Statistics Refresher

The slides contain just the important concepts. It does not involve any in-depth Mathematics

Expectation of a Random Variable

The expectation of a random variable or its *expected value* is the mean or average value that random variable takes and is defined as

$$\mathbb{E}[X] = \sum_{x \in S_X} x \cdot \mathbb{P}[X = x]$$

Sometimes the notation used is just $\mathbb{E}X$ *i.e. brackets are omitted*

The name suggests that the r.v. is expected to take this value

Some truth to this: if we sample from $\mathbb{P}_X$*, "likely" to get a value "close" to* $\mathbb{E}X$

What "close", "likely" mean are topics in a learning theory course (e.g. CS777)

However, can be misleading – be careful not to read too much into it

$\mathbb{E}X$ *need not be most likely value for* X *i.e.* $\mathbb{E}X \neq \arg\max \mathbb{P}[X = x]$ *possible*

In fact, there are r.v. X *which can never take this value i.e.* $\mathbb{P}[X = \mathbb{E}X] = 0$

Rules of Expectation: Sum Rule

Linearity of Expectation: given two r.v. X, Y, no matter how they are defined, no matter whether independent or not, we always have

$$\mathbb{E}[X+Y] = \mathbb{E}X + \mathbb{E}Y$$

Proof: *Let* $Z \triangleq X + Y$ *be a new r.v.. We have* $\mathbb{E}[Z] = \sum_{z \in S_Z} z \cdot \mathbb{P}[Z = z]$. *Now the only possible values for z are of the form* $(x + y)$ *where* $x \in S_X, y \in S_Y$.

Thus, we have $\mathbb{E}Z = \sum_{x \in S_X} \sum_{y \in S_Y} (x + y) \cdot \mathbb{P}[X = x, Y = y]$. *Note that even if multiple ways of getting a value z, all have been taken into account.*

$$\sum_x \sum_y (x+y) \cdot \mathbb{P}[x,y] = \sum_x \sum_y x \cdot \mathbb{P}[x,y] + \sum_x \sum_y y \cdot \mathbb{P}[x,y]$$
$$= \sum_x x \sum_y \mathbb{P}[x,y] + \sum_y y \sum_x \mathbb{P}[x,y] = \sum_x x\mathbb{P}[x] + \sum_y y\mathbb{P}[y] = \mathbb{E}X + \mathbb{E}Y$$

Note that the only result we used in our proof is the law of total probability in the second last step above which always holds no matter which r.v.s we have

Note: *the same proof shows that* $\mathbb{E}[X - Y] = \mathbb{E}X - \mathbb{E}Y$

Rules of Expectation: Scaling Rule

Given a r.v. X and a constant c, define a new r.v. $Y = c \cdot X$ i.e. on any outcome $\omega \in \Omega, Y(\omega) = c \cdot X(\omega)$, then $\mathbb{E}Y = c \cdot \mathbb{E}X$

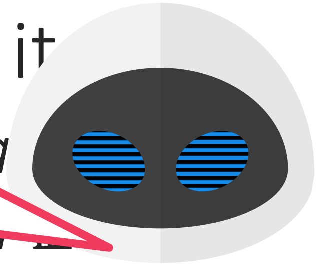
Proof: *any value y that Y takes is cx for some $x \in S_X$. Thus, we get* $\mathbb{E}Y = \sum_{y \in S_Y} y \cdot \mathbb{P}[Y = y] = \sum_{x \in S_X} cx \cdot \mathbb{P}[X = x] = c \cdot \mathbb{E}X$

The e ... nt it

If v ... ha

ou ... ien $\mathbb{E}$

For example, if we have a fair coin and create a r.v. X s.t. $X = 1$ for heads and $X = 0$ for tails, then $\mathbb{E}X = 0.5$ (since coin is fair) and clearly $\mathbb{E}X$ is a constant that does not depend on the outcome of any toss



For any r.v. X, we always have $\mathbb{E}[X - \mathbb{E}X] = 0$

Proof: *Create a dummy random variable Z that always takes the value $\mathbb{E}X$. Note that $\mathbb{E}X$ is a constant (does not depend on the outcome $\omega \in \Omega$) . Linearity gives us* $\mathbb{E}[X - Z] = \mathbb{E}X - \mathbb{E}Z = \mathbb{E}X - \mathbb{E}X = 0$

Note: *notation is horrible here. In the expression $\mathbb{E}[X - \mathbb{E}X]$, X and X do not refer to two r.v.s or the same r.v. repeated. Instead, just read $\mathbb{E}X$ as constant.*

Rules of Expectation

Law of the Unconscious Statistician (LOTUS)

Helps calculate expectations for complicated random variables easily

Suppose we have random variable X whose PMF we know $\mathbb{P}_X$

Suppose there is a weird function $g: S_X \to \mathbb{R}$ and we define a new random variable $Y \triangleq g(X)$. Can we calculate $\mathbb{E}Y$?

Calculating $\mathbb{E}Y$ directly would require us to first get hold of $\mathbb{P}_Y$ – difficult!

LOTUS gives us a way to use $\mathbb{P}_X$ itself to calculate $\mathbb{E}Y$

$$\mathbb{E}Y = \mathbb{E}g(X) = \sum_{x \in S_X} g(x) \cdot \mathbb{P}[X = x]$$

***Proof**: much the same way we proved linearity of expectation*

Works no matter what r.v. X we have, no matter how complicated g is

The function g does need to satisfy some very easy conditions – all functions we will look at in this course will satisfy these conditions

Rules of Expectation: Product Rule

If X, Y are two independent random variables, then we have stronger results on them $\mathbb{E}[X \cdot Y] = \mathbb{E}X \cdot \mathbb{E}Y$

Proof: *Let* $Z \triangleq XY$ *be a new r.v.. We have* $\mathbb{E}[Z] = \sum_{z \in S_Z} z \cdot \mathbb{P}[Z = z]$. *Now the only possible values for z are of the form* xy *where* $x \in S_X, y \in S_Y$.

Thus, we have $\mathbb{E}Z = \sum_{x \in S_X} \sum_{y \in S_Y} xy \cdot \mathbb{P}[X = x, Y = y]$. *Note that even if multiple ways of getting a value z, all have been taken into account.*

Using independence gives us $\mathbb{E}Z = \sum_{x \in S_X} \sum_{y \in S_Y} xy \cdot \mathbb{P}[X = x] \cdot \mathbb{P}[Y = y]$

$= \left(\sum_x x \cdot \mathbb{P}[x]\right) \cdot \left(\sum_y y \cdot \mathbb{P}[y]\right) = \mathbb{E}X \cdot \mathbb{E}Y$

Warning: this result crucially uses independence: may fail if X, Y are not independent

Samp

Indeed! If we ask 1000 random Indians, how many children they have, the sample mean might come out to be 2.35. However, no Indian can have 2.35 children since number of children has to be an integer!

Suppos ... y ... s

E.g. we have a dice/coin and we throw/toss it again and again

Make sure

Yes, that is why we warned not to take expectation/sample mean literally. All that your experiment tells you is that most Indians have *around* 2.35 children. Some may have much more (e.g. 7) or much less (e.g. 0) but they are usually rarer

For examp ... tos ... d
then blind

Using the ... x_n, *we can*
get a
Calle

Interesting fact: the sample mean is the point which is the closest to all samples in terms of squared distance (**Proof**: use first order optimality)

$$\hat{\mathbb{E}}X = \arg\min_c \sum_{i=1}^n (x_i - c)^2$$

Note: s ... i

Interesting fact: even the mean itself satisfies the nice property

$$\mathbb{E}X = \arg\min_c \mathbb{E}[(X - c)^2]$$

Mode of a Random Variable

The mode of a random variable is simply the value(s) that the r.v. takes with highest probability

Warning: a r.v. may have more than one mode value

$$\text{mode}(X) \triangleq \arg\max_{x \in S_X} \mathbb{P}[X = x]$$

Recall that $\mathbb{I}\{\mathbf{blah}\} = 1$ if blah is true (or blah happens) else if blah does not happen or is false, $\mathbb{I}\{\mathbf{blah}\} = 0$

he *empirical mode* similarly

y in the samples

$$\text{mode}(x_1, x_2, \ldots, x_n) \triangleq \arg\max_{x \in S_X} \sum_{i=1}^{n} \mathbb{I}\{x_i = x\}$$

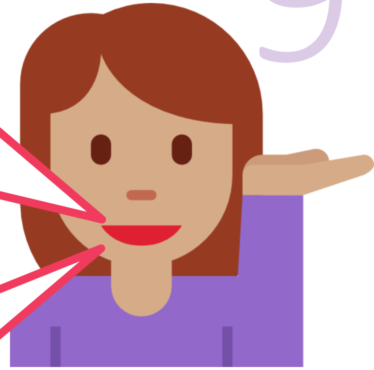
$$= \arg\max_{x \in \{x_1, \ldots, x_n\}} \sum_{i=1}^{n} \mathbb{I}\{x_i = x\}$$

Note: mode of a random variable (or even samples) is always in S_X i.e. always a valid value that the r.v. can actually take (unlike expectation)

Medi

Interesting Fact: The empirical median is the point which is the closest to all samples in terms of absolute distance (**Proof**: in notes)

$$\hat{\mathbb{E}}X = \arg\min_c \sum_{i=1}^{n} |x_i - c|$$



The me

$\mathbb{P}[X \leq$

Interesting fact: even the median itself satisfies the nice property

$$\mathbb{E}X = \arg\min_c \mathbb{E}[|X - c|]$$

The *em* $, \ldots, x_n$

of a ran as many samples are greater than or equal to m as are less than or equal to m

Often we talk about median income *of a country – this is a value such that half the population earns at least that much value as income*

To find the empirical median, first arrange samples in increasing order i.e. $x_1 \leq x_2 \leq \cdots \leq x_n$

If n is odd, then $m = x_{\frac{n+1}{2}}$. If n is even, then may be (infinitely) many empirical medians but we often take $m = \frac{1}{2}\left(x_{\frac{n}{2}} + x_{\frac{n}{2}+1}\right)$

The empirical median gives a good estimate of median of the r.v. if n is large

Variance

Tells us how "spread out" are the values that an r.v. takes. Specifically, how far away from its expectation does the r.v. often take values

For a random variable X with expectation $\mu \triangleq \mathbb{E}X$, its variance, denoted as $\mathbb{V}[X]$ or $\mathrm{Var}[X]$ or often just as σ^2 can be defined as

$$\sigma^2 = \mathbb{E}[(X-\mu)^2] = \mathbb{E}[(X-\mathbb{E}X)^2] = \sum_{x \in S_X} (x-\mu)^2 \cdot \mathbb{P}[X=x]$$

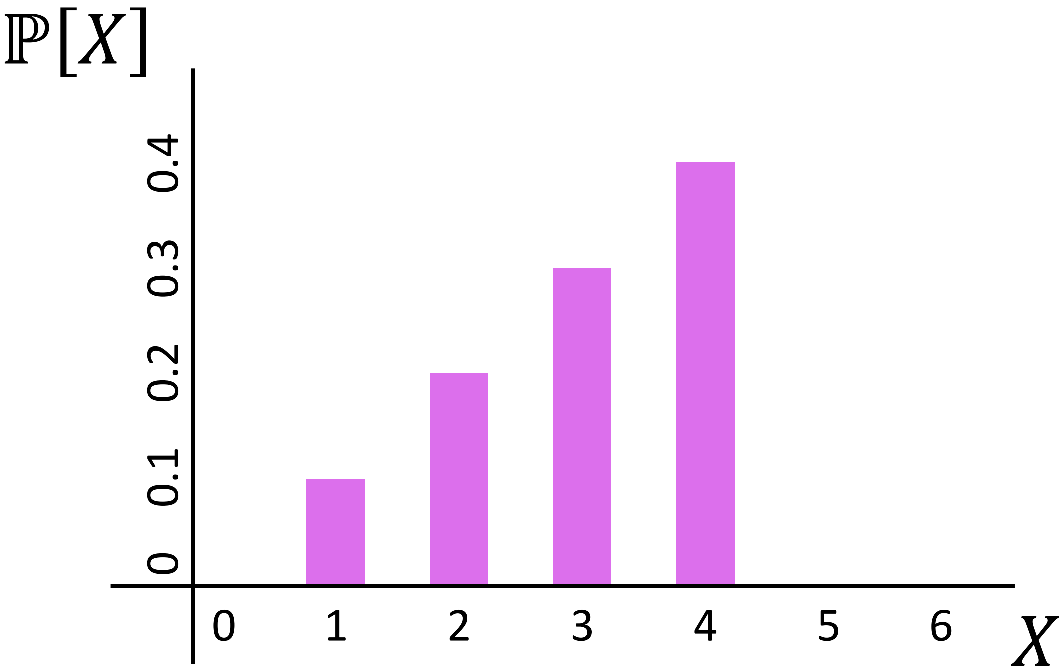
Can be simplified to obtain another (equivalent) definition

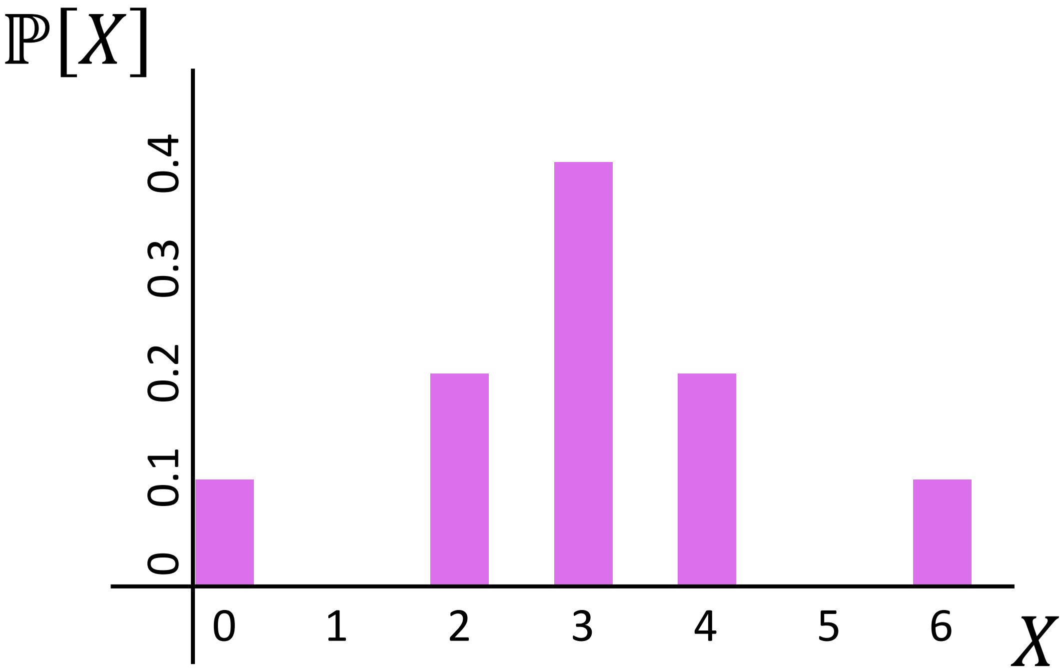
$$\mathbb{E}[(X-\mu)^2] = \mathbb{E}[X^2+\mu^2-2\mu\cdot X] = \mathbb{E}[X^2]+\mathbb{E}[\mu^2]-2\mu\cdot\mathbb{E}[X]$$
$$= \mathbb{E}[X^2]-\mu^2 = \mathbb{E}[X^2]-(\mathbb{E}[X])^2$$

Notice: $(x-\mu)^2 \geq 0$ *for all* $x \in S_X$ *which means* $\mathbb{E}[(X-\mu)^2] \geq 0$ *which means that* $\mathbb{E}[X^2] \geq (\mathbb{E}[X])^2$ *for all r.v.* X*. Also* $\mathbb{V}[X] \geq 0$ *for all r.v.* X

Standard deviation: the square root of the variance (denoted σ)

Example





$\mathbb{E}X = 3, \mathrm{med}(X) = 3,$
$\mathrm{mode}(X) = 4, \mathbb{V}X = 1$

$\mathbb{E}X = 3, \mathrm{med}(X) = 3,$
$\mathrm{mode}(X) = 3, \mathbb{V}X = 2.2$

This distribution has the same mean and median as the first one but is more “spread out” hence larger variance

Sample Variance

Given n independent samples $x_1, \ldots, x_n$ of a random variable X, the empirical variance can be calculated in two (equivalent) ways

First find the empirical mean $\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} x_i$

Method 1*: Calculate* $\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \hat{\mu})^2$

$^2 = \hat{s} - \hat{\mu}^2$

rflow errors occur)

ince it can be

wo passes over data

An effect called *catastrophic cancellation*. Basically, on computers, due to finite precision, for example, if we have $\hat{s} = 1000000000001$ and $\hat{\mu}^2 = 1000000000000$, then clearly $\hat{\sigma}^2 = 1$ but our computers may store $\hat{s} = 1000000000000$ to save space and ignore the error and cause us to get $\hat{\sigma}^2 = 0$

However, method 2 can be bad if $\hat{s}$ *and* $\hat{\mu}^2$ *are both very large and close*

As before, if n is large, empirical variance is a good estimate of $\mathbb{V}[X]$

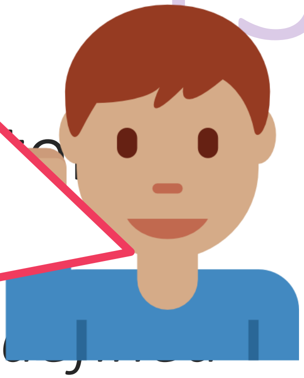
We can estimate covariance using samples too. Suppose we are given values of X, Y on n outcomes (i.e. we sampled n outcomes $\omega_1, \dots, \omega_n$ and on each outcome ω_i, we return $(x_i, y_i) = \left(X(\omega_i), Y(\omega_i)\right)$). Then sample covariance can be computed in two ways. First calculate empirical mean of X and Y

$\hat{\mu}_X = \frac{1}{n}\sum_{i=1}^n x_i$ and $\hat{\mu}_Y = \frac{1}{n}\sum_{i=1}^n y_i$

Method 1: Calculate $\widehat{\text{Cov}}(X,Y) = \frac{1}{n}\sum_{i=1}^n (x_i - \hat{\mu}_X)\,(y_i - \hat{\mu}_Y)$

Method 2: First calculate $\hat{c} = \frac{1}{n}\sum_{i=1}^n x_i y_i$ and get $\widehat{\text{Cov}}(X,Y) = \hat{c} - \hat{\mu}_X\hat{\mu}_Y$

Just as before, both methods always give the same answer. Method 2 useful when data not available all at once but can be bad if $\hat{c}$ and $\hat{\mu}_X\hat{\mu}_Y$ are both very large in magnitude but close together as well



...es defined ...n of a ...ir income ...xpected ...n the

person would sleep fewer than the average number of hours (since children typically sleep more and old people tend to sleep less)

$$\text{Cov}(X,Y) \triangleq \mathbb{E}[(X - \mathbb{E}X)\cdot(Y - \mathbb{E}Y)] = \mathbb{E}[XY] - \mathbb{E}X \cdot \mathbb{E}Y$$
$$= \sum_{\omega\in\Omega}(X(\omega) - \mathbb{E}X)\cdot(Y(\omega) - \mathbb{E}Y)\cdot p_\omega$$

Note that $\text{Cov}(X,X) = \mathbb{V}[X]$

Note that $\text{Cov}(X,Y)$ *may be positive, negative or zero*

Rules of Variance

Suppose $b, c \in \mathbb{R}$ are any two constants and X, Y are any two r.v.s, then

Constant Rule: $\mathbb{V}[c] = 0$ i.e. if $Z \equiv c$ is a constant r.v. then $\mathbb{V}[Z] = 0$

Se ... *ary at all i.e. zero variance*

Often used to deal with catastrophic cancellation by shifting the data to make it smaller in magnitude but leaving variance unchanged

Scali ...

Shift Rule: $\mathbb{V}[X + c] = \mathbb{V}[X]$ i.e. if $W \triangleq X + c$ then $\mathbb{V}[W] = \mathbb{V}[X]$

Shifting a random variable does not change its "spread"

Sum Rule: $\mathbb{V}[X + Y] = \mathbb{V}[X] + \mathbb{V}[Y] + 2\mathrm{Cov}[X, Y]$

Difference Rule: $\mathbb{V}[X - Y] = \mathbb{V}[X] + \mathbb{V}[Y] - 2\mathrm{Cov}[X, Y]$

In books/papers, you may come across a term called *correlation* which is a normalized version of covariance.

$$\rho_{X,Y} = \text{Corr}(X, Y) = \frac{\text{Cov}(X, Y)}{\sqrt{\mathbb{V}[X] \cdot \mathbb{V}[Y]}}$$

For any two r.v.s X, Y, we always have $\rho_{X,Y} \in [-1,1]$. If $\rho_{X,Y} = 0$ then the two r.v.s are said to be *uncorrelated*. Note that if X, Y are uncorrelated, then also we have $\mathbb{V}[X+Y] = \mathbb{V}[X] + \mathbb{V}[Y]$. Warning: independent r.v.s are always uncorrelated but not all uncorrelated r.v.s need be independent.

Can estimate $\rho_{X,Y}$ using samples as well $\hat{\rho}_{X,Y} = \frac{\widehat{\text{Cov}}(X,Y)}{\sqrt{\hat{\sigma}_X^2 \hat{\sigma}_Y^2}}$

If X, Y are independent then $\text{Cov}[X, Y] = 0$

If $\rho_{X,Y} < 0$, this means that typically, whenever X takes larger values than its own mean, Y takes smaller values than its own mean and vice versa. If $\rho_{X,Y} > 0$, then this means that both r.v.s take values larger or smaller than their respective means together. $\rho_{X,Y} = 0$ means that typically, even if X takes a value larger than its mean, Y may take smaller or larger values than its own mean

Conditional Statistics

The notation $[\cdot \mid \cdot]$ is used to express how one quantity behaves when some other quantities are fixed to some given values

These "other" quantities could be random variables themselves, or even constants. Sometimes we condition just to clarify exactly what those constants are

For example we could ask, what is the probability of me misclassifying a test data point $(\mathbf{x}, y) \sim \mathcal{D}$ *if I use a model* $\mathbf{w}$ *i.e.* $\mathbb{P}[y \cdot \mathbf{w}^\top \mathbf{x} < 0 \mid \mathbf{w}]$

Here $\mathbf{w}$ *is not a random variable (it could be in other settings but here it is not)*

We previously saw conditional probabilities $\mathbb{P}[X = 1 \mid Y = 2] = \frac{\mathbb{P}[X=1, Y=2]}{\mathbb{P}[Y=2]}$

Let us see other quantities that can be defined conditionally

Conditional Statistics

Conditional Expectation $\mathbb{E}[X \mid Y = y_0] \triangleq \sum_{x \in S_X} x \cdot \mathbb{P}[X = x \mid Y = y_0]$

Conditional Variance $\mathbb{V}[X \mid Y = y_0] \triangleq \mathbb{E}[(X - \mu)^2 \mid Y = y_0]$ where we have $\mu = \mathbb{E}[X \mid Y = y_0]$

Conditional Covariance $\mathbf{Cov}[X, Y \mid Z = z_0]$

$= \mathbb{E}[(X - \mu_X) \cdot (Y - \mu_Y) \mid Z = z_0] = \mathbb{E}[XY \mid Z = z_0] - \mu_X \cdot \mu_Y$ where $\mu_X = \mathbb{E}[X \mid Z = z_0]$ and $\mu_Y = \mathbb{E}[Y \mid Z = z_0]$

Conditional Mode $\mathbf{mode}[X \mid Y = y_0] = \arg\max_{x \in S_X} \mathbb{P}[X = x \mid Y = y_0]$

Similarly we can define conditional median etc but not very popular

Note: these definitions do not require X, Y, Z to be independent at all!

Conditional Statistics

Rules of expectation (sum, scaling, LOTUS, product) all continue to hold even with conditional except that all expectation are conditional

$$\mathbb{E}[X + Y \mid Z = z_0] = \mathbb{E}[X \mid Z = z_0] + \mathbb{E}[Y \mid Z = z_0]$$

$$\mathbb{E}[c \cdot X \mid Z = z_0] = c \cdot \mathbb{E}[X \mid Z = z_0]$$

$$\mathbb{E}[g(X) \mid Z = z_0] = \sum_{x \in S_X} g(x) \cdot \mathbb{P}[X = x \mid Z = z_0]$$

If $X \perp\!\!\!\perp Y \mid Z$ then $\mathbb{E}[X \cdot Y \mid Z = z_0] = \mathbb{E}[X \mid Z = z_0] \cdot \mathbb{E}[Y \mid Z = z_0]$

Rules of variance and covariance also continue to hold if we systematically condition all expressions involved in those rules

Note: conditioning must be the same everywhere, i.e. may happen that

$$\mathbb{E}[X + Y \mid Z = z_0] \neq \mathbb{E}[X \mid Z = z_1] + \mathbb{E}[Y \mid Z = z_2]$$

Statistics of Random Vectors

Expectation of a random vector is simply another vector (of same dim) of the expectations of the individual random variables

$$\mathbb{E}\mathbf{X} = [\mathbb{E}X_1, \mathbb{E}X_2, \dots \mathbb{E}X_d]^\top$$

Linearity of expectation continues to hold: if $\mathbf{X}, \mathbf{Y}$ any two vector r.v. (not necessarily independent, then $\mathbb{E}[\mathbf{X} + \mathbf{Y}] = \mathbb{E}\mathbf{X} + \mathbb{E}\mathbf{Y}$

Scaling Rule: If $c \in \mathbb{R}$ is a constant then $\mathbb{E}[c \cdot \mathbf{X}] = c \cdot \mathbb{E}\mathbf{X}$

Dot Product Rule: If $\mathbf{a} \in \mathbb{R}^d$ is a constant vector, then $\mathbb{E}[\mathbf{a}^\top\mathbf{X}] = \mathbf{a}^\top\mathbb{E}\mathbf{X}$

Proof: $\mathbb{E}[\mathbf{a}^\top\mathbf{X}] = \mathbb{E}\left[\sum_{i=1}^d a_i X_i\right] = \sum_{i=1}^d \mathbb{E}[a_i X_i] = \sum_{i=1}^d a_i \cdot \mathbb{E}[X_i] = \mathbf{a}^\top\mathbb{E}\mathbf{X}$

Matrix Product Rule: If $A \in \mathbb{R}^{n\times d}$ is a constant matrix then

$$\mathbb{E}[A\mathbf{X}] = A\mathbb{E}\mathbf{X}$$

Proof: *Use Dot Product Rule n times*

Statistics of Random Vectors

Mode easy to define: $\arg\max_{X_1,\ldots,X_d} \mathbb{P}[X_1, \ldots, X_d]$

Median not easy to define – no unique definition

Definition 1: $\text{med}(\mathbf{X}) = [\text{med}(X_1), \text{med}(X_2), \ldots, \text{med}(X_d)]^\top$

Definition 2: minimizer of absolute distance (in this case L1 norm)

$$\text{med}(\mathbf{X}) = \arg\min_{\mathbf{v}\in\mathbb{R}^d} \mathbb{E}[\|\mathbf{X}-\mathbf{v}\|_2]$$

Note: even here we still have $\mathbb{E}[\mathbf{X}] = \arg\min_{\mathbf{v}\in\mathbb{R}^d} \mathbb{E}[\|\mathbf{X}-\mathbf{v}\|_2^2]$

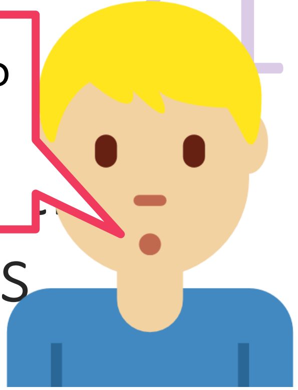
Proof: $\mathbb{E}[\|\mathbf{X}-\mathbf{v}\|_2^2] = \mathbb{E}[\|\mathbf{X}\|_2^2] + \mathbb{E}[\|\mathbf{v}\|_2^2] - 2\cdot\mathbf{v}^\top\mathbb{E}[\mathbf{X}]$

Taking derivative w.r.t $\mathbf{v}$ *and using first order optimality does the trick*

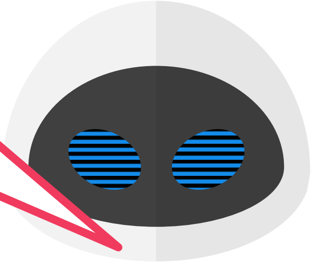
Statistics of Random Vectors

Since random vectors are a bunch of [partly covered by a speech box] variance of this collection, need to have all pairwise covariances

$$\text{Cov}(\mathbf{X}) = \begin{bmatrix} \mathbb{V}X_1 & \text{Cov}(X_1, X_2) & \cdots & \text{Cov}(X_1, X_d) \\ \text{Cov}(X_2, X_1) & & & \\ \vdots & & & \\ \text{Cov}(X_d, X_1) & \text{Cov}(X_d, X_2) & \dots & \mathbb{V}X_d \end{bmatrix}$$



If $\mathbf{X}$ is a vector, isn't $\mathbf{X}\mathbf{X}^\top$ a matrix?
What does $\mathbb{E}[\mathbf{X}\mathbf{X}^\top]$ even mean?

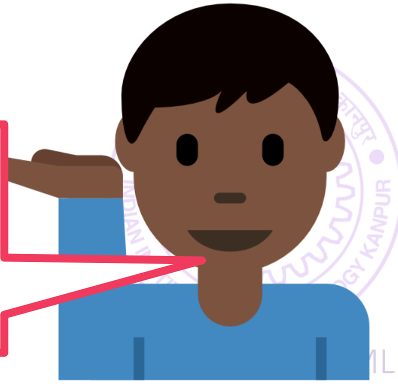


Just as a random vector is a collection of random variables arranged as a 1D array, a random matrix is a collection of r.v.s arranged as a 2D array!

Another cute formula

$$\text{Cov}(\mathbf{X}) = \mathbb{E}[(\mathbf{X} - \boldsymbol{\mu})(\mathbf{X} - \boldsymbol{\mu})^\top] = \mathbb{E}[\mathbf{X}\mathbf{X}^\top] - \boldsymbol{\mu}\boldsymbol{\mu}^\top, \text{ where } \boldsymbol{\mu} = \mathbb{E}\mathbf{X}$$

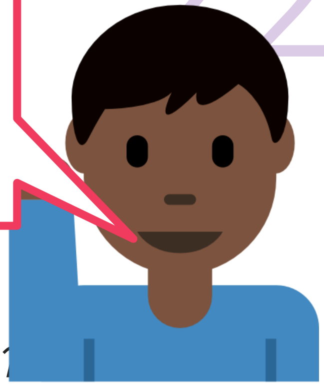
$$\text{Cov}(c \cdot \mathbf{X}) = c^2 \cdot \text{Cov}(\mathbf{X})$$



Note that (i, j)-th entry of matrix $(\mathbf{X} - \boldsymbol{\mu})(\mathbf{X} - \boldsymbol{\mu})^\top$ is $(X_i - \mu_i)(X_j - \mu_j)$. Thus, (i, j)-th entry of $\mathbb{E}[(\mathbf{X} - \boldsymbol{\mu})(\mathbf{X} - \boldsymbol{\mu})^\top]$ is $\mathbb{E}\left[(X_i - \mu_i)(X_j - \mu_j)\right] = \text{Cov}(X_i, X_j)$

Useful Operations on V

> Can you prove that the covariance matrix of any random vector is always a PSD matrix?



If $\mathbf{X} \in \mathbb{R}^m, \mathbf{Y} \in \mathbb{R}^n$ are two random
independent), then

$$\text{Cov}(\mathbf{X}, \mathbf{Y}) = \mathbb{E}[(\mathbf{X} - \boldsymbol{\mu}_\mathbf{X})(\mathbf{Y} - \boldsymbol{\mu}_\mathbf{Y})^\top] = \mathbb{E}[\mathbf{X}\mathbf{Y}^\top] - \boldsymbol{\mu}_\mathbf{X}\boldsymbol{\mu}_\mathbf{Y}^\top \in \mathbb{R}$$

where $\boldsymbol{\mu}_\mathbf{X} = \mathbb{E}\mathbf{X}$ and $\boldsymbol{\mu}_\mathbf{Y} = \mathbb{E}\mathbf{Y}$, $\text{Cov}(\mathbf{X}, \mathbf{Y})$

Dot Product Rule: If $\mathbf{a} \in \mathbb{R}^d$ is a constant vector, then $\mathbb{V}[\mathbf{a}^\top\mathbf{X}] = \mathbf{a}^\top\text{Cov}[\mathbf{X}]\mathbf{a}$

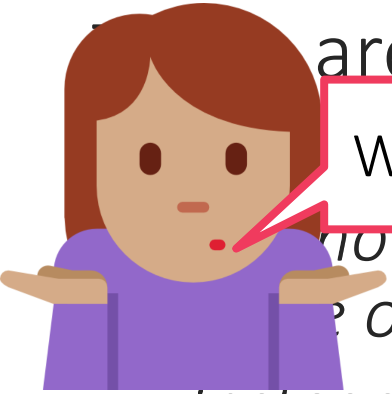
Proof: $\mathbb{V}[\mathbf{a}^\top\mathbf{X}] = \mathbb{E}[(\mathbf{a}^\top\mathbf{X})^2] - (\mathbf{a}^\top\boldsymbol{\mu}_\mathbf{X})^2 = \mathbb{E}[\mathbf{a}^\top\mathbf{X}\mathbf{X}^\top\mathbf{a}] - \mathbf{a}^\top\boldsymbol{\mu}_\mathbf{X}\boldsymbol{\mu}_\mathbf{X}^\top\mathbf{a}$
$= \mathbf{a}^\top\mathbb{E}[\mathbf{X}\mathbf{X}^\top]\mathbf{a} - \mathbf{a}^\top\boldsymbol{\mu}_\mathbf{X}\boldsymbol{\mu}_\mathbf{X}^\top\mathbf{a} = \mathbf{a}^\top\left(\mathbb{E}[\mathbf{X}\mathbf{X}^\top] - \boldsymbol{\mu}_\mathbf{X}\boldsymbol{\mu}_\mathbf{X}^\top\right)\mathbf{a} = \mathbf{a}^\top\text{Cov}[\mathbf{X}]\mathbf{a}$

Matrix Product Rule: If $A \in \mathbb{R}^{n \times d}$ is a constant matrix then

$$\text{Cov}[A\mathbf{X}] = A\text{Cov}[\mathbf{X}]A^\top \in \mathbb{R}^{n \times n}$$

Proof: *Try arguing similarly as the dot product rule*

Continuous Random Variables



... are r.v. that can take infinitely many possibly values that are not ... i.e. support is $\mathbb{R}$ or some subset of $\mathbb{R}$

Why $\approx$ why not $=$?

...notion of PMF which tells us with what probability does the r.v. take this ... or that value does not make sense when support is continuous

Instead of PMF, we use a PDF (probability density function) in such cases

Consi...

War...

To s...

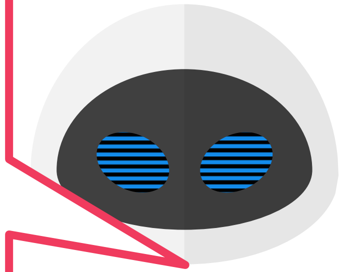
The PD... ...ive number

... be negative

...is X to take a

We do have an exact formula too $\mathbb{P}\left[X \in [x-\delta, x+\delta]\right] = \int_{x-\delta}^{x+\delta} f_X(t)dt$

In general, if the r.v. X has a PDF f_X, then for any interval within its support $[a,b] \subseteq S_X$, we have $\mathbb{P}\left[X \in [a,b]\right] = \int_a^b f_X(t)dt$



However, if the interval is “small”, then we can often get a good and simple approximation $\mathbb{P}\left[X \in [a,b]\right] \approx f_X(c)\cdot(b-a)$ where $c = \frac{(b+a)}{2}$. How small is “small” enough depends on the PDF f_X

$$\mathbb{P}\left[X \in [x-\delta, x+\delta]\right] \approx f_X(x)\cdot 2\delta$$

Continuous R.V.s– the Rules Revisited

PDF f_X of a r.v. X satisfies $f_X(x) \geq 0$ for all $x \in S_X$ and $\int_{S_X} f_X(t)dt = 1$

Expectation of a continuous R.V. X is $\mathbb{E}X \triangleq \int_{S_X} t \cdot f_X(t)dt$

LOTUS: $\mathbb{E}[g(X)] = \int_{S_X} g(t) \cdot f_X(t)dt$

Variance of a continuous R.V. X is $\mathbb{V}X \triangleq \int_{S_X} (t - \mathbb{E}X)^2 \cdot f_X(t)dt$

$\mathbb{V}X = \mathbb{E}[X^2] - (\mathbb{E}X)^2 = \int_{S_X} t^2 \cdot f_X(t)dt - (\mathbb{E}X)^2$

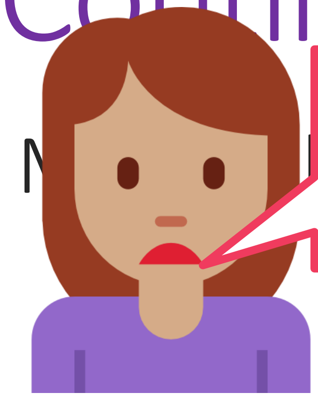
Joint PDFs make sense too $f_{X,Y}$: suppose $[a, b] \subseteq S_X$ and $[p, q] \subseteq S_Y$

$$\mathbb{P}\left[X \in [a, b], Y \in [p, q]\right] = \int_p^q \int_a^b f_{X,Y}(s, t)\, ds\, dt$$

They make sense even if X is continuous and Y is discrete and vice versa

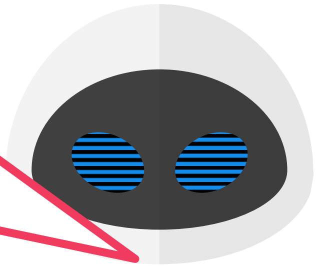
Details of these constructions, however, are beyond the scope of CS771

Continuous R.V.s– the Rules Revisited



N ... to make sense

$f_X(x) = \int f_{X,Y}(x,t)\,dt$



Conditional pro...

When X, Y are ...

$\mathbb{P}\big[X \in [a,b] \mid Y \in [p,q]\big] \triangleq \mathbb{P}\big[X \in [a,b], Y \in [p,q]\big] / \mathbb{P}\big[Y \in [p,q]\big]$

Actually, even $\mathbb{P}[X \in [a,b] \mid Y = r]$ makes sense in this case – details beyond CS771

When X is discrete but Y is continuous $\mathbb{P}\big[X = c \mid Y \in [p,q]\big]$

Actually, even $\mathbb{P}[X = c \mid Y = r]$ makes sense in this case – details beyond CS771

When X is continuous but Y is discrete $\mathbb{P}[X \in [a,b] \mid Y = r]$

Conditional expectations, (co)variances also defined similarly

Tricky to define in some cases as above – details beyond scope of CS771

Continuous R.V.s– the Rules Revisited

Rules of Probability: All rules Sum, Product, Chain, Bayes, Complement, Union continue to hold

If X, Y are independent continuous R.V. then $f_{X,Y}(x,y) = f_X(x) \cdot f_Y(y)$

For independent continuous R.V. we continue to have

$$\mathbb{P}\big[X \in [a,b] \mid Y \in [p,q]\big] = \mathbb{P}\big[X \in [a,b]\big]$$
$$\mathbb{P}\big[X \in [a,b] \mid Y \in r\big] = \mathbb{P}\big[X \in [a,b]\big]$$

Rules of Expectation: All rules Linearity, Scaling, Product still hold

Rules of (co)Variance: All rules Constant, Scaling, Shift, Sum still hold

Probability Distributions

- Some distributions popularly used in ML
- Discrete distributions: Bernoulli, Rademacher
- Continuous distributions: Uniform, Gaussian, Laplacian
- Some of their properties such as mean, median etc
- Notion of a parametric distribution

Bernoulli Distributions

These are probability distributions over the support $\{0,1\}$

Very useful in binary classification as labels are often named $\{0,1\}$

Arguably the simplest of all distributions. PMF of a r.v. Y with Bernoulli distribution is uniquely specified by just specifying $\mathbb{P}[Y = 1] = p$

Using complement rule we automatically get $\mathbb{P}[Y = 0] = 1 - p$

p *called "success probability" or "bias"*

Do not confuse this with the bias of linear model – not the same thing!

Mean: p

Mode: 1 if $p > 0.5$, 0 if $p < 0.5$, $\{0,1\}$ if $p = 0.5$

Variance: $p(1 - p)$

Rademacher Distributions

These are probability distributions over the support $\{-1,1\}$

Very similar to Bernoulli distributions except that support is different

If X is distributed as Bernoulli then $2X - 1$ is distributed as Rademacher

If Y is distributed as Rademacher then $(Y + 1)/2$ is distributed as Bernoulli

Also extremely simple distribution. PMF of a r.v. Y with Rademacher distribution is uniquely specified by just specifying $\mathbb{P}[Y = 1] = p$

Using complement rule we automatically get $\mathbb{P}[Y = -1] = 1 - p$

Often, papers refer to Rademacher distribution only in special case $p = 0.5$

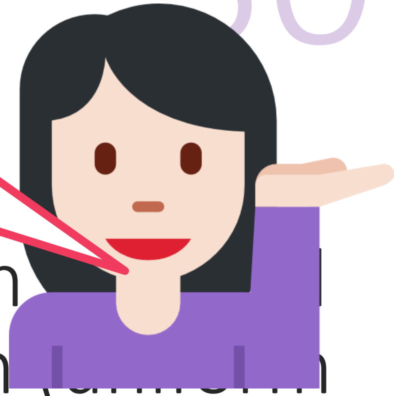
Mean: $2p - 1$ (**Hint**: use scaling and sum rules for expectation)

Mode: 1 if $p > 0.5$, -1 if $p < 0.5$, $\{-1,1\}$ if $p = 0.5$

Variance: $4 \cdot p(1 - p)$ (**Hint**: use scaling and shift rules for variance)

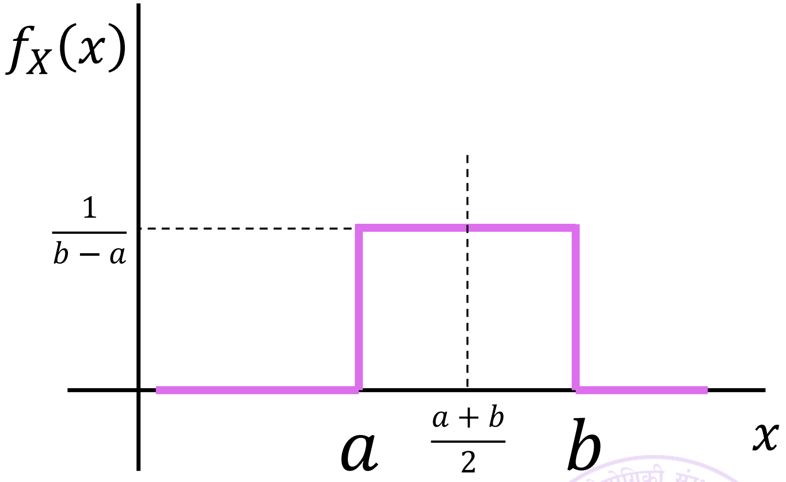
Uniform Distribution

Recall that we commented that although we must have $f_X(x) > 0$, we need not have $f_X(x) \leq 1$. Note that if in the uniform case, if we have $b - a < 1$ then indeed $f_X(x) > 1$ and its perfectly fine



Can be c

Let X be a continuous r.v. with support $S_X = [a, b] \in \mathbb{R}$. Then to have a uniform distribution if its PDF is a constant function (uniform density) i.e. $f_X(x) = \begin{cases} \frac{1}{b-a} & \text{if } x \in [a,b] \\ 0 & \text{if } x \notin [a,b] \end{cases}$



Note: $f_X(x) \geq 0$ if $x \in S_X$ and $\int_{S_X} f_X(t)\, dt = 1$

Mean: $\mathbb{E}[X] = (a+b)/2$

Variance: $\mathbb{V}[X] = (b-a)^2/12$

Note: variance increases as $b - a \uparrow$ since r.v. more "spread out"

Notation: Often we use $\mathrm{UNIF}([a, b])$ to denote uniform dist over $[a, b]$

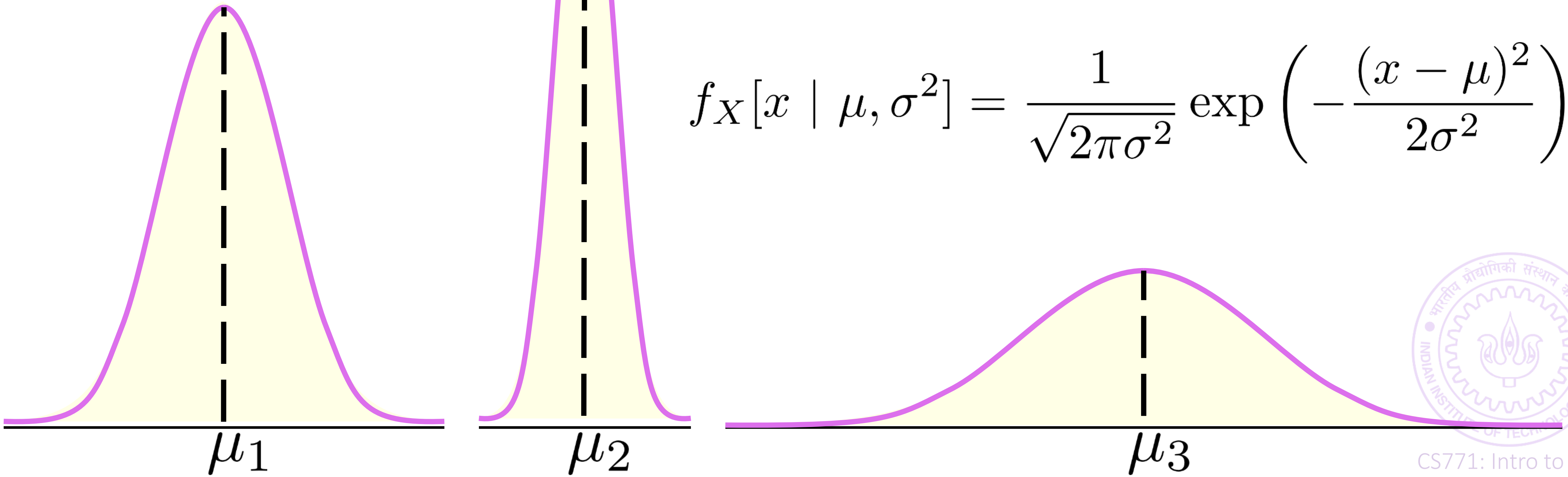
Gaussian (aka Normal) Distributions

Arguably one of the most popular of all probability distributions

Models our intuitive assumption that in real life, data often takes values around its mean value and it gets unlikely to witness extreme values

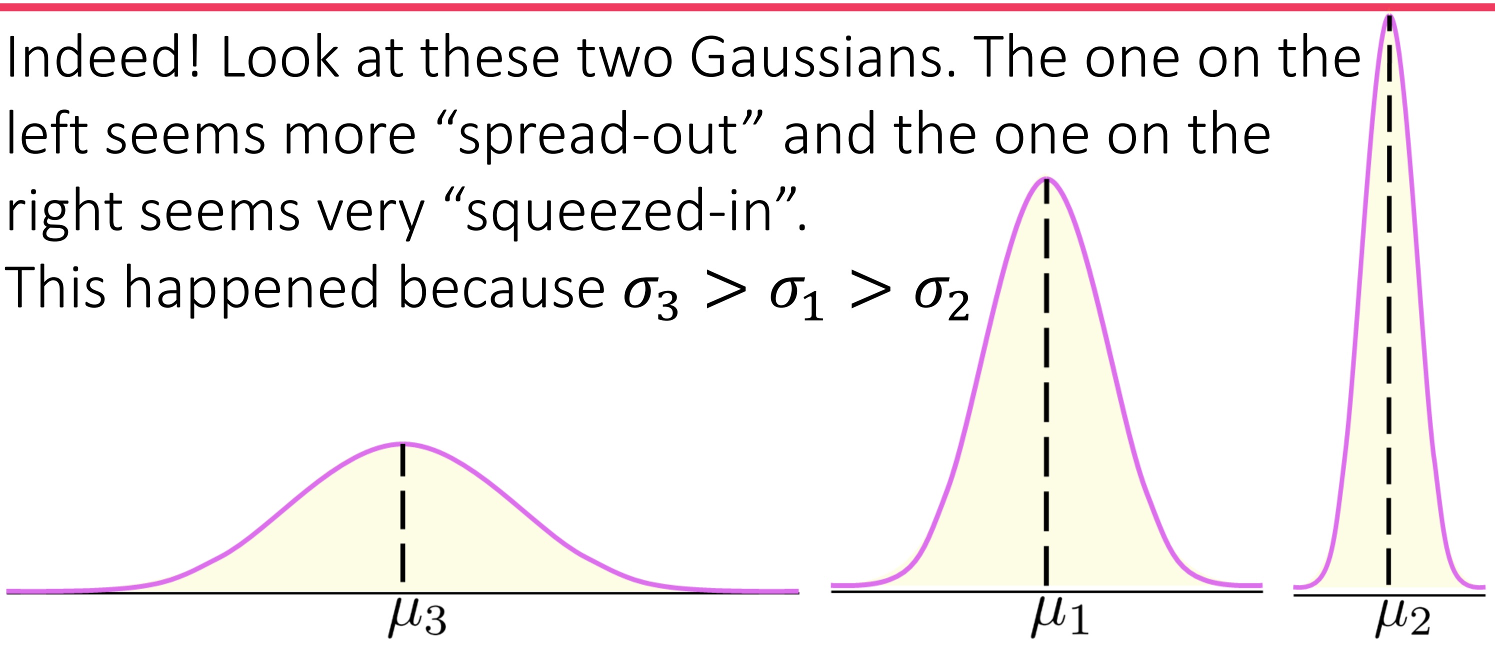
A fundamental result in probability theory – the law of large numbers –*shows that some form of this is indeed true*

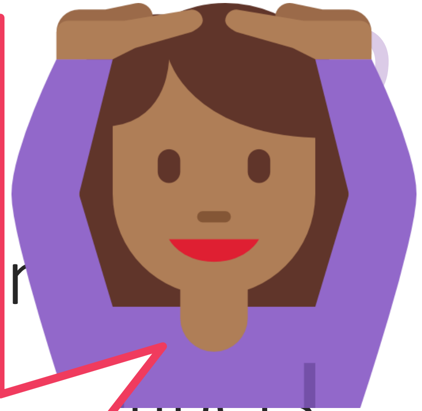
$$f_X[x \mid \mu, \sigma^2] = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$$



Gaussian

Indeed! Look at these two Gaussians. The one on the left seems more "spread-out" and the one on the right seems very "squeezed-in".
This happened because $\sigma_3 > \sigma_1 > \sigma_2$





Specifying a Be ... r

Specifying a cat ... mbers

Specifying a Ga ... ers

μ: must be a r ... ro)

σ^2: must be a non-negative real number

Notation: PDF for a Gaussian r.v. X i.e. $f_X[X \mid \mu, \sigma^2]$ is often written as $\mathcal{N}_X(x; \mu, \sigma^2)$ or simply as $\mathcal{N}(x; \mu, \sigma^2)$

Notice that even here we condition on constants (either using | or ; symbol)

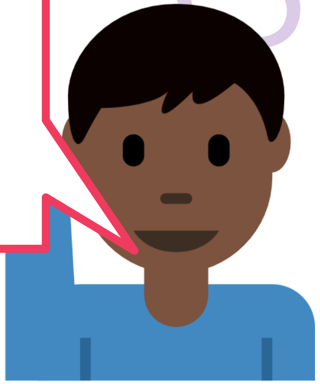
The notation is no accident – if the PDF of a r.v. X is $\mathcal{N}_X(x; \mu, \sigma^2)$, then

$\mathbb{E}[X] = \mu$ *= Mode = Median, as well as* $\mathbb{V}[X] = \sigma^2$

Requires a bit of integration to prove these results ☺

O

Note: we can derive results such as $\mathbb{E}W = \mu_X + \mu_Y$ and $\mathbb{V}W = \sigma_X^2 + \sigma_Y^2$ using rules we studied earlier. However, those rules do not assure us that W must be Gaussian (they just assure us that W is some r.v. with such and such mean and variance. It takes special analysis to show that Z, W, V etc are Gaussian r.v. too!



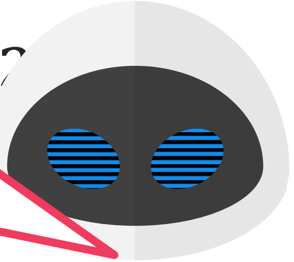
Si

Let X, Y be two **independent** r.v. whose PDF is Gaussian i.e.

$\mathcal{N}_X(\cdot\,; \mu_X, \sigma_X^2)$

The colloquial "**68-95-99.7 rule**" describes this more generally

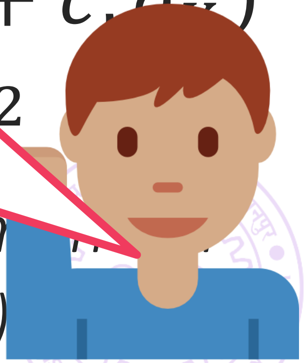
$$\mathbb{P}[|X - \mu_X| \leq \sigma_X] \approx 0.68$$
$$\mathbb{P}[|X - \mu_X| \leq 2 \cdot \sigma_X] \approx 0.95$$
$$\mathbb{P}[|X - \mu_X| \leq 3 \cdot \sigma_X] \approx 0.997$$



Scaling Rule: If ... c^2

Sum Rule: If W

Be careful that this rule apples only to the Gaussian distribution. A random variable sampled from some other distribution may very well violate this rule. People often cite the 68-95-99.7 rule to make real-life predictions. This is merely an approximation (possibly a good one, possibly a bad one) based on an ***assumption*** that the real life distribution is approximately Gaussian



Shif ... $+ c, \sigma_X^2)$

Tail ... $/2$

It ... *n*

For $t = 5$, we have $\mathbb{P}[|X - \mu_X| \geq 5 \cdot \sigma_X] < 0.000004$ (5-sigma rule)

As $\sigma_X \downarrow$ the r.v. gets more and more concentrated around its mean

Gaussian Random Vector

As in the scalar case, the *multivariate* Gaussian requires just the mean $\boldsymbol{\mu} \in \mathbb{R}^d$ and the covariance $\Sigma \in \mathbb{R}^{d \times d}$ to be specified $\mathcal{N}(\boldsymbol{\mu}, \Sigma)$

$$\mathbb{P}[\mathbf{x} \mid \boldsymbol{\mu}, \Sigma] = \frac{1}{\sqrt{(2\pi)^d |\Sigma|}} \exp\left(-\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu})^\top \Sigma^{-1} (\mathbf{x} - \boldsymbol{\mu})\right)$$

Special case $\boldsymbol{\mu} = \mathbf{0}$ and $\Sigma = I_d$ called *standard Gaussian/Normal dist*

$$\mathbb{P}[\mathbf{x} \mid \mathbf{0}, I_d] = \frac{1}{\sqrt{(2\pi)^d}} \exp\left(-\frac{1}{2}\|\mathbf{x}\|_2^2\right) = \prod_{i=1}^d \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}x_i^2\right)$$

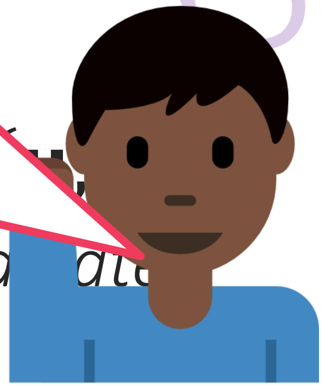
However, $\frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}x_i^2\right)$ is simply $\mathcal{N}(0,1)$ i.e. we indeed have

$$\mathbb{P}[x_1, \ldots, x_d \mid \mathbf{0}, \mathrm{I}] = \prod_{i=1}^d \mathbb{P}[x_i \mid 0,1]$$

All d coordinates of a standard Gaussian r.v. are independent!

Gaus

Note: Just as before, we can derive results such as $\mathbb{E}[\mathbf{a}^\top \mathbf{x}] = \boldsymbol{\mu}^\top \mathbf{x}$ and $\mathbb{V}[\mathbf{a}^\top \mathbf{x}] = \mathbf{a}^\top \Sigma \mathbf{a}$ using rules we studied earlier. However, those rules do not assure us that $\mathbf{a}^\top \mathbf{x}$ or $A\mathbf{x}$ must be Gaussian (they just assure us that these are some r.v./r.vec. with such and such mean and (co)-variance. It takes a more detailed analysis to show that these are actually Gaussian.



Given a

Every ... *ate need not be independent if the Gaussian is non-standard*

The above holds true even if conditioned on all other coordinates of $\mathbf{x}$

Consider any coordinate of the vector, say $j \in [d]$

$\mathbf{x}_j$ *is distributed as the Gaussian* $\mathcal{N}\left(\mu_j, \Sigma_{jj}\right)$

Given values $\mathbf{x}_k = v_k$ *for all other coordinates* $k \neq j$, $\mathbf{x}_j$ *is still Gaussian*

Expression a bit complicated – refer to DFO Sec 6.5.1 (see the reference section on the course webpage)

If $\mathbf{a} \in \mathbb{R}^d$ is a constant vector, then $\mathbb{R} \ni \mathbf{a}^\top \mathbf{x} \sim \mathcal{N}(\boldsymbol{\mu}^\top \mathbf{a}, \mathbf{a}^\top \Sigma \mathbf{a})$

If $A \in \mathbb{R}^{n \times d}$ is a constant matrix then $\mathbb{R}^d \ni A\mathbf{x} \sim \mathcal{N}(A\boldsymbol{\mu}, A\Sigma A^\top)$

Laplacian Distribution

If X is a r.v. with a Laplacian PDF with parameters μ_X, σ_X, then $Y = a \cdot X + b$ (where a, b are constants) is also a Laplacian r.v. but with parameters $\mu_Y = a \cdot \mu_X + b$ and $\sigma_Y = a \cdot \sigma_Y$

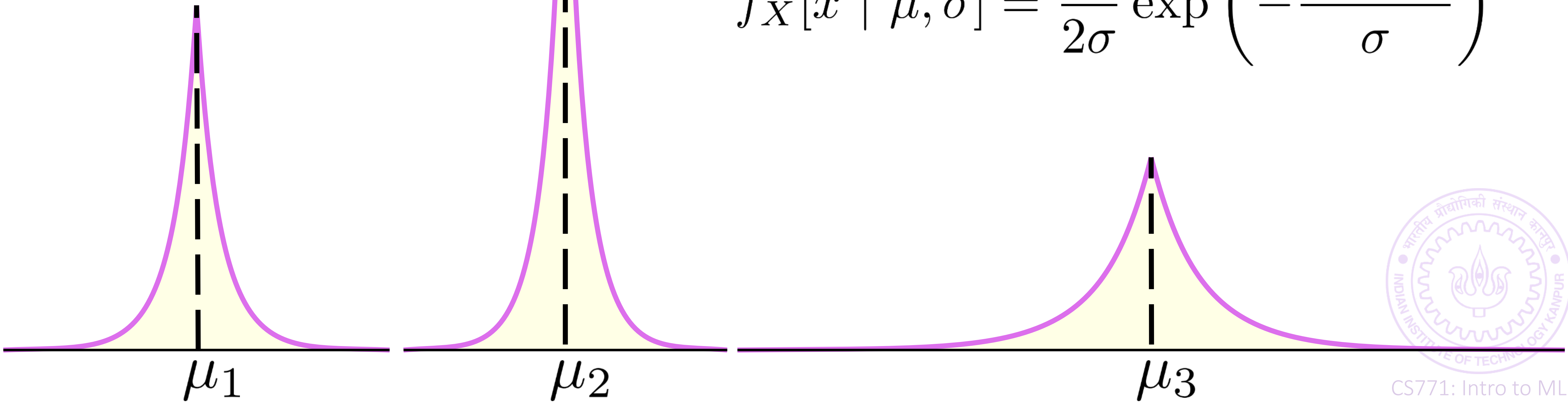
...acian r.v.

...rates much more strongly around its mean than a Gaussian r.v.

Also require two parameters to be specified $\mu \in \mathbb{R}, \sigma \geq 0$

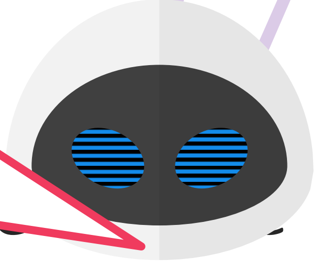
Mean = Mode = Median: μ, **Variance:** $2\sigma^2$

$$f_X[x \mid \mu, \sigma] = \frac{1}{2\sigma} \exp\left(-\frac{|x-\mu|}{\sigma}\right)$$



Parame

Parametric distributions are extremely important for ML. We will next learn about algorithms that try to make realistic predictions by first learning a parametric distribution that mimics reality. This is done by learning the parameters of the distribution using data



Certain distr ... number of *parameters* that completely describe the distribution

- *Similar to parametric models like LwP or linear models where a finite number of parameters (e.g. a model vector and a bias value) describe the model fully*
- *There exist non-parametric distributions too (beyond scope of CS771)*
- *Bernoulli/Rademacher* (p)*, Uniform* (a, b)*, Gaussian* (μ, σ^2)*, Laplacian* (μ, σ)

Statistics: apart from the name of a branch of mathematics, the word "statistic" also refers to some quantity we calculate using samples

- *E.g. sample mean, sample variance, sample mode, sample median*
- *A common usage of statistics is to estimate parameters of the distribution that generated those samples*
- *E.g. under some mild conditions, sample mean/variance is a good estimate of the expectation/variance of distribution that generated the samples*