Probability Theory

Motivation —— Measurements always have uncertainty.
↳ Probability is the mathematics of uncertain quantities

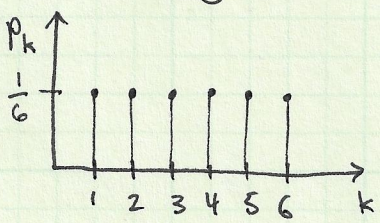
Two types of uncertainty/randomness : dice example

1. Lack of knowledge — if we were to measure more precisely, there would be no randomness.
2. True randomness — QM

↳ In practice, we will use the same mathematics for both — random variables.

Ex. : casting a die.

- Six possibilities. Assign a probability to each.
  ↳ "Probability mass function" ← discrete possibilities



Properties of p_k:

$0 \leq p_k \leq 1$,

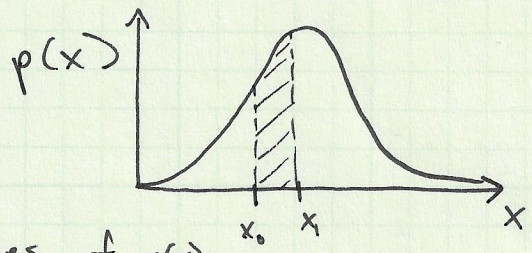
$\sum_k p_k = 1$.

- What does p_k mean?
  ↳ In this case, it means that if we repeat the exact same experiment infinitely many times, the fraction of times we'll get outcome k is p_k. (Frequentism)
  ↳ Later on, we'll see that probabilities are often assigned to things that are not experiments (e.g., competing theories, distances to stars). Here, probability has to do with plausibility, usually given some data. (Bayesianism)

  ⇒ This fundamental issue in probability theory is the fuzziest / least well defined aspect of the theory, strange as that may seem. In practice, however, it doesn't matter.

- Random variables
  - Instead of having a single value, they are described by distributions of possible values.
    ↳ In the above example, k is the random variable, and p_k is its distribution.

- Probability density functions
  - Some variables are <u>continuous</u>, not discrete.
    - ↪ position, velocity, height, ...
  - Described by a "<u>density</u>", not a "mass":



Probability of being btw/ x_0 and x_1:

$$\int_{x_0}^{x_1} p(x)\,dx.$$

  - Properties of $p(x)$:

$$p(x) \geq 0,$$

$$\int_{-\infty}^{\infty} p(x)\,dx = 1.$$

  - Very similar to a PMF, and much of what we'll learn about ~~PDF~~s & PMFs is interchangeable.
    - ↪ One difference: units.

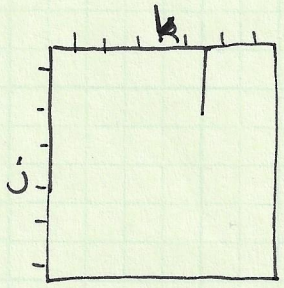
$p(x)$ has units of $\frac{1}{x}$.

p_k is unitless.

If you integrate a PDF, you get a PMF.
(over an interval)

- Joint distributions:
  - Multiple random variables.
  - Ex.: two dice.

p_{jk} ↑ ↑ (die #1, die #2)



$6 \times 6 = 36$ entries

$$\sum_{j,k} p_{jk} = 1.$$

For PDFs, write $p(x,y)$.

$$\int p(x,y)\,dx\,dy = 1.$$

- Conditional distributions:
  - If we know variable Y, what is the distribution of X?

$p(X \mid Y)$ "Probability of X given Y" / " " " conditional on Y"

  - The joint distribution can be calculated by

$$p(x,y) = p(y)\,p(x \mid y). \quad (*) \qquad = p(x)\,p(y \mid x).$$

  - Ex.: Native language given country of origin.
- Independence:
  - X is independent of Y if $p(X \mid Y) = p(x)$, i.e., knowing Y does not affect the distribution of X.
  - Independence is mutual: if X is indep. of Y, then Y is indep of X. (easy to show from (*))
  - For independent variables X and Y,

$$p(x,y) = p(x)\,p(y) \quad \text{(see (*))}.$$

↳ Ex: casting two die.

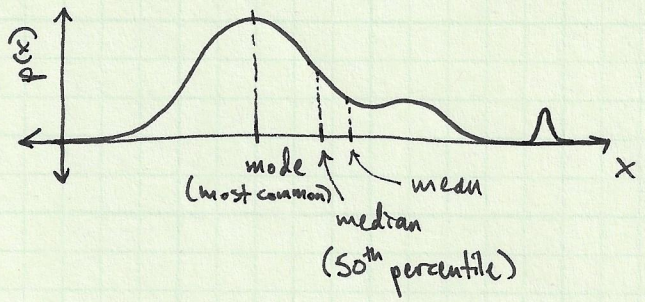
- Marginal distributions: if we know $p(x,y)$, we can find $p(x)$ alone by summing (or integrating) over all possible values of Y:

$$p(x) = \int p(x,y)\,dy. \quad \text{(for PDFs)}$$

$$p_x = \sum_y p_{x,y}. \quad \text{(for PMFs)}$$

Probability Theory

- How to characterize a distribution?



- mean: $\langle x \rangle = \int x\, p(x)\, dx$. ($\langle x \rangle$ also often called μ)

  Also called the "expectation value of x."

- Expectation values: we can actually take the expectation value of any function of x. This is the mean of that function over infinitely many draws of x from p(x).

$$\langle f(x) \rangle = \int f(x)\, p(x)\, dx.$$

- Variance: measures "width" of distribution.

$$\begin{aligned} \mathrm{Var}(x) &= \langle (x - \langle x \rangle)^2 \rangle \\ &= \int (x - \langle x \rangle)^2 p(x)\, dx. \\ &= \int (x^2 - 2x\langle x \rangle + \langle x \rangle^2)\, p(x)\, dx \\ &= \langle x^2 \rangle - 2\langle x \rangle\langle x \rangle + \langle x \rangle^2 \\ &= \langle x^2 \rangle - \langle x \rangle^2. \end{aligned}$$

(Var(x) also often called σ_x^2)

↳ Standard deviation = $\sqrt{\mathrm{Var}(x)}$, often called σ_x.

- Moments: $\langle x^n \rangle$ is called the n^{th} "moment" of the distribution p(x). The mean is the "first moment," while the variance is related to the 2nd moment.

Bernoulli trial

- Simplest probability distribution: "yes" or "no" questions.

$$\left.\begin{aligned} p(\text{"yes"}) &\equiv p \\ p(\text{"no"}) &\equiv q \end{aligned}\right\} \quad p+q=1.$$

- Examples:
  - flipping a coin ($p=q=\frac{1}{2}$).
  - boy or girl? ($p=0.517$, $q=0.483$)
  - does a team win a game? ($p=?$)

- Expectation value (mean):

Call "no" $k=0$, and "yes" $k=1$.

$$p(k=0)=q, \quad p(k=1)=p.$$

$$\langle k\rangle = \sum_{k=0}^{1} k p(k) = 0\cdot q + 1\cdot p = p.$$ Pretty simple.

- Variance:

$$\langle k^2\rangle = \sum_{k=0}^{1} k^2 p_k = 0^2\cdot q + 1^2\cdot p = p.$$

$$\Rightarrow \mathrm{Var}(k) = \langle k^2\rangle - \langle k\rangle^2 = p - p^2 = p(1-p) = pq.$$

Binomial distribution

- If we conduct n indep. Bernoulli trials, what is the probability that k of them are successes ("yes")?
  ↳ Depends on p, n and k.

- Simple example: $n=3$, $k=2$.

  ⊗⊗○
  ⊗○⊗
  ○⊗⊗
  } 3 ways to select 2 items.

  Each way has probability $p^2 q = p^2(1-p)$.

  ~~[illegible]~~

  $\Rightarrow p(k=2 \mid n=3) = 3\,p^2(1-p)$.

- In general,

$$p(k \mid n) = \underbrace{\binom{n}{k}}_{\text{"Binomial coefficient"}} p^{k} (1-p)^{n-k}$$

(# of successes: k; # of failures: $n-k$)

("Binomial coefficient" = # of ways of selecting k out of n items)

- How to calculate $\binom{n}{k}$?

  First, imagine selecting k items in order:

  First item: ○○⊗○○○○ (1) — n choices
  2nd item: ○○⊗○○○⊗ (2) — $n-1$ choices
  ⋮ ⋮
  k^{th} item: ○○⊗○⊗⊗⊗ (k) — $n-k+1$ choices

  $\Rightarrow n(n-1)\cdots(n-k+1) = \dfrac{n!}{(n-k)!}$ ways to select k items in order.

  But the order doesn't matter! All that matters is the total number of successes, not the order in which they are selected.

  ○○⊗(1)○⊗(3)⊗(4)⊗(2) = ○○⊗(4)○⊗(1)⊗(3)⊗(2) = ... $\Rightarrow$ $k!$ orderings of k selected items.

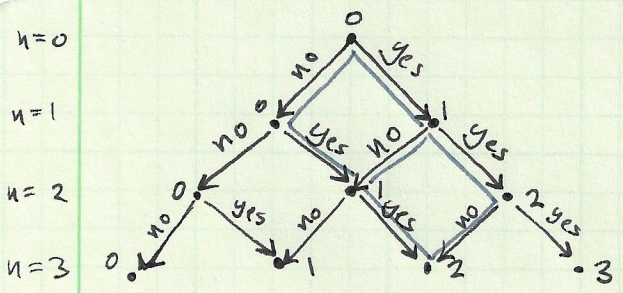
  Divide # of ways of selecting k items in order by the number of orderings of k items.

$$\Rightarrow \boxed{\binom{n}{k} = \frac{n!}{(n-k)!\,k!}}.$$

Note the symmetry $k \leftrightarrow n-k$ (successes vs. failures, "yes" vs. "no").

Binomial distribution, cont'd

There's a nice graphical representation of a binomial process, as a tree:

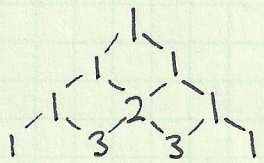


three routes to arrive at $n=3,\ k=2$.

We can calculate the # of routes recursively, from the leaf node:

# of ways to reach node = Sum of # of ways to reach parent nodes.

Using this principle, we can propagate downwards from the top node ($n=0, k=0$):



"Pascal's triangle".

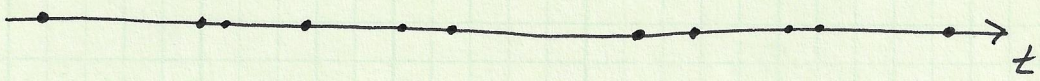
- Expectation values:
  - mean: $\langle k \rangle$. Since a Binomial process is just a sum of n independent Bernoulli trials, its mean is the sum of the means of n Bernoulli trials.

    $\Rightarrow \boxed{\langle k \rangle = np.}$ This also makes intuitive sense.
  - variance: $\mathrm{Var}(k)$ = Sum of variances of n Bernoulli trials.

    $\Rightarrow \boxed{\mathrm{Var}(k) = npq = np(1-p).}$

Poisson distribution

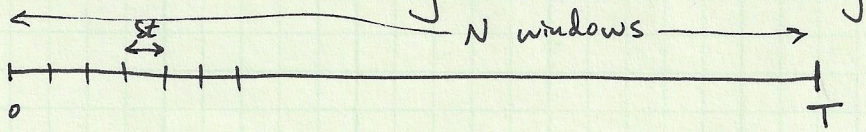
- Events that occur at a <u>constant rate</u>, and which are <u>independent of one another</u>.



Probability of event occurring in some time interval does not depend on when last event occurred.

- Examples:
  - Photons hitting a camera.
  - Radioactive decay.
  - Visitors to a website
  - People arriving at a train station
    - ↳ What if they arrive in bunches?
      - ⇒ Not independent – not a Poisson process!
      - (could group people → Poisson process again)

- Derivation: how many events in a given time interval?



Divide interval up into very short windows (δt).

↳ As $\delta t \to 0$, we either get 0 or 1 events.
↳ Probability of getting events in different windows indep.
↳ Rate: probability should be proportional to δt.

⇒ $p(\text{event in window } i) = \frac{\lambda}{N} \propto \delta t$. ← "rate parameter"

$\lambda / N \equiv p_w$

How many events total? Count up # of windows w/ event.

↳ Each window is an independent trial, w/ prob. p_w of success.

⇒ Binomial distribution

Probability of k events? ⇒ Probability of k successes.

$$p(k \text{ events}) = \lim_{N \to \infty} \binom{N}{k} p_w^k (1 - p_w)^{N-k}.$$

$$p(k \text{ events}) = \lim_{N\to\infty} \frac{N!}{k!\,(N-k)!} \left(\frac{\lambda}{N}\right)^k \left(1-\frac{\lambda}{N}\right)^{N-k}$$

$$= \frac{\lambda^k}{k!} \lim_{N\to\infty} \underbrace{\frac{N!}{(N-k)!}\frac{1}{N^k}} \left(1-\frac{\lambda}{N}\right)^{N-k}$$

$$= \frac{N(N-1)\cdots(N-k+1)}{N^k} = \frac{N}{N}\cdot\frac{N-1}{N}\cdots\frac{N-k+1}{N} \to 1$$

$$= \frac{\lambda^k}{k!} \lim_{N\to\infty} \underbrace{\left(1-\frac{\lambda}{N}\right)^N}_{\to e^{-\lambda}} \underbrace{\left(1-\frac{\lambda}{N}\right)^{-k}}_{\to 1}$$

$$= \boxed{\frac{\lambda^k e^{-\lambda}}{k!}}.$$

- What does λ mean?

Find the expected (mean) number of events:

$$\langle k\rangle = \sum_{k=0}^{\infty} k\,p_k = \sum_{k=0}^{\infty} \frac{k\lambda^k e^{-\lambda}}{k!} = e^{-\lambda}\sum_{k=0}^{\infty}\frac{k\lambda^k}{k!} \qquad (*)$$

$$= e^{-\lambda}\sum_{k=1}^{\infty}\frac{\lambda^k}{(k-1)!} = \lambda e^{-\lambda}\sum_{k=1}^{\infty}\frac{\lambda^{k-1}}{(k-1)!} = \lambda e^{-\lambda}\underbrace{\sum_{m=0}^{\infty}\frac{\lambda^m}{m!}}_{=e^{\lambda}\ \text{(Taylor's.)}}$$

(change of lower limit)

$$= \lambda. \quad \Rightarrow \boxed{\lambda \text{ is the mean \# of events.}}$$

What about the variance of k?

$$\mathrm{Var}(k) = \langle k^2\rangle - \langle k\rangle^2.$$

$$\langle k^2\rangle = \sum_{k=0}^{\infty} k^2\frac{\lambda^k e^{-\lambda}}{k!} = e^{-\lambda}\sum_{k=1}^{\infty} k^2\frac{\lambda^k}{k!}$$

(change limit)

$$= \lambda e^{-\lambda}\sum_{k=1}^{\infty} k\frac{\lambda^{k-1}}{(k-1)!} \leftarrow (\text{compare to } *)$$

$$= \lambda e^{-\lambda}\sum_{k=1}^{\infty}\left[\underbrace{(k-1)\frac{\lambda^{k-1}}{(k-1)!}}_{\text{same as } (*)} + \underbrace{\frac{\lambda^{k-1}}{(k-1)!}}_{\to e^{\lambda}}\right] = \lambda e^{-\lambda}\left(\lambda e^{\lambda} + e^{\lambda}\right)$$

$$= \lambda(\lambda+1).$$

$$\Rightarrow \mathrm{Var}(k) = \lambda(\lambda+1) - \lambda^2 = \boxed{\lambda}.$$

Poisson distribution, cont'd

$\Rightarrow$ Standard deviation $= \sqrt{\lambda}$.

Consequence: if we observe a Poisson process for 4x as long, the fractional uncertainty in the rate drops by a factor of 2.

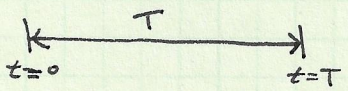
$$\frac{\sigma}{\mu} = \frac{1}{\sqrt{\lambda}}.$$

↳ Example: observe a star for 4x as long, determine flux to 2x the precision.
(photons/second)

Poisson inter-arrival times

• How long between events in a Poisson process?

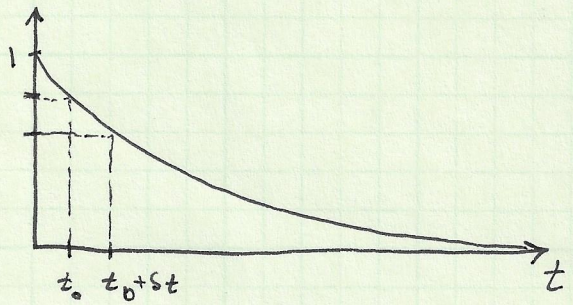
Begin at some time, which we'll call $t=0$. Then look for time T:



we expect λT events.

$$p(k \text{ events}) = \frac{(\lambda T)^k e^{-\lambda T}}{k!}.$$

$$\Rightarrow p(0 \text{ events}) = e^{-\lambda T}.$$



(note that this is essentially a decay process)

What fraction of the time does the next event arrive between times t_0 and $t_0+\delta t$?

$$e^{-\lambda t_0} - e^{-\lambda(t_0+\delta t)}$$

Divide by δt to get the probability density of the next arrival time:

$$\lim_{\delta t \to 0} \frac{e^{-\lambda t_0} - e^{-\lambda(t+\delta t)}}{\delta t}.$$

$$\Rightarrow p(t) = \frac{d}{dt}\left(-e^{-\lambda t}\right) = \boxed{\lambda e^{-\lambda t}}. \quad \text{("exponential distribution")}$$

If you start looking at a Poisson process, this is the probability density of the time of the next event. It doesn't matter when you start looking, or when the last event was ("memorylessness").

This is a continuous distribution (PDF) (function of a real number, not an integer). It is related to the Poisson distribution, which is a discrete distribution (PMF).

How long btw/ events?

$$\langle t \rangle = \int_0^\infty t\, p(t)\, dt = \int_0^\infty \lambda t\, e^{-\lambda t}\, dt \qquad \text{(integrate by parts)}$$

$$= \frac{1}{\lambda}, \quad \text{as expected.}$$

Moment-generating functions

Not something you necessarily have to know for this course, but very cool, so I'll throw it in anyways.

A general method of determining the moments of a distribution.

↳ Leads into interesting/powerful area of probability theory (characteristic functions).

Given a probability mass (or density) function $p(x)$, define

$$M_x(t) \equiv \langle e^{xt} \rangle. \qquad (\text{"Moment-generating function"})$$

This is a transformation of $p(x)$, sort of like a Fourier transform. The derivatives of $M_x(t)$ have an interesting property:

$$\frac{dM_x}{dt} = \frac{d}{dt}\langle e^{xt}\rangle = \frac{d}{dt}\sum_x e^{xt}p(x) = \sum_x x e^{xt} p(x)$$

$$= \langle x e^{xt} \rangle.$$

$$\Rightarrow \left.\frac{dM_x}{dt}\right|_{t=0} = \langle x \rangle. \quad \leftarrow \text{Derivatives of } M_x \text{ linked to moments of } p(x).$$

In general,

$$\left.\frac{d^n M_x}{dt^n}\right|_{t=0} = \langle x^n \rangle.$$

Ex.: Bernoulli trial

$$p_k = \begin{cases} p, & k=1 \\ 1-p, & k=0 \end{cases}. \qquad M_k(t) = \langle e^{kt} \rangle = \sum_{k=0}^{1} p_k e^{kt}$$

$$= 1 - p + pe^t.$$

Derivatives: $M_k'(t) = pe^t. \Rightarrow \langle k \rangle = M_k'(0) = p.$

$\vdots$

$M_k^{(n)}(t) = pe^t.$ $\quad \mathrm{Var}(k) = \langle k^2 \rangle - \langle k \rangle^2 = p - p^2 = p(1-p).$

$\Rightarrow$ All the moments are p.

Moment-generating f'ns, cont'd

Ex.: Binomial distribution

$$M_k(t) = \sum_{k=0}^{\infty} \binom{n}{k} p^k (1-p)^{n-k} e^{kt}$$

$$= \sum_{k=0}^{\infty} \binom{n}{k} (pe^t)^k (1-p)^{n-k}$$

$$= (1 - p + pe^t)^n. \qquad \text{(by the Binomial theorem)}$$

Derivatives:

$$M_k'(t) = n(1-p+pe^t)^{n-1} pe^t. \Rightarrow \langle k \rangle = M_k'(0) = np.$$

$$M_k''(t) = npe^t(1-p+pe^t)^{n-1} + n(n-1)(pe^t)^2(1-p+pe^t)^{n-2}.$$

$$\Rightarrow \langle k^2 \rangle = M_k''(0) = n(n-1)p^2 + np.$$

$$\Rightarrow \mathrm{Var}(k) = \langle k^2 \rangle - \langle k \rangle^2 = n(n-1)p^2 + np - (np)^2$$

$$= np - np^2 = np(1-p).$$

We could calculate higher moments too.

Notice that the moment-generating f'n (MGF) of the Binomial distribution is the MGF of the Bernoulli distribution raised to the n^{th} power. This is not a coincidence!

The Binomial distribution results from n Bernoulli trials.

$k_1, k_2, \ldots, k_n$ ← outcomes of independent Bernoulli trials.

$k = k_1 + k_2 + \cdots + k_n$. ← # of successes.

In general, the MGF of the sum of two <u>independent</u> variables, x and y, is given by

$$M_{x+y}(t) = \langle e^{(x+y)t} \rangle = \langle e^{xt} e^{yt} \rangle = \langle e^{xt} \rangle \langle e^{yt} \rangle \qquad \text{(b/c } x \text{ \& } y \text{ are independent)}$$

$$= M_x(t) M_y(t).$$

Thus, the Binomial distribution has the MGF

$$M_k(t) = M_{k_1+k_2+\cdots+k_n}(t) = \underbrace{M_{k_1}(t)}_{\text{all}} \underbrace{M_{k_2}(t)}_{\text{identical}} \cdots M_{k_n}(t) = (1-p+pe^t)^n.$$

Moment-generating f'ns, cont'd

Ex. Poisson MGF

$$p_k = \frac{\lambda^k e^{-\lambda}}{k!}.$$

$$M_k(t) = \langle e^{kt} \rangle = \sum_{k=0}^{\infty} \frac{\lambda^k e^{-\lambda}}{k!} e^{kt} = e^{-\lambda} \underbrace{\sum_{k=0}^{\infty} \frac{(\lambda e^t)^k}{k!}}_{= e^{\lambda e^t}}$$

$$= e^{-\lambda} e^{\lambda e^t}$$

$$= e^{\lambda(e^t - 1)}.$$

$$M_k'(t) = \lambda e^t \, e^{\lambda(e^t-1)} = \lambda e^{\lambda(e^t-1)+t}. \Rightarrow \langle k \rangle = \lambda.$$

$$M_k''(t) = \lambda(\lambda e^t + 1) e^{\lambda(e^t-1)+t}. \Rightarrow \langle k^2 \rangle = \lambda^2 + \lambda = \lambda(\lambda+1).$$

$$\Rightarrow \mathrm{Var}(k) = \langle k^2 \rangle - \langle k \rangle^2 = \lambda(\lambda+1) - \lambda^2 = \lambda.$$

We could also have calculated the MGF by going all the way back to the definition of the Poisson distribution as the sum of N Bernoulli trials, each w/ probability λ/N, as $N \to \infty$. We don't even have to know the Poisson probability mass f'n.

$$M_k(t) = \lim_{N\to\infty} M_{k_1}(t) M_{k_2}(t) \cdots M_{k_N}(t) \quad \leftarrow \text{Bernoulli trials}$$

$$= \lim_{N\to\infty} \left(1 - \frac{\lambda}{N} + \frac{\lambda}{N} e^t\right)^N$$

$$= \lim_{N\to\infty} \left[1 + \frac{1}{N}\lambda(e^t - 1)\right]^N$$

$$= e^{\lambda(e^t-1)}.$$

Moment-generating functions make a lot of proofs easy. For example, how is the sum of N independent Poisson-distributed variables distributed?

X_i has rate λ_i, $i = 1, \ldots, N$.

$$Z = X_1 + X_2 + \ldots + X_N.$$

$$M_Z(t) = M_{X_1}(t) M_{X_2}(t) \cdots M_{X_N}(t)$$

$$= e^{\lambda_1(e^t-1)} e^{\lambda_2(e^t-1)} \cdots e^{\lambda_N(e^t-1)}$$

$$= e^{(\lambda_1+\lambda_2+\ldots+\lambda_N)(e^t-1)},$$

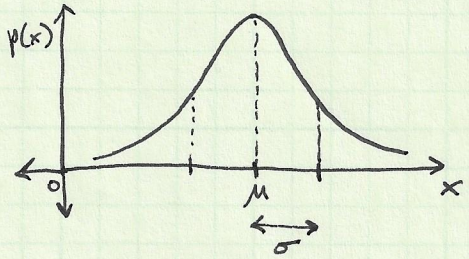
$\Rightarrow$ Z is a Poisson process with rate $\lambda_1 + \lambda_2 + \ldots + \lambda_N$.

Gaussian distribution

- Probably the most important distribution. → Pops up everywhere.
  ↳ So important, its also called the "normal distribution".

$$p(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$

μ = mean, σ = standard deviation.



$\uparrow$

$\langle (x-\mu)^2 \rangle = \sigma^2$.

- Many quantities in nature are "normally" distributed, or nearly so:
  - Human height (given sex)
  - Temperature at a given location on given day of year
  - Flipping a coin a large number of times: number of tails (or heads).

  ↳ Important for physics/astronomy: measurement errors are often normally distributed.

Central Limit Theorem

- Roughly: if you add up a large number of random variables, the result looks Gaussian.

  ↳ Ex.: Adding up a large number of sources of error in a measurement. → Errors look Gaussian.

- More formally:

  If $x_1, x_2, \ldots, x_n$ are independent random variables with means $\mu_1, \ldots, \mu_n$ and variances $\sigma_1^2, \ldots, \sigma_n^2$, then

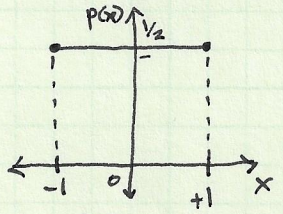
$$z \equiv \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \frac{x_i - \mu_i}{\sigma_i}$$

  approaches $\underbrace{\mathcal{N}(0,1)}$ as $n \to \infty$.

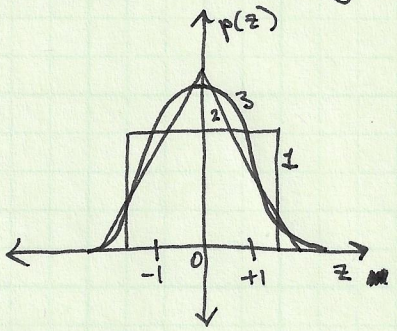
  normal distribution w/ mean 0 and variance 1 ("unit normal distribution")

- Ex.: Adding up uniform distributions.

  $x_i \sim U(-1, 1)$. (x_i is distributed uniformly btw/ -1 and +1).



As we add up more variables:



- Not difficult to prove using moment-generating functions.

  - Calculate MGF of x as Taylor series, and then show that MGF of z approaches

$$\left(1 - \frac{t^2}{2n}\right)^n \quad \text{as } n \to \infty.$$

  This goes to $e^{-\frac{1}{2}t^2}$, the MGF of $\mathcal{N}(0,1)$.

Inference, Introduction

- In science, we often have theories (or models) that predict what data we should observe.

$$\text{theory} \longmapsto \text{data}$$

This prediction is often (or almost always) probabilistic.

- Ex.: I have a coin. My theory is that it is a fair coin ($p = q = \frac{1}{2}$). I flip it 1000 times, how many heads should I observe?
  - ↳ Expectation: 500, but there's actually a PMF of the theoretical outcome.
  - ↳ Binomial distribution: $p(k \mid n, p = \frac{1}{2})$.

- However, we generally want to use data to see if our theories are correct, or to constrain parameters in our theories.

$$\text{data} \longmapsto \text{theory}$$

- Very roughly, we can write

$p(\text{data} \mid \text{theory})$ ≡ "likelihood", normally calculable.

$p(\text{theory} \mid \text{data})$ ≡ "posterior", what we generally want to know.

  - ↳ How to relate the likelihood and posterior?

- Ex.: If we observe 551 heads out of 1000 flips, can we determine the posterior distribution of p?

→ Bayes' Theorem

Bayes' Theorem

Consider the joint probability of two variables, A and B:

$$p(A, B).$$

We learned earlier that

$$\left.\begin{aligned} p(A,B) &= p(A|B)\,p(B) \\ p(A,B) &= p(B|A)\,p(A) \end{aligned}\right\} \text{ symmetry in } A, B$$

Equating these two expansions of $p(A,B)$ and rearranging,

$$p(B|A) = \frac{p(A|B)\,p(B)}{p(A)}. \qquad \text{"Bayes' Theorem"}$$

Inference, Part II

- Apply Bayes' Theorem to inference.

We have $p(\text{data} \mid \text{theory})$.

We want $p(\text{theory} \mid \text{data})$.

$$\Rightarrow \underbrace{p(\text{theory} \mid \text{data})}_{\text{"posterior"}} = \frac{\overbrace{p(\text{data} \mid \text{theory})}^{\text{"likelihood"}} \overbrace{p(\text{theory})}^{\text{"prior"}}}{\underbrace{p(\text{data})}_{\text{"evidence"}}}.$$

- This might all seem too theoretical, so let's look at a concrete example: coin-flipping, again.

theory = probability p of heads.

data = number of heads observed, k.

$p(\text{theory} \mid \text{data}) = p(p \mid k, n)$

$p(\text{data} \mid \text{theory}) = p(k \mid p, n)$ (n = # of flips) = Binomial distribution.

$p(\text{theory}) = p(p) = ?$ Our prior expectation. Let's be completely agnostic, and assign constant probability density between 0 and 1.

$p(\text{data}) = p(k \mid n)$. Probability of getting k heads, averaged over all our possible values of p.

$$p(p \mid k, n) = \frac{p(k \mid p, n)\, p(p)}{\int_0^1 p(k \mid p', n)\, p(p')\, dp'}$$

Inference, Part II, cont'd

Coin-flipping posterior probability:

$$p(p|k,n) = \frac{p(k|p,n)\overbrace{p(p)}^{=1 \text{ (because const. from 0 to 1, and } \int p(p)dp = 1)}}{\int_0^1 p(k|p',n)p(p')dp'}$$

$$= \frac{\binom{n}{k} p^k (1-p)^{n-k}}{\int_0^1 \binom{n}{k} p'^k (1-p')^{n-k} dp'}$$

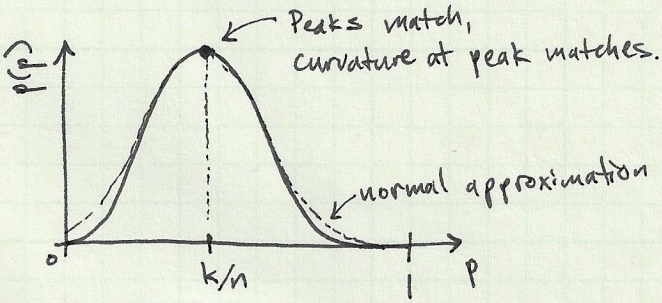
} doesn't depend on p, just a normalizing constant.

$$= \frac{(n+k+1)!}{n!\,k!} p^k (1-p)^{n-k}.$$

We didn't even have to do the integral in the denominator in order to see shape of the posterior distribution. That integral, the "evidence," just provides a normalizing constant.

- As n and k become large, incredibly, this distribution begins to look Gaussian.

↳ Gaussian approximation of the posterior.



- Find a Gaussian w/ the same peak, and which has the same curvature at the peak.
- Log(prob) of Gaussian: $\ln p(x) = -\frac{1}{2}\ln(2\pi\sigma^2) - \frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2$.

$$\frac{d}{dx}\ln p(x) = -\left(\frac{x-\mu}{\sigma^2}\right). \Rightarrow \text{Peak at } x = \mu.$$

$$\frac{d^2}{dx^2}\ln p(x) = -\frac{1}{\sigma^2}. \leftarrow \text{Curvature}$$

$\Rightarrow$ Gaussian approximation: μ = peak,

$$\sigma^2 = -\left[\frac{d^2}{dx^2}\ln p(x)\right]^{-1}_{x=\mu}.$$

Inference, Part II. Coin-flipping, cont'd.

Gaussian approx. of coin-flipping posterior:

$$\ln p(p \mid k,n) = k \ln p + (n-k) \ln(1-p) + \text{const.}$$

$$\frac{d}{dp} \ln p(p \mid k,n) = \frac{k}{p} - \frac{n-k}{1-p} = \frac{k(1-p) - p(n-k)}{p(1-p)}$$

$$= \frac{k - np}{p(1-p)} . \quad \Rightarrow 0 \text{ when } p = \frac{k}{n},$$

as expected.

$$\frac{d^2}{dp^2} \ln p(p \mid k,n) = -\frac{k}{p^2} + \frac{n-k}{(1-p)^2} .$$

$\Rightarrow$ Plug in $p = \frac{k}{n}$ to find curvature at peak.

$$\left.\frac{d^2}{dp^2} \ln p(p \mid k,n)\right|_{p=\frac{k}{n}} = -\left[\frac{k}{(k/n)^2} + \frac{n-k}{(1-k/n)^2}\right] = -n\left[\frac{(k/n)}{(k/n)^2} - \frac{1-(k/n)}{(1-k/n)^2}\right]$$

$$= -\left[\frac{1}{(k/n)} + \frac{1}{1-(k/n)}\right] = -n\left[\frac{1-\cancel{(k/n)} + \cancel{(k/n)}}{(k/n)(1-k/n)}\right]$$

$$= -\frac{n}{\frac{k}{n}\left(1-\frac{k}{n}\right)} .$$

$$\Rightarrow \mu = \frac{k}{n}, \quad \sigma^2 = \frac{\frac{k}{n}\left(1-\frac{k}{n}\right)}{n} = \frac{\mu(1-\mu)}{n} .$$

Becomes narrower as we observe more coin flips.

4x more flips. $\Rightarrow$ 2x smaller uncertainty on p.

Inference, part II, cont'd

Infer the rate of a Poisson process. (eg, how bright is a star, given an observed # of photons?)

- Observe k counts. $\rightarrow$ Infer λ.
- Bayes:

$$p(\lambda \mid k) = \frac{p(k \mid \lambda)\, p(\lambda)}{p(k)}.$$

- Likelihood (Poisson distribution):

$$p(k \mid \lambda) = \frac{\lambda^k e^{-\lambda}}{k!}.$$

- Prior: Different possible choices. We could say there's constant probability from 0 to some max. rate (uniform prior).

Instead, assume the probability of lower rates is higher (e.g., more faint than bright stars):

$$p(\lambda) \propto \begin{cases} \frac{1}{\lambda}, & \lambda_0 < \lambda < \lambda_1 \\ 0, & \text{otherwise} \end{cases}.$$

Normalize: $\int_{\lambda_0}^{\lambda_1} p(\lambda)\, d\lambda = \int_{\lambda_0}^{\lambda_1} \frac{d\lambda}{\lambda} = \ln\left(\frac{\lambda_1}{\lambda_0}\right)$.

$$\Rightarrow p(\lambda) = \begin{cases} \frac{1}{\ln(\lambda_1/\lambda_0)} \frac{1}{\lambda}, & \lambda_0 < \lambda < \lambda_1 \\ 0, & \text{otherwise} \end{cases}.$$

- Evidence: We don't need to calculate it to see shape of posterior. It's just a normalizing constant.

$$p(k) = \int p(k \mid \lambda)\, p(\lambda)\, d\lambda$$

$$= \int_{\lambda_0}^{\lambda_1} \frac{\lambda^k e^{-\lambda}}{k!} \frac{1}{\ln(\lambda_1/\lambda_2)} \frac{1}{\lambda} d\lambda$$

$$= \frac{1}{k! \ln(\lambda_1/\lambda_0)} \underbrace{\int_{\lambda_0}^{\lambda_1} \lambda^{k-1} e^{-\lambda} d\lambda}$$

"incomplete gamma function": $\int_0^a x^{n-1} e^{-x} dx = \gamma(n, a)$

$\gamma(n, \infty) = \Gamma(n)$.

$$= \frac{\gamma(k, \lambda_1) - \gamma(k, \lambda_0)}{k! \ln(\lambda_1/\lambda_0)}.$$

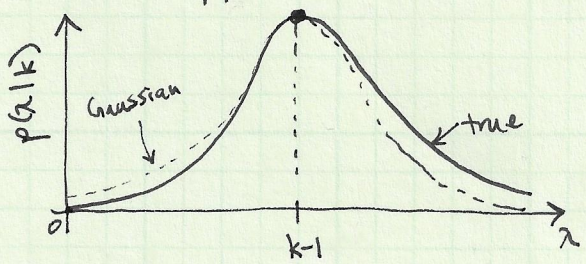
Inference, part II, cont'd

- Posterior: Putting everything together,

$$p(\lambda \mid k) = \begin{cases} \dfrac{\lambda^{k-1} e^{-\lambda}}{\Gamma(k,\lambda_1) - \Gamma(k,\lambda_0)}, & \lambda_0 < \lambda < \lambda_1 \\ 0, & \text{otherwise} \end{cases}$$

↳ Looks like the Poisson distribution, but note that this is a <u>density in λ</u> not a <u>mass in k</u>.

- Gaussian approximation:



Find the peak and curvature of $\ln p(\lambda \mid k) = (k-1)\ln\lambda - \lambda + \text{const.}$:

$$\left.\frac{d}{d\lambda} \ln p(\lambda \mid k)\right|_{\lambda=\mu} = \left.\left(\frac{k-1}{\lambda} - 1\right)\right|_{\lambda=\mu} = \frac{k-1}{\mu} - 1 = 0. \Rightarrow \mu = k-1.$$

$$\left.\frac{d^2}{d\lambda^2} \ln p(\lambda \mid k)\right|_{\lambda=\mu} = \left.-\frac{k-1}{\lambda^2}\right|_{\lambda=\mu} = -\frac{k-1}{\mu^2} = -\frac{1}{\sigma^2}.$$

$$\Rightarrow \sigma^2 = k-1 = \mu.$$

The peak probability density is at $k-1$, not k, because of our prior that lower rates are more probable.

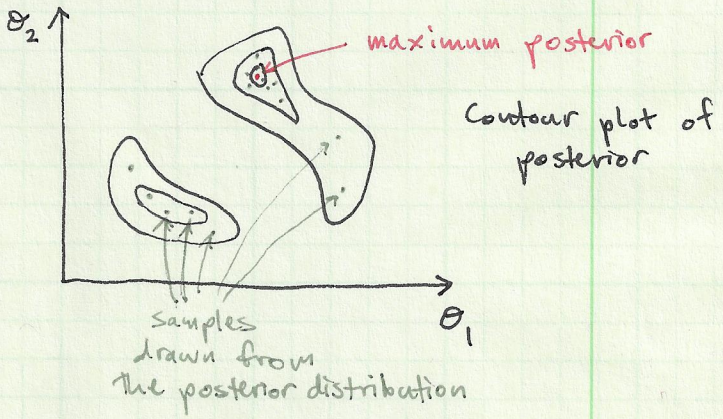
Also, note that $\frac{\sigma}{\mu} = \frac{1}{\sqrt{\mu}}$, so the larger k is, the lower the fractional uncertainty on the rate is. This is typical of Poisson processes.

Dealing with complicated posterior distributions

- So far, I've given simple examples that can be analyzed quite effectively with pen and paper.
- However, in real-world problems, working directly with the posterior distribution can often be difficult.
  - The likelihood may involve a large amount of data, and could be computationally expensive to calculate.
  - The posterior may not look like a Gaussian, so the Gaussian approximation trick we've been using may be inadequate.
- Let's imagine a complicated 2-dimensional posterior,

$$p(\theta_1, \theta_2 \mid D)$$

(θ_1, θ_2: 2 model parameters; D: data)



This distribution has two "modes" (regions of high probability density), so it cannot be well approximated by a single Gaussian.

Ways to deal with this posterior distribution:

1. Find θ_1, θ_2 that maximize the posterior.
   → We lose all information about uncertainty.
2. Draw N samples from the posterior.

$$\vec{\theta}_i \sim p(\vec{\theta} \mid D), \quad i = 1, 2, \ldots, N.$$

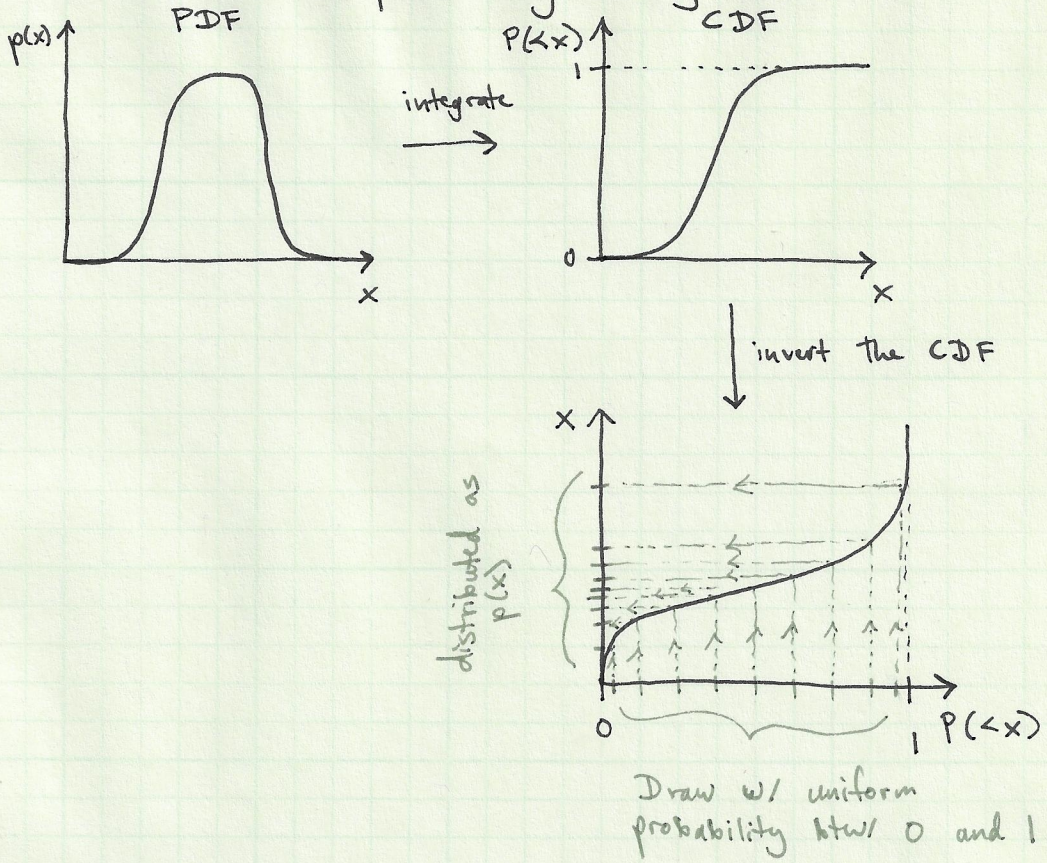
($\sim$: "drawn from" or "distributed according to"; represent (θ_1, θ_2) by a vector, $\vec{\theta}$.)

3. Try to approximate the posterior by a sum ("mixture") of Gaussians.

All 3 approaches are used, but we'll focus on #2: sampling.

Sampling 1D posterior distributions

- There's a simple algorithm for drawing samples from a 1-dimensional probability density function.



Construct the CDF, then invert it (this can be done numerically). Then, draw

$$u_i \sim U(0,1), \quad i = 1, 2, \ldots, N. \leftarrow N \text{ samples}$$

↑ uniform distribution from 0 to 1

Put these random variables through CDF^{-1} to obtain samples of x:

$$x_i = CDF^{-1}(u_i).$$

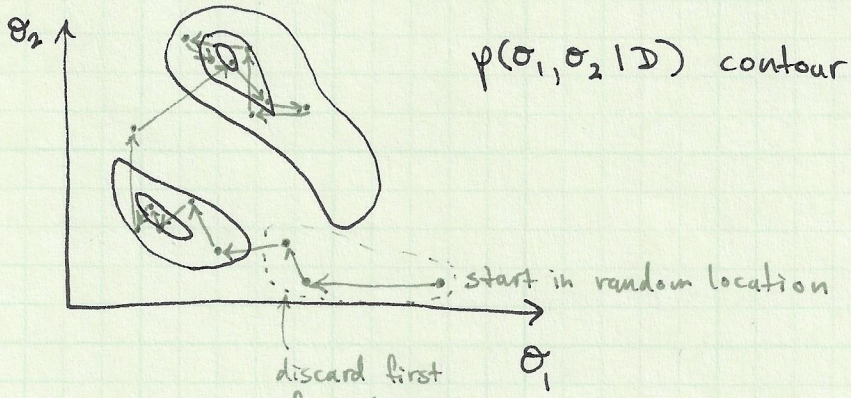
$$\Rightarrow x_i \sim p(x).$$

This is great for 1D distributions, but it doesn't easily extend to higher-dimensional distributions, because there's no concept of a CDF in >1 dimension (you can't order >1-dimensional vectors).

Markov Chain Monte Carlo (MCMC)

- This is where the idea of "Markov Chain Monte Carlo" (MCMC) comes in.

- MCMC is a way of taking a <u>random walk</u> through space, guided by a probability distribution. In the end, the points that have been visited are distributed according to the probability distribution.



$p(\theta_1, \theta_2 | D)$ contour plot.

⇒ Remaining points distributed according to $p(\theta_1, \theta_2 | D)$.

- Once we have our Markov Chain samples, we can do all of our further analysis on them, rather than on the full distribution.

- <u>Ex.</u>: Calculating the mean of the distribution.

$$\langle \theta_1 \rangle = \int p(\vec{\theta}|D)\,\theta_1\, d^2\vec{\theta} \approx \frac{1}{N}\sum_{i=1}^{N} \theta_{i,1}, \quad \text{where } \vec{\theta}_i \sim p(\vec{\theta}|D).$$

An integral over the distribution turns into a sum over samples drawn from the distribution.

- <u>Ex.</u>: Calculating the expectation value of any function of $\vec{\theta}$.

$$\langle f(\vec{\theta}) \rangle = \int p(\vec{\theta}|D) f(\vec{\theta})\, d^2\vec{\theta} \approx \frac{1}{N}\sum_{i=1}^{N} f(\vec{\theta}_i).$$

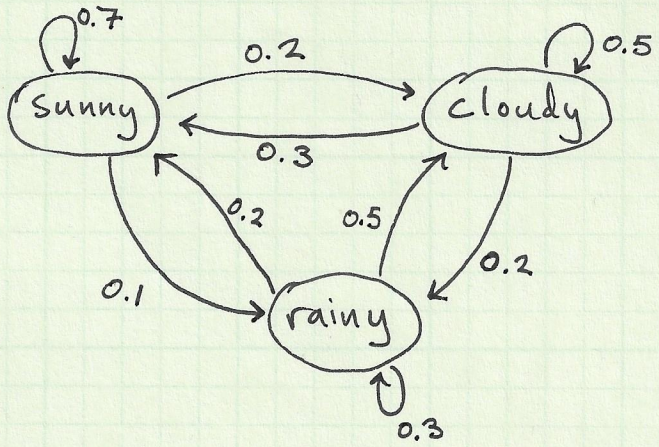
↳ We could use this to calculate the (co)variance of $\vec{\theta}_1$, or even expectation values of various quantities in our theoretical model.

Markov Chains on discrete spaces

Markov Chains are easiest to understand on discrete spaces (with probability mass functions).

Consider a simple model of the weather. Each day is sunny, cloudy or rainy.

- Begin in a random state.
- Transition probability: $p(\text{next state} \mid \text{current state})$.
  - ↳ "Memoryless": Tomorrow's weather depends only on today's weather. You do not need to know the history of weather to predict tomorrow's weather.



Represent these probabilities as a matrix:

"Transition matrix" P_{ij}

current state \ next state	s	c	r
s	0.7	0.2	0.1
c	0.3	0.5	0.2
r	0.2	0.5	0.3

← Each row sums to 1

All entries: $0 \leq P_{ij} \leq 1$.

Let's say we begin with an ensemble of states at day 0.

$\vec{n}^0 = (n_s^0 \quad n_c^0 \quad n_r^0)$. (number that are sunny, cloudy or rainy on day 0).

We transition to the next day:

$$\vec{n}^1 = \vec{n}^0 P. \quad \rightarrow \quad \vec{n}^2 = \vec{n}^1 P = (\vec{n}^0 P) P = \vec{n}^0 P^2.$$

$$\Rightarrow \vec{n}^k = \vec{n}^0 P^k.$$

We can also divide $\vec{n}$ by the total number of states, N, to get the fraction of our ensemble that is in each state:

$$\vec{\pi}^k \equiv \frac{1}{N} \vec{n}^k, \quad N \equiv n_s^k + n_c^k + n_r^k. \text{ (N doesn't change with k).}$$

The same transition rule holds for $\vec{\pi}^k$:

$$\vec{\pi}^{k+1} = \vec{\pi}^k P.$$

Most transition matrices have an interesting property:

$$\lim_{k\to\infty} P^k = \begin{pmatrix} P_0 & P_1 & P_2 \\ P_0 & P_1 & P_2 \\ P_0 & P_1 & P_2 \end{pmatrix}.$$

↑ Each column is filled with a constant.

For our transition matrix,

$$\lim_{k\to\infty} P^k \approx \begin{pmatrix} 0.47 & 0.36 & 0.17 \\ 0.47 & 0.36 & 0.17 \\ 0.47 & 0.36 & 0.17 \end{pmatrix}.$$

This means that regardless of $\vec{\pi}^0$, we always end up approaching the same $\vec{\pi}^k$ as $k\to\infty$:

$$\lim_{k\to\infty} \vec{\pi}^k = \vec{\pi}^0 \lim_{k\to\infty} P^k = (\pi_s^0\ \pi_c^0\ \pi_r^0) \begin{pmatrix} P_s & P_c & P_r \\ P_s & P_c & P_r \\ P_s & P_c & P_r \end{pmatrix}$$

$$= \left(\underbrace{(\pi_s^0+\pi_c^0+\pi_r^0)}_{=1} P_s \quad (\pi_s^0+\pi_c^0+\pi_r^0) P_c \quad (\pi_s^0+\pi_c^0+\pi_r^0) P_r \right)$$

$$= (P_s\ P_c\ P_r). \quad \leftarrow \text{"stationary distribution" of } P.$$

↑ Probability of it being sunny far in the future, regardless of the weather today.

The weather "forgets" its initial state, over the long run.

↳ Common property of Markov Chains.

We went from transition probabilities to a probability distribution over all the states.

↳ The point of MCMC is to engineer a transition matrix for our system that we know will generate the probability distribution we want.

Markov Chains on discrete spaces, cont'd

If $\vec{\pi}$ is the "stationary distribution" that a Markov Chain asymptotes to, then transitioning does not change the distribution:

$$\vec{\pi} P = \vec{\pi}. \quad \leftarrow \text{"detailed balance"}$$

$$\Rightarrow \vec{\pi} \text{ is a left eigenvector of } P \text{ w/ eigenvalue of } 1.$$

This condition is called "detailed balance", and it means that the flux into any state is equal to the flux out of that state.

Metropolis-Hastings algorithm

This is a clever algorithm to generate a transition matrix that maintains detailed balance for an arbitrary probability distribution.

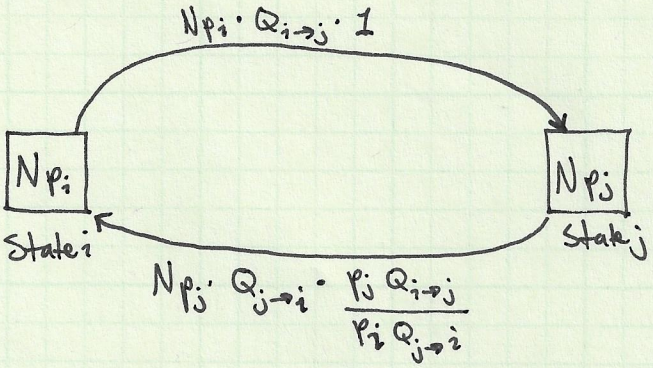
Metropolis-Hastings algorithm:

Beginning in state i,

1. Propose a new candidate j, using a "proposal distribution" $Q_{i \to j}$ that depends on i.
   - ↳ The proposal distribution might choose candidates that are close to or similar to state i in some sense.
2. If $p_j \geq p_i$, then accept the candidate. This is then the new state. If $p_j < p_i$, then accept the candidate with probability
$$\alpha = \frac{p_j}{p_i} \frac{Q_{j \to i}}{Q_{i \to j}}.$$
Otherwise, remain at state i.

This algorithm maintains detailed balance by ensuring that if an ensemble of states are in the desired "target distribution", then the flux between any pair of states is equal in each direction.

Consider 2 states, i and j, with $p_i \geq p_j$. Assume we have an ensemble of N Metropolis-Hastings "walkers" stepping through state space.



$= N p_i Q_{i \to j}$

Flux from $i \to j$:

$N p_i \cdot Q_{i \to j} \cdot 1$

- $N p_i$: # in state i
- $Q_{i \to j}$: proposal prob.
- 1: acceptance prob.

Flux from $j \to i$:

$N p_j \cdot Q_{j \to i} \cdot \frac{p_j Q_{i \to j}}{p_i Q_{j \to i}} = N p_i Q_{i \to j}$

- $N p_j$: # in state j
- $Q_{j \to i}$: proposal prob.
- $\frac{p_j Q_{i \to j}}{p_i Q_{j \to i}}$: acceptance prob.

If the flux btw/ any pair of states is balanced, then there is no net flux into or out of any state.

Metropolis-Hastings algorithm, cont'd

We've generated a way of stepping through state space that maintains detailed balance for an arbitrary probability distribution.

If we begin in a random state and start stepping using this algorithm, we'll eventually visit each state in proportion to its probability.

↳ We discard the first few states in the chain, because it takes some time for the chain to "forget" its initial state.

Metropolis-Hastings was originally invented for thermodynamics computations.

- A thermodynamic system with states $i = 1, 2, \ldots$, with energies $\varepsilon_1, \varepsilon_2, \ldots$, and temperature T. The probability of being in state i is proportional to

$$p_i \propto e^{-\varepsilon_i / k_B T}.$$

(k_B: Boltzmann's constant)

→ Normalize probabilities:

$$1 = \sum_i p_i = \frac{1}{Z} \sum_i e^{-\varepsilon_i / k_B T}.$$

$$\Rightarrow Z = \sum_i e^{-\varepsilon_i / k_B T}.$$

$$\Rightarrow p_i = \frac{1}{Z} e^{-\varepsilon_i / k_B T}.$$

What if we wanted to calculate the average energy of the system at temperature T?

$$\langle \varepsilon \rangle = \sum_i \varepsilon_i p_i = \frac{1}{Z} \sum_i \varepsilon_i e^{-\varepsilon_i / k_B T}.$$

Often, the number of states is so large that computing this sum directly is infeasible.

↳ Estimate it by drawing N states, according to their probabilities p_i, and computing the average energy of that subsample of states.

Metropolis-Hastings on continuous spaces

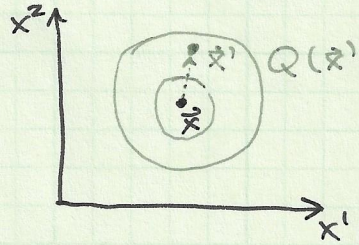
So far, we've only considered Markov chains on discrete spaces. However, they also carry over to continuous spaces (w/ probability densities):

Transition matrix: $P_{i,j}$ (conditional PMF) $\longrightarrow$ $P(\vec{x}' \mid \vec{x})$ (conditional PDF), where $\vec{x}'$ = next state, $\vec{x}$ = current state.

Stationary distribution: $\vec{\pi}$ (PMF) $\longrightarrow$ $\pi(\vec{x})$ (PDF).

Proposal distribution: $Q_{i \to j}$ (conditional PMF) $\longrightarrow$ $Q(\vec{x}' \mid \vec{x})$ (conditional PDF), where $\vec{x}'$ = proposed state, $\vec{x}$ = current state.

For Metropolis-Hastings on a continuous space, we first propose a new candidate state $\vec{x}'$, based on the current state $\vec{x}$:



$Q(\vec{x}' \mid \vec{x})$ is typically some simple distribution centered on $\vec{x}$ (e.g., a Gaussian distribution).

If $p(\vec{x}') \geq p(\vec{x})$, we take the step. If $p(\vec{x}') < p(\vec{x})$, we take the step w/ probability

$$\alpha = \frac{p(\vec{x}')}{p(\vec{x})} \frac{Q(\vec{x} \mid \vec{x}')}{Q(\vec{x}' \mid \vec{x})}.$$

Otherwise, we remain at $\vec{x}$.

Applying MCMC for Inference

With this overview of MCMC, let's go back to Bayesian inference.

We typically have data D, theoretical parameters θ, and various quantities dependent on our theoretical parameters that we would like to calculate averages (or variances, ...) of: $f(\theta)$.

$$p(\theta|D) = \frac{p(D|\theta)\,p(\theta)}{p(D)}$$

$$= \frac{p(D|\theta)\,p(\theta)}{\underbrace{\int p(D|\theta')\,p(\theta')\,d\theta'}_{\equiv Z(D)}}$$

↑ the evidence serves as a normalization constant that does not depend on θ.

$$= \frac{1}{Z(D)}\,p(D|\theta)\,p(\theta).$$

To calculate $\langle f(\theta)\rangle$, we could attempt to calculate the integral

$$\langle f(\theta)\rangle = \int \frac{1}{Z(D)}\,p(D|\theta)\,p(\theta)\,f(\theta)\,d\theta,$$

but this is often infeasible.

Instead, we sample from $p(\theta|D)$ using MCMC, and then use the estimate

$$\langle f(\theta)\rangle \approx \frac{1}{N}\sum_{\text{sample } i} f(\theta_i).$$

The nice thing is that we don't even have to calculate $Z(D)$. Note that in the Metropolis-Hastings algorithm, we only use ratios of probability densities:

$$\frac{p(\theta'|D)}{p(\theta|D)} = \frac{\frac{1}{Z(D)}\,p(D|\theta')\,p(\theta')}{\frac{1}{Z(D)}\,p(D|\theta)\,p(\theta)} = \frac{p(D|\theta')\,p(\theta')}{p(D|\theta)\,p(\theta)}.$$