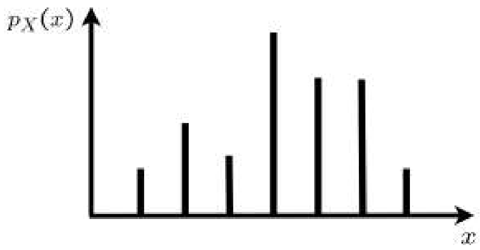
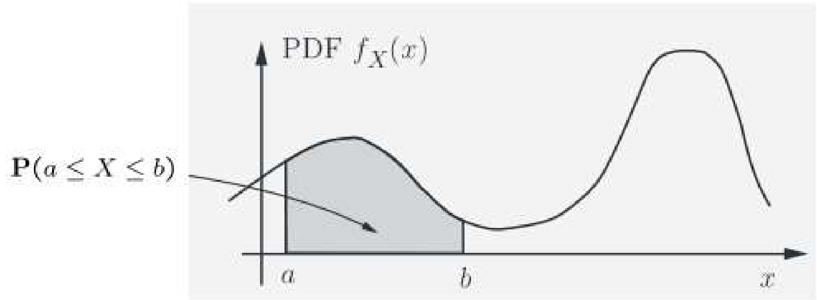
Probability slides from the course MITx: 6.041x Introduction to Probability - The Science of Uncertainty

By

Professor Nikos Tsitsiklis

Expectation/mean of a continuous random variable





$$\mathbf{E}[X] = \sum_x x p_X(x)$$

$$\mathbf{E}[X] = \int_{-\infty}^{\infty} x f_X(x)\,dx$$

- **Interpretation:** Average in large number of independent repetitions of the experiment

Fine print:
Assume $\int_{-\infty}^{\infty} |x|\, f_X(x)\,dx < \infty$

Properties of expectations

- If $X \geq 0$, then $\mathbf{E}[X] \geq 0$
- If $a \leq X \leq b$, then $a \leq \mathbf{E}[X] \leq b$
- Expected value rule:

$$\mathbf{E}[g(X)] = \sum_x g(x) p_X(x)$$

$$\boxed{\mathbf{E}[g(X)] = \int_{-\infty}^{\infty} g(x) f_X(x)\, dx}$$

- Linearity

$$\boxed{\mathbf{E}[aX + b] = a\mathbf{E}[X] + b}$$

Variance and its properties

- **Definition of variance:** $\text{var}(X) = \mathbf{E}[(X-\mu)^2]$

- Calculation using the expected value rule, $\mathbf{E}[g(X)] = \int_{-\infty}^{\infty} g(x) f_X(x)\, dx$

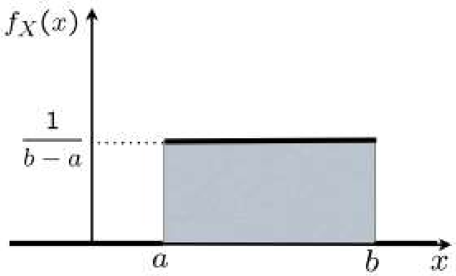
$$\text{var}(X) =$$

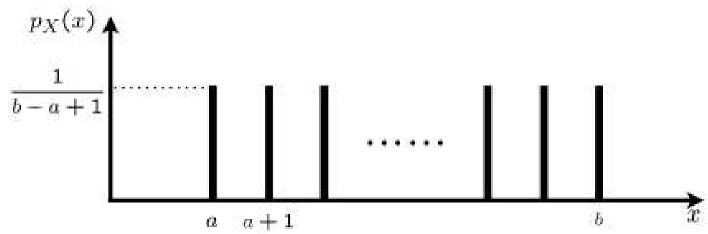
Standard deviation: $\sigma_X = \sqrt{\text{var}(X)}$

$\text{var}(aX+b) = a^2\text{var}(X)$

A useful formula: $\text{var}(X) = \mathbf{E}[X^2] - \big(\mathbf{E}[X]\big)^2$

Continuous uniform random variable; parameters a, b





$$\mathbf{E}[X] = \int_{-\infty}^{\infty} x f_X(x)\, dx$$

$$\mathbf{E}[X] = \frac{a+b}{2}$$

$$\text{var}(X) = \frac{1}{12}(b-a)(b-a+2)$$

$$\mathbf{E}[X^2] =$$

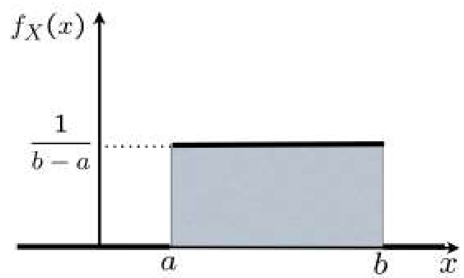
$$\text{var}(X) =$$

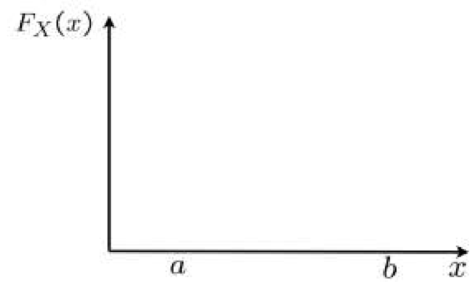
Cumulative distribution function (CDF)

CDF definition: $F_X(x) = \mathbf{P}(X \le x)$

- Continuous random variables:

$$F_X(x) = \mathbf{P}(X \le x) = \int_{-\infty}^{x} f_X(t)\, dt$$





Normal (Gaussian) random variables

- Important in the theory of probability
  - Central limit theorem

- Prevalent in applications
  - Convenient analytical properties
  - Model of noise consisting of many, small independent noise terms

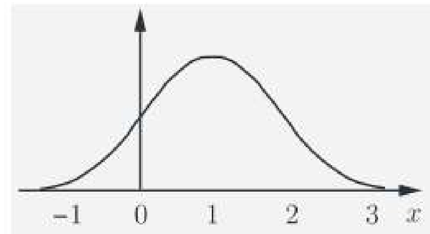
Standard normal (Gaussian) random variables

- Standard normal $N(0,1)$: $f_X(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}$

- $\mathbf{E}[X] =$

- $\text{var}(X) = 1$

General normal (Gaussian) random variables



- General normal $N(\mu, \sigma^2)$: $f_X(x) = \dfrac{1}{\sigma\sqrt{2\pi}} e^{-(x-\mu)^2/2\sigma^2}$

- $\mathbf{E}[X] =$

- $\mathrm{var}(X) = \sigma^2$

Linear functions of a normal random variable

- Let $Y = aX + b \quad X \sim N(\mu, \sigma^2)$

  $\mathbf{E}[Y] =$

  $\mathsf{Var}(Y) =$

- Fact (will prove later in this course):

  $\boxed{Y \sim N(a\mu + b, a^2\sigma^2)}$

- Special case: $a = 0$?

Calculating normal probabilities

- Express an event of interest in terms of standard normal

	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
0.0	.5000	.5040	.5080	.5120	.5160	.5199	.5239	.5279	.5319	.5359
0.1	.5398	.5438	.5478	.5517	.5557	.5596	.5636	.5675	.5714	.5753
0.2	.5793	.5832	.5871	.5910	.5948	.5987	.6026	.6064	.6103	.6141
0.3	.6179	.6217	.6255	.6293	.6331	.6368	.6406	.6443	.6480	.6517
0.4	.6554	.6591	.6628	.6664	.6700	.6736	.6772	.6808	.6844	.6879
0.5	.6915	.6950	.6985	.7019	.7054	.7088	.7123	.7157	.7190	.7224
0.6	.7257	.7291	.7324	.7357	.7389	.7422	.7454	.7486	.7517	.7549
0.7	.7580	.7611	.7642	.7673	.7704	.7734	.7764	.7794	.7823	.7852
0.8	.7881	.7910	.7939	.7967	.7995	.8023	.8051	.8078	.8106	.8133
0.9	.8159	.8186	.8212	.8238	.8264	.8289	.8315	.8340	.8365	.8389
1.0	.8413	.8438	.8461	.8485	.8508	.8531	.8554	.8577	.8599	.8621
1.1	.8643	.8665	.8686	.8708	.8729	.8749	.8770	.8790	.8810	.8830
1.2	.8849	.8869	.8888	.8907	.8925	.8944	.8962	.8980	.8997	.9015
1.3	.9032	.9049	.9066	.9082	.9099	.9115	.9131	.9147	.9162	.9177
1.4	.9192	.9207	.9222	.9236	.9251	.9265	.9279	.9292	.9306	.9319
1.5	.9332	.9345	.9357	.9370	.9382	.9394	.9406	.9418	.9429	.9441
1.6	.9452	.9463	.9474	.9484	.9495	.9505	.9515	.9525	.9535	.9545
1.7	.9554	.9564	.9573	.9582	.9591	.9599	.9608	.9616	.9625	.9633
1.8	.9641	.9649	.9656	.9664	.9671	.9678	.9686	.9693	.9699	.9706
1.9	.9713	.9719	.9726	.9732	.9738	.9744	.9750	.9756	.9761	.9767
2.0	.9772	.9778	.9783	.9788	.9793	.9798	.9803	.9808	.9812	.9817
2.1	.9821	.9826	.9830	.9834	.9838	.9842	.9846	.9850	.9854	.9857
2.2	.9861	.9864	.9868	.9871	.9875	.9878	.9881	.9884	.9887	.9890
2.3	.9893	.9896	.9898	.9901	.9904	.9906	.9909	.9911	.9913	.9916
2.4	.9918	.9920	.9922	.9925	.9927	.9929	.9931	.9932	.9934	.9936
2.5	.9938	.9940	.9941	.9943	.9945	.9946	.9948	.9949	.9951	.9952
2.6	.9953	.9955	.9956	.9957	.9959	.9960	.9961	.9962	.9963	.9964
2.7	.9965	.9966	.9967	.9968	.9969	.9970	.9971	.9972	.9973	.9974
2.8	.9974	.9975	.9976	.9977	.9977	.9978	.9979	.9979	.9980	.9981
2.9	.9981	.9982	.9982	.9983	.9984	.9984	.9985	.9985	.9986	.9986

Conditional expectation of X, given an event

$$\mathbf{E}[X] = \sum_x x p_X(x) \qquad \mathbf{E}[X] = \int x f_X(x)\, dx$$

$$\mathbf{E}[X \mid A] = \sum_x x p_{X|A}(x) \qquad \mathbf{E}[X \mid A] = \int x f_{X|A}(x)\, dx$$

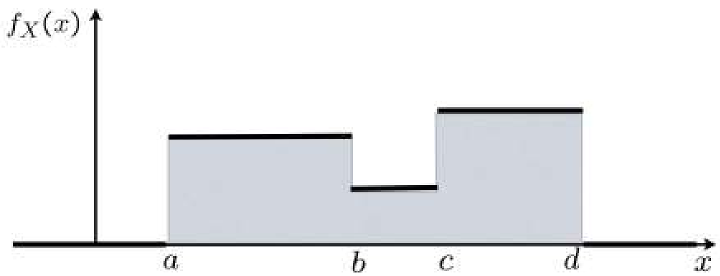
Expected value rule:

$$\mathbf{E}[g(X)] = \sum_x g(x) p_X(x) \qquad \mathbf{E}[g(X)] = \int g(x) f_X(x)\, dx$$

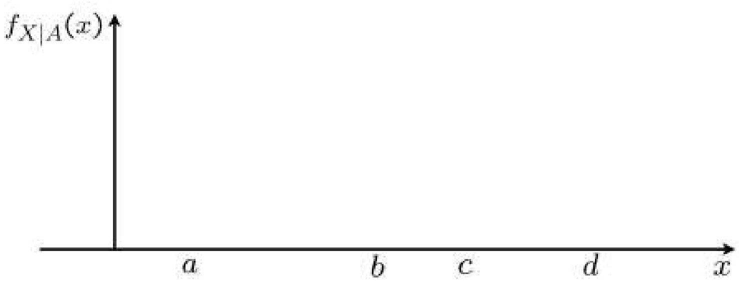
$$\mathbf{E}[g(X) \mid A] = \sum_x g(x) p_{X|A}(x) \qquad \mathbf{E}[g(X) \mid A] = \int g(x) f_{X|A}(x)\, dx$$

Example

$$A: \quad \frac{a+b}{2} \le X \le b$$

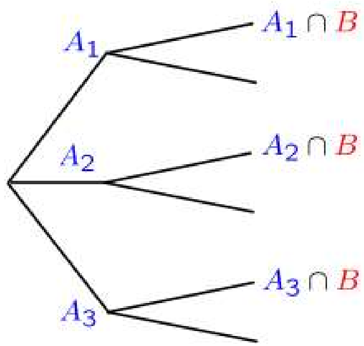


$$\mathbf{E}[X \mid A] =$$



$$\mathbf{E}[X^2 \mid A] =$$

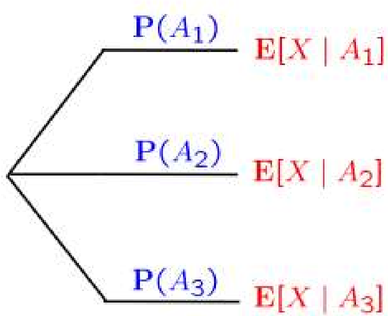
Total probability and expectation theorems



$$\mathbf{P}(B) = \mathbf{P}(A_1)\mathbf{P}(B \mid A_1) + \cdots + \mathbf{P}(A_n)\mathbf{P}(B \mid A_n)$$

$$p_X(x) = \mathbf{P}(A_1)\, p_{X|A_1}(x) + \cdots + \mathbf{P}(A_n)\, p_{X|A_n}(x)$$

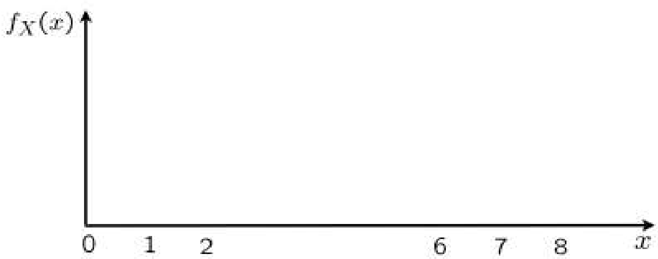
$$\boxed{f_X(x) = \mathbf{P}(A_1) f_{X|A_1}(x) + \cdots + \mathbf{P}(A_n) f_{X|A_n}(x)}$$



$$\boxed{\mathbf{E}[X] = \mathbf{P}(A_1)\mathbf{E}[X \mid A_1] + \cdots + \mathbf{P}(A_n)\mathbf{E}[X \mid A_n]}$$

Example

- Bill goes to the supermarket shortly, with probability 1/3,
  at a time uniformly distributed between 0 and 2 hours from now;
  or with probability 2/3, later in the day
  at a time uniformly distributed between 6 and 8 hours from now



$$f_X(x) = \mathbf{P}(A_1) f_{X|A_1}(x) + \cdots + \mathbf{P}(A_n) f_{X|A_n}(x)$$

$$\mathbf{E}[X] = \mathbf{P}(A_1)\mathbf{E}[X \mid A_1] + \cdots + \mathbf{P}(A_n)\mathbf{E}[X \mid A_n]$$

Total probability and expectation theorems

$$p_X(x) = \sum_y p_Y(y) p_{X|Y}(x \mid y)$$

$$f_X(x) = \int_{-\infty}^{\infty} f_Y(y) f_{X|Y}(x \mid y)\, dy$$

$$\mathbf{E}[X \mid Y = y] = \sum_x x p_{X|Y}(x \mid y)$$

$$\mathbf{E}[X \mid Y = y] = \int_{-\infty}^{\infty} x f_{X|Y}(x \mid y)\, dx$$

$$\mathbf{E}[X] = \sum_y p_Y(y) \mathbf{E}[X \mid Y = y]$$

$$\mathbf{E}[X] = \int_{-\infty}^{\infty} f_Y(y) \mathbf{E}[X \mid Y = y]\, dy$$

- Expected value rule...

Independence

$p_{X,Y}(x,y) = p_X(x)\, p_Y(y)$, for all x, y

$f_{X,Y}(x,y) = f_X(x)\, f_Y(y)$, for all x and y

$f_{X,Y}(x,y) = f_{X|Y}(x\,|\,y)\, f_Y(y)$

- equivalent to: $f_{X|Y}(x\,|\,y) = f_X(x)$, for all y with $f_Y(y) > 0$ and all x

If X, Y are **independent:** $\mathbf{E}[XY] = \mathbf{E}[X]\mathbf{E}[Y]$

$\text{var}(X+Y) = \text{var}(X) + \text{var}(Y)$

$g(X)$ and $h(Y)$ are also independent: $\mathbf{E}[g(X)h(Y)] = \mathbf{E}[g(X)] \cdot \mathbf{E}[h(Y)]$

Covariance properties

$$\boxed{\text{cov}(X,Y) = \mathbf{E}\Big[(X - \mathbf{E}[X]) \cdot (Y - \mathbf{E}[Y])\Big]}$$

$\text{cov}(X,X) =$

$\text{cov}(aX + b, Y) =$

$\text{cov}(X, Y + Z) =$

$$\boxed{\text{cov}(X,Y) = \mathbf{E}[XY] - \mathbf{E}[X]\mathbf{E}[Y]}$$

The variance of a sum of random variables

$\mathsf{var}(X_1 + X_2) =$

The variance of a sum of random variables

$\mathrm{var}(X_1 + X_2) = \mathrm{var}(X_1) + \mathrm{var}(X_2) + 2\,\mathrm{cov}(X_1, X_2)$

$\mathrm{var}(X_1 + \cdots + X_n) =$

$$\mathrm{var}(X_1 + \cdots + X_n) = \sum_{i=1}^{n} \mathrm{var}(X_i) + \sum_{\{(i,j)\,:\, i \neq j\}} \mathrm{cov}(X_i, X_j)$$

The Correlation coefficient

- Dimensionless version of covariance:

$$\begin{aligned}\rho(X,Y) &= \mathbf{E}\left[\frac{(X-\mathbf{E}[X])}{\sigma_X}\cdot\frac{(Y-\mathbf{E}[Y])}{\sigma_Y}\right]\\ &= \frac{\mathrm{cov}(X,Y)}{\sigma_X\sigma_Y}\end{aligned}$$

$$\boxed{-1 \le \rho \le 1}$$

- Measure of the degree of "association" between X and Y

- Independent $\Rightarrow$ $\rho = 0$, "uncorrelated" (converse is not true)

- $\rho(X,X) =$

- $|\rho| = 1 \iff (X-\mathbf{E}[X]) = c(Y-\mathbf{E}[Y])$ (linearly related)

- $\mathrm{cov}(aX+b,Y) = a\cdot\mathrm{cov}(X,Y) \quad \Rightarrow \quad \rho(aX+b,Y) =$

Interpreting the correlation coefficient

$$\rho(X,Y) = \frac{\text{cov}(X,Y)}{\sigma_X \sigma_Y}$$

- Association does not imply causation or influence

  X: math aptitude
  Y: musical ability

- Correlation often reflects underlying, common, hidden factor

  – Assume, Z, V, W are independent

  $X = Z + V \qquad\qquad Y = Z + W$

  Assume, for simplicity, that Z, V, W have zero means, unit variances

The Bayes rule — a theme with variations

$$\begin{tikzcd}[column sep=huge] X \arrow[r] & Y \arrow[l, bend right, "\text{Inference}\quad p_{X|Y}(\cdot\,|\,y)"'] \end{tikzcd}$$

unobserved value x | observed value y

prior $p_X(\cdot)$ | model $p_{Y|X}(\cdot\,|\,\cdot)$

$$\begin{aligned} p_{X,Y}(x,y) &= p_X(x)\, p_{Y|X}(y\,|\,x) \\ &= p_Y(y)\, p_{X|Y}(x\,|\,y) \end{aligned}$$

$$\begin{aligned} f_{X,Y}(x,y) &= f_X(x)\, f_{Y|X}(y\,|\,x) \\ &= f_Y(y)\, f_{X|Y}(x\,|\,y) \end{aligned}$$

$$p_{X|Y}(x\,|\,y) = \frac{p_X(x)\, p_{Y|X}(y\,|\,x)}{p_Y(y)}$$

$$f_{X|Y}(x\,|\,y) = \frac{f_X(x)\, f_{Y|X}(y\,|\,x)}{f_Y(y)}$$

$$p_Y(y) = \sum_{x'} p_X(x')\, p_{Y|X}(y\,|\,x')$$

$$f_Y(y) = \int f_X(x')\, f_{Y|X}(y\,|\,x')\, dx'$$

The Bayes rule — discrete unknown, continuous measurement

- unkown K: equally likely to be -1 or $+1$
- measurement Y: $\quad Y = K + W;\ \ W \sim \mathcal{N}(0,1)$

- Probability that $K = 1$, given that $Y = y$?

$$\boxed{p_{K|Y}(k \,|\, y) = \frac{p_K(k)\, f_{Y|K}(y \,|\, k)}{f_Y(y)}}$$

$p_K(k) =$ $\qquad\qquad f_{Y|K}(y \,|\, k) =$

$f_Y(y) =$

$$f_Y(y) = \sum_{k'} p_K(k')\, f_{Y|K}(y \,|\, k')$$

$p_{K|Y}(1 \,|\, y) =$

The Bayes rule — continuous unknown, discrete measurement

- measurement K: Bernoulli with parameter Y

$$\boxed{f_{Y|K}(y\,|\,k) = \frac{f_Y(y)\,p_{K|Y}(k\,|\,y)}{p_K(k)}}$$

- unkown Y: uniform on $[0,1]$

$$p_K(k) = \int f_Y(y')\,p_{K|Y}(k\,|\,y')\,dy'$$

- Distribution of Y given that $K = 1$?

$$f_Y(y) =$$

$$p_{K|Y}(1\,|\,y) =$$

$$p_K(1) =$$

$$f_{Y|K}(y\,|\,1) =$$

Hypothesis testing versus estimation

- Hypothesis testing:
  - unknown takes one of few possible values
  - aim at small probability of incorrect decision

  Is it an airplane or a bird?

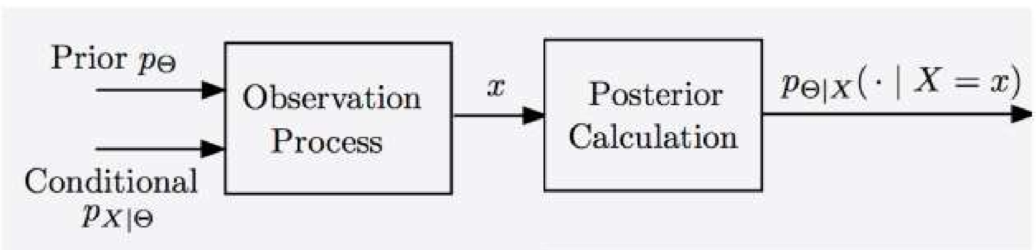
- Estimation:
  - numerical unknown(s)
  - aim at an estimate that is "close" to the true but unknown value

The Bayesian inference framework

- Unknown Θ
  - treated as a random variable
  - prior distribution p_Θ or f_Θ

- Observation X
  - observation model $p_{X|\Theta}$ or $f_{X|\Theta}$

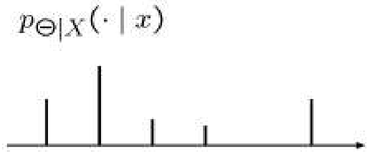
- Use appropriate version of the Bayes rule to find $p_{\Theta|X}(\cdot \mid X=x)$ or $f_{\Theta|X}(\cdot \mid X=x)$

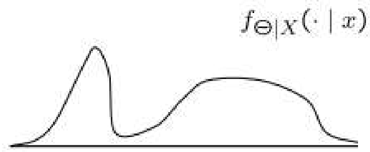


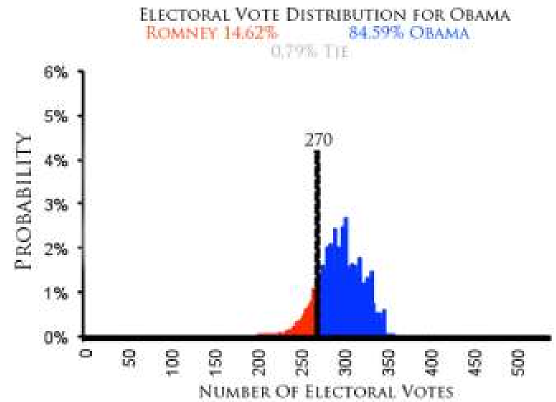
- Where does the prior come from?
  - symmetry
  - known range
  - earlier studies
  - subjective or arbitrary

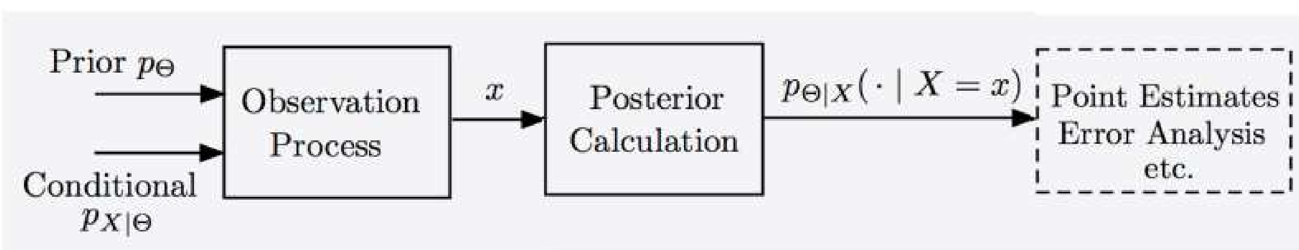
The output of Bayesian inference

The complete answer is a posterior distribution:
PMF $p_{\Theta|X}(\cdot \mid x)$ or PDF $f_{\Theta|X}(\cdot \mid x)$



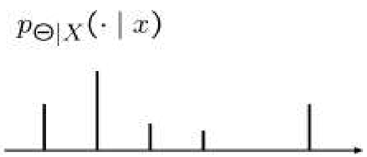


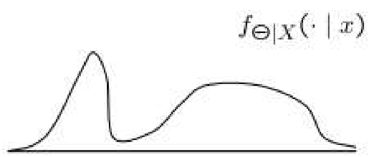




Point estimates in Bayesian inference

The complete answer is a posterior distribution:
PMF $p_{\Theta|X}(\cdot \mid x)$ or PDF $f_{\Theta|X}(\cdot \mid x)$





estimate: $\hat{\theta} = g(x)$
(number)

estimator: $\widehat{\Theta} = g(X)$
(random variable)

- Maximum a posteriori probability (MAP):

$$p_{\Theta|X}(\theta^* \mid x) = \max_\theta p_{\Theta|X}(\theta \mid x)$$

$$f_{\Theta|X}(\theta^* \mid x) = \max_\theta f_{\Theta|X}(\theta \mid x)$$

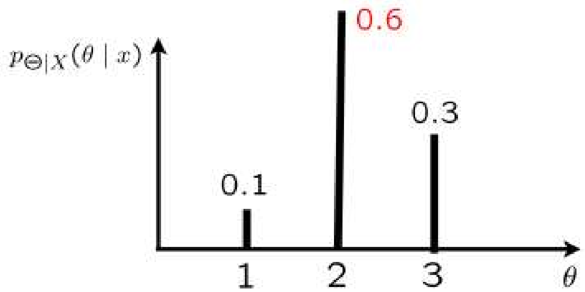
- Conditional expectation: $\mathbf{E}[\Theta \mid X = x]$ (LMS: Least Mean Squares)

Discrete Θ, discrete X

- values of Θ: alternative hypotheses

$$p_{\Theta|X}(\theta \mid x) = \frac{p_\Theta(\theta)\, p_{X|\Theta}(x \mid \theta)}{p_X(x)}$$

$$p_X(x) = \sum_{\theta'} p_\Theta(\theta')\, p_{X|\Theta}(x \mid \theta')$$



- MAP rule: $\hat{\theta} =$

- conditional prob of error:

  $\mathbf{P}(\hat{\theta} \neq \Theta \mid X = x)$

  smallest under the MAP rule

- overall probability of error:

$$\mathbf{P}(\widehat{\Theta} \neq \Theta) = \sum_x \mathbf{P}(\widehat{\Theta} \neq \Theta \mid X = x)\, p_X(x)$$

$$= \sum_\theta \mathbf{P}(\widehat{\Theta} \neq \Theta \mid \Theta = \theta)\, p_\Theta(\theta)$$

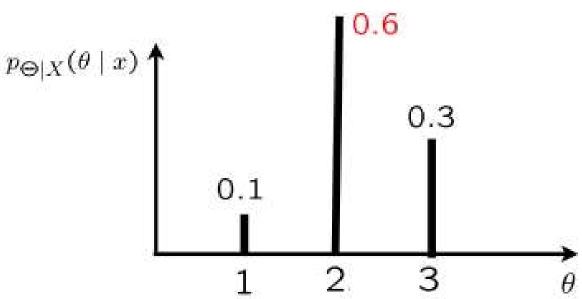
Discrete Θ, continuous X

- Standard example:
  - send signal $\Theta \in \{1, 2, 3\}$

    $X = \Theta + W$

    $W \sim N(0, \sigma^2)$, indep. of Θ

    $f_{X|\Theta}(x \mid \theta) = f_W(x - \theta)$



- MAP rule: $\hat{\theta} =$

$$p_{\Theta|X}(\theta \mid x) = \frac{p_\Theta(\theta)\, f_{X|\Theta}(x \mid \theta)}{f_X(x)}$$

$$f_X(x) = \sum_{\theta'} p_\Theta(\theta') f_{X|\Theta}(x \mid \theta')$$

- conditional prob of error:

  $\mathbf{P}(\hat{\theta} \neq \Theta \mid X = x)$

  smallest under the MAP rule

- overall probability of error:

$$\mathbf{P}(\widehat{\Theta} \neq \Theta) = \int \mathbf{P}(\widehat{\Theta} \neq \Theta \mid X = x) f_X(x)\, dx$$

$$= \sum_{\theta} \mathbf{P}(\widehat{\Theta} \neq \theta \mid \Theta = \theta)\, p_\Theta(\theta)$$

Continuous Θ, continuous X

$$f_{\Theta|X}(\theta \mid x) = \frac{f_\Theta(\theta)\, f_{X|\Theta}(x \mid \theta)}{f_X(x)}$$

$$f_X(x) = \int f_\Theta(\theta') f_{X|\Theta}(x \mid \theta')\, d\theta'$$

- linear normal models

  estimation of a noisy signal

  $X = \Theta + W$

  Θ and W: independent normals

  multi-dimensional versions (many normal parameters, many observations)

- estimating the parameter of a uniform

  X: uniform$[0, \Theta]$

  Θ: uniform $[0, 1]$

- $\widehat{\Theta} = g(X)$
- interested in:

  $\mathbf{E}\big[(\widehat{\Theta} - \Theta)^2 \mid X = x\big]$

  $\mathbf{E}\big[(\widehat{\Theta} - \Theta)^2\big]$

Linear models with normal noise

$$X_i = \sum_{j=1}^{m} a_{ij}\Theta_j + W_i \qquad\qquad W_i,\ \Theta_j\text{: independent, normal}$$

- Very common and convenient model
- Bayes' rule: normal posteriors
- MAP and LMS estimates coincide
  - simple formulas
    (linear in the observations)
- Many nice properties
- Trajectory estimation example

Recognizing normal PDFs

$X \sim N(\mu, \sigma^2)$ $$f_X(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-(x-\mu)^2/2\sigma^2}$$

$c \cdot e^{-8(x-3)^2}$

$f_X(x) = c \cdot e^{-(\alpha x^2 + \beta x + \gamma)}$ $\alpha > 0$ Normal with mean $-\beta/2\alpha$ and variance $1/2\alpha$

Estimating a normal random variable in the presence of additive normal noise

$$f_{\Theta|X}(\theta \mid x) = \frac{f_\Theta(\theta)\, f_{X|\Theta}(x \mid \theta)}{f_X(x)}$$

$$f_X(x) = \int f_\Theta(\theta) f_{X|\Theta}(x \mid \theta)\, d\theta$$

$X = \Theta + W$ $\quad\quad$ $\Theta, W : N(0,1)$, independent

$f_{X|\Theta}(x \mid \theta)$:

$f_{\Theta|X}(\theta \mid x) =$

$\widehat{\theta}_{\mathsf{MAP}} = \ \widehat{\theta}_{\mathsf{LMS}} = \mathbf{E}[\Theta \mid X = x] =$

$\widehat{\Theta}_{\mathsf{MAP}} = \mathbf{E}[\Theta \mid X] =$

Estimating a normal random variable in the presence of additive normal noise

$$f_{\Theta|X}(\theta \mid x) = \frac{f_\Theta(\theta)\, f_{X|\Theta}(x \mid \theta)}{f_X(x)}$$

$$f_X(x) = \int f_\Theta(\theta) f_{X|\Theta}(x \mid \theta)\, d\theta$$

$X = \Theta + W \qquad \Theta, W : N(0,1), \quad$ independent

$$\widehat{\Theta}_{\text{MAP}} = \widehat{\Theta}_{\text{LMS}} = \mathbf{E}[\Theta \mid X] = \frac{X}{2}$$

- Even with general means and variances:
  - posterior is normal
  - LMS and MAP estimators coincide
  - these estimators are "linear," of the form $\widehat{\Theta} = aX + b$

The case of multiple observations

$$X_1 = \Theta + W_1 \qquad \Theta \sim N(x_0, \sigma_0^2) \qquad W_i \sim N(0, \sigma_i^2)$$

$$\vdots$$

$$X_n = \Theta + W_n \qquad \Theta, W_1, \ldots, W_n \text{ independent}$$

$$f_{\Theta|X}(\theta \mid x) = \frac{f_\Theta(\theta)\, f_{X|\Theta}(x \mid \theta)}{f_X(x)}$$

$$f_X(x) = \int f_\Theta(\theta) f_{X|\Theta}(x \mid \theta)\, d\theta$$

$$f_{X_i|\Theta}(x_i \mid \theta) =$$

$$f_{X|\Theta}(x \mid \theta) =$$

$$f_{\Theta|X}(\theta \mid x) =$$

The case of multiple observations

$$f_{\Theta|X}(\theta \mid x) = c \cdot \exp\Big\{ -\text{quad}(\theta) \Big\} \qquad \text{quad}(\theta) = \frac{(\theta - x_0)^2}{2\sigma_0^2} + \frac{(\theta - x_1)^2}{2\sigma_1^2} + \cdots + \frac{(\theta - x_n)^2}{2\sigma_n^2}$$

$$\boxed{\widehat{\theta}_{\text{MAP}} = \widehat{\theta}_{\text{LMS}} = \mathbf{E}[\Theta \mid X = x] = \frac{\displaystyle\sum_{i=0}^{n} \frac{x_i}{\sigma_i^2}}{\displaystyle\sum_{i=0}^{n} \frac{1}{\sigma_i^2}}}$$

The case of multiple observations

- Key conclusions:
  - posterior is normal
  - LMS and MAP estimates coincide
  - these estimates are "linear," of the form $\hat{\theta} = a_0 + a_1 x_1 + \cdots + a_n x_n$
- Interpretations:
  - estimate $\hat{\theta}$: weighted average of x_0 (prior mean) and x_i (observations)
  - weights determined by variances

$$\boxed{\hat{\theta}_{\text{MAP}} = \hat{\theta}_{\text{LMS}} = \mathbf{E}[\Theta \mid X = x] = \frac{\displaystyle\sum_{i=0}^{n} \frac{x_i}{\sigma_i^2}}{\displaystyle\sum_{i=0}^{n} \frac{1}{\sigma_i^2}}}$$

The mean squared error

$$\widehat{\theta} = \frac{\sum_{i=0}^{n} \frac{x_i}{\sigma_i^2}}{\sum_{i=0}^{n} \frac{1}{\sigma_i^2}}$$

$$f_{\Theta|X}(\theta \,|\, x) = c \cdot \text{exp}\Big\{ -\text{quad}(\theta) \Big\}$$

$$\text{quad}(\theta) = \frac{(\theta - x_0)^2}{2\sigma_0^2} + \frac{(\theta - x_1)^2}{2\sigma_1^2} + \cdots + \frac{(\theta - x_n)^2}{2\sigma_n^2}$$

- Performance measures:

$$\mathbf{E}\Big[(\Theta - \widehat{\Theta})^2 \mid X = x\Big] = \mathbf{E}\Big[(\Theta - \widehat{\theta})^2 \mid X = x\Big] = \text{var}(\Theta \mid X = x) = 1 \Big/ \sum_{i=0}^{n} \frac{1}{\sigma_i^2}$$

$$\mathbf{E}\Big[(\Theta - \widehat{\Theta})^2\Big] =$$

$f_X(x) = c \cdot e^{-(\alpha x^2 + \beta x + \gamma)}$ $\quad \alpha > 0 \quad$ Normal with mean $-\beta/2\alpha$ and variance $1/2\alpha$

The mean squared error

$$\mathbf{E}\Big[(\Theta - \widehat{\Theta})^2 \mid X = x\Big] = \mathbf{E}\Big[(\Theta - \widehat{\Theta})^2\Big] = 1 \Big/ \sum_{i=0}^{n} \frac{1}{\sigma_i^2}$$

$$\hat{\theta} = \frac{\displaystyle\sum_{i=0}^{n} \frac{x_i}{\sigma_i^2}}{\displaystyle\sum_{i=0}^{n} \frac{1}{\sigma_i^2}}$$

- Example: $\sigma_0^2 = \sigma_1^2 = \cdots = \sigma_n^2 = \sigma^2$

- conditional mean squared error same for all x

- Example: $X = \Theta + W \qquad \Theta \sim N(0,1), \qquad W \sim N(0,1)$

  independent $\Theta, W \qquad \boxed{\widehat{\Theta} = X/2} \qquad \mathbf{E}\Big[(\Theta - \widehat{\Theta})^2 \mid X = x\Big] =$

Linear normal models

- Θ_j and X_i and are linear functions of independent normal random variables

- $f_{\Theta|X}(\theta \mid x) = c(x) \exp\Big\{ - \text{quadratic}(\theta_1, \ldots, \theta_m)\Big\}$

- MAP estimate: maximize over $(\theta_1, \ldots \theta_m)$;
  (minimize quadratic function)

  $\widehat{\Theta}_{\text{MAP},j}$: linear function of $X = (X_1, \ldots, X_n)$

- Facts:

  - $\widehat{\Theta}_{\text{MAP},j} = \mathbf{E}[\Theta_j \mid X]$

  - marginal posterior PDF of Θ_j: $f_{\Theta_j|X}(\theta_j \mid x)$, is normal

  - MAP estimate based on the joint posterior PDF:
    same as MAP estimate based on the marginal posterior PDF

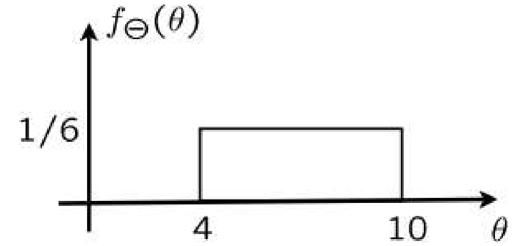
  - $\mathbf{E}\Big[(\widehat{\Theta}_{i,\text{MAP}} - \Theta_i)^2 \mid X = x\Big]$: same for all x

Least mean squares (LMS) estimation

- minimize (conditional) mean squared error $\mathbf{E}\Big[(\Theta - \hat{\theta})^2 \,|\, X = x\Big]$
  - solution: $\hat{\theta} = \mathbf{E}[\Theta \,|\, X = x]$
  - general estimation method
- Mathematical properties
- Example

LMS estimation in the absence of observations

- unknown Θ; prior $p_\Theta(\theta)$
  - interested in a point estimate $\hat{\theta}$
  - no observations available
  - MAP rule:
  - (Conditional) expectation:



- Criterion: Mean Squared Error (MSE): $\mathbf{E}\left[(\Theta - \hat{\theta})^2\right]$

  minimize mean squared error

LMS estimation in the absence of observations

- Least mean squares formulation:

minimize mean squared error (MSE), $\mathbf{E}\left[(\Theta - \widehat{\theta})^2\right]$: $\widehat{\theta} = \mathbf{E}[\Theta]$

- Optimal mean squared error: $\mathbf{E}\left[(\Theta - \mathbf{E}[\Theta])^2\right] = \mathrm{var}(\Theta)$

LMS estimation of Θ based on X

- unknown Θ; prior $p_\Theta(\theta)$
  - interested in a point estimate $\widehat{\theta}$
- observation X; model $p_{X|\Theta}(x\,|\,\theta)$
  - observe that $X = x$

minimize mean squared error (MSE), $\mathbf{E}\left[(\Theta - \hat{\theta})^2\right]$: $\hat{\theta} = \mathbf{E}[\Theta]$

minimize conditional mean squared error, $\mathbf{E}\left[(\Theta - \hat{\theta})^2 \,|\, X = x\right]$: $\hat{\theta} = \mathbf{E}[\Theta \,|\, X = x]$

- LMS estimate: $\widehat{\theta} = \mathbf{E}[\Theta \,|\, X = x]$

  estimator: $\widehat{\Theta} = \mathbf{E}[\Theta \,|\, X]$

LMS estimation with multiple observations or unknowns

- unknown Θ; prior $p_\Theta(\theta)$
  - interested in a point estimate $\widehat{\theta}$
- observations $X = (X_1, X_2, \ldots, X_n)$; model $p_{X|\Theta}(x \,|\, \theta)$
  - observe that $X = x$
  - new universe: condition on $X = x$
- LMS estimate: $\mathbf{E}[\Theta \mid X_1 = x_1, \ldots, X_n = x_n]$

- If Θ is a vector, apply to each component separately

Some challenges in LMS estimation

$$f_{\Theta|X}(\theta \mid x) = \frac{f_\Theta(\theta)\, f_{X|\Theta}(x \mid \theta)}{f_X(x)}$$

$$f_X(x) = \int f_\Theta(\theta') f_{X|\Theta}(x \mid \theta')\, d\theta'$$

- Full correct model, $f_{X|\Theta}(x \mid \theta)$, may not be available
- Can be hard to compute/implement/analyze

Linear least mean squares (LLMS) estimation

- Conditional expectation $\mathbf{E}[\Theta \mid X]$ may be hard to compute/implement
- Restrict to estimators $\widehat{\Theta} = aX + b$
  - minimize mean squared error
- Simple solution
- Mathematical properties

- Minimize $\mathbf{E}\left[(\widehat{\Theta} - \Theta)^2\right]$
- Estimators $\widehat{\Theta} = g(X) \longrightarrow \widehat{\Theta}_{\text{LMS}} = \mathbf{E}[\Theta \mid X]$
- Consider estimators of Θ, of the form $\widehat{\Theta} = aX + b$
- Minimize $\mathbf{E}\left[(\Theta - aX - b)^2\right]$, w.r.t. a, b
- If $\mathbf{E}[\Theta \mid X]$ is linear in X, then $\widehat{\Theta}_{\text{LMS}} = \widehat{\Theta}_{\text{LLMS}}$

Solution to the LLMS problem

- Minimize $\mathbf{E}\left[(\Theta - aX - b)^2\right]$, w.r.t. a, b
  - suppose a has already been found:

$$\boxed{\widehat{\Theta}_L = \mathbf{E}[\Theta] + \frac{\mathsf{Cov}(\Theta, X)}{\mathsf{var}(X)}\big(X - \mathbf{E}[X]\big) = \mathbf{E}[\Theta] + \rho\frac{\sigma_\Theta}{\sigma_X}\big(X - \mathbf{E}[X]\big)}$$

Remarks on the solution and on the error variance

$$\boxed{\widehat{\Theta}_L = \mathbf{E}[\Theta] + \frac{\text{Cov}(\Theta, X)}{\text{var}(X)}\big(X - \mathbf{E}[X]\big) = \mathbf{E}[\Theta] + \rho\frac{\sigma_\Theta}{\sigma_X}\big(X - \mathbf{E}[X]\big)}$$

- Only means, variances, covariances matter

- $\rho > 0$:

- $\rho = 0$:

$$\boxed{\mathbf{E}\Big[(\widehat{\Theta}_L - \Theta)^2\Big] = (1 - \rho^2)\,\text{var}(\Theta)}$$

- $|\rho| = 1$: