STA 602 – Intro to Bayesian Statistics

Lecture 6

Li Ma

Point estimation

- Let us come back to the point estimation problem—the "guessing" of the value of a parameter θ *based on observed data* $\mathbf{X} = (X_1, X_2, \ldots, X_n)$.
- Such guesses, which are *functions of the data*, are called *estimators* for the parameter. Common notations: $\hat{\theta}(\mathbf{X})$, $\delta(\mathbf{X})$, etc. Estimators are essentially "rules" that map data to guesses.
- If the observed data is $\mathbf{X} = \mathbf{x}$, where $\mathbf{x} = (x_1, x_2, \ldots, x_n)$, the realized value of an estimator $\delta(\mathbf{X})$ is $\delta(\mathbf{x})$, which is called an *estimate*. Estimates are the actual guesses.
- What is a criterion for "good" estimators?
- A good estimator should be such that the estimate and the actual parameter θ are *"likely to be close"*.

What does “likely to be close” mean?

2. The sampling view:
   - The parameter is a *fixed* unknown number. The only random quantities are the data.
   - After data is observed, however, nothing is random. No matter what estimator $\delta(\mathbf{X})$ we are considering, we cannot judge how close the parameter θ is to a *single realization of* the estimate $\delta(\mathbf{x})$ after $\mathbf{X} = \mathbf{x}$ is observed.
   - In this case, we want to choose an estimator $\delta(\mathbf{X})$ that will “with high probability” take values close to the underlying fixed θ.
   - Such a probabilistic statement can only be made *before the experiment*, or under repeated experiment.

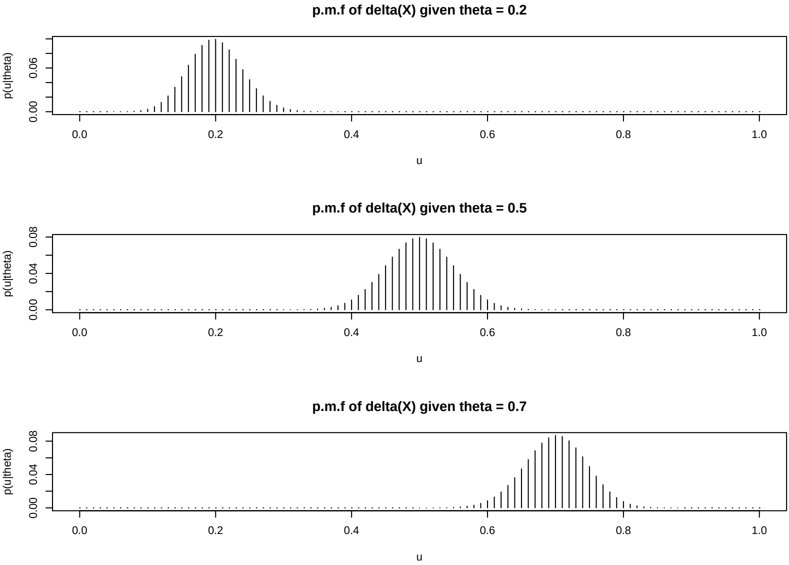
Note that this is a “before the experiment” *view, in contrast to the “after the experiment” view taken by the Bayesian perspective.*

The sampling distribution of an estimator

- Now we take the view of a *fixed* but *unknown* parameter θ.
- The data $\mathbf{X} = (X_1, X_2, \ldots, X_n)$ are jointly distributed random variables with a joint distribution $p(\mathbf{x}|\theta)$, presumably different for each θ.
- Any estimator $\delta(\mathbf{X})$ is also a random variable, and will have a sampling distribution under repeated experiments.

Example: the political poll revisited

- Sampling distribution of the estimator $\delta(X) = X/n$ for $n = 100$.



p.m.f of delta(X) given theta = 0.2

p.m.f of delta(X) given theta = 0.5

p.m.f of delta(X) given theta = 0.7

Example: measuring air pollutant

If n independent readings $X_1, X_2, \ldots, X_n$ are taken from the distribution $\mathrm{N}(\theta, \sigma^2)$ with known σ, then what is the sampling distribution of the *sample mean estimator* (given θ)?

$$\delta(\mathbf{X}) = \delta(X_1, X_2, \ldots, X_n) = \bar{X} = \frac{\sum_{i=1}^{n} X_i}{n}.$$

Note that

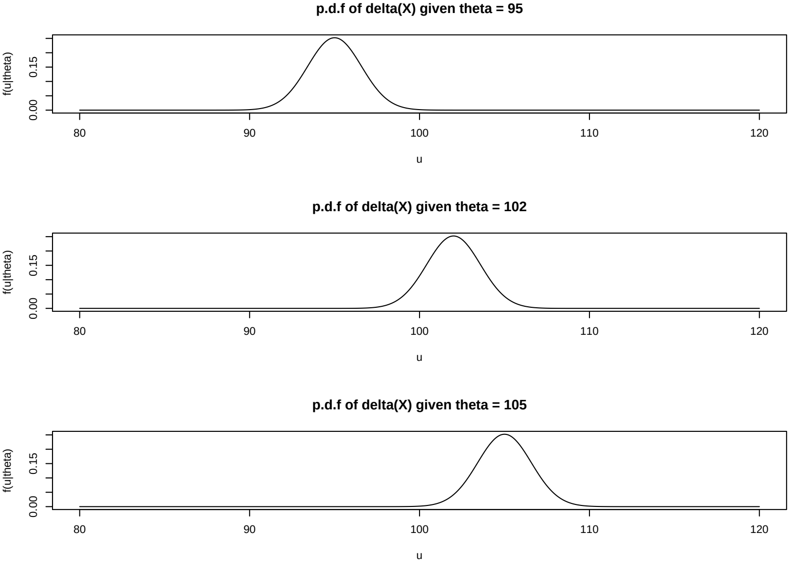
$$\mathrm{E}[\bar{X}|\theta] = \frac{\sum_{i=1}^{n} \mathrm{E}(X_i|\theta)}{n} = \frac{n\theta}{n} = \theta$$

and

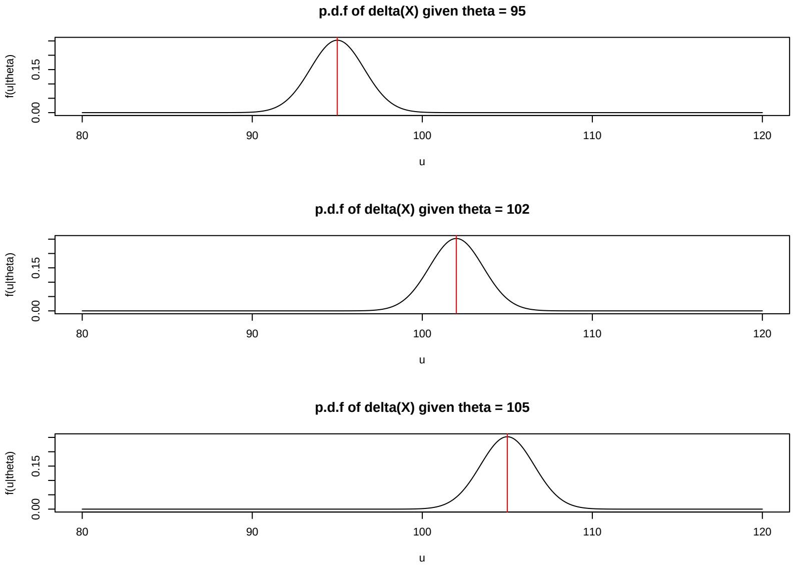
$$\mathrm{Var}(\bar{X}|\theta) = \mathrm{Var}\left(\frac{\sum_{i=1}^{n} X_i}{n}\middle|\theta\right) = \frac{\sum_{i=1}^{n} \mathrm{Var}(X_i|\theta)}{n^2} = \frac{n\sigma^2}{n^2} = \frac{\sigma^2}{n}.$$

Since the sum of independent normal random variables are still normal random variables, its sampling distribution is $\mathrm{N}(\theta, \sigma^2/n)$.

Sampling distribution of $\delta(\mathbf{X})$ for $n = 10$, $\sigma = 5$



Sampling distribution of $\delta(\mathbf{X})$ for $n = 10$, $\sigma = 5$



p.d.f of delta(X) given theta = 95

p.d.f of delta(X) given theta = 102

p.d.f of delta(X) given theta = 105

Criteria for selecting “good” estimators

Two desirable properties of estimators under repeated experiments:

- “Overall accuracy”: $\mathrm{E}(\delta(X)|\theta)$ is close to θ.
- “Precision”: $\mathrm{Var}(\delta(X)|\theta)$ is small.

Bias of an estimator

An estimator for θ is said to be *unbiased* if its overall accurate for all possible values of θ. That is

$$\mathrm{E}(\delta(X)|\theta) = \mathrm{E}_\theta(\delta(X)) = \mathrm{E}_\theta(\delta) = \theta \quad \textit{for all } \theta.$$

The *bias* of $\delta(X)$ given θ is defined to be

$$\mathrm{Bias}_\delta(\theta) = \mathrm{E}(\delta|\theta) - \theta.$$

- $\mathrm{Bias}_\delta(\theta) > 0$: $\delta(X)$ tends to overestimate θ.
- $\mathrm{Bias}_\delta(\theta) < 0$: $\delta(X)$ tends to underestimate θ.

Back to the political poll example

- The estimator $\delta_1(X) = \frac{X}{n}$ is unbiased, i.e., $\text{Bias}_{\delta_1}(\theta) = 0$ for all θ or
$$\text{E}(\delta_1|\theta) = \theta \quad \text{for all } \theta.$$
- The estimator $\delta_2(X) = \frac{X+12}{n+24}$ is not unbiased.
$$\text{Bias}_{\delta_2}(\theta) = \text{E}(\delta_2|\theta) - \theta = \frac{n\theta + 12}{n+24} - \theta = \frac{12 - 24\theta}{n+24}.$$
So for $\theta > 1/2$, $\text{Bias}_{\delta_2}(\theta) < 0$ and for $\theta < 1/2$, $\text{Bias}_{\delta_2}(\theta) > 0$.

Question: Is $\delta_1(X)$ always more desirable than $\delta_2(X)$?

Decision theoretic criteria for “good” estimators

Recall from last time the goal of constructing an estimator $\delta(\mathbf{X})$ so that $\delta(\mathbf{X})$ will *likely to be close to* θ.

- Again, we need a notion of distance between the estimate and the parameter.
- Can again use a loss function just like before such as
  1. The absolute error loss: $L(\theta, a) = |\theta - a|$.
  2. The squared error loss: $L(\theta, a) = (\theta - a)^2$.
  3. The step error loss: $L(\theta, a) = \mathbf{1}(|\theta - a| > \Delta)$.

  So we can again try to choose an estimator that minimizes the *expected loss*.

But now the expectation is taken over the distribution of the estimator given θ:

$$\mathrm{E}(L(\theta,\delta(\mathbf{X}))|\theta) = \int_{-\infty}^{\infty} L(\theta,\delta(\mathbf{x}))p(\mathbf{x}|\theta)d\mathbf{x}.$$

This expectation is also called the *(sampling) risk function* of estimator δ:

$$R_\delta(\theta) = R(\delta,\theta) := \mathrm{E}(L(\theta,\delta(\mathbf{X}))\,|\,\theta). \quad \text{(Here } \delta(\mathbf{X}) \text{ is random.)}$$

Contrast this with the Bayes risk for an estimator δ,

$$r(\delta,\mathbf{x}) = \mathrm{E}(L(\theta,\delta(\mathbf{x}))|\mathbf{x}). \quad \text{(Given } \mathbf{x}, \ \delta(\mathbf{x}) \text{ is a specific estimate.)}$$

Note that for any given estimator δ, $R(\delta,\theta)$ depends on the sampling model alone, while $r(\delta,\mathbf{x})$ depends also on the prior.

Two different objectives in finding “optimal” decisions

- Frequentist:
  - Find a *decision rule* $\delta : S \to \mathscr{A}$ that has small $R(\delta, \theta)$—that is, small average loss under repeated experiments.

    $$\delta^*(\cdot) := \text{argmin}_\delta \, R_\delta(\theta)$$

    But for what θ? For all θ?
  - That is often not possible for any given n.
  - Three common strategies
    - Place additional constraints on δ, e.g., unbiased. (Occassionally useful.)
    - Minimize the worst case risk rather than minimize the risk function over all θ, e.g., minimax. (Broadly useful.)

      $$\delta^{**}(\cdot) = \text{argmin}_\delta \, \text{max}_\theta \, R_\delta(\theta)$$

    - Assume $n \to \infty$. (Broadly useful.)

Two different objectives in finding "optimal" decisions

- Bayesian:
  - Find a specific decision a that minimizes $r(a, \mathbf{x})$—that is, small posterior average loss for the particular data set at hand, and call that our estimate

    $$\delta^*(\mathbf{x}) := \text{argmin}_\delta r(\delta, \mathbf{x}) = \text{argmin}_a r(a, \mathbf{x}).$$

  - Just as the Bayes risk $r(a, \mathbf{x})$ depends on the prior of θ (since the posterior depends on the prior), so does the minimizer $\delta^*(\mathbf{x})$. So to be explicit, one often writes δ_Λ as the Bayes rule for a prior Λ.
  - This is a much simpler objective and generally has a solution under common models and losses.
- The corresponding mapping $\delta^* : S \to \mathscr{A}$ is the *Bayes (decision) rule*. It's an estimator, which represents what a Bayesian with a given prior would do if one analyzes multiple data sets that arise from this experiment.
- One can show that it actually minimizes the weighted average of the frequentist risk function where the weight is the prior.

$$\delta^* = \text{argmin}_\delta \mathrm{E} R_\delta(\theta) = \int R_\delta(\theta) p(\theta) d\theta.$$

Frequentist properties of estimators

- Whether one takes the Bayesian or frequentist approach, the frequentist risk function is a helpful measure of the performance under *repeated applications*.
- That is, one can evaluate how a Bayesian with a given prior will do if the experiment is repeated!
- Surprisingly, Bayes rules often fare well even from the repeated experiment perspective! (Topics to be covered in STA 532/732).
    - Interestingly, the Bayes rule often solves the frequentist problem when $n \to \infty$!
    - Moreover, essentially all *admissible* decision rules are Bayesian under some priors! (Complete class theorems.)
    - A solution to the minimax problem is often a Bayes rule or the limit of a sequence of Bayes rule!

Examples of risk functions

- For absolute error loss

$$\begin{aligned} R_\delta(\theta) = \mathrm{MAE}_\delta(\theta) :&= \mathrm{E}\left[|\delta(X) - \theta||\theta\right] \\ &= \mathrm{E}_\theta\left[|\delta - \theta|\right] \\ &= \int_{-\infty}^{\infty} |\delta(x) - \theta| p(x|\theta) dx. \end{aligned}$$

  This is called the *mean absolute error (risk)*.

- For squared error loss

$$\begin{aligned} R_\delta(\theta) = \mathrm{MSE}_\delta(\theta) :&= \mathrm{E}\left[(\delta(X) - \theta)^2|\theta\right] \\ &= \mathrm{E}_\theta\left[(\delta - \theta)^2\right] \\ &= \int_{-\infty}^{\infty} (\delta(x) - \theta)^2 p(x|\theta) dx. \end{aligned}$$

  This is called the *mean squared error (risk)*.

Note that these expectations are computed using only information available *before the experiment*, not using the observed value of the data.

- These are the *average* distances between the estimator $\delta(\mathbf{X})$ and the parameter θ *if the experiment is repeated many times* under fixed parameter value θ. (Recall the “frequentist” viewpoint.)
- Given an observation $\mathbf{X} = \mathbf{x}$, the corresponding estimate is $\delta(\mathbf{x})$. Note that the estimator $\delta(\mathbf{X})$ is chosen using information available *before the experiment*. The estimate given data $\mathbf{X} = \mathbf{x}$ is simply the plug-in value of that estimator.
- Contrast this with the Bayesian approach to estimation, where we use the posterior given the data to find the estimate, and then find the corresponding estimator.
- Under the sampling perspective, we know nothing about how far our realized estimate $\delta(\mathbf{x})$ is from the underlying θ.
- This is a price to pay when treating parameters as fixed quantities.

While the mean absolute error (MAE) seems to be the most natural criterion for judging “average closeness”, the mean squared error (MSE) is the most popular one to use due to ease of computation.

$$
\begin{aligned}
\mathrm{MSE}_\delta(\theta) &= \mathrm{E}\left((\delta(X)-\theta)^2|\theta\right) \\
&= \mathrm{E}[(\delta-\mathrm{E}(\delta|\theta))+(\mathrm{E}(\delta|\theta)-\theta)\,|\,\theta]^2 \\
&= \mathrm{E}(\delta-\mathrm{E}(\delta|\theta)|\theta)^2+2(\mathrm{E}(\delta|\theta)-\theta)\mathrm{E}(\delta-\mathrm{E}(\delta|\theta)|\theta) \\
&\qquad\qquad +(\mathrm{E}(\delta|\theta)-\theta)^2 \\
&= \mathrm{E}(\delta-\mathrm{E}(\delta|\theta)|\theta)^2+(\mathrm{E}(\delta|\theta)-\theta)^2 \\
&= \mathrm{Var}(\delta|\theta)+\mathrm{Bias}_\delta(\theta)^2.
\end{aligned}
$$

This is sometimes referred to as the *Bias-Variance trade-off*. We want estimators that strike a balance between small bias and small variability.

- It may be worth it to take a biased estimator if its variance is much smaller than an alternative unbiased one. (Draw a figure.)

Example: Political poll

We have data $X \sim \text{Binomial}(n, \theta)$. Let us now consider three estimators for θ.

$$\delta_1(X) = \frac{X}{n} \quad \text{(The sample mean. It's unbiased.)}$$

$$\delta_2(X) = \frac{1}{2} \quad \text{(A ``stubborn'' estimator. It has zero variance.)}$$

$$\delta_3(X) = \frac{X+12}{n+24} \quad \text{(What esimator is this?)}$$

$$\delta_4(X) = \frac{X+1}{n+2} \quad \text{(What estimator is this?)}$$

Recall that the Bayes estimators δ_3 and δ_4 are weighted averages of δ_1 and δ_2. What are the weights?

Next, let us evaluate and compare these three esimators in their mean squared error.

$\delta_1(X) = \frac{X}{n}$ (The sample mean.)

$$\begin{aligned}\mathrm{MSE}_{\delta_1}(\theta) &= \mathrm{E}_\theta(\delta_1(X) - \theta)^2 = \mathrm{E}_\theta\left(\frac{X}{n} - \theta\right)^2 \\ &= \mathrm{Var}_\theta\left(\frac{X}{n}\right) = \frac{\mathrm{Var}_\theta(X)}{n^2} = \frac{n\theta(1-\theta)}{n^2} = \frac{\theta(1-\theta)}{n}.\end{aligned}$$

$\delta_2(X) = \frac{1}{2}$ (A "stubborn" estimator.)

$$\mathrm{MSE}_{\delta_2}(\theta) = \mathrm{E}_\theta\left(\frac{1}{2} - \theta\right)^2 = \left(\frac{1}{2} - \theta\right)^2 = \frac{1}{4} - \theta(1-\theta).$$

- Note that this is $\mathrm{Bias}_{\delta_2}(\theta)^2$ as this estimator has variance 0.

$\delta_3(X) = \frac{X+12}{n+24}$: Bayes estimator under Beta(12,12) prior.

$$\begin{aligned}
\mathrm{MSE}_{\delta_3}(\theta) &= \mathrm{E}_\theta\Big(\delta_3(X)-\theta\Big)^2 = \mathrm{Var}_\theta(\delta_3) + \mathrm{Bias}_{\delta_3}(\theta)^2 \\
&= \mathrm{Var}_\theta\left(\frac{X+12}{n+24}\right) + \left(\frac{12-24\theta}{n+24}\right)^2 \\
&= \frac{\mathrm{Var}_\theta(X)}{(n+24)^2} + \frac{144(1-2\theta)^2}{(n+24)^2} \\
&= \frac{n\theta(1-\theta)+144(1-2\theta)^2}{(n+24)^2}
\end{aligned}$$

Note that

$$\delta_3 = \frac{n}{n+24}\delta_1 + \frac{24}{n+24}\delta_2.$$

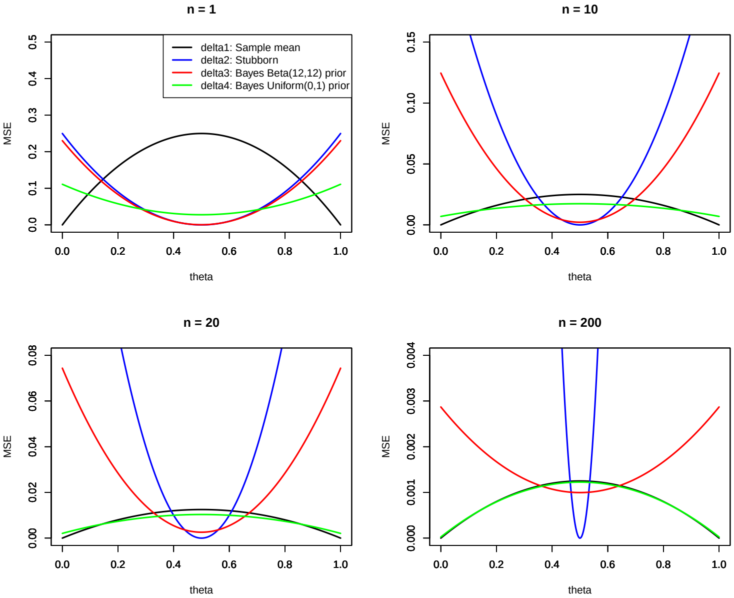
$\delta_4(X) = \frac{X+1}{n+2}$: Bayes estimator under Uniform(0,1) prior.

$$\mathrm{MSE}_{\delta_4}(\theta) = \frac{n\theta(1-\theta) + (1-2\theta)^2}{(n+2)^2} \quad \text{(Exercise.)}$$

Note that

$$\delta_4 = \frac{n}{n+2}\delta_1 + \frac{2}{n+2}\delta_2.$$

Plots of MSE(θ) for the four estimators



- There is no champion over all possible values of θ.
- The range of θ values over which the stubborn estimate does better than the sample mean shrinks as n increases. The more data you have, the more costly it is to ignore them.
- The Bayes estimators are a compromise between sample mean and the stubborn estimator. Depending on the strength of the prior belief, the Bayes estimators will be closer to the sample mean or the stubborn estimator.
- For the binomial experiment, if no prior information is incorporated, then it is "hardest" to estimate θ is when θ is about 1/2.
- Bayes estimators tend to behave better for θ values that is likely according to prior knowledge.
- *Bayesian "shrinkage" is particularly helpful when there is limited amount of data.* In this case, unbiasedness is essentially not useful at all, as the risk is dominated by variance (so-called overfitting).
- *Question:* What happens in high-dimensional problems where the number of parameters are numerous?