Statistical Inference and Data Analysis: G0P75b

Lael John

October 18, 2022

Preface

This set of notes deals with SIDA, my second course in statistics and probability. I'm looking to solidify concepts in statistics, and gain some insight into thinking about the correct statistical model(s) to use in a given situation. These notes are also going to be as detailed as I can make them, because they also help in filling in the gaps in my knowledge of probability and statistics. If it seems like I'm stating the obvious, feel free to skip ahead.

Contents

Chapter 1

Statistical Models and Estimators

What are these things? Firstly, statistics deals with stochastic (seemingly random) data, to try and better understand what is happening/why such data was generated in the first place. There's two major tools at a statisticians disposal

- Probability Theory: Gives us a framework/way of thinking about random phenomena, allows us to study interactions between sets of random events

- Random Samples: Gives us a foundation upon which to apply our probability theory knowledge.

1.1 Probability Theory

Probability theory must be studied in some "space" so to speak. We deal with classes of events and how those "events" or classes can potentially distort each other. Further, because we deal with such an arbitrary definition of what events are, it becomes necessary to first transport every event to some portion of the REAL NUMBER interval $[0, 1]$, allowing us to then begin working from a numerical foundation.
That being said, we denote a probability space as $(\Omega, \mathcal{A}, P)$ where the first element denotes the space of all outcomes, the second element giving us what

we will later come to know as a σ-algebra, and the last giving us a probability function/MEASURE.

Remark. Note that when talking about the universe of all outcomes, no mention is made of the fact that those outcomes may or may not be possible. That falls into the realm of our second element, the σ-algebra. This term seems to deal with a class of subsets, each of which can be measured?

Thoughts. The concept of a "measure" is one that is slightly abstract. Such a function returns a sort of volume? Or assigns a number to a certain region of the space we're considering, but that number also seems to be determined by a function that's already on that space (like the definition of the Lebesgue measure.)

1.1.1 Random Variables and Vectors

Now when we talk about a function, $X : \Omega \to \mathbb{R}$, that can be considered a random variable if it is a "measureable" function. Again, take this at face value, but the underlying principle seems to be that elements of a σ-algebra on(in) our target space are pulled back to elements of our σ-algebra in the probability space.
NOW, because X is a random variable, mapping the probability space to the set of measurable subsets in $\mathbb{R}$, we can also see that P induces P_X onto the reals, where $P_X(B) = P(X^{-1}(B)) = P(\{\omega \in \Omega | X(\omega) \in B\})$. Here the second of our 3 equivalent expressions shows us the reality of what this induced probability really is. P gives us the measure of the set that maps to B in the target space. The map that gets us to the target space is X, so that map that we need to use to get the set in the domain is precisely X^{-1}
Because we have access to our random variable now, we can begin to talk about other functions that involve the random variable, namely

1. Cumulative distribution function: $F_X(x) = P(X \leq x) = P(\{\omega \in \Omega | X(\omega) \leq x\})$

2. Density function: $f_X(x) = \frac{dF_X(x)}{dx} = P(X = x) = P(\{\omega \in \Omega | X(\omega) = x\})$ (Note that this only works if X is absolutely continuous? Will need to review why this is the case.)

3. Moment Generating Function: A holdover from using mechanics terms within mathematics (curse you Newton), $M_X(t) = E[e^{tX}]$, i.e. the

weighted average of all the e^{tx} for all x, and a particular t. Note also that this function may in fact, not exist at all, but the next one, our *characteristic function* will, for all random variables.

4. Characteristic Function: Similar to the above except $\phi_X(t) = E[e^{iXt}] = E[cos(Xt)] + iE[sin(Xt)]$

A random vector therefore is basically a collection of random variables from the probability space to the real numbers. Here the probability measure that is introduced becomes the measure of the intersection of the preimages of all the values given by each random variable. Similarly the cumulative distribution function can be defined, though now when we talk about the density function, we take the derivative of the CDF with respect to all our random variables, not just one. (NOTE: some property of higher order derivatives needs to be fulfilled in order for this to happen, needs a little more research). The moment generating function and characteristic functions again, remain the same, with their respective modifications.

Conclusion

In essence when dealing with probability theory, we have a well defined measure that we can apply onto our universal space, and we study the property of that space, with its sigma algebra all while using our trusty known measure. THIS DOES NOT HAVE TO BE THE CASE WITH MATHEMATICAL STATISTICS.

1.2 Mathematical Statistics

When we begin to talk about mathematical statistics, we begin simply with our universal set of data, and we assume various probability models (or spaces in our previous terminology). Thus our statistical model becomes either one of the following

- $(\Omega, \mathcal{A}, \{P_\theta | \theta \in \Theta\})$
- $(\Omega, \mathcal{A}, \{F_\theta | \theta \in \Theta\})$

Here Θ denotes the parameter space, within which θ varies, and P, F correspond to various probability measures or density functions respectively.

1.2.1 Types of Statistical Models (based on parameter space)

One way of classifying our statistical models is with the relative size or properties of our parameter space.

1. Parametric statistics: When Θ has finite dimension (i.e. when it can be thought of as a vector space, each n-dimensional vector corresponding to n different parameters to be used to create the current probability measure in question).

2. Non-parametric Statistics: Here, stupidly almost Θ must have an infinite dimension. Therefore it makes no sense for us to even think about specifying a θ. In this case, the different measures correspond to different sequences (countably infinite) or functions (possibly uncountably infinite).

3. Semi-parametric: What you get when you decide that some parameters can be specified in finite dimensions, but you need to append an infinite dimensional parameter anyway.

1.2.2 Statistical Inference

Inference, or figuring out what θ seems to have generated the data that we're looking at. Looking at what we've covered already, we're studying a particular value of our statistical model in consideration, knowing that we already have some data to work with. In essence, given a large family of models, we're trying to find a particular parameter that seems closest to reality. (This is a little confusing, given that when we want to talk about reality, we are also using another probability function? even if it's an approximation??? You're basically approximating an approximation, what is happening in the world.) To end up being sucessful in our process of inference, we must necessarily make some estimates about the state of our data. Those estimates (or estimators that generate our estimates) can be classified simply into the following categories:

- Point Estimators: Where we're trying to find a relatively good approximation of θ.

- Confidence intervals: Trying to determine some subset of Θ, that with a reasonably high confidence, we can say θ lies.

- Hypothesis test: Basically partition the space into two, and classify θ as lying in one part of the space, or its complement.

1.3 Random Samples

A random sample, is a collection of *independent* random variables, all having the same distribution as X. These are then called individually independent random variables. (Essentially a tuple, each using the same map on the coordinates.) The way one writes a random sample $(X_1, X_2 \dots X_n)$
A *statistic* is now a function from a random sample $(X_1, X_2 \dots X_n) \to T(X_1, X_2 \dots X_n)$. Now if we go from an n-dimensional universe of outcomes (n different universes of outcomes) to a simple real number, then $T(X_1, X_2 \dots X_n)$ is simply a random variable. If however, we go to p-dimensional real space, then each function $T_i(X_1, X_2 \dots X_n)$ from $i = 1 \dots p$ is component of $T(X_1, X_2 \dots X_n)$, where each is a random variable, resulting in a random vector.

1.3.1 Point Estimators

Now when we consider a statistic $T_n = T(X_1, X_2 \dots X_n)$ which for the random sample $(X_1, X_2 \dots X_n)$ in consideration, helps us approximate θ, it becomes known as an *estimator* of θ. A specific value of T_n i.e. t_n is then what gets called an estimate of θ.
It may be helpful to consider some examples that help illustrate this concept.

1. $\theta = E[X]$ and we consider our estimator to be $\overline{X_n} = \frac{1}{n}\sum_{i=1}^{n} X_i$, the sample mean.

2. $\theta = \mathrm{Var}[X] = E[(X - E[X])^2]$ and our estimator to be $S_n^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X_n})^2$, the sample variance.

In general though, one could have multiple estimators for a particular value of θ, which brings us to the question of how we find an estimator that is preferable over others in consideration.

1.4 Performance Measures

Once we do have a statisctical model on a particular random sample, we look to see how our estimators behave given certain performance measures. These could be any of the following

- **Bias**: Unsurprisingly, this tells us how accurate our expectation of the estimator is i.e. How far away from the true value of θ our estimator seems to be. Here $b_\theta(T_n) = E_\theta[T_n(\mathbf{X})] - \theta$. Here b_θ is our performance measure for each θ, and we look to minimise it.
  We end up calling our estimator "unbiased" if $b_\theta(T_n) = 0 \forall \theta \in \Theta \iff E_\theta[T_n(\mathbf{X})] = \theta$.

- **Mean Squared Error**: This is a much more straightforward function, essentially returning the mean of the sum of squared deviations from the actual data. (Squaring allows us to work only with positive values, and look at the cumulative deviation from the sample.). $MSE_\theta = E_\theta[(T_n(\mathbf{X}) - \theta)^2]$

- **Mean Absolute Deviation Error**: Rather than introducing a squaring term as in the previous measure, we could also just consider the absolute values of all the deviations. The formula for this becomes pretty simple $ABS_\theta = E_\theta[|T_n(\mathbf{X}) - \theta|]$

- **General Expected Loss/Risk**: Such a function assigns a loss function to each estimator, parameter pair, and the goal then becomes to minimise the expectation of such a function. $R_\theta(T_n) = E_\theta[L(T_n(\mathbf{X}), \theta)]$

Remark. Here the goal in all these examples is to minimise the value of the functions. In particular, the general expected risk function allows for the definition of Loss to mirror either the mean absolute deviation, mean squared error. However, it could also set a threshold for acceptable levels of loss, before considering all those models where loss would be higher than that set threshold.

1.4.1 Asymptotic properties of a sequence of statistics

When considering a sequence of statistics, we try and look at the asymptotic properties of such a sequence. In particular, if $< T_n = T(X_1, X_2 \dots X_n) >$ then when $n \to \infty$

- T_n is said to be asymptotically unbiased, if $\text{lt}_{n\to\infty}(b_\theta(T_n)) \to 0$. Basically, as $n \to \infty$, the expected value of the estimator tends to the actual value of the parameter.

- T_n is weakly consistent, if the estimator in consideration converges in probability to the actual value of our parameter, for all values of θ. Mathematically,
$$\forall\theta \in \Theta, T_n \to^P \theta, n \to \infty$$
Remark. Here note that convergence in probabilty simply states that for a given value of ϵ, the probability of a member of a sequence of random variables differs at most by ϵ from some fixed random variable, as $n \to \infty$

- T_n is strongly consistent if
$$\forall\theta \in \Theta, T_n \to^{a.s} \theta, n \to \infty$$
Remark. Almost sure convergence is defined by the fact
$$P(\{s \in S| \lim_{n\to\infty} X_n(s) = X(s)\}) = 1$$
where S is the sample space. The probability of, for each outcome, the sequence of random variables tending to a fixed random variable, is exactly 1. This chain implies weak consistency.

- Finally, T_n can be said to be Mean Square Consistent, if the Mean Square Error (for an arbitrary value of θ) of T_n tends to 0 as $n \to \infty$.

1.4.2 Asymptotic Normality

We're now looking at the behaviour of certain sequences of statistics, and seeing if they resemble normal distributions as $n \to \infty$. What fresh hell is this.

Definition. A univariate estimator T_n for $\theta \in \Theta$ is said to be *asymptotically normally distributed* if, $\forall\theta \in \Theta, \exists V_\theta > 0$ such that
$$\sqrt{n}(T_n - \theta) \to^D N(0, V_\theta), n \to \infty$$
Here V_θ is called the asymptotic variance of T_n, and $T_n \approx N(\theta, \frac{V_\theta}{n})$

Remark. Here to understand such a definition, when trying to estimate a point value for our parameter, we check the difference between the estimator and the sample value as we go further along the sequence. As that starts to resemble a normal distribution with mean 0 and some positive variance, the defnition is satisfied. The symbol symbolises a convergence in distribution.

Remark. We can extend such a definition to multivariate estimators, where now instead of considering just one V_θ, instead we look for a positive-definite, symmetric matrix Σ_θ, called the asymptotic variance-covariance matrix as $n \to \infty$ if it exists.

If T_n is asymptotically normal, then we consider $\sqrt{n}(T_n - \theta)$ to be bounded in probability.

1.4.3 Functions of estimators

Now that we have a basic understanding of estimators and how to measure their performance and long term behaviour, we can begin to talk about functions on given estimators. Very simply put, if T_n is an estimator of θ, and g is a function, then $g(T_n)$ is an estimator of $g(\theta)$.

The *Delta Method* tells us that given an asymptotically normal univariate estimator for the parameter, and g being differentiable on $\mathbb{R}$, never 0, then

$$\sqrt{n}(g(T_n) - g(\theta)) \to^D N(0, g'(\theta)^2 V_\theta)$$

Remark. Note though, that V_θ does seem to depend on θ for an asymptotically normal estimator. Now the question arises wheter we can find an estimator for which this is not the case, i.e if there exists a function g such that when applied on our estimator, the resulting estimator tends in distribution towards a normal distribution around 0, with a constant positive variation, for all θ. Clearly this means $g'(\theta)^2 V_\theta = c^2$ which boils down to solving the differential equation $\frac{dg(\theta)}{d\theta} = \frac{c}{\sqrt{V_\theta}}$

In similar fashion, we can also talk about a function of a multivariate estimator being asymptotically normal, and a similar remark also follows.

1.5 Optimal Estimators

When considering a model $(\Omega, \mathcal{A}, \{F_\theta | \theta \in \Theta\})$, then how does one begin to find a 'best' estimator for θ, based on a random sample given? In general

the answer would be to either remove bias entirely, or minimise one of the performance measures that deals with error. In particular, this might also be a minimisation of our general expected loss/risk function.

Such a minimisation may not always be possible, hence our restricting the class of estimators in consideration, to better arrive at an answer.

1.5.1 UMVUE (Uniform Minimum Variance Unbiased Estimator)

When considering the class of unbiased estimators, with a finite variance, then an estimator is said to be an UMVUE of θ if

- $E_\theta[T_n] = \theta \forall \theta \in \Theta$
- For any other S_n unbiased estimator of θ, $\text{Var}_\theta[T_n] \leq \text{Var}_\theta[S_n]$

A characterisation of such an UMVUE given by CR Rao tells us that if we consider an unbiased estimator with finite variance, for all parameter values, then it is an UMVUE $\iff$ $E_\theta[T_n U_n] = 0 \forall \theta \in \Theta$, and $\forall U_n$, an unbiased estimator of 0 with finite variance.

An UMVUE is unique (slightly obvious)

1.5.2 BLUE (Best Linear Unbiased Estimators)

Sometimes when considering classes of estimators, we can restrict our set even further to just those estimators that are unbiased, and a linear combination of the observations. T_n is a BLUE if

- $E_\theta[T_n] = \theta \forall \theta \in \Theta$
- T_n is a linear estimator ($T_n = \sum_{i=1}^n c_i X_i, c_i \in \mathbb{R}$)
- For any other linear unbiased estimator, the variance of T_n is lesser or equal to it, for all values of θ.

Remark. UMVUE $\implies$ BLUE. Also, there do exist many distributions where it may not be possible to find linear unbiased estimators at all.

1.5.3 Minimax Estimators

We could also search for the estimators whose maximal possible risk is the least. Such an estimator becomes a Minimax Estimator.

1.5.4 Bayesian Estimator

In the Bayesian framework, here our parameters are also random variables/vectors, distributed over the space Θ. First a prior distribution is assumed (based on past data/current beliefs, but an assumption is made regardless.)