Probability Theory and Mathematical Statistics

Jingcheng Liu

Outline

Many conceptual ideas, minimal proofs and derivations

- Estimation theory
  - Comparison between Bayesian and Frequentist approach
  - Confidence interval
- Hypothesis testing
  - Significance and power
  - P-values
- Linear regression

Estimation theory

We saw two estimators for the parameter p given n iid samples from $Bernoulli(p)$:

- MLE:
  - Frequentists approach
  - Inference based on likelihood
  - p is an unknown parameter, we estimate it purely based on data

Parameter: fixed
Data: random

- MAP:
  - Bayesian approach
  - p is unknown, but it follows a prior distribution
  - Inference based on posterior distribution
  - we estimate it based on the observed data and our prior belief

Parameter: random
Data: fixed

- How do we compare different estimators?
  - Bayesian: mean squared error;

Compared with Bayesian MMSE: expectation is taken over prior!

Frequentists risk

Consider n iid samples from $Bernoulli(p)$ with an unknown parameter p:

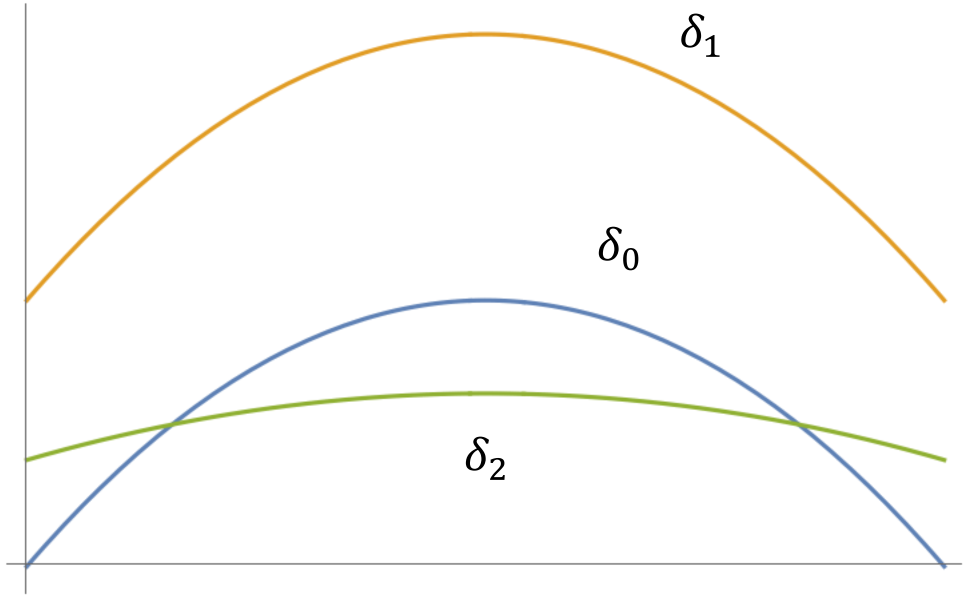
- Loss: $L(p, \delta)$ measures how bad an estimate is
  - $L(p, \delta) = (p - \delta)^2$ is known as the squared loss
- Risk of an estimator:
  - Expected loss, where expectation is taken over the distribution of data

Example

- $\delta_0(X_1, X_2, \dots, X_n) = \sum_i \frac{X_i}{n}$
- $\mathbb{E}\delta_0(X_1, X_2, \dots, X_n) = p,$ so unbiased
- Risk under mean squared loss: $\mathbb{E}(p - \delta_0)^2 = Var(\delta_0) = \frac{p(1-p)}{n}$

Consider two other estimators: $\delta_1 = \frac{1+\sum_i X_i}{n}$, $\delta_2 = \frac{5+\sum_i X_i}{10+n}$

Let's plot their risk functions



Compared with Bayesian MMSE: expectation is taken over prior!

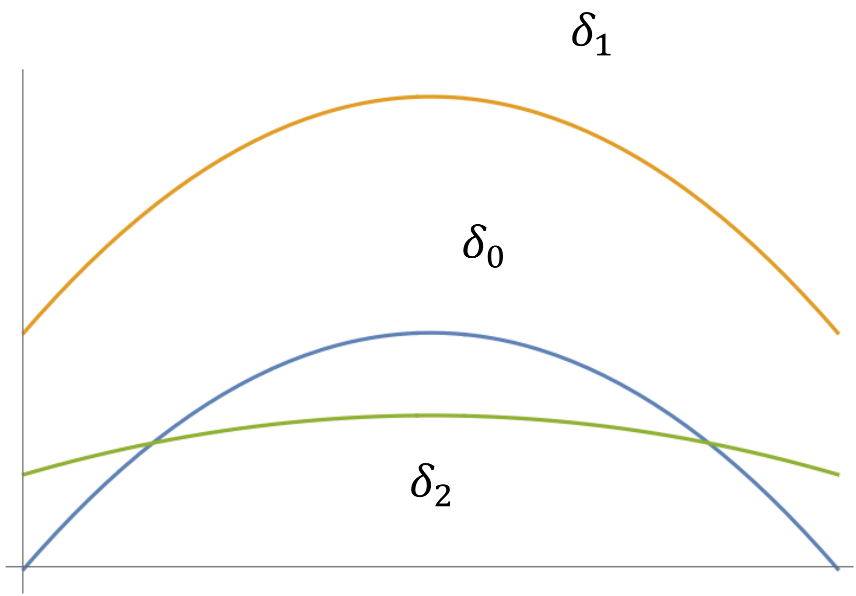
Frequentists risk

Example

- $\delta_0(X_1, X_2, \ldots, X_n) = \sum_i \frac{X_i}{n}$
- $\mathbb{E}\delta_0(X_1, X_2, \ldots, X_n) = p,$ so unbiased
- Risk under mean squared loss: $\mathbb{E}(p - \delta_0)^2 = Var(\delta_0) = \frac{p(1-p)}{n}$

Consider two other estimators: $\delta_1 = \frac{1+\sum_i X_i}{n}$, $\delta_2 = \frac{5+\sum_i X_i}{10+n}$

δ_1 may look stupid. But δ_0 vs δ_2 is trickier…



Rules for choosing THE BEST one:

- Average risk: choose a prior over p $\rightarrow$ Bayesian!
- Worst-case risk: minimax estimator
- Only consider unbiased estimator: (see next)

Sufficient statistics

Suppose $X_1, \dots, X_n \sim Bernoulli(p)$:

Consider $T(X) \coloneqq X_1 + \cdots + X_n \sim Bin(n, p)$

$X_1, \dots, X_n \;\rightarrow\; T(X)$ can throw away information

To estimate p however, $T(X)$ is just as informative as $X_1, \dots, X_n$

$$\Pr[X = x | T = t] = \frac{\Pr[X = x, T = t]}{\Pr[T = t]}$$

Definition. $T(X)$ is a ***sufficient statistic*** for a parameter p, if the distribution of X does not depend on p given T

Sufficient statistics are the only information needed to build an estimator

Parameter → Sufficient statistics → Samples

Minimal sufficiency

There are many sufficient statistics for our toy model:

- $X_1, \dots, X_n$
- $X_{\sigma(1)}, \dots, X_{\sigma(n)}$
- $X_1 + \cdots + X_n$

Definition. $T(X)$ is a ***minimal sufficient statistic*** for a parameter p, if T is sufficient, and any other sufficient statistic $S(X)$, $T(X) = f(S(X))$ for some f

Intuitively, minimal sufficient statistics are the most efficient statistics capturing all the information about the parameter

Roughly speaking, if T determines the likelihood ratio in a "one-to-one fashion", then T is minimal sufficient.

See also: Fisher's factorization theorem.

Sufficiency principle: Rao-Blackwellization

Let $T(X)$ be a sufficient statistic, and $\delta_0(X)$ an estimator.
Consider a new estimator $\delta_1(T(X)) \coloneqq \mathbb{E}[\delta_0(X) \mid T(X)]$

For convex losses, the Rao–Blackwell estimator δ_1 is at least as good as δ_0

In practice, can lead to enormous difference.

See Textbook [BT] page 426 Exercises for examples

Minimum variance unbiased estimator (optional)

Lehmann–Scheffé theorem roughly says that any unbiased estimator through a *complete* and sufficient statistic, is the ***unique*** minimum variance unbiased estimator.

Complete statistic
Roughly, T is complete if there is no non-trivial estimate of 0 through T
Different estimates of T lead to different distributions

See also: Cramér–Rao bound, which gives a bound on how efficient an unbiased estimator can be.

Caution about unbiasedness (optional topic)

Not always a good idea to insist unbiasedness, because Cramér–Rao bound may not be achievable

Example:

Data samples $X \sim Bin(1000, p)$, want to estimate $\Pr[X \geq 500]$.

One can show that the minimum variance unbiased estimator is just $\mathbb{I}[X \geq 500]$

- This means that if $X = 500$, our estimate is 1
- if $X = 499$, our estimate is 0

Confidence interval

How do you interpret the results of an estimation?

- By LLN/CLT, any (asymptotically) unbiased estimator converges to the true parameter as the sample size tends to infinity
- By Chernoff-Hoeffding bound, we also get a finite size bound

Suppose $X_1, \ldots, X_n \sim Bernoulli(p)$ are iid r.v. , and $S_n = \sum_i X_i$ then for any $t > 0$

$$\Pr[|S_n - np| \geq t] \leq 2\mathrm{e}^{-\frac{2\mathrm{t}^2}{\mathrm{n}}}$$

Setting $\alpha = 2\mathrm{e}^{-\frac{2\mathrm{t}^2}{\mathrm{n}}}$, we have $t = \sqrt{\frac{n \ \ln(2/\alpha)}{2}}$.

This means that with probability $1 - \alpha$,

$$p \in \left(\frac{S_n}{n} - \sqrt{\frac{\ln\left(\frac{2}{\alpha}\right)}{2n}}, \quad \frac{S_n}{n} + \sqrt{\frac{\ln(2/\alpha)}{2n}} \right).$$

It is important to note that this probability is **over the distribution of $\boldsymbol{S_n}$**

Confidence interval: interpretations

A 95% confidence interval is NOT an interval that contains the true parameter with probability at least 95%

The confidence interval is a function of the data

After observing the data, the confidence interval is a fixed interval

It either contains the true parameter, or not

To bring back probabilistic interpretation:

- Consider repeating the experiments, over and over again
  - Now you have new, fresh, random data, so that the confidence interval can be treated as a random object over *future repeated experiments*
  - In particle physics, usually a five-sigma rule, unless ground-breaking discovery
- Bayesian approach: credible region
  - Only way to conclude from what we have already observed

Recall Probability vs. Statistics

In probability: Compute probabilities from a parametric model with known parameters

Previous studies found the treatment is 80% effective. Then we expect that for a study of 100 patients, on average 80 will be cured. And the probability that at least 65 will be cured is at least 99.99%.

In statistics: Estimate the probability of parameters given a parametric model and collected data from it

Observe that 78/100 patients were cured. We will be able to conclude that: **if we repeat this experiment**, then we are 95% confident that the number of cured patients are between 69 to 87.

Bayesian vs. frequentist

Bayesian

- Inference based on posterior
- A feature or a bug: Prior
- Probabilities can be interpreted
- Prior is made explicit
- Prior can be subjective
- No canonical prior: can change under re-parameterization
- Hierarchical Bayesian, graphical model
- Computation/sampling of posterior can be hard
  - Frontiers of many research

Frequentist

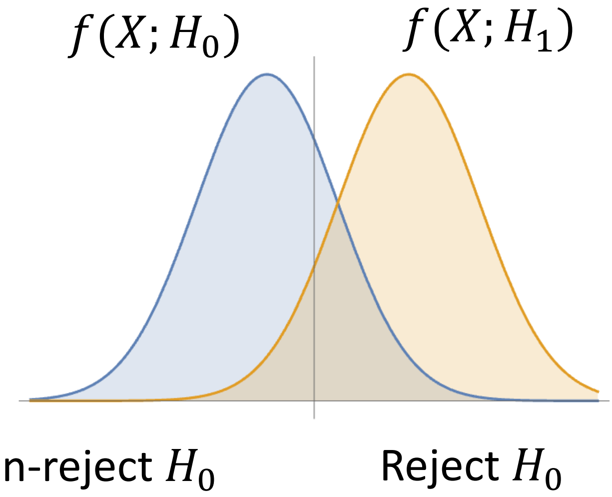
- Inference based on likelihood
- No prior
- Objective – everyone gets the same answer
- Often gets mis-interpreted
- Needs to completely specify an experiment AND the data analysis, before collecting data and actually doing the analysis
- No adaptive re-use of the same dataset
  - There is an entire field for systematically coping with [adaptive data analysis]

Hypothesis testing

Given data X, which of the two (sub)-models generated X ?

Models $P_\theta : \theta \in \Theta$

- Null hypothesis: $H_0 := \{\theta \in \Theta_0\}$
- Alternative hypothesis: $H_1 := \{\theta \in \Theta_1\}$



H_0 is the default/fallback choice

- Fail to reject H_0, no definite conclusion
- Reject H_0 (conclude that H_0 is false, H_1 is true)

If X is a test statistic, the <u>**rejection region**</u> is the set of values to reject H_0 in favor of H_1 if X belongs to it.

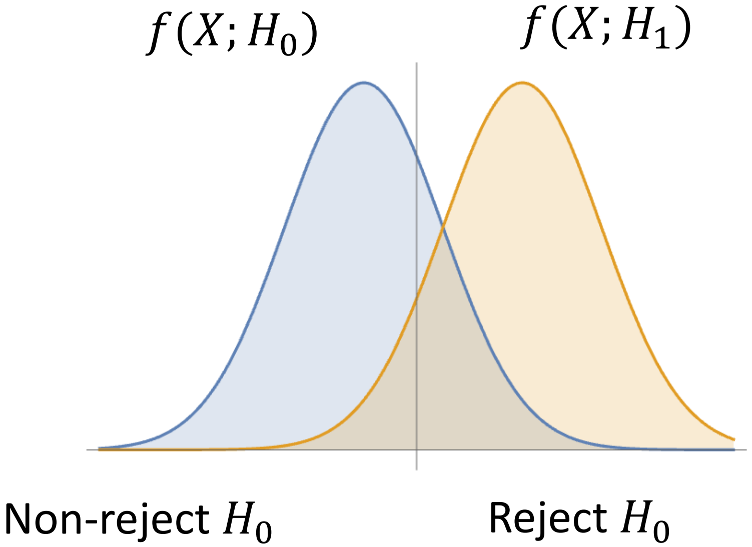
Hypothesis testing

Example: $X_1, \ldots, X_n \sim Bernoulli(\theta)$

Test statistic: the number of heads $S_n = \sum_i X_i$

- Null hypothesis: fair coin $H_0 \coloneqq \{\theta = 0.5\}$
- Alternative hypothesis: biased coin $H_1 \coloneqq \{\theta \neq 0.5\}$

Ideally, would like to choose critical value ξ, so that we reject H_0 whenever $|S_n - 0.5n| > \xi$



Type I, Type II errors

		True answer	
We report		H_0	H_1
	Reject H_0	Type I error	Correct
	Don't reject H_0	Correct	Type II error

Significance and power

- Significance level = $\Pr[\text{type I error}] = \Pr[\text{false positive}]$
  = probability of incorrectly rejecting H_0

- Power = probability of correctly rejecting H_0
  = $1 - \Pr[\text{type II error}]$

Ideally, want significance level near 0 and power near 1

“P-hacking”

P-values

Instead of choosing significance level and power, one often simply reports a single p-value

Say x is a test statistic

- Right sided p-value: $\Pr[X > x; H_0]$
- Two sided: $\Pr[|X| > x; H_0]$

$\Pr[x;\ H_0]$ vs $\Pr[x|H_0]$

Interpretations: how likely are your data (or something more extreme) under null hypothesis?

Likelihood ratio test (LRT)

More generally, this is a common test known as ***likelihood ratio test***

- $L(x) := \frac{\Pr[x; H_1]}{\Pr[x; H_0]}$
- If $L(x) > \xi$, then reject H_0

See also **Neyman-Pearson Lemma**: for any fixed level of $\Pr[\text{type I error}]$ that can be achieved by an LRT, there is an LRT that achieves the smallest $\Pr[\text{type II error}]$ among all (randomized) predictors.

* One way to prove this lemma is to use the *Lagrange multiplier method*

* A key insight is that for any LRT, we can find a loss function for which it is optimal

Linear regression

Why least squares make sense in linear regression

- Assume independent Gaussian noise are added to the data

$$y_i = \beta_0 + \beta_1 x_i + N(0,1)$$

- Given data $\{(x_i, y_i)\}_{i=1}^n$
- Want to find MLE estimate for (β_0, β_1)

This gives precisely the formula of minimizing $\sum_i (y_i - \beta_0 - \beta_1 x_i)^2$