1

Mathematical Tools
for Neural and Cognitive Science

Fall semester, 2018

Probability & Statistics:
Intro, summary statistics, probability

2

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, *Introduction to the Bootstrap*, 1998

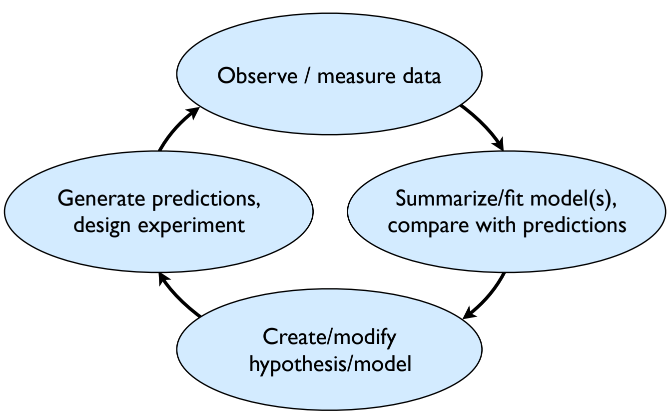
3

Some history…

- 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)
- Since 1950's! : statistical neural/cognitive models

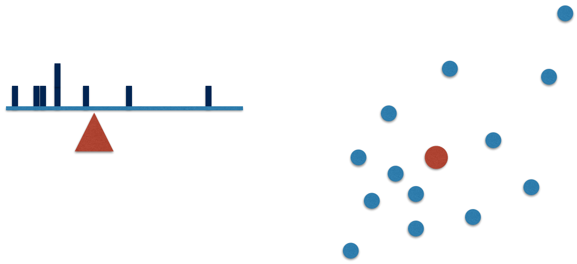
4

Scientific process



5

Descriptive statistics: Central tendency



6

Descriptive statistics: Central tendency

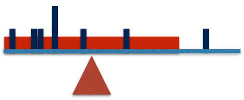
- We often summarize data with the average. Why?
- Average minimizes the squared error (as in regression!):

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

- Generalize: minimize L_p norm: $\arg\min_{c} \left[\frac{1}{N} \sum_{n=1}^{N} |x_n - c|^p \right]^{1/p}$
  - minimize L_1 norm: median, $m(\vec{x})$
  - minimize L_0 norm: mode
  - minimize L_∞ norm: midpoint of range
- Issues: outliers, asymmetry, bimodality
- How do we choose?

7

Descriptive statistics: Dispersion



8

Descriptive statistics: Dispersion

- Sample standard deviation

$$\sigma(\vec{x}) = \min_c \left[\frac{1}{N} \sum_{n=1}^{N} (x_n - c)^2 \right]^{1/2}$$
$$= \left[\frac{1}{N} \sum_{n=1}^{N} (x_n - \mu(\vec{x}))^2 \right]^{1/2}$$

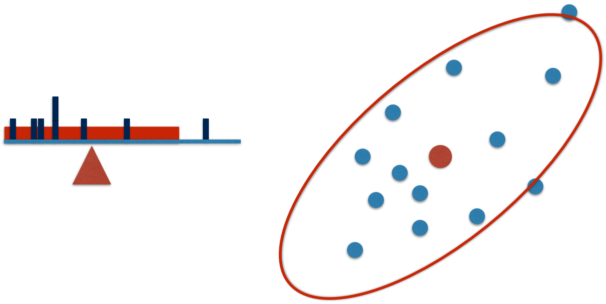
- Mean absolute deviation (MAD) about the median

$$d(\vec{x}) = \frac{1}{N} \sum_{n=1}^{N} \left| x_n - m(\vec{x}) \right|$$

- Quantiles

9

Descriptive statistics: Dispersion

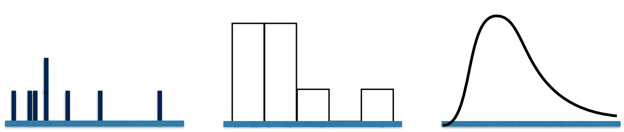


Summary statistics (eg: sample mean/var) can be interpreted as *estimates of model parameters*

To formalize this, we need tools from *probability*…

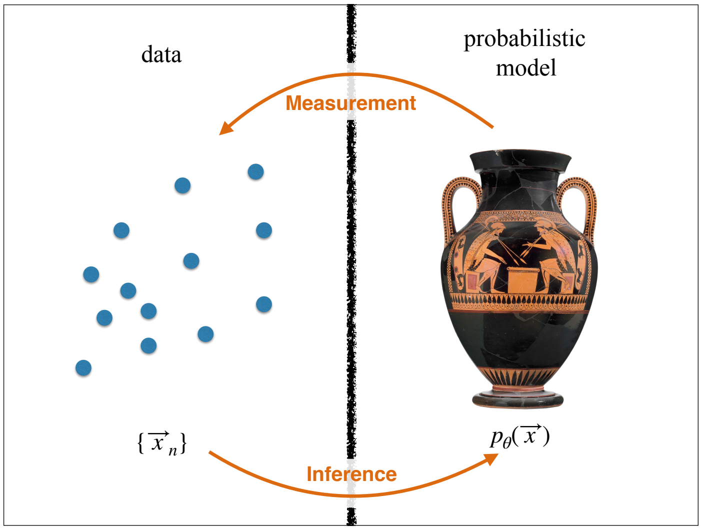
10

data → histogram → probability distribution



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

11



12

Probabilistic Middleville

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

The stork delivers boys and girls randomly, with family probability {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.

data

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

inference

13

Statistical Middleville

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

~~The stork delivers boys and girls randomly, with family probability {BB,BG,GB,GG}={0.2,0.3,0.2,0.3}~~

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

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

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

14

Probability basics (outline)

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

15

Probability: Definitions/not

Useful to have this notation up on slid, while introducing concepts on board

let *X, Y, Z* be random variables

they can take on values (like 'heads' or 'tails'; or integers 1-6; or real-valued numbers)

let *x, y, z* stand generically for values they can take, and denote events such as $X = x$

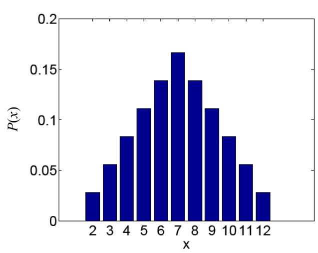
write the probability that *X* takes on value *x* as $P(X = x)$, *or* $P_X(x)$, or sometimes just $P(x)$

$P(x)$ is a function over values x, which we call the probability "distribution" function (pdf) (for continuous variables, "density")

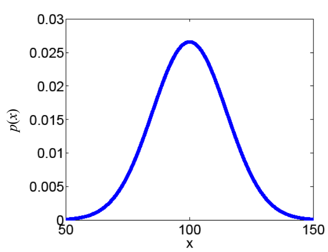
16

Probability distributions

Discrete random variable



Continuous random variable



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

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

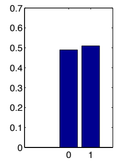
$$0 < p(x)$$

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

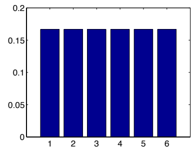
Example distributions

17

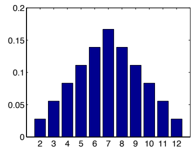
a not-quite-fair coin



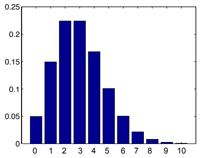
roll of a fair die



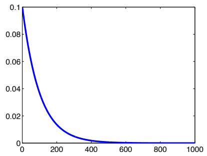
sum of two rolled fair dice



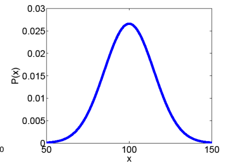
clicks of a Geiger counter, in a fixed time interval



... and, time between clicks



horizontal velocity of gas molecules exiting a fan

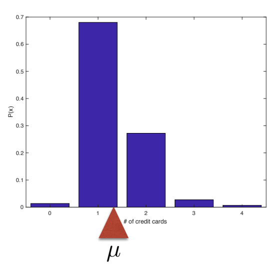


Expected value - discrete

18

$$E(X) = \sum_{i=1}^{N} x_i p(x_i) \qquad \text{[the mean, } \mu\text{]}$$

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



Expected value - continuous

19

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

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

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

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

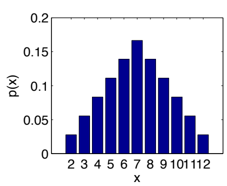
$$E\left(f(x)\right) = \int f(x) \; p(x) \; dx \qquad \text{[“expected value of } f\text{”]}$$

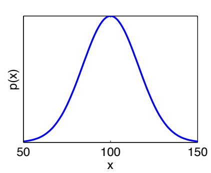
Note: this is an inner product, and thus *linear:*

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

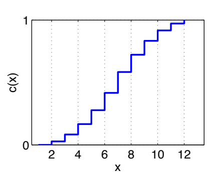
Cumulatives

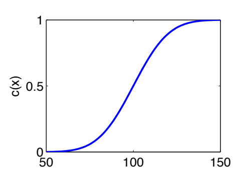
20



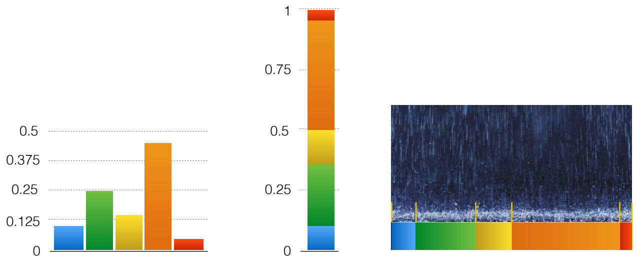


$$c(y) = \int_{-\infty}^{y} p(x)dx$$





Drawing samples - discrete



21

Multi-variate probability

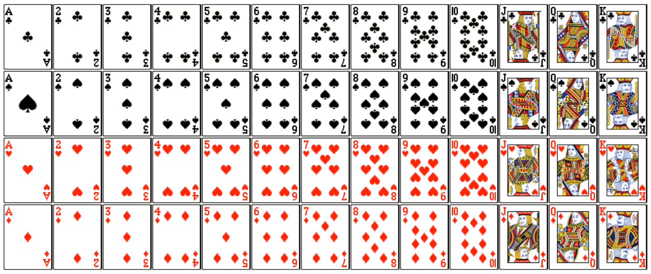
- joint distributions
- marginals (integrating)
- conditionals (slicing)
- Bayes' rule (inverse probability)
- statistical independence (separability)
- linear transformations

[on board]

22

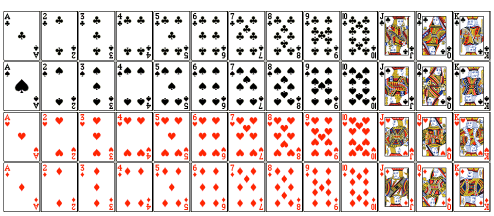
23

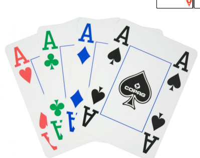
Joint and conditional probability - discrete



24

Joint and conditional probability - discrete



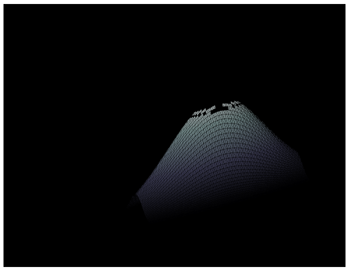


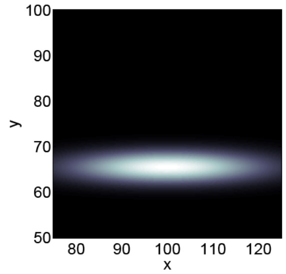
P(Ace)
P(Heart)
P(Ace & Heart)
P(Ace | Heart)
P(not Jack of Diamonds)
P(Ace | not Jack of Diamonds)

“Independence”

25

Joint distribution (continuous)





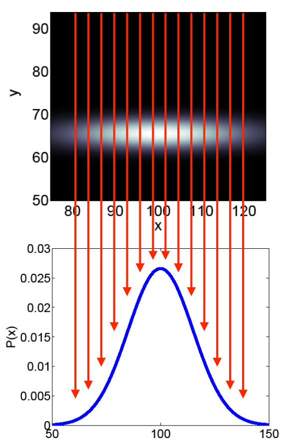
$p(x, y)$

26

Marginal distribution

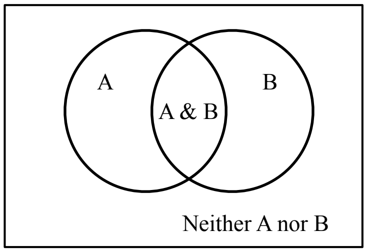
$p(x, y)$

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



27

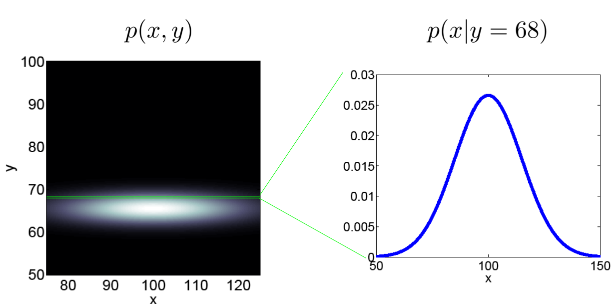
Conditional probability



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

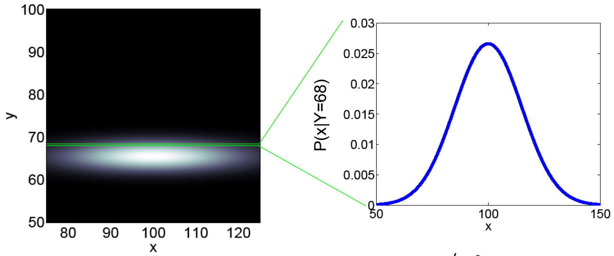
28

Conditional distribution



29

Conditional distribution



$$p(x|y=68) = p(x, y=68) \Big/ \int p(x, y=68)dx$$

$$= p(x, y=68) / p(y=68)$$

slide joint distribution

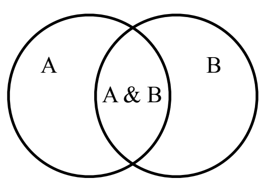
normalize (by marginal)

More generally:

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

30

Bayes' Rule



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

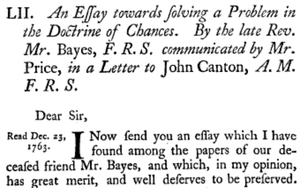
$$p(A \,\&\, B) = p(B)p(A \mid B)$$

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

$$\Rightarrow p(A \mid B) = \frac{p(B \mid A)p(A)}{p(B)}$$

31

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.

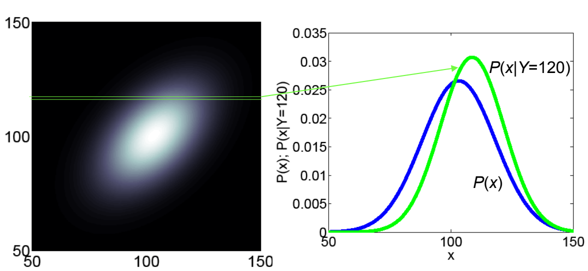
I Now send you an essay which I have found among the papers of our deceased friend Mr. Bayes, and which, in my opinion, has great merit, and well deserves to be preserved.

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

(a direct consequence of the definition of conditional probability)

32

Conditional vs. marginal

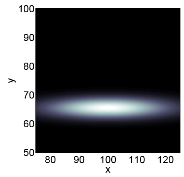


In general, the marginals for different Y values differ.

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

33

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:

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

34

Sums of RVs

Let $Z = X + Y$. Since expectation is linear:

$$E(X+Y) = E(X) + E(Y)$$

In addition, if X and Y are independent, then

$$E(XY) = E(X)E(Y)$$

$$\sigma_Z^2 = E\left(\left(\left(X+Y\right)-\left(\mu_X+\mu_Y\right)\right)^2\right) = \sigma_X^2 + \sigma_Y^2$$

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

[on board]

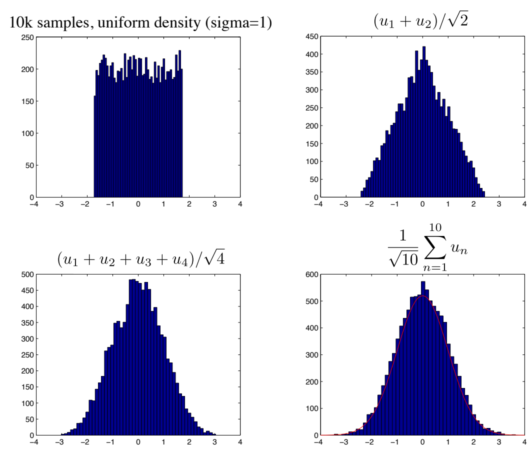
35

Mean and variance

- Mean and variance summarize the centroid/width
- Translation and rescaling of random variables
- Mean/variance of weighted sum of random variables
- The **sample average**
  - ... converges to true mean (except for bizarre distributions)
  - ... with variance σ^2/N
  - ... most common common choice for an **estimate** **...**

36

Central limit for a uniform distribution...



Central limit for a binary distribution...