Mathematical Tools
for Neural and Cognitive Science

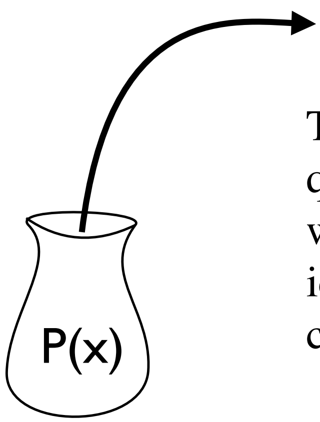
Fall semester, 2016

Section 4:
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.**

Statistics as a form of summary



0, 1, 0, 0, 0, 1, 0, 1, ...

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

Statistics for Data Summary...

- Sample average (minimizes mean squared error)
- Sample median (minimizes mean absolute deviation)
- Least-squares regression - summarizes relationships between controlled and measured quantities
- TLS regression - summarizes relationships between measured quantities

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.

Statistical theory attempts to answer three basic questions:

(1) How should I collect my data?

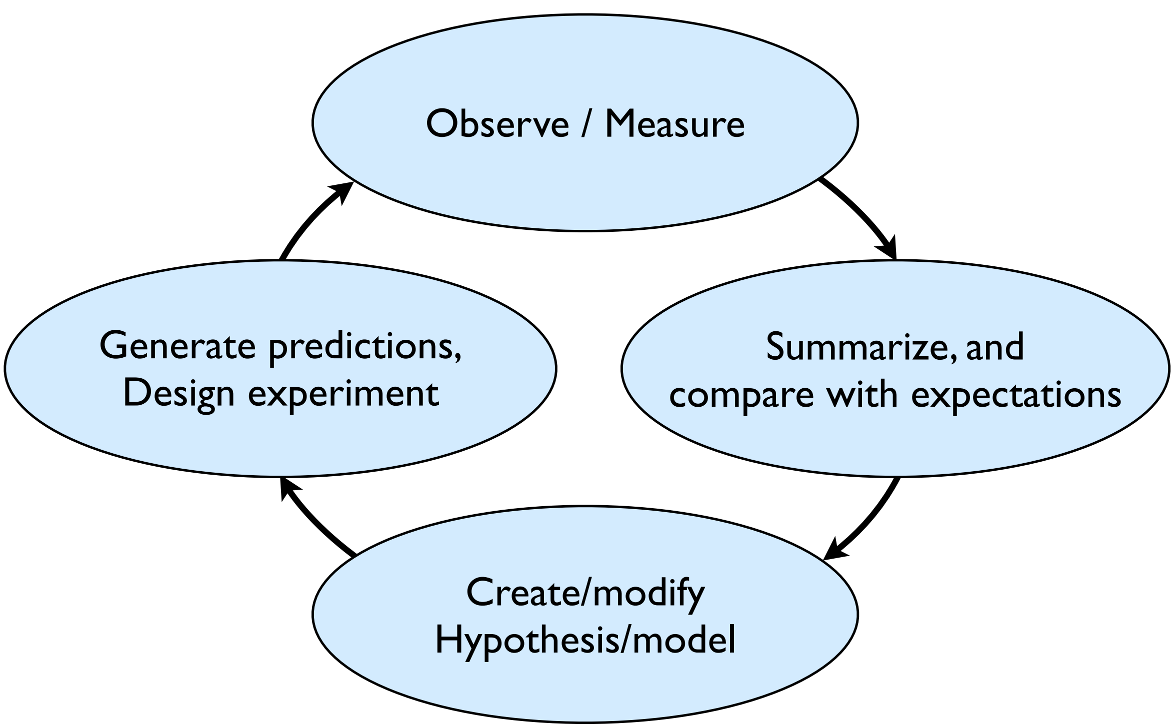
(2) How should I analyze and summarize the data that I've collected?

(3) How accurate are my data summaries?

Question 3 constitutes part of the process known as statistical inference.

- Efron & Tibshirani, Introduction to the Bootstrap

Scientific process



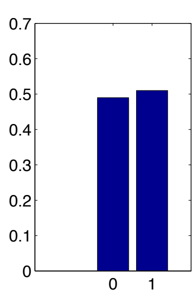
Probability basics

- discrete probability distributions
- continuous probability densities
- cumulative distributions
- translation and scaling of distributions (adding or multiplying by a constant)
- monotonic nonlinear transformations
- drawing samples from a distribution via inverse cumulative mapping
- example densities/distributions

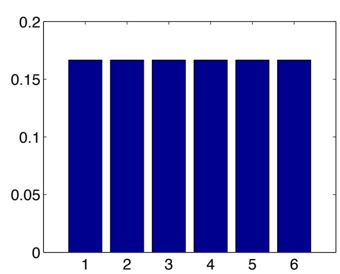
[on board]

Example distributions

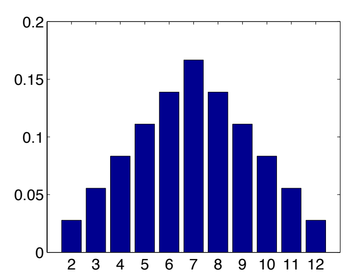
a not-quite-fair coin



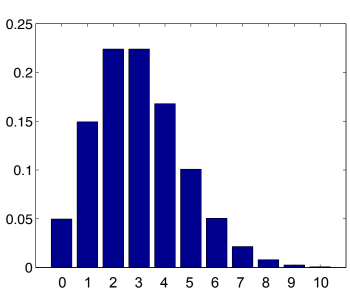
roll of a fair die



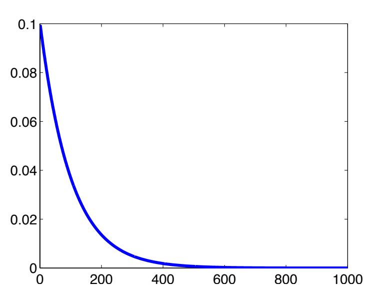
sum of two rolled fair dice



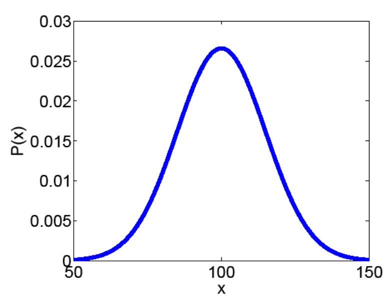
clicks of a Geiger counter, in a fixed time interval



... and, time between clicks



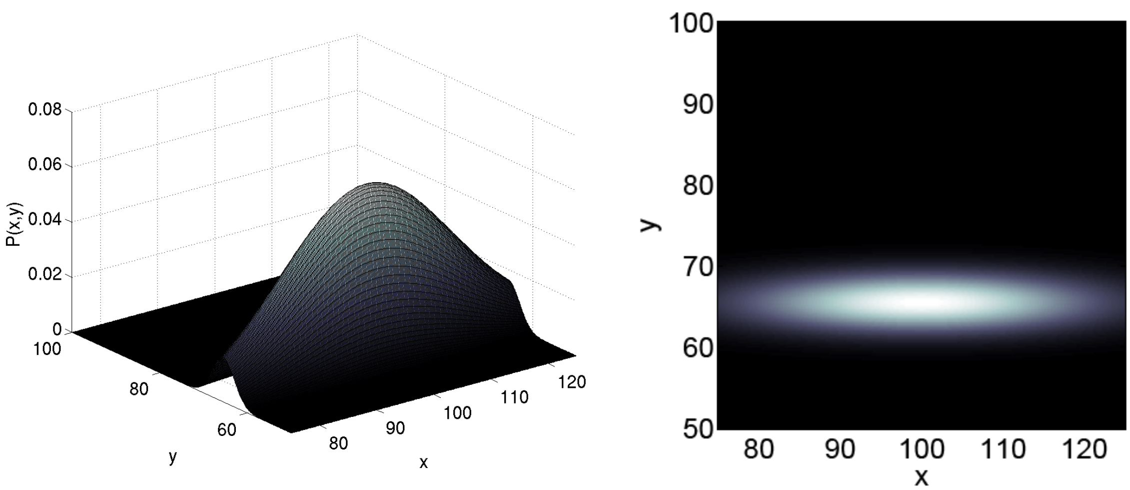
horizontal velocity of gas molecules exiting a fan



Multi-dimensional random variables

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

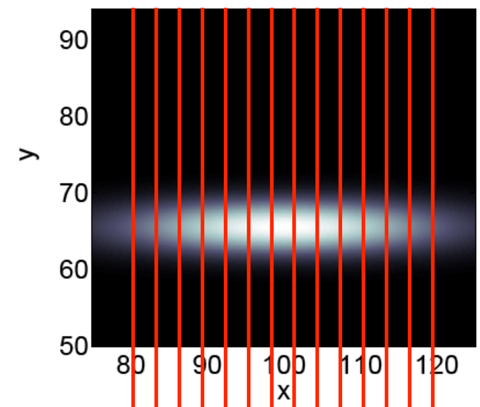
Joint distribution



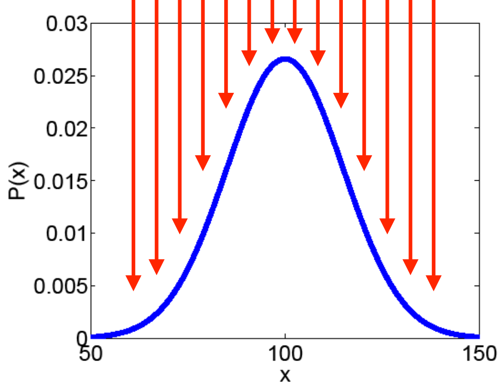
$$p(x, y)$$

Marginal distribution

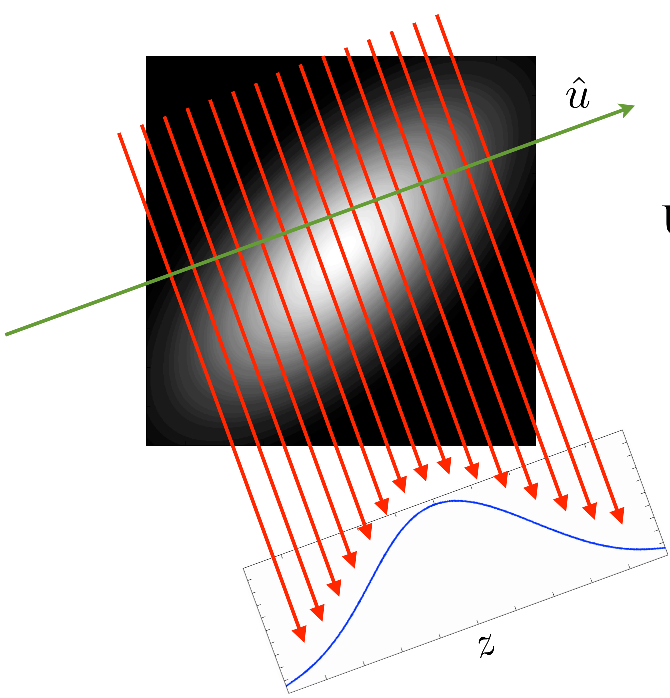
$$p(x, y)$$



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



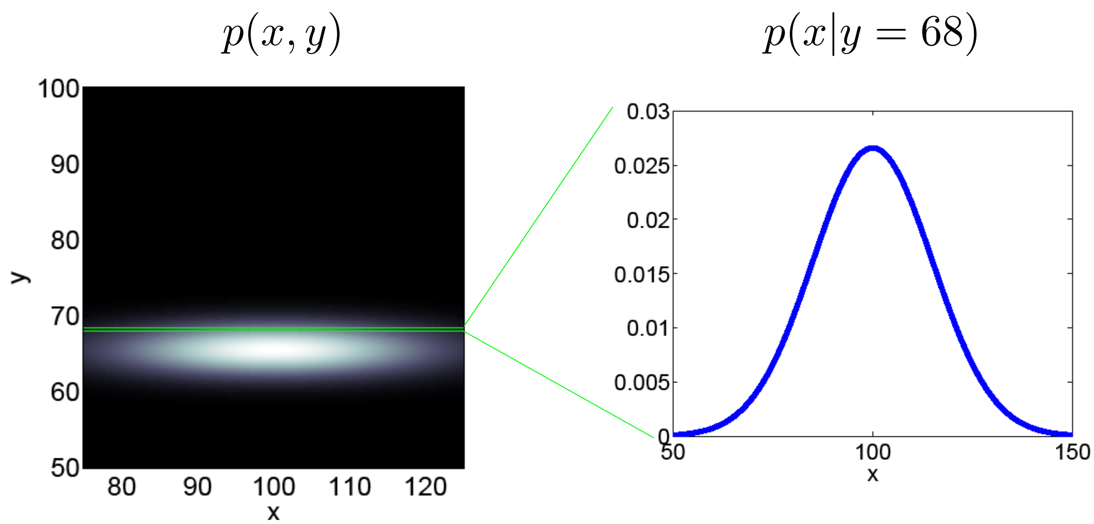
Generalized marginal distribution



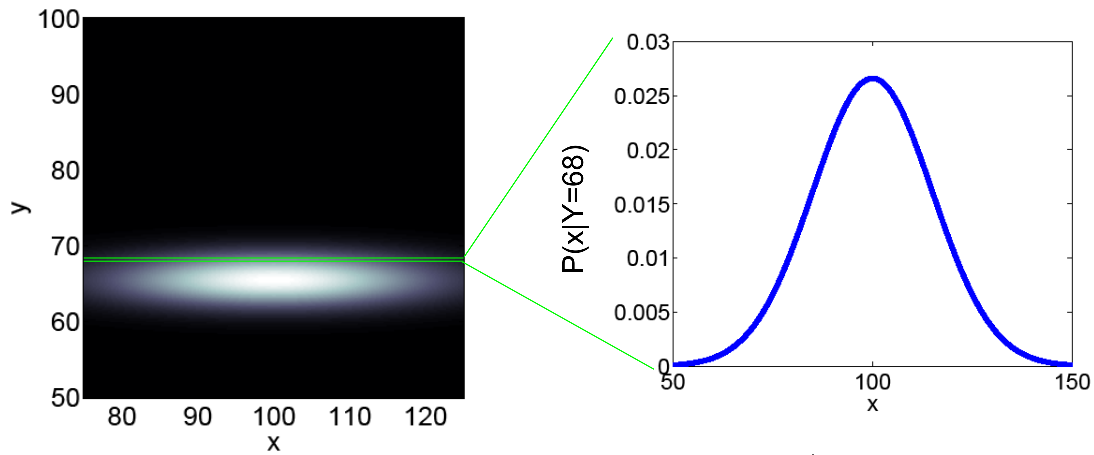
Using vector notation:

$$p(z) = \int_{\vec{x} \cdot \hat{u} = z} p(\vec{x}) d\vec{x}$$

Conditional distribution



Conditional distribution



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

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

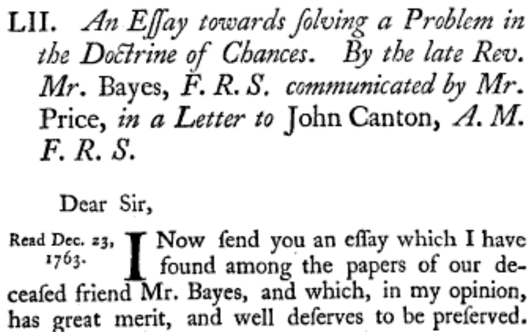
slice joint distribution

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,

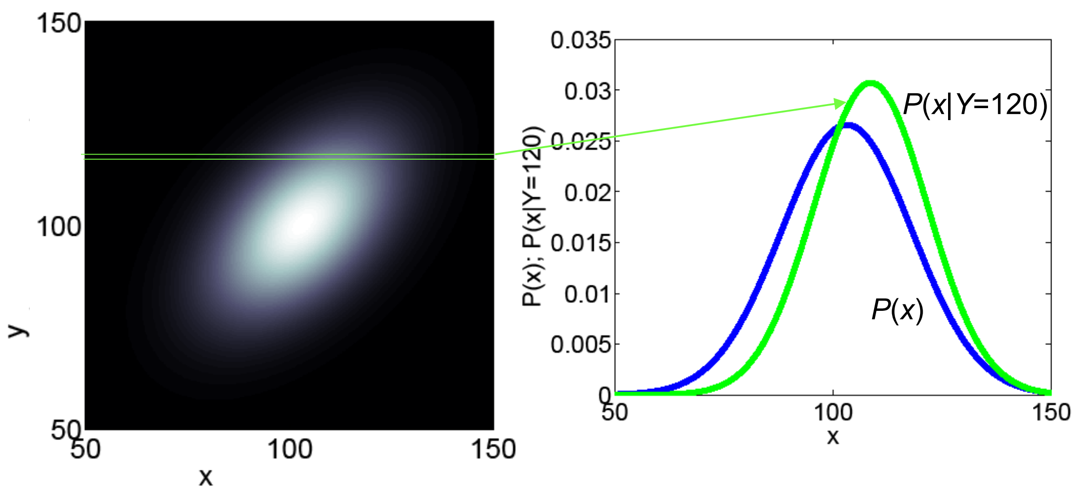
Read Dec. 23, 1763.

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)

Conditional vs. marginal



In general, these differ.

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

Statistical independence

Variables x and y are statistically independent if (and only if):

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

Independence implies that all condionals are equal to the corresponding marginal:

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

Uncorrelated doesn't mean independent...

Statistical independence a stronger assumption uncorrelatedness

⇒All independent variables are uncorrelated

⇒Not all uncorrelated variables are independent:

Expected value

$$E(x) = \int x \; p(x) \; dx$$ [the mean, μ]

$$E(x^2) = \int x^2 \; p(x) \; dx$$ [the "second moment"]

$$E\left((x-\mu)^2\right) = \int (x-\mu)^2 \; p(x) \; dx$$ [the variance, σ^2]

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

$$E\left(f(x)\right) = \int f(x) \; p(x) \; dx$$ [note: an inner product, and thus *linear*!]

Mean and (co)variance

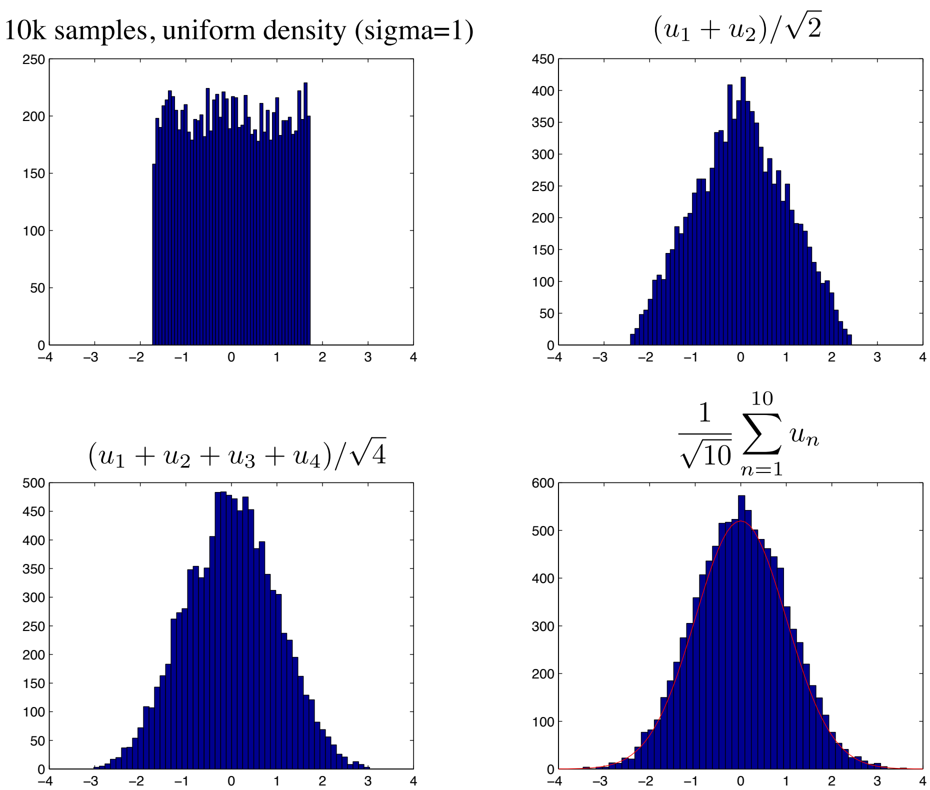
- One-D: mean and covariance summarize centroid/width
  - translation and rescaling of random variables
  - nonlinear transformations - "warping"
- Multi-D: vector mean and *covariance* matrix, elliptical geometry
- Mean/variance of weighted sum of random variables
- The **sample average**
  - ... converges to true mean (except for bizarre distributions)
  - ... with variance σ^2/N
  - ... most common common choice for an **estimate ...**
- Correlation

Distribution of a sum of independent R.V.'s - the return of convolution

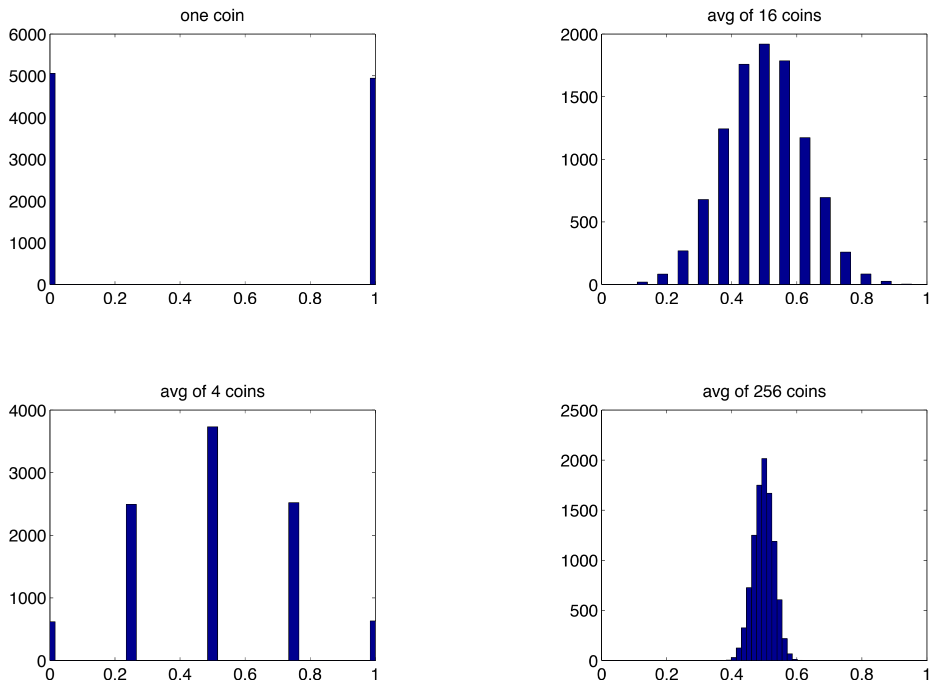
The Central Limit Theorem

[on board]

Central limit for a uniform distribution...

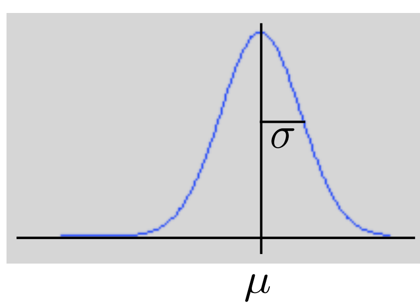


Central limit for a binary distribution...



The Gaussian

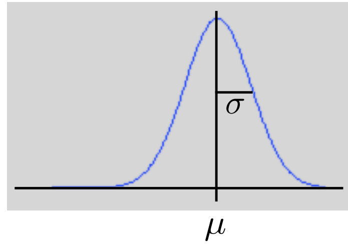
$$p(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \; e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

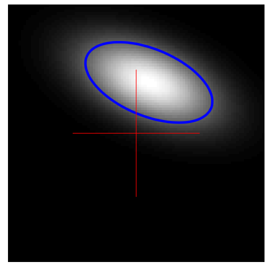


- parameterized by mean and stdev (position / width)
- joint density of two indep Gaussian RVs is circular! [easy]
- product of two Gaussians is Gaussian! [easy]
- conditionals of a Gaussian are Gaussian! [easy]
- sum of Gaussian RVs is Gaussian! [moderate]
- marginals of a Gaussian are Gaussian! [moderate]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

Gaussian densities

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





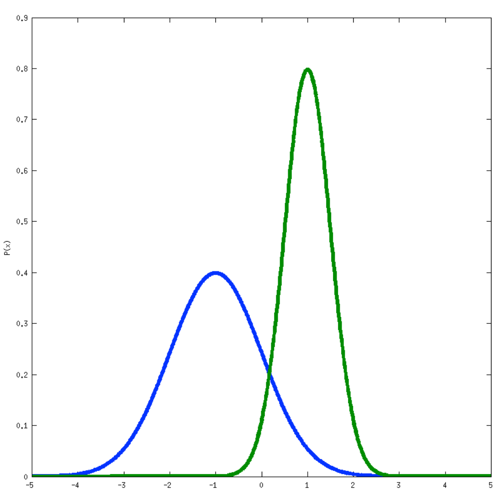
mean: $[0.2, 0.8]$
cov: [1.0 -0.3;
-0.3 0.4]

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

Product of Gaussians is Gaussian

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

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



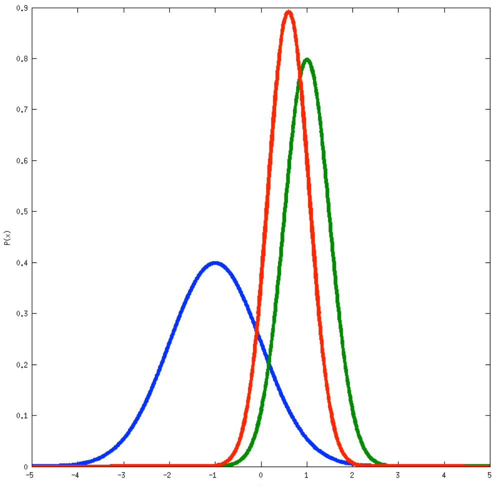
Product of Gaussians is Gaussian

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

$$p(x|y) \quad \propto \quad p(y|x)p(x)$$

$$\propto \quad e^{-\frac{1}{2}\left[\frac{1}{\sigma_n^2}(x-y)^2\right]} e^{-\frac{1}{2}\left[\frac{1}{\sigma_x^2}(x-\mu_x)^2\right]}$$

$$= \quad e^{-\frac{1}{2}\left[\left(\frac{1}{\sigma_n^2}+\frac{1}{\sigma_x^2}\right)x^2 - 2\left(\frac{y}{\sigma_n^2}+\frac{\mu_x}{\sigma_x^2}\right)x + \ldots\right]}$$



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

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

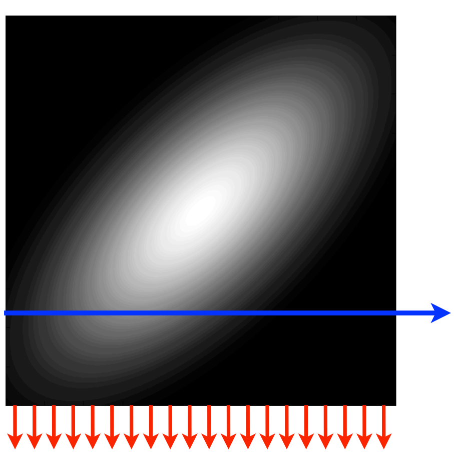
$$\mu = \left(\frac{y}{\sigma_n^2} + \frac{\mu_x}{\sigma_x^2}\right) \Bigg/ \left(\frac{1}{\sigma_n^2} + \frac{1}{\sigma_x^2}\right)$$

(average, weighted by *inverse* variances!)

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

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

$$= \quad e^{-\frac{1}{2}\left[P_{11}x_1^2 - 2(P_{11}\mu_1 + P_{12}(a-\mu_2))x_1 + \ldots\right]}$$

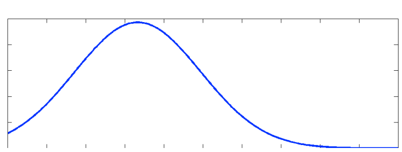


Gaussian, with:

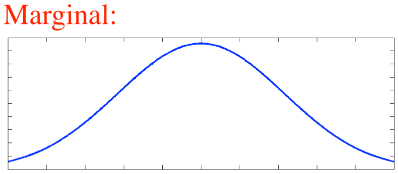
$$\mu = \mu_1 + \frac{P_{12}}{P_{11}}(a - \mu_2)$$

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

Conditional:



Marginal:



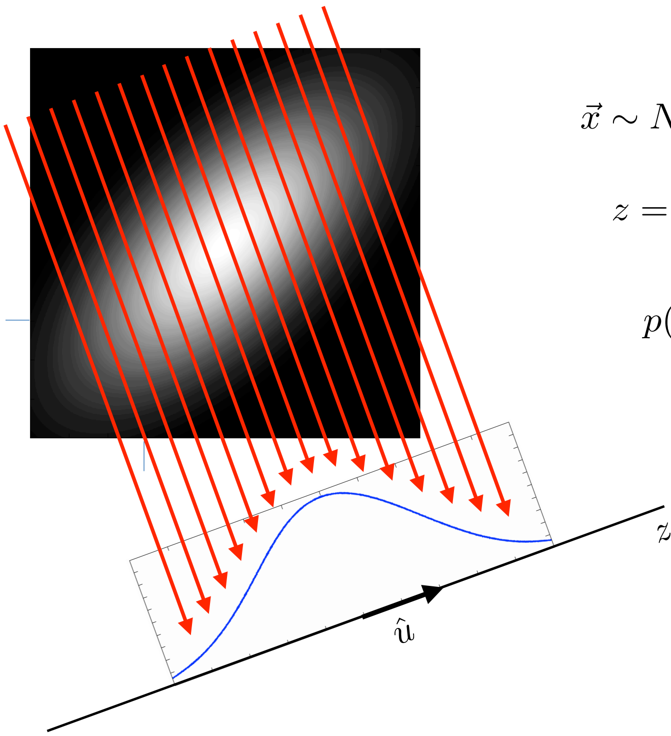
$$p(x_1) = \int p(\vec{x}dx_2$$

Gaussian, with:

$$\mu = \mu_1$$

$$\sigma^2 = C_{11}$$

Generalized marginals of a Gaussian

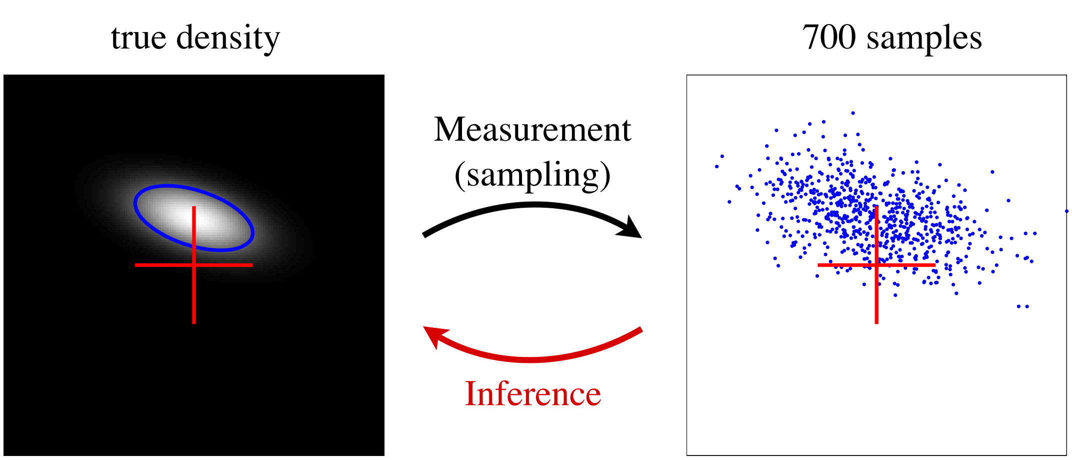


$$\vec{x} \sim N(\vec{\mu}_x, C_x)$$

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

$p(z)$ is Gaussian, with:

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



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

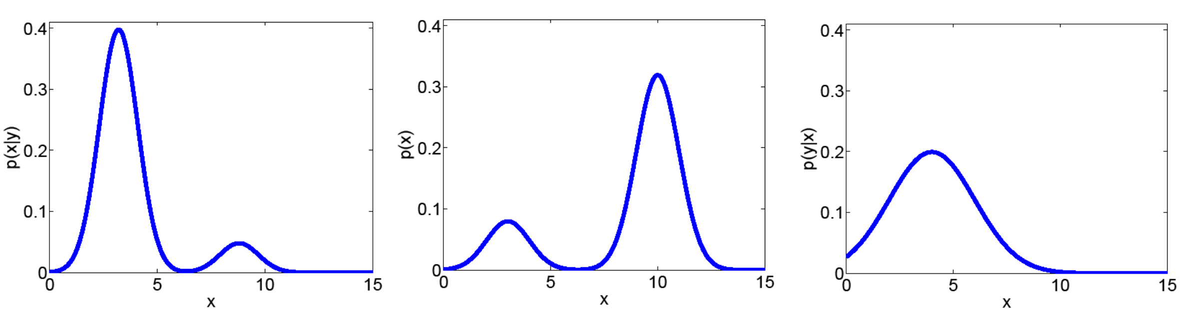
sample mean: [-0.05 0.83]
sample cov: [0.95 -0.23
-0.23 0.29]

Point Estimates

- Estimator: Any function of the data, intended to represent the best approximation of the true value of a parameter
- Most common estimator is the sample average
- Statistically-motivated examples:
  - Maximum likelihood (ML): $\hat{x}(\vec{d}) = \arg\max_{x} p(\vec{d}|x)$
  - Max a posteriori (MAP): $\hat{x}(\vec{d}) = \arg\max_{x} p(x|\vec{d})$
  - Min Mean Squared Error (MMSE):

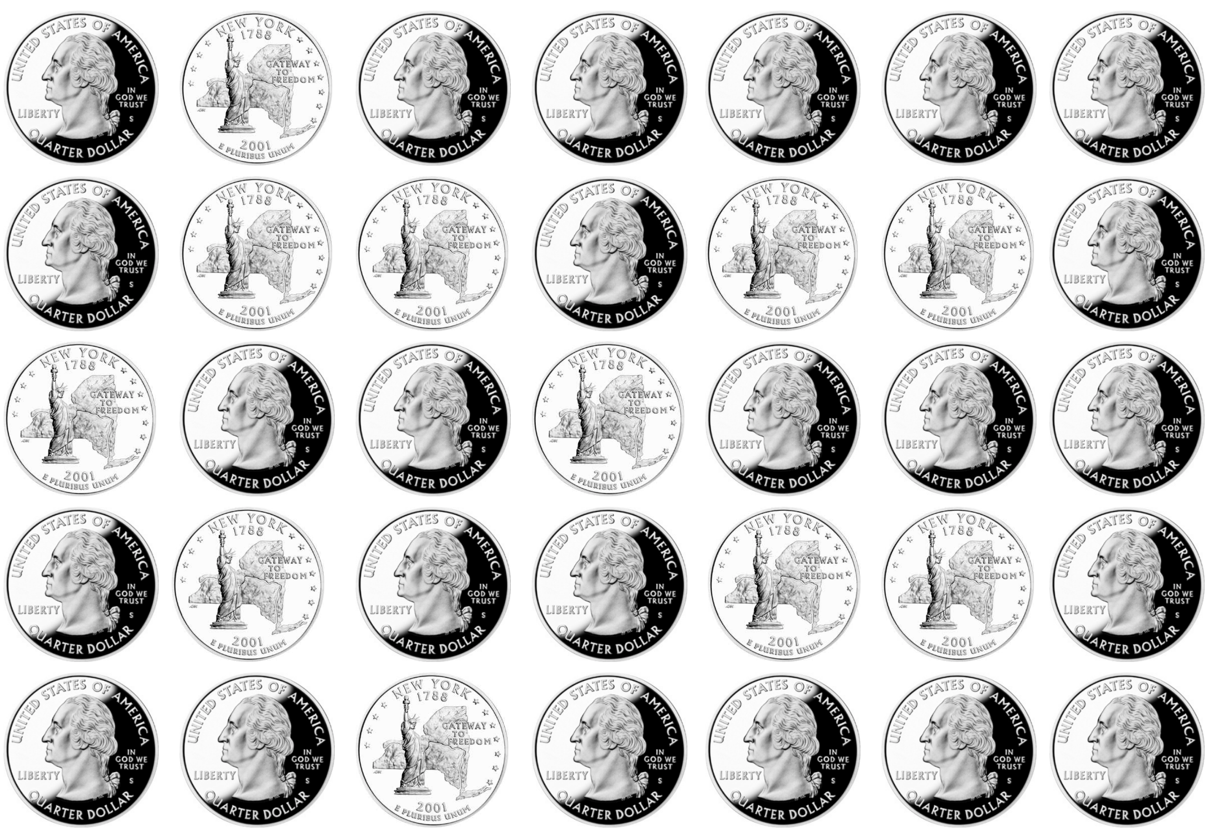
$$\hat{x}(\vec{d}) = \arg\min_{\hat{x}} \mathbf{E}\left((x-\hat{x})^2|\vec{d}\right)$$
$$= \mathbf{E}\left(x|\vec{d}\right)$$

p(x|y) proportional to p(x) * p(y|x)

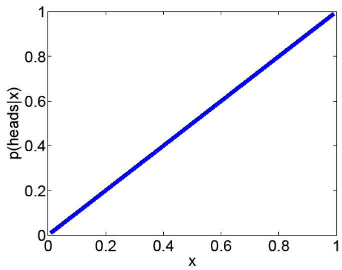


- why must both prior and likelihood be taken into account?
- why doesn't data dominate?
- when would it? when would prior dominate?
- what if prior and likelihood are incompatible?

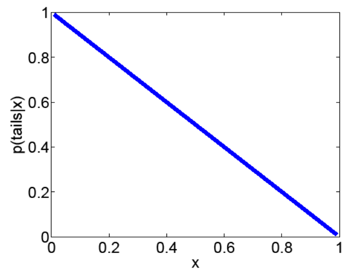

UNITED STATES OF AMERICA
IN GOD WE TRUST
LIBERTY
S
QUARTER DOLLAR


UNITED STATES OF AMERICA
IN GOD WE TRUST
LIBERTY
S
QUARTER DOLLAR
NEW YORK
1788
GATEWAY TO FREEDOM
2001
E PLURIBUS UNUM

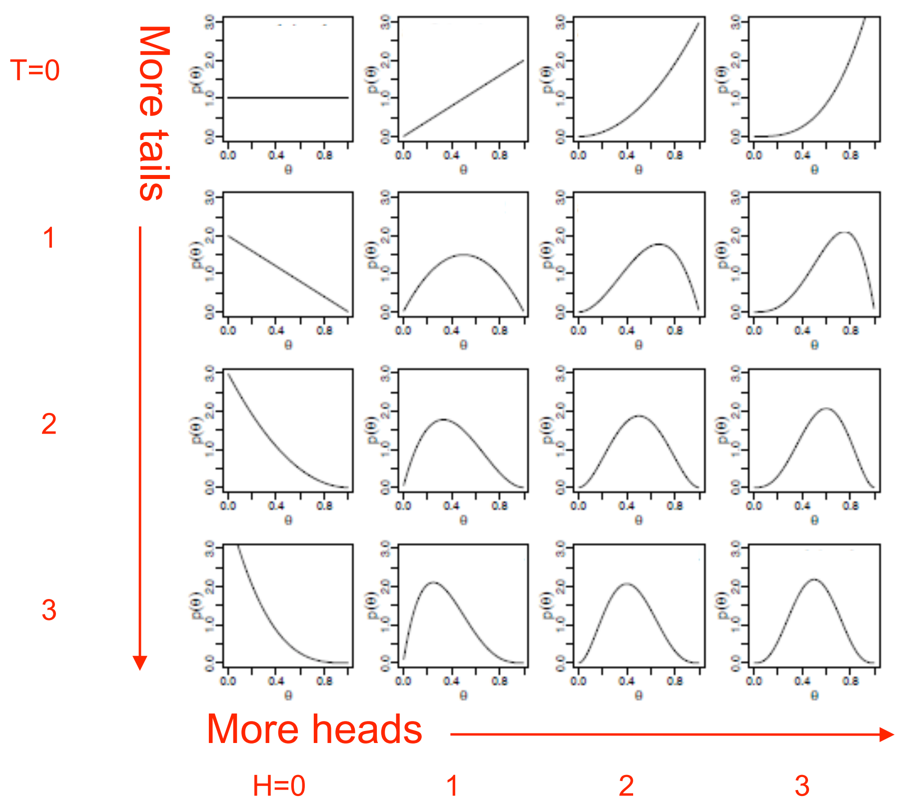
Likelihood: 1 head



Likelihood: 1 tail



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

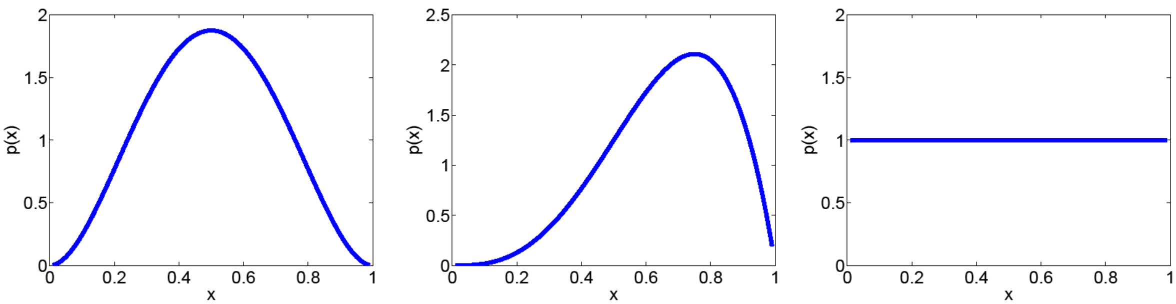


example

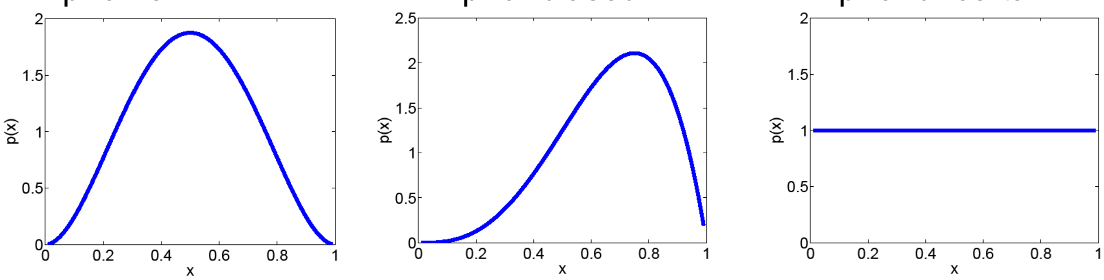
infer whether a coin is fair by flipping it repeatedly
here, x is the probability of heads (50% is fair)
$y_{1 \ldots n}$ are the outcomes of flips

Consider three different priors:

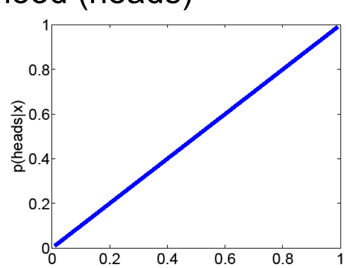
suspect fair | suspect biased | no idea



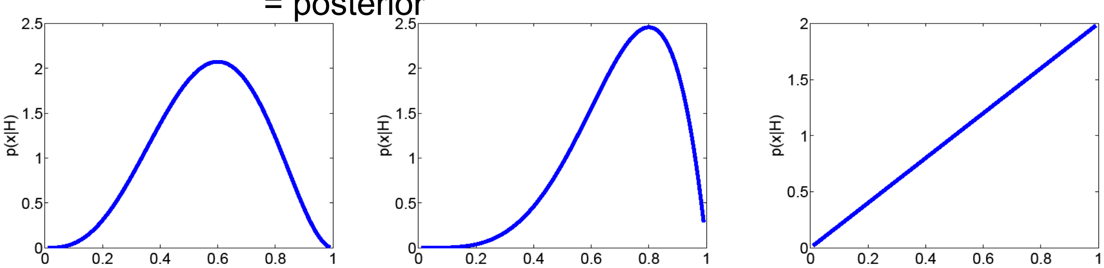
prior fair | prior biased | prior uncertain

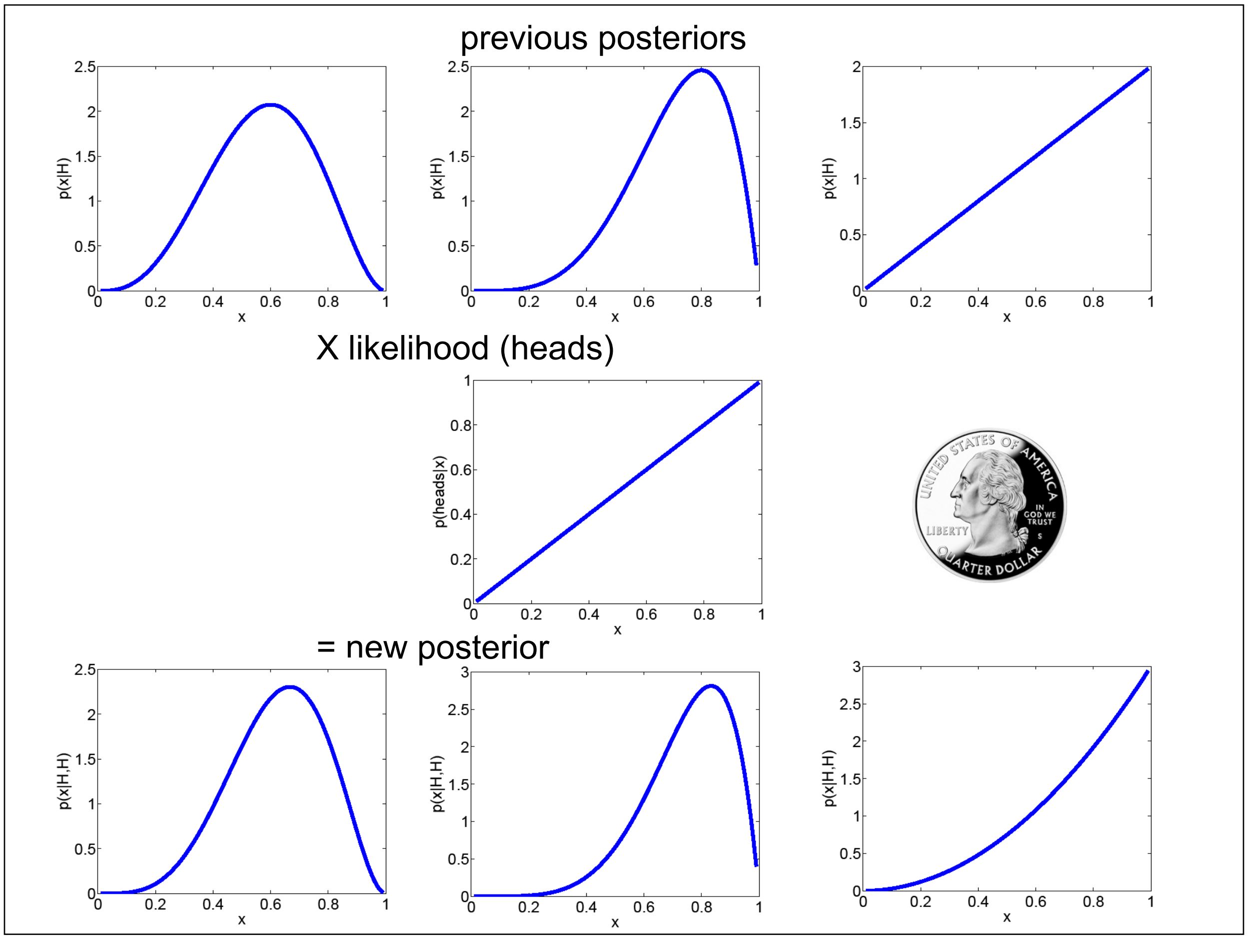


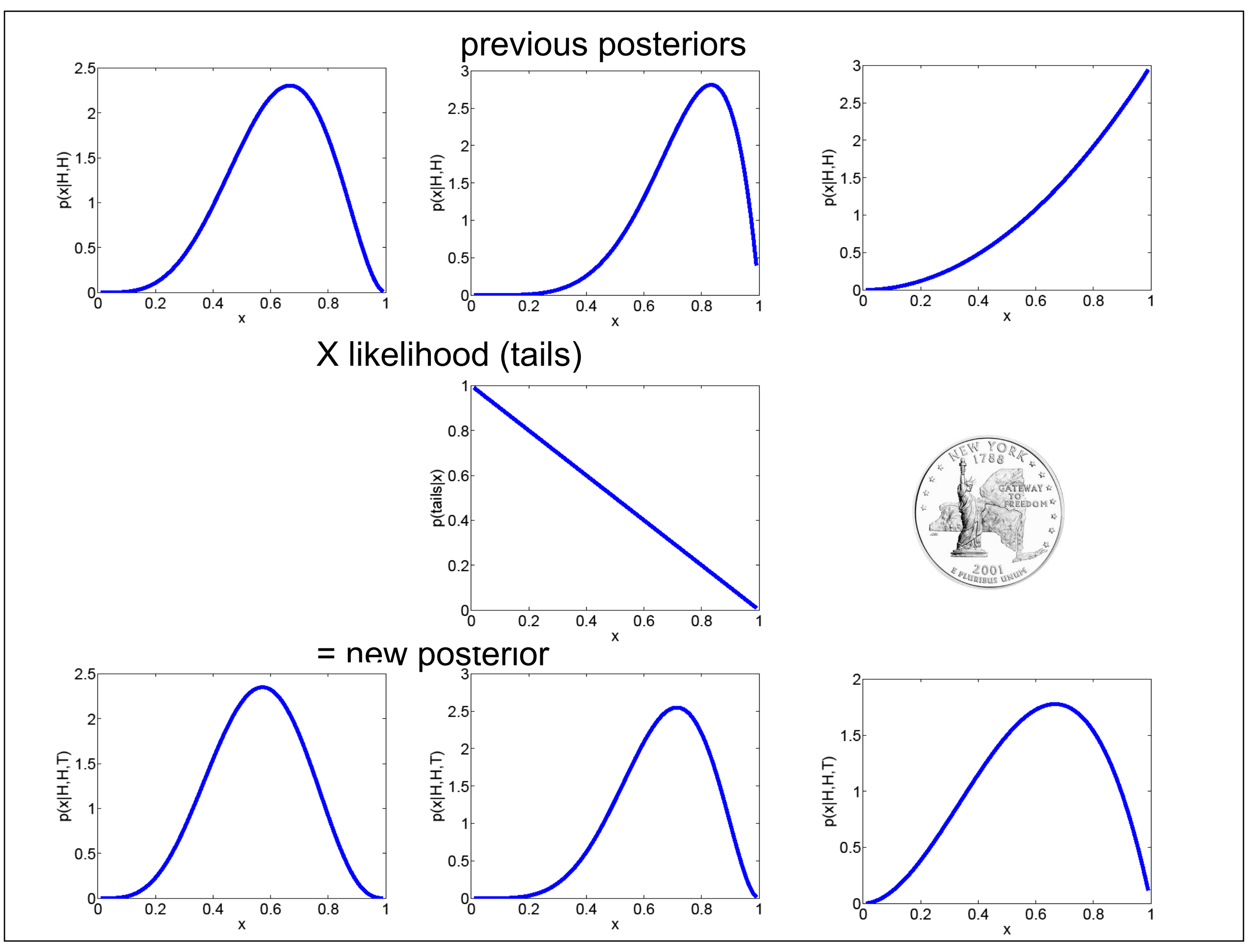
X likelihood (heads)



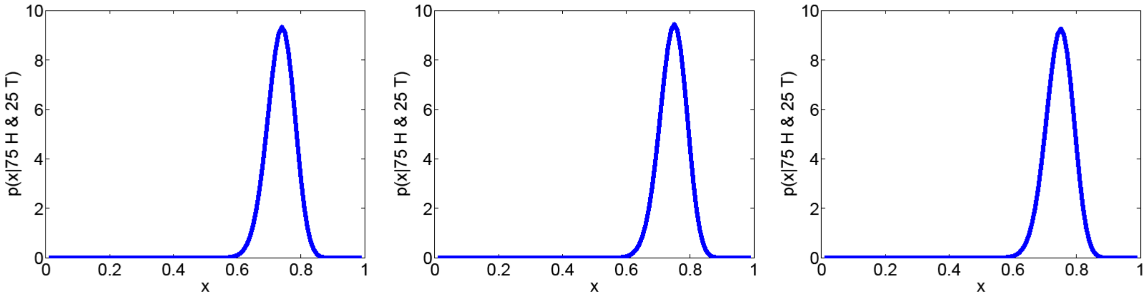
= posterior




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


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

Posteriors after observing 75 heads, 25 tails

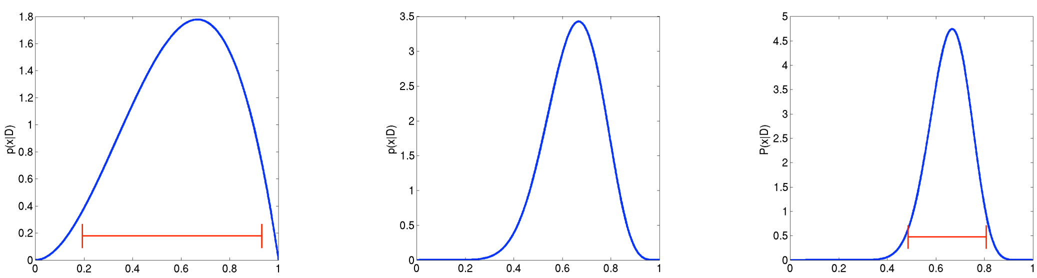


→prior differences are ultimately overwhelmed by data

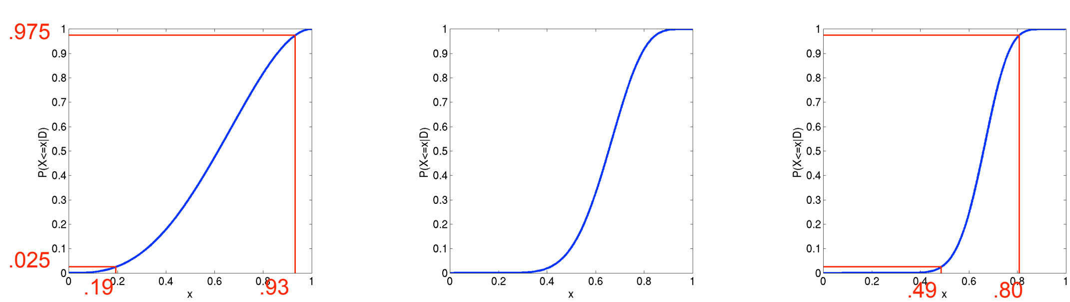
Confidence

PDFs

2H / 1T | 10H / 5T | 20H / 10T



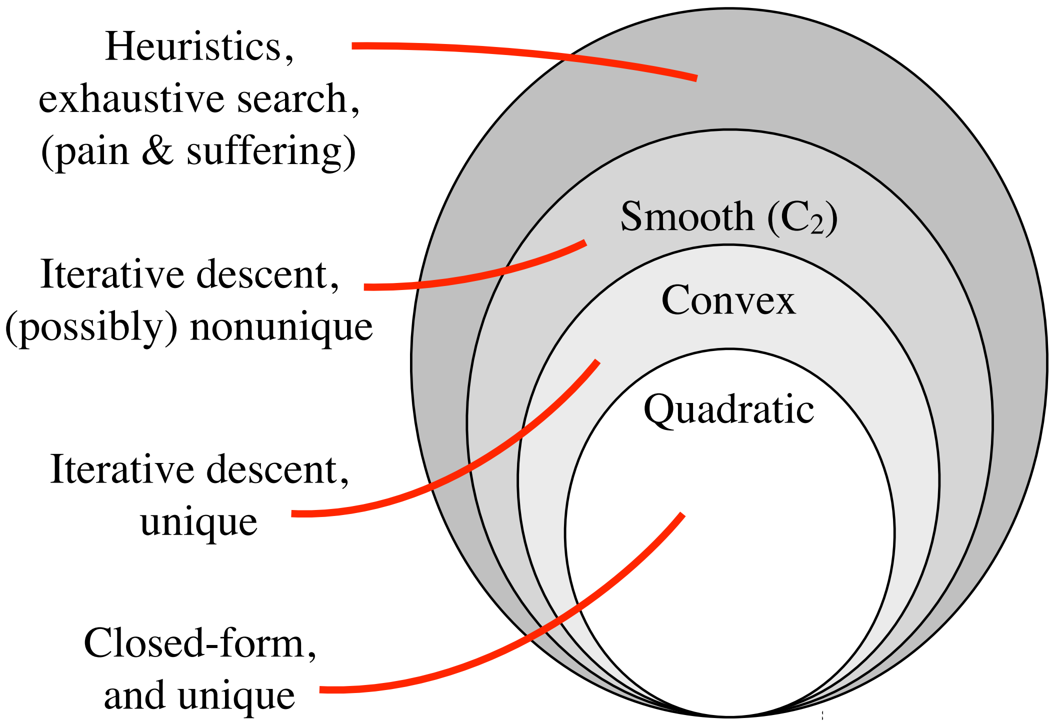
CDFs



Bias & Variance

- MSE = bias^2 + variance
- Bias is difficult to assess (since requires knowing the "true" value). But variance is easier.
- Classical statistics generally aims for an unbiased estimator, with minimal variance
- 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.

Optimization...



Bootstrapping

- "The Baron had fallen to the bottom of a deep lake. Just when it looked like all was lost, he thought to pick himself up by his own boostraps" [Adventures of Baron von Munchausen, by Rudolph Erich Raspe]
- A **resampling** method for computing estimator distribution (incl. stdev or error bars)
- Idea: instead of running experiment multiple times, resample from existing data (with replacement). Compute estimates from these "bootstrap" data sets.

HEART ATTACK RISK FOUND TO BE CUT BY TAKING ASPIRIN

LIFESAVING EFFECTS SEEN

Study Finds Benefit of Tablet Every Other Day Is Much Greater Than Expected

[New York Times, 27 Jan 1987]

The summary statistics in the newspaper article are very simple:

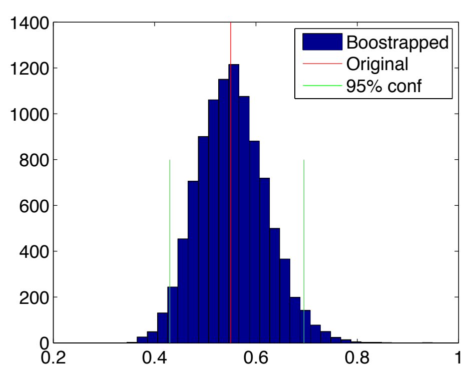
	heart attacks (fatal plus non-fatal)	subjects
aspirin group:	104	11037
placebo group:	189	11034

$$\hat{\theta} = \frac{104/11037}{189/11034} = .55. \qquad (1.1)$$

If this study can be believed, and its solid design makes it very believable, the aspirin-takers only have 55% as many heart attacks as placebo-takers.

Of course we are not really interested in $\hat{\theta}$, the estimated ratio. What we would like to know is θ, the true ratio.

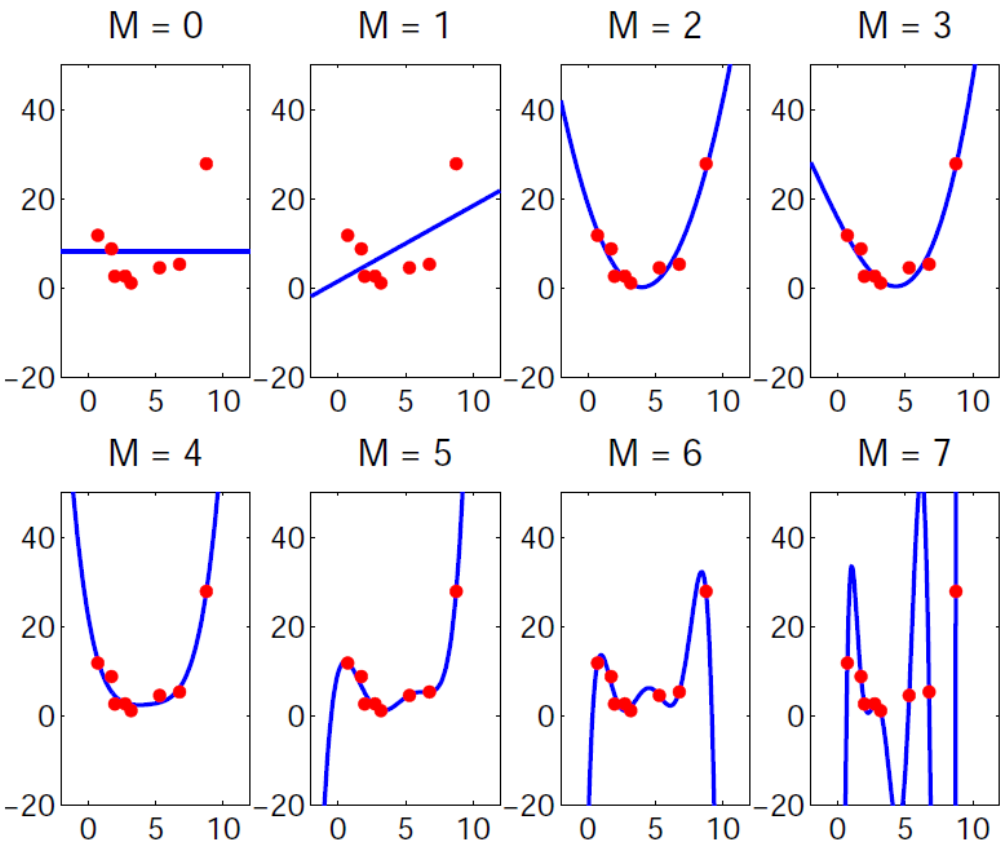
Histogram of bootstrap estimates



=> with 95% confidence,

$$0.43 < \theta < 0.7$$

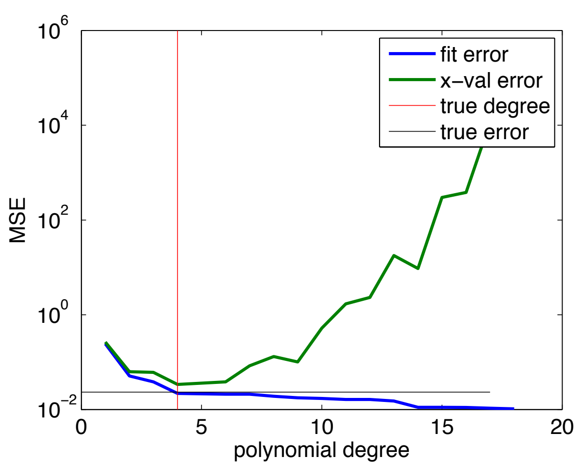
[Efron & Tibshirani '98]



Cross-validation

A resampling method for determining predictive power of a model.
Widely used to identify/avoid over-fitting.

(1) Randomly partition your data into a “training” set, and a “test” set.
(2) Fit model to training set. Measure error on test set.
(3) Repeat (many times)



Ridge regression

(a.k.a. Tikhonov regularization, or linear regularization)

Ordinary least squares regression:

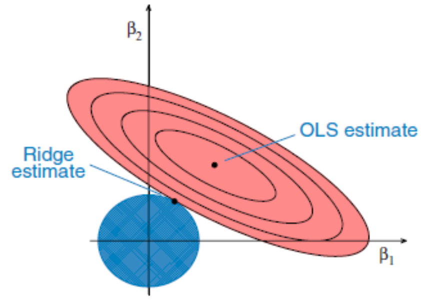
$$\arg\min_{\vec{\beta}} ||\vec{y} - X\vec{\beta}||^2$$

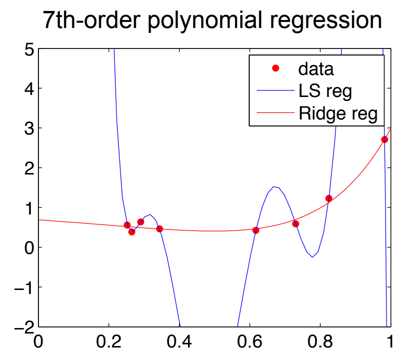
"Regularized" least squares regression:

$$\arg\min_{\vec{\beta}} ||\vec{y} - X\vec{\beta}||^2 + \lambda||\vec{\beta}||^2$$

Note: negative log posterior, assuming Gaussian likelihood & prior

Choose lambda by cross-validation





L_1 regularization

(a.k.a. LASSO - least absolute shrinkage and selection operator)

$$\arg\min_{\vec{\beta}} ||\vec{y} - X\vec{\beta}||^2 + \lambda \sum_k |\beta_k|$$

L_1 norm

Using an absolute error regularization term promotes *selection* of regressors: