Mathematical Tools
for Neural and Cognitive Science

Fall semester, 2024

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)
- 2000's: Machine learning (statistical inference with large-scale computing + lots of data)
- Also (since 1950's): statistical neural/cognitive models!

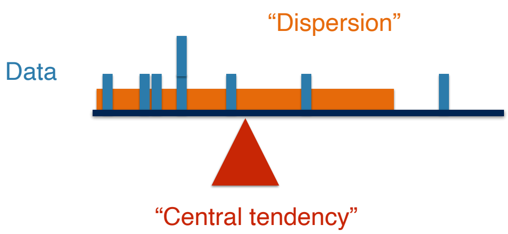
Statistics as summary description

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



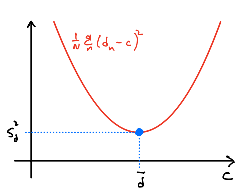
Descriptive statistics: mean and variance

Often use average & variance for central tendency & dispersion:

- Sample mean ***minimizes*** the squared error

$$\bar{d} = \arg\min_c \frac{1}{N}\sum_{n=1}^{N}(d_n - c)^2 = \arg\min_c \frac{1}{N}\|\vec{d} - c\vec{1}\|^2$$

$$= \frac{\vec{d}^T\vec{1}}{\vec{1}^T\vec{1}} = \frac{\vec{d}^T\vec{1}}{N} \qquad \text{(regression!)}$$



- Sample variance ***is*** the squared error

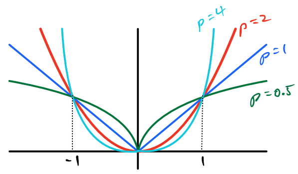
$$s_d^2 = \min_c \frac{1}{N}\sum_{n=1}^{N}(d_n - c)^2$$

$$= \frac{1}{N}\sum_{n=1}^{N}(d_n - \bar{d})^2 = \frac{1}{N}\sum_{n=1}^{N} d_n^2 - \bar{d}^2 = \frac{\|\vec{d}\|^2}{N} - \bar{d}^2$$

Descriptive statistics: alternatives

More generally, define **dispersion** in terms of the "L_p norm":

$$\arg\min_{c} \left[\frac{1}{N} \sum_{n=1}^{N} |d_n - c|^p \right]^{1/p}$$

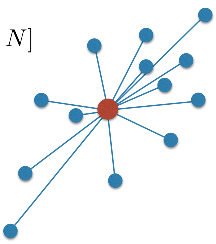


Different p values give different measures of **central tendency**:

- $p = 2$: mean (standard choice)
- $p = 1$: median
- $p \to 0$: mode (location of maximum)
- $p \to \infty$: midpoint of range

Descriptive statistics: 2-D

- Data points: $\vec{d}_n = \begin{bmatrix} x_n \\ y_n \end{bmatrix}, \quad n \in [1 \ldots N]$
- Sample mean: as in 1-D, define as vector $\vec{c}$ that minimizes average squared distance to data points:



$$\bar{d} = \arg\min_{\vec{c}} \sum_n \|\vec{d}_n - \vec{c}\|^2, \quad \vec{c} = \begin{bmatrix} c_x \\ c_y \end{bmatrix}$$

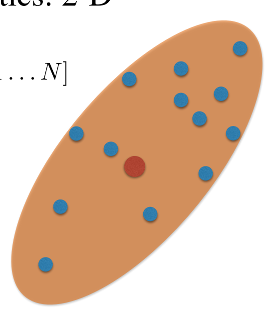
$$= \arg\min_{c_x, c_y} \sum_n (x_n - c_x)^2 + (y_n - c_y)^2$$

$$= \frac{1}{N} \sum_{n=1}^{N} \vec{d}_n = \begin{bmatrix} \bar{x} \\ \bar{y} \end{bmatrix}$$

Descriptive statistics: 2-D

- Data points: $\vec{d}_n = \begin{bmatrix} x_n \\ y_n \end{bmatrix}, \quad n \in [1 \ldots N]$
- Sample mean:

$$\bar{d} = \frac{1}{N} \sum_{n=1}^{N} \vec{d}_n = \begin{bmatrix} \bar{x} \\ \bar{y} \end{bmatrix}$$



- Sample covariance:

$$C_d = \frac{1}{N} \sum_{n=1}^{N} (\vec{d}_n - \bar{d})(\vec{d}_n - \bar{d})^T = \frac{1}{N} \sum_{n=1}^{N} \vec{d}_n \vec{d}_n^T - \bar{d}\bar{d}^T$$

$$= \frac{1}{N} \begin{bmatrix} ||\vec{x}||^2 & \vec{x}^T \vec{y} \\ \vec{y}^T \vec{x} & ||\vec{y}||^2 \end{bmatrix} - \begin{bmatrix} \bar{x}^2 & \bar{x}\bar{y} \\ \bar{y}\bar{x} & \bar{y}^2 \end{bmatrix} = \begin{bmatrix} s_x^2 & s_{xy} \\ s_{xy} & s_y^2 \end{bmatrix}$$

Descriptive statistics: multi-D

Data points: matrix D (N data vectors on rows)

Sample mean:

$$\bar{d} = \frac{D^T \vec{1}}{N}$$

Sample covariance:

$$C_d = \frac{\left(D - \vec{1}\bar{d}^T\right)^T \left(D - \vec{1}\bar{d}^T\right)}{N}$$

$$= \frac{D^T D}{N} - \bar{d}\,\bar{d}^T$$

[compare to 1-D case]

Affine transformations

If $\vec{b}_n = M\left(\vec{d}_n - \vec{a}\right)$ (translate, then rotate-stretch-rotate)

then $\bar{b} = M\left(\bar{d} - \vec{a}\right)$

$C_b = M C_d M^T$

(mean and covariance transform according to simple rules)

Standard case: "re-center" and "normalize" the components:

Let $\vec{a} = \begin{bmatrix} \bar{x} \\ \bar{y} \end{bmatrix}$ $\quad M = \begin{bmatrix} \frac{1}{s_x} & 0 \\ 0 & \frac{1}{s_y} \end{bmatrix}$

then $\bar{b} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$ $\quad C_b = \begin{bmatrix} 1 & \frac{s_{xy}}{s_x s_y} \\ \frac{s_{xy}}{s_x s_y} & 1 \end{bmatrix}$

"r" (Pearson correlation coefficient)

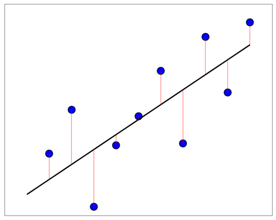
[on board]

Correlation (r) captures dependency

… but not slope!

Regression (revisited)

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



Optimal regression line slope:

$$\beta = \frac{\vec{x}^T\vec{y}}{\vec{x}^T\vec{x}} = \frac{s_{xy}}{s_x^2}$$

Error variance:

$$s_e^2 = s_y^2 - 2\beta s_{xy} + \beta^2 s_x^2$$

$$= s_y^2 - \frac{s_{xy}^2}{s_x^2}$$

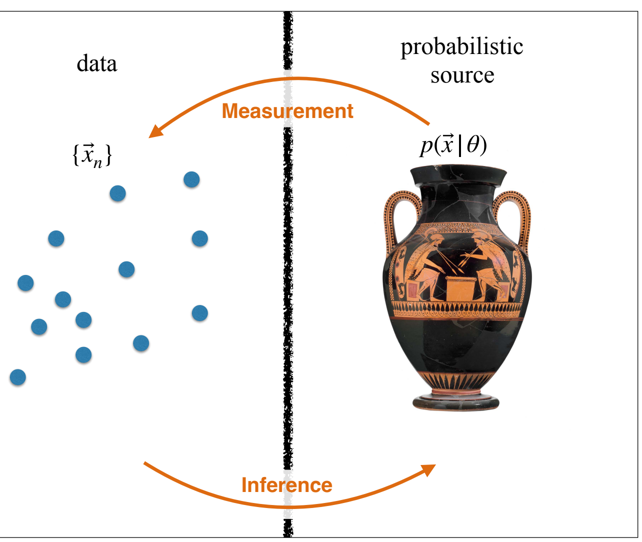
Expressed as a proportion of σ_y^2:

$$\frac{s_e^2}{s_y^2} = 1 - \frac{s_{xy}^2}{s_x^2 s_y^2} = 1 - r^2$$

"r-squared" (proportion of variance explained)

Probability: an abstract mathematical framework for describing random quantities

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

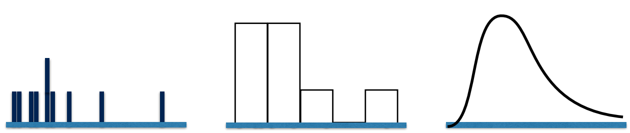


Univariate Probability (outline)

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

Frequentist view of probability: limit of infinite data

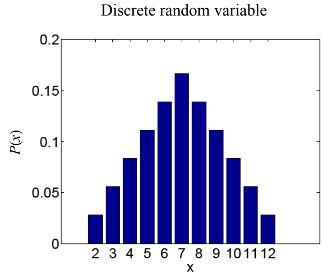
data → histogram → probability distribution



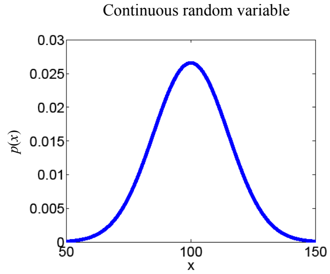
$\{x_n\}$ $\{c_k, h_k\}$ $p(x)$

Probability distributions

Discrete random variable



Continuous random variable



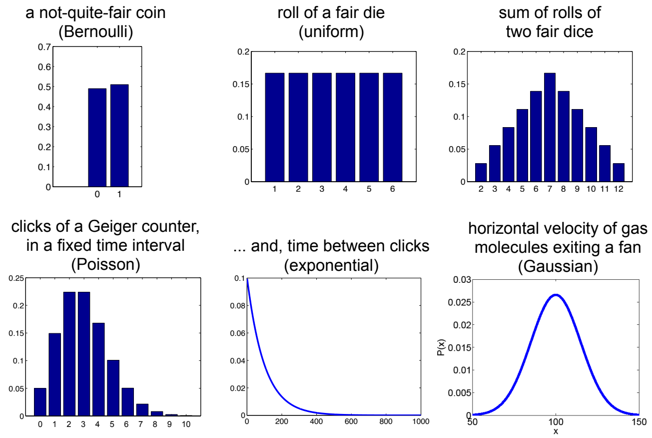
$$0 \le P(x_i) \le 1, \quad \forall i$$

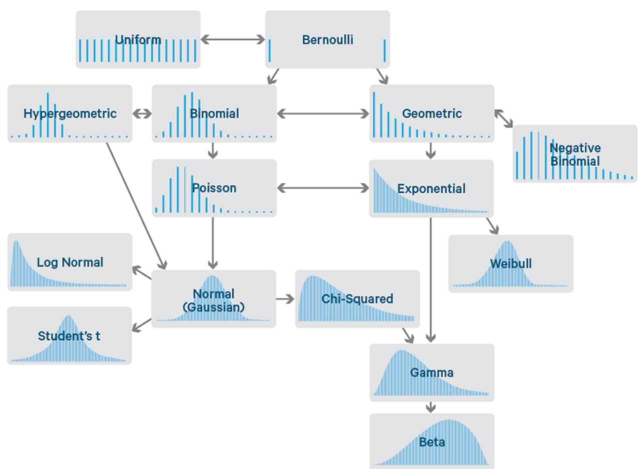
$$\sum_i P(x_i) = 1$$

$$0 \le p(x)$$

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

Example distributions



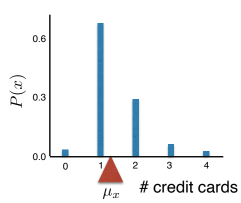


[Figure: Sean Owen, Cloudera Engineering]

Expected value (for a discrete random variable)

$$\mu_x = \mathbb{E}(x) = \sum_{k=1}^{K} x_k \; P(x_k)$$

a weighted sum over the discrete values



More generally: $$\mathbb{E}\left(f(x)\right) = \sum_{k=1}^{K} f(x_k) P(x_k)$$ (sum over values of R.V.)

Sample average: an estimate of the expected value:

$$\bar{f}(x) = \frac{1}{N} \sum_{n=1}^{N} f(x_n)$$ (sum over data samples)

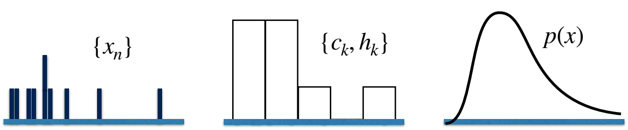
Sample average converges to expected value as one gathers more data…

A note on notation

- We have, and will continue to use the notation for a “sample mean” ($\bar{x}$) and a “sample standard deviation” (s) or variance (s^2).
- Statistics makes a distinction between these sample values and the corresponding “population” values of mean (μ) and variance (σ^2).

Expected value (continuous random variable)

data → histogram → probability distribution



{center, normalized height} of kth bin

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

Expected value (continuous)

$$\mathbb{E}(x) = \int x \; p(x) \; dx$$ [“mean”, μ]

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

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

$$= \int x^2 p(x) \; dx - \mu^2$$ [m_2 *minus* μ^2]

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

Note: expectation is an integral, and thus *linear*, so:

$$\mathbb{E}\left(af(x) + bg(x)\right) = a\mathbb{E}\left(f(x)\right) + b\mathbb{E}\left(g(x)\right)$$

Transformations of scalar random variables

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

Analogous to sample mean/covariance:

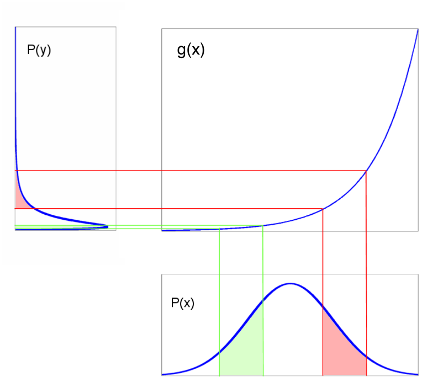
$$\mu_Y = \mathbb{E}(Y) = a\mathbb{E}(X) + b = a\mu_X + b$$

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

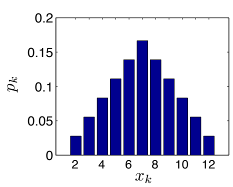
Full distribution: $p_Y(y) = \frac{1}{a}\, p_X\left(\frac{y-b}{a}\right)$

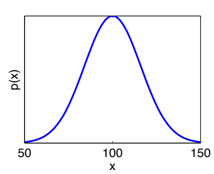
$Y = g(X)$ (assume g is "monotonic" - i.e., derivative > 0)

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



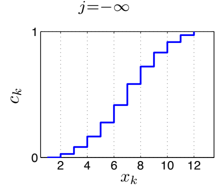
Cumulative distributions

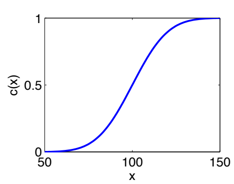




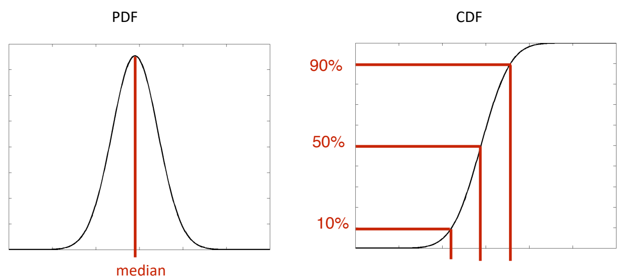
$$c_k = \sum_{j=-\infty}^{k} p_j$$

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

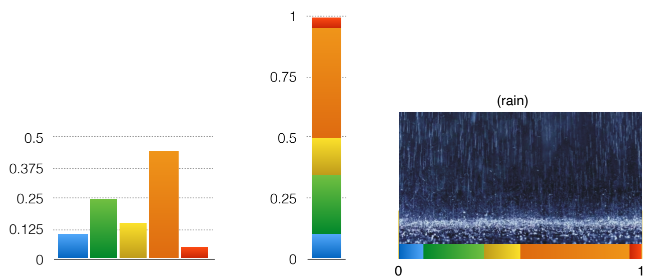




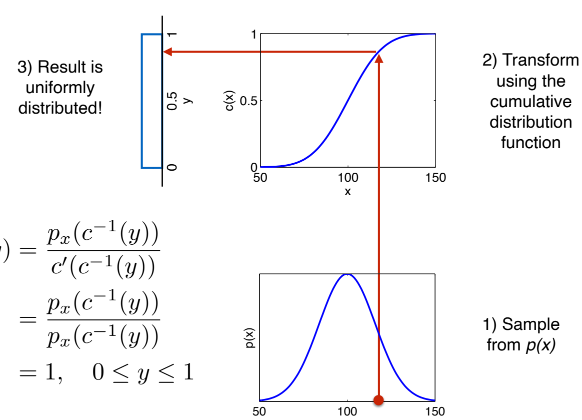
Quantiles



Drawing samples - discrete

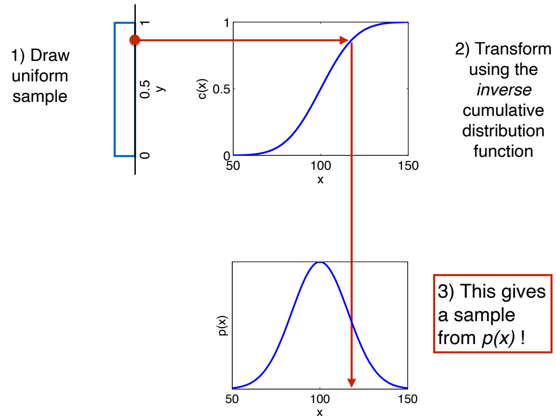


Drawing samples - continuous



$$p_y(y) = \frac{p_x(c^{-1}(y))}{c'(c^{-1}(y))}$$
$$= \frac{p_x(c^{-1}(y))}{p_x(c^{-1}(y))}$$
$$= 1, \quad 0 \leq y \leq 1$$

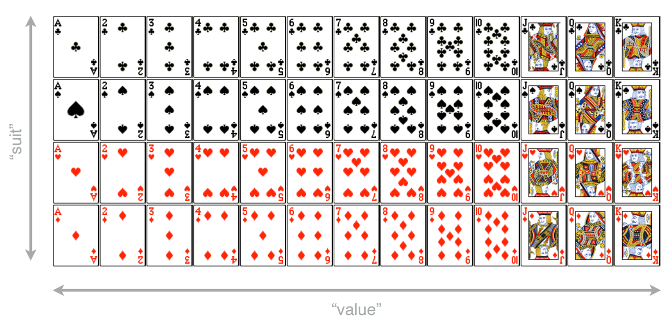
Drawing samples - continuous



Multi-variate probability (outline)

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

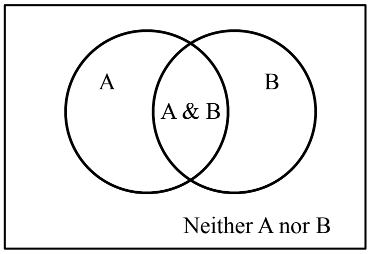
Joint probability - discrete



P(Ace)
P(Heart)

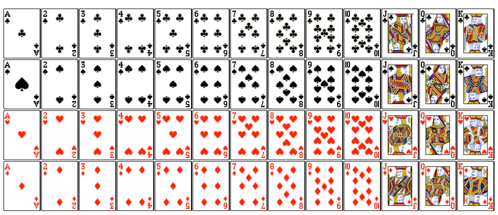
P(Ace & Heart)
P(Ace & red)
P(number & Heart)

Conditional probability



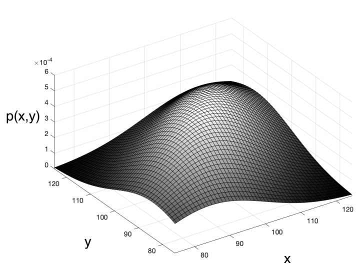
$$p(A \mid B) = \text{probability of } A \text{ given that } B \text{ is asserted to be true} = \frac{p(A \,\&\, B)}{p(B)}$$

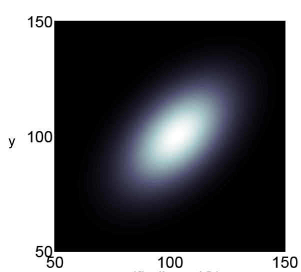
Conditional probability - discrete



P(Ace | Heart)
P(Ace | red)
P(number | Heart)

Joint distribution (continuous)



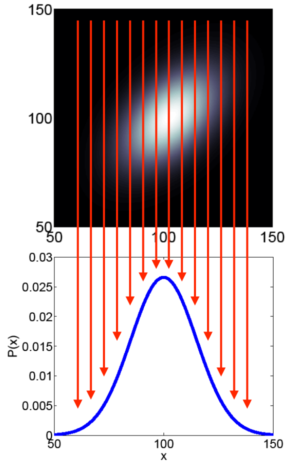


$p(x, y)$

Marginal distribution

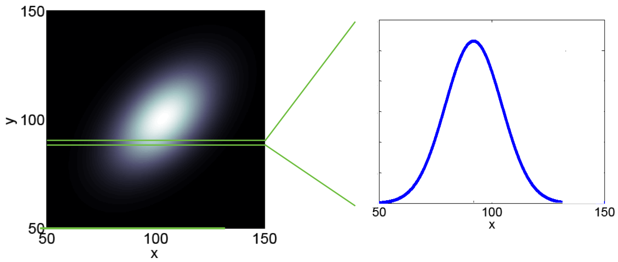
$p(x, y)$

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

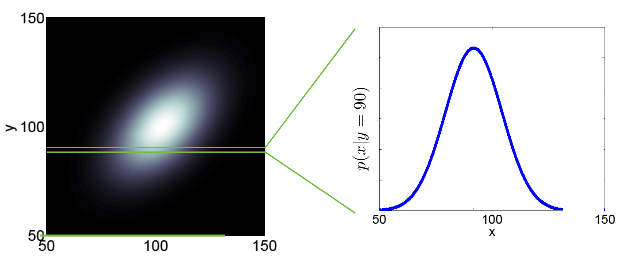


Conditional distribution

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



Conditional distribution



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

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

slice joint distribution

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.*

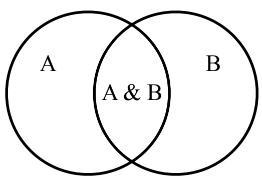
Dear Sir,

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)

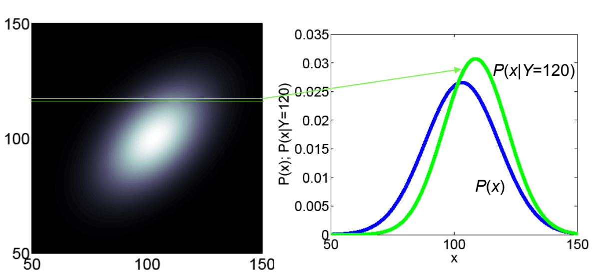
Bayes' Rule



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

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

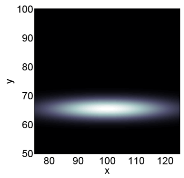
Conditional vs. marginal



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:

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

Mean, covariance, affine transformations

For R.V. $\vec{x}$, $\vec{\mu}_x = \mathbb{E}(\vec{x})$, $C_x = \mathbb{E}\left((\vec{x} - \vec{\mu}_x)(\vec{x} - \vec{\mu}_x)^T\right)$

For R.V. $\vec{y} = M(\vec{x} - \vec{a})$,

analogous to results for sample mean/covariance:

$$\begin{aligned} \vec{\mu}_y &= \mathbb{E}\left(M(\vec{x} - \vec{a})\right) \\ &= M\left(\mathbb{E}(\vec{x}) - \vec{a}\right) \\ &= M\left(\vec{\mu}_x - \vec{a}\right) \end{aligned}$$

$$\begin{aligned} C_y &= \mathbb{E}\left((M(\vec{x} - \vec{\mu}_x))(M(\vec{x} - \vec{\mu}_x))^T\right) \\ &= M\mathbb{E}\left((\vec{x} - \vec{\mu}_x))(\vec{x} - \vec{\mu}_x))^T\right) M^T \\ &= MC_xM^T \end{aligned}$$

Special case: Sum of two RVs

Let $Z = X + Y$, or $Z = \vec{1}^T \begin{bmatrix} X \\ Y \end{bmatrix}$

$$\mu_Z = \mu_X + \mu_Y$$

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

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

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

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

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

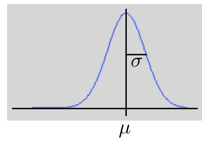
[on board]

Gaussian (a.k.a. “Normal”) densities

One-dimensional:

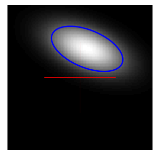
$$p(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \; e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

Alt. notation: $x \sim N(\mu, \sigma^2)$



Multi-dimensional:

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

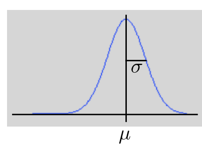


mean: $[0.2, 0.8]$
cov: [1.0 -0.3;
-0.3 0.4]

Gaussian properties

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

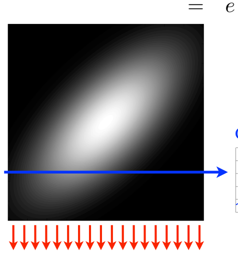
$$p(\vec{x}) = \frac{1}{\sqrt{(2\pi)^N |C|}} \; e^{-(\vec{x}-\vec{\mu})^T C^{-1} (\vec{x}-\vec{\mu})/2}$$



- 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 independent Gaussian RVs is Gaussian [moderate]
- the most random (max entropy) density of given variance [moderate]
- central limit theorem: sum of many indep. RVs is Gaussian [hard]

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

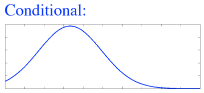
$$\begin{aligned} p(x_1 | x_2 = a) &\propto e^{-\frac{1}{2}\left[P_{11}(x_1-\mu_1)^2 + 2P_{12}(x_1-\mu_1)(a-\mu_2) + \ldots\right]} \\ &= e^{-\frac{1}{2}\left[P_{11}x_1^2 + 2(P_{12}(a-\mu_2) - P_{11}\mu_1)x_1 + \ldots\right]} \\ &= e^{-\frac{1}{2}\left(x_1-\mu_1+\frac{P_{12}}{P_{11}}(a-\mu_2)\right)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



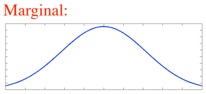
Gaussian, with: $\mu = \mu_1 - \frac{P_{12}}{P_{11}}(a - \mu_2)$

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

Conditional:



Marginal:



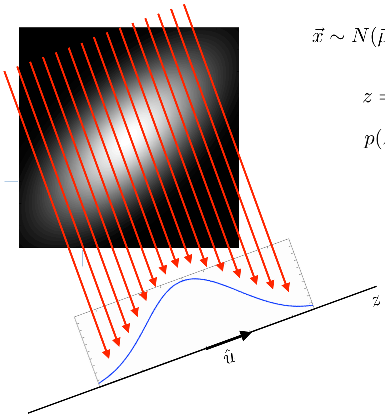
$$p(x_1) = \int p(\vec{x}) \; dx_2$$

[on board]

Gaussian, with:

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

Generalized marginals of a Gaussian



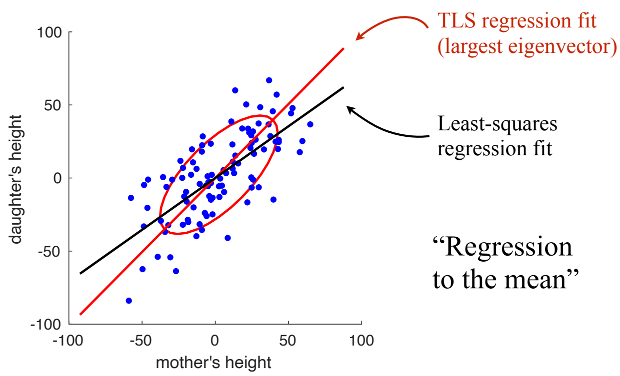
$\vec{x} \sim N(\vec{\mu}_x, C_x)$

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

$p(z)$ is Gaussian, with:

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

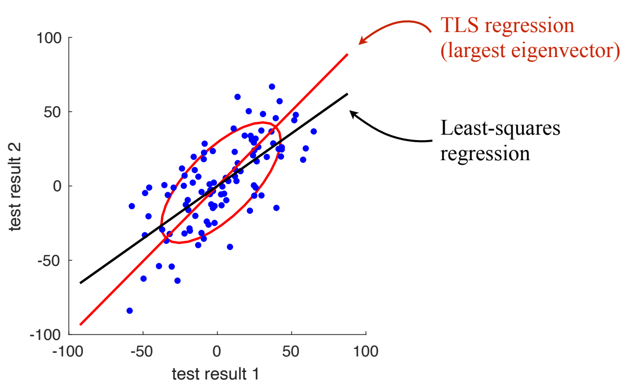
Correlation and regression



"Regression to the mean"

Francis Galton (1886). "Regression towards mediocrity in hereditary stature"

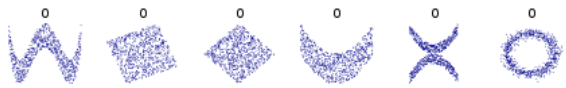
Correlation and regression



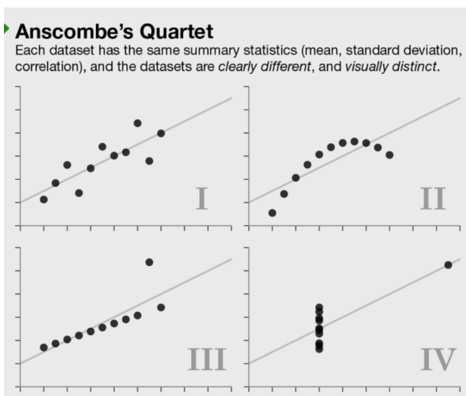
Correlation implies dependency

… but not slope

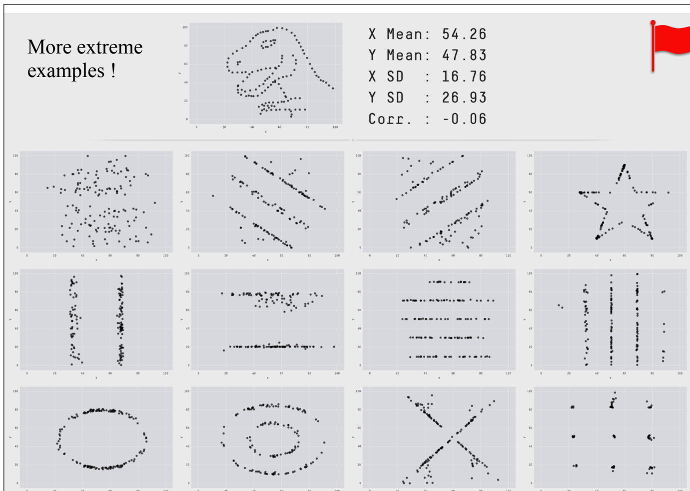
… and its absence does not imply independence!



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



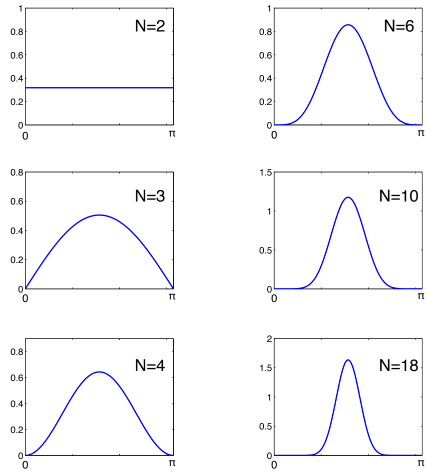
More extreme examples !



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

Distribution of angles of pairs of random unit vectors

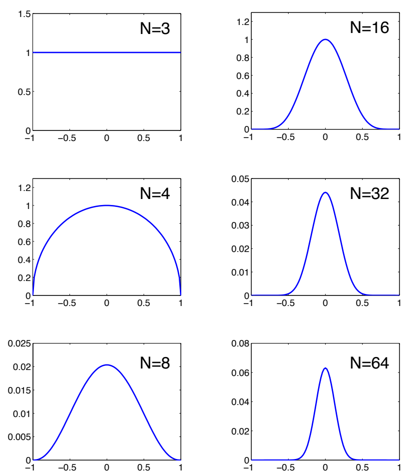
$$p(\theta) \propto \sin(\theta)^{(N-2)}$$



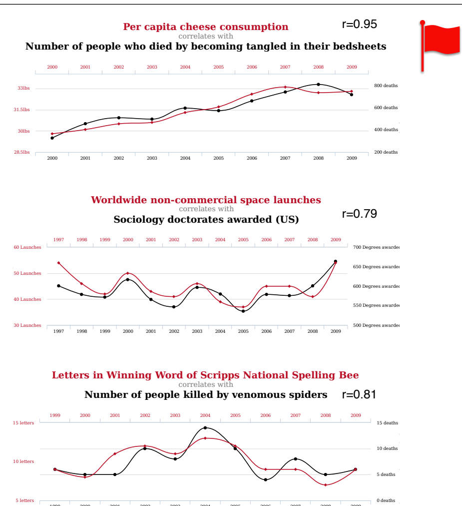
Lack of correlation is favored in N>3 dimensions

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

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



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



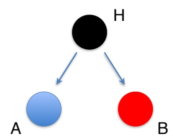
http://www.tylervigen.com/spurious-correlations

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…

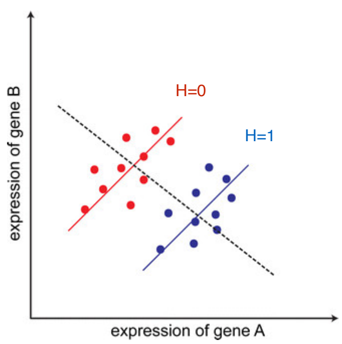
Example: conditional independence:

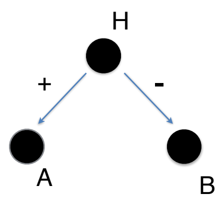
$$p(A,B\,|\,H) = p(A\,|\,H)p(B\,|\,H)$$



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

Another example: "Simpson's paradox"

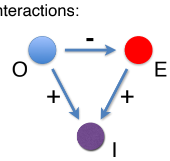




Milton Friedman's Thermostat

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

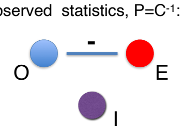
True interactions:



Statistical observations:

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

Observed statistics, $P=C^{-1}$:



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 implies dependency, but does *not* imply data lie near a line/plane/hyperplane.
- Independent implies uncorrelated. But uncorrelated does *not* imply independent.
- Correlation does *not* imply causation (and often arises from hidden common factors).
- Correlation is a **descriptive statistic**, and does not eliminate the need for reasoning/experiments/models!