# Notes on intermediate statistics

Aodong Li

January 4 - 20, 2018

**This document only serves as notes on intermediate statistics 36-705 CMU.**

## 1 Probability Review

*Random variable* is a *function* that maps the outcomes in sample space into $\mathbb{R}$.

Conditional expectation is a *random variable*.

An important theorem is *iterated expectations*

$$E[E[f(X)|Y]] = E[f(X)].$$

The *moment generating function* $E[e^{tX}]$ completely characterizes a random variable. It means that if $E[e^{tX}] = E[e^{tY}]$, then $X$ and $Y$ have the same distribution.

To generate a moment,

$$E[X^k] = \frac{d^k E[e^{tX}]}{dt^k}|_{t=0}.$$

**Exercise 3.16.** Show that $E[r(X)s(Y)|X] = r(X)E[s(Y)|X]$ and that $E[r(X)|X] = r(X)$.

$$\begin{aligned} h(x) &= E[r(X)s(Y)|X = x] \\ &= E[r(x)s(Y)|X = x] \\ &= r(x)E[s(Y)|X = x] \\ h(X) &= r(X)E[s(Y)|X] \end{aligned}$$

$$\begin{aligned} f(x) &= E[r(X)|X = x] \\ &= E[r(x)|X = x] \\ &= r(x) \\ f(X) &= r(X) \end{aligned}$$

**Exercise 3.18.** Show that if $E[X|Y = y] = c$ for some constant $c$, then $X$ and $Y$ are uncorrelated.

$$\begin{aligned}
Cov(XY) &= E[XY] - E[X]E[Y] \\
&= E[E[XY|Y]] - E[E[X|Y]]E[Y] \quad \text{(by iterated expectations)} \\
&= E[YE[X|Y]] - E[E[X|Y]]E[Y] \quad \text{(by exercise 3.16)} \\
&= cE[Y] - cE[Y] \\
&= 0
\end{aligned}$$

**Exercise Moments generating function.** Show that if $X_1, ..., X_n \sim N(\mu, \sigma^2)$, then $\bar{X}_n \sim N(\mu, \sigma^2/n)$.

$$\begin{aligned}
E[e^{t\bar{X}_n}] &= E[\prod_{i=1}^{n} e^{tX_i/n}] \\
&= \prod_{i=1}^{n} E[e^{tX_i/n}] \\
&= \prod_{i=1}^{n} \exp(\frac{\mu t}{n} + \frac{\sigma^2 t^2}{2n^2}) \\
&= \exp(\mu t + \frac{\sigma^2 t^2}{2n})
\end{aligned}$$

This means that $\bar{X}_n \sim N(\mu, \sigma^2/n)$.

## 2 Inequalities

The Gaussian Tail Inequality, Markov's Inequality, and Chebyshev's Inequality show similar properties of distributions–to what probability that the random variable lies in a far-away area $P(|X| > t)$.

The Gaussian Tail Inequality is much more tighter since it introduces much more information about the details of the distribution.

We can use more moments to better bound the probability, which comes to the *Chernoff's method.* Because the moments generating function encodes all the moments of a random variable (Taylor expansion shows all the moments), it's more appropriate to use $E[e^{tX}]$

$$P(X > \epsilon) = P(e^{tX} > e^{t\epsilon}) \leq \inf_{t \geq 0} e^{-t\epsilon} E[e^{tX}]$$

where $t > 0$.

The derivatives show an very important trick called **variational trick** by introducing more variables, here it's $t$. This renders much more flexibility, for example, minimizing the right size with respect to $t$ gives a much tighter bound.

We will not be able to compute the moments generating function for all different distributions, so we might as well bound it.

Note that $a \leq X \leq b$ implies that all of $X$'s moments are bounded! The important result is

$$E[e^{tX}] \leq e^{t\mu} e^{\frac{t^2(b-a)^2}{8}}$$

where $\mu = E[X]$.

**Hoeffding's Inequality** gives a very much tight bound on sum of independent variables and shows that the convergence is very quick. Hoeffding's Inequality also defines the confidence interval.

A function of a set of independent variables also shows such properties if the function is wiggling within some constant. That is demonstrated by The Bounded Difference Inequality or McDiarmid's inequality.

log function in KL distance is a concave function from which we can utilize the Jensen's inequality.

The cdf of the maximum of a set of random variables can be calculated by requiring every random variable to be less than $m$. But this is often not possible in practice.

The expectation of the maximum of a set of random variables grows at $\log n$ rate by Theorem 16.

$a_n = o_P(1)$ means that $a_n$ eventually **converges** to zero as $n \to \infty$. $a_n = O_P(1)$ means $a_n$ is eventually **bounded**, i.e., $|a_n| \leq C$ for some $C > 0$.

For probabilistic versions, if $Y_n = o_P(1)$, then for every $\epsilon > 0$,

$$P(|Y_n| > \epsilon) \to 0.$$

It says that as $n$ gets large, the probability mass concentrates around zero.

if $Y_n = O_P(1)$, then for every $\epsilon > 0$, there exists a $C > 0$ such that

$$P(|Y_n| > C) \leq \epsilon.$$

It says that as $n$ gets large, the probability mass is bounded. It can also be explained as that first choose an arbitrary $\epsilon$, we can find a $C$ satisfying the inequality. Consider the shifted probability mass sequence $Y_n \sim N(n, 1)$. It is bounded and $Y_n = O_P(1)$.

$\hat{p}_n - p = O_P(1/\sqrt{n})$ by Hoeffding's Inequality. The standard deviation $\sim 1/\sqrt{n}$.

$O_P(\cdot)$ gives us more information.

**Exercise $O_P(1)$ and $o_P(1)$.** Suppose $x_n = O_P(1)$, $y_n = o_P(1)$, and $z_n = x_n y_n$, all of these can be proved.

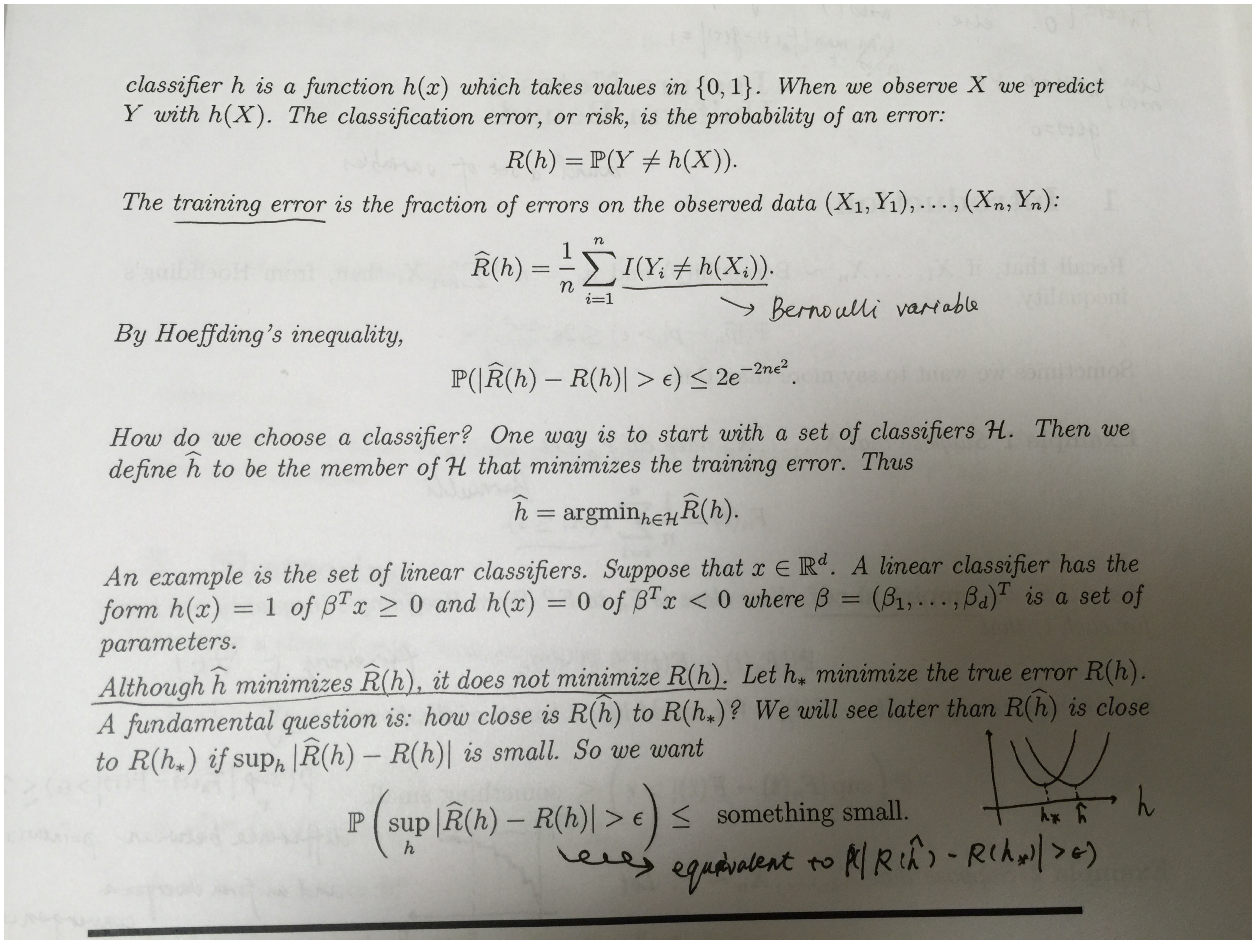
$$\begin{aligned} O_P(1)o_P(1) &= o_P(1) \\ O_P(1)O_P(1) &= O_P(1) \\ o_P(1) + O_P(1) &= O_P(1) \\ O_P(a_n)o_P(b_n) &= o_P(a_n b_n) \\ O_P(a_n)O_P(b_n) &= O_P(a_n b_n) \end{aligned}$$

# 3 Uniform bound and VC dimension

Empirical cdf can be seen as a Bernoulli variable. So we can use Hoeffding's Inequality.

$P(|F_n(t) - F(t)| > \epsilon) \leq 2e^{-2n\epsilon^2}$ **for every t** $\nRightarrow P(\sup_t |F_n(t) - F(t)| > \epsilon) \leq \textit{something}$. This is an example of the difference between pointwise convergence and uniform convergence. It relates to the issue of overfitting[1]. Check infimum speed over all points to determine whether it is uniform convergence.

[1] https://math.stackexchange.com/questions/597765/pointwise-vs-uniform-convergence In general, pointwise convergence is looser than uniform convergence. Pointwise convergence says at every point the sequence of functions has its own speed of convergence (that can be very fast at some points and very very very very slow at others), but uniform convergence says there is an overall speed of convergence.



*classifier $h$ is a function $h(x)$ which takes values in $\{0,1\}$. When we observe $X$ we predict $Y$ with $h(X)$. The classification error, or risk, is the probability of an error:*

$$R(h) = \mathbb{P}(Y \neq h(X)).$$

*The training error is the fraction of errors on the observed data $(X_1, Y_1), \ldots, (X_n, Y_n)$:*

$$\widehat{R}(h) = \frac{1}{n}\sum_{i=1}^{n} \underbrace{I(Y_i \neq h(X_i))}_{\rightarrow \text{Bernoulli variable}}.$$

*By Hoeffding's inequality,*

$$\mathbb{P}(|\widehat{R}(h) - R(h)| > \epsilon) \leq 2e^{-2n\epsilon^2}.$$

*How do we choose a classifier? One way is to start with a set of classifiers $\mathcal{H}$. Then we define $\widehat{h}$ to be the member of $\mathcal{H}$ that minimizes the training error. Thus*

$$\widehat{h} = \text{argmin}_{h \in \mathcal{H}} \widehat{R}(h).$$

*An example is the set of linear classifiers. Suppose that $x \in \mathbb{R}^d$. A linear classifier has the form $h(x) = 1$ of $\beta^T x \geq 0$ and $h(x) = 0$ of $\beta^T x < 0$ where $\beta = (\beta_1, \ldots, \beta_d)^T$ is a set of parameters.*

*Although $\widehat{h}$ minimizes $\widehat{R}(h)$, it does not minimize $R(h)$. Let $h_*$ minimize the true error $R(h)$. A fundamental question is: how close is $R(\widehat{h})$ to $R(h_*)$? We will see later than $R(\widehat{h})$ is close to $R(h_*)$ if $\sup_h |\widehat{R}(h) - R(h)|$ is small. So we want*

$$\mathbb{P}\left(\sup_h |\widehat{R}(h) - R(h)| > \epsilon\right) \leq \text{ something small.}$$

$\rightsquigarrow$ equivalent to $\mathbb{P}(|R(\hat{h}) - R(h_*)| > \epsilon)$

$$P(sub_{A \in \mathcal{A}}|P_n(A) - P(A)| > \epsilon) = P(\cup_{i=1}^{N}|P_n(A_i) - P(A_i)| > \epsilon)$$

The classes $\mathcal{A}$ can have many different forms.

The situation is separated from finite classes and infinite classes.

# 4 Convergence

The convergence is all about a sequence of statistics. The sequence does not need to be iid, but the underlying sequence is iid. They are not independent.

In probabilistic realm, $\lim_{n\to\infty} X_n = c$ is an **event**, so we have to bound it using probability.

The convergence in probability says $X_n$ is concentrating around $X$ but not $X_n = X$ numerically. In particular, if $X_n \to c$, then the distribution of $X_n$ concentrates around $c$ and is sharp.

$$X_n - c = o_P(1)$$

Convergence in probability is the central role.

Convergence in distribution holds at all $t$ where $F(t)$ is continuous. *It doesn't matter where $F(t)$ is not continuous.*

Convergence in distribution is another central role.

Sometimes it is easier to prove the convergence in quadratic mean, which implies the convergence in probability.

If a random variable converges to a point mass in distribution, then the convergence in probability also holds.

Point mass distribution $\delta_c(t)$ 0 if $t < c$ and 1 if $t \geq c$.

Let $X_n = \sqrt{n} I(0 < U < \frac{1}{n})$, $X_n$ can only take two values 0 or $\sqrt{n}$.

Convergence in quadratic mean is stronger in that it requires the probability mass doesn't go far away while convergence in probability implies probability concentrates around some point (and doesn't care if the probability goes far away). More precisely, convergence in probability does not consider the moments. Any statement that involves moments is stronger.

Let $X_n = -X$ for $n = 1, 2, 3, \ldots$, this is a well defined sequence but highly correlated and it's a sequence of a single value. $X_n$ is defined over a common sample space and remember that random variable is a function that maps every outcome into $\mathcal{R}$. $X_n$ for all $n$ are always the same function, so whenever an outcome is revealed the mapped value is determined. Make sure to distinguish the difference between $X_n = X$ and $X_n$ equals $X$ in distribution.

Convergence in distribution can be used to calculate the probabilistic statements by substituting a convergence distribution.

**Exercise Convergence is preserved under transformations.** If $X_n \to^P X$ and $Y_n \to^P Y$, then $X_n + Y_n \to^P X + Y$.

$$\begin{aligned}
P(|X_n + Y_n - X - Y| \geq \epsilon) &\leq P(|X_n - X| \geq \frac{\epsilon}{2} \cup |Y_n - Y| \geq \frac{\epsilon}{2}) \\
&\leq P(|X_n - X| \geq \frac{\epsilon}{2} + P(|Y_n - Y| \geq \frac{\epsilon}{2} \quad \text{(by union bound)} \\
&\to 0.
\end{aligned}$$

Not all random variables have a moment generating function.

Moments generating function is smooth.

CLT can be proved by moments generating function. It turns out that the mgf of $\bar{X}_n$ has the form of Gaussian mgf.

Berry-Esseen Theorem tells how close the distribution of $\bar{X}$ is to the Normal distribution and why it has $\sqrt{n}$ term.

If we replace $\sigma$ with sample standard deviation, CLT still holds.

$$\frac{\sqrt{n}(\bar{X}_n - \mu)}{S_n} \sim N(0,1)$$

This can be proved by Slutzky's theorem and continuous mapping theorem through

$$\frac{\sqrt{n}(\bar{X}_n - \mu)}{\sigma}\frac{\sigma}{S_n}$$

To measure the distance between two distributions, we can use

$$\sup_z |X(z) - Y(z)|$$

For multivariate central limit theorem, the vectors are iid instead of elements in the vector. Instead, the elements can be highly correlated, like, height and weight of a person.

The distribution of $g(\bar{X})$ cannot be directly cast into any distribution scaled by $g(\cdot)$ because there is a scaling issue caused by the sum of one constraint.

**Exercise See why the Delta Method holds.** By Taylor's theorem

$$g(\bar{x}) \approx g(\mu) + (\bar{x} - \mu)g'(\mu).$$

$$\sqrt{n}(g(\bar{x}) - g(\mu)) \approx \sqrt{n}(\bar{x} - \mu)g'(\mu)$$
$$\sqrt{n}(\bar{X} - \mu) \to N(0, \sigma^2)$$
$$\sqrt{n}(\bar{x} - \mu)g'(\mu) \to N(0, g'(\mu)^2\sigma^2)$$

# 5 Sufficiency

Statistics is kind of the reverse of probability. It is about given data and infer the distribution.

Parametric model can be parameterized by a finite number of parameters.

Parametric models is simple but can de dangerous since we assume a distribution that might not be the true case.

Parametric model is a good start because it can be seen as building blocks for complicated nonparametric models.

$\bar{X} - \mu$ is *not* a statistic because it is not a function of data, instead, it is a function of unknown parameters.

$f(X;\theta)$ is a family of density indexed by parameter $\theta$ (Note the semicolon).

Sufficiency is only useful for parametric model.

$P(X_1 = x_1, ..., X_n = x_n, T = t) = P(X_1 = x_1, ..., X_n = x_n)$ because the intersection of the set $\{X^n | T = t\}$ and the set $\{X^n | X_1 = x_1, ..., X_n = x_n\}$ is just the latter, which represents a point.

Statistics are equivalent if they produce the same partitions of the data.

If we can derive a conclusion from other statistics, it can be seen that other statistics contain more information or redundant information.

Minimal sufficient statistic generates the coarsest sufficient partition.

If we have a sufficient statistic, we can compute the likelihood function from it.

# 6 The likelihood function

The likelihood function serves the purpose of generating estimators, Bayesian inference, and sufficiency.

**Exercise Proof of Theorem 4.** if $L(\theta|x^n) \propto L(\theta|y^n)$, then the likelihood function has the same shape no matter what the parameter is. **This means $x^n$ and $y^n$ are in the same partition such that $T(x^n)$ and $T(y^n)$ should be equal.** Intuitively, this introduces a minimal sufficient partition. This is indeed the case.

Technically,

$$\begin{aligned} R(x^n, y^n; \theta) &= \frac{p(x^n;\theta)}{p(y^n;\theta)} \\ &= \frac{L(\theta|x^n)}{L(\theta|y^n)} \\ &= c \end{aligned}$$

$R(x^n, y^n; \theta)$ does not depend on $\theta$. By Theorem 10 in lecture note 5, this is indeed the case.

Then we can say that the likelihood function contains the information of minimal sufficiency.

# 7 Parametric point estimation

In frequentist view, the parameter is fixed, unknown constant but the estimator is a random variable, a function of sampled data.

Estimating "mixture of Gaussian" using the method of moments is tractable than MLE.

Minimax theory is perhaps the most important field of defining the what an optimal estimator is and it provides a formal way.

**Consistency of the estimator should be a minimal requirement!**

Method of moments: Equate the sample moments with the theoretical moments by the law of large number.

The estimator of $k$ for Binomial distribution might be a bad estimator since the denominator can be negative.

The log-likelihood changes the shape of the function but the maximum is the same place.

MLE under certain conditions is the optimal.

Sometimes MLE need to be found by numerical methods.

Find the profile likelihood $\sup_{\xi} L(\eta, \xi)$ implies that finding the maximum along $\eta$. The result is still a set. Maximizing the set gives $\hat{\eta}$.

MLE has a very important property called equivariant. This might be used to construct concave function. Even though the parameter is transformed, the maximum of the likelihood function does not change.

**Equivariance** for MLE is a very good property that does not share with other estimators.

Unbiased estimator does *not* have the equivariant property.

Bayes estimator is not Bayes inference.

Every different prior distribution gives a different Bayes estimator.

We can also take the moments of the posterior distribution.

Mean square error is the overall property of the estimators.

Nowadays unbiased estimators are not taken so important.

Convergence of MSE to zero with speed of $O(\frac{1}{n})$ is the property of some parametric models. But nonparametric models always have a slower convergence speed.

Computing MSE is a useful thing but it does not quite solve the problem since it always include the unknown parameter.

# 8 Minimax theory

Classification usually uses zero-one loss.

MSE is a special case of a more general concept.

The risk of an estimator is a function of parameter $\theta$.

$$R(\theta, \hat{\theta}) = g(\theta)$$

Maximum risk is the worst case of the risk.

First we can find the minimax risk and then find an estimator equal to the minimax risk. This estimator is a minimax estimator.

Instead of the supremum of the risk, we can also take the expected value of it.

For the estimator $\frac{\sum X+\alpha}{\alpha+\beta+n}$, every $\alpha$ and $\beta$ gives rise to a valid Bayes estimator.

Different from Maximum risk, which uses the maximum of the risk, Bayes risk takes the weighted average of the risk. But Bayes risk is dependent on the choice of prior distribution.

$$B_\pi(\hat{\theta}) = \int R(\theta, \hat{\theta})\pi(\theta)d\theta$$

and Bayes estimator minimizes Bayes risk.

**Unlike the minimax risk, Bayes risk is easier to calculate and if we choose $\pi(\cdot)$ carefully, Bayes estimator is minimax estimator. So the process is that first we calculate a Bayes estimator and then choose a prior distribution that makes it a minimax estimator.**

Minimax risk $R_n$ usually reaches zero as $n$ approaches $\infty$, but we consider more on at what rate it goes to zero. It can be seen that for parametric learning the speed that $R_n$ reaches zero is faster than nonparametric learning.

The main focus is that choosing the Bayes estimator as a root for minimax estimator with a proper prior.

Posterior risk takes the expected value with respect to $f_{\theta|x^n}$ while the risk takes the expected value with respect to $f_{x^n|\theta}$.

Bayes risk can also be written as

$$B_\pi(\hat{\theta}) = \int r(\hat{\theta}|x^n)\pi(x^n)dx^n$$

which is nonnegative. This gives a way to compute Bayes estimator.

For theoretical reason, we always use the formal definition of $B_\pi(\hat{\theta})$ but for computation aspect, we always use the posterior form of $B_\pi(\hat{\theta})$.

A Bayes estimator minimizes the posterior risk. It's kind of like if I want to minimize the sum, minimizing each element of the sum minimizes the whole sum.

The mean of the posterior $E[\theta|X = x^n]$ is the Bayes estimator for $L_2$ loss. This is still a function of the prior.

**Theorem The bound for the minimax risk.** Suppose $R_n = \inf_{\tilde{\theta}} \sup_\theta R(\theta, \tilde{\theta})$ is the minimax risk.

$$B_\pi(\hat{\theta}) \le R_n \le \sup_\theta R(\theta, \tilde{\theta}_0)$$

where $\hat{\theta}$ is an arbitrary Bayes estimator for any given prior $\pi(\theta)$ and $\tilde{\theta}_0$ is an arbitrary estimator.

*Proof.* The right-hand side of the inequality is obvious, since the minimum of a function is always less than or equal to values at any other point.

The left-hand side of the inequality can be proved.

$$\begin{aligned} \inf_{\tilde{\theta}} \sup_\theta R(\theta, \tilde{\theta}) &\ge \inf_{\tilde{\theta}} \int R(\theta, \tilde{\theta})\pi(\theta)d\theta \\ &= B_\pi(\hat{\theta}) \end{aligned}$$

since maximum of a set is always larger than a weighted average of the set. □

**MLE is approximately minimax for parametric models if the number of parameters is fixed. MLE is approaching minimax as $n$ is increasing. For now, we can take that MLE is minimax.**

Bayes estimator with a constant Bayes risk function is minimax.

**Theorem $\bar{X}_n$ is minimax under squared error loss.** Suppose $X_1, ..., X_n \sim N(\theta, 1)$ and $\theta \sim N(0, c^2)$, then $\theta = \bar{X}_n$ is minimax under squared error loss.

*Proof.* Take the estimator $\hat{\theta}_0 = \bar{X}_n$,

$$\sup_\theta R(\theta, \hat{\theta}_0) = Var[\bar{X}_n] = \frac{1}{n}$$

then $R_n(\hat{\theta}) \leq \sup_\theta R(\theta, \hat{\theta}_0) = \frac{1}{n}$.

Then take the Bayes estimator under squared error loss, which is the posterior mean

$$\hat{\theta}_1 = E[\theta | X = x^n] = \frac{nc^2 \bar{x}}{1 + nc^2}$$

$$\begin{aligned}
Var[\hat{\theta}_1] &= \frac{nc^4}{(1+nc^2)^2} \\
bias(\hat{\theta}_1) &= E_\theta[\hat{\theta}_1] - \theta \\
&= \frac{\theta}{1+nc^2} \\
R(\theta, \hat{\theta}_1) &= E[(\theta - \hat{\theta}_1)^2] \\
&= Var[\hat{\theta}_1] + bias(\hat{\theta}_1)^2 \\
&= \frac{\theta^2 + nc^4}{(1+nc^2)^2} \\
B_\pi(\hat{\theta}_1) &= \int R(\theta, \hat{\theta}_1)\pi(\theta) d\theta \\
&= \frac{c^2(1+nc^2)}{(1+nc^2)^2} \\
&= \frac{c^2}{1+nc^2}
\end{aligned}$$

$$R_n(\hat{\theta}) \geq B_\pi(\hat{\theta}_1) \to \frac{1}{n}$$

where $c \to \infty$ because it holds for every $c$.

So $R_n = \frac{1}{n}$. We can find an estimator $\theta$ whose minimax risk is $\frac{1}{n}$.

$$\sup_\theta R(\theta, \bar{X}_n) = \frac{1}{n}$$

thus by the definition of the minimax estimator, $\bar{X}_n$ is minimax under squared error loss. □

# 9 Asymptotic theory

Probably the most important distance is Kullback-Leibler distance

$$KL(P||Q) = \int p(x) \log \frac{p(x)}{q(x)} dx$$

and Hellinger distance

$$h(P,Q) = \sqrt{\int (\sqrt{p(x)} - \sqrt{q(x)})^2 dx}$$

To make MLE consistent, the premise is the model is **parametric**. And **(1) the dimension of the parameter space does not change with $n$; (2) $p(x;\theta)$ is a smooth function of $\theta$.**

MLE is very naturally related to KL distance.

The likelihood is bigger at the point of the true parameter than other points.

$$P(\frac{L(\hat{\theta})}{L(\theta)} > 1) \to 1$$

Thus Maximum Likelihood Estimator is a consistent estimator, which means it always gives the true estimate under regularity conditions.

MLE is effective for situations like the number of data is large and the number of parameter is small.

MLE converges to Normal distribution. In fact, other estimators also converge to Normal distribution but with a larger variance. That is why we call the mle is *optimal or efficient.*

*Score function and Fisher information* comes up very often in maximum likelihood theory.

Take the score function, the derivative of log likelihood, $S_n(\theta) = 0$ usually gives the mle.

Score function is a random function. It is a function of $\theta$ but also depends on the data. So given any $\theta$, $S_n(\cdot)$ is a random variable of $X$ and we can take the mean or variance on it.

$$Var(\hat{\theta}) \approx \frac{1}{I_n(\theta)}$$

Why $I_n(\theta)$ is called information is that if we have a lot of information, $I_n(\theta)$ is large, the variance is small and the estimator is precise.

For $Uniform(0,\theta)$, its support of $p$ changes with $\theta$.

Since the Fisher information $I_n(\theta) = nI(\theta)$, as we get more and more data, the information adds up.

If without regularity conditions, such as $Uniform(0,\theta)$, the Score function and Fisher information are useless. They are only dealing with well-behaved parametric models.

We usually use asymptotic normality to compute the confidence interval.

An counterexample: $X_1, ..., X_n \sim Uniform(0, \theta)$, then mle is $\hat{\theta}$. It turns out that

$$\sqrt{n}(\theta - \hat{\theta}) \to Exponential$$

The estimated standard error for the mle estimator

$$\hat{se} = \sqrt{\frac{1}{I_n(\hat{\theta})}}$$

is probably the most important approximation.

Relative efficiency is used to compare estimators, because some estimators are easier to compute than MLE.

As we can see from the Example 17, different estimators might be valid and mle is always the more efficient. This in only true when the model is correct. **But** if the true model is not what you assume, mle may gives the biased estimate while estimate from LLN gives unbiased result. **Moral: there is a trade-off between precise and efficiency. It depends on the extent to which you trust your model.**

MLE is optimal only in the sense that the model is true.

The real stuff is the nonparametric statistics.

All the above facts lie in that the parametric model is correct. But the parametric model is not correct.

We can trade efficiency for robustness.

# 10 Hypothesis testing

The question is whether there is sufficient evidence to reject $H_0$.

In general, false positive error are worse and we do not want to make false positive error.

First and foremost, we need to choose an error rate for **Type I error** based on the cases that we are concerned. For example, the null hypothesis is the mountain spring is harmful for students. We definitely want to reject it not easily. That corresponds to a low Type I error.

1. Choose a proper test statistc and 2. choose a proper rejection region.

Neyman-Pearson test is not usually used in practice but it is very important for theoretical reason.

$\alpha$ is the amount of Type I error that we can tolerate.

Null hypothesis is more specific. In Neyman-Pearson test, the alternative hypothesis is also specific, like two sides coin. That is counter to practical situations.

For the Neyman-Pearson test, it says if we have another size $\alpha$ test with power function $\beta$

$$\beta_{NP}(\theta_1) > \beta(\theta_1)$$

Wald test, likelihood ratio test, and permutation test is more common in practice.

For the Wald Test, the goal is to test a scalar and note that $\hat{\theta}$ is the mle estimator under null hypothesis. There are two different ways to compute se. You can use the estimated standard error of $\hat{\theta}$, or on the other hand, you can use the standard error under $H_0$, because $\hat{\theta}$ converges to $\theta$ asymptotically.

If the true value is $\theta$, then we can rewrite $T_n$

$$\begin{aligned}\frac{\hat{\theta} - \theta_0}{\hat{se}} &= \sqrt{nI(\theta)}(\hat{\theta} - \theta_0) \\ &= \sqrt{nI(\theta)}(\hat{\theta} - \theta) + \sqrt{nI(\theta)}(\theta - \theta_0)\end{aligned}$$

The first term is converging to $N(0, 1)$ and the second term is getting larger and larger such that $\beta(\theta) \to 1$ for $\theta \in H_1$.

For the likelihood ratio test, it can be used to test vectors. And unlike Neyman-Pearson test, it can be used for composite null hypothesis and composite alternative hypothesis.

Finding the maximum of the likelihood function can use numerical methods by sampling all values of $\theta$.

Rearrange the reject region for the likelihood ratio test $\lambda(x_1, ..., x_n) < c$ into $|T_n| > k$ where

$$T_n = \frac{\sqrt{n}(\bar{X} - \theta_0)}{S}$$

This is kind of the Wald test. Under $H_0$, $T_n$ has a t-distribution with n-1 degrees of freedom. The Student's t-distribution is more accurate when $n$ is small. But it is useless we won't conduct such computation when $n$ is small.

$p$ value basically can be seen as cdf. And if $H_0$ holds, p conforms to Uniform distribution.

Permutation test is distribution free. So sometimes working out the distribution under null hypothesis is very hard and we can use permutation instead.

Any continuous cdf conforms to uniform distribution [0,1]. This can be demonstrated by the inverse transform theorem $Y = F_X^{-1}(U) = F_Y(y)$.

$$F_{F_Y}(y) = F_U(y)$$

# 11 Confidence sets

Remember, the parameter is not random, instead, the procedure of generating the sets is random.

What we usually do is to construct a test first, and then convert the hypothesis testing into a confidence set.

**Theorem Confidence interval for MLE.**

$$\frac{\sqrt{n}(\hat{\theta}-\theta)}{\hat{\sigma}} \sim N(0,1)$$

The confidence interval is $C_n = \hat{\theta} \pm \hat{\sigma} z_{\alpha/2}\sqrt{n}$ such that

$$P(\theta \in C_n) \to 1-\alpha$$

The confidence intervals constructed by probability inequalities are quite conservative and so they're not used much since the interval is quite wide and uninformative.

The Wald Interval under the asymptotic approximate is the most commonly used interval.

When $\alpha = .05$, $z_{\alpha/2} \approx 1.96 \approx 2$, if $\hat{\theta}_n$ is the MLE, then 95% confidence interval is

$$\hat{\theta}_n \pm 2se$$

Confidence intervals are more informative. Once we have a confidence interval, we can test a hypothesis.

The Wald test interval is symmetric but the likelihood test interval is not. In some sense, the likelihood test interval is more accurate than the Wald interval.

# 12 Nonparametric Inference

For histogram density estimator, the convergence speed is $O(\frac{1}{n^{-2/3}})$. For the kernel density estimator, the risk is $\frac{C_1}{n^{4/5}}$.

For high dimension situation, say $d$-dimension, the kernel density estimator is

$$\hat{p}(x) = \frac{1}{n}\sum \frac{1}{h^d} K(\frac{x-X_i}{h})$$

and the risk function is

$$R = C_1 h^4 + \frac{C_2}{nh^d}$$

the minimum risk is

$$R = \frac{C_1}{n^{4/(4+d)}}$$

So we can see that as the $d$ gets larger. The estimator gets poor very quickly. This is a glimpse of the curse of the dimension.

The empirical cdf can be seen as putting mass $1/n$ at each $X_i$

$$F_n(x) = \frac{1}{n} \sum I(X_i \leq x)$$

PCA is a statistical functional of distribution–calculate the covariance matrix and do the eigendecomposition.

Plug-in estimator is a way to compute the statistical functional from empirical cdf.

To decide when a plug-in estimator is a good estimator is a very complicated subject. A lot of time they are not good estimators.

**Now to use the plug-in estimator to estimate the mean, we use $\int x dF_n(x)$. But the empirical cdf $F_n(x)$ is discrete, so the integral should be interpreted as $\sum xp(x)$ where $p(x)$ is the probability mass function. The key here is the empirical cdf $F_n(x)$ can be interpreted as putting mass $1/n$ at each data point, which is the pmf.** The final result is

$$\int x dF_n(x) = \sum X_i \frac{1}{n} = \bar{X}$$

For most cases, the statistical functional $\hat{\theta}_n = T(P_n)$, then

$$\frac{\hat{\theta}_n - \theta}{\hat{se}} \to N(0,1)$$

so we can use the nonparametric version of Wald test.

The plug-in estimator might be useful for the simple functional like the mean $\hat{X}$ and the variance$\frac{S^2}{n}$, but how about the complicated functional like the maximum eigenvalue and the covariance matrix. That needs bootstrap.

The plug-in estimator for the mean and the variance happens to be the MLE estimator for parametric model.

Plug-in estimator for computing the pdf from empirical cdf fails and that is a case when plug-in estimator fails.

# 13 Bootstrap

If we know the true distribution $P$ of data, then we can simulate the variance by the Law of Large Number Theorem.

$$\begin{aligned}
\frac{1}{N}\sum_{i=1}^{N}\theta_i &\to E[\theta] \\
\frac{1}{N}\sum_{i=1}^{N}\theta_i^2 &\to E[\theta^2] \\
\frac{1}{N}\sum_{i=1}^{N}(\theta_i-\bar{\theta})^2 &= \frac{1}{N}\sum_{i=1}^{N}\theta_i^2 - (\frac{1}{N}\sum_{i=1}^{N}\theta_i)^2 \\
&\to E[\theta^2] - (E[\theta])^2 \\
&= Var[\theta]
\end{aligned}$$

where $\theta_i = g(X_{i1}, X_{i2}, ..., X_{im})$. In fact, $\theta_i$ means the $i_{th}$ $m$-number sample from the data. In total, we sample $N$ set of samples of size $m$. But we do not know the true distribution $P$. **Use the empirical distribution $P_n$! That's the bootstrap.**

In all, we estimate $S_n(P)$ with $S_n(P_n)$. The simulation part is not difficult, the real difficult part is replace $P$ with $P_n$.

Drawing data *with replacement* ensures the data are iid.

For $O_p(1/\sqrt{n})$, we cannot control. But for $O_p(1/\sqrt{B})$, we can increase $B$ to decrease the error. So in all, $O_p(1/\sqrt{n})$ dominates.

Estimate the cdf of the estimator using the bootstrap!

# 14 Bayesian inference

The prior distribution in Bayesian inference is different from the one used for deriving the minimax estimator. Bayesian prior is used to express the subjective belief while the one used in minimax theory might be uninformative and fails to capture the subjective belief.

For Bayesian inference, you don't have to use Bayesian theorem. You can just look at the data and come up with the posterior distribution.

In contrast, in Frequentist inference, you can also use Bayes theorem to find a frequentist estimator like Minimax estimator, which is also a Bayes estimator.

Most of time, the integral in Bayesian inference is intractable. In practice, we resort to Monte Carlo methods.

As $n$ gets larger and larger, the prior disappears.

Usually, ”non-informative priors” is not transformation invariant, but Jeffreys priors is.

If you come up with a 95% Bayesian confidence interval, it does not mean you can capture the true parameter with 95% with repeated procedures. Instead, every time the parameter might be very different.

In Bayesian inference, the consistency means the posterior distribution converges to or concentrates to the true value.