THEORY OF STATISTICS AND PROBABILITY

MINSEOK SONG

Statistics

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

1 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
- Causal inference

(6) Evaluation of models
- Cross-validation

(7) Communication of models

1.1 Types of Missing Data

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

2 Outliers

3 Inference on statistics

4 p-value

Definition 1. *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).

5 t-distribution

Definition 2. *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, $\frac{\frac{\hat{\beta}_1-\beta_1}{\frac{\sigma}{\sqrt{\sum_i (X_i-\bar{X})^2}}}}{\sqrt{(n-2)\cdot\frac{\hat{\sigma}^2}{\sigma^2}/(n-2)}} \sim t_{n-2}$ under normal assumption of residual.

6 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.

7 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.

7.1 Fisher information

Definition 3. *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 1 (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.

Lienar Model

Probability Theory

Definition 4. ***Distribution Function (F):*** *A function* $F : \mathbb{R} \to [0,1]$ *that:*

(1) Is non-decreasing.
(2) Is right-continuous.
(3) Limits to 0 as $x \to -\infty$ *and 1 as* $x \to \infty$*.*

Random Variable (X): *A measurable function* $X : \Omega \to \mathbb{R}$ *where:*

(1) Ω *is a sample space equipped with a probability measure.*
(2) $\mathbb{R}$ *is the real line equipped with the Lebesgue measure.*

Probability Measure (P): *A measure defined on a sample space* Ω *whose total measure is 1.*

Distribution (D): *Given a random variable* X*, its distribution is the measure induced by* X *on* $\mathbb{R}$*.*

Density Function (f): *For a random variable* X*, its density function* f *is any measurable function that satisfies:*

$$\Pr(X \in A) = \int_A f \, d\mu$$

for every measurable set $A \subset \mathbb{R}$*, where* μ *is the Lebesgue measure.*

Some theorems and properties may appear simple at first glance, but upon closer examination, they are actually quite complex.

- The Skorokhod Representation Theorem states that for every CDF, there exists a canonical probability space and a random variable on that space with the given CDF.
  - You can consider either $\sup\{x \in \mathbb{R} : F(x) \leqslant w\}$ or $\sup\{x \in \mathbb{R} : F(x) < w\}$

  - This shows that distribution function fully characterizes random dsitribution.
- Kolmogorov's Existence threorem states that given finite dimensional sets of distribution, under some conditions, there exists a canonical probability space with random variables whose corresponding distributions coincide with our original distributions.
- The existence of distribution function does not guarantee the existence of density function. By Radon-Nikodym, there is a nice characterization when this happens: when distribution (which is a measure defined on R) corresponding to distribution function is absolutely continuous with respect to Lebesgue measure; this is a iff condition, that is, absolutely continuous if and only if distribution exists.
- The corresponding distribution is connected to the notion of generalized function (linear functional operator).
- The previous discussion illustrates that Kolmogorov's existence theorem is more general than Skorokhod Representation theorem.

Throughout the theory of probability, we usually assume that the probability measure is complete, because of the following proposition.

Proposition 2. *Let us assume that probability measure is complete. Let $X_t = Y_t$ almost everywhere for every $t \geqslant 0$. Then there exists $\tilde{Y_t} = Y_t$ almost everywhere for every t such that $X_t(w) = \tilde{Y_t}(w)$ for every $w \in P$ for every t for measurable set P whose measure is 1.*

8 Covariance, Independence

- Covariance only gives the information about linear relationship between two random variables. It has a limit of capturing non-linear relationship.
- Non-linear activation function in deep neural network is an example of such an attempt to leverage non-linearities.
- Further, as correlation gives only the linear relationship, zero correlation does not necessarily imply independence.
- Independence does not necessarily mean that, in a strict sense, two variables have no "relationsihps." It simply means that the information **about the value of the random variable** does not give any information of **the value of the other random variable**. For example, we may have $Y = X^2$ where $X = -1$ or 1.
- Also note that two random variables may have different mappings while having the same distribution: standard example is $1 - X$ and $X \in [0, 1]$.

9 Distribution function

- If we were to calculate the density function of $X + Y$, we may use convolution.
- If we were interested in $\frac{X}{Y}$, we may use transformation and integrate out.
- Another analytic approach is to use characteristic functions.
- We can use Monte Carlo method numerically.

10 Convergence

11 Markov Chain

Fact 1. The reverse of Markov chain is also Markov chain.

Proof. We want to show

$$P(T_k | T_{k+1}, \dots, T_N) = P(T_k | T_{k+1})$$

This is equivalent to

$$P(T_{k+2}, \dots, T_N | T_k, T_{k+1}) = P(T_{k+2}, \dots, T_N | T_{k+1})$$

Well, it holds because

$$\begin{aligned} LHS &= P(T_{k+2}|T_k, T_{k+1}) \times P(T_{k+3}|T_k, T_{k+1}, T_{k+2}) \times \cdots \times P(T_N|T_k, \ldots, T_{N-1}) \\ RHS &= P(T_{k+2}|T_{k+1}) \times P(T_{k+3}|T_{k+1}, T_{k+2}) \times \cdots \times P(T_N|T_{k+1}, \ldots, T_{N-1}) \end{aligned}$$

□

by the following fact.

Fact 2. Markov Chain X satisfies $P(X_k|X_{k-1}, \ldots, X_0) = P(X_k|X_{k-1}) = P(X_k|X_{k-1}, \text{anything from } X_{k-2} \text{ to } X_0)$

Proof. For convenience, denote $P(X_k|X_{k-1}, \text{anything from } X_{k-2} \text{ to } X_0)$ by $P(X_k|X_{k-1}, \ldots)$. We have two facts...

(1) $P(X_k|X_{k-1})$ is $\sigma(X_{k-1}, \ldots)$-measurable.
(2)
$$\int_A 1_{X_k=s} = \int_A P(X_k|X_{k-1}, \ldots, X_0) = \int_A P(X_k|X_{k-1}) \quad \forall A \in \sigma(X_{k-1}, \ldots)$$

This shows that $P(X_k|X_{k-1}) = P(X_k|X_{k-1}, \ldots)$. □

Theorem 3 (Good Markov Property)**.** *For any given $N \in \mathbb{N}$ and $x \in S$, and conditioned on the event $\{X_N = x\}$, the sequence of random variables $\{X_N, X_{N+1}\}$ is a Markov Chain and is independent of $X_0, \ldots, X_{N-1}$.*

Theorem 4 (Strong Markov Property)**.** *Let T be a stopping time. Then, conditional on the event $\{T < \infty\}$ and $\{X_T = x\}$ and for any $x \in S$ such that $P(\{T < \infty\} \cap \{X_T = x\}) > 0$, the random sequence $\hat{X} = \{X_{T+n}\}_{n \geqslant 0}$, defined by X_{T+n} for each $n \in \mathbb{N}_0$, is a Markov Chain with transition probability matrix P and initial distribution δ_x.*

Furthermore, the random sequences T and $\{X_{T+n}\}_{n \geqslant 0}$, are then conditionally independent, given the event $\{T < \infty\}$ and $\{X_T = x\}$.

- The event $\{T < \infty\} \cap \{X_T = x\}$ represents the event that the stopping time is finite and the process is in state x at this time.
- Conditional independence given the above event means that once we know two things - first, that the stopping time T is finite, and second, that the process is in state x in this stopping time - then the sequence of states after stopping time T does not depend on the sequence of states up to the stopping time T.

12 Martingale

Definition 5. *A sequence $Y_i, i = 1, 2, 3, \ldots$ is said to be a martingale with respect to the filtration $\mathcal{F}$ if $\forall n$,*

- $E(|Y_n|) < \infty$ and $E(Y_{n+1}|\mathcal{F}_n) = Y_n$
- This is a fair game. It says that expected future value given all the information up to now doesn't give any edge and should be equal to the present value.
- In this context, we want submartingale, where we have $E(Y_{n+1}|\mathcal{F}_n) = X_n$

Theorem 5. *(Option Sampling Theorem) Suppose T is a stopping time and M_n is a martingale with respect to $\{\mathcal{F}_n\}$. Then $Y_n = M_{n \wedge T}$ is a martingale. In particular, for each n,*

$$E(M_{n \wedge T}) = E(M_0).$$

If T is bounded, that is, if there exists $k < \infty$ such that $P(T \leqslant k) = 1$, then

$$E(M_T) = E(M_0)$$

- This says that expected value of the martingale at the stopping time is the same as its initial value, where you might be tempted to think you can choose an advantageous time to stop betting or investing to make a profit.

13 Brownian Motion

Definition 6. *It is a stochastic process with*

(1) $B_0 = 0$*,*
(2) $B_t - B_s$ *is same as* B_{t-s} *(stationary),*
(3) for each s*,* $\{B_{t+s} - B_s, t \geqslant 0\}$ *is independent of* $\{B_r, r \leqslant s\}$*,*
(4) and $t \mapsto B_t$ *is a continuous function of* t *almost surely.*

- (check) We can show that for every $s < t$, we have $B_t - B_s \sim N(m(t-s), \sigma^2(t-s))$
- When $m = 0$ and $\sigma = 1$, we call it Standard Brownian Motion.

13.1 Two constructions

(1) First construction
  - First, we use Daniel-Kolmogorov theorem to get the following.

    Fact 3. There is a probability measure P on $(\mathbb{R}_{[0,\infty)}, B(\mathbb{R}_{[0,\infty)}))$, under which $X_t(w) = w(t)$ (w is a path) has stationary, independent increments. Further, $X_t - X_s$ is normal with mean zero and variance $t - s$.

  - Now, we want to say that $P(C([0,\infty))) = 1$. The problem is, $C([0,\infty))$ is not even measurable.
  - To mitigate this, one might need some theorem extending to continuity.

(2) Second construction
  - We can do like this.
  - Define on Dyadic rationals
  - Prove that B_t is uniformly continuous on the dyadic rationals with probability one.
  - Extend B_t to t in $[0,1]$ by continuity.
  - Check that this works.

14 Relationship with PDE

15 Relationship with Complex Analysis

16 Misecelleneous

Definition 7. *A sequence of random variables* $(X_1, \ldots, X_n)$ *has a joint normal distribution if*

$$X_j = m_j + a_{j1}Z_1 + a_{j2}Z_2 + \ldots + a_{jm}Z_m$$

where $Z_i \sim N(0,1)$*, i.i.d.*

- Clearly, $E(X_j) = m_j$. Assume mean zero. Then we can express $X = AZ$ where A is an $n \times m$ matrix with entries a_{jk}. Each X_j is a normal with mean zero and variance $E(X_j^2) = a_{j1}^2 + \ldots + a_{jm}^2$.
- Further, covariance is given by $\mathrm{Cov}(X_j, X_k) = E(X_j X_k) = \sum_{l=1}^{m} a_{jl}a_{kl}$.
- Letting $\Gamma = AA^T$, we get $\Gamma_{jk} = E(X_j X_k)$, and this is called covariance matrix.