

# CS599—Statistics Refresher

---

- Goal: Refresh Statistics & Probability Theory

- u random variables (discrete & continuous)
- u distributions (discrete & continuous)
- u expected values, moments
- u joint distributions, conditional distributions, independence
- u Bayes Rule

- If you are interested in statistics books:

- u Intro. book: Rice, J.A. (1987): Mathematical Statistics and Data Analysis. Wadsworth.
- u Great book with very advanced concepts, but heavy notation: Papoulis, A. (1991). Probability, Random Variables, and Stochastic Processes. McGraw-Hill.

# Random Variables

---

- A random variable is a random number determined by chance, or more formally, drawn according to a probability distribution
  - u the probability distribution can be given by the physics of an experiment (e.g., throwing dice)
  - u the probability distribution can be synthetic
  - u discrete & continuous random variables
- Typical random variables in Neural Nets:
  - u the input data
  - u the output data
  - u noise
- Important concept in learning: The data generating model
  - u e.g., what is the data generating model for: i) throwing dice, ii) regression, iii) classification, iv) for visual perception?
- Problems:
  - u on which time scale does one observe a distribution

# Discrete Probability Distributions

---

- The random variables only take on discrete values

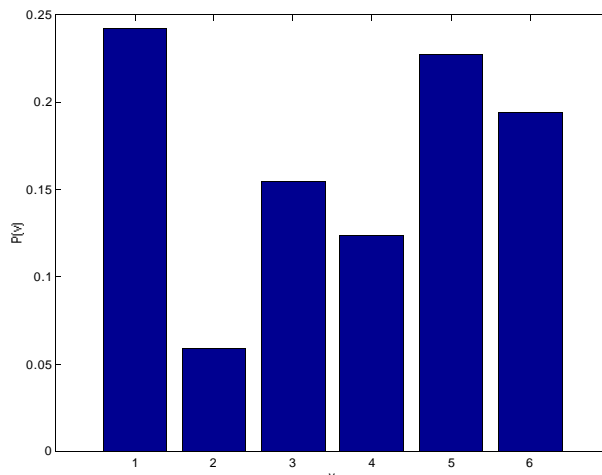
+ e.g., throwing dice: possible values

$$v_i \in \{1, 2, 3, 4, 5, 6\}$$

- The probabilities sum to 1

$$\sum_i P(v_i) = 1$$

- Discrete distributions are particularly important in classification
- Probability Mass Function or Frequency Function (normalized histogram)



A “non fair” die

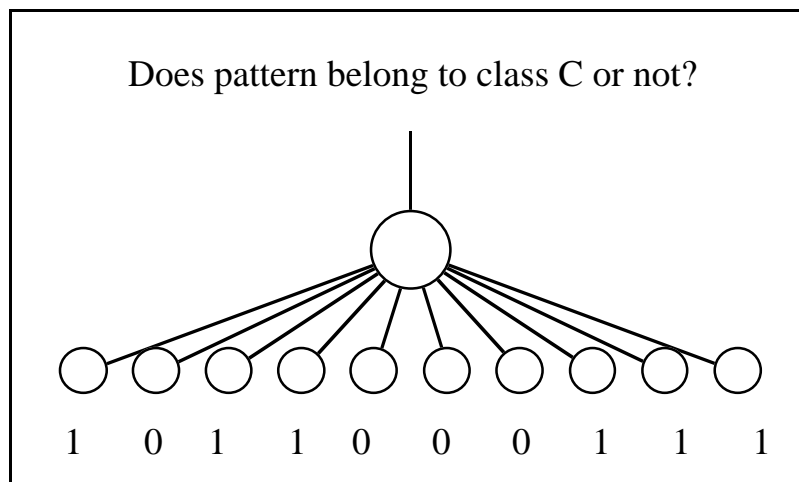
# Classic Discrete Distributions: The Bernoulli Distribution

---

- A Bernoulli random variable takes on only two values, i.e., 0 and 1
- $P(0)=p$  and  $P(1)=1-p$ , or in compact notation:

$$P(x) = \begin{cases} p^x(1-p)^{1-x}, & \text{if } x = 0 \text{ or } x = 1 \\ 0, & \text{otherwise} \end{cases}$$

- As we will see later, Bernoulli distributions are naturally modeled by sigmoidal activation functions in neural networks (Bishop, Ch.1 & h.3) with binary inputs.



# Classic Discrete Distributions: The Binomial Distribution

---

- Like Bernoulli distribution: binary input variables: 0 or 1, and probability  $P(0)=p$  and  $P(1)=1-p$
- What is the probability of  $k$  successes,  $P(k)$ , in a series of  $n$  independent trials? ( $n \geq k$ )
- $P(k)$  is a binomial random variable:

$$P(k) = \binom{n}{k} p^k (1-p)^{n-k} \text{ where } \binom{n}{k} = \frac{n!}{k!(n-k)!}$$

- Binomial variables are important for density estimation networks, e.g. “what is the probability that  $k$  data points fall into region  $R$ ?” (Bishop, Ch.2)
- Bernoulli distribution is a subset of binomial distribution (i.e.,  $n=1$ )

# Classic Discrete Distributions: The Multinomial Distribution

---

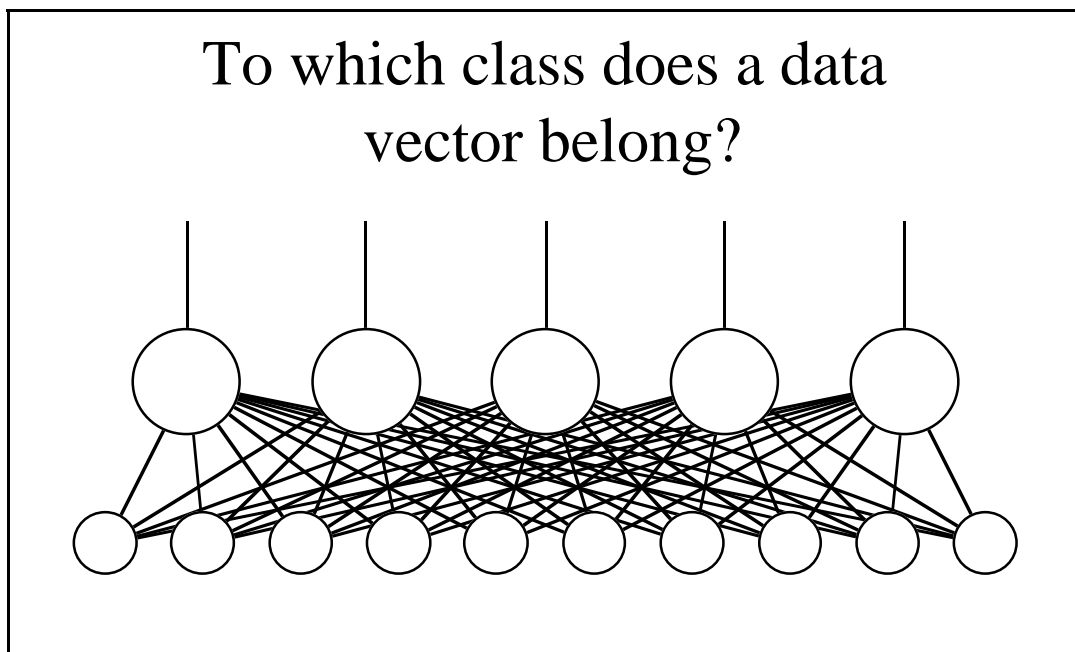
- A generalization of the binomial distribution to multiple outputs (i.e., multiple classes can be categorized instead of just one class).
- $n$  independent trials can result in one of  $r$  types of outcomes, where each outcome  $c_r$  has a probability  $P(c_r)=p_r$  ( $\sum p_r=1$ ).
- What is the probability  $P(n_1, n_2, \dots, n_r)$ , i.e., the probability that in  $n$  trials, the frequency of the  $r$  classes is  $(n_1, n_2, \dots, n_r)$ ? This is a multinomial random variable:

$$P(n_1, \dots, n_r) = \binom{n}{n_1 n_2 \dots n_r} p_1^{n_1} p_2^{n_2} \dots p_r^{n_r} \text{ where } \binom{n}{n_1 n_2 \dots n_r} = \frac{n!}{n_1! n_2! \dots n_r!}$$

# The Multinomial Distribution (cont'd)

---

- The multinomial distribution plays a most important role in multi-class classification (where  $n=1$ )



# Classic Discrete Distributions: The Poisson Distribution

---

- The Poisson distribution is binomial distribution where the number of trials  $n$  goes to infinity, and the probability of success on each trial,  $p$ , goes to zero, such that  $np = \lambda$

$$P(k) = \frac{\lambda^k}{k!} e^{-\lambda}$$

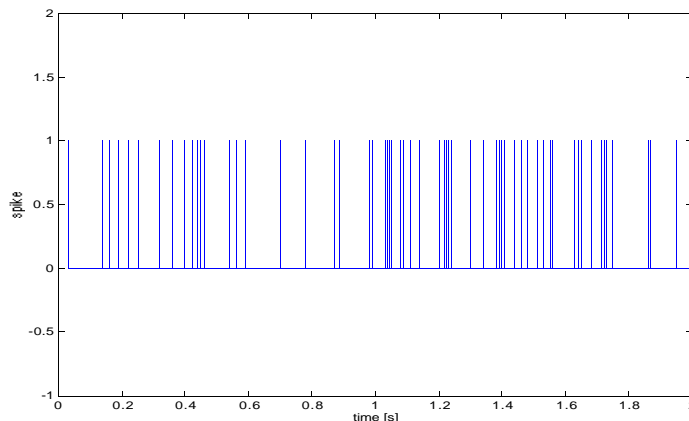
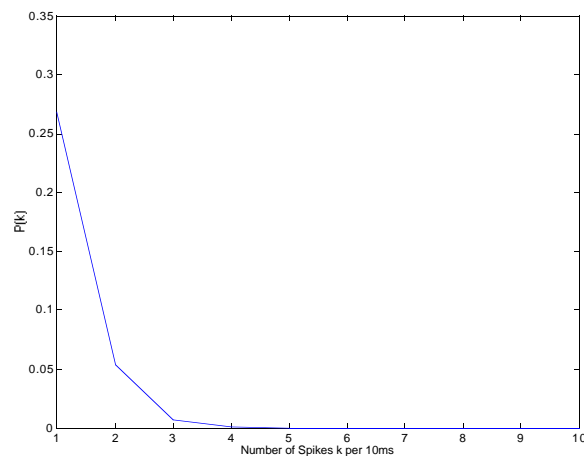
- Poisson distributions are an important model for the firing characteristics of biological neurons. They are also used as an approximation to binomial variables with small  $p$ .

# Poisson Distribution (cont'd)

---

- Example: What is the Poisson distribution of neuronal firing of a cerebellar Purkinje cell in a 10ms interval?

- u we know that the average firing rate of a pyramidal cell is 40Hz
- u  $\lambda = 40\text{Hz} * 0.01\text{s} = 0.4$
- u note that approximation only works if probability of spiking is small in the considered interval



# Continuous Probability Distributions

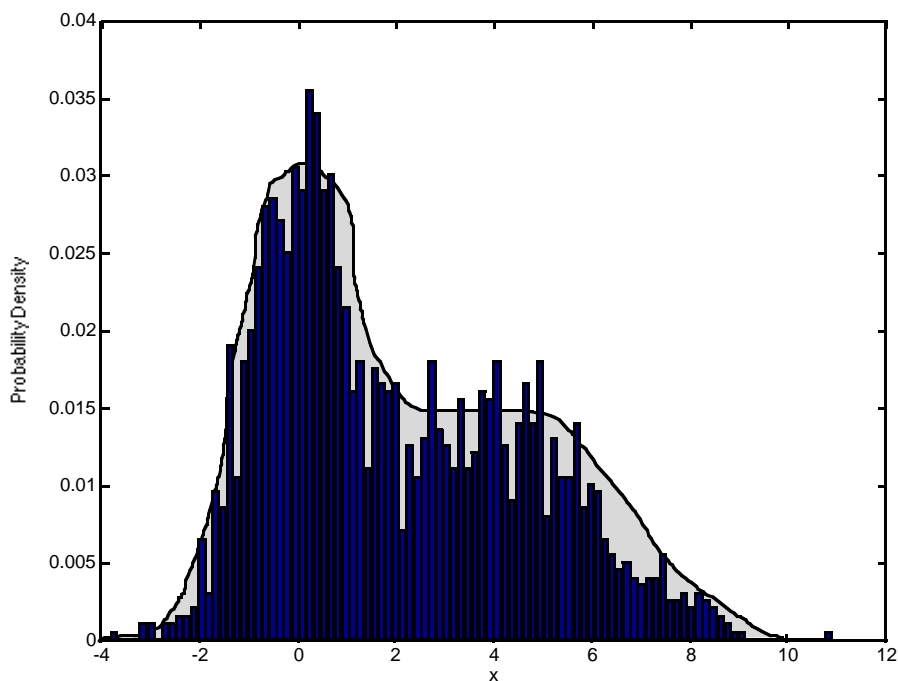
---

- Random variables take on real values
- Continuous distributions are discrete distributions where the number of discrete values goes to infinity while the probability of each discrete value goes to zero.
- Probabilities become densities
- Probability density integrates to 1
$$\int_{-\infty}^{+\infty} p(x)dx = 1$$
- Continuous distributions are particularly important in regression

# Continuous Probability Distributions (cont'd)

---

- Probability Density Function  $p(x)$



- Probability of an event:

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

# Classic Cont. Distributions: The Normal Distribution

---

- The most important continuous distribution

$$p(\mathbf{x}) = \frac{1}{\sqrt{(2\pi)^d |\Sigma|}} \exp\left(-\frac{1}{2}(\mathbf{x} - \mu)^T \Sigma^{-1}(\mathbf{x} - \mu)\right)$$

- Also called Gaussian distribution after C.F.Gauss who proposed it
- Justified by the Central Limit Theorem:
  - u roughly: “if a random variable is the sum of a large number of independent random variables, it is approximately normally distributed”
  - u Many observed variables are the sum of several random variables
- Shorthand:  $x \sim N(\mu, \Sigma)$



# The Normal Distribution (cont'd)

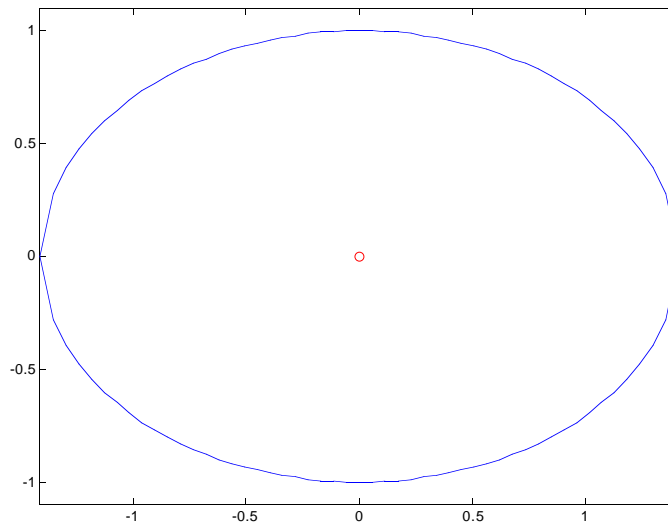
---

- Mahalanobis Distance:

$$r^2 = (\mathbf{x} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\mathbf{x} - \boldsymbol{\mu})$$

- u points with equal  $r$  are points of constant density
- u the contours of constant density are (hyper-) ellipsoids
- u the principle axes of these hyperellipsoids are given by the eigenvectors of .
- u Example 1: Draw  $r = 1$  for

$$\boldsymbol{\mu} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}, \quad \boldsymbol{\Sigma} = \begin{bmatrix} 2 & 0 \\ 0 & 1 \end{bmatrix}$$

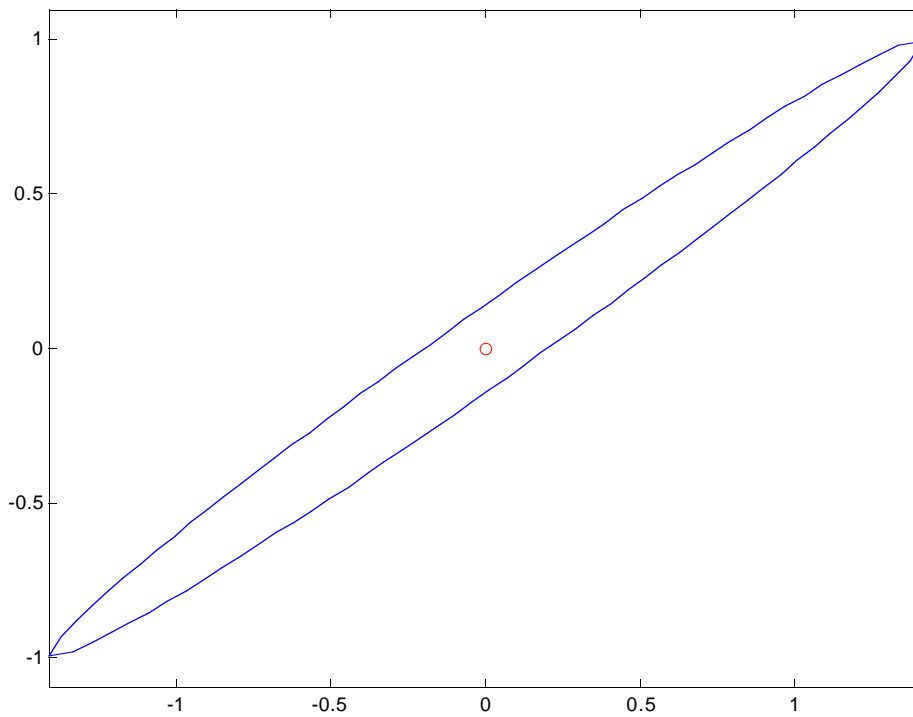


# Mahalanobis Distance (cont'd)

---

u Example 2: Draw  $r=1$  for

$$\mu = \begin{bmatrix} 0 \\ 0 \end{bmatrix}, \quad \Sigma = \begin{bmatrix} 2 & 1.4 \\ 1.4 & 1 \end{bmatrix}$$



# Classic Cont. Distributions: The Exponential Family

---

- A large class of distributions that are all analytically appealing. Why? Because taking the  $\log()$  of them decomposes them into simple terms.

$$p(\mathbf{x}) = \exp\left(\frac{x\theta - b(\theta)}{a(\phi)} + c(x, \phi)\right)$$

for some specific functions  $a()$ ,  $b()$ , and  $c()$ , and parameter vectors  $\theta$  and  $\phi$ .

- All members are unimodal.
- However, there are many “daily”-life distributions that are not captured by the exponential family.

# Exponential Family (cont'd)

---

- Example: Univariate Gaussian

$$p(\mathbf{x}) = \exp\left(\frac{x\theta - b(\theta)}{a(\phi)} + c(x, \phi)\right)$$

$$\theta = \mu, \quad \phi = \sigma^2$$

$$a(\phi) = \phi, \quad b(\theta) = \frac{\theta^2}{2},$$

$$c(x, \phi) = -\frac{1}{2}\left(\frac{x^2}{\phi} + \log(2\pi\phi)\right)$$

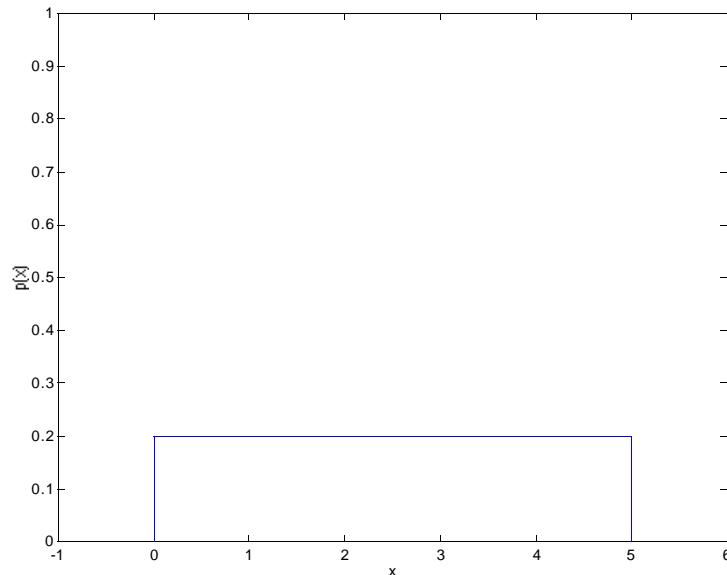
- Other members of the family include:

- u Exponential distribution
- u Rayleigh distribution
- u Maxwell distribution
- u Gamma distribution
- u Beta distribution
- u Poisson distribution
- u Binomial distribution
- u Multinomial distribution

# Classic Cont. Distributions: The Uniform Distribution

---

- All data is equally probable within a bounded region  $R$ ,  $p(\mathbf{x})=1/R$ .



Uniform distributions play a very important role in neural networks based on information theory and entropy methods.

# Expected Values

---

- Definition for discrete random variables:

$$E\{\mathbf{x}\} = \sum_i \mathbf{x}_i P(\mathbf{x}_i) = \langle \mathbf{x} \rangle$$

- Definition for continuous random variables:

$$E\{\mathbf{x}\} = \int_{-\infty}^{+\infty} \mathbf{x}_i p(\mathbf{x}_i) d\mathbf{x} = \langle \mathbf{x} \rangle$$

- $E\{\mathbf{x}\}$  is often called the MEAN of  $\mathbf{x}$ .
- $E\{\mathbf{x}\}$  is the “Center of Mass” of the distribution.

+ Example I: What is the mean of a normal distribution?

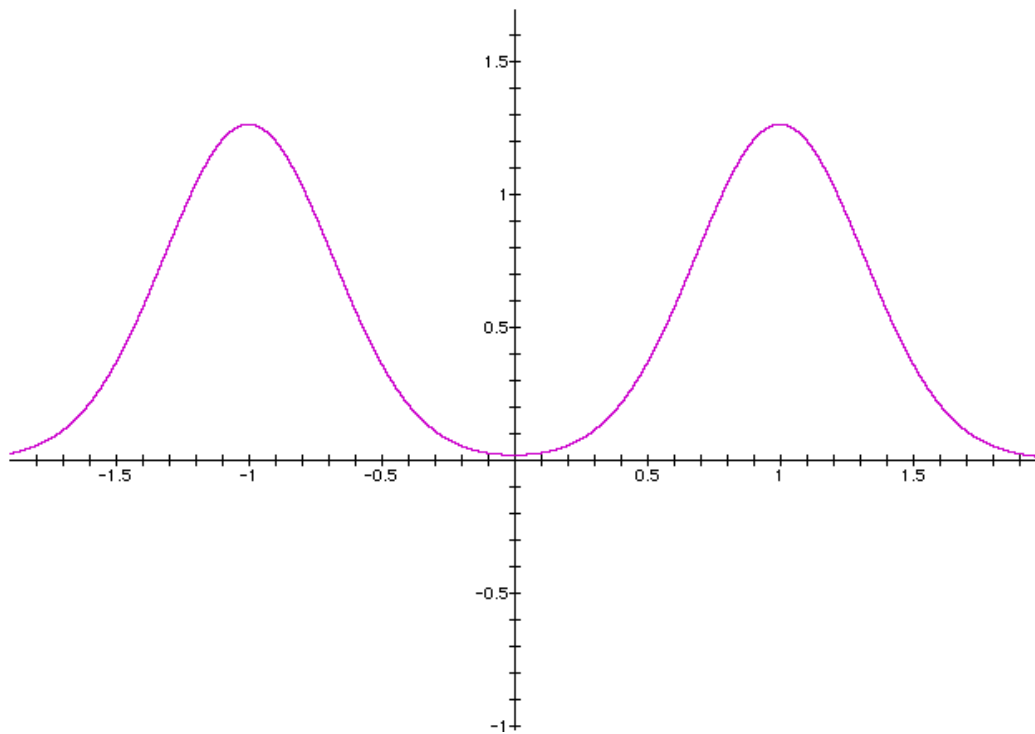
$$p(\mathbf{x}) = \frac{1}{\sqrt{(2\pi)^d |\Sigma|}} \exp\left(-\frac{1}{2}(\mathbf{x} - \mu)^T \Sigma^{-1}(\mathbf{x} - \mu)\right)$$

Note: The Expectation of a variable is often assume to be the most probable value of the variable -- but this may go wrong!

# Expected Values (cont'd)

---

+ Example II: What is the mean of the distribution below?



# Sample Expectation

---

- Given a FINITE sample of data, what is the Expectation?

$$E\{\mathbf{x}\} = \frac{1}{N} \sum_{i=1}^N \mathbf{x}_i$$

- What happened to the probabilities?

# Expectation of Functions of Random Variables

---

- $E\{g(\mathbf{x})\} = ?$

+ as long as sum (or integral) remain bounded, just replace  $\mathbf{x} \cdot p(\mathbf{x})$  with  $g(\mathbf{x}) \cdot p(\mathbf{x})$  in  $E\{\}$

- Note: in general:

$$E\{g(\mathbf{x})\} \neq g(E\{\mathbf{x}\})$$

- Other rules:

$$E\{a \mathbf{x}\} = a E\{\mathbf{x}\}$$

$$E\{\mathbf{x} + \mathbf{y}\} = E\{\mathbf{x}\} + E\{\mathbf{y}\}$$

$$E\left\{\sum_i a_i \mathbf{x}_i\right\} = \sum_i a_i E\{\mathbf{x}_i\}$$

~~$$E\{\mathbf{x} \mathbf{y}\} = E\{\mathbf{x}\} E\{\mathbf{y}\}$$~~

# Variance and Standard Deviation

---

- Definition:

$$\text{Var}\{x\} = E\{(x - E\{x\})^2\}$$

$$\text{Std}\{x\} = \sqrt{\text{Var}\{x\}}$$

+ the Var gives a measure of dispersion of the data

+ Example I: What is the variance of a normal distribution?

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

+ Example II: What is the variance of a uniform distribution  $x \in [0, r]$  ?

$$\text{Var}\{x\} = \frac{r^2}{12}$$

+ A most important rule (but numerically dangerous):

$$\text{Var}\{x\} = E\{x^2\} - (E\{x\})^2$$

# Sample Variance and Covariance

---

- Sample Variance

$$\text{Var}\{x\} = \frac{1}{N-1} \sum_{i=1}^N (x_i - E\{x\})^2$$

u Why division by (N-1)? This is to obtain an unbiased estimate of the variance.

- Covariance

$$\text{Cov}\{x, y\} = E\{(x - E\{x\})(y - E\{y\})\}$$

- Sample Covariance

$$\text{Cov}\{x, y\} = \frac{1}{N-1} \sum_{i=1}^N (x_i - E\{x\})(y_i - E\{y\})$$

$$\text{Cov}\{\mathbf{x}\} = \frac{1}{N-1} \sum_{i=1}^N (\mathbf{x}_i - E\{\mathbf{x}\})(\mathbf{x}_i - E\{\mathbf{x}\})^T$$

# Moments of a Random Variable

---

- Definition of Moments

$$m_n = E\{x^n\}$$

- Definition of Central Moments

$$cm_n = E\{(x - \mu)^n\}$$

- Useful moments:

- +  $m_1$ =Mean

- +  $cm_2$ =Variance

- +  $cm_3$ =Skewness (measure of asymmetry of a distribution)

- +  $cm_4$ =Kurtosis (detects heavy and light tails and deformations of a distribution; important in computer vision)

# Joint Distributions

---

- Joint distributions are distributions of several random variables, stating the probability that event\_1 AND event\_2 occur simultaneously.

- u Example 1: Generic 2 dimensional joint distribution

$$\int_{-\infty}^{\infty} p(x, y) dx dy = 1$$

- u Example 2: Multivariate normal distribution in vector notation

$$p(\mathbf{x}) = \frac{1}{\sqrt{(2\pi)^d |\Sigma|}} \exp\left(-\frac{1}{2}(\mathbf{x} - \mu)^T \Sigma^{-1}(\mathbf{x} - \mu)\right)$$

- Marginal Distributions: Integrate out some variables (this can be computationally very expensive)

$$p(x) = \int_{-\infty}^{\infty} p(x, y) dy$$

# Probabilistic Independence

---

- Definition:

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

- Knowledge about independence is VERY powerful since it simplifies the evaluation of equations a lot.

- u Example 1: Marginal distribution of independent variables

$$\begin{aligned} p(x) &= \int_{-\infty}^{\infty} p(x, y) dy = \int_{-\infty}^{\infty} p(x)p(y) dy = \\ &= p(x) \int_{-\infty}^{\infty} p(y) dy = p(x) \end{aligned}$$

- u Example 2: The multivariate normal distribution for independent variables

$$\begin{aligned} p(\mathbf{x}) &= \frac{1}{\sqrt{(2\pi)^d |\Sigma|}} \exp\left(-\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu})^T \Sigma^{-1}(\mathbf{x} - \boldsymbol{\mu})\right) = \\ &= \prod_i \frac{1}{\sqrt{2\pi\sigma_i^2}} \exp\left(-\frac{1}{2} \frac{(x_i - \mu_i)^2}{\sigma_i^2}\right) \end{aligned}$$

# Conditional Distributions

---

- Definition:

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

- Since conditional distributions are more “intuitive”, some people believe that joint distributions should be defined through the more atomic conditions distribution

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

- What does independence mean for conditional distributions?

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

# Conditional Distributions (cont'd)

---

- The Chain Rule of Probabilities

$$P(x_1, x_2, \dots, x_n) = P(x_1 | x_2, \dots, x_n)P(x_2 | x_3, \dots, x_n) \\ \dots P(x_{n-1} | x_n)P(x_n)$$

# Bayes Rule

---

- Definition

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

- Verification from:

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

- Interpretation:

- u  $P(y)$  is the PRIOR knowledge about  $y$
- u  $x$  is new evidence to be incorporated to update my belief about  $y$
- u  $P(x|y)$  is the LIKELIHOOD of  $x$  given that  $y$  was observed.
- u Both prior and likelihood can often be generated beforehand, e.g., by histogram statistics.
- u  $P(x)$  is a normalizing factor, corresponding to the marginal distribution of  $x$ . Often it need not be evaluated explicitly. But it can become a great computational burden. “ $P(x)$  is an enumeration of all possible combinations in which  $x$  and  $y$  can occur”.
- u  $P(y|x)$  is the POSTERIOR probability of  $y$ , i.e., the belief in  $y$  after one discovered  $x$ .

# Bayes Rule

## (cont'd)

---

- Example

- + There are 3 doors
- + Behind one is a treasure
- + You get one chance to pick one door (but do not open it yet)
- + The Quizmaster tells you which of the two doors you did not pick does NOT contain the treasure.
- + **THE BIG QUESTION:** Do you stick to your original door or do you switch to the other one?