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

By

Professor Nikos Tsitsiklis

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

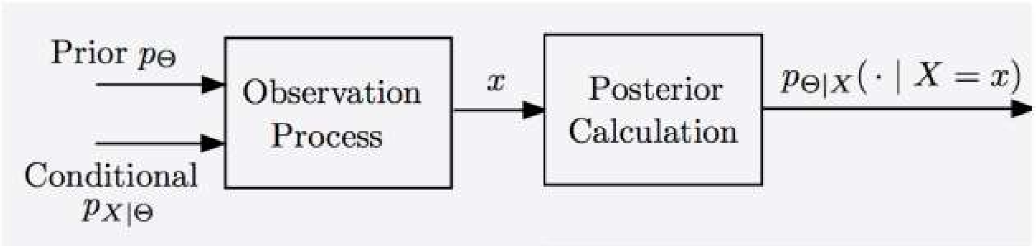
- Hypothesis testing:
  - unknown takes one of few possible values
  - aim at small probability of incorrect decision

  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

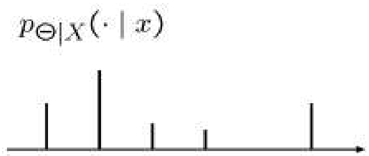
- Unknown Θ
  - treated as a random variable
  - prior distribution p_Θ or f_Θ
- Observation X
  - observation model $p_{X|\Theta}$ or $f_{X|\Theta}$
- Use appropriate version of the Bayes rule to find $p_{\Theta|X}(\cdot \mid X = x)$ or $f_{\Theta|X}(\cdot \mid X = x)$

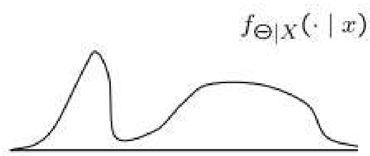


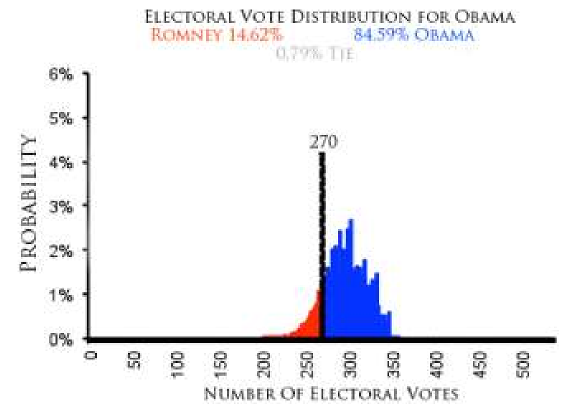
- Where does the prior come from?
  - symmetry
  - known range
  - earlier studies
  - subjective or arbitrary

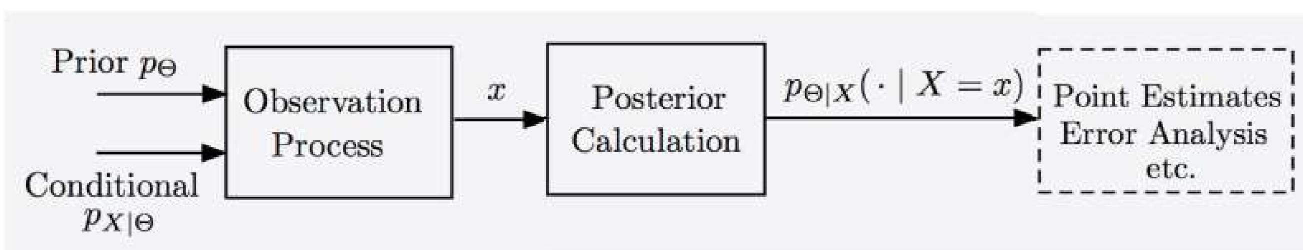
The output of Bayesian inference

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



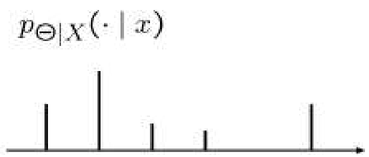


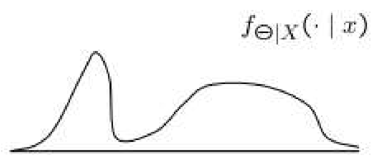




Point estimates in Bayesian inference

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):

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

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

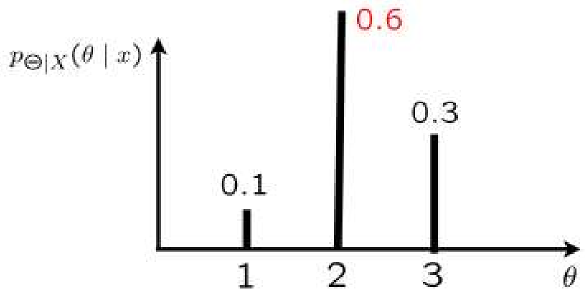
- Conditional expectation: $\mathbf{E}[\Theta \mid X = x]$ (LMS: Least Mean Squares)

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:

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

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

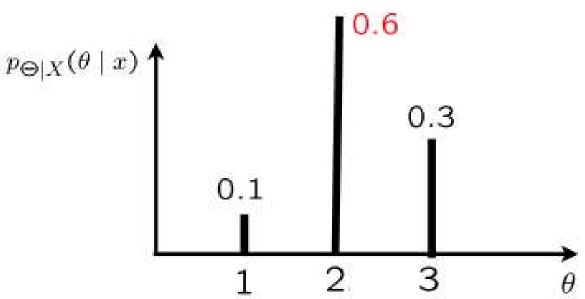
Discrete Θ, continuous X

- 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

$$X_i = \sum_{j=1}^{m} a_{ij}\Theta_j + W_i$$

W_i, Θ_j: 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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

The case of multiple observations

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

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

The mean squared error

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

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

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

- Performance measures:

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

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

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

The mean squared error

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

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

- 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 Θ, 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

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

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

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

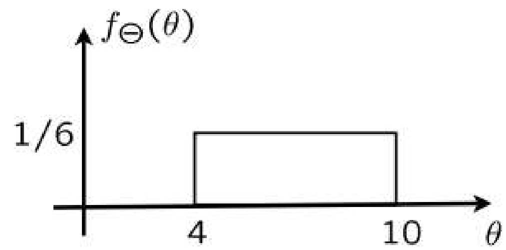
- Facts:
  - $\widehat{\Theta}_{\text{MAP},j} = \mathbf{E}[\Theta_j \,|\, X]$
  - marginal posterior PDF of Θ_j: $f_{\Theta_j|X}(\theta_j \,|\, x)$, is normal
  - MAP estimate based on the joint posterior PDF:
    same as MAP estimate based on the marginal posterior PDF
  - $\mathbf{E}\Big[(\widehat{\Theta}_{i,\text{MAP}} - \Theta_i)^2 \,|\, X = x\Big]$: same for all x

Least mean squares (LMS) estimation

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

- 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 $p_{X|\Theta}(x \mid \theta)$
  - observe that $X = x$

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

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

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

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

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 $p_{X|\Theta}(x \mid \theta)$
  - 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 $\mathbf{E}\Big[(\widehat{\Theta} - \Theta)^2\Big]$
- Estimators $\widehat{\Theta} = g(X) \quad \longrightarrow \quad \widehat{\Theta}_{\text{LMS}} = \mathbf{E}[\Theta \mid X]$
- 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:

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

Remarks on the solution and on the error variance

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

- Only means, variances, covariances matter

- $\rho > 0$:

- $\rho = 0$:

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

- $|\rho| = 1$:

LECTURE 19: The Central Limit Theorem (CLT)

- WLLN: $\dfrac{X_1 + \cdots + X_n}{n} \longrightarrow \mathbf{E}[X]$

- CLT: $X_1 + \cdots + X_n \approx$ normal
  - precise statement
  - universality, usefulness
  - many examples
  - refinement for discrete r.v.s
  - application to polling

Different scalings of the sum of i.i.d. random variables

- $X_1, \ldots, X_n$ i.i.d., finite mean μ and variance σ^2

- $S_n = X_1 + \cdots + X_n$

  variance: $n\sigma^2$

- $M_n = \dfrac{S_n}{n} = \dfrac{X_1 + \cdots + X_n}{n}$

  variance: $\dfrac{\sigma^2}{n}$

- $\dfrac{S_n}{\sqrt{n}} = \dfrac{X_1 + \cdots + X_n}{\sqrt{n}}$

  variance: σ^2

The Central Limit Theorem (CLT)

- $X_1, \ldots, X_n$ i.i.d., finite mean μ and variance σ^2
- $S_n = X_1 + \cdots + X_n$ variance: $n\sigma^2$
- $\dfrac{S_n}{\sqrt{n}} = \dfrac{X_1 + \cdots + X_n}{\sqrt{n}}$ variance: σ^2

$Z_n = \dfrac{S_n - n\mu}{\sqrt{n}\,\sigma}$ $\mathbf{E}[Z_n] =$ $\mathrm{var}(Z_n) =$

- Let Z be a standard normal r.v. (zero mean, unit variance)

Central Limit Theorem: For every z: $\lim_{n\to\infty} \mathbf{P}(Z_n \le z) = \mathbf{P}(Z \le z)$

- $\mathbf{P}(Z \le z)$ is the standard normal CDF, $\Phi(z)$, available from the normal tables

Usefulness of the CLT

$S_n = X_1 + \cdots + X_n$ $\qquad$ $Z_n = \dfrac{S_n - n\mu}{\sqrt{n}\,\sigma}$ $\qquad$ $Z \sim N(0,1)$

Central Limit Theorem: For every z: $\lim_{n\to\infty} \mathbf{P}(Z_n \le z) = \mathbf{P}(Z \le z)$

- universal and easy to apply; only means, variances matter
- fairly accurate computational shortcut
- justification of normal models

What exactly does the CLT say? — Practice

$S_n = X_1 + \cdots + X_n$ $\qquad$ $Z_n = \dfrac{S_n - n\mu}{\sqrt{n}\,\sigma}$ $\qquad$ $Z \sim N(0,1)$

Central Limit Theorem: For every z: $\lim_{n\to\infty} \mathbf{P}(Z_n \le z) = \mathbf{P}(Z \le z)$

- The **practice** of normal approximations:
  - treat Z_n as if it were normal
  - hence treat S_n as if normal: $\mathcal{N}(n\mu, n\sigma^2)$

- Can we use the CLT when n is "moderate"?
  - usually, yes
  - symmetry and unimodality help

Example 1

- $\mathbf{P}(S_n \le a) \approx b$ given two parameters, find the third
- Package weights X_i, i.i.d. exponential, $\lambda = 1/2$;
- Load container with $n = 100$ packages

$$Z_n = \frac{S_n - n\mu}{\sqrt{n}\,\sigma}$$

$$\mu = \sigma = 2$$

$$\mathbf{P}(S_n \ge 210)$$

	.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

Example 2

- $\mathbf{P}(S_n \leq a) \approx b$ given two parameters, find the third
- Package weights X_i, i.i.d. exponential, $\lambda = 1/2$;
- Let $n = 100$. Choose the "capacity" a, so that $\mathbf{P}(S_n \geq a) \approx 0.05$.

$$Z_n = \frac{S_n - n\mu}{\sqrt{n}\,\sigma}$$

$$\mu = \sigma = 2$$

	.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

Example 3

- $\mathbf{P}(S_n \le a) \approx b$ given two parameters, find the third
- Package weights X_i, i.i.d. exponential, $\lambda = 1/2$;
- How large can n be, so that $\mathbf{P}(S_n \ge 210) \approx 0.05$?

$$Z_n = \frac{S_n - n\mu}{\sqrt{n}\,\sigma}$$

$$\mu = \sigma = 2$$

	.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

Example 4

- $\mathbf{P}(S_n \le a) \approx b$ given two parameters, find the third

$$Z_n = \frac{S_n - n\mu}{\sqrt{n}\,\sigma}$$

- Package weights X_i, i.i.d. exponential, $\lambda = 1/2$;

$$\mu = \sigma = 2$$

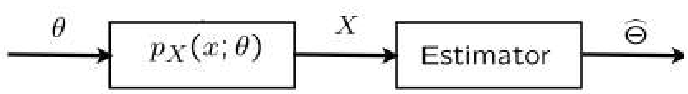
- Load container until weight exceeds 210
  N: number of packages loaded

- $\mathbf{P}(N > 100)$

	.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

Classical statistics

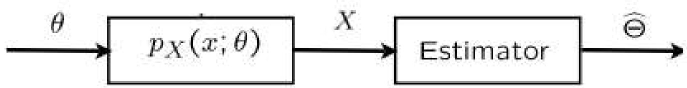
- Inference using the Bayes rule:
  unknown Θ and observation X are both random variables
  - Find $p_{\Theta|X}$
- Classical statistics: unknown constant θ



  - also for vectors X and θ: $p_{X_1,\ldots,X_n}(x_1,\ldots,x_n;\theta_1,\ldots,\theta_m)$
  - $p_X(x;\theta)$ are NOT conditional probabilities; θ is NOT random
  - mathematically: many models, one for each possible value of θ

Problem types in classical statistics

- Classical statistics: unknown constant θ



- Hypothesis testing: $H_0 : \theta = 1/2$ versus $H_1 : \theta = 3/4$

- Composite hypotheses: $H_0 : \theta = 1/2$ versus $H_1 : \theta \neq 1/2$

- Estimation: design an **estimator** $\widehat{\Theta}$, to "keep estimation **error** $\widehat{\Theta} - \theta$ small"

Estimating a mean

- $X_1, \ldots, X_n$: i.i.d., mean θ, variance σ^2

$\widehat{\Theta}_n =$ sample mean $= M_n = \dfrac{X_1 + \cdots + X_n}{n}$ $\qquad$ $\widehat{\Theta}_n$: estimator (a random variable)

Properties and terminology:

- $\mathbf{E}[\widehat{\Theta}_n] = \theta$ (unbiased)

- WLLN: $\widehat{\Theta}_n \to \theta$ (consistency)

- mean squared error (MSE): $\mathbf{E}\Big[(\widehat{\Theta}_n - \theta)^2\Big]$

On the mean squared error of an estimator

- For any estimator, using $\mathbf{E}[Z^2] = \text{var}(Z) + (\mathbf{E}[Z])^2$:

$$\mathbf{E}[(\widehat{\Theta} - \theta)^2] \;=\; \text{var}(\widehat{\Theta} - \theta) + \left(\mathbf{E}[\widehat{\Theta} - \theta]\right)^2 \;=\; \text{var}(\widehat{\Theta}) + (\text{bias})^2$$

- $\sqrt{\text{var}(\widehat{\Theta})}$ is called the standard error

Confidence intervals (CIs)

- The value of an estimator $\widehat{\Theta}$ may not be informative enough

- An $1-\alpha$ **confidence interval** is an interval $[\widehat{\Theta}^-, \widehat{\Theta}^+]$,

  s.t. $\mathbf{P}(\widehat{\Theta}^- \leq \theta \leq \widehat{\Theta}^+) \geq 1-\alpha, \quad$ for all θ

  - often $\alpha = 0.05$, or 0.025, or 0.01

  - interpretation is subtle

CI for the estimation of the mean

$$\widehat{\Theta}_n = \text{sample mean} = M_n = \frac{X_1 + \cdots + X_n}{n}$$

normal tables: $\Phi(1.96) = 0.975 = 1 - 0.025$

$$\mathbf{P}\left(\frac{|\widehat{\Theta}_n - \theta|}{\sigma/\sqrt{n}} \leq 1.96\right) \approx 0.95 \quad \text{(CLT)}$$

$$\boxed{\mathbf{P}\left(\widehat{\Theta}_n - \frac{1.96\,\sigma}{\sqrt{n}} \leq \theta \leq \widehat{\Theta}_n + \frac{1.96\,\sigma}{\sqrt{n}}\right) \approx 0.95}$$

Confidence intervals for the mean when σ is unknown

$$\mathbf{P}\left(\widehat{\Theta}_n - \frac{1.96\,\sigma}{\sqrt{n}} \le \theta \le \widehat{\Theta}_n + \frac{1.96\,\sigma}{\sqrt{n}}\right) \approx 0.95$$

- **Option 3:** Use sample mean estimate
  of the variance

Start from $\sigma^2 = \mathbf{E}[(X_i - \theta)^2]$

$$\frac{1}{n}\sum_{i=1}^{n}(X_i - \theta)^2 \quad \to \quad \sigma^2$$

(but do not know θ)

$$\frac{1}{n}\sum_{i=1}^{n}(X_i - \widehat{\Theta}_n)^2 \quad \to \quad \sigma^2$$

- Two approximations involved here:
  - CLT: approximately normal
  - using estimate of σ
- correction for second approximation (t-tables) used when n is small

Other natural estimators

- $\theta_X = \mathbf{E}[X]$ $\quad\quad \widehat{\Theta}_X = \frac{1}{n}\sum_{i=1}^{n} X_i$

- $\theta = \mathbf{E}[g(X)]$ $\quad\quad \widehat{\Theta} = \frac{1}{n}\sum_{i=1}^{n} g(X_i)$

- $v_X = \text{var}(X) = \mathbf{E}\Big[(X - \theta_X)^2\Big]$ $\quad\quad \widehat{v}_X = \frac{1}{n}\sum_{i=1}^{n}\left(X_i - \widehat{\Theta}_X\right)^2$

- $\text{cov}(X,Y) = \mathbf{E}\Big[(X - \theta_X)(Y - \theta_Y)\Big]$ $\quad\quad \widehat{\text{cov}}(X,Y) = \frac{1}{n}\sum_{i=1}^{n}\left(X_i - \widehat{\Theta}_X\right)\left(Y_i - \widehat{\Theta}_Y\right)$

- $\rho = \dfrac{\text{cov}(X,Y)}{\sqrt{v_x}\cdot\sqrt{v_Y}}$ $\quad\quad \widehat{\rho} = \dfrac{\widehat{\text{cov}}(X,Y)}{\sqrt{\widehat{v}_x}\cdot\sqrt{\widehat{v}_Y}}$

- next steps: find the distribution of $\widehat{\Theta}$, MSE, confidence intervals,...

Maximum Likelihood (ML) estimation

- $\theta = \mathbf{E}[g(X)]$ $\qquad$ $\widehat{\Theta} = \dfrac{1}{n}\displaystyle\sum_{i=1}^{n} g(X_i)$

- Pick θ that "makes data most likely"

  $$\hat{\theta}_{\mathrm{ML}} = \arg\max_{\theta}\ p_X(x;\theta)$$

  – also applies when x, θ are vectors or x is continuous

- compare to Bayesian posterior: $\qquad p_{\Theta|X}(\theta\,|\,x) = \dfrac{p_{X|\Theta}(x\,|\,\theta)\,p_{\Theta}(\theta)}{p_X(x)}$

  – interpretation is very different

Comments on ML

- maximize $p_X(x;\theta)$

- maximization is usually done numerically

- if have n i.i.d. data drawn from model $p_X(x;\theta)$, then, under mild assumptions:

  - consistent: $\widehat{\Theta}_n \to \theta$

  - asumptotically normal: $\dfrac{\widehat{\Theta}_n - \theta}{\sigma(\widehat{\Theta}_n)} \longrightarrow N(0,1)$ (CDF convergence)

- analytical and simulation methods for calculating $\hat{\sigma} \approx \sigma(\widehat{\Theta}_n)$

  - hence confidence intervals $\mathbf{P}\Big(\widehat{\Theta}_n - 1.96\,\hat{\sigma} \leq \theta \leq \widehat{\Theta}_n + 1.96\,\hat{\sigma}\Big) \approx 0.95$

  - asymptotically "efficient" ("best")

ML estimation example: parameter of binomial

- K: binomial with parameters n (known), and θ (unknown)

$$p_K(k;\theta) = \binom{n}{k}\theta^k(1-\theta)^{n-k}$$

$$\widehat{\theta}_{\text{ML}} = \frac{k}{n} \qquad \widehat{\Theta}_{\text{ML}} = \frac{K}{n}$$

- same as MAP estimator with uniform prior on θ

ML estimation example — normal mean and variance

- $X_1, \ldots, X_n$: i.i.d., $N(\mu, v)$ $\qquad f_X(x; \mu, v) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi v}} \exp\left\{ - \frac{(x_i - \mu)^2}{2v} \right\}$

  minimize $\quad \frac{n}{2} \log v + \sum_{i=1}^{n} \frac{(x_i - \mu)^2}{2v}$

  - minimize w.r.t. μ: $\qquad \hat{\mu} = \frac{x_1 + \cdots + x_n}{n}$

  - minimize w.r.t. v: $\qquad \hat{v} = \frac{1}{n} \sum_{i=1}^{n} (x_i - \hat{\mu})^2$