Statistics in Data Mining

Seminar Data Mining

Ege Onur Taga
Department of Informatics
Technische Universität München
Email: egetaga@gmail.com

Abstract— **Data mining is a cross-disciplinary field at the intersection of computer science and statistics aiming to make valuable inferences and predictions from data. Statistical methods provide data mining with invaluable tools to understand and interpret the data better. This paper aims to present statistics in data mining with an overview of some of the use cases and concludes with a short comparison of these two fields. The emphasis is on classical tools of statistics such as estimation, sampling and hypothesis testing.**

Keywords— **Data Mining, Statistics, Sampling, Estimation, Hypothesis Testing, Probability Distributions, Maximum Likelihood Estimation (MLE)**

I. Introduction

Data mining is the process of making inferences, predictions, and discovering patterns in big data [1]. Being at the intersection of machine learning, statistics, and database systems, it utilizes various tools from different disciplines. Statistics has a central role in data mining. In fact, before the advent of powerful computers and machine learning, the role of making useful inferences was the task of statistics, albeit with much smaller data [2]. Statistics flourished during the 20th century and is concerned with analyzing, collecting, and interpreting data. Mathematical rigor plays a pivotal role in statistics. The properties of the methods developed in statistics are proved diligently in a formal manner. In data mining, however, empirical results are also acceptable in the evaluation of methods. Although data mining emphasizes empirical evaluation rather than mathematical rigor, the role of statistics in the development of data mining is undeniable. Many machine learning methods, such as bayes classifiers, have assumptions regarding the distribution of data. Such assumptions and their implications are essential for developing data mining methods and for appropriately using them. Moreover, many machine learning techniques in use, such as linear regression and logistic regression, were developed previously in the domain of statistics. Understanding the statistical point of view in the development of such models may improve the quality of our understanding of data mining.

Many papers [3] [4] [5] [6] shedding light on the relationship between statistics and data mining are present. Benjamini and Leshno [3] argue that a better understanding of statistics in data mining has a vast potential for data mining research and discuss issues in data mining where statistical approaches address fundamental problems such as scalability, the curse of dimensionality, modeling relationships, and assessing uncertainty. They argue that despite the terminological differences, some of the approaches are similar. For example, they claim "learning" in neural networks has counterparts in statistics as "estimation". On the other hand, Hand [4] focuses more on the differences between these two fields, albeit agreeing that there is considerable overlap between them. He devotes two sections to the nature of statistics and that of data mining, in which he argues that statistics is more concerned with model fitting, whereas data mining is concerned with learning in the areas they overlap. He draws attention to the discrepancy between these two fields regarding the validation methods, the size of the data sets, and the way data is collected. He argues that the process of discovering patterns from large data sets is thrilling, yet with a risk of discovering misleading patterns. He proposes that joint research of data miners and statisticians is essential in this regard. Hosking, Pednault and Sudan [5] present a more detailed overview of statistics and data mining. As in previous papers, they highlight the methodological discrepancies, albeit with focusing more on the parallelity. They illustrate three approaches to machine learning: classical statistical modeling, Vapnik's statistical learning theory, and computational learning theory and PAC learning. They argue that data mining and classical statistical inference have much to offer to each other. For example, they suggest that statisticians inspect the asymptotic properties of models less and focus more on predictive accuracies. Similarly, they suggest that data miners make use of the experiences of statisticians regarding outliers and influential observations. Concerning the real-world problems, they advocate the combination of methods in statistical inference and that of data mining.

The primary purpose of this paper is to explain the use of statistics in data mining focusing more on classical statistics rather than comparing these two fields. Occasionally, however, we extend the ideas presented and explain how a particular statistical approach affects data mining. In section II, we introduce probability distributions with an emphasis on the ones prevalent in machine learning. In sections III-V, we explain sampling, estimation, and hypothesis testing, respectively. These sections draw attention to concepts such as maximum likelihood estimation and their implications for data mining. The last section summarizes all these statistical concepts and discusses their role in data mining with a short comparison of these two fields.

II. Probability Distributions

Probability distributions play a central role in data mining since both interpreting the data and developing machine learning algorithms require an understanding of the assumptions on the distribution. There are mainly two types of probability distributions as continuous distributions and discrete distributions. The random variable X has a continuous probability distribution if X takes uncountably many possible values. If X takes countably many values, we say that X has a discrete distribution. For example, a feature variable denoting eye color has a discrete distribution in a dataset, whereas a feature variable denoting the height of a person has a continuous distribution. In the following subsections, we will use basic probabilistic facts and definitions. The following references cover them thoroughly [7] [8, chapter 1].

A. Continuous Probability Distributions

There are many well-studied continuous probability distributions. Some of them, such as continuous uniform, normal and exponential distribution, have principal importance in data mining.

1) Continuous Uniform Distribution: A random variable X has a continuous uniform distribution when it is equally probable to take any value between the finite limits. Formally, when the probability density function is as follows:

$$f(x) = \begin{cases} \dfrac{1}{b-a} & \text{if } x \in [a,b] \\ 0 & \text{otherwise} \end{cases}. \tag{1}$$

With trivial computations, it is easy to see that the $\mu_x = \frac{a+b}{2}$, and $\sigma_x^2 = \frac{(b-a)^2}{12}$. A uniform distribution is generally used when there is little to no information about the distribution [8, chapter 2]. It is important to note that the minimum bound a and the maximum bound b fully parameterize the distribution.

2) Normal Distribution: The normal distribution, also known as Gaussian distribution, is one of the most fundamental distributions in statistics and other fields. A random variable X has a normal distribution when the probability density function is as follows:

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right) \tag{2}$$

The mean and variance of X is μ and σ^2, respectively. One of the most important properties of normal distribution is that its mean, median and mode are equal [8, chapter 8]. If random variable X follows normal distribution, it is denoted as follows:

$$X \sim N\left(\mu, \sigma^2\right) \tag{3}$$

When $\mu_z = 0$ and $\sigma_z^2 = 1$, the normal random variable Z has standard normal distribution and is denoted as $Z \sim N(0,1)$. The random variable X could be converted to a standard normal random variable using the equation [8, chapter 8]:

$$z = \frac{x - \mu_x}{\sigma_x} \tag{4}$$

Since there is no closed form solution for the cumulative density function of normal distribution, this transformation is widely used when using z-tables or software packages. The following figure[1] demonstrates an important property about normal distributions:

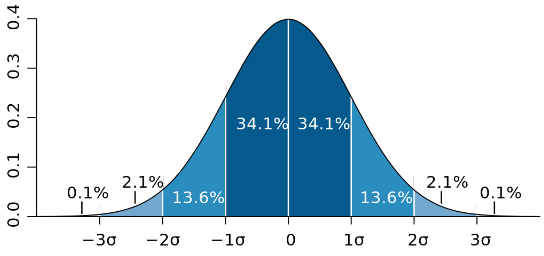

Fig. 1. The figure shows that approximately 99.7% of the values lie within 3 standard deviations.

3) Exponential Distribution: The exponential distribution is of great importance to queuing theory, reliability, and survival analysis. The probability density function of exponentially distributed random variable X is as follows:

$$f(x) = \begin{cases} \lambda e^{-\lambda x} & \text{if } x \geq 0 \\ 0 & \text{otherwise} \end{cases}. \tag{5}$$

It is easy to show that $\mu_x = \frac{1}{\lambda}$ and $\sigma_x^2 = \frac{1}{\lambda^2}$. Moreover, the cumulative density function $F(x)$ has a closed form.

$$F(x) = \int_0^x \lambda e^{-\lambda x}\, dx = 1 - e^{-\lambda x} \tag{6}$$

The distribution has an important property called memorylessness, meaning that the probability of an event happening in a certain time interval does not depend on the starting time. Stated formally, $P(X > a + b | X > a) = P(X > b)$. In fact, it is the only continuous distribution with memorylessness. For the proof of this property, please refer to [8, chapter 3]. The relation between Poisson distribution and exponential is also interesting. If the probability mass function of a Poisson distribution is $g(n) = \frac{\lambda^n e^{-\lambda n}}{n!}$ and the probability density of the random variable X is $f(x) = \lambda e^{-\lambda x}$ the random variable X refers to the time interval between poisson events [8, chapter 3].

There are many other continuous probability distributions which we did not cover such as Erlang, Gamma, Beta, Weibull etc. with important applications in data mining. We could not introduce multivariate distributions. We suggest Thompoulos [8] for the facts and applications about these distributions. For more diligent approach with mathematical rigor, we suggest Ross [9] and Hoel, Port and Stone [10, chapter 5].

[1] File:Standard deviation diagram.svg, https://commons.wikimedia.org/wiki/File:Standard_deviation_diagram.svg

B. Discrete Probability Distributions

Just as continuous probability distributions, there are many well-studied discrete probability distributions. Some of them, such as discrete uniform, binomial, geometric, and discrete distributions, have a wide range of natural science and data mining applications.

1) Discrete Uniform Distribution: The random variable X has a discrete uniform distribution when it is equally probable for X to take any integer value between a and b. Formally, if the probability mass function of X is as follows:

$$p_X(x) = \frac{1}{b-a+1} \text{ for } x \in \{a, a+1, ..., b-1, b\} \tag{7}$$

It is trivial to compute that $\mu_x = \frac{a+b}{2}$ and $\sigma_x^2 = \frac{(b-a+1)^2-1}{12}$.

2) Binomial Distribution: The random variable X has a binomial distribution when X denotes the number of successes in n identically and independently run experiments, each of which has two outcomes as success and fail with probability p and $1-p$, respectively. For example, if $n=3$, x can take values $0, 1, 2, 3$. The probability of X successes in n trials is:

$$p_X(x) = \binom{n}{x} p^x (1-p)^{n-x}, \text{for } x \in \{0, ..., n\} \tag{8}$$

The mean of X is $\mu_x = np$ and the variance $\sigma_x^2 = np(1-p)$ [8, chapter 14]. A special case of binomial distribution occurs when $n = 1$ and is called Bernoulli distribution. The random variable X in such distribution is called Bernoulli random variable and the experiment Bernoulli trial.

It is possible to approximate the binomial distribution with normal distribution when n is large. The approximation works if $p \leq 0.5$ and $np > 5$ or if $p > 0.5$ and $n(1-p) > 5$. Such an approximation is particularly important because it reduces computational complexities of many tasks. When n is large and p is small but the normal approximation is not suitable, it is possible to use Poisson distribution to approximate with $\lambda = np$ [8, chapter 14]. The following figure[2] demonstrates the approximation:

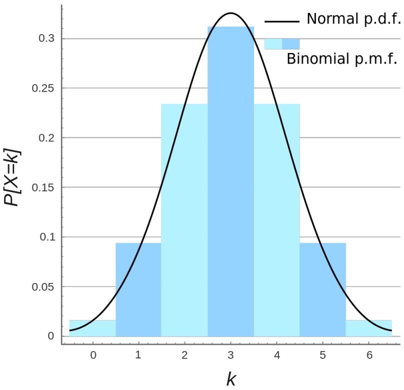

Fig. 2. A figure showing normal approximation to binomial distribution.

3) Geometric Distribution: The random variable X has a geometric distribution when X denotes the number of trials until a successful outcome is observed in identically and independently run experiments, each of which has two outcomes as success and fail with probability p and $1-p$. There are also other definitions where X refers to the number of fails until a successful outcome is observed [8, chapter 15]. In the following, we will stick with the first definition. The probability mass function of geometric distribution is as follows:

$$p_X(x) = p(1-p)^{x-1}, \text{ for } x \in N^+ \tag{9}$$

The mean of X is $\mu_x = \frac{1}{p}$ and the variance $\sigma_x^2 = \frac{1-p}{p^2}$ and the cumulative mass function has a closed form as,

$$F(x) = 1 - (1-p)^x \tag{10}$$

Just like the exponential distribution, geometric distribution has memorylessness property [8, chapter 15], meaning that formally,

$$P(X > a + b | X > a) = P(X > b) \tag{11}$$

It is easy to prove that this equality holds using the cumulative mass function.

4) Poisson Distribution: The random variable X has a Poisson distribution when it refers to the number of events happening in a specified unit interval such as a minute or a room with a specified size, in which the mean rate of events(λ) that are occurring is constant [8, chapter 17]. Poisson distribution has a wide range of applications in queuing systems. The probability mass function of Poisson distribution is as follows:

$$p_X(x) = \lambda^x \frac{e^{-\lambda}}{x!}, \text{ for } x \in N \tag{12}$$

The mean of X is $\mu_x = \lambda$, and the variance is $\sigma_x^2 = \lambda$ [8, chapter 17]. As mentioned above in exponential distribution, there is an interesting relation between Poisson distribution and exponential distribution. As found above, the average rate of events in unit interval is $\mu_x = \lambda$, implying that the interval between events is exponentially distributed with a probability density function $g(x) = \lambda e^{-\lambda x}$ [8, chapter 17].

There are many other discrete probability distributions such as negative binomial distribution, hyper-geometric distribution, multinomial distribution etc. which we couldn't cover here. Thompoulos [8] provides a relatively detailed overview for discrete distributions, providing use cases and important formulas without diving too much into detail. Also, Hoel, Port and Stone [10, chapter 3] has mathematical formality but lacks use cases. DasGupta [11, chapter 1], on the other hand offers a more balanced approach between formality and applications.

C. Probability Distributions in Data Mining

Probability distributions play a central role in data mining, both in understanding data and developing algorithms. A straightforward illustration is about inferring information about the heights of the male primary school students in a city. There are tests that determine the normality of a population.

[2] File:Binomial Distribution.svg, `https://commons.wikimedia.org/wiki/File:Binomial_Distribution.svg`

If such tests yield a positive result for the normality, we can estimate mean and variance, which fully parameterize the normal distribution. Then the practitioners or software can detect height anomalies in school children. Another illustration is linear regression. The linear regression model assumes that the output variable y is equal to the linear function applied to x with Gaussian noise, that is,

$$y = \langle x, \theta \rangle + z, \ \textit{where } z \sim N\left(0, \sigma^2\right) \tag{13}$$

A linear regression model may fail in the learning task, for example, if the noise is exponentially distributed. Without understanding the importance of Gaussian noise with 0 mean in such a model, a data miner can not fully grasp the reason for the failure.

III. Sampling

We made heavy use of the term population and the fully parametrized probability distributions in the previous section. The real-world applications, however, differ in that population parameters are unknown. We need to estimate population parameters from a small subset of the population in many cases, which we call sample. There are mainly two reasons behind this approach: it may be impossible to list all the population, i.e., the population is infinite, and it may not be feasible to list all the members of a finite population [12, chapter 1]. There are, for example, a finite number of images on the Internet at time t. Nevertheless, collecting them is an impossible task considering our current skill-set, which obliges us to work with a subset of them, a sample.

Sampling is a delicate task and requires great diligence. The sample should represent the population accurately. For example, assume that there is a machine learning task to classify cats according to their breeds. If the data set, which is more or less a sample, does not contain some breeds and over-represent the others, the model to be trained is inherently erroneous. It is crucial to emphasize the principle: "garbage in, garbage out". The data sets and samples should be representative of the population they are drawn from.

A. Terminology

A sampling unit is the entity that a statistician chooses for the sample. The sampling unit is a single entity or may comprise of more than one element. If it contains multiple elements, it is said to be a cluster. A unit may be a cat picture, for example. A sampling frame is the list or enumeration of all sampling units. It is possible that the sampling frame does not contain all the elements in the universe [12, chapter 12]. In a sampling process, for example, all the cat images that we can include in the sampling frame are publicly available ones, i.e, a subset of all cat images. The sample consists of sampling units chosen from the sampling frame. The process of choosing sampling units from the sampling frame is related to sampling design [12, chapter 12]. There are many sampling designs that characterize the sampling process, such as random sampling, systematic sampling, and stratified sampling, etc. We will stick to random sampling in the following subsections since it is one of the most basic and used sampling designs.

Random Sampling: Random sampling has two subcategories: simple random sampling and random sampling with variable probability. For a sampling design to be considered as random sampling, it is enough to satisfy the following equality where x_i's denote sampling units [12, chapter 12].

$$P(x_1, ..., x_n) = P(x_1) \cdot ... \cdot P(x_n) \tag{14}$$

When the probability of choosing each sample unit from the sample frame is the same, then the sampling design has simple random sampling property. In more rigorous way, there is a simple random sampling when,

$$P(x_i) = P(x_j) \ \textit{for all } x_i, x_j \in \textit{Sampling Frame} \tag{15}$$

B. Sampling Distributions

The reason for the sampling in the first place was to gain information about the population parameters. Accordingly, sample mean, sample variance, and sample covariance are unbiased estimators of their population counterparts [12, chapter 12]. An estimator t is unbiased when the real value of the estimation is θ and $E[t] = \theta$. That is,

$$E[\bar{x}] = \mu_x \ \textit{,where } \bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} \tag{16}$$

$$E[s_X^2] = \sigma_x^2 \ \textit{,where } s_x^2 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n-1} \tag{17}$$

$$E[s_{XY}] = \sigma_{XY} \ \textit{,where } s_{XY} = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{n-1} \tag{18}$$

Another important property about sample mean is that the variance of sample mean is

$$\sigma_{\bar{X}}^2 = \frac{\sigma_X^2}{n}. \tag{19}$$

For the proofs of these propositions, please refer to Johnson [12, chapter 12].

Central Limit Theorem: An interesting phenomenon occurs when the sample size n is large: the distribution of $\bar{X}$ approximates the normal distribution [12, chapter 12], which is called central limit theorem. Formally, for sufficiently large n ,

$$\bar{X} \approx \left(\mu_X, \frac{\sigma_X^2}{n}\right) \tag{20}$$

A rule of thumb is that the normal approximation can be applied when $n \geq 30$. Yet, even for smaller sample sizes, if the population is infinite, the approximation works well in the centre of the distribution but may diverge in the other parts [12, chapter 12].

C. Further Reading

There are many other topics that we did not cover thoroughly, such as systematic sampling in sampling designs and student's t-distributions in sampling distributions. We suggest Johnson [12] as a complete reference for sampling. Casella and Berger [13, chapter 5] provide a detailed and rigorous approach to random sampling. DasGupta [11, chapter 5.4] details sampling distributions, a crucial topic on which we could not elaborate.

IV. Estimation

One of the main reasons in the first place for developing sampling methods was to estimate the population parameters. In general, the population parameters are unknown, and the only way to know about the population is by estimating such parameters from a sample drawn from the population. Moreover, many data mining and machine learning methods involve estimators in their learning process, such as maximum likelihood estimators. In fact, point estimators have a wide range of applications in machine learning, while interval estimators are one of the backbones of hypothesis testing, a topic which we will introduce in the following section. Learning to evaluate such estimators is essential to analyze and develop machine learning methods. We will introduce the basic terminology of estimators, maximum likelihood estimators and point estimators in the following subsections.

A. Terminology

An estimator υ estimating the population parameter ϑ is a function of the sample, $\{x_1, ..., x_n\}$ randomly drawn from the population [12, chapter 13]. We assume for the sake of simplicity that x_i's are independently and identically distributed (i.i.d.) random variables, even though it is not a strict requirement in sampling theory. Formally,

$$\upsilon = \upsilon(x_1, ..., x_n) \tag{21}$$

There are mainly two categories of estimators, point estimators and interval estimators. Point estimators estimate a single point, whereas interval estimators estimate an interval [12, chapter 13]. Sample mean, for example, is the point estimate of population mean as we found in the previous section.

An essential notion in estimation is unbiasedness. An estimator is said to be unbiased if the estimator's expected value is equal to the population parameter that is estimated. That is,

$$E[\upsilon] = \vartheta \tag{22}$$

That means, intuitively, on average the estimator estimates the population parameter correctly. Another important term about the estimators is consistency. The consistency of an estimator relates to the behavior of the expected value and the variance of it as $n \to \infty$. An estimator is consistent if σ_υ^2 converges to 0 and $E[\upsilon]$ converges in probability to ϑ as $n \to \infty$ [12, chapter 13]. That is, for all $\epsilon > 0$:

$$\lim_{n \to +\infty} E[|\upsilon - \vartheta| > \epsilon] = 0 \tag{23}$$

B. Evaluation of Estimators

For a population parameter, there may be many estimators. For example, the sample mean and the sample median are both the estimators for the population mean in a normal population [12, chapter 13]. They differ, however, in the effectiveness of the estimation. Moreover, there may be a population parameter for which there is no unbiased estimator. Lehmann and Casella [14, chapter 2] introduce such an example. All in all, we need to have evaluation criteria for assessing the effectiveness of the estimator in question.

To assess the effectiveness of an estimator, we should consider unbiasedness, the variance the estimator and consistency of the estimator in the following ways [12, chapter 13]. An unbiased estimator is preferable to the biased counterpart. This intuitively makes sense because we want estimators to be correct on average. Moreover, when estimators are unbiased, the one with smaller variance is preferable. Again, that makes sense because we want estimators not to fluctuate much. Also, as Johnson [12, chapter 13] suggests, another way of thinking about the smaller variance requirement is in terms of efficiency. He considers, for example, the one with smaller variance as more efficient in the following way:

$$\sigma_{x_{median}}^2 = 1.57\frac{\sigma_x^2}{n} \tag{24}$$

$$\sigma_{\bar{x}}^2 = \frac{\sigma_x^2}{n} \tag{25}$$

When the population in question is normal, both x_{median} and $\bar{x}$ are unbiased estimators of the population mean. However, for a fixed variance of an estimator, the sample size should be bigger for x_{median}. Moreover, the estimators with consistency are preferable, since with growing n, we would like to have better estimates of the population parameter in question [12, chapter 13].

C. Maximum Likelihood Estimators

One estimator widely used in statistics and machine learning is the maximum likelihood estimator. As its name suggests, the main goal is to maximize the likelihood of the sample. Given a sample consisting of $\{x_1, ..., x_n\}$, the likelihood is $f(x_1, ..., x_n; \upsilon)$. In many applications, x_i's are i.i.d., hence $f(x_1, ..., x_n; \upsilon) = f(x_1; \upsilon) \cdot ... \cdot f(x_n; \upsilon)$. The maximum likelihood estimate of population parameter ϑ is:

$$\hat{\vartheta} = \underset{\upsilon}{\operatorname{argmax}}\ f(x_1, ..., x_n; \upsilon) \tag{26}$$

If x_i's are i.i.d., then it is easier to work with the logarithm of the likelihood function because of simplification and computational convenience. Taking the logarithm of likelihood does not change $\hat{\vartheta}$ since the logarithm is a monotone increasing function. $\hat{\vartheta}$ can be calculated using the gradient descent algorithm.

There are remarkable properties of the maximum likelihood estimator that make it desirable for statisticians and data miners. Johnson [12, chapter 13] summarizes them as follows. They are consistent in general, and when the sample size is large, efficient. Also, they are the best in their class, i.e., sufficient, in that no other estimator is estimating more in terms of information than the maximum likelihood estimator. However, the maximum likelihood estimator may be biased. Also, they have approximately a normal distribution for large samples and invariance property. For the proofs and complete definitions, please refer to Johnson [12, chapter 13].

D. Further Reading

We could not cover the interval estimations, which have significant implications for hypothesis testing. In addition to being an easy-to-understand reference to estimators, Johnson [12, chapter 13] covers interval estimations in an informative way. For more rigorous and complete covering, please refer to Lehmann and Casella [14].

V. Hypothesis Testing

Sometimes, the main concern is not directly related to estimating the parameters of the population, but to decision making. A scientist studying coronaviruses, for example, may want to test whether there is any effect of vitamin deficiency on the severity of COVID-19 infection. Since hypothesis testing is not directly related to machine learning or data mining but still important for decision making, we will give basic definitions and leave the rest as further reading.

A. Terminology

Before everything else, a statistician needs to define what hypotheses are. These hypotheses are the null hypothesis, H_0 and the alternative hypothesis, H_1, where the null hypothesis is the status quo and the alternative hypothesis is the refutation of the null hypothesis. Following the COVID-19 example, the null and alternative hypotheses may be as follows:

H_0: Vitamin deficiency has no effect on the severity of COVID-19 infection

H_1: Vitamin deficiency has an effect on the severity of COVID-19 infection

Depending on the outcome of the testing procedure, we may fail to reject the null hypothesis or reject the null hypothesis in favor of the alternative hypothesis. However, the decision procedure may be erroneous. The rejection of the true hypothesis is called a type 1 error, α , while the non-rejection of a false hypothesis is called a type 2 error, β [12, chapter 14]. The rejection of a true hypothesis is intuitively more severe, since the non-rejection of a false hypothesis just preserves the status quo. The maximum probability of committing a type 1 error is defined as the significance level [12, chapter 14]. In certain hypothesis testing schemes, such as in the ones concerned with population mean, the significance level determines the critical regions.

B. Further Reading

Since we could not explain the details of hypothesis testing in the scope of this paper, we introduced the fundamental terms and left the rest as a reading material. We suggest that you read on z-tests and t-tests, two widely used testing schemes. Johnson [12, chapter 14] provides a concise material. Massey and Miller [15] has a more detailed and thorough text, covering even non-parametric tests.

VI. Conclusion

As the previous chapters demonstrate, statistics has much to offer to data mining. Many methods in data mining make use of statistical notions directly or indirectly. A thorough comprehension of data in data mining requires a clear understanding of probability distributions. Additionally, samples in statistics correspond to data sets in machine learning, albeit indirectly. Understanding sampling procedures benefits data miners in the age of big data, where data are vast, resources are comparably limited, and many data sets are biased [16]. It is important to emphasize that some machine learning methods such as logistic regression utilize estimators. Learning the theoretical framework behind estimators is valuable in that it paves the way for the development of new machine learning methods.

The discrepancies between the two fields are noticeable. The mathematical rigor and diligence in statistics differ from the approaches of data mining. However, such discrepancies should not be viewed as discouraging by data miners. On the contrary, these discrepancies are a great source of inspiration. Considering that the ten most commonly used datasets contain label errors [17], data mining may benefit from the more rigorous and diligent approach of statisticians.

References

[1] C. Clifton, "Data mining," Dec 2019, accessed 6 May 2021. [Online]. Available: https://www.britannica.com/technology/data-mining

[2] D. R. Anderson, D. J. Sweeney, and T. A. Williams, "Statistics," Oct 2020, accessed online on 6 May 2021. [Online]. Available: https://www.britannica.com/science/statistics

[3] Y. Benjamini and M. Leshno, "Statistical methods for data mining," in *Data mining and knowledge discovery handbook*. Springer, 2005, pp. 565–587.

[4] D. J. Hand, "Statistics and data mining: Intersecting disciplines," *SIGKDD Explor. Newsl.*, vol. 1, no. 1, p. 16–19, Jun. 1999. [Online]. Available: https://doi.org/10.1145/846170.846171

[5] J. R. M. Hosking, E. P. D. Pednault, and M. Sudan, "A statistical perspective on data mining," *Future Gener. Comput. Syst.*, vol. 13, no. 2–3, p. 117–134, Nov. 1997. [Online]. Available: https://doi.org/10.1016/S0167-739X(97)00016-2

[6] N. M. Adams, "Perspectives on data mining," *International Journal of Market Research*, vol. 52, no. 1, pp. 11–19, 2010. [Online]. Available: https://doi.org/10.2501/S147078531020103X

[7] S. M. Ross, "Introduction to probability theory," in *Introduction to Probability Models (Twelfth Edition)*, twelfth edition ed., S. M. Ross, Ed. Academic Press, 2019, pp. 1–21. [Online]. Available: https://www.sciencedirect.com/science/article/pii/B9780128143469000068

[8] N. T. Thomopoulos, *Statistical Distributions*. Springer International Publishing, 2017.

[9] S. M. Ross, "Random variables," in *Introduction to Probability Models (Twelfth Edition)*, twelfth edition ed., S. M. Ross, Ed. Academic Press, 2019, pp. 23–99. [Online]. Available: https://www.sciencedirect.com/science/article/pii/B978012814346900007X

[10] P. Hoel, S. Port, and C. Stone, *Introduction to Probability Theory*, ser. Houghton Mifflin series in statistics. Houghton Mifflin, 1971. [Online]. Available: https://books.google.de/books?id=2xLvAAAAMAAJ

[11] A. DasGupta, *Probability for Statistics and Machine Learning: Fundamentals and Advanced Topics*, 1st ed. Springer Publishing Company, Incorporated, 2011.

[12] E. Johnson, *Forest Sampling Desk Reference*. CRC Press, 2000. [Online]. Available: https://books.google.de/books?id=AITLBQAAQBAJ

[13] G. Casella and R. Berger, *Statistical Inference*, ser. Duxbury advanced series. Duxbury Thomson Learning, 2002. [Online]. Available: https://books.google.de/books?id=ZpkPPwAACAAJ

[14] E. Lehmann and G. Casella, *Theory of Point Estimation*, ser. Springer texts in statistics. Springer, 1998. [Online]. Available: https://books.google.de/books?id=YElFQwAACAAJ

[15] A. Massey and S. J. Miller, "Tests of hypotheses using statistics," accessed online on 1 June 2021. [Online]. Available: https://web.williams.edu/Mathematics/sjmiller/public_html/BrownClasses/162/Handouts/StatsTests04.pdf

[16] J. Buolamwini and T. Gebru, "Gender shades: Intersectional accuracy disparities in commercial gender classification," in *Proceedings of the 1st Conference on Fairness, Accountability and Transparency*, ser. Proceedings of Machine Learning Research, S. A. Friedler and C. Wilson, Eds., vol. 81. New York, NY, USA: PMLR, 23–24 Feb 2018, pp. 77–91. [Online]. Available: http://proceedings.mlr.press/v81/buolamwini18a.html

[17] C. G. Northcutt, A. Athalye, and J. Mueller, "Pervasive label errors in test sets destabilize machine learning benchmarks," 2021. [Online]. Available: https://arxiv.org/abs/2103.14749