RICE CRUMB #3 – v2.0

- How can we write the distribution of the parameters estimated by MLE for $N \rightarrow +\infty$?

… keep the central limit theorem in mind …

Central limit theorem

Let's consider n independent random variables $x_1, \dots x_n$
with mean and variances μ_i and σ_i^2

Let's consider a new random variable

$$z = \sum_{i=1}^{n} x_i$$

z has $\longrightarrow$ mean $\mu = \sum_{i=1}^{n} \mu_i$

$\searrow$ variance $\sigma^2 = \sum_{i=1}^{n} \sigma_i^2$

Central limit theorem

CENTRAL LIMIT THEOREM:

$$\mathrm{pdf}\left(q = \frac{z-\mu}{\sigma}\right) \xrightarrow{n \to +\infty} \mathcal{N}(0,1)$$

↳ CONSEQUENCE:

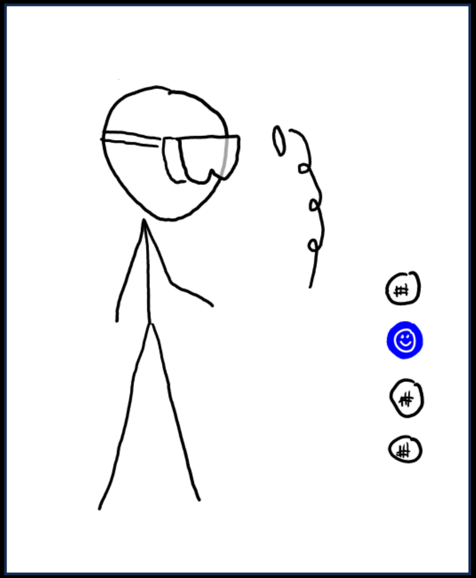
for large enough n, we can consider z to be approximately Gaussian with mean μ and variance σ^2

RICE CRUMB #3 – v2.0

- How can we write the distribution of the parameters estimated by MLE for N→+∞ ?

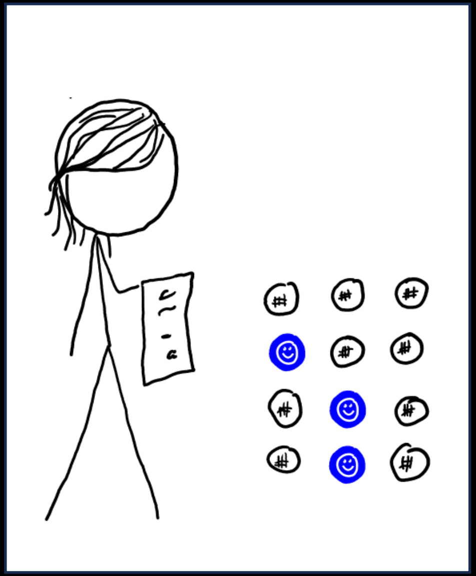
$$\underline{\hat{\theta}}_{ML} \sim \mathcal{N}(\underline{\hat{\theta}}_0, \underline{\underline{\Sigma}}_{\underline{\hat{\theta}}})$$

The ML estimate is related to the sum of random variables



I tossed four coins
and got one head.

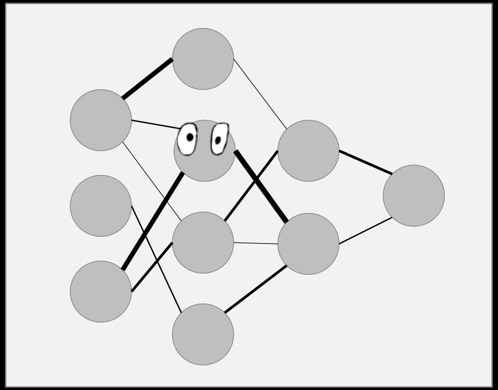
Using a $\text{Bin}(n, p)$ model, I estimate
the probability of heads is $\hat{p} = 25\%$



I tossed twelve coins
and got three heads.

Using a $\text{Bin}(n, p)$ model, I estimate
the probability of heads is $\hat{p} = 25\%$

But surely, the more data we have, the more confident we should be!



"This is a 40mph speed limit, with probability 98%."

Neural networks tell us *probabilities*, but they don't tell us their *confidence.*

No one has worked out how to extract confidences from neural networks. But, in Bayesian statistics, we do know how to …

Bayes's rule

Data from a population sample of 100,000 people:

	test +ve	test -ve	total
got COVID	376	24	400
not got COVID	996	98,604	99,600

Let's rewrite this data as a probability model:

Let $X = 1_{\text{have COVID}}$ and let $Y = 1_{\text{test+ve}}$

$X \sim \text{Bin}(1, 0.004)$ 400 / 100 000 = 0.004
if X == 1:
    $Y \sim \text{Bin}(1, 0.94)$ 376 / 400 = 0.94
else:
    $Y \sim \text{Bin}(1, 0.01)$ 996 / 99 600 = 0.01

What are these probabilities?

- $\mathbb{P}(\text{have COVID} \mid \text{test} +\text{ve})$
- $\mathbb{P}(\text{have COVID} \mid \text{test} -\text{ve})$

$$\mathbb{P}(X = 1 \mid Y = 1)$$

$$= \frac{\mathbb{P}(X = 1)\, \mathbb{P}(Y = 1 | X = 1)}{\mathbb{P}(Y = 1)}$$

$$= \frac{0.004 \times 0.94}{0.004 \times 0.94 + 0.996 \times 0.01}$$

Credits: IEEE GRSS Data Fusion contest 2018

Thomas Bayes (1701-1761)

Bayes's rule

For two **discrete** random variables X and Y,

$$\mathbb{P}(X = x|Y = y) = \frac{\mathbb{P}(X = x)\mathbb{P}(Y = y|X = x)}{\mathbb{P}(Y = y)} \quad \text{when } \mathbb{P}(Y = y) > 0$$

For two **discrete or continuous** random variables X and Y,

$$\Pr_X(x|Y = y) = \frac{\Pr_X(x)\Pr_Y(y|X = x)}{\Pr_Y(y)} \quad \text{when } \Pr_Y(Y) > 0$$

Recap

Marginal Probability

- It is the probability of an event irrespective of any other factor/event/circumstance. Basically, you ‘marginalize’ other events and hence the name. It is denoted by P(A) and read as “probability of A”.

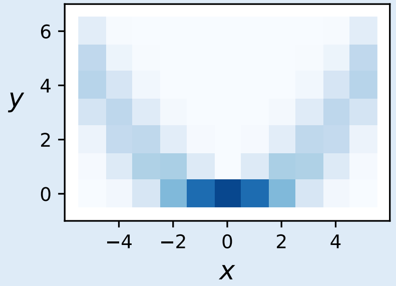
Conditional Probability

- Conditional probability is when the occurrence of an event is wholly or partially affected by other event(s). Alternatively put, it is the probability of occurrence of an event A when an another event B has already taken place. It is denoted by P(A|B) and read as “probability of A given B”.

Joint Probability

- Joint probability is calculated when we are interested in the occurrence of two different events simultaneously. It is also often referenced as probability of intersection of two events. It is denoted by P(A, B) and read as “probability of A and B”.

Joint distribution

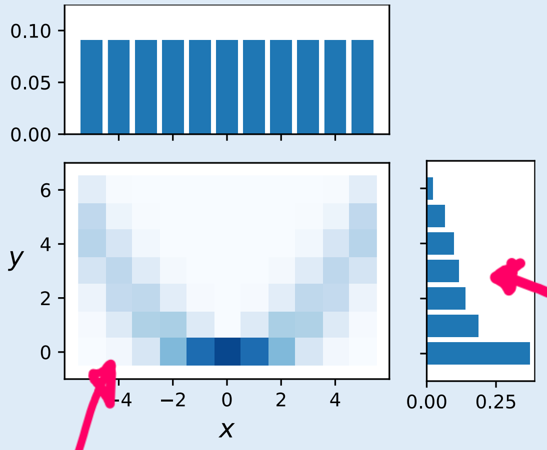



```
def rxy():
    x = np.random.randint(low=-5, high=6)  # from -5 to +5 inclusive
    y = np.random.binomial(n=6, p=(x/6)**2)
    return (x,y)
```

The joint pmf of (X,Y)
$\Pr_{X,Y}(x,y) = \mathbb{P}(X = x, Y = y)$ $=$ $\underbrace{\mathbb{P}(X=x)\,\mathbb{P}(Y=y \mid X=x)}^{\text{defn. of cond. prob.}}$ $= \frac{1}{11} \times \binom{6}{y} \left[\left(\frac{x}{6}\right)^2\right]^y \left[1-\left(\frac{x}{6}\right)^2\right]^{6-y}$

Code to plot the joint pmf
```
xy_samp = [rxy() for _ in range(1000)]
plt.hist2d(xy_samp)
```

Marginal random variables




```
def rxy():
    x = np.random.randint(low=-5, high=6)  # from -5 to +5 inclusive
    y = np.random.binomial(n=6, p=(x/6)**2)
    return (x,y)
```

The joint pmf of (X, Y)

$\Pr_{X,Y}(x,y) = \mathbb{P}(X = x, Y = y)$

The marginal of Y

$$\begin{aligned}\Pr_Y(y) &= \mathbb{P}(Y = y) \\ &= \sum_x \mathbb{P}(X = x, Y = y) \quad \text{by the Sum Rule} \\ &= \sum_x \Pr_{X,Y}(x,y)\end{aligned}$$

Code to plot the joint pmf

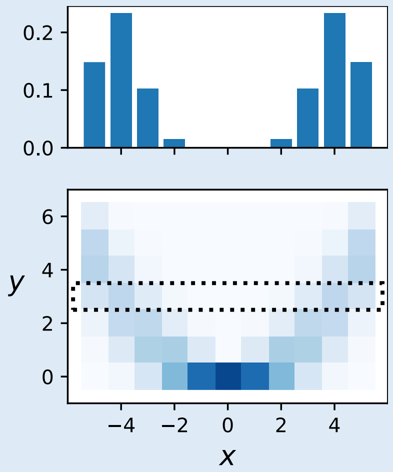
```
xy_samp = [rxy() for _ in range(1000)]
plt.hist2d(xy_samp)
```

Code to plot the marginal pmf

```
y_samp = [y for (x,y) in xy_samp]
plt.hist(y_samp)
```

← i.e. just throw away the x values

Conditional random variables




```
def rxy():
    x = np.random.randint(low=-5, high=6)  # from -5 to +5 inclusive
    y = np.random.binomial(n=6, p=(x/6)**2)
    return (x,y)
```

X conditional on $Y=3$

$$\mathbb{P}(X=x|Y=3)=\frac{\mathbb{P}(X=x,Y=3)}{\mathbb{P}(Y=3)}=\frac{\Pr_{X,Y}(x,3)}{\Pr_Y(3)}$$

$\mathrm{pmf}_3(x)$ //

i.e. take the Y=3 row, then rescale it to sum to 1

We can think of "X conditional on $Y=3$" as a random variable …

We've provided a valid probability mass function:

$\mathrm{pmf}_3(\cdot)\geq 0$ $\quad\sum_x \mathrm{pmf}_3(x)=1$

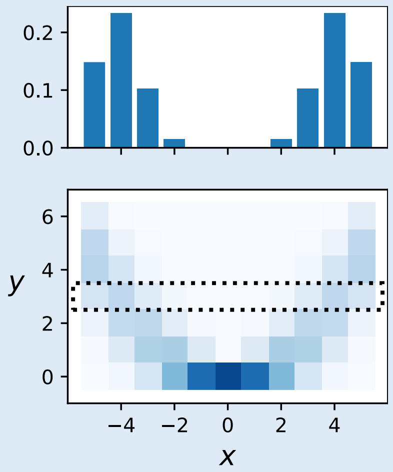
Sample space: $\Omega=\{-5,-4,\ldots,4,5\}$ same as for x.

Code to generate values from it:

```
def rx_given_y():
    while True:
        x,y = rxy()
        if y == 3: break
    return x
```

```
def rx_given_y():
    Ω = {-5,...,5}
    p = [pmf(x) for x in Ω]
    return np.random.choice(Ω, p=p)
```

Conditional random variables




```
def rxy():
    x = np.random.randint(low=-5, high=6)   # from -5 to +5 inclusive
    y = np.random.binomial(n=6, p=(x/6)**2)
    return (x,y)
```

$\Theta \sim U[0,1]$

$X \sim Bin(4, \Theta)$

We *define* the **conditional random variable**, written $(X|Y=y)$, by specifying its likelihood:

$$\boxed{\Pr_{(X|Y=y)}(x)} = \frac{\Pr_{X,Y}(x,y)}{\Pr_Y(y)}$$

Taking the $Y=y$ row from joint pmf, rescale it

commonly written $\Pr_X(x \mid Y=y)$

```
def rx_given_y():
    Ω = {-5,...,5}
    p = [pmf(x) for x in Ω]
    return np.random.choice(Ω, p=p)
```

Given a problem with M classes
$y_1 \ldots y_M$

we want to classify an unknown RV X
into the most likely / probable class

Consider $P(y_i | X)$ for each class $i = 1 \ldots M$
[A POSTERIORI PROBABILITIES]
that represent the unknown RV
to belong to class y_i given the value
that X assumes.

BAYES CLASSIFICATION RULES

Assign an unknown RV [FEATURE] X to the class y_i from M classes $y_1 \dots y_M$

s.t. $P(y_i | x) > P(y_j | x) \quad \forall j \neq i \in \{1 \dots M\}$

→ PROBLEM: how do we determine $P(y_i | x)$?

THINGS WE MIGHT KNOW (OR CAN DETERMINE) FROM TRAINING DATA

— a priori probabilities $P(\omega_i)$

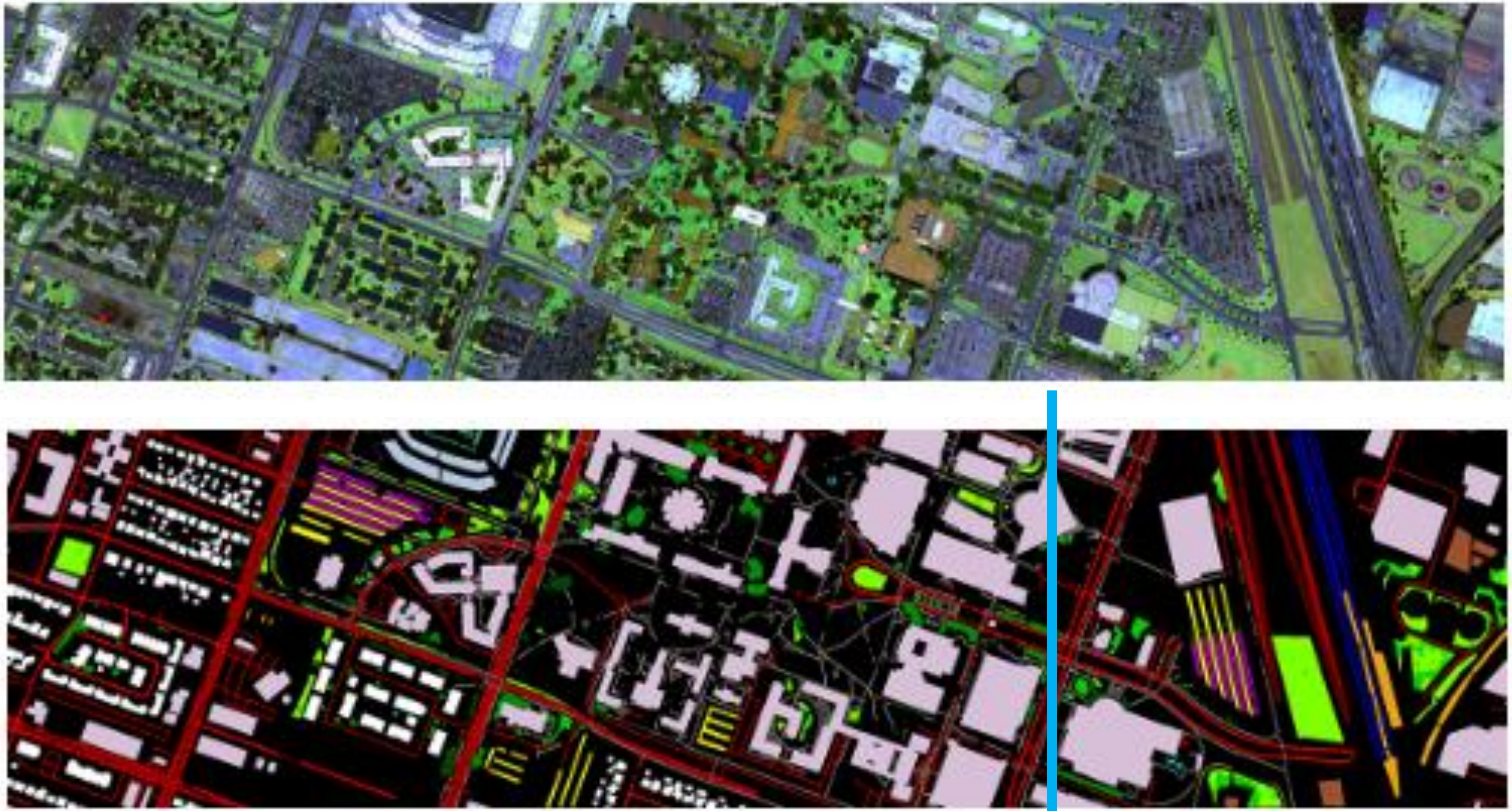
↳ either
- known
- assumed
- estimated (e.g. from training data proportions)

- class conditional probability density functions

$P(x|y_i)$ → describe distribution of x in each class

→ if unknown, estimate from data (TRAINING)

The quality of the analysis (and of the statistics) depends on the quality of the data and of the design choices we make



← Training | Test →

Why this is not a good choice?

Credits: IEEE GRSS Data Fusion contest 2018

BAYES RULE $$P(y_i \mid x) = \frac{P(x \mid y_i)\, P(y_i)}{P(x)}$$

$$P(x) = \sum_{i=1}^{m} P(x \mid y_i)\, P(y_i)$$

Recall: pdf and cdf for continuous random variables

Definition of continuous RV

Continuous random variable

A random variable X is continuous if there is a probability density function (PDF), $f(x) \geq 0$ such that for $-\infty < x < \infty$:

$$\mathbf{P}[a \leq X \leq b] = \int_a^b f(x)dx$$

To preserve the axioms that guarantee that $\mathbf{P}[a \leq X \leq b]$ is a probability, the following properties must hold:

$$0 \leq \mathbf{P}[a \leq X \leq b] \leq 1$$

$$\mathbf{P}[-\infty < X < \infty] = 1 \qquad \left(= \int_{-\infty}^{\infty} f(x)dx\right)$$

- Note: we also write $f(x)$ as $f_X(x)$.
- In continuous world, every RV has a PDF: its relative value wrt to other possible values.
- Integrate $f(x)$ to get probabilities.

Joint Distributions of Continuous Variables

Definition

Random variables X and Y have a joint continuous distribution if for some function $f : \mathbb{R}^2 \to \mathbb{R}$ and for all numbers $a_1 \leq b_1$ and $a_2 \leq b_2$,

$$\mathbf{P}[a_1 \leq X \leq b_1, a_2 \leq Y \leq b_2] = \int_{a_1}^{b_1} \int_{a_2}^{b_2} f(x,y)\,dx\,dy.$$

The function f has to satisfy $f(x,y) \geq 0$ for all x and y, and $\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f(x,y)dxdy = 1$. We call f the joint probability density.

As in one-dimensional case we switch from F to f by differentiating (or integrating):

$$F(a,b) = \int_{-\infty}^{a}\int_{-\infty}^{b} f(x,y)dx\,dy \qquad \text{and} \qquad f(x,y) = \frac{\partial^2}{\partial x \partial y} F(x,y)$$

For a continuous random variable X

$$\mathbb{P}(x_1 \leq X \leq x_2) = \int_{x=x_1}^{x_2} \Pr_X(x)\,dx$$

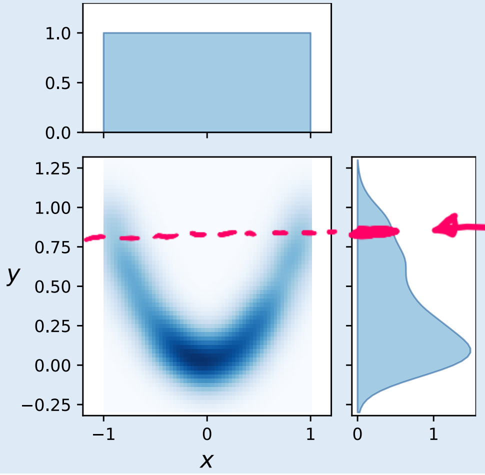
$$\Pr_X(x) = \frac{d}{dx}\mathbb{P}(X \leq x)$$

For a pair of continuous random variable X and Y

$$\mathbb{P}(x_1 \leq X \leq x_2 \text{ and } y_1 \leq Y \leq y_2) = \int_{x=x_1}^{x_2}\int_{y=y_1}^{y_2} \Pr_{X,Y}(x,y)\,dx\,dy$$

$$\Pr_{X,Y}(x,y) = \frac{\partial^2}{\partial x\,\partial y}\mathbb{P}(X \leq x \text{ and } Y \leq y)$$

Joint distribution and marginals (continuous case)




```
def rxy():
    x = np.random.uniform(-1,1)
    y = np.random.normal(loc=x**2, scale=0.1)
    return (x,y)
```

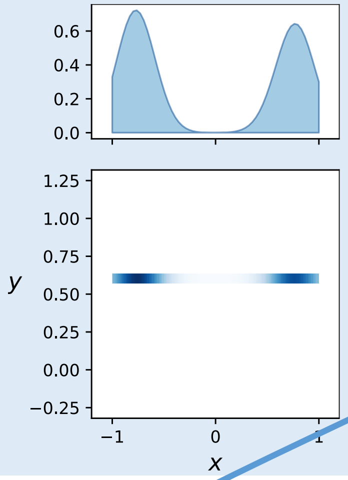
The joint pdf of (X, Y)

$\Pr_{X,Y}(x, y)$

The marginal of Y

$$\Pr_Y(y) = \int_x \Pr_{X,Y}(x, y)\ dx$$

Conditional random variables (continuous case)



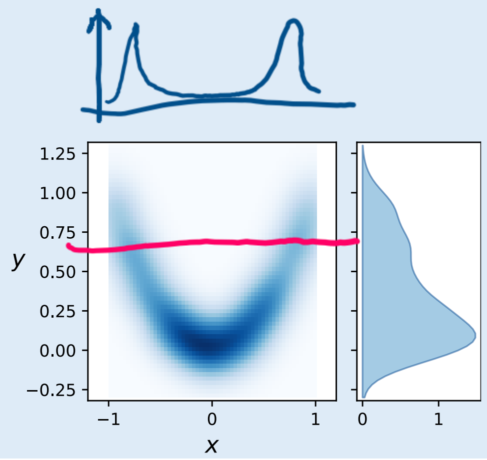

```
def rxy():
    x = np.random.uniform(-1,1)
    y = np.random.normal(loc=x**2, scale=0.1)
    return (x,y)
```

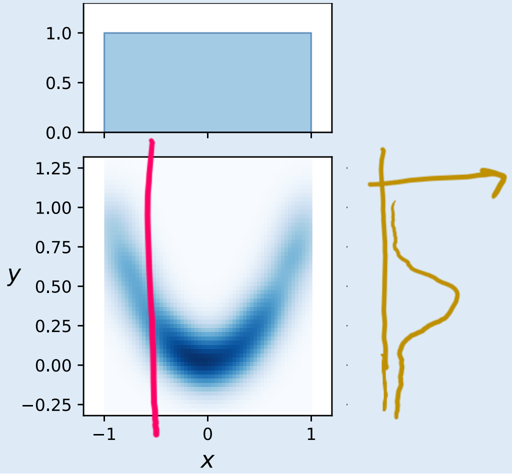
Take the $Y = 0.6$ slice of the joint pdf,
then rescale it so it integrates to 1
i.e. so we get a legitimate pdf.

We define the conditional random variable $(\boldsymbol{X}|\boldsymbol{Y}=\boldsymbol{y})$ by specifying its likelihood:

$$\Pr_X(x|Y=y) = \frac{\Pr_{X,Y}(x,y)}{\Pr_Y(y)}$$

Bayes's rule





$$\Pr_X(x \mid Y=y) = \frac{\Pr_{X,Y}(x,y)}{\Pr_Y(y)}$$

$$\Pr_Y(y \mid X=x) = \frac{\Pr_{X,Y}(x,y)}{\Pr_X(x)}$$

$$\Pr_X(x \mid Y=y) = \frac{\Pr_{X,Y}(x,y)}{\Pr_Y(y)} = \frac{\Pr_X(x)\,\Pr_Y(y \mid X=x)}{\Pr_Y(y)}$$

Bayes's rule is true for any pair of random variables X, Y.

It's only useful in "sequential models" i.e. when the question tells us $\Pr_X(x)$ and $\Pr_Y(y|X=x)$.

Bayes's rule for discrete or continuous random variables

For two random variables X and Y,

$$\Pr_X(x|Y=y) = \frac{\Pr_X(x)\,\Pr_Y(y|X=x)}{\Pr_Y(y)} \quad \text{when} \quad \Pr_Y(y) > 0$$

In practice, we use it as

$$\Pr_X(x|Y=y) = \kappa \Pr_X(x)\,\Pr_Y(y|X=x)$$

$\Pr_{(X|Y=y)}(x)$

constant that doesn't involve x

$\int_x \Pr_X(x|Y=y)\,dx = 1$ or $\sum_x \Pr_X(x|Y=y) = 1$

then figure out κ so that $\Pr_X(\cdot\,|Y=y)$ is a legitimate likelihood function

ie so that $\int_x \Pr_{(X|Y=y)}(x)\,dx = 1.$

Exercise 5.2.1

Consider the pair of random variables (X, Y) generated by

```
def rxy(σ):
    x = np.random.uniform(-1,1)
    y = np.random.normal(loc=x**2, scale=σ)
    return (x,y)
```

Or, in maths notation,

$$X \sim U[-1,1], \qquad Y \sim N(X^2, \sigma^2)$$

Calculate $\Pr_X(x \mid Y = y)$.

$\Pr_X(x) = \frac{1}{2}$ since $X \sim U[-1,1]$

$\Pr_Y(y|X = x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(y-x^2)^2/2\sigma^2}$

$\Pr_X(x|Y = y) = \kappa \Pr_X(x) \Pr_Y(y|X = x) = \kappa \frac{1}{2} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(y-x^2)^2/2\sigma^2}$

function of x

$= \kappa' e^{-(y-x^2)^2/2\sigma^2}$ where κ' has non-x terms.

$= \kappa' e^{-(x^2-y)^2/2\sigma^2}$ to remind me it's a function of x.

$$\int_{-1}^{1} \kappa' e^{-(x^2-y)^2/2\sigma^2} dx = 1$$

$$\Rightarrow \kappa' = \frac{1}{\int_{-1}^{1} e^{-(x^2-y)^2/2\sigma^2} dx}$$

= <yuck />

2 CLASS EXAMPLE

- if $P(y_1|x) > P(y_2|x)$ then assign x to y_1
- if $P(y_1|x) < P(y_2|x)$ then assign x to y_2

$$\Downarrow$$

$$P(y_1|x) \underset{y_2}{\overset{y_1}{\gtrless}} P(y_2|x)$$

Bayes classification rule is optimal w.r.t. minimising the classification error probability

Is it always the best choice?

Hint

Try to perform classification on a dataset used to determine whether a landslide is occurring or not