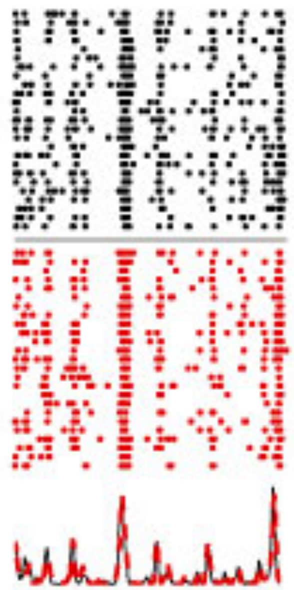
# Probability Theory Intro



Jonathan Pillow

Mathematical Tools for Neuroscience (NEU 314)
Fall, 2021

lecture 13

# PCA review

## the data

$$X = \left.\overbrace{\begin{bmatrix} — \ \vec{x}_1 \ — \\ — \ \vec{x}_2 \ — \\ \vdots \\ — \ \vec{x}_N \ — \end{bmatrix}}^{d}\right\} N$$

collection of N data vectors

## goal

Find a subspace (spanned by columns of B) that captures the maximum projected sum-of-squares

$$\arg\max_B \underbrace{||XB||_F^2} \quad \text{such that} \quad B^T B = I$$

squared Frobenius norm (sum-of-squares of data projected onto subspace)

columns of B are orthogonal unit vectors

## Solution

2nd moment matrix

$$C = X^\top X$$

do SVD

$$C = USU^\top$$

$$\{u_1, \ldots u_k\}$$ first k PCs

$$\frac{s_1 + \cdots + s_k}{s_1 + \cdots + s_N}$$ fraction of sum of squares

# Least Squares Regression review

**the data**

$$X = \overbrace{\begin{bmatrix} — \vec{x}_1 — \\ — \vec{x}_2 — \\ \vdots \\ — \vec{x}_N — \end{bmatrix}}^{d} \quad Y = \left.\overbrace{\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_N \end{bmatrix}}^{1}\right\} N$$

inputs outputs

**goal**

Find weight vector $\overrightarrow{w}$ that minimizes sum of squared residuals

$$\arg\min_{\vec{w}} ||\underbrace{Y - X\vec{w}}_{\text{residuals}}||^2$$

residuals

(difference between observed $y_i$ and linear prediction $\overrightarrow{x}_i \cdot \overrightarrow{w}$ )

**Solution**

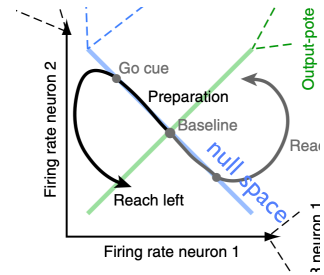
$$\hat{w} = (X^T X)^{-1} X^T Y$$

proof based on:

residuals $(Y - X\overrightarrow{w})$ should be orthogonal to every column of $X$

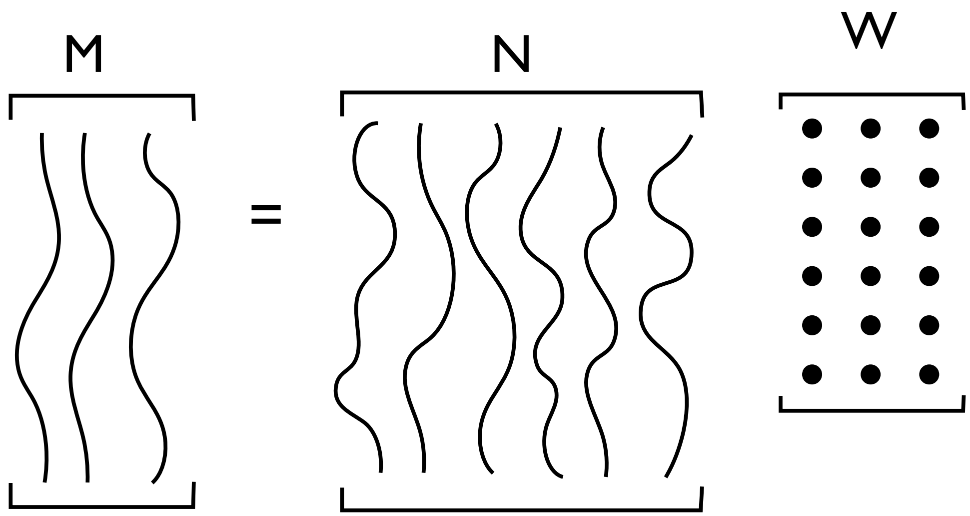
# **Call-back:**

Cortical activity in the null space (Kaufman 2014)




## **Principal components regression (PCR)**

1) Do PCA to reduce dimensionality
2) Then do least squares to estimate weights




Top 3 PCs

Top 6 PCs of neural activity

(each column of W has weights for a single muscle) now: begin probability!

# neural coding problem

$x$

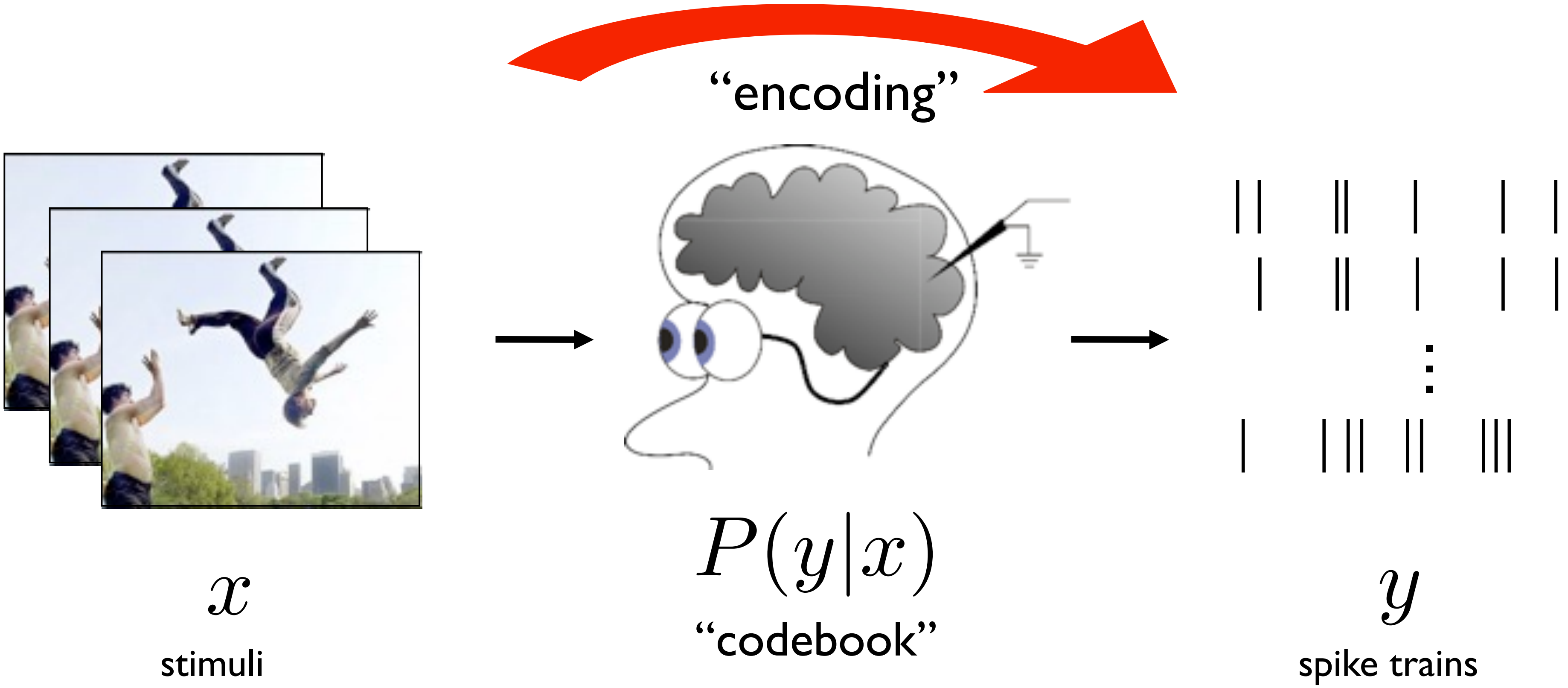
stimuli

$y$

spike trains

- what is the probabilistic relationship between stimuli and spike trains?

# neural coding problem




- what is the probabilistic relationship between stimuli and spike trains?

# neural coding problem

"encoding"

novel stimulus

(Aditi Jha,
Cosyne 2020)

$x$

$\longrightarrow$

$P(y|x)$

"codebook"

$\longrightarrow$

?

$y$

# neural coding problem

"decoding"

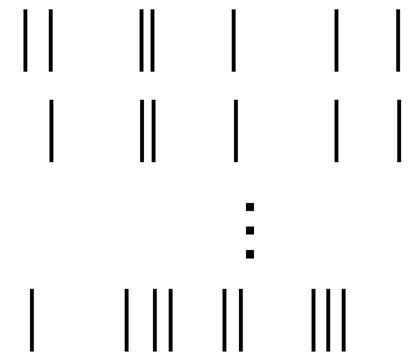
?

"what was that"?

$x$

$P(y|x)$

"codebook"



$y$

posterior likelihood prior

Bayes' Rule: $P(x|y) \propto P(y|x)P(x)$

# Goals for today

- basics of probability
- probability vs. statistics
- continuous & discrete distributions
- joint distributions
- marginalization
- conditionalization
- expectations & moments

- “probability distribution”

- “events”
- “random variables”

parameter

$\theta$

model

$$P_\theta(x)$$

samples

$X$

also written:

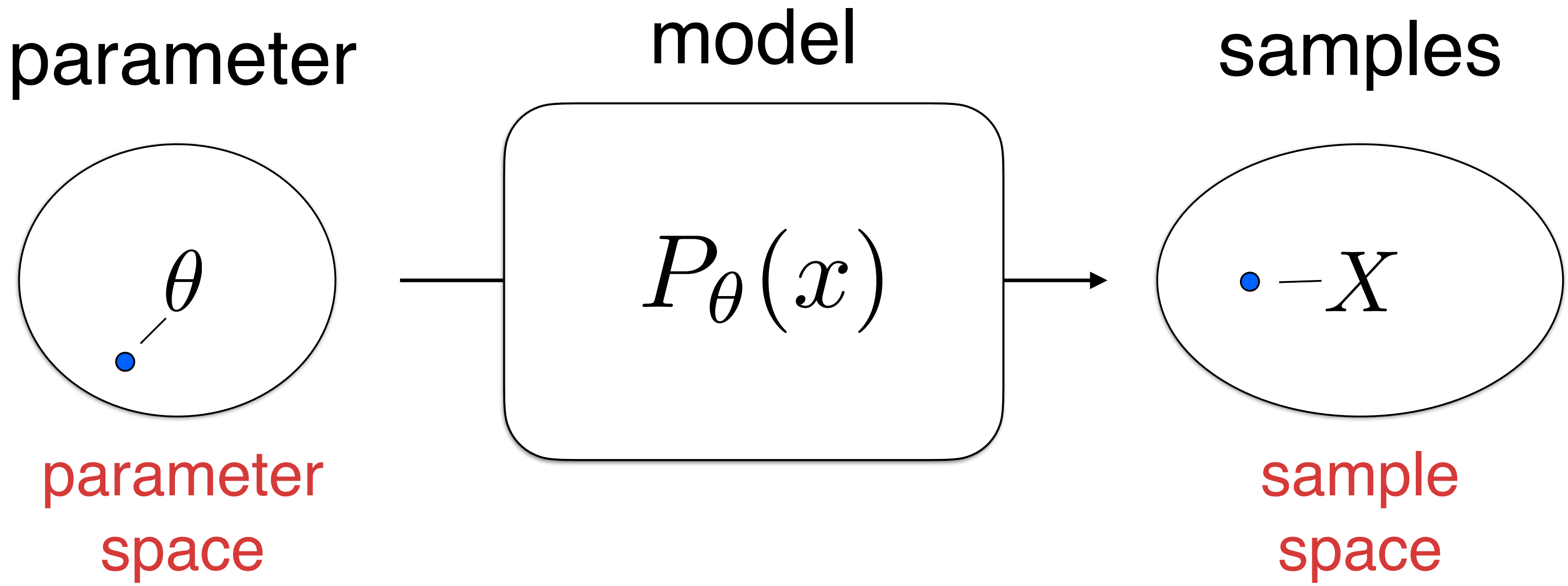
$$P(x|\theta)$$

or

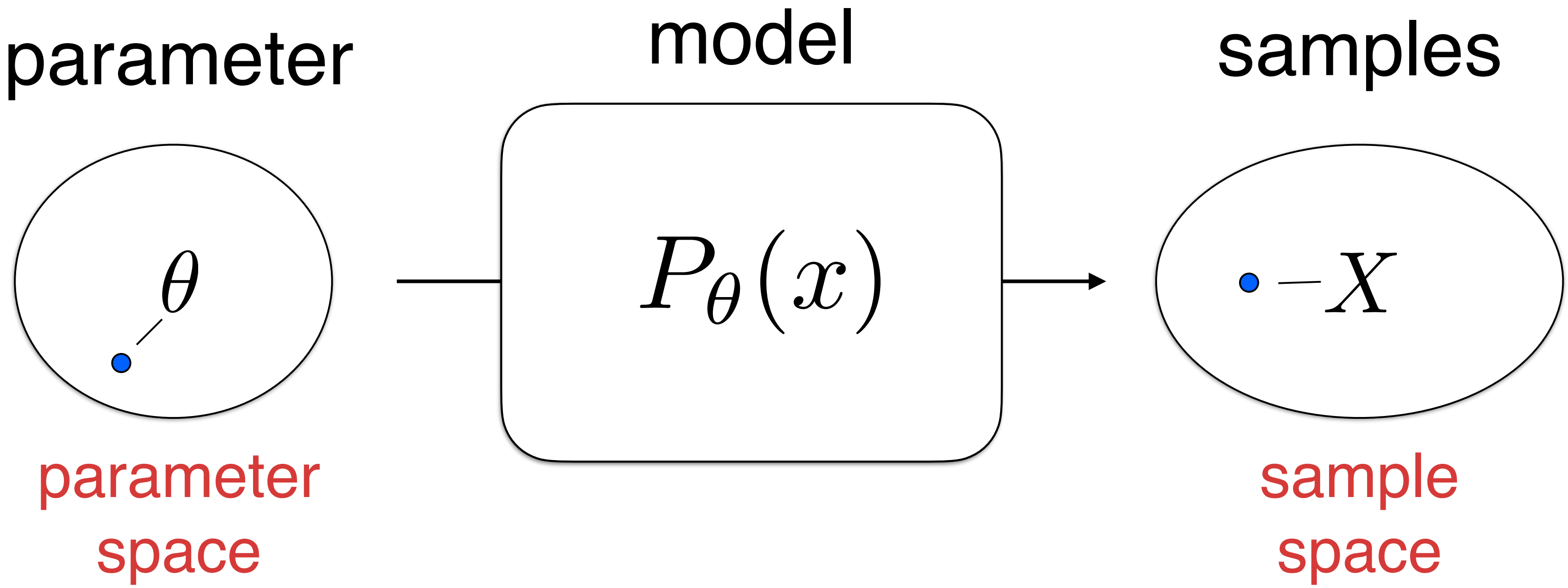
$$P(x;\theta)$$

- “probability distribution”

- “events”
- “random variables”



- “probability distribution”
- “events”
- “random variables”




<u>examples</u>

**1. coin flipping**

$$\theta = p(\text{“heads”})$$

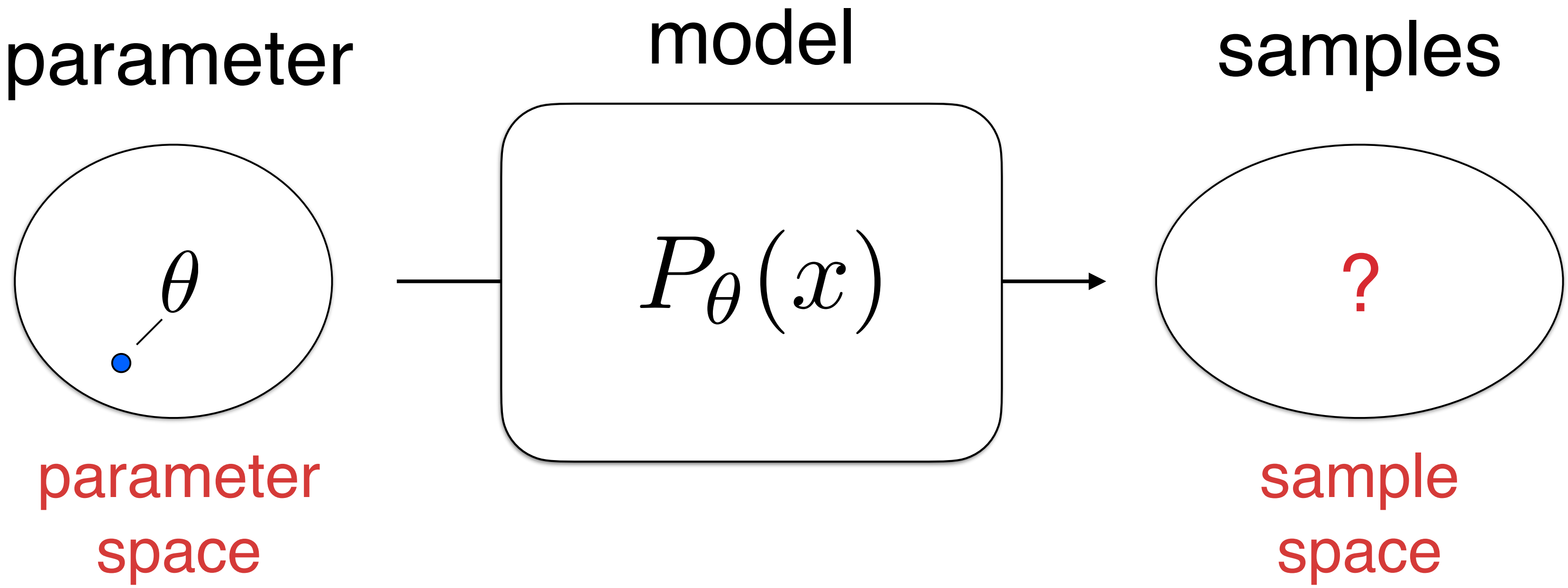
X = “H” or “T”

**2. spike counts**

$\theta =$ mean spike rate

$$X \in \{0, 1, \ldots \ldots\}$$

# Probability vs. Statistics



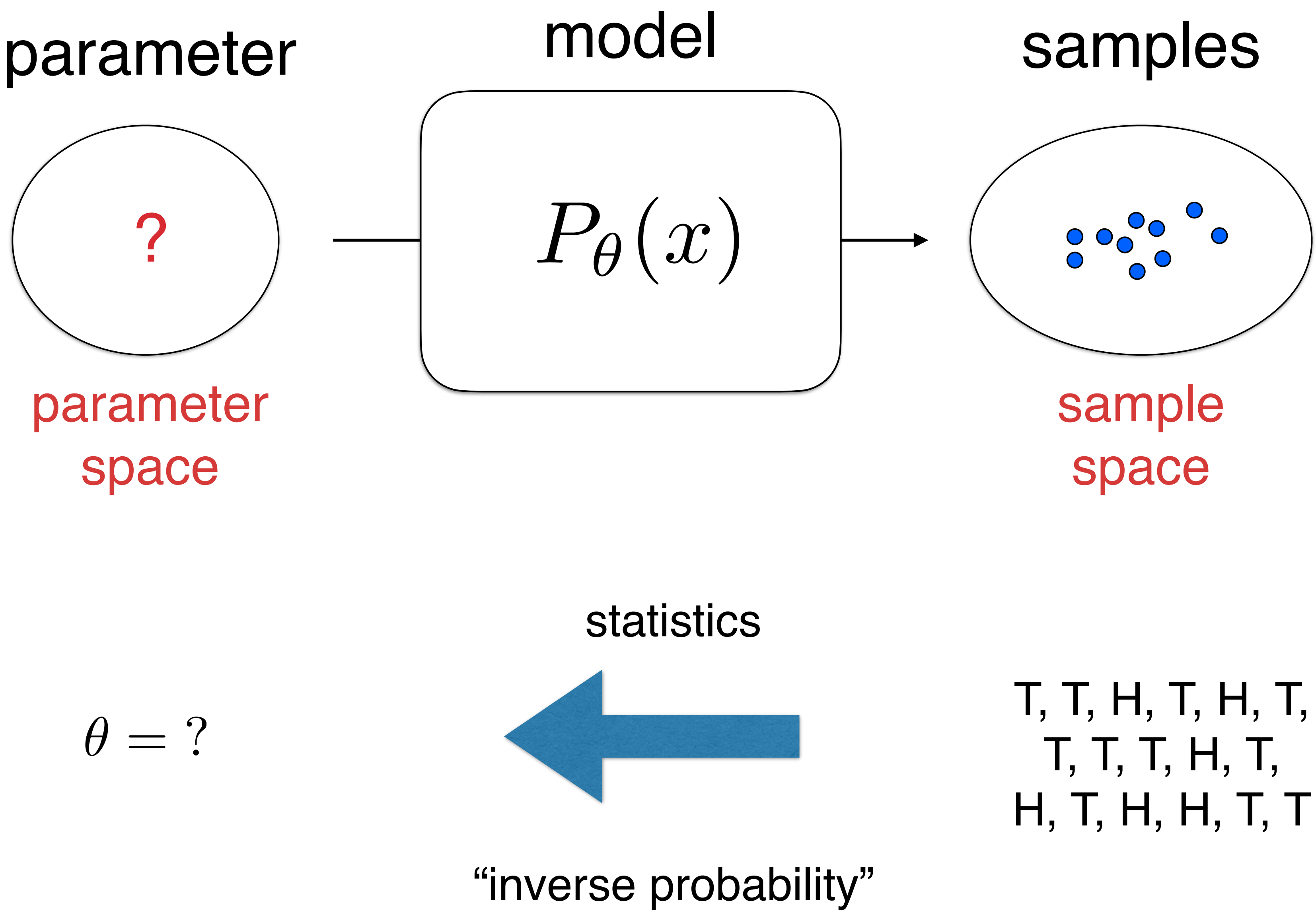

coin flipping

$\theta = 0.3$

probability

T, T, H, T, H, T,
T, T, T, ....

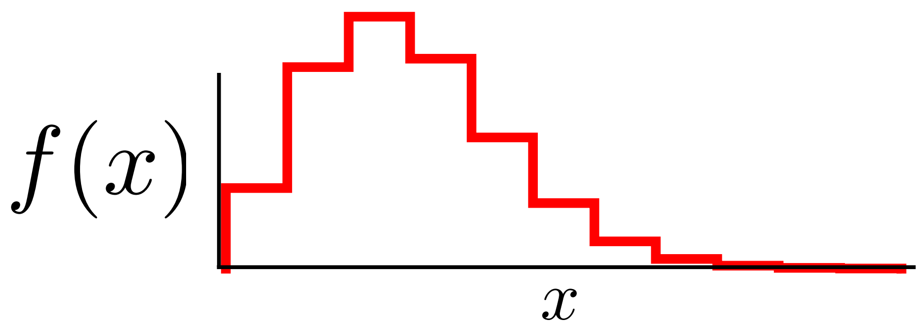
# Probability vs. Statistics



# discrete probability distribution

takes finite (or countably infinite) number of values, eg $x \in \mathbb{N}$

**probability mass function (pmf):**



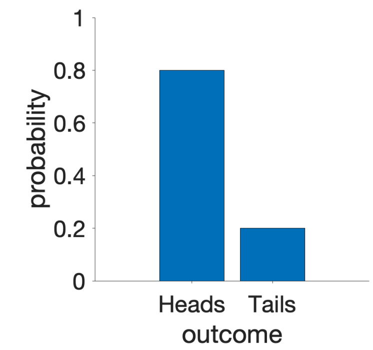

- $f(x_i) \geq 0$ for all i — non-negative
- $\sum_{i=1}^{N} f(x_i) = 1$ — sums to 1
- $P(x = a) = f(a)$ — gives probability of observing a particular value of x

# some friendly neighborhood probability distributions

## Discrete

**Bernoulli**
(coin flipping)

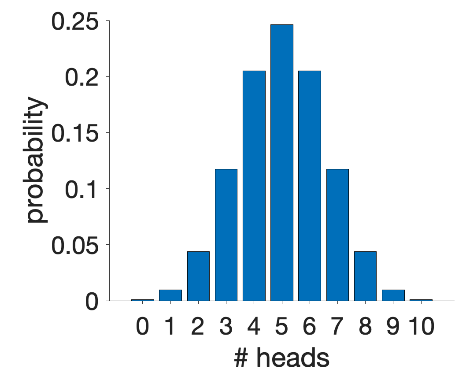
$$P(x|p) = p^x \cdot (1-p)^{(1-x)}$$




**binomial**

$$P(k|\,n,p) = \binom{n}{k} p^k (1-p)^{n-k}$$

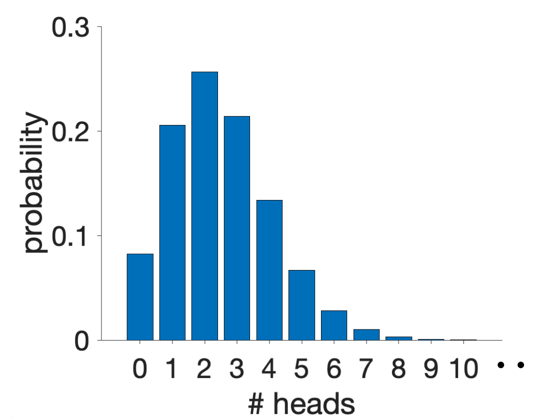
(sum of n coin flips)




**Poisson**

$$P(k|\,\lambda) = \frac{\lambda^k}{k!} e^{-\lambda}$$

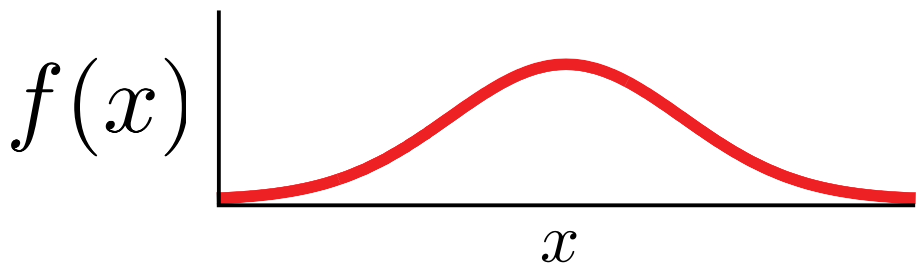
(sum of n coin flips with
P(heads)=λ/n, in limit n→∞)



# continuous probability distribution

takes values in a continuous space, e.g., $x \in \mathbb{R}$

**probability density function (pdf)**:




- $f(x) \geq 0$ for all x — non-negative
- $\displaystyle\int_{-\infty}^{\infty} f(x)\, dx = 1$ — integrates to 1

- $P(x = a) = 0$
- $P(a < x < b) = \displaystyle\int_a^b f(x)\, dx$

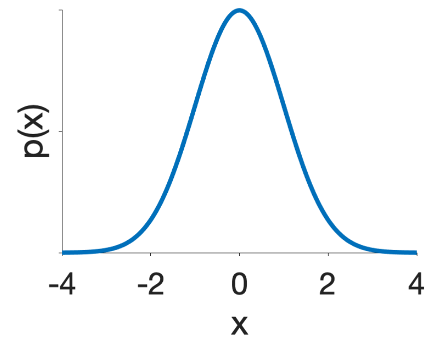
gives probability of x falling within some interval

# some friendly neighborhood probability distributions

## Continuous

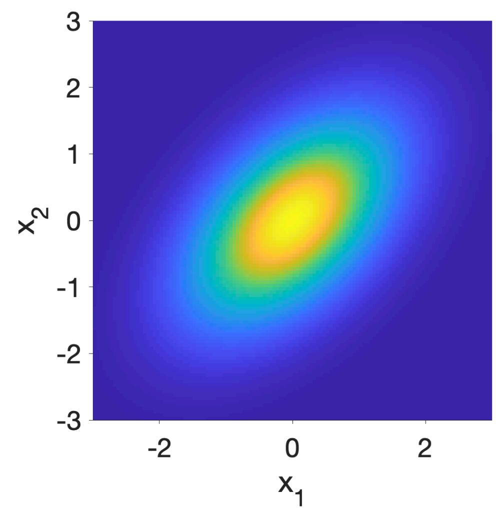
**Gaussian**

$$P(x|\mu, \sigma) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left[-\frac{(x-u)^2}{2\sigma^2}\right]$$



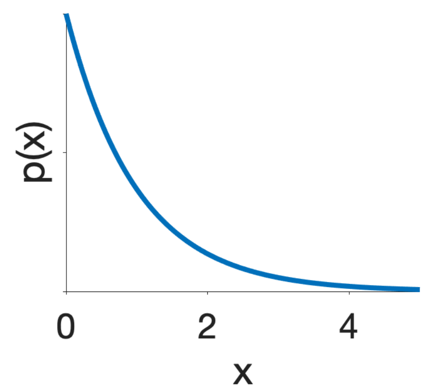

**multivariate Gaussian**

$$P(\mathbf{x} \,|\, \mu, \Lambda) = \frac{1}{(2\pi)^{\frac{n}{2}} \, |\Lambda|^{\frac{1}{2}}} \exp\left[-\tfrac{1}{2}(\mathbf{x}-\mu)^T \Lambda^{-1} (\mathbf{x}-\mu)\right]$$



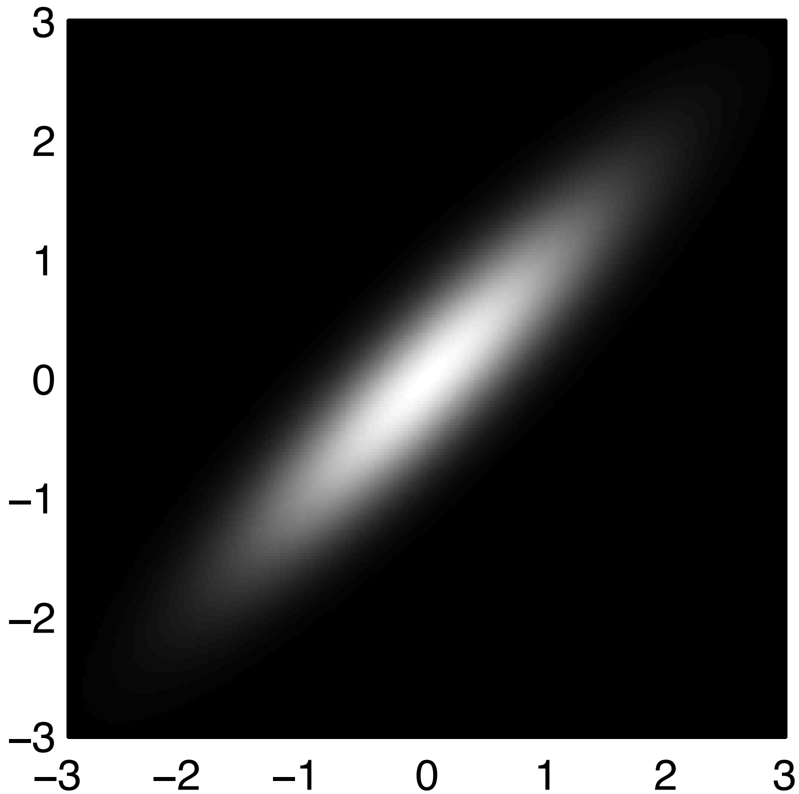

**exponential**

$$P(x|a) = ae^{-ax}$$



# joint distribution

$$P(x, y)$$




$y$

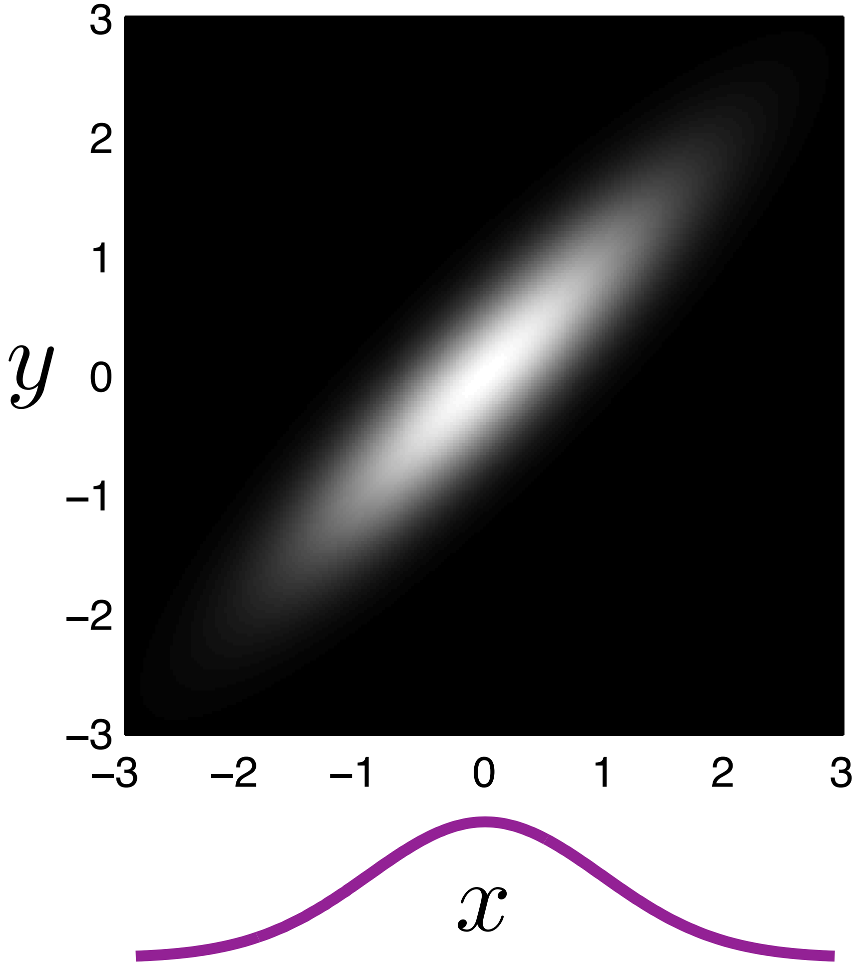
$x$

- positive
- sums to 1

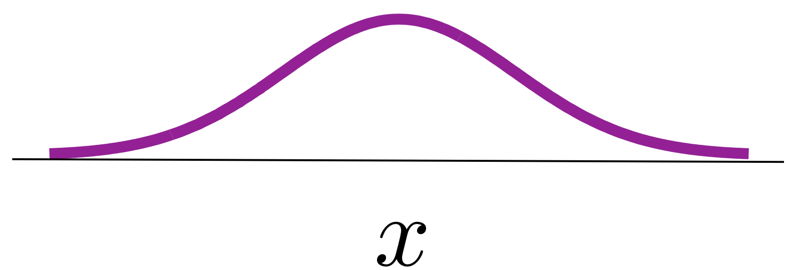
$$\iint P(x, y)\, dx\, dy = 1$$

# marginalization ("integration")

$P(x, y)$

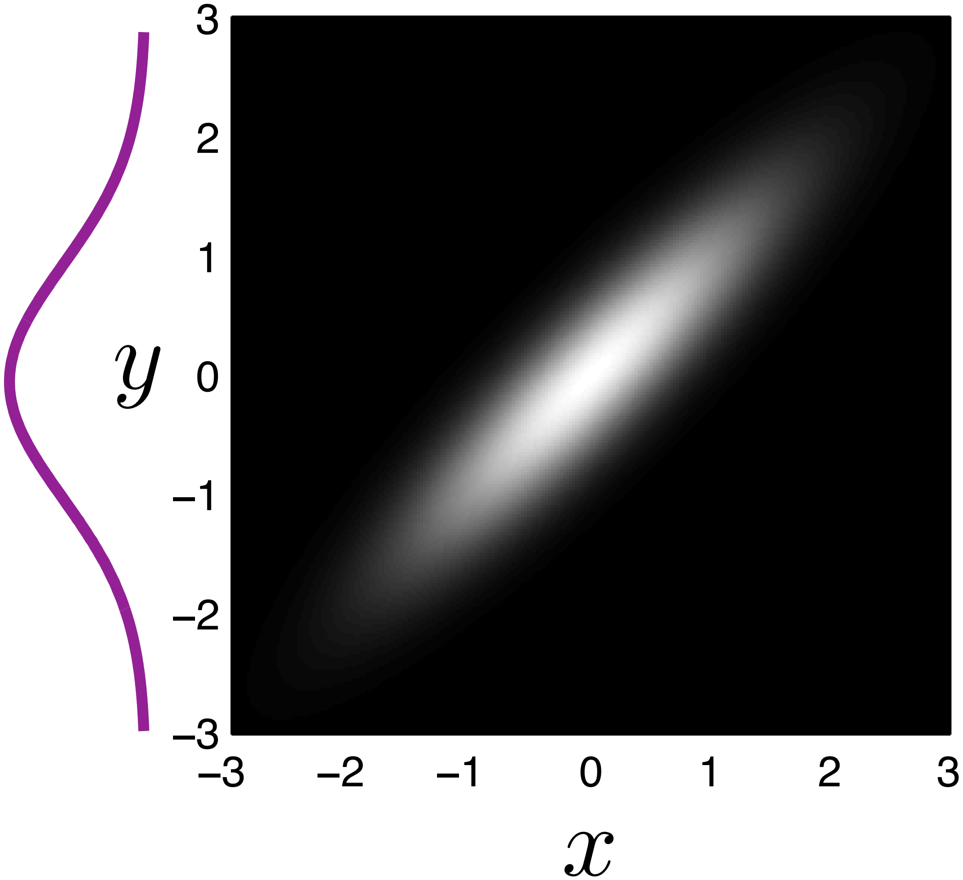



$$P(x) = \int P(x, y)dy$$

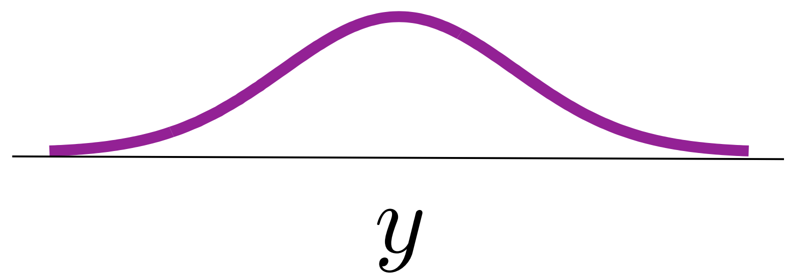


# marginalization (“integration”)

$P(x, y)$

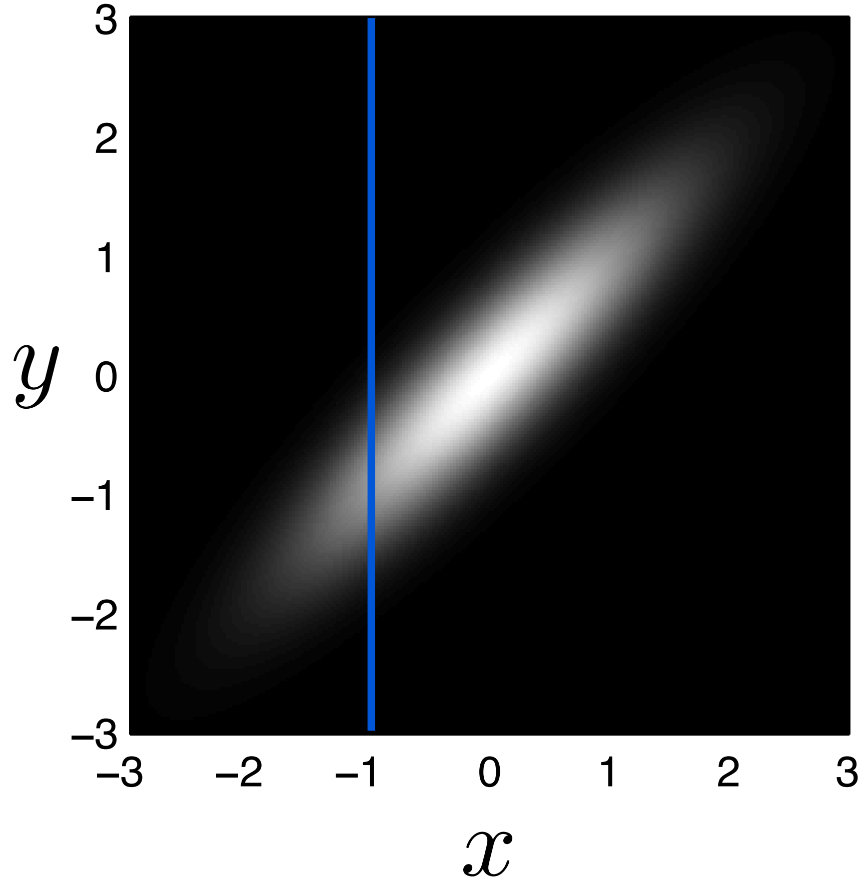



$$P(y) = \int P(x, y)dx$$



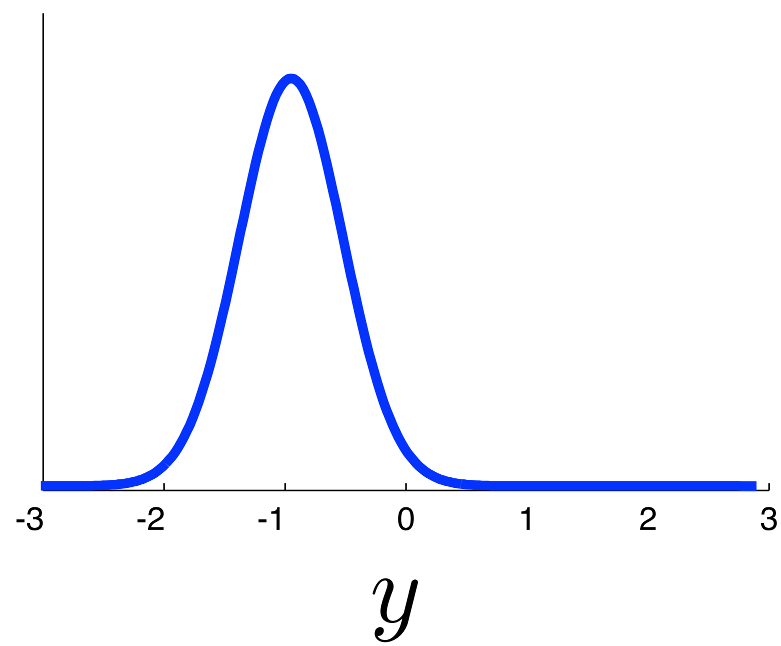
# conditionalization ("slicing")

$$P(x, y)$$



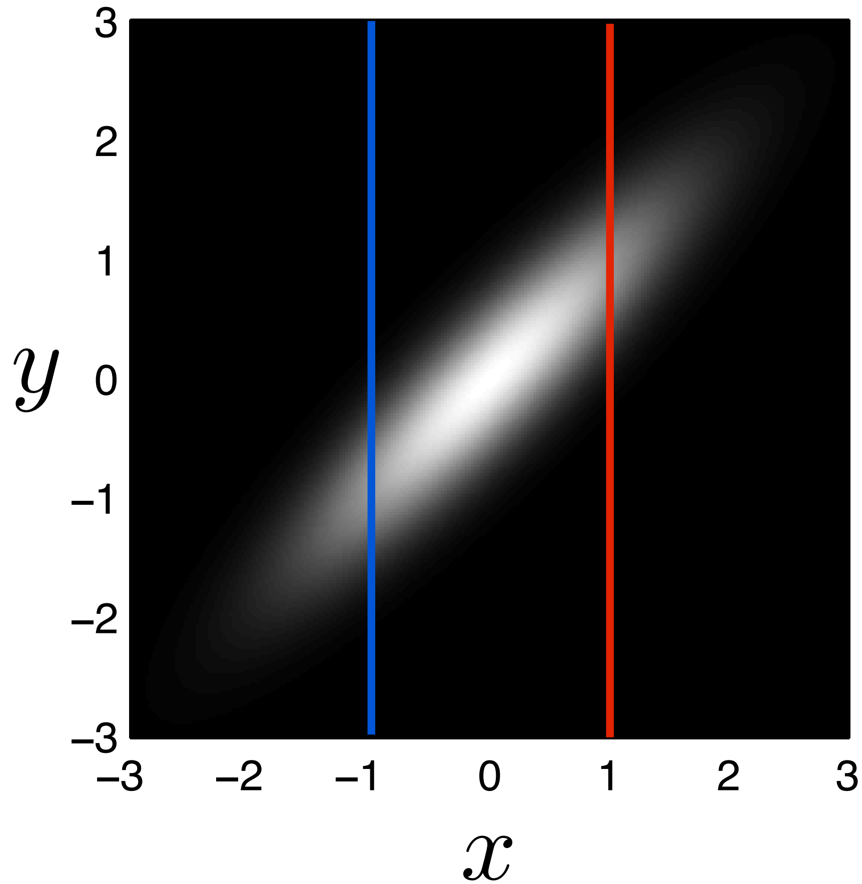

$$P(y|x = -1) = \frac{P(y, x = -1)}{P(x = -1)}$$

("joint divided by marginal")



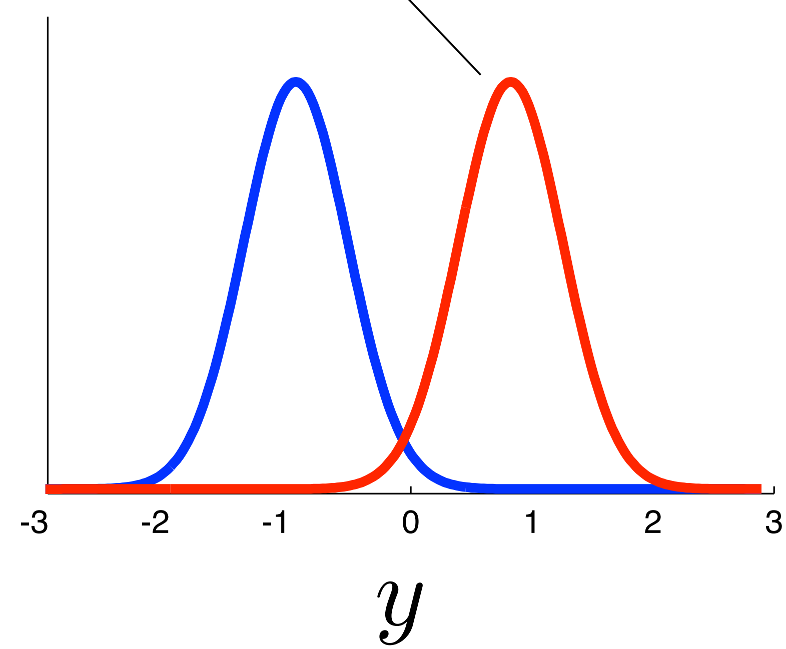
# conditionalization ("slicing")

$$P(x, y)$$



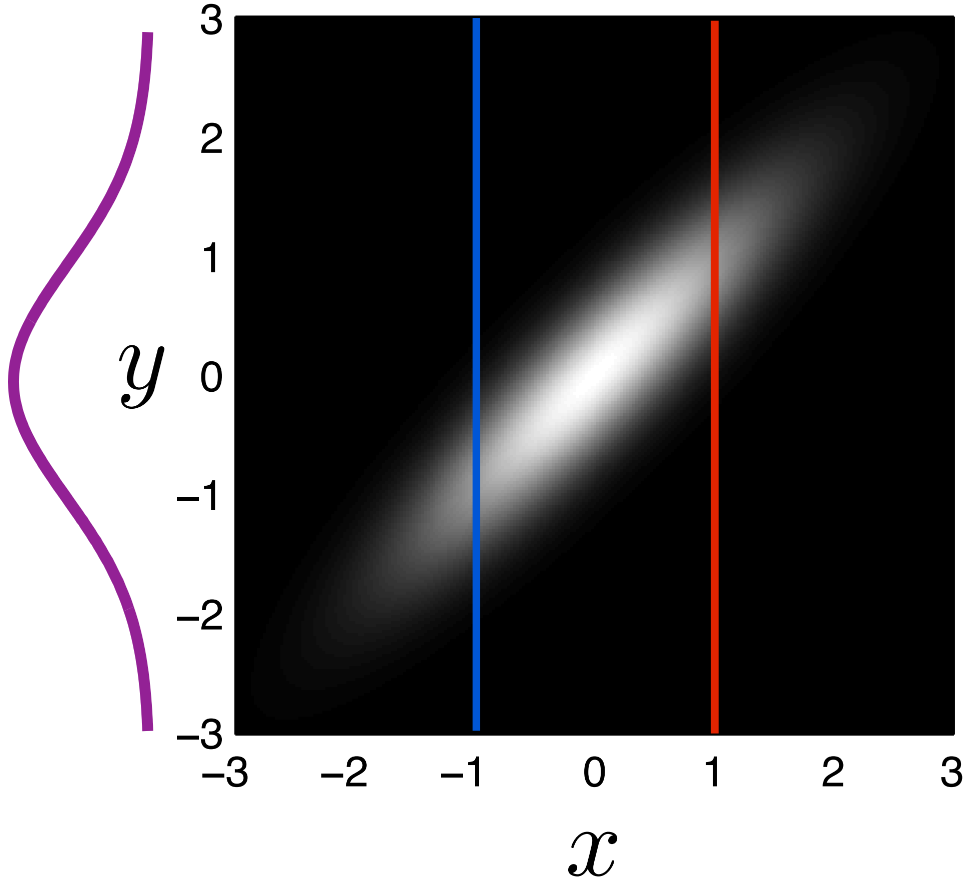

$$P(y|x=1) = \frac{P(y, x=1)}{P(x=1)}$$

("joint divided by marginal")



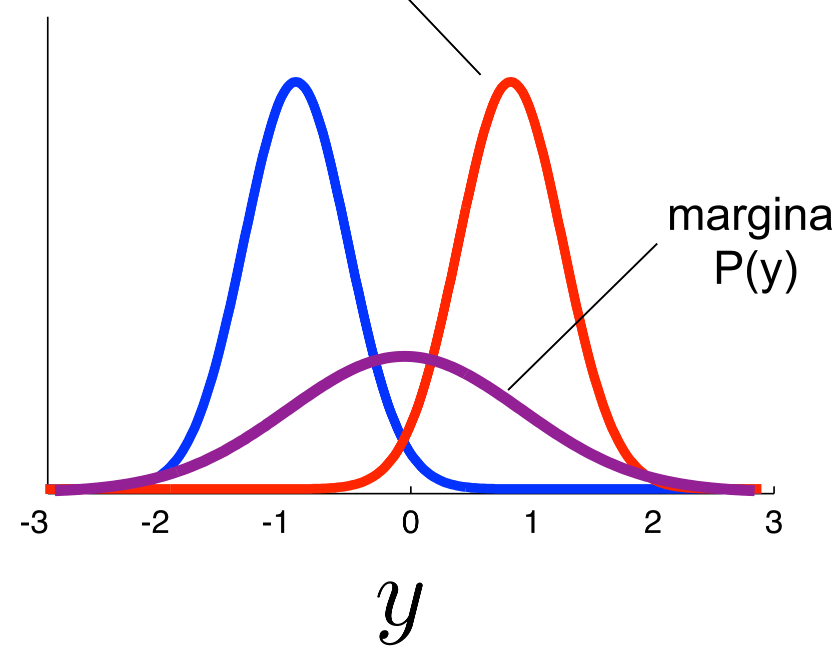
# conditionalization ("slicing")

$$P(x, y)$$



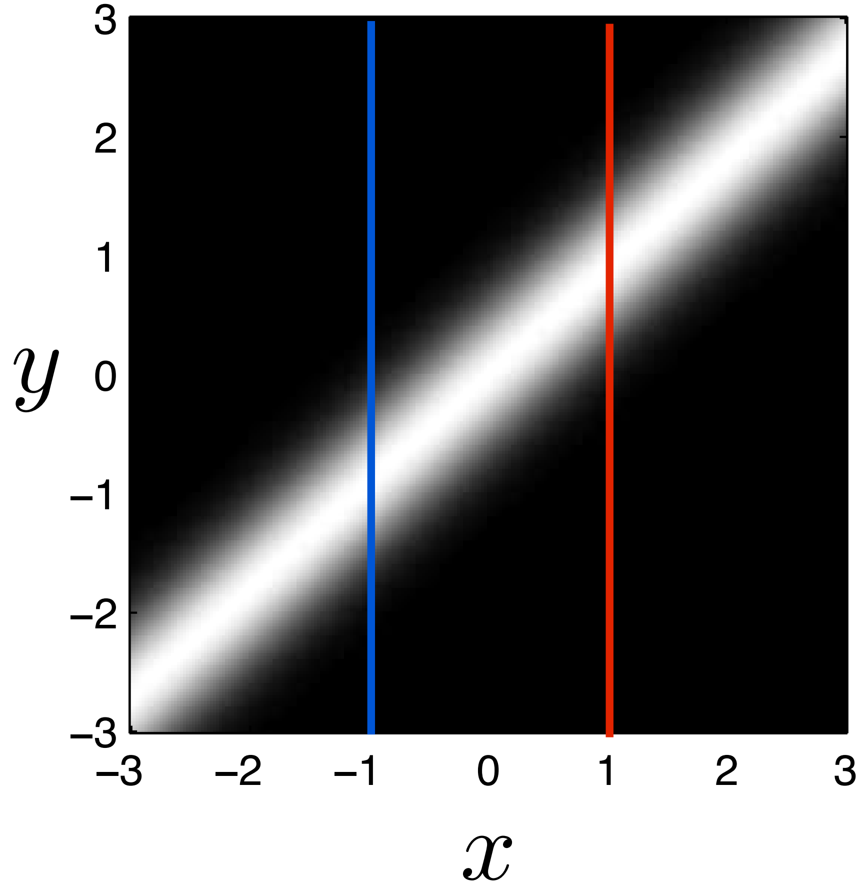

conditional

$$P(y|x=1) = \frac{P(y, x=1)}{P(x=1)}$$

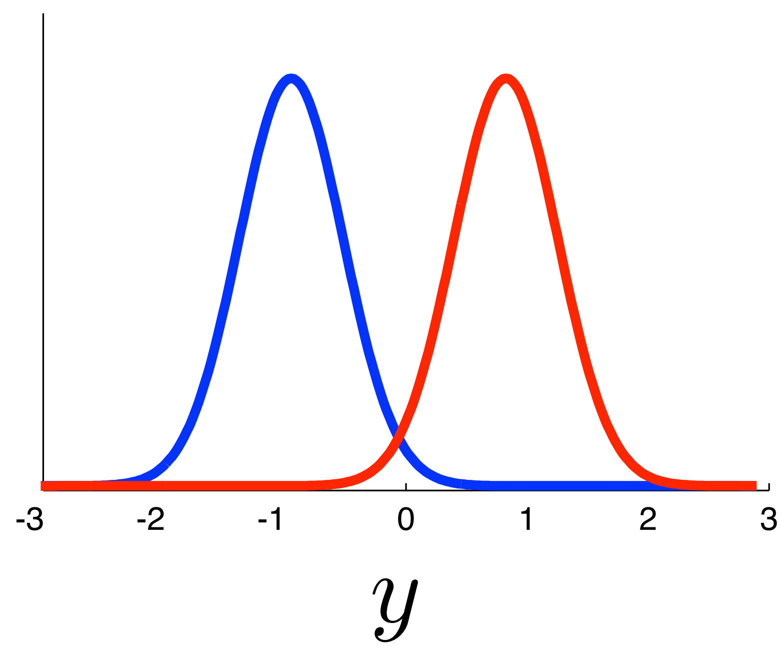


# conditional densities

$$P(y|x)$$

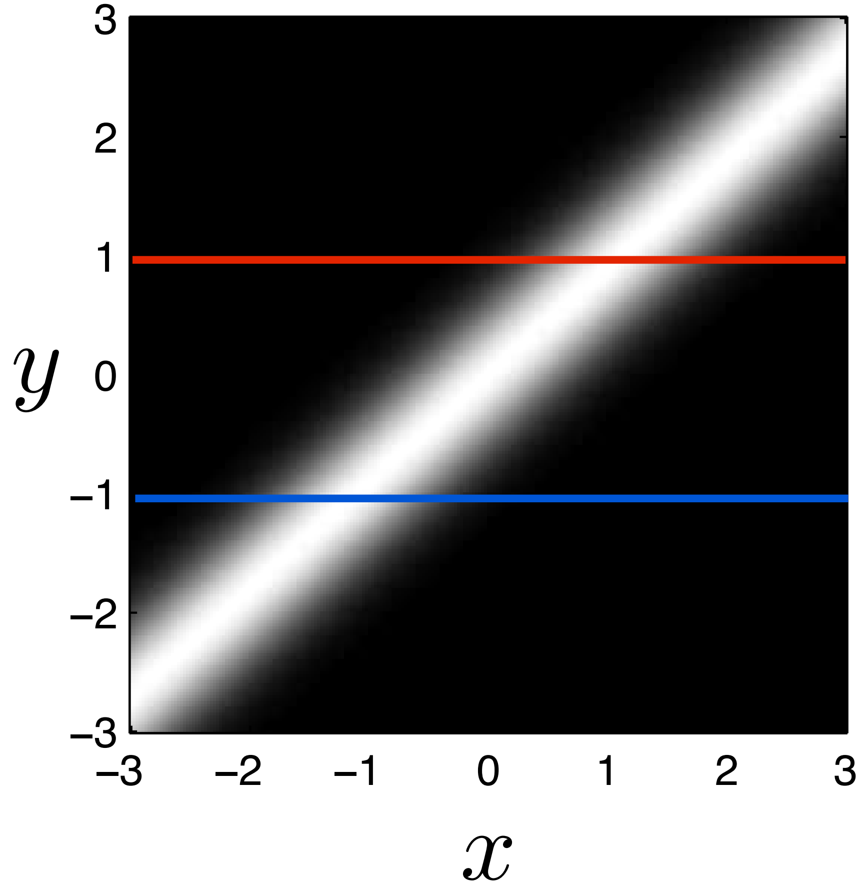



$$P(y|x) = \frac{P(x,y)}{P(x)}$$

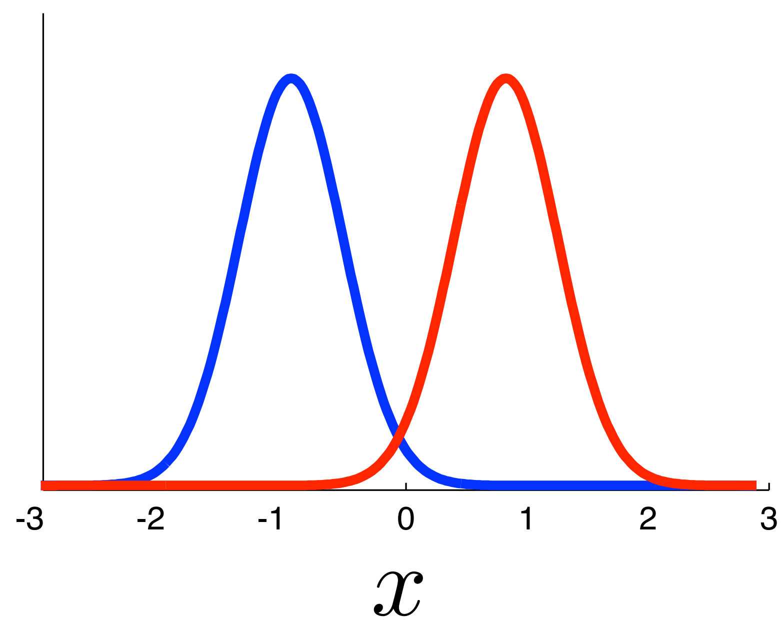


# conditional densities

$$P(x|y)$$




$$P(x|y) = \frac{P(x,y)}{P(y)}$$



# Bayes' Rule

**Conditional Densities**

$$P(y|x) = \frac{P(x,y)}{P(x)} \qquad P(x|y) = \frac{P(x,y)}{P(y)}$$

$$P(x,y) = P(y|x)P(x) = P(x|y)P(y)$$

rearranging gives:

**Bayes' Rule**

$$P(x|y) = \frac{P(y|x)P(x)}{P(y)}$$

# Expectations ("averages")

Expectation is the weighted average of a function (of a random variable) according to the distribution (of that random variable)

**discrete**

pmf

$$\mathbb{E}[f(x)] = \sum_i f(x_i) P(x_i)$$

or

**continuous**

pdf

$$\mathbb{E}[f(x)] = \int f(x) P(x) dx$$

Corresponds to taking weighted average of f(X), weighted by how probable they are under P(x).

# Expectations ("averages")

Expectation is the weighted average of a function (of a random variable) according to the distribution (of that random variable)

**discrete** or **continuous**

pmf

$$\mathbb{E}[f(x)] = \sum_i f(x_i) P(x_i)$$

pdf

$$\mathbb{E}[f(x)] = \int f(x) P(x) dx$$

It's really just a dot product! $\mathbb{E}[f(x)] = \vec{P} \cdot \vec{f}$ $\quad \vec{P} = \begin{bmatrix} P(x_1) \\ \vdots \\ P(x_m) \end{bmatrix} \quad \vec{f} = \begin{bmatrix} f(x_1) \\ \vdots \\ f(x_m) \end{bmatrix}$

# Several important expectations:

1) Mean: $\mathbb{E}[x]$ - the average value of a random variable

"1st moment" (here we have simply f(x) = x)

if x is discrete, taking on N values: $$\mathbb{E}[x] = \sum_{i=1}^{N} x_i P(x_i)$$

example

$$\begin{array}{cc} x & P(x) \\ \begin{bmatrix} 1 \\ 2 \\ 3 \end{bmatrix} & \begin{bmatrix} 0.5 \\ 0.3 \\ 0.2 \end{bmatrix} \end{array}$$

$$\mathbb{E}[x] = \vec{x} \cdot \vec{P}$$

$$= 1(0.5) + 2(0.3) + 3(0.2) = 1.7$$

# Several important expectations:

1) Mean: $\mathbb{E}[x]$ - the average value of a random variable

"1st moment" (here we have simply f(x) = x)

if x is continuous: $$\mathbb{E}[x] = \int x P(x)\, dx$$

• can still think of this as a dot product between two (infinitely tall) vectors of x values and probilities

$$\mathbb{E}[x] = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ \\ \end{bmatrix} \bullet \begin{bmatrix} P(x_1) \\ P(x_2) \\ \vdots \\ \\ \end{bmatrix}$$

# Several important expectations:

2) $\mathbb{E}[x^2]$ - the average value of squared random variable

"2nd moment" (here $f(x) = x^2$)

if x is discrete, taking on N values: $$\mathbb{E}[x^2] = \sum_{i=1}^{N} x_i^2 P(x_i)$$

example

$$\begin{array}{ccc} x & x^2 & P(x) \\ \begin{bmatrix} 1 \\ 2 \\ 3 \end{bmatrix} & \begin{bmatrix} 1 \\ 4 \\ 9 \end{bmatrix} & \begin{bmatrix} 0.5 \\ 0.3 \\ 0.2 \end{bmatrix} \end{array}$$

$$\mathbb{E}[x^2] = \overrightarrow{x^2} \cdot \overrightarrow{P}$$

$$= 1(0.5) + 4(0.3) + 9(0.2) = 3.5$$

# Several important expectations:

3) variance: $\mathbb{E}[(x - \mathbb{E}[x])^2]$ (average squared difference between x and its mean)

mean $\mathbb{E}[x]$

if x is discrete: $$\mathrm{var}(x) = \sum_{i=1}^{N} (x_i - \mu)^2 P(x_i)$$

if x is continuous: $$\mathrm{var}(x) = \int (x - \mu)^2 P(x)\, dx$$

# Note: expectations don't always exist!

e.g.

the Cauchy distribution: $$P(x) = \frac{1}{\pi(1+x^2)}$$ has no mean!

$$\mathbb{E}[x] = \int_{-\infty}^{\infty} xP(x) = \boxed{\int_{-\infty}^{\infty} \frac{x}{\pi(1+x)^2} dx}$$

undefined / does not exist

# Summary

- basics of probability
- probability vs. statistics
- continuous & discrete distributions
- joint distributions
- marginalization (splatting)
- conditionalization (slicing)
- Bayes’ rule (for relating conditionals)
- expectations & moments