Introduction to Probability and Statistics (Continued)

Prof. Nicholas Zabaras
Center for Informatics and Computational Science
https://cics.nd.edu/
University of Notre Dame
Notre Dame, Indiana, USA

Email: nzabaras@gmail.com
URL: https://www.zabaras.com/

August 27, 2018

Contents

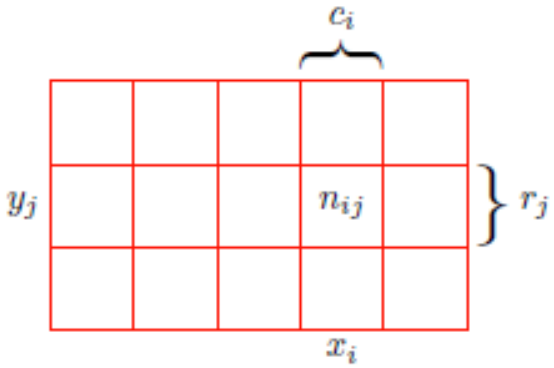
References

- Following closely Chris Bishops' PRML book, Chapter 2
- Kevin Murphy's, Machine Learning: A probablistic perspective, Chapter 2
- Jaynes, E. T. (2003). Probability Theory: The Logic of Science. Cambridge University Press.
- Bertsekas, D. and J. Tsitsiklis (2008). Introduction to Probability. Athena Scientific. 2nd Edition
- Wasserman, L. (2004). All of statistics. A Concise Course in Statistical Inference. Springer.

Joint Probability

- Probability theory provides a consistent framework for the quantification and manipulation of uncertainty.
- It forms one of the central foundations for pattern recognition and machine learning.
- The probability that the event X will take the value x_i and event Y will take the value y_j is written $P(X = x_i, Y = y_j)$ and is called the joint probability of $X = x_i$ and $Y = y_j$.

$$P(X = x_i, Y = y_j) = \frac{n_{ij}}{N}$$



- Here n_{ij} is the number of times (in N trials) that the event $X = x_i, Y = y_j$ occurs. Similarly c_i is the number of times that $X = x_i$ occurs.

The Sum and Product Rules

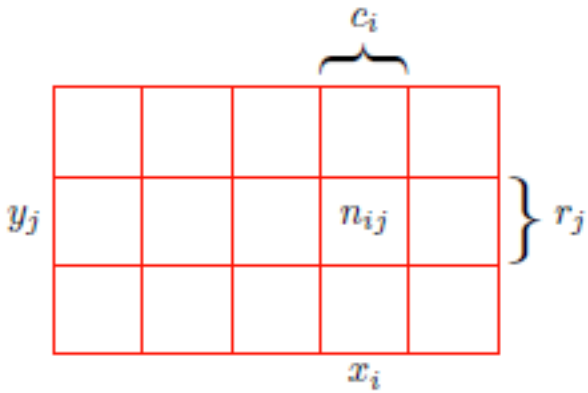
- Even complex calculations in probability are simply derived from the sum and product rules of probability.

- Sum Rule:

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

- Product Rule:

$$P(X = x_i, Y = y_j) = \frac{n_{ij}}{N} = \frac{n_{ij}}{c_i}\frac{c_i}{N}$$
$$= P(Y = y_j \mid X = x_i)P(X = x_i)$$



- The product rule leads to the Chain Rule:

$$P(X_{1:D}) = P(X_1)P(X_2 \mid X_1)P(X_3 \mid X_1, X_2)...P(X_D \mid X_{1:D-1})$$

Conditional Probability and Bayes' Rule

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

- *Bayes'* theorem plays a central role in pattern recognition and machine learning

- The normalizing factor $P(Y)$ is given as:

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

Example: Generative Classifier

$$P(y=c \mid \boldsymbol{x}, \theta) = \frac{P(y=c \mid \theta) P(\boldsymbol{x} \mid y=c, \theta)}{\sum_{x'} P(y=c' \mid \theta) P(\boldsymbol{x} \mid y=c', \theta)}$$

- Here we classify feature vectors $\boldsymbol{x}$ to classes using the above posterior.
- It is a generative classifier as it specifies how to generate data $\boldsymbol{x}$ using *the class-conditional probabilities* $P(\boldsymbol{x} \mid y=c, \theta)$ *and class priors* $P(y=c \mid \theta)$.

  θ are (unknown) parameters defining these distributions.
- *In a discriminative setting,* the posterior $P(y=c \mid \boldsymbol{x})$ is directly fitted.

Independency, Conditional Probability

- Two events A *and* B are independent (written as $A \perp B$) if

$$P(A \cap B) = P(A)P(B)$$

- Conditional probability: Probability that A happens provided that B happens,

$$P(A \mid B) = \frac{P(A \cap B)}{P(B)} \quad \left(Note: P(A \mid B) \geq P(A \cap B) \right)$$

- Using the above Eqs, we see that for independent events,

$$P(A \mid B) = P(A)$$

Independence, Conditional Independence

- X and Y are *unconditionally independent or marginally independent*, denoted $X \perp Y$, if we can represent the joint as the product of the two marginals

$$X \perp Y \Leftrightarrow P(X,Y) = P(X)P(Y)$$

- X and Y are *conditionally independent (CI)* given Z iff the conditional joint can be written as a product of conditional marginals:

$$X \perp Y \mid Z \Leftrightarrow P(X,Y \mid Z) = P(X \mid Z)P(Y \mid Z)$$

Pairwise Vs. Mutual Independence

- Pairwise independence does not imply mutual independence.

  - ✓ Consider 4 balls (numbered $1,2,3,4$) in a box. You draw one at random. Define the following events:

$$X_1 : ball\ 1\ or\ 2\ is\ drawn$$
$$X_2 : ball\ 2\ or\ 3\ is\ drawn$$
$$X_3 : ball\ 1\ or\ 3\ is\ drawn$$

  - ✓ Note that the three events are pairwise independent, e.g.

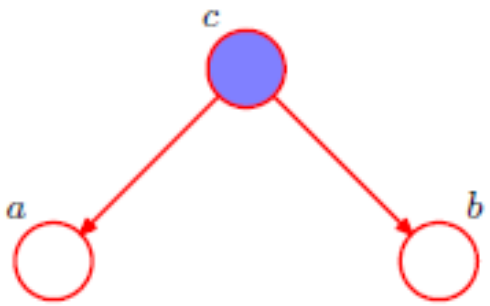
$$X_1 \perp X_2 \Leftrightarrow P(X_1, X_2) = \underbrace{P(X_1)}_{1/2}\underbrace{P(X_2)}_{1/2} = \frac{1}{4}$$

  - ✓ However:

$$P(X_1, X_2, X_3) = 0, \underbrace{P(X_1)}_{1/2}\underbrace{P(X_2)}_{1/2}\underbrace{P(X_3)}_{1/2} = \frac{1}{8}$$

Independence, Conditional Independence

- Consider the following example. Define:



- Event a = `it will rain tomorrow'
- Event b =`the ground is wet today' and
- Event c =`raining today'.
- c causes both a and b – thus given c we don't need to know about b to predict a

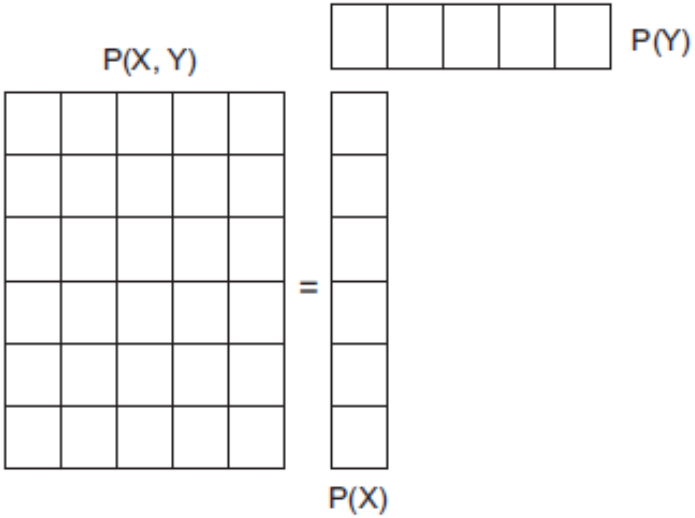
$$a \perp b \mid c \Leftrightarrow P(a,b \mid c) = P(a \mid c)P(b \mid c)$$

$$a \perp b \mid c \Leftrightarrow P(a \mid b,c) = P(a \mid c)$$

- Observing a "root node", separates "the children"!

Independence, Conditional Independence

- Assume unconditional independence $X \perp Y$. Let us assume that X takes 6 values and Y takes 5 values. The cost for defining $p(X, Y)$ is drastically reduced if $X \perp Y$.



- The parameters are reduced from $29\ (= 30 - 1)$ to $9 = 5 + 4 = (6 - 1) + (5 - 1)$.[a] *Independence is key to efficient probabilistic modeling* (naïve Bayes classifiers, Markov Models, graphical models, etc.).

[a] We subtract one on the counts to account for the sum-to-one probability constraint rule.

Conditional Independence Theorem

- $X \perp Y \mid Z$ iff there exist functions g and h such that the following holds

$$X \perp Y \mid Z \Leftrightarrow P(x, y \mid z) = g(x, z)h(y, z) \; \forall x, y, z \; such \; that \; P(z) > 0$$

- Independence implies the factorization:

$$X \perp Y \mid Z \Leftrightarrow P(x, y \mid z) = \underbrace{P(x \mid z)}_{g(x,z)} \underbrace{P(y \mid z)}_{g(y,z)} = g(x, z)h(y, z)$$

- Factorization implies independence:

$$P(x, y \mid z) = g(x, z)h(y, z) \Rightarrow 1 = \sum_{x,y} P(x, y \mid z) = \sum_{x,y} g(x, z)h(y, z) = \sum_{x} g(x, z) \sum_{y} h(y, z) \; (1)$$

$$P(x \mid z) = \sum_{y} g(x, z)h(y, z) = g(x, z) \sum_{y} h(y, z) \; (2)$$

$$P(y \mid z) = \sum_{x} g(x, z)h(y, z) = h(y, z) \sum_{x} g(x, z) \; (3)$$

$$(2,3) \Rightarrow P(x \mid z)P(y \mid z) = g(x, z)h(y, z) \sum_{x} g(x, z) \sum_{y} h(y, z) \overset{(1)}{=} g(x, z)h(y, z) = P(x, y \mid z)$$

Independence Relations: Decomposition

- The following decomposition independence relation holds:

$$X \perp Y,W \mid Z \Rightarrow X \perp Y \mid Z$$

- Proof:

$$X \perp Y,W \mid Z \Rightarrow P(X,Y,W \mid Z) = P(X \mid Z)P(Y,W \mid Z) \Rightarrow \quad \text{(independence)}$$

$$P(X,Y \mid Z) = \sum_{w} P(X \mid Z)P(Y,w \mid Z) \Rightarrow \quad \text{(sum rule)}$$

$$P(X,Y \mid Z) = P(X \mid Z)\sum_{w} P(Y,w \mid Z) = P(X \mid Z)P(Y \mid Z) \Rightarrow X \perp Y \mid Z$$

Independence Relations: Weak Union

- The following Weak Union independence relation holds:

$$X \perp Y,W \mid Z \Rightarrow X \perp Y \mid Z,W$$

- Proof:

$$X \perp Y,W \mid Z \Rightarrow P(X,Y,W \mid Z) = P(X \mid Z)P(Y,W \mid Z) \Rightarrow$$

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

$$P(X,Y \mid Z,W) = P(X \mid Z,W)P(Y \mid Z,W) \Rightarrow X \perp Y \mid Z,W$$

- Note that in the last step, we used the decomposition independence relation:

$$X \perp Y,W \mid Z \Rightarrow X \perp W \mid Z \Rightarrow P(X,W \mid Z) = P(X \mid Z)P(W \mid Z) \Rightarrow$$

$$P(X \mid Z) = \frac{P(X,W \mid Z)}{P(W \mid Z)} = P(X \mid Z,W) \Rightarrow P(X \mid Z) = P(X \mid Z,W)$$

Independence Relations: Contraction

- The following Contraction independence relation holds:

$$(X \perp W \mid Z,Y)\&(X \perp Y \mid Z) \Rightarrow X \perp Y,W \mid Z$$

- Proof:

$$\left.\begin{array}{ll} X \perp W \mid Y,Z \Rightarrow P(X,W \mid Y,Z) = P(X \mid Y,Z)P(W \mid Y,Z) & (a) \\ X \perp Y \mid Z \Rightarrow P(X \mid Y,Z) = P(X \mid Z) & (b) \end{array}\right\} \Rightarrow$$

$$P(X,Y,W \mid Z) = P(X,W \mid Y,Z)P(Y \mid Z) =$$

$$\overset{(a)}{=} P(X \mid Y,Z)P(W \mid Y,Z)P(Y \mid Z)$$

$$\overset{(b)}{=} P(X \mid Z)P(W,Y \mid Z)$$

$$\Rightarrow X \perp W,Y \mid Z$$

Random Variables

- Define Ω to be a probability space equipped with a probability measure P that measures the probability of events $E \subset \Omega$. Ω *contains* all possible events in the form of its own subsets.
- A real valued random variable X *is a mapping* $X : \Omega \to \mathbb{R}$
- We call $x = X(\omega), \omega \in \Omega,$ a realization of X.
- Probability distribution of X: For $B \subset \mathbb{R}$,

$$P_X(B) = P\left(X^{-1}(B)\right) = P\left\{X(\omega) \in B\right\}$$

- Probability density

$$P_X(B) = \int_B p_X(x)dx$$

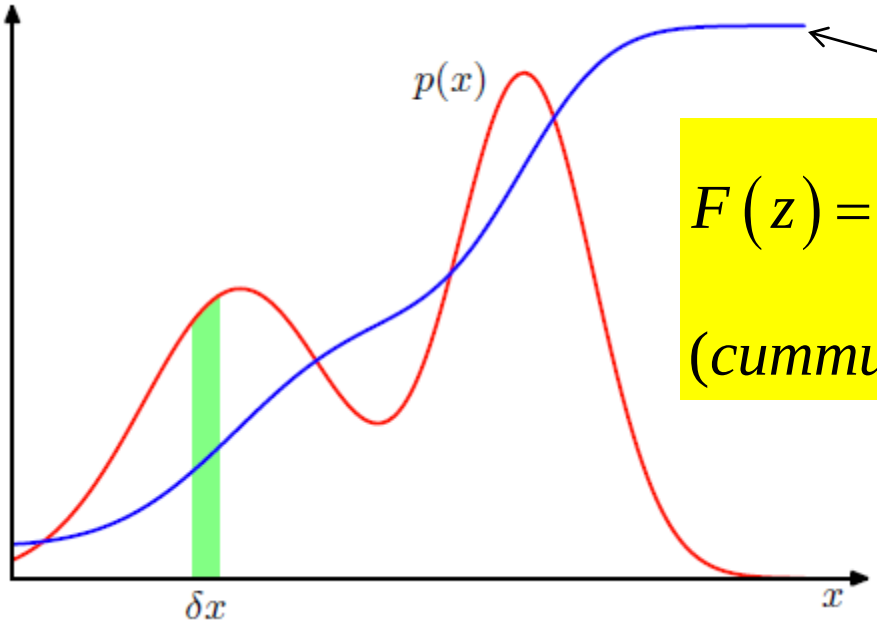
- We often write:

$$p_X(x) = p(x)$$

Cummulative Density Function

$$p(x) \geq 0$$

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



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

(*cummulative density function*)

$$P(x \in (a,b)) = \int_{a}^{b} p(x)dx = F(b) - F(a)$$

- The CDF for a random variable X is the function $F(x)$ that returns the probability that X is less than x

Mean and Variance

- The expected (mean) value of X is:

$$\mathbb{E}[X] = \int_{-\infty}^{+\infty} x p_X(x) dx$$

- The variance of X is defined as:

$$\text{var}[X] = \mathbb{E}\left[\left(X - \mathbb{E}[X] \right)^2 \right] = \mathbb{E}\left[X^2 \right] - \mathbb{E}[X]^2$$

- The standard deviation is often useful defined as

$$std[X] = \sqrt{\text{var}[X]}$$

Mean and Variance

- If $T(X)$ is a function of X (called a statistic), the mean of $T(X)$ is given by

$$\mathbb{E}\left[T(X)\right] = \int_{-\infty}^{+\infty} T(x)\, p_X(x) dx$$

- The variance of $T(X)$ is also defined as:

$$\text{var}\left[T(X)\right] = \mathbb{E}\left[\left(T(X) - \mathbb{E}(T(X))\right)^2\right] = \mathbb{E}\left[T(X)^2\right] - \left(\mathbb{E}\left[T(X)\right]\right)^2$$

- The k^{th} moment is defined as: ($k = 3$, skewness, $k = 4$, kurtosis)

$$\mathbb{E}\left[\left(X - \mathbb{E}[X]\right)^k\right] = \int_{-\infty}^{+\infty} (x - \mathbb{E}[X])^k\, p_X(x) dx$$

Expectation

- For discrete case

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

- For continuous case

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

- Conditional expectation

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

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

Expectation

- *The expectation of a random variable is not necessarily the value that we should expect a realization to have.* Let $\Omega = [-1,1]$ with the probability P being uniform:

$$P(I) = \frac{1}{2}\int_I dx = \frac{1}{2}|I|, \; I \subset [-1,1]$$

- Consider the following two random variables:

$$X_1 : [-1,1] \to \mathbb{R}, \quad X_1(\omega) = 1 \; \forall \omega$$

$$X_2 : [-1,1] \to \mathbb{R}, \quad X_2(\omega) = \begin{cases} 2 & \omega \geq 0 \\ 0 & \omega < 0 \end{cases}$$

- We can see that

$$\mathbb{E}[X_1] = \int_{-1}^{1} X_1(\omega)\frac{1}{2}dx = \int_{-1}^{1} 1\frac{1}{2}dx = 1,$$

$$\mathbb{E}[X_2] = \int_{0}^{1} X_2(\omega)\frac{1}{2}dx = \int_{0}^{1} 2\frac{1}{2}dx = 1$$

although $X_2(\omega) \neq 1 \; \forall \omega \in \mathbb{R}.$

Uniform Random Variable

- Consider the Uniform distribution

$$\mathcal{U}(x \mid a,b) = \frac{1}{b-a}\mathbb{I}(a \le x \le b)$$

- What is the CDF of a uniform random variable $\mathcal{U}(0,1)$?

$$P_U(x) = \begin{cases} x \quad for\ 0 \le x \le 1, \\ 1, \ for\ x \ge 1 \\ 0, \ otherwise \end{cases}$$

- You can show easily that the mean and variance of $\mathcal{U}(x \mid a,b)$ are:

$$\mathbb{E}\left[x \mid a,b\right] = \frac{a+b}{2}, \ \mathbb{E}\left[x^2 \mid a,b\right] = \frac{a^2+ab+b^2}{3}, \ \mathrm{var}\left[x \mid a,b\right] = \frac{(b-a)^2}{12}$$

- Note that it is possible for $p(x) > 1$ but the density still needs to integrate to 1. For example, note that

$$\mathcal{U}(x \mid 0,1/2) = 2 > 1 \ \forall x \in \left[0, \frac{1}{2}\right]$$

The Gaussian Distribution

- A random variable $X \in \mathbb{R}$ is Gaussian or normally distributed $X \sim \mathcal{N}(\mu, \sigma^2)$ if:

$$P\{X \leq t\} = \frac{1}{\sqrt{2\pi\sigma^2}} \int_{-\infty}^{t} \exp\left(-\frac{1}{2\sigma^2}(x-\mu)^2\right) dx$$

- The probability density is:

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

- To show that this density is normalized note the following trick:

$$Let\ I \equiv \int_{-\infty}^{+\infty} \exp\left(-\frac{1}{2\sigma^2}(x-\mu)^2\right) dx,\ then: I^2 = \int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} \exp\left(-\frac{1}{2\sigma^2}(y-\mu)^2\right) \exp\left(-\frac{1}{2\sigma^2}(x-\mu)^2\right) dxdy$$

$$Set\ r^2 = (x-\mu)^2 + (y-\mu)^2,\ then: I^2 = \int_{0}^{+\infty}\int_{0}^{2\pi} \exp\left(-\frac{1}{2\sigma^2}r^2\right) r dr d\theta = 2\pi \int_{0}^{\infty} \exp\left(-\frac{1}{2\sigma^2}r^2\right) r dr$$

$$Thus: I^2 = \pi \int_{0}^{\infty} \exp\left(-\frac{1}{2\sigma^2}u\right) du = -\left(2\pi\sigma^2\right)\left(e^{-\infty} - 1\right) = 2\pi\sigma^2$$

The Gaussian Distribution

- A random variable $X \in \mathbb{R}$ is Gaussian or normally distributed $X \sim \mathcal{N}(\mu, \sigma^2)$ if:

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

- The following can be shown easily with direct integration:

$$\mathbb{E}[X] = \frac{1}{\sqrt{2\pi\sigma^2}} \int_{-\infty}^{+\infty} \exp\left(-\frac{1}{2\sigma^2}(x-\mu)^2\right) x dx = \mu,$$

$$\mathbb{E}\left[X^2\right] = \frac{1}{\sqrt{2\pi\sigma^2}} \int_{-\infty}^{+\infty} \exp\left(-\frac{1}{2\sigma^2}(x-\mu)^2\right) x^2 dx = \mu^2 + \sigma^2, \quad \text{var}[X] = \mathbb{E}\left[\left(X-\mu\right)^2\right] = \sigma^2$$

- The following integrals are useful in these derivations :

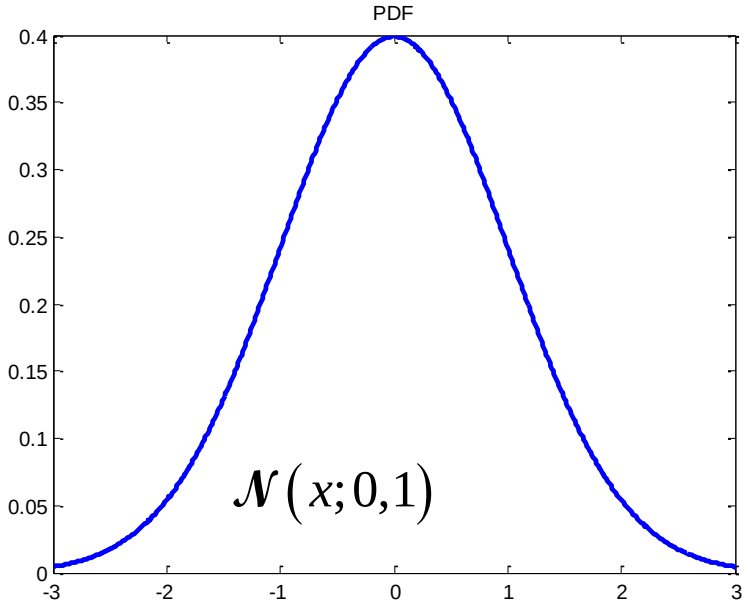
$$\int_{-\infty}^{+\infty} \exp\left(-u^2\right) du = \sqrt{\pi}, \quad \int_{-\infty}^{+\infty} u \exp\left(-u^2\right) du = 0, \quad \int_{-\infty}^{+\infty} u^2 \exp\left(-u^2\right) du = \frac{\sqrt{\pi}}{2}$$

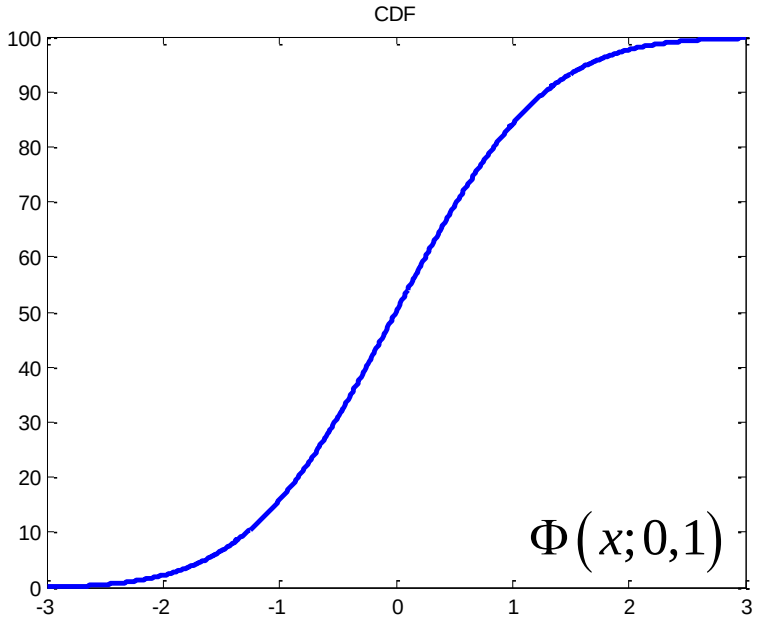
- We often work with *the precision of a Gaussian $\lambda = 1/\sigma^2$. The higher the λ the narrower the distribution is*.

CDF of a Gaussian

- Plot of the Standard Normal $\mathcal{N}(0,1)$ and CDF.

Run MatLab function *gaussPlotDemo* from Kevin Murphys' PMTK





$$F\left(x;\mu,\sigma^2\right)=\int_{-\infty}^{x}\mathcal{N}(z\mid\mu,\sigma^2)dz$$

$$F\left(x;\mu,\sigma^2\right)=\frac{1}{2}\left[1+erf\left(z/\sqrt{2}\right)\right],z=(x-\mu)/\sigma$$

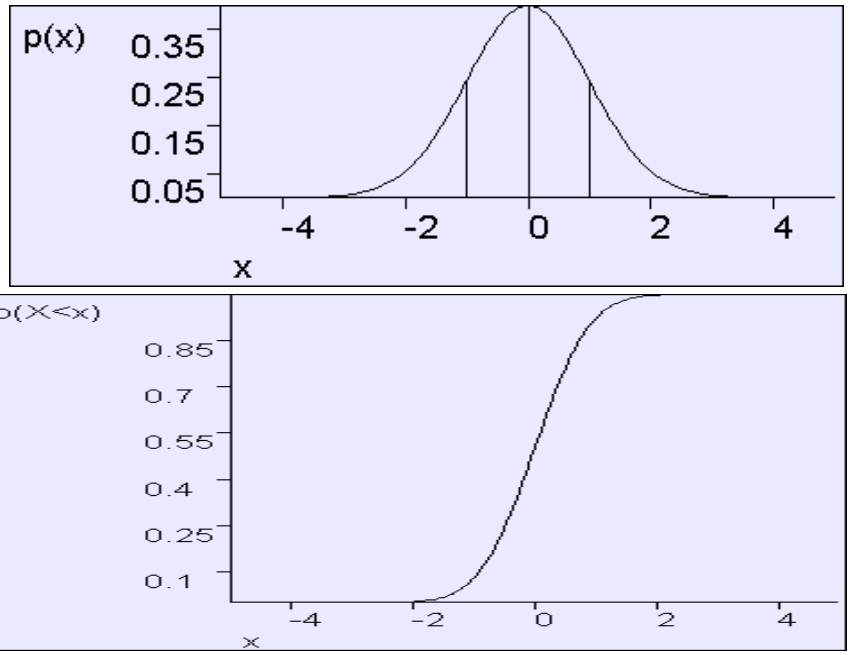
$$erf\left(x\right)=\frac{2}{\sqrt{\pi}}\int_{0}^{x}e^{-t^2}dt$$

CDF and the Standard Normal

- Assume $X \sim \mathcal{N}(0,1)$
- Define $\Phi(x) = P(X < x)$ = Cumulative Distribution of X

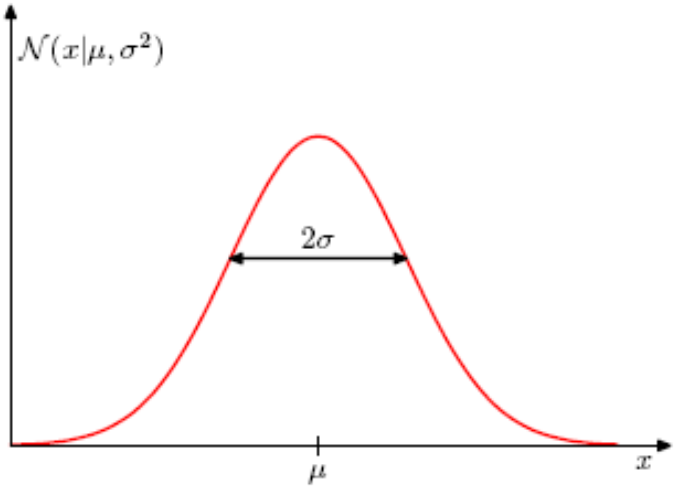
$$\Phi(x) = \int_{z=-\infty}^{x} p(z)dz$$

$$= \frac{1}{\sqrt{2\pi}} \int_{z=-\infty}^{x} \exp\left(-\frac{z^2}{2}\right)dz$$



- Assume $X \sim \mathcal{N}(\mu, \sigma^2)$
- $F(x) = P(X < x|\mu, \sigma^2) = \Phi\left(\frac{x-\mu}{\sigma}\right)$

Univariate Gaussian



$$\mathbb{E}[X] = \mu$$

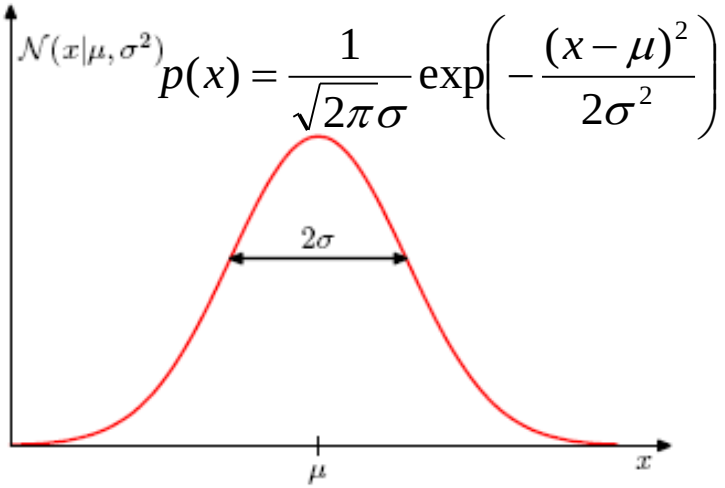
$$var[X] = \sigma^2$$

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

- Shorthand: We say $X \sim \mathcal{N}(\mu, \sigma^2)$ to mean "X is distributed as a Gaussian with parameters μ and σ^2".

Univariate Gaussian

- Representation of symmetric phenomena without long tails.
- ***Inappropriate for skewness, fat tails, multimodality,*** etc.
- The popularity of the Gaussian is related to facts as:



  - Completely defined in terms of the mean and the variance
  - The Central Limit Theorem shows that the sum of i.i.d. random variables has approximately a Gaussian distribution making it an appropriate choice for modeling noise (limit of additive small effects)
  - The Gaussian distribution makes the least assumptions (maximum entropy) from all distributions with given mean and variance.

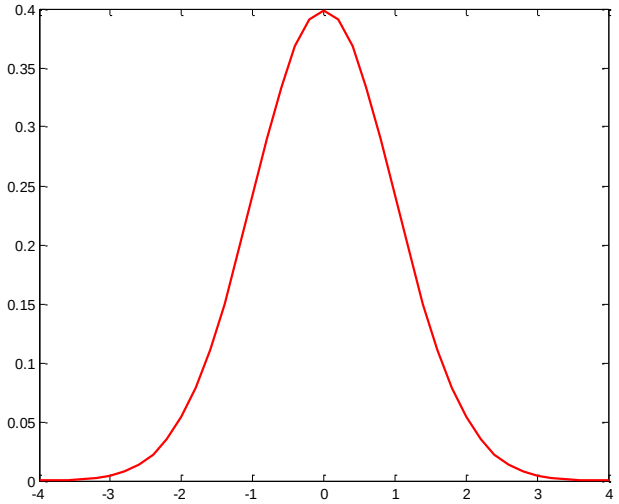
The Normal Model

- The normal (or Gaussian) distribution $\mathcal{N}(\mu, \sigma^2)$

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

is one of the most studied and one of the most used distributions

- Sample $x_1, \ldots, x_n$ from a normal distribution $\mathcal{N}(\mu, \sigma^2)$.
  In a typical inference problem, we are interested in:



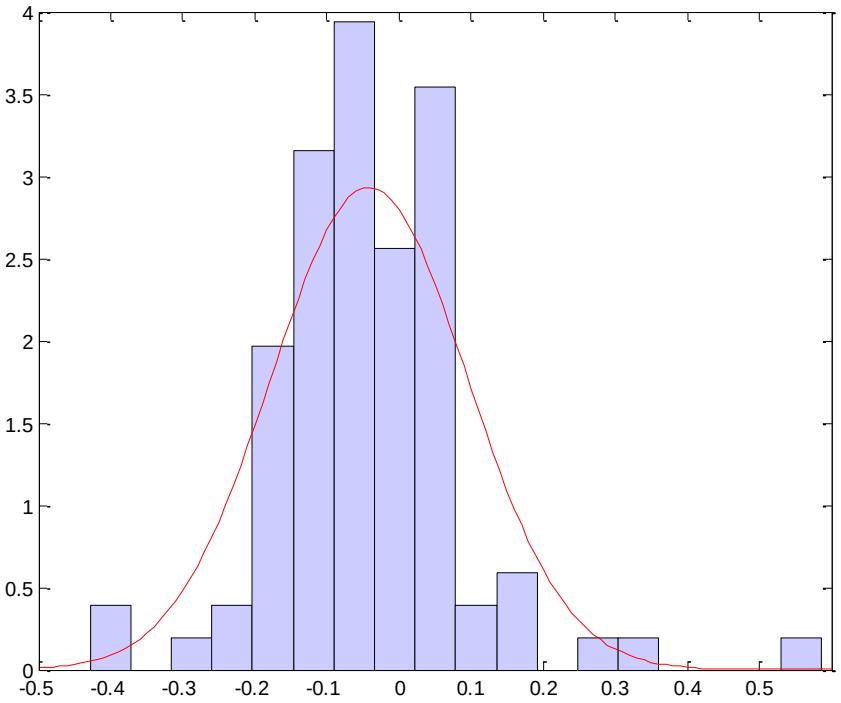
  - Estimation of (μ, σ)
  - Confidence regions on (μ, σ)
  - Test on (μ, σ) and comparison with other samples

Datasets: Normal Data

- Normaldata : Relative changes in reported larcenies between 1991 and 1995 (relative to 1991) for the 90 most populous US counties (Source: FBI)[1]

Normal data

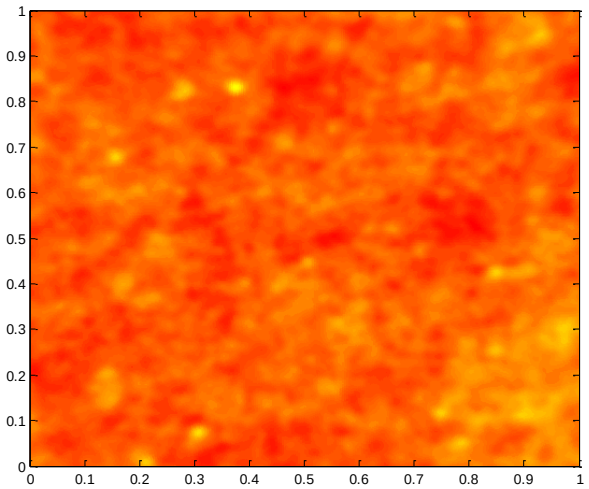
Matlab implementation



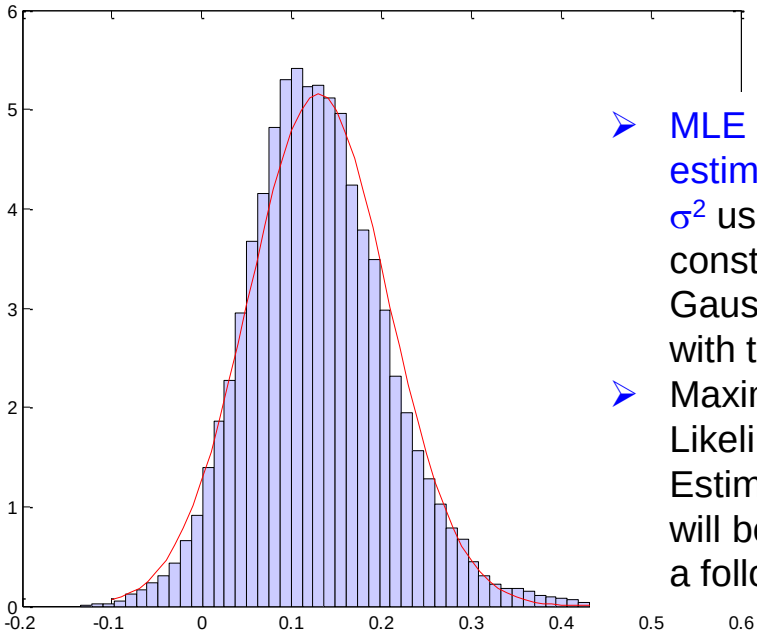
From Bayesian Core, J.M. Marin and C.P. Roberts, Chapter 2 (available online)

Datasets: CMBData

- CMBdata: Spectral representation of the cosmological microwave background (CMB), i.e. electromagnetic radiation from photons back to 300,000 years after the Big Bang, expressed as difference in apparent temperature from the mean temperature.[1]



CMBdata



- MLE based estimates of μ and σ^2 used in constructing the Gaussian shown with the solid line
- Maximum Likelihood Estimators (MLE) will be discussed in a follow up lecture

Matlab implementation

Normal estimation

From Bayesian Core, J.M. Marin and C.P. Roberts, Chapter 2 (available online)

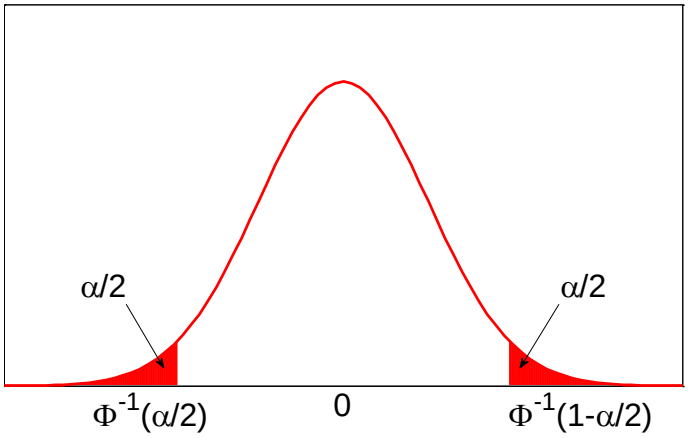
Quantiles

- Recall that the probability density function $p_X(\cdot)$ of a random variable X is defined as the derivative of the cumulative density function, so that

$$F(x_0) = \int_{-\infty}^{x_0} p_X(x)dx$$

- The value $y_{(\alpha)}$ such that $F(y_{(\alpha)}) = \alpha$ is called the α-quantile of the distribution with CDF F. The median is of course $F(y_{(0.5)}) = 0.5$

- Note that $$P_X(\infty) = \int_{-\infty}^{\infty} p_X(x)dx = 1$$



- One can define ***tail area probabilities***. The shaded regions each contain $\alpha/2$ of the probability mass. For $\mathcal{N}(0, 1)$, the leftmost cutoff point is $\Phi^{-1}(\alpha/2)$, where Φ is the cdf of $\mathcal{N}(0, 1)$.
- If $\alpha = 0.05$, ***the central interval*** is 95%, and the left cutoff is -1.96 and the right is 1.96. Figure generated by quantileDemo from PMTK
- For $\mathcal{N}(\mu,\sigma^2)$, the 95% interval becomes $(\mu-1.96\sigma, \mu+1.96\sigma)$ – often approximated as $(\mu\pm2\sigma)$

Binary Variables

- Consider a coin flipping experiment with heads $= 1$ and tails $= 0$. With $\mu \in [0,1]$

$$p(x=1 \mid \mu) = \mu$$
$$p(x=0 \mid \mu) = 1-\mu$$

- This defines the Bernoulli distribution as follows:

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

- Using the indicator function, we can also write this as:

$$\mathcal{Bern}(x \mid \mu) = \mu^{\mathbb{I}(x=1)}(1-\mu)^{\mathbb{I}(x=0)}$$

Bernoulli Distribution

- Recall that in general

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

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

- For the Bernoulli distribution $\mathcal{Bern}(x \mid \mu) = \mu^x (1-\mu)^{1-x}$, we can easily show from the definitions:

$$\mathbb{E}[x] = \mu$$

$$\text{var}[x] = \mu(1-\mu)$$

$$\mathbb{H}[x] \equiv -\sum_{x \in \{0,1\}} p(x \mid \mu) \ln p(x \mid \mu) = -\mu \ln \mu - (1-\mu) \ln(1-\mu)$$

- Here $\mathbb{H}[x]$ is the "entropy of the distribution"

Likelihood Function for Bernoulli Distribution

- Consider the data set

$$\mathcal{D} = \{x_1, x_2, ..., x_N\}$$

in which we have m heads ($x = 1$), and $N - m$ tails ($x = 0$)

- The likelihood function takes the form:

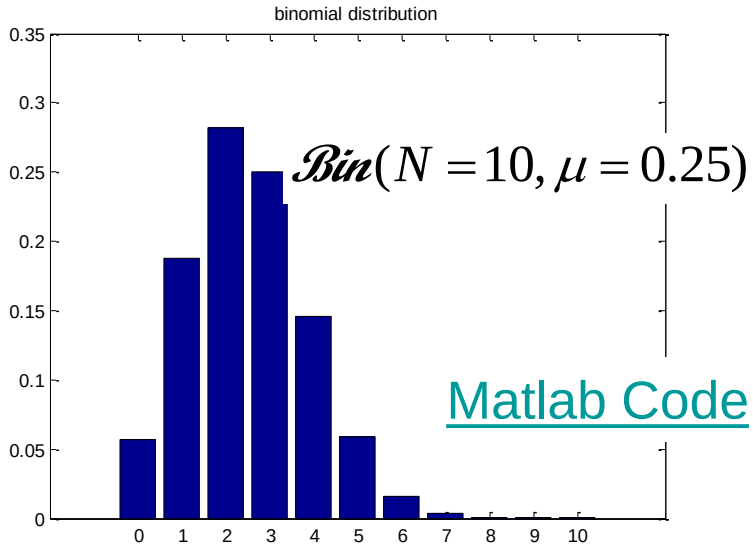
$$p(\mathcal{D} \mid \mu) = \prod_{n=1}^{N} p(x_n \mid \mu) = \prod_{n=1}^{N} \mu^{x_n} (1-\mu)^{1-x_n} = \mu^m (1-\mu)^{N-m}$$

Binomial Distribution

- Consider the discrete random variable $X \in \{0,1,2,...,N\}$

- We define the Binomial distribution as follows:

$$\mathcal{Bin}(X = m \mid N, \mu) = \binom{N}{m} \mu^m (1-\mu)^{N-m}$$



Matlab Code

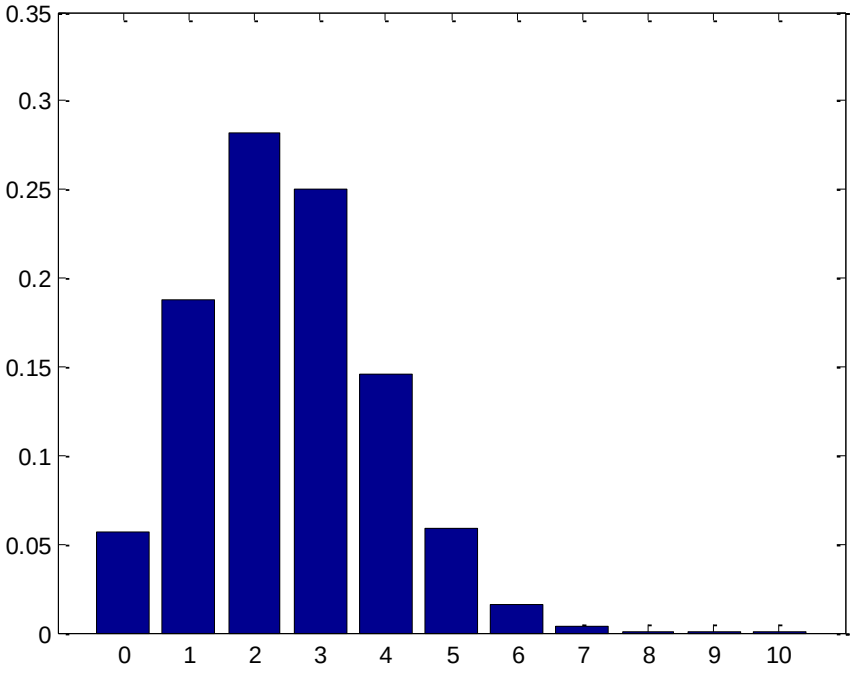
- In our coin flipping experiment, it gives the probability in N flips to get m heads with μ being the probability getting heads in one flip.

- It can be shown (see S. Ross, Introduction to Probability Models) that the limit of the binomial distribution as $N \to \infty, N\mu \to \lambda$, is the $\mathcal{Poisson}(\lambda)$ distribution.

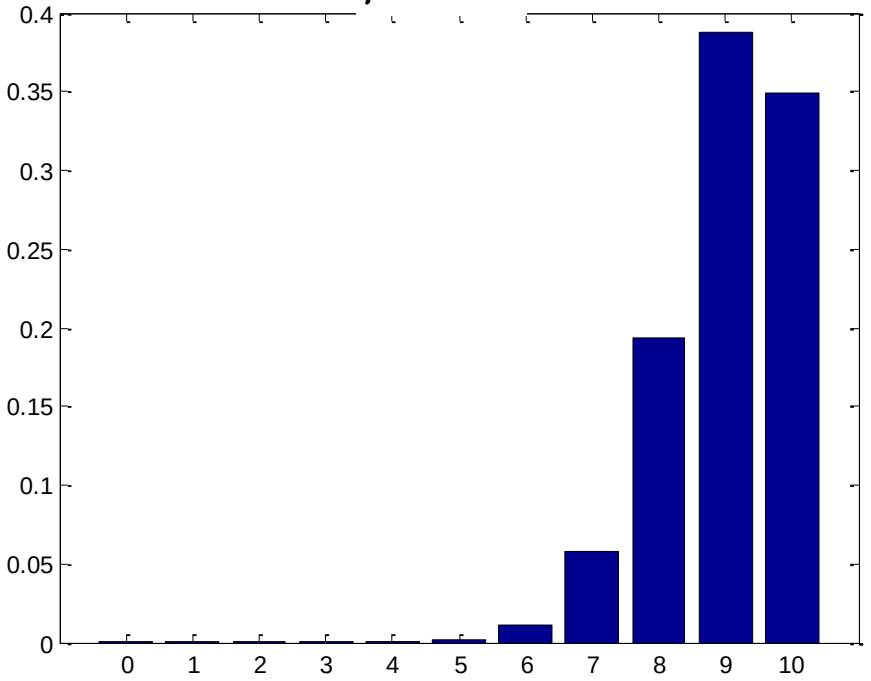
Binomial Distribution

- The Binomial distribution for $N = 10$, and $\mu \in \{0.25, 0.9\}$ is shown below using MatLab function *binomDistPlot* from Kevin Murphys' PMTK

$\mu = 0.25$



$\mu = 0.9$



$\mathcal{Bin}(N, \mu)$

Mean, Variance of the Binomial Distribution

- Consider for independent events the mean of the sum is the sum of the means, and the variance of the sum is the sum of the variances.

- Because $m = x_1 + \ldots + x_N$, and for each observation the mean and variance are known from the Bernoulli distribution:

$$\mathbb{E}[m] \equiv \sum_{m=0}^{N} m\,\mathcal{Bin}(m \mid N, \mu) = \mathbb{E}[x_1 + \ldots + x_N] = N\mu$$

$$var[m] \equiv \sum_{m=0}^{N} \left(m - \mathbb{E}[m]\right)^2 \mathcal{Bin}(m \mid N, \mu) = var[x_1 + \ldots + x_N] = N\mu(1-\mu)$$

- One can also compute $\mathbb{E}[m], \mathbb{E}[m^2]$ by differentiating (twice) the identity $\sum_{m=1}^{N} \binom{N}{m} \mu^m (1-\mu)^{N-m} = 1$ wrt μ. Try it!

Binomial Distribution: Normalization

To show that the Binomial is correctly normalized, we use the following identities:

- Can be shown with direct substitution: $\binom{N}{n}+\binom{N}{n-1}=\binom{N+1}{n}$ (*)
- Binomial theorem: $(1+x)^N=\sum_{m=0}^{N}\binom{N}{m}x^m$ (**)

This theorem is proved by induction using (*) and noticing:

$$(1+x)^{N+1}=\sum_{m=0}^{N}\binom{N}{m}x^m(1+x)=\sum_{m=0}^{N}\binom{N}{m}x^m+\sum_{m=0}^{N}\binom{N}{m}x^{m+1}=\sum_{m=0}^{N}\binom{N}{m}x^m+\sum_{m=1}^{N+1}\binom{N}{m-1}x^m=$$

$$\left(1+\sum_{m=1}^{N}\binom{N}{m}x^m\right)+\left(\sum_{m=1}^{N}\binom{N}{m-1}x^m+x^{N+1}\right)\overset{*}{=}1+\sum_{m=1}^{N}\binom{N+1}{m}x^m+x^{N+1}=\sum_{m=0}^{N+1}\binom{N+1}{m}x^m$$

- To finally show normalization using (**):

$$\sum_{m=0}^{N}\binom{N}{m}\mu^m(1-\mu)^{N-m}=(1-\mu)^N\sum_{m=0}^{N}\binom{N}{m}\left(\frac{\mu}{1-\mu}\right)^m=(1-\mu)^N\left(1+\frac{\mu}{1-\mu}\right)^N=1$$

Generalization of the Bernoulli Distribution

- We are now looking at discrete variables that can take on one of K possible mutually exclusive states.

- The variable is represented by a K-dimensional vector $\boldsymbol{x}$ in which one of the elements x_k equals 1, and all remaining elements equal *0*: $\boldsymbol{x} = (0,0,\ldots,1,0,\ldots,0)^T$

  These vectors satisfy: $\sum_{k=1}^{K} x_k = 1$

- Let the probability of $x_k = 1$ be denoted as μ_k. Then

$$p(\boldsymbol{x} \mid \boldsymbol{\mu}) = \prod_{k=1}^{K} \mu_k^{x_k} = \prod_{k=1}^{K} \mu_k^{\mathbb{I}(x_k=1)}, \quad \sum_{k=1}^{K} \mu_k = 1, \mu_k \geq 0$$

where $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_K)^T$.

Multinoulli/Categorical Distribution

- The distribution is already normalized:

$$\sum_{\boldsymbol{x}} p(\boldsymbol{x} \mid \boldsymbol{\mu}) = \sum_{\boldsymbol{x}} \prod_{k=1}^{K} \mu_k^{x_k} = \sum_{k=1}^{K} \mu_k = 1$$

- The mean of the distribution is computed as:

$$\mathbb{E}[\boldsymbol{x} \mid \boldsymbol{\mu}] = \sum_{\boldsymbol{x}} \boldsymbol{x} p(\boldsymbol{x} \mid \boldsymbol{\mu}) = \left(\mu_1, \ldots, \mu_K\right)^T = \boldsymbol{\mu}$$

similar to the result for the Bernoulli distribution.

- The Multinoulli also known as the *Categorical distribution* often denoted as ($\mathcal{Mu}$ here is the multinomial distribution):

$$\mathcal{Cat}(\boldsymbol{x} \mid \boldsymbol{\mu}) = \mathcal{Multinoulli}(\boldsymbol{x} \mid \boldsymbol{\mu}) = \mathcal{Mu}(\boldsymbol{x} \mid 1, \boldsymbol{\mu})$$

- The parameter 1 stands to emphasize that we roll a K-sided dice once ($N = 1$) – see next for the multinomial distribution.

Likelihood: Multinoulli Distribution

- Let us consider a data set $\mathcal{D} = (\boldsymbol{x}_1, ..., \boldsymbol{x}_N)$. The likelihood becomes:

$$p(\mathcal{D} \mid \boldsymbol{\mu}) = \prod_{n=1}^{N} \prod_{k=1}^{K} \mu_k^{x_{nk}} = \prod_{k=1}^{K} \mu_k^{\sum_{n=1}^{N} x_{nk}} = \prod_{k=1}^{K} \mu_k^{m_k}, \quad where: m_k = \sum_{n=1}^{N} x_{nk}$$

  is the # of observations of $x_k = 1$.

- m_k is the "***sufficient statistic" of the distribution***.

MLE Estimate: Multinoulli Distribution

- To compute the maximum likelihood (MLE) estimate of $\boldsymbol{\mu}$, we maximize an augmented log-likelihood

$$\ln p(\mathcal{D} \mid \boldsymbol{\mu}) + \lambda \left(\sum_{k=1}^{K} \mu_k - 1 \right) = \sum_{k=1}^{K} m_k \ln \mu_k + \lambda \left(\sum_{k=1}^{K} \mu_k - 1 \right)$$

- Setting the derivative wrt μ_k equal to zero: $\mu_K = -\dfrac{m_k}{\lambda}$

- Substitution into the constraint

$$\sum_{k=1}^{K} \mu_k = 1 \Rightarrow -\frac{\sum_{k=1}^{K} m_k}{\lambda} = 1 \Rightarrow \lambda = -\sum_{k=1}^{K} m_k \Rightarrow$$

$$\mu_K = \frac{m_k}{\sum_{k=1}^{K} m_k} = \frac{m_k}{N}$$

As expected, this is the fraction in the N observations of $x_k = 1$

Multinomial Distribution

- We can also consider *the joint distribution of $m_1, \ldots, m_K$ in N observations* conditioned on the parameters $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_K)$.

- From the expression for the likelihood given earlier

$$p(\mathcal{D} \mid \boldsymbol{\mu}) = \prod_{k=1}^{K} \mu_k^{m_k}$$

the multinomial distribution $\mathcal{Mu}(m_1, \ldots, m_K \mid N, \boldsymbol{\mu})$ with parameters N and $\boldsymbol{\mu}$ takes the form:

$$p(m_1, m_2, \ldots, m_K \mid N, \mu_1, \mu_2, \ldots, \mu_K) = \frac{N!}{m_1! m_2! \ldots m_k!} \mu_1^{m_1} \mu_2^{m_2} \ldots \mu_K^{m_k} \quad \text{where} \sum_{k=1}^{K} m_k = N$$