MATHEMATICAL STATISTICS

BENJAMIN KONSTAN

Abstract. This paper serves as an introduction to mathematical statistics. It will precisely define key objects and tools, and then sequentially prove key theorems that take the reader on a journey from those basic definitions to some of the main theorems used in the field of statistics. Every statement along the way will be accompanied by rigorous proof and conceptual discussion, leaving behind a strong understanding of the mechanisms behind applied statistics. Important concepts covered include: Moment-generating functions; Chebyshev's Inequality; Law of Large Numbers.

Contents

1. Introduction

In high school statistics, students are taught t-tests, p-tests, and even chi-squared test. They may be told that "in the long run," a repeated event with probability one half will result in a success half the time, or that for $n \geq 30$, the distribution of sample means approaches a normal distribution. However, all these statements fail spectacularly to demonstrate how statistics is a branch of mathematics. They lack the rigor and precision of mathematical rules, and they are often presented to students without any proof whatsoever.

The sub-field of mathematical statistics is the theory underlying these statements that addresses those issues. In it, rigorous and precise definitions provide the framework for proving all of the elementary statistical nuggets we are taught as fact. This paper will provide an introduction to mathematical statistics, proving from scratch a chain of statements that help us better and more precisely understand probability and statistics.

Date: August, 2019.

2. Laying the Groundwork

We need to begin by laying out a sequence of definitions to frame everything we're going to be talking about. We denote the "Event" or "Outcome" space Ω as the set of possible events; its elements, the elementary events, are denoted by ω. An event is a subset of the event space, denoted $C \subset \Omega$

Example 2.1. Consider a random number generator (RNG) that produces an integer between 1 and k (inclusive). Its event space is $\{\omega_1, \omega_2, \omega_3, ..., \omega_k\}$ where ω_i represents an output from the RNG of integer i.

We can then define a probability measure as a function:

$$P : \Omega \to [0, 1]$$

satisfying several key properties:

- $\forall_{C \subset \Omega} P(C) \geq 0$
- $P(\Omega) = 1$
- $\forall_{C_i \neq C_j} P(C_i \bigcup C_j) = P(C_i) + P(C_j)$

Next, we define a random variable as a function:

$$f : \Omega \to R$$

This is particularly useful, as the values of f form a set χ, and if we have a probability metric P defined over the domain of f, we call P the probability distribution of the random variable f.

Moving forward in our study of random variables, x will often refer to the value of the random variable, and $P(x)$ its probability. However, one last definition we need is that of a distribution function. Let x be a continuous variable. A distribution of this variable will be a function $f(x)$ such that:

$$\int_a^b f(x)dx = P(a < x < b)$$

As R is the event space, it follows from the second property above that $\int_{-\infty}^{\infty} f(x)dx = 1$.

Example 2.2. I now present the binomial distribution. Let's say we have a probabilistic process with two possible outcomes: success and failure, each occurring with probabilities $p \in [0, 1]$ and $q = 1 - p$, respectively. Now consider running this process n times, independently (such as flipping a coin 3 times where heads is considered a success). The number of successes over those n trials is a random variable χ. Its event space consists of all the possible numbers of successes i.e. $0, 1, 2, 3, ..., n$, and each has an assigned probability. Since one of those numbers of successes will take place, it is easy to understand that $P(\Omega) = 1$. We will examine the binomial distribution in greater detail later, but for now let's calculate the $P(\omega = n)$. As you learned in a high school statistics course, the probabilities of independent events multiply. Thus, getting n successes requires multiplying the probability of success, p, n times, thereby getting us p^n.

3. Moment Generating Functions

We define the k^{th} moment of a random variable about the origin as follows:

$$m'_k = \sum_{x=0}^{\infty} x^k P(x)$$

Or the analogous version in the continuous case:

$$m'_k = \int_{-\infty}^{\infty} x^k f(x)dx$$

Clearly, the first moment about the origin is what we commonly call the "mean" or "expected value" of a random variable, m. We can then similarly define the k^{th} moment about the mean of a random variable as:

$$m_k = \sum_{x=0}^{\infty} (x-m)^k P(x)$$

Or, for continuous variables,

$$m_k = \int_{-\infty}^{\infty} (x-m)^k f(x)dx$$

As it turns out, m_2 measures the spread of a distribution about its mean - we call it the variance. Its square root is the standard deviation. Further moments about the mean measure skewness, peakedness, and many other qualities of a distribution. In fact, the distribution of a random variable is completely determined by its moments.

While it is theoretically possible to compute any individual moment of a distribution by carrying out the necessary summation or integration, there is an easier technique called the Moment-Generating Function.

Definition 3.1. We define the moment-generating function for a random variable X with distribution function $f(x)$:

$$M_x(\theta) = \int_{-\infty}^{\infty} e^{x\theta} f(x)dx$$

Or, in the discrete case:

$$M_x(\theta) = \sum_{x=0}^{\infty} e^{x\theta} P(x)$$

To demonstrate that the moment-generating function really generates moments, we can expand the exponential as follows:

$$M_x(\theta) = \int_{-\infty}^{\infty} (1 + x\theta + \frac{x^2\theta^2}{2!} + ...)f(x)dx$$

$$= \int_{-\infty}^{\infty} f(x)dx + \theta \int_{-\infty}^{\infty} xf(x)dx + \frac{\theta^2}{2!} \int_{-\infty}^{\infty} x^2 f(x)dx + ...$$

$$= m'_0 + m'_1\theta + m'_2 \frac{\theta^2}{2!} + ...$$

We can retrieve these coefficiencts through repeated differentiation. It is easily verified that:

$$m'_k = \frac{d^k M}{d\theta^k}|_{\theta=0}$$

It is not hard to see that this holds in the discrete case as well.

Example 3.2. Let's do an example with the normal distribution - i.e. a bell curve.

Definition 3.3. The normal distribution (centered around t, variance 1) is defined by the distribution

$$f(x) = \frac{1}{2\pi} e^{-\frac{(x-t)^2}{2}}$$

We can use this and the above definition to calculate its moment-generating function for the case of $t = 0$:

$$\begin{aligned}
&\frac{1}{2\pi}\int_{-\infty}^{\infty} e^{x\theta} e^{-\frac{x^2}{2}} dx \\
&= \frac{1}{2\pi}\int_{-\infty}^{\infty} e^{-\frac{x^2}{2}+x\theta} dx \\
&= \frac{1}{2\pi}\int_{-\infty}^{\infty} e^{-\frac{1}{2}(x-\theta)^2+\frac{1}{2}\theta^2} \\
&= \frac{1}{2\pi}\int_{-\infty}^{\infty} e^{-\frac{1}{2}(x-\theta)^2} e^{\frac{1}{2}\theta^2} \\
&= e^{\frac{1}{2}\theta^2}\int_{-\infty}^{\infty} \frac{1}{2\pi} e^{-\frac{1}{2}(x-\theta)^2} \\
&= e^{\frac{1}{2}\theta^2}
\end{aligned}$$

The last equality holds because the integrand is just a normal probability distribution function whose integral must be 1 (this can also be verified using either series or change of variables). Thus, the result we get is that:

$$M_x(\theta) = e^{\frac{1}{2}\theta^2}$$

We can use this result to calculate some moments.

$$m'_1 = \frac{dM}{d\theta}|_{\theta=0} = (0)e^{\frac{0}{2}} = 0$$

This tells us that the mean is 0, exactly as we defined it to be.

$$m'_2 = \frac{d^2 M}{d\theta^2}|_{\theta=0} = (0^2+1)e^{\frac{0}{2}} = 1$$

Since our distribution has mean 0, $m'_2 = m_2$, so our variance is 1 just like we expected.

$$m'_3 = \frac{d^3 M}{d\theta^3}|_{\theta=0} = (0^3+3(0))e^{\frac{0}{2}} = 0$$

The third moment - which measures how skewed a distribution is - also leads to zero (the normal distribution has no skew - it's perfectly symmetric!). This pattern reveals an interesting fact about normal distribution - all odd moments obtain a value of 0.

This example was used specifically because the normal distribution is incredibly powerful in the field of statistics. The Central Limit Theorem states that the sum of independent random variables will converge to the normal distribution irrespective of the initial distributions. Moreover, the normal distribution is a powerful tool used to approximate binomial distributions which were discussed earlier and will continue to be useful to us. Before that however, we want to prove one more thing about moments that will come in handy:

Lemma 3.4. *We claim that* $m_2 = m_2' - m_1'^2$.

Proof.

$$m_2 = \sum_{x=0}^{\infty}(x-m)^2 P(x)$$

$$= \sum_{x=0}^{\infty} x^2 P(x) - 2xmP(x) + m^2 P(x)$$

$$= \sum_{x=0}^{\infty} x^2 P(x) + -2m\sum_{x=0}^{\infty} xP(x) + m^2 \sum_{x=0}^{\infty} P(x)$$

$$= m_2' - 2mm + m^2 == m_2' - m_1'^2$$

The last step is clear, as $m = m_1'$ and by definition, $\sum_{x=0}^{\infty} P(x) = 1$. □

4. Moments of a Binomial Distribution

We want to find the first and second moments of the binomial distribution. To do so, we can use its moment-generating function. From the definition above, we know that

$$M_x(\theta) = \sum_{x=0}^{n} e^{\theta x} P(x)$$

To find $P(x)$, we see that in n trials, the probability of k successes is the number of arrangements of k successes (and n-k failures) times the probability of each arrangement i.e. $P(k) = (nCk)p^k q^{n-k}$. Thus, we get:

$$M_x(\theta) = \sum_{x=0}^{n} e^{\theta x} \frac{x!}{x!(n-x)!} p^x q^{n-x} = \sum_{x=0}^{n} \frac{x!}{x!(n-x)!} (pe^{\theta})^x q^{n-x}$$

This is just a binomial expansion. We can rewrite it as:

$$M_x(\theta) = (q + pe^{\theta})^n$$

To find the desired moments, we can differentiate:

$$M'(\theta) = npe^{\theta}(q + pe^{\theta})^{n-1}$$

$$M''(\theta) = npe^{\theta}(q + pe^{\theta})^{n-2}(q + npe^{\theta})$$

Evaluating these at $\theta = 0$, we get $\mu_1' = np$ and $\mu_2' = npq + (np)^2$. Using lemma 1, we then get $\mu_2 = \mu_2' - \mu_1'^2 = npq$. Thus, we have:

$$m = np$$

$$\sigma = \sqrt{npq}$$

5. Theorems

We're now going to use the resources we've built up to prove some useful theorems in probability.

Theorem 5.1. *Chebyshev's Inequality: Assume a continuous distribution function with finite variance. Then,*

$$P(|x-m| > k\sigma) \leq \frac{1}{k^2}$$

.

Proof. Let us start with the integral defining variance (i.e. the second moment about the mean) as:

$$\sigma^2 = \int_{-\infty}^{\infty} (x-m)^2 f(x)dx$$

Given $k > 0$, this can be broken down into three integrals to get:

$$\int_{-\infty}^{m-k\sigma} (x-m)^2 f(x)dx + \int_{m-k\sigma}^{m+k\sigma} (x-m)^2 f(x)dx + \int_{m+k\sigma}^{\infty} (x-m)^2 f(x)dx$$

Examining the middle quantity, we see that the integrand is always positive and that the lower bound is less than the upper bound. Therefore, we can conclude that:

$$\sigma^2 \geq \int_{-\infty}^{m-k\sigma} (x-m)^2 f(x)dx + \int_{m+k\sigma}^{\infty} (x-m)^2 f(x)dx$$

To further simplify this inequality, we can see that the minimum value of the $(x-m)^2$ term takes place at the limit closest to the mean (i.e. the upper limit for the left integral and the lower limit for the right integral). Thus, we get:

$$\sigma^2 \geq \int_{-\infty}^{m-k\sigma} (m-k\sigma-m)^2 f(x)dx + \int_{m+k\sigma}^{\infty} (x+k\sigma-m)^2 f(x)dx$$

This is equal to:

$$(k\sigma)^2 \left[\int_{-\infty}^{m-k\sigma} f(x)dx + \int_{m+k\sigma}^{\infty} f(x)dx) \right]$$

Notice that these two integrals give the probability that x will be at least k standard deviations from the mean (left and right, respectively). Thus, it is equivalent to saying:

$$\sigma^2 \geq k^2\sigma^2 P(|x-m| > k\sigma)$$

This can be written as above:

$$P(|x-m| > k\sigma) \leq \frac{1}{k^2}$$

□

Example 5.2. We'll start with a relatively familiar example with a twist (we don't get to assume a normal distribution!). Let's say that a factory worker suspects that a new machine isn't working as quickly as the others at the factory. She had already measured the average speed down the production line to be 4.5 hours, and that the standard deviation for the different machines (all of which perform the same task -

this is a BIG factory) is about 6 minutes. She then tested this new machine, and found it took a whopping 5 hours. Would she be right, within 95 percent certainty, to claim that there is something wrong with this new machine?

To answer this question, we need to identify how many standard deviations from the mean this new machine was. Since it took 30 minutes longer and the standard deviation is 6 minutes, that means it is 5 standard deviations away from the mean. The upper bound on the probability of this occurring is given by Chebyshev's inequality:

$$\frac{1}{k^2} = \frac{1}{5^2} = .04$$

An event this extreme would only occur at most 4 percent of the time! That means we would expect it not to occur in 96 percent of machines in the distribution, so with 95 percent confidence, we can say something is wrong with this one! She deserves a raise.

Example 5.3. Let's say you're a biologist working in a laboratory with bacteria. You know that this strand of bacteria reproduces an average of 5 times per minute, at a rate that is intrinsic to the individual bacterium. While you don't know the complete shape of the distribution, you know that the standard deviation is .5 babies per minute. What percent, at most, of the population can you expect to reproduce less than 2 or more than 8 times per minute?

This is a straightforward application of Chebyshev's inequality. We're looking at a situation where $k = 6$ since we're looking at cases where the observation is 6 standard deviations from the mean. The $|x - m| > k\sigma$ is equivalent to our "less than 2 or more than 8" condition. We know that its probability is less than or equal to $\frac{1}{k^2}$, or, $\frac{1}{36} = 2.7\%$. Thus, at most 2.7 percent of our bacteria will be in this special category.

Example 5.4. Consider the following: you're a professor writing an exam. You are confident in your ability to ensure that the class mean will be 65 percent, and you have the ability to approximately create the standard deviation for the score distribution. Lastly, you want at least 99 percent of the class to score above a 40. What should you set the standard deviation as?

This question requires two steps: first we must find out how many deviations are guaranteed to be within the 99 percent threshold, and second we must figure out what value the standard deviation must be for that number of standard deviations to get us to 40 percent. The first step requires applying Chebyshev's inequality. We know that 99 percent of observations are contained in the interval leaving 1 percent out; therefore, we set

$$\frac{1}{k^2} = .01 \to k = 10$$

This implies that within 10 standard deviations, we are guaranteed to find 99 percent of all observations. Now we want those 10 standard deviations to correspond to the 25 percentage point difference between the mean, 60, and our lower bound, 40. Dividing this out, we get a standard deviation of 2.5 percentage points.

Theorem 5.5. *Law of Large Numbers: The probability of a sample success ratio differing from its expected value by any amount goes to zero as the sample size approaches infinity*

Proof. Consider n trials of an event whose probability of success in a single trial is p (and whose probability of failure is $q = 1 - p$). Moreover, let x be the random variable representing the success ratio over those n trials. This is just a binomial distribution (with everything divided by n to get a ratio). Clearly, $m = p$ and $\sigma = \sqrt{\frac{pq}{n}}$. We can apply Theorem 1 to see that:

$$P(|x - p| > k\sqrt{\frac{pq}{n}}) \leq \frac{1}{k^2}$$

Now fix any $\epsilon > 0$ and let $k = \frac{\epsilon}{\sqrt{\frac{pq}{n}}}$. The inequality reduces to:

$$P(|x - p| > \epsilon) \leq \frac{pq}{n\epsilon}$$

This shows that for any fixed $\epsilon > 0$, we can take a large enough number of trials n, and the probability of x being ϵ away from the mean p converges to 0. □

Example 5.6. Let's say you're throwing a dinner party for a political fundraising event. You know that some people will show up having eaten, whereas some will want to eat food. From prior experience, you can say that 70%, on average, of your 200 guests will be hungry. Naturally, you decide to cook a guarantee of 140 meals. However, as a statistician, you know that you can calculate how many extra meals you might need, within 95 percent of cases. How many meals should you have in the back, ready to heat up in case you're hosting a hungry audience? Note, you can assume that everyone will show up alone, not having talked to each other (i.e. the variables are independent).

This question is just an exercise of applying the Law of Large Numbers. This is a binomial distribution where a success represents a guest arriving hungry. We know that $p = .7$, $q = 1 - .7 = .3$, and $n = 210$. Since we're concerned with 95 percent of cases, we want $P(|x - p| > \epsilon) \leq .05$. Thus, we can plug in

$$\frac{pq}{n\epsilon} = \frac{.21}{200\epsilon} = .05 \rightarrow \epsilon = .021$$

To see how many meals this is, we multiply the ratio by the number of guests:

$$.021(200) = 4.2$$

We know that the point estimate of 140 meals will be off by less than 4.2 meals in 95 percent of meals! We'll leave it then to your personal philosophy whether to keep 4 or 5 on your back stove.

Example 5.7. Let's say you have a suspicion that a certain coin lands on heads more than it lands on tails. You want to go ahead and actually test this out, so you design an experiment where you'll flip the coin n times, and then if the resulting proportion of heads varies from .5 by more than .01, you conclude that it's biased. How many times should you flip the coin so that you'd only diagnose a fair coin as biased 1 percent of the time? As you may have learned in high school statistics, this is what we call Type 1 Error.

Once again, this is an application of the Law of Large Numbers. We are looking for

$$P(|x - .5| > .01) \leq \frac{pq}{n\epsilon} = \frac{.5(.5)}{.01n}$$

which we want to be less than .01. Thus, we want:

$$\frac{.25}{.01n} = .01 \rightarrow 2500 = n$$

Therefore, if we do 2500 trials, we get the desired precision for our study!

Example 5.8. Let's say you're assigning binary identification codes (sequences of 0s and 1s). You do this by choosing a probability for a 1 and randomly spitting out each digit according to that probability. Moreover, you want to be able to test whether a given sequence is one you assigned. These sequences are of length 100, and you want to be able to say with 95 percent certainty that if a sequence varies from its expected number of 1s by more than 1 element, then that sequence is not one of yours. What should you make the probability of a 1, assuming you want a probability as close to .5 as possible to minimize the likelihood of randomly repeating an identification code?

Following a similar pattern, we're looking for

$$\frac{pq}{n\epsilon} = .05$$

This time, we know that $n = 100$ and $\epsilon = .01$, which sets up the equation:

$$\frac{p(1-p)}{100(.01)} = .05$$

At this point, we can note the symmetry between p and q because there's no fundamental difference between 1s and 0s. Solving the equation above, we get the result that $p = .947$ or $p = .053$. Indeed, in each case, q will take on the other value. Thus, we should set the probability of a 1 to be either .947 or .053.

6. Concluding Thoughts

This paper introduced a theoretical, proof-based approach to statistics. Precise definition and theorem statements were laid out and proven. Fundamental facts of probability were built from the ground up, in what should have been a relatively accessible manner. However, I hope the power of these statements we generated isn't lost on you. With only a few pages and a couple lines of proof, we have built all the machinery necessary to conduct a "p-test." We have been able to put bounds on probability distributions we know very little about, and we have shown rigorously that our intuition was true: when you flip a coin one million times, the probability that the proportion of heads varies by any meaningful amount from one half is very low. These very general statements, proven through calculus alone, were then applied to real-world problems ranging from hosting political fundraisers to handing out exams to working in a scientific lab. I thank you for joining me on this adventure today and hope that you have a slightly greater appreciation for both the rigor of the theory behind seemingly obvious statements and the broadness of application of ostensibly inconsequential theory. Because at the intersection of application and theory, that's where the field of mathematical statistics lies.

7. Acknowledgments*

I would first and foremost like to thank my mentor, Elia Portnoy, for encouraging me to explore my mathematical interests, helping me to understand tougher concepts, and supporting me as I researched and wrote this paper. Secondly, I would like to thank Daniil Rudenko, the instructor of the apprentice lecture, for teaching me the joy of problem solving and amazing me with the depths and interconnectedness of mathematics. Lastly, I would like to thank Aryan Kejriwal for being an amazing friend and answering any mathematical inquiries I had.

References

[1] Paul G. Hoel. Introduction to Mathematical Statistics. Third Printing. 1947.
[2] Yakov G. Sinai. Probability Theory. MGU. 1985.