Bayes rule (recap): let A, B be random events

$$\boxed{\Pr(A|B) =} \frac{\Pr(A\cap B)}{\Pr(B)}, \text{ assuming } \Pr(B)>0$$

$$= \frac{\Pr(B|A)\cdot\Pr(A)}{\Pr(B\cap(A\cup A^c))} = \frac{\Pr(B|A)\cdot\Pr(A)}{\Pr((B\cap A)\cup(B\cap A^c))}$$

$$= \frac{\Pr(B|A)\cdot\Pr(A)}{\Pr(B\cap A)+\Pr(B\cap A^c)}$$

since $B\cap A$ and $B\cap A^c$ are disjoint

$$= \boxed{\frac{\Pr(B|A)\cdot\Pr(A)}{\Pr(B|A)\cdot\Pr(A)+\Pr(B|A^c)\cdot\Pr(A^c)}}$$

Bayes rule updates $\Pr(A)$ (prior knowledge) with $\Pr(B|A)$ ("experimental data"); to get $\Pr(A|B)$ ("posterior probability / knowledge / information)

Pr(X = x) : prior prob. of [X=x]

p(Y = y | X = x): model for experimental data

MLE: Intuition via Bayes Rule I

$$\underbrace{Pr(\overbrace{X = x}^{A}|\overbrace{Y = y}^{B})}_{\text{posterior prob. } [X=x] \text{ when the experimental outcome is } [Y=y].} = \frac{Pr(X = x, Y = y)}{Pr(Y = y)}$$

$$= \frac{Pr(Y = y, X = x)}{\sum_{t \in \mathcal{X}} Pr(Y = y, X = t)}$$

the likelihood

$$= \frac{\boxed{Pr(Y = y|X = x)}\, Pr(X = x)}{\sum_{t \in \mathcal{X}} Pr(Y = y|X = t) Pr(X = t)}.$$

Flip a coin 100 times independently with the probability of success X; observe $Y = y$ successes (e.g., $y = 67$).

Game against the Nature: Nature chooses X (prob. of success in coin flip) from prior distribution. Statistican observes the event [Y = y]. Goal: deduce the value of X that Nature chose.

<u>**Q**</u>: What is your best guess about the true probability of success X, given that you observed y successes?

MLE: Intuition via Bayes Rule II

Flip a coin 100 times independently with the probability of success X; observe $Y = y$ successes (e.g., $y = 67$). Suppose X is a rv such that $Pr(X = i/100) = 1/101$ for $i = 0, 1, \dots, 100$.

"likelihood".

Q: What is the most likely value of X, given that you observed y successes?

$$Pr(X = x \mid Y = \overset{=67}{y}) = \frac{\boxed{Pr(Y=y \mid X=x)} \cdot Pr(X=x)}{\text{const}}$$

prior knowledge

$$= \binom{n}{y} \cdot x^y \cdot (1-x)^{n-y} \cdot \frac{1}{101} \cdot \frac{1}{\text{const}}.$$

$$= \binom{100}{67} \cdot x^{67} (1-x)^{33} \cdot \frac{1}{101 \cdot \text{const}}.$$

MLE is the Bayesian estimator of X that assumes that every value of X is equally likely a priori, $Pr[X=x] = c$ for all x.

want to maximize this expression wrt $x \in [0,1]$.

Solution: maximum likelihood estimator - value of x that maximizes the likelihood

Principle of Maximum Likelihood Estimation: discrete rvs

Let $x_1, x_2, \ldots, x_n$ be the observed values of iid rvs $X_1, X_2, \ldots, X_n$.

When the X_i's are discrete rvs,

$$Pr(X_1 = x_1, \ldots, X_n = x_n) = \prod_{i=1}^{n} Pr(X_i = x_i) = \prod_{i=1}^{n} f(x_i|\theta) > 0$$

is the probability of observing the vector of outcomes $[x_1, x_2, \ldots, x_n]$.

Since the events of high probability are more likely to occur than the events of low probability, and the event $[X_1 = x_1, \ldots, X_n = x_n]$ has occurred, it is sensible to estimate the unknown parameter θ using the value $\widehat{\theta}(x_1, \ldots, x_n)$ that makes $P(X_1 = x_1, \ldots, X_n = x_n|\theta)$ as high as possible.

Principle of ML Estimation: continuous rvs

Let $x_1, x_2, \ldots, x_n$ be the observed values of iid rvs $X_1, X_2, \ldots, X_n$.

When the X_i's are continuous rvs, $Pr(\bigcap_{i=1}^n [X_i = x_i] | \theta) = 0$. However, in this case

$$\prod_{i=1}^n f(x_i|\theta) \approx \frac{Pr(\bigcap_{i=1}^n [x_i - \delta/2 \leq X_i \leq x_i + \delta/2] | \theta)}{\delta^n} > 0.$$

Hence maximization of $\prod_{i=1}^n f(x_i|\theta)$ wrt θ is equivalent to maximization wrt θ the probability of the event that $\bigcap_{i=1}^n \{X_i \in [x_i - \delta/2, x_i + \delta/2]\}$.

Method of ML Estimation: Preliminaries

Likelihood function is the joint probability density or mass function of the data, treated as a function of θ, i.e.,

$$L(\theta|x_1, \ldots, x_n) = \prod_{i=1}^{n} f(x_i|\theta).$$

Notice that, in $L(\theta|x_1, \ldots, x_n)$, θ is the variable, and the sample $[x_1, \ldots, x_n]$ is treated as fixed.

Recall that in the pdf/pmf, θ is held fixed, and the x_i's vary. For convenience, $L(\theta|x_1, \ldots, x_n)$ will be abbreviated as $L(\theta)$.

Log-likelihood function is $l(\theta) = \log L(\theta)$. Here, $\log$ is typically taken to be the natural logarithm, $\ln$.

Example: Write down the likelihood and log-likelihood functions when $X_1, \ldots, X_n$ are iid $Bernoulli(p)$ rvs.

MLE: Procedure

Step 1: Write down the likelihood as a function of the parameter (vector) θ.

Step 2: Write down the log-likelihood as a function of the parameter (vector) θ, call it $l(\theta)$.

Step 3: Maximize the log-likelihood function with respect to θ. Often, but not always, this amounts to

Step 3a: solving for θ the *score equation*

$$S(\theta) = \frac{\partial l(\theta)}{\partial \theta} = 0.$$

Step 3b: if $\widehat{\theta}$ is the solution, checking that $\widehat{\theta}$ is indeed the maximizer of $l(\theta)$. Often, this amounts to checking that

$$\left.\frac{\partial^2 l(\theta)}{\partial \theta^2}\right|_{\theta=\widehat{\theta}} < 0.$$

Point and Interval Estimation: Some Motivation; I

In point estimation, the goal is to find an estimator $\widehat{\theta}_n$ for θ that has nice properties such as low MSE, low bias and consistency: $\Pr\left(|\widehat{\theta}_n - \theta| \leq \delta\right)$ is large.

However, no matter how large n is, usually $\Pr\left(\widehat{\theta}_n = \theta\right) = 0$, e.g., when $\widehat{\theta}_n$ is a continuous rv.

Suppose instead of "hitting" θ exactly, we constructed a random set (e.g., a *confidence interval*)

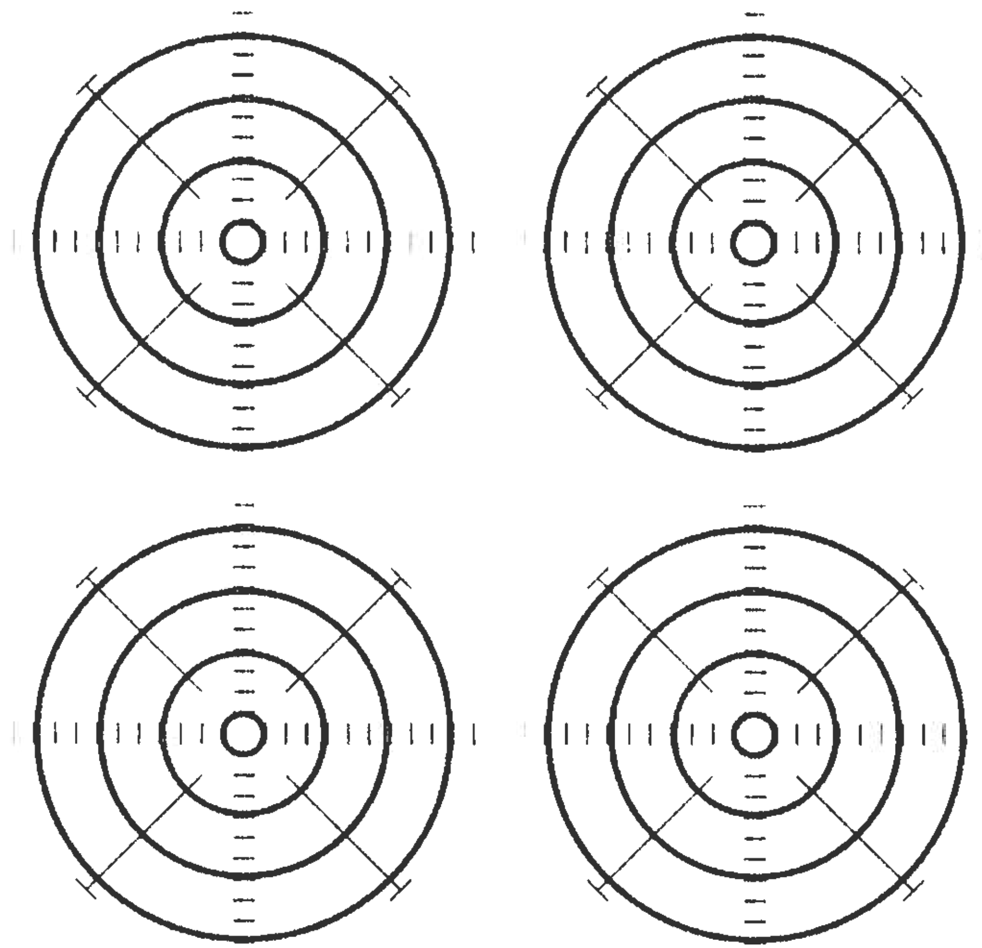
$$[L(X_1, \ldots, X_n|\alpha), U(X_1, \ldots, X_n|\alpha)]$$

such that

$$\Pr\left(\theta \in [L(X_1, \ldots, X_n|\alpha), U(X_1, \ldots, X_n|\alpha)]\right) = 1 - \alpha.$$

Here, α is some small number, e.g., 0.05 or 0.01.

Point and Interval Estimation: Some Motivation; II



Ingredients of the Confidence Intervals (CIs)

- L : lower bound, U : upper bound.

- $1-\alpha$: *confidence coefficient*
  $\equiv$ probability of coverage $\equiv$ level of the CI.

- If L and U are both finite, $[L, U]$ is called a 2-sided interval.

- If $|L|$ or U (but not both) is infinity, the CI is called one-sided. If $L = -\infty$, then the CI is called left-sided. If $U = \infty$, then the CI is called right-sided.

- Note that $[L(X_1, \ldots, X_n|\alpha), U(X_1, \ldots, X_n|\alpha)]$ is a random interval; this is an *interval estimatOR* of θ.

- When $(X_1, \ldots, X_n) = (x_1, \ldots, x_n)$,
  $[L(x_1, \ldots, x_n|\alpha), U(x_1, \ldots, x_n|\alpha)]$ is the *interval estimatE* of θ (a particular realization of the random interval estimator).

Large-sample justification of a level $(1-\alpha)$ CI

Let $X_{i,j}$ for $i=1,\ldots,m$ and $j=1,\ldots,n$ be iid from F_θ. Let $L_i = L\left(X_{i,1}, X_{i,2}, \ldots, X_{i,n}\right)$ and $U_i = U\left(X_{i,1}, X_{i,2}, \ldots, X_{i,n}\right)$.

Write the $X_{i,j}$'s in a matrix and compute $[L_i, U_i]$ for each row:

$$\begin{bmatrix} X_{1,1} & X_{1,2} & \cdots & X_{1,n} \\ X_{2,1} & X_{2,2} & \cdots & X_{2,n} \\ \vdots & \vdots & \ddots & \vdots \\ X_{m,1} & X_{m,2} & \cdots & X_{m,n} \end{bmatrix}$$

Interpretation: If the "experiment" $(X_1, X_2, \ldots, X_n)$ is independently replicated m times and the CI is computed each time, then the frequency of coverage $\frac{S_m}{m}$ of θ by the random intervals $[L_1, U_1], \ldots, [L_m, U_m]$ tends to $1-\alpha$ as $m \to \infty$.