Some Concepts of Probability (Review)

Volker Tresp
Summer 2016

Ockham chooses a razor

Summary

- Conditional probability:

$$P(y|x) = \frac{P(x,y)}{P(x)} \quad \text{with} \quad P(x) > 0$$

- Product rule

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

- Chain rule

$$P(x_1, \ldots, x_M) = P(x_1)P(x_2|x_1)P(x_3|x_1,x_2)\ldots P(x_M|x_1,\ldots,x_{M-1})$$

- Bayes theorem

$$P(y|x) = \frac{P(x,y)}{P(x)} = \frac{P(x|y)P(y)}{P(x)} \quad P(x) > 0$$

- Marginal distribution

$$P(x) = \sum_y P(x, y)$$

- Independent random variables

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

Discrete Random Variables

- A **random variable** $X(c)$ is a variable (more precisely a function), whose value depends on the result of a random process
- Examples:
  - c is a coin toss and $X(c) = 1$ if the result is head
  - c is a person, randomly selected from the University of Munich. $X(c)$ is the height of that person
- A **discrete random variable** X can only assume a countable number of states. Thus $X = x$ with $x \in \{x_1, x_2, \ldots\}$

Discrete Random Variables (2)

- A probability distribution specifies with which probability a random variable assumes a particular state
- A probability distribution of X can be defined via a **probability function** $f(x)$:

$$P(X = x) = P(\{c : X(c) = x\}) = f(x)$$

- $f(x)$ is the probability function and x is a realisation of X
- One often writes

$$f(x) = P_X(x) = P(x)$$

Elementary / Atomic Events

- In statistics, one attempts to derive the probabilities from data (machine learning)
- In probability one assumes either that some probabilities are known, or that they can be derived from some atomic events
- **Atomic event**: using some basic assumptions (symmetry, neutrality of nature, fair coin, ...) one assumes the probabilities for some elementary events

Example: Toss of a Fair Coin

- Atomic events: $c = \{h, t\}$
- The probability of each elementary event is $1/2$
- $X(c)$ is a random variable that is equal to one if the result is head and is zero otherwise
- $P(X = 1) = 1/2$

Random Variables

- From now on we will not refer to any atomic event; for complex random variables like the height or the weight of a person, it would be pretty much impossible to think about the atomic events that produced height and weight

- We directly look at the random variables and their dependencies

- The running example will be the distribution of height H and weight W of students in Munich. For simplicity we assume that there are only two states for either variables: $H = t$ for a tall person and $H = s$ for a small person. Similarly, $W = b$ for a big person and $W = l$ for a light person

Univariate Probabilities

Sample size 100

P(H=t) = 0.5	(in the sample, 50 persons were tall)
P(H=s)=0.5	(in the sample, 50 persons were small)
P(W=b)=0.6	(in the sample, 60 persons were big)
P(W=l)=0.3	(in the sample, 30 persons were light)

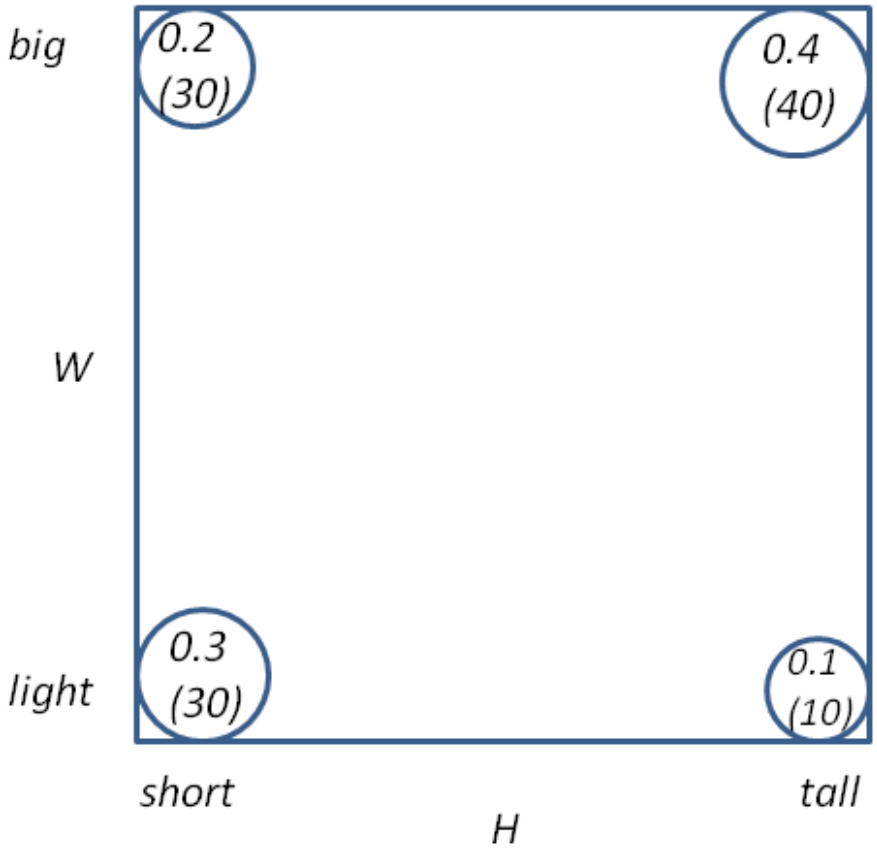
Multivariate Probability Distributions

- Define two random variables $X(c)$ and $Y(c)$. A **multivariate distribution** is defined as:

$$P(x, y) = P(X = x, Y = y) = P(X = x \wedge Y = y)$$

- Note that defines the probability of a *conjunction*!

Multivariate Probabilities



P(H=t,W=b) = 0.4 (in the sample, 40 persons were tall and big)

Special Cases

- If two random variables are independent, then $P(X, Y) = P(X)P(Y)$. This is not the case in our example since $P(t, b) = 0.4 \neq P(t)P(b) = 0.5 \times 0.6 = 0.3$
- Two random variables can be mutually exclusively true: $P(X = 1, Y = 1) = 0$. Also not the case in our example (we identify b and t with true)
- If M binary random variables $X_1, \ldots, X_M$ are all mutually exclusive and collectively exhaustive (i.e., exactly one variable assumes the state 1 in a given sample), then the M variables can be represented by one random variable with M states

Mutual Exclusive and Exhaustive Random Variables

A person belongs to exactly one age class

- *4 binary random variables that are mutually exclusive and collectively exhaustive*
  - *Teen=false, YoungAdult=true, Adult=false, MiddleAge=false*

- *1 discrete random variable with 4 states*
  - *Age=YoungAdult*

Which Random Variables?

- It should be clear from the discussion that the definition of random variables in a domain is up to the researcher, although there is often a "natural" choice (height of a person, income of a person, age of a person, ...)

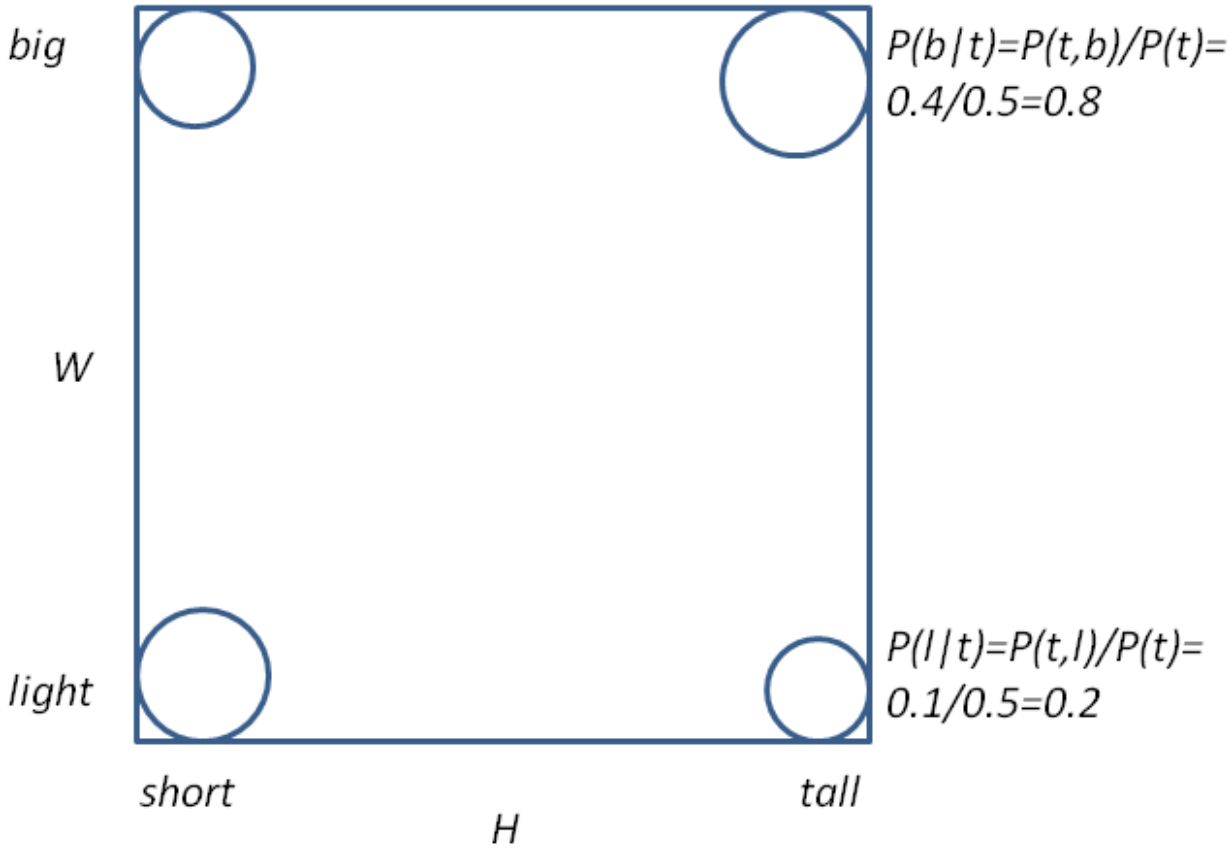
Conditional Distribution

- I am interested in the probability distribution of the random variable Y but consider only atomic events, where $X = x$
- *Definition* of a **conditional probability distribution**

$$P(Y = y|X = x) = \frac{P(X = x, Y = y)}{P(X = x)} \text{ with } P(X = x) > 0$$

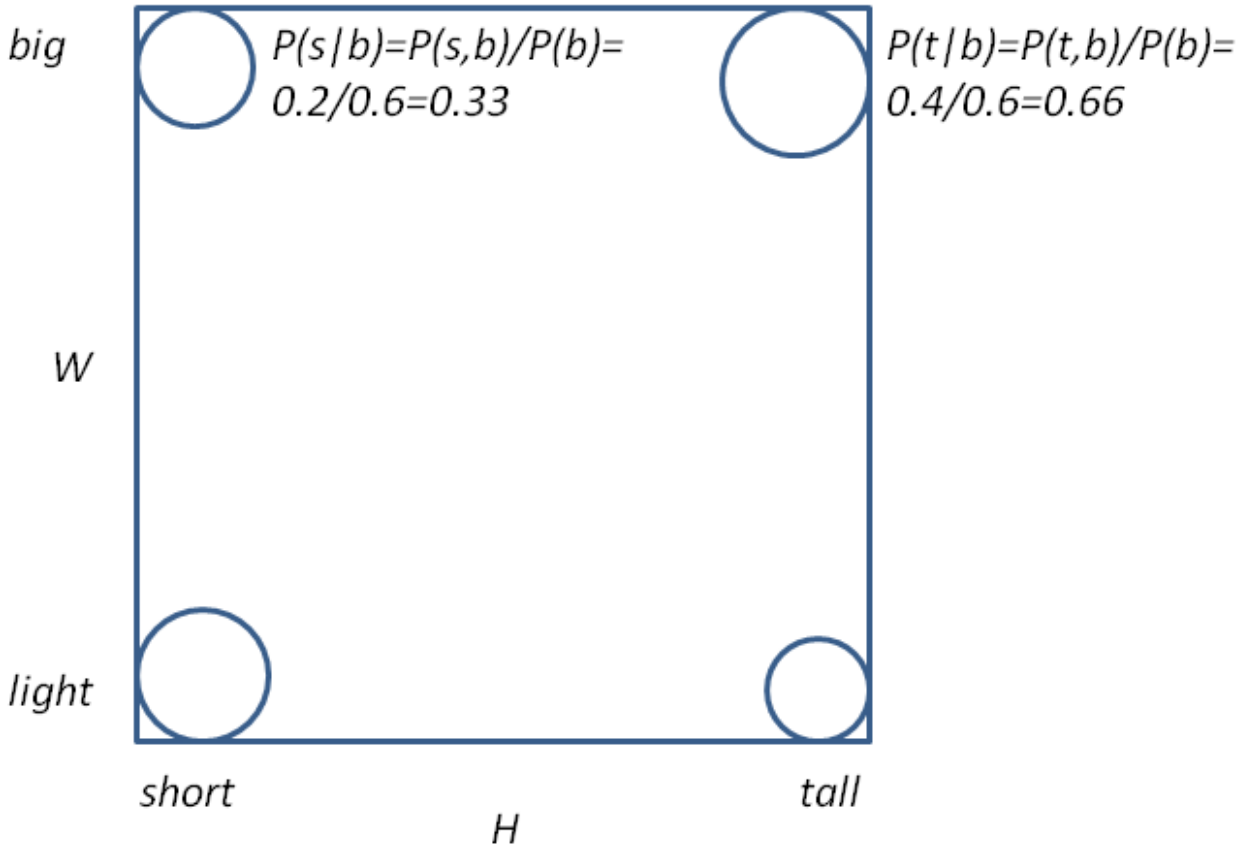
- The distribution is identical to the one for the unconditional case, only that I have to divide by $P(X = x)$ (re-normalize)

Conditional Probabilities *P(W|H)*



The probability that a person is big, given that this person is tall, is 0.8

Conditional Probabilities *P(H|W)*



The probability that a person is tall, given that this person is big, is 0.66

Product Rule and Chain Rule

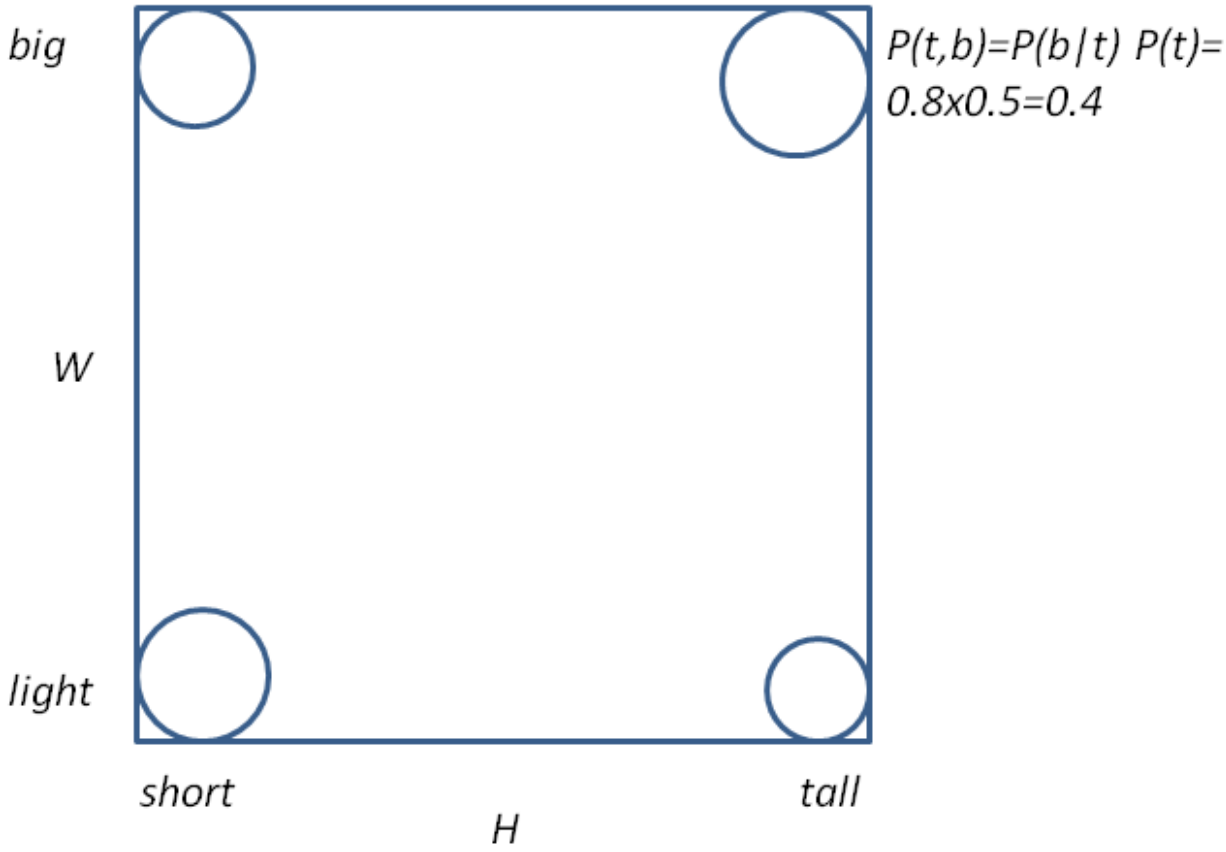
- It follows: **product rule**

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

- and **chain rule**

$$P(x_1, \ldots, x_M) = P(x_1)P(x_2|x_1)P(x_3|x_1, x_2) \ldots P(x_M|x_1, \ldots, x_{M-1})$$

Product Rule



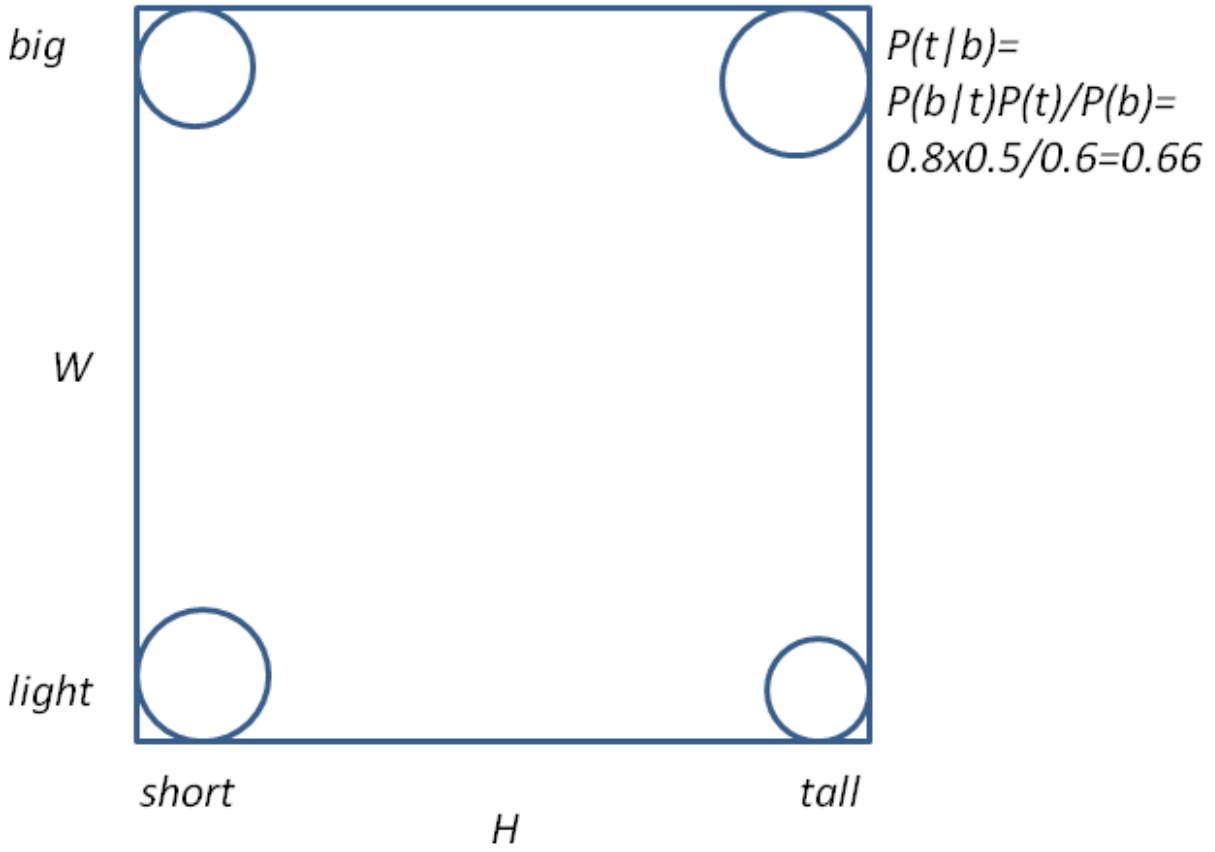
The probability that a person is tall and that this person is big --- is the same as the probability that a person is big, given that this person is tall, times the probability that this person is tall

Bayes Theorem

- **Bayes Theorem**

$$P(y|x) = \frac{P(x,y)}{P(x)} = \frac{P(x|y)P(y)}{P(x)} \quad P(x) > 0$$

Bayes Theorem



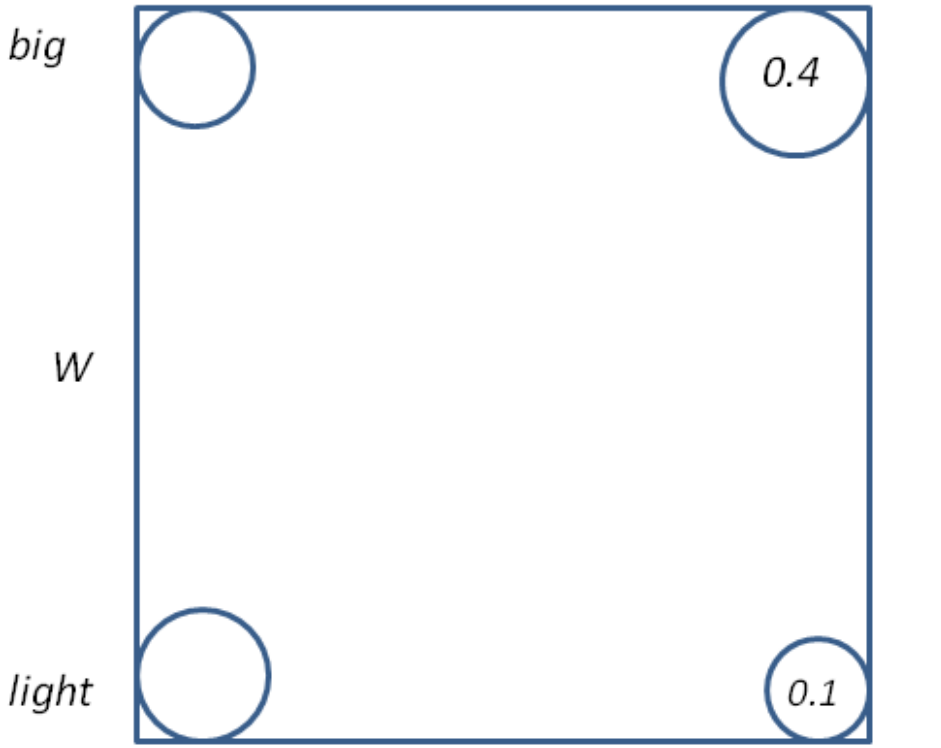
The probability that this person is tall, given that this person is big --- is the same as the probability that someone is big given that this person is tall, multiplied by the probability that this person is tall divided by the probability that this person is big

Marginal Distribution

- The **marginal distribution** can be calculated from a joint distribution as:

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

Marginal Probability



$P(t) = P(t, b) + P(t,l) = 0.4+0.1=0.5$

The probability that a person is tall --- is the probability that someone is tall and big plus the probability that someone is tall and light

General (Logical) Expression (Query)

- Example: $\Phi = X \vee Y \wedge Z$. What is $P(\Phi = true)$?
- We can write the joint as: $P(\Phi, X, Y, Z) = P(\Phi|X, Y, Z)P(X, Y, Z)$
- The **marginal distribution** can be calculated from a joint distribution as:

$$P(\Phi = true) = \sum_{x,y,z} P(\Phi = true|x, y, z)P(x, y, z)$$

$$\sum_{x,y,z:\Phi=true} P(x, y, z)$$

Special Case: Disjunction

- We get for the **disjunction**

$$P(X = 1 \ \vee \ Y = 1) =$$

$$P(X = 1, Y = 0) + P(X = 0, Y = 1) + P(X = 1, Y = 1) =$$

$$(P(X = 1, Y = 0) + P(X = 1, Y = 1)) + (P(X = 0, Y = 1) + P(X = 1, Y = 1)$$

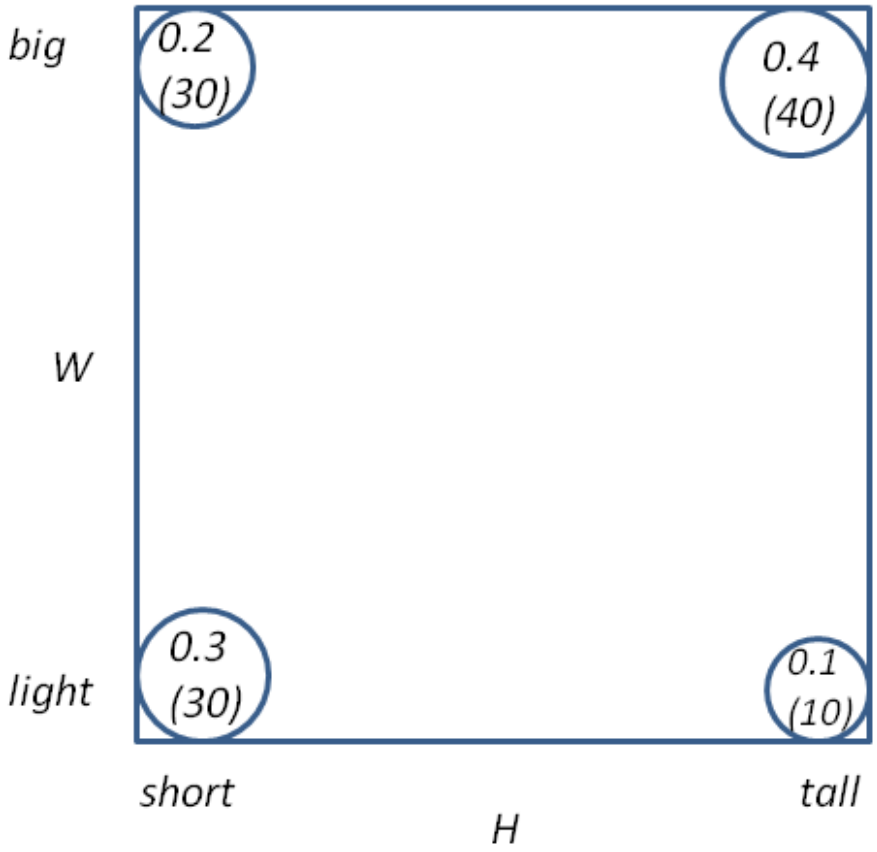
$$-P(X = 1, Y = 1)$$

$$= P(X = 1) + P(Y = 1) - P(X = 1, Y = 1)$$

- Only if states are **mutually exclusive**, $P(X = 1, Y = 1) = 0$, then

$$P(X = 1 \vee Y = 1) = P(X = 1) + P(Y = 1)$$

Disjunction



P((H=t) OR (W=b)) =0.2+0.4+0.1=0.7 =P(t)+P(b)-P(t, b)= 0.5+0.6-0.4 = 0.7

The probability that a person is tall OR that a person is big--- is the probability that someone is short and big plus the probability that someone is tall and big plus the probability that someone is tall and light

Marginalization and Conditioning: Basis for Probabilistic Inference

- $P(I, F, S)$ where $I = 1$ stands for influenza, $F = 1$ stands for fever, $S = 1$ stands for sneezing
- What is the probability for influenza, when the patient is sneezing, but temperature is unknown, $P(I|S)$?
- Thus I need (conditioning) $P(I = 1|S = 1) = P(I = 1, S = 1)/P(S = 1)$
- I calculate via marginalization

$$P(I = 1, S = 1) = \sum_f P(I = 1, F = f, S = 1)$$

$$P(S = 1) = \sum_i P(I = i, S = 1)$$

Independent Random Variables

- **Independence**: two random variables are independent, if,

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

- It follows for independent random variables,

$$P(X = x \vee Y = y) = P(X = x) + P(Y = y) - P(X = x)P(Y = y)$$

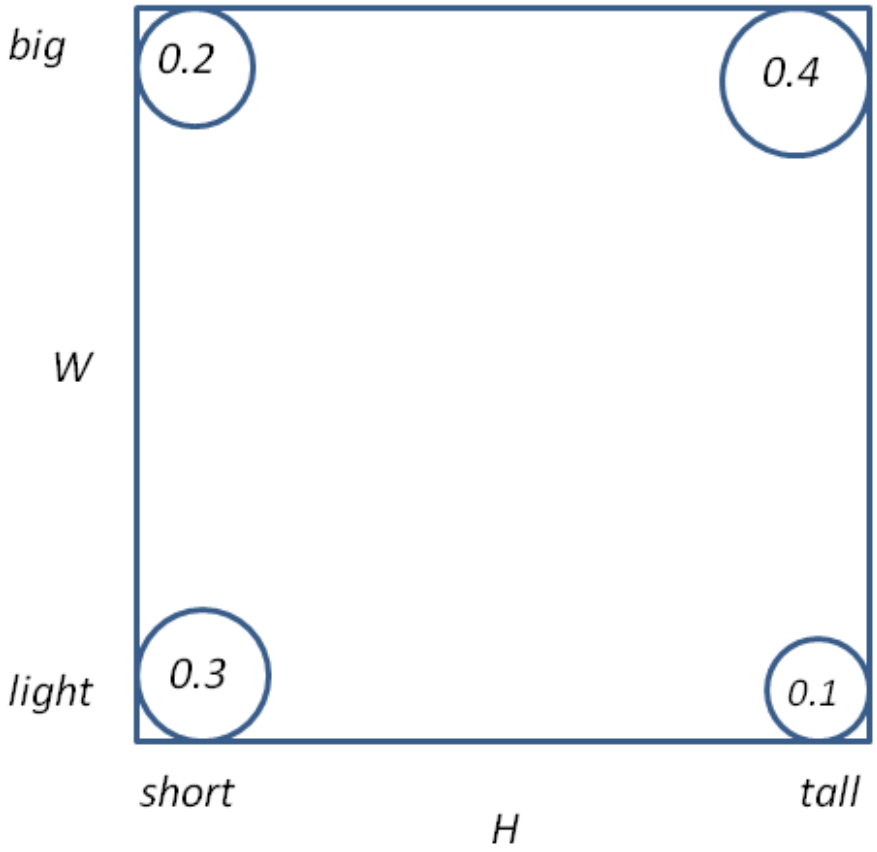
Expected Values

- **Expected value**

$$E(X) = E_{P(x)}(X) = \sum_i x_i P(X = x_i)$$

Expected Value

Let's associate with tall 180cm, with short 150cm, with big 100kg, and with light 50kg.



E(Height) = 0.5x180cm+0.5x150cm=165cm
E(Weight)=0.6x100kg+0.4x50kg= 80kg
We can also calculate E(Weight|H=t)= 0.8x100kg+0.2x50kg= 90kg

Variance

- The **Variance** of a random variable is:

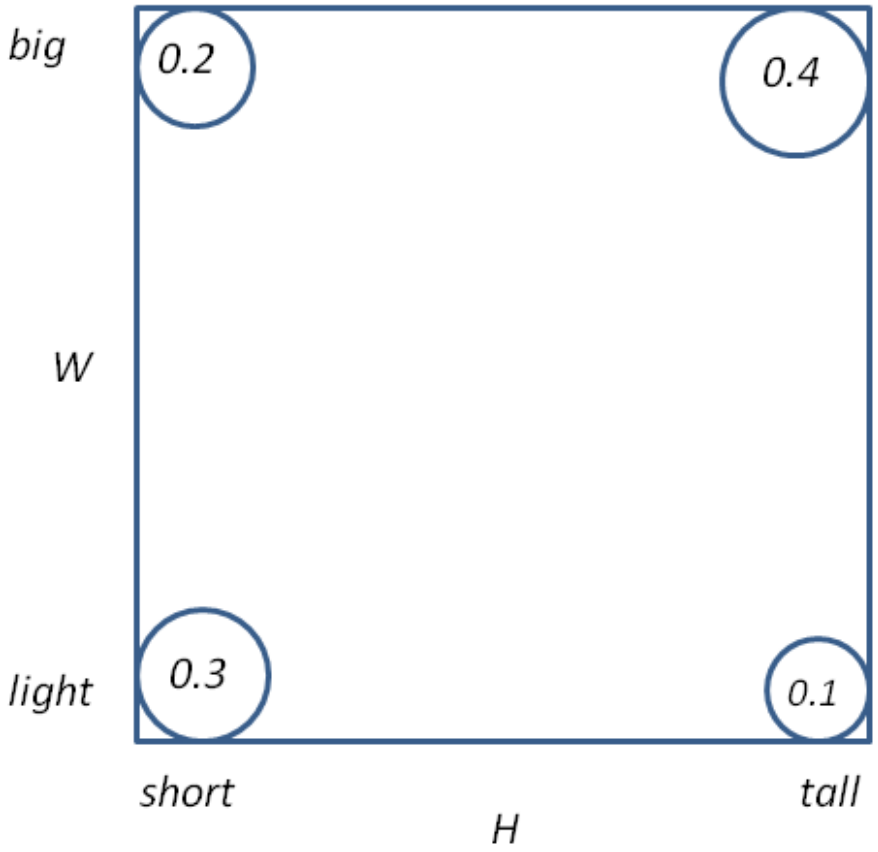
$$var(X) = \sum_i (x_i - E(X))^2 P(X = x_i)$$

- The **Standard Deviation** is its square root:

$$stdev(X) = \sqrt{Var(x)}$$

Variance

Let's associate with tall 180cm, with short 150cm, with big 100kg, and with light 50kg.



$Var(Height) = 0.5x(180cm-165cm)^2+0.5x(150cm-165cm)^2=400.50cm^2$ $stdev(Height)=20.0cm$

$Var(Weight)=0.6x(100kg-80kg)^2+0.4x(50kg-80kg)^2= 600kg^2$ $stdev(Weight)=24.5kg$

$Var(Weight|H=t)= 0.8x(100kg-90kg)^2+0.2x(50kg-90kg)^2 = 400kg^2$ $stdev(Weight|H=1)=20kg$

Covariance

- **Covariance**:

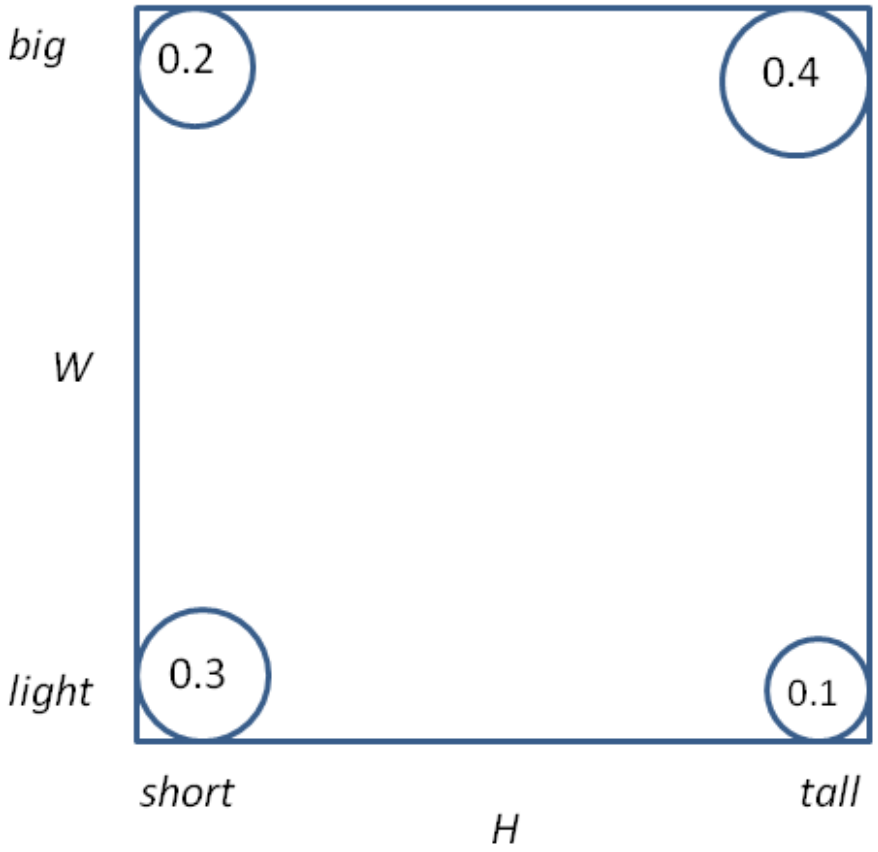
$$cov(X,Y) = \sum_i \sum_j (x_i - E(X))(y_j - E(Y))P(X = x_i, Y = y_j)$$

- **Covariance matrix**:

$$\Sigma_{[XY],[XY]} = \begin{pmatrix} var(X) & cov(X,Y) \\ cov(Y,X) & var(Y) \end{pmatrix}$$

Covariance

Let's associate with tall 180cm, with short 150cm, with big 100kg, and with light 50kg.



$$Cov(Height, Weight) = 0.4(180-165)(100-80)+0.1(180-165)(50-80)$$
$$+0.2(150-165)(100-80)+0.3(150-165)(50-80) = 150$$

Correlation

- Useful identity:

$$cov(X, Y) = E(XY) - E(X)E(Y)$$

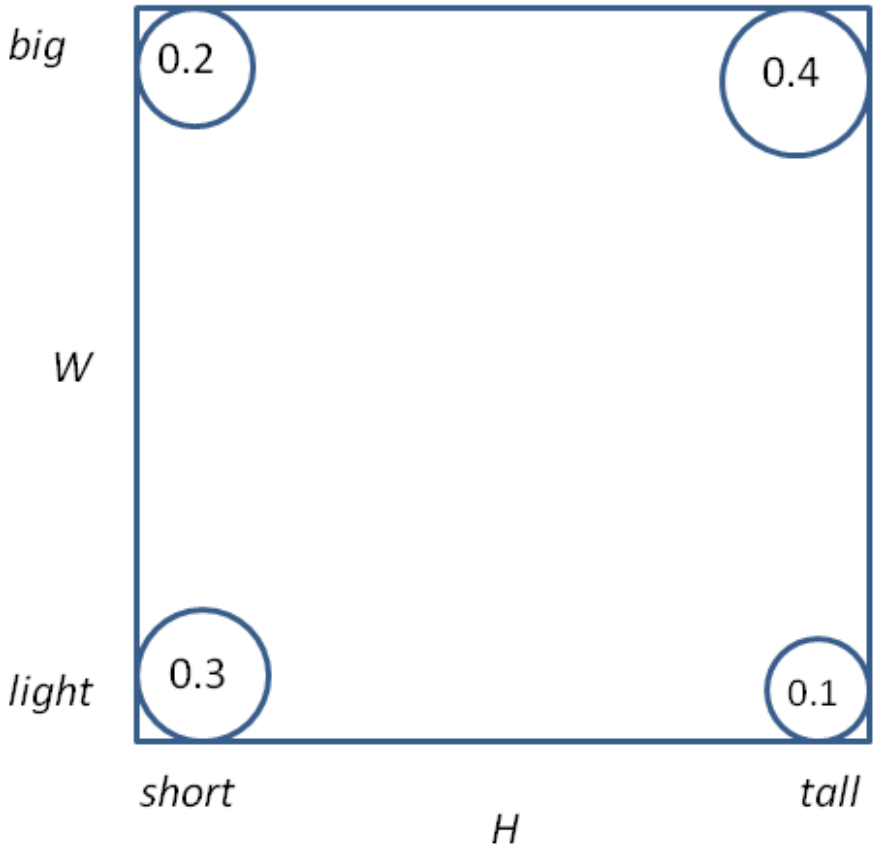
where $E(XY)$ is the **correlation**.

Correlation coefficient (confusing naming!) is

$$r = \frac{cov(X, Y)}{\sqrt{var(X)}\sqrt{var(Y)}}$$

Correlation Coefficient

Let's associate with tall 180cm, with short 150cm, with big 100kg, and with light 50kg.



r = Cov(Height, Weight)/(Stdev(Height)xStdev(Height))=150/(20x24.5)=0.3

More Useful Rules

- We have, independent of the correlation between X and Y,

$$E(X+Y) = E(X) + E(Y)$$

and thus also

$$E(X^2+Y^2) = E(X^2) + E(Y^2)$$

- For the sum of covariances,

$$var(X+Y) = E(X+Y-(E(X)+E(Y))) = E((X-E(X))+(Y-E(Y)))^2$$

$$= E((X-E(X))^2) + E((Y-E(Y))^2) - 2E(X+E(X))(Y-E(Y))$$

$$= var(X) + var(Y) - 2cov(X,Y)$$

- If $\mathbf{w}$ is a random vector with covariance matrix $\mathbf{cov}(\mathbf{w})$ and $\mathbf{y} = A\mathbf{w}$ where A is a fixed matrix. Then

$$\mathbf{cov}(\mathbf{y}) = A\mathbf{cov}(\mathbf{w})A^T$$

Continuous Random Variables

- **Probability density**

$$f(x) = \lim_{\Delta x \to 0} \frac{P(x \leq X \leq x + \Delta x)}{\Delta x}$$

- Thus

$$P(a < x < b) = \int_a^b f(x)dx$$

- The **distribution function** is

$$F(x) = \int_{-\infty}^{x} f(x)dx = P(X \leq x)$$

Expectations for Continuous Variables

- Expected value

$$E(X) = E_{P(x)}(X) = \int xP(x)dx$$

- Variance

$$var(X) = \int (x - E(x))^2 P(x)dx$$

- Covariance:

$$cov(X, Y) = \int \int (x - E(X))(y - E(Y))P(x, y)dxdy$$