MATHEMATICAL STATISTICS

MINSEOK SONG

Mathematical Statistics

1 Sufficiency and Exponential family

Definition 1. *Statistic T is called sufficient if $X \mid T$ does not depend on θ.*

- $(*)$We need **measure theoretic** definition of sufficiency. What exactly is meant that samples from $X \mid T$ follow the same distribution as X?
- $(*)$What I am getting at is that $P(X \in A) = P(X \in A \mid \sigma(w : T(X(w)) \in T(A)))(w)$ for $w \in \{w : T(X(w)) \in T(A)\}$. To be continued (check!)
- Well, I think it is a direct consequence of $P(X = x) = \int P(X = x \mid T(x) = t)P(T(x) = t)dt$. So it indeed follows from definition.
- Here, T being sufficient is irrelevant. However, we'll easily sample from $P(X = x \mid T(x) = t)$ because the distribution $X \mid T$ doesn't depend on θ.
- The setting here is that we are trying to infer θ based on data X. Bayesian definition would be phrased as $X \mid T \perp \theta$
- The intuition here is that having X given T does not give further information about θ. Just by definition, X (n samples) has the same distribution as a sample from $X|T$ (also n samples).

  Example 1.
  For example $X|T$ has a distribution depending on $T(X)$ such as

  $$N(1^T\bar{X}, \begin{pmatrix} 1-1/n & -1/n & \dots \\ -1/n & 1-1/n & \dots \\ \dots & & 1-1/n \end{pmatrix})$$

  and if we sample $\tilde{X}_i$'s from this distribution, it does follow $N(\theta, 1)$ (the conditional distribution follows that above distribution).
  Another example would be sum for Bernoulli, and maximum value for uniform distribution. By the way, maximum and minimum of samples from uniform distribution are both sufficient, illustrating that 1d sufficient statistic is not necessarily minimal.

- Any one to one function of minimal sufficient statistic is also minimal sufficient statistic. For example, $(T(x), T(x))$ would do.
- Lehmann-Scheffe theorem establishes the relationship between unbiasedness and sufficiency of statistic.

Theorem 1. *(factorization theorem) Sufficient statistic T satisfies $P(x) = f_\theta(T(x))g(x)$, that is, θ depends only through $T(x)$.*

Definition 2. *A sufficient statistic T is called minimal if any other sufficient statistic M is a function of T.*

Lemma 2. *Suppose $H_0 \subset H$. Suppose S is minimal sufficient for $P_\theta, \theta \in H_0$ and sufficient for $P_\theta, \theta \in H$. Then it is minimal sufficient for $P_\theta, \theta \in H$.*

- This is because any sufficient statistic in a larger parameter space is sufficient in a smaller set of parameter space by definition.

Theorem 3. *For $P_\theta : \theta \in \{\theta_1, \dots, \theta_d\}$, $T(x) = (\frac{P_{\theta_1(x)}}{P_{\theta_0}(x)}, \frac{P_{\theta_2(x)}}{P_{\theta_0}(x)}, \dots \frac{P_{\theta_d(x)}}{P_{\theta_0}(x)})$ is a minimal sufficient statistic.*

Proof. Sufficiency by definition. Minimality due to factorization theorem. □

Definition 3. ***Exponential family*** *is of the form $p(x \mid \theta) = \exp(\sum\limits_{j=1}^{d} \eta_j(\theta)T_j(x) - B(\theta))h(x)$.*

- We can see that $B(\theta) = \log \int e^{\sum_{j=1}^{d} \eta_j(\theta)T_j(x)} h(x)d\mu(x)$, which can be viewed as a function of η.
- Canonical form is of the form $p(x \mid \eta) = \exp(\sum_{j=1}^{d} \eta_j T_j(x) - A(\eta))h(x)$. We can see that $A(\eta)$ is convex function and parameter space is convex set.
- Using natural parameter and normalizing constant, we can easily get expected value and variance of $T(X)$, i.e. $E_\theta[T(x)] = J^{-1}\nabla B$.

Theorem 4. *(Rau-Blackwell) Assume $L(\hat{\theta}, \theta)$ is convex in $\hat{\theta}$ for any $\hat{\theta}$ and any sufficient statistic T. Define $\hat{\theta} = E_\theta(\hat{\theta} \mid T)$. Then $R(\hat{\theta}, \theta) \leqslant R(\hat{\theta}, \theta)$.*

Proof. $L(\hat{\theta}, \theta) = L(E_\theta(\hat{\theta} \mid T), \theta) \leqslant E_\theta(L(\hat{\theta}, \theta) \mid T)$ by Jensen inequality. Take expected value on both sides and we get the result. □

- I don't think we used sufficiency in the proof?

Definition 4. *Statistic T is called ancillary if it doesn't depend on θ. If the expected value doesn't depend on θ, it's called first order ancillary.*

Definition 5. *$T(X)$ is complete iff $E_\theta(f(T(x))) = 0, \forall \theta \in H$ imples $f(T(x)) = 0$ almost surely.*

- Ancillary statistic is something that is not informative about θ.
- In words, completeness means that if the function of statistic is first-order ancillary, it should better be constant.
- So, since T already has much information about θ, if $f(T)$ has hint of having no information about θ (first order ancillary), it surely doesn't have information about $f(T)$.
- Few ways to prove completeness:
  1. use the idea of MGF
  2. use measure theory result: if the integral on measurable set is same then the integrand is same almost everywhere,
  3. For full rank exponential family, T is complete.
- In order to complete non-completeness, one uses counter-exmple of non-zero f. One uses symmetry; for example one may use $\sum_{i=1}^{n} x_{(i)}$-median.

Theorem 5. *(Bahador) If T is sufficient and complete, then T is minimal sufficient.*

- T may well be minimal sufficient but not complete.

Theorem 6. *(Basu) If T is sufficient and complete, and if A is ancillary, then T and A are independent.*

- T is like a perfect summary of the parameter, while A doesn't contain information on θ so it makes sense that T and A are independent of each other.

2 Decision Theory

Definition 6. *The risk is defined by $E_\theta(l(\hat{\theta}(X),\theta))$. It is a function of θ.*

- If we have a prior on θ, we can further define Bayes estimator, that is,

$$\arg\min_{\hat{\theta}(X)\in H}\int R(\hat{\theta}(X),\theta)\pi(\theta)d\theta$$

- Minimax is defined as $\arg\min_{\hat{\theta}(X)\in H}\sup_{\theta\in H}R(\hat{\theta}(X),\theta)$
- Note that this is a function of X, that is, statistic.

Fact 1. $\hat{\theta}_\pi(x)=\arg\min_a\int L(a,\theta)\pi(\theta\mid x)d\theta$ is Bayes.

Proof. Fubini theorem. □

Fact 2. For a squared loss, Bayes estimator is given by $E(\theta\mid X)$. This is essentially due to bias-variance trade-off!

Fact 3. $\arg\min_a E(|X-a|)$ = median of X where median is defined as $P(X\leqslant m)\geqslant 1/2$ and $P(X\geqslant m)\geqslant 1/2$.

Fact 4. The median is given by a closed interval $[m_0,m_1]$.

- The inequality is important, because we are excluding the edge case where for example $P(X=0)=2/3$ and $P(X=1)=1/3$.
- The above fact is due to the following string of (in)equality, given $m_1<c$.

$$E(|X-c|)-E(|X-m|)=(c-m)[P(X\leqslant m)-P(X>m)]+2\int_{m<x<c}(c-x)dP(x)\geqslant 0$$

- As you can see the right hand side is some kind of deviation term, and the left hand side used the definition of median.
- $\arg\min_{\hat{\theta}(X)\in H}E_\theta(|\hat{\theta}(X)-\theta|)$ is given by $\hat{\theta}_\pi(x)=\arg\min_a\int|a-\theta|\pi(\theta\mid x)$, according to the previous argument, posterior median.

Theorem 7. *If for some π, $\hat{\theta}$ satisfies $\sup_{\theta\in H}R(\hat{\theta},\theta)=\inf_{\hat{\theta}}\int R(\tilde{\theta},\theta)\pi(\theta)d\theta$, then $\hat{\theta}$ is minimax.*

Proof. $\forall\tilde{\theta}$,

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

□

- This says that if worst possible risk for an estimator $\hat{\theta}$ is best possible average risk, then $\hat{\theta}$ is performing as well as possible, even in the worst-case scenario.

Corollary 8. *If $\hat{\theta}=\hat{\theta}_\pi$ for some π and $R(\hat{\theta}_\pi,\theta)$ is constant over $\theta\in H$ then $\hat{\theta}$ is minimax.*

Proof. The first inequality becomes equality due to constant assumption, and the second inequality becomes equality due to $\hat{\theta}$ being Bayes. □

Fact 5. The minimax estimator for squared loss is unbiased.

DS procedure

The dichotomy of Bayesian and Frequentist approach is mostly a matter of division based on foundational philosophical differences, and in some situation is not necessarily natural; each instance can be formally simply seen as a choice of analysis. The dichotomy has mainly developed based on two philosophical school of thoughts: subjective belief (via prior distribution) vs. long-run frequency (via sampling distribution).

(1) Frequentist
- They avoid making probabilistic claims about model parameters and hypothesss. Instead, they describe the behavior of statistics and procedures over many hypothetical repeated samples.
- They use Statistics derived from data, and use deterministic approach for inference, with methods like hypothesis testing and confidence intervals to draw conclusions about population parameters based on sample data.
- examples: T-test, linear regression, etc

(2) Bayesian
- They assign prior beliefs (distribution function) on parameters of model.
- They use probabilistic argument on specific hypothesis or parameter values.
- examples: MCMC, Bayesian hierarchical modeling, etc

3 Data Science Procedure

[list of topics that will be discussed]

(1) Preparation of data
- handling missing data (imputation, etc)
- transformation
- Box-Cox transformation
- Outliers

(2) Interpret descriptive statistics about data
- plots
- t-distribution
- F-distribution

(3) Modeling (parametric, nonparametric)
- ANOVA
- Ensemble methods
- Boosting
- AIC-BIC
- Regularization

(4) Predict based on model
- Confidence/Prediction intervals
- Real-time prediction

(5) Inference on statistics
- Hypothesis testing
- p-value
- Type I/II error
  - (a) Type I: reject null when true: for example, it's like wrongly assuming that the medication has an effect, when it doesn't.
  - (b) Type II: fail to reject null when false: for example, it's like failing to deciding that the medication has an effect, when it has an effect.
- Causal inference

(6) Evaluation of models
   - Cross-validation

(7) Communication of models

3.1 Types of Missing Data

(1) MCAR
(2) MAR
(3) MNAR

4 Outliers

5 Inference on statistics

6 p-value

Definition 7. *Given some hypothesis, the p-value is the probability of obtaining test results at least as extreme as the result actually observed, under the assumption that the null hypothesis is correct.*

- If p-value is less than significance level, we reject the null hypothesis. Significance level is a threshold we set to decide when we have enough evidence to reject the null hypothesis in favor of the alternative hypothesis.
- In statistics, we do not quite "prove" that the hypothesis is true. We can only reject the false statements or assertions (the theory of falsification, proposed by Karl Popper).

7 t-distribution

Definition 8. *t-distribution is defined as* $\frac{Z}{\sqrt{\frac{V}{k}}}$ *where* $Z \perp V$*,* $Z \sim N(0,1)$*, and* $V \sim \chi^2_k$*.*

- The t-distribution has the thicker tails than the normal distribution.
- Note that the χ^2_k distribution has mean k and variance $2k$. Hence as k gets larger the t-distribution looks like standard normal distribution due to CLT.
- In multiple regression setting, $\frac{\frac{\frac{\hat{\beta}_1-\beta_1}{\sigma}}{\sqrt{\sum_i (X_i-\bar{X})^2}}}{\sqrt{(n-2)\cdot\frac{\hat{\sigma}^2}{\sigma^2}/(n-2)}} \sim t_{n-2}$ under normal assumption of residual.

8 Chi-squared distribution

- It is possible to have $f(X)$ and $g(X)$ be independent. The key example concerns chi-distribution: Say θ is uniformly distributed in $[0, 2\pi]$. Let us say R^2 follows *chi*-squared distribution with degree 2: this acts like a distance square in the 2d space. Then $X = r\cos(\theta)$ and $Y = r\sin(\theta)$ are independent.
- The definition of chi-square looks contrived, in particular compared to normal distribution, but we use it a lot in the context of sum of squares.
- Under moment assumption, it's approximately true for large n via CLT.
- Under moment assumption, it's approximately true for large n via CLT.

9 Prediction interval, Confidence interval

- In the context of linear regression, a **confidence interval** provides a range of values within which we expect the true regression coefficient (e.g., β_0 or β_1) to lie with a certain level of confidence. This reflects our uncertainty about the value of the coefficient based on the data we have observed.

- On the other hand, a **prediction interval** provides a range of values within which a new observation y (given a particular x) is expected to fall with a certain level of confidence. We have $y = \beta_0 + \beta_1 x + \epsilon$. Note the additional uncertainty introduced by the random error ϵ.

9.1 Fisher information

Definition 9. *Fisher information is defined by $\mathcal{I}(\theta) = E[(\frac{\partial}{\partial\theta}\log f(X;\theta))^2|\theta]$ for appropriately regular f.*

- Note that $E[\frac{\partial}{\partial\theta}\log f(X;\theta)|\theta] = 0$.
- Think of this as how much (the logarithm of) likelihood varies in the vicinity of θ.
- Large variation of likelihood we have more "information." (for example, if a deviation causes a significant drop in likelihood, it means that the data is most probable under the original θ value.)

Theorem 9 (Cramer-Rao bound)**.** *Say $\hat{\theta}(X)$ is an unbiased estimator. We have $Var(\hat{\theta}) \geqslant \frac{1}{\mathcal{I}(\theta)}$.*

- This says that as the fisher information gets larger, we have better precision.
- There is a similar result for biased estimator.

10 Consistency

Definition 10. *We say $\hat{\theta}$ is consistent to θ if $P(|\hat{\theta} - \theta| < c) \to 1$ as $n \to \infty, \forall c > 0$.*

Example 2. Say X_i and Y_i are sampled from $N(\mu_i, \sigma^2)$. We can see that MLE $\hat{\sigma}^2$ is not consistent because of growing number of nuisance parameters. Essentially, as we add more groups and calculate $\hat{\mu}_i$, even if we have more data, it doesn't help us get a better estimate of the variance σ^2 due to "nuisance parameters" μ_i's.

- So how do we deal with it? There are two ways.
  1. Integrate out all the nuisance parameters and use MLE on that.
  2. Condition on the minimal sufficient statistic of nuisance parameters and maximize the conditional distribution.

11 Location family

Definition 11. *It is a family of probability distribution parametrized by a location parameter μ and a non-negative scale parameter.*

- The distribution function is of the form $F(\frac{x-a}{b})$.
- We can check that $(X_1 - X_i)/b$ is ancillary when $X_i's$ are i.i.d. for a known b because b will be cancelled.
- Elaboration: say $X \sim F(x)$ and $X = a + bY$. Then $P(X \leqslant x) = P(a + bY \leqslant x) = P(Y \leqslant \frac{x-a}{b}) = G(\frac{x-a}{b}) = F(x)$

12 Uniformly most powerful test(UMP)

Definition 12. *A test function $\phi(x)$ is UMP of size α if for any other test function $\phi'(x)$ satisfying*

$$\sup_{\theta\in\Theta_0} E(\phi'(X)|\theta) = \alpha' \leqslant \alpha = \sup_{\theta\in\Theta_0} E(\phi(X)|\theta)$$

we have

$$\forall\theta \in \Theta_1, \quad E(\phi'(X)|\theta) = 1 - \beta'(\theta) \leqslant 1 - \beta(\theta) = E(\phi(X)|\theta)$$

- The first term is type I error, and the second term is the power of test function.

- It says that UMP test has the highest power among all tests whose type I error is less than or equal to α, a maximum type I error for our chosen test function.
- The assumption is needed because we can artificially make the test function to have th highest power by rejecting all the time.

13 unbiasedness of test function

Definition 13. *Define $\phi(X) = 1$ if $X \in R$ and 0 if $X \in R^c$ where R is a reject region. If a test function $\phi(X)$ satisfies $E_\theta \phi(X) \leqslant \alpha$ if $\theta \in \Omega_H$ and $E_\theta \phi(X) \geqslant \alpha$ if $\theta \in \Omega_K$ then it is said to be unbiased.*

- UMP is unbiased. Think of a test function $\phi = \alpha$.

14 distributions

14.1 Exponential distribution

- The distribution function is $F_X(x) = 1 - e^{-\lambda t}, t \geqslant 0$. λ here is like a rate.
- When the rate is λ, we are asking how long do we wait until we first see the bus.
- Mathematically, this is like calculating $F(t) = 1 - \lim_{\Delta t \to 0} (1 - \lambda \Delta t)^{\frac{t}{\Delta t}} = 1 - e^{-\lambda t}$.
- So we have $f(t) = \lambda e^{-\lambda t}$. Mean value is shown to be $\frac{1}{\lambda}$

  Example 3. As an exercise, let us check the distribution of $x_{(1)}$ from n samples. We have $P(x_{(1)} \geqslant t) = P(x_i \geqslant t, i = 1, 2, \ldots, n) = e^{-\lambda n t}$, so that $P(x_{(1)} \leqslant t) = 1 - e^{-\lambda n t}$.

14.2 Poisson distribution

- It is given by $P(X = k) = \lim\limits_{n \to \infty} \binom{n}{k} (\frac{\lambda}{n})^k (1 - \frac{\lambda}{n})^{n-k}$
- So exponential distribution is about time that we first see an event happening with the rate λ(essentially success probability over unit time), and poisson distribution is about finding out for a fixed time when the rate is λ, the probability of having k events.

14.3 Beta distribution

- It is given by $\frac{1}{B(\alpha,\beta)} x^{\alpha-1} (1-x)^{\beta-1}$ where $B(\alpha, \beta) = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}$
- It can be conjugate prior with respect to Bernoulli by the following formula:

$$p(\theta \mid x) \propto p^{\sum x_i} (1-p)^{n - \sum x_i} \times p^{\alpha} (1-p)^{n-\alpha} \sim B(\sum x_i + \alpha, \sum (1 - x_i) + \beta)$$

Linear Model

- Two assumptions: normality($Y_i \sim N(\beta_0 + \beta_1 X_i, \sigma^2)$) and moment assumption.

  Fact 6. For Normal distribution, pair-wise independence implies mutual independence.

  Proof. It comes from the fact that the mutual independence holds if and only if density function factorises. In particular, multivariate normal distribution with diagonal matrix covariate satisfies this property. □

- For 1d case, we could calculate $\hat{\beta}$ based on $\partial / \partial \beta (RSS)$...
- After computing β_0 and β_1, we can also compute their respective variance and covariance between them. I need to use the fact that $\bar{Y}$ and $\hat{\beta}_1$, which could also be easily calculated.
- We could further define $\hat{\epsilon}$, and unbiased estimator $\hat{\sigma}$. Showing that it is unbiased and is independent to $\hat{\beta}_0$ and $\hat{\beta}_1$ is more involved (need to check).

- Knowing that $\frac{\hat{\beta}_1-\beta_1}{\sigma/\sqrt{\sum_i(X_i-\bar{X})^2}} \sim N(0,1) \perp$ (because $\hat{\beta}_1 \perp \hat{\sigma}^2$) $(n-2)\cdot\frac{\hat{\sigma}^2}{\sigma^2} \sim \chi^2_{n-2}$ we can do some more analysis involving confidence interval of predictive value and mean value at a given point.
- Some **intuition** about why the variance of β is inversely proportional to dispersion: well, as the data gets concentrated, the prediction is rather **unreliable**.
- The reason we used t-distribution is to cancel out the **unknown value** σ.
  - RSS = portion that is not explained by the model
  - TSS = total sum of squares = $\|Y-\bar{Y}1_n\|^2$. Note that $Y-\bar{Y}1_n \in span(X)$
  - ESS = explained sum of squares = $\|\hat{Y}-\bar{Y}1_n\|^2$
- RSS + ESS = TSS. The role of $\bar{y}$ is like the "standardized line." Refer to here.
- R square is ESS/TSS.
- We would use F test... F statistic is given by $F = \frac{\|U_2^TY\|^2/k}{\|U_3^TY\|^2/(n-p)} = \frac{RSS_{\text{partial model}}-RSS_{\text{full model}}}{RSS_{\text{full model}}/(n-p)}$
- We want F to be as small as possible for the hypothesis $\beta=0$ to be true, to accept partial model. In other words, $RSS_{\text{partial model}}$ tends to be larger if $\beta_S \neq 0$.