Speech and Language Processing

Lecture 1
Basics of probability distributions

Information and Communications Engineering Course

Takahiro Shinozaki
Manabu Okumura

2024/9/30

Basics of probability distributions

Probability Space

- Sample space (Ω)
  - Set of all possible outcomes of an experiment
- Probability function ($f(x)$)
  - A function that maps each outcome to a probability

$$f(x) \in [0,1] \quad \text{for all } x \in \Omega$$

$$\sum_{x \in \Omega} f(x) = 1$$

- Event (E)
  - Subset of the sample space

Probability of an event E is : $$P(E) = \sum_{x \in E} f(x)$$

Random Variable

- A function that maps an outcome of an experiment to a value
  - Notation:
    "$P[X=x] = p$" means "the probability of a random variable X takes a value x is p"

Example

X: The value of a die

Y: Whether the value of a die is odd number or not

Z: Whether the value of a die is larger than 2 or not

Random Variable	1	2	3	4	5	6
X	1	2	3	4	5	6
Y	1	0	1	0	1	0
Z	0	0	1	1	1	1

Joint Probability

- Probability that more than one events jointly occur

Example

X : Dice 1

Y : Dice 2

$P(X = i, Y = j)$: Probability that the value of X is i and the value of Y is j

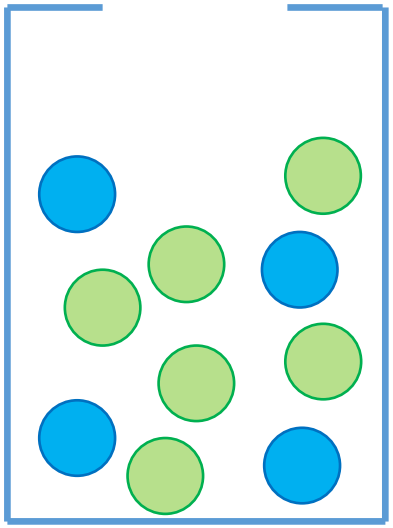
Note: $P(X=i, Y=j) = P(Y=j, X=i)$

Conditional Probability

- Probability of an event given that another event has occurred

Example

Randomly picks up two balls sequentially from a box containing 4 blue and 6 green balls



$$P(\text{"second ball is blue"}|\text{"first ball is geen"}) = \frac{4}{9}$$

$$P(\text{"second ball is blue"}|\text{"first ball is blue"}) = \frac{3}{9}$$

Two Principal Rules

- Sum rule
  - Summing joint probability $P(X,Y)$ for all possible values of Y gives probability of $P(X)$
  - $P(X)$ is called the marginal probability

$$P(X = x_i) = \sum_j P(X = x_i, Y = y_j)$$

- Product rule
  - Product of probability $P(X)$ and conditional probability $P(Y|X)$ is equal to joint probability $P(X,Y)$

$$P(X = x_i, Y = y_j) = P(Y = y_j \mid X = x_i) P(X = x_i)$$

Bayes' Theorem

- From the product rule, we obtain:

$$P(Y = y_j \mid X = x_i) = \frac{P(X = x_i \mid Y = y_j)P(Y = y_j)}{P(X = x_i)} \quad \text{for } \forall x_i, y_i$$

If we simplify the notation, we have:

$$P(Y \mid X) = \frac{P(X \mid Y)P(Y)}{P(X)}$$

(Bayes' theorem)

Using the sum rule, $P(Y|X)$ is obtained from joint probability as:

$$P(Y \mid X) = \frac{P(X \mid Y)P(Y)}{P(X)} = \frac{P(X,Y)}{\sum_Y P(X,Y)}$$

Independence

- If the joint distribution of two variables X and Y factorizes into the product of the marginals, then X and Y are said to be "independent"

$$P(X,Y) = P(X \mid Y)P(Y) = P(X)P(Y)$$ ⇒ X and Y are independent

$$P(X,Y) = P(X \mid Y)P(Y) \neq P(X)P(Y)$$ ⇒ X and Y are not independent

Exercises 1.1,1.2

Joint probability of random variables A and B are given in the following table. According to the table, for example,
$P(A = 0, B = 0) = 0.2$

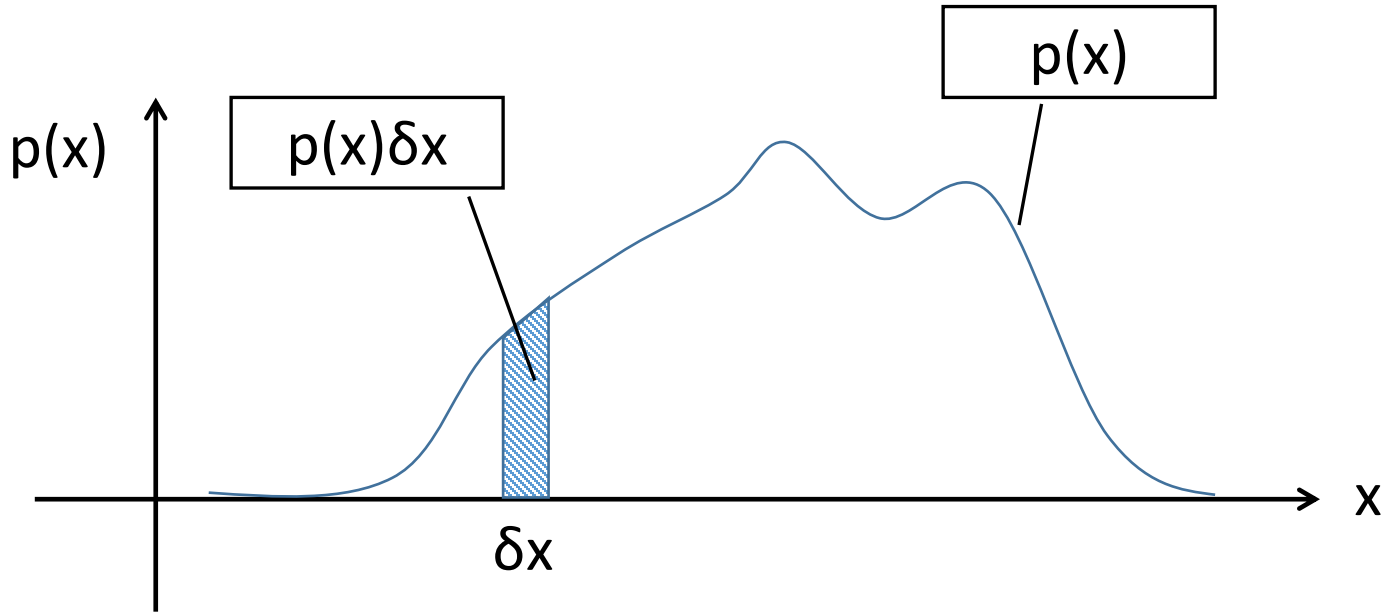
$P(A, B)$	B=0	B=1	B=2
A=0	0.2	0.1	0.1
A=1	0.3	0.2	0.1

1. Obtain $P(A = 0)$
2. Obtain $P(A = 0|B = 0)$

Probability Densities

- If the probability of a real-valued variable x falling in the interval $(x, x+\delta x)$ is given by $p(x)\delta x$ when $\delta x \to 0$, $p(x)$ is called the probability density of x

p(x)δx is probability ➡ $p(x) \geq 0$ and $\int_{-\infty}^{\infty} p(x)dx = 1$



The Sum and The Product Rules For Continuous Variable

$$P(X = x_i) = \sum_j P(X = x_i, Y = y_j)$$

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

$$P(X = x_i, Y = y_j) = P(Y = y_j \mid X = x_i)P(X = x_i)$$

$$\Rightarrow \quad p(x, y) = p(y \mid x)p(x)$$

Expectation

- Expectation of a function $f(x)$ under a probability distribution $p(x)$ is denoted by $E[f]$

$$E[f] = \sum_x p(x) f(x)$$ (x is discrete)

$$E[f] = \int p(x) f(x) dx$$ (x is continuous)

Mean and Variance

- Mean
  - Synonym of the expectation $E[f(x)]$
- Variance
  - A measure of how much variability there is in $f(x)$ around its mean value $E[f(x)]$

$$\mathrm{var}[f] \equiv E\left[\left(f(x) - E[f(x)]\right)^2\right] = E\left[f(x)^2\right] - E[f(x)]^2$$

  - In particular, the variance of the variable x itself is:

$$\mathrm{var}[x] = E\left[\left(x - E[x]\right)^2\right]$$

Covariance

- Covariance
  - The extent to which x and y vary together

$$\begin{aligned}\mathrm{cov}[x, y] &\equiv E_{p(x,y)}[(x - E[x])(y - E[y])] \\ &= E_{p(x,y)}[xy] - E[x]E[y]\end{aligned}$$

Expectation with respect to joint probability of x and y

Entropy

- Amount of randomness in the random variable

$$H[p] = E\big[-\log\big(p(x)\big)\big]$$

Example

x	0	1
p(x)	0.5	0.5

$$H[x] = -\sum_x p(x)\log p\,(x)$$
$$= -0.5\log(0.5) - 0.5\log(0.5)$$
$$= 0.693$$

x	0	1
p(x)	0.1	0.9

$$H[x] = -\sum_x p(x)\log p\,(x)$$
$$= -0.1\log(0.1) - 0.9\log(0.9)$$
$$= 0.325$$

Relative Entropy

- A measure of dissimilarity of two distributions p and q
  - Also called as kullback-Leibler (KL) divergence

$$KL[p||q] = E_p\left[\log\left(\frac{p(x)}{q(x)}\right)\right] = -\int p(x)\log\left(\frac{q(x)}{p(x)}\right)dx$$

- $KL[p||q]$ is nonnegative.
  $KL[p||q] = 0$ if and only if $p(x) = q(x)$
- In general, $KL[p||q] \neq KL[q||p]$

Cross Entropy

- The cross-entropy between p and q over the same underlying set of events measures the average number of bits needed to identify an event drawn from the set if a coding scheme used for the set is optimized for an estimated probability distribution q rather than the true distribution p.

$$CE[p,q] = E_p[-\log q(x)] = -\int p(x) \log q(x)\, dx$$

- Relationship with entropy and KL divergence

$$CE[p,q] = H[p] + KL[p||q]$$

Categorical distribution

- A probability model for a categorical random variable
- The distribution is represented by a table
- An example is the probability distribution of a skewed die

Number	1	2	3	4	5	6
Probability	0.3	0.1	0.1	0.1	0.3	0.1

1-of-K Representation

- The same probability as the table description can be expressed as an equation by using 1-of-K representation

Value v	1-of-K representation $X_v=(x_{v,1}, x_{v,2}, x_{v,3}, x_{v,4}, x_{v,5})$	Probability $\boldsymbol{\alpha}=(\alpha_1, \alpha_2, \alpha_3, \alpha_4, \alpha_5)$
1 （あ）	1,0,0,0,0	$Pr(X=1)=\alpha_1=0.3$
2 （い）	0,1,0,0,0	$Pr(X=2)=\alpha_2=0.1$
3 （う）	0,0,1,0,0	$Pr(X=3)=\alpha_3=0.2$
4 （え）	0,0,0,1,0	$Pr(X=4)=\alpha_4=0.1$
5 （お）	0,0,0,0,1	$Pr(X=5)=\alpha_5=0.3$

$$p(v|\boldsymbol{\alpha}) = p(X_v|\boldsymbol{\alpha}) = \prod_{k=1}^{K} {\alpha_k}^{x_{v,k}}$$

Exercise 1.3

- When $P(v|\boldsymbol{\alpha})$ is given as follows, obtain P("i"$|\boldsymbol{\alpha}$)

Value v	1-of-K representation $X_v=(x_{v,1}, x_{v,2}, x_{v,3}, x_{v,4}, x_{v,5})$
1 （a）	1,0,0,0,0
2 （i）	0,1,0,0,0
3 （u）	0,0,1,0,0
4 （e）	0,0,0,1,0
5 （o）	0,0,0,0,1

$$P(v|\boldsymbol{\alpha}) = \prod_{k=1}^{5} {\alpha_k}^{x_{v,k}}$$

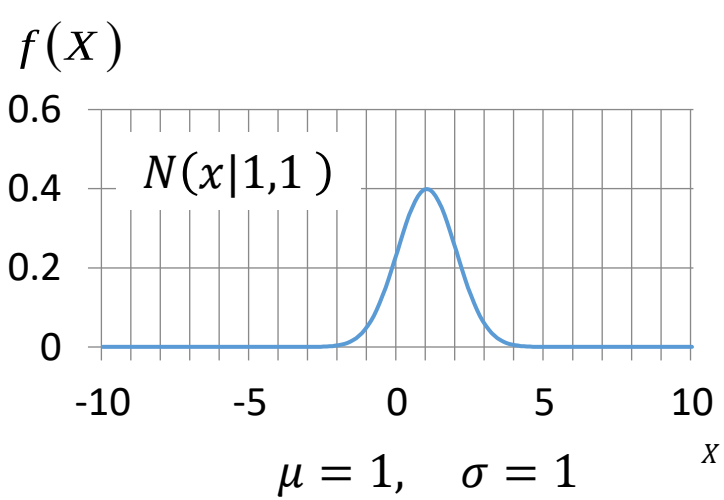
$$\boldsymbol{\alpha} = (0.3,\ 0.2,\ 0.1,\ 0.1,\ 0.3)$$

Gaussian Distribution

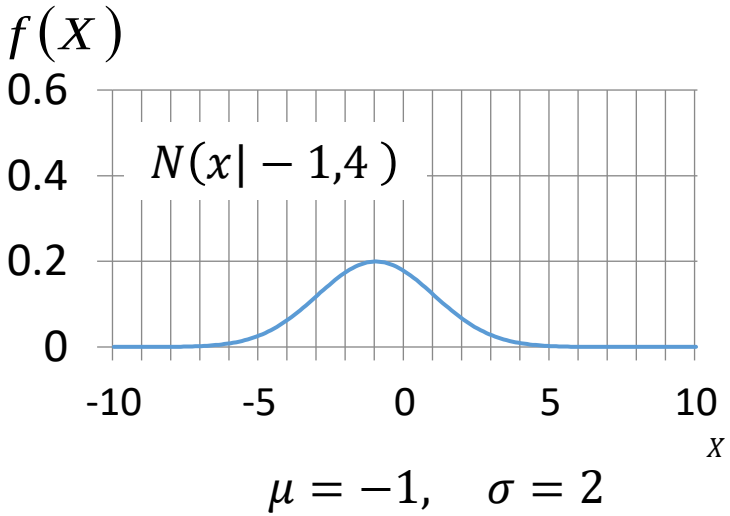
- Defined by two parameters mean μ and standard deviation σ (σ^2 is variance)

$$N(x|\mu,\sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}}\exp\left\{-\frac{1}{2\sigma^2}(x-\mu)^2\right\}$$

It satisfies: $0 < N(x|\mu,\sigma^2), \quad \int_{-\infty}^{\infty} N(x|\mu,\sigma^2)dx = 1$



$\mu = 1, \quad \sigma = 1$



$\mu = -1, \quad \sigma = 2$

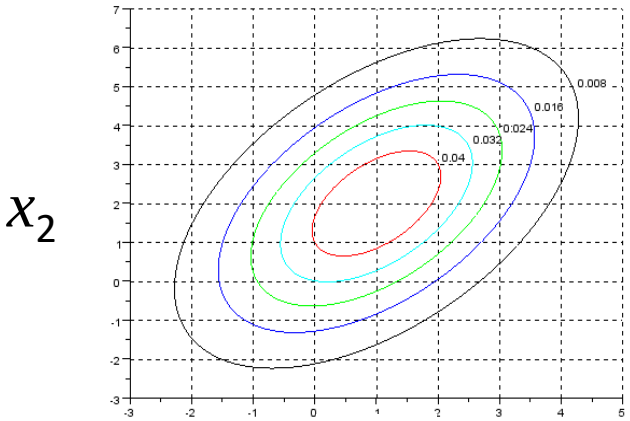
Multivariate Gaussian Distribution

- For D-dimensional vector $\boldsymbol{x}$, it is defined using a mean vector $\boldsymbol{\mu}$ and a covariance matrix $\boldsymbol{S}$:

$$N(\boldsymbol{x}|\boldsymbol{\mu},\boldsymbol{S}) = \frac{1}{\sqrt{(2\pi)^D|\boldsymbol{S}|}}\exp\left\{-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^T\boldsymbol{S}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right\}$$

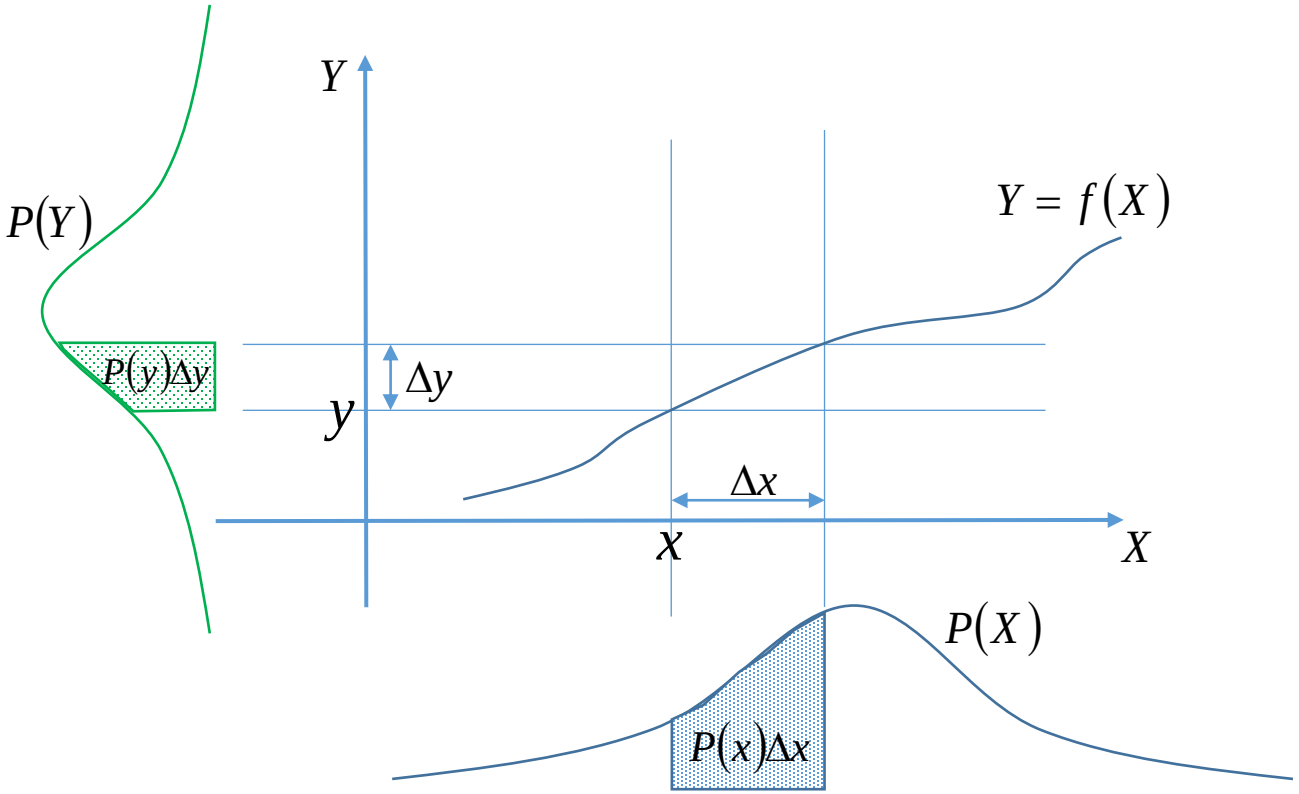
$|\boldsymbol{S}|$ denotes determinant of $\boldsymbol{S}$

Contour plot of an example of a two dimensional Gaussian distribution



Transformation of Continuous Random Variable

- Suppose *X* and *Y* are random variables and *Y=f(X)*



$$P_X(x)\Delta x \approx P_Y(y)\Delta y$$

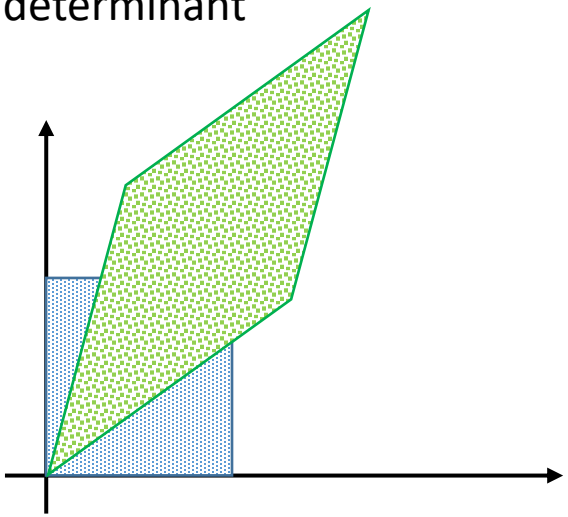
$$P_Y(y) \approx \frac{\Delta x}{\Delta y} P_X(x)$$

$$P_Y(y) = \left|\frac{dx}{dy}\right| P_X(x)$$

Transformation of Vector Variable

$$Y = f(X)$$

The volume (area) ratio by a linear transformation is equal to the absolute value of determinant



$$P_Y(Y)\Delta V_Y = P_X(X)\Delta V_X$$

$$P_Y(Y) = \left|\frac{\Delta V_X}{\Delta V_Y}\right| P_X(X)$$

$$P_Y(y_1, y_2, \cdots, y_N)$$

$$= \left|\left|\frac{\partial(x_1, x_2, \cdots, x_N)}{\partial(y_1, y_2, \cdots, y_N)}\right|\right| P_X(x_1, x_2, \cdots, x_N)$$

$$= \left|\begin{vmatrix} \frac{\partial x_1}{\partial y_1} & \frac{\partial x_1}{\partial y_2} & \cdots & \frac{\partial x_1}{\partial y_N} \\ \frac{\partial x_2}{\partial y_1} & \ddots & & \vdots \\ \vdots & & \ddots & \\ \frac{\partial x_N}{\partial y_1} & \cdots & & \frac{\partial x_N}{\partial y_N} \end{vmatrix}\right| P_X(x_1, x_2, \cdots, x_N)$$

Absolute value

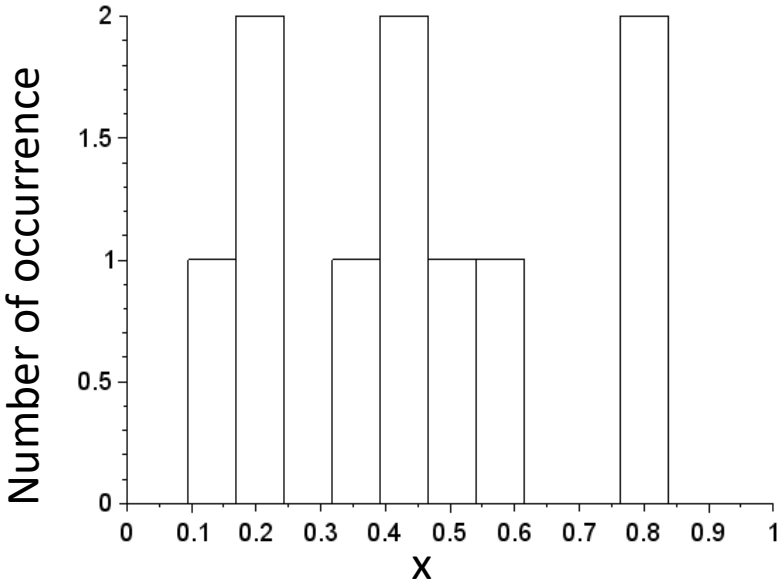
Jacobian determinant of the inverse function

Sampling From a Uniform Distribution

- Samples distribute uniformly over some region

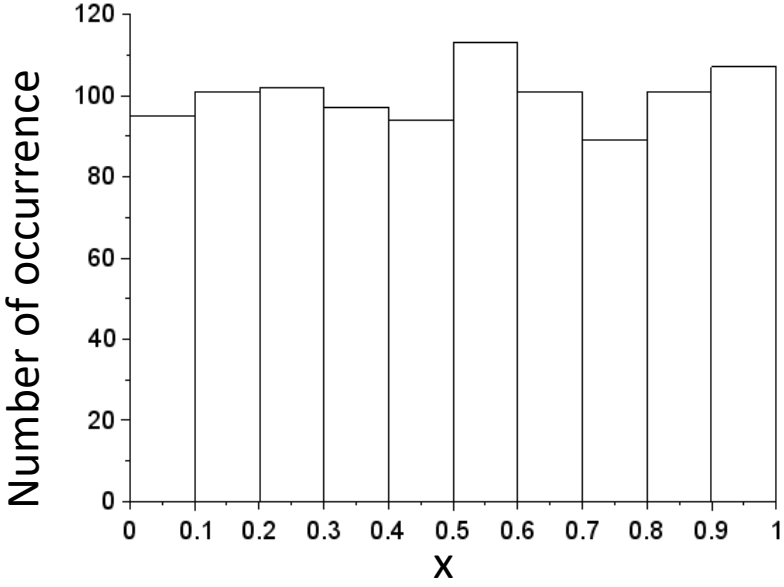
  Example:
  Histogram of samples obtained from a uniform distribution over (0, 1)

10 samples



histplot(10,rand(1:10),
normalization=%f)

1000 samples

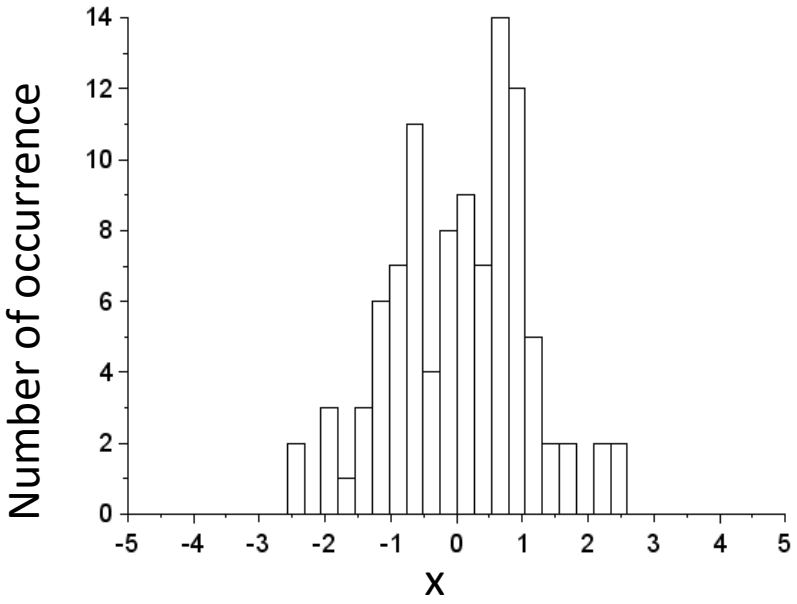


histplot(10,rand(1:1000),
normalization=%f)

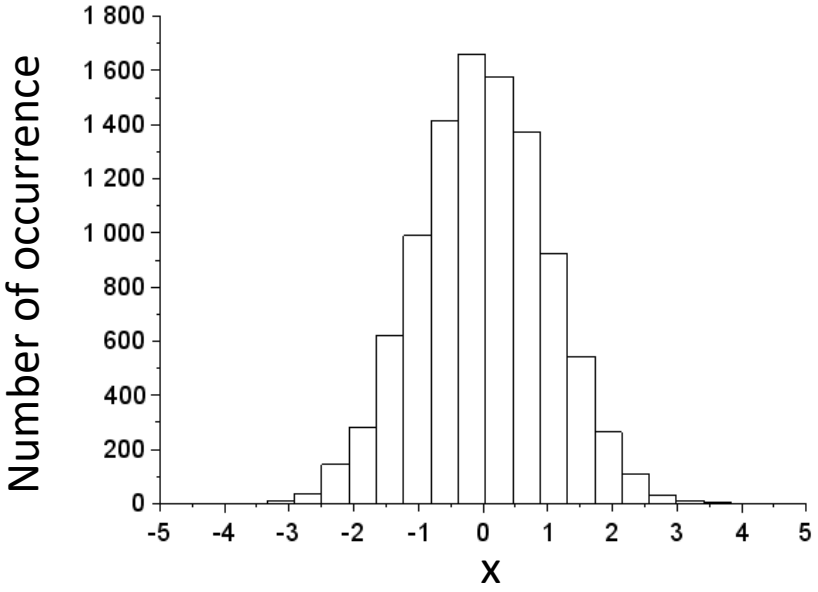
Sampling From a Gaussian Distribution

- Standard normal (Gaussian) distribution has a mean 0.0 and a variance 1.0

100 samples



10000 samples



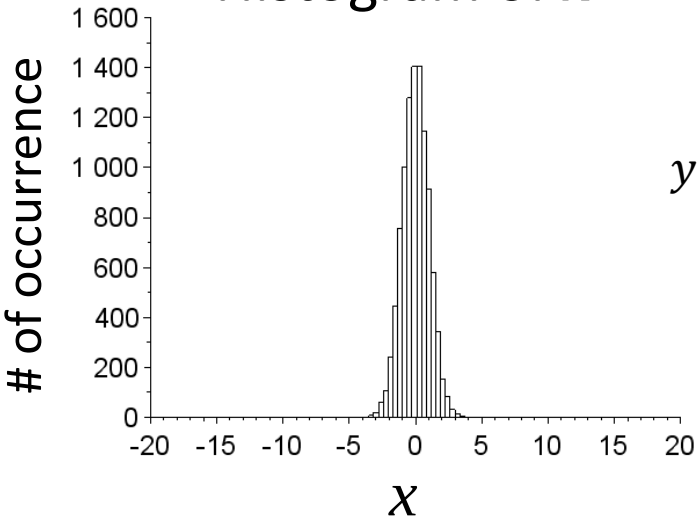
Exercise 1.4

Assume $p(x)$ and $y = f(x)$ are given as follows

$$p(x) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{x^2}{2}\right) = N(x|0,1) \quad x \in (-\infty, \infty), \qquad y = 3x + 4$$

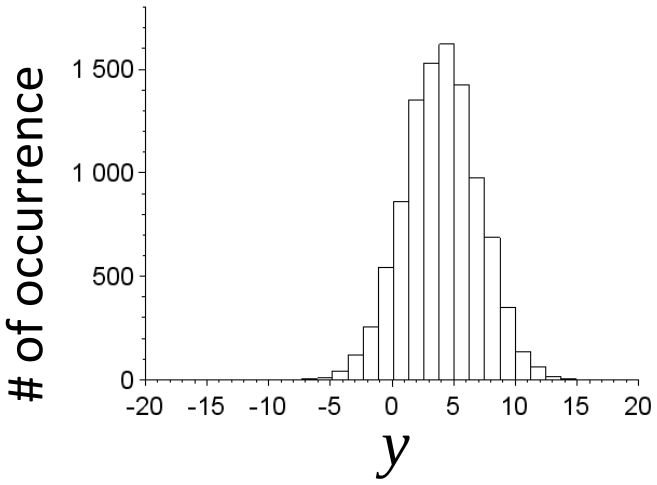
Then, $q(y)$ becomes a Gaussian distribution $N(y|E, 3^2)$. Obtain E

Histogram of x



$y = 3x + 4$

Histogram of y



Approximating Expectation with Sampling

When $x_1, \mathrm{x}_2, \cdots, x_N$ are samples independently drawn from a distribution $p(x)$

$$E[f] = \sum_x p(x) f(x) \approx \frac{1}{N} \sum_{n=1}^{N} f(x_n) \qquad (x \text{ is discrete})$$

$$E[f] = \int p(x)\, f(x) dx \approx \frac{1}{N} \sum_{n=1}^{N} f(x_n) \qquad (x \text{ is continuous})$$

Advanced

Maximum Likelihood (ML) Method

- Assume that we have a set of samples $D=\{x_1, x_2, \ldots x_n\}$ drawn from a distribution $p(x|\boldsymbol{\theta})$ with unknown parameters $\boldsymbol{\theta}$, and we want to estimate $\boldsymbol{\theta}$

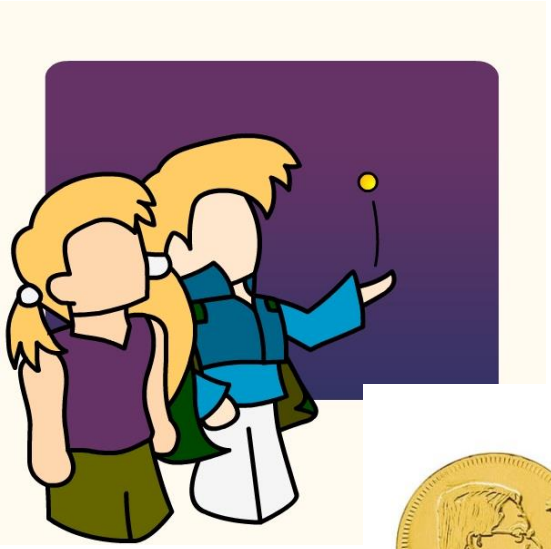
- Maximum likelihood method estimates the parameters by maximizing likelihood $p(D|\theta)$

$$\hat{\theta} = \arg\max_{\theta} p(D \mid \theta) = \arg\max_{\theta} \prod_{i=1}^{n} p(x_i \mid \theta)$$

Probability of the data set D is decomposed to a product of samples when they are drawn independently

Bernoulli Distribution

- Probability distribution of a binary random variable which takes value 1 with probability μ and value 0 with probability 1-μ



x	0	1
Bern(x)	1-μ	μ

$$Bern(x) = \mu^{x}(1-\mu)^{1-x}$$

Is the result Head or Tail?

ML Estimation for Bernoulli Distribution

- Parameter θ in this case is : μ
- Training sample $x_i = 0 \; or \; 1$

$$\hat{\mu} = \arg\max_{\mu} p(D \mid \mu) = \arg\max_{\mu} \prod_{i=1}^{n} \mu^{x_i} (1-\mu)^{1-x_i}$$

$$= \arg\max_{\mu} \log\left(\prod_{i} \mu^{x_i} (1-\mu)^{1-x_i} \right)$$

$$= \arg\max_{\mu} \left\{ \sum_i x_i \log(\mu) + \sum_i (1-x_i) \log(1-\mu) \right\}$$

$$\frac{\partial}{\partial \mu} \left(\sum_i x_i \log(\mu) + \sum_i (1-x_i) \log(1-\mu) \right) = 0$$

$$\Leftrightarrow \quad \mu = \frac{1}{n} \sum_i x_i$$

Taking log does not change the problem and makes the equation a bit easier

n: the number of samples

Categorical Distribution

- As a generalization of the Bernoulli Distribution, lets consider a discrete random variable X that takes K values

X	1	2	…	K
1-of-K $<x_1,x_2,\ldots,x_K>$	$<1,0,..,0>$	$<0,1,..,0>$	…	$<0,0,..,1>$
$p(X)$	μ_1	μ_2	…	μ_K

$$p(\mathbf{x}\,|\,\boldsymbol{\mu}) = \prod_{k=1}^{K} \mu_k^{\,x_k}$$

ML for Categorical Distribution

- Parameter θ in this case is : $\boldsymbol{\mu} = \{\mu_1, \mu_2, \ldots, \mu_K\}$
- Training sample x_i is a vector of 1-of-K representation. When x_i represents k-th value, $x_{i,k}$=1, and $x_{i,j}$=0 for $j \neq k$

$$\hat{\boldsymbol{\mu}} = \arg\max_{\boldsymbol{\mu}} p(D \mid \boldsymbol{\mu}) = \arg\max_{\boldsymbol{\mu}} \prod_{i=1}^{n} \prod_{k=1}^{K} \mu_k^{x_{i,k}} = \arg\max_{\boldsymbol{\mu}} \prod_{k=1}^{K} \mu_k^{m_k}$$

$$\sum_{k=1}^{K} \mu_k = 1$$

Constraint

m_k is the number of the occurrence of k-th value, where n is the number of samples

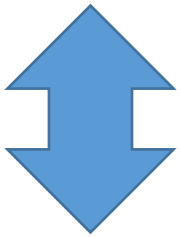
$$m_k = \sum_{i=1}^{n} x_{i,k}$$

This is a maximization problem with a constraint

➡ Use the method of Lagrange multiplier

Method of Lagrange Multiplier

Maximize $f(X)$ subject to $g(X)=0$



Equivalent

Maximize $f(X)$-$\lambda g(X)$
with respect to X and λ,
where λ is a new parameter

Solution

$$\hat{\boldsymbol{\mu}} = \arg\max_{\boldsymbol{\mu}} \prod_{k=1}^{K} \mu_k^{\,m_k} = \arg\max_{\boldsymbol{\mu}} \sum_{k=1}^{K} m_k \log(\mu_k)$$

$$\sum_{k=1}^{K} \mu_k = 1$$

$$\Rightarrow \quad \hat{\boldsymbol{\mu}} = \arg\max_{\boldsymbol{\mu}} \left\{ \sum_{k=1}^{K} m_k \log \mu_k + \lambda \left(\sum_{k=1}^{K} \mu_k - 1 \right) \right\}$$

$$\Rightarrow \quad \mu_k = \frac{m_k}{n}$$

Exercise 1.5

- Show the derivation process of obtaining $\mu_k = \frac{m_k}{n}$

  for the categorical distribution by maximizing

$$L(\boldsymbol{\mu}, \lambda) = \sum_{k=1}^{K} m_k \log \mu_k + \lambda \left(\sum_{k=1}^{K} \mu_k - 1 \right)$$

  where λ is the Lagrange multiplier.