Lecture 12

- **Read:** Chapter 9.1-9.3.

Statistical Inference

- Linear Estimation of X Given Y
- MAP and ML Estimation
- Estimation of Model Parameters
  - Properties of Estimates
  - Estimation of the Expected Value of a Random Variable
  - Estimation of the Variance of a Random Variable
  - Confidence Interval Estimation

Statistical Inference

- Need to be able to reason in the presence of uncertainty
- Analyze observations of an experiment to reach conclusions with some assessment of the quality or risk associated with these conclusions

Statistical Inference: Various Scenarios

- **Significance/Hypothesis Testing:**
  - **Significance:** Given a single hypothesis H_0, figure out if it holds.
  - **Hypothesis:** Given several hypotheses $H_1, H_2, ..., H_n$, which one is "best"?
- **Estimation:**
  - **random variable:** Goal is to estimate an RV X, based on some observation (e.g., an event or another random variable).
  - **parameter estimation:** Here, X, might be modeled using some parametric distribution (e.g., $X \sim exp(\lambda)$), and based on observation we wish to estimate the true parameter λ.
    1. point estimate $\rightarrow \hat{\lambda}$
    2. confidence interval estimate
       $\rightarrow [a, b]$
       where, $P[\lambda \in [a, b]] \geq \underbrace{1 - \alpha}_{\text{confidence}}$

Statistical Inference: Point Estimation of a Parameter

- **Conclusion:** The value of a parameter of a probability model (e.g., expected value) is $\hat{c}$.
- **Accuracy Measure:** The mean square error: $E[(c - \hat{c})^2]$ (where, c is the true value.)
- **Question:** Assuming λ is a constant, what is $\hat{\lambda}$, the best estimate of λ?

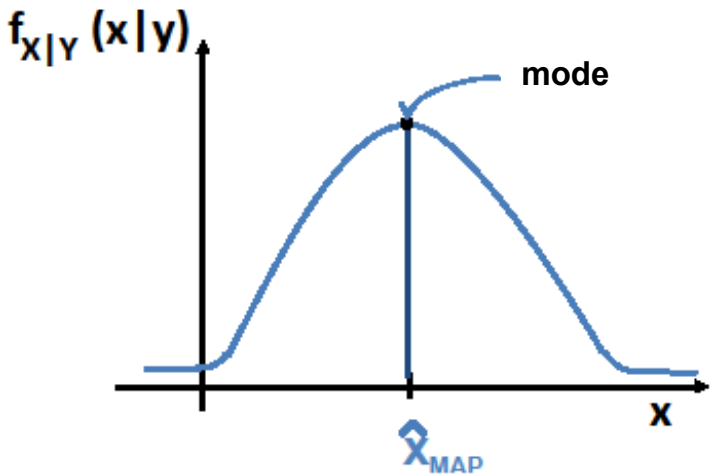
Statistical Inference: Confidence Interval Estimation Example

- **Conclusion:** The value, c, of a parameter of an RV is in the interval $a \leq c \leq b$.
- **Accuracy Measure:** The interval size $b - a$ and α, the probability that the conclusion is false.
- **Question:** Assuming λ is constant, what values of λ_1 and λ_2 satisfy $P[\lambda_1 \leq \lambda \leq \lambda_2] \geq 0.95$?

Estimation

- **Goal:** Estimate X given an observation of Y.
- **Solutions:**
  - MMSE estimator $\rightarrow \hat{X}_M(Y) = E[X|Y]$
  - LMSE estimator $\rightarrow \hat{X}_L(Y) = a^* Y + b^*$
    (a^* and b^* selected appropriately)

$$\min E\left[\left|X - \hat{X}_M(Y)\right|^2\right]$$



Linear Estimation of X Given Y

- For all $y \in S_Y$, the linear estimate is a single function $\hat{x}_L(y) = ay + b$.
- Linear estimates are easy to compute, sometimes optimal, require only means, variances, covariances.
- Linear mean square error (LMSE): $e_L = E[(X - \hat{X}_L(Y))^2]$
  - In this formula, we use $\hat{X}_L(Y)$ and not $\hat{x}_L(y)$ because the expected value in the formula is an unconditional expected value in contrast to the conditional expected value that is the quality measure for $\hat{x}_M(y)$.
  - Minimum mean square error estimation in principle uses a different calculation for each $y \in S_Y$.
  - In contrast, a linear estimator uses the same coefficients a and b for all y.

Estimation of a Random Variable: Case 4, What Is The "Best" Linear Estimator?

- **"Best" Linear Estimator:** $\min_{a,b} E[(X-(aY+b))^2]$
  - a, b depend only on $E[X]$, $E[Y]$, $Var[X]$, $Var[Y]$, and $Cov[X,Y]$.
  - Therefore, it is not necessary to know the complete probability model of X and Y.
- **Key Fact:** If X and Y are bivariate Gaussian:

$$\hat{X}_L(Y) = E[X|Y] = \hat{X}_M(Y)$$

- **Summary:** "best"= minimum mean square error (MMSE)
  - What is the best constant estimate, $\hat{x}_B$ for x?

$$\hat{x}_B = E[X]$$

  - What is the best estimate for X given $Y = y$?

$$\hat{x}_M(Y) = E[X|Y = y]$$

  - What is the best linear estimator, $\hat{X}_L(Y)$, of X given Y?

$$\hat{X}_L(Y) = a^*Y + b^*$$

Linear Mean Square Error (LMSE) Properties

- LMSE

$$\hat{X}_L(Y) = \rho_{X,Y}\sigma_X \frac{(Y - E[Y])}{\sigma_Y} + E[X]$$

As σ_Y increases, effect of observation $Y \neq E[Y]$ decreases

$$\rho_{X,Y} = \frac{Cov[X,Y]}{\sigma_X \sigma_Y} \Rightarrow \text{ if } \rho_{X,Y} = 0, \text{ ignore } Y \text{ (blind estimate)}$$

$$a^* = \frac{Cov[X,Y]}{Var[Y]}\ ,\ b^* = E[X] - a^* E[Y]$$

$$e_L^* = Var[X](1 - \rho_{X,Y}^2) \leftarrow \text{minimum mean square estimation error}$$

- What is the best estimator of X given Y?

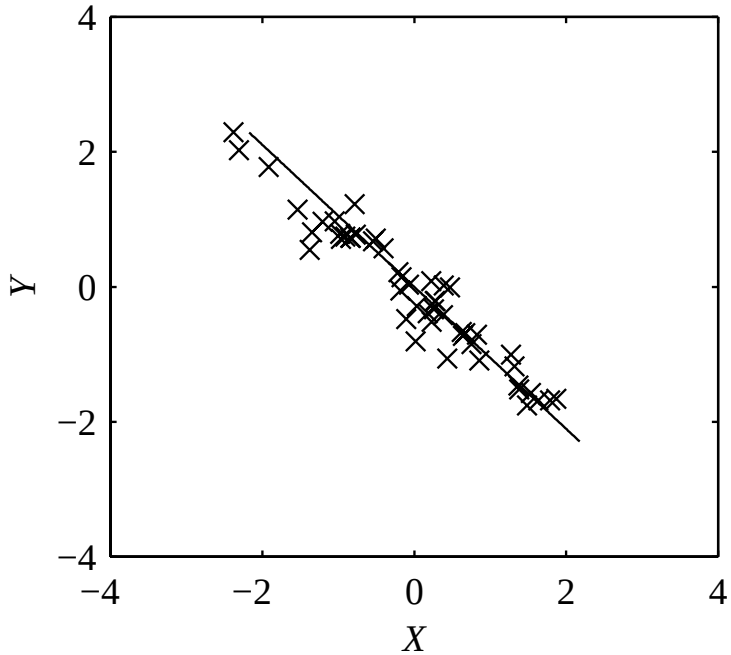
$$\hat{X}_M(Y) = E[X|Y]$$

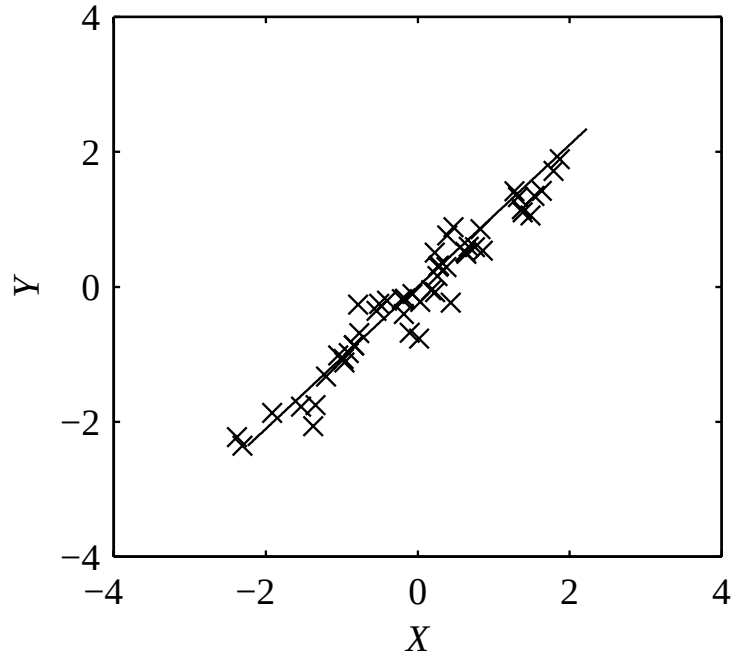
- When can you say for sure that $\hat{X}_L(Y) = \hat{X}_M(Y)$?
  - When X and Y are bivariate Gaussian random variables.

LMSE Examples I

- The magnitude of the correlation coefficient indicates the extent to which observing Y improves our knowledge of X, and the sign of $\rho_{X,Y}$ indicates whether the slope of the estimate is positive, negative, or zero.
- We will look at three different pairs of random variables X and Y.
- In each graph, the crosses are 50 outcomes (x, y) of the underlying experiment, and the line is the optimum linear estimate of X.
- In all cases, $E[X] = E[Y] = 0$ and $Var[X] = Var[Y] = 1$.
- Then, the optimum linear estimator of X given Y is the line $\hat{X}_L(Y) = \rho_{X,Y} Y$.
- For each pair (x, y), the estimation error equals the vertical distance to the estimator line.

LMSE Examples (II)

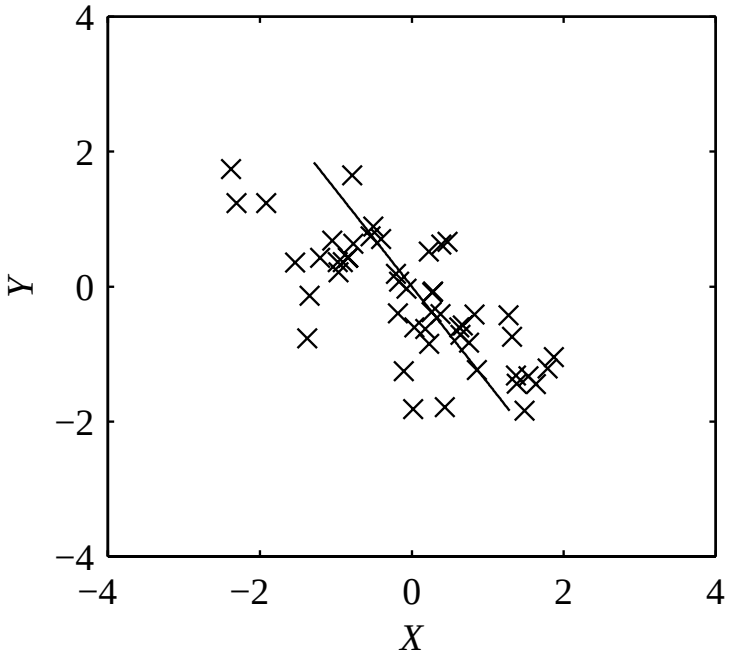


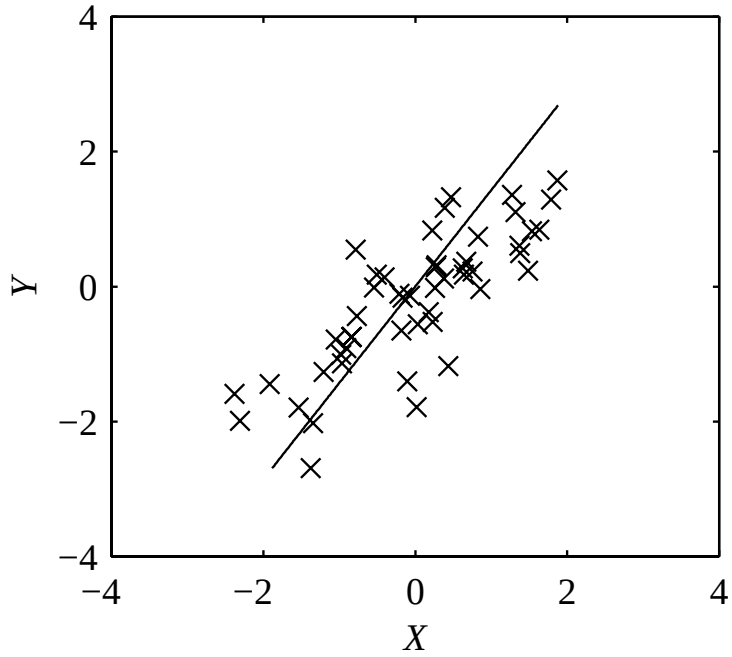


$$\rho_{X,Y} = -0.95 \qquad \rho_{X,Y} = 0.95$$

- All the observations are close to the estimate, which has a slope of -0.95 (left-hand graph) or 0.95 (right-hand graph).
- $e_L^* = Var[X](1-\rho_{X,Y}^2) = 1 \cdot (1-0.95^2) = 1-0.9025 = 0.0975$

LMSE Examples (III)

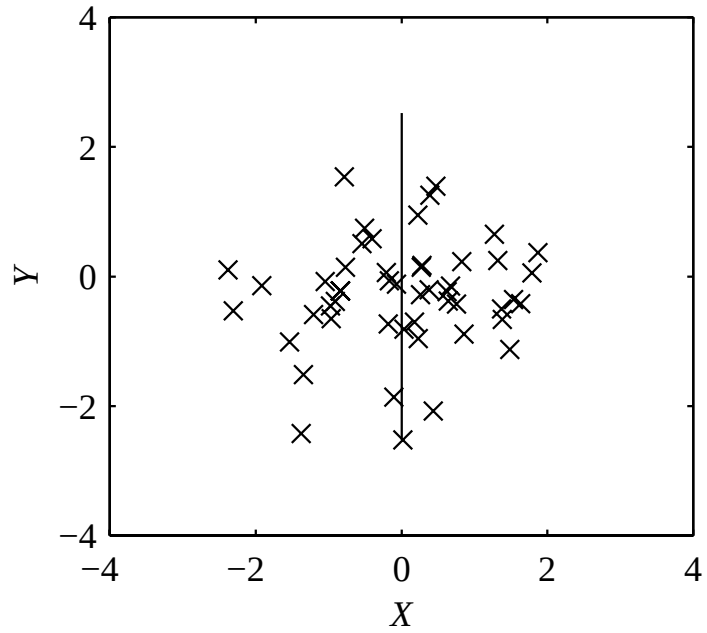




$\rho_{X,Y} = -0.70$ $\rho_{X,Y} = 0.70$

- In the left-hand graph, $\rho_{X,Y} = -0.7$, and the observations, on average, follow the estimator $\hat{X}_L(Y) = -0.7Y$, although the estimates are less accurate than those in the previous two graphs.
- $e_L^* = Var[X](1 - \rho_{X,Y}^2) = 1 \cdot (1 - 0.7^2) = 1 - 0.49 = 0.51$

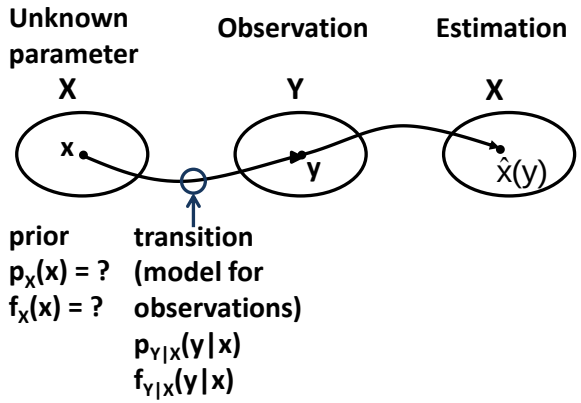
LMSE Examples (IV)



$\rho_{X,Y} = 0$

- With X and Y uncorrelated, the points are scattered randomly in the x-y plane and $e_L^* = Var[X] = 1$.

Estimation



Problems: $\hat{x}(y)$ estimate of X given observation $Y = y$.

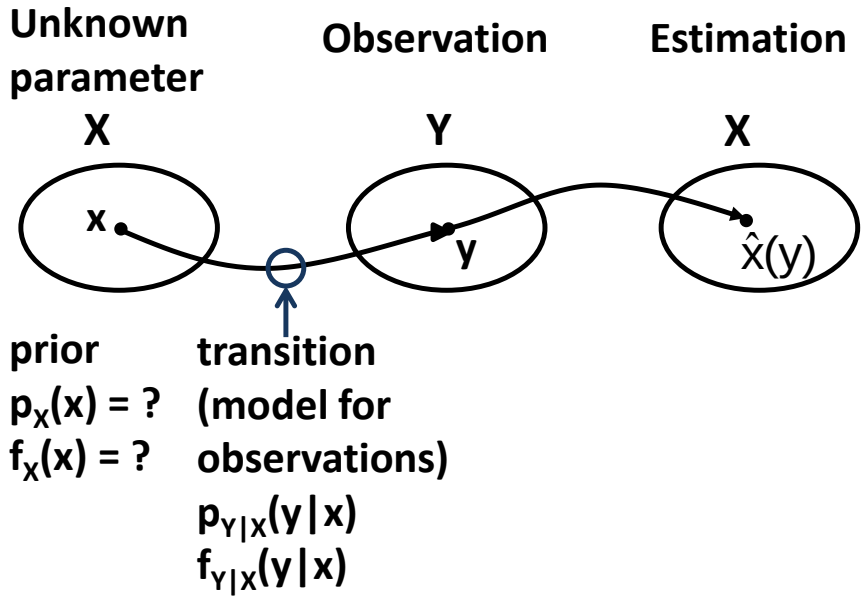
1. No prior (prior distribution is not known or modeled), x interpreted as an unknown parameter (parameter estimation, ML estimation)
2. Given prior distribution ($p_X(x)$ or $f_X(x)$) of unknown X (MMSE estimation, MAP estimation)

What are measures for goodness of estimators?

Applications:

- Measurement, approximations, model fitting
- Pattern matching, learning theory, tracking, control, ...

Parameter Estimation Problem

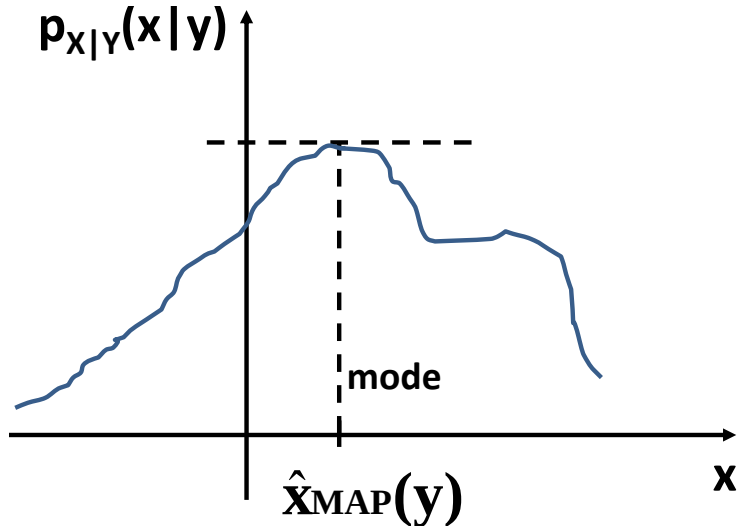


- **Setup:** Suppose the distribution of Y depends on some unknown parameter X. We will abbreviate this as $f_{Y|X}(y|x)$ if it is continuous or $p_{Y|X}(y|x)$ if it is discrete.
- **Goal:** Given observation of $Y = y$, find an estimator for the unknown parameter X.

Maximum a Posteriori (MAP) and ML Estimation

- We described methods for minimizing the mean square error in estimating a random variable X given a sample value of another random variable Y.
- We will now look at the maximum a posteriori probability (MAP) estimator and the maximum likelihood (ML) estimator.
- Although neither of these estimates produces the minimum mean square error, they are convenient to obtain in some applications, and they often produce estimates with errors that are not much higher than the minimum mean square error.
- As you might expect, MAP and ML estimation are closely related to MAP and ML hypothesis testing.
- We will describe these methods in the context of continuous random variables X and Y.

Maximum a Posteriori (MAP) Estimate



- **Setup:** Given prior $p_X(x)$ (or $f_X(x)$)
- **MAP Estimate:** $\hat{x}_{MAP}(y)$

$$\hat{x}_{MAP}(y) = \arg\max_x p_{X|Y}(x|y)$$

- **Remarks:**
  - Picks mode of $p_{X|Y}(x|y)$ (or $f_{X|Y}(x|y)$)
  - Often gives same estimate as MMSE

MAP Estimate

- MAP (Maximum probability of X given Y)

$$\hat{x}_{MAP}(y) = \arg\max_x f_{X|Y}(x|y)$$

- arg max: the argument x that gives the maximum for the function
- Essentially, MAP hypothesis testing over continuous range of hypotheses
- Does not minimize MSE but often very good
- Recall:

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

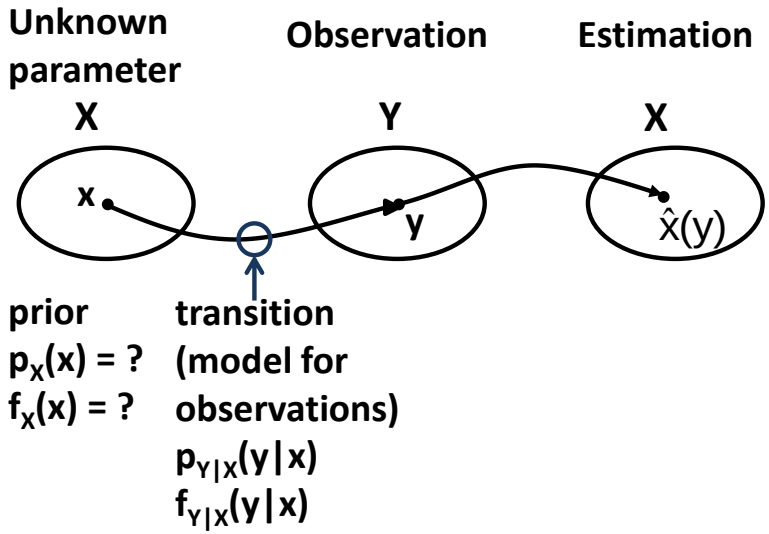
- Using that, MAP estimate (version 2):

$$\begin{aligned}\hat{x}_{MAP}(y) &= \arg\max_x f_{Y|X}(y|x)f_X(x) \\ &= \arg\max_x f_{X,Y}(x,y)\end{aligned}$$

MAP Estimation

- We see that the MAP estimation procedure requires that we know the PDF $f_X(x)$.
- That is, the MAP procedure needs the a priori probability model for random variable X.
- This is analogous to the requirement of the MAP hypothesis test that we know the a priori probabilities $P[H_i]$.
- In the absence of this a priori information, we can instead implement a maximum likelihood estimator.

Maximum Likelihood (ML) Estimation



- **Setup:** No prior specified (i.e., $p_X(x)$ (or $f_X(x)$) unknown)
- **ML Estimator:** $\hat{x}_{ML}(y)$

$$\hat{x}_{ML}(y) = \arg\max_x p_{Y|X}(y|x)$$

- **Remarks:**
  - no underlying probabilistic model for X needed
  - X could represent an unknown parameter of a system
    - e.g., estimate unknown mean of Y

Maximum Likelihood (ML) Estimate

- The ML estimate of X given $Y = y$:

$$\hat{x}_{ML}(y) = \arg\max_{x} f_{Y|X}(y|x)$$

- Can simply take derivative of $f_{Y|X}(y|x)$ with respect to x, set it equal to 0, and solve.
- Compared to MAP, ML ignores $f_X(x)$.
- Same as MAP if $f_X(x)$ is uniform over some range

Example: Unknown Success Probabilities of Coin Flips (I)

- Experiment produces a Bernoulli RV with success probability q.
- q is a sample of RV $Q \sim beta(2, 2)$.

$$f_Q(q) = \begin{cases} 6q(1-q) & , 0 \leq q \leq 1 \\ 0 & , \text{otherwise} \end{cases}$$

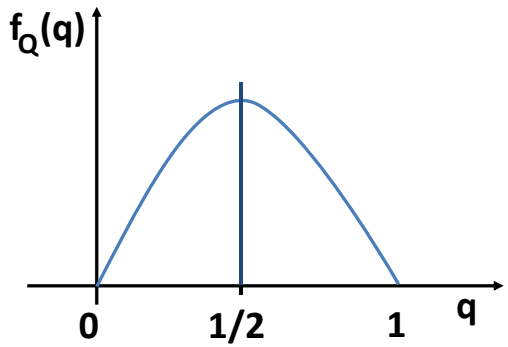
- Estimate Q using n trials ($K =$ number of successes):

$$p_{K|Q}(k|q) = \begin{cases} \binom{n}{k} q^k (1-q)^{n-k} & , k = 0, ..., n \\ 0 & , \text{otherwise} \end{cases}$$

Example: Unknown Success Probabilities of Coin Flips (II)

- Blind estimate:

$$\hat{q}_B = E[Q] = \frac{i}{i+j} = \frac{1}{2}$$



- ML: Choose q that maximizes $p_{K|Q}(k|q)$:

$$\hat{q}_{ML}(k) = \arg\max_{q \in [0,1]} p_{K|Q}(k|q)$$

Example: Unknown Success Probabilities of Coin Flips (III)

ML continued:

$$\hat{q}_{ML}(k) = \arg\max_{q \in [0,1]} p_{K|Q}(k|q)$$

$$\frac{\partial}{\partial q}\left[p_{K|Q}(k|q)\right] = 0$$

$$\frac{\partial}{\partial q}\left[\binom{n}{k} q^k (1-q)^{n-k}\right] = 0$$

$$k q^{k-1}(1-q)^{n-k} - q^k (n-k)(1-q)^{n-k-1} = 0$$

$$q^{k-1}(1-q)^{n-k-1}\left[k(1-q) - q(n-k)\right] = 0$$

$$k(1-q) = q(n-k)$$

$$k - kq = qn - kq$$

$$q = \frac{k}{n}$$

$$\therefore \hat{q}_{ML} = \frac{k}{n}$$

Example: Unknown Success Probabilities of Coin Flips (IV)

- Blind estimate: $1/2$, independent of outcomes
- ML: k/n, relative frequency, independent of prior information

Estimation of Model Parameters

- Parameter estimation
  - point estimators
  - interval estimators
- Experiments are performed in order to obtain information about parameters of an unknown probability model.
- Typical parameters: expected value, variance

Parameter Estimation

- **Problem:**
  - Make a sequence of experiments: $X_1, X_2, X_3, ...$
  - **Our goal:** is to estimate some parameter, r, of the underlying distribution for $X_1, X_2, ...$
  - **Example:** Estimate the mean $\mu_X = E[X_i]$.
- **Definition:** An estimator $\hat{R}$ is a function of the observations that are made.
- **Definition:** A sequence of estimators $\hat{R}_n = f(X_1, ..., X_n)$ can be considered for our problem.

Properties of Estimators

- **Definition:** An estimator $\hat{R}$ is unbiased if $E[\hat{R}] = r$ (the true parameter value).
- **Definition:** A sequence of estimators $\hat{R}_1, \hat{R}_2, \ldots$ is consistent if $\hat{R}_n \xrightarrow{p} r$, as $n \to \infty$
  - i.e., $\forall \epsilon > 0$, $P\left[\left|\hat{R}_n - r\right| \geq \epsilon\right] \to 0$, as $n \to \infty$.
- **Definition:** A sequence of estimators $\hat{R}_1, \hat{R}_2, \ldots$ is asymptotically unbiased if $\lim_{n\to\infty} E[\hat{R}_n] = r$.

Estimation of the Mean, μ_X

- **Theorem:** $M_n(X) = \frac{X_1+X_2+...+X_n}{n}$ is an unbiased, consistent estimator for $E[X] = \mu_X$.
- **Proof:** $E[M_n(X)] = \mu_X$: That is, $M_n(X)$ is an unbiased estimator for $E[X]$.

  Chebyshev inequality applied to $M_n(X)$:

$$P\left[|M_n(X) - \mu_X| \geq \epsilon\right] \leq \frac{Var[M_n(X)]}{\epsilon^2}$$

  Weak Law of Large Numbers says that for $M_n(X)$:

$$\lim_{n\to\infty} P\left[|M_n(X) - \mu_X| \geq \epsilon\right] = 0$$

  Combining the two:

$$P\left[|M_n(X) - \mu_X| \geq \epsilon\right] \leq \frac{Var[M_n(X)]}{\epsilon^2} = \frac{\sigma_X^2}{n\epsilon^2} \underset{n\to\infty}{\longrightarrow} 0$$

  So, $M_n(X)$ is consistent.

Estimation of the Variance (When μ_X Is Known)

- Unknown: $r = Var[X]$
- μ_X known. Let $W = (X - \mu_X)^2$. Then, $E[W] = \sigma_X^2$

  Weak Law of Large Numbers says that an unbiased, consistent estimator for σ_X^2 (or $E[W] = E[(X - \mu_X)^2]$) is the sample mean:

$$M_n(W) = \frac{1}{n}\sum_{i=1}^{n}(X_i - \mu_X)^2$$

Estimation of the Variance (When μ_X Is Unknown)

- Try using $M_n(X)$ (the estimator of μ_X) in place of μ_X

$$V_n(X) = \frac{1}{n}\sum_{i=1}^{n}(X_i - M_n(X))^2$$

- **Theorem:** $E[V_n(X)] = \frac{n-1}{n} Var[X]$
- $V_n(X)$ is biased, but asymptotically unbiased.
- An unbiased estimator for this case is:

$$V'_n(X) = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - M_n(X))^2$$

  - We lose one degree of freedom (because we use the sample to estimate the mean); that is why we divide the above sum by "n-1" instead of "n".

Point versus Interval Estimators

- We have discussed point estimators for unknown parameters, and shown various criteria for how good they are as $n \to \infty$.
- But, what if n is finite, i.e., we have to stop at some point?
- A confidence interval estimator is an interval that can be said to contain the true value with some degree of confidence. Its size gives an idea of how close we are to the true value.

Confidence Interval Estimates

- **Question:** Why do we need them?
- **Answer:** The previous point estimators for the parameters of the underlying distribution do not give any quality assessment of how good the estimate is.
- **Confidence intervals:** An interval estimator for an unknown parameter, r, is a pair of RVs, A, B, such that the probability that r is between a and b is greater than or equal to α.

$$P[r \in [A, B]] = P[A \leq r \leq B] \geq 1 - \alpha$$

  - $B - A$ is the confidence interval.
  - $1 - \alpha$ is the confidence coefficient.
  - An accurate estimate: low value of $B - A$, high value of $1 - \alpha$
  - The idea is to specify an interval or range and the probability that it will contain the true value r.

Finding Confidence Intervals (I)

- **Setup:** X_i are iid with $E[X_i] = \mu_X$ and $Var[X_i] = \sigma_X^2$.
- **Goal:** Estimate a confidence interval on μ_X.
- **Analysis:** $M_n(X) = \frac{1}{n}\sum X_i \xrightarrow{p} \mu_X$ as $n \to \infty$ (WLLN)

$$\frac{M_n(X)-\mu_X}{\sigma_X/\sqrt{n}} = \frac{\sum X_i - n\mu_X}{\sqrt{n}\sigma_X} \xrightarrow[n\to\infty]{d} Z \sim N(0,1) \text{ (CLT)}$$

$$\text{So, } P\left[|M_n(X) - \mu_X| \geq c\frac{\sigma_X}{\sqrt{n}}\right] \approx P\left[|Z| \geq c\right]$$

$$= 2[1 - \Phi(c)]$$

- **Note:** If $c = 1.96$, then $2[1 - \Phi(c)] = 0.05$.

Finding Confidence Intervals (II)

- **Rewriting the above:**

$$P\left[\mu_X \notin \left[M_n(X) - \frac{c\sigma_X}{\sqrt{n}}, M_n(X) + \frac{c\sigma_X}{\sqrt{n}}\right]\right] = 2[1 - \Phi(c)]$$

$$P\left[\mu_X \in \left[M_n(X) - \frac{c\sigma_X}{\sqrt{n}}, M_n(X) + \frac{c\sigma_X}{\sqrt{n}}\right]\right] = 1 - 2[1 - \Phi(c)]$$

- Suppose $c = 1.96$. Then, we can interpret this result as saying that with probability 0.95,

$$\mu_X \in \left[M_n(X) - \frac{c\sigma_X}{\sqrt{n}}, M_n(X) + \frac{c\sigma_X}{\sqrt{n}}\right]$$

- Thus, if we make $n = 10$ measurements, obtaining $X_1 = x_1, ..., X_{10} = x_{10}$, then we compute $\frac{1}{n}\sum x_i = \bar{x}$, our confidence interval is $\left[\bar{x} - \frac{c\sigma_X}{\sqrt{10}}, \bar{x} + \frac{c\sigma_X}{\sqrt{10}}\right]$.

Finding Confidence Intervals: Problem 1

- **Problem:** If we do not know σ_X^2, we usually just estimate it using

$$V_n'(X) = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - M_n(X))^2$$

- **Strategy:**
  1. Decide on confidence level, e.g., $0.95 \rightarrow c = 1.96$.
  2. Decide on an acceptable relative error (e.g., $d = 0.02\mu_X \rightarrow$ i.e., a % error you are willing to tolerate).
  3. Take sample of the RV X_i $(i = 1, ..., n)$ until

$$\frac{c\sigma_X}{\sqrt{n}} \leq d$$

  4. Confidence interval given by:

$$\left[\frac{1}{n}\sum x_i - d, \frac{1}{n}\sum x_i + d\right]$$

Finding Confidence Intervals: Problem 2

- **Problem:** We are estimating the mean of a Bernoulli RV X_i.

$$X_i = \begin{cases} 1 & \text{, with probability } p \\ 0 & \text{, with probability } 1-p \end{cases}$$

$E[X_i] = p,\ Var[X_i] = p(1-p)$

How many samples do I need before I have a 0.95 confidence interval with relative error 20%?

. . . . . .

- **Answer:** $\frac{1.96p(1-p)}{\sqrt{n}} \leq 0.20p \Rightarrow n \geq 100\frac{1}{p}$
(You will not be tested on this, just for your information)