86-631/42-631 Neural Data Analysis
Lecture 04: Review of probability II

1) Expectation
    a. Linearity, expectation of a linear function of X
    b. Expectation of a function
2) Variance
3) Sample mean and variance
4) Continuous random variables
    a. CDFs and PDFs
5) Some useful continuous distributions
    a. Uniform
    b. Gaussian
    c. Exponential
    d. Poisson
6) Random vectors
    a. Joint distributions
    b. Marginal distributions
    c. Independence of joint distributions
7) ***Slides: Fano factor in practice (Churchland et al. 2011)***

Before we go on to defining some useful distributions, there are two crucial ideas to learn: *expectation*, and *variance*.

Expectation and Variance

Imagine performing an experiment where you were trying to measure how many times a neuron would fire in response to a particular stimulus – like the brief tone pips we looked at last class. Say you record from the neuron on 50 separate trials, and you get the following responses:

# of spikes	3	4	5	6	7	8
Count	3	7	12	14	10	4

Since we ran the experiment 50 times, we could come up with a probability distribution that represents the probability of observing a particular number of spikes on any given trial, by dividing the count by 50:

X	3	4	5	6	7	8
P(x)	.06	.14	.24	.28	.20	.08

If we were to create a bar plot showing the number of measurements of x (against x), we'd have a *histogram* of the count distribution. If we were to do the same thing with p(x), we would, instead, have a plot of the *relative frequencies*. A plot of the probability distribution is the same as a histogram, except that the total amount of probability must equal 1.

This brings up a rather important point. One way to define the probability of an event is through its relative frequency:

Frequency histogram (adds to 1)
Count histogram (adds to n)

$$P(E) \equiv \lim_{n\to\infty} \frac{n(E)}{n}$$

$n(E)$: # of events occurring in trial; n: total # of trials ran

where n(E) is the number of times the event was observed and n is the total number of times the experiment was performed. Although this is a nice intuitive definition for the probability of an event, it is difficult to use in practice because the limit can never be observed. We need theoretical tools like we develop through the calculus of probability to understand the behavior and convergence properties of these limits under certain conditions.

Now, how would we calculate the average number of spikes we observed during our experiment? Typically, we would compute the following:

$$mean = \frac{3(3)+7(4)+12(5)+14(6)+10(7)+4(8)}{50} = 5.66$$

An alternative way to write this equation is:

probability

$$mean = 3\left(\frac{3}{50}\right)+4\left(\frac{7}{50}\right)+5\left(\frac{12}{50}\right)+6\left(\frac{14}{50}\right)+7\left(\frac{10}{50}\right)+8\left(\frac{4}{50}\right)=5.66$$

Which is the same thing as:

$$mean = 3p(3)+4p(4)+5p(5)+6p(6)+7p(7)+8p(8)=5.66$$

This is one example of the *mean* or *expected value* or *expectation* of the random variable X:

mean

$$\mu_X = E[X] \equiv \sum_{x \in X} xp(x)$$

range of x

One way to think of expectation is as an operation: essentially, it means "take a weighted average of", where the weight is determined by the probability. As an example:

$$E[f(x)] = \sum_{x \in X} f(x)p(x)$$

gives you a centroid of data

Now, the mean is only one way to summarize a probability distribution. Another way to summarize it is by its *variance*:

Variance

$$\sigma_X^2 = Var[X] \equiv \sum_{x \in X} (x-\mu_X)^2 p(x)$$

result is non-negative

=0 when all the's equal the mean

gives you a spread of data

The *standard deviation* of X is σ_X. It summarizes, roughly, the average amount of deviation around the mean, ie, the spread. Note that this means:

$$Var[X]=E[(X-\mu_X)^2].$$

Some properties of expectation:

Some properties of expectation prove useful. (Prove these on board.)

prove these yourselves

(1) Linearity. $E[aX+b]=aE[X]+b$. → useful for derivations

(2) $Var[aX]=a^2Var[X]$.

(3) $Var[X]= E[(X-\mu_X)^2 = E[X^2]-E^2[X]$.

(1) $E[aX+b] = \sum_{x \in X} (ax+b)p(x)$

$= \sum_x aXp(x) + \sum_x bp(x)$

$= a\sum_x xp(x) + b\sum_x p(x)$

$= aE[X] + b$

Example: mean and variance of the Bernoulli distribution.

Recall the Bernoulli distribution: P(0)=(1-q), P(1)=q. We can easily compute the mean of this distribution:

μ=0(1-q)+1(q)=q. $\mu = \sum_{x \in X} xp(x)$

Similarly,

$\sigma_X^2 = \sum_{x \in X} (x - \mu_X)^2 \, p(x)$ $\sigma^2=(0-q)^2(1-q)+(1-q)^2q=q^2-q^3+q(1-2q+q^2)=q^2-q^3+q-2q^2+q^3=q-q^2=q(1-q)$.

Example: mean of the Binomial distribution.

$$E[k] = \sum_{k=0}^{n} k \binom{n}{k} q^k (1-q)^{n-k} = \sum_{k=1}^{n} k \binom{n}{k} q^k (1-q)^{n-k}$$

Note: $$\binom{n}{k} = \frac{n!}{k!(n-k)!} = \frac{n(n-1)!}{k(k-1)!(n-1-k+1)!} = \frac{n}{k}\binom{n-1}{k-1}$$

Let k'=k-1, n'=n-1 $$E[k] = \sum_{k=1}^{n} n \binom{n-1}{k-1} q^k (1-q)^{n-k} = nq \sum_{k=1}^{n} \binom{n-1}{k-1} q^{k-1} (1-q)^{n-k}$$

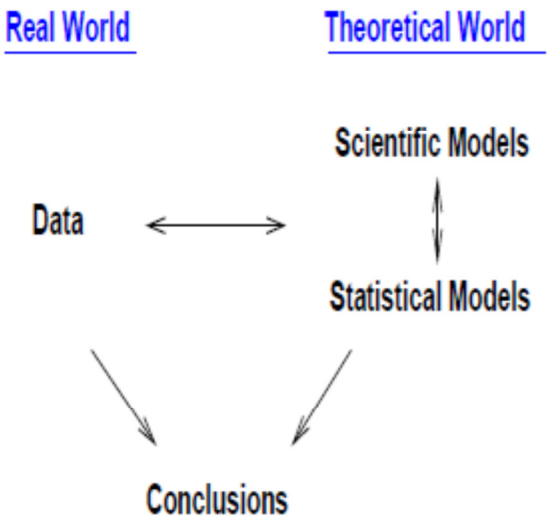
$$E[k] = nq \sum_{k'=0}^{n'} \binom{n'}{k'} q^{k'} (1-q)^{n'-k'} = nq$$

theoretical values (∞ samples)

outcomes of experiment (n samples)

Mean, variance, sample mean, and sample variance.

A point worth stressing is that the mean μ and standard deviation σ are both theoretical quantities of a probability distribution. When we measure data, we may assume it comes from a particular distribution, but the data itself is not the distribution; the distribution is a theoretical construct that describes the data. The data are, in some sense, a *sample* of the population distribution. μ and σ define the *population mean* and *population standard deviation*. When we compute these quantities from the data, we are really computing the *sample mean* and *sample standard deviation*.



The sample mean is computed as:

Sample mean

$$\bar{X} = \frac{1}{N}\sum_{i=1}^{N} x_i$$

The sample variance is computed as:

Sample variance

$$s^2 = \frac{1}{N-1}\sum_{i=1}^{N}\left(x_i - \bar{X}\right)^2$$

Note that the sample mean and sample variance are computed in almost the same way as the population (or theoretical) mean and variance. In fact, when we introduced expectation, we saw that if the data come out with the same relative frequencies as the distribution, the population and sample means will be the same. For rather esoteric reasons, the sample variance and population variance are slightly different. This is because when the sample mean is used in place of the true mean, it turns out dividing the sum by 1/N would result in a *biased* estimate of the variance. (This is because the sample mean will be, in a sense, closer to the centroid of the data than the population mean would be. This will result in an underestimate of the spread. To correct for this, we divide by N-1 instead of N, which returns an *unbiased* estimate of the variance that is typically closer to the true value of the population.

Population

- Mean

summing over all possible outcomes

$$M_X = E[X] = \sum_{x \in X} x\,p(x)$$

- Variance

$$\sigma_X^2 = Var[X] = \sum_{x \in X} (x - M_X)^2 p(x)$$

Sample

- Sample Mean

sum over outcomes

$$\bar{X} = \frac{1}{N}\sum_{i=1}^{N} x_i$$

- Sample Variance

$$s^2 = \frac{1}{N-1}\sum_{i=1}^{N} (x - \bar{X})^2$$

accounting for natural bias

N → Biased Smaller than pop Var

Discrete → values on # line

Continuous → real line

PMFs, PDFs, and CDFs

Last class, we talked about the binomial distribution as one example of a random variable. Recall that this distribution counts the number of times an event that occurs independently on every trial with probability p when the test is independently performed n times. The probability of observing the event occur exactly k times is

Probability of each point

If K~Binomial(n,q), then

$$P(K=k)=\binom{n}{k}q^k(1-q)^{n-k}$$

This is known as the *probability mass function* (pmf) of the binomial distribution.

discrete random variables

Of course, random variables could be discrete, like the binomial distribution, or they could be continuous. Continuous distributions take on values over the real line; an example of a very popular continuous random variable is the Gaussian distribution. In the case of continuous random variables, the pmf is called a *probability density function* (pdf), and is denoted f(x). It has the following relationship: Suppose X is a random variable on an interval (A,B), with A=-∞ and B=∞ both being possible. Then the pdf f(x) is:

Smeared probability over real line

$$P(a\le X\le b)=\int_a^b f(x)dx \qquad \int_A^B f(x)dx=1$$

What are units of p(x)? → Unitless

of f(x)? → x^{-1} (ex: ms^{-1})

Note that for continuous distributions, P(a≤X) = P(a<X) (b/c P(X=a)=0). We should think of f(x) as the probability per unit of x; f(x)dx is the probability that X will lie in some infinitesimal interval about x. As a convention, when the random variable doesn't actually go out to +/- ∞, we simply extend it to infinity and define f(x)=0 where there is no support.

(Sometimes, the term probability density function is applied to both continuous and discrete distributions.) Random variables are also described by their *cumulative distribution function* (cdf), which is defined as F(x)≡P(X≤x). For continuous distributions,

For any value a, CDF tells you P(X≤a):

$$F(a)=\int_{-\infty}^{a} f(x)dx$$

(plotted as fxn of a)

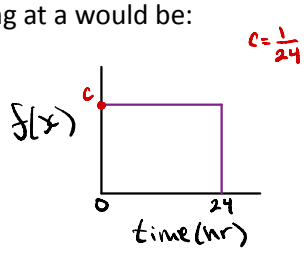
for discrete distributions, the integration is replaced by a sum.

Example: Uniform distribution. ⟶

The simplest example of a continuous random variable is the uniform distribution. (Eg, if the time of day when births occurred followed a uniform distribution, then the probability of a mother giving birth in any given 30 minute interval would be the same as any other given 30 minute interval.) In this case, f(x) would be constant over the interval from 0 to 24 hours. Because it must integrate to 1, we must have f(x)=1/24, and the probability of a birth in any given 30 minute interval starting at a would be:

pdf is flat (constant)

$$P(a \le X \le a+.5) = \int_a^{a+.5} \frac{1}{24} dx = \frac{1}{48}$$



When a random variable has a uniform distribution over the interval (A,B), we write this as X~U(A,B), and the pdf is f(x)=1/(B-A). Why? $\int_A^B f(x)\,dx = 1$

Note: ~ means "distributed as". $X \sim \text{Uniform}(A,B)$

Note: the matlab function rand returns a random variable X~U(0,1).

(How would you generate a uniform random variable between 5 and 10 in Matlab?)

Expectation and variance are defined on continuous distributions in a similar way as with discrete distributions, except using integrals instead of sums:

$$\mu_X = E[X] \equiv \int_A^B x f(x) dx \qquad \sigma_X^2 = Var[X] \equiv \int_A^B (x-\mu_X)^2 f(x) dx$$

For the Uniform distribution U(A,B),

$$\mu_X = \int_A^B \frac{x}{B-A} dx = \frac{1}{B-A}\frac{1}{2}x^2\Big|_A^B = \frac{1}{2}\frac{(B^2-A^2)}{B-A} = \frac{1}{2}\frac{(B-A)(B+A)}{B-A} = \frac{(B+A)}{2}$$

$$Var[X] = E[X^2] - E^2[X].$$

$$E[X^2] = \int_A^B \frac{x^2}{B-A} dx = \frac{1}{B-A}\frac{1}{3}x^3\Big|_A^B = \frac{1}{3}\frac{(B^3-A^3)}{B-A} = \frac{1}{3}\frac{(B-A)(B^2+AB+A^2)}{B-A} = \frac{1}{3}(B^2+AB+A^2)$$

$$Var[X] = \frac{1}{3}(B^2+AB+A^2) - \frac{(B+A)^2}{4} = \frac{4B^2+4AB+4A^2-3B^2-6AB-3A^2}{12} = \frac{(B-A)^2}{12}$$

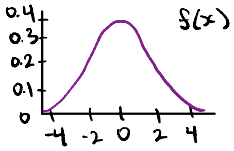
Example: Gaussian distribution (Normal)

Perhaps the most ubiquitous continuous distribution in all of probability and statistics is the normal distribution, also called the Gaussian distribution. This is not because of its ability to describe data; in fact, most data are detectably non-normal in some way or another. Rather, the normal distribution is useful because it is often an accurate description of the variability of quantities *derived from* the data as functions of the sample mean (sample means are often normally distributed, because of the Central Limit Theorem). Further, it turns out to be an extremely useful distribution for doing ordinary linear regression, as the way you fit a line to data (by minimizing the squared difference between predicted values and observed data) turns out to be a provably optimal process when the noise in the data is normally distributed. We will get in to this more later in the class.

assumption made on noise values

Quantities that you derive from your data are gaussian

The normal distribution is characterized by two parameters (the mean and variance), and when a random variable X follows a normal distribution we write $X \sim N(\mu,\sigma^2)$. Its pdf is given by



pdf:

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

μ = mean
σ = Standard deviation
range = $(-\infty, \infty)$

(Draw picture on board, show mean and variance.) Because of its shape, the distribution is often called the "bell-shaped curve". Note the 68-95-99.7 rule:

For normal distributions:

$$P(\mu-\sigma \le x \le \mu+\sigma) \approx 0.68$$

$$P(\mu-2\sigma \le x \le \mu+2\sigma) \approx 0.95$$

$$P(\mu-3\sigma \le x \le \mu+3\sigma) \approx 0.997$$

As we said earlier, the mean of the normal distribution is μ, and the variance is σ^2.

Other things to note:

If $X \sim N(\mu,\sigma^2)$, $aX+b \sim N(a\mu+b, a^2\sigma^2)$, and

$aX_1+bX_2 \sim N(a\mu_1+b\mu_2, a^2\sigma_1^2+b^2\sigma_2^2)$, **IF** X_1 and X_2 are independent.

And, calling randn in Matlab returns a random variable $\sim N(0,1)$.

(How would you get generate an N(2,2) r.v. in Matlab?)

Example: Exponential distribution. (Inter spike intervals or Length of channels open)

A random variable X is said to follow an exponential distribution with parameter λ when its pdf is:

λ = rate parameter of exponential distribution

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

mean = $\frac{1}{\lambda}$

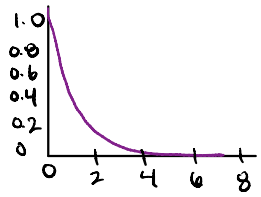
Var = $\frac{1}{\lambda^2}$

It's cdf is given by

$$f(x;\lambda) = \begin{cases} \lambda e^{-\lambda x} & x \geq 0 \\ 0 & x < 0 \end{cases}$$

$$F(x) = \int_0^x \lambda e^{-\lambda t}\, dt = -e^{-\lambda t}\Big|_0^x = 1 - e^{-\lambda t}$$

(Draw picture on board.) Note 1/e=0.3679.



We say this random variable is X~Exp(λ).

λ = range = $(0,\infty)$

The exponential distribution:

$$E[X]=1/\lambda$$

$$Var[X]=1/\lambda^2$$

Turns out that the length of time an ion channel stays open will often follow an exponential distribution. The exponential distribution has extremely close ties to the Poisson process, which we will discuss in the next class. Finally, the exponential distribution is considered to be *memoryless*. That is, if we know the open time of an ion channel follows an exponential distribution with parameter λ, how long would we expect a given channel to remain open if we know that it has already stayed open for time 1/λ?

(What units does f(t) have for the exponential distribution? What units does F(t) have?)

The Poisson Distribution (discrete)

Probably the most useful probability distribution in all of neuroscience is the Poisson distribution. Why? Because it turns out that it's a pretty good description of the probability associated with seeing a particular number of spikes from a neuron. spikes occurring independently (not perfect because of refractory periods)

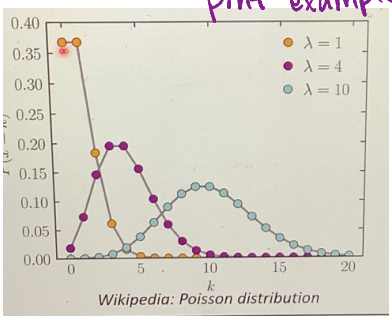
We say that a random variable X is Poisson distributed with rate parameter λ (written X~Poisson(λ)) when the probability of observing k events (spikes) is:

prove this!! mean = ↗, var = ↗, range = non-negative integers

↗ = rate parameter

$$P(K=k)=\frac{\lambda^k e^{-\lambda}}{k!}$$ (k = #spikes)

pmf example



As ↗ ↑ = looks more gaussian

First, let's verify that this is a proper probability distribution.

$$\sum_{k=0}^{\infty} p(k)=1 \qquad \sum_{k=0}^{\infty}\frac{\lambda^k e^{-\lambda}}{k!}=e^{-\lambda}\sum_{k=0}^{\infty}\frac{\lambda^k}{k!}$$

Taylor series:

Recall the Taylor series of a function f(x) about the point a:

$$f(x)=\sum_{k=0}^{\infty}\frac{(x-a)^k f^{(n)}(a)}{k!}$$

So, expanding f(x) about 0 gives $f(x)=f(0)+xf'(0)+x^2f''(0)/2!+x^3f'''(0)/3!+\ldots$

The Taylor expansion of e^x about 0 is therefore $1+x+x^2/2!+x^3/3!+\ldots$ $\qquad e^{\lambda}=\sum_{k=0}^{\infty}\frac{\lambda^k}{k!}$

Therefore,

$$\sum_{k=0}^{\infty} p(k)=e^{-\lambda}\sum_{k=0}^{\infty}\frac{\lambda^k}{k!}=e^{-\lambda}e^{\lambda}=1$$

Mean of the Poisson distribution:

$$E[K]=\sum_{k=0}^{\infty}kp(k)=\sum_{k=0}^{\infty}k\frac{\lambda^k e^{-\lambda}}{k!}=\sum_{k=1}^{\infty}k\frac{\lambda^k e^{-\lambda}}{k!}=\lambda\sum_{k=1}^{\infty}\frac{\lambda^{k-1}e^{-\lambda}}{(k-1)!}$$

Let k'=k-1

$$E[K]=\lambda\sum_{k'=0}^{\infty}\frac{\lambda^{k'}e^{-\lambda}}{(k')!}=\lambda$$

Random vectors

In most experimental settings, data are collected simultaneously on many variables, and the statistical problem is to describe the *joint* variation, meaning their tendency to vary together. The starting point involves m-dimensional *random vectors* (where m is some positive integer), which are the natural multivariate extension of random variables.

An example: say we are able to record from two neurons simultaneously, and are interested in whether their firing in any given period of time is related. We can describe their *joint distribution* from their *joint pdf* or *pmf*:

P(X=x,Y=y).

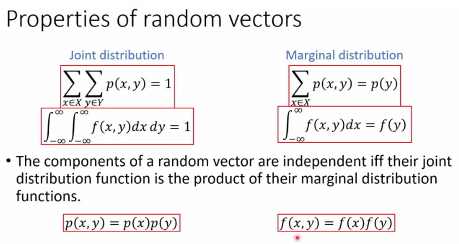
$$P(a \le X \le b, c \le Y \le d) = \int_c^d \int_a^b f(x,y)dxdy$$

The joint pdf has the following properties:

$$\sum_{x\in X}\sum_{y\in Y} p(x,y) = 1 \qquad \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f(x,y)dxdy = 1$$

The *marginal* distribution results when summing (or integrating) over the other interval. For example, if we want to know p(X=x), regardless of y, we may compute:

$$p(x) = \sum_{y\in Y} p(x,y) \qquad f(x) = \int_{-\infty}^{\infty} f(x,y)dy$$



(Show multi-dimensional case.)

Independence of random vectors:

Finally, the components of a random vector are considered independent if and only if their joint distribution function is the product of the marginal distribution functions:

$$p(x,y) = p(x)p(y) \qquad f(x,y) = f(x)f(y)$$

More generally, a set of random variables $X_1, \ldots, X_n$ are independent iff

$$f(x_1,\ldots,x_n) = f(x_1)\ldots f(x_n)$$

Example: spike counts measured on two electrodes:

joint prob distribution

Y				
	2	.05	.05	.1
	1	.05	.15	.1
	0	.3	.15	.05
		0	**1**	**2**
			X	

probability of joint outcomes

P(X)=[.4,.35,.25]; P(Y)=[.5,.3,.2]

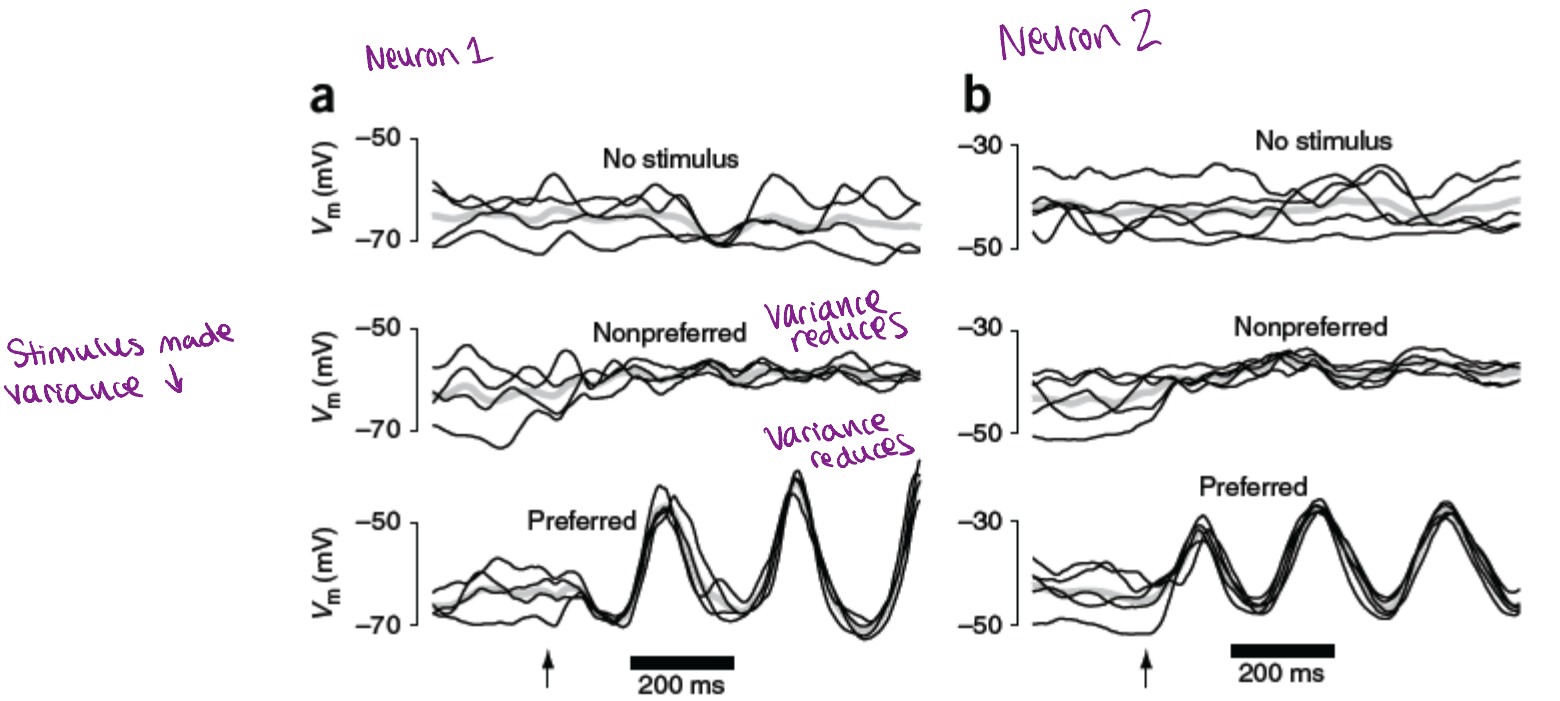
(Are X and Y independent? P(X)P(Y)=.5*.4=.2 ≠.3, so, no, they aren't!)

Do they all sum to 1? yes, good check

$P(X) = [0.3+0.05+0.05,\ 0.15+0.15+0.05,\ 0.05+0.1+0.1] = [0.4, 0.35, 0.25]$

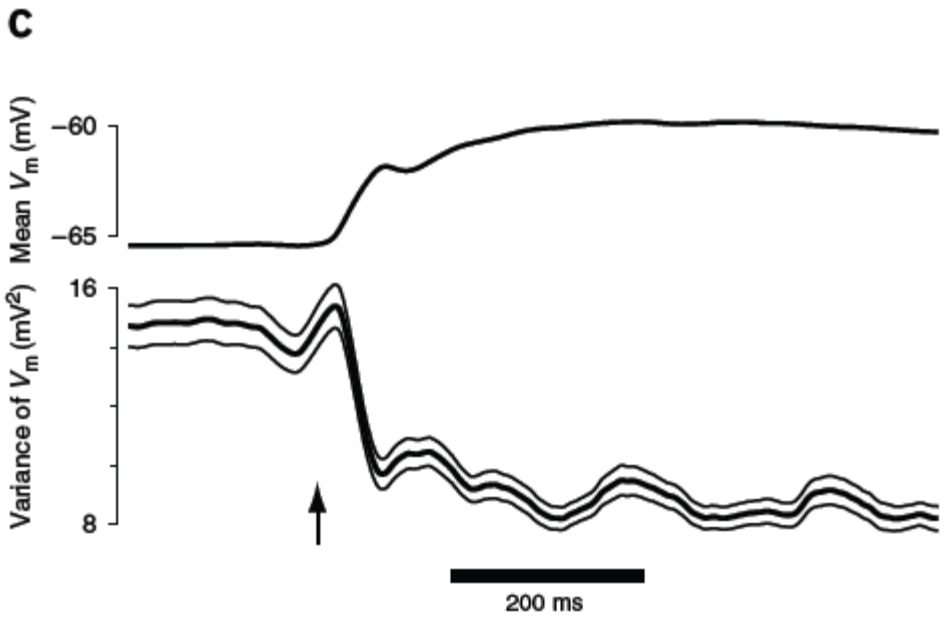
$P(Y) = [0.3+0.15+0.05,\ 0.05+0.15+0.05,\ 0.05+0.05+0.5] = [0.5, 0.3, 0.2]$

Are X & Y independent? No! $P(X=0) * P(Y=0) = (0.5)(0.4) \neq 0.3$



Intracellularly recorded membrane potentials in cat V1 in response to drifting sine wave gratings at various frequencies and orientations.

Churchland et al. (2010) Nat. Neurosci. Fig. 2a,b

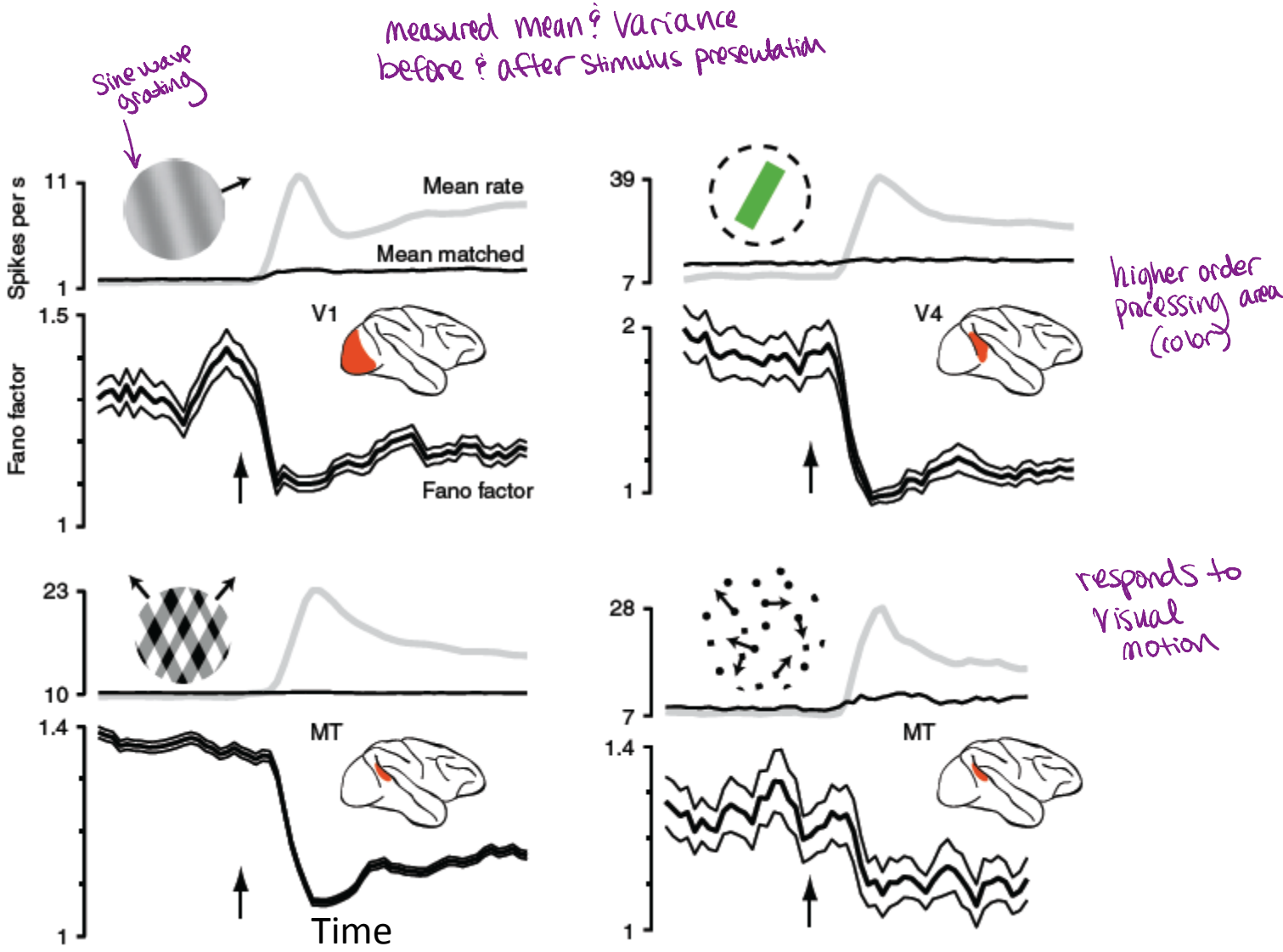


Churchland et al. (2010) Nat. Neurosci. Fig. 2c

Fano Factor

$$FF = \frac{variance}{mean}$$

FF went down



Fano factor v time in various regions of the brain, in response to stimuluation (not necessarily preferred) that the brain area cares about.

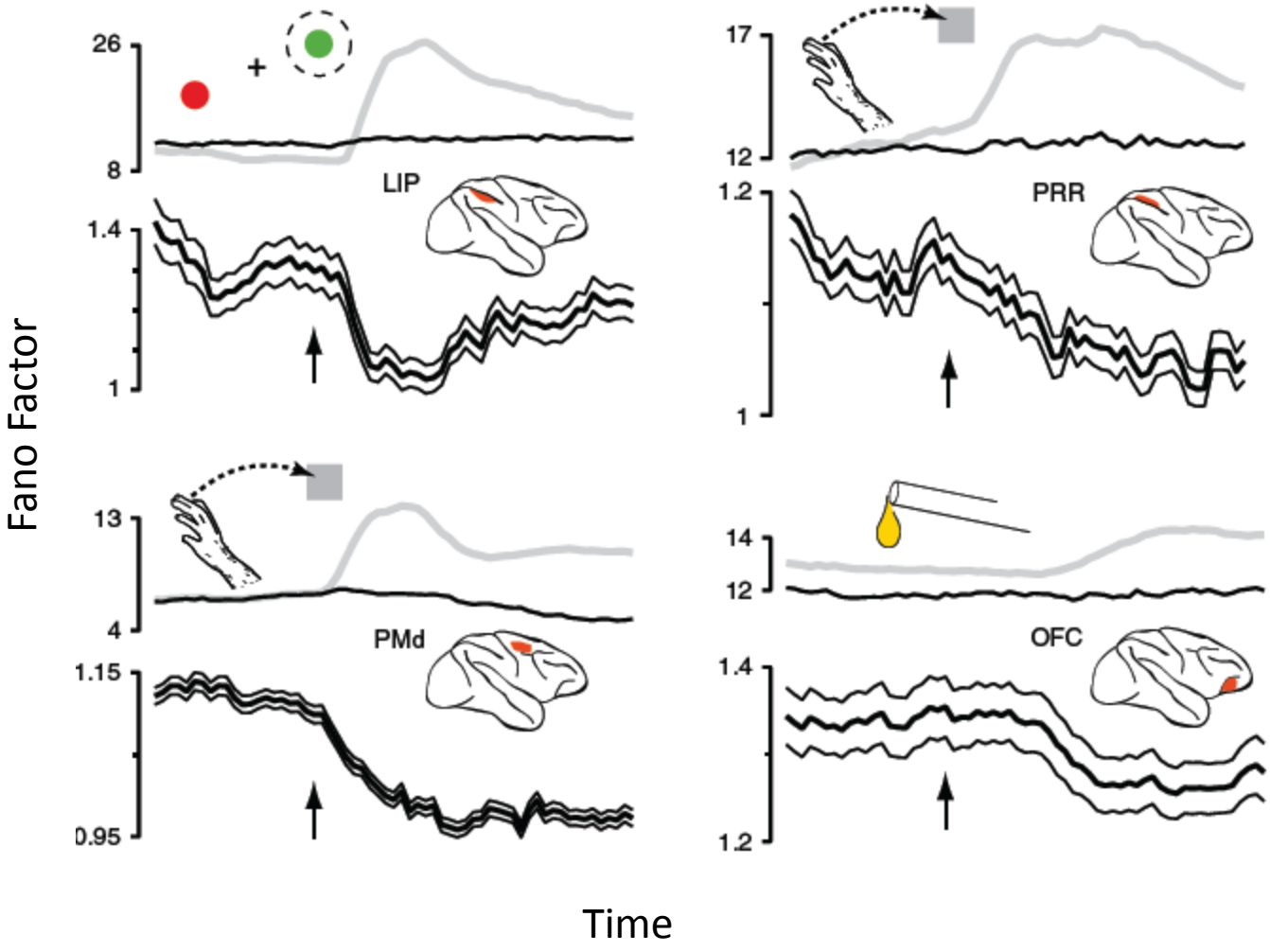
Churchland et al. (2010) Nat. Neurosci. Fig. 3

LIP for planning saccades

PPR plans reaches

PMd reach planning

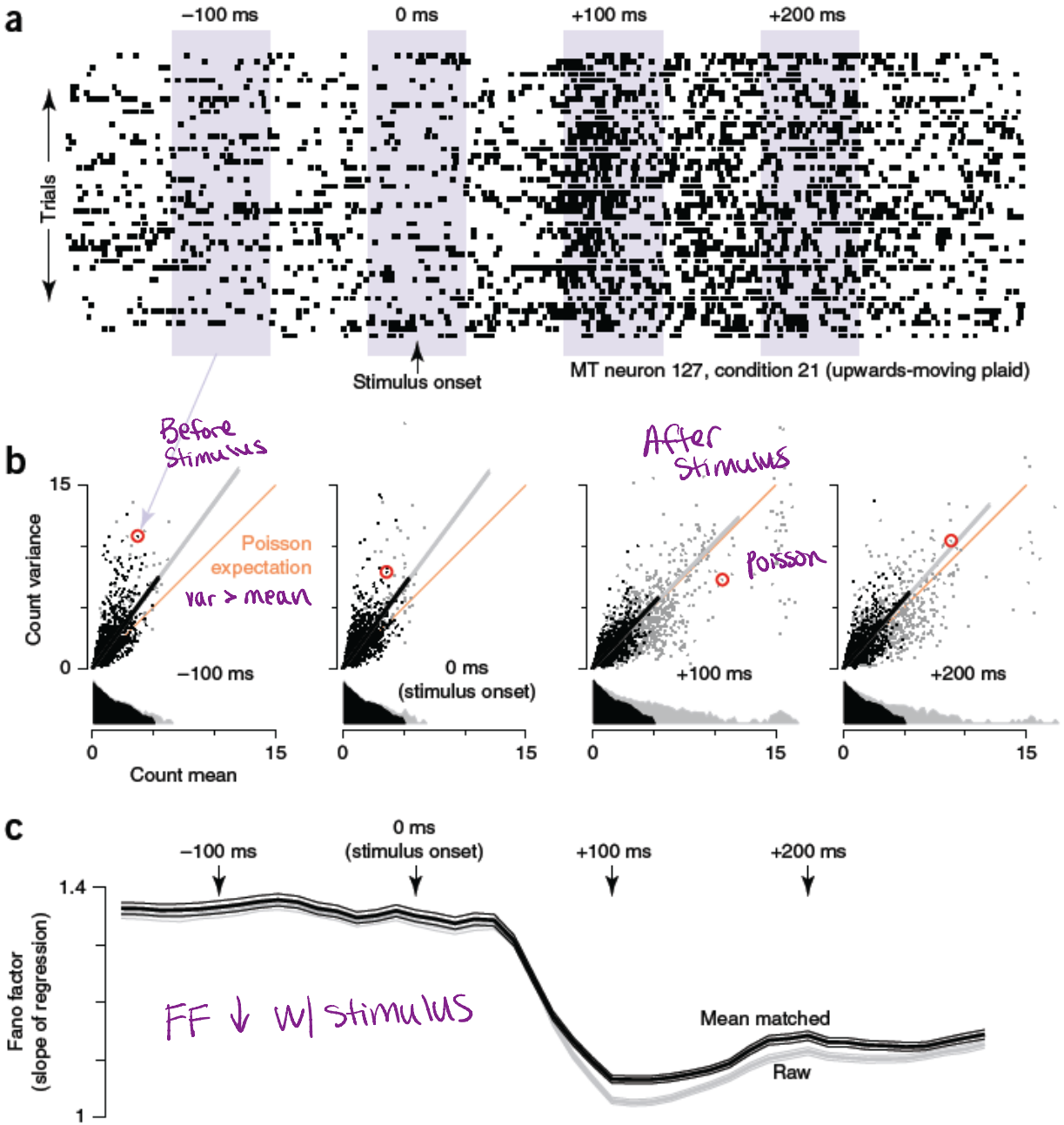
OFC responds to rewards of different types



Churchland et al. (2010) Nat. Neurosci. Fig. 3 (cont.)

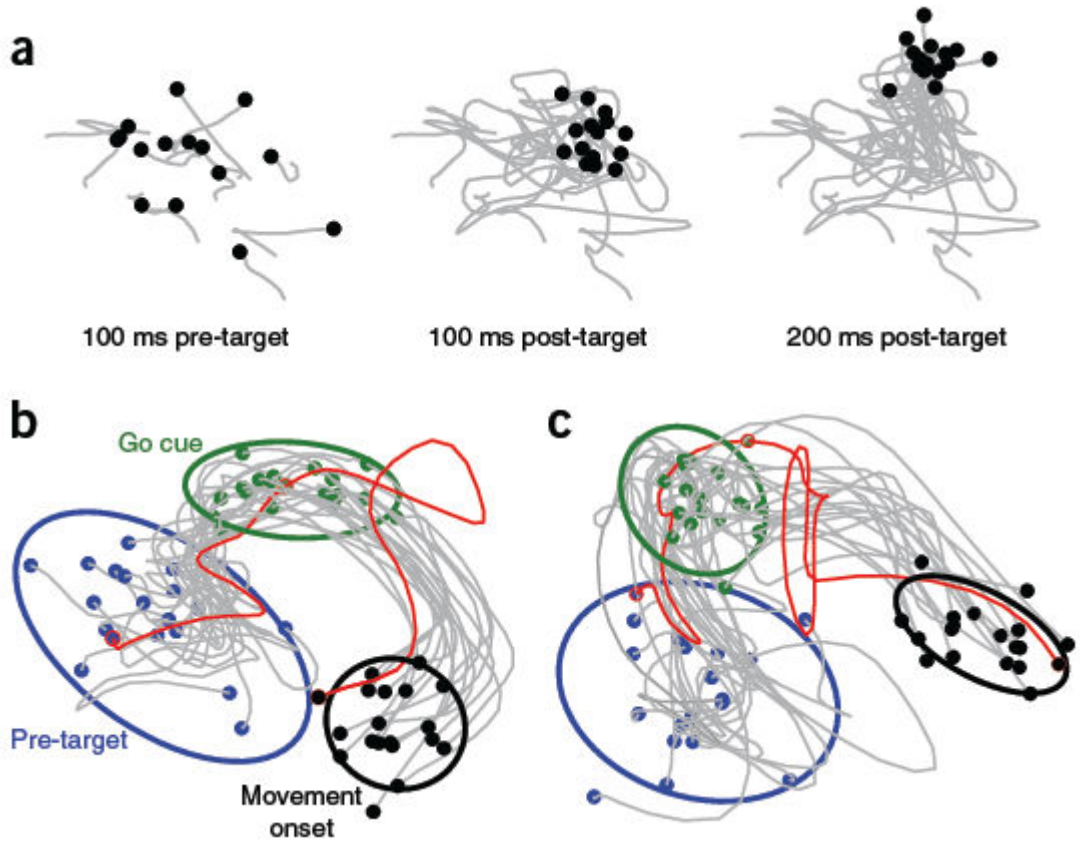
Stimulus quenches neural variability in the brain

Super poisson
↓
poisson



See if you can derive mean of poisson distribution!

Churchland et al. (2010) Nat. Neurosci. Fig. 4



Churchland et al. (2010) Nat. Neurosci. Fig. 7