Mathematical Tools
for Neural and Cognitive Science

Fall semester, 2019

Section 4:
Summary Statistics & Probability

Statistics is the science of learning from experience, especially experience that arrives a little bit at a time. The earliest information science was statistics, originating in about 1650. This century has seen statistical techniques become the analytic methods of choice in biomedical science, psychology, education, economics, communications theory, sociology, genetic studies, epidemiology, and other areas. Recently, traditional sciences like geology, physics, and astronomy have begun to make increasing use of statistical methods as they focus on areas that demand informational efficiency, such as the study of rare and exotic particles or extremely distant galaxies.

Most people are not natural-born statisticians. Left to our own devices we are not very good at picking out patterns from a sea of noisy data. To put it another way, we are all too good at picking out non-existent patterns that happen to suit our purposes. Statistical theory attacks the problem from both ends. It provides optimal methods for finding a real signal in a noisy background, and also provides strict checks against the overinterpretation of random patterns.

[Efron & Tibshirani, 1998]

Historical context

- 1600’s: Early notions of data summary/averaging
- 1700’s: Bayesian prob/statistics (Bayes, Laplace)
- 1920’s: Frequentist statistics for science (e.g., Fisher)
- 1940’s: Statistical signal analysis and communication, estimation/decision theory (e.g., Shannon, Wiener, etc)
- 1950’s: Return of Bayesian statistics (e.g., Jeffreys, Wald, Savage, Jaynes…)
- 1970’s: Computation, optimization, simulation (e.g,. Tukey)
- 1990’s: Machine learning (large-scale computing + statistical inference + lots of data)
- Also, since 1950’s: statistical neural/cognitive models!

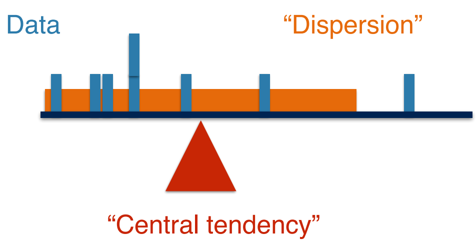
Statistics as summary

0.1, 4.5, -2.3, 0.8, -1.1, 3.2, …

"The purpose of statistics is to replace a quantity of data by relatively few quantities which shall ... contain as much as possible, ideally the whole, of the relevant information contained in the original data"

- R.A. Fisher, 1934

Descriptive statistics



Descriptive statistics: Central tendency

- We often summarize data with averages. Why?
- Average minimizes the squared error (as in regression!):

$$\mu_x = \arg\min_c \frac{1}{N}\sum_{n=1}^{N}(x_n - c)^2 = \frac{1}{N}\sum_{n=1}^{N} x_n = \frac{1}{N}\vec{1}^T\vec{x}$$

- Generalize: minimize L_p norm: $\arg\min_c \left[\frac{1}{N}\sum_{n=1}^{N}|x_n - c|^p\right]^{1/p}$
  - $p = 1$: median, m_x
  - $p \to 0$: mode (location of maximum)
  - $p \to \infty$: midpoint of range
- Issues: outliers, asymmetry, bimodality

Descriptive statistics: Dispersion

- Sample variance (squared standard deviation):

$$\sigma_x^2 = \min_c \frac{1}{N}\sum_{n=1}^{N}(x_n - c)^2 = \frac{1}{N}\sum_{n=1}^{N}(x_n - \mu_x)^2$$
$$= \frac{1}{N}\sum_{n=1}^{N} x_n^2 - \mu_x^2 = \frac{1}{N}||\vec{x}||^2 - \mu_x^2$$

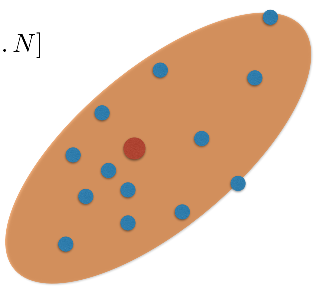
- Mean absolute deviation (MAD) about the median:

$$d_x = \frac{1}{N}\sum_{n=1}^{N}|x_n - m_x|$$

- Quantiles (eg: "90% of data lie in range [1.5 8.2]")

Descriptive statistics: Multi-D

Data points: $\left\{\vec{d}_n\right\} \quad n \in [1 \ldots N]$



Sample mean:

$$\vec{\mu}_d = \frac{1}{N}\sum_{n=1}^{N}\vec{d}_n = \begin{bmatrix}\mu_x \\ \mu_y\end{bmatrix}$$

Sample covariance:

$$C_d = \frac{1}{N}\sum_{n=1}^{N}(\vec{d}_n - \vec{\mu}_d)(\vec{d}_n - \vec{\mu}_d)^T = \frac{1}{N}\sum_{n=1}^{N}\vec{d}_n\vec{d}_n^T - \vec{\mu}_d\vec{\mu}_d^T$$
$$= \frac{1}{N}\begin{bmatrix}||\vec{x}||^2 & \vec{x}^T\vec{y} \\ \vec{y}^T\vec{x} & ||\vec{y}||^2\end{bmatrix} - \begin{bmatrix}\mu_x^2 & \mu_x\mu_y \\ \mu_y\mu_x & \mu_y^2\end{bmatrix} = \begin{bmatrix}\sigma_x^2 & \sigma_{xy} \\ \sigma_{xy} & \sigma_y^2\end{bmatrix}$$

Affine transformations

If $\vec{b}_n = M\left(\vec{d}_n - \vec{a}\right)$

then $\vec{\mu}_b = M\left(\vec{\mu}_d - \vec{a}\right)$

$C_b = M C_d M^T$

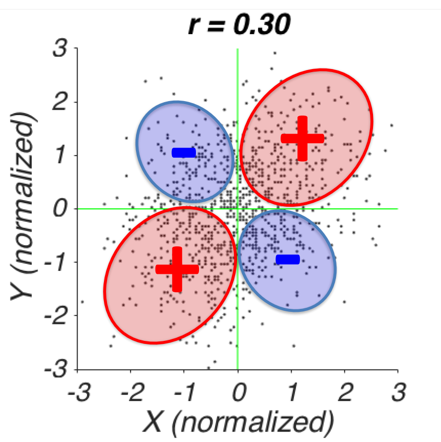
Special case: "center" and "normalize" the data:

$$\vec{a} = \begin{bmatrix}\mu_x \\ \mu_y\end{bmatrix} \qquad M = \begin{bmatrix}\frac{1}{\sigma_x} & 0 \\ 0 & \frac{1}{\sigma_y}\end{bmatrix}$$

"r" (Pearson correlation coefficient)

then $\vec{\mu}_b = \begin{bmatrix}0 \\ 0\end{bmatrix} \qquad C_b = \begin{bmatrix}1 & \boxed{\frac{\sigma_{xy}}{\sigma_x\sigma_y}} \\ \frac{\sigma_{xy}}{\sigma_x\sigma_y} & 1\end{bmatrix}$

[on board]



Correlation r captures dependency

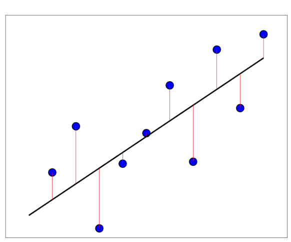
… but not slope!

Regression (revisited)

$$\vec{y} = \beta\vec{x} + \vec{e}$$

Optimal regression line slope:

$$\beta = \frac{\vec{x}^T\vec{y}}{\vec{x}^T\vec{x}} = \frac{\sigma_{xy}}{\sigma_x^2}$$



Error variance:

$$\sigma_e^2 = \sigma_y^2 - 2\beta\sigma_{xy} + \beta^2\sigma_x^2$$

$$= \sigma_y^2 - \frac{\sigma_{xy}^2}{\sigma_x^2}$$

Partition of variance:
error variance = data variance - explained variance

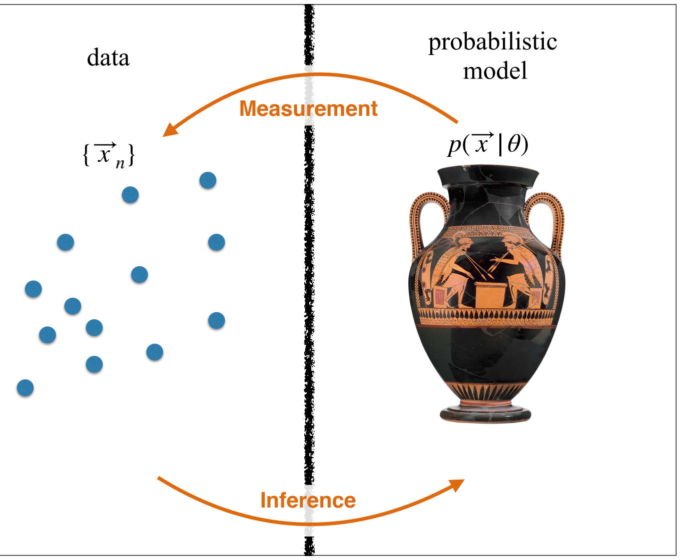
Expressed as a proportion of σ_y^2 :

$$\frac{\sigma_e^2}{\sigma_y^2} = 1 - \frac{\sigma_{xy}^2}{\sigma_x^2\sigma_y^2} = 1 - \boxed{r^2}$$

"r-squared" (proportion of variance explained)

Probability: an abstract mathematical framework for describing random quantities, or stochastic models of the world

Statistics: use of probability to summarize, analyze, interpret data. **Fundamental to all experimental science.**



Probabilistic Middleville

In Middleville, every family has two children, brought by the stork.

The stork delivers boys and girls randomly, with family probabilities {BB,BG,GB,GG}={0.2,0.3,0.2,0.3}
probabilistic model

You pick a family at random and discover that one of the children is a girl.
new data

What are the chances that the other child is a girl?
inference

Statistical Middleville

In Middleville, every family has two children, brought by the stork.

In a survey of 100 of the Middleville families, 32 have two girls, 23 have two boys, and the remainder one of each.
prev data

You pick a family at random and discover that one of the children is a girl.
new data

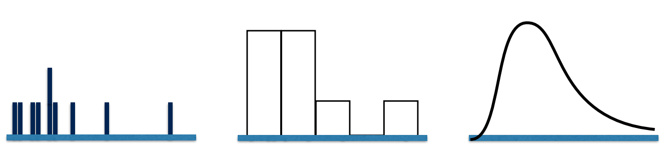
What are the chances that the other child is a girl?
inference

Univariate Probability (outline)

- distributions: discrete and continuous
- expected value, moments
- transformations: affine, monotonic nonlinear
- cumulative distributions. Quantiles, drawing samples

Frequentist view of probability: limit of infinite data

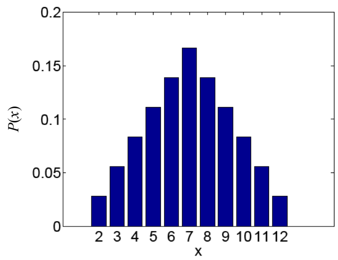
data → histogram → probability distribution



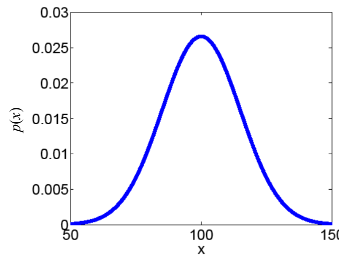
$\{x_n\}$ $\{c_k, h_k\}$ $p(x)$

Probability distributions

Discrete random variable



Continuous random variable



$$0 \le P(x_i) \le 1, \quad \forall i$$

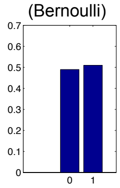
$$\sum_i P(x_i) = 1$$

$$0 \le p(x)$$

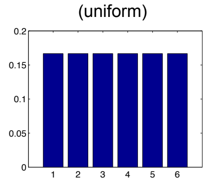
$$\int_{-\infty}^{\infty} p(x)\, dx = 1$$

Example distributions

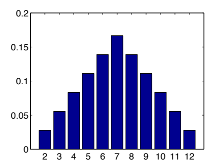
a not-quite-fair coin (Bernoulli)



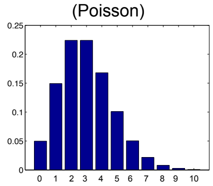
roll of a fair die (uniform)



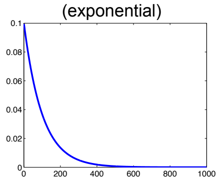
sum of rolls of two fair dice



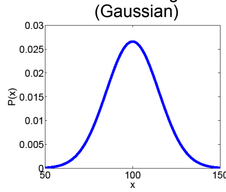
clicks of a Geiger counter, in a fixed time interval (Poisson)

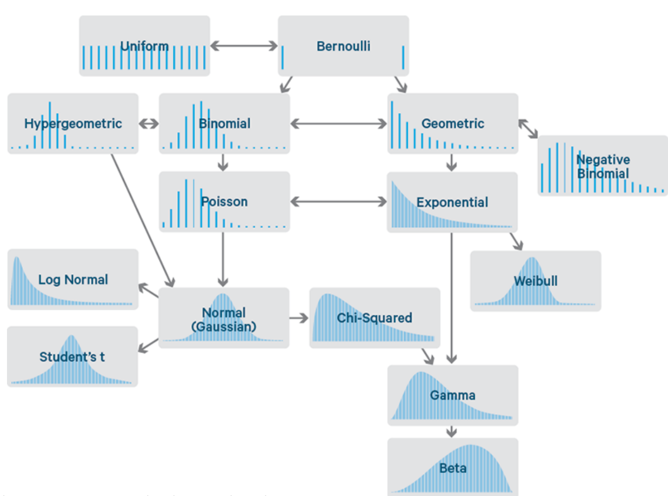


... and, time between clicks (exponential)



horizontal velocity of gas molecules exiting a fan (Gaussian)

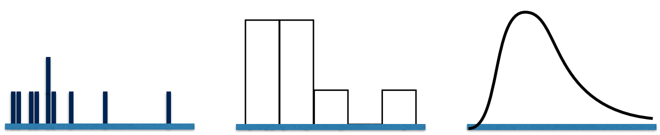




[Figure: Sean Owen, Cloudera Engineering]

Frequentist view of probability: limit of infinite data

data → histogram → probability distribution



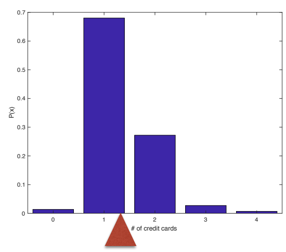
$\{x_n\}$ $\{c_k, h_k\}$ $p(x)$

$\mu_x = \frac{1}{N}\sum_n x_n$ $\mu_x \approx \frac{1}{K}\sum_k c_k h_k = \vec{c}^T\vec{h}$ $\mu_x = \int x\ p(x)\ dx$

Expected value (discrete)

$$E(X) = \sum_{n=1}^{N} x_n p(x_n)$$

(the mean, μ)



μ

More generally: $E(f(X)) = \sum_{n=1}^{N} f(x_n)p(x_n)$

Expected value (continuous)

$$E(x) = \int x\ p(x)\ dx \qquad [\text{“mean”}, \mu]$$

$$E(x^2) = \int x^2\ p(x)\ dx \qquad [\text{“second moment”}, m_2]$$

$$E\left((x-\mu)^2\right) = \int (x-\mu)^2\ p(x)\ dx \qquad [\text{“variance”}, \sigma^2]$$

$$= \int x^2\ p(x)\ dx - \mu^2 \qquad [m_2 \textit{ minus } \mu^2]$$

$$E\left(f(x)\right) = \int f(x)\ p(x)\ dx \qquad [\text{“expected value of } f\text{”}]$$

Note: expectation is an inner product, and thus *linear*, so:

$$E\left(af(x) + bg(x)\right) = aE\left(f(x)\right) + bE\left(g(x)\right)$$

Transformations of random variables

$Y = aX + b$ "affine" (linear plus constant)

Analogous to sample mean/covariance:

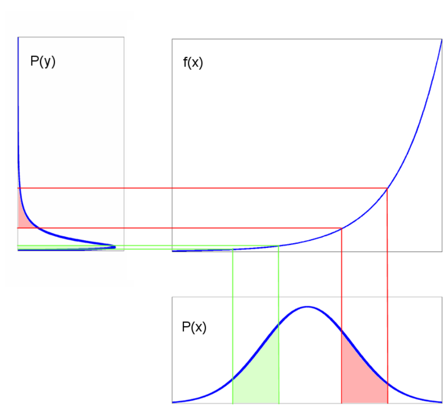
$$\mu_Y = E(Y) = aE(X) + b = a\mu_X + b$$

$$\sigma_Y^2 = E\left(\left(Y - \mu_Y\right)^2\right) = E\left(\left(aX - a\mu_X\right)^2\right) = a^2\sigma_X^2$$

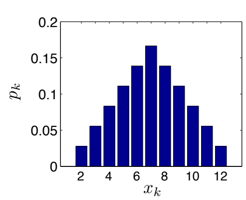
Full distribution: $p_Y(y) = \frac{1}{a}\ p_X\left(\frac{y-b}{a}\right)$

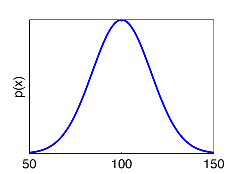
$Y = g(X)$ "monotonic" (derivative > 0)

$$p_Y(y) = \frac{p_X\left(g^{-1}(y)\right)}{g'\left(g^{-1}(y)\right)}$$



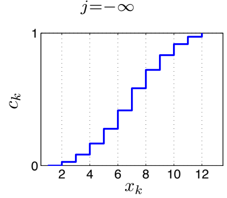
Cumulative distributions

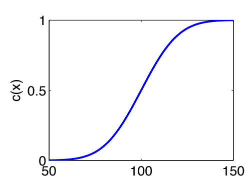




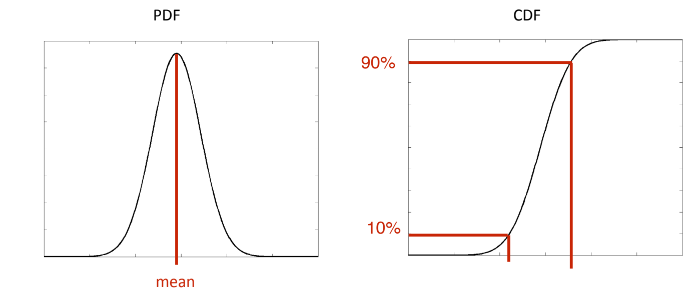
$$c_k = \sum_{j=-\infty}^{k} p_j$$

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

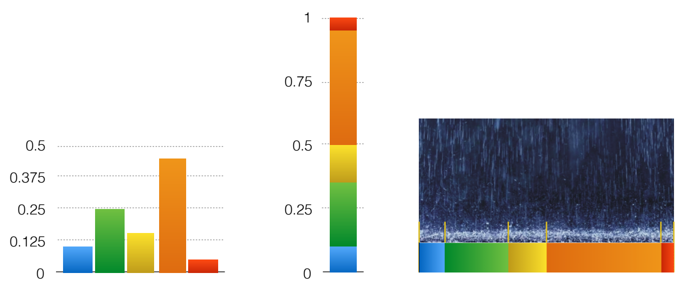




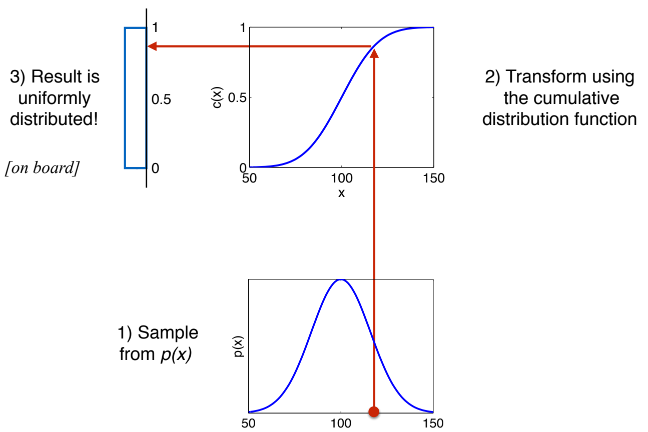
Confidence intervals



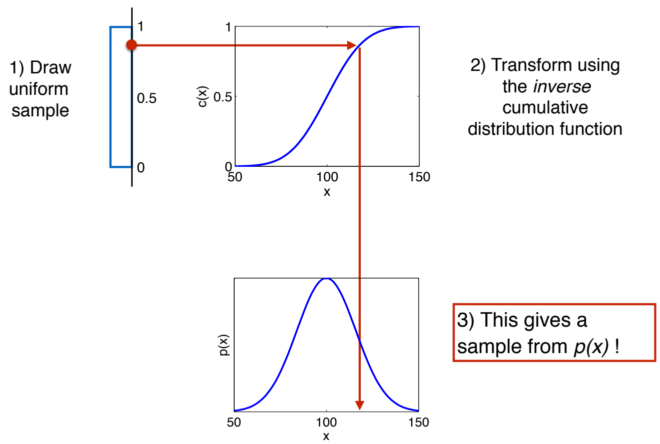
Drawing samples - discrete



Drawing samples - continuous



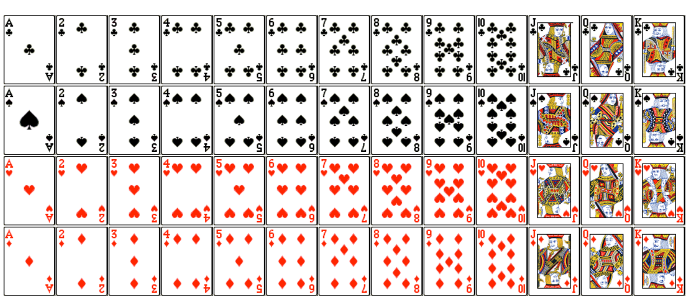
Drawing samples - continuous



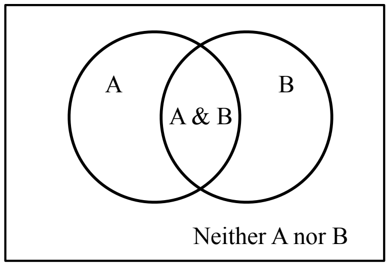
Multi-variate probability (outline)

- Joint distributions
- Marginals (integrating)
- Conditionals (slicing)
- Bayes' rule (inverse probability)
- Statistical independence (separability)
- Mean/Covariance
- Linear transformations

Joint and conditional probability - discrete

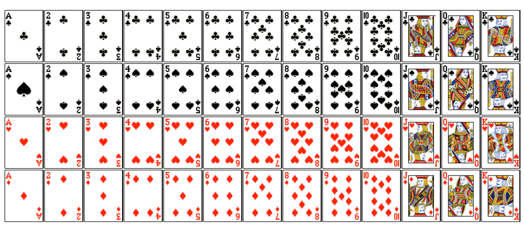


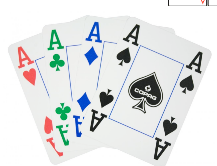
Conditional probability



$p(A \mid B)$ = probability of A given that B is asserted to be true $= \dfrac{p(A \,\&\, B)}{p(B)}$

Joint and conditional probability - discrete

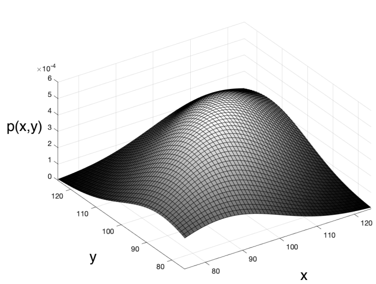


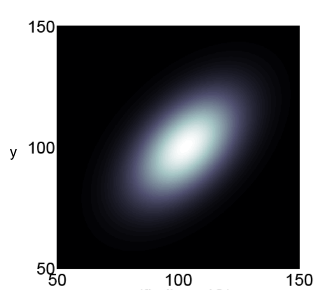


P(Ace)
P(Heart)
P(Ace & Heart)
P(Ace | Heart)
P(not Jack of Diamonds)
P(Ace | not Jack of Diamonds)

"Independence"

Joint distribution (continuous)

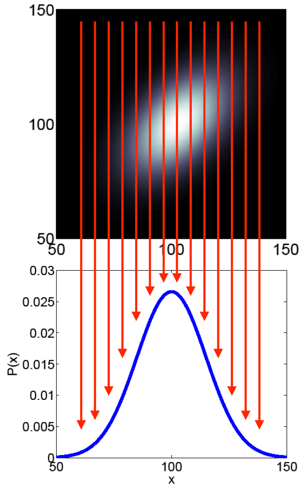




$$p(x, y)$$

Marginal distribution

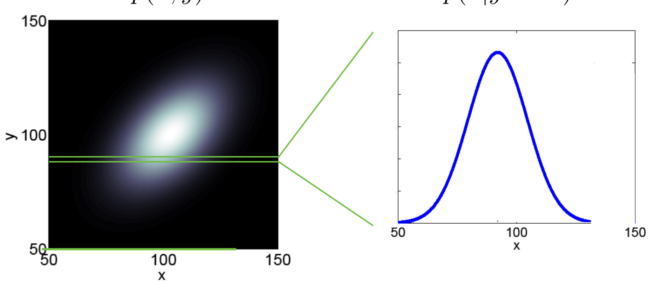
$$p(x, y)$$



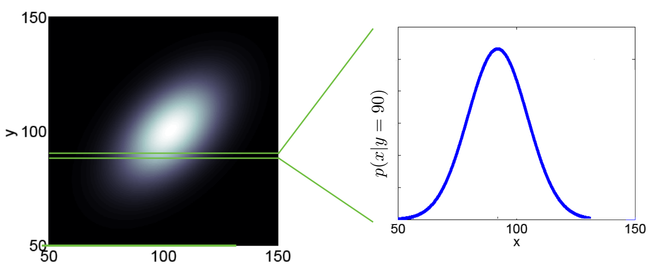
$$p(x) = \int p(x, y)dy$$

Conditional distribution

$p(x, y)$ $\quad$ $p(x|y = 90)$



Conditional distribution



$$p(x|y = 90) = p(x, y = 90) \Big/ \int p(x, y = 90)dx$$

$$= p(x, y = 90) / p(y = 90)$$

slide joint distribution $\quad$ normalize (by marginal)

More generally:

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

Bayes' Rule

LII. *An Essay towards solving a Problem in the Doctrine of Chances. By the late Rev. Mr.* Bayes, *F. R. S. communicated by Mr.* Price, *in a Letter to* John Canton, *A. M. F. R. S.*

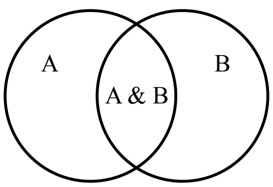
Dear Sir,

Read Dec. 23, 1763. I Now send you an essay which I have found among the papers of our deceased friend Mr. Bayes, and which, in my opinion, has great merit, and well deserves to be preserved.

$$p(x|y) = p(y|x)\,p(x)/p(y)$$

(a direct consequence of the definition of conditional probability)

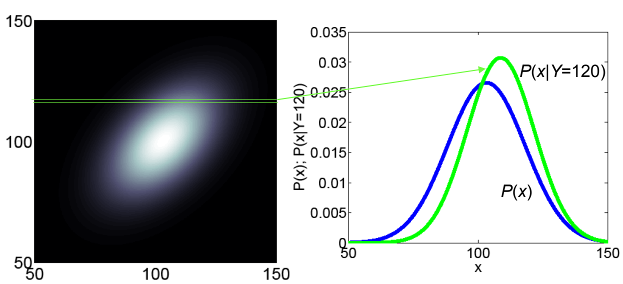
Bayes' Rule



$p(A\,|\,B) =$ probability of A given that B is asserted to be true $= \dfrac{p(A\,\&\,B)}{p(B)}$

$$p(A\,\&\,B) = p(B)\,p(A\,|\,B)$$
$$= p(A)\,p(B\,|\,A)$$
$$\Rightarrow p(A\,|\,B) = \frac{p(B\,|\,A)\,p(A)}{p(B)}$$

Conditional vs. marginal



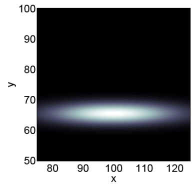
In general, the marginals for different Y values differ.

When are they they same? In particular, when are all conditionals equal to the marginal?

Statistical independence

Random variables X and Y are statistically independent if (and only if):

$$p(x,y) = p(x)p(y) \quad \forall\, x,y$$



(note: for discrete distributions, this is an outer product!)

Independence implies that *all* conditionals are equal to the corresponding marginal:

$$p(x \mid y) = p(x,y)\,/\,p(y) = p(x) \quad \forall\, x,y$$

Special case: Sum of two RVs

Let $Z = X + Y$. From rules for affine transforms:

$$\mu_Z = E(Z) = E(X) + E(Y)$$

$$\sigma_Z^2 = \sigma_X^2 + 2\sigma_{XY} + \sigma_Y^2$$

Special case: if X and Y are independent, then:

$$E(XY) = E(X)E(Y) \quad \text{and thus} \quad \sigma_{XY} = 0$$

$$\sigma_Z^2 = \sigma_X^2 + \sigma_Y^2$$

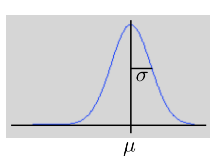
$p_Z(z)$ is the *convolution* of $p_X(x)$ and $p_Y(y)$

[on board]

Gaussian (a.k.a. “Normal”) densities

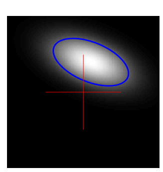
Parameterized by mean and stdev:

$$p(x) = \frac{1}{\sqrt{2\pi\sigma^2}}\, e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$



Multi-dimensional generalization:

$$p(\vec{x}) = \frac{1}{\sqrt{(2\pi)^N |C|}}\, e^{-(\vec{x}-\vec{\mu})^T C^{-1} (\vec{x}-\vec{\mu})/2}$$

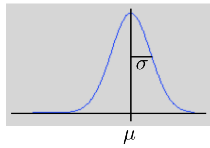


mean: [0.2, 0.8]
cov: [1.0 -0.3;
-0.3 0.4]

Gaussian properties

$$p(x) = \frac{1}{\sqrt{2\pi\sigma^2}}\, e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

$$p(\vec{x}) = \frac{1}{\sqrt{(2\pi)^N |C|}}\, e^{-(\vec{x}-\vec{\mu})^T C^{-1} (\vec{x}-\vec{\mu})/2}$$

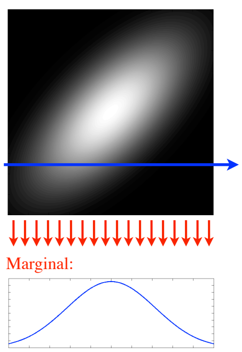


- joint density of indep Gaussian RVs is elliptical [easy]
- conditionals of a Gaussian are Gaussian [easy]
- marginals of a Gaussian are Gaussian [easy]
- product of two Gaussian dists is Gaussian [easy]
- sum of Gaussian RVs is Gaussian [moderate]
- the most random (max entropy) density of given variance [moderate]
- central limit theorem: sum of many RVs is Gaussian [hard]

---

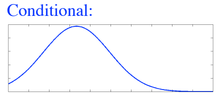
let $P = C^{-1}$ (the "precision" matrix)

$$\begin{aligned} p(x_1 | x_2 = a) &\propto e^{-\frac{1}{2}\left[P_{11}(x_1-\mu_1)^2 + 2P_{12}(x_1-\mu_1)(a-\mu_2)+\ldots\right]} \\ &= e^{-\frac{1}{2}\left[P_{11}x_1^2 + 2(P_{12}(a-\mu_2)-P_{11}\mu_1)x_1+\ldots\right]} \\ &= e^{-\frac{1}{2}\left(x_1-\mu_1+\frac{P_{12}}{P_{11}}(a-\mu_2)\right)P_{11}\left(x_1-\mu_1+\frac{P_{12}}{P_{11}}(a-\mu_2)\right)+\ldots} \end{aligned}$$



Gaussian, with: $\mu = \mu_1 - \frac{P_{12}}{P_{11}}(a-\mu_2)$

$\sigma^2 = \frac{1}{P_{11}}$



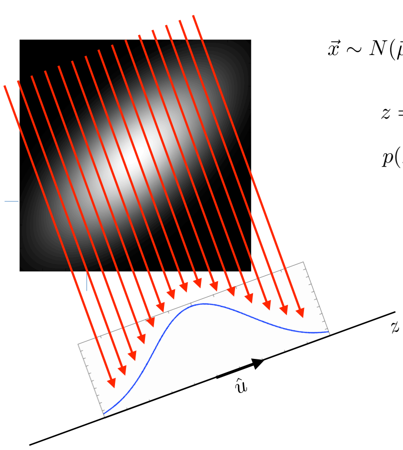
$$p(x_1) = \int p(\vec{x})\; dx_2$$

[on board]

Gaussian, with:

$$\begin{aligned} \mu &= \mu_1 \\ \sigma^2 &= C_{11} \end{aligned}$$

Generalized marginals of a Gaussian



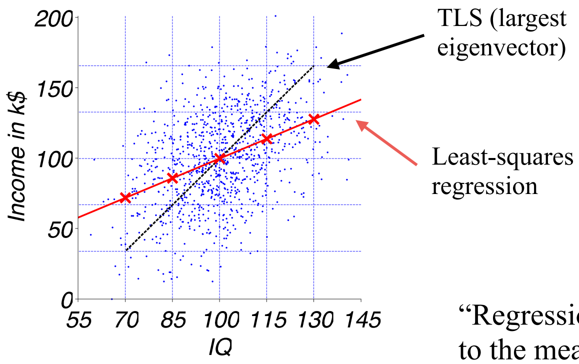
$\vec{x} \sim N(\vec{\mu}_x, C_x)$

$z = \hat{u}^T \vec{x}$

$p(z)$ is Gaussian, with:

$$\begin{aligned} \mu_z &= \hat{u}^T \vec{\mu}_x \\ \sigma_z^2 &= \hat{u}^T C_x \hat{u} \end{aligned}$$

Correlation and regression



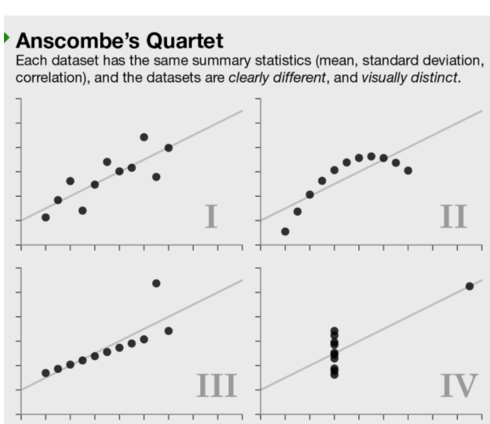
"Regression to the mean"

Correlation implies dependency

… but not slope

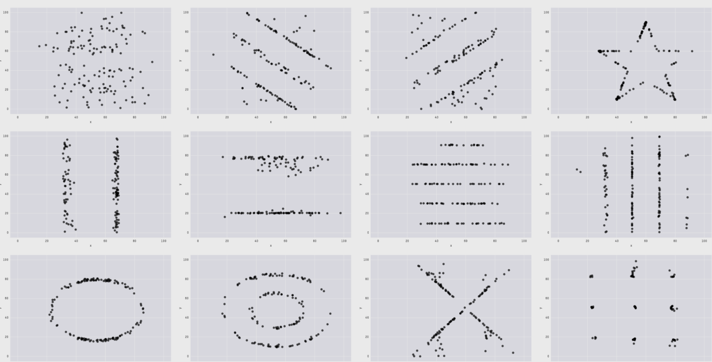
… and its absence does not imply independence!

Correlation between variables does not uniquely indicate the shape of their joint distribution



More extreme examples !

X Mean: 54.26
Y Mean: 47.83
X SD : 16.76
Y SD : 26.93
Corr. : -0.06

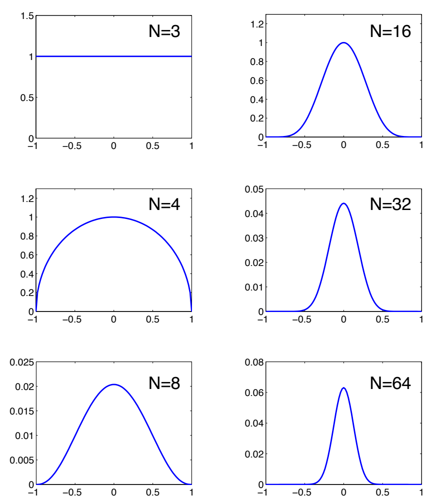


https://www.autodeskresearch.com/publications/samestats

Lack of correlation is favored in N>3 dimensions

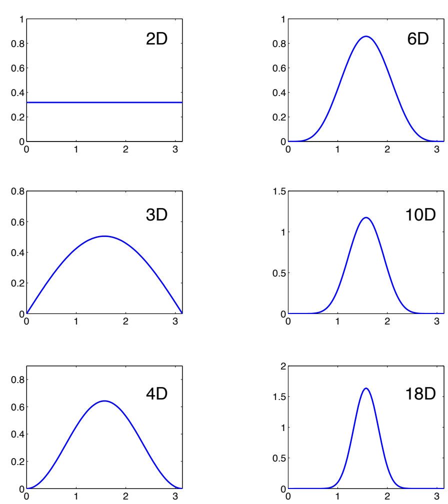
Null Hypothesis: Distribution of normalized dot product of pairs of Gaussian vectors in N dimensions:

$$(1-d^2)^{\frac{N-3}{2}}$$

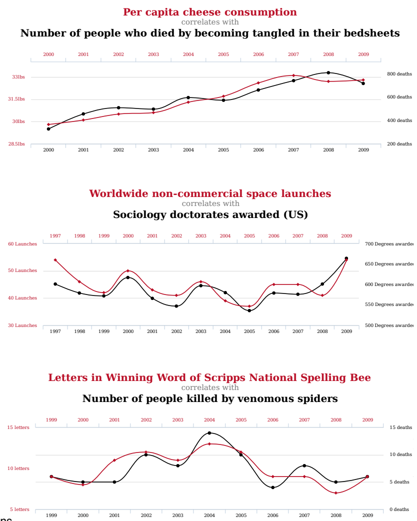


Distribution of angles of pairs of Gaussian vectors

sin(theta)^(N-2)



Nevertheless, one can find correlation if one looks for it!



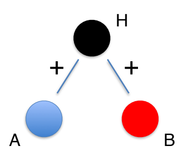
http://www.tylervigen.com/spurious-correlations

Covariation/correlation does not imply causation

- Correlation does not provide a direction for causality. For that, you need additional (temporal) information.
- More generally, correlations are often a result of hidden (unmeasured, uncontrolled) variables…

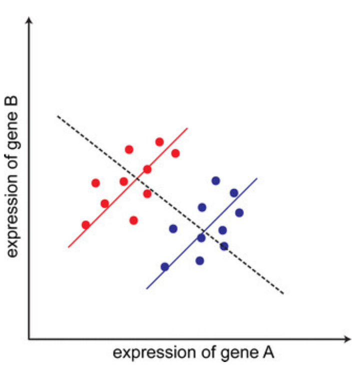
Example: conditional independence:

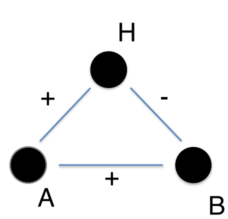
$$p(A,B \mid H) = p(A \mid H)\, p(B \mid H)$$



[On board: in Gaussian case, connections are explicit in the *precision matrix*]

Another example: Simpson's paradox

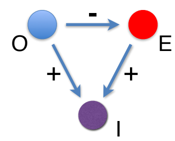




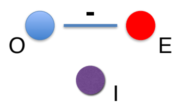
Milton Friedman's Thermostat

O = outside temperature (assumed cold)
I = inside temperature (ideally, constant)
E = energy used for heating

True interactions:



Statistical interactions, $P=C^{-1}$:



Statistical observations:

- O and I uncorrelated
- I and E uncorrelated
- O and E anti-correlated

Some nonsensical conclusions:

- O and E have no effect on I, so shut off heater to save money!
- I is irrelevant, and can be ignored. Increases in E cause decreases in O.

Statistical summary cannot replace scientific reasoning/experiments!

Summary: Correlation misinterpretations

- Correlation does not imply data lie near a line/plane/hyperplane (subspace), with simple noise perturbations
- Correlation implies dependency, but lack of correlation does *not* imply independence
- Correlation does *not* imply causation (temporally, or by direct influence/connection)
- Correlation is a descriptive statistic, and does not eliminate the need for scientific reasoning/experiment!