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

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

Types of Missing Data

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

Outliers

Inference on statistics

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).

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

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.

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.

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.

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.*

There are several theorems/properties that look trivial but are not really trivial.

- 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.
- 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 $\mathbb{R}$) corresponding to distribution function is absolutely continuous with respect to Lebesgue measure.
- The existence of distribution function guarantees the existence of distribution, and vice versa.
- However, the corresponding distribution always exists (also called generalized function).
- 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.*

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 $X \in [0, 1]$.

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.

Convergence

Markov Chain

Martingale

Brownian Motion

Relationship with PDE

Relationship with Complex Analysis