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
  - NHST
  - Significance and power
  - P-values
  - Neyman-Pearson’s optimal test

Recall: 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;

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*** of the assumed statistical/generative model
  - 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.

Note: we are repeating an idealized statistical experiment

Bayesian vs. frequentist

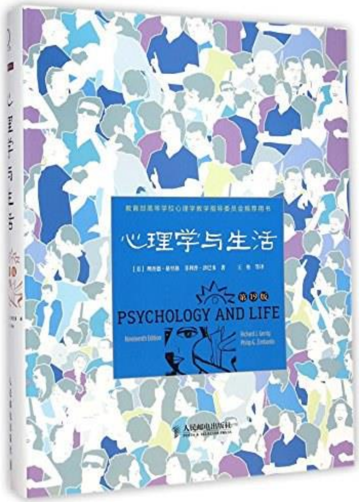
Bayesian

- 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

Null hypothesis significance testing (NHST)



Considered as the “backbone of psychological research”

One might think hypothesis testing “should” work like this:

- Say you want to know if a treatment is effective
- You perform a randomized controlled experiment, with or without the treatment
- Look at the collected data
- Decide if they provide convincing evidence for or against the hypothesis

In other words: estimate the likelihood that “the treatment is effective”, given the data and all the context (e.g., experimental setup)

Null hypothesis significance testing (NHST)

Instead, this is how NHST actually works:

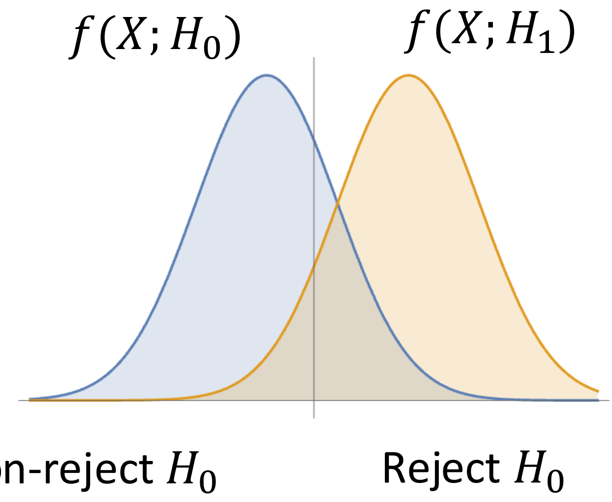
- Say you want to know if a treatment is effective
- Create a negated hypothesis, called <u>**null hypothesis**</u>: “the treatment is not effective” (AKA nil hypothesis)
- We must assume the null hypothesis is true.
- Then look at the data, and decide how likely is it to see the data under the null hypothesis
- If the data are sufficiently unlikely under null hypothesis
    - Reject the null in favor of the <u>**alternative hypothesis**</u> “the treatment is effective”
- Otherwise, there is insufficient evidence
    - Retain (or “fail to reject”) the null hypothesis, falling back to the default assumption

Hypothesis testing

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

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

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



H_0 is the default/fallback choice

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

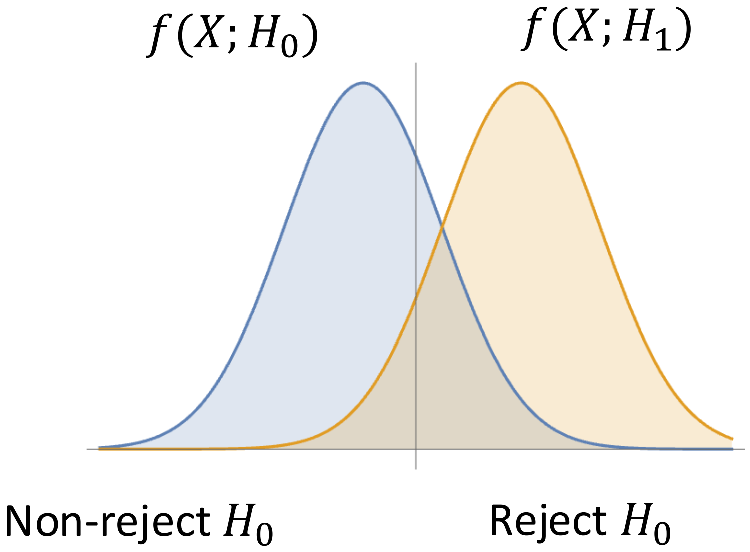
If X is a ***test statistic***, the **rejection region** is the set of values to reject H_0 in favor of H_1 if X belongs to it.

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?

Mis-interpretations of P-values

Say you find a test statistic with a p-value 0.01

Which, if any, of the following statements are true?

1. You have absolutely disproved the null hypothesis
2. You have absolutely proved the alternative hypothesis
3. You have found the probability of the null hypothesis being true
4. You can deduce the probability of the alternative hypothesis being true
5. You now know, if you decide to reject the null hypothesis, the probability that you are making the wrong decision
6. You have a reliable experimental finding in the sense that if, hypothetically, the experiment were repeated a great many times, you would obtain a significant result on 99% of occasions.

Mis-use of NHST

So far we talked about one test.

What if we use computers to search for “significance discovery”?

- In Genome-wide association study (GWAS), there are millions of locations of genomes.
- Brain imaging collects many locations in the brain at once (~ 10^5)

Imagine we simply perform a hypothesis testing at each individual location, and report back if the test is significant at $p < 0.05$

What’s wrong with this approach?

A simple fix is known as **Bonferroni correction**, essentially a union bound.

Recap: Null hypothesis significance testing

Instead, this is how NHST actually works:

- Formulate a hypothesis that embodies our prediction (*before seeing the data*)
- Say you want to know if a treatment is effective
- Specify null and alternative hypotheses
- “the treatment is not effective” vs. “the treatment is effective”
- Collect some data relevant to the hypothesis
- Fit a model to the data and compute a test statistic that quantifies the amount of evidence for or against the null hypothesis
- If the data are sufficiently unlikely under null hypothesis
  - Reject the null in favor of the **alternative hypothesis** “the treatment is effective”
- Otherwise, there is insufficient evidence
  - Retain (or “fail to reject”) the null hypothesis, falling back to the default assumption

Hypothesis testing as decision making

- Instead of inferring the significance of one hypothesis test
- Neyman and Pearson suggest that we should think of hypothesis testing as "repeated" decision making
  - Minimize the error rate in the long run
  - In other words, we don't know which of our decisions are right or wrong
  - If we follow the same rule, we can still know how often our decisions are right or wrong
- Trade-off between $\Pr[\text{type I error}]$ and $\Pr[\text{type II error}]$
  - Always reject: $\Pr[\text{type I error}] = 1$ but $\Pr[\text{type II error}] = 0$
  - Always retain: $\Pr[\text{type I error}] = 0$ but $\Pr[\text{type II error}] = 1$

- For further readings on Fisher's take vs. Neyman-Pearson's take on hypothesis testing, see Section 3 of [Mindless Statistics, by Gerd Gigerenzer]

Hypothesis testing as decision making

To compare different decision making, one considers expected loss

- Say we make a decision/prediction of $Y \in \{0,1\}$ based on observing X
- The $loss(\hat{Y}, Y)$ is a loss function for predicting $\hat{Y}$ while the truth is Y

Example: $loss(\hat{Y}, Y) = 1[\hat{Y} \neq Y]$

The expected loss is also known as the ***risk*** of a predictor

$$\mathbb{E}_{X,Y}\left[loss\left(\hat{Y}(X), Y\right)\right]$$

Lemma. The optimal decision rule minimizing the expected loss is given by:

$$\hat{Y}(x) = 1\left[\frac{\Pr[Y=1|X=x]}{\Pr[Y=0|X=x]} \geq \frac{loss(1,0) - loss(0,0)}{loss(0,1) - loss(1,1)}\right]$$

Assuming Bayesian inference

Hypothesis testing as decision making

Lemma. The optimal decision rule minimizing the expected loss is given by:

$$\hat{Y}(x) = 1\left[\frac{\Pr[Y = 1|X = x]}{\Pr[Y = 0|X = x]} \geq \frac{loss(1,0) - loss(0,0)}{loss(0,1) - loss(1,1)}\right]$$

Proof. $\mathbb{E}_{X,Y}\left[loss\left(\hat{Y}(X), Y\right)\right] = \mathbb{E}_X\left[\mathbb{E}_{Y|X}\left[loss\left(\hat{Y}(X), Y\right)|X\right]\right]$

For any fixed value of x

$$\mathbb{E}_{Y|X}[loss(0, Y)|X = x] = loss(0,0)\Pr[Y = 0|X = x] + loss(0,1)\Pr[Y = 1|X = x]$$
$$\mathbb{E}_{Y|X}[loss(1, Y)|X = x] = loss(1,0)\Pr[Y = 0|X = x] + loss(1,1)\Pr[Y = 1|X = x]$$

So, the optimal decision rule is to predict $\hat{Y}(x) = 0$ if the first is smaller
and predict $\hat{Y}(x) = 1$ if the second is smaller

Rearranging these inequalities gives the optimal decision rule

Assuming Bayesian inference

Likelihood ratio test (LRT)

Lemma. The optimal decision rule minimizing the expected loss is given by:

$$\hat{Y}(x) = 1\left[\frac{\Pr[Y=1|X=x]}{\Pr[Y=0|X=x]} \geq \frac{loss(1,0) - loss(0,0)}{loss(0,1) - loss(1,1)}\right]$$

Note that $\Pr[Y = \cdot\, |X = x]$ is the posterior probability

Let $p_0 = \Pr[Y = 0]$ and $p_1 = \Pr[Y = 1]$ be the prior probability

Then the (Bayesian) optimal decision rule is equivalent to a ***likelihood ratio test***:

$$\hat{Y}(x) = 1\left[\frac{\Pr[X=x|Y=1]}{\Pr[X=x|Y=0]} \geq \frac{p_0}{p_1} \cdot \frac{loss(1,0) - loss(0,0)}{loss(0,1) - loss(1,1)}\right]$$

where $\mathfrak{L}(x) = \frac{\Pr[X=x|Y=1]}{\Pr[X=x|Y=0]}$ is known as the ***likelihood ratio***

and $\hat{Y}(x) = 1[\mathfrak{L}(x) \geq \eta]$ of this form is known as ***likelihood ratio test***

Maximum a posteriori as LRT

Recall the MAP in Bayesian inference

$$\hat{Y}(x) = \arg\max_{y \in \{0,1\}} \Pr[Y = y | X = x]$$

By setting $loss(1,0) = loss(0,1) = 1$, and $loss(0,0) = loss(1,1) = 0$,

$$\hat{Y}(x) = 1\left[\frac{\Pr[Y = 1 | X = x]}{\Pr[Y = 0 | X = x]} \geq \frac{loss(1,0) - loss(0,0)}{loss(0,1) - loss(1,1)}\right]$$

simplifies to

$$\hat{Y}(x) = 1\left[\frac{\Pr[Y = 1 | X = x]}{\Pr[Y = 0 | X = x]} \geq 1\right] = \arg\max_{y \in \{0,1\}} \Pr[Y = y | X = x]$$

Maximum likelihood as LRT

Recall the MLE in Frequentist inference

$$\hat{Y}(x) = \arg\max_{y \in \{0,1\}} \Pr[X = x | Y = y]$$

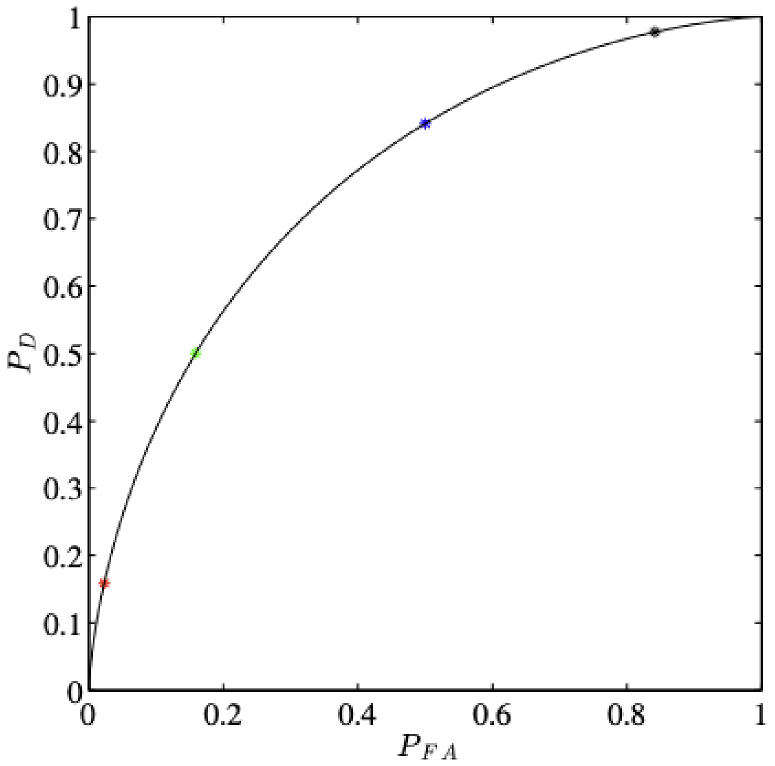
Let $loss(1,0) = loss(0,1) = 1$, $loss(0,0) = loss(1,1) = 0$, and $p_0 = p_1$,

$$\hat{Y}(x) = 1\left[\frac{\Pr[X = x | Y = 1]}{\Pr[X = x | Y = 0]} \geq \frac{p_0}{p_1} \cdot \frac{loss(1,0) - loss(0,0)}{loss(0,1) - loss(1,1)}\right]$$

simplifies to

$$\hat{Y}(x) = 1\left[\frac{\Pr[X = x | Y = 1]}{\Pr[X = x | Y = 0]} \geq 1\right] = \arg\max_{y \in \{0,1\}} \Pr[X = x | Y = y]$$

Receiver operating characteristic (ROC) curves



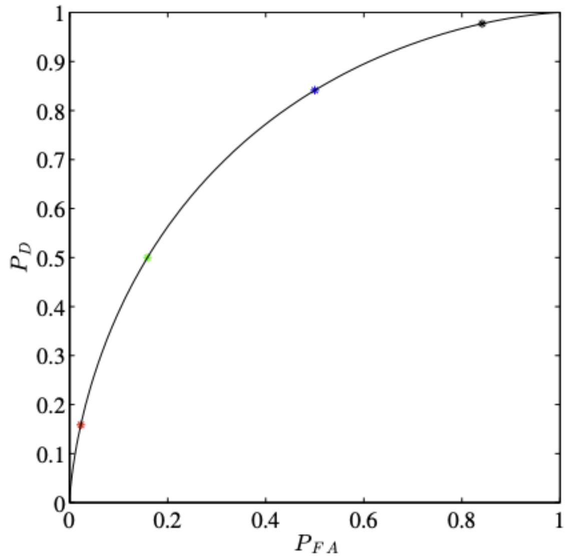
The probability of a false-positive is also called the probability of false-alarm, denoted by

$$P_{FA} = \Pr\left[\hat{Y}(x) = 1 | Y = 0\right]$$

The probability of detection (1− probability of a false-negative) is denoted by

$$P_D = \Pr\left[\hat{Y}(x) = 1 | Y = 1\right]$$

The most powerful predictor has concave ROC



Given two predictors ϕ_1, ϕ_2, consider their "convex combination"
by selecting ϕ_1 with prob. p
and ϕ_2 with prob. $1 - p$

What is the expected power and significance of the combined predictor?

Optimality of Likelihood ratio test (LRT)

<u>**Neyman-Pearson Lemma**</u> for simple hypotheses: for any fixed level of $P_{FA} = \Pr[\text{type I error}]$ that can be achieved by an LRT, there is an LRT that achieves the smallest $1 - P_D = \Pr[\text{type II error}]$ among all (randomized) predictors.

- One way to prove this lemma is to use the <u>*Lagrange multiplier method*</u>
- A key insight is that for any LRT, we can find a loss function for which it is optimal

Put differently, LRT gives the optimal ROC curve by varying threshold

Proof of Neyman-Pearson (Assume the likelihoods are continuous functions)

For any fixed level of $P_{FA} = \alpha$

Let η be the threshold where the LRT $Q_\eta \coloneqq 1[\mathfrak{L}(x) \geq \eta]$ achieves $P_{FA} = \alpha$. Let $\beta = P_D$

Consider the loss function

$$loss(1,0) = \frac{\eta p_1}{p_0}\ , \qquad loss(0,1) = 1, \qquad loss(0,0) = loss(1,1) = 0$$

Note that the expected loss of any predictor Q is given by

$$\begin{aligned}\mathbb{E}_{X,Y}[loss(Q(X),Y)] &= p_0 P_{FA}(Q) loss(1,0) + p_1\big(1 - P_D(Q)\big) loss(0,1) \\ &= p_1 \eta P_{FA}(Q) + p_1\big(1 - P_D(Q)\big)\end{aligned}$$

And Q_η minimizes the expected loss among them (verify!), so we have

$$p_1 \eta \alpha + p_1(1 - \beta) \leq p_1 \eta P_{FA}(Q) + p_1\big(1 - P_D(Q)\big)$$

Consider any other predictor Q with $P_{FA}(Q) \leq \alpha$, this implies $P_D(Q) \leq \beta$.

Put differently, at any point P_{FA}, there's an LRT giving the optimal P_D

Statistical limits in binary hypothesis testing

Le Cam's inequality

Let P, Q be distributions defined on Ω, then

$$\inf_{T:\Omega\to\{0,1\}} P(T(X) \neq 0) + Q(T(X) \neq 1) = 1 - \frac{\|P - Q\|_1}{2},$$

where the infimum is taken over all predictors $T: \Omega \to \{0,1\}$

Proof. Any (deterministic) predictor has an **acceptance** region, say $A \subseteq \Omega$ where it outputs 0, and a **rejection** region A^c where it outputs 1

$$P(T(X) \neq 0) + Q(T(X) \neq 1) = P(A^c) + Q(A) = 1 - (P(A) - Q(A))$$

Optimizing over all predictor is the same as optimizing over A, and

$$\sup_{A\subseteq\Omega} P(A) - Q(A) = \frac{\|P - Q\|_1}{2}$$

Statistical limits in binary hypothesis testing

How many iid samples do we need to reliably distinguish between

- $\mathrm{Bernoulli}(p)$ for $p > 1/2$
- $\mathrm{Bernoulli}(q)$ for $q < 1/2$

Probability amplification, or error reduction in randomized algorithm:

we repeat an algorithm n independent rounds, and take majority

Concentration inequality tells us that roughly $n = \mathrm{O}\left(\frac{1}{(p-q)^2}\right)$ suffices

Is this also necessary?

To apply Le Cam, let $P = \mathrm{Binomial}(n, p)$, $Q = \mathrm{Binomial}(n, q)$

How do you control $\|P - Q\|_1$ for product distributions?

Statistical limits in binary hypothesis testing*

How do you control $\|P-Q\|_1$ for product distributions?

Idea: relate to another distance that "tensorizes"

Popular choice: KL-divergence and Hellinger distance

$$D_{KL}(P||Q) = \sum_{x\in\Omega} P(x) \ln \frac{P(x)}{Q(x)}$$

$$D_{KL}(\mathrm{Binomial}(n,p)||\mathrm{Binomial}(n,q)) = n\, D_{KL}(\mathrm{Bernoulli}(p)||\mathrm{Bernoulli}(q))$$
$$= n\left(p \ln\frac{p}{q} + (1-p)\ln\frac{1-p}{1-q}\right) \approx \frac{n(p-q)^2}{2p(1-p)}$$

Pinsker's inequality:

$$\frac{1}{2}\|P-Q\|_1 \le \sqrt{\frac{1}{2} D_{KL}(P||Q)}$$

Combined, to get a small probability of error, one also needs $n = \Omega\left(\frac{1}{(p-q)^2}\right)$

Bretagnolle–Huber inequality: $\frac{1}{2}\|P-Q\|_1 \le \sqrt{1-\exp(-D_{KL}(P||Q))}$

Bonus material: 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 $\sum_i (y_i - \beta_0 - \beta_1 x_i)^2$

Quick Recap

Basic probabilistic models: for example,

- Balls into bins, Monty Hall, coin flipping, card drawing, dice rolling, Buffon's needle, coupon collector
- sampling with or without replacement
- Erdos-Renyi random graph

Basic notions: for example,

- Probability measures, sigma algebra
- Independence, conditional independence, correlation
- Moments, mean, median, variance, covariance, expectation
- Binomial, multinomial, Poisson, Gaussian, exponential, geometric
- Convergence and limit theorems

Basic techniques: for example,

- Inclusion-Exclusion, Union bound
- probabilistic method
- linearity of expectation
- Chernoff bound, Martingale and bounded difference, optional stopping