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.

Binary Variables

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

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

- This defines the Bernoulli distribution as follows:

$$\mathcal{B}ern(x|\mu)=\mu^{x}(1-\mu)^{1-x}$$

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

$$\mathcal{B}ern(x|\mu)=\mu^{\mathbb{I}(x=1)}(1-\mu)^{\mathbb{I}(x=0)}$$

Bernoulli Distribution

- Recall that in general

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

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

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

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

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

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

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

Likelihood Function for Bernoulli Distribution

- Consider the data set

$$\mathscr{D} = \left\{ x_1, x_2, \ldots, x_N \right\}$$

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

- The likelihood function takes the form:

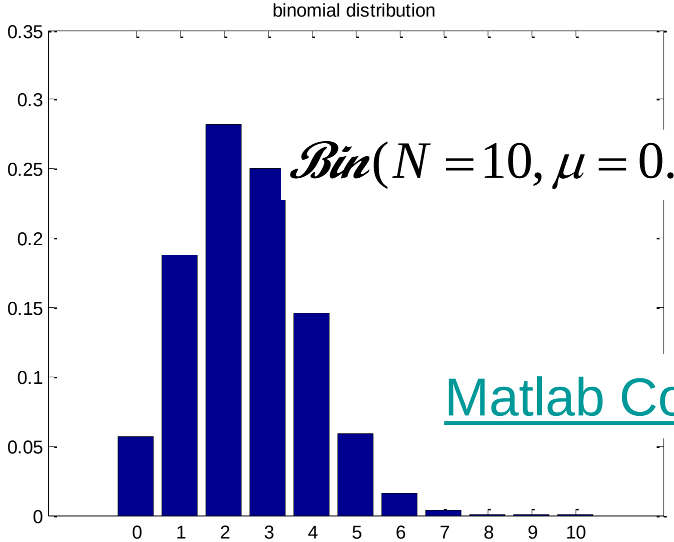
$$p(\mathscr{D} \mid \mu) = \prod_{n=1}^{N} p(x_n \mid \mu) = \prod_{n=1}^{N} \mu^{x_n} (1-\mu)^{1-x_n} = \mu^m (1-\mu)^{N-m}$$

Binomial Distribution

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

- 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

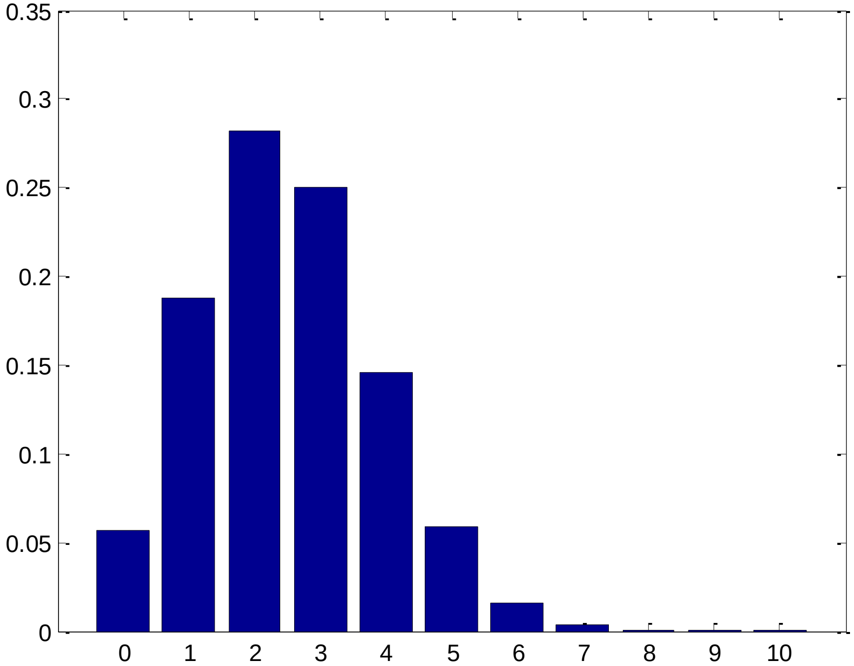
- In our coin flipping experiment, it gives the probability in N flips to get m heads with μ being the probability getting heads in one flip.

- 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.

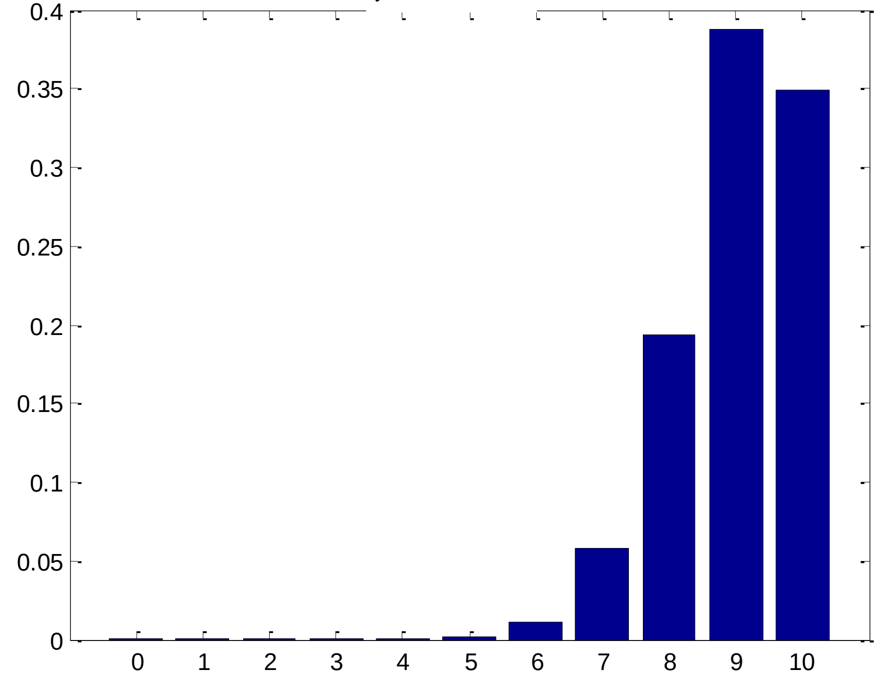
Binomial Distribution

- 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:

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

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

- 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*: $\boldsymbol{x} = (0,0, \dots, 1,0, \dots, 0)^T$

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

Multinoulli/Categorical Distribution

- The distribution is already normalized:

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

- The mean of the distribution is computed as:

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

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):

$$\mathcal{Cat}\left(\mathbf{x}|\boldsymbol{\mu}\right) = \mathcal{Multinoulli}\left(\mathbf{x}|\boldsymbol{\mu}\right) = \mathcal{Mu}\left(\mathbf{x}|1, \boldsymbol{\mu}\right)$$

- 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 $\mathcal{D}=\left(\boldsymbol{x}_1,\ldots,\boldsymbol{x}_N\right)$. The likelihood becomes:

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

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

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

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

- Substitution into the constraint

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

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

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

Multinomial Distribution

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

- 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 $\mathcal{M}u(m_1, \dots, m_K \mid N, \boldsymbol{\mu})$ with parameters N and $\boldsymbol{\mu}$ takes the form:

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

Example: Biosequence Analysis

- Consider a set of DNA sequences where there are 10 rows (sequences) and 15 columns (locations along the genome).

```
c g a t  a c g  g g g t c g a a
c a a t  c c g  a g a t c g c a
c a a t  c c g   t g t t g g g a
c a a t  c g g  c a t g c g g g
c g a g c c g  c g t a c g a a
c a t  a c g g  a g c a c g a a
t  a a t  c c g  g g c a t g t a
c g a g c c g  a g t a c a g a
c c a t  c c g  c g t a a g c a
g g a t  a c g  a g a t g a c a
```

Sequences ↓

N=1:10

Location along the genome ⟶

- Several locations are conserved by evolution (e.g., because they are part of a gene coding region), since the corresponding columns tend to be pure e.g., column 7 is all g's.

- To visuallize the data (**sequence logo**), we plot the letters A, C, G and T with a font size proportional to their empirical probability, and with the most probable letter on the top.

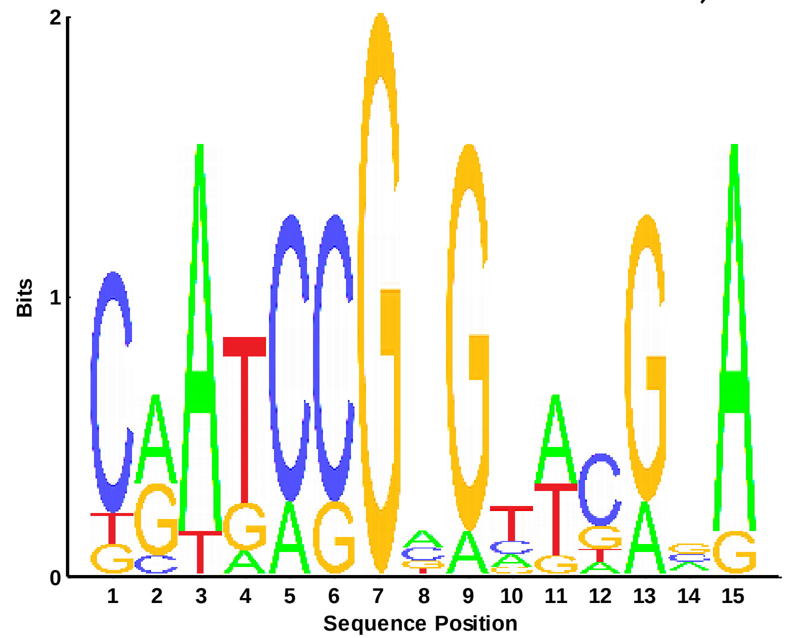
Example: Biosequence Analysis

- The empirical probability distribution at location t, is obtained by normalizing the vector of counts (see MLE estimate)

$$\hat{\boldsymbol{\theta}}_t = \frac{1}{N}\left(\sum_{i=1}^{N}\mathbb{I}(X_{it}=1), \sum_{i=1}^{N}\mathbb{I}(X_{it}=2), \sum_{i=1}^{N}\mathbb{I}(X_{it}=3), \sum_{i=1}^{N}\mathbb{I}(X_{it}=4)\right)$$

- *This distribution is known as a **motif**.*

- Can also compute the *most probable letter in each location*; this is the *consensus sequence.*



Use MatLab function *seqlogoDemo* from Kevin Murphys' PMTK

Summary of Discrete Distributions

- A summary of the multinomial and related discrete distributions is summarized below on a Table from Kevin Murphy's textbook

Name	n	K	x
Multinomial	-	-	$\mathbf{x} \in \{0, 1, \ldots, n\}^K, \sum_{k=1}^K x_k = n$
Multinoulli	1	-	$\mathbf{x} \in \{0, 1\}^K, \sum_{k=1}^K x_k = 1$ (1-of-K encoding)
Binomial	-	1	$x \in \{0, 1, \ldots, n\}$
Bernoulli	1	1	$x \in \{0, 1\}$

- $n = 1$ (one roll of the dice), $n = -$ (N rolls of the dice)

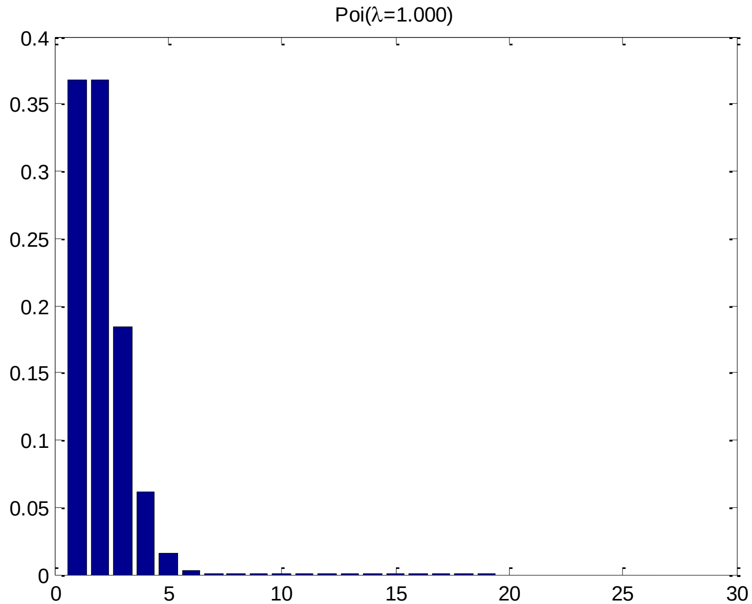
- $K = 1$ (binary variables), $K = -$ (1-of-K encoding)

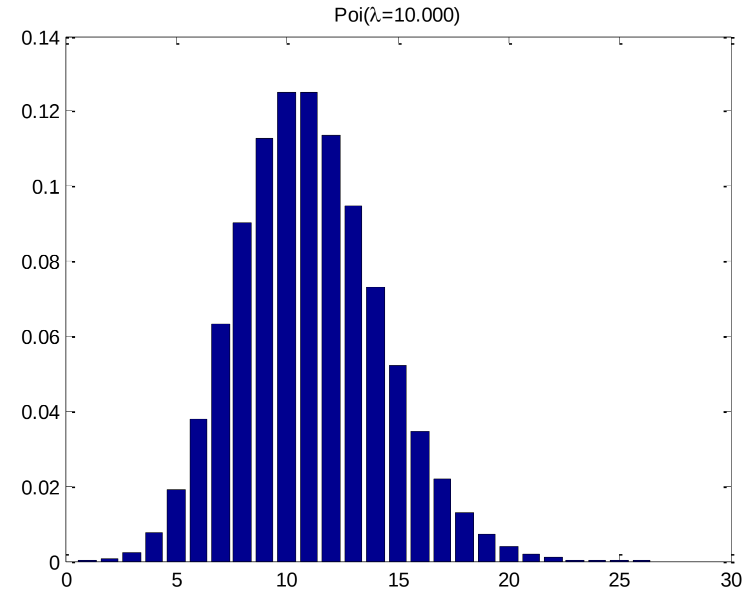
The Poisson Distribution

➢ We say that $X \in \{0,1,2,3,\ldots\}$ has a Poisson distribution with parameter $\lambda >0$, if its pmf is

$$X \sim \mathcal{Poi}(\lambda): \mathcal{Poi}(x \mid \lambda) = e^{-\lambda} \frac{\lambda^x}{x!}$$

➢ This is a model for counts of rare events.





Use MatLab function poissonPlotDemo from Kevin Murphys' PMTK

The Empirical Distribution

- Given data, $\mathcal{D} = \{x_1, \dots, x_N\}$, we define the empirical distribution as:

$$p_{emp}(A) = \frac{1}{N}\sum_{i=1}^{N}\delta_{x_i}(A),\ Dirac\ Measure: \delta_{x_i}(A) = \begin{cases} 1\ if\ \ x_i \in A \\ 0\ if\ \ x_i \notin A \end{cases}$$

- We can also associate weights with each sample:

$$Generalize\ \ p_{emp}(x) = \frac{1}{N}\sum_{i=1}^{N}\delta_{x_i}(x)\ \Rightarrow p_{emp}(x) = \sum_{i=1}^{N} w_i\delta_{x_i}(x)\ ,\ 0 \le w_i \le 1, \sum_{i=1}^{N} w_i = 1$$

- This corresponds to a histogram with spikes at each sample point with height equal to the corresponding weight. This distribution assigns zero weight to any point not in the dataset.

- Note that the "sample mean of $f(x)$" is the expectation of $f(x)$ under the empirical distribution:

$$\mathbb{E}[f(x)] = \int f(x)\sum_{i=1}^{N}\frac{1}{N}\delta_{x_i}(x)dx = \frac{1}{N}\sum_{n=1}^{N} f(x_i)$$

Student's $\mathcal{T}$ Distribution

$$p(x|\mu,\lambda,\upsilon)=\frac{\Gamma(\frac{\upsilon}{2}+\frac{1}{2})}{\Gamma(\frac{\upsilon}{2})}\left(\frac{\lambda}{\pi\upsilon}\right)^{1/2}\left[1+\frac{\lambda\left(x-\mu\right)^2}{\upsilon}\right]^{-\upsilon/2-1/2}$$

- The parameter λ is called the precision of the $\mathcal{T}$-distribution, even though it is not in general equal to the inverse of the variance (see below on behavior as $\upsilon \to \infty$).

- The parameter υ is called the degrees of freedom.

- For the particular case of $\upsilon = 1$, the $\mathcal{T}$-distribution reduces to the Cauchy distribution.

- In the limit $\upsilon \to \infty$, the $\mathcal{T}$-distribution $\mathcal{T}(x|\mu,\lambda,\upsilon)$ becomes a Gaussian $\mathcal{N}(x|\mu,\lambda^{-1})$ with mean μ and precision λ.

For $\upsilon \to \infty$, $\mathcal{T}(x|\mu,\lambda,\upsilon)$ Becomes a Gaussian

$$p(x|\mu,\lambda,\upsilon)=\frac{\Gamma(\frac{\upsilon}{2}+\frac{1}{2})}{\Gamma(\frac{\upsilon}{2})}\left(\frac{\lambda}{\pi\upsilon}\right)^{1/2}\left[1+\frac{\lambda\left(x-\mu\right)^2}{\upsilon}\right]^{-\upsilon/2-1/2}$$

- We first write the distribution as follows:

$$\mathcal{T}(x|\mu,\lambda,\upsilon)\propto\left[1+\frac{\lambda\left(x-\mu\right)^2}{\upsilon}\right]^{-\upsilon/2-1/2}=\exp\left\{-\frac{\upsilon+1}{2}\ln\left[1+\frac{\lambda\left(x-\mu\right)^2}{\upsilon}\right]\right\}$$

- For large υ, we can approximate the log as follows:

$$\mathcal{T}(x|\mu,\lambda,\upsilon)\propto\exp\left\{-\frac{\upsilon+1}{2}\left[\frac{\lambda\left(x-\mu\right)^2}{\upsilon}+O(\upsilon^{-2})\right]\right\}=\exp\left\{-\frac{\lambda\left(x-\mu\right)^2}{2}+O(\upsilon^{-1})\right\}$$

- In the limit $\upsilon \to \infty$, the $\mathcal{T}$-distribution $\mathcal{T}(x|\mu,\lambda,\upsilon)$ is indeed a Gaussian $\mathcal{N}(x|\mu,\lambda^{-1})$ with mean μ and precision λ. The normalization of the $\mathcal{T}$ is valid in this limit as well (so the Gaussian obtained is normalized).

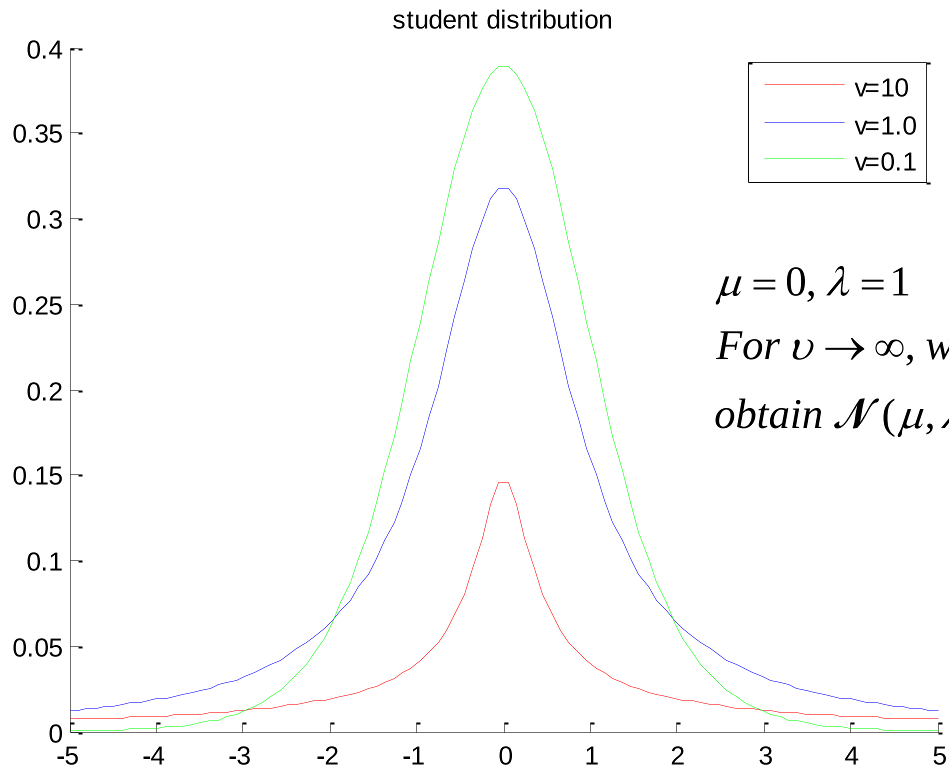
Student's $\mathcal{T}$ Distribution

$$p(x|\mu,\lambda,\upsilon)=\frac{\Gamma(\frac{\upsilon}{2}+\frac{1}{2})}{\Gamma(\frac{\upsilon}{2})}\left(\frac{\lambda}{\pi\upsilon}\right)^{1/2}\left[1+\frac{\lambda\left(x-\mu\right)^{2}}{\upsilon}\right]^{-\upsilon/2-1/2}$$

$$Mean:\mu,\upsilon>1$$

$$Mode:\mu$$

$$Var:\frac{\upsilon}{\lambda\left(\upsilon-2\right)},\upsilon>2$$



$\mu=0,\lambda=1$

For $\upsilon\to\infty$, we obtain $\mathcal{N}(\mu,\lambda^{-1})$

MatLab Code

Student's $\mathcal{T}$ Vs the Gaussian

- We plot:

$$\mathcal{N}\left(x\,|\,0,1\right), \mathcal{T}(x\,|\,0,1,1), \mathcal{Lap}\left(x\,|\,0,1/\sqrt{2}\right)$$

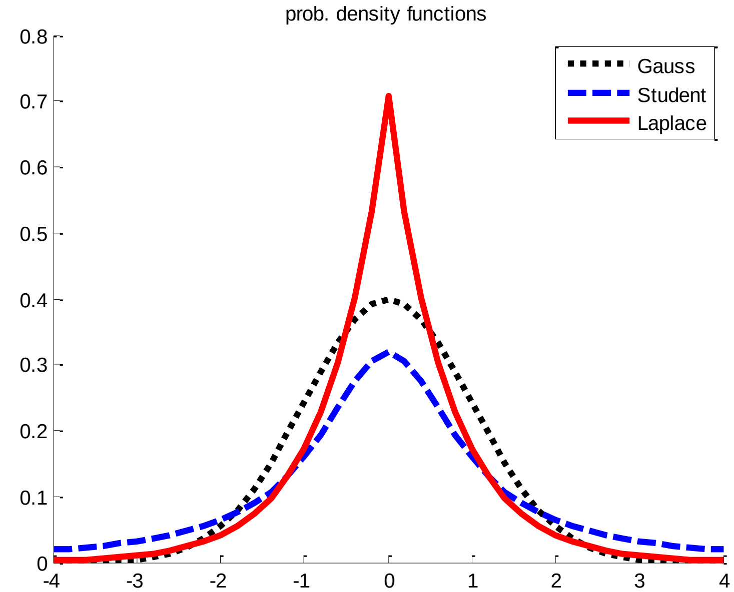
- *The mean and variance of the Student's is undefined for $\upsilon = 1$.*

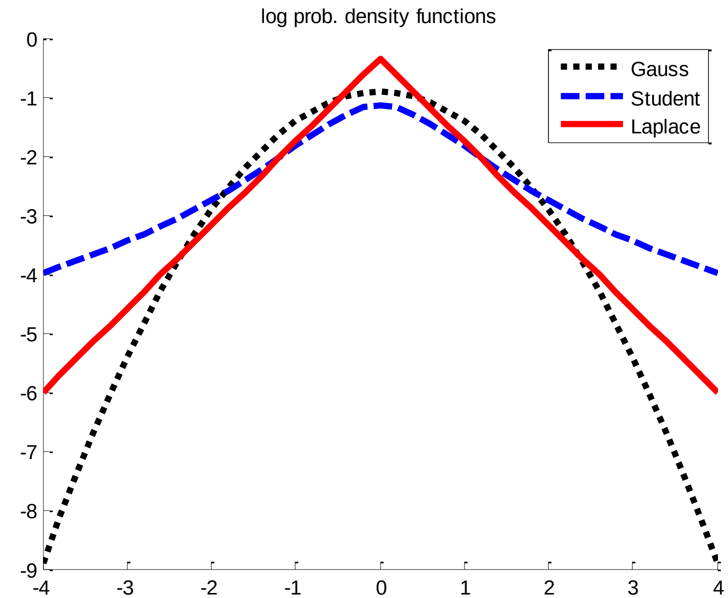
- Logs of the PDFs*. The Student's is NOT log concave.*

Run MatLab function *studentLaplacePdfPlot* from Kevin Murphys' PMTK

- *When $\upsilon = 1$, the distribution is known as Cauchy or Lorentz. Due to its heavy tails, the mean does not converge.*

- ***Recommended to use $\upsilon = 4$.***





Student's $\mathcal{T}$ Distribution

$$p(x \mid \mu, a, b) = \int_0^\infty \mathcal{N}\left(x \mid \mu, \tau^{-1}\right) \mathcal{Gamma}\left(\tau \mid a, b\right) d\tau =$$

$$= \int_0^\infty \left(\frac{\tau}{2\pi}\right)^{1/2} \exp\left(-\frac{\tau}{2}\left(x-\mu\right)^2\right) \frac{b^a}{\Gamma(a)} \tau^{a-1} e^{-b\tau} d\tau$$

- The Student's $\mathcal{T}$ distribution can be seen from the equation above (see following two slides for proof) as an infinite ***mixture of Gaussians each of them with different precision*** (governed by a Gamma distribution)
- The result is a distribution that in general has longer 'tails' than a Gaussian.
- This gives the $\mathcal{T}$-distribution robustness, i.e. the $\mathcal{T}$-distribution is much less sensitive than the Gaussian to the presence of outliers.

Appendix: Student's $\mathcal{T}$ as a Mixture of Gaussians

- If we have a univariate Gaussian $\mathcal{T}(x|\mu,\tau^{-1})$ together with a prior $\mathcal{G}amma(\tau\,|a,b)$ and we integrate out the precision, we obtain the marginal distribution of x

$$p(x\,|\,\mu,a,b)=\int_0^\infty \mathcal{N}\left(x\,|\,\mu,\tau^{-1}\right)\mathcal{G}amma\left(\tau\,|\,a,b\right)d\tau=$$

$$=\int_0^\infty\left(\frac{\tau}{2\pi}\right)^{1/2}\exp\left(-\frac{\tau}{2}\left(x-\mu\right)^2\right)\frac{b^a}{\Gamma(a)}\tau^{a-1}e^{-b\tau}d\tau=$$

- Introduce the transformation $z=\underbrace{\left[b+\frac{1}{2}\left(x-\mu\right)^2\right]}_{A}\tau$ to simplify as:

$$p(x\,|\,\mu,a,b)=\frac{b^a}{\Gamma(a)}\left(\frac{1}{2\pi}\right)^{1/2}\int_0^\infty\tau^{1/2}\exp\left(-z\right)\tau^{a-1}d\tau=$$

$$=\frac{b^a}{\Gamma(a)}\left(\frac{1}{2\pi}\right)^{1/2}\frac{1}{A^{1/2+a-1+1}}\int_0^\infty z^{1/2}\exp\left(-z\right)z^{a-1}dz$$

Appendix: Student's $\mathcal{T}$ as a Mixture of Gaussians

$$p(x|\mu,a,b)=\frac{b^a}{\Gamma(a)}\left(\frac{1}{2\pi}\right)^{1/2}\frac{1}{A^{1/2+a-1+1}}\int_0^\infty z^{1/2}\exp(-z)\,z^{a-1}dz$$

$$=\frac{b^a}{\Gamma(a)}\left(\frac{1}{2\pi}\right)^{1/2}\left[b+\frac{1}{2}(x-\mu)^2\right]^{-a-1/2}\int_0^\infty \exp(-z)\,z^{a-1/2}dz$$

- Recalling the definition of the Gamma function: $\Gamma(a)=\int_0^\infty \exp(-z)\,z^{a-1}dz$

$$p(x|\mu,a,b)=\frac{b^a}{\Gamma(a)}\left(\frac{1}{2\pi}\right)^{1/2}\left[b+\frac{1}{2}(x-\mu)^2\right]^{-a-1/2}\Gamma(a+\frac{1}{2})$$

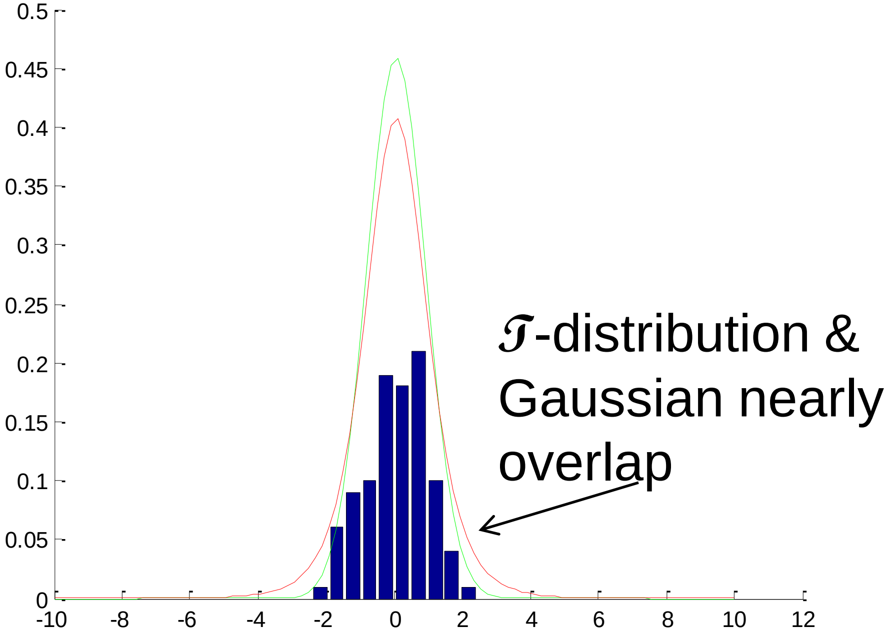
- It is common to redefine the parameters in this distribution as: $\upsilon=2a,\ \lambda=\frac{a}{b}$

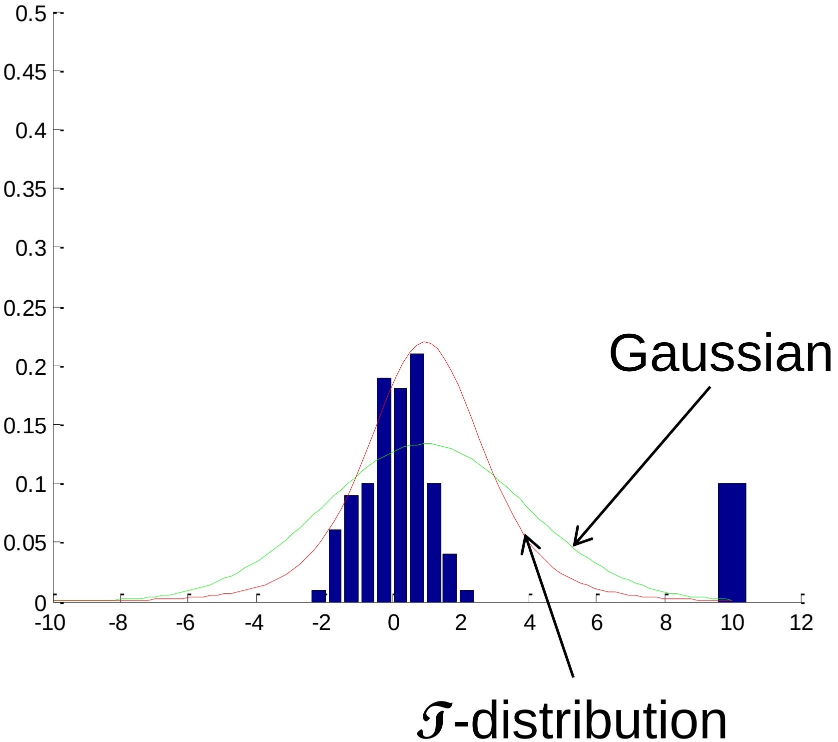
$$p(x|\mu,\lambda,\upsilon)=\frac{\Gamma(\frac{\upsilon}{2}+\frac{1}{2})}{\Gamma(\frac{\upsilon}{2})}\left(\frac{\lambda}{\pi\upsilon}\right)^{1/2}\left[1+\frac{\lambda(x-\mu)^2}{\upsilon}\right]^{-\upsilon/2-1/2}$$

Robustness of Student's $\mathcal{T}$ Distribution

- The robustness of the $\mathcal{T}$-distribution is illustrated here by comparing the "maximum likelihood solutions" for a Gaussian and a $\mathcal{T}$-distribution (30 data points from the Gaussian are used).

- The effect of a small number of outliers (Fig. on the right) is less significant for the $\mathcal{T}$-distribution than for the Gaussian.

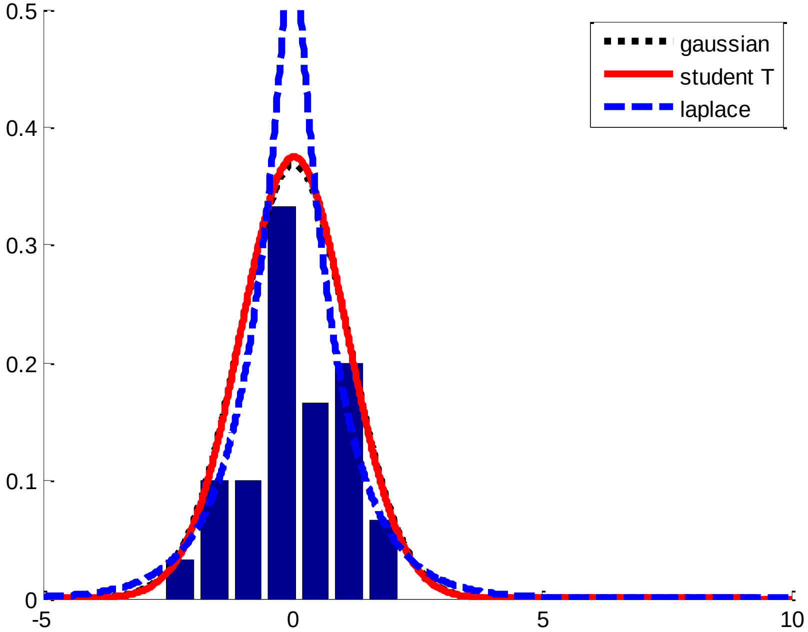


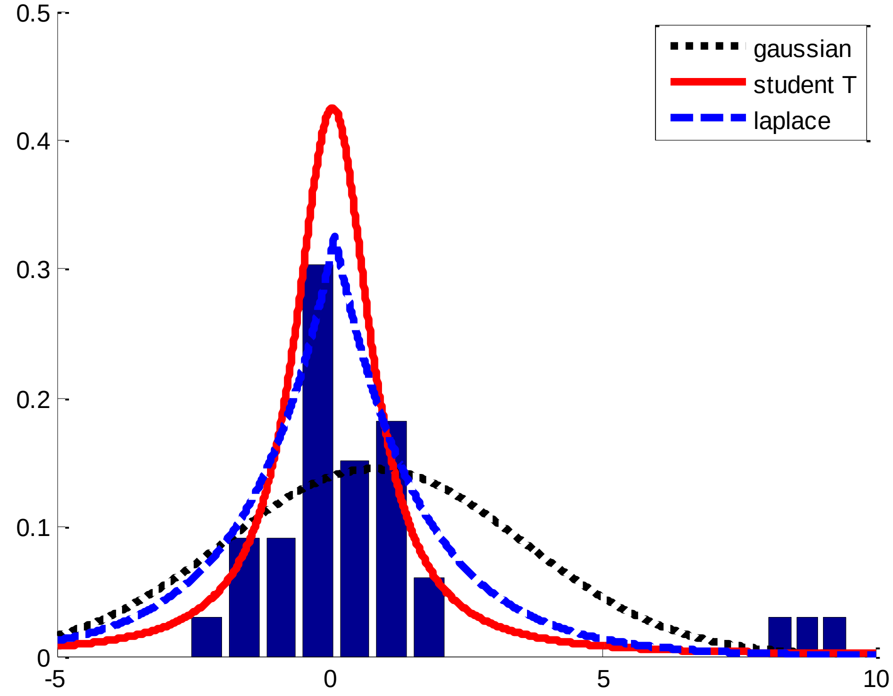


[MatLab Code]

Robustness of Student's $\mathcal{T}$ Distribution

- The earlier simulation is repeated here with the PMTK toolbox.





Run MatLab function *robustDemo*
from Kevin Murphys' PMTK

The Laplace Distribution

- Another distribution with heavy tails is the Laplace distribution, also known as the *double sided exponential distribution*. It has the following pdf:

$$\mathcal{Lap}\left(x \mid \mu, b\right) = \frac{1}{2b} e^{-\frac{|x-\mu|}{b}}$$

- μ is a location parameter and $b > 0$ is a scale parameter

$$Mean = \mu, Mode = \mu, Var = 2b^2$$

- Its robust to outliers (see earlier demonstration).

- *It puts mores probability density at* 0 *than the Gaussian. This property is a useful way to encourage sparsity in a model*.

Beta Distribution

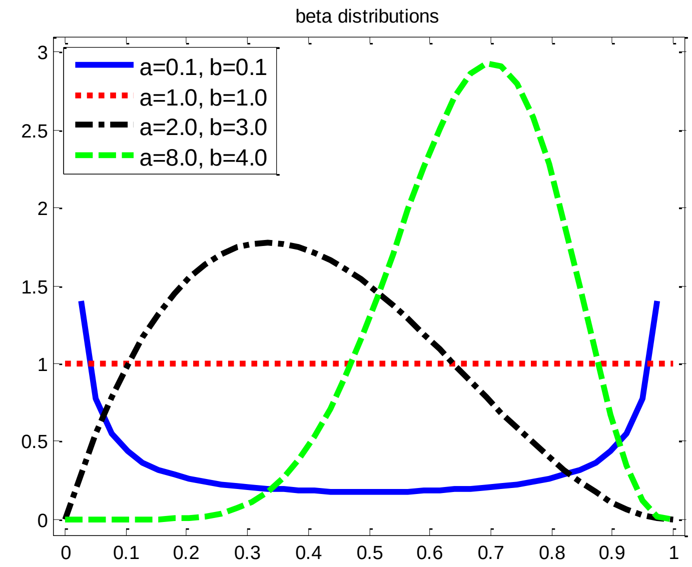
- The $\mathcal{Beta}(\alpha,\beta)$ distribution with $x \in [0,1], \alpha, \beta > 0$ is defined as follows:

$$Beta(x) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} x^{\alpha-1}(1-x)^{\beta-1} = \frac{x^{\alpha-1}(1-x)^{\beta-1}}{\underbrace{beta(\alpha,\beta)}_{\text{Normalizing factor}}}, \quad beta(\alpha,\beta) = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}$$

- The expected value, mode and variance of a $\mathcal{Beta}$ random variable x with (hyper-) parameters α and β :

$$\mathbb{E}[x] = \frac{\alpha}{\alpha+\beta}, \quad \text{mode}[x] = \frac{a-1}{a+b-2}$$

$$var[x] = \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}$$



- For more information visit this link.

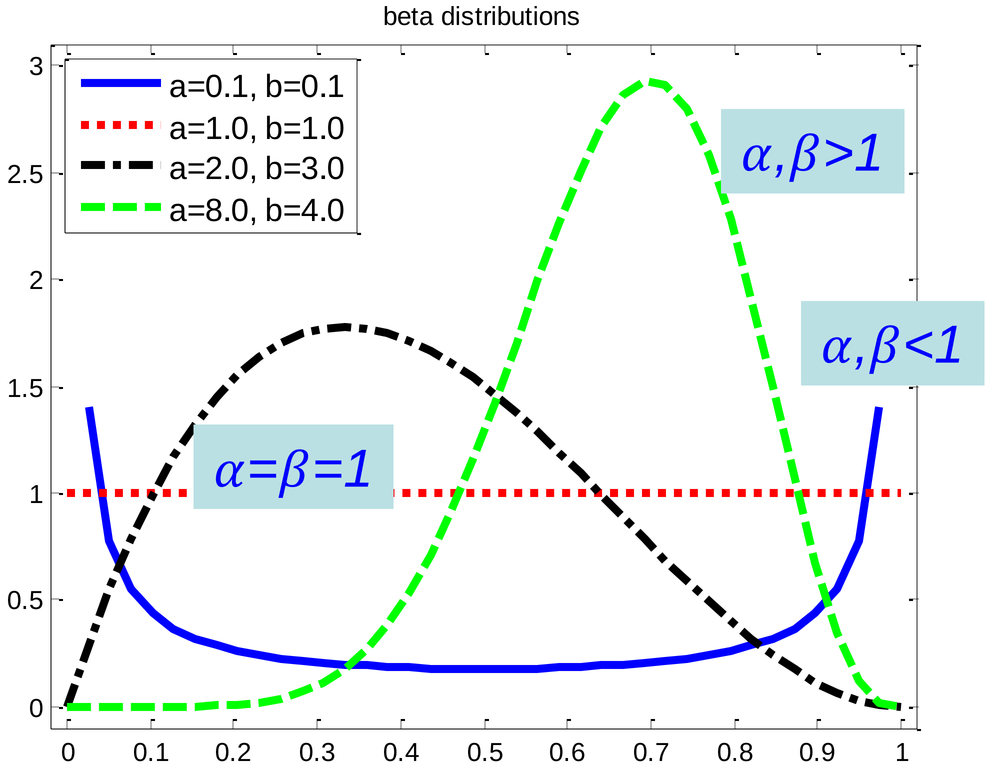
Beta Distribution

- *If $\alpha = \beta = 1$, we obtain a uniform distribution.*

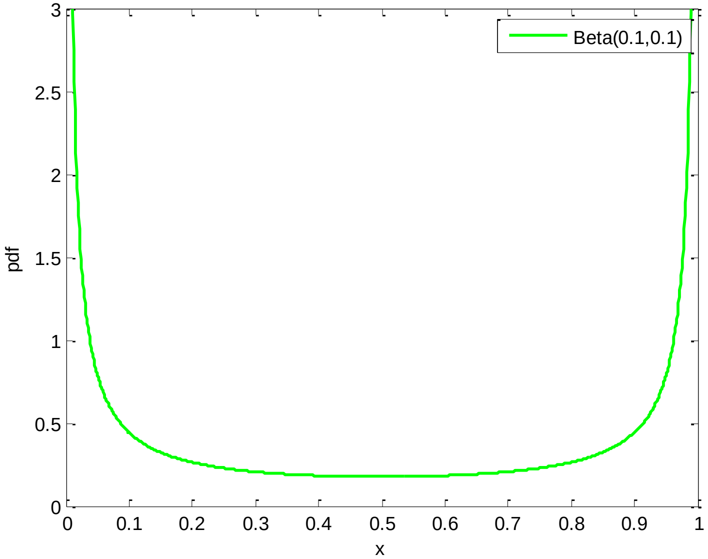
- If α and β are both less than 1, we get a bimodal distribution with spikes at 0 and 1.

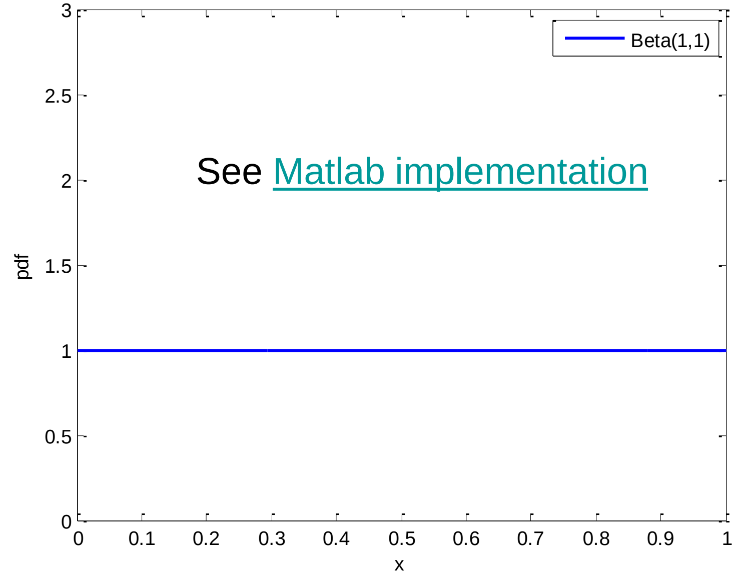
- If α and β are both greater than 1, the distribution is unimodal.

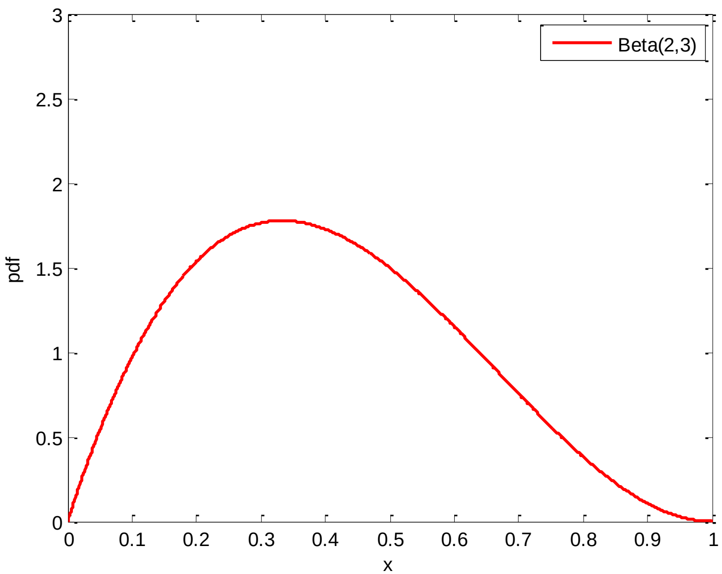
Run *betaPlotDemo*
from PMTK

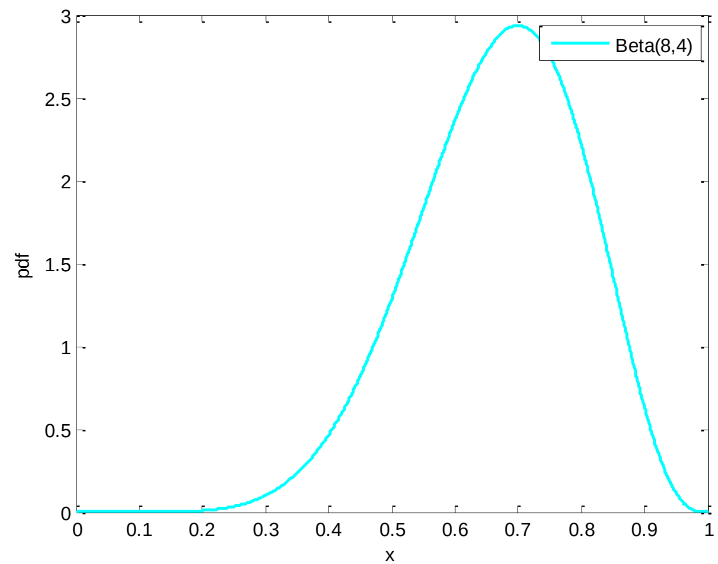


Beta Distribution









Gamma Function

$$Beta(x) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} x^{\alpha-1}(1-x)^{\beta-1}$$

- The gamma function extends the factorial to real numbers:

$$\Gamma(x) = \int_0^\infty u^{x-1} e^{-u} du$$

- With integration by parts:

$$\Gamma(x+1) = x\Gamma(x)$$

- For integer n:

$$\Gamma(n) = (n-1)!$$

- For more information visit this link.

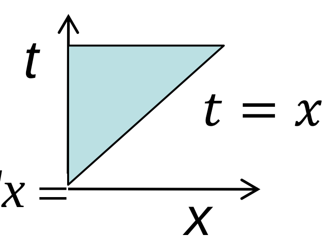
Beta Distribution: Normalization

- Showing that the $\mathcal{B}eta(\alpha,\beta)$ distribution is normalized correctly is a bit tricky. We need to prove that:

$$\Gamma(\alpha)\Gamma(\beta)=\Gamma(\alpha+\beta)\int_0^1 \mu^{\alpha-1}(1-\mu)^{\beta-1}\,d\mu$$

Follow the steps:

- ✓ (a) change the variable y below to $y = t - x$;
- ✓ (b) change the order of integration in the shaded triangular region;
- ✓ and (c) change x to *n* via $x = t\mu$:



$$\Gamma(\alpha)\Gamma(\beta)=\left(\int_0^\infty x^{\alpha-1}e^{-x}dx\right)\left(\int_0^\infty y^{\beta-1}e^{-y}dy\right)\underset{y=t-x}{=}\int_0^\infty x^{\alpha-1}\left(\int_x^\infty e^{-t}(t-x)^{\beta-1}\,dt\right)dx=$$

$$=\int_0^\infty\left(\int_0^t x^{\alpha-1}e^{-t}(t-x)^{\beta-1}\,dx\right)dt=\int_0^\infty t^{\alpha-1}e^{-t}t^{\beta-1}t\,dt\int_0^1 \mu^{\alpha-1}(1-\mu)^{\beta-1}\,d\mu=$$

$$=\Gamma(\alpha+\beta)\int_0^1 \mu^{\alpha-1}(1-\mu)^{\beta-1}\,d\mu$$

Gamma Distribution- Rate Parametrization

- It is frequently a model for waiting times. For important properties see here.

- It is more often parameterized in terms of a shape parameter a and an inverse scale parameter $b = 1/\theta$, called *a rate parameter*:

$$p(x|a,b) = \frac{b^a}{\Gamma(a)} x^{a-1} e^{-bx}, \; x \in [0,\infty], \; \Gamma(a) = \int_0^\infty u^{a-1} e^{-u} du$$

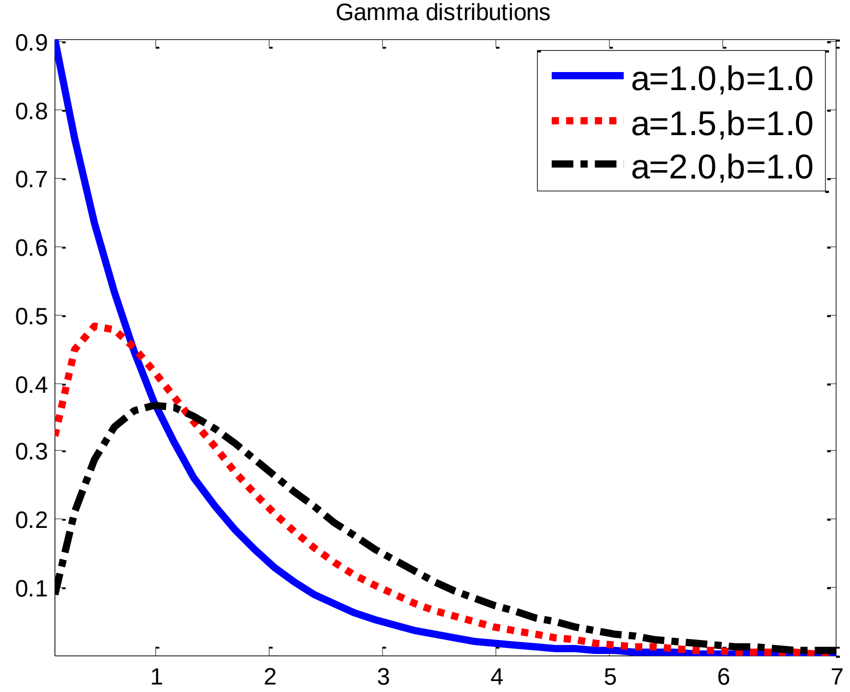
- The mean, mode and variance with this parametrization are:

$$\mathbb{E}[x] = \frac{\alpha}{b} \qquad \mathrm{mod}\, e[x] = \begin{cases} \frac{a-1}{b}, & for\ a > 1 \\ 0 & otherwise \end{cases} \qquad var[x] = \frac{\alpha}{b^2}$$

Gamma Distribution

- Plots of $$\mathcal{Gamma}(X \mid a,b) = \frac{b^a}{\Gamma(a)} x^{a-1} \exp(-xb), b=1$$

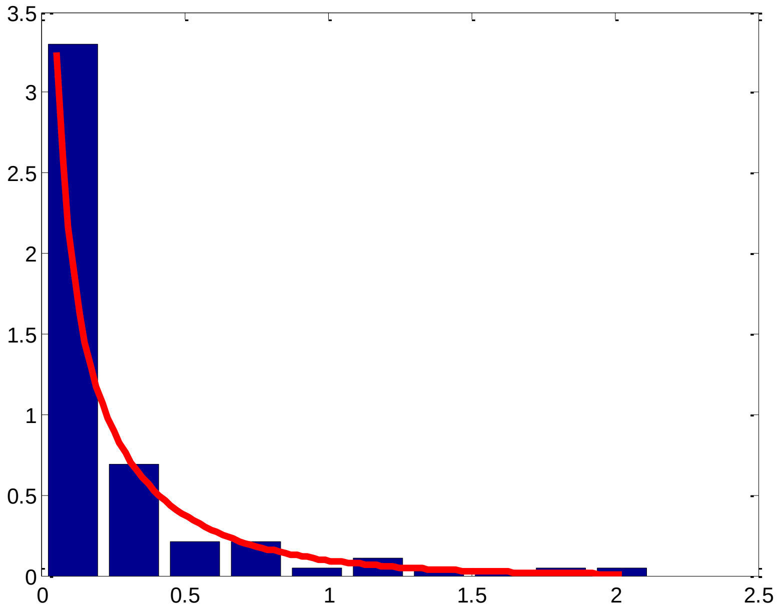
- As we increase the rate b, the distribution squeezes leftwards and upwards. For $a < 1$, the mode is at zero.

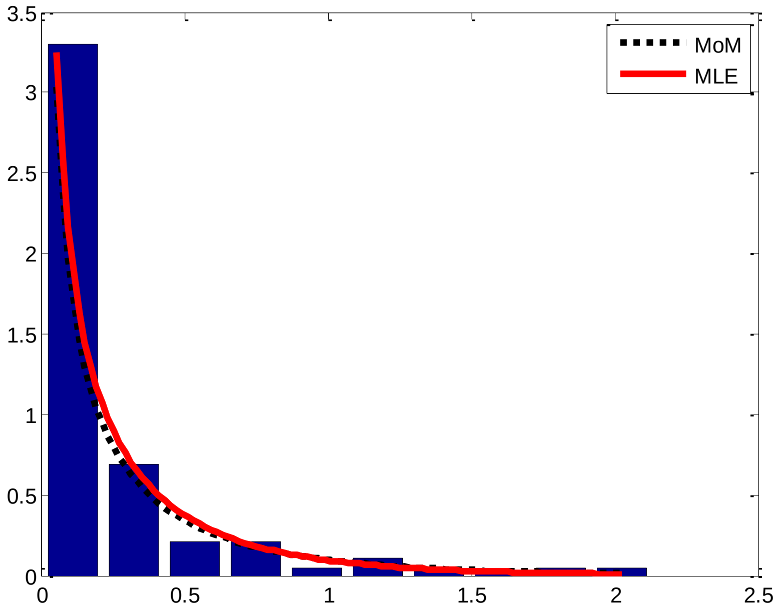


Run gammaPlotDemo
from PMTK

Gamma Distribution

- An empirical PDF of rainfall data fitted with a Gamma distribution.





Run MatLab function [*gammaRainfallDemo*]
from [PMTK]

Exponential Distribution

- This is defined as

$$\mathcal{E}xpon(X \mid \lambda) = \mathcal{G}amma(X \mid 1, \lambda) = \lambda \exp(-x\lambda),\ x \in [0, \infty]$$

-
- Here λ is the rate parameter.

- This *distribution describes the times between events in a Poisson process, i.e. a process in which events occur continuously and independently at a constant average rate* λ*.*

Chi-Squared Distribution

- This is defined as

$$\chi^2(X \mid \nu) = \mathcal{G}amma(X \mid \frac{\nu}{2}, \frac{1}{2}) = \frac{\left(\frac{1}{2}\right)^{\frac{\nu}{2}}}{\Gamma\left(\frac{\nu}{2}\right)} x^{\frac{\nu}{2}-1} \exp(-\frac{x}{2}), x \in [0, \infty]$$

- *This is the distribution of the sum of squared Gaussian random variables.*

- More precisely,

$$Let\ Z_i \sim \mathcal{N}(0,1)\ and\ S = \sum_{i=1}^{\nu} Z_i^2,\ then: S \sim \chi_\nu^2$$

Inverse Gamma Distribution

- This is defined as follows:

$$If\ \ X \sim \mathcal{Gamma}(X \mid a,b) \Rightarrow X^{-1} \sim \mathcal{InvGamma}(X \mid a,b)$$

where:

$$\mathcal{InvGamma}(X \mid a,b) = \frac{b^a}{\Gamma(a)} x^{-(a+1)} \exp(-b/x),\ x \in [0,\infty]$$

- *a is the shape and b the scale parameters.*

- Note that b is a scale parameter since:

$$\mathcal{InvGamma}(X \mid a,b) = \frac{\mathcal{InvGamma}(\frac{X}{b} \mid a,1)}{b}$$

- It can be shown that:

$$Mean = \frac{b}{a-1}\ (exists\ \ for\ a>1),\ Mode = \frac{b}{a+1},\ \mathrm{var} = \frac{b^2}{(a-1)^2(a-2)}\ (exists\ \ for\ a>2)$$

The Pareto Distribution

- Used to model the distribution of quantities that exhibit **long tails** (heavy tails)

$$\mathcal{Pareto}(X|\mathrm{k},m) = km^k x^{-(k+1)}\mathbb{I}(x \geq m)$$

- This density asserts that x must be greater than some constant m, but not too much greater, k controls what is “too much”.

- Modeling the frequency of words vs. their rank (e.g. “the”, “of”, etc.) or the wealth of people.*

- As $k \to \infty$, the distribution approaches $\delta(x - m)$.

- On a log-log scale, the pdf forms a straight line of the form $\log p(x) = a\log x + c$ for some constants a and c (power law, Zipf’s law).

* Basis of the distribution: a high proportion of a population has low income and only few have very high incomes.

The Pareto Distribution

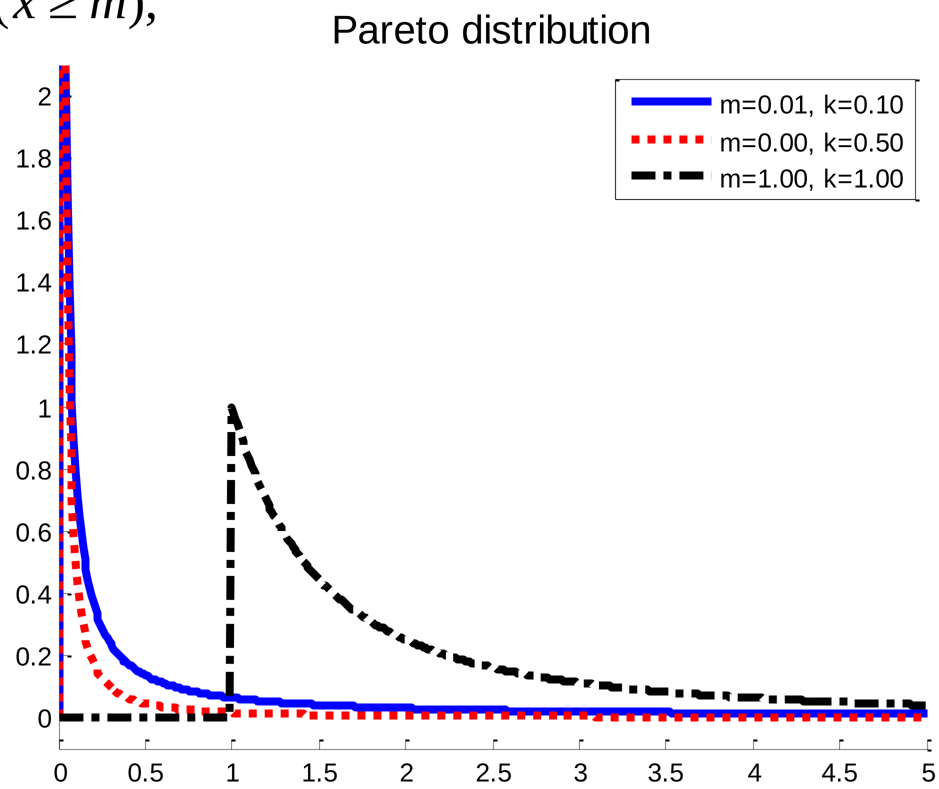
➢ Applications: Modeling the frequency of words vs their rank, distribution of wealth (k =Pareto Index), etc.

$$\mathscr{Pareto}(X \mid k, m) = k m^k x^{-(k+1)} \mathbb{I}(x \geq m),$$

$$Mean = \frac{km}{k-1} \ (if \ \ k > 1),$$

$$Mode = m,$$

$$\text{var} = \frac{m^2 k}{(k-1)^2 (k-2)} \ (if \ \ k > 2)$$



ParetoPlot from PMTK

Covariance

- Consider two random variables $X, Y : \Omega \to \mathbb{R}.$

- The joint probability distribution is defined as:

$$P\left\{X \in A, Y \in B\right\} = P\left\{X^{-1}(A) \bigcap Y^{-1}(B)\right\} = \iint_{A\times B} p(x,y)dxdy$$

- Two random variables are independent if

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

- The covariance of *X* and *Y* is defined as:

$$\text{cov}(X,Y) = \mathbb{E}\left[(X - \mathbb{E}(X))(Y - \mathbb{E}(Y))\right]$$

- It is straight forward to verify that: $\text{cov}(X,Y) = \mathbb{E}\left[XY\right] - \mathbb{E}\left[X\right]\mathbb{E}\left[Y\right]$

Correlation, Center Normalized Random Variables

- Consider two random variables $X, Y : \Omega \to \mathbb{R}$.

- The correlation coefficient of X and Y is defined as:

$$corrc(X,Y) = \frac{\text{cov}(X,Y)}{\sigma_X \sigma_Y}$$

where the standard deviations of X and Y are

$$\sigma_X = \sqrt{\text{cov}(X)}, \sigma_Y = \sqrt{\text{cov}(Y)}$$

- The center normalized random variables are defined as:

$$\begin{cases} \tilde{X} = \dfrac{X - \mathbb{E}[X]}{\sigma_X} \\ \tilde{Y} = \dfrac{Y - \mathbb{E}[Y]}{\sigma_Y} \end{cases}$$

- It is straight forward to verify that:

$$\mathbb{E}\left[\tilde{X}\right] = \mathbb{E}\left[\tilde{Y}\right] = 0 \qquad \text{var}\left[\tilde{X}\right] = \text{var}\left[\tilde{Y}\right] = 1$$