Some Concepts of Probability (Review)

Volker Tresp
Summer 2019

Definition

- There are different way to define what a probability stands for
- Mathematically, the most rigorous definition is based on Kolmogorov axioms and probability theory is a mathematical discipline
- For beginners it is more important to obtain an intuition, and the definition I present is based on a relative frequency; in statistics, probability theory is applied to problems of the real world
- We start with an example

Example: Students in Munich

- Let's assume that there are $\tilde{N} = 50000$ students in Munich. This set is called the *population*
- $\tilde{N}$ is the size of the population, often assumed to be infinite
- Formally, I put the all 50000 students in an urn (bag)
- I randomly select a student: this is called an *(atomic) event* or an *experiment* and defines a *random process*
- The selected student is an *outcome* of the experiment

Sample

- A particular student will be picked with elementary probability $1/\tilde{N}$
- Performing the experiment N times produces a sample (training data set) D of size N
- An analysis of the sample can give us insight about the population (statistical inference)
- Sampling *with replacement*: I return the student to the urn after the experiment
- Sampling *without replacement*: I do not return the student to the urn after the experiment

Random Variable

- A particular student has a height attribute *(tiny, small, medium, large, huge)*
- The height H is called a *random variable* with states $h \in \{$*tiny, small, medium, large, huge*$\}$
- A random variable is a variable (more precisely a function of the outcome of the random experiment), whose value depends on the result of a random process
- Thus at each experiment I measure a particular h

Probability

- Then the *probability* that a randomly picked student has height $H = h$ is defined as

$$P(H = h) = \lim_{N \to \infty} \frac{N_h}{N}$$

  with $0 \leq P(H = h) \leq 1$

- N_h is the number of times that a selected student is observed to have height $H = h$

Sample / Training Data

- I can estimate

$$\hat{P}(H = h) = \frac{N_h}{N} \approx P(H = h)$$

- In statistics one is interested in how well $\hat{P}(H = h)$ (the probability estimate derived from the sample) approximates $P(H = h)$ (the probability in the population)
- Note the importance of the definition of a population: $P(H = h)$ might be different, when I consider individuals in Munich or Germany
- Thus the population play an important role in a statistical analysis

Statistics and Probability

- *Probability* is a mathematical discipline developed as an abstract model and its conclusions are *deductions* based on *axioms* (Kolmogorov axioms)
- *Statistics* deals with the application of the theory to real problems and its conclusions are *inferences* or *inductions*, based on observations (Papoulis: Probability, Random variables, and Stochastic Processes)
- *Frequentist or classical statistics* and *Bayesian statistics* apply probability in slightly different ways

Joint Probabilities

- Now assume that we also measure weight (size) S with weight attributes *very light, light, normal, heavy, very heavy*. Thus S is a second random variable

- Similarly

$$P(S = s) = \lim_{N \to \infty} \frac{N_s}{N}$$

- We can also count co-occurrences

$$P(H = h, S = s) = \lim_{N \to \infty} \frac{N_{h,s}}{N}$$

This is called the *joint probability distribution* of H and S

Marginal Probabilities

- It is obvious that we can calculate the *marginal probability* $P(H = h)$ from the joint probabilities

$$P(H = h) = \lim_{N \to \infty} \frac{\sum_s N_{h,s}}{N}$$

$$= \sum_s P(H = h, S = s)$$

- This is called marginalization; I can calculate the marginal probability from the joint probability (without gong back to the counts)

Conditional Probabilities

- One is often interested in the *conditional probability*. Let's assume that I am interested in the probability distribution of S for a given height $H = h$. Since I need a different normalization I get

$$P(S = s|H = h) = \lim_{N \to \infty} \frac{N_{h,s}}{N_h}$$

  So I count the co-occurrences, but I normalize by N_h Then,

$$P(S = s|H = h) = \frac{P(H = h, S = s)}{P(H = h)}$$

- Relationship to machine learning: $H = h$ is the *input* and $S = s$ is the *output*
- Conditioning is closely related to the definition of a population: $P(S = s|H = h)$ is the same as $P(S = s)$ in a population which is restricted to students with $H = h$

Conditional Probabilities and IDs

- Each student in the sample has an ID, e.g. $\mathit{ID} = \mathit{Jane}$
- $P(S = s|H = h)$ gives me a probability
- But, $P(S = s(\mathit{Jane})|\mathit{ID}{=}\mathit{Jane}) = 1$
- So conditioning can also be applied to IDs
- If *Jane* is in the sample, we know $s(\mathit{Jane}$
- To be more precise, one also needs to specify a time stamp for the sampling, since some attributes (e.g., job status, color of the shirt) might be time dependent

What if I only have IDs

- Embeddings are attributes derived from data that cannot directly be measured

Bayes Formula

- If I know $P(S = s|H = h)$, does it tell me anything about $P(H = h|S = s)$?
- We use the definition of a conditional probability,

$$P(H = h|S = s) = \frac{P(H = h, S = s)}{P(S = s)}$$

$$P(S = s|H = h) = \frac{P(H = h, S = s)}{P(H = h)}$$

- Thus we get *Bayes' formula*

$$P(H = h|S = s) = \frac{P(S = s|H = h)P(H = h)}{P(S = s)}$$

or

$$P(H = h|S = s) = P(S = s|H = h)\frac{P(H = h)}{P(S = s)}$$

Curse of Dimensionality

- This is all great; so why do we need neural networks and the likes?
- If the number of inputs is large I would need to estimate

$$\widehat{P}(Y = y|X_1 = x_1, ..., X_M = x_M) = \frac{N_{y,x_1,...,x_M}}{N_{x_1,...,x_M}}$$

- When the random variables $X_1, \ldots, X_M$ are binary, I would need to estimate 2^M quantities
- Another encounter of the "curse of dimensionality": the number of parameters and the required sample size N are $\mathcal{O}(2^M)$

Supervised Learning with a Linear Model

- Fortunately, in reality the dependencies are often simpler and one might get good conditional models with simplified assumptions

- Linear logistic regression assumes

$$\hat{P}(Y=1|X_1=x_1,...,X_M=x_M)=\mathit{sig}\left(w_0+\sum_{j=1}^{M}w_jx_j\right)$$

Instead of estimating 2^M quantities I only need to estimate $M+1$ quantities

Supervised Learning with a Neural Network

- If logistic regression is too simplistic, I can also assume a neural network with weight vector $\mathbf{w}$

$$\widehat{P}(Y = 1|X_1 = x_1, ..., X_M = x_M) = \mathit{sig}\,(\mathit{NN}_{\mathbf{w}}(x_1, \ldots, x_M))$$

- We can tune the number of hidden units in the neural network to match the requires complexity (recall the discussion on modelling complexity)

Classical Statistics

- We can find a good estimate of $\mathbf{w}$ by minimizing a cost function defined on the training data D

- A *maximum likelihood* approach would maximize

$$P(D|\mathbf{w})$$

- The negative log likelihood, i.e., $-\log P(D|\mathbf{w})$, is also called cross entropy cost function

- Thus the maximum likelihood estimate is

$$\mathbf{w}_{ML} = \arg\max_{\mathbf{w}} P(D|\mathbf{w}) = \arg\max_{\mathbf{w}} \log P(D|\mathbf{w})$$

  The optimal $\mathbf{w}$ is a so-called *point estimate*

Bayesian Statistics

- Now assume a second urn (bag); each ball has attached a vector $\mathbf{w}$
- Again the random process selects random balls out off the urn
- This defines a second random process with random variable with joint probability distribution $P(W = \mathbf{w})$
- If I could sample many times I could get a good estimate of $P(W = \mathbf{w})$

Bayesian Statistics, cont'd

- Unfortunately, nature only samples one $\mathbf{w}$ (e.g., the parameters in a logistic regression model), so I, the researcher, simply assumes a $P(\mathbf{w})$

- A common assumption is a multivariate Gaussian distribution with zero mean and variance α^2,

$$P(\mathbf{w}) = \mathcal{N}(0, \alpha^2 I)$$

  $P(\mathbf{w})$ is called the *a priori weight distribution*, or simply the *prior*

- After nature selects a $\mathbf{w}$, it produces a population and a sample according to $P(D|\mathbf{w})$

- With a given $P(\mathbf{w})$ and $P(D|\mathbf{w})$ we can simply apply Bayes' formula to get the *a posteriori weight distribution*, of simply the *posterior*

$$P(\mathbf{w}|D) = \frac{P(D|\mathbf{w})P(\mathbf{w})}{P(D)}$$

- $P(D) = \int P(D|\mathbf{w})P(\mathbf{w})d\mathbf{w}$ is called the *evidence* or *marginal likelihood*

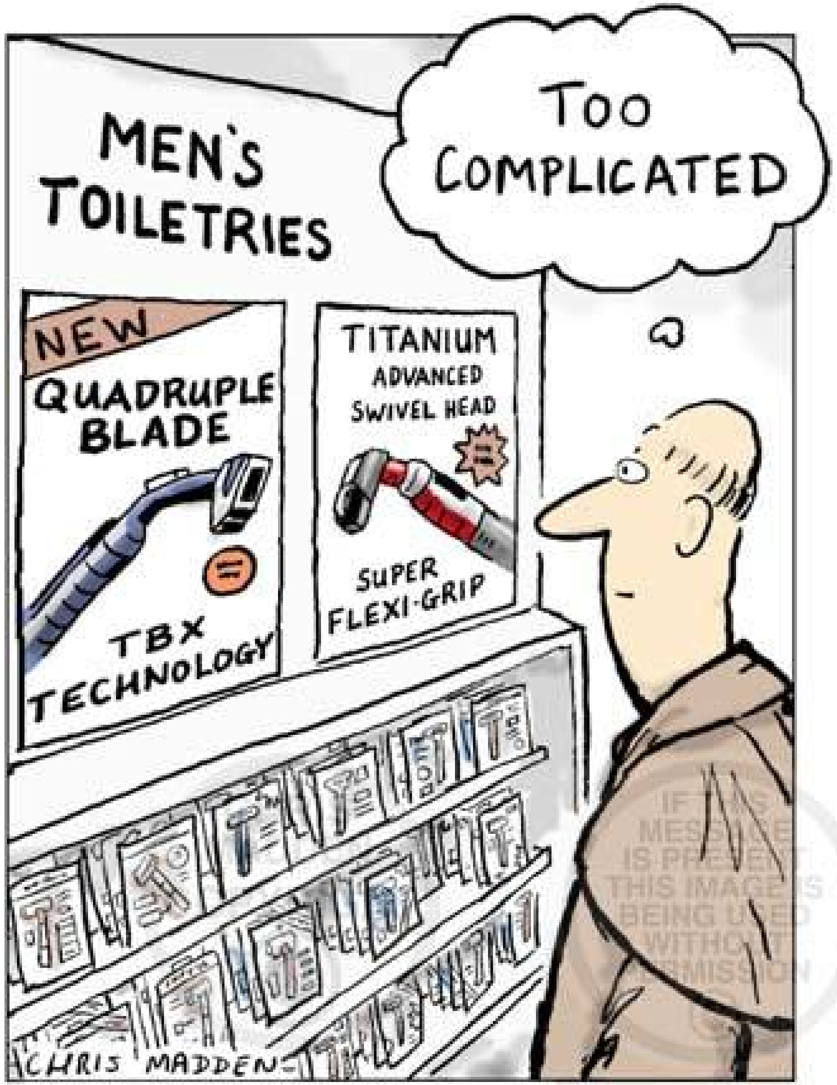
Posterior Distributions and Point Estimates

- Note that a Bayesian treatment provides a posterior distribution $P(\mathbf{w}|D)$ which in many ways is more informative than a point estimate

- We can also derive a point estimate: The *maximum a posteriori (MAP)* estimate is

$$\mathbf{w}_{MAP} = \arg\max_{\mathbf{w}} P(D|\mathbf{w})P(\mathbf{w})$$

$$= \arg\max_{\mathbf{w}}(\log P(D|\mathbf{w}) + \log P(\mathbf{w}))$$



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

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

Some Concepts of Probability (Another Review with a Few Additional Concepts)

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)
$P(W=l)=0.4$ (in the sample, 30 persons were light)

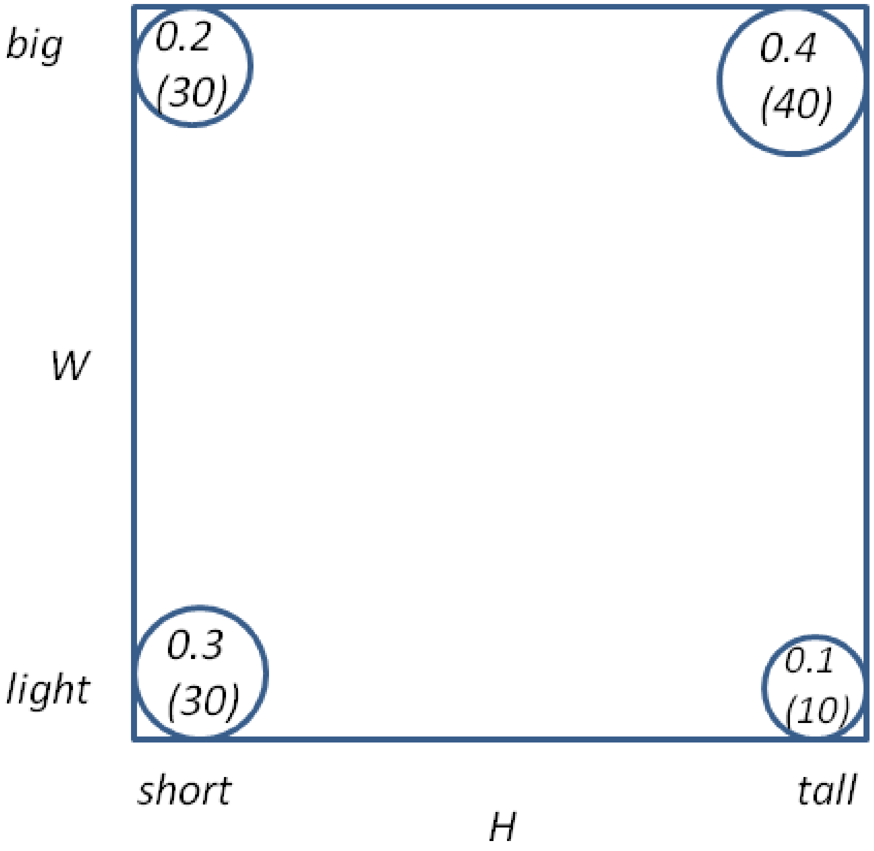
Multivariate Probability Distributions

- Define two random variables X and Y. 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



W \ H	short	tall
big	0.2 (30)	0.4 (40)
light	0.3 (30)	0.1 (10)

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

Special Cases

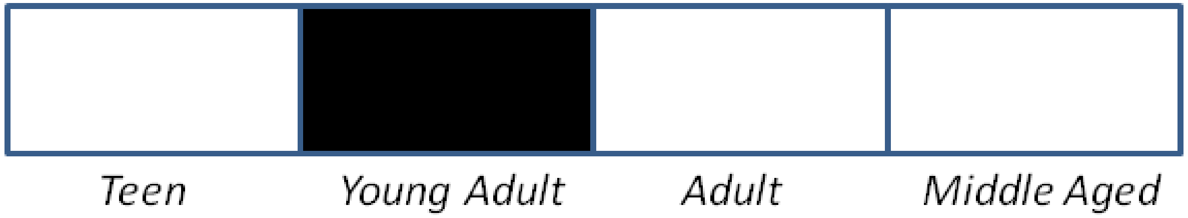
- If two random variables are independent, then $P(X, Y) = P(X)P(Y)$. This is not the case in our example since $P(t, b) = 0.4 \neq P(t)P(b) = 0.5 \times 0.6 = 0.3$

- 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 binary 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



Teen *Young Adult* *Adult* *Middle Aged*

- *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, ...)

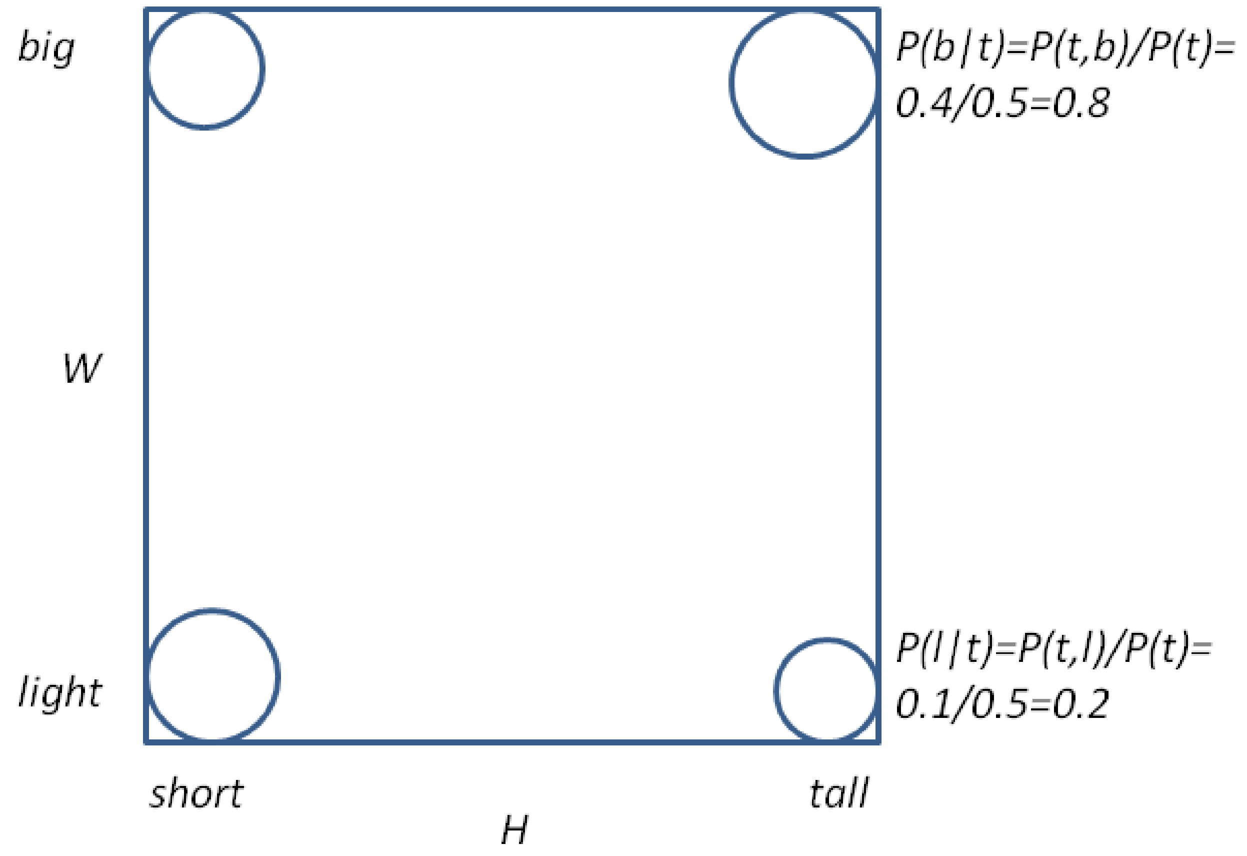
Conditional Distribution

- 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**

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

- The distribution is identical to the one for the unconditional case, only that I have to divide by $P(X = x)$ (re-normalize)
- Example: The probability that you have flu is small ($P(Y = 1)$) but the probability that you have the flu, given that fever is high, might be high ($P(Y = 1|X = 1)$) (the filter selects only individuals in the population with high fever)

Conditional Probabilities *P(W|H)*

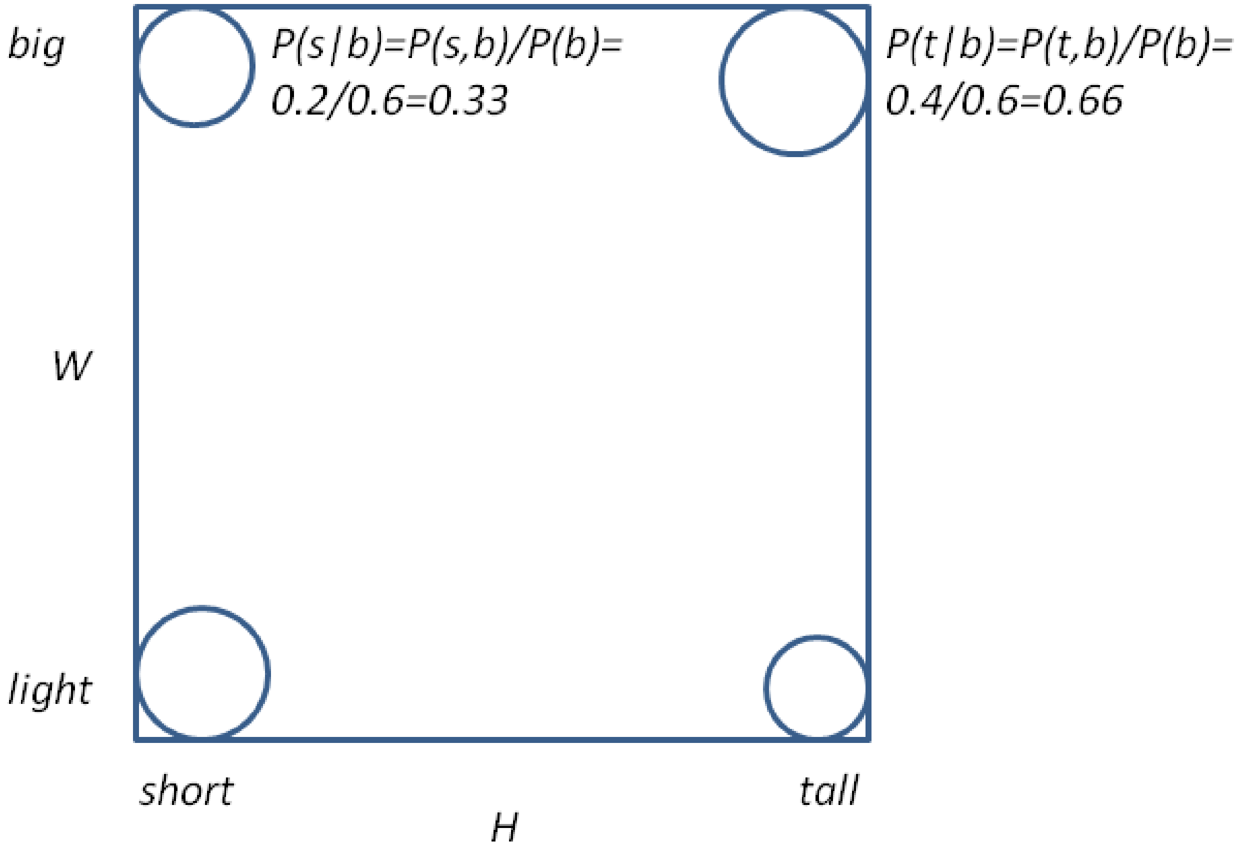


P(b|t)=P(t,b)/P(t)= 0.4/0.5=0.8

P(l|t)=P(t,l)/P(t)= 0.1/0.5=0.2

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

Conditional Probabilities *P(H|W)*



P(s|b)=P(s,b)/P(b)=
0.2/0.6=0.33

P(t|b)=P(t,b)/P(b)=
0.4/0.6=0.66

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

Product Rule and Chain Rule

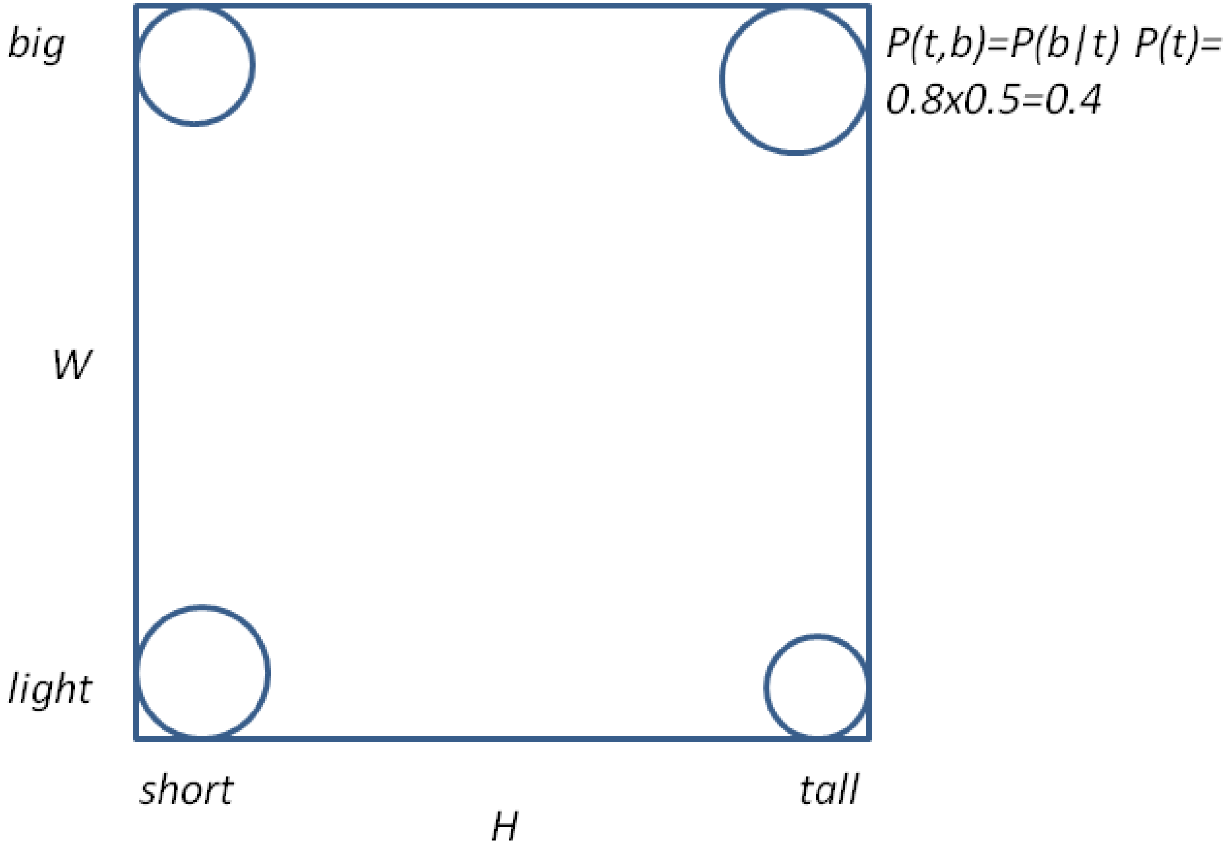
- It follows: **product rule**

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

- 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

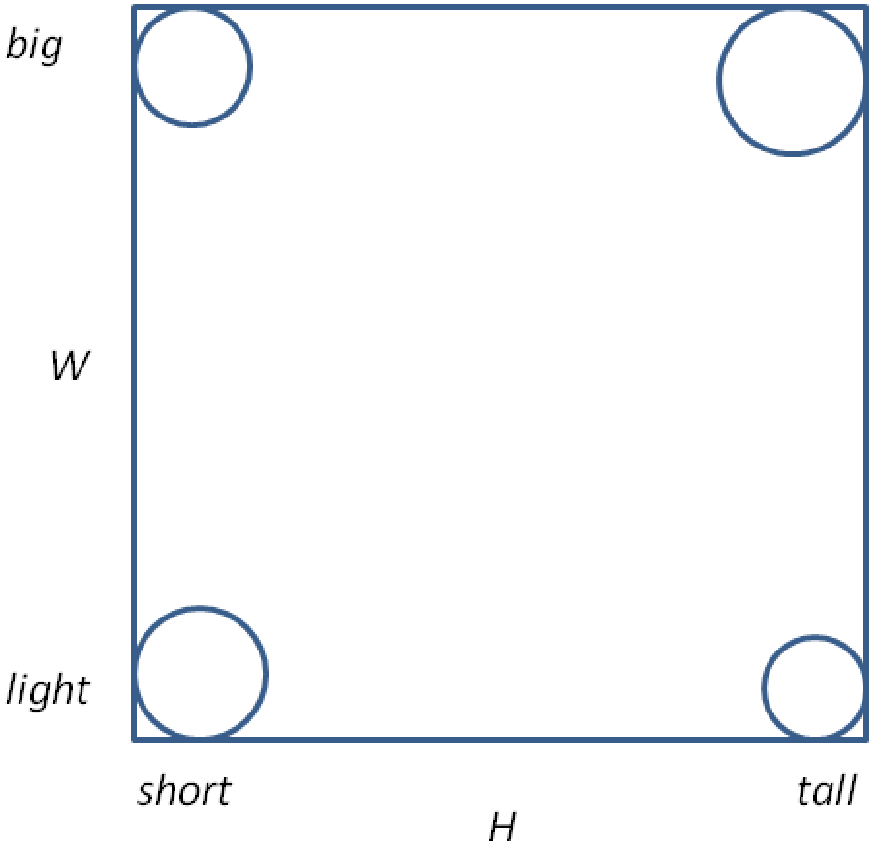
Bayes Theorem

- In general, $P(Y|X) \neq P(X|Y)$. Bayes' theorem gives the true relationship
- **Bayes Theorem**

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

$$P(y) > 0$$

Bayes Theorem



P(t|b)=
P(b|t)P(t)/P(b)=
0.8x0.5/0.6=0.66

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

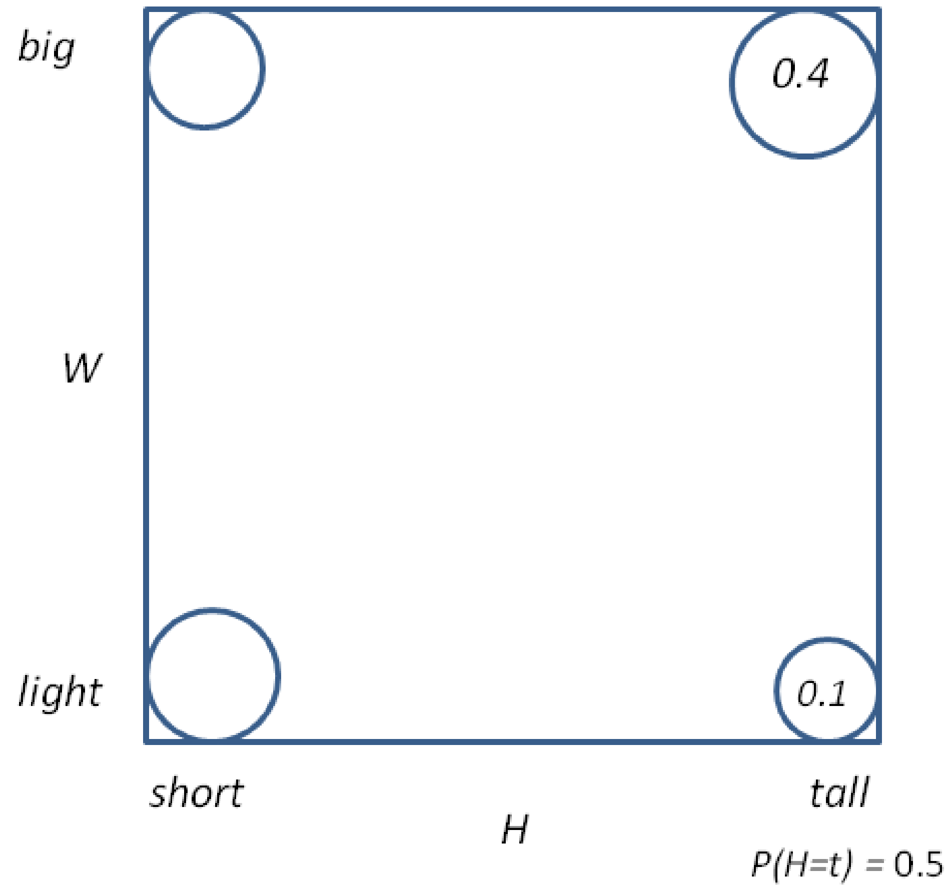
Marginal Distribution

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

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

Also called: law of total probability

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**

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

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

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

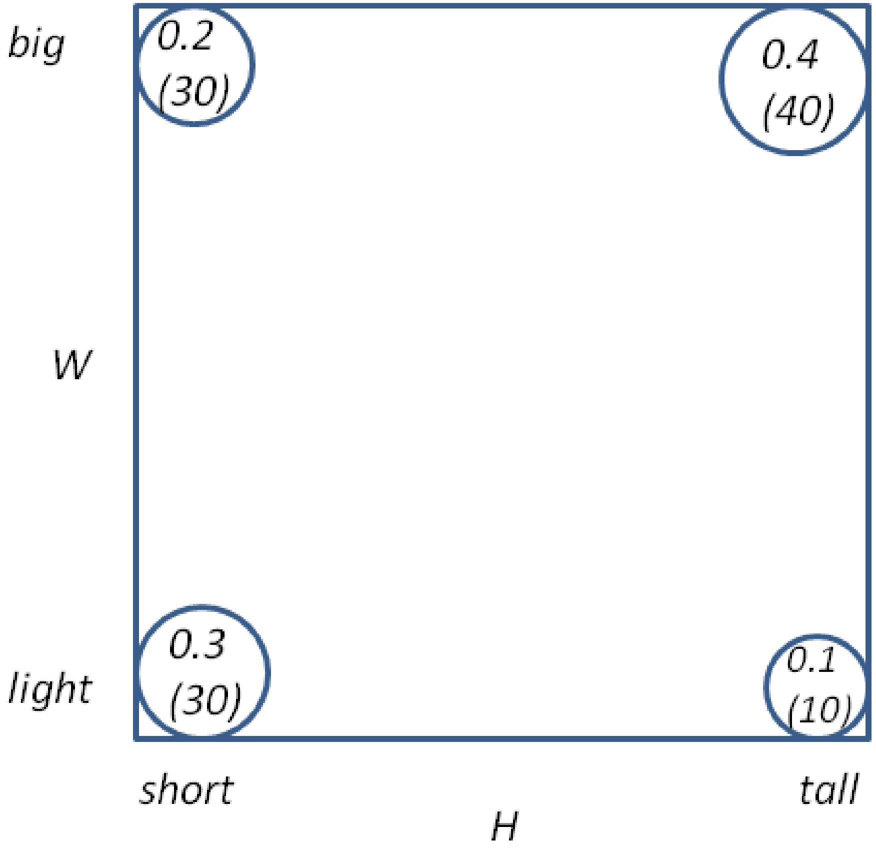
$$-P(X = 1, Y = 1)$$

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

- 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,

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

- Example: The probability that you have flu does not change if I know your age: $P(Y = 1|Age = 38) = P(Y = 1)$

- (It follows for independent random variables that $P_{x,y} = P_x \otimes P_y$, where $P_{x,y}$ is the matrix of joint probabilities and P_x and P_y are vectors of marginal probabilities and $\otimes$ is the Kronecker product)

Joint Tables and Marginals

Independent random variables:

0.5	0.25	0.25
0.5	0.25	0.25
	0.5	0.5

0	0	0
1	1	0
	1	0

0.3	0.12	0.18
0.7	0.28	0.42
	0.4	0.6

Dependent random variables:

0.5	0.5	0
0.5	0	0.5
	0.5	0.5

0.5	0	0.5
0.5	0.5	0
	0.5	0.5

0.3	0.10	0.2
0.7	0.30	0.4
	0.4	0.6

Note that diagonal tables indicate **dependencies**!

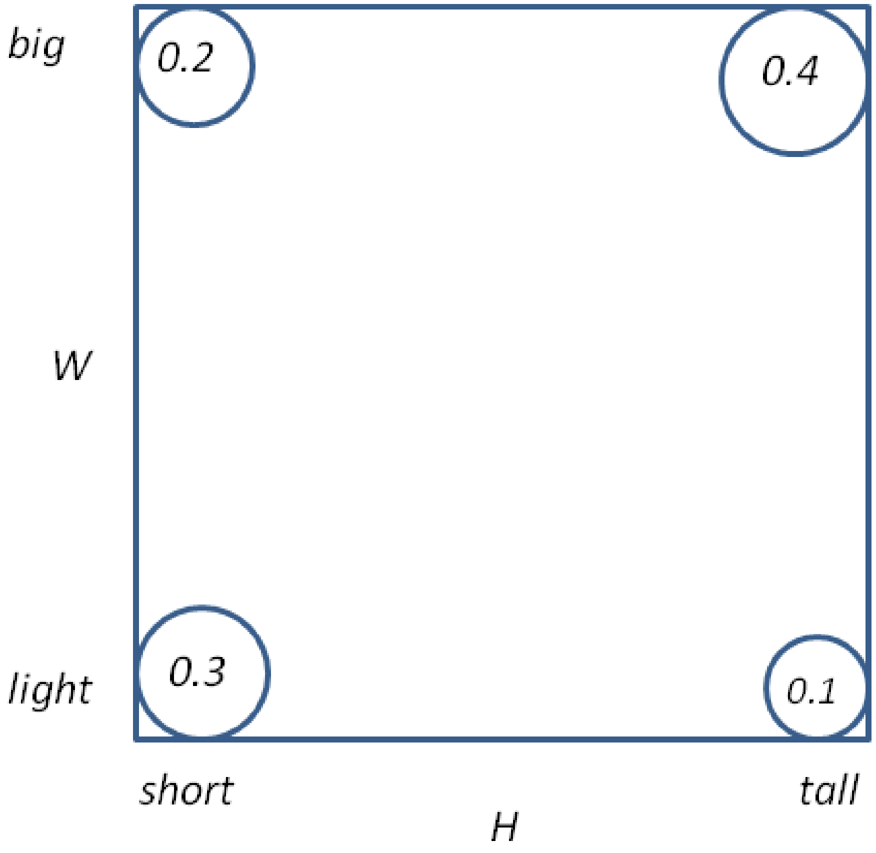
Expected Values

- **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:

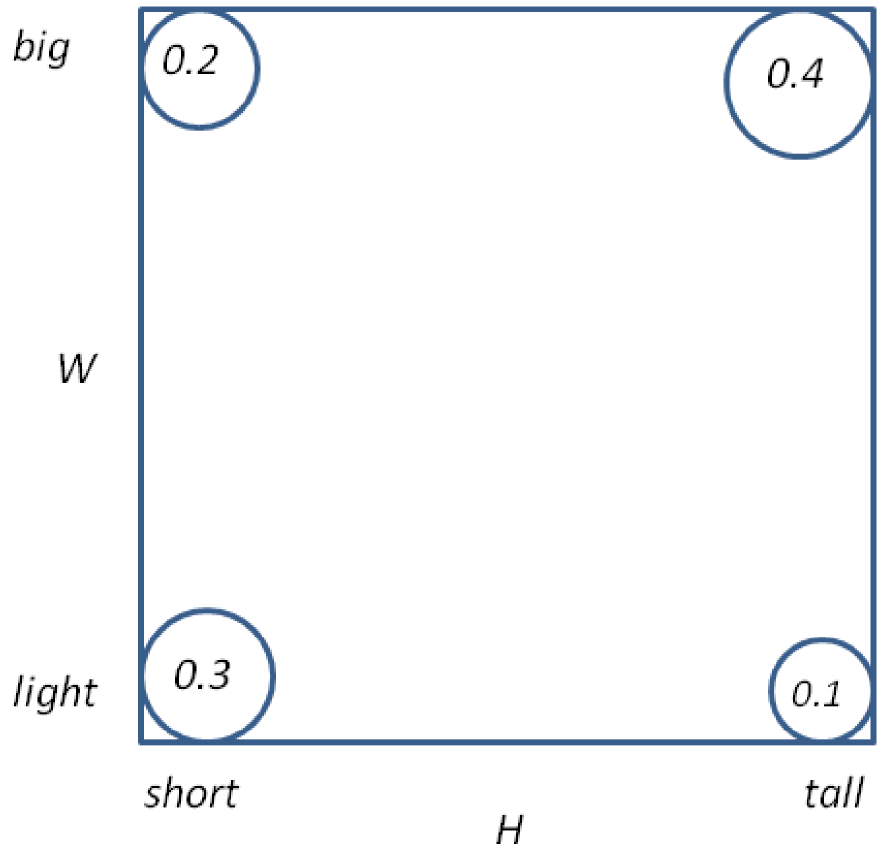
$$var(X) = \sum_i (x_i - E(X))^2 P(X = x_i)$$

- 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**:

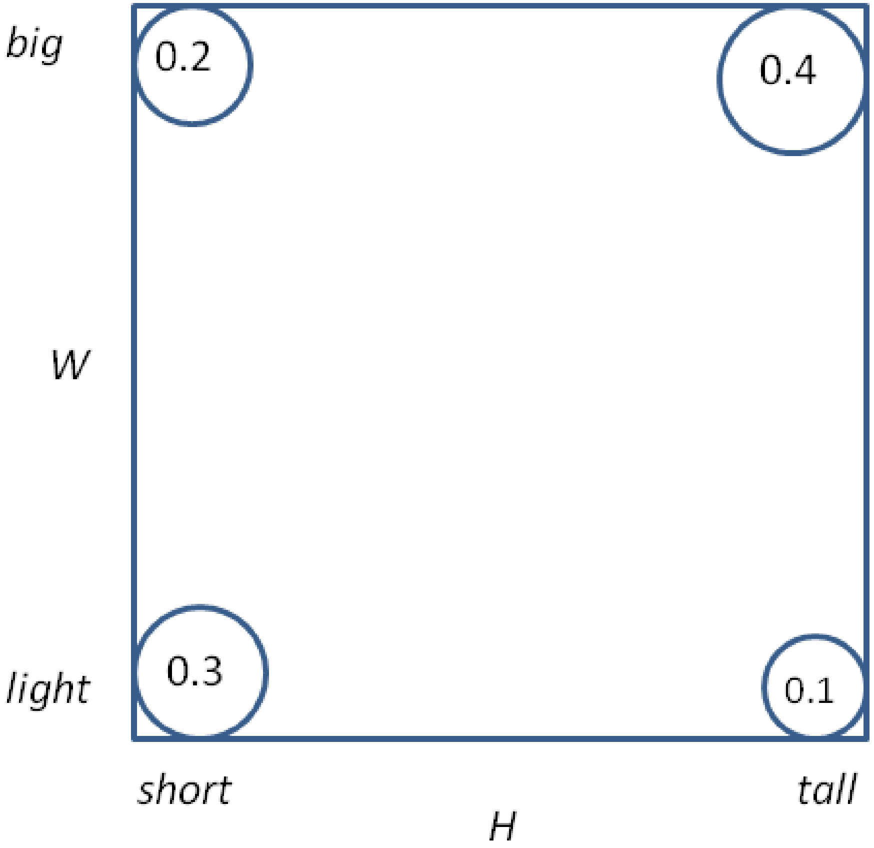
$$cov(X,Y) = \sum_i \sum_j (x_i - E(X))(y_j - E(Y))P(X = x_i, Y = y_j)$$

- **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

Covariance, Correlation, and Correlation Coefficient

- Useful identity:

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

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

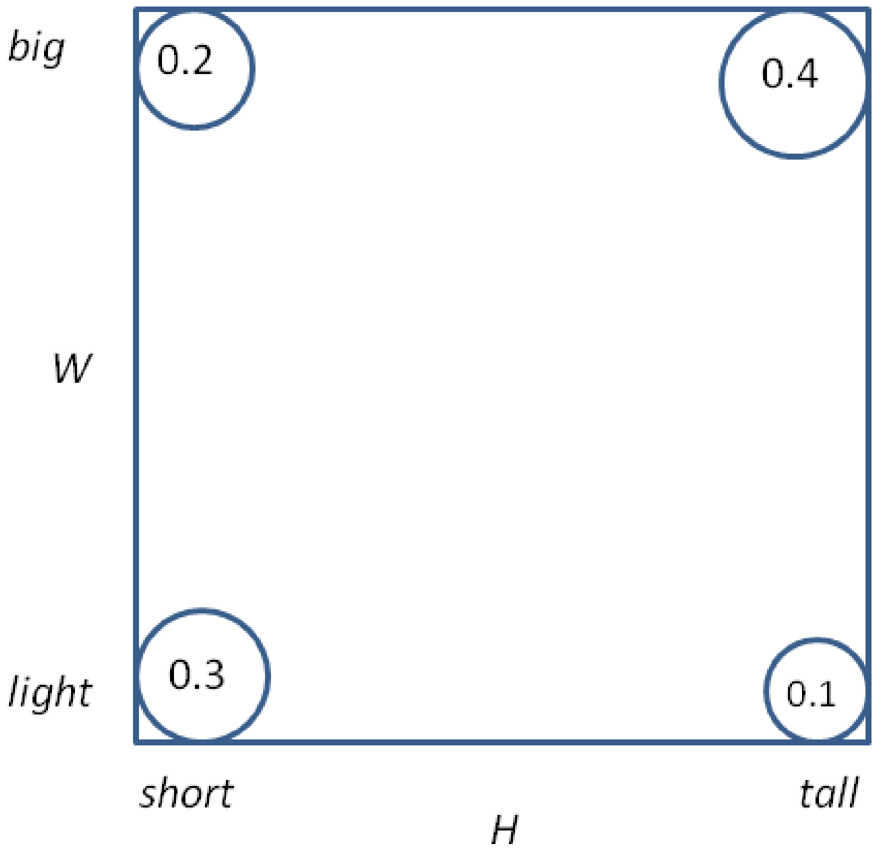
- The **(Pearson) correlation coefficient** (confusing naming!) is

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

- It follows that $var(X) = E(X^2) - (E(X))^2$ and

$$var(f(X)) = E(f(X)^2) - (E(f(X)))^2$$

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



The ***correlation coefficient*** *is:*

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 variance of the sum of random variables,

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

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

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

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

- Similarly,

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

Covariance Matrix of Linear Transformation

- Let $\mathbf{w}$ be a random vector with covariance matrix $\textbf{cov}(\mathbf{w})$
- Let

$$\mathbf{y} = A\mathbf{w}$$

where A is a fixed matrix. Then

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

Continuous Random Variables

- **Probability density**

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

- Thus

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

- The **distribution function** is

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

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:

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