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

**Statistics**: use of probability to analyze, interpret, summarize data. Fundamental to *essentially all experimental science*.
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.
What do hypotheses predict about potential data?

How does data support (or undermine) hypotheses?

-from a course on statistics, by M. West
Probability basics

• discrete probability distributions
• continuous probability densities
• cumulative distributions
• transformations on random variables
• multi-dimensional densities: marginals, conditionals, Bayes Rule
• independence

[on board]
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 just \(P(x)\)

that is: \(P(x)\) is a function over \(x\) (‘distribution’, ‘density’)

example probability distributions

not-quite-fair coin

roll of a die

sum of two dice

velocity of gas molecules

and, time between clicks
Joint distribution
Conditional
Conditional

$P(x | y)$: conditional distribution over $X$ given that $Y = y$
For a particular $y$, this is a function of $x$

e.g: someone’s height is 68 inches; what is their IQ?

$P(x | Y = y) = P(x, Y = y) / P(y) = P(x, Y = y) / \sum_x P(x, Y = y)$

select
renormalize
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, 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.
\[ p(x|y) \text{ is proportional to } \frac{p(x)}{p(y|x)} \]

- 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?
The fairness of a coin might be hugely consequential for high stakes games, but it isn’t often in life that we flip coins. So why bother studying the statistics of coin flips?

Because coin flips are a surrogate for myriad other real-life events that we care about. For a given type of heart surgery, what is the probability that patients survive more than 1 year? For a given type of drug, what is the probability of headache as a side effect? For a particular training method, what is the probability of at least 10% improvement? For a survey question, what is the probability that people will agree or disagree? In a two candidate election, what is the probability that a person will vote for each candidate?

Whenever we are discussing coin flips, keep in mind that we could be talking about some domain in which you are actually interested! The coins are merely a generic representative of a universe of analogous applications.
Likelihood: 1 heads

Likelihood: 1 tails
posteriors $p(\theta|a,b)$
(equivalently: rescaled likelihoods $p(a,b|\theta)$ assuming $P(\theta)=1$)

$b=0$

More tails

1

2

3

More heads

a=0 1 2 3

Tuesday, 17 November 2015
Last time:
• inference via inverting noise model
• estimating proportion (e.g. $a$ heads, $b$ tails)
  - examine “generative” likelihood $P(a,b|x)$
  - invert with Bayes rule: $P(x|a,b)$

today:
  - priors
  - central tendencies (ML, MAP, mean)
  - uncertainty (variance, confidence intervals, CDF)
  - obvious estimator: $a/(a+b)$
  - hypothesis testing
example

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

Three initial priors:
suspect fair
suspect biased
uncertain
prior fair

prior biased

prior uncertain

X likelihood (heads)

= posterior

Tuesday, 17 November 2015
previous posteriors

$X$ likelihood (heads)

= new posterior
previous posteriors

X likelihood (tails)

= new posterior
coins

Posteriors after 75 heads, 25 tails

→prior differences ultimately overwhelmed by data
Mean and (co)variance

- one-D: mean and covariance summarize position/width
- multi-D: vector mean and covariance matrix, elliptical geometry
- marginals: mean/variance of sum of random variables
- generally: mean/covariance of linear transformations of RVs
- what about nonlinear transformations? 1D easy...

- sample mean
  - ... almost always converges to true mean
  - ... with variance $\sigma^2/N$
  - common choice for an estimator ...

[on board]
Estimators

• Any function of the data, intended to represent your best guess of the true value of a parameter

• 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):
    \[ \hat{x}(\vec{d}) = \arg \min_{\hat{x}} \mathbb{E} \left( (x - \hat{x})^2 | \vec{d} \right) \]
    \[ = \mathbb{E} \left( x | \vec{d} \right) \]

• Bias and Variance of an estimator

• Sample mean is often used as an estimator, (e.g., it’s the MLE for coin-tossing example)...

Tuesday, 17 November 2015
Gaussian densities

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

\[ p(\mathbf{x}) = \frac{1}{\sqrt{(2\pi)^N |C|}} e^{- (\mathbf{x} - \mathbf{\mu})^T C^{-1} (\mathbf{x} - \mathbf{\mu}) / 2} \]

mean: [0.2, 0.8]

cov: [1.0 -0.3; -0.3 0.4]
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]
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]
The Gaussian

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

- parameterized by mean and variance (shifting / rescaling)
- joint density of two indep Gaussian RVs is circular! [easy]
- product of two Gaussians is Gaussian! [easy] [ex. Bayes]
- 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

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

\[ p(x|y) = p(y|x)p(x) \]
Product of Gaussians is Gaussian

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

\[
p(x|y) = p(y|x)p(x)
\]

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

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

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

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

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

(average, weighted by inverse variances!)
Product of Gaussians is Gaussian

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

\[
p(x|y) = p(y|x)p(x)
\]

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

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

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

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

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

(average, weighted by inverse variances!)
\[ \tilde{x} \sim N(\tilde{\mu}, \tilde{C}), \quad P = C^{-1} \]
\[ \vec{x} \sim N(\vec{\mu}, C), \quad P = C^{-1} \]

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

Gaussian, with:

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

\[
\sigma^2 = \frac{1}{P_{11}}
\]

Conditional:
\[ \vec{x} \sim N(\vec{\mu}, C), \quad P = C^{-1} \]

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

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

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

\[
\sigma^2 = \frac{1}{P_{11}}
\]

Conditional:

Marginal:

Gaussian, with:
\[
\mu = \mu_1
\]

\[
\sigma^2 = C_{11}
\]
Generalized marginals of a Gaussian

\[ \tilde{x} \sim N(\tilde{\mu}_x, C_x) \]

\[ y = \tilde{w}^T \tilde{x} \]

Gaussian, with:

\[ \mu_y = \tilde{w}^T \tilde{\mu}_x \]

\[ \sigma^2_y = \tilde{w}^T C_x \tilde{w} \]
Central limit for a uniform distribution...

10k samples, uniform density (sigma=1)

$(u_1 + u_2)/\sqrt{2}$

$(u_1 + u_2 + u_3 + u_4)/\sqrt{4}$

$\frac{1}{\sqrt{10}} \sum_{n=1}^{10} u_n$
Central limit for a binary distribution...

avg of 4 coins

avg of 16 coins

avg of 64 coins
Central limit for a binary distribution...

one coin

avg of 4 coins

avg of 16 coins

avg of 256 coins
PDFs

2H / 1T

10H / 5T

20H / 10T
PDFs

2H / 1T

10H / 5T

20H / 10T

CDFs
power

\[ P(a|N=4, x=0.5) \]

\[ P(a|N=500, x=0.5) \]

\[ P(a|N=4, x=0.75) \]

\[ P(a|N=500, x=0.75) \]
joint posterior

mean

variance
joint posterior
Confidence intervals

PDF

\[ P(\text{mean} \mid \text{data, variance}) \]

mean

CDF

\[ P(\text{mean} < x \mid \text{data, variance}) \]

mean
joint posterior
Confidence intervals

\begin{align*}
P(\text{mean} \mid \text{data}) & \quad \text{PDF} \\
P(\text{mean} < x \mid \text{data}) & \quad \text{CDF}
\end{align*}
On-line inference

\[ P(\text{mean}) \]

mean
On-line inference
On-line inference
On-line inference

\[ P(\text{mean}) \]

\[ \text{mean} \]
On-line inference
On-line inference

\[ P(\text{mean}) \]

mean
multivariate Gaussian

\[ p(\vec{x}) = \frac{1}{\sqrt{(2\pi)^N |C|}} e^{- (\vec{x} - \vec{\mu})^T C^{-1} (\vec{x} - \vec{\mu}) / 2} \]
Figure 1 Graphs illustrating the various types of situations in which the error of comparing significance levels occurs. (a) Comparing effect sizes in an experimental group/condition and a control group/condition. (b) Comparing effect sizes during a pre-test and a post-test. (c) Comparing several brain areas and claiming that a particular effect (property) is specific for one of these brain areas. (d) Data presented in a,