Probability Theory

Sadegh Eskandari
Department of Computer Science, University of Guilan

Probability Theory

- A key concept in data science is that of uncertainty.
  - Noise on measurements
  - Finite size of data sets
- Probability theory provides a consistent framework for the quantification and manipulation of uncertainty in data
- Probability theory + Decision Theory = Optimal Predictions given all the information available to us

Probability Theory

- Let X is random variable with possible values $\{x_1, x_2, \dots, x_N\}$
- Let Y is random variable with possible values $\{y_1, y_2, \dots, y_M\}$

- The marginal probability

$$p(X = x_i) = \sum_{j=1}^{M} p(X = x_i, Y = y_j)$$

$$p(Y = y_j) = \sum_{i=1}^{N} p(X = x_i, Y = y_j)$$

The sum rule

$$p(X) = \sum_{Y} p(X, Y)$$

- The conditional probability

$$p(X = x_i | Y = y_j) = \frac{p(X = x_i, Y = y_j)}{p(Y = y_j)}$$

$$p(Y = y_j | X = x_i) = \frac{p(X = x_i, Y = y_j)}{p(X = x_i)}$$

The product rule

$$p(X, Y) = p(X|Y)p(Y) = p(Y|X)p(X)$$

Probability Theory

- The Bayes' Theorem

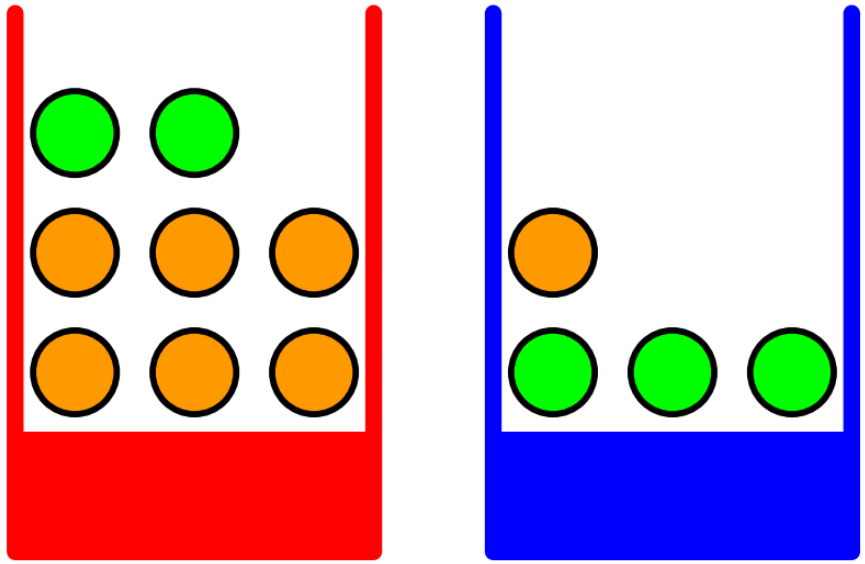
$$p(X = x_i | Y = y_j) = \frac{P(Y = y_j | X = x_i) P(X = x_i)}{P(Y = y_j)}$$

The Bayes' theorem

$$p(X|Y) = \frac{P(Y|X)P(X)}{P(Y)} = \frac{P(Y|X)P(X)}{\sum_X p(X,Y)} = \frac{P(Y|X)P(X)}{\sum_X p(Y|X)p(X)}$$

Probability Theory

- Two boxes: red and blue
- The red box: 2 apples, 6 oranges
- The blue box: 3 apples, 1 oranges



- Suppose we randomly pick one of the boxes and from that box we randomly select an item of fruit, and having observed which sort of fruit it is we replace it in the box from which it came. We could imagine repeating this process many times.

- Red box **40%** - Blue box **60%**
- Equally likely to select any of the pieces of fruit in each box
- Random Variables:
  - B: The identity of the box. $B = r$ (the red box), $B = b$ (the blue box)
  - F: The identity of the fruit. $F = a$ (the apple), $F = o$ (orange)

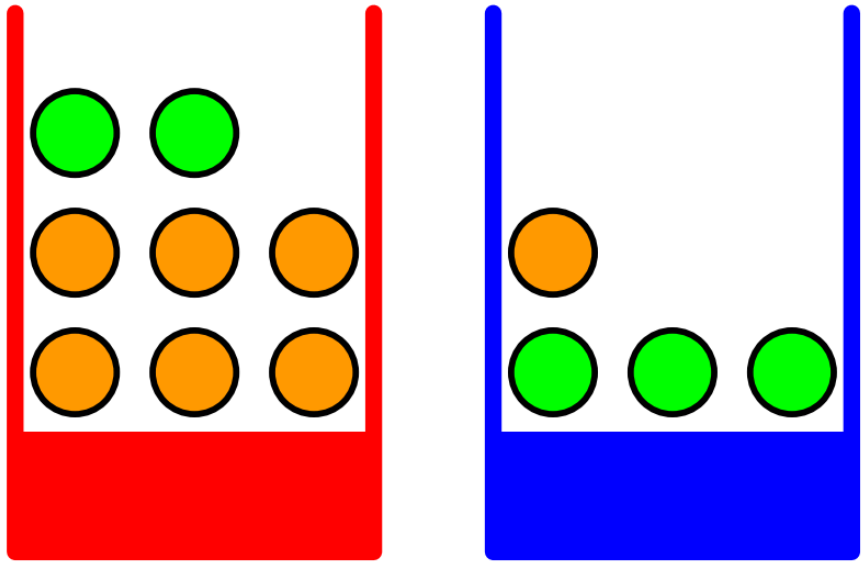
Probability Theory

- The available information:

$p(B = r) = 0.4$ $\quad$ $p(B = b) = 0.6$

$p(F = a|B = r) = \frac{1}{4}$ $\quad$ $p(F = o|B = r) = \frac{3}{4}$

$p(F = a|B = b) = \frac{3}{4}$ $\quad$ $p(F = o|B = b) = \frac{1}{4}$



what is the overall probability that the selection procedure will pick an apple? $p(F = a) = ?$

Using the sum rule: $p(F) = \sum_B p(F, B)$ $\quad \Rightarrow p(F = a) = p(F = a, B = r) + p(F = a, B = b)$

Using the product rule: $p(F, B) = p(F|B)p(B)$

$$\Rightarrow p(F = a) = p(F = a|B = r)p(B = r) + p(F = a|B = b)p(B = b)$$

$$= \frac{1}{4} \times 0.4 + \frac{3}{4} \times 0.6 = 0.55$$

Find $p(F = o) = ?$

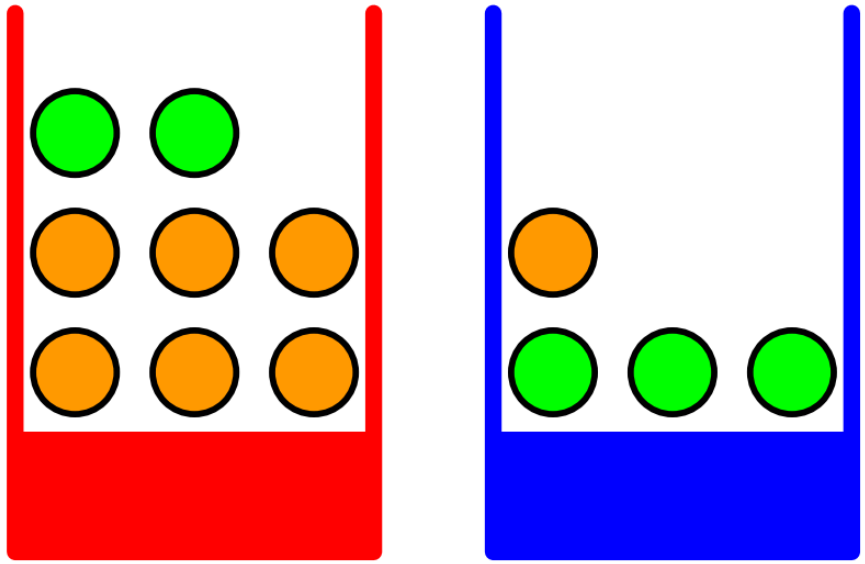
Probability Theory

- The available information:

$p(B = r) = 0.4$ $\quad$ $p(B = b) = 0.6$

$p(F = a|B = r) = \frac{1}{4}$ $\quad$ $p(F = o|B = r) = \frac{3}{4}$

$p(F = a|B = b) = \frac{3}{4}$ $\quad$ $p(F = o|B = b) = \frac{1}{4}$



given that we have chosen an apple, what is the probability that the box we chose was the blue one?

$$p(B = b|F = a) = ?$$

Using the Bayes' theorem:

$$p(B = b|F = a) = \frac{p(F = a|B = b)p(B = b)}{p(F = a)}$$

$$= \frac{\frac{3}{4} \times 0.6}{0.55} \approx 0.82$$

Find $p(B = \mathrm{b}|F = o) = ?$

Probability Theory

Another Example

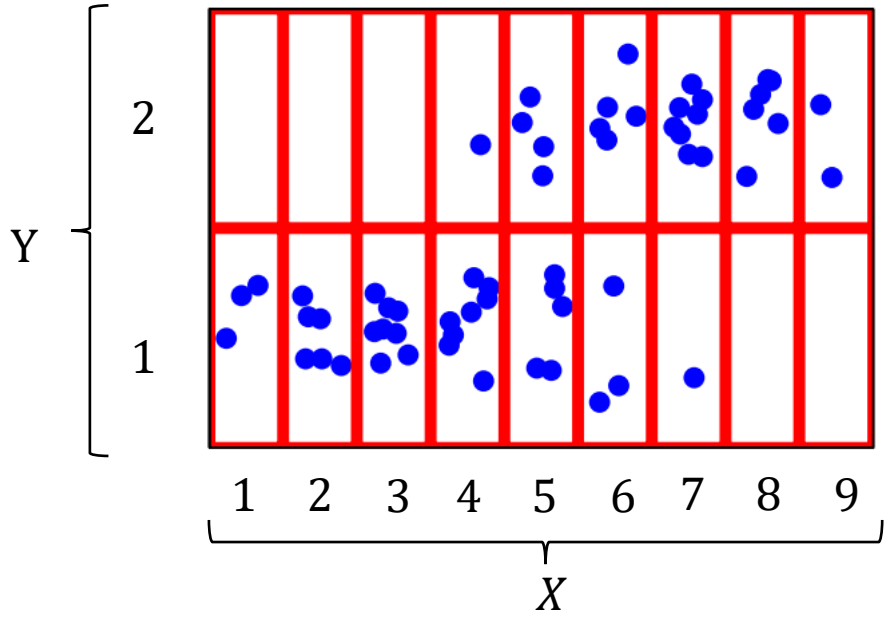
Total number of trials = 60

- $p(X = 5, Y = 1) = \frac{5}{60}$
- $p(X = 3, Y = 2) = 0$
- $p(X = 6) = ?$

$$p(X = 6) = p(X = 6, Y = 1) + p(X = 6, Y = 2) = \frac{3}{60} + \frac{5}{60} = \frac{8}{60}$$

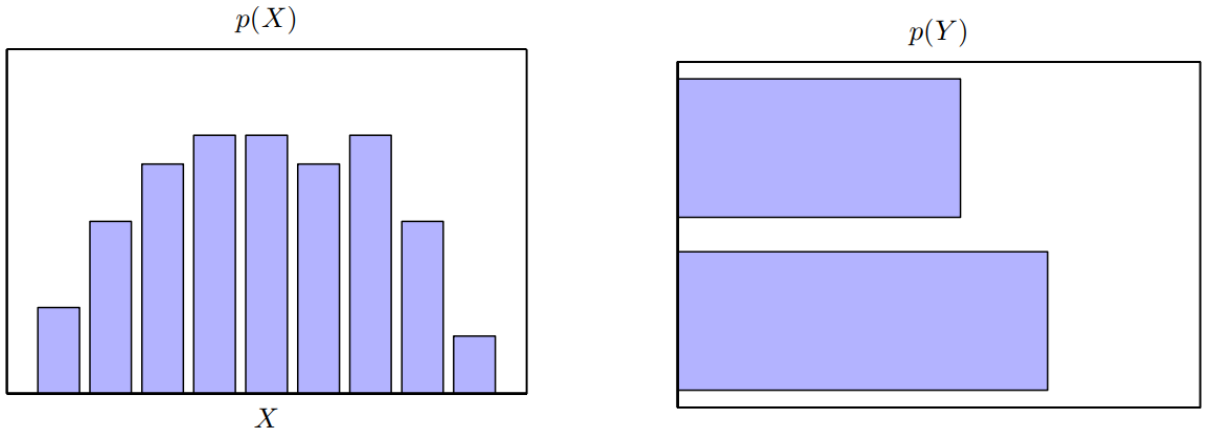
The sum rule

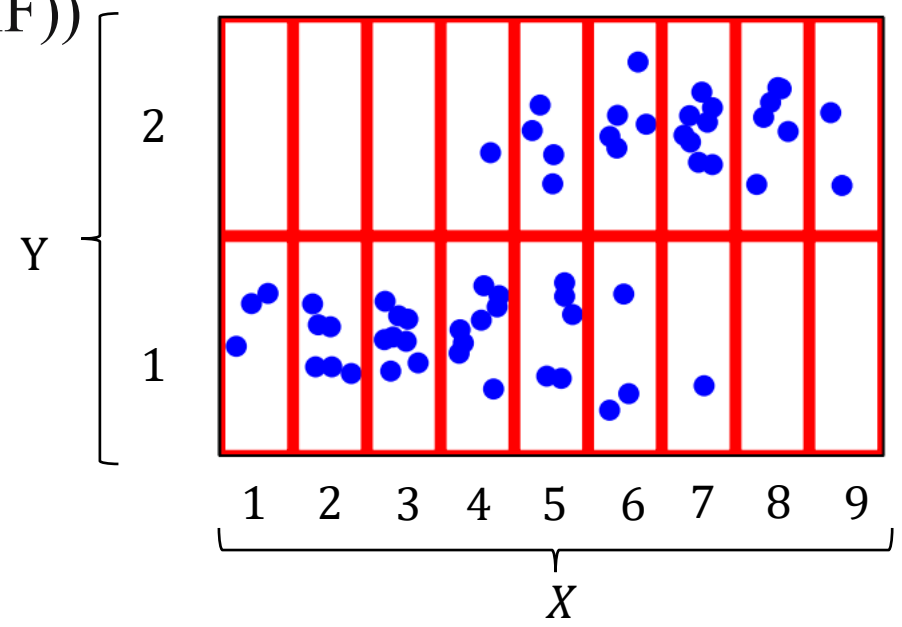
- $p(Y = 1) = \frac{3+6+8+8+5+3+1+0+0}{60} = \frac{34}{60}$
- $p(X = 6|Y = 1) = \frac{p(X=6,Y=1)}{p(Y=1)} = \frac{\left(\frac{3}{60}\right)}{\left(\frac{34}{60}\right)} = \frac{3}{34}$

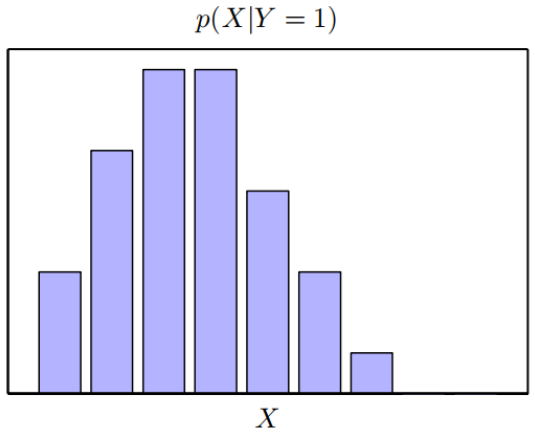


Probability Theory

Probability Distributions (Probability Mass Function (PMF))







Independent Variables

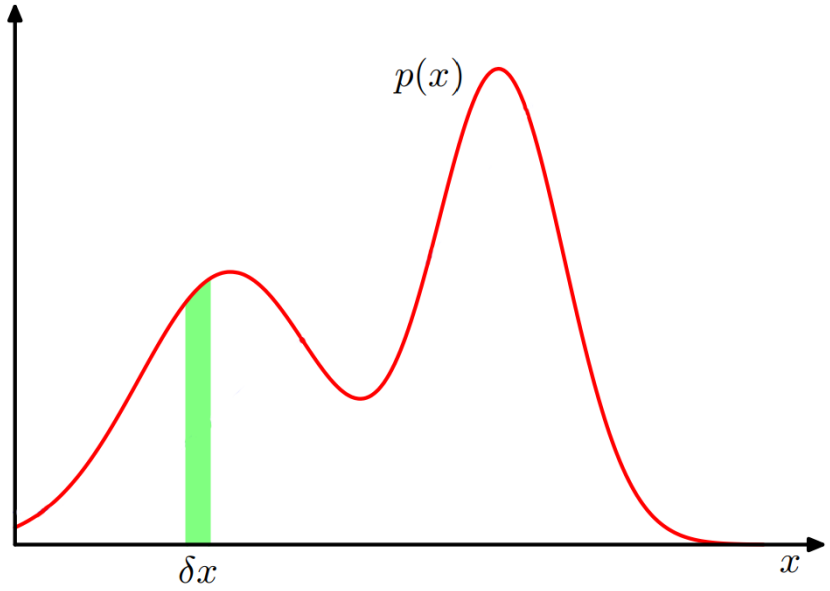
$$p(X,Y) = p(X)p(Y)$$

Probability Densities

- We also wish to consider probabilities with respect to continuous variables.
- However, the PMF does not work for continuous random variables, because for a continuous random variable X, $p(X = x) = 0$, for all $x \in \mathbb{R}$
- Instead, we define the probability density function (PDF) over a continues variable

$$p_X(x) = \lim_{\delta x \to 0} \frac{p\big(x \in (x, x + \delta x)\big)}{\delta x}$$

- Where there is no ambiguity, we simply use $p(x)$ instead of $p_X(x)$



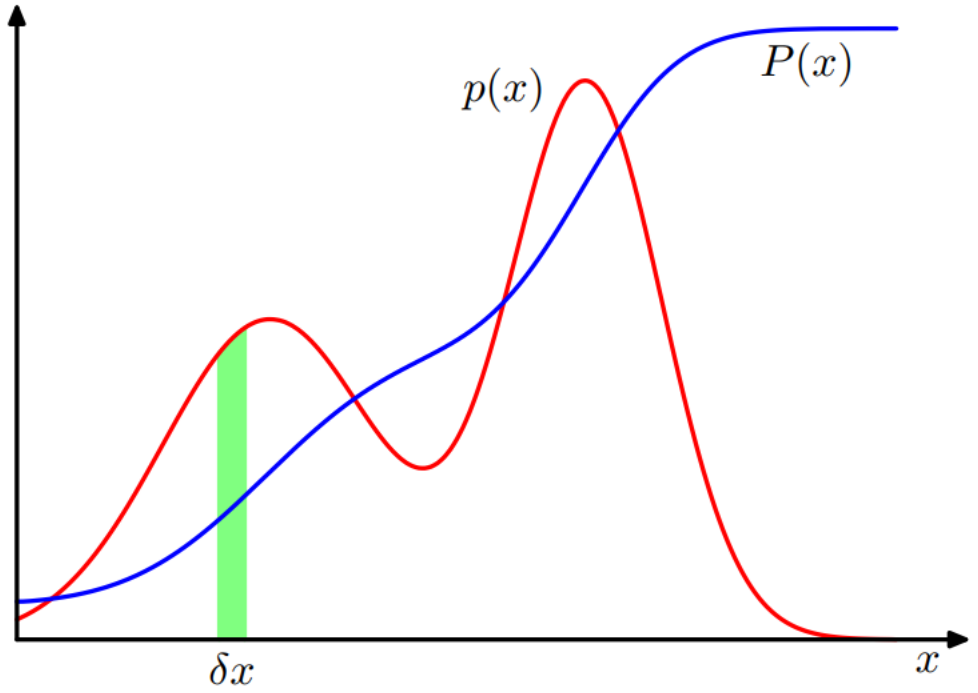
- Therefor the probability that x will lie in an interval (a, b) is then given by

$$p\big(x \in (a, b)\big) = \int_a^b p(x)dx$$

Probability Densities

- The probability density $p(x)$ must satisfy the two conditions:

$$p(x) \geq 0 \qquad \int_{-\infty}^{+\infty} p(x)dx = 1$$



- The probability that x lies in the interval $(-\infty, z)$ is given by the cumulative distribution function (CDF) defined by

$$P(z) = \int_{-\infty}^{z} p(x)dx$$

- Note

$$p(x) = P'(x)$$

Probability Densities

- The sum and product rules of probability, as well as Bayes' theorem, apply equally to the case of probability densities

The sum rule

$$p(x) = \int p(x, y)dy$$

The product rule

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

The Bayes' theorem

$$p(x|y) = \frac{p(y|x)p(x)}{p(y)} = \frac{p(y|x)p(x)}{\int p(x, y)dx} = \frac{p(y|x)p(x)}{\int p(y|x)p(x)dx}$$

Expectations and Covariances

- One of the most important operations involving probabilities is that of finding weighted averages of functions.
- The average value of function $f(x)$ under a probability distribution $p(x)$ is called the expectation of $f(x)$ and will be denoted by $E[f]$

$$\mathbb{E}[f] = \sum_x p(x)f(x)$$

Discrete Random Variable

$$\mathbb{E}[f] = \int p(x)f(x)dx$$

Continues Random Variable

- Let X represent the outcome of an unbiased six-sided dice. Suppose that in a sequence of ten rolls of the dice, the outcomes are 5, 2, 6, 2, 2, 1, 2, 3, 6, 1. Then calculate the expected and average values.

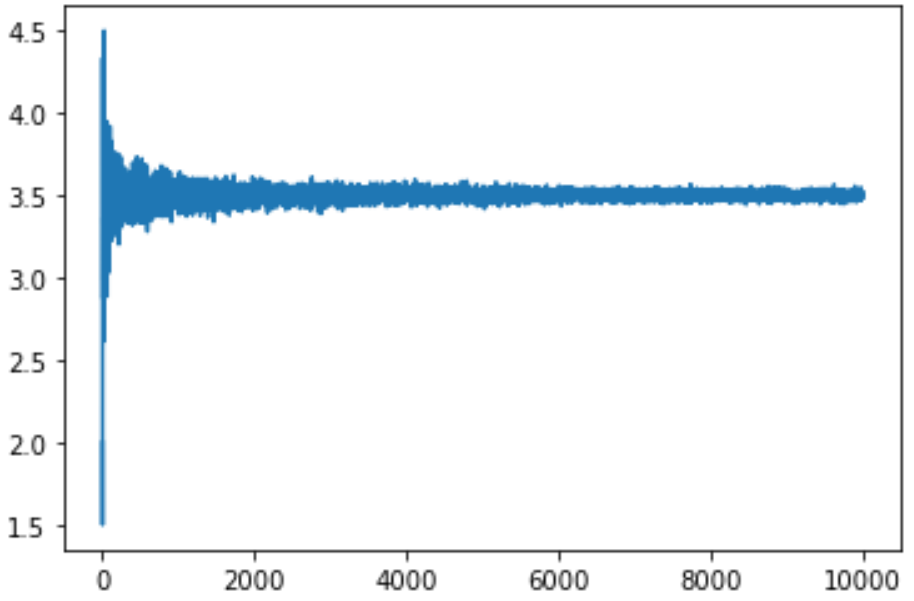
$$\mathbb{E}[x] = \frac{1}{6}(1+2+3+4+5+6) = 3.5$$

$$\bar{X} = \frac{1}{10}(5+2+6+2+2+1+2+3+6+1) = 3$$

Expectations and Covariances

- If we are given a finite number N of points drawn from the probability distribution or probability density, then the expectation can be approximated as a finite sum over these points

$$\mathbb{E}[f] \simeq \frac{1}{N}\sum_{n=1}^{N} f(x_n)$$



- Sometimes we will be considering expectations of functions of several variables, in which case we can use a subscript to indicate which variable is being averaged over

$$\mathbb{E}_x[f(x,y)] = \sum_x p(x) f(x,y)$$

- We can also consider a conditional expectation with respect to a conditional distribution:

$$\mathbb{E}_x[f|y] = \sum_x p(x|y) f(x)$$

Expectations and Covariances

- The variance of f(x) provides a measure of how much variability there is in f(x) around its mean value Type equation here.

$$var[f] = \mathbb{E}[(f(x) - \mathbb{E}[f])^2]$$

- Expanding out the square, we see that the variance can also be written in terms of the expectations of $f(x)$ and $f(x)^2$

$$var[f] = \mathbb{E}[f^2] - \mathbb{E}[f]^2$$

Proof: Homework

- For two random variables x and y, the covariance expresses the extent to which x and y vary together.

$$cov[x, y] = \mathbb{E}_{x,y}[\{x - \mathbb{E}[x]\}\{y - \mathbb{E}[y]\}]$$
$$= \mathbb{E}_{x,y}[xy] - \mathbb{E}[x]\mathbb{E}[y]$$

Proof: Homework

Expectations and Covariances

- Example

$$\mathbb{E}[x] = 0 \times \frac{1}{4} + 1 \times \frac{1}{2} + 2 \times \frac{1}{4} = 1$$

$$\mathbb{E}[y] = 0 \times \frac{1}{4} + 1 \times \frac{1}{2} + 2 \times \frac{1}{4} = 1$$

$$var[x] = \mathbb{E}[(x - \mathbb{E}[x])^2] = \frac{1}{4}(0-1)^2 + \frac{1}{2}(1-1)^2 + \frac{1}{4}(2-1)^2$$

$$= \frac{1}{4} + 0 + \frac{1}{4} = \frac{1}{2}$$

$$var[x] = \mathbb{E}[x^2] - \mathbb{E}[x]^2 = \left(0^2 \times \frac{1}{4} + 1^2 \times \frac{1}{2} + 2^2 \times \frac{1}{4}\right) - (1)^2$$

$$= \left(\frac{1}{2} + 1\right) - 1 = \frac{1}{2}$$

$$var[y] = \mathbb{E}[(y - \mathbb{E}[y])^2] = \frac{1}{2}$$

		y = 0	1	2	$p(x)$
x	0	$\frac{1}{8}$	$\frac{1}{8}$	0	$\frac{1}{4}$
	1	$\frac{1}{8}$	$\frac{2}{8}$	$\frac{1}{8}$	$\frac{1}{2}$
	2	0	$\frac{1}{8}$	$\frac{1}{8}$	$\frac{1}{4}$
$p(y)$		$\frac{1}{4}$	$\frac{1}{2}$	$\frac{1}{4}$	

Expectations and Covariances

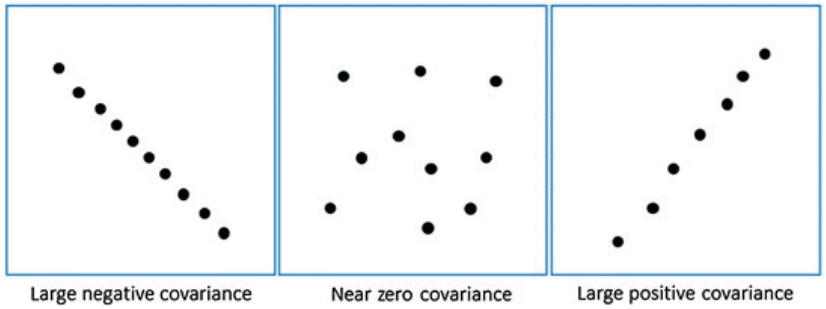
- Example

$$cov[x,y] = \mathbb{E}_{x,y}[\{x-\mathbb{E}[x]\}\{y-\mathbb{E}[y]\}] = \sum_x \sum_y p(x,y)\,(\{x-1\}\{y-1\}) =$$

$$\frac{1}{8}(\{0-1\}\times\{0-1\}) + \frac{1}{8}(\{1-1\}\times\{0-1\}) + 0(\{2-1\}\times\{0-1\})$$
$$+\frac{1}{8}(\{0-1\}\times\{1-1\}) + \frac{2}{8}(\{1-1\}\times\{1-1\}) + \frac{1}{8}(\{2-1\}\times\{1-1\})$$
$$+0(\{0-1\}\times\{2-1\}) + \frac{1}{8}(\{1-1\}\times\{2-1\}) + \frac{1}{8}(\{2-1\}\times\{2-1\})$$
$$= \frac{1}{8}+0+0+0+0+0+0+0+\frac{1}{8} = \frac{1}{4}$$

$$cov[x,y] = \mathbb{E}_{x,y}[xy] - \mathbb{E}[x]\mathbb{E}[y]$$
$$= \sum_x \sum_y xy\,p(x,y) \; - 1 = 0+0+0+0+\frac{2}{8}+\frac{2}{8}+0+\frac{2}{8}+\frac{4}{8}-1 = \frac{1}{4}$$

x \ y	0	1	2	$p(x)$
0	$\frac{1}{8}$	$\frac{1}{8}$	0	$\frac{1}{4}$
1	$\frac{1}{8}$	$\frac{2}{8}$	$\frac{1}{8}$	$\frac{1}{2}$
2	0	$\frac{1}{8}$	$\frac{1}{8}$	$\frac{1}{4}$
$p(y)$	$\frac{1}{4}$	$\frac{1}{2}$	$\frac{1}{4}$	



Expectations and Covariances

Estimating Covariance From Data

$$\boldsymbol{x} = [x_1, x_2, \dots, x_N] \qquad \boldsymbol{y} = [y_1, y_2, \dots, y_N]$$

a. Find the sample mean:

$$\mathbb{E}[x] \simeq \frac{1}{N}\sum_{n=1}^{N} x_n \qquad \mathbb{E}[y] \simeq \frac{1}{N}\sum_{n=1}^{N} y_n$$

b. Calculate $cov(x, y)$ as follows:

$$cov(x, y) = \frac{1}{N}[x_1 - \mathbb{E}[x] \quad x_2 - \mathbb{E}[x] \quad \dots \quad x_N - \mathbb{E}[x]] \begin{bmatrix} y_1 - \mathbb{E}[y] \\ y_2 - \mathbb{E}[y] \\ \vdots \\ y_N - \mathbb{E}[y] \end{bmatrix}$$

Expectations and Covariances

- In the case of two vectors of random variables $\boldsymbol{x}$ and $\boldsymbol{y}$, the covariance is a matrix

$$cov[\boldsymbol{x}, \boldsymbol{y}] = \mathbb{E}_{\boldsymbol{x},\boldsymbol{y}}[\{\boldsymbol{x} - \mathbb{E}[\boldsymbol{x}]\}\{\boldsymbol{y}^T - \mathbb{E}[\boldsymbol{y}^T]\}] = \mathbb{E}_{\boldsymbol{x},\boldsymbol{y}}[\boldsymbol{x}\boldsymbol{y}^T] - \mathbb{E}[\boldsymbol{x}]\mathbb{E}[\boldsymbol{y}^T]$$

$$\boldsymbol{x} = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix} \quad \boldsymbol{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} \quad \Longrightarrow \quad cov(\boldsymbol{x}, \boldsymbol{y}) = \begin{bmatrix} c_{11} & c_{12} & \cdots & c_{1n} \\ c_{21} & c_{22} & \cdots & c_{2n} \\ \vdots & \vdots & \vdots & \vdots \\ c_{n1} & c_{n2} & \cdots & c_{nn} \end{bmatrix}$$

- $c_{ij} = cov(x_i, y_j) = \mathbb{E}[\{x_i - \mathbb{E}[x_i]\}\{y_j - \mathbb{E}[y_j]\}]$
- $c_{ij} = c_{ji}$ (the covariance matrix is symmetric)
- Its diagonal elements are the individual variances: $c_{ii} = var(x_i)$

Expectations and Covariances

Estimating Covariance Matrix From Data

$$\boldsymbol{x} = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1N} \\ x_{21} & x_{22} & \cdots & x_{2N} \\ \vdots & \vdots & \vdots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{nN} \end{bmatrix} \quad \boldsymbol{y} = \begin{bmatrix} y_{11} & y_{12} & \cdots & y_{1N} \\ y_{21} & y_{22} & \cdots & y_{2N} \\ \vdots & \vdots & \vdots & \vdots \\ y_{n1} & y_{n2} & \cdots & y_{nN} \end{bmatrix}$$

a. Find the sample mean vector:

$$\mathbb{E}[\boldsymbol{x}] = \begin{bmatrix} \mathbb{E}[x_1] \simeq \frac{1}{N}\sum_{n=1}^{N} x_1 \\ \mathbb{E}[x_2] \simeq \frac{1}{N}\sum_{n=1}^{N} x_2 \\ \vdots \\ \mathbb{E}[x_n] \simeq \frac{1}{N}\sum_{n=1}^{N} x_n \end{bmatrix} \quad \mathbb{E}[\boldsymbol{y}] = \begin{bmatrix} \mathbb{E}[y_1] \simeq \frac{1}{N}\sum_{n=1}^{N} y_1 \\ \mathbb{E}[y_2] \simeq \frac{1}{N}\sum_{n=1}^{N} y_2 \\ \vdots \\ \mathbb{E}[y_n] \simeq \frac{1}{N}\sum_{n=1}^{N} y_n \end{bmatrix}$$

b. Calculate $cov(x, y)$ as follows:

$$cov(x,y) = \frac{1}{N} \begin{bmatrix} x_{11} - \mathbb{E}[x_1] & x_{12} - \mathbb{E}[x_1] & \dots & x_{1N} - \mathbb{E}[x_1] \\ x_{21} - \mathbb{E}[x_2] & x_{22} - \mathbb{E}[x_2] & \dots & x_{2N} - \mathbb{E}[x_2] \\ & \vdots & & \\ x_{n1} - \mathbb{E}[x_n] & x_{n2} - \mathbb{E}[x_n] & \dots & x_{nN} - \mathbb{E}[x_n] \end{bmatrix} \begin{bmatrix} y_{11} - \mathbb{E}[y_1] & y_{21} - \mathbb{E}[y_2] & & y_{n1} - \mathbb{E}[y_n] \\ y_{12} - \mathbb{E}[y_1] & y_{22} - \mathbb{E}[y_2] & \dots & y_{n2} - \mathbb{E}[y_n] \\ \vdots & \vdots & & \vdots \\ y_{1N} - \mathbb{E}[y_1] & y_{2N} - \mathbb{E}[y_2] & & y_{nN} - \mathbb{E}[y_n] \end{bmatrix}$$

Bayesian Probability

- So far, we have viewed probabilities in terms of the frequencies of random, repeatable events. We shall refer to this as the classical or frequentist interpretation of probability.
- Bayesian probability provides a more general tool for quantification of uncertainty, even on unrepeatable events.
- Repeatable events: selecting a red or blue box, selecting a fruit from a box
- Unrepeatable events: whether the moon was once in its own orbit around the sun, whether the Arctic ice cap will have disappeared by the end of the century

Bayesian Probability

- For example, in the case of fruits problem we know the probability of selecting each box ($p(B = r) = 0.4$ and $p(B = b) = 0.6$). We call $p(B)$ the prior probability because it is the probability available before we observe the identity of the fruit.
- Suppose that we are told that the fruit is an orange (new evidence), then we can use this new evidence to update our idea of the color box ($p(B|F)$). We call $p(B|F)$ the posterior probability of the box color.
- we can then use Bayes' theorem to compute the probability $p(B|F)$

Likelihood

Posterior Probability

Prior Probability

$$p(B|F) = \frac{p(F|B)\ p(B)}{p(F)}$$

$$\sum_B p(F|B)p(B)$$

Bayesian Probability

- For the case of polynomial curve fitting, we capture our assumptions about $\boldsymbol{w}$, before observing the data, in the form of a prior probability distribution $p(\boldsymbol{w})$.
- The effect of the observed data $\mathcal{D} = \{t_1, t_2, \ldots, t_N\}$ is expressed through the conditional probability $p(\mathcal{D}|\boldsymbol{w})$.
- Then the Bayes' theorem allows us to evaluate the uncertainty in $\boldsymbol{w}$ after we have observed $\mathcal{D}$ in the form of the posterior probability $p(\boldsymbol{w}|\mathcal{D})$.

$$p(\boldsymbol{w}|\mathcal{D}) = \frac{p(\mathcal{D}|\boldsymbol{w}) \quad p(\boldsymbol{w})}{p(\mathcal{D})}$$