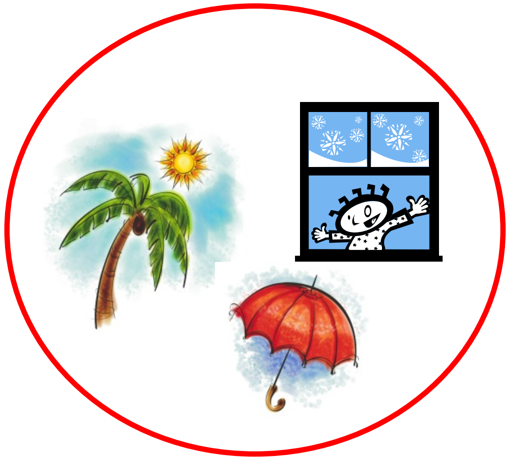
Brief Review of Probability

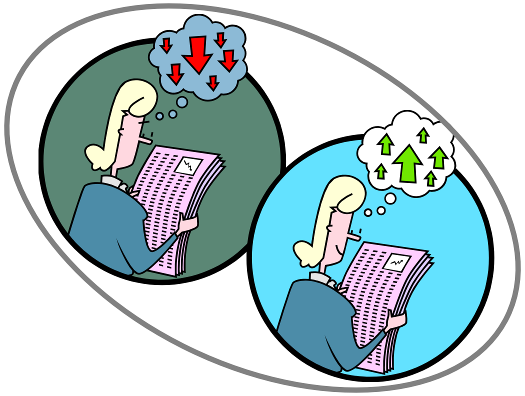
Nuno Vasconcelos
(Ken Kreutz-Delgado)

ECE Department, UCSD

Probability

- Probability theory is a mathematical language to deal with ***processes*** or ***experiments*** that are non-deterministic



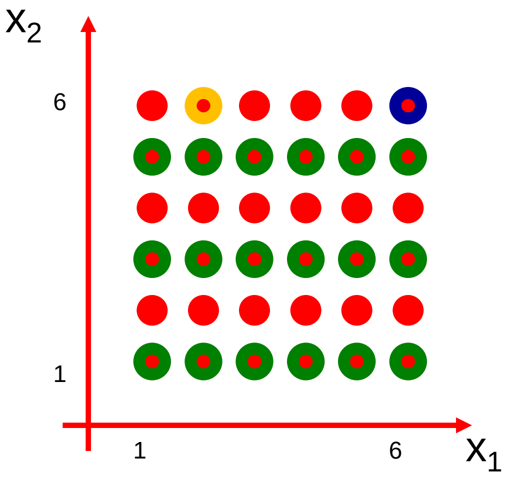


- Examples:
  - If I flip a coin 100 times, how many can I expect to see heads?
  - What is the weather going to be like tomorrow?
  - Are my stocks going to be up or down in value?

Sample Space = Universe of Outcomes

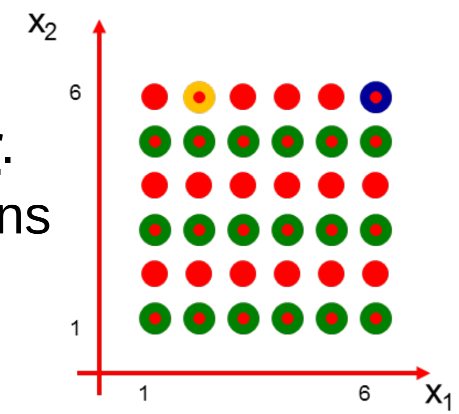
- The most fundamental concept is that of a ***Sample Space*** (denoted by Ω or ***S*** or **U**), also called the ***Universal Set***.
- A ***Random Experiment*** takes values in a set of ***Outcomes***
  - The outcomes of the random experiment are used to define ***Random Events***
    - ***Event = Set of Possible Outcomes***

- Example of a ***Random Experiment***:
  - Roll a single die ***twice consecutively***
  - call the value on the up face at the n^{th} toss x_n for $n = 1,2$
  - E.g., two possible experimental **outcomes**:
    - **two sixes** ($x_1 = x_2 = 6$)
    - $x_1 = 2$ **and** $x_2 = 6$
- Example of a ***Random Event***:
  - **An odd number occurs on the 2nd toss**.



Sample Space = Universal Event

- The **sample space U** is a set of experimental outcomes that **must** satisfy the following two properties:
  - **Collectively Exhaustive**: ***all*** possible outcomes are listed in **U** ***and*** when an experiment is performed ***one of these outcomes must occur***.
  - **Mutually Exclusive**: ***only one*** outcomes happens and no other can occur (if $x_1 = 5$ it cannot be anything else).



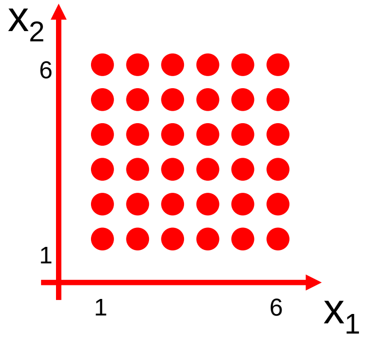
- The mutually exclusive property of *outcomes* simplifies the calculation of the probability of *events*
- Collectively Exhaustive means that there is no possible event to which we cannot assign a probability
- The **Universe U** (= sample space) of possible experimental outcomes is equal to the **event "Something Happens"** when an experiment is performed. Thus we also call **U** the ***Universal Event***

Probability Measure

- ***Probability*** of an ***event*** **:**
  - A positive real number between 0 and 1 expressing the chance that the **event** will occur when a random experiment is performed.
- A ***probability measure*** satisfies the

  Three Kolmogorov Axioms:

  - $P(A) \geq 0$ for any event A (every event A is a **subset** of **U**)
  - $P(\mathbf{U}) = P(\text{Universal Event}) = 1$ (because "something must happen")
  - $\text{if } A \cap B = \varnothing$, then $P(A \cup B) = P(A) + P(B)$

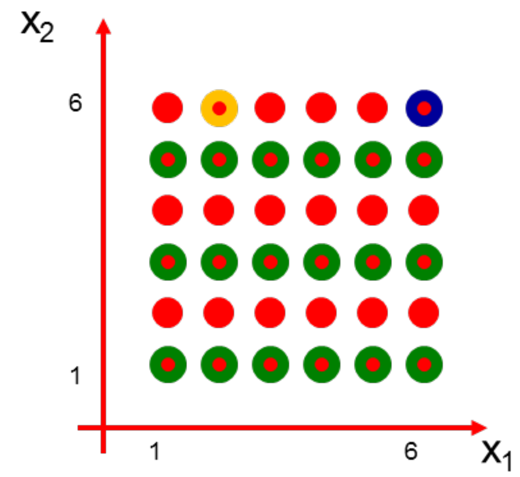


- e.g.
  - $P\left(\{x_1 \geq 0\}\right) = 1$
  - $P\left(\{\, x_1\, even\, \} \cup \{\, x_1\, odd\, \}\right) = P\left(\{\, x_1\, even\, \}\right) + P\left(\{\, x_1\, odd\, \}\right)$

Probability Measure

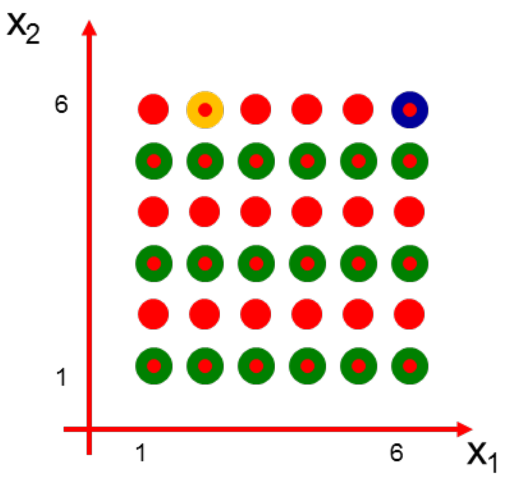
- The last axiom of the three, when combined with the mutually exclusive property of the sample set,
  - allows us to easily assign probabilities to all possible events if the probabilities of ***atomic events***, aka ***elementary events***, are known

- Back to our dice example:
  - Suppose that the probability of the ***elementary event*** consisting of ***any*** single outcome-pair, $A = \{(x_1,x_2)\}$, is $P(A) = 1/36$
  - We can then compute the probabilities of all events, including ***compound events***:
    - P(**x_2 odd**) = 18x1/36 = 1/2
    - P(**U**) = 36x1/36 = 1
    - P(**two sixes**) = 1/36
    - P(**x_1 = 2 and x_2 = 6**) = 1/36



Probability Measure

- Note that there are many ways to decompose the universal event **U** (the "ultimate" compound event) into the disjoint union of simpler events:



  – E.g. if $A = \{x_2 \text{ odd}\}$, $B = \{x_2 \textit{ even}\}$, then $\mathbf{U} = A \cup B$

  – on the other hand
  $\mathbf{U} = \{(1,1)\} \cup \{(1,2)\} \cup \{(1,3)\} \cup \cdots \cup \{(6,6)\}$

  – The fact that the sample space is exhaustive and mutually exclusive, combined with the three probability measure (Kolmogorov) axioms makes the whole procedure of computing the probability of a compound event from the probabilities of simpler events consistent.

Random Variables

- A ***random variable X***
  - is a ***function*** that assigns a real value to each sample space outcome
  - we have already seen one such function: $P_{\mathbf{X}}(\{x_1,x_2\}) = 1/36$ for all outcome-pairs (x_1,x_2) (viewing an outcome as an atomic event)

- Most Precise Notation:
  - Specify both the random variable, ***X***, and the value, ***x***, that it takes in your probability statements. E.g., $\boldsymbol{X}(u) = \boldsymbol{x}$ for any outcome u in $\mathbf{U}$.
  - In a *probability measure*, specify the random variable as a subscript, $P_{\boldsymbol{X}}(x)$,and the value x as the argument.
    For example

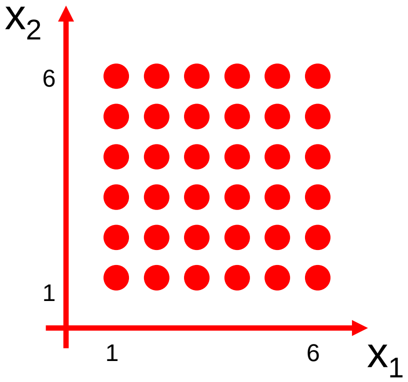
$$P_{\boldsymbol{X}}(x) = P_{\boldsymbol{X}}(x_1,x_2) = 1/36$$

    means $\mathrm{Prob}[\boldsymbol{X} = (x_1,x_2)] = 1/36$
  - Without such care, probability statements can be hopelessly confusing

Random Variables

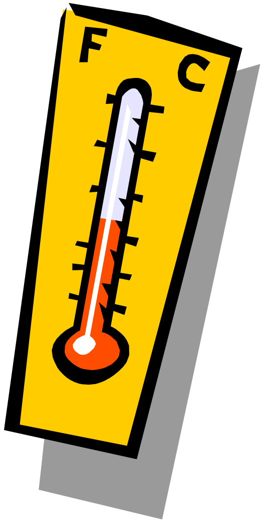
- Types of random variables:
  - ***discrete*** and ***continuous*** (and sometimes ***mixed***)
  - Terminology relates to what types of values the RV can take

- If the RV can take only one of a ***finite or*** at most ***countable*** set of possibilities, we call it ***discrete.***
  - If there are furthermore ***only*** a **finite** set of possibilities, the discrete RV is ***finite***. For example, in the two-throws-of-a-die example, there are only (at most) 36 possible values that an RV can take:



Random Variables

- If an RV can take arbitrary values in a ***real interval*** we say that the random variable is ***continuous***
- E.g. consider the sample space of weather temperature
  - we know that it could be any number between -50 and 150 degrees Celsius
  - random variable $T \in [-50,150]$
  - note that the extremes do not have to be very precise, we can just say that $P(T < -45^o) = 0$

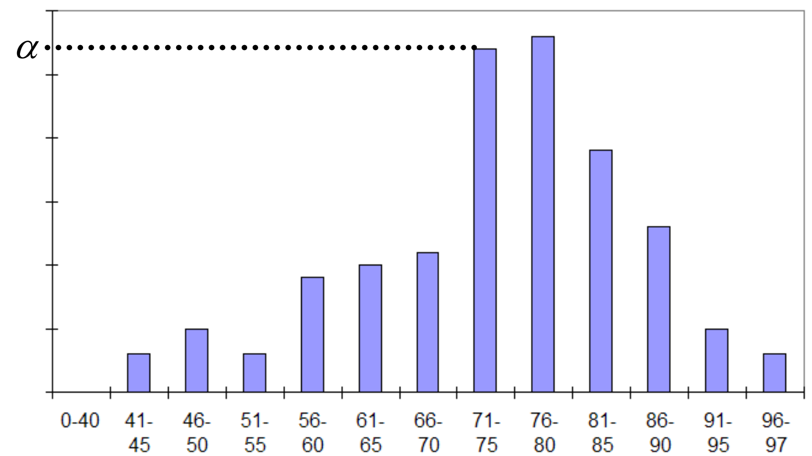


- Most probability notions apply equal well to discrete and continuous random variables

Discrete RV

- For a discrete RV the probability assignments given by a ***probability mass function*** (**pmf**)
  - this can be thought of as a normalized histogram
  - satisfies the following properties

$$0 \le P_X(a) \le 1, \quad \forall a$$
$$\sum_a P_X(a) = 1$$

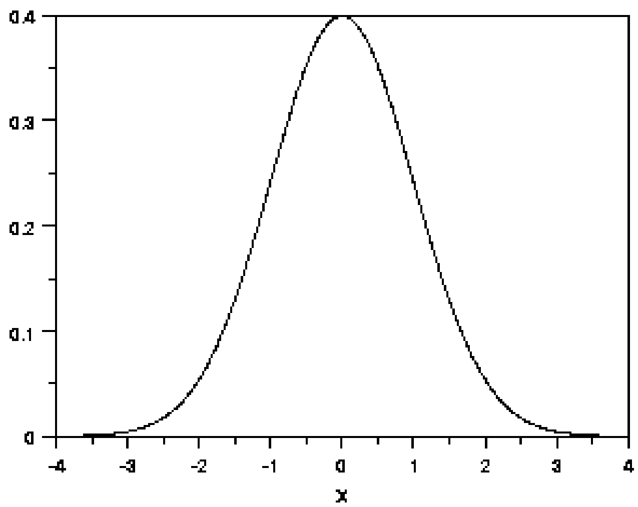


- Example of a discrete (and finite) random variable
  - $X \in \{1,2,3, \dots , 20\}$ where $X = i$ if the grade of student z on class is greater than $5(i-1)$ and less than or equal to $5i$
  - We see from the discrete distribution plot that $P_{\mathbf{X}}(15) = \alpha$

Continuous RV

- For a continuous RV the probability assignments are given by a ***probability density function*** (**pdf**)
  - this is a piecewise continuous function that satisfies the following properties

$$0 \le P_X(a) \quad \forall a$$
$$\int P_X(a)\,da = 1$$



- Example for a Gaussian random variable of mean μ and variance σ^2

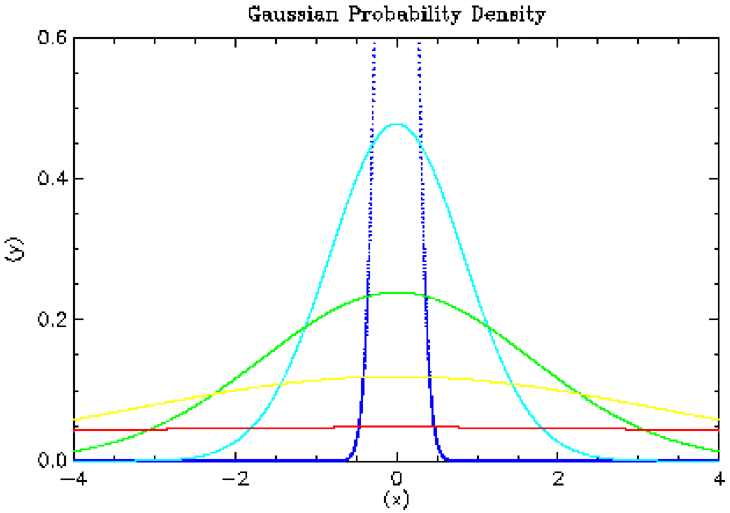
$$P_X(a) = \frac{1}{\sqrt{2\pi}\sigma}\exp\left\{-\frac{(a-\mu)^2}{2\sigma^2}\right\}$$

Discrete vs Continuous RVs

- In general the math is the same, up to replacing summations by integrals
- Note that **pdf** means "density of the probability",
  - This is probability per unit "area" (e.g., ***length*** for a scalar rv).
  - The probability of a particular value $X = t$ of a ***continuous*** RV X is always zero
    - Nonzero probabilities arise as:

$$\Pr(t \le X \le t + dt) = P_X(t)dt$$

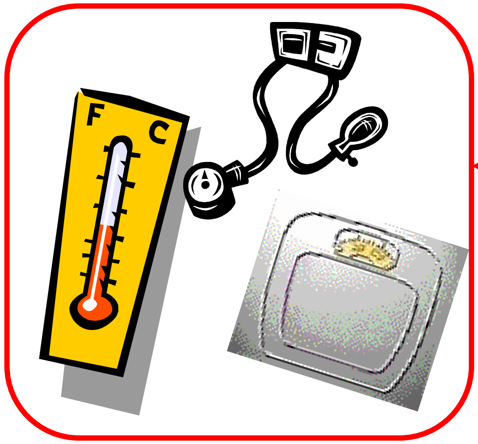
$$\Pr(a \le X \le b) = \int_a^b P_X(t)dt$$

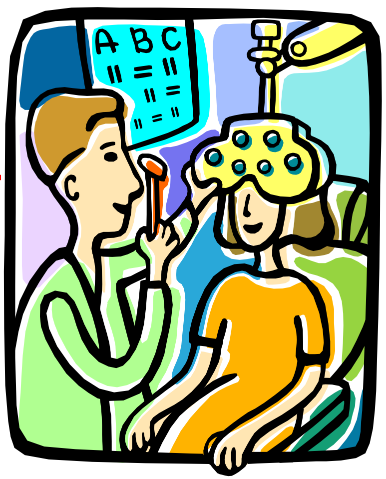


  - Note also that pdfs are *not* necessarily upper bounded
    - e.g. Gaussian goes to Dirac delta function when variance goes to zero

Multiple Random Variables

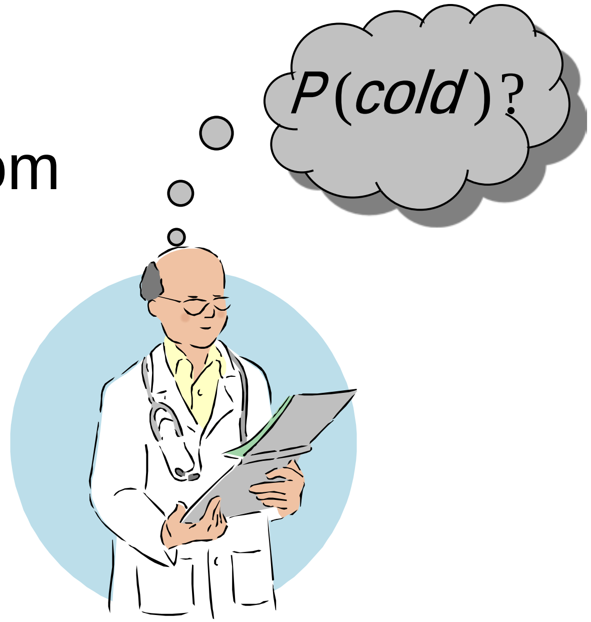
- Frequently we have to deal with ***<u>multiple random variables</u>*** aka ***<u>random vectors</u>***
  - e.g. a doctor's examination measures a ***<u>collection</u>*** of random variable values:
    - x_1: temperature
    - x_2: blood pressure
    - x_3: weight
    - x_4: cough
    - …





- We can summarize this as
  - a vector $\boldsymbol{X} = (X_1, \ldots, X_n)^T$ of n random variables
    - $P_X(x_1, \ldots, x_n)$ is the ***<u>joint</u> probability distribution***

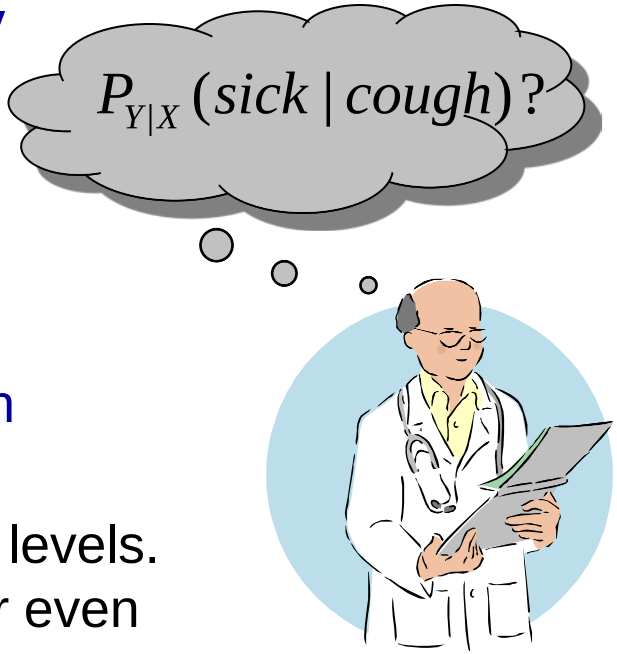
Marginalization



- An important notion for multiple random variables is ***marginalization***
  - e.g. having a cold does not depend on blood pressure and weight
  - all that matters are fever and cough
  - that is, we only need to know $P_{X1,X4}(a,b)$
- We ***marginalize*** with respect to a subset of variables
  - (in this case X_1 and X_4)
  - this is done by ***summing (or integrating) the others out***

$$\boxed{\begin{aligned} P_{X_1,X_4}(x_1,x_4) &= \sum_{x_3,x_4} P_{X_1,X_2,X_3,X_4}(x_1,x_2,x_3,x_4) \\ P_{X_1,X_4}(x_1,x_4) &= \int\!\!\int P_{X_1,X_2,X_3,X_4}(x_1,x_2,x_3,x_4)\,dx_2\,dx_3 \end{aligned}}$$

Conditional Probability

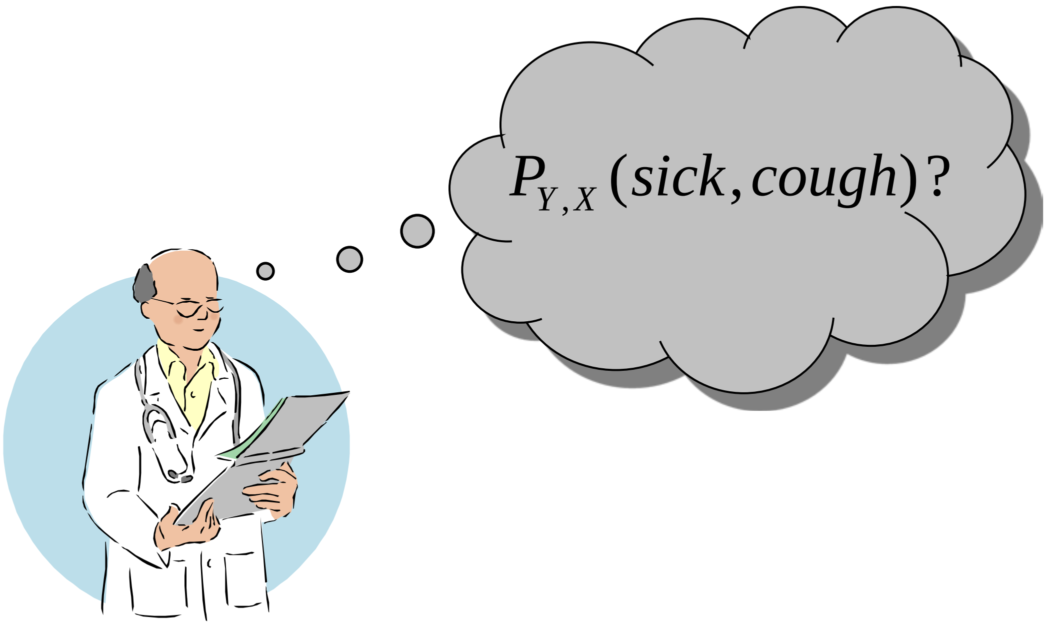


- Another ***very important*** notion:
  - So far, doctor has $P_{X1,X4}(fever,cough)$
  - Still does not allow a diagnosis
  - For this we need a new variable Y with two states $Y \in \{sick, not\ sick\}$
  - Doctor ***measures*** the fever and cough levels. These are now no longer unknowns, or even (in a sense) random quantities.
  - The question of interest is "what is the probability that patient is sick ***given*** the measured values of fever and cough?"
- This is exactly the definition of ***conditional probability***
  - E.g., what is the probability that "Y = sick" ***given*** observations "X_1 = 98" and "X_4 = high"? We write this probability as:

$$\boxed{P_{Y|X_1,X_4}(sick \mid 98, high)}$$

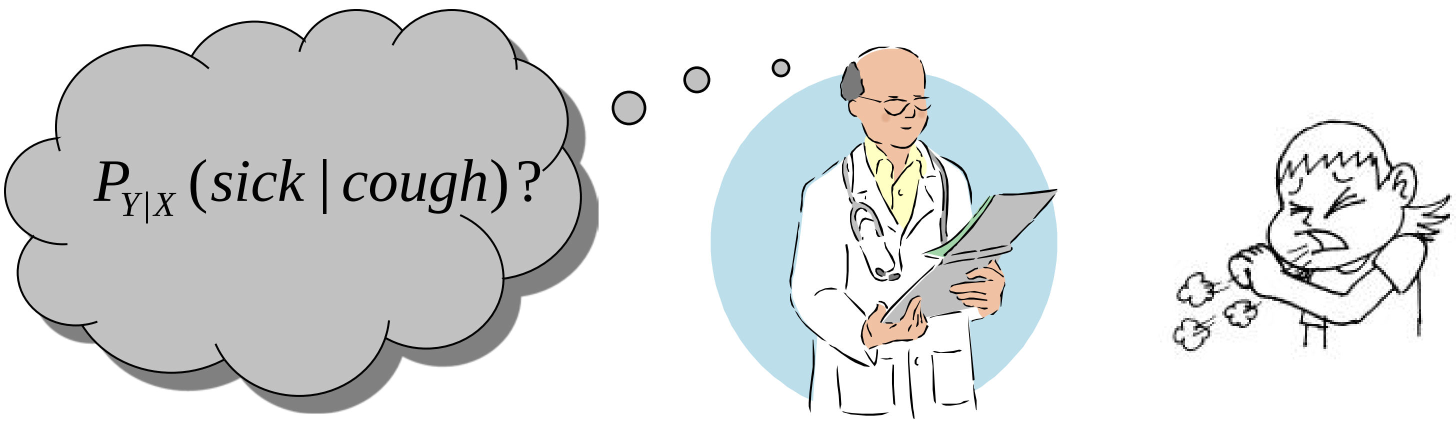
Joint versus Conditional Probability

- Note the very important difference between conditional and joint probability
- Joint probability corresponds to an hypothetical question about probability over ***all*** random variables
  - E.g., what is the probability that you will be sick ***and*** cough a lot?



Conditional Probability

- Conditional probability means that you ***<u>know</u> the values of <u>some</u> variables***, while ***the <u>remaining</u> variables are <u>unknown</u>***.
  - E.g., this leads to the question: what is the probability that you are sick ***given*** that you cough a lot?



  - “given” is the key word here
  - **<u>conditional probability is very important</u>** because it allows us to structure our thinking
  - shows up again and again in design of intelligent systems

Conditional Probability

- Fortunately it is easy to compute (via a consistent ***definition***)
  - We simply normalize the joint by the probability of what we know

$$\boxed{P_{Y|X_1}(sick \mid 98) = \frac{P_{Y,X_1}(sick, 98)}{P_{X_1}(98)}}$$

  - Makes sense since the conditional probability is then nonnegative, and

$$P_{Y|X_1}(sick \mid 98) + P_{Y|X_1}(not\ sick \mid 98) = 1$$

  as a consequence of the definition and the marginalization equation,

$$P_{Y,X_1}(sick, 98) + P_{Y,X_1}(not\ sick, 98) = P_{X_1}(98)$$

  - The definition of conditional probability is such that
    - Conditioned on what we know, we ***still*** have a valid probability measure
    - In particular, the ***new (restricted) universal event*** of interest, {*sick*} *U* {*not sick*}, has probability 1 after conditioning on the temperature observation

The Chain Rule of Probability

- An important consequence of the definition of conditional probability
  - note that the definition can be ***equivalently written*** as

$$P_{X_1,X_2}(x_1,x_2) = P_{X_2|X_1}(x_2 \mid x_1)P_{X_1}(x_1)$$

  - By recursion on this definition, more generally we have the product ***chain rule of probability***:

$$\boxed{\begin{aligned} P_{X_1,\dots,X_n}(x_1,\dots,x_n) = {} & P_{X_1|X_2,\dots,X_n}(x_1 \mid x_2,\dots,x_n)\times \\ & P_{X_2|X_3\dots,X_n}(x_2 \mid x_3,\dots,x_n)\times\dots \\ & \times\dots\times P_{X_{n-1}|X_n}(x_{n-1} \mid x_n)P_{X_n}(x_n) \end{aligned}}$$

- Combining this rule with the marginalization procedure allows us to make difficult probability questions simpler

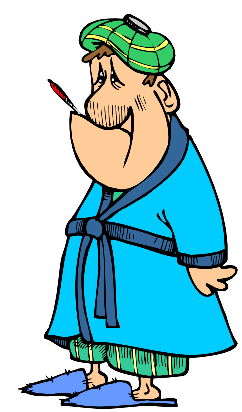
The Chain Rule of Probability

- E.g. what is the probability that you will be sick and have 104º F of fever?

$$P_{Y,X_1}(sick,104) = P_{Y|X_1}(sick\,|\,104)P_{X_1}(104)$$

  - breaks down a hard question (prob of sick and 104) into two easier questions
  - Prob (sick|104): everyone knows that this is close to one

$$P_{Y|X}(sick\,|\,104) = 1!$$

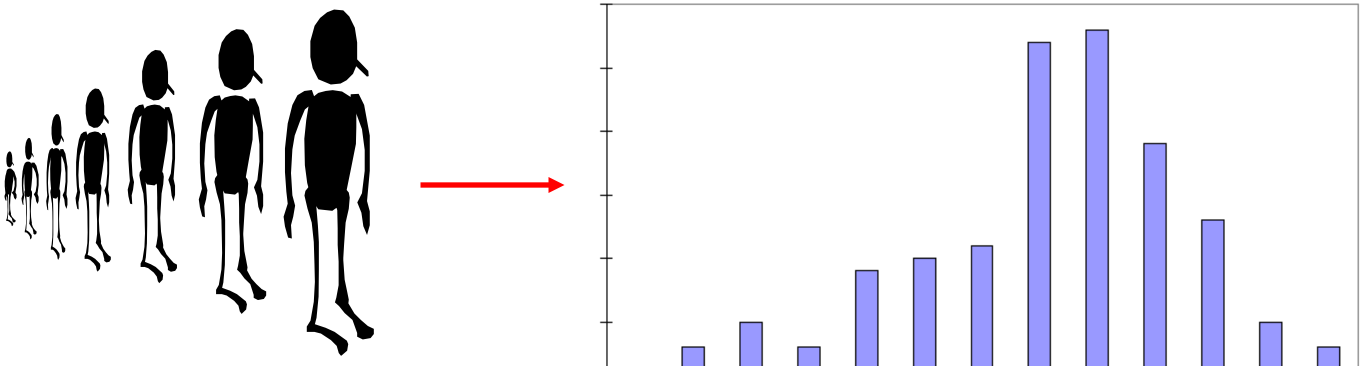


The Chain Rule of Probability

- E.g. what is the probability that you will be sick and have 104º of fever?

$$P_{Y,X_1}(sick,104) = P_{Y|X_1}(sick\,|\,104)P_{X_1}(104)$$

  – Computing P(104) is still hard, but easier than P(sick,104) since we now only have *one* random variable (temperature)
    - P(104) does not depend on sickness, it is just the question "what is the probability that someone will have 104º?"
    - gather a number of people, measure their temperatures and make an histogram that everyone can use after that



The Chain Rule of Probability

- In fact, the chain rule is so handy, that most times we use it to compute marginal probabilities
  - e.g. $P_Y(sick) = \int P_{Y,X_1}(sick, t)dt$ (marginalization)

$$= \int P_{Y|X_1}(sick \mid t) P_{X_1}(t) dt$$

  - in this way we can get away with knowing
    - ***$P_{X_1}(t)$***, which we may know because it was needed for some other problem
    - ***$P_{Y|X_1}($ sick $\mid t)$***, we can ask a doctor (a so-called ***domain expert***), or approximate with a rule of thumb

$$P_{Y|X_1}(sick \mid t) \approx \begin{cases} 1 & t > 102 \\ 0.5 & 98 < t < 102 \\ 0 & t < 98 \end{cases}$$

Independence

- Another fundamental concept for multiple variables
  - Two variables are ***independent*** if the joint is the product of the marginals:

$$\boxed{P_{X_1,X_2}(a,b) = P_{X_1}(a)P_{X_2}(b)}$$

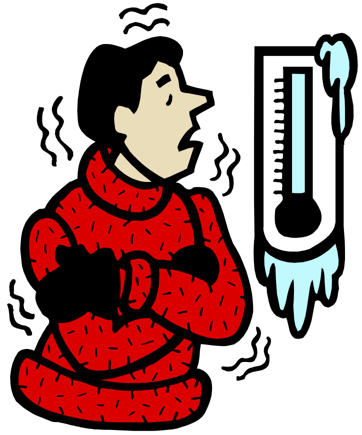
  - Note: This is ***equivalent*** to the statement:

$$\boxed{P_{X_1|X_2}(a|b) = \frac{P_{X_1,X_2}(a,b)}{P_{X_2}(b)} = P_{X_1}(a)}$$

  which captures the intuitive notion:

  - "if X_1 is ***independent*** of X_2, knowing X_2 does ***not*** change the probability of X_1"
    - e.g. knowing that it is sunny today does not change the probability that it will rain in three years

Conditional Independence



- ***Extremely useful*** in the design of intelligent systems
  - Sometimes knowing X makes Y independent of Z
  - E.g. consider the shivering symptom:
    - if you have temperature you sometimes shiver
    - it is a symptom of having a cold
    - but once you measure the temperature, the two become independent

$$
\begin{aligned}
P_{Y,X_1,S}(sick, 98, shiver) &= P_{Y|X_1,S}(sick \mid 98, shiver) \times \\
&\quad P_{S|X_1}(shiver \mid 98) P_{X_1}(98) \\
&= P_{Y|X_1}(sick \mid 98) \times \\
&\quad P_{S|X_1}(shiver \mid 98) P_{X_1}(98)
\end{aligned}
$$

- Simplifies considerably the estimation of probabilities

Independence

- Useful property: if you ***add*** two ***independent*** random variables their probability distributions convolve
  - I.e. if *Z = X + Y* and *X,Y* are independent then

$$\boxed{P_Z(z) = P_X(z) * P_y(z)}$$

where * is the convolution operator

  - For discrete random variables, this is:

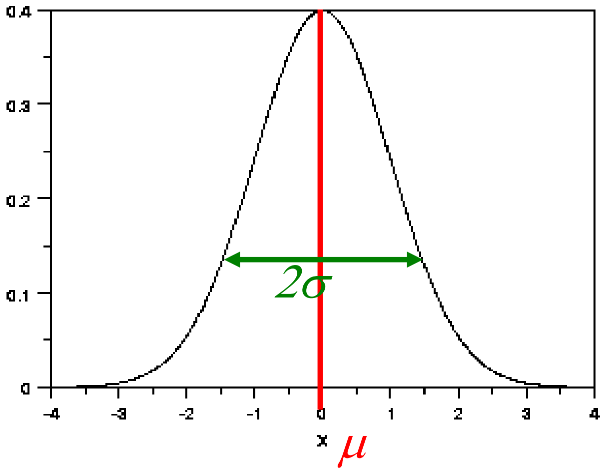
$$\boxed{P_Z(z) = \sum_k P_X(k) P_Y(z-k)}$$

  - For continuous random variables, this is

$$\boxed{P_Z(z) = \int P_X(t) P_Y(z-t) dt}$$

Moments

- Moments are important properties of random variables
  - They summarize the distribution



- The two most Important moments
  - mean: $\mu = E[X]$
  - variance: $\sigma^2 = Var(X) = E[(X-\mu)^2]$

	discrete	continuous
mean	$\mu = \sum_k P_X(k)\,k$	$\mu = \int P_X(k)\,k\,dk$
variance	$\sigma^2 = \sum_k P_X(k)(k-\mu)^2$	$\sigma^2 = \int P_X(k)\,(\mathrm{k}-\mu)^2\,dk$

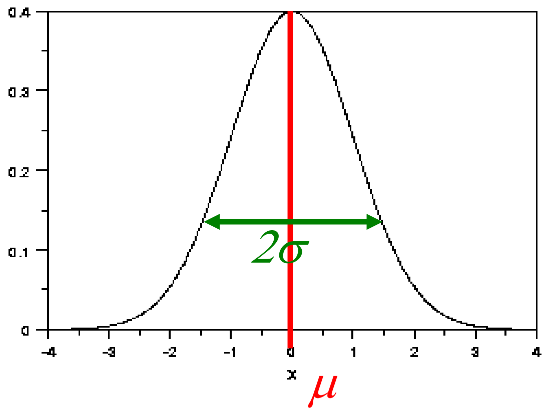
- "Nice" distributions are completely specified by a very few moments. E.g., the Gaussian by the mean and variance.

Mean

- $\mu = E[x]$, is the center of probability mass of the distribution

	discrete	continuous
mean	$\mu = \sum_k P_X(k)\, k$	$\mu = \int P_X(k)\, k\, dk$

- Mean is a linear function of its argument
  - if $Z = X + Y$, then $E[Z] = E[X] + E[Y]$
  - this does *not* require any special relation between X and Y
  - always holds



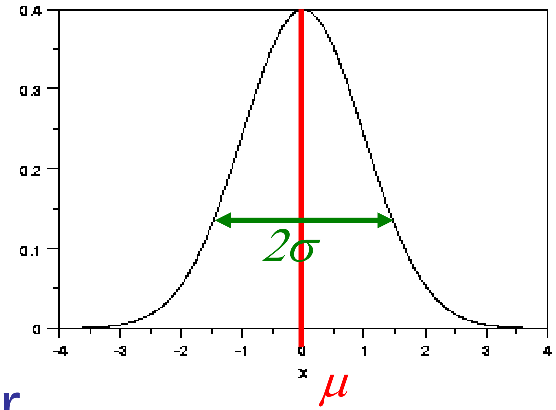
- The other moments are the ***mean of the powers*** of X
  - n^{th} order (***non-central***) moment is $E[X^n]$
  - n^{th} ***central*** moment is $E[(X-\mu)^n]$

Variance

- $\sigma^2 = E[(x-\mu)^2]$ measures the dispersion around the mean (= 2nd central moment)

	Discrete	Continuous
variance	$\sigma^2 = \sum_k P_X(k)(k-\mu)^2$	$\sigma^2 = \int P_X(k)\,(k-\mu)^2\,dk$

- **in general**, it is ***not*** a linear function
  – if $Z = X + Y$, then $\text{Var}[Z] = \text{Var}[X] + \text{Var}[Y]$ only holds **if** X and Y are **independent**



- The variance is related to the 2nd order non-central moment by $\sigma^2 = E\left[(x-\mu)^2\right] = E\left[x^2 - 2x\mu + \mu^2\right]$

$$= E\left[x^2\right] - 2E[x]\mu + \mu^2 = E\left[x^2\right] - \mu^2$$

Any questions?