Mathematical Tools
for Neural and Cognitive Science

Fall semester, 2017

Probability, Statistics and Inference

Probability: an abstract mathematical framework for describing random quantities (e.g., measurements)

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 equal probability. probabilistic model

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

What is the probability that the other child is a girl? statistical inference

Statistical Middleville

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

~~The stork delivers boys and girls randomly, with equal probability.~~

In a survey of 100 Middleville families, 32 have two girls, 24 have two boys, and the remainder have one of each.

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

What is the probability that the other child is a girl?

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.

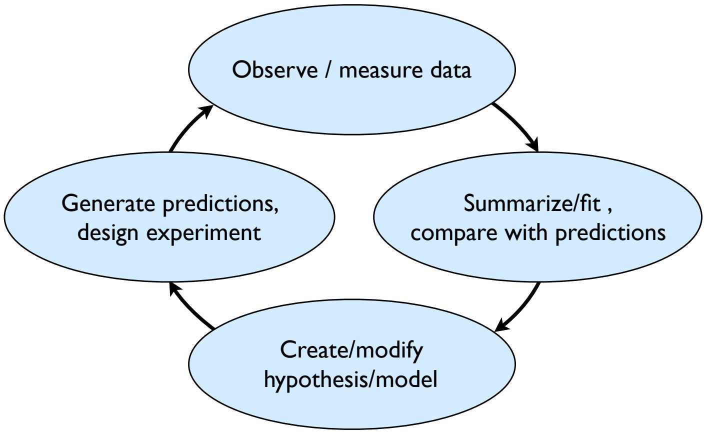
Most people are not natural-born statisticians. Left to our own devices we are not very good at picking out patterns from a sea of noisy data. To put it another way, we are all too good at picking out non-existent patterns that happen to suit our purposes. Statistical theory attacks the problem from 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

Some 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 (Shannon, Wiener, etc)
- 1970's: Computational optimization and simulation (e.g,. Tukey)
- 1990's: Machine learning (large-scale computing + statistical inference + lots of data)
- Since 1950's: statistical neural/cognitive models

Scientific process



Estimating model parameters

- How do I compute the estimate?
  (mathematics vs. numerical optimization)
- How “good” are my estimates?
- How well does my model explain the data?
  Future data (prediction/generalization)?
- How do I compare two (or more) models?

Outline of what's coming

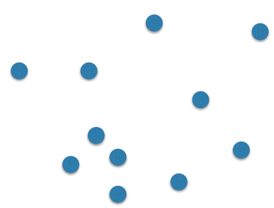
Themes:

- Uni-variate vs. multi-variate
- Discrete vs. continuous
- Math vs. simulation
- Bayesian vs. frequentist inference

Topics:

- Descriptive statistics
- Basic probability theory: univariate, multivariate
- Model parameter estimation
- Hypothesis testing / model comparison

Example: Localization



Issues: Mean and variability (accuracy and precision)

Descriptive statistics: Central tendency

- We often summarize data with the *average*. Why?
- Average minimizes the squared error (think regression!)

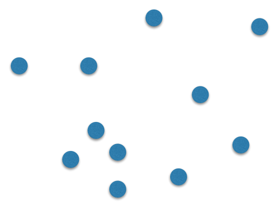
$$\arg\min_{\hat{x}} \frac{1}{N}\sum_{n=1}^{N}(x_n - \hat{x})^2 = \frac{1}{N}\sum_{n=1}^{N} x_n$$

- More generally, for L_p norms: $$\left[\frac{1}{N}\sum_{i=1}^{N}|x_n - \hat{x}|^p\right]^{1/p}$$
- minimum L_1 norm: median
- minimum L_0 norm: mode
- Issues: Data from a common source, outliers, asymmetry, bimodality

Descriptive statistics: Dispersion

- Sample variance $$s^2 = \frac{1}{N-1}\sum_{i=1}^{N}\left(x_i - \overline{x}\right)^2$$
- Why N-1?
- Sample standard deviation
- Mean absolute deviation $$\frac{1}{N}\sum_{i=1}^{N}\left|x_i - \overline{x}\right|$$

Example: Localization



I find that $\bar{x} \neq 0$. Is that convincing? Is the apparent bias real?

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

Probability: notation

let X, Y, Z be random variables

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

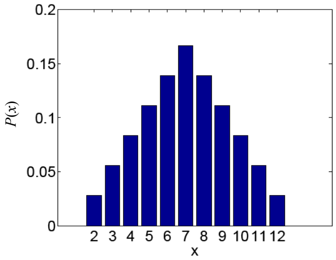
let x, y, z stand generically for values they can take,
and also, in shorthand, for events like $X = x$

we 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 x, which we call the probability "distribution" function (pdf) (or, for continuous variables only, "density")

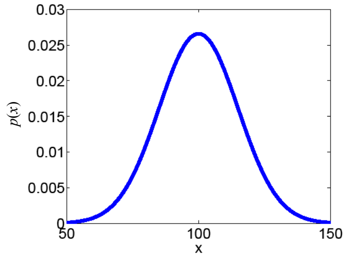
Discrete pdf

A distribution
(the sum of 2 dice rolls)

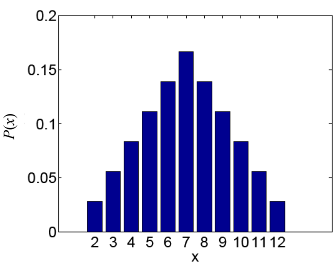


Continuous pdf

Another distribution
(IQ or a randomly chosen person)

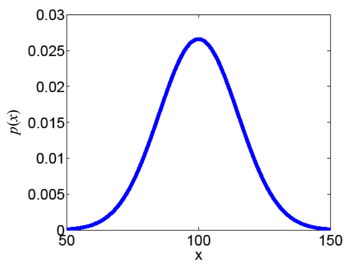


Normalization



$$0 < P(x) < 1$$

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



$$0 < p(x)$$

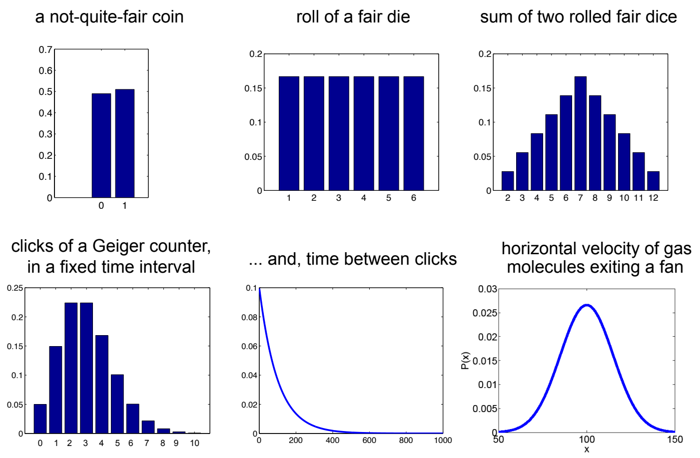
$$\int_{-\infty}^{\infty} p(x)\,dx = 1$$

Probability basics

- discrete probability distributions
- continuous probability densities
- cumulative distributions
- translation and scaling of distributions
- monotonic nonlinear transformations
- drawing samples from a distribution. Uniform. Inverse cumulative mapping
- example densities/distributions

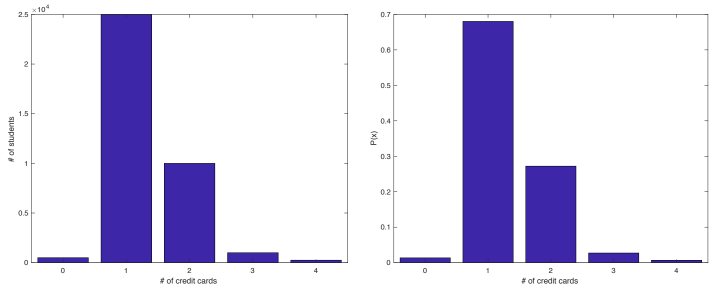
[on board]

Example distributions



Expected value - discrete

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



Expected value - continuous

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

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

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

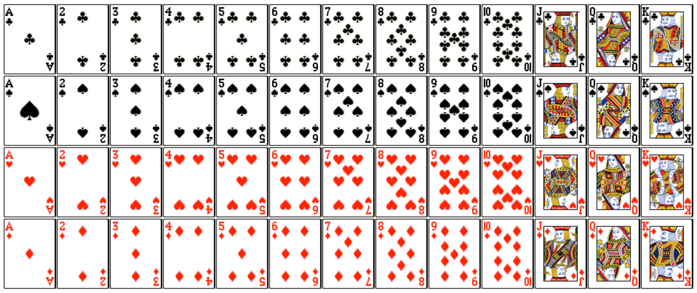
$$= \int x^2 \ p(x) \ dx - \mu^2$$

$$E\left(f(x)\right) = \int f(x) \ p(x) \ dx$$

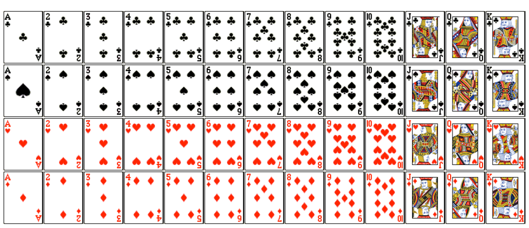
note: an inner product, and thus *linear,* i.e.,

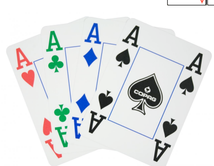
$$E(af(X) + bg(X)) = aE(f(X)) + bE(g(X))$$

Joint and conditional probability - discrete



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)

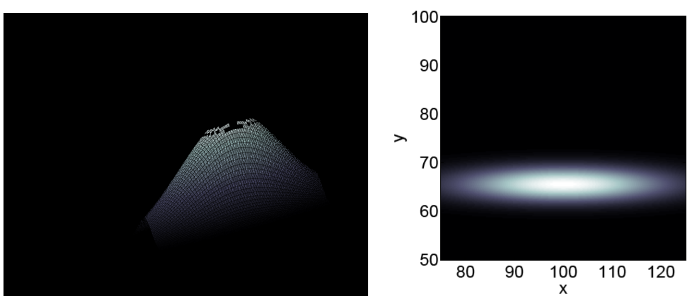
"Independence"

Multi-variate probability

- Joint distributions
- Marginals (integrating)
- Conditionals (slicing)
- Bayes' Rule (inverting)
- Statistical independence (separability)

[on board]

Joint distribution

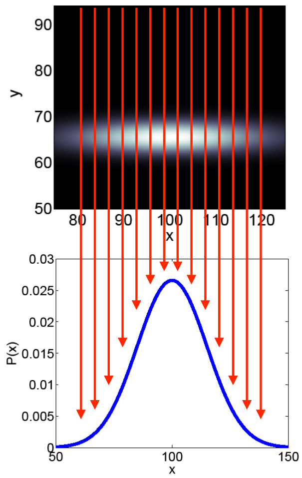


$p(x, y)$

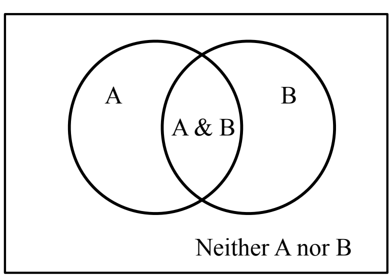
Marginal distribution

$p(x, y)$

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



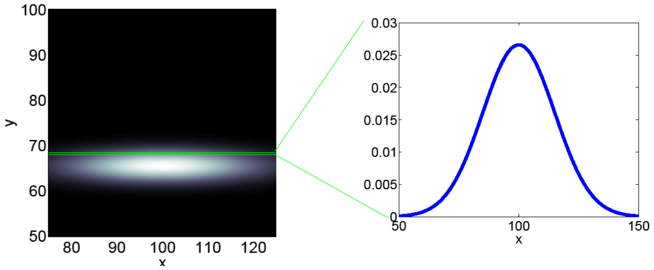
Conditional probability



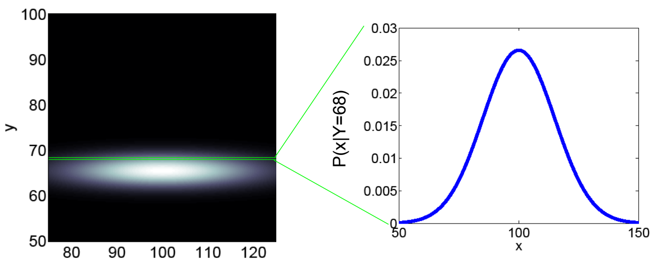
$p(A \mid B) =$ probability of A given that B is asserted to be true $= \dfrac{p(A \& B)}{p(B)}$

Conditional distribution

$p(x, y)$ $\qquad\qquad$ $p(x|y = 68)$



Conditional distribution



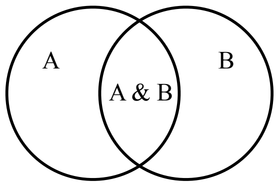
$$p(x|y = 68) = p(x, y = 68) \Big/ \int p(x, y = 68)dx$$

$$= \underbrace{p(x, y = 68)}_{\text{slice joint distribution}} \Big/ \underbrace{p(y = 68)}_{\text{normalize (by marginal)}}$$

More generally:

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

Bayes' Rule



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

$$\begin{aligned} p(A \,\&\, B) &= p(B)\,p(A \mid B) \\ &= p(A)\,p(B \mid A) \\ \Rightarrow p(A \mid B) &= \frac{p(B \mid A)\,p(A)}{p(B)} \end{aligned}$$

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,

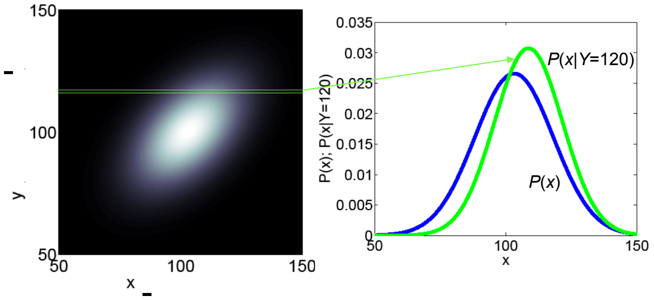
Read Dec. 23, 1763.

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)

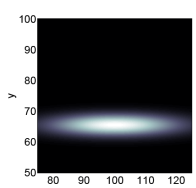
Conditional vs. marginal



In general, these 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:

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

Sums of independent RVs

For *any* two random variables (independent or not):

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

Suppose X and Y are independent, then

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

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

and $p_{X+Y}(z)$ is a convolution

Implications: (1) Sums of Gaussians are Gaussian,
(2) Properties of the sample average

Mean and variance

- Mean and variance summarize centroid/width
  - translation and rescaling of random variables
  - nonlinear transformations - "warping"
- 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** **...**

Point Estimates

- Estimator: Any function of the data, intended to compute an estimate of the true value of a parameter

- The most common estimator is the sample average, used to estimate the true mean of the distribution.

- Statistically-motivated examples: $\hat{x}(\vec{d}) = \arg\max_x p(\vec{d}|x)$
  - Maximum likelihood (ML): $\hat{x}(\vec{d}) = \arg\max_x p(x|\vec{d})$
  - Max a posteriori (MAP): $\hat{x}(\vec{d}) = \arg\min_{\hat{x}} \mathbf{E}\left((x-\hat{x})^2|\vec{d}\right)$
  - Min Mean Squared Error (MMSE): $= \mathbf{E}\left(x|\vec{d}\right)$

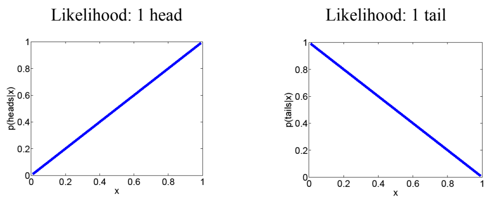
Example: Estimate the bias of a coin

Bayes' Rule and Estimation

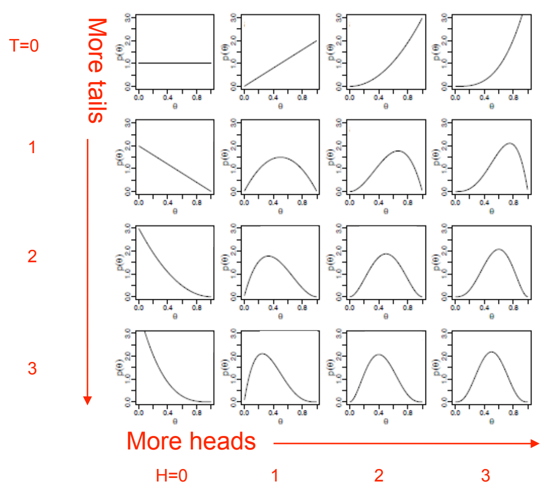
Posterior Likelihood Prior

$$p(\text{parameter value} \mid \text{data}) = \frac{p(\text{data} \mid \text{parameter value})\, p(\text{parameter value})}{p(\text{data})}$$

Nuisance normalizing term



Posteriors, p(H,T|x), assuming prior p(x)=1

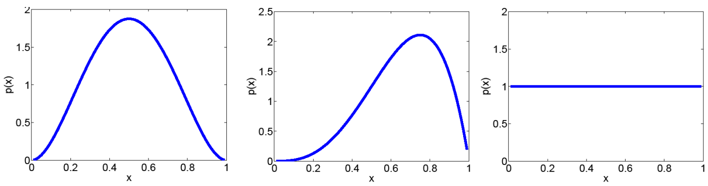


example

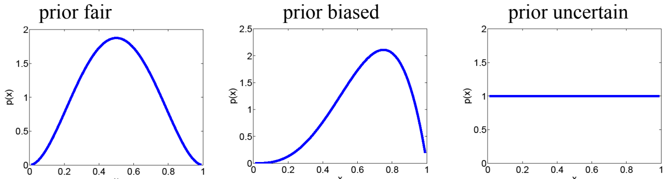
infer whether a coin is fair by flipping it repeatedly
here, x is the probability of heads (50% is fair)
$y_{1 \dots n}$ are the outcomes of flips

Consider three different priors:

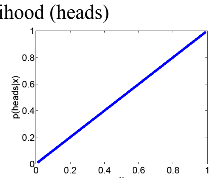
suspect fair | suspect biased | no idea



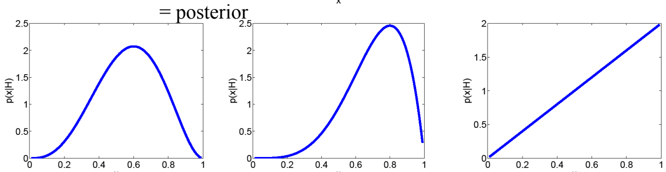
prior fair | prior biased | prior uncertain

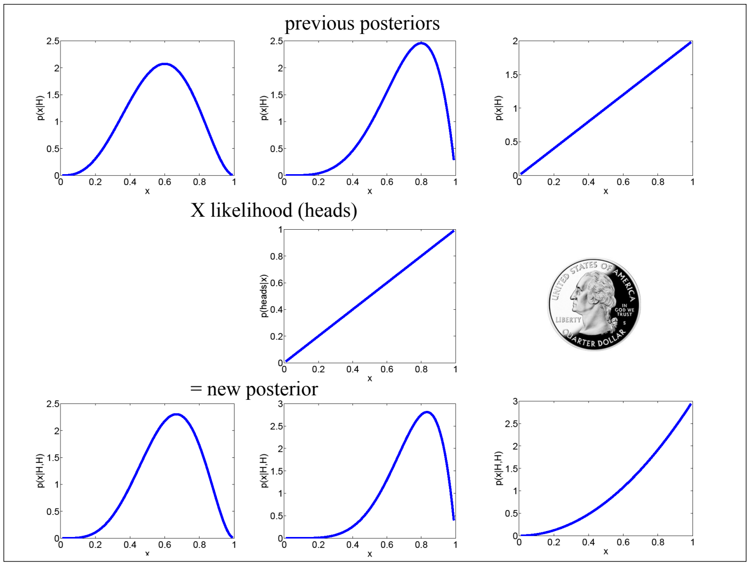


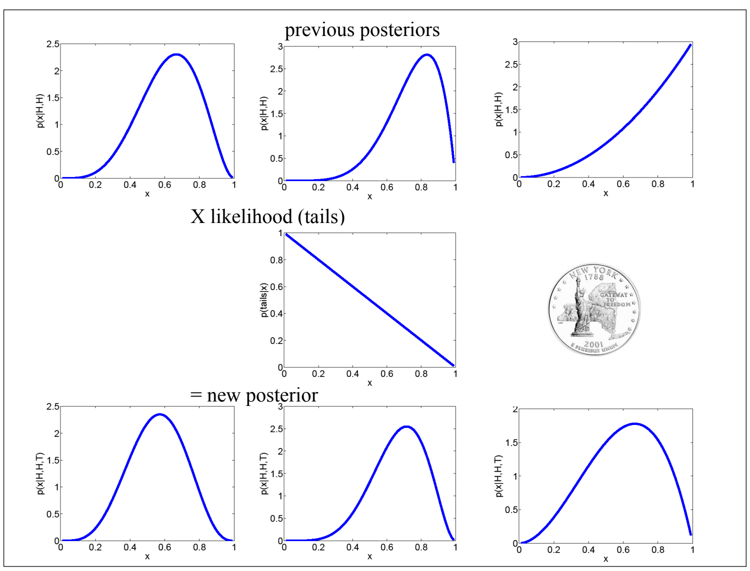
X likelihood (heads)



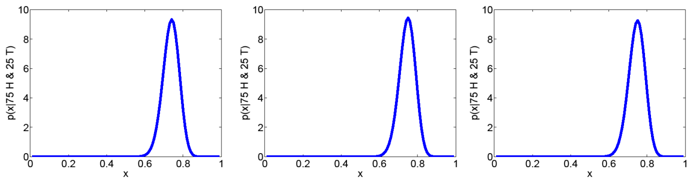
= posterior




previous posteriors
p(x|H)
X likelihood (heads)
p(heads|x)
UNITED STATES OF AMERICA
LIBERTY
IN GOD WE TRUST
QUARTER DOLLAR
= new posterior
p(x|H,H)
x


previous posteriors
p(x|H,H)
X likelihood (tails)
p(tails|x)
NEW YORK
1788
GATEWAY TO FREEDOM
2001
E PLURIBUS UNUM
= new posterior
p(x|H,H,T)
x

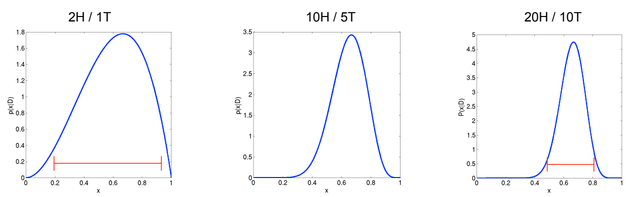
Posteriors after observing 75 heads, 25 tails



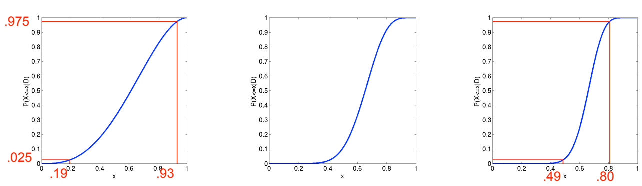
→ prior differences are ultimately overwhelmed by data

Confidence

PDFs



CDFs



Bias & Variance

- Mean squared error = bias^2 + variance
- Bias is difficult to assess (requires knowing the "true" value). Variance is easier.
- Classical statistics generally aims for an unbiased estimator, with minimal variance ("MVUE").
- The MLE is *asymptotically* unbiased (under fairly general conditions), but this is only useful if
  - the likelihood model is correct
  - the optimum can be computed
  - you have enough data
- More general/modern view: estimation is about trading off bias and variance, through model selection, "regularization", or Bayesian priors.

Bayesian Model Comparison

- Is the coin fair? Compared to what?
- Point hypotheses: $M_1 : p = p_1 = 0.5 \quad M_2 : p = p_2 = 0.6$

$$p(M_1 \mid D) = \frac{p(D \mid M_1)P(M_1)}{p(D)} = \frac{p(D \mid M_1)P(p_1)}{p(D)}$$

Assuming equal priors over models the *Bayes factor* is

$$\frac{p(M_1 \mid D)}{p(M_2 \mid D)} = \frac{p(D \mid M_1)P(M_1)}{p(D \mid M_2)P(M_2)} = \frac{p(D \mid M_1)P(p_1)}{p(D \mid M_2)P(p_2)}$$

Bayesian Model Comparison

- Is the coin fair? Compared to what?
- Alternative hypothesis: $M_1 : p = p_1 = 0.5 \quad M_2 : p \neq 0.5$

$$p(M_2 \mid D) = \frac{p(D \mid M_2)p(M_2)}{p(D)}$$

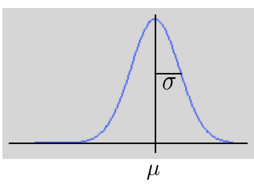
$$= \int_0^1 p(p_{\text{coin}} \mid D)p(p_{\text{coin}})\,dp_{\text{coin}}$$

$$= \frac{\int_0^1 p(D \mid M_2, p_{\text{coin}})p(p_{\text{coin}})\,dp_{\text{coin}} P(M_2)}{p(D)}$$

Compute *Bayes factor* as before.

The Gaussian

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

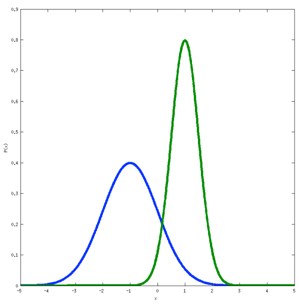


- parameterized by mean and stdev (position / width)
- joint density of two indep Gaussian RVs is circular! [easy]
- product of two Gaussians is Gaussian! [easy]
- conditionals of a Gaussian are Gaussian! [easy]
- sum of Gaussian RVs is Gaussian! [moderate]
- marginals of a Gaussian are Gaussian! [moderate]
- central limit theorem: sum of many RVs is Gaussian! [hard]
- most random (max entropy) density with this variance! [moderate]

Product of Gaussians is Gaussian

$$y = x + n, \quad x \sim N(\mu_x, \sigma_x), \; n \sim N(0, \sigma_n)$$

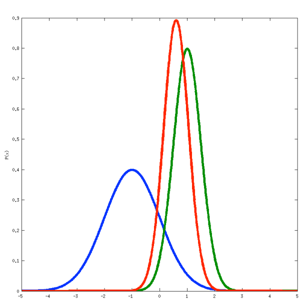
$$p(x|y) \;\propto\; \underline{p(y|x)}\,\underline{p(x)}$$



Product of Gaussians is Gaussian

$$y = x + n, \quad x \sim N(\mu_x, \sigma_x), \; n \sim N(0, \sigma_n)$$

$$\begin{aligned} \underline{p(x|y)} &\propto \underline{p(y|x)}\,\underline{p(x)} \\ &\propto e^{-\frac{1}{2}\left[\frac{1}{\sigma_n^2}(x-y)^2\right]} e^{-\frac{1}{2}\left[\frac{1}{\sigma_x^2}(x-\mu_x)^2\right]} \\ &= e^{-\frac{1}{2}\left[\left(\frac{1}{\sigma_n^2}+\frac{1}{\sigma_x^2}\right)x^2 - 2\left(\frac{y}{\sigma_n^2}+\frac{\mu_x}{\sigma_x^2}\right)x + \ldots\right]} \end{aligned}$$



Completing the square shows that this posterior is also Gaussian, with

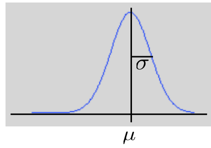
$$\sigma^2 = 1 \Big/ \left(\frac{1}{\sigma_n^2} + \frac{1}{\sigma_x^2}\right)$$

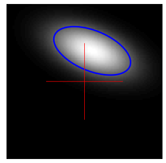
$$\mu = \left(\frac{y}{\sigma_n^2} + \frac{\mu_x}{\sigma_x^2}\right) \Big/ \left(\frac{1}{\sigma_n^2} + \frac{1}{\sigma_x^2}\right)$$

(average, weighted by *inverse* variances!)

Gaussian densities

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



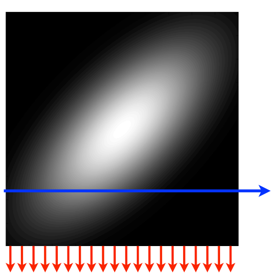


$$p(\vec{x}) = \frac{1}{\sqrt{(2\pi)^N |C|}}\, e^{-(\vec{x}-\vec{\mu})^T C^{-1} (\vec{x}-\vec{\mu})/2}$$

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

$\vec{x} \sim N(\vec{\mu}, C)$, let $P = C^{-1}$ (known as the "precision" matrix)

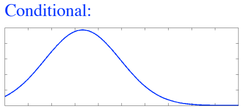
$$\begin{aligned} p(x_1 | x_2 = a) &\propto e^{-\frac{1}{2}\left[P_{11}(x_1-\mu_1)^2 - 2P_{12}(x_1-\mu_1)(a-\mu_2) + \ldots\right]} \\ &= e^{-\frac{1}{2}\left[P_{11}x_1^2 - 2(P_{11}\mu_1 + P_{12}(a-\mu_2))x_1 + \ldots\right]} \end{aligned}$$



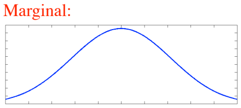
Gaussian, with: $\mu = \mu_1 + \frac{P_{12}}{P_{11}}(a - \mu_2)$

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

Conditional:



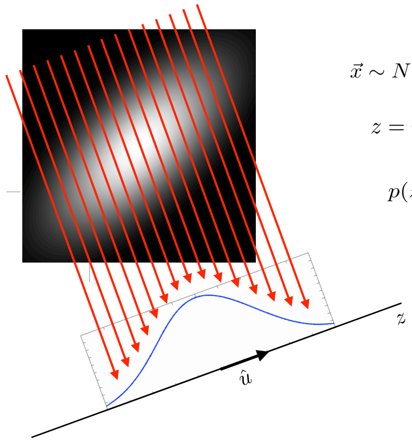
Marginal:



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

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

Generalized marginals of a Gaussian

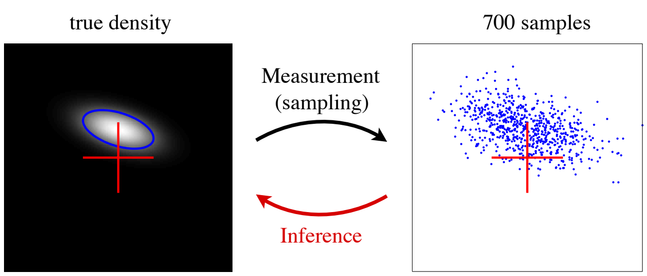


$$\vec{x} \sim N(\vec{\mu}_x, C_x)$$

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

$p(z)$ is Gaussian, with:

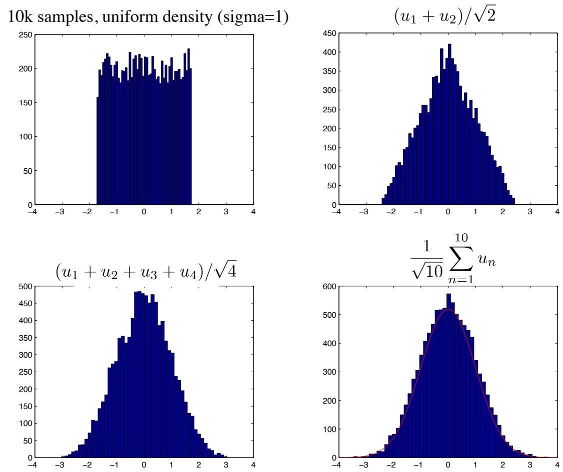
$$\begin{aligned} \mu_z &= \hat{u}^T \vec{\mu}_x \\ \sigma_z^2 &= \hat{u}^T C_x \hat{u} \end{aligned}$$



true mean: [0 0.8]
true cov: [1.0 -0.25
-0.25 0.3]

sample mean: [-0.05 0.83]
sample cov: [0.95 -0.23
-0.23 0.29]

Central limit for a uniform distribution...



Central limit for a binary distribution...