Advanced mathematical statistics August 30, 2025

Part II. Information Theoretic Aspects of Statistics

Instructor: Lin Liu

1 Key Concepts and Philosophies in Statistics

Statistics can be roughly categorized into the following paradigms:

1. Start from a concrete scientific question abstracted as a parameter θ of some statistical model $\{\mathbb{P}_\theta, \theta \in \Theta\}$, collect relevant data $\mathcal{D}$, and try to study the following aspects of θ:

   - hypothesis testing on θ: e.g. null hypothesis $H_0 : \theta = 0$ or $H_0 : \theta \in \Theta_0$
     find a measurable function of the data and a nominal level α (maximum type-I error allowed), $T_\alpha(\mathcal{D}) \in \{0, 1\}$ (a nominal level α test statistic for H_0), such that

     $$\sup_{\theta \in H_0} \mathbb{P}_\theta(T_\alpha(\mathcal{D}) = 1) \leq \alpha \tag{1}$$

     If (1) is satisfied, then the nominal level α test statistic $T_\alpha(\mathcal{D})$ is said to be (uniformly) valid. If we find $T_\alpha(\mathcal{D}) = 1$ for the given data $\mathcal{D}$, then we say we reject H_0 at level α.

     If one also has an alternative hypothesis $H_a : \theta \in \Theta_a$, then one at least hopes that under the constraint that (1) is satisfied, the following also holds for some $\beta \in [0, 1]$ (best if $\beta = 1$)

     $$\inf_{\theta \in H_a} \mathbb{P}_\theta(T_\alpha(\mathcal{D}) = 1) \geq \beta. \tag{2}$$

     When $\beta = 1$, then a valid test $T_\alpha(\mathcal{D})$ is said to be (uniformly) powerful; when $\beta = 0$, then a valid test $T_\alpha(\mathcal{D})$ is said to be powerless (in worst case); when $\beta \in (0, 1)$, then a valid test $T_\alpha(\mathcal{D})$ is said to have non-trivial power. If the power of a valid test $T_\alpha(\mathcal{D})$ cannot be improved, then we say $T_\alpha(\mathcal{D})$ is a uniformly optimal valid test.

   - estimation of θ: learn the value of θ from data
     find a measurable function of the data $T(\mathcal{D})$ (estimator) such that $\|T(\mathcal{D}) - \theta\|$ is close to 0. If we index $T(\mathcal{D})$ as a sequence of estimators $T_n(\mathcal{D})$ where n could be the sample size of $\mathcal{D}$, then if $\|T_n(\mathcal{D}) - \theta\| \to 0$ in $\mathbb{P}_\theta$-probability as $n \to \infty$, then $T_n(\mathcal{D})$ is said to be a consistent estimator of θ. If the following stronger condition holds: for every $\epsilon > 0$,

     $$\sup_{\theta \in \Theta} \mathbb{P}_\theta(\|T_n(\mathcal{D}) - \theta\| > \epsilon) \to 0 \text{ as } n \to \infty \tag{3}$$

     then $T_n(\mathcal{D})$ is said to be a uniformly consistent estimator of θ over the class Θ. Consistency does not have practical relevance but uniform consistency does, which is one of the common theme that we will emphasize throughout this semester.

   - uncertainty quantification (aka statistical inference) of θ: learn a spectrum of values of θ that incorporates the uncertainty in the data (e.g. noise, measurement error, systematic bias, contamination, adversarial attack, or even intrinsic quantum stochasticity)

find a set of measurable functions of the data and a nominal confidence level α, $\{T(\mathcal{D})\}_\alpha$ (nominal $1-\alpha$ confidence set), such that

$$\mathbb{P}_\theta\left[\theta \in \{T(\mathcal{D})\}_\alpha\right] \geq 1-\alpha \tag{4}$$

and the Lebesgue measure or any appropriate measure of the volume of $\{T(\mathcal{D})\}_\alpha$ is as small as possible. Any nominal $1-\alpha$ confidence set $\{T(\mathcal{D})\}_\alpha$ satisfying (4) is said to be a (point-wise) valid nominal $1-\alpha$ confidence set. If

$$\inf_{\theta\in\Theta} \mathbb{P}_\theta\left[\theta \in \{T(\mathcal{D})\}_\alpha\right] \geq 1-\alpha \tag{5}$$

then $\{T(\mathcal{D})\}_\alpha$ is said to be an honest (uniformly valid) nominal $1-\alpha$ confidence set of θ over the class Θ.

Folklore: hypothesis testing $\prec$ estimation $\prec$ statistical inference, where the partial ordering $\prec$ means increasing level of difficulty. This is a philosophy for viewing statistical problems, not an actual mathematical theorem.

2. The above paradigm works well in classical scientific inquiry, but not so much in modern big data era, when the scientific question of interest can be quite nebulous, or even worse, does not even exist before the data is collected. In this case, the above problems are still of interest, but the requirements are much more difficult to satisfy, as one needs to take the data-exploration step into account. Popular statistical problems include large-scale multiple hypothesis testings, statistical inference validity after model selection/data dredging. For simplifying our terminology's sake, we categorize this type of problems as "model selection", which is orthogonal to, but at the same time related to, hypothesis testing, estimation, and statistical inference.

Before large-scale computing becomes routine, mathematical theorems and the philosophy behind the theorems are the major topics statisticians care about. We will also mention them for pedagogical purposes.

1.1 Bayesian vs. Frequentists [vs. Fiducialists] (BFFs)

Recall from above, we define honest uncertainty quantification/confidence sets as the long-run behavior of a statistical procedure. Related to frequentists' confidence sets is fiducial inference, which was one of RA Fisher's major contribution and gradually got lost in history because of its numerous issues. For more detail, you can look up in [**?**]. But there is a recent surge of interest in fiducial inference due to the assumption-free pursuit in statistics.

In contrast, Bayesian inference treats θ as random and starts with a prior probability measure $\mathrm{d}\Pi(\theta)$ on θ. Conditioning on $\theta = \theta'$, the data $\mathcal{D}$ is a random drawn from a probabilistic model $\mathrm{d}\mathbb{P}(\mathcal{D}|\theta)$. Then we "update" our belief on θ after seeing the data by Bayes' rule:

$$\Pi(\theta \in B|\mathcal{D}) = \frac{\int_B \mathrm{d}\mathbb{P}(\mathcal{D}|\theta)\mathrm{d}\Pi(\theta)}{\int_\Theta \mathrm{d}\mathbb{P}(\mathcal{D}|\theta)\mathrm{d}\Pi(\theta)}$$

where $B \in \mathcal{B}$ and $\mathcal{B}$ is a σ-field on Θ.

When we attach sample size n to the prior and the likelihood (and the data), in Bayesian inference, consistency "becomes" posterior consistency and confidence sets "becomes" credible sets.

Definition 1 (Posterior consistency)**.** The posterior distribution $\Pi_n(\theta \in B|\mathcal{D}_n)$ is said to be consistent at $\theta_0 \in \Theta$ if $\Pi_n(\theta \in N^c|\mathcal{D}_n) \to 0$ in $\mathbb{P}_{\theta_0}^{(n)}$-probability, as $n \to \infty$, for every neighborhood N of θ_0. Uniform posterior consistency is then obvious to define.

Definition 2 (Bayesian credible sets)**.** A nominal $1-\alpha$ Bayesian credible set $C_\alpha(\mathcal{D}_n)$ for $\theta \in \Theta$ is a subset of Θ such that

$$\Pi_n(\theta \in C_\alpha(\mathcal{D}_n)|\mathcal{D}_n) \geq 1-\alpha. \tag{6}$$

But as frequentists, we would like to establish the frequentist property of a Bayesian credible set, which we termed as honest Bayesian credible set (Bayesians generally do not like the name)

$$\inf_{\theta\in\Theta} \Pi_n(\theta \in C_\alpha(\mathcal{D}_n)) \geq 1-\alpha. \tag{7}$$

We may come back to this in the end of this semester.

Finally, there is an extremely small group of statisticians named Dempster-Shafer (Arthur Dempster and Glenn Shafer) school. They believe in the "belief function" and "plausibility function" and operate with Dempster-Shafer calculus [**?**], arguably a more systematic way of evaluating the strength of evidence from data.

Definition 3 (Belief function and Plausibility function [**?**])**.** For any subset $A \subset \Omega$ of a probability space $(\Omega, \mathcal{F}, \mathbb{P})$, where $\mathbb{P} \in \mathcal{P}$

$$\begin{aligned} \text{Bel}(A) &= \inf_{\mathbb{P}\in\mathcal{P}} \mathbb{P}(A), \\ \text{Pl}(A) &= \sup_{\mathbb{P}\in\mathcal{P}} \mathbb{P}(A). \end{aligned} \tag{8}$$

Belief function and Choquet capacity [**?**] are sometimes used interchangeably. Interestingly, Dempster is one of the most famous statisticians in history because of his work on EM algorithm with Donald B. Rubin and Nan Laird. But the belief function, though his signature invention, is not well received by his fellow statisticians in general. It has, however, received a huge attention in artificial intelligence (the old school AI like expert systems, not today's AI based exclusively on large-scale optimization and deep neural networks). Few statistics departments in the world cover Dempster-Shafer calculus in any class nowadays.

In terms of practical relevance, confidence sets from frequentists and credible sets from Bayesian are still the most widely used uncertainty quantification methods.

Remark 4. Debates among BFFs still go on till today. There is a BFF conference held every year since 2014. Chinese statisticians in general do not show much interest in studying statistical philosophy. One of the few exceptions is Xiao-Li Meng (and his pupils), who is famous for writing statistics research papers in an artistic style: e.g. [**? ? ? ?**].

2 Statistical estimation: Decision- and information-theoretic perspectives

Since most of you are familiar with sufficient statistic, I only provide a brief summary of the key points on sufficiency, completeness, and complete sufficiency.

2.1 Sufficient statistic vs. Information theory

Definition 5 (Sufficient statistic (SS), minimal sufficient statistic (MSS))**.** A statistic $T(X)$ is said to be sufficient for θ is the law of $X|T(X)$ does not depend on θ. A minimal sufficient statistic $T^*(X)$ is a sufficient statistic that can be written as a function of any other sufficient statistic.

You are expected to be familiar with the following theorems on sufficient statistic: First, Fisher-Neyman factorization is a procedure of finding sufficient statistics:

Theorem 6 (Neyman-Fisher factorization)**.** *$T(X)$ sufficient for $\theta \Leftrightarrow f_X(x;\theta) = g(T(x);\theta)h(x)$ for some functions $g(\cdot,\theta)$ depending on X only through $T(X)$ and h free of θ.*

The proof of the above theorem can be found in standard statistical textbooks, hence omitted. Sufficient condition for an MSS

Theorem 7. *$T(X)$ is sufficient for θ. T is an MSS if the likelihood ratio $\frac{f_X(x;\theta)}{f_X(x';\theta)}$ is free of θ for any $x' \neq x$ implies $T(x) = T(x')$.*

Proof. Let T' be any other SS. If $T'(x) = T'(x') \Rightarrow T(x) = T(x')$ for any $x \neq x'$, then $T = g(T')$. We are done. So we need to show $T'(x) = T'(x') \Rightarrow T(x) = T(x')$. Take $x \neq x'$, but $T'(x) = T'(x')$, then the likelihood ratio

$$\frac{f_X(x;\theta)}{f_X(x';\theta)} = \frac{g(T'(x);\theta)h(x)}{g(T'(x');\theta)h(x')} = \frac{h(x)}{h(x')}$$

is free of θ. By the assumption in the theorem, this implies $T(x) = T(x')$, which means we establish the logical chain:

$$T'(x) = T'(x') \Rightarrow T(x) = T(x').$$

□

Sufficiency is closely related to information theoretic aspect of statistics. First though, we recall the following famous Data Processing Inequality (DPI)

Lemma 8 (DPI)**.** *For data processing depicted in Figure 1, we have*

$$I(\theta;T(X)) \leq I(\theta;X)$$

that is, data processing diminishes the information on the source θ. Here $I(X;Y)$ is the mutual information (MI) between two random variables X and Y, defined as

$$I(X;Y) = \mathsf{H}(X) - \mathsf{H}(X|Y) \tag{9}$$

where $\mathsf{H}(X) = -\int \log \mathrm{d}\mathbb{P}(x)\mathrm{d}\mathbb{P}(x)$ is the Shannon entropy of X, and $\mathsf{H}(X|Y)$ is the conditional Shannon entropy of $X|Y$, which is a deterministic quantity and defined as

$$\mathsf{H}(X|Y) = \int \mathsf{H}(X|Y=y)\mathrm{d}\mathbb{P}(y). \tag{10}$$

Note that $\mathsf{H}(X,Y) = \mathsf{H}(X) + \mathsf{H}(Y|X) = \mathsf{H}(Y) + \mathsf{H}(X|Y)$. *Similarly one can define the conditional MI between* X *and* Y *conditional on* Z *as*

$$I(X;Y|Z) = \mathsf{H}(X|Z) - \mathsf{H}(X|(Y,Z)) \tag{11}$$

If $X \perp\!\!\!\perp Y$ *then* $I(X;Y) = 0$*; if* $X \perp\!\!\!\perp Y|Z$ *then* $I(X;Y|Z) = 0$.

Proof.

$$\begin{aligned} I(\theta;(X,T(X))) &= \mathsf{H}(\theta) - \mathsf{H}(\theta|(X,T(X))) \\ &= \mathsf{H}(\theta) - \mathsf{H}(\theta|T(X)) + \mathsf{H}(\theta|T(X)) - \mathsf{H}(\theta|(X,T(X))) \\ &= I(\theta;T(X)) + I(\theta;X|T(X)) \\ \text{By symmetry } &= I(\theta;X) + I(\theta;T(X)|X). \end{aligned}$$

By Figure 1, we immediately have $I(\theta;T(X)|X) = 0$ so

$$I(\theta;X) = I(\theta;T(X)) + I(\theta;X|T(X)) \geq I(\theta;T(X))$$

where in the last inequality we use the fact that mutual information is non-negative, which can be proved by Jensen's inequality. □

Data processing gives us the following Markov chain:

Figure 1: Markov chain of data processing

If T happens to be a sufficient statistic of θ, then a part of the above Markov chain can be reversed:

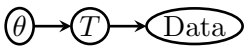

Figure 2: Markov chain under sufficiency

By DPI, we have, when $T(X)$ is a sufficient statistic, $I(\theta;X) = I(\theta;T(X))$. Minimal sufficient statistic T^* is much clearer to us if we define it in information-theoretic parlance:

$$T^* \coloneqq \arg\min_{\widetilde{T}} I(X;\widetilde{T}(X)) \text{ s.t. } I(\theta;X) = I(\theta;\widetilde{T}(X)).$$

In words, a minimal sufficient statistic is the sufficient statistic that is the "furthest" from the data X.

Definition 9. $T(X)$ is said to be a complete statistic for a family distribution indexed by the parameter θ if it is impossible to construct a non-trivial unbiased estimator of 0 from $T(X)$, i.e. $\mathbb{E}_\theta(h(T(X))) = 0 \quad \forall\, \theta \implies h(T) = 0$ w.p.1.

Theorem 10. *Any CSS is also minimal as long as* $\exists$ *at least one MSS.*

Proof. Let T be a CSS, M be a MSS and $h(T) = \mathbb{E}_\theta(T|M) - T$. We can show (a) $h(T)$ is free of θ; (b) $h(T)$ is a function of T; and (c) $h(T)$ is an unbiased estimator of 0.

From (a) (b) (c) and the definition of CSS, $h(T) = 0$ w.p.1 i.e. $\mathbb{E}_\theta(T|M) = T$ w.p.1 $\implies$ T is a function of M on a set of measure 1 $\implies$ T is a function of every other SS $\implies$ T is a MSS $\square$

For exponential family distributions, $f_X(x;\eta) = \exp\{\eta T(x) - \psi(\eta)\}h(x)$ then $T(X)$ is CSS for η. CSS is useful for constructing not-so-bad estimators.

Theorem 11 (Rao-Blackwell)**.** *$\widehat{T}$ unbiased for $T(\theta)$. S is any SS. Take $\phi(S) = \mathbb{E}_\theta(\widehat{T}|S)$. Then (1) $\phi(S)$ is also unbiased for $T(\theta)$ and (2) $\mathsf{var}_\theta(\phi(S)) \le \mathsf{var}_\theta(\widehat{T})$ for any θ.*

The proof is trivial. But I want to remark that since S is SS, conditioning on S the law of $\widehat{T}$ does not depend on θ. Thus $\phi(S)$ is a statistic (i.e. a function of the data).

With Rao-Blackwell, one can show

Theorem 12 (Lehmann-Scheffé)**.** *An unbiased estimator of $T(\theta)$ that is a function of a CSS is the UMVUE of $T(\theta)$ (uniformly minimum variance unbiased estimator).*

Proof. Take T, T' two unbiased estimators of $T(\theta)$. S is CSS for θ. $\phi(S) = \mathbb{E}_\theta(T|S)$ $\phi'(S) = \mathbb{E}_\theta(T'|S)$. Rao-Blackwell: $\mathsf{var}_\theta(\phi(S)) \le var_\theta(T)$ $\mathsf{var}_\theta(\phi'(S)) \le \mathsf{var}_\theta(T')$ $h(S) = \phi(S) - \phi'(S)$. Then $\mathbb{E}_\theta(h(S)) = \mathbb{E}_\theta(\phi(S)) - \mathbb{E}_\theta(\phi'(S)) = 0 \implies h(S) = 0$ w.p.1 $\implies \phi(S) = \phi'(S)$ w.p.1. If T is a function of S and T' has smaller variance, then $\mathsf{var}_\theta(\phi'(S)) \le \mathsf{var}_\theta(T') \le \mathsf{var}_\theta(T) = \mathsf{var}_\theta(\phi(S))$. This is a contradiction.(Because $\phi(S) = \phi'(S)$w.p.1.) Thus $\phi(S)$ is a UMVUE. $\square$

Remark 13. $X \sim Pois(\lambda)$, $g(\lambda) = e^{-2\lambda}$, $T(X) = (-1)^X$ is unbiased for $g(\lambda)$. This is because $\mathbb{E}_\lambda(T(X)) = \mathbb{E}_\lambda((-1)^X) = \sum_{x=0}^{\infty} \frac{e^{-\lambda}(-\lambda)^x}{x!} = e^{-\lambda}\sum_{x=0}^{\infty}\frac{(-\lambda)^x}{x!} = e^{-\lambda}e^{-\lambda} = g(\lambda)$. And X is a CSS $\implies$ $T(X)$ is a UMVUE of $e^{-2\lambda}$. But $T(X) = 1$ when X is even, -1 when X is odd. The key reason is that "unbiasedness" is too restrictive!

Definition 14 (Ancillary statistic (AS))**.** A statistic $A(X)$ is said to be ancillary for θ if the law of $A(X)$ does not depend on θ.

Theorem 15 (Basu's theorem)**.** *If T and A are CSS and AS for θ, then $T \perp\!\!\!\perp A$.*

Proof. For any measurable set B, define $h_B(T) \coloneqq \mathbb{P}_\theta(A \in B|T) - \mathbb{P}_\theta(A \in B)$ is a function of T not depending on θ by ancillarity of A and sufficiency of T. T is complete so $\mathbb{E}_\theta h_B(T) = 0$ implies $h_B(T) \equiv 0$ so $A \perp\!\!\!\perp T$. $\square$

Do not confuse ancillary statistics with "pivotal quantities". A pivotal quantity is a function of the data and the unknown parameters such that its distribution does not depend on the underlying distribution. So an ancillary statistic is a pivotal quantity but without depending on the unknown parameter.

Both ancillary statistics and pivotal quantities can be used to construct confidence intervals.

Example 1. *$X_1, \cdots, X_n \overset{iid}{\sim} N(\mu, \sigma^2)$ then $\frac{\bar{X}-\mu}{s_n/\sqrt{n}} \sim t_{n-1}$, so it is a pivotal quantity and we can construct a valid $(1-\alpha)$ CI for μ by inverting the quantile of the Student's T distribution with $n-1$ degree of freedom. If normality is not assumed, we can use CLT and Slutsky theorem to show it is an asymptotic pivotal quantity and we can construct an asymptotically valid $(1-\alpha)$ CI for μ by inverting the quantile of the standard normal.*

2.2 Distribution-free statistics

2.2.1 Conformal inference

A related modern statistical concept is conformal inference [**?** **?**] and inference after model selection [**?** **?** **?**].

First let's consider the simple setting. $Y_1, \cdots, Y_n, Y_{n+1} \sim \mathbb{P}$ and we want to build a confidence interval for the unseen Y_{n+1}. and the only assumption we make is that the joint distribution $Y_1, \cdots, Y_{n+1}$ is exchangeable:

$$\mathbb{P}(Y_1, \cdots, Y_{n+1}) = \mathbb{P}(Y_{\sigma(1)}, \cdots, Y_{\sigma(n+1)})$$

for any permutation $\sigma(\{1, \cdots, n+1\}) = \{\sigma(1), \cdots, \sigma(n+1)\}$.

Now we instead consider the regression setup. Given n i.i.d. pairs (X_i, Y_i), divide the data into two disjoint groups, with sizes n_1 and n_2. Use sample 1 $\mathcal{D}_1$ to fit deep neural nets and obtain $\widehat{f}_1$. Denote the new data point as (X', Y'). We would like to construct a prediction interval $\widehat{\mathsf{PI}}_\alpha$ around $\widehat{f}_1(X')$ such that with probability at least $1-\alpha$ $Y' \in \widehat{\mathsf{PI}}_\alpha$. Now we can use sample 2 $\mathcal{D}_2$ because $\mathcal{D}_1 \perp\!\!\!\perp \mathcal{D}_2 \perp\!\!\!\perp (X', Y')$. How to make no assumptions except i.i.d./exchangeability? The vanilla conformal inference goes as follows: Compute the residuals $\mathcal{E}_i = |Y_i - \widehat{f}_1(X_i)|$ for every $(X_i, Y_i) \in \mathcal{D}_2$. Then find out the order statistics $\mathcal{E}_{(\lceil (1-\alpha)(n_2+1) \rceil)}$. Finally, define $\widehat{\mathsf{PI}} = (\widehat{f}_1(X') + \mathcal{E}_{(\lceil (1-\alpha)(n_2+1) \rceil)})$. One can easily show

$$\mathbb{P}\left(Y' \in \widehat{\mathsf{PI}}\right) = \mathbb{P}\left(Y' - \widehat{f}_1(X') \le \mathcal{E}_{(\lceil (1-\alpha)(n_2+1) \rceil)}\right) \ge 1 - \alpha.$$

If the residuals $\{\mathcal{E}_i\}, i = 1, \cdots, n_2$ have a continuous distribution (so no ties), and $n_1 = n_2 = n/2$, one can easily show [**?**]

$$\mathbb{P}\left(Y' \in \widehat{\mathsf{PI}}\right) = \mathbb{P}\left(Y' - \widehat{f}_1(X') \le \mathcal{E}_{(\lceil (1-\alpha)(n_2+1) \rceil)}\right) \le 1 - \alpha + \frac{1}{n/2+1}$$

2.2.2 Rank-based statistic

But there is one caveat about rank – it is tricky to generalize to multivariate settings. But it is still possible. Can you think of some reasonable strategies?

2.2.3 Permutation-based statistic

Consider $X_1, \cdots, X_n \sim \mathbb{P}$ and $Y_1, \cdots, Y_m \sim \mathbb{Q}$ both are exchangeable. We want to test $H_0 : \mathbb{P} = \mathbb{Q}$. Let $Z = (Z_1, \cdots, Z_N) = (X_1, \cdots, X_n, Y_1, \cdots, Y_m)$. Under H_0, Z is exchangeable. Denote the permutation group Π_N over $\{1, \cdots, N\}$. Then $Z \sim \pi \circ Z$ for any $\pi \in \Pi_N$.

We want to construct a test $T_\alpha(Z) \in \{0, 1\}$ such that under H_0, the nominal type-I error is guaranteed: $\mathbb{P}_{H_0}(T_\alpha(Z) = 1) \le \alpha$. To get a test, one needs a test statistic $S_{n,m} = s(Z)$. Denote $S^\pi_{n,m} = s(\pi \circ Z)$ for $\pi \in \Pi_N$. Under H_0, $S^\pi_{n,m}$ are identically distributed over all $\pi \in \Pi_N$ and there are $N!$ many of them. Say we did the permutation B times with $B < N!$ (including the unpermuted one) and label them as $S^{\pi_0}_{n,m} \equiv S_{n,m}, S^{\pi_1}_{n,m}, \cdots, S^{\pi_{B-1}}_{n,m}$. Take $k = \lceil (1-\alpha)B \rceil$. Let $S^{(k)}_{n,m}$ be the k-th smallest among $S^{\pi_0}_{n,m} \equiv S_{n,m}, S^{\pi_1}_{n,m}, \cdots, S^{\pi_{B-1}}_{n,m}$. Then we consider the following test:

$$T_\alpha = \begin{cases} 0 & S_{n,m} \le S^{(k)}_{n,m} \\ 1 & S_{n,m} > S^{(k)}_{n,m} \end{cases}$$

It is easy to see that

$$\mathbb{P}_{H_0}(T_\alpha = 1) \leq \alpha.$$

Now think about the following problem: what if we instead testing $H_0 : \mathbb{E}X = \mathbb{E}Y$? Will permutation test still have the desired type-I error?

2.3 Decision theoretic aspects of statistics

From a societal perspective, statistical analysis is an interplay/game played between an experimenter/statistician and the nature (or nowadays even some adversary). This is the viewpoint taken by David Blackwell [**?**][1] and Lucien Le Cam [**?**].

Blackwell and Le Cam begins with the following setup:

Given a set of data $\mathcal{D}$, we have a scientific problem of interest abstracted as a parameter θ, and if you construct from $\mathcal{D}$ an estimator $\widehat{\theta}$ of θ, how do you tell if it is a good estimator? One approach is through decision theory by introducing a loss function ℓ and its associated risk R. To be more precise

Definition 16. Let $\widehat{T} : \mathcal{D} \to \widehat{\mathcal{Y}}$ be a decision procedure. The parameter of interest is actually $T(\theta)$ where $T : \Theta \to \mathcal{Y}$. In the most general form, $\widehat{\mathcal{Y}}$ and $\mathcal{Y}$ can be different. Given a loss function $\ell : \mathcal{Y} \times \widehat{\mathcal{Y}} : \mathbb{R}_+$, we define the risk of $\widehat{T}$ at parameter θ as

$$R_\theta(\widehat{T}) = \mathbb{E}_\theta \ell(\widehat{T}, T(\theta)) \tag{12}$$

Remark 17. $\widehat{T}$ can be a randomized algorithm e.g. $\widehat{T}(\mathcal{D}, U)$ where U is a random variable independent of the data. Randomized algorithm is extremely useful – it often can improve the computational speed of an algorithm by orders of magnitude by trading-off some accuracy (though still accurate with high probability). Jelani Nelson, James Lee, David Woodruff, Ryan O'Donnell and Edgar Dobriban (the only statistician in this list) are the people you should search for if interested in such problems.

With the definition of loss function and its associated risk, we can start to compare estimators. The very first criterion that statisticians came up with is admissibility of a decision procedure.

Definition 18. $\widehat{T}$ is said to be inadmissible if $\exists\ \widehat{T}'$, such that $R_\theta(\widehat{T}') \leq R_\theta(\widehat{T})\ \forall \theta \in \Theta$ and for some $\theta_0 \in \Theta$, $R_{\theta_0}(\widehat{T}') < R_{\theta_0}(\widehat{T})$.

Example 2. *This is not a perfect criterion. Consider $X \sim$ Bernoulli(p). Take $\widehat{T} = 0.5$, it has 0 risk at $p = 0.5$ but lousy at other p's. It is immediate to see that for any other $\widehat{T}'$, if we require $R_p(\widehat{T}') \leq R_p(\widehat{T})$ at all p, then it must hold that $R_{0.5}(\widehat{T}') \leq R_{0.5}(\widehat{T}) = 0$. But $R_{0.5}(\widehat{T}') \equiv 0$ if and only if $\widehat{T}' \equiv 0.5$.*

For more details on admissibility, you should read "Theory of Point Estimation" by Erich Lehmann and George Casella [**?**], which contains all the details on this topic.

Other reasonable approaches of comparing estimator include the following:

[1] David Blackwell was an expert in probability theory, information theory, game theory and statistics. He was the first Black member elected into the American Academy of Arts and Sciences.

1. Restrict the type of estimators (e.g. linear estimator, quadratic estimator, equivariant estimator, rotation-invariant estimator [**?**], etc.)

2. Average-case risk: First define the posterior risk

$$R_\pi(\widehat{T}) = \mathbb{E}_{\theta\sim\pi} R_\theta(\widehat{T})$$

With that, define the Bayes risk as $R^*_\pi = \inf_{\widetilde{T}} R_\pi(\widetilde{T})$. We may also define the worst-case Bayes risk: $R^*_B = \sup_\pi R^*_\pi$.

3. Worst-case risk (Minimax)

$$R^* = \inf_{\widetilde{T}} \sup_{\theta\in\Theta} R_\theta(\widetilde{T}).$$

We will study minimax risk frequently during this course.

We have the following connection between Bayes risk and admissibility:

Theorem 19. *A unique Bayes estimator is admissible.*

Proof. If $\widehat{T}_\Pi$ is a Bayes estimator of $T(\theta)$ for the prior Π and is not admissible, then for some other $\widehat{T}'$, $R_\theta(\widehat{T}') \le R_\theta(\widehat{T}_\Pi)$ for all $\theta \in \Theta$. Then

$$\int_{\theta\in\Theta} R_\theta(\widehat{T}')\mathrm{d}\Pi(\theta) \le \int_{\theta\in\Theta} R_\theta(\widehat{T}_\Pi)\mathrm{d}\Pi(\theta).$$

But since $\widehat{T}_\Pi$ is a Bayes estimator, $\widehat{T}'$ is also a Bayes estimator. By uniqueness, we have $\widehat{T}_\Pi$ must be admissible. □

In fact, admissible estimator is either a Bayes estimator or a limit of a sequence of Bayes estimators only under certain extra conditions. For the complete statement and a proof of this claim, see the Appendix for Chapter 4 of [**?**].

An obvious connection between minimax risk and worst-case Bayes risk is

$$R^* \ge R^*_B = \sup_\pi R^*_\pi.$$

Example 3. *Player A (statistician) guesses any $\widehat{\theta}$ from all natural numbers and compares with Player B (nature)'s choice. The loss for player A is $\mathbb{1}\{\widehat{\theta} < \theta\}$. Here $R^* \ge \lim_{\theta\to\infty} \mathbb{P}_\theta(\widehat{\theta} < \theta) = 1$. But for any prior Π on all natural numbers, $R_\Pi(\widehat{\theta}) = \int \mathbb{P}_\theta(\widehat{\theta} < \theta)\mathrm{d}\Pi(\theta)$ and if we let $\widehat{\theta} \to \infty$, $R_\Pi(\widehat{\theta}) = 0$. This is true for any prior so $R^* \ge 1 > 0 = R^*_B$.*

Remark 20. The above inequality has a game-theoretic interpretation in the context of a min-max game. Player A tries to minimize $R_\theta(\widehat{T})$ whereas Player B tries to maximize $R_\theta(\widehat{T})$. Whoever goes first has a disadvantage: in worst-case setting, statistician goes first and nature goes second

Then the question is when equality holds. This can often be answered by viewing minimax risk and Bayes risk in optimization lens. Let us look at minimax risk but for the moment assuming (1) our goal is to estimate the parameter θ itself, (2) Θ is finite hence inf and sup are simply min and max, and (3) the loss function $\ell(\theta, \widehat{\theta})$ is convex in the argument $\widehat{\theta}$:

$$\begin{aligned} R^* &= \min_{\widehat{\theta}} \max_{\theta\in\Theta} \mathbb{E}_\theta[\ell(\widehat{\theta}, \theta)] \\ &= \min_{\widehat{\theta}, v} v, \text{ subject to } \mathbb{E}_\theta[\ell(\widehat{\theta}, \theta)] \le v, \theta \in \Theta. \end{aligned} \tag{13}$$

R^* is in fact a convex program: the mapping $\mathbb{P}_{\widehat{\theta}|\text{Data}} \mapsto \mathbb{E}_\theta[\ell(\widehat{\theta}, \theta)]$ is simultaneously convex and concave (so affine) and $\max_{\theta\in\Theta}$ over affine functions is a convex function.

Now look at the Lagrangian of (13):

$$\begin{aligned}\mathsf{Lagrangian}(\widehat{\theta}, v, \lambda) &= v + \sum_{\theta\in\Theta} \lambda_\theta \left(\mathbb{E}_\theta[\ell(\widehat{\theta}, \theta)] - v\right) \\ &= \left(1 - \sum_{\theta\in\Theta} \lambda_\theta\right) v + \sum_{\theta\in\Theta} \lambda_\theta \mathbb{E}_\theta[\ell(\widehat{\theta}, \theta)]\end{aligned} \tag{14}$$

where λ's are the Lagrangian multipliers. Then the Lagrangian-Fenchel dual is

$$\max_{\lambda} \min_{\widehat{\theta}, v} \mathsf{Lagrangian}(\widehat{\theta}, v, \lambda).$$

But note that unless $\sum_{\theta\in\Theta} \lambda_\theta = 1$, $\min_{\widehat{\theta},v} \mathsf{Lagrangian}(\widehat{\theta}, v, \lambda) = -\infty$, which is moot. Thus λ forms a probability measure over Θ. But the dual problem under this constraint becomes

$$\max_{\lambda\in\text{Probability Measures}(\Theta)} \min_{\widehat{\theta}} R_\lambda(\widehat{\theta}) \equiv R_B^*.$$

So whenever strong duality holds, we have Bayes risk equals minimax risk. One special sufficient condition is when both Θ and the data are finite sets.

Example 4. *Estimating a normal mean μ from a single observation $X \sim N(\mu, 1)$. An obvious estimator for μ is X. Is it minimax optimal under squared L_2 risk? First compute its risk: $R_\mu(X) = \mathbb{E}(X - \mu)^2 = 1$. The supremum over $R_\mu(X)$ is always 1. Is it tight i.e. $R^* \geq 1$? In this simple example, we can look at the Bayes risk:*

$$R^* \geq \sup_\pi R_\pi^*.$$

Choose a prior $\pi = N(0, \sigma^2)$ on μ. Then the Bayes estimator in this case is

$$\widehat{T}_{Bayes} = \mathbb{E}[\mu|X] = X\frac{\sigma^2}{1+\sigma^2}$$

and $R_\pi(\widehat{T}_{Bayes})$ is

$$\mathbb{E}_{\mu\sim N(0,\sigma^2)}\mathbb{E}_\mu(\mu - \mathbb{E}[\mu|X])^2 = \mathbb{E}_{\mu\sim N(0,\sigma^2)}\mathbb{E}_\mu\left(\mu - \frac{1}{1+\sigma^2}X\right)^2 = \frac{\sigma^2}{1+\sigma^2} \leq 1.$$

Here a worst-case prior is quite obvious: $R_B^* = \sup_\pi R_\pi(\widehat{T}_{Bayes}) \geq \lim_{\sigma^2\to\infty} \frac{\sigma^2}{1+\sigma^2} = 1$. So we conclude that X is minimax optimal under squared L_2 loss. But minimax estimator need not be unique: As we will see shortly after, another estimator called James-Stein estimator dominates X in the squared L_2 risk so it is also minimax optimal.

Example 5. *We can generalize the above calculations to many normal mean models: $\boldsymbol{X} \sim MVN(\boldsymbol{\mu}, \sigma^2 I)$ with d dimensions. Then the minimax risk is $d\sigma^2$.*

Example 6. *Now if we observe n independent samples rather than one sample: $\boldsymbol{X}_i = \boldsymbol{\mu} + \boldsymbol{Z}_i$, $\boldsymbol{Z}_i \overset{iid}{\sim} MVN(\boldsymbol{0}, I)$ again with $\boldsymbol{\mu} \in \mathbb{R}^d$. Then $R^* = \frac{d}{n}$.*

Proof. Obviously the empirical mean $\bar{\boldsymbol{X}}$ is a sufficient statistic of $\boldsymbol{\mu}$, which reduces the model to $\bar{\boldsymbol{X}} \sim MVN(\boldsymbol{\mu}, \frac{1}{n}I)$. Using the minimax risk of many normal mean models, we have $R^* = \frac{d}{n}$. □

Example 7. *Now what if the normality assumption is dropped?* $\boldsymbol{X}_i = \boldsymbol{\mu} + \boldsymbol{Z}_i$, $\boldsymbol{Z}_i \overset{iid}{\sim} (0, Id)$ *where* $(0, Id)$ *is a abuse of notation of mean zero and identity covariance matrix. What is* R^* *under squared* L_2 *risk?*

Proof. In terms of the upper bound, let us try $\bar{\boldsymbol{X}}$.

$$\begin{aligned}
R_{\boldsymbol{\mu}}(\bar{\boldsymbol{X}}) &= \mathbb{E}\|\bar{\boldsymbol{X}} - \boldsymbol{\mu}\|^2 = \mathbb{E}\bar{\boldsymbol{X}}^\top \bar{\boldsymbol{X}} - \boldsymbol{\mu}^\top \boldsymbol{\mu} \\
&= \sum_{j=1}^{d} \left(\mathbb{E}[\bar{\boldsymbol{X}}_j^2] - \boldsymbol{\mu}_j^2\right) \\
&= \sum_{j=1}^{d} \left(\frac{1}{n^2}\mathbb{E}\left[\left(\sum_{i=1}^{n} \boldsymbol{X}_{i,j}\right)^2\right] - \boldsymbol{\mu}_j^2\right) \\
&= \sum_{j=1}^{d} \left(\frac{1}{n}\mathbb{E}[\boldsymbol{X}_{1,j}^2] + \frac{n-1}{n}\mathbb{E}[\boldsymbol{X}_{1,j}\boldsymbol{X}_{2,j}] - \boldsymbol{\mu}_j^2\right) \\
&= \sum_{j=1}^{d} \left(\frac{1}{n}(\boldsymbol{\mu}_j^2 + 1) + \frac{n-1}{n}\boldsymbol{\mu}_j^2 - \boldsymbol{\mu}_j^2\right) \\
&= \frac{d}{n}.
\end{aligned}$$

How about the lower bound? Since in this example, we are working on a much more general setting than the multivariate Gaussian example. By plain logic, the lower bound in a general setting must not be smaller than the lower bound in a special setting. Thus $R^* \geq \frac{d}{n}$ still holds. □

The above setting is what statisticians actually care about. Given a statistical model and a statistical problem, how many samples should I collect to achieve a desirable statistical accuracy. When the data is i.i.d., then the space of probability measures of n data is

$$\mathcal{P}_n = \left\{\mathbb{P}_\theta^{\otimes n} : \theta \in \Theta\right\} \tag{15}$$

Definition 21 (Sample complexity)**.** The sample complexity of a statistical model is, given an error tolerance $\epsilon > 0$,

$$n^*(\epsilon) = \min\{n \in \mathbb{N} : R^*(\mathcal{P}_n) \leq \epsilon\}.$$

In research, people sometimes want high-probability guarantee instead of average guarantee: given an error tolerance $\epsilon > 0$ and a confidence level $0 < \delta < 1$, the sample complexity $n^*(\epsilon, \delta)$ under loss ℓ is the smallest natural number such that

$$\inf_{\widehat{\theta}_n} \sup_{\theta \in \Theta} \mathbb{P}_\theta \left(\ell(\theta, \widehat{\theta}_n) \leq \epsilon\right) \geq 1 - \delta.$$

Remark 22 (Relation with PAC learnability)**.** In 1984, now Turing award winner Leslie G Valiant [**?**] created a concept called PAC (Probably Approximately Correct) learnability. It is viewed as the theoretical foundation of modern machine learning research because it is closely related to the generalization error. PAC learnability is defined as: for any $\delta > 0$ and $\widehat{\theta}$

$$\sup_{\theta \in \Theta} \mathbb{P}_\theta \left(\ell(\widehat{\theta}_n, \theta) \leq \epsilon \right) \geq 1 - \delta \tag{16}$$

but ϵ only depends on the data and δ, and most importantly, $\widehat{\theta}$ is usually over poly-time computable quantities at least in theoretical computer scientists' mind.

An interesting anecdote: Larry Wasserman wrote a blog in 2013 on PAC learnability and, just like most statisticians, tried to protect our own territory by claiming PAC learnability was what statisticians have been doing for the last century. This inevitably stimulated a lot of debate between him and theoretical computer scientists, who were not as onto statistical problems as they are today.

2.4 Stein's paradox

For the many normal mean model, we have seen that $\boldsymbol{X}$ is the MLE and the minimax optimal estimator. But is it admissible? Interestingly, when $d = 1, 2$, it is admissible and proved by Charles Stein himself in Sections 2-4 of [**?**]. But Charles Stein shocked the statistics world in that same 1956 paper by showing that it is inadmissible when $d \geq 3$ and completely dominated by a nonlinear estimator, later called James-Stein estimator [**?** **?**] under the quadratic loss. This estimator is the precursor of ridge regression, Lasso, and many other popular estimators in high-dimensional statistics.

We consider the following setup: $\boldsymbol{X}_{d\times 1} \sim N(\boldsymbol{\mu}_{d\times 1}, \sigma^2 \mathsf{Id}_{d\times d})$, with σ^2 known to us and the unknown parameter is $\boldsymbol{\mu}$.

First, observe that

$$R_{\boldsymbol{\mu}}(\boldsymbol{X}) = \mathbb{E}_{\boldsymbol{\mu}} \|\boldsymbol{\mu} - \boldsymbol{X}\|_2^2 = \sum_{i=1}^{d} \mathbb{E}_{\mu_i} (X_i - \mu_i)^2 = d\sigma^2.$$

Stein considers spherically symmetric estimators (lying on the line passing through $\boldsymbol{X}$ and the distance to 0 only depends on $\|\boldsymbol{X}\|_2$ or $\Gamma^{-1}\widehat{\boldsymbol{\mu}}(\Gamma \boldsymbol{X})$ for any orthogonal transformation Γ) of the following form:

$$\widehat{\boldsymbol{\mu}}(\boldsymbol{X}) = g(\boldsymbol{X})\boldsymbol{X} \text{ i.e. } \begin{pmatrix} \widehat{\mu}_1(\boldsymbol{X}) \\ \vdots \\ \widehat{\mu}_d(\boldsymbol{X}) \end{pmatrix} = \begin{pmatrix} g(\boldsymbol{X}) X_1 \\ \vdots \\ g(\boldsymbol{X}) X_d \end{pmatrix}.$$

g is usually called Stein's shrinker. Then our goal is to find a Stein's shrinker g s.t.

$$R_{\boldsymbol{\mu}}(\widehat{\boldsymbol{\mu}}(\boldsymbol{X})) - R_{\boldsymbol{\mu}}(\boldsymbol{X}) = R_{\boldsymbol{\mu}}(\widehat{\boldsymbol{\mu}}(\boldsymbol{X})) - d\sigma^2 < 0.$$

Let us calculate the risk of $\widehat{\boldsymbol{\mu}}(\boldsymbol{X})$ first.

$$R_{\boldsymbol{\mu}}(\widehat{\boldsymbol{\mu}}(\boldsymbol{X})) = \mathbb{E}_{\boldsymbol{\mu}} \|\widehat{\boldsymbol{\mu}}(\boldsymbol{X}) - \boldsymbol{\mu}\|_2^2$$

$$
\begin{aligned}
&= \mathbb{E}_{\boldsymbol{\mu}} \|\widehat{\boldsymbol{\mu}}(\boldsymbol{X}) - \boldsymbol{X} + \boldsymbol{X} - \boldsymbol{\mu}\|_2^2 \\
&= \mathbb{E}_{\boldsymbol{\mu}} \|\boldsymbol{\mu} - \boldsymbol{X}\|_2^2 + \mathbb{E}_{\boldsymbol{\mu}} \|\widehat{\boldsymbol{\mu}}(\boldsymbol{X}) - \boldsymbol{X}\|_2^2 + 2\mathbb{E}_{\boldsymbol{\mu}} \left(\widehat{\boldsymbol{\mu}}(\boldsymbol{X}) - \boldsymbol{X}\right)^\top (\boldsymbol{X} - \boldsymbol{\mu}) \\
&= d\sigma^2 + \mathbb{E}_{\boldsymbol{\mu}} \|\widehat{\boldsymbol{\mu}}(\boldsymbol{X}) - \boldsymbol{X}\|_2^2 + 2\mathbb{E}_{\boldsymbol{\mu}} \widehat{\boldsymbol{\mu}}(\boldsymbol{X})^\top (\boldsymbol{X} - \boldsymbol{\mu}) - 2\mathbb{E}_{\boldsymbol{\mu}} \boldsymbol{X}^\top (\boldsymbol{X} - \boldsymbol{\mu}) \\
&= d\sigma^2 + \mathbb{E}_{\boldsymbol{\mu}} \|\widehat{\boldsymbol{\mu}}(\boldsymbol{X}) - \boldsymbol{X}\|_2^2 + 2\mathbb{E}_{\boldsymbol{\mu}} \widehat{\boldsymbol{\mu}}(\boldsymbol{X})^\top (\boldsymbol{X} - \boldsymbol{\mu}) - 2\mathbb{E}_{\boldsymbol{\mu}} (\boldsymbol{X} - \boldsymbol{\mu})^\top (\boldsymbol{X} - \boldsymbol{\mu}) \\
&= d\sigma^2 + \mathbb{E}_{\boldsymbol{\mu}} \|\widehat{\boldsymbol{\mu}}(\boldsymbol{X}) - \boldsymbol{X}\|_2^2 + 2\mathbb{E}_{\boldsymbol{\mu}} \widehat{\boldsymbol{\mu}}(\boldsymbol{X})^\top (\boldsymbol{X} - \boldsymbol{\mu}) - 2\mathbb{E}_{\boldsymbol{\mu}} \|\boldsymbol{X} - \boldsymbol{\mu}\|_2^2 \\
&= -d\sigma^2 + \mathbb{E}_{\boldsymbol{\mu}} \|\widehat{\boldsymbol{\mu}}(\boldsymbol{X}) - \boldsymbol{X}\|_2^2 + 2\sum_{i=1}^d \mathbb{E}_{\boldsymbol{\mu}} \widehat{\mu}_i(\boldsymbol{X})(X_i - \mu_i).
\end{aligned}
$$

Then recall the famous Stein's identity for normal distribution, we have

$$
\mathbb{E}_{\boldsymbol{\mu}} \widehat{\mu}_i(\boldsymbol{X})(X_i - \mu_i) = \sigma^2 \mathbb{E}_{\boldsymbol{\mu}} \left[\frac{\partial \widehat{\mu}_i(\boldsymbol{X})}{\partial X_i}\right] = \sigma^2 \mathbb{E}_{\boldsymbol{\mu}} \left[\frac{\partial g(\boldsymbol{X})}{\partial X_i} X_i + g(\boldsymbol{X})\right]
$$

Plugging in $R_{\boldsymbol{\mu}}(\widehat{\boldsymbol{\mu}}(\boldsymbol{X}))$, we have

$$
\begin{aligned}
R_{\boldsymbol{\mu}}(\widehat{\boldsymbol{\mu}}(\boldsymbol{X})) &= -d\sigma^2 + \mathbb{E}_{\boldsymbol{\mu}} \|g(\boldsymbol{X})\boldsymbol{X} - \boldsymbol{X}\|_2^2 + 2\sigma^2 \sum_{i=1}^d \mathbb{E}_{\boldsymbol{\mu}} \left[\frac{\partial g(\boldsymbol{X})}{\partial X_i} X_i + g(\boldsymbol{X})\right] \\
&= d\sigma^2 \mathbb{E}_{\boldsymbol{\mu}} \left[2g(\boldsymbol{X}) - 1\right] + \mathbb{E}_{\boldsymbol{\mu}} \|(g(\boldsymbol{X}) - 1)\boldsymbol{X}\|_2^2 + 2\sigma^2 \sum_{i=1}^d \mathbb{E}_{\boldsymbol{\mu}} \left[\frac{\partial g(\boldsymbol{X})}{\partial X_i} X_i\right].
\end{aligned}
$$

Then Stein guess the following ansatz:

$$
g(\boldsymbol{x}) = 1 - \frac{c}{\|\boldsymbol{x}\|_2^2}
$$

with partial derivative, for $i = 1, \cdots, d$,

$$
\frac{\partial g(\boldsymbol{x})}{\partial x_i} = \frac{2cx_i}{\|\boldsymbol{x}\|_2^4}.
$$

Then we want to tune the value of c such that

$$
\begin{aligned}
&R_{\boldsymbol{\mu}}(\widehat{\boldsymbol{\mu}}(\boldsymbol{X})) - R_{\boldsymbol{\mu}}(\boldsymbol{X}) = d\sigma^2 \mathbb{E}_{\boldsymbol{\mu}} \left[2(g(\boldsymbol{X}) - 1)\right] + \mathbb{E}_{\boldsymbol{\mu}} (g(\boldsymbol{X}) - 1)^2 \|\boldsymbol{X}\|_2^2 + 2\sigma^2 \sum_{i=1}^d \mathbb{E}_{\boldsymbol{\mu}} \left[\frac{\partial g(\boldsymbol{X})}{\partial X_i} X_i\right] \\
&= -2d\sigma^2 \mathbb{E}_{\boldsymbol{\mu}} \left[\frac{c}{\|\boldsymbol{X}\|_2^2}\right] + \mathbb{E}_{\boldsymbol{\mu}} \left[\frac{c^2 \|\boldsymbol{X}\|_2^2}{\|\boldsymbol{X}\|_2^4}\right] + 4\sigma^2 \mathbb{E}_{\boldsymbol{\mu}} \left[\frac{c \sum_{i=1}^d X_i^2}{\|\boldsymbol{X}\|_2^4}\right] \\
&= -2d\sigma^2 \mathbb{E}_{\boldsymbol{\mu}} \left[\frac{c}{\|\boldsymbol{X}\|_2^2}\right] + \mathbb{E}_{\boldsymbol{\mu}} \left[\frac{c^2}{\|\boldsymbol{X}\|_2^2}\right] + 4\sigma^2 \mathbb{E}_{\boldsymbol{\mu}} \left[\frac{c}{\|\boldsymbol{X}\|_2^2}\right] \\
&= \mathbb{E}_{\boldsymbol{\mu}} \left[\frac{1}{\|\boldsymbol{X}\|_2^2} \left(c^2 + 4c\sigma^2 - 2d\sigma^2 c\right)\right] < 0.
\end{aligned}
$$

Consider the function $h(c) = c^2 + 4c\sigma^2 - 2d\sigma^2 c$. The above desiderata is satisfied if the minimizer of h is negative:

$$
h^\circ = \min_{c \in \mathbb{R}} h(c) : 2c^\circ + 4\sigma^2 - 2d\sigma^2 = 0 \Rightarrow c^\circ = \sigma^2(d-2)
$$

$$\Rightarrow h^\circ = \sigma^4(d-2)^2 + 4\sigma^4(d-2) - 2d\sigma^4(d-2) = \sigma^4(d-2)^2 - 2\sigma^4(d-2)^2 = -\sigma^4(d-2)^2 < 0$$

if $d \geq 3$. Hence when we choose $g^\circ(\boldsymbol{X}) = 1 - \frac{\sigma^2(d-2)}{\|\boldsymbol{X}\|_2^2}$, $R_{\boldsymbol{\mu}}(\widehat{\boldsymbol{\mu}}(\boldsymbol{X})) - R_{\boldsymbol{\mu}}(\boldsymbol{X}) < 0$ provided that $\mathbb{E}_{\boldsymbol{\mu}}\left[\frac{1}{\|\boldsymbol{X}\|_2^2}\right]$ exists.

Luckily, for $d \geq 3$, we have the following lemma:

Lemma 23. *For $d \geq 3$, for any $\boldsymbol{\mu} \in \mathbb{R}^d$, $0 < \mathbb{E}_{\boldsymbol{\mu}}\left[\frac{1}{\|\boldsymbol{X}\|_2^2}\right] < \infty$.*

Proof. This lemma can be proved by exploiting the rotation invariance of isotropic Gaussian. First, we observe

$$\mathbb{E}\left[\frac{1}{\|\boldsymbol{X}\|_2^2}\right] = \mathbb{E}\left[\frac{1}{\|\sigma Z + \boldsymbol{\mu}\|_2^2}\right] = \mathbb{E}\left[\frac{1}{\sigma^2\|Z + \sigma^{-1}\boldsymbol{\mu}\|_2^2}\right].$$

By rotation invariance, we know $\|Z + \boldsymbol{\nu}\|_2^2 \sim \|Z + \boldsymbol{\nu}'\|_2^2$ for any $\boldsymbol{\nu}, \boldsymbol{\nu}'$ such that $\|\boldsymbol{\nu}\|_2 = \|\boldsymbol{\nu}'\|_2$ because one can always find a $\Gamma \in \mathcal{O}$ such that $\Gamma\boldsymbol{\nu} = \boldsymbol{\nu}'$. So we can create a new vector $\boldsymbol{\nu} = \left(\sigma^{-1}\|\boldsymbol{\mu}\|_2, 0, \cdots, 0\right)^\top$ and have

$$\begin{aligned}
\mathbb{E}\left[\frac{1}{\|\boldsymbol{X}\|_2^2}\right] &= \mathbb{E}\left[\frac{1}{\sigma^2\|Z + \boldsymbol{\nu}\|_2^2}\right] \\
&= \frac{1}{\sigma^2(\sqrt{2\pi})^d}\int \exp\left(-\frac{\|z\|_2^2}{2}\right)\|\boldsymbol{\nu} + z\|_2^{-2}\mathrm{d}z \\
&= \frac{1}{\sigma^2(\sqrt{2\pi})^d}\int \exp\left(-\frac{\|x - \boldsymbol{\nu}\|_2^2}{2}\right)\|x\|_2^{-2}\mathrm{d}x \\
&= \frac{1}{\sigma^2(\sqrt{2\pi})^d}\exp\left(-\frac{\sigma^{-2}\|\boldsymbol{\mu}\|_2^2}{2}\right)\int \exp\left(\frac{x_1\|\boldsymbol{\mu}\|_2}{\sigma} - \frac{\|x\|_2^2}{2}\right)\|x\|_2^{-2}\mathrm{d}x \\
&\leq \frac{1}{\sigma^2(\sqrt{2\pi})^d}\exp\left(-\frac{\sigma^{-2}\|\boldsymbol{\mu}\|_2^2}{2}\right)\int \exp\left(\frac{3\|\boldsymbol{\mu}\|_2^2}{\sigma^2} + \frac{\|x\|_2^2}{3} - \frac{\|x\|_2^2}{2}\right)\|x\|_2^{-2}\mathrm{d}x \\
&= \frac{1}{\sigma^2(\sqrt{2\pi})^d}\exp\left(\frac{5\sigma^{-2}\|\boldsymbol{\mu}\|_2^2}{2}\right)\int \exp\left(-\frac{\|x\|_2^2}{6}\right)\|x\|_2^{-2}\mathrm{d}x \\
&\lesssim \frac{1}{\sigma^2(\sqrt{2\pi})^d}\exp\left(\frac{5\sigma^{-2}\|\boldsymbol{\mu}\|_2^2}{2}\right)\int_0^\infty \exp\left(-\frac{r^2}{6}\right)r^{d-3}\mathrm{d}r
\end{aligned}$$

where in the first inequality we use Young's inequality $|xy| \leq 3x^2 + y^2/3$ and in the second inequality we use the polar transformation. □

Why intuitively $d = 3$ is the critical dimension and why spherically symmetric estimators? Or even more simply put, why shrinkage at all? Stein's original proof has provided a great deal of insights already and it is deeply connected to the rotation invariance of Gaussian. First, why shrinkage? The following heuristic argument might be helpful: Using standard chi-squared concentration bound

$$\|\boldsymbol{X}\|_2^2 = \|\boldsymbol{\mu}\|_2^2 + d + O_p(\sqrt{d})$$

which gives

$$\|\boldsymbol{\mu}\| = \sqrt{\|\boldsymbol{X}\|_2^2 - d - O_p(\sqrt{d})}$$

$$\approx \|\boldsymbol{X}\|_2 - \frac{d + O_p(\sqrt{d})}{2\|\boldsymbol{X}\|_2}.$$

In terms of why only spherically symmetric estimators? I will provide a partial argument given by Stein and you are suggested to read Section 4 of [**?**] to finish the argument. Stein first show that there does not exist estimators other than spherically symmetric estimators that can achieve a better squared L_2 risk than $\boldsymbol{X}$ itself. Suppose on the contrary, there exists such an estimator $\widetilde{\boldsymbol{\mu}}$. Then

$$R_{\boldsymbol{\mu}}(\widetilde{\boldsymbol{\mu}}(\boldsymbol{X})) < R_{\boldsymbol{\mu}}(\widehat{\boldsymbol{\mu}}(\boldsymbol{X})).$$

By continuity of the risk, if the above holds, then it must hold for an open set. Choose some $\Gamma \in \mathcal{O}$, then

$$R_{\boldsymbol{\mu}}(\Gamma^{-1}\widetilde{\boldsymbol{\mu}}(\Gamma\boldsymbol{X})) = \mathbb{E}_{\boldsymbol{\mu}}\left[\left(\Gamma^{-1}\widetilde{\boldsymbol{\mu}}(\Gamma\boldsymbol{X}) - \boldsymbol{\mu}\right)^2\right] = \mathbb{E}_{\boldsymbol{\mu}}\left[\left(\widetilde{\boldsymbol{\mu}}(\Gamma\boldsymbol{X}) - \Gamma\boldsymbol{\mu}\right)^2\right] < R_{\boldsymbol{\mu}}(\widehat{\boldsymbol{\mu}}(\boldsymbol{X})).$$

Thus the above display must also hold for an open set including Γ in $\mathcal{O}$. Take λ to be the Haar measure on $\mathcal{O}$ (an invariant measure). Then define a (Bayesian) spherically symmetric estimator $\widetilde{\boldsymbol{\mu}}'(\boldsymbol{X}) \coloneqq \int \Gamma^{-1}\widetilde{\boldsymbol{\mu}}(\Gamma\boldsymbol{X})\mathrm{d}\lambda(\Gamma)$. Finally, by the convexity of the risk, we have

$$R_{\boldsymbol{\mu}}(\widetilde{\boldsymbol{\mu}}'(\boldsymbol{X})) \le \int R_{\boldsymbol{\mu}}(\Gamma^{-1}\widetilde{\boldsymbol{\mu}}(\Gamma\boldsymbol{X}))\mathrm{d}\lambda(\Gamma) < R_{\boldsymbol{\mu}}(\widetilde{\boldsymbol{\mu}}(\boldsymbol{X})).$$

Lawrence (aka Larry) Brown (passed in 2018), a legendary statistician at Wharton statistics department, spent a large part of his career trying to understand this and provided a probably much deeper reason. As a real mathematician, Brown [**?**] amazingly noticed the connection of this problem to the recurrence of Brownian motion at $d \le 2$ and the transience at $d \ge 3$, and the existence of solution to the exterior Dirichlet problem only if $d \le 2$. The exterior Dirichlet problem is

$$\nabla u = 0, |x| > 1, u = \begin{cases} 1 & |x| = 1 \\ 0 & |x| \to \infty. \end{cases}$$

Brown's resolution has a very strong Bayesian flavor. You can read a summary of the above results in [**?**].

James-Stein estimator has recently motivated a lot of interesting papers in large dimensional ($d \asymp n$) covariance/Gram matrix estimation [**?** **?** **?**]. The original idea (once again) goes back to Charles Stein's paper [**?** **?**].

2.4.1 Empirical Bayes interpretation

James-Stein estimator also goes beyond the shrinkage paradigm. It has a Bayesian interpretation [**?**] first realized by Herbert Robbins, who is more famous for his contribution to sequential decision making. Machine learning people often pay homage to Robbins as the first person working on reinforcement learning.

Consider putting a prior Π on μ_j and $X_j|\mu_j \sim N(\mu_j, 1)$. Then X_j has marginal distribution with density

$$f_{X_j}(x) = \int \varphi(x - \mu_j)\mathrm{d}\Pi(\mu_j)$$

where φ denotes the standard normal cdf. Then we have the following famous Tweedie's formula[2] for the Bayes estimator under the normal model:

$$\mathbb{E}[\mu_j|X_j = x] = x + \frac{\mathrm{d}}{\mathrm{d}x}\log f_{X_j}(x) = x + \frac{f'_{X_j}(x)}{f_{X_j}(x)} \tag{17}$$

So if we choose a prior $\mu_j \sim N(0, \nu^2)$, then marginally $X_j \sim N(0, \kappa^2 = \nu^2 + 1)$ for $j = 1, \cdots, d$. Then

$$\frac{f'_{X_j}(x)}{f_{X_j}(x)} = -\frac{x}{\kappa^2} \text{ so } \mathbb{E}[\mu_j|X_j = x] = x - \frac{x}{\kappa^2}.$$

This is now extremely similar to the James-Stein estimator except that the hyperparameter κ^2 of the prior is not specified so κ^2 is unknown. What to do? Because we have i.i.d. over $j = 1, \cdots, d$, we can try to estimate κ^2 by $\|\boldsymbol{X}\|_2^2/(d-2)$, which reproduces the James-Stein shrinkage estimator. But this strategy of estimating hyperparameters in the prior from data has a much more profound impact – it becomes a new school of statistical philosophy – the Empirical Bayesianism. It also becomes a very powerful framework for dealing with model selection, multiple testing, and adaptive estimation in practice. Empirical Bayes is also the underlying philosophy of mixed-effect/random-effect models, multi-level modeling, and hierarchical Bayes[3]. In terms of machine learning, the above philosophy is essentially what meta or multi-task learning (learning a common model from multiple tasks) is trying to achieve. But meta or multi-task learning focuses more on the algorithmic aspects.

Theoretical analysis of empirical nonparametric Bayesian procedure is notoriously difficult. For example, in nonparametric Bayesian, people (such as Aad van der Vaart, Ismael Castillo) often prove certain theorems when the prior hyperparameters are tuned with searching, which is almost never used in practice. Empirical Bayesian is the way to go in practice, but few can theoretically show why it works in practice.

Along this line, James-Stein estimator has also been generalized to the many Poisson mean model in [**?**]. Robbins' estimator for many Poisson mean model is again inspired from the Bayesian interpretation (Tweedie's formula for Poisson):

$$\widetilde{\lambda}_j = (Y_j + 1)\frac{f_\Pi(Y_j + 1)}{f_\Pi(Y_j)} \tag{18}$$

where $f_\Pi(\cdot)$ is the marginal of Y_j when we choose a prior over $\lambda_j \sim \Pi$. Robbins simply used the empirical distribution in the data directly to replace the prior-dependent marginal:

$$\widehat{\lambda}_j = (Y_j + 1)\frac{N(Y_j + 1)}{N(Y_j)} \tag{19}$$

where $N(y)$ is the number of counts equal to y.

[2] Check on your own at least for once!

[3] Applied statisticians love these complex models before deep neural nets showed up. The most famous Bayesian statistician nowadays is Andrew Gelman. He basically called empirical Bayesian school as the Stanford School of Statistics. You can look up related topics in his blog.

2.4.2 More on Tweedie's formula

For normal, Tweedie's formula is telling us that the posterior mean is determined by the data and the marginal density of the data under a given prior Π. When the normal likelihood has variance s^2, Tweedie's formula (also including the posterior variance) becomes

$$\begin{aligned} \mathbb{E}_{\mu|X}[\mu|X=x] &= x + s^2 \frac{f'_{\Pi,s^2}(x)}{f_{\Pi,s^2}(x)} \\ \mathrm{var}_{\mu|X}[\mu|X=x] &= s^2 \left\{ 1 + s^2 \left(\frac{f''_{\Pi,s^2}(x)}{f_{\Pi,s^2}(x)} - \frac{f'_{\Pi,s^2}(x)^2}{f_{\Pi,s^2}(x)^2} \right) \right\}. \end{aligned} \tag{20}$$

One may ask the following question: What is the class of marginal distributions that can be represented as the convolution between a Gaussian density and an arbitrary prior probability measure? The answer is quite remarkable:

Theorem 24 (Guo, McQueen and Richardson 2020 [**?**])**.** *When the prior* Π *has a density* π*,*

$$f_{\pi,t}(x) = \int \varphi((x-\mu)/\sqrt{t})\pi(\mu)\mathrm{d}\mu$$

is a density if and only if $f_{\pi,t}$ *is a solution to heat equation*

$$\frac{\partial}{\partial t} f_{\pi,t}(x) = \frac{1}{2}\frac{\partial^2}{\partial x^2} f_{\pi,t}(x), t \geq 0, x \in \mathbb{R} \tag{21}$$

with the boundary condition $f_{\pi,0}(x) = \pi(x)$*.*

Why this is so remarkable? Think of the following question: if the prior π is not the true probability measure of μ, we will easily encounter model misspecification bias. But this heat equation characterization provides a foundation for nonparametric estimation of the posterior mean and variance of μ. First, trigonometric polynomials (sin and cos) are eigenfunctions of the heat equation and form an orthogonal basis system for representing $f_{\pi,t}$ so one can simply write $f_{\pi,t}$ as an infinite Fourier series. Estimating a function non-parametrically is estimating a truncated infinite series. This is exactly the route that [**?**] took.

2.5 Alternative criterion exists: A case-study in robust statistics

2.5.1 Median-of-Mean (MoM) estimators

Risk defined in the above form is not the only reasonable criterion. There might be other options. For example: let scalar random variables $X_1, \cdots, X_n \overset{iid}{\sim} \mathbb{P}$ with $\mathbb{E}X^2 < \infty$ and say $\mathrm{var}X < \sigma^2$. Here we do not assume $\mathbb{E}|X|^r < \infty$ for any $r > 2$ (heavy-tailed distribution). Our goal is to recover/estimate $\mu = \mathbb{E}X$. Under the usual squared L_2 risk, the sample average $\bar{X}$ is the "optimal" estimator (no other estimator beats it in terms of the speed at which the risk converges to 0 as n increases). But if we look at its concentration around μ, since we only assume $\mathbb{E}X^2 < \infty$, the best one can do is Chebyshev inequality

$$\mathbb{P}\left(\bar{X} - \mu > t\right) \leq \left(\frac{nt^2}{\sigma^2}\right)^{-1}.$$

One can actually show that this tail bound cannot be improved [**?**]. However, recall that when $X_1, \cdots, X_n \overset{iid}{\sim} N(\mu, \sigma^2)$, we have a much better concentration of $\bar{X}$ around μ:

$$\mathbb{P}\left(\bar{X} - \mu > t\right) \leq \exp\left\{-\frac{nt^2}{2\sigma^2}\right\}. \tag{22}$$

Then it makes perfect sense to use (22) as a criterion to evaluate estimators/decision procedures. One would simply ask the question, if it is possible to obtain such exponential concentration when $\mathbb{E}|X|^r < \infty$ is not assumed for $r > 2$. Interestingly, there exists such an estimator called "median-of-mean" estimator. The idea is plain and simple: divide the whole data into K equal-sized groups and compute the sample mean of each group $\bar{X}_k$ for $k = 1, \cdots, K$. So for each group-wise sample mean, we have by Chebyshev inequality: for each k

$$\mathbb{P}\left(\bar{X}_k - \mu > t\right) \leq \left(\frac{nt^2}{K\sigma^2}\right)^{-1}.$$

Choose $t = 2\sigma/\sqrt{n/K}$, then

$$\mathbb{P}\left(\bar{X}_k - \mu > 2\sigma/\sqrt{n/K}\right) \leq 1/4.$$

How to combine these sample averages? Let us try their median: $\text{median}\left(\bar{X}_1, \cdots, \bar{X}_K\right)$. In the analysis below, we use the definition of a median

$$\begin{aligned}
&\mathbb{P}\left\{\text{median}\left(\bar{X}_1, \cdots, \bar{X}_K\right) - \mu > 2\sigma/\sqrt{n/K}\right\} \\
&= \mathbb{P}\left\{\sum_{k=1}^{K} \mathbb{1}\{\bar{X}_k - \mu > 2\sigma/\sqrt{n/K}\} > \frac{K}{2}\right\} \\
&\leq \mathbb{P}\left(\text{Binom}(K, 1/4) > \frac{K}{2}\right) \\
&\leq \mathbb{P}\left(\sum_{k=1}^{K}(B_k - 1/4) > \frac{K}{4}\right) \quad B_k \sim \text{Bernoulli}(1/4) \\
&\overset{\star}{\leq} e^{-K/8}.
\end{aligned}$$

If we choose $K = \frac{n\epsilon^2}{4\sigma^2}$, then

$$\mathbb{P}\left\{\text{median}\left(\bar{X}_1, \cdots, \bar{X}_K\right) - \mu > \epsilon\right\} < e^{-n\epsilon^2/(32\sigma^2)}.$$

For $\star$, we use Hoeffding inequality:

Theorem 25 (Hoeffding inequality: General form)**.** *$X_1, \cdots, X_n$ are independent random variables with mean 0, and $X_i \in [a_i, b_i]$. Denote $S_n = \sum_{i=1}^n X_i$. Then for any $\lambda \geq 0$,*

$$\mathbb{E}e^{\lambda S_n} \leq e^{\lambda^2 \sum_{i=1}^n (b_i - a_i)^2/8}.$$

Then by Chernoff bound

$$\begin{aligned}
\mathbb{P}(S_n \geq t) &\leq \exp\left(-\frac{2t^2}{\sum_{i=1}^n (b_i - a_i)^2}\right) \\
\mathbb{P}(S_n \leq -t) &\leq \exp\left(-\frac{2t^2}{\sum_{i=1}^n (b_i - a_i)^2}\right)
\end{aligned}$$

Proof. For a random variable X bounded between $[a, b]$, its MGF has the following bound:

$$\mathbb{E}e^{\lambda(X-\mu)} \leq e^{\frac{\lambda^2}{8}(b-a)^2} \tag{23}$$

Then obviously when $X_1, \cdots, X_n$ have mean zero,

$$\mathbb{E}e^{\lambda S_n} \leq e^{\lambda^2 \sum_{i=1}^n (b_i-a_i)^2/8}$$

and the rest follows. We are left to prove (23). To this end, define $g(\lambda) = \log \mathbb{E}e^{\lambda X}$. Then

$$g'(\lambda) = \frac{\mathbb{E}Xe^{\lambda X}}{\mathbb{E}e^{\lambda X}}, g''(\lambda) = \frac{\mathbb{E}X^2e^{\lambda X}}{\mathbb{E}e^{\lambda X}} - \left(\frac{\mathbb{E}Xe^{\lambda X}}{\mathbb{E}e^{\lambda X}}\right)^2.$$

By Taylor expansion, for some $\epsilon \in (0, \lambda)$

$$\begin{aligned} g(\lambda) &= g(0) + g'(0)\lambda + \frac{\lambda^2}{2}g''(\epsilon) \\ &= 0 + 0 + \frac{\lambda^2}{2}g''(\epsilon) \\ &= \frac{\lambda^2}{2}\mathsf{var}_{\epsilon,*}(X) \leq \frac{\lambda^2}{8}(b-a)^2. \end{aligned}$$

In the last line, we use the fact for any random variable bounded between a and b, regardless of the underlying probability measure, we have

$$\mathsf{var}_*(X) = \mathbb{E}_*(X-\mu_*)^2 \leq \mathbb{E}_*(X - \frac{b+a}{2})^2 \leq \frac{1}{4}(b-a)^2.$$

□

The above idea can be extended from heavy-tailed scalar random variables to $\mathbb{R}^d$, by defining median in high dimension appropriately [**?** **?**].

2.5.2 Trimmed mean estimators

Another robust mean estimator is simply removing the extreme data points:

$$\widehat{\mu}_{trim} = \frac{1}{n}\sum_{i=1}^n X_i \mathbb{1}\{|X|_i \leq |X|_{(c\log(1/\delta)/n)}\} \tag{24}$$

where δ is the desired confidence level (i.e. $\mathbb{P}(\widehat{\mu}_{trim} - \mu \geq ...) \leq 1-\delta$). The analysis is somewhat more involved due to the involvement of order statistics. We will leave it to your own reading [**?**].

3 Hypothesis testing

Hypothesis testing is the simplest statistical problem but it still provides a very deep understanding on the difficulty of a statistical problem. Because studying estimation is often reduced to studying hypothesis testing, some theoretical-oriented statisticians simply study hypothesis testing problem

in their whole life. But on the other hand, "trained to reject null hypothesis" or simply "trained to reject applied people's finding" is one stigma attached to statisticians[4].

Yuri Ingster, a Russian mathematical statistician, is the grandmaster on hypothesis testing. His book [**?**] co-authored with his student contains almost everything you want to know about testing hypothesis and even more.

3.1 Minimax hypothesis testing

Let $O_1, \cdots, O_n \sim \mathbb{P}$ where $\mathbb{P} \in \mathcal{P}$. Recall that in general, the goal is to test

$$H_0 : \mathbb{P} = \mathbb{P}_0 \text{ vs. } H_a : \mathbb{P} \neq \mathbb{P}_0.$$

A valid level α test T_n is a measurable function of the data and α to $\{0, 1\}$ such that

$$\sup_{\mathbb{P} \sim H_0} \mathbb{P}^{\otimes n}(T_n = 1) \equiv \mathbb{P}_0^{\otimes n}(T_n = 1) \leq \alpha.$$

Minimax criterion for hypothesis testing is about the power, or the Type-II error:

$$\beta_n(\epsilon) = \inf_{T_n} \sup_{\mathbb{P} \in \mathcal{P}(\epsilon)} \mathbb{P}^{\otimes n}(T_n = 0) \tag{25}$$

where

$$\mathcal{P}(\epsilon) = \{\mathbb{P} \in \mathcal{P} : d(\mathbb{P}_0, \mathbb{P}) > \epsilon\} \tag{26}$$

with some distance measure d (e.g. metrics or divergences). Define the critical signal strength-/minimax separation rate $\epsilon_n(\delta)$ as

$$\epsilon_n(\delta) = \inf \{\epsilon : \beta_n(\epsilon) \leq \delta\}. \tag{27}$$

Remark 26. When the problem at hand is difficult, one can relax the above criterion and consider the minimax risk for hypothesis testing as follows:

$$R_n(\epsilon)^* = \inf_{T_n} \left\{ \mathbb{P}_0(T_n = 1) + \sup_{\mathbb{P} \in \mathcal{P}(\epsilon)} \mathbb{P}(T_n = 0) \right\}. \tag{28}$$

Test T_n is said to be asymptotically powerful if $\lim R_{T_n}(\epsilon) = 0$; asymptotic powerless if $\lim R_{T_n}(\epsilon) \geq 1$. The critical signal strength/minimax separation rate is the $\epsilon^\dagger$ such that if $\epsilon \ll \epsilon^\dagger$ $R_n(\epsilon)^* \to 1$ and if $\epsilon \gg \epsilon^\dagger$ $R_n(\epsilon)^* \to 0$.

3.2 Some classical results

Neyman-Pearson lemma is one of the most important classical results in statistics. It tells us for single vs. single hypothesis testing problems, likelihood ratio test (LRT) is the optimal α-level test (most powerful or smallest Type-II error).

[4] See Larry Wasserman's roundtable talk at the University of Chicago.

Theorem 27 (Neyman-Pearson lemma)**.** *$X \sim \mathbb{P} \in \mathcal{P}$. $H_0 : \mathbb{P} = \mathbb{P}_0$ vs. $H_a : \mathbb{P} = \mathbb{P}_1$. Then the following α-level test*

$$\widehat{T}_c = \mathbb{1}\left\{\frac{\mathrm{d}\mathbb{P}_1}{\mathrm{d}\mathbb{P}_0}(X) \geq c\right\} \tag{29}$$

with c chosen such that

$$\mathbb{P}_0(\widehat{T}_c = 1) = \alpha$$

is the most powerful test for H_0 vs. H_a.

Proof. Suppose there is another α-level test $\widehat{T}'$. The rejection region of $\widehat{T}_c$ (the sample space for which $\widehat{T}_c = 1$) is denoted as R. The rejection region of $\widehat{T}'$ is denoted as R'. Denote $R_1 = R \setminus R'$, $R_2 = R' \setminus R$, $R_3 = R' \cap R$. See Figure 3 below.

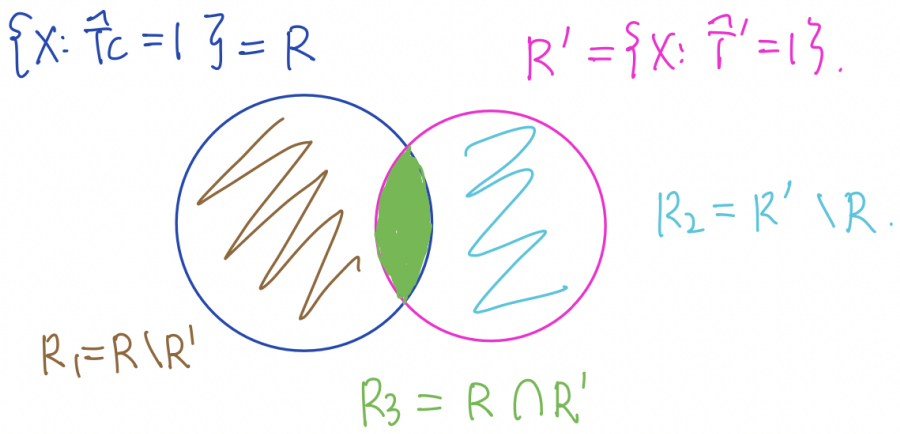

Figure 3: picture for Neyman-Pearson lemma

Then by the premise on $\widehat{T}_c$ and $\widehat{T}'$, we have

$$\begin{aligned} &\mathbb{P}_0(R) = \alpha \geq \mathbb{P}_0(R') \text{ by the level-}\alpha\text{-ness of } \widehat{T}' \\ \Rightarrow\ &\mathbb{P}_0(R_1) \geq \mathbb{P}_0(R_2) \text{ by the commonality of } R_3. \end{aligned}$$

Now we look at the power or $1 -$ Type II error: because $R_1 \subseteq R$ and $R_2 \not\subseteq R$, we have

$$\mathbb{P}_1(R_1) = \int_{R_1} \mathrm{d}\mathbb{P}_1 = \int_{R_1} \frac{\mathrm{d}\mathbb{P}_1}{\mathrm{d}\mathbb{P}_0}\mathrm{d}\mathbb{P}_0 \geq c \int_{R_1} \mathrm{d}\mathbb{P}_0 = c\mathbb{P}_0(R_1) \geq c\mathbb{P}_0(R_2) = c \int_{R_2} \mathrm{d}\mathbb{P}_0$$

and

$$c \int_{R_2} \mathrm{d}\mathbb{P}_0 = c \int_{R_2} \frac{\mathrm{d}\mathbb{P}_0}{\mathrm{d}\mathbb{P}_1}\mathrm{d}\mathbb{P}_1 \geq c \cdot c^{-1} \int_{R_2} \mathrm{d}\mathbb{P}_1 = \mathbb{P}_1(R_2).$$

So we conclude under H_a, $\mathbb{P}_1(R_1) \geq \mathbb{P}_1(R_2)$, and finally we add the common part R_3 back to the space, we have $\mathbb{P}_1(\widehat{T}_c = 1) \geq \mathbb{P}_1(\widehat{T}' = 1)$, i.e. $\widehat{T}_c$ is more powerful than any other level-α test. □

Neyman-Pearson lemma has been generalized in several directions: e.g. Karlin-Rubin theorem for monotone likelihood ratio classes (see [**?**]). For composite vs. composite testing problem, Neyman-Pearson becomes less useful. But when dealing with complicated hypothesis testing problems, people still often start with LRT anyway.

Another important theorem is the following Wilks' theorem. We will provide a heuristic argument and in the Chapter of M-estimation, you will be able to derive the results more rigorously.

Theorem 28 (Wilks' theorem)**.** *$X_1, \cdots, X_n \overset{iid}{\sim} \mathbb{P}_\theta$ for $\theta \in \Theta$. $H_0 : \theta \in \Theta_0$ and $H_a : \theta \in \Theta_1 = \Theta \setminus \Theta_0$. We also assume all sorts of nice regularity conditions on Θ_0 and Θ_1 and the MLE $\widehat{\theta}_n$. Then define the LRT with n i.i.d. data as follows:*

$$\Lambda_n = \frac{\sup_{\theta\in\Theta} \mathrm{d}\mathbb{P}_\theta^{\otimes n}(X_1, \cdots, X_n)}{\sup_{\theta\in\Theta_0} \mathrm{d}\mathbb{P}_\theta^{\otimes n}(X_1, \cdots, X_n)}.$$

As $n \to \infty$, we have

$$2\log\Lambda_n \xrightarrow{d} \chi^2_{dim(\Theta)-dim(\Theta_0)}. \tag{30}$$

Heuristic arguments. Let us only consider the simplest scenario: $\Theta = \mathbb{R}$ (so the ambient dimension is 1) and $\Theta_0 = \{\theta_0\}$ so the dimension for H_0 is 0. We expect the log-likelihood ratio test statistic converges to χ^2_1. We write down $2\log\Lambda_n$ explicitly below: by the i.i.d. assumption, and recall that $\widehat{\theta}_n$ is the MLE so $\sup_{\theta\in\Theta} \mathrm{d}\mathbb{P}_\theta^{\otimes n}(X_1, \cdots, X_n) = \mathrm{d}\mathbb{P}_{\widehat{\theta}_n}^{\otimes n}(X_1, \cdots, X_n)$

$$2\log\Lambda_n = 2\left\{\sum_{i=1}^n \text{log-likelihood}(X_i; \widehat{\theta}_n) - \sum_{i=1}^n \text{log-likelihood}(X_i; \theta_0)\right\}$$

Then we do Taylor expansion: define $\ell(\theta) = \sum_{i=1}^n \text{log-likelihood}(X_i; \theta)$, so $\ell'(\theta)$ is the score function

$$2\log\Lambda_n = 2(\widehat{\theta}_n - \theta)\ell'(\widehat{\theta}_n) + (\widehat{\theta}_n - \theta_0)^2\ell'(\theta_n^*).$$

But recall that $\widehat{\theta}_n$ is the MLE, when the log-likehood function is differentiable, we have $\ell'(\widehat{\theta}_n) = 0$. So that leaves us

$$2\log\Lambda_n = (\sqrt{n}(\widehat{\theta}_n - \theta_0))^2 \frac{\ell'(\theta_n^*)}{n}.$$

For MLE, we usually have, again under regularity conditions and when H_0 is true,

$$\sqrt{n}(\widehat{\theta}_n - \theta_0) \xrightarrow{d} N(0, \sigma^2)$$

and here σ^2 is the inverse Fisher information, i.e. the expectation of the Hessian (second-derivative) of the log-likelihood function. But $\frac{\ell'(\theta_n^*)}{n}$ is very close to a sample average, so if weak law of large number can be used, we have $\frac{\ell'(\theta_n^*)}{n} \to \sigma^{-2}$ in $\mathbb{P}_0$-probability. Combining the above heuristic arguments and Slutsky's theorem, we argue

$$2\log\Lambda_n \xrightarrow{d} N(0, \sigma^2)^2\sigma^{-2} \sim N(0,1)^2 \sim \chi^2_1.$$

□

In homework, we will derive Wilks' theorem for some simple distributions for you to get a sense what is actually going on.

Remark 29. Wilks' theorem has been generalized in many directions: high dimension or irregular case. In the above formulation, we have assumed the dimension d is not diverging as $n \to \infty$. In fact, if we let $d \to \infty$ but $d = o(n^{1/6})$, this theorem still holds. However, when $d/n \to c \in (0, 1)$, $2 \log \Lambda_n$ no longer converges to a chi-square distribution [**?**]. The proof technique will involve approximate message passing (AMP) originated in statistical physics and spin glass theory and Random Matrix Theory (RMT). We will cover AMP during this semester.

3.3 f-divergences

Before we really talk about hypothesis testing, let us take a detour and introduce another set of tools from information theory – f-divergences – that will be very important for deriving lower bounds (lower bounds are essentially about information-theoretical limit of a statistical problem).

f-divergences are a class of divergence measures between probability measures, generalizing the KL divergence. One may prefer different f-divergences depending on the application contexts to ease computation. f-divergences are very important in quantifying the difficulty.

Definition 30. f-divergence between two probability measures $\mathbb{P}$ and $\mathbb{Q}$ satisfying $\mathbb{P} \ll \mathbb{Q}$[5] are defined as

$$D_f(\mathbb{P}||\mathbb{Q}) = \mathbb{E}_{\mathbb{Q}} f\left(\mathrm{d}\mathbb{P}/\mathrm{d}\mathbb{Q}\right) \tag{31}$$

where $f : (0, \infty) \to \mathbb{R}$ is convex and strictly convex at 1, and $f(1) = 0$. We also define the following convention to handle singularities:

$$f(0) = \lim_{x \downarrow 0} f(x), 0f\left(\frac{0}{0}\right) = 0.$$

Remark 31. In fact, $\mathbb{P} \ll \mathbb{Q}$ is not necessary. We can always find another probability measure μ (e.g. $\mu = (P + Q)/2$) such that $\mathbb{P} \ll \mu$ and $\mathbb{Q} \ll \mu$ and redefine $D_f(\mathbb{P}||\mathbb{Q})$ as

$$D_f(\mathbb{P}||\mathbb{Q}) = \int \frac{\mathrm{d}\mathbb{Q}}{\mathrm{d}\mu} f\left(\frac{\mathrm{d}\mathbb{P}/\mathrm{d}\mu}{\mathrm{d}\mathbb{Q}/\mathrm{d}\mu}\right).$$

But in this course, we ignore this subtlety.

f-divergences that we might encounter in this course:

name	f	formula
KL divergence	$f(x) = x \log x$	$D_{KL}(\mathbb{P}\|\|\mathbb{Q}) = \mathbb{E}_{\mathbb{Q}}\left[\frac{\mathrm{d}\mathbb{P}}{\mathrm{d}\mathbb{Q}} \log\left(\frac{\mathrm{d}\mathbb{P}}{\mathrm{d}\mathbb{Q}}\right)\right] = \int \mathrm{d}\mathbb{P} \log\left(\frac{\mathrm{d}\mathbb{P}}{\mathrm{d}\mathbb{Q}}\right)$
Total variation (TV) distance	$f(x) = \frac{1}{2}\lvert 1 - x\rvert$	$d_{TV}(\mathbb{P}, \mathbb{Q}) = \frac{1}{2}\mathbb{E}_{\mathbb{Q}}\left[\left\lvert 1 - \frac{\mathrm{d}\mathbb{P}}{\mathrm{d}\mathbb{Q}}\right\rvert\right] = \frac{1}{2}\int \lvert \mathrm{d}\mathbb{P} - \mathrm{d}\mathbb{Q}\rvert$
χ^2 divergence	$f(x) = (1 - x)^2$	$\chi^2(\mathbb{P}\|\|\mathbb{Q}) = \mathbb{E}_{\mathbb{Q}}\left[\left(1 - \frac{\mathrm{d}\mathbb{P}}{\mathrm{d}\mathbb{Q}}\right)^2\right] = \int \frac{(\mathrm{d}\mathbb{P} - \mathrm{d}\mathbb{Q})^2}{\mathrm{d}\mathbb{Q}} = \int \frac{(\mathrm{d}\mathbb{P})^2}{\mathrm{d}\mathbb{Q}} - 1$
squared Hellinger distance	$f(x) = (1 - \sqrt{x})^2$	$H^2(\mathbb{P}, \mathbb{Q}) = \mathbb{E}_{\mathbb{Q}}\left[\left(1 - \sqrt{\frac{\mathrm{d}\mathbb{P}}{\mathrm{d}\mathbb{Q}}}\right)^2\right] = \int \left(\sqrt{\mathrm{d}\mathbb{P}} - \sqrt{\mathrm{d}\mathbb{Q}}\right)^2$

[5] P dominated by $\mathbb{Q}$

Remark 32. Mutual information vs. KL divergence:

$$I(X;Y) = D_{KL}(\mathbb{P}_{X,Y}||\mathbb{P}_X \otimes \mathbb{P}_Y).$$

Theorem 33 (Key properties of f-divergences)**.**

1. *(Monotonicity)* $D_f(\mathbb{P}_{X,Y}||\mathbb{Q}_{X,Y}) \geq D_f(\mathbb{P}_X||\mathbb{Q}_X)$.
2. *(Data Processing Inequality) Suppose* $\mathbb{P}_Y$ *is the marginal distribution of* Y *of the joint* $\mathbb{P}_{Y,X} = \mathbb{P}_X\mathbb{P}_{Y|X}$ *and* $\mathbb{Q}_Y$ *is the marginal distribution of* Y *of the joint* $\mathbb{Q}_{Y,X} = \mathbb{Q}_X\mathbb{P}_{Y|X}$*, then*

   $$D_f(\mathbb{P}_X||\mathbb{Q}_X) \geq D_f(\mathbb{P}_Y||\mathbb{Q}_Y),$$

   that is, processing blurs the difference between the sources.
3. $D_f(\mathbb{P}||\mathbb{Q}) \geq 0$ *and "=" holds if and only if* $\mathbb{P} = \mathbb{Q}$ *a.s.*
4. $D_f(\mathbb{P}||\mathbb{Q})$ *is jointly convex in both* $\mathbb{P}$ *and* $\mathbb{Q}$ *arguments*
5. *Given two joint* $\mathbb{P}_{X,Y} = \mathbb{P}_X\mathbb{P}_{Y|X}$ *and* $\mathbb{Q}_{X,Y} = \mathbb{Q}_X\mathbb{P}_{Y|X}$ *with different marginal distributions on* X *but the same conditional law of* Y *given* X*, we have* $D_f(\mathbb{P}_{X,Y}||\mathbb{Q}_{X,Y}) = D_f(\mathbb{P}_X||\mathbb{Q}_X)$.
6. *Define the conditional* f*-divergence:*

   $$D_f(\mathbb{P}_{Y|X}||\mathbb{Q}_{Y|X} \mid X) := \mathbb{E}_{X\sim\mathbb{P}_X}\left[D_f(\mathbb{P}_{Y|X}||\mathbb{Q}_{Y|X})\right]. \tag{32}$$

   Suppose $\mathbb{P}_Y$ *is the marginal distribution of* Y *of the joint* $\mathbb{P}_{Y,X} = \mathbb{P}_X\mathbb{P}_{Y|X}$ *and* $\mathbb{Q}_Y$ *is the marginal distribution of* Y *of the joint* $\mathbb{Q}_{Y,X} = \mathbb{P}_X\mathbb{Q}_{Y|X}$*, then*

   $$D_f(\mathbb{P}_{Y|X}||\mathbb{Q}_{Y|X} \mid X) \geq D_f(\mathbb{P}_Y||\mathbb{Q}_Y),$$

   that is, conditioning increases divergences.

Proof.

1. By Jensen inequality

$$\begin{aligned} D_f(\mathbb{P}_{X,Y}||\mathbb{Q}_{X,Y}) &= \mathbb{E}_{X\sim\mathbb{Q}_X}\left[\mathbb{E}_{Y\sim\mathbb{Q}_{Y|X}} f\left(\frac{\mathrm{d}\mathbb{P}_{X,Y}}{\mathrm{d}\mathbb{Q}_{X,Y}}\right)\right] \\ &\geq \mathbb{E}_{X\sim\mathbb{Q}_X}\left[f\left(\mathbb{E}_{Y\sim\mathbb{Q}_{Y|X}}\frac{\mathrm{d}\mathbb{P}_{X,Y}}{\mathrm{d}\mathbb{Q}_{X,Y}}\right)\right] \\ &= \mathbb{E}_{X\sim\mathbb{Q}_X}\left[f\left(\frac{\mathrm{d}\mathbb{Q}_X}{\mathrm{d}\mathbb{Q}_X}\right)\right] = D_f(\mathbb{P}_X||\mathbb{Q}_X). \end{aligned}$$

2.

$$\begin{aligned} D_f(\mathbb{P}_X||\mathbb{Q}_X) &= \int \mathrm{d}\mathbb{Q}_X f\left(\frac{\mathrm{d}\mathbb{P}_X}{\mathrm{d}\mathbb{Q}_X}\right) \\ &= \int_X\int_Y \mathrm{d}\mathbb{Q}_{X,Y} f\left(\frac{\mathrm{d}\mathbb{P}_{X,Y}}{\mathrm{d}\mathbb{Q}_{X,Y}}\right) \text{ by Theorem 33.5} \\ &\geq \int_Y \mathrm{d}\mathbb{Q}_Y f\left(\frac{\mathrm{d}\mathbb{P}_Y}{\mathrm{d}\mathbb{Q}_Y}\right) = D_f(\mathbb{P}_Y||\mathbb{Q}_Y) \text{ by monotonicity.} \end{aligned}$$

3. The first statement again follows from Jensen inequality. For the second statement, suppose there exists $\mathbb{P} \neq \mathbb{Q}$ but still $D_f(\mathbb{P}||\mathbb{Q}) = 0$. There exists a measurable set A such that $p := \mathbb{P}(A) \neq \mathbb{Q}(A) =: q > 0$. Take $Y = \mathbb{1}\{X \in A\}$, then by DPI, $D_f(\mathbb{P}||\mathbb{Q}) \geq D_f(\text{Bern}(p)||\text{Bern}(q)) = 0$. Since $\mathbb{P} \ll \mathbb{Q}$, we must have $0 < \mathbb{Q}(A) < 1$. Therefore

$$0 = D_f(\text{Bern}(p)||\text{Bern}(q)) = qf\left(\frac{p}{q}\right) + (1-q)f\left(\frac{1-p}{1-q}\right).$$

So we can find ρ, x, x' such that $\rho f(x) + (1-\rho)f(x') = 0$ and $\rho x + (1-\rho)x' = 1$ so f is not strictly convex at 1, a contradiction.

4. Define a mapping $g(a,b) : (a,b) \mapsto bf(a/b)$. We can compute the Hessian of $g(a,b)$ and check it is positive semi-definite, which implies the joint convexity of $D_f(\mathbb{P}||\mathbb{Q})$.

5. $$\begin{aligned} D_f(\mathbb{P}_{X,Y}||\mathbb{Q}_{X,Y}) &= \int_X \int_Y \mathrm{d}\mathbb{Q}_{X,Y} f\left(\frac{\mathrm{d}\mathbb{P}_{X,Y}}{\mathrm{d}\mathbb{Q}_{X,Y}}\right) \\ &= \int_X \int_Y \mathrm{d}\mathbb{Q}_X \mathrm{d}\mathbb{P}_{Y|X} f\left(\frac{\mathrm{d}\mathbb{P}_X \mathrm{d}\mathbb{P}_{Y|X}}{\mathrm{d}\mathbb{Q}_X \mathrm{d}\mathbb{P}_{Y|X}}\right) \\ &= \int_X \mathrm{d}\mathbb{Q}_X \underbrace{\int_Y \mathrm{d}\mathbb{P}_{Y|X}}_{\equiv 1} f\left(\frac{\mathrm{d}\mathbb{P}_X}{\mathrm{d}\mathbb{Q}_X}\right) \\ &= D_f(\mathbb{P}_X||\mathbb{Q}_X). \end{aligned}$$

6. By the joint convexity of $D_f(\mathbb{P}||\mathbb{Q})$:

$$\begin{aligned} D_f(\mathbb{P}_{Y|X}||\mathbb{Q}_{Y|X} \mid X) &:= \mathbb{E}_{X\sim\mathbb{P}_X}\left[D_f(\mathbb{P}_{Y|X}||\mathbb{Q}_{Y|X})\right] \\ &\geq D_f\left(\mathbb{E}_{X\sim\mathbb{P}_X}\mathbb{P}_{Y|X}||\mathbb{E}_{X\sim\mathbb{P}_X}\mathbb{Q}_{Y|X}\right) \\ &= D_f(\mathbb{P}_Y||\mathbb{Q}_Y). \end{aligned}$$

□

Remark 34. For deeper results of f-divergence, a very good resource will be Yihong Wu and Yury Polyanskiy's lecture notes on "Information-Theoretic Methods for High-Dimensional Statistics". The structure of their notes is very similar to a large part of our course, but we emphasize much less on information theory. But be cautious while reading their notes as they were scribed by students and might contain numerous typos/errors.

There is also a strong connection between f-divergence and sampling. See the famous f-GAN paper: https://arxiv.org/abs/1606.00709.

Theorem 35 (Connection between d_{TV} and hypothesis testing; Scheffé's theorem)**.** *Suppose $X \sim \mathbb{P}'$ and we want to test $H_0 : \mathbb{P}' = \mathbb{P}$ vs. $H_a : \mathbb{P}' = \mathbb{Q}$*

$$d_{TV}(\mathbb{P},\mathbb{Q}) = 1 - \int (\mathrm{d}\mathbb{P} \wedge \mathrm{d}\mathbb{Q});$$

$$d_{TV}(\mathbb{P},\mathbb{Q}) = \sup_A \int_A \mathrm{d}\mathbb{P} - \mathrm{d}\mathbb{Q} \equiv \sup_A \mathbb{P}(A) - \mathbb{Q}(A);$$

and

$$d_{TV}(\mathbb{P},\mathbb{Q}) = 1 - \inf_{T:\mathbb{X}\to\{0,1\}} \left(\underbrace{\mathbb{P}(T=1)}_{\textit{Type-I error}} + \underbrace{\mathbb{Q}(T=0)}_{\textit{Type-II error}} \right).$$

Proof. The first statement follows from the definition.

For the second statement, consider the measurable set $A = \{x : \mathbb{P}(x) > \mathbb{Q}(x)\}$. Then

$$\begin{aligned} d_{TV}(\mathbb{P},\mathbb{Q}) &= \frac{1}{2}\int |\mathrm{d}\mathbb{P} - \mathrm{d}\mathbb{Q}| \\ &= \frac{1}{2}\int_A |\mathrm{d}\mathbb{P} - \mathrm{d}\mathbb{Q}| + \frac{1}{2}\int_{A^c} |\mathrm{d}\mathbb{P} - \mathrm{d}\mathbb{Q}| \\ &= \frac{1}{2}\int_A \mathrm{d}\mathbb{P} - \mathrm{d}\mathbb{Q} + \frac{1}{2}\int_{A^c} \mathrm{d}\mathbb{Q} - \mathrm{d}\mathbb{P} \\ &\le \sup_A \int_A \mathrm{d}\mathbb{P} - \mathrm{d}\mathbb{Q}. \end{aligned}$$

Next, for any measurable A', we have

$$\begin{aligned} & \left| \int_{A'} \mathrm{d}\mathbb{P} - \mathrm{d}\mathbb{Q} \right| \\ =\ & \max\left\{ \int_{A'} \mathrm{d}\mathbb{P} - \mathrm{d}\mathbb{Q}, \int_{A'} \mathrm{d}\mathbb{Q} - \mathrm{d}\mathbb{P} \right\} \\ \le\ & \max\left\{ \int_{A'\cap A} \mathrm{d}\mathbb{P} - \mathrm{d}\mathbb{Q}, \int_{A'\cap A^c} \mathrm{d}\mathbb{Q} - \mathrm{d}\mathbb{P} \right\} \\ \le\ & \max\left\{ \int_{A} \mathrm{d}\mathbb{P} - \mathrm{d}\mathbb{Q}, \int_{A^c} \mathrm{d}\mathbb{Q} - \mathrm{d}\mathbb{P} \right\} \\ =\ & \int_A \mathbb{P} - \mathbb{Q} = \frac{1}{2}\int |\mathrm{d}\mathbb{P} - \mathrm{d}\mathbb{Q}| = d_{TV}(\mathbb{P},\mathbb{Q}). \end{aligned}$$

The third statement follows from the first statement combined with Neyman-Pearson lemma. You will complete the proof in your homework. □

Now we look at n i.i.d. data setting. So $X_1, \cdots, X_n \overset{iid}{\sim} \mathbb{P}'$ and we still want to test $H_0 : \mathbb{P}' = \mathbb{P}$ and $H_1 : \mathbb{P}' = \mathbb{Q}$. Now the sum of Type-I error and Type-II error for some test statistic T_n becomes:

$$\inf_{T_n:\mathbb{X}^n\to\{0,1\}} \underbrace{\mathbb{P}^{\otimes n}(T_n=1)}_{\text{Type-I error}} + \underbrace{\mathbb{Q}^{\otimes n}(T_n=0)}_{\text{Type-II error}}.$$

By Theorem 35, we know, if assuming that probability measures $\mathbb{P}$ and $\mathbb{Q}$ have densities p and q

$$\inf_{T_n:\mathbb{X}^n\to\{0,1\}} \mathbb{P}^{\otimes n}(T_n=1) + \mathbb{Q}^{\otimes n}(T_n=0) = 1 - d_{TV}(\mathbb{P}^{\otimes n},\mathbb{Q}^{\otimes n}) = \int (\mathrm{d}\mathbb{P}^{\otimes n} \wedge \mathrm{d}\mathbb{Q}^{\otimes n})$$

$$= \int \left(\prod_{i=1}^{n} p(x_i) \wedge \prod_{i=1}^{n} q(x_i) \right) \mathrm{d}x_1 \cdots \mathrm{d}x_n.$$

So an exact computation of d_{TV} is difficult to carry out for product measures. This is one of the reasons why we need some many f-divergences. In fact, we can sandwich d_{TV} by squared Hellinger distances as follows

Lemma 36 (Le Cam's inequality[6])**.**

$$0 \leq \frac{1}{2} H^2(\mathbb{P}, \mathbb{Q}) \leq d_{TV}(\mathbb{P}, \mathbb{Q}) \leq H(\mathbb{P}, \mathbb{Q}) \left(1 - \frac{H^2(\mathbb{P}, \mathbb{Q})}{4} \right)^{1/2} \leq 1.$$

Proof. For the first inequality, notice:

$$\frac{1}{2} H^2(\mathbb{P}, \mathbb{Q}) = 1 - \int \sqrt{\mathrm{d}\mathbb{P}} \sqrt{\mathrm{d}\mathbb{Q}} \leq 1 - \int \mathrm{d}\mathbb{P} \wedge \mathrm{d}\mathbb{Q} \equiv d_{TV}(\mathbb{P}, \mathbb{Q}).$$

For the second inequality, notice:

$$\begin{aligned}
& H^2(\mathbb{P}, \mathbb{Q}) \left(1 - \frac{H^2(\mathbb{P}, \mathbb{Q})}{4} \right) \\
&= 2 \left(1 - \int \sqrt{\mathrm{d}\mathbb{P}} \sqrt{\mathrm{d}\mathbb{Q}} \right) \left[1 - \frac{2}{4} \left(1 - \int \sqrt{\mathrm{d}\mathbb{P}} \sqrt{\mathrm{d}\mathbb{Q}} \right) \right] \\
&= \left(1 - \int \sqrt{\mathrm{d}\mathbb{P}} \sqrt{\mathrm{d}\mathbb{Q}} \right) \left(1 + \int \sqrt{\mathrm{d}\mathbb{P}} \sqrt{\mathrm{d}\mathbb{Q}} \right) \\
&= 1 - \left(\int \sqrt{\mathrm{d}\mathbb{P}} \sqrt{\mathrm{d}\mathbb{Q}} \right)^2 \\
&\equiv 1 - \left(\int \sqrt{\mathrm{d}\mathbb{P} \wedge \mathrm{d}\mathbb{Q}} \sqrt{\mathrm{d}\mathbb{P} \vee \mathrm{d}\mathbb{Q}} \right)^2 \\
&\geq 1 - \int \mathrm{d}\mathbb{P} \wedge \mathrm{d}\mathbb{Q} \int \mathrm{d}\mathbb{P} \vee \mathrm{d}\mathbb{Q} \text{ [by Cauchy-Schwarz]} \\
&= 1 - \int \mathrm{d}\mathbb{P} \wedge \mathrm{d}\mathbb{Q} \left[2 - \int \mathrm{d}\mathbb{P} \wedge \mathrm{d}\mathbb{Q} \right] \text{ [because } \int \mathrm{d}\mathbb{P} \vee \mathrm{d}\mathbb{Q} + \int \mathrm{d}\mathbb{P} \wedge \mathrm{d}\mathbb{Q} = 2 \text{]} \\
&= \left(1 - \int \mathrm{d}\mathbb{P} \wedge \mathrm{d}\mathbb{Q} \right)^2 \equiv d_{TV}(\mathbb{P}, \mathbb{Q})^2.
\end{aligned}$$

□

We can also compare other f-divergences with KL divergence in the following results.

Lemma 37.

$$H^2(\mathbb{P}, \mathbb{Q}) \leq D_{KL}(\mathbb{P} || \mathbb{Q}).$$

[6] This is the first time we encounter results attributed to Lucien Le Cam. His name will show up many times in the M-estimation and MLE chapter.

Proof. When $x > -1$, we have $-\log(1+x) \geq -x$. So

$$\begin{aligned} D_{KL}(\mathbb{P}||\mathbb{Q}) &= \int \mathrm{d}\mathbb{P} \log\left(\frac{\mathrm{d}\mathbb{P}}{\mathrm{d}\mathbb{Q}}\right) = 2\int \mathrm{d}\mathbb{P}\left(-\log\sqrt{\frac{\mathrm{d}\mathbb{Q}}{\mathrm{d}\mathbb{P}}}\right) \\ &= 2\int \mathrm{d}\mathbb{P}\left\{-\log\left[\left(\sqrt{\frac{\mathrm{d}\mathbb{Q}}{\mathrm{d}\mathbb{P}}}-1\right)+1\right]\right\} \\ &\geq 2\int \mathrm{d}\mathbb{P}\left[\left(\sqrt{\frac{\mathrm{d}\mathbb{Q}}{\mathrm{d}\mathbb{P}}}-1\right)\right] \\ &= 2\left(\int \sqrt{\mathrm{d}\mathbb{Q}}\sqrt{\mathrm{d}\mathbb{P}} - 1\right) \equiv H^2(\mathbb{P},\mathbb{Q}). \end{aligned}$$

□

Lemma 38 (Pinsker's inequality[7])**.**

$$d_{TV}(\mathbb{P},\mathbb{Q}) \leq \sqrt{D_{KL}(\mathbb{P}||\mathbb{Q})/2} \wedge \left(1 - \frac{1}{2}\exp\left\{-D_{KL}(\mathbb{P}||\mathbb{Q})\right\}\right)$$

Proof. We first prove the first part. Define a function $g(x) = x\log x - x + 1$ for $x \geq 0$ and $f(x) = (x-1)^2 - \left(\frac{4}{3} + \frac{2}{3}x\right)g(x)$ (which can be proved by Taylor expansion). So $f(x) \leq 0$ for all $x \geq 0$.

$$\begin{aligned} d_{TV}(\mathbb{P},\mathbb{Q}) &= \frac{1}{2}\int |\mathrm{d}\mathbb{P} - \mathrm{d}\mathbb{Q}| = \frac{1}{2}\int \left|\frac{\mathrm{d}\mathbb{P}}{\mathrm{d}\mathbb{Q}} - 1\right| \mathrm{d}\mathbb{Q} \\ &\leq \frac{1}{2}\int \mathrm{d}\mathbb{Q}\sqrt{\left(\frac{4}{3} + \frac{2}{3}\frac{\mathrm{d}\mathbb{P}}{\mathrm{d}\mathbb{Q}}\right) g\left(\frac{\mathrm{d}\mathbb{P}}{\mathrm{d}\mathbb{Q}}\right)} \\ &\leq \frac{1}{2}\sqrt{\int \mathrm{d}\mathbb{Q}\left(\frac{4}{3} + \frac{2}{3}\frac{\mathrm{d}\mathbb{P}}{\mathrm{d}\mathbb{Q}}\right)}\sqrt{\int \mathrm{d}\mathbb{Q} g\left(\frac{\mathrm{d}\mathbb{P}}{\mathrm{d}\mathbb{Q}}\right)} \\ &\leq \frac{1}{2}\sqrt{\int \left(\frac{4}{3}\mathrm{d}\mathbb{Q} + \frac{2}{3}\mathrm{d}\mathbb{P}\right)}\sqrt{\int \mathrm{d}\mathbb{Q}\left[\frac{\mathrm{d}\mathbb{P}}{\mathrm{d}\mathbb{Q}}\log\left(\frac{\mathrm{d}\mathbb{P}}{\mathrm{d}\mathbb{Q}}\right) - \frac{\mathrm{d}\mathbb{P}}{\mathrm{d}\mathbb{Q}} + 1\right]} \\ &= \frac{1}{\sqrt{2}}\sqrt{\int \mathrm{d}\mathbb{Q}\frac{\mathrm{d}\mathbb{P}}{\mathrm{d}\mathbb{Q}}\log\left(\frac{\mathrm{d}\mathbb{P}}{\mathrm{d}\mathbb{Q}}\right)} \\ &= \sqrt{D_{KL}(\mathbb{P}||\mathbb{Q})/2}. \end{aligned}$$

The second part is relatively easy:

$$\begin{aligned} \int \mathrm{d}\mathbb{P} \wedge \mathrm{d}\mathbb{Q} &\geq \frac{1}{2}\left(\int \sqrt{\mathrm{d}\mathbb{P}}\sqrt{\mathrm{d}\mathbb{Q}}\right)^2 \\ &= \frac{1}{2}\exp\left(2\log\int \sqrt{\mathrm{d}\mathbb{P}}\sqrt{\mathrm{d}\mathbb{Q}}\right) \end{aligned}$$

[7] You might also encounter Pinsker's constant in the future. It is the optimal constant in front of minimax risk convergence rate. Pinsker's constant is almost hopeless to obtain in most problems.

$$
\begin{aligned}
&= \frac{1}{2}\exp\left(2\log\int \mathrm{d}\mathbb{P}\sqrt{\frac{\mathrm{d}\mathbb{Q}}{\mathrm{d}\mathbb{P}}}\right)\\
&\geq \frac{1}{2}\exp\left(2\int \mathrm{d}\mathbb{P}\log\sqrt{\frac{\mathrm{d}\mathbb{Q}}{\mathrm{d}\mathbb{P}}}\right)\\
&= \frac{1}{2}\exp\left(-\int \mathrm{d}\mathbb{P}\log\frac{\mathrm{d}\mathbb{P}}{\mathrm{d}\mathbb{Q}}\right)\\
&= \frac{1}{2}\exp\left(-D_{KL}(\mathbb{P}||\mathbb{Q})\right).
\end{aligned}
$$

□

Lemma 39.

$$D_{KL}(\mathbb{P}||\mathbb{Q}) \leq \log\left(1+\chi^2(\mathbb{P}||\mathbb{Q})\right) \leq \chi^2(\mathbb{P}||\mathbb{Q}).$$

Next, we observe that Hellinger distance of product measures tensorizes in a very general form:

Lemma 40. *Given two sequences of probability measures $\{\mathbb{P}_n\}, \{\mathbb{Q}_n\}$ so the probability measure for each data can depend on the index n (think of high dimensional random vectors where the dimension grows with the sample size; so does the probability measure for each random vector). Then*

$$H^2(\mathbb{P}_n^{\otimes n}, \mathbb{Q}_n^{\otimes n}) = 2 - 2\left(1 - \frac{1}{2}H^2(\mathbb{P}_n, \mathbb{Q}_n)\right)^n \tag{33}$$

or more generally

$$H^2\left(\bigotimes_{i=1}^n \mathbb{P}_{i,n}, \bigotimes_{i=1}^n \mathbb{Q}_{i,n}\right) = 2 - 2\prod_{i=1}^n\left(1 - \frac{1}{2}H^2(\mathbb{P}_{i,n}, \mathbb{Q}_{i,n})\right). \tag{34}$$

Proof.

$$
\begin{aligned}
H^2(\mathbb{P}_n^{\otimes n}, \mathbb{Q}_n^{\otimes n}) &= 2 - 2\underbrace{\int\cdots\int}_{\text{repeat } n \text{ times}}\sqrt{\mathrm{d}\mathbb{P}_n^{\otimes n}}\sqrt{\mathrm{d}\mathbb{Q}_n^{\otimes n}}\\
&= 2 - 2\underbrace{\int\sqrt{\mathrm{d}\mathbb{P}_n}\sqrt{\mathrm{d}\mathbb{Q}_n}\cdots\int\sqrt{\mathrm{d}\mathbb{P}_n}\sqrt{\mathrm{d}\mathbb{Q}_n}}_{\text{again, repeat } n \text{ times}}\\
&= 2 - 2\left(\int\sqrt{\mathrm{d}\mathbb{P}_n}\sqrt{\mathrm{d}\mathbb{Q}_n}\right)^n.
\end{aligned}
$$

Finally, recall that $H^2(\mathbb{P}_n, \mathbb{Q}_n) = 2 - 2\int\sqrt{\mathrm{d}\mathbb{P}_n}\sqrt{\mathrm{d}\mathbb{Q}_n}$. □

Now let us go back to the hypothesis testing problem when we observe n i.i.d. data points. We can now use the squared Hellinger distance as a surrogate for total variation distance because squared Hellinger distance tensorizes, making the analysis much simpler.

Theorem 41 (Non-asymptotic version of Theorem 42)**.** *For any $0 < \delta < 1/2$, for any test statistic with testing error not surpassing δ, to distinguish between $\mathbb{P}$ and $\mathbb{Q}$ with $H^2(\mathbb{P}, \mathbb{Q}) \leq 1$, we need at least*

$$\frac{1}{H^2(\mathbb{P}, \mathbb{Q})} \log\left(\frac{1}{\delta}\right)$$

many independent samples.

Proof sketch. Applying Lemma 36 and Lemma 40, together with the observation that

$$1 - \frac{1}{2}x \geq e^{-x}, \text{ if } 0 < x < 1.$$

□

Theorem 42. *Given two sequences of probability measures $\{\mathbb{P}_n\}, \{\mathbb{Q}_n\}$. As $n \to \infty$,*

$$\begin{aligned} d_{TV}(\mathbb{P}_n^{\otimes n}, \mathbb{Q}_n^{\otimes n}) \to 0 \Leftrightarrow H^2(\mathbb{P}_n, \mathbb{Q}_n) = o\left(\frac{1}{n}\right), \\ d_{TV}(\mathbb{P}_n^{\otimes n}, \mathbb{Q}_n^{\otimes n}) \to 1 \Leftrightarrow H^2(\mathbb{P}_n, \mathbb{Q}_n) = \omega\left(\frac{1}{n}\right). \end{aligned} \tag{35}$$

Proof sketch. A direct consequence of the above theorem. The conclusion also follows by noticing $\left(1 - \frac{C}{n}\right)^n \to e^{-C}$ as $n \to \infty$. □

Theorem 42 entails the following "intuition" any statisticians should always bear in mind: In general, for parametric statistical models (i.e. the probability measure $\mathbb{P}_n \equiv \mathbb{P}$ invariant to n), if the difference between two hypotheses is around or below order $1/\sqrt{n}$, the two hypotheses are essentially hard to distinguish.

In words, computing the Hellinger distance between probability measures for one data point suffices to tell if the optimal test is asymptotically powerful or powerless.

Similar results hold for χ^2-divergence, which will be used in the example we describe next.

Lemma 43.

1. $2d_{TV}(\mathbb{P}, \mathbb{Q}) \leq \sqrt{\chi^2(\mathbb{P}||\mathbb{Q})}$
2. $\chi^2(\mathbb{P}_n^{\otimes n}||\mathbb{Q}_n^{\otimes n}) = \left(1 + \chi^2(\mathbb{P}_n||\mathbb{Q}_n)\right)^n - 1$, *or more generally,*

$$\chi^2\left(\bigotimes_{i=1}^n \mathbb{P}_{i,n} || \bigotimes_{i=1}^n \mathbb{Q}_{i,n}\right) = \prod_{i=1}^n \left(1 + \chi^2(\mathbb{P}_{i,n}||\mathbb{Q}_{i,n})\right) - 1.$$

In words, χ^2-divergence being $O(1)$ provides a sufficient condition for the testing risk to be close to 0. Apart from tensorization, χ^2-divergence can be handy when one compares a mixture distribution with a single probability measure or two mixture distributions.

Lemma 44 (χ^2-divergence of mixtures)**.** *Given a class of probability measures parameterized by* Θ*:* $\{\mathbb{P}_\theta, \theta \in \Theta\}$*. Define the mixture distribution mixed over a prior* Π *as*

$$\mathbb{P}_\Pi \coloneqq \int_{\theta\in\Theta} \mathbb{P}_\theta \mathrm{d}\Pi(\theta).$$

Then

$$\chi^2(\mathbb{P}_\Pi||\mathbb{Q}) = \mathbb{E}_{\theta,\theta' \overset{iid}{\sim} \Pi}\left[\int \frac{\mathrm{d}\mathbb{P}_\theta \mathrm{d}\mathbb{P}_{\theta'}}{\mathrm{d}\mathbb{Q}}\right] - 1. \tag{36}$$

Proof.

$$\begin{aligned}
\chi^2(\mathbb{P}_\Pi||\mathbb{Q}) &= \int \frac{(\mathrm{d}\mathbb{P}_\Pi)^2}{\mathrm{d}\mathbb{Q}} - 1 \\
&= \int \frac{\left(\int_{\theta\in\Theta} \mathrm{d}\mathbb{P}_\theta \mathrm{d}\Pi(\theta)\right)^2}{\mathrm{d}\mathbb{Q}} - 1 \\
&= \int \frac{\int_{\theta\in\Theta}\int_{\theta'\in\Theta} \mathrm{d}\mathbb{P}_\theta \mathrm{d}\mathbb{P}_{\theta'} \mathrm{d}\Pi(\theta)\mathrm{d}\Pi(\theta')}{\mathrm{d}\mathbb{Q}} - 1 \text{ "replica trick"} \\
&= \int_{\theta\in\Theta}\int_{\theta'\in\Theta} \left\{\int \frac{\mathrm{d}\mathbb{P}_\theta \mathrm{d}\mathbb{P}_{\theta'}}{\mathrm{d}\mathbb{Q}}\right\} \mathrm{d}\Pi(\theta)\mathrm{d}\Pi(\theta') - 1 \text{ Fubini} \\
&\equiv \mathbb{E}_{\theta,\theta' \overset{iid}{\sim} \Pi}\left[\int \frac{\mathrm{d}\mathbb{P}_\theta \mathrm{d}\mathbb{P}_{\theta'}}{\mathrm{d}\mathbb{Q}}\right] - 1.
\end{aligned}$$

□

In general, χ^2-divergence between two mixtures does not have nice closed-form formula, and is usually bounded by Hellinger distance between two mixtures. See e.g. [**?**] for Hellinger distance between two mixtures.

3.4 A non-trivial single parametric testing example: Erdös-Renyi Random Graph vs. Stochastic Block Models (SBM)

Consider two different random graph models

- One is the famous Erdös-Renyi (ER) Random Graph $G(n, r)$, r being the probability that any two vertices out of n total vertices connected by an edge.
- The other is the SBM $\mathrm{SBM}(n, p, q)$, where the graph G can be decomposed into two clusters '-1' and '$+1$'. The membership of each vertex is an unbiased Rademacher random variable $\varepsilon \sim \mathrm{Rad}(1/2)$ over the cluster labels. Within clusters, the probability that any two vertices connected by an edge is p; across clusters, the probability that any two vertices connected by an edge is q. Let us focus on the case where $r = (p+q)/2$ and bounded degree graph i.e. $p = a/n$ and $q = b/n$ for some $a, b = \Theta(1)$.

We are interested in the following testing problem:

$$H_0 : G \sim G(n, r) \text{ vs. } H_a : G \sim \mathrm{SBM}(n, p, q).$$

We will partially prove the following theorem:

Theorem 45. *There is a phase transition: as $n \to \infty$*

- *If* $\frac{(a-b)^2}{2(a+b)} \le 1$,
$$d_{TV}(H_0, H_a) < 1 - \Omega(1);$$
- *If* $\frac{(a-b)^2}{2(a+b)} > 1$,
$$d_{TV}(H_0, H_a) \to 1.$$

Proof. In the proof, we do not consider the critical threshold $\frac{(a-b)^2}{2(a+b)} = 1$ as it is way too technical.

It suffices to show $\chi^2(H_0||H_a) = O(1)$. For a graph G, denote its corresponding adjacency matrix as $A_{n\times n}$. For a random graph G, A is a random matrix. Under H_0, the probability measure over A is simply

$$\mathbb{P}_{A,0} = \bigotimes_{1\le i<j\le n} \mathbb{P}_{\mathrm{Bern}(r)} = \bigotimes_{1\le i<j\le n} \frac{\mathbb{P}+\mathbb{Q}}{2}, \quad r = \frac{p+q}{2}$$

where $\mathbb{P}$ and $\mathbb{Q}$ are the short-hand notation for $\mathrm{Bern}(p)$ and $\mathrm{Bern}(q)$.

Under H_a, conditioning on the memberships of all vertices $\boldsymbol{\varepsilon} = \{\varepsilon_i; i = 1, \cdots, n\}$ the probability measure over A is a mixture distribution

$$\begin{aligned}
\mathbb{P}_{A,a,\boldsymbol{\varepsilon}} &= \bigotimes_{1\le i<j\le n} \left(\mathbb{P}_{\mathrm{Bern}(p)} \mathbb{1}\{\varepsilon_i = \varepsilon_j\} + \mathbb{P}_{\mathrm{Bern}(q)} \mathbb{1}\{\varepsilon_i \ne \varepsilon_j\}\right) \\
&= \bigotimes_{1\le i<j\le n} \left(\mathbb{P}_{\mathrm{Bern}(r)} + \mathbb{P}_{\mathrm{Bern}(t=\frac{p-q}{2})} \varepsilon_i\varepsilon_j\right) \\
&= \bigotimes_{1\le i<j\le n} \left(\frac{\mathbb{P}+\mathbb{Q}}{2} + \frac{\mathbb{P}-\mathbb{Q}}{2}\varepsilon_i\varepsilon_j\right).
\end{aligned}$$

To apply Lemma 44, we first compute

$$\begin{aligned}
&\int \frac{\mathrm{d}\mathbb{P}_{A,a,\boldsymbol{\varepsilon}} \mathrm{d}\mathbb{P}_{A,a,\boldsymbol{\varepsilon}'}}{\mathrm{d}\mathbb{P}_{A,0}} \\
&= \prod_{1\le i<j\le n} \int \frac{\mathrm{d}\left(\frac{\mathbb{P}+\mathbb{Q}}{2} + \frac{\mathbb{P}-\mathbb{Q}}{2}\varepsilon_i\varepsilon_j\right) \mathrm{d}\left(\frac{\mathbb{P}+\mathbb{Q}}{2} + \frac{\mathbb{P}-\mathbb{Q}}{2}\varepsilon'_i\varepsilon'_j\right)}{\mathrm{d}\frac{\mathbb{P}+\mathbb{Q}}{2}} \\
&= \prod_{1\le i<j\le n} \int \mathrm{d}\frac{\mathbb{P}+\mathbb{Q}}{2} + \mathrm{d}\frac{\mathbb{P}-\mathbb{Q}}{2}\varepsilon'_i\varepsilon'_j + \mathrm{d}\frac{\mathbb{P}-\mathbb{Q}}{2}\varepsilon_i\varepsilon_j + \frac{\left(\mathrm{d}\frac{\mathbb{P}-\mathbb{Q}}{2}\right)^2}{\mathrm{d}\frac{\mathbb{P}+\mathbb{Q}}{2}}\varepsilon_i\varepsilon_j\varepsilon'_i\varepsilon'_j \\
&= \prod_{1\le i<j\le n} \left[1 + \underbrace{\int \frac{(\mathrm{d}\mathbb{P} - \mathrm{d}\mathbb{Q})^2}{2(\mathrm{d}\mathbb{P} + \mathrm{d}\mathbb{Q})}}_{=:\varrho} \varepsilon_i\varepsilon_j\varepsilon'_i\varepsilon'_j\right] \\
&= \prod_{1\le i<j\le n} \left[1 + \varrho\varepsilon_i\varepsilon_j\varepsilon'_i\varepsilon'_j\right] \\
&\le \prod_{1\le i<j\le n} \exp\{\varrho\varepsilon_i\varepsilon_j\varepsilon'_i\varepsilon'_j\}
\end{aligned}$$

$$= \exp\left\{\varrho \sum_{1\le i<j\le n} \varepsilon_i\varepsilon_j\varepsilon'_i\varepsilon'_j\right\} \le \exp\left\{\frac{1}{2}\varrho\left(\boldsymbol{\varepsilon}^\top\boldsymbol{\varepsilon}'\right)^2\right\}.$$

Then

$$\chi^2(H_0||H_a) \le \mathbb{E}_{\boldsymbol{\varepsilon},\boldsymbol{\varepsilon}'}\left[\exp\left\{\frac{1}{2}\varrho\left(\boldsymbol{\varepsilon}^\top\boldsymbol{\varepsilon}'\right)^2\right\}\right] - 1.$$

Here ϱ is easy to calculate:

$$\varrho = \frac{(p-q)^2}{2(p+q)} + \frac{(p-q)^2}{2(2-p-q)} = \frac{1}{n}\left(\frac{(a-b)^2}{2(a+b)} + o(1)\right).$$

Now we can in turn bound the χ^2-divergence as follows:

$$\begin{aligned}
\chi^2(H_0||H_a) &\le \mathbb{E}_{\boldsymbol{\varepsilon},\boldsymbol{\varepsilon}'}\left[\exp\left\{\frac{1}{2}\left[\frac{(a-b)^2}{2(a+b)} + o(1)\right]\left(\frac{1}{\sqrt{n}}\boldsymbol{\varepsilon}^\top\boldsymbol{\varepsilon}'\right)^2\right\}\right] - 1 \\
&\le \mathbb{E}_{\boldsymbol{\varepsilon},\boldsymbol{\varepsilon}'}\left[\exp\left\{\frac{1}{2}\left[\frac{(a-b)^2}{2(a+b)} + o(1)\right]\left(\underbrace{\frac{1}{\sqrt{n}}\sum_{i=1}^n \varepsilon_i\varepsilon_{i'}}_{\xrightarrow{d} N(0,1)\sim Z}\right)^2\right\}\right] - 1 \\
&\to \underbrace{\mathbb{E}\left[\exp\left\{\frac{1}{2}\left[\frac{(a-b)^2}{2(a+b)} + o(1)\right]\chi_1^2\right\}\right]}_{\text{MGF of } \chi_1^2} - 1 \\
&= \begin{cases} +\infty & \dfrac{(a-b)^2}{2(a+b)} > 1, \\ \dfrac{1}{\sqrt{1-\frac{(a-b)^2}{2(a+b)}}} & \dfrac{(a-b)^2}{2(a+b)} < 1. \end{cases}
\end{aligned}$$

□

Remark 46. In actual proof, we also need to show the existence of a test. But due to time limit, this part will not be covered in this course. If you are interested, the algorithm that tightly achieves the information-theoretical limit given in the above theorem is proved in [**?**], by Elchanan Mossel, Joe Neeman, and Alan Sly in 2012. Their algorithm is based on short-cycle counting or k-cycle counting. In particular, they showed that under H_0, the number of k-cycles in G should be close to $\mathsf{Pois}\left(\frac{1}{k}\left(\frac{a+b}{2}\right)^k\right)$, whereas under H_a, the number of k-cycles in G should be close to $\mathsf{Pois}\left(\frac{1}{k}\left(\frac{a+b}{2}\right)^k + \frac{1}{k}\left(\frac{a-b}{2}\right)^k\right)$. By comparing the number of k-cycles taking into account the variance of these k-cycles, the upper bound matches the hardness threshold. In fact, [**?**] were inspired by the non-rigorous yet deep and insightful statistical physics calculations done by [**?**] in 2011, using the so-called "replica symmetric cavity method", or "Belief Propagation (BP)", or "message passing". Essentially this is a technique that, by believing that long-range interactions

do not matter, uses mean-field limit plus some short-range interactions to approximate the whole thermodynamic system (in SBM case, the graph). In fact, BP is exact for trees and close to being exact for tree-like graphs. "Methods from statistical physics" is one of the hottest topic in statistics and machine learning in recent years. You are strongly recommended to read papers and listen to some related lectures given by Florent Krzakala, Lenka Zdeborová, David Gamarnik, and Andrea Montanari.

3.4.1 A nonparametric example: Uniformity testing on $[0,1]$

For $X \in \mathbb{P}([0,1])$ where $\mathbb{P}$ has absolutely continuous density f, we want to test

$$H_0 : F(t) = t \ \forall t \in [0,1] \text{ vs. } H_a : F(t) \neq t \ \exists t \in [0,1].$$

Recall that for minimax testing problem, we only consider the alternative class to be sufficiently far away from the null (otherwise it is impossible to obtain uniform result). We have the following theorem:

Theorem 47.

$$H_0 : F(t) = t \ \forall t \in [0,1] \text{ vs. } H_a(r_n) : F \in \left\{ F : \sup_{t\in[0,1]} |F(t) - t| \gtrsim r_n \right\}$$

Then the minimax separation rate for testing the above hypothesis is $p_n \asymp \frac{1}{\sqrt{n}}$, *in the following sense: there exists a test statistic* T_n : *i.i.d.* $X_1, \cdots, X_n \mapsto \{0,1\}$ *such that for any* $\alpha > 0$

$$\begin{aligned} &\mathbb{P}_{H_0}^{\otimes n}(T_n = 1) + \sup_{F \in H_a(p_n)} \mathbb{P}_F^{\otimes n}(T_n = 0) \leq \alpha \\ &\liminf_n \inf_{\widetilde{T}_n} \mathbb{P}_{H_0}^{\otimes n}(\widetilde{T}_n = 1) + \sup_{F \in H_a(r_n)} \mathbb{P}_F^{\otimes n}(\widetilde{T}_n = 0) > 0, \ \textit{if } r_n = o(p_n) \end{aligned} \tag{37}$$

Remark 48. During the lecture, I rushed through a lot of details. Please see the proof below.

Proof. Since the alternative is a composite hypothesis, for the lower bound, we need to exhibit a "worst-case" instance in the alternative to show the hardness of the problem. This is a general scheme for lower bound proofs. First, denote $f_0 = 1$, the p.d.f. under H_0. We define a p.d.f. f_1 as follows:

$$f_1 = f_0 + \Delta_n \tag{38}$$

where $\Delta_n = r_n \Delta$ with Δ any bounded function supported in $[0,1]$ such that $\int_0^1 \Delta(x)\mathrm{d}x = 0$ and $\int_0^1 \Delta(x)^2 \mathrm{d}x = 1$. Then obviously f_1 is a p.d.f. with support $[0,1]$ and the total variation distance between f_0 and f_1 is, from the second representation of total variation distance given in Theorem 35,

$$r_n \sup_t \left| \int_0^t \Delta(x)\mathrm{d}x \right| \asymp r_n.$$

So $F_1 \in H_a(r_n)$.

Now let us again compute $\chi^2(H_a||H_0)$ with product measures:

$$\chi^2(\mathbb{P}_1^{\otimes n}||\mathbb{P}_0^{\otimes n}) = \left(1 + \chi^2(\mathbb{P}_1||\mathbb{P}_0\right)^n - 1$$

$$= \left(1 + \int \frac{f_1^2}{f_0} - 1\right)^n - 1$$
$$= \left(\int \frac{(f_0 + r_n\Delta)^2}{f_0}\right)^n - 1$$
$$= \left(1 + r_n^2 \int \Delta^2\right)^n - 1 \le e^{r_n^2 n} - 1.$$

So $r_n = o(1/\sqrt{n})$ suggests non-separation (i.e. $\chi^2(\mathbb{P}_1^{\otimes n} || \mathbb{P}_0^{\otimes n}) \to 0$).[8]

For the achievability when $r_n \asymp p_n$, we need to exhibit a test. The test statistic is quite natural – it is the famous Kolmogorov-Smirnov goodness-of-fit test:

$$T_n = \mathbb{1}\left\{\sqrt{n} \sup_{t\in[0,1]} |F_n(t) - t| > z_\alpha\right\} \tag{39}$$

where $F_n(t) = \frac{1}{n}\sum_{i=1}^n \mathbb{1}_{(0,t]}(X_i)$ is the empirical c.d.f., and z_α is the upper α-quantile of the distribution of $\sup_{t\in[0,1]} |G(t)|$ where G is the standard Brownian bridge. This part looks quite involved because of that supremum in the test statistic. But later in this course, we will actually show the indicator function belongs to the so-called Donsker class (the class functions for which Donsker's theorem holds). One can also develop a non-asymptotic test by using Dvoretzky-Kiefer-Wolfowitz inequality:

$$\mathbb{P}\left(\sqrt{n}\|F_n - F\|_\infty \ge t\right) \le 2e^{-2t^2} \tag{40}$$

so choose t such that $2e^{-2t^2} = \alpha$. □

3.5 Property testing

Property testing is a general class of problems originated from theoretical computer science literature. Initially, computer scientists (e.g. Andrew Yao) worked on computer-aided proof checker by testing certain clauses in the proof, which gradually evolves into testing properties. Property testing concerns problem like testing uniformity, identity, monotonicity, junta (i.e. sparse Boolean functions) [**?**], etc.

Property testing is closely related to hypothesis testing in statistics, but the probabilistic models are usually discrete or combinatorial and the requirement on the test statistic can be somewhat relaxed because it is not concerned with life and death situation as real-life statistics/biostatistics do. For a comprehensive survey, please take a look at [**? ?**].

3.5.1 Uniformity testing

Uniformity testing for discrete distributions is an important primitive for property testing. Many problems, e.g. identity testing, can be broken into simpler components which are themselves uniformity testing. You can look at related discussions in [**?**].

Consider a multinomial distribution $\mathsf{Multinom}(n, p_1, \cdots, p_k)$. In theoretical computer science community, they are interested in the following question: Can we find a test statistic such that

- If $H_0 : p_1 = \cdots = p_k = 1/k$, not to reject H_0 with probability at least 3/4.

[8] The blue part was rushed through in the lecture.

- If $H_a(r_{n,k}) : 2d_{TV}(\mathsf{Multinom}(1, p_1, \cdots, p_k), \mathsf{Multinom}(1, 1/k, \cdots, 1/k)) > r_{n,k}$, reject H_0 with probability at least 3 / 4.

Note that the choice of 3 / 4 is for convenience and it can be any number between $(1/2, 1)$. As you can see, such requirement is somewhat relaxed compared to hypothesis testing in statistics.

Then we have the following lower bound result:

Theorem 49 (Hardness of discrete uniformity test)**.** *The minimax separation rate is at least* $r_{n,k} = \Omega\left(\frac{k^{1/4}}{n^{1/2}}\right)$*; or in other words, to test a distribution away from uniform with distance* r*, we need at least* $\Omega(\sqrt{k}/r^2)$ *i.i.d. samples.*

Proof. Again we want to construct a worst-case instance that is just close enough to uniformity. But let us focus on the Poissonized version to avoid handling dependencies of multinomial distributions. We create a random variable $N \sim \mathsf{Pois}(n)$. Under Poissonization, the number of samples in category j is denoted as $N_j \sim \mathsf{Pois}(np_j = n/k)$ under H_0 uniformity. We denote this distribution as $U_n[k]$. So $H_0 : U_1[k]$.

Now let us construct the worst-case instance from non-uniform multinomials. For simplicity, take k to be an even integer so $k/2$ is still an integer. We follow a similar idea to the perturbation conducted in the continuous case: Generate $\epsilon_1, \cdots, \epsilon_{k/2} \overset{iid}{\sim} \text{Rademacher}(1/2)$. Then for $j = 1, \cdots, k/2$,

$$N_{2j-1} \sim \mathsf{Pois}\left(\tfrac{n}{k}(1 + Cr\epsilon_j)\right), N_{2j} \sim \mathsf{Pois}\left(\tfrac{n}{k}(1 - Cr\epsilon_j)\right)$$

For convenience, we denote this distribution after marginalizing over the randomness of ϵ's as $M_n[k]$. To simplify a bit, we have

$$U_n[k](N_1 = n_1, \cdots, N_k = n_k) = \prod_{j=1}^{k} \frac{(n/k)^{n_j} e^{-n/k}}{n_j!} = \frac{(n/k)^{\sum_{j=1}^k n_j} e^{-n}}{n_1! \cdots n_k!}$$

and

$$\begin{aligned}
&M_n[k](N_1 = n_1, \cdots, N_k = n_k) \\
&= \frac{1}{2^{k/2}} \sum_{(\epsilon_1, \cdots, \epsilon_{k/2}) \in \{-1,+1\}^{k/2}} M_n[k](N_1 = n_1, \cdots, N_k = n_k | \epsilon_1, \cdots, \epsilon_{k/2}) \\
&\text{where } M_n[k](N_1 = n_1, \cdots, N_k = n_k | \epsilon_1, \cdots, \epsilon_{k/2}) \\
&= \prod_{j=1}^{k/2} \frac{(n(1 + Cr\epsilon_j)/k)^{n_{2j-1}} e^{-n(1+Cr\epsilon_j)/k}}{n_{2j-1}!} \frac{(n(1 - Cr\epsilon_j)/k)^{n_{2j}} e^{-n(1-Cr\epsilon_j)/k}}{n_{2j}!} \\
&= \frac{(n/k)^{\sum_{j=1}^k n_j} e^{-n}}{n_1! \cdots n_k!} \prod_{j=1}^{k/2} (1 + Cr\epsilon_j)^{n_{2j-1}} (1 - Cr\epsilon_j)^{n_{2j}}
\end{aligned}$$

so

$$M_n[k](N_1 = n_1, \cdots, N_k = n_k)$$

$$= \frac{(n/k)^{\sum_{j=1}^k n_j} e^{-n}}{n_1! \cdots n_k!} \prod_{j=1}^{k/2} \left\{ \frac{1}{2} (1+Cr)^{n_{2j-1}} (1-Cr)^{n_{2j}} + \frac{1}{2} (1-Cr)^{n_{2j-1}} (1+Cr)^{n_{2j}} \right\}$$

By construction, $d_{TV}(U_1[k], M_1[k]) = Cr$[9], which means $M_1[k] \in H_a(r)$. To figure out the testability, we compute the chi-square divergence between $M_n[k]$ and $U_n[k]$:

$$\begin{aligned}
&\chi^2 \left(M_n[k] || U_n[k] \right) \\
&= \mathbb{E}_{M_n[k]} \left[\frac{\mathrm{d}M_n[k]}{\mathrm{d}U_n[k]} \right] - 1 \\
&= \mathbb{E}_{M_n[k]} \left[\prod_{j=1}^{k/2} \left\{ \frac{1}{2} (1+Cr)^{N_{2j-1}} (1-Cr)^{N_{2j}} + \frac{1}{2} (1-Cr)^{N_{2j-1}} (1+Cr)^{N_{2j}} \right\} \right] - 1 \\
&= \prod_{j=1}^{k/2} \mathbb{E}_{M_n[k]} \left[\left\{ \frac{1}{2} (1+Cr)^{N_{2j-1}} (1-Cr)^{N_{2j}} + \frac{1}{2} (1-Cr)^{N_{2j-1}} (1+Cr)^{N_{2j}} \right\} \right] - 1 \\
&= \prod_{j=1}^{k/2} \frac{1}{2}\frac{1}{2} \sum_{\epsilon \in \{-1,+1\}} \mathbb{E}_{M_n[k]|\epsilon} \left[(1+Cr)^{N_1} (1-Cr)^{N_2} + (1-Cr)^{N_1} (1+Cr)^{N_2} \,|\epsilon \right] - 1 \\
&= \prod_{j=1}^{k/2} \frac{1}{4} \left(\begin{array}{c} G_{\frac{n(1+Cr)}{k}}(1+Cr) G_{\frac{n(1-Cr)}{k}}(1-Cr) + G_{\frac{n(1+Cr)}{k}}(1-Cr) G_{\frac{n(1-Cr)}{k}}(1+Cr) \\ + G_{\frac{n(1-Cr)}{k}}(1+Cr) G_{\frac{n(1+Cr)}{k}}(1-Cr) + G_{\frac{n(1-Cr)}{k}}(1-Cr) G_{\frac{n(1+Cr)}{k}}(1+Cr) \end{array} \right) - 1 \\
&= \prod_{j=1}^{k/2} \frac{1}{4} \left(2e^{\frac{n(1+Cr)Cr}{k}} e^{-\frac{n(1-Cr)Cr}{k}} + 2e^{-\frac{n(1+Cr)Cr}{k}} e^{\frac{n(1-Cr)Cr}{k}} \right) - 1 \\
&= \left(\frac{e^{\frac{2n(Cr)^2}{k}} + e^{-\frac{2n(Cr)^2}{k}}}{2} \right)^{k/2} - 1 \\
&\leq \left(e^{\frac{4n^2(Cr)^4}{k^2}} \right)^{k/2} - 1 = e^{\frac{2n^2(Cr)^4}{k}} - 1
\end{aligned}$$

where $G_\lambda(s) = \mathbb{E}_{X \sim \mathsf{Pois}(\lambda)}(s^X)$ is the probability generating function of a Poisson distribution with mean λ which equals $e^{\lambda(s-1)}$ and the last inequality follows from $e^x + e^{-x} \leq 2e^{x^2/2}$. So we can take $n^2 r^4 / k = \Omega(1)$ to get $n = \Omega\left(\frac{\sqrt{k}}{r^2}\right)$.

At last, we perform de-Poissonization as in lecture1.pdf. □

Remark 50.

1. What is the intuition for sample complexity being $\sqrt{k}$, which means on average, each category has $1/\sqrt{k}$ samples, a tiny number? This is based on an intuition from the famous birthday

[9] Here $n = 1$. First, conditioning on the random signs, for each category,

$$d_{\mathsf{TV}}\left(\text{multinom}(1, p_1, \cdots, p_k), \text{multinom}(1, k^{-1}, \cdots, k^{-1})\right) = \Theta(r).$$

problem. In its essence, it says the following:

$$\mathbb{P}\left(\text{no collision}\right) = \prod_{i=1}^{n}\left(1-\frac{i-1}{k}\right) \le e^{-\sum_{i=1}^{n}\frac{i-1}{k}} \asymp e^{-\frac{n^2}{k}}.$$

by "no collision", we mean no two samples belong to the same category. When $\frac{n}{k} = o(1)$, we also have a corresponding lower bound:

$$\mathbb{P}\left(\text{no collision}\right) = \prod_{i=1}^{n}\left(1-\frac{i-1}{k}\right) \ge e^{-(1+o(1))(\sum_{i=1}^{n}\frac{i-1}{k})} \asymp e^{-\frac{n^2}{k}}.$$

Therefore, when the upper and lower bounds match, we have an exact formula for "no collision" probability $e^{-n^2/k}$ and if $n^2/k = O(1)$ i.e. $n = O(\sqrt{k})$. Only if this probability is higher than some constant, we can distinguish between $U_1[k]$ versus $U_1[k/2]$, whose TV distance is 0.5, quite large. Otherwise we only see a lot of singletons and hence cannot tell them apart. There is actually a way to construct "worst-case" instance by constructing $U_n[k/2]$ with additional random Rademacher multipliers. But we will not go into the detail on this alternative construction.

2. Uniformity testing is a primitive problem in property testing. Many other testing problems can be reduced to uniformity testing. For example, in "identity testing" ($H_0 : D_1 = D_2$ vs. $H_a : d(D_1, D_2) > r$), it is often broken into small pieces of uniformity testing.

3. Property testing has also been extended to the quantum case: see [**?**].

3.6 Some final comments on hypothesis testing

Suppose you just refuse to agree that hypothesis testing can be useful in data analysis. Then why bother to study hypothesis testing problems?

As we mentioned in the beginning of this chapter, there is a natural ordering among different statistical problems. For example, hypothesis testing is "easier" than estimation. So there is a natural reduction from estimation to hypothesis testing. If you are interested in estimation problems, hypothesis testing is almost an unavoidable intermediate step. We will use such reduction quite often in future lectures. Moreover, the type of thinking in hypothesis testing is strongly related to many concepts in other fields, such as machine learning and cryptography. For example:

3.6.1 Hypothesis testing and differential privacy

The concept of "Differential Privacy" (DP) is invented by Cynthia Dwork [**?**], who started her career as a cryptographer. The definition is as follows:

Definition 51. Let $\epsilon > 0$. A randomized algorithm $\mathsf{Alg} : \mathcal{X} \mapsto \mathcal{P}(\mathcal{Y})$, where $\mathcal{X}$ is the input space, $\mathcal{Y}$ is the output space and $\mathcal{P}(\mathcal{Y})$ is the space of probability measures on the output space $\mathcal{Y}$, is ϵ-DP if for every two datasets D_0 and D_1 that differs only by one individual and for every set $S \subseteq \mathcal{Y}$, we have

$$\mathbb{P}\left(\mathsf{Alg}(D_1) \in S\right) \le e^{\epsilon}\mathbb{P}\left(\mathsf{Alg}(D_0) \in S\right) \tag{41}$$

uniformly.

Remark 52. This is a very stringent requirement! I will copy the following comments from Larry Wasserman's lecture note on DP here (which might be a bit obsolete):

Strengths of DP:

1. DP gives a very rigorous, precise notion of privacy.
2. Many methods in machine learning and statistics can be made differentially private.
3. DP can be used for other purposes. For example, Dwork et al. 2015 [**?**] (a Science paper!) created a method called *reusable holdout* that allows an interactive approach to data analysis while making repeated looks at the data without introducing too much bias. The heart of the method is to impose a sort of differential privacy on each step of the analysis.

Weaknesses of DP:

1. DP has dominated the research in privacy. It seems that there is not much research in other approaches.
2. DP is very strong. You need to add a lot of noise to the data.
3. When there is a structure in the data, such as voids, manifolds etc., it is destroyed by DP.
4. I have not seen it really used in much practical data analysis.

Almost every statistical or machine learning problems can be extended to their corresponding DP versions. DP is related to hypothesis testing in the following sense:

Theorem 53 (Wasserman and Zhou 2010 [**?**])**.** *Denote $y \in \mathcal{Y}$ as an output of an ϵ-DP randomized algorithm* Alg*. Consider the following hypothesis testing problem:*

$$H_0 : y \text{ comes from } \mathsf{Alg}(D_0) \quad \text{vs.} \quad H_a : y \text{ comes from } \mathsf{Alg}(D_1).$$

For any rejection region $S \subseteq \mathcal{Y}$ of the above testing problem:

$$\text{if: } \underbrace{\mathbb{P}(S \ni \mathsf{Alg}(D_0))}_{\text{Type I error}} \leq \alpha,$$

$$\text{then: } \underbrace{\mathbb{P}(S \not\ni \mathsf{Alg}(D_1))}_{\text{Type II error}} \geq 1 - e^{\epsilon}\alpha.$$

The proof is a trivial application of Definition 51 so omitted.

3.6.2 Hypothesis testing, model selection, and adaptivity

Whenever you are doing model selection, you are essentially asking a hypothesis testing question: which model (hypothesis) cannot be rejected by the data?

An interesting direction in recent years is adaptive/interactive data analysis: the goal is to propose a data-analytic method that is always statistically rigorous and valid but allows users to perform all kinds of crazy analysis on data. In fact, DP has shown to be a very good strategy for adaptive data analysis [**?**].

In later lectures, we will encounter a (theoretical[10]) model selection method called "Lepskii's method", which is essentially a multiple testing procedure.

[10] This is because real data application never uses Lepskii's method.