) = \begin{cases} \frac{1}{\pi}, \text{ for } x^2+y^2 \le 1 \\ \\ 0, \text{ for } x^2+y^2 > 1 \end{cases} .$$





top view:

$$p(x) = \int_{-\infty}^{\infty} dy\, p(x,y) = \begin{cases} \int_{-\sqrt{1-x^2}}^{\sqrt{1-x^2}} dy \frac{1}{\pi} = \frac{2}{\pi}\sqrt{1-x^2}, \text{ for } x \le 1 \\ \\ 0, \text{ for } x > 1 \end{cases}$$



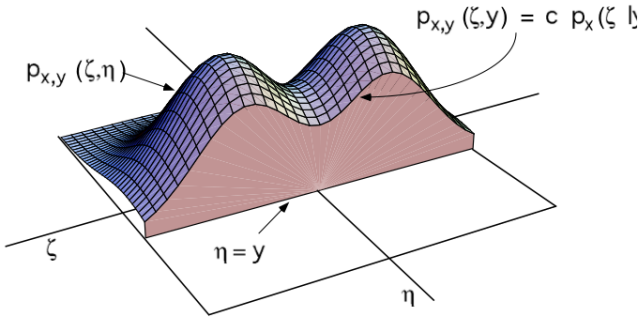
Here is a second way to get a probability distribution for one variable from a joint distribution for two. It is a little more delicate.

Conditional probability density

Suppose we *do* know something about y. For example, suppose we have a probability distribution $p(x, y)$ and suddenly discover a new way to measure the value of y; then we'd like to know what information that gives us about x:

$$p_x(\zeta|y)d\zeta \equiv \text{Prob}\big(\zeta \le x < \zeta + d\zeta \text{ given that } \eta = y\big)$$

is a single-variable probability density for x; ζ is a dummy variable as usual. y is a parameter of the distribution $p_x(\zeta|y)$, not a random variable anymore. The operation of forming the conditional probability can be described as "slicing the mountain": since we are specifying y, we only care about the possibilities along the slice $\eta = y$ in the picture.



This means that $p(\zeta|y)$ must be proportional to $p_{x,y}(\zeta, y)$. There is no reason for the latter quantity to be normalized as a distribution for ζ, however:

$$\underbrace{p_{x,y}(\zeta, y)}_{\text{a slice: not normalized}} = \underbrace{c}_{\text{to be determined}} \underbrace{p_x(\zeta|y)}_{\text{a normalized prob density for } \zeta}$$

The condition that $p_x(\zeta|y)$ be normalized determines the constant c. As follows:

$$\underbrace{\int_{-\infty}^{\infty} d\zeta p_{x,y}(\zeta, y)}_{p_y(\eta=y)} = c \underbrace{\int_{-\infty}^{\infty} d\zeta p_x(\zeta|y)}_{=1}$$

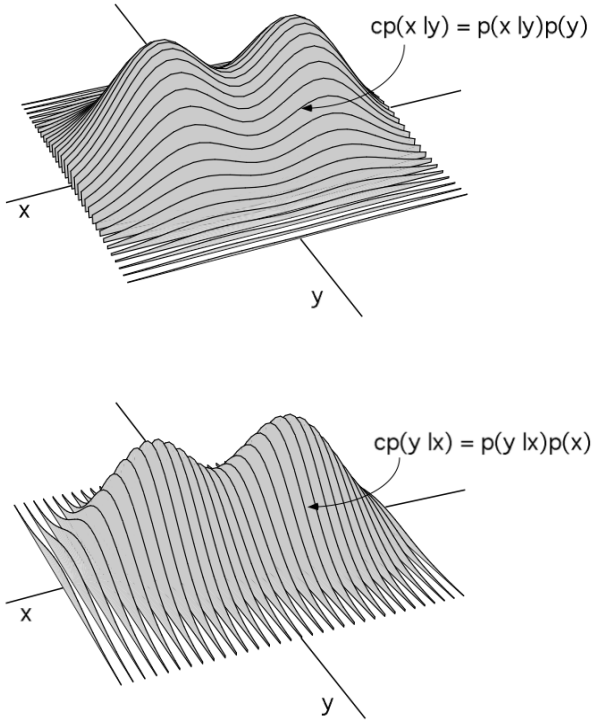
$$\Longrightarrow \qquad c = p_y(\eta = y).$$

$$p_{x,y}(\zeta, y) = p_y(\eta = y)p_x(\zeta|y)$$

Translation into human language:

$$\boxed{p(x|y) = \frac{p(x,y)}{p(y)}} \qquad \text{“Bayes' Theorem”}$$

The only content of this Theorem is: Slice the mountain, then normalize.

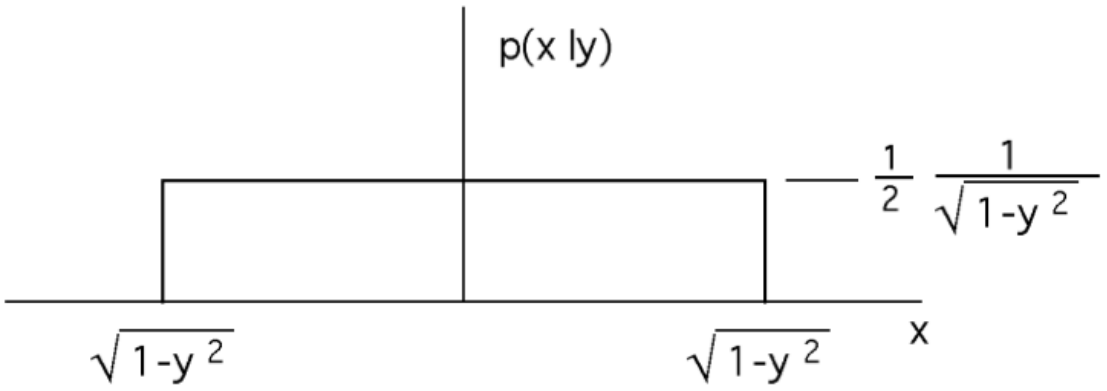


Slicing the other way allows Bayes to put his name on another formula:

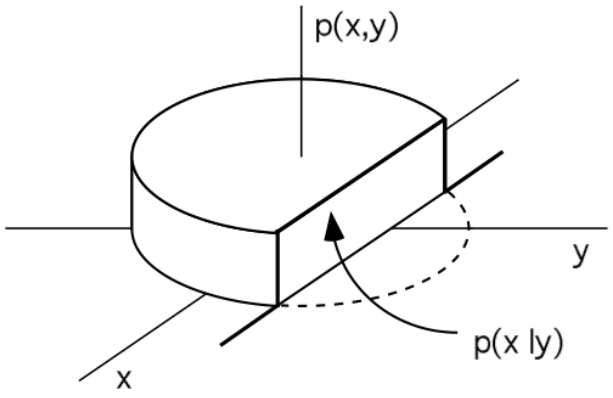
$$p(x,y) = p(x|y)p(y) = p(y|x)p(x) \quad .$$

Looked at this way, this relation is a prescription for building the joint probability mountain from conditional probabilities $p(x|y)$ and the single-variable probability $p(y)$ for the ‘conditioning variable’ y.

Returning to the hockey puck example:



$p(x|y)$ gets skinnier and (therefore) higher as $y \to 1$.



Notice that as the conditioning variable y approaches 1, we become more and more certain about the value of x.

With these defs, we can make a useful characterization of a joint distribution:

Statistical Independence

Two random variables are *statistically independent* (SI) if the joint distribution *factorizes*:

$$p_{x,y}(\zeta, \eta) = p_x(\zeta) p_y(\eta)$$

or, equivalently, if

$$p(x|y) = \frac{p(x,y)}{p(y)} = p(x), \quad \text{independent of } y$$

(in words: the conditional probability for x given y is independent of the choice of y. Telling me y leaves me as ignorant of x as before.) and

$$p(y|x) = \frac{p(x,y)}{p(x)} = p(y), \quad \text{independent of } x$$

So: for SI RVs, knowing something about one variable gives no additional information about the other. You are still just as ignorant about the other as you were before you knew anything about the first one.

Hockey puck example: x and y are NOT SI.

Example: Deriving the Poisson density.

Now you have all the ingredients necessary to follow the discussion of the derivation of the Poisson density in Prof. Greytak's notes. You have seen a different derivation of this distribution in recitation, which is reviewed in the final sub section of these notes (2.5).

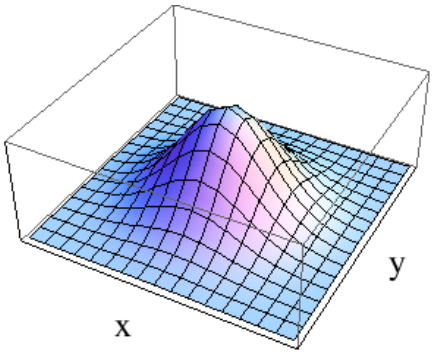
Example: Jointly gaussian random variables

$$p(v_1, v_2) = \frac{1}{2\pi\sigma^2\sqrt{1-\rho^2}} \exp[-\frac{v_1^2 - 2\rho v_1 v_2 + v_2^2}{2\sigma^2(1-\rho^2)}]$$

v_1, v_2: random variables. v is for 'voltage' as we'll see later.
ρ, σ: parameters specifying the density. They satisfy $\sigma > 0; -1 \leq \rho \leq 1$.

Start by analyzing the special case $\rho = 0$:

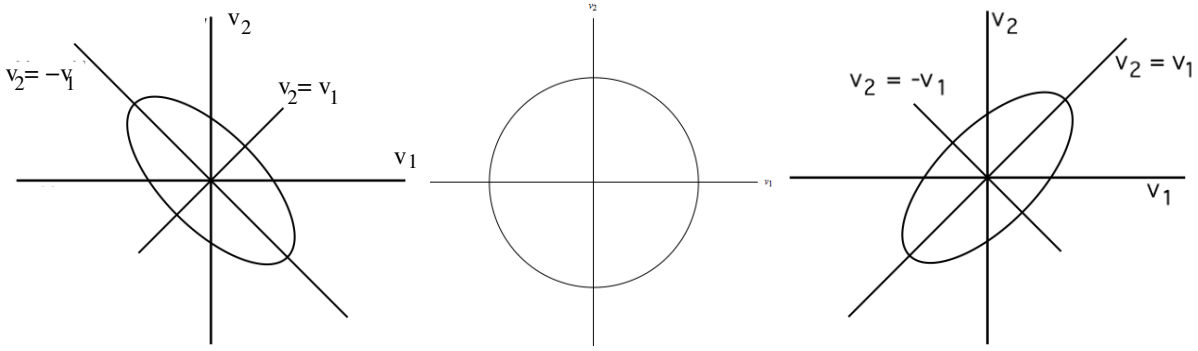
$$p(v_1, v_2) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{v_1^2}{2\sigma^2}} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{v_2^2}{2\sigma^2}} \quad = p(v_1) \cdot p(v_2),$$



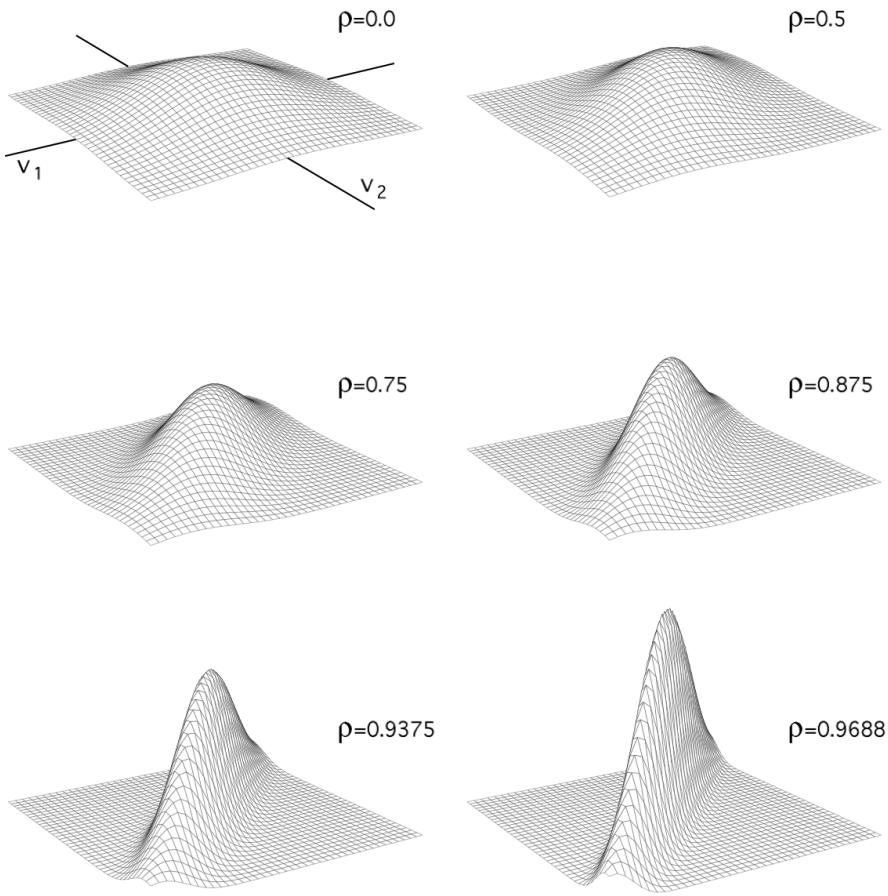
a circularly symmetric gaussian mountain. In this case, v_1 and v_2 are SI, as demonstrated by the last equality above. Slicing a gaussian mountain gives gaussian slices.

Now consider $\rho \neq 0$.

Start by plotting. Contours of constant probability are ellipses in the (v_1, v_2) plane:



$\rho = -1 + \epsilon$ $\rho = 0$ $\rho = 1 - \epsilon$



Reduction to a single variable *i.e.* "squashing the mountain"

$$\begin{aligned} p(v_1) &= \int_{-\infty}^{\infty} dv_2 p(v_1, v_2) \\ &= \frac{1}{2\pi\sigma^2\sqrt{(1-\rho^2)}} \exp[-\frac{v_1^2}{2\sigma^2}] \int_{-\infty}^{\infty} dv_2 \exp[-\frac{(v_2-\rho v_1)^2}{2\sigma^2(1-\rho^2)}] \end{aligned} \tag{1}$$

Here we completed the square in the exponent of $p(v_1, v_2)$:

$$v_1^2 - 2\rho v_1 v_2 + v_2^2 = (v_2 - \rho v_1)^2 - \rho^2 v_1^2 - v_1^2 = (v_2 - \rho v_1)^2 - (1-\rho^2)v_1^2$$

But then this is just a gaussian integral for v_2:

$$p(v_1) = \frac{1}{2\pi\sigma^2\sqrt{(1-\rho^2)}} \exp[-\frac{v_1^2}{2\sigma^2}] \underbrace{\int_{-\infty}^{\infty} dv_2 \exp[-\frac{(v_2-\rho v_1)^2}{2\sigma^2(1-\rho^2)}]}_{\sqrt{2\pi\sigma^2(1-\rho^2)}}$$

$$= \frac{1}{\sqrt{2\pi\sigma^2}} \exp[-\frac{v_1^2}{2\sigma^2}] \tag{2}$$

This is independent of ρ. Similarly for $p(v_2)$.

Statistical Independence? The distribution only factorizes if $\rho = 0$. If $\rho \neq 0$, not SI.

Information about the correlations between v_1 and v_2 (i.e. all data about the effect of ρ on the joint distribution) is lost in the squashing process. This information is hiding in the ...

Conditional probability: (i.e. "slicing the mountain")

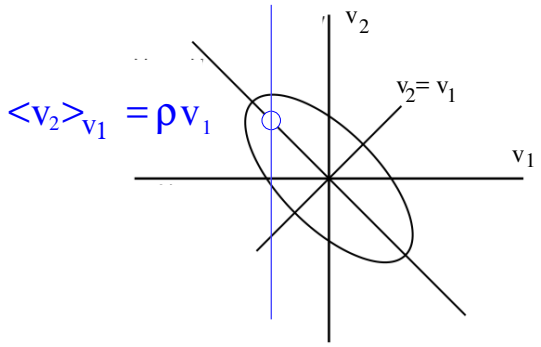
$$\begin{aligned} p(v_2|v_1) &= \frac{p(v_1, v_2)}{p(v_1)} \\ &= \frac{\sqrt{2\pi\sigma^2}}{2\pi\sigma^2\sqrt{1-\rho^2}} \exp\left[-\frac{v_1^2 - 2\rho v_1 v_2 + v_2^2}{2\sigma^2(1-\rho^2)} + \frac{v_1^2(1-\rho^2)}{2\sigma^2(1-\rho^2)}\right] \end{aligned} \tag{3}$$

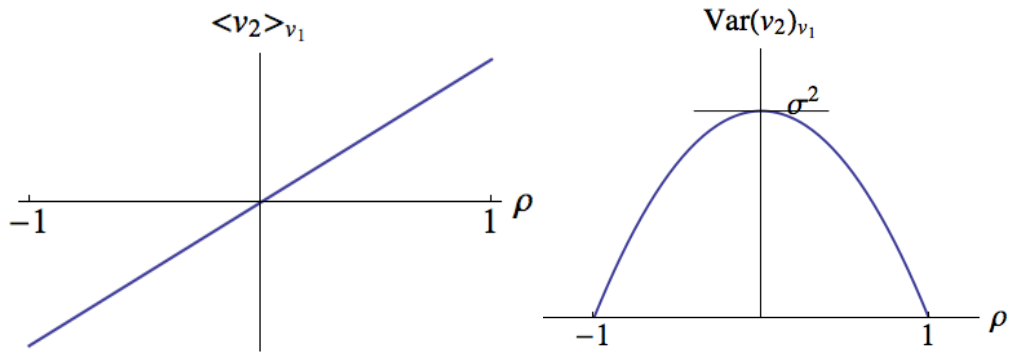
But now the same manipulation of completing the square as above shows that this is

$$p(v_2|v_1) = \frac{1}{\sqrt{2\pi\sigma^2}} \frac{1}{\sqrt{1-\rho^2}} \exp\left[-\frac{(v_2 - \rho v_1)^2}{2\sigma^2(1-\rho^2)}\right] \tag{4}$$

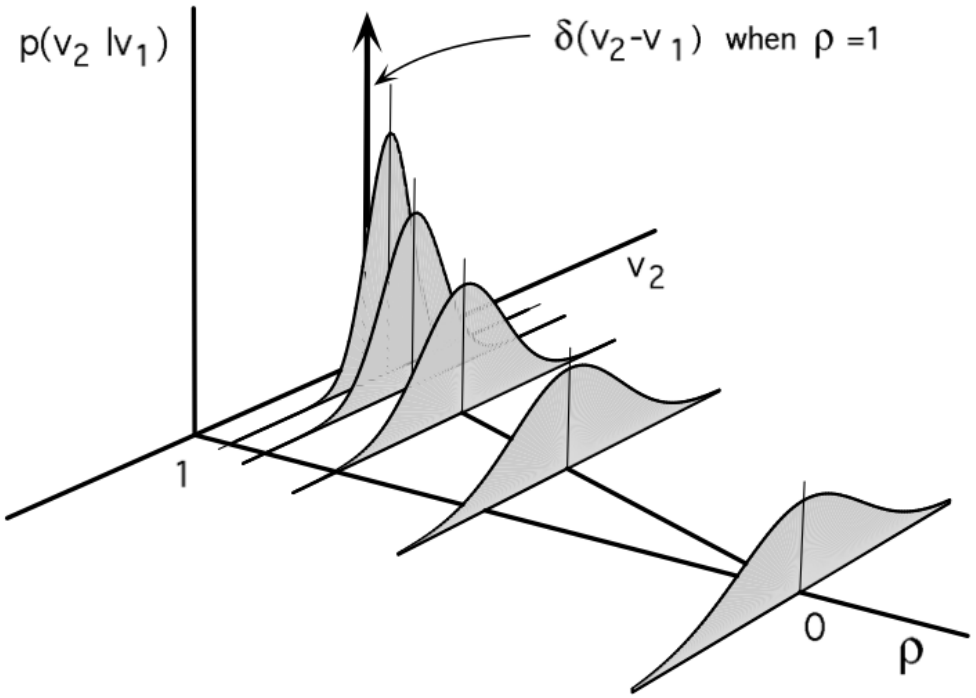
This is a probability density for v_2 which is gaussian, with mean ρv_1 (remember v_1 is a parameter labelling this distribution for the random variable v_2), with stdev $= \sigma\sqrt{1-\rho^2}$.

Look at the contour plots again. Pick a v_1. This determines a distribution for v_2 with $\langle v_2 \rangle_{v_1} = \rho v_1$:





Plots of $p(v_2|v_1)$ for various v_1:



Note: as $\rho \to 1$, $p(v_2|v_1) \to \delta(v_2 - \rho v_1)$.

Correlation function:

$$\equiv \langle v_1 v_2 \rangle = \int_{-\infty}^{\infty} dv_1 \int_{-\infty}^{\infty} dv_2 \; v_1 v_2 p(v_1, v_2)$$

We could just calculate, but we've already done the hard part. Use Bayes here:

$$\langle v_1 v_2 \rangle = \int_{-\infty}^{\infty} dv_1 \int_{-\infty}^{\infty} dv_2 \; v_1 v_2 p(v_2|v_1) p(v_1)$$

$$\langle v_1 v_2 \rangle = \int_{-\infty}^{\infty} dv_1 v_1 p(v_1) \underbrace{\int_{-\infty}^{\infty} dv_2 \; v_2 p(v_2|v_1)}_{\text{conditional mean}=\rho v_1}$$

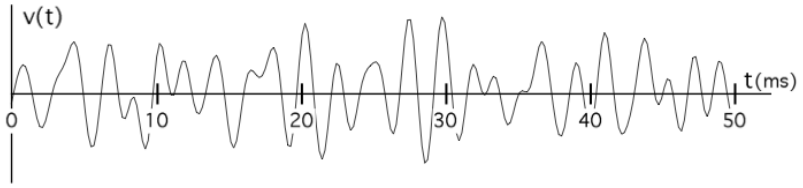
in our joint gaussian example

$$\langle v_1 v_2 \rangle = \rho \underbrace{\int_{-\infty}^{\infty} dv_1 v_1^2 p(v_1)}_{=\sigma^2} = \rho \sigma^2.$$

$\rho > 0 : v_1, v_2$ correlated
$\rho < 0 : v_1, v_2$ anticorrelated

Whence this probability distribution?

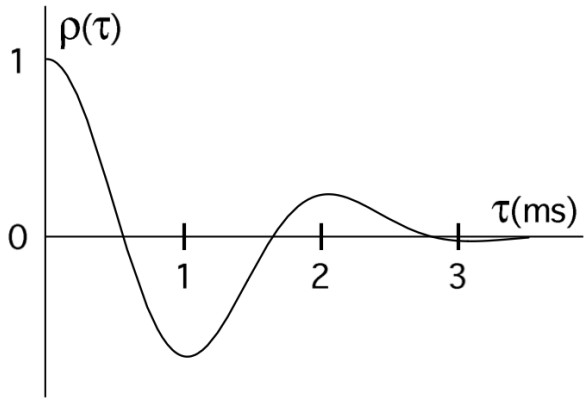
Johnson noise: thermal noise in the voltage across an electric circuit with L, R, C (but no batteries) at room temperature.



This is an attempt at a random curve with $\langle v \rangle = 0$, $\langle v^2 \rangle - \langle v \rangle^2 = \sigma^2$, something which depends on temperature, but with some excess power at a certain frequency (500 Hz). To see this excess power:

Define: $v_1 \equiv V(t), v_2 \equiv V(t+\tau)$. What do we expect as τ varies?

- For large τ: v_1, v_2 are uncorrelated $\implies \rho \to 0$. Each is gaussian. σ depends on T.
- For $\tau \to 0$: v_1 and v_2 are highly correlated: $\rho \to 1$.
- For intermediate τ, they can be (but need not be) anticorrelated:



e.g.:

Different circuits (different L, R, C in different arrangements) will have different functions $\rho(\tau)$.

$$\langle v_1 v_2 \rangle = \langle V(t) V(t+\tau) \rangle \propto \rho(\tau)$$

is called the correlation function. It characterizes the noise in a circuit, and can be used to diagnose features of an unknown circuit. [Note: σ depends on temperature, but ρ does not.]

Why $\rho < 0$? Consider a circuit with a sharp resonance. The fluctuations can get a large contribution from a particular resonant frequency ω. Then at $\tau = \frac{\text{period}}{2} = \frac{2\pi}{\omega} \cdot \frac{1}{2}$, $v_1 > 0$ means $v_2 < 0$. Hence a negative correlation function: $\langle v_1 v_2 \rangle \propto \rho < 0$.

2.3 Functions of a random variable

Consider a gas in thermal equilibrium. Suppose you know $p(v)$, where v is the speed of *one* atom.

$$\text{Kinetic energy of an atom } \equiv \text{KE} = \frac{1}{2}mv^2$$

So: what is $p(KE)$?

In general, given $p(x)$ what is $p(f(x))$?

Various methods, some fancier than others. Here is a very robust 3-step pictorial method:

$\boxed{A}$ Sketch $f(x)$. Find where $f(x) < \eta$: $R_\eta \equiv \{x | f(x) < \eta\}$

$\boxed{B}$ Integrate $p_x(\zeta)$ over regions found in $\boxed{A}$. This gives the cumulative probability for f:

$$P_f(\eta) = \int_{R_\eta} d\zeta p_x(\zeta).$$

$\boxed{C}$ As usual, we can get the probability distribution by:

$$p_f(\eta) = \frac{d}{d\eta} P_f(\eta) \ .$$

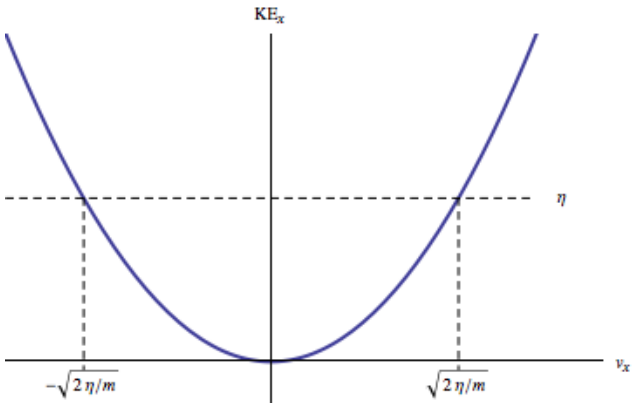
Example 1: Kinetic energy of an ideal gas.

We'll see much later in the course that a molecule or atom in an ideal gas has a velocity distribution of the form:

$$p(\underbrace{v_x}_{x-\text{component of velocity}}) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp[-\frac{v_x^2}{2\sigma^2}] \qquad \text{with } \sigma = \sqrt{kT/m}$$

(same for v_y, v_z. let's ignore those for now)

Define $\text{KE}_x \equiv \frac{1}{2}mv_x^2$. What's the resulting $p(\text{KE}_x)$?



$\boxed{A}$ $R_\eta = [-\sqrt{2\eta/m}, \sqrt{2\eta/m}]$.

$\boxed{B}$

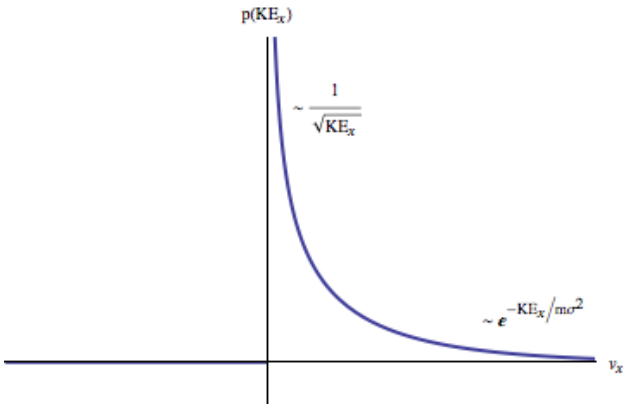
$$P_{\text{KE}_x}(\eta) = \int_{-\sqrt{2\eta/m}}^{\sqrt{2\eta/m}} p_{v_x}(\zeta)d\zeta$$

And: DON'T DO THE INTEGRAL!

$\boxed{C}$

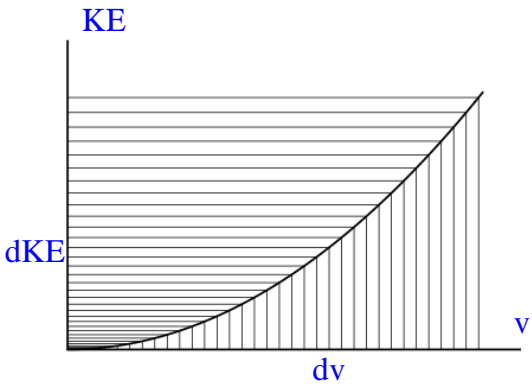
$$p_{\text{KE}_x}(\eta) = \frac{d}{d\eta}P_{\text{KE}_x} = \frac{1}{\sqrt{2m\eta}}p_{v_x}(\sqrt{2\eta/m}) - (-\frac{1}{\sqrt{2m\eta}})p_{v_x}(\sqrt{2\eta/m})$$

$$p(\mathrm{KE}_x) = \begin{cases} \frac{1}{\sqrt{\pi m\sigma^2 \mathrm{KE}_x}} \exp[-\frac{\mathrm{KE}_x}{m\sigma^2}], & \text{for } \mathrm{KE}_x > 0 \\ \\ 0 & \text{for } \mathrm{KE}_x < 0 \end{cases}$$



Not gaussian at all! Completely different shape from $p(v_x)$.

Whence the divergence? As $v \to 0$, $\frac{dv}{dKE_x} \to \infty$. $\Longrightarrow$ Pileup at $v = 0$.



[End of Lecture 5.]

Note that $p_f(\eta)$ is in general a completely different function from $p_x(\zeta)$. This is the purpose of the pedantic notation with the subscripts and the dummy variables. A more concise statement of their relationship is

$$p_f(\eta) = \int_{-\infty}^{\infty} d\zeta p_x(\zeta)\delta(\eta - f(\zeta)).$$

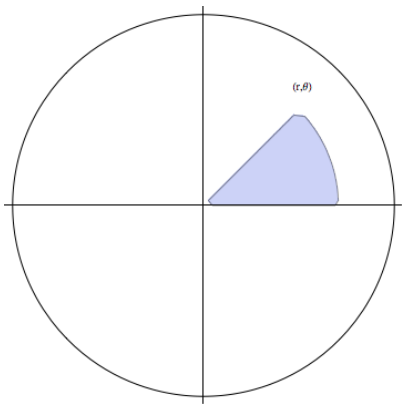
Unpacking this neat formula is what's accomplished by the 3-step graphical method. Let's work another example with that method.

Example 2: Hockey puck distribution again.

$$p(x,y) = \begin{cases} \frac{1}{\pi}, \text{ for } x^2+y^2 \leq 1 \\ \\ 0, \text{ for } x^2+y^2 > 1 \end{cases} .$$

This the same as the paint droplet problem on pset 2.

Find $p(r,\theta)$.



$\boxed{A}$: pick some r, θ. Sketch the region $R_{r,\theta} \equiv \{(r',\theta')|r' < r, \theta' < \theta\}$.

$\boxed{B}$:

$$P(r,\theta) = \int_{R_{r,\theta}} dxdyp(x,y) = \begin{cases} \int_0^r r'dr \int_0^\theta d\theta' \frac{1}{\pi} = \frac{1}{\pi}\pi r^2 \frac{\theta}{2\pi} & \text{for } r < 1 \\ \\ \frac{\theta}{2\pi} & \text{for } r > 1. \end{cases}$$

It's just the fraction of the area of the disc taken up by $R_{r,\theta}$.

$\boxed{C}$:

$$p(r,\theta) = \partial_r \partial_\theta P(r,\theta) = \begin{cases} \frac{r}{\pi} & r < 1 \\ 0 & r > 1 \end{cases}$$

Check against pset 2:

$$p(r) = \int_0^{2\pi} p(r,\theta)d\theta = \begin{cases} 2r & r < 1 \\ 0 & r > 1 \end{cases}$$

$$p(\theta) = \int_0^{2\pi} p(r,\theta)dr = \frac{r^2}{2\pi}|_0^1 = \frac{1}{2\pi}$$

Note: $p(r,\theta) = p(r)p(\theta)$. With this joint distribution r, θ are SI, although x, y are not.

2.4 Sums of statistically-independent random variables

Variables describing the macrostate of a gas in the thermodynamic limit = sums of 10^{23} variables describing the individual particles.
Energy ideal gas: $E = \sum_{i=1}^{10^{23}} \mathrm{KE}_i$

Claim: "all of thermodynamics is about sums of SI RVs". (a more accurate claim: we can get all of the effects of thermodynamics from sums of SI RVs.)

Consider N SI RVs $x_i, i = 1..N$. Let $S_N \equiv \sum_{i=1}^{N} x_i$. (Note: $p_{x_i}(\zeta)$ need not be independent of i.)

$$\begin{aligned} \langle S_N \rangle &= \int dx_1 ... dx_N \underbrace{S_N}_{=x_1+x_2+...+x_N} p(x_1, x_2 ... x_N) \\ &= \sum_{i=1}^{N} \int dx_i x_i p(x_i) = \sum_{i=1}^{N} \langle x_i \rangle \end{aligned} \tag{5}$$

All the other integrals are just squashing in $N-1$ dimensions.
Mean of sum is sum of means (whether or not RVs are SI).

$$\begin{aligned} \mathrm{Var}(S_N) &= \langle (S_N - \langle S_N \rangle)^2 \rangle \\ &= \langle (x_1 - \langle x_1 \rangle + x_2 - \langle x_2 \rangle + ... + x_N - \langle x_N \rangle)^2 \rangle \\ &= \sum_{ij} \langle (x_i - \langle x_i \rangle)(x_j - \langle x_j \rangle) \rangle \\ &= \sum_{ij} (\langle x_i x_j \rangle - \langle x_i \rangle \langle x_j \rangle) \\ &= \sum_{i \neq j} \underbrace{0}_{\text{by SI}} + \sum_{i=j} \left(\langle x_i^2 \rangle - \langle x_i \rangle^2 \right) \\ \mathrm{Var}(S_N) &= \sum_{i=1}^{N} \mathrm{Var}(x_i) \end{aligned} \tag{6}$$

More on that crucial step:

$$\begin{aligned} \langle x_i x_j \rangle &= \int dx_1 .. dx_N \; p(x_1 ... x_N) x_i x_j \\ &\overset{SI}{=} \int .. p(x_i) ... p(x_j) ... x_i x_j \end{aligned}$$

$$= (\int dx_i p(x_i) x_i)(\int dx_j p(x_j) x_j)$$
$$= \langle x_i \rangle \langle x_j \rangle \tag{7}$$

Correlation functions of SI RVs factorize.

SO:

$$\boxed{\text{Var(sum)} = \text{sum(Variances)}} \quad \text{if RVs are SI}$$

Crucial note: where correlations, this statement is *not* true. Suppose given distributions for two variables (results of squashing): $p(x_1), p(x_2)$.

example 1: Suppose perfect correlation: $p(x_1, x_2) = \delta(x_1 - x_2) p(x_1)$.
In every copy of the ensemble, $x_1 = x_2$.
Claim: then $\text{Var(sum)} = 4\ \text{Var}(x_1)$.

example 2: Suppose perfect anticorrelation: $p(x_1, x_2) \propto \delta(x_1 + x_2)$.
The sum is *always* zero! So $\text{Var}(x_1 + x_2) = 0$.

"iid" RVs

Consider n SI RVs, $x_1..x_n$. If all $p(x_i)$ are the same, and the x_i are SI, these are (naturally) called "independent and identically distributed" ("iid"). Let $\langle x_i \rangle \equiv \langle x \rangle$, $\text{Var}(x_i) \equiv \sigma^2$. Then

$$\langle S_n \rangle = n \langle x \rangle, \quad \text{Var}(S_n) = n\sigma^2$$

As $n \to \infty$,

$$\langle S_n \rangle \propto n, \quad \text{Var}(S_n) \propto n, \quad \text{stdev}(S_n) \propto \sqrt{n}\ .$$

$$\implies \quad \frac{\text{stdev}(S_n)}{\langle S_n \rangle} \propto \frac{1}{\sqrt{n}}. \quad \text{The distribution gets narrower as } n \to \infty.$$

Notes:

- This statement is still true if the $p(x_i)$ are different, as long as no subset dominates the mean. (i.e. it's not like: 1000 vars, 999 with $\langle x_i \rangle = 1$ and one with $x_\star = 10^{52}$.)
- We assumed means and variances are all (nonzero and) finite.
- This applies for both continuous and discrete RVs.
- This is the basis of statements like: "The statistical error of the opinion poll was x". That means that x^2 people were polled.

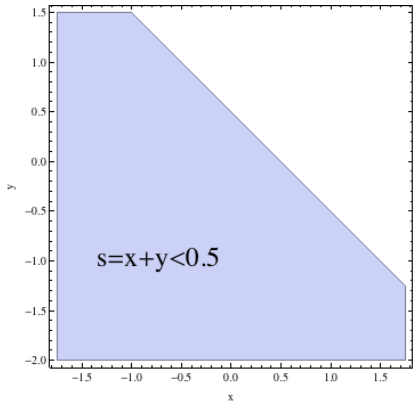
Towards the central limit theorem.

So far we know $\langle S_n \rangle, \mathrm{Var} S_n$. What can we say about the shape of $p(S_n)$?

Answer first for two vars x, y. $s \equiv x + y$. Given $p_{x,y}(\zeta, \eta)$, what's $p(s)$? One way to do it is:

$$p(s) = \int d\zeta d\eta \; p_{x,y}(\zeta, \eta) \delta(s - (\zeta + \eta)).$$

But let's use the graphical method:



$\boxed{A}$: Shade the region where $s = \zeta + \eta \leq \alpha$:

$\boxed{B}$: Write an expression for the cumulative probability for s:

$$P_s(\alpha) = \int_{-\infty}^{\infty} d\zeta \int_{-\infty}^{\alpha - \zeta} d\eta \; p_{x,y}(\zeta, \eta)$$

Don't do the integral.

$\boxed{C}$: The reason we don't do the integral is because we can use the Fundamental Theorem of Calculus here: $p_s(\alpha) = \frac{d}{d\alpha} P_s(\alpha) \stackrel{FTC}{=} \int_{-\infty}^{\infty} d\zeta p_{x,y}(\zeta, \alpha - \zeta)$. Note that this result is completely general – we haven't assumed anything about the joint distribution.

In the special case that x and y are SI:

$$\boxed{p_s(\alpha) = \int_{-\infty}^{\infty} d\zeta \; p_x(\zeta) p_y(\alpha - \zeta)} \quad (\star)$$

In words: The probability distribution for the sum of SI RVs is the *convolution* of the two original distributions.

Note that the apparent asymmetry between x and y in $(\star)$ is an illusion:

$$\zeta' \equiv \alpha - \zeta \quad \Longrightarrow \; p_s(\alpha) = \int_{-\infty}^{\infty} d\zeta' \; p_x(\alpha - \zeta') p_y(\zeta')$$

[Mathy aside on convolutions: given two functions $f(x), g(x)$, their convolution is defined to be

$$(f \otimes g)(x) \equiv \int_{-\infty}^{\infty} dz f(z) g(x - z).$$

A few useful properties of this definition that you can show for your own entertainment:

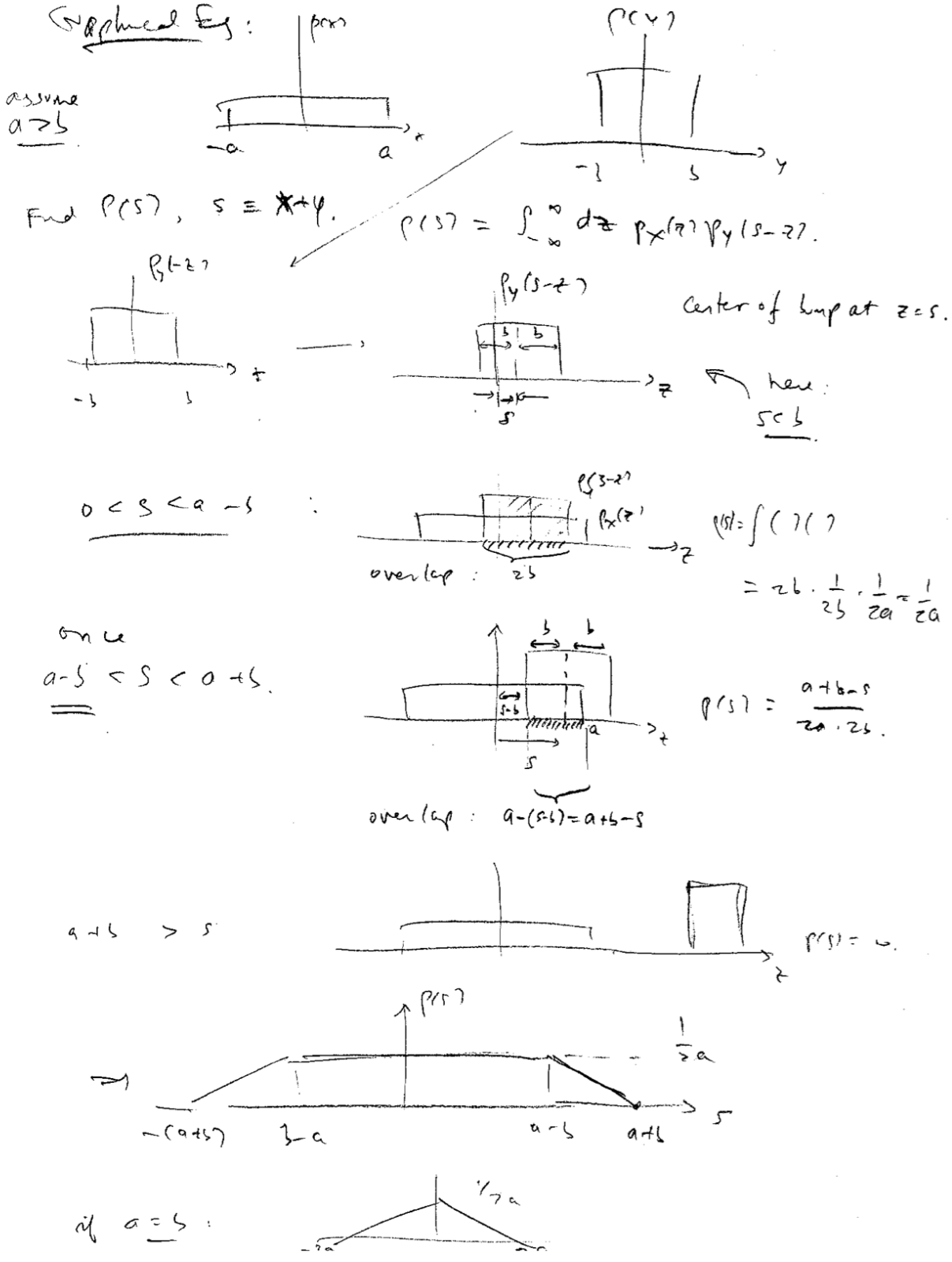
$$f \otimes g = g \otimes f$$

$$f \otimes (g + h) = f \otimes g + f \otimes h$$

$$f \otimes (g \otimes h) = (f \otimes g) \otimes h$$

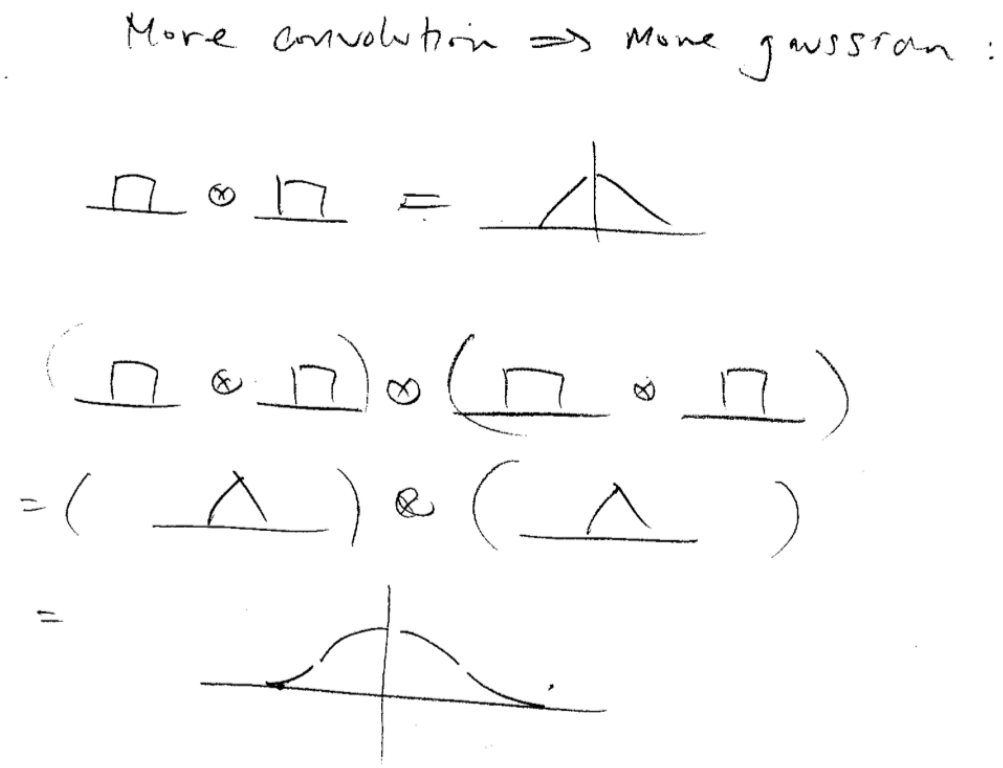
Fourier transform of convolution is multiplication.]

Graphical example of convolution, with steps.



Lessons

Lesson 1: convolution changes the shape.
Lesson 2: convolution makes the result more gaussian.
If you don't believe it, do it again:



See Greytak for more details on p(sum of 4 RVs).

Three famous exceptions to lessons:

1) Gaussians: Consider two gaussian RVs:

$$p_x(x) = \frac{1}{\sqrt{2\pi\sigma_x^2}} e^{-\frac{(x-a)^2}{2\sigma_x^2}}, \qquad p_y(y) = \frac{1}{\sqrt{2\pi\sigma_y^2}} e^{-\frac{(y-b)^2}{2\sigma_y^2}} .$$

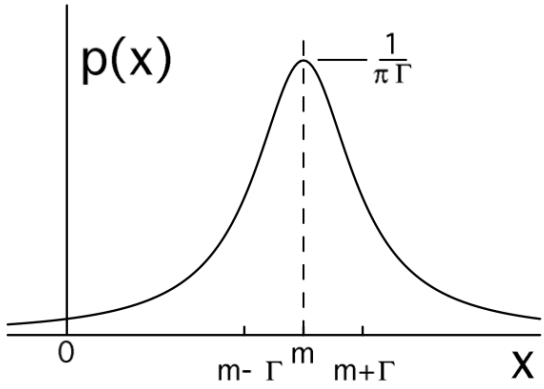
Claim [algebra, Greytak]: $p(s = x + y) = (p_x \otimes p_y)(s)$ is gaussian. By our previous analysis of SI RVs, we know its $\langle s \rangle, \mathrm{Var}(s)$. So if it's gaussian:

$$p(s) = \frac{1}{\sqrt{2\pi \underbrace{(\sigma_x^2 + \sigma_y^2)}_{\text{Vars add}}}} \exp[-\frac{(s - (a+b))^2}{2(\sigma_x^2 + \sigma_y^2)}]$$

2) Poisson: Claim [algebra, Greytak]: Sum of two SI Poisson RVs is also Poisson distributed. with mean = sum of means = variance. (Recall that a Poisson distribution is fully specified by the mean.)

3) Lorentzian:

$$p(x) = \frac{1}{\pi}\frac{1}{(x-m)^2+\Gamma^2}, \qquad \Gamma, m \text{ parameters}$$



Check normalization $\int_{-\infty}^{\infty} dx\ p(x) = 1$.
$\langle x \rangle = m$.
$p(m \pm \Gamma) = p(m)/2$. So: $\Gamma =$ "half-width at half-height" (or just "width").

Claim: sum of two Lorentzian-distributed SI RVs is also Lorentzian-distributed.

$$\Longrightarrow \langle s \rangle = m_1 + m_2, \quad \Gamma_{\text{sum}} = \Gamma_1 + \Gamma_2$$

What's the variance?

$$\langle x^2 \rangle = \int_{-\infty}^{\infty} dx \frac{x^2}{x^2+\Gamma^2} = \infty$$

$\infty + \infty = \infty$ so the variances do add. This is a distribution with "fat tails": it only falls off like $1/x^2$ away from the peak. If it fell off much slower it wouldn't even be normalizable.

This is a useful exception to remember when listening to people who work in finance blather on about how your money is safe because fluctuations will average out (for example when they bundle together lots of mortgages each with a fat-tailed distribution of outcomes...) and in the following.

Central Limit Theorem (CLT)

Let s_n = sum of n SI RVs, which are identically distributed (so: iid) ($\star$)
with mean $\langle x \rangle$ and variance σ_x^2 (which must be finite ($\star\star$)).
For large n ($\star\star\star$),
$p(s_n)$ can often be well-represented by ($\star\star\star\star$)
a gaussian, with mean $n\langle x \rangle$, and variance $n\sigma_x^2$.

$$i.e. \quad p(s_n) \underbrace{\approx}_{\text{can be well-rep. by}} \frac{1}{\sqrt{2\pi n\sigma_x^2}} \exp[-\frac{(s_n - n\langle x \rangle)^2}{2n\sigma_x^2}]$$

Fine print

($\star$): The theorem also works if not identical, as long as no subset dominates s.

($\star\star$): Think of Lorentzians. Convolving many distributions with fat tails gives another distribution with fat tails, *i.e.* the fluctuations about the mean are still large.

($\star\star\star$): The mathematicians among us would like us to say how large n has to be in order to achieve a given quality of approximation by a guassian. This is not our problem, because $n \sim 10^{23}$ gives an approximation which is good enough for anyone.

($\star\star\star\star$): The reason for vagueness here is an innocent one; it's so that we can incorporate the possibility of discrete variables. For example, flip a coin 1000 times. The probability distribution for the total number of heads looks like a gaussian if you squint just a little bit: the envelope of the histogram is gaussian. This is what we mean by "can be well-represented by". (More on this next.)

This is why we focus on gaussians.

An example: consider a sum of many independent Poisson-distributed variables. I already claimed that convolutions of Poissons is again Poisson. How is this consistent with the CLT which says that the distribution for the sum of Poisson-distributed SI RVs should be gaussian? When we sum more such (positive) RVs the mean also grows. The CLT means that (the envelope for) the Poisson distribution with large mean must be gaussian. (Look back at Poisson with $\langle n \rangle = rL = 25$ and you'll see this is true.)

We will illustrate the origin of the CLT with an example. We'll prove (next) that it works for many binomially-distributed SIRVs.

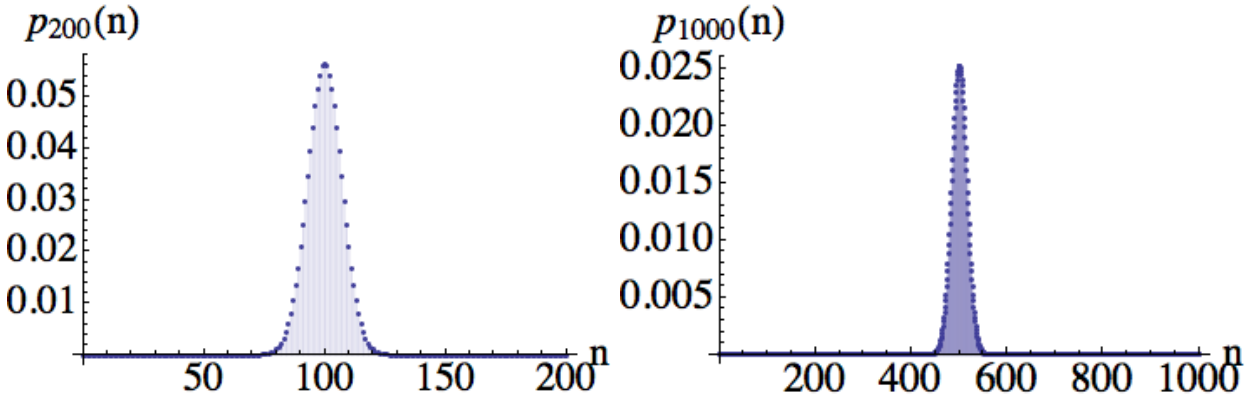
[End of Lecture 6.]

Flipping a coin (pset 1) A discrete example.

$N = 1000$ flips $p_N(n) \equiv \text{Prob}(n \text{ heads})$. $\langle n \rangle = 500 = N/2$.

$$\text{pset 1:} \qquad p_N(n) = \underbrace{\frac{1}{2^N}}_{\text{total \# of possible outcomes}} \times \underbrace{\frac{N!}{n!(N-n)!}}_{\text{\# of ways to get } n \text{ heads}}$$

Claim: This (binomial distribution) becomes a gaussian in the appropriate circumstances. More precisely, it becomes a comb of δ-functions with a Gaussian envelope.



Let $n = \frac{N}{2} + \epsilon$. $\epsilon \equiv$ deviation from mean. Since the distribution only has support when $\epsilon \ll N/2$, we will make this approximation everywhere below.

$$p\left(\frac{N}{2} + \epsilon\right) = \frac{1}{2^N} \frac{N!}{\left(\frac{N}{2} + \epsilon\right)! \left(\frac{N}{2} - \epsilon\right)!}$$

The log of a function is always more slowly-varying than the function itself; this means that it is better-approximated by its Taylor series; for this reason let's take the log of both sides. Use Stirling's approximation ($\ln N! \overset{N \gg 1}{\approx} N \ln N - N$, valid for $N \gg 1$):

$$\begin{aligned} \ln p\left(\frac{N}{2} + \epsilon\right) \overset{N \gg 1}{\approx} & N \ln 1/2 + N \ln N & & -N \\ & & -\left(\tfrac{N}{2} + \epsilon\right) \ln\left(\tfrac{N}{2} + \epsilon\right) & +N/2 + \epsilon \\ & & -\left(\tfrac{N}{2} - \epsilon\right) \ln\left(\tfrac{N}{2} - \epsilon\right) & +N/2 - \epsilon \end{aligned}$$

Now expand about the mean (which is also the maximum) of the distribution $\epsilon = 0$, using $\ln(N/2 \pm \epsilon) \approx \ln N/2 \pm 2\epsilon/N - (2\epsilon/N)^2/2 + \mathcal{O}\left(\frac{\epsilon}{N}\right)^3$[1]

$$\begin{aligned} \ln p\left(\frac{N}{2}+\epsilon\right) \overset{\epsilon \ll N/2, N \gg 1}{\approx} & \; N \ln N/2 \\ & -(N/2+\epsilon)\left(\ln N/2 + 2\epsilon/N - \frac{1}{2}(2\epsilon/N)^2\right) \\ & -(N/2-\epsilon)\left(\ln N/2 - 2\epsilon/N - \frac{1}{2}(2\epsilon/N)^2\right) \end{aligned} \tag{8}$$

Collect terms by powers of ϵ:

$$\begin{aligned} \ln p\left(\frac{N}{2}+\epsilon\right) \overset{\epsilon \ll N/2, N \gg 1}{\approx} & \; \epsilon^0 \left(N \ln N/2 - \frac{N}{2}\ln N/2 - \frac{N}{2}\ln N/2\right) \\ + & \; \epsilon^1 \left(-\ln N/2 + \ln N/2 + \frac{2\epsilon}{N}\frac{N}{2} - \frac{2\epsilon}{N}\frac{N}{2}\right) \\ + & \; \epsilon^2 \left(-\frac{4}{N} + \frac{1}{2}\left(\frac{2}{N}\right)^2 \frac{N}{2} + \frac{1}{2}\left(\frac{2}{N}\right)^2 \frac{N}{2}\right) \\ + & \; \epsilon^3 \left(\ldots \frac{1}{N^2}\right) \end{aligned}$$

Therefore:

$$\ln p\left(\frac{N}{2}+\epsilon\right) \approx -\frac{2\epsilon^2}{N}\left(1 + \mathcal{O}\left(\frac{\epsilon}{N}\right)\right).$$

Comments: First, the reason that the order-ϵ terms cancel is that we are expanding around the *maximum* of the distribution. The statement that it is the maximum means that the derivative vanishes there – that derivative is exactly the linear term in the Taylor expansion – and that the second derivative is negative there; this is an important sign. Second, the nontrivial statement of the Central Limit Theorem here is not just that we can Taylor expand the log of the distribution about the maximum. The nontrivial statement is that the *coefficients* of terms of higher order than ϵ^2 in that Taylor expansion *become small* as $N \to \infty$. It is crucial here that the terms we are neglecting go like $\frac{\epsilon^3}{N^2}$.

Therefore: the probability distribution is approximately Gaussian:

$$p\left(\frac{N}{2}+\epsilon\right) \approx \exp\left(-\frac{\epsilon^2}{N/2}\right) \quad .$$

These expressions are valid for ϵ small compared to the mean, but indeed the mean is

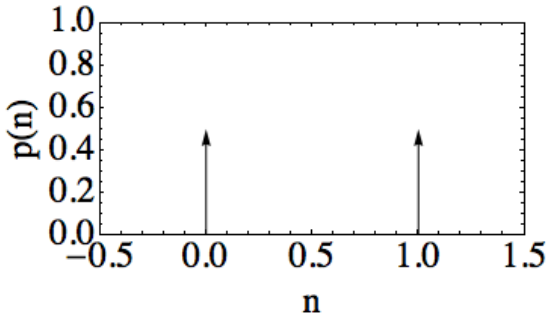
[1] (by $\mathcal{O}(x)^3$ I mean "terms of order x^3 which we are ignoring")

$N/2 \gg 1$.

$$p(n) \approx \exp\left(-\frac{(n-N/2)^2}{N/2}\right) = \underbrace{\mathcal{N}}_{\text{fixed by normalization}} \exp\left(-\frac{(n-\langle n\rangle)^2}{2\underbrace{\sigma_N^2}_{\text{Var}_N(n)}}\right)$$

It's a gaussian with mean $N/2$ and variance $\text{Var}_N(n) = \sigma_N^2 = N/4$.

This is consistent with the CLT quoted above. For one coin flip, $\langle n\rangle = 1/2, \sigma_1^2 = \text{Var}(n) = 1/4$.



The variance of the distribution of n after N flips is $\sigma_n^2 = N\sigma_1^2 = N/4$ (Variance adds for IID RVs).

On pset 4 you'll look at where the half-maximum occurs:

$$\frac{1}{2} = e^{-\frac{(n-N/2)^2}{N/2}}$$

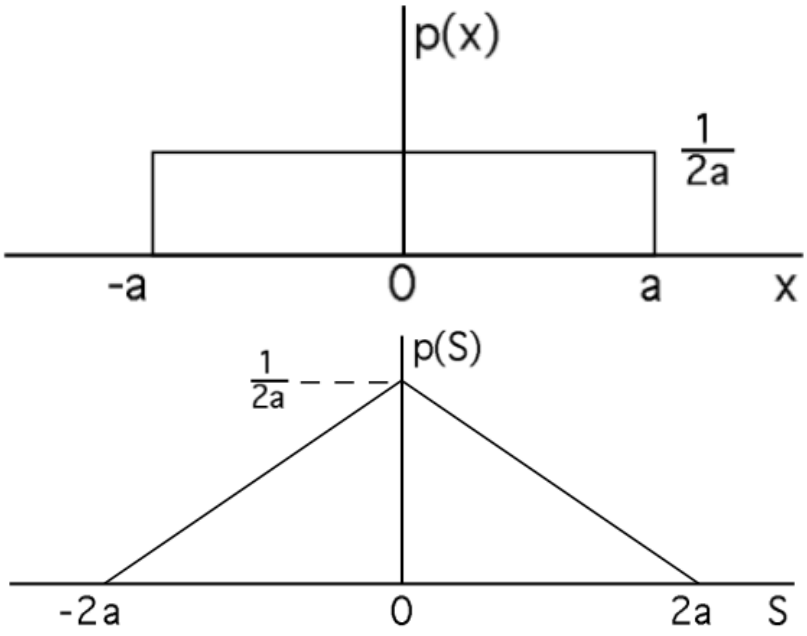
$$\ln 1/2 = -\frac{(n-N/2)^2}{N/2}$$

$$n - N/2 = \sqrt{N/2}\ln 2 = \sqrt{\log 2/2}\sqrt{N} = \sqrt{0.3466N}.$$

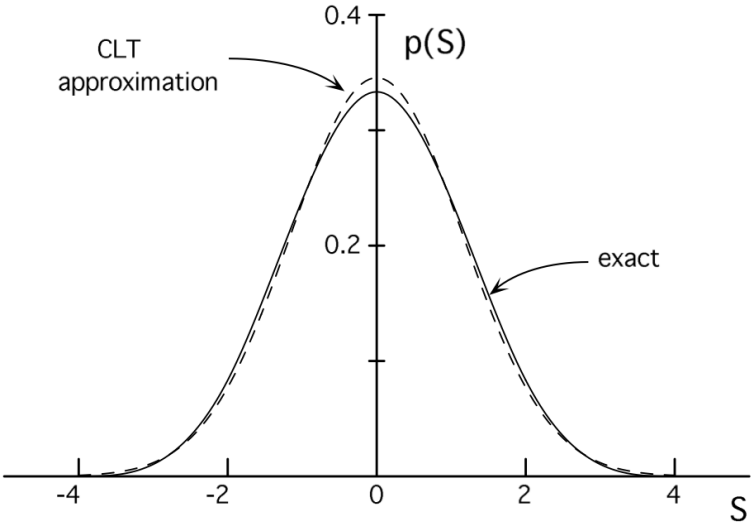
and compare with your approximate treatment on pset 1 which gave $\sqrt{0.333N}$.

How quickly does the CLT become true as we increase N?

Two pictorial examples from Tom Greytak's notes:



$N = 2 : S = x_1 + x_2$



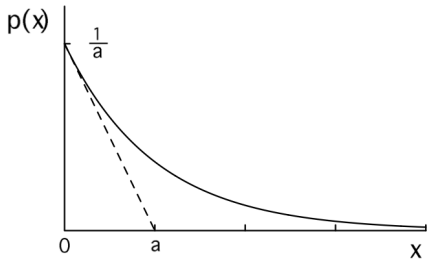
$N = 4 : S = x_1 + x_2 + x_3 + x_4$

This amazing convergence is not quite typical.

Some more convolutions, in Tom Greytak's notes. Consider the one-variable distribution:

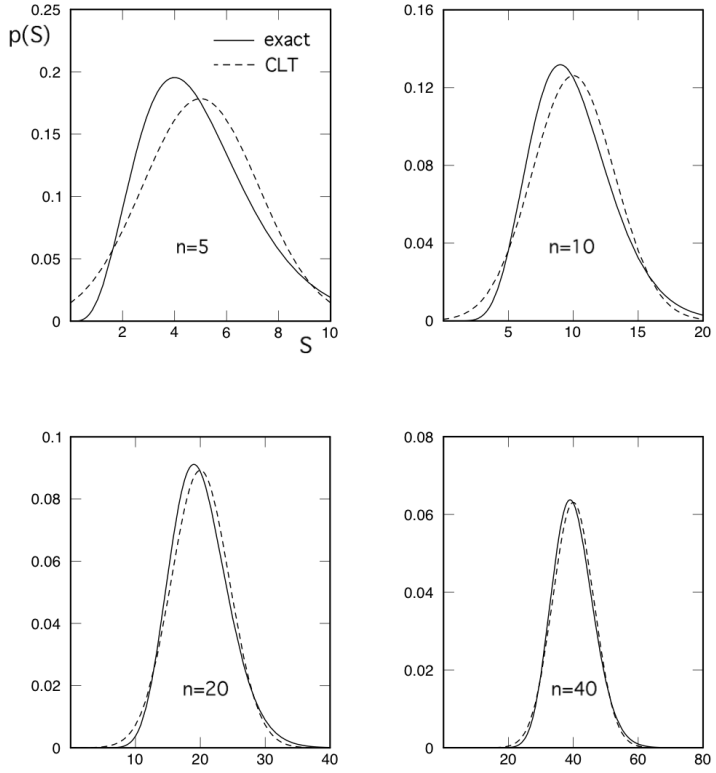
$$p(x) = \begin{cases} \frac{1}{a} e^{-x/a}, & \text{for } x \geq 0 \\ \\ 0, & \text{for } x < 0 \end{cases}$$

Let $S = \sum_{i=1}^{n} x_i$ where each x_i is governed by the density $p(x)$. For simplicity, set $a = 1$.

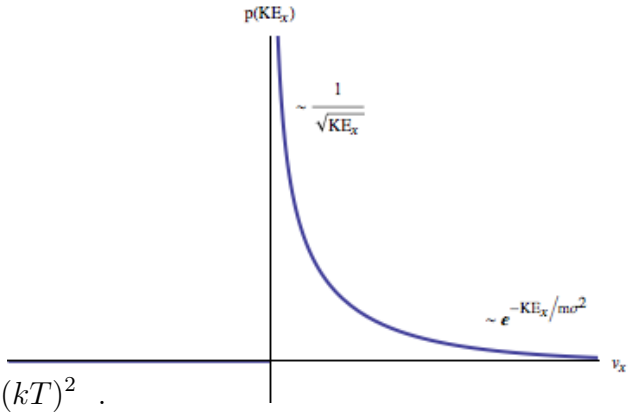


Claim [do convolutions]:

$$p_n(S) = \begin{cases} \frac{S^{n-1} e^{-S}}{(n-1)!}, & \text{for } S \geq 0 \\ \\ 0, & \text{for } S < 0 \end{cases}$$



An example from physics: energy of molecules in ideal gas



Previously we considered the distribution for 'kinetic energy of one atom due to motion in the x direction':

$$p(\mathrm{KE}_x) = \frac{1}{\sqrt{\pi m\sigma^2 \mathrm{KE}_x}} e^{-\frac{\mathrm{KE}_x}{m\sigma^2}} \qquad \sigma^2 = kT/m$$

We'll derive this in a few weeks. For now, it's our starting point. On pset 4 you'll play with this distribution and in particular will show that

$$\langle \mathrm{KE}_x \rangle = \frac{1}{2}kT \qquad \mathrm{Var}(\mathrm{KE}_x) = \frac{1}{2}(kT)^2 \quad .$$

Consider N such atoms in a (3-dimensional) room, each atom with this distribution for its $\mathrm{KE}_x, \mathrm{KE}_y, \mathrm{KE}_z$. We ignore interactions between the atoms.

Energy of N atoms in 3d: $\quad E = (\mathrm{KE}_x + \mathrm{KE}_y + \mathrm{KE}_z)_{\text{atom 1}} + (\mathrm{KE}_x + \mathrm{KE}_y + \mathrm{KE}_z)_{\text{atom 2}} + \ldots$

Assume: each of these variables is statistically independent from all the others. (That is, $(\mathrm{KE}_x)_{\text{atom 1}}$ and $(\mathrm{KE}_y)_{\text{atom 1}}$ are SI AND $(\mathrm{KE}_x)_{\text{atom 1}}$ and $(\mathrm{KE}_x)_{\text{atom 2}}$ are SI and so on....) Then we have for sums of RVs that the means add:

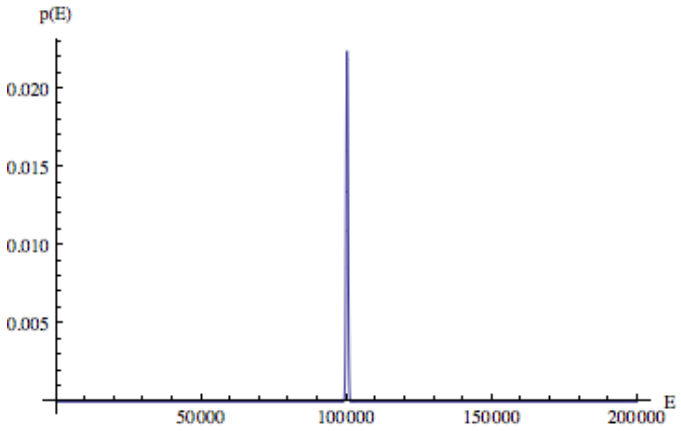
$$\langle E \rangle = \sum \langle \text{each KE} \rangle = 3N \cdot \frac{1}{2}kT = \frac{3}{2}NkT$$

And for sums of SI RVs that the variances add as well:

$$\mathrm{Var}(E) = \frac{3}{2}N(kT)^2.$$

Then the CLT tells us what the whole distribution for E is:

$$p(E) = \frac{1}{\sqrt{2\pi \frac{3}{2}N(kT)^2}} \exp\left[-\frac{(E - \frac{3}{2}NkT)^2}{2 \cdot \frac{3}{2}N(kT)^2}\right]$$



This is with $N = 10^5$ (in units where $\frac{3}{2}kT = 1$).

Very small fluctuations about the mean!

$$\frac{\text{width}}{\text{mean}} \sim \frac{\sqrt{\text{Var}}}{\text{mean}} = \frac{kT\sqrt{\frac{3}{2}N}}{\frac{3}{2}N} \sim \frac{1}{\sqrt{N}} \sim 10^{-25/2}.$$

Our discussion of the CLT here assumed that the vars we were summing were SI. The CLT actually still applies, as long as correlations are small enough.

Atoms in a real fluid are not SI. An occasion where correlations lead to important fluctuations is at a critical point. See 8.08 or 8.333.

2.5 Epilog: an alternate derivation of the Poisson distribution

Here is a discussion of an origin of the Poisson distribution complementary to the careful one in Prof. Greytak's notes.

2.5.1 Random walk in one dimension

A drunk person is trying to get home from a bar at $x = 0$, and makes a series of steps of length L down the (one-dimensional) street. Unfortunately, the direction of each step is random, and uncorrelated with the previous steps: with probability p he goes to the right and with probability $q = 1 - p$ he goes to the left. Let's ask: after N steps, what's his probability P(m) of being at $x = mL$?

Note that we've assumed all his steps are the same size, which has the effect of making space discrete. Let's restrict ourselves to the case where he moves in one dimension. This already has many physical applications, some of which we'll mention later.

What's the probability that he gets $|m| > N$ steps away? With N steps, the farthest away he can get is $|m| = N$, so for $|m| > N$, $P(m) = 0$.

Consider the probability of a particular, ordered, sequence of N steps, $x_i = L$ or R:

$$P(x_1, x_2 ... x_N) = P(x_1)P(x_2) \cdots P(x_N) = p^{n_R} q^{n_L}.$$

In the second step here we used the fact that the steps are statistically independent, so the joint probability factorizes. n_R is the number of steps to the right, i.e. the number of the x_i which equal R. Since the total number of steps is N, $n_L + n_R = N$, the net displacement (in units of the step length L) is

$$m = n_R - n_L = 2N_R - N.$$

Note that $m = N$ mod two.

In asking about the drunk's probability for reaching some location, we don't care about the order of the steps. There are many more ways to end up near the starting point than close by. For example, with $N = 3$, the possibilities are

$$LLL \qquad m = -3$$

$$RLL, LRL, LLR \quad m = -1$$

$$RRL, RLR, LRR \quad m = 1$$

$$RRR \qquad m = 3$$

What's the number of sequences for a given n_L, n_R? The sequence is determined if we say which of the steps is a R, so we have to choose n_R identical objects out of N. The number of ways to do this is

$$\binom{N}{n_R} = \frac{N!}{n_R! n_L!} = \binom{N}{n_L}.$$

A way to think about this formula for the number of ways to arrange $N = n_R + n_L$ of which n_R are indistinguishably one type and n_L are indistinguishably another type, is: $N!$ is the total number of orderings if all the objects can be distinguished. Redistributing the n_R R-steps amongst themselves doesn't change the pattern (there are $n_R!$ such orderings), so we must divide by this overcounting. Similarly redistributing the n_L L-steps amongst themselves doesn't change the pattern (there are $n_L!$ such orderings).

So

$$P(n_L, n_R) = \frac{N!}{n_R! n_L!} p^{n_R} q^{n_L}.$$

Note that the binomial formula is

$$(p+q)^N = \sum_{n=0}^{N} \frac{N!}{n!(N-n)!} p^n q^{N-n}.$$

Since we have $p + q = 1$, this tells us that our probability distribution is normalized:

$$\sum_{n_R=0}^{N} \frac{N!}{n_R!(N-n_R)!} p^{n_R} q^{N-n_R} = 1^N = 1.$$

The probability for net displacement m is

$$P(m) = \frac{N!}{\left(\frac{N+m}{2}\right)! \left(\frac{N-m}{2}\right)!} p^{\frac{N+m}{2}} q^{\frac{N-m}{2}}$$

for $N \pm m$ even, and zero otherwise.

2.5.2 From binomial to Poisson

We have shown that the probability that an event with probability p occurs n times in N (independent) trials is

$$W_N(n) = \frac{N!}{n!(N-n)!} p^n (1-p)^{N-n} \quad ;$$

this is called the binomial distribution. $1 - p$ here is the probability that *anything* else happens. So the analog of "step to the right" could be "a particular song is played on your ipod in shuffle mode" and the analog of "step to the left" is "any other song comes on".

For example, suppose you have 2000 songs on your ipod and you listen on shuffle by song; then the probability of hearing any one song is $p = \frac{1}{2000}$. Q: If you listen to $N = 1000$ songs on shuffle, what's the probability that you hear a particular song n times?

The binomial distribution applies. But there are some simplifications we can make. First, p itself is a small number, and N is large. Second, the probability will obviously be very small for $n \sim N$, so let's consider the limit $n \ll N$. In this case, we can apply Sterling's formula to the factorials:

$$W_N(n) \approx \frac{1}{n!} \frac{N^N}{(N-n)^{N-n}} p^n (1-p)^{N-n}$$

We can use $N - n \sim N$ except when there is a cancellation of order-N terms:

$$W_N(n) \approx \frac{1}{n!} \frac{N^N}{(N)^{N-n}} p^n (1-p)^{N-n} = \frac{1}{n!} N^n p^n (1-p)^{N-n}$$

Now we can taylor expand in small p, using $\ln(1-x) \approx -x + x^2/2 - x^3/3 + ...$

$$W_N(n) \approx \frac{1}{n!} (Np)^n e^{(N-n)\ln(1-p)} \approx \frac{1}{n!} (Np)^n e^{-Np}.$$

This is called the Poisson distribution,

$$\text{Poisson}_\mu(n) = \frac{1}{n!} \mu^n e^{-\mu}.$$

Note that it only depends on the product $\mu = pN$, which for our example is $pN = \frac{1}{2000} 1000 = 1/2$. In this case, it looks like in the figure 2.5.2.

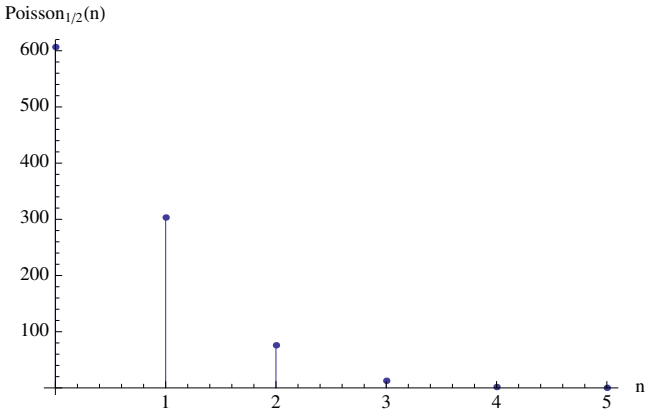


Figure 1: The Poisson distribution for $pN = 1/2$, $\text{Poisson}_{1/2}(n)$.

It may seem like your ipod is conspiring to play some songs multiple times and not play others at all (I had this impression too until I thought about it), but it's just because we don't have much intuition yet about the Poisson distribution. In fact, if we vary $\mu = Np$, we can make the probability that a given song is never heard much larger than the probability that it is heard once; see figure 2.5.2.

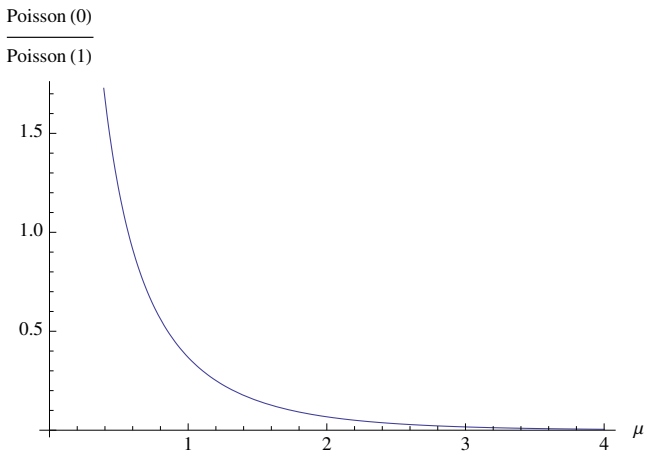


Figure 2: The ratio of the poisson distribution at $n = 0$ to $n = 1$ as we vary the parameter μ. (Note that this figure is not a probability distribution.)