Probability

ddebarr@uw.edu

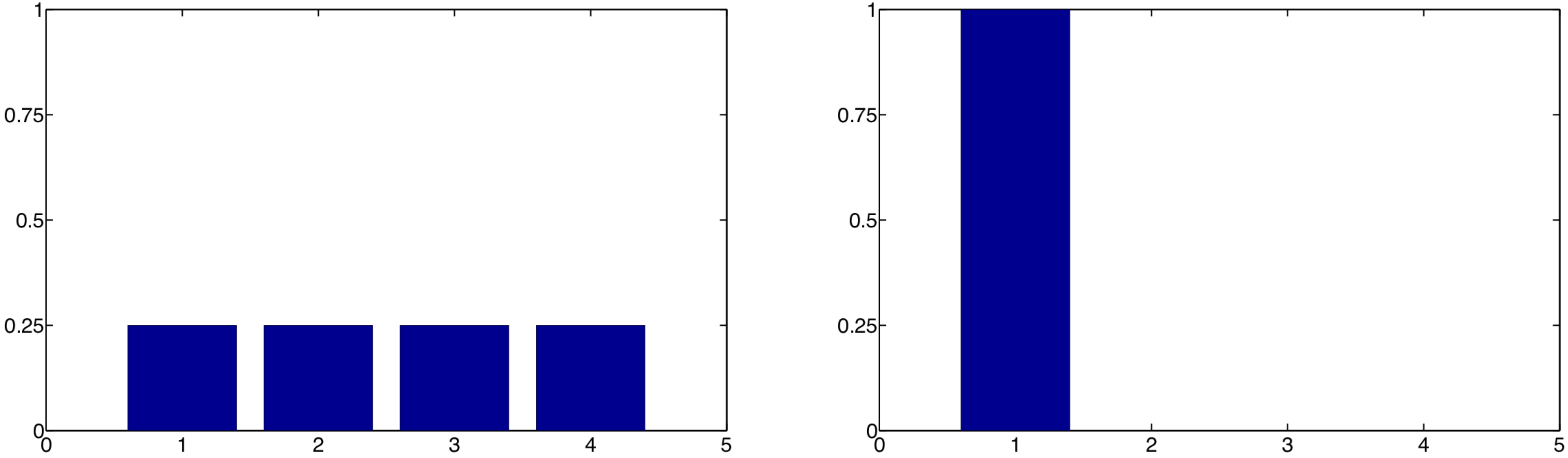
2016-04-14

Agenda

- Fundamentals
- Discrete Distributions
- Continuous Distributions
- Joint Distributions
- Transformations
- Monte Carlo Approximation
- Information Theory

Discrete Random Variables

PMF: Probability Mass Function



maximum entropy
(“uniform” distribution)

minimum entropy
(“degenerate distribution”: constant)

Bayes' Rule: Conditional Probability

$$p(X = x|Y = y) = \frac{p(X = x, Y = y)}{p(Y = y)} = \frac{p(X = x)p(Y = y|X = x)}{\sum_{x'} p(X = x')p(Y = y|X = x')}$$

Can we describe precision and recall as probabilities?

Medical Diagnosis Example

- Given likelihood of cancer prediction and prior for actual cancer …
  - p(prediction = cancer | actual = cancer) = 0.8
  - p(prediction = cancer | actual = not cancer) = 0.1
  - p(actual = cancer) = 0.004
- Derive posterior …
  - p(actual = cancer | prediction = cancer)
- Bayes' rule says posterior is …

  = (likelihood * prior) / evidence

  = (0.8 * 0.004) / (0.8 * 0.004 + 0.1 * 0.996)

  = 0.0032 / (0.0032 + 0.0996)

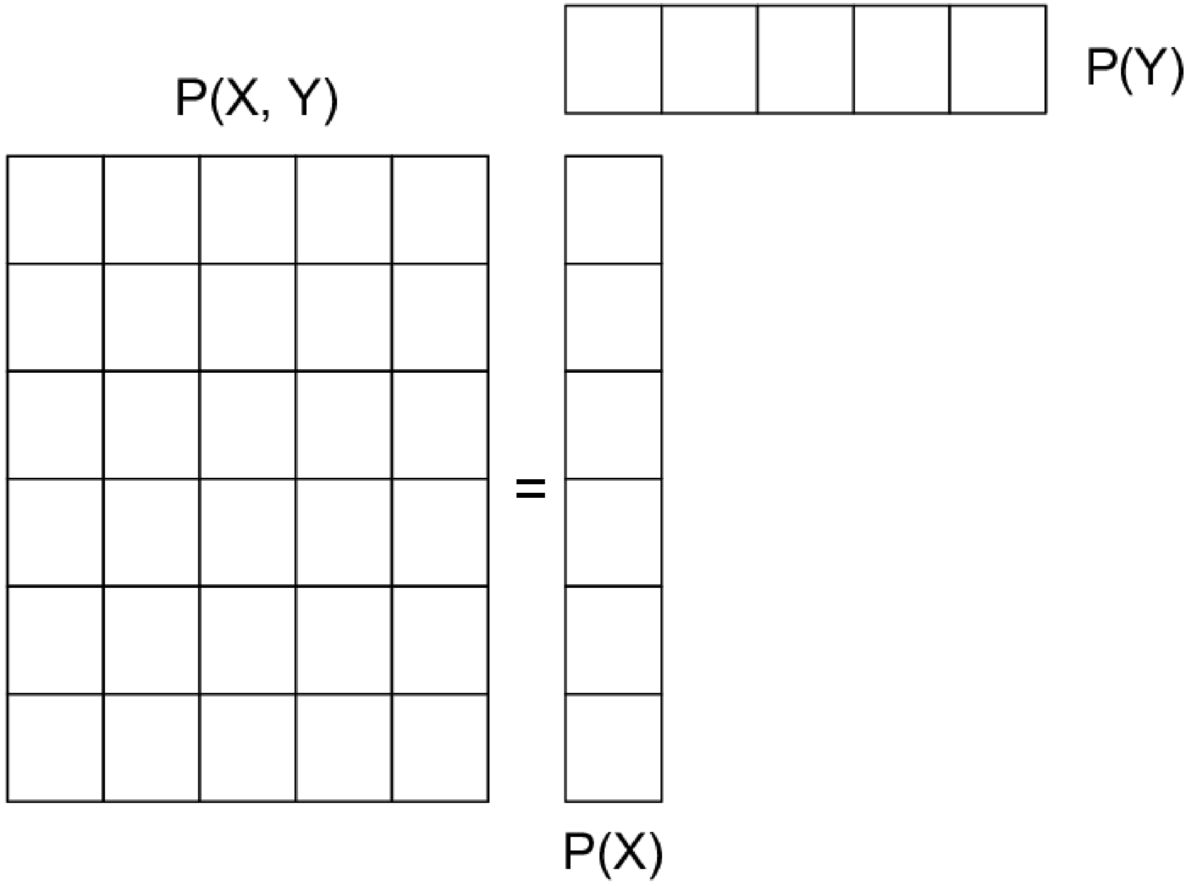
  = 0.0312
- Generative classifier:

$$p(y = c|\mathbf{x}) = \frac{p(y = c)p(\mathbf{x}|y = c)}{\sum_{c'} p(y = c'|\boldsymbol{\theta})p(\mathbf{x}|y = c')}$$

Independence and Conditional Independence

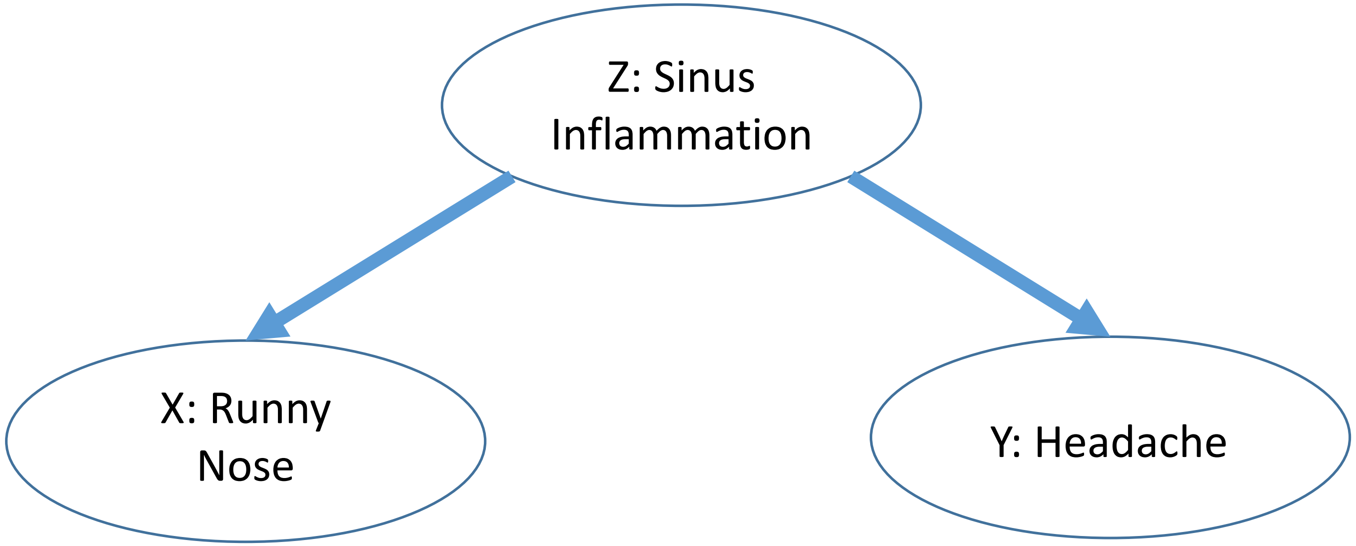
Unconditionally independent:

$$X \perp Y \iff p(X, Y) = p(X)p(Y)$$

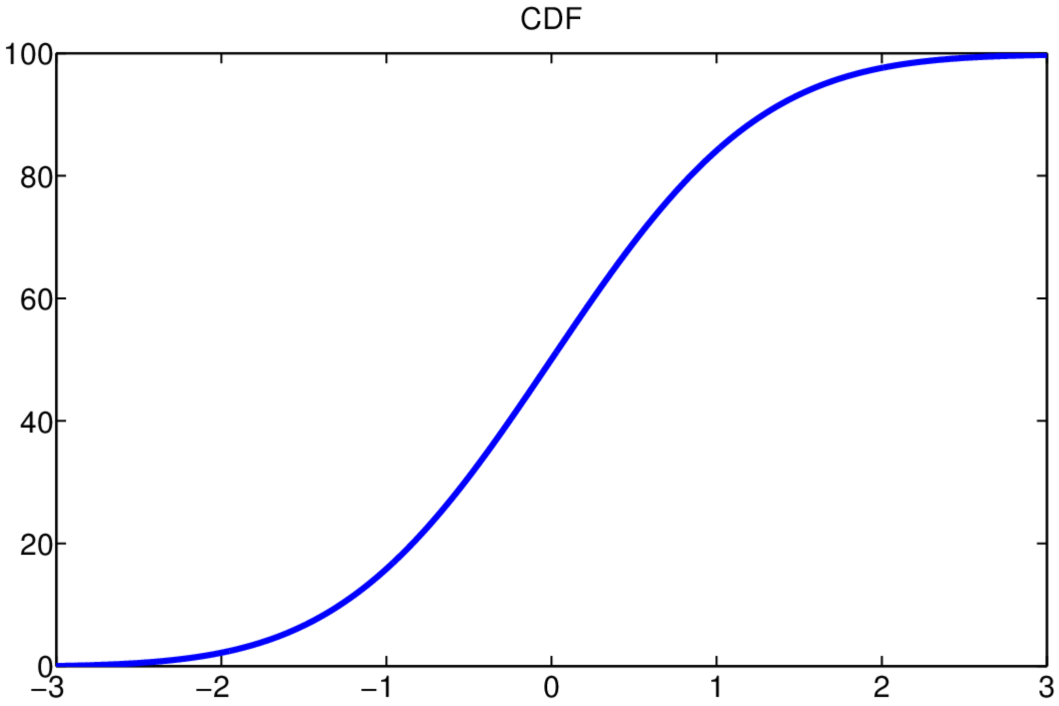


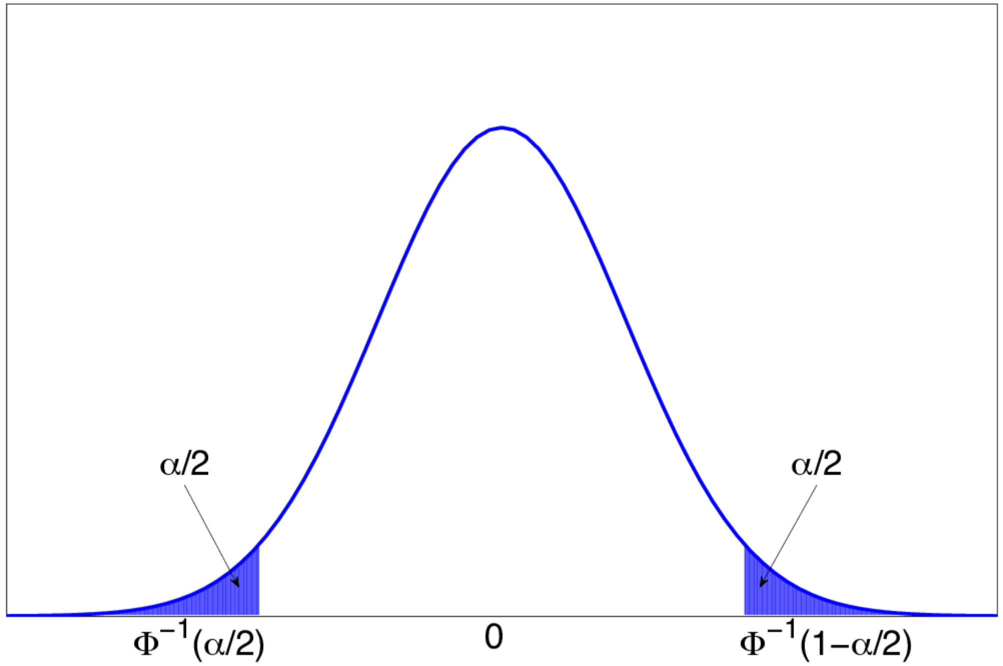
Conditionally independent:

$$X \perp Y | Z \iff p(X, Y | Z) = p(X | Z)p(Y | Z)$$



Conditional Random Variables

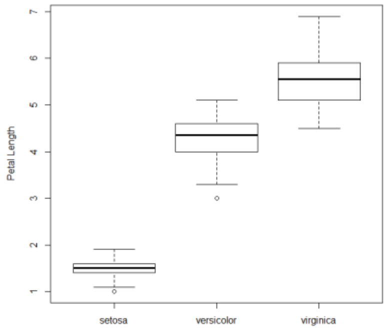




CDF: Cumulative Distribution Function

PDF: Probability Density Function

Quantiles



- Median (aka 2nd quartile)
  - 50th percentile: at least 50% of the values are less than or equal to the median; and at least 50% of the values are greater than or equal to the median [what happens if all values are the same?]
  - More robust measure of location [compared to mean]
- 1st and 3rd Quartile: 25th and 75th percentiles respectively
  - InterQuartile Range (IQR): more robust measure of dispersion [comared to standard deviation]
- Quantiles are also useful for confidence intervals
  - The capital “phi” (pronounced “fee”, by me) is commonly used to denote the Gaussian CDF; so the inverse can be used to denote the bounds of a 95% confidence interval for a sample mean

$$(\Phi^{-1}(0.025), \Phi^{-1}(0.975)) = (-1.96, 1.96)$$

Mean and Variance

- Mean [aka expected value]

$$\mathbb{E}[X] \triangleq \sum_{x \in \mathcal{X}} x \ p(x) \qquad \mathbb{E}[X] \triangleq \int_{\mathcal{X}} x \ p(x) dx$$

- Variance

$$\begin{aligned} \text{var}[X] &\triangleq \mathbb{E}\left[(X-\mu)^2\right] = \int (x-\mu)^2 p(x) dx \\ &= \int x^2 p(x) dx + \mu^2 \int p(x) dx - 2\mu \int x p(x) dx = \mathbb{E}\left[X^2\right] - \mu^2 \end{aligned}$$

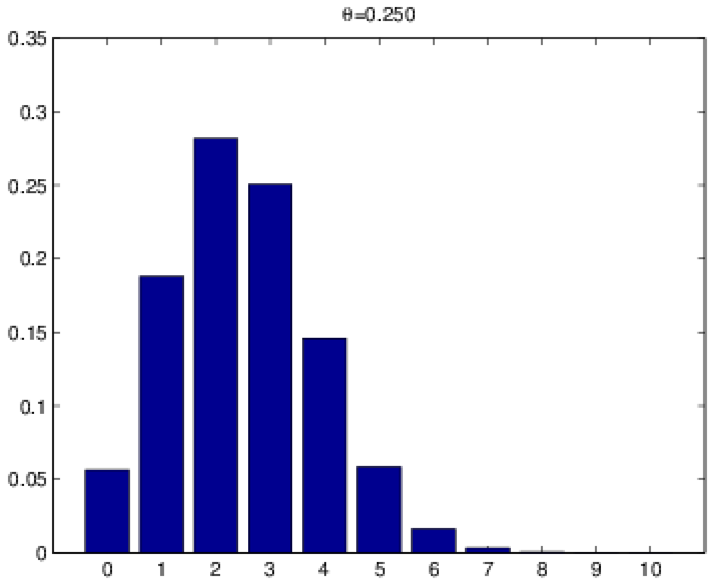
- Variance of a mean

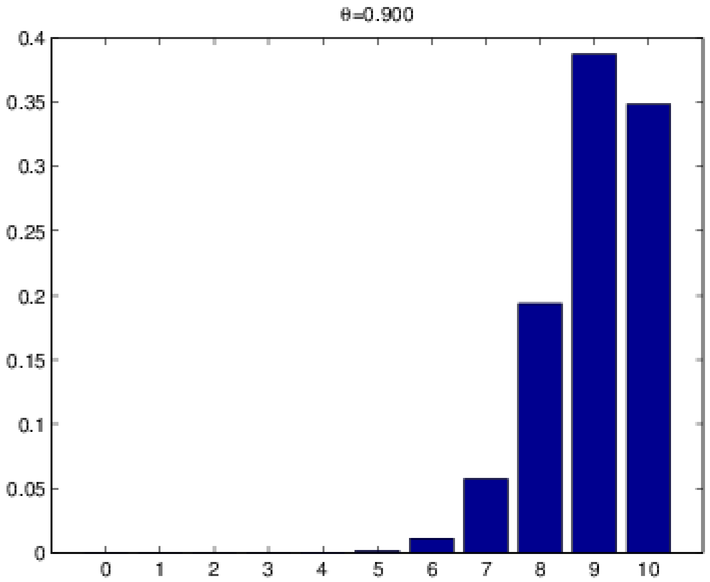
$$var\left[\frac{\sum_{i=1}^{n} x_i}{n}\right] = \frac{var[\sum_{i=1}^{n} x_i]}{n^2} = \frac{n * var[x_i]}{n^2} = \frac{var[x_i]}{n}$$

iid

Binomial and Bernoulli Distributions

Independent trials with two possible outcomes; e.g. flipping a coin





$$\mathrm{Bin}(k|n,\theta) \triangleq \binom{n}{k} \theta^k (1-\theta)^{n-k}$$

$$\binom{n}{k} \triangleq \frac{n!}{(n-k)!k!}$$

$$\mathrm{Ber}(x|\theta) = \theta^{\mathbb{I}(x=1)}(1-\theta)^{\mathbb{I}(x=0)}$$

Multinomial and Multinoulli Distributions

Independent trials with more than two possible outcomes; e.g. rolling a die

$$\mathrm{Mu}(\mathbf{x}|n,\boldsymbol{\theta}) \triangleq \binom{n}{x_1 \dots x_K} \prod_{j=1}^{K} \theta_j^{x_j}$$

$$\binom{n}{x_1 \dots x_K} \triangleq \frac{n!}{x_1! x_2! \cdots x_K!}$$

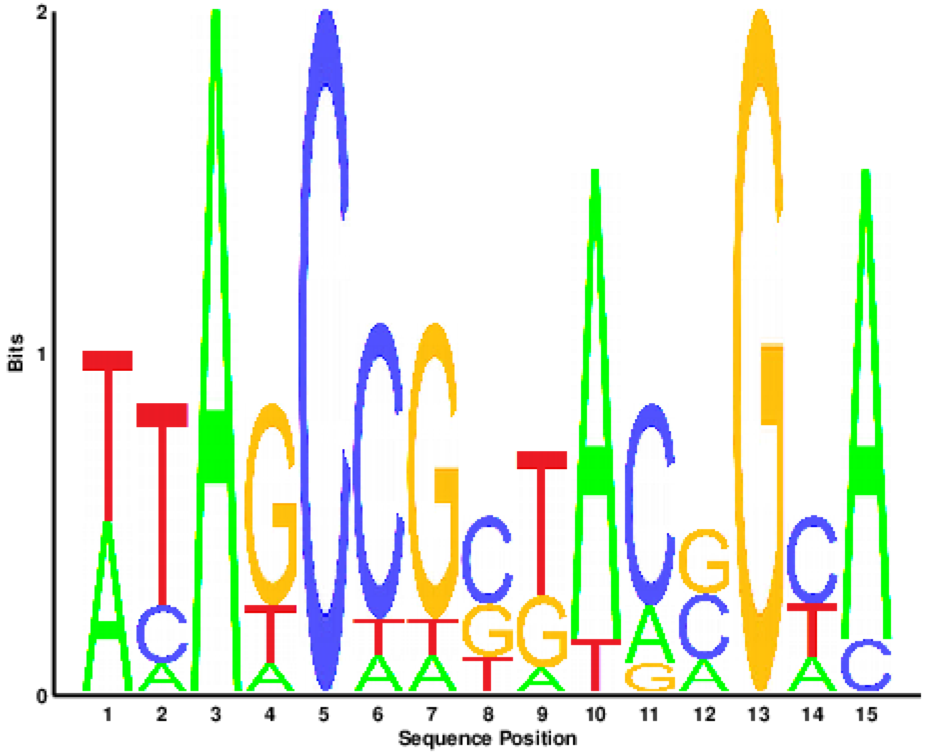
$$\mathrm{Mu}(\mathbf{x}|1,\boldsymbol{\theta}) = \prod_{j=1}^{K} \theta_j^{\mathbb{I}(x_j=1)}$$

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

DNA Sequence Motifs

← sequence →

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



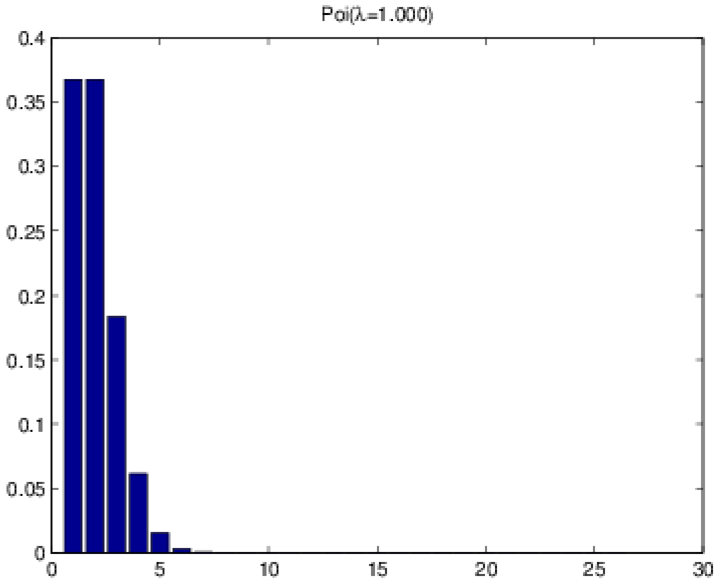
Nucleotides: (a)denine, (c)ytosine, (g)uanine, (t)hymine

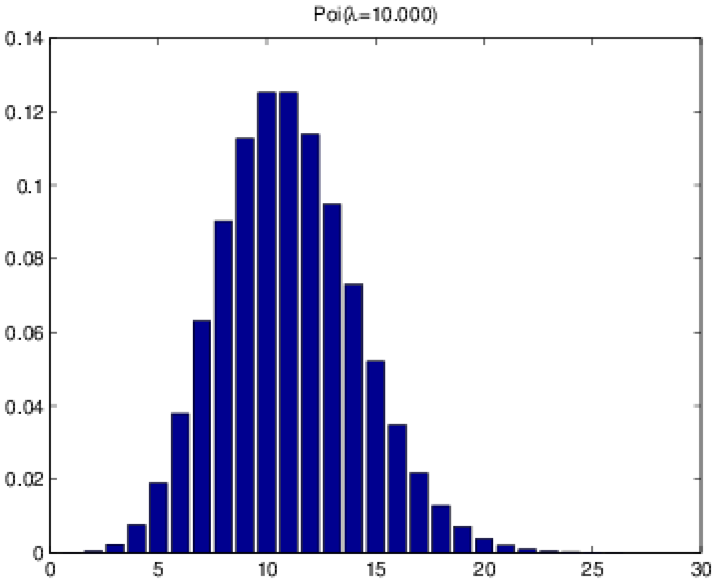
$$\mathbf{N}_t = \left(\sum_{i=1}^{N} \mathbb{I}(X_{it} = 1), \sum_{i=1}^{N} \mathbb{I}(X_{it} = 2), \sum_{i=1}^{N} \mathbb{I}(X_{it} = 3), \sum_{i=1}^{N} \mathbb{I}(X_{it} = 4) \right)$$

$$\hat{\boldsymbol{\theta}}_t = \mathbf{N}_t / N$$

Poisson Distribution

Count of independent events during some time interval; e.g. number of goals during a soccer match





$$\mathrm{Poi}(x|\lambda) = e^{-\lambda} \frac{\lambda^x}{x!}$$

Empirical Distribution

$$p_{\text{emp}}(A) \triangleq \frac{1}{N} \sum_{i=1}^{N} \delta_{x_i}(A)$$

$$\delta_x(A) = \begin{cases} 0 & \text{if } x \notin A \\ 1 & \text{if } x \in A \end{cases}$$

$$p(x) = \sum_{i=1}^{N} w_i \delta_{x_i}(x)$$

$$0 \leq w_i \leq 1 \text{ and } \textstyle\sum_{i=1}^{N} w_i = 1$$

Gaussian, Student, and Laplace Distribution

- Gaussian

$$\mathcal{N}(x|\mu,\sigma^2) \triangleq \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}(x-\mu)^2}$$

$$\Phi(x;\mu,\sigma^2) \triangleq \int_{-\infty}^{x} \mathcal{N}(z|\mu,\sigma^2)dz$$

$$z = (x-\mu)/\sigma$$

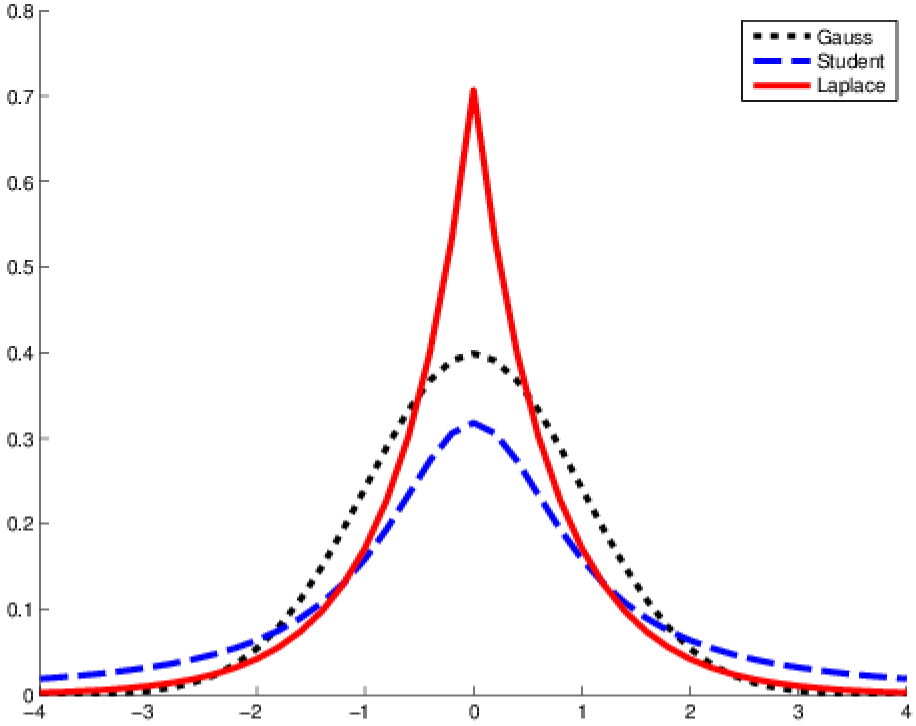
- Student's t

$$\mathcal{T}(x|\mu,\sigma^2,\nu) \propto \left[1+\frac{1}{\nu}\left(\frac{x-\mu}{\sigma}\right)^2\right]^{-\left(\frac{\nu+1}{2}\right)}$$

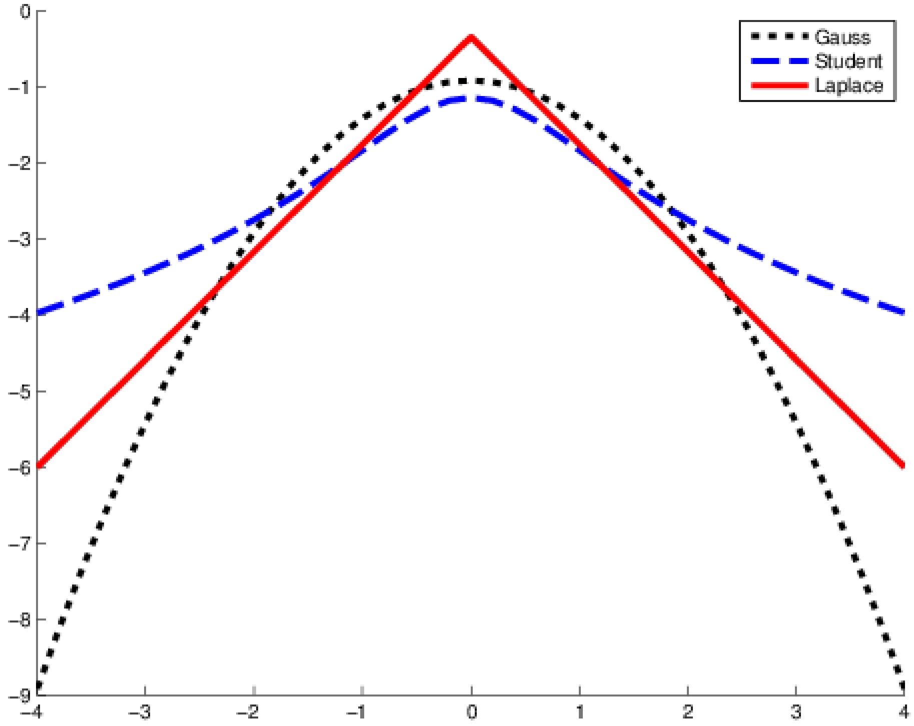
- Laplace

$$\mathrm{Lap}(x|\mu,b) \triangleq \frac{1}{2b}\exp\left(-\frac{|x-\mu|}{b}\right)$$

Gaussian, Student, and Laplace Distribution

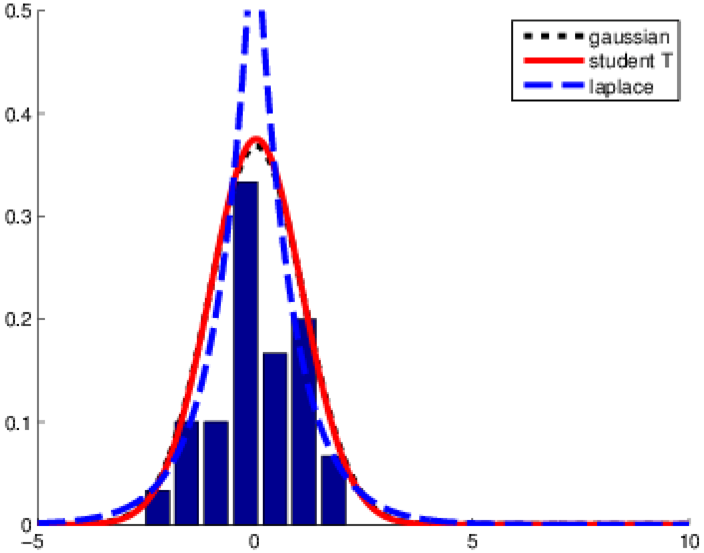


$\mathcal{N}(0,1)$, $\mathcal{T}(0,1,1)$ and $\mathrm{Lap}(0,1/\sqrt{2})$

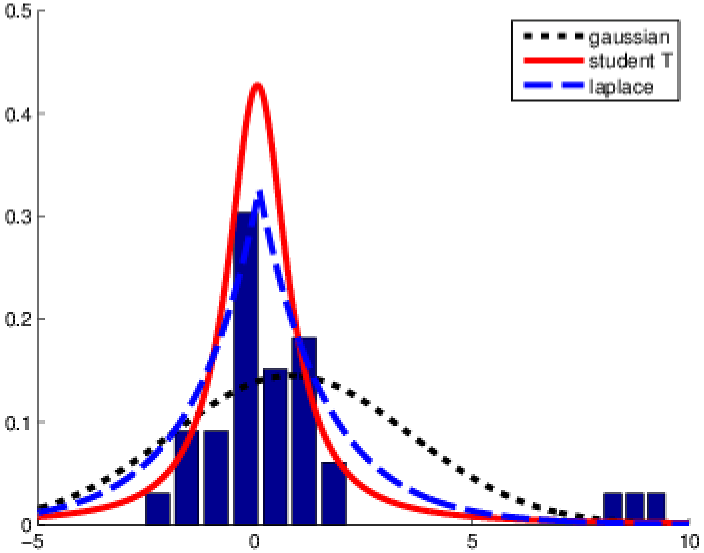


log(X)

Effect of Outliers

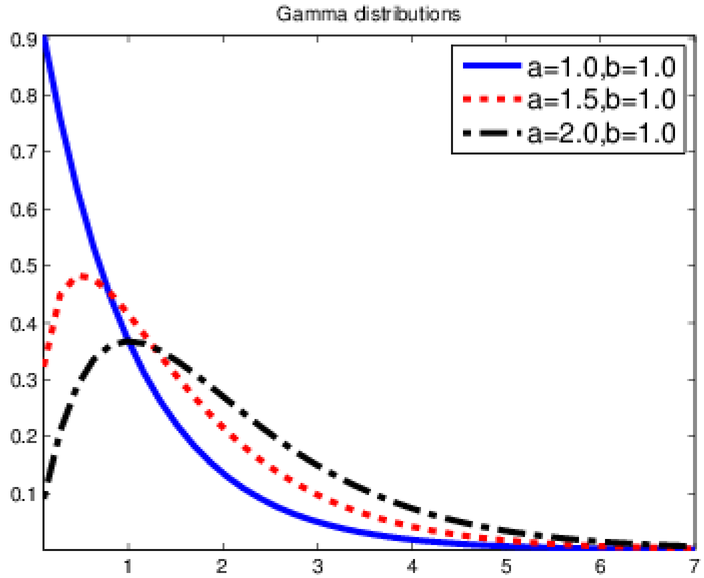


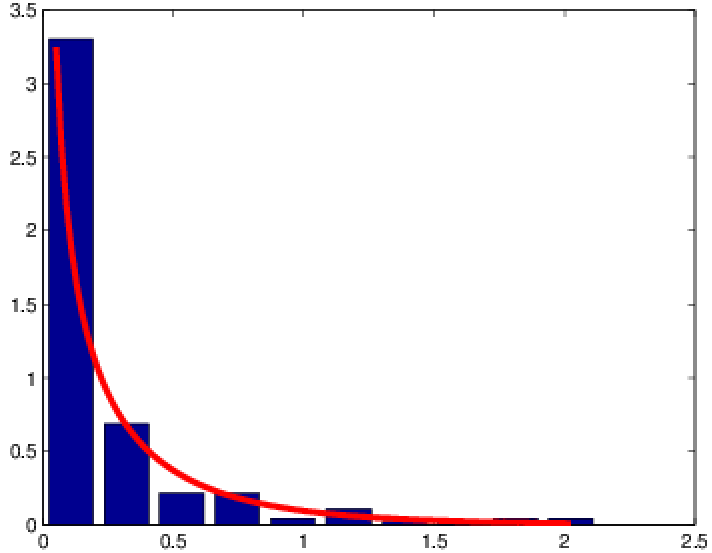
without outliers



with outliers …
Gaussian: location affected
Student t and Laplace: more robust

Gamma Distribution





$$\mathrm{Ga}(T|\text{shape} = a, \text{rate} = b) \triangleq \frac{b^a}{\Gamma(a)} T^{a-1} e^{-Tb}$$

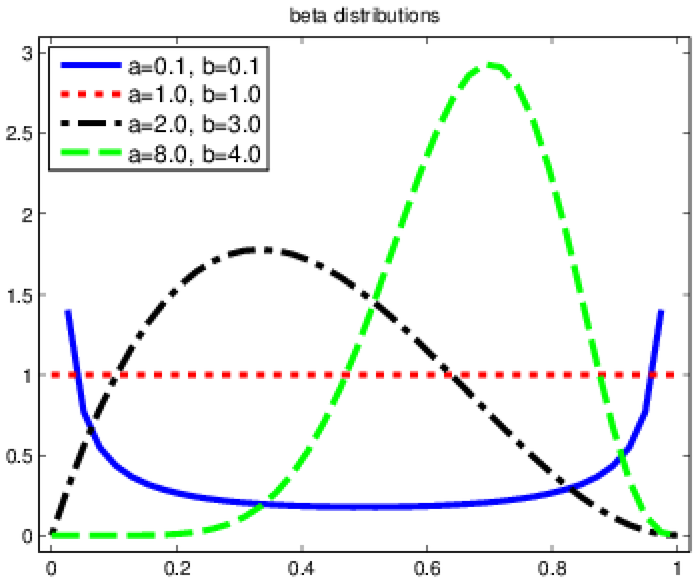
rainfall ~ Ga(0.44, 1.97)

$$\mathrm{Expon}(x|\lambda) \triangleq \mathrm{Ga}(x|1, \lambda)$$

$$\mathrm{Erlang}(x|\lambda) = \mathrm{Ga}(x|2, \lambda)$$

$$\chi^2(x|\nu) \triangleq \mathrm{Ga}(x|\tfrac{\nu}{2}, \tfrac{1}{2})$$

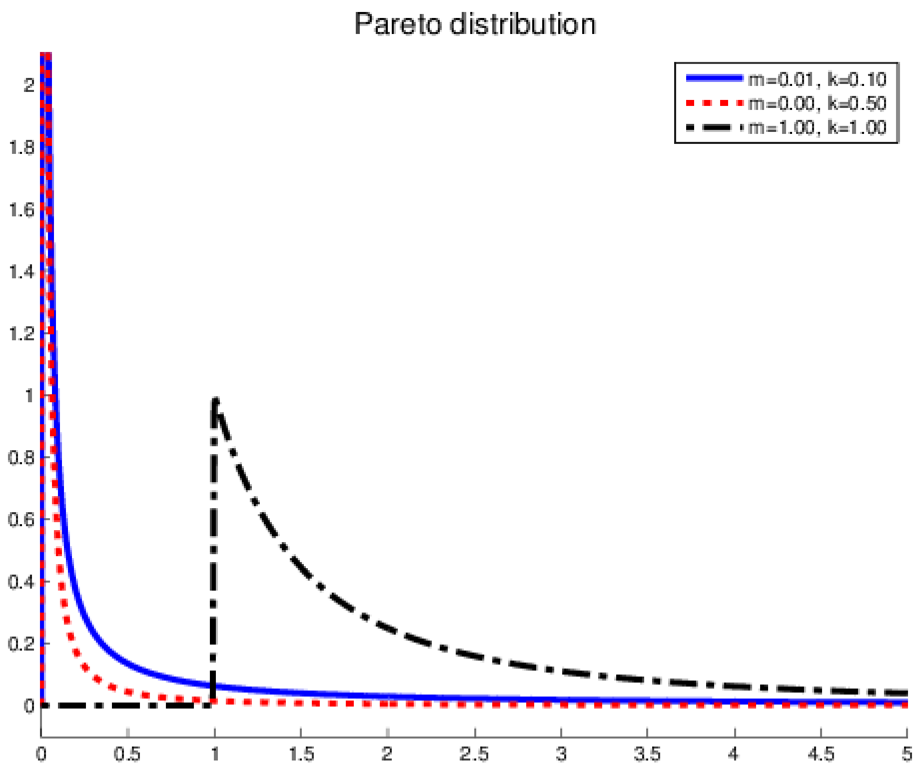
Beta Distribution

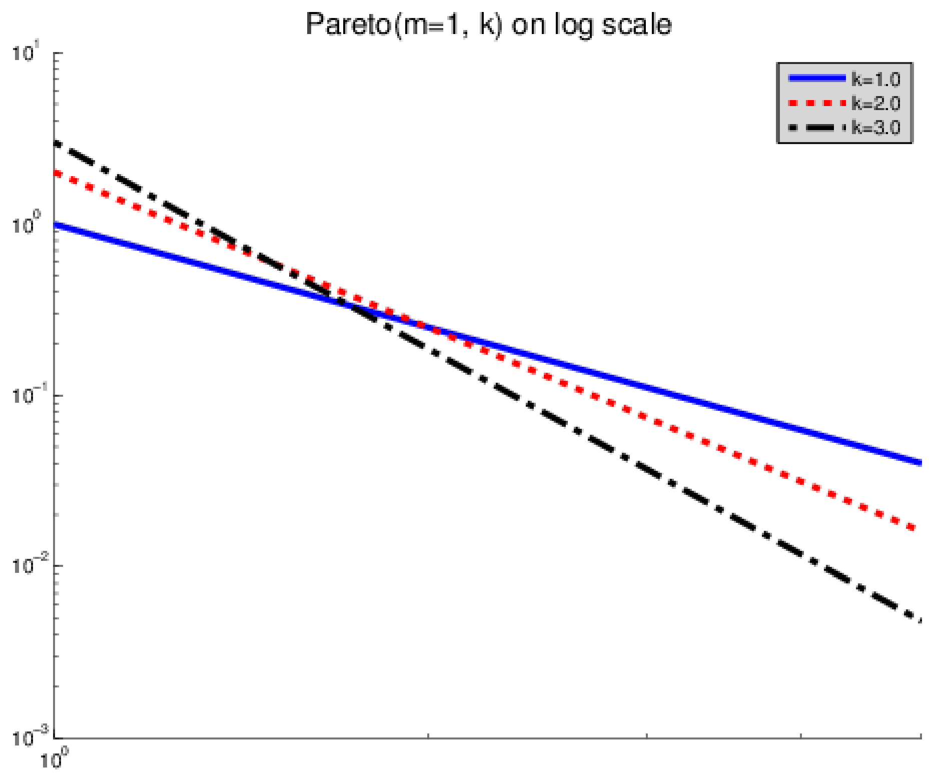


$$\mathrm{Beta}(x|a,b) = \frac{1}{B(a,b)} x^{a-1}(1-x)^{b-1} \qquad B(a,b) \triangleq \frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)}$$

Used to model the probability parameter for a Bernoulli trial:
* ‘a’ and ‘b’ can be used as weights for successful and unsuccessful outcomes
* larger weights yield a more concentrated distribution

Pareto (Power Law) Distribution





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

$$\log p(x) = a \log x + k$$

long, heavy tail

Degenerate PDF

- It's a constant [not actually a random variable]

$$\lim_{\sigma^2 \to 0} \mathcal{N}(x|\mu, \sigma^2) = \delta(x - \mu)$$

- Dirac delta function

$$\delta(x) = \begin{cases} \infty & \text{if } x = 0 \\ 0 & \text{if } x \neq 0 \end{cases}$$

Covariance

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

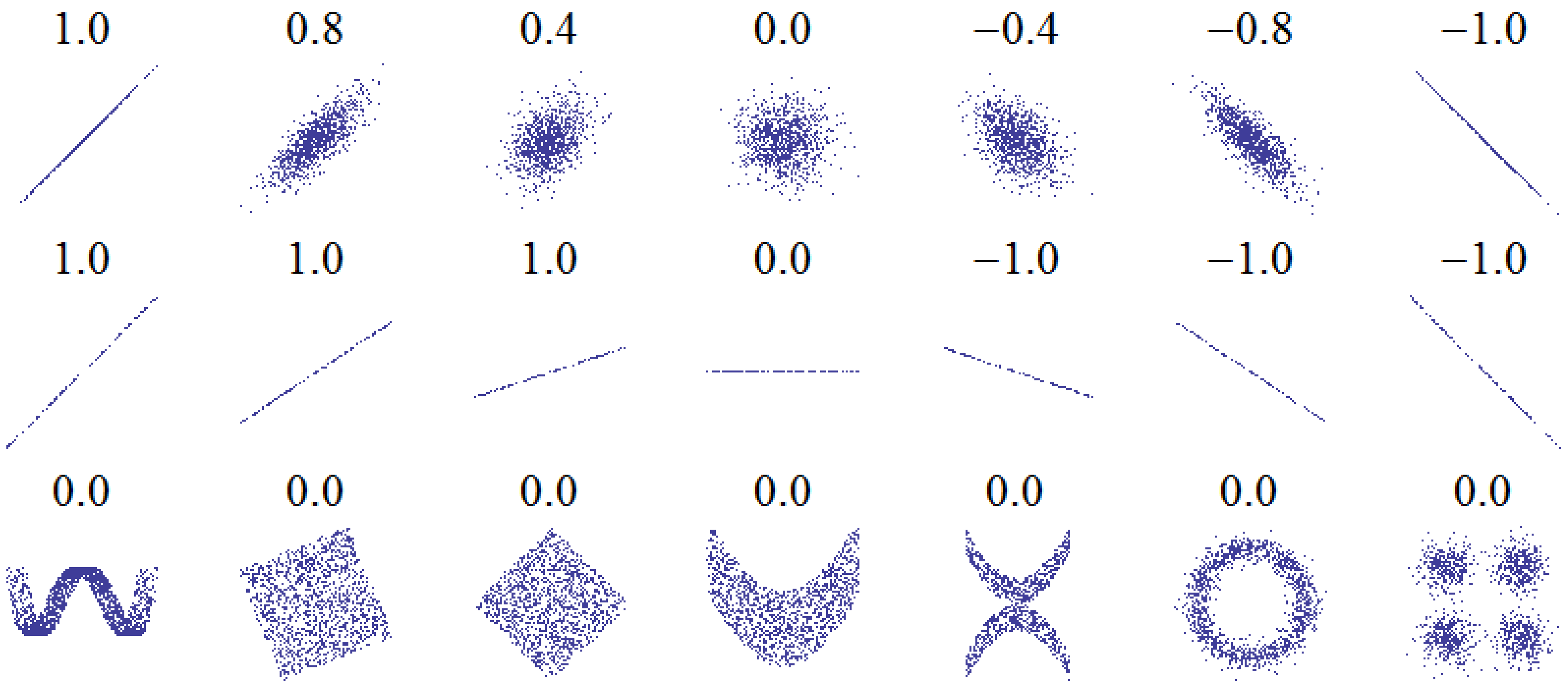
$$\begin{aligned}
\text{cov}[\mathbf{x}] &\triangleq \mathbb{E}\left[(\mathbf{x} - \mathbb{E}[\mathbf{x}])(\mathbf{x} - \mathbb{E}[\mathbf{x}])^T\right] \\
&= \begin{pmatrix}
\text{var}[X_1] & \text{cov}[X_1, X_2] & \cdots & \text{cov}[X_1, X_d] \\
\text{cov}[X_2, X_1] & \text{var}[X_2] & \cdots & \text{cov}[X_2, X_d] \\
\vdots & \vdots & \ddots & \vdots \\
\text{cov}[X_d, X_1] & \text{cov}[X_d, X_2] & \cdots & \text{var}[X_d]
\end{pmatrix}
\end{aligned}$$

Correlation

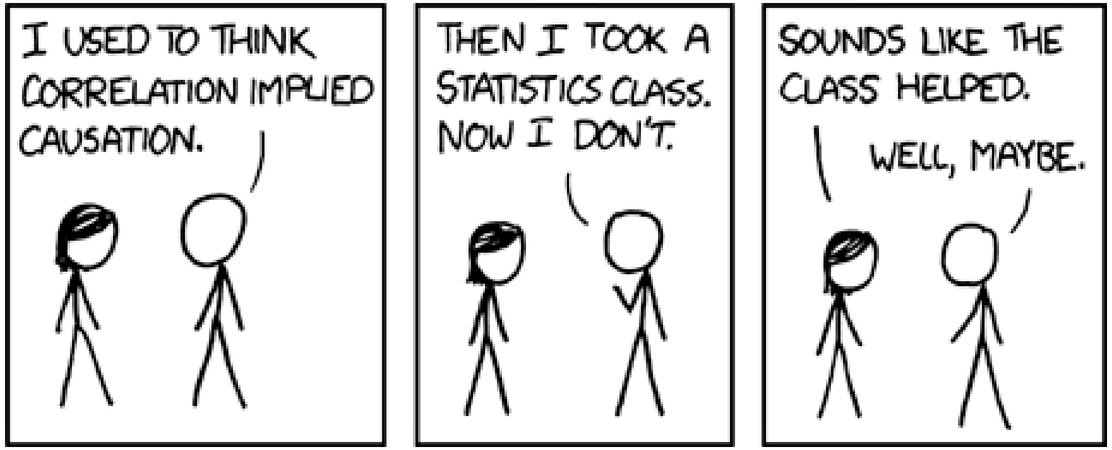
$$\mathrm{corr}\left[X,Y\right] \triangleq \frac{\mathrm{cov}\left[X,Y\right]}{\sqrt{\mathrm{var}\left[X\right]\mathrm{var}\left[Y\right]}}$$

$$\mathbf{R} = \begin{pmatrix} \mathrm{corr}\left[X_1,X_1\right] & \mathrm{corr}\left[X_1,X_2\right] & \cdots & \mathrm{corr}\left[X_1,X_d\right] \\ \vdots & \vdots & \ddots & \vdots \\ \mathrm{corr}\left[X_d,X_1\right] & \mathrm{corr}\left[X_d,X_2\right] & \cdots & \mathrm{corr}\left[X_d,X_d\right] \end{pmatrix}$$

Correlation Examples



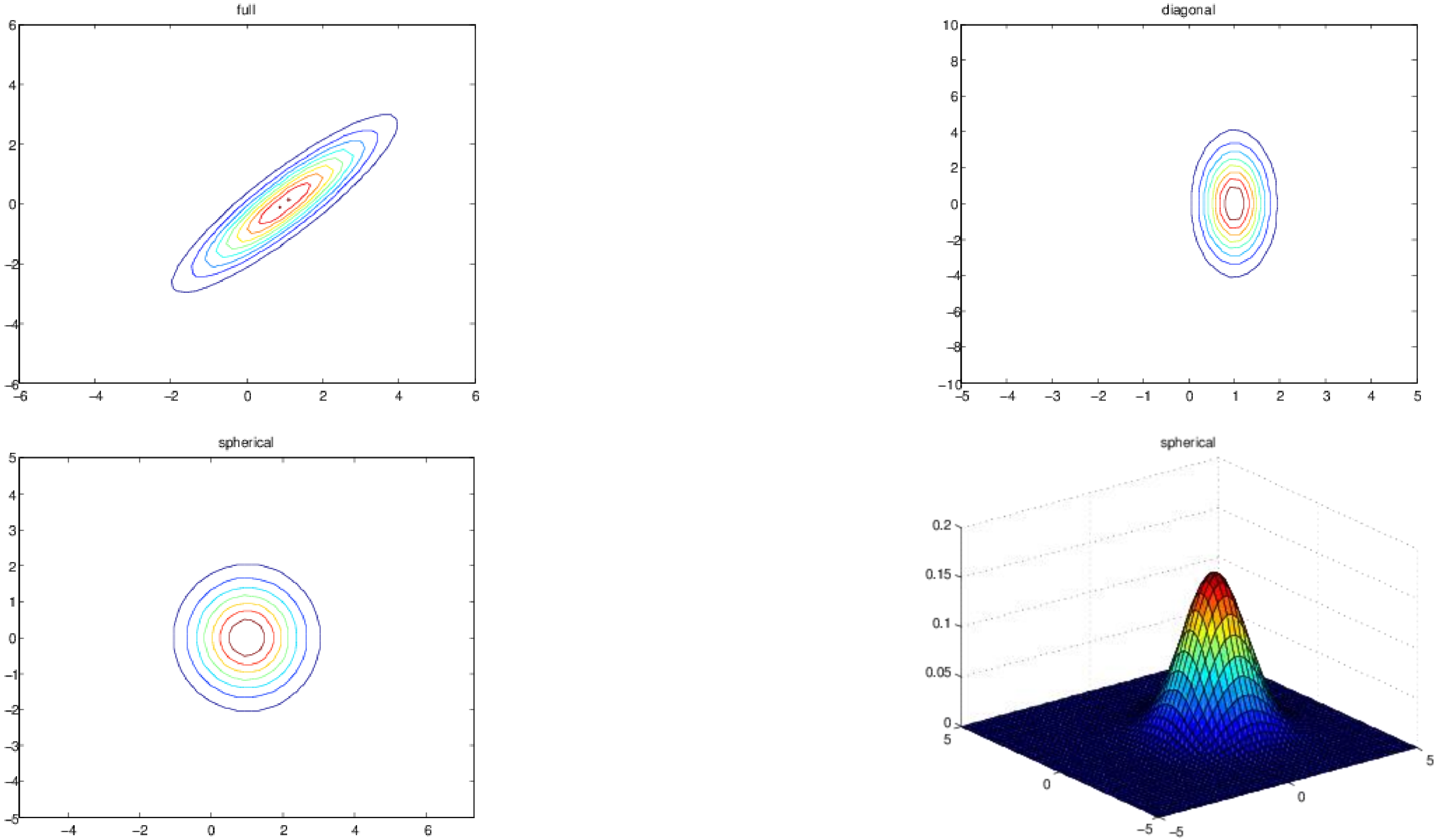
Correlation Doesn't Imply Causation!



"Correlation doesn't imply causation, but it does waggle its eyebrows suggestively and gesture furtively while mouthing 'look over there'."

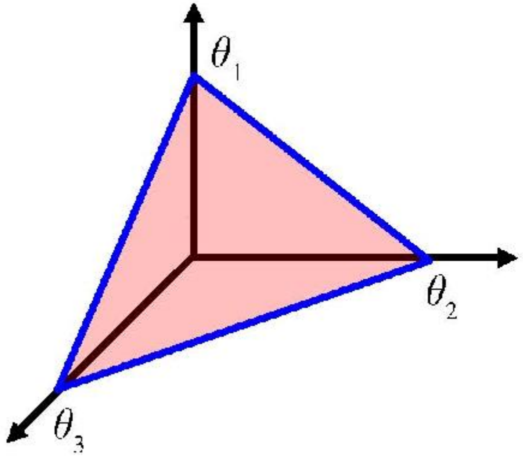
http://xkcd.com/552/

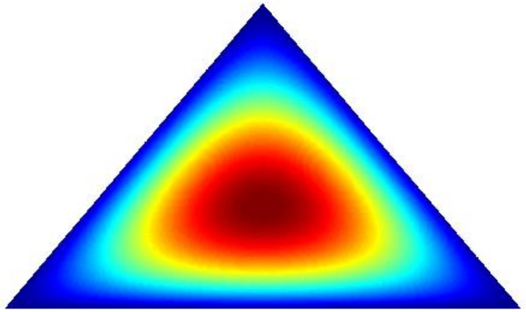
Multi-variate Gaussian



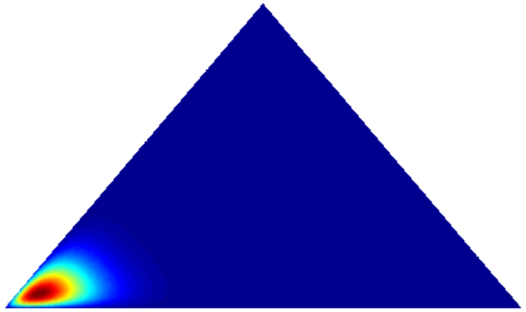
$$\mathcal{N}(\mathbf{x}|\boldsymbol{\mu},\boldsymbol{\Sigma}) \triangleq \frac{1}{(2\pi)^{D/2}|\boldsymbol{\Sigma}|^{1/2}} \exp\left[-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right]$$

Dirichlet Distribution [3 outcomes]

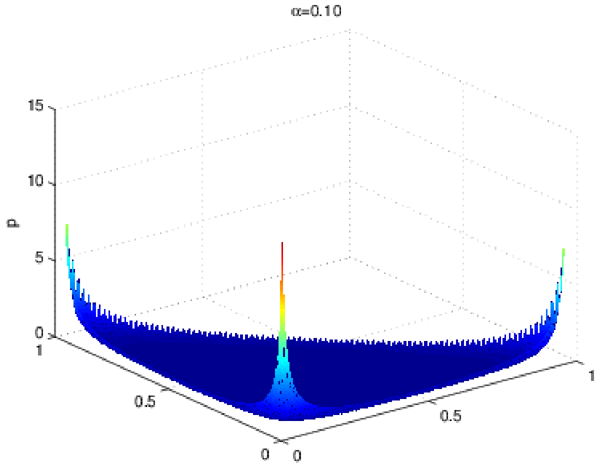




$\boldsymbol{\alpha} = (2, 2, 2)$

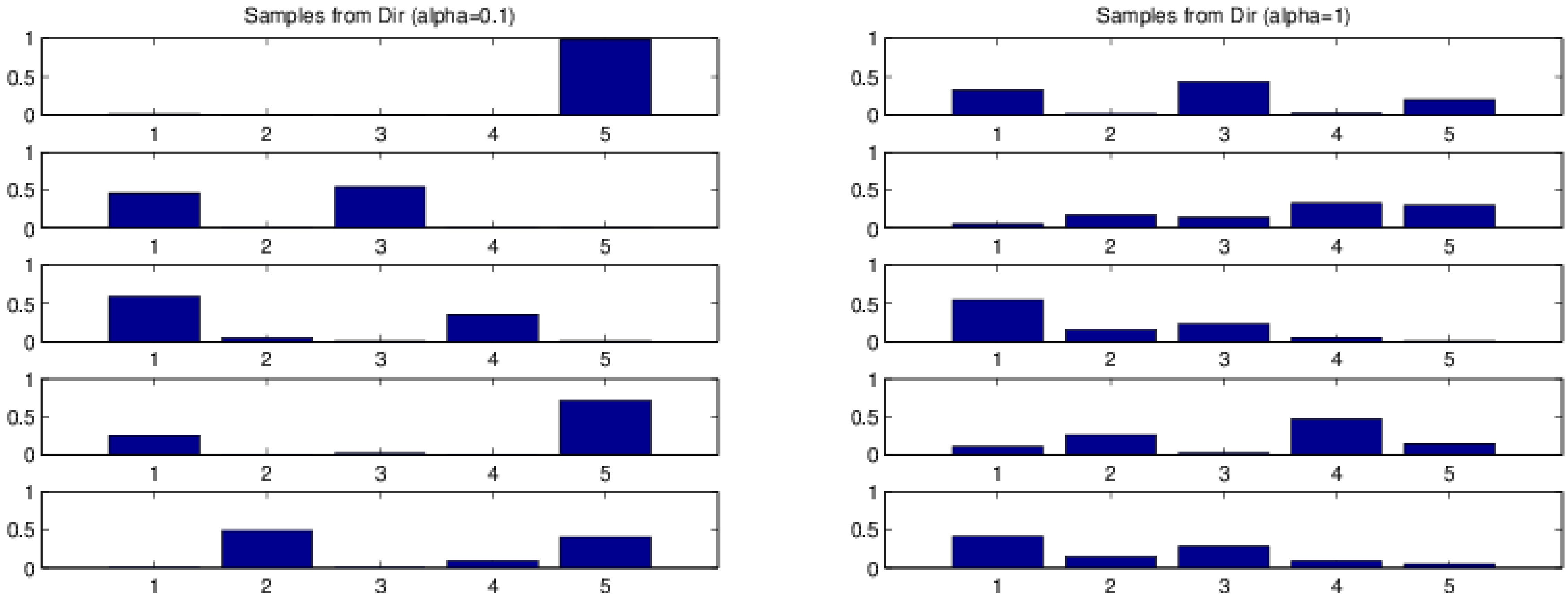


$\boldsymbol{\alpha} = (20, 2, 2)$



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

Samples from Dirichlet Distribution



5 outcomes; “alpha” assigned to all 5 parameters

Transformation: Univariate Change of Variable

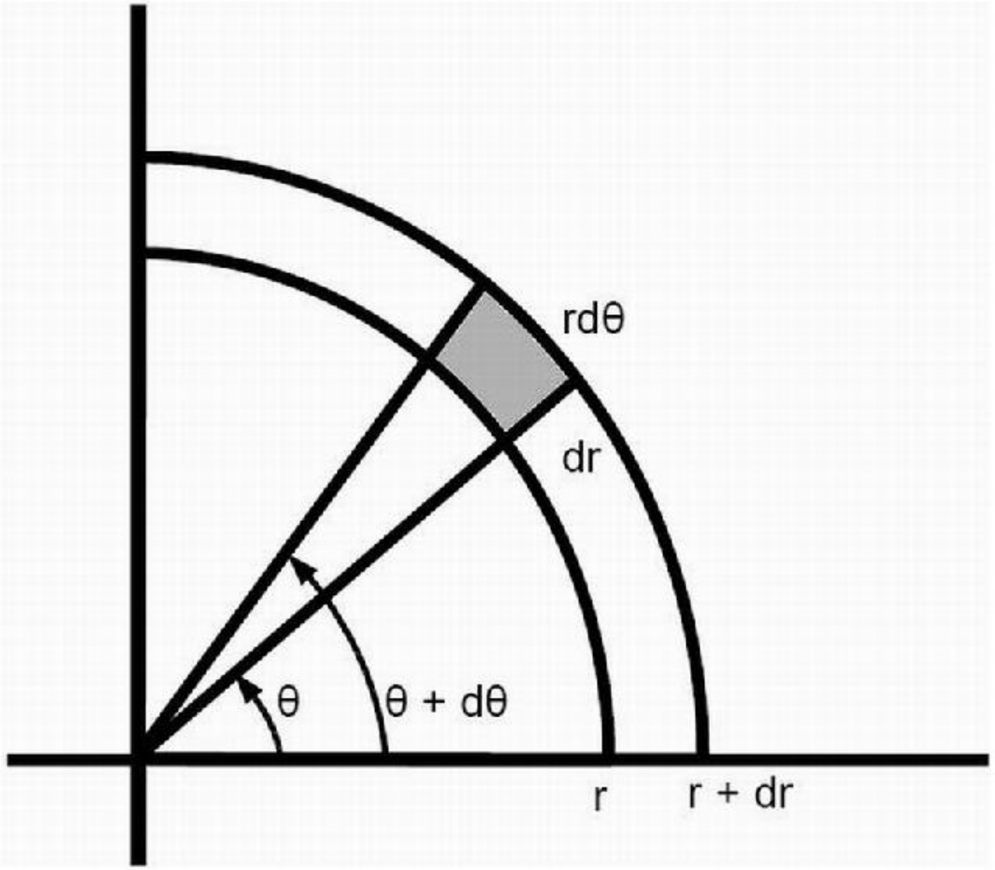
- Density function for transformed variable:

$$p_y(y) = p_x(x)\left|\frac{dx}{dy}\right|$$

- Example:

$$Y = X^2 \qquad X \sim U(-1,1) \qquad p_y(y) = \tfrac{1}{2}y^{-\frac{1}{2}}$$

Transformation: Multivariate Change of Variables

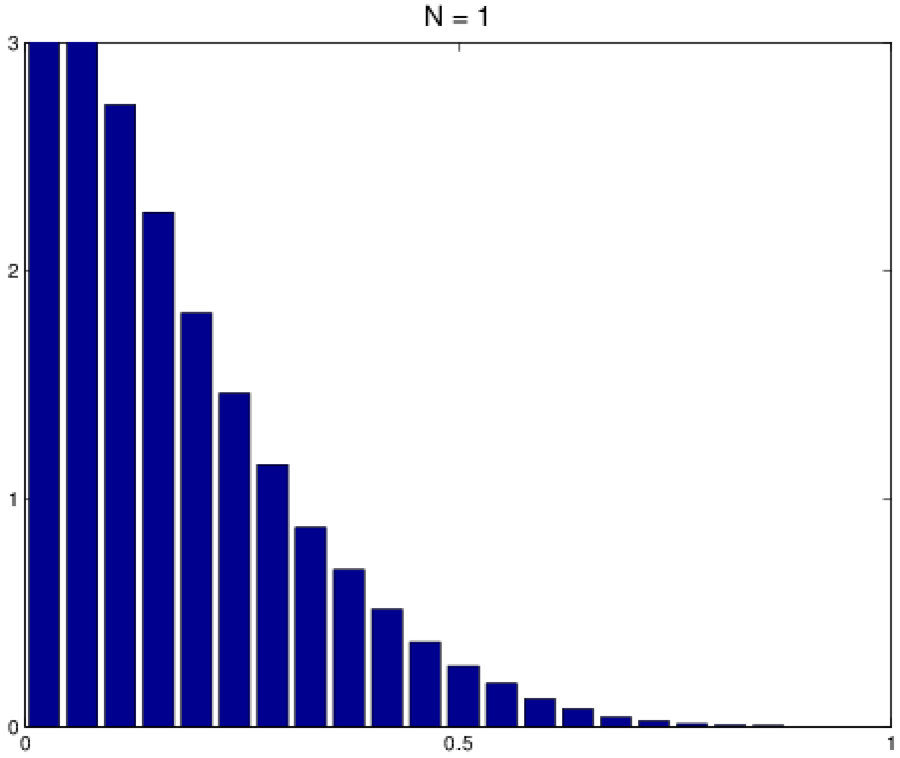


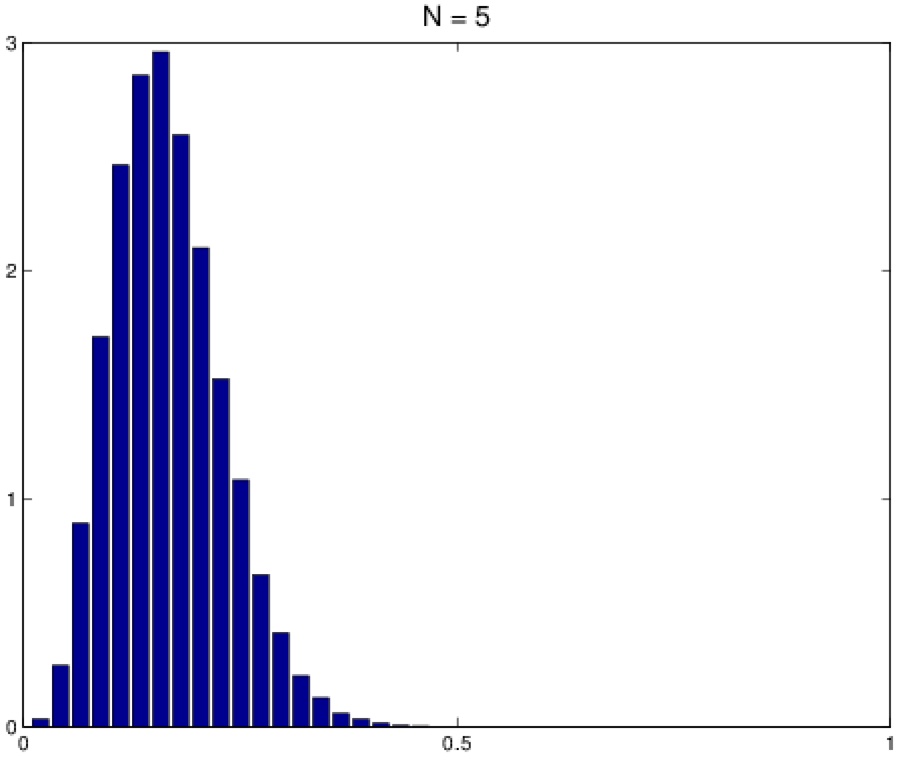
Transformation of x,y to r,θ
area of the patch is r * $\mathrm{d}\theta$ * dr, where r * $\mathrm{d}\theta$ is the length of the arc

$$p_{r,\theta}(r,\theta)drd\theta \quad = \quad p_{x_1,x_2}(r\cos\theta, r\sin\theta)r\ dr\ d\theta$$

Central Limit Theorem

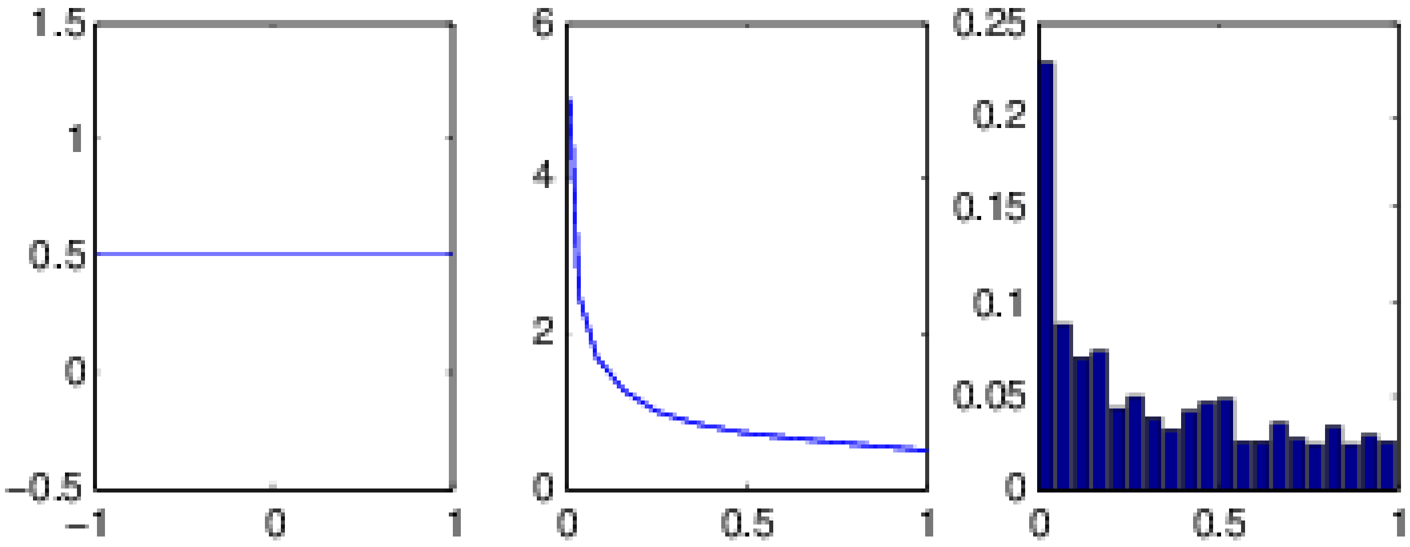
if $S_N = \sum_{i=1}^{N} X_i$ then $Z_N \triangleq \frac{S_N - N\mu}{\sigma\sqrt{N}} = \frac{\overline{X} - \mu}{\sigma/\sqrt{N}}$ converges to Gaussian(0, 1) as n goes to infinity





The sampling distribution of the mean value rapidly converges to a Gaussian distribution

Monte Carlo Integration



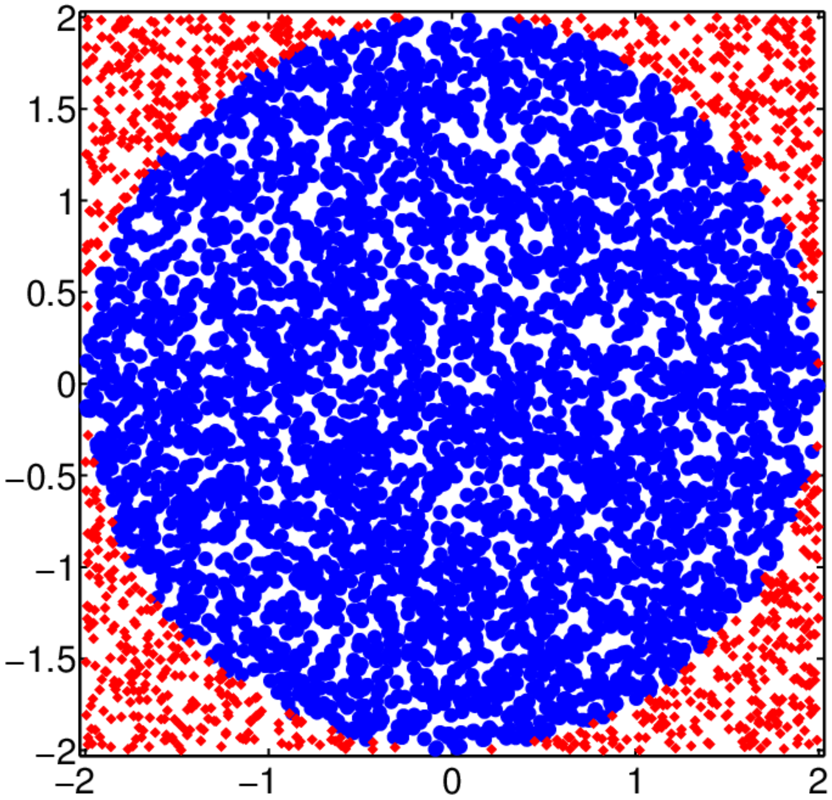
$$\int \frac{1}{2\sqrt{y}} dy = \sqrt{y}$$

http://integrals.wolfram.com/

Example: Instead integrating the density function, we can generate random samples of the transformed variable …

$$\frac{1}{S}|\{x_s \leq c\}| \rightarrow P(X \leq c)$$

Monte Carlo Approximation of π



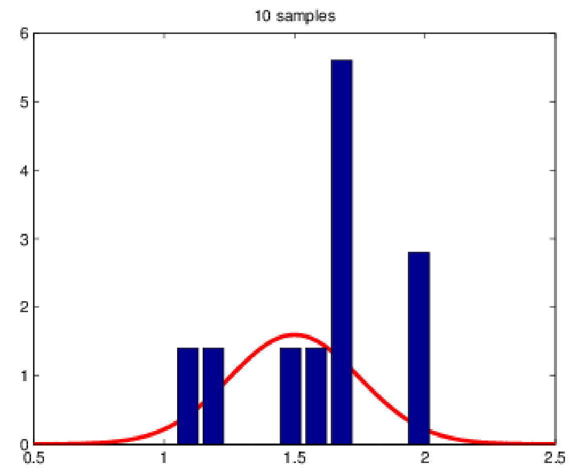
$$I = \int_{-r}^{r}\int_{-r}^{r} \mathbb{I}(x^2 + y^2 \le r^2)dxdy$$

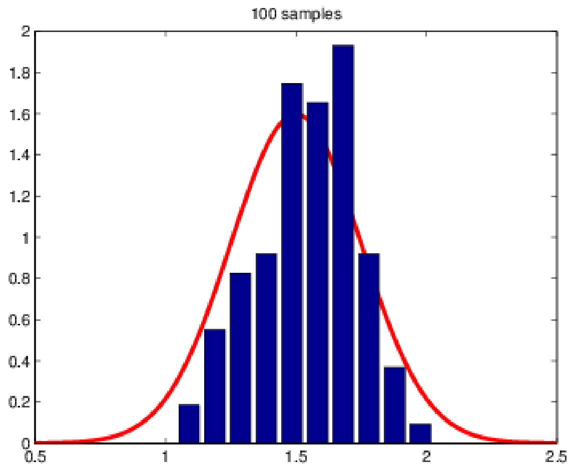
$$\begin{aligned} I &= (2r)(2r)\int\int f(x,y)p(x)p(y)dxdy \\ &= 4r^2\int\int f(x,y)p(x)p(y)dxdy \\ &\approx 4r^2\frac{1}{S}\sum_{s=1}^{S} f(x_s, y_s) \\ \pi &= I/(r^2) \end{aligned}$$

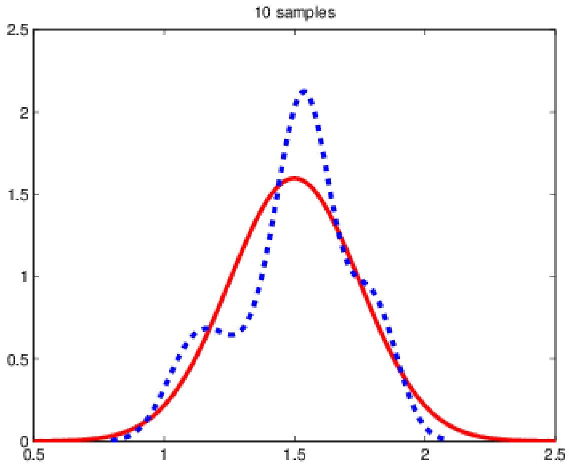
$$\hat{\sigma}^2 = \frac{1}{S}\sum_{s=1}^{S}(f(x_s) - \hat{\mu})^2$$

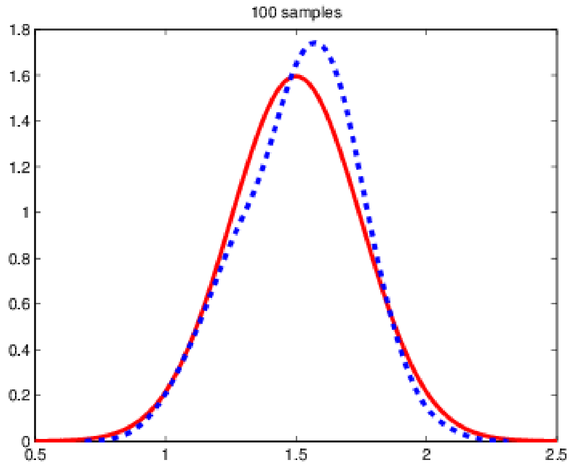
$$P\left\{\mu - 1.96\frac{\hat{\sigma}}{\sqrt{S}} \le \hat{\mu} \le \mu + 1.96\frac{\hat{\sigma}}{\sqrt{S}}\right\} \approx 0.95$$

Monte Carlo Approximation for Gaussian(1.5, .25)

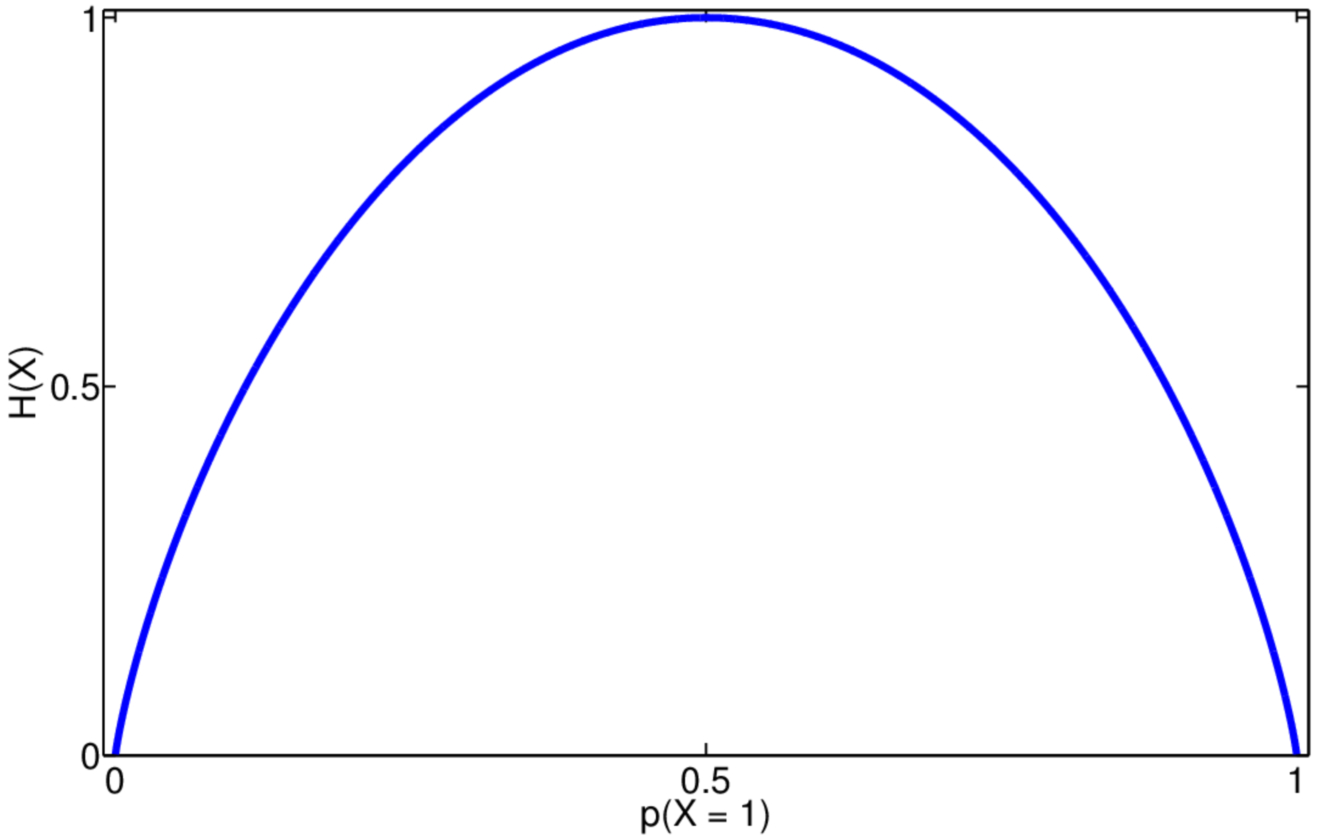








Entropy



$$\mathbb{H}(X) \triangleq -\sum_{k=1}^{K} p(X = k) \log_2 p(X = k)$$

Kullback-Leibler (KL) Divergence

- Defined as the average number of bits needed to encode data, caused by using distribution "q" to encode the data rather than distribution "p"

$$\mathbb{KL}\left(p||q\right) = \sum_k p_k \log p_k - \sum_k p_k \log q_k = -\mathbb{H}\left(p\right) + \mathbb{H}\left(p, q\right)$$

- The second term is known as cross entropy: measuring the dissimilarity of the two distributions

Mutual Information

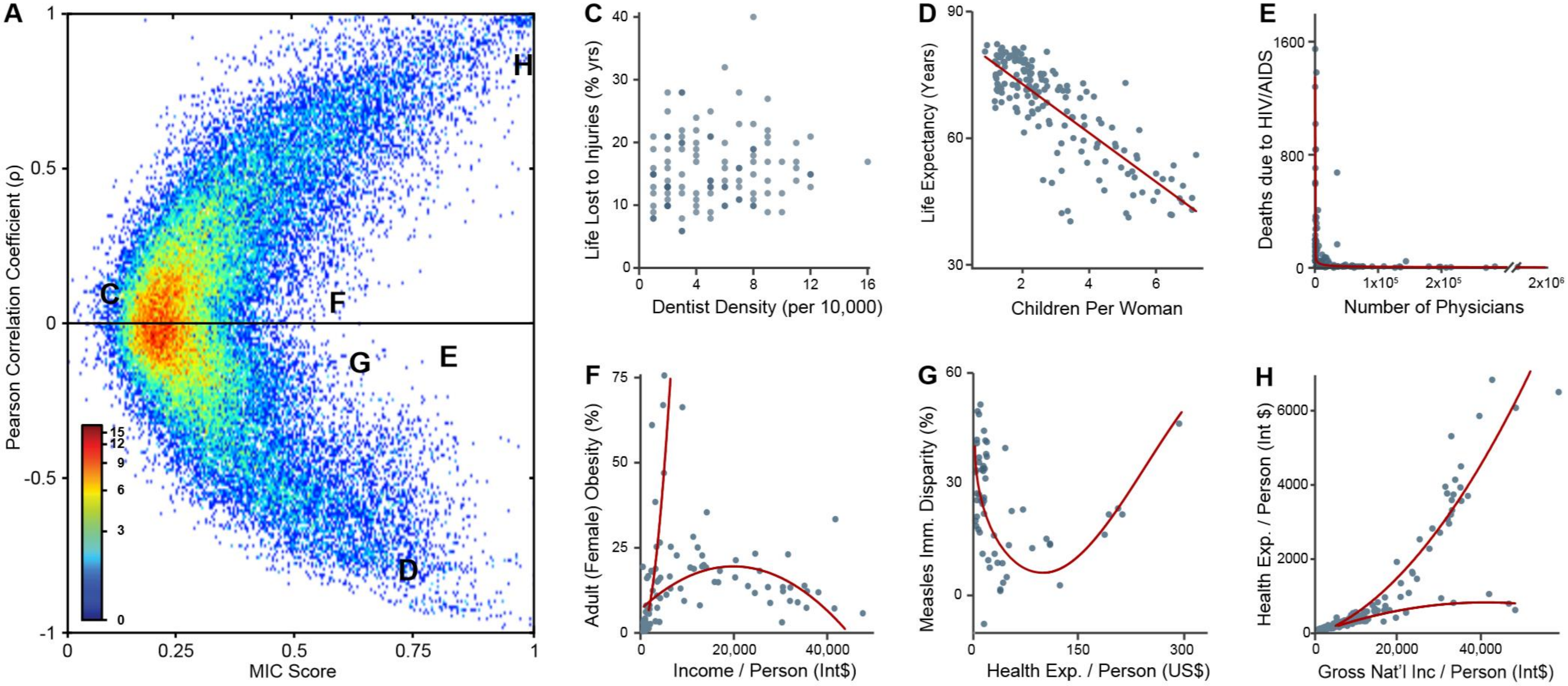
- Measures the strength of the relationship between variables
  - Expected value of the ratio of the joint probability to the product of priors

$$\mathbb{I}(X;Y) \triangleq \mathbb{KL}\left(p(X,Y)||p(X)p(Y)\right) = \sum_x \sum_y p(x,y) \log \frac{p(x,y)}{p(x)p(y)}$$

$$\mathbb{I}(X;Y) \geq 0$$

- Mutual information can capture relationships missed by Pearson's correlation coefficient

Maximal Information Coefficient (MIC)



Spectral Clustering

Spectral Clustering References

- Section 25.4 of our text
  - PMTK demo: http://pmtk3.googlecode.com/svn/trunk/docs/demoOutput/bookDemos/(25)-Clustering/spectralClusteringDemo.html
- Ng, Jordan, Weiss NIPS 2001 paper
  - http://ai.stanford.edu/~ang/papers/nips01-spectral.pdf
- Scikit-learn documentation
  - http://scikit-learn.org/stable/modules/generated/sklearn.cluster.SpectralClustering.html#sklearn.cluster.SpectralClustering

Spectral Clustering Algorithm

Given a set of points $S = \{s_1, \ldots, s_n\}$ in $\mathbb{R}^l$ that we want to cluster into k subsets:

1. Form the affinity matrix $A \in \mathbb{R}^{n \times n}$ defined by $A_{ij} = \exp(-||s_i - s_j||^2/2\sigma^2)$ if $i \neq j$, and $A_{ii} = 0$.
2. Define D to be the diagonal matrix whose (i,i)-element is the sum of A's i-th row, and construct the matrix $L = D^{-1/2}AD^{-1/2}$.[1]
3. Find $x_1, x_2, \ldots, x_k$, the k largest eigenvectors of L (chosen to be orthogonal to each other in the case of repeated eigenvalues), and form the matrix $X = [x_1 x_2 \ldots x_k] \in \mathbb{R}^{n \times k}$ by stacking the eigenvectors in columns.
4. Form the matrix Y from X by renormalizing each of X's rows to have unit length (i.e. $Y_{ij} = X_{ij}/(\sum_j X_{ij}^2)^{1/2}$).
5. Treating each row of Y as a point in $\mathbb{R}^k$, cluster them into k clusters via K-means or any other algorithm (that attempts to minimize distortion).
6. Finally, assign the original point s_i to cluster j if and only if row i of the matrix Y was assigned to cluster j.

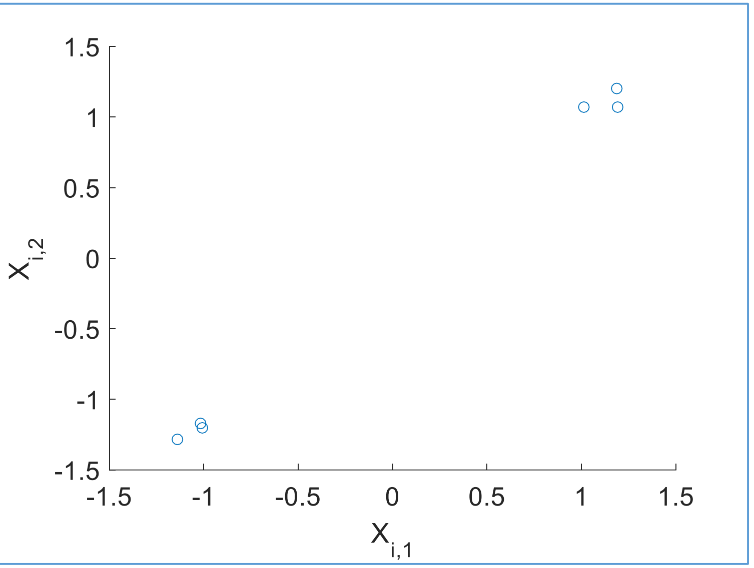
Compared to Other Methods

- Primary Advantage
  - Can be used to generate clusters with arbitrary shapes
- Primary Disadvantage
  - Cannot generalize to unseen data
- Focus is on spectral representation
  - K-means is often used for clustering the spectral representation; but other clustering methods can be used

Matlab Code for Spectral Representation

```
sigma = 0.1;
num_clusters = 2;
S1S2 = -2 * data * data';
SS = sum(data.^2,2);
A = exp(- (S1S2+repmat(SS,1,length(SS))+repmat(SS',length(SS),1)) / (2*sigma^2));
D = diag(1 ./ sqrt(sum(A, 2)));
L = D * A * D;
[X, D] = eigs(L, num_clusters);
Y = X ./ repmat(sqrt(sum(X.^2, 2)), 1, num_clusters);
```

Spectral Representation for 6 Observations




```
>> A
A =
    1.0000    0.2132    0.4195    0.0000    0.0000    0.0000
    0.2132    1.0000    0.0933    0.0000    0.0000    0.0000
    0.4195    0.0933    1.0000    0.0000    0.0000    0.0000
    0.0000    0.0000    0.0000    1.0000    0.2516    0.3107
    0.0000    0.0000    0.0000    0.2516    1.0000    0.9560
    0.0000    0.0000    0.0000    0.3107    0.9560    1.0000
>> D
D =
    0.7826         0         0         0         0         0
         0    0.8748         0         0         0         0
         0         0    0.8130         0         0         0
         0         0         0    0.8000         0         0
         0         0         0         0    0.6730         0
         0         0         0         0         0    0.6642
>> L
L =
    0.6125    0.1460    0.2669    0.0000    0.0000    0.0000
    0.1460    0.7654    0.0664    0.0000    0.0000    0.0000
    0.2669    0.0664    0.6610    0.0000    0.0000    0.0000
    0.0000    0.0000    0.0000    0.6401    0.1355    0.1651
    0.0000    0.0000    0.0000    0.1355    0.4530    0.4274
    0.0000    0.0000    0.0000    0.1651    0.4274    0.4412
```

```
>> X
X =
   -0.6019    0.0336
   -0.5385    0.0300
   -0.5794    0.0323
    0.0557   -0.5080
    0.0662   -0.6038
    0.0671   -0.6118
>> Y
Y =
   -0.9985    0.0557
   -0.9985    0.0557
   -0.9985    0.0557
    0.1090   -0.9940
    0.1090   -0.9940
    0.1090   -0.9940
```