Statistical Inference: A Bedtime Story[*]

Linxi Cindy Zeng

December 8, 2022

"In the middle of the journey of our life, I came to myself, in a dark wood, where the direct way was lost. It is a hard thing to speak of, how wild, harsh and impenetrable that wood was, so that thinking of it recreates the fear. It is scarcely less bitter than death: but, in order to tell of the good that I found there, I must tell of the other things I saw there." (Dante, *The Divine Comedy - Inferno: Canto I*)

1 Preface

Its presence befalls you. You don't want *it.* You want something *in* it, something you can only get *from* it - to understand what begets it.

In the realm of statistical inference, "it" is data, and that something is information about the parameters (which are assumed to govern the data generating process under the parametric assumption).

2 The Chisel and the Luminite - Principles of Data Reduction: Sufficient, Ancillary, and Complete Statistics, and The Likelihood Principle

Before you is a boulder. Luminite, an indigenous gemstone on Cind's planet that glows in the dark, is contained in the boulder yet mingled with useless rocky substance that we call "impurities." You carry on your shoulder a bag of disposable chisels. You want to chip away the impurities and find the luminite - your light in the dark.

The boulder here is the sample ($X = X_1, ..., X_n$). A chisel is a statistic ($T(X)$). And the luminite is the part of the data that contains information about the parameters (θ). A statistic defines a form of data reduction and is a partition of the sample space.

2.1 The Sufficiency Principle

A sufficient statistic chips away some rock and preserves all the luminite, i.e., it achieves some summarization of the data without harming the information about θ contained in it. Formally, $T(X)$ is a sufficient statistic for θ if $P(X|T(X))$ is independent of θ, i.e., $\frac{p(x|\theta)}{q(T(x)|\theta)}$ is independent of θ. Any inference about θ should depend on the sample X only through $T(X)$, and this is embodied by the factorization theorem:

A statistic $T(X)$ is a sufficient statistic for θ iff $\exists g(t|\theta)$ & $h(x)$ $s.t., \forall x$ & $\theta, f(x|\theta) = g(T(x)|\theta)h(x)$.

Once $T(X)$ is found, the original sample can provide no new information regarding θ and is thus rendered negligible. Poor data, exploited then abandoned.

A more-chiseled chunk containing all the luminite ensures that a less-chiseled chunk contains all the luminite as well: $r(T(X))$ is sufficient $\Rightarrow T(X)$ is sufficient.

When no more chipping is possible without harming the luminite in the boulder, the composite of all the chisels you have used, represented by a single chisel (a composite function), is called the "Minimal Sufficient Statistic (MSS)." $T(X)$ is an MSS if, for any other sufficient statistic $T'(X)$, $T(x)$ is a function of $T'(x)$.

An alternative way to think about this chipping-away-until-you-can't business is to return to the interpretation of a statistic being a partition of the sample space. Instead of a boulder, you are now confronted by a huge Schwarzwälder Kirschtorte (Black Forest Cake, to shove it into our context) with different kinds of berries as the topping, and you use your chisel to slice it into pieces (after sterilizing the chisel, of course). Your objective is to slice the cake in such a way that each slice has the same kind of berry topping, so that you only need one bite to know what how slice tastes, which is desirable for a busy Gourmet-of-the-World like you. Let $\mathcal{T} = \{t : t = T(x)$ for some $x \in \mathcal{X}\}$ be the image of $\mathcal{X}$ under $T(x)$. Then $T(x)$ partitions the sample space into

[*] This note is an extension of Prof. Jianfeng Mao's lectures and notes for STA3020 Statistical Inference at CUHKSZ in Fall 2022 and its textbook Statistical Inference (Casella & Berger, 2001). The narration is mostly original but also much inspired by Prof. Mao's lectures (e.g., Prof. Mao used the analogy of stone & gold and the phrase "KPI" in class). Suggestions and corrections are welcome - please email the author at linxizeng@link.cuhk.edu.cn.

sets $A_t, t \in \mathcal{T}$, defined by $A_t = \{x : T(x) = t\}$. If T is a sufficient statistic, each A_t shares the same conclusion of θ.

Suddenly, a little sprite springs onto the cake and proudly declares that she feeds on "what is not there," which is in our case, the voids left by the activity of your chisel. She is the reverse of your work. You separate, she unites. Any function $s(T'(X))$ of a statistic can be viewed as such a sprite. A sufficient statistic $T(X) = s(T'(X))$ is an MSS if our sprite s has eaten all borders that could be eaten to still preserve the homogeneity of berries on each slice. The partition associated with an MSS is the coarsest possible partition for a sufficient statistic.

$T(X)$ is an MSS iff "$\frac{f(x|\theta)}{f(y|\theta)}$ is independent of $\theta \Leftrightarrow T(x) = T(y)$. The "$\Leftarrow$" guarantees that $T(X)$ is sufficient, and "$\Rightarrow$" guarantees that $T(X)$ is a function of any sufficient statistic. "$\Leftarrow$" draws correct borders (same berries on the same slice), "$\Rightarrow$" makes sure the borders aren't too many (different berries on different slices).

2.2 Ancillary Statistics

Returning to our boulder, it is possible for you to chip away all the luminite and bestow upon yourself pure impurities. Formally, $S(X)$ is called an ancillary statistic if $P(S(X)|\theta)$ is independent of θ. Alone, an ancillary statistic contains no information about θ. However, when it is used in conjunction with other statistics, it sometimes can increase the precision of inferences about θ. It relies on others to shine.

2.3 Complete Statistics

The luminite can do more than glow. It is able to possess and to suffuse a (chunk of a) boulder, like a phantom does an opera. Formally, let $f(t|\theta)$ be a family of pdfs or pmfs for a statistic $T(X)$. $T(X)$ is a complete statistic if $E(g(T)|\theta) = 0, \forall\theta \in \Theta \Rightarrow P(g(T) = 0|\theta) = 1, \forall\theta$. In words, $T(X)$ is completely determined by θ. The information of θ cannot be removed from $T(X)$ by applying any $g(\cdot)$ unless g is a constant function. If you want to remove luminite, you'd have to throw everything away because everything is already luminite. Geometrically, completeness means if a vector $g(T)$ is orthogonal to the pdf $f_\theta(T)\forall\theta$, then $g(T) = 0$, i.e., the functions $f_\theta(T)$ for varying θ span the whole space of functions of T.

In a somewhat opposite fashion to sufficient statistics, a less-chiseled chunk containing only luminite ensures that a more-chiseled chunk contains only luminite as well: $T(X)$ is complete $\Rightarrow r(T(X))$ is complete.

No ancillary statistics can be constructed based on a complete statistic. And a sufficient statistic, if complete, would be independent of any ancillary statistic (Basu's Theorem).

A complete and sufficient statistic (CSS) must be an MSS. But an MSS doesn't have to be a CSS because a CSS might not exist. However, if it does exist, it must be an MSS (Bahadur's Theorem). It's all or nothin'.

With the terms (and hopefully the intuition of) "sufficient statistics," "ancillary statistics," and "complete statistics" in mind, you now have a better sense of the chisels you carry. Knowing when to use what make them weapons at your call, otherwise they are just weights.

2.4 The Likelihood Principle

Recall how it all began. Before you is a boulder, the sample $(X = X_1, ..., X_n)$, containing luminite, information about the parameters (θ). You have a bag of chisels, statistics $(T(X))$, and among them is a critical one called the likelihood function, which is defined by $L(\theta|X) = f(X|\theta)$ where $f(X|\theta)$ is the joint pdf or pmf of the sample X. Your focus is now shifted to θ (how likely is it?), given X.

The likelihood principle states that if two sample points have only proportional likelihoods $(L(\theta|x) = C(x,y)L(\theta|y), \forall\theta)$, they contain equivalent information about θ. Write $\frac{L(\theta|x)}{L(\theta|y)}$, catch a glimpse of a sufficient statistic criterion?

3 Point Estimation: Methods of Finding and Evaluating Estimators

3.1 MoM, MLE, and Bayesian Estimators

Previously we have focused on $X \xrightarrow{\text{data reduction}} T(X)$. Now we go one step further and one step closer to our dream: $X \xrightarrow{W(X)} \hat{\theta}$. A point estimator is any function $W(\cdot)$ of the sample, and any statistic is a point estimator. $T(X)$ and $W(X)$ are essentially the same thing but... new name, new journey!

You have the sample X (flesh, no skeleton, for the mundane), you "know" the form of the DGP (skeleton, no flesh, for the divine), and moments can be used to describe random variables. Then a straightforward way to infer about θ is to equate the population moments with the sample moments, and solve the simultaneous equations.

$$m_k = \frac{1}{n}\sum_{i=1}^{n} X_i^k \approx \mu'_k(\theta_1, ..., \theta_d) = EX^k, k = 1, 2, ..., d$$

Method of Moments (MoM) has many shortcomings despite its innocent appearance. More popular is Maximum Likelihood Estimators (MLE). Recall how likelihood is defined by

$$L(\theta|x) = L(\theta_1, ..., \theta_k|x_1, ..., x_n) = \prod_{i=1}^{n} f(x_i|\theta_1, ..., \theta_k)$$

Naturally we want to find a $\hat{\theta}$ that is *most likely* to have been *the* θ that generated this boulder before us. Therefore, it is no surprise that the MLE is defined to be $\hat{\theta}(X) = \arg\max_{\theta\in\Theta} L(\theta|X)$. Then everything is reduced to an optimization problem. The invariance property of MLEs is merely a ball-passing result (If $\hat{\theta}$ is the MLE of θ, the MLE of $\tau(\theta)$ is $\tau(\hat{\theta}), \forall\tau(\cdot)$.

In the minds of frequentists, θ is an unknown constant. This is the mindset we have been swimming in until now. Bayesians, on the other hand, view θ as a random variable that reflects belief. Therefore, in addition to data X and the form of underlying distribution $f(x|\theta)$, we have one more piece of raw material to base our inference upon: prior belief of θ, $\pi(\theta)$. Posterior distribution:

$$\pi(\theta|x) = \frac{f(x|\theta)\pi(\theta)}{m(x)} = \frac{f(x|\theta)\pi(\theta)}{\int f(x|\theta)\pi(\theta)d\theta}$$

Since a sufficient statistic $T(X)$ contains all the information about θ in the data, $\pi(\theta|x) = \pi(\theta|T(X))$.

3.2 Methods of Evaluating Estimators

Now you have proposed some estimators that at least make some sense - you have created a reservoir of eligible candidates. But which one should you choose? More importantly, how and why should one be chosen? Recall your ideal image of $\hat{\theta}$, who derives her sole meaning of existence from some other being, θ.

Firstly, you want $\hat{\theta}$ to be as close to θ as possible. How can closeness be measured? You take the difference, formally, this is called "bias." The bias of a point estimator W of a parameter θ is the difference between the expected value of W and θ; that is, $Bias_\theta W = E_\theta W - \theta$. If $E_\theta W = \theta, \forall\theta$, we say W is unbiased.

Secondly, you want $\hat{\theta}$ to be as close to itself EW as possible. Lower volatility, higher reliability. The variance of the estimator addresses this concern.

Formally, after the data $X = x$ are observed, where $X \sim f(x|\theta), \theta \in \Theta$, a decision regarding θ is made. The set of allowable decisions is the action space, denoted by $\mathscr{A}$. Loss function is a nonnegative function that generally increases with the distance between an action, a, and θ. More is lost when more is missed. Two commonly used loss functions are:

$$\text{absolute error loss}, L(\theta, a) = |a - \theta|$$

$$\text{squared error loss}, L(\theta, a) = (a - \theta)^2$$

You can tinker with the functional form of L to penalize as you wish.

In decision theoretic analysis, the quality of an estimator, $\delta(x)$, is quantified by its risk function,

$$R(\theta, \delta) = E_\theta L(\theta, \delta(X))$$

i.e., at a given θ, the risk is the average loss that will be incurred if the estimator $\delta(x)$ is used.

MSE, which will be extensively used, is an example of a risk function w.r.t. the squared error loss $R(\theta, \delta) = E_\theta L(\theta, \delta(X)) = E_\theta(\theta - \delta(X))^2 = Var_\theta\delta(X) + (Bias_\theta\delta(X))^2$. A decision theoretic analysis would be comprehensive in that both the variance and bias are incorporated in the risk and will be considered simultaneously.

However, often times it is easier to deal with stuff one by one instead of everything all at once. If you can first shrink our candidate pool to only unbiased estimators (this is really not too much to ask for from an estimator), you can then make them compete against each other and select the one with the smallest variance. This is the idea behind best unbiased estimators. Formally, W^* is a best unbiased estimator of $\tau(\theta)$ if it satisfies $E_\theta W^* = \tau(\theta), \forall\theta$ and, for any other estimator W with $E_\theta W = \tau(\theta))$, we have $Var_\theta W^* \leq Var_\theta W, \forall\theta$. W^* is also called a uniform minimum variance unbiased estimator (UMVUE) of $\tau(\theta)$.

But how do you know if the variance has reached its minimal? It would be wonderful if we can specify a theoretic lower bound which, once attained, can be guaranteed as the lowest. Thus, the Cramer-Rao Lower Bound (CRLB) is born. Let $X_1, .., X_n$ be a sample with pdf $f(x|\theta)$, and let $W(X) = W(X_1, ..., X_n)$ be any estimator satisfying

$$\frac{d}{d\theta}E_\theta W(X) = \int_X \frac{\partial}{\partial\theta}[W(x)f(x|\theta)]dx \qquad \text{and} \qquad Var_\theta W(X) < \infty$$

Then

$$Var_\theta(W(X)) \geq \frac{(\frac{d}{d\theta}E_\theta W(X))^2}{E_\theta((\frac{\partial}{\partial\theta}\log f(X|\theta))^2)}$$

The denominator of CRLB, $E_\theta((\frac{\partial}{\partial\theta}\log f(X|\theta))^2)$, is called Fisher information of the sample - the more information you have about θ, the smaller the variance can possibly be.

But it is not always the case that the CRLB can be attained. Take a step back, if the best cannot be immediately found, is there a way to at least be better? Yes. Let W be any unbiased estimator of $\tau(\theta)$, then conditioning W on anything with a result independent of θ is an improvement (it sucks information about θ from what it conditions on). Sufficient statistics should come to mind. If T is a sufficient statistic for θ, let $\phi(T) = E(W|T)$, then $E_\theta\phi(T) = \tau(\theta)$ and $Var_\theta\phi(T) \leq Var_\theta W, \forall\theta$; that is, $\phi(T)$ is a uniformly better unbiased estimator of $\tau(\theta)$ (Rao-Blackwell Theorem). So the best must be an estimator based on a sufficient statistic. It can be shown that the best unbiased estimator of $\tau(\theta)$ is unique. Therefore the best unbiased estimator must be based on the "best" of sufficient statistics (MSS).

So an estimator can be improved upon by conditioning on something that contains all the information about θ. What if it can be improved upon something that contains no information about θ? Then it is definitely not the best. $E_\theta W = \tau(\theta), E_\theta U = 0 \Rightarrow W$ is the best unbiased estimator iff $Cov_\theta(W, U) = 0$.

Now, what if a family $f(x|\theta)$ has no unbiased estimator of 0 (other than 0 itself) in the first place? Then if $E_\theta W = \tau(\theta)$, W would automatically be the best unbiased estimator. This whole business with 0... doesn't it remind you somewhat of complete statistics?

Yup, any estimator $\phi(T)$ based on a CSS T is the unique best unbiased estimator of its expected value.

The Bayesian approach again would incorporate the prior-posterior framework. For $\mathbf{X} \sim f(\mathbf{x}|\theta)$ and $\theta \sim \pi$, BayesRisk=$\int_\Theta R(\theta,\delta)\pi(\theta)d\theta = \int_\Theta(\int_\mathcal{X} L(\theta,\delta(\mathbf{x}))f(\mathbf{x}|\theta)d\mathbf{x})\pi(\theta)d\theta = \int_\mathcal{X}[\int_\Theta L(\theta,\delta(\mathbf{x}))\pi(\theta|\mathbf{x})d\theta]m(\mathbf{x})d\mathbf{x}$, where the stuff in [] is called "posterior expected loss (or simply, BayesLoss in the course)." It is a function only of $\mathbf{x}$ (or δ, your decision), and not a function of θ. So BayesRisk can be viewed as the average performance of δ. Thus, for each $\mathbf{x}$, if we choose the action $\delta(\mathbf{x})$ to minimize the posterior expected loss, we will minimize the Bayes risk. We have a clear cut, whereas a clear cut may not be achieved under a frequentist framework.

4 Cinderella and the Glass Slippers - Hypothesis Testing: Methods of Finding and Evaluating Tests

4.1 Finding Hypothesis Tests

Previously we have been talking about point estimation: coming up with a point estimator (statistic) $W(X)$ that is stably close to θ. Now we introduce another inference method, hypothesis testing. (At this point I asked myself, what exactly do they mean by "another"? Something distinctly parallel? If so, how is hypothesis testing different from point estimation? Then hours went by. What I wrote I deleted. And I decide that it does not behoove me to answer this question - I know too little and can say even less. All I can say is, point estimation feels like the behavior of a starving cat - it is a direct hunt - and hypothesis testing that of a playful cat - she is more interested in whether she is able to catch it than actually catching it. Point estimation asks, "what is it?" And hypothesis testing asks, "is it this(/here)?" To find and to prove are different motives, but in no way am I saying one is more subjective than the other, as the terms might suggest. These are profound questions that I just can't begin to answer. And it is not always advisable to let them occupy the mind.)

Two roads diverged in the dark wood, and you took the other one. You come to a pair of glowing glass slippers, hidden beneath a bush. You are the king's son, remembering Cinderella's hurried look. Forlorn was that encounter, and Cinderella shall be, who fits these slippers good.

The glass slippers is the sample ($X = X_1, ..., X_n$). Cinderella is the θ (an unknown constant, according to the frequentists) you try to find. All the girls in town constitute the parameter space Θ. You take these slippers and go from door to door to let each girl ($\forall\theta \in \Theta$) try them on and see how they fit: the better the fit, the higher the likelihood $L(\theta|x)$.

The town is divided into two districts: O Square and I Street. Since you want to take it slow, it currently suffices to just locate her district. You sincerely believe (which is better, believe or hope) Cinderella lives on I Street for a secret reason. You posit a null and an alternative hypothesis, "Cinderella lives on O Square," and "Cinderella lives on I Street":

$$H_0 : \theta \in \Theta_0$$

$$H_1 : \theta \in \Theta \setminus \Theta_0$$

You want to convince your council that Cinderella indeed lives on I St. The question is what criteria of rejection of H_0 can you specify to make your belief reasonable (to both yourself and your council).

There is only one real Cinderella, just like there is only one θ from the frequentist viewpoint. Therefore if you need to come up with a measure to represent the performance of each district (recall that your primary interest at this stage is not Cinderella's exact location, but her district), it would be $\max L$, not average, not mode. So define $L_{H_0} = \max_{\theta\in\Theta_0} L(\theta|x)$, $L_{H_1} = \max_{\theta\in\Theta_0^c} L(\theta|x)$. If $L_{H_1} >> L_{H_0}$, your rejection of L_{H_0} would be quite reasonable. Calculating the quotient is the easiest way to represent relative magnitude, thus $\lambda(\mathbf{x}) = \frac{\max_{\theta\in\Theta_0} L(\theta|\mathbf{x})}{\max_{\theta\in\Theta} L(\theta|\mathbf{x})}$, this is formally defined as the likelihood ratio test statistic. Then a likelihood ratio test (LRT) is any test that has a rejection region of the form $\{\mathbf{x} : \lambda(\mathbf{x}) \leq c\}$, where $0 \leq c < 1$. Why do we have $\max_{\theta\in\Theta} L(\theta|\mathbf{x})$ instead of $\max_{\theta\in\Theta_0^c} L(\theta|\mathbf{x})$? Because $P(H_0)+P(H_1) = 1$ anyways and unconstrained optimization is generally easier to do than constrained ones. The LRT statistic is basically a restricted-range MLE over an unrestricted MLE.

By the same ole' reasoning, LRT based on $X \Leftrightarrow$ LRT based on sufficient statistic $T(X)$.

If you switch to the Bayesian viewpoint, with an additional piece of raw material $\pi(\theta)$, the posterior probability $(X|\theta)$ can be computed. Previously you didn't have probability, you only had likelihood $(\theta|X)$. And you can reject H_0 if $P(\theta \in \Theta_0|\mathbf{x}) << 0.5$ or $P(\theta \in \Theta_0^c|\mathbf{x}) >> 0.5$

4.2 Evaluating Hypothesis Tests

Just like what you did in point estimation, you have now created a reservoir of candidate tests. Bias and variance were your 2 KPIs for a point estimators, what should be the KPI for tests? Since each test yields a yes/no answer, it naturally follows that this answer could be right/wrong. The degree of "wrongness" is formalized to be the Type I and Type II Error. Type I Error occurs when $\theta \in \Theta_0$ yet you reject H_0. Type II Error occurs when $\theta \in \Theta_0^c$ yet you accept H_0.

	$x \in R$	$x \notin R$
$\theta \in \Theta_0$	Type I Error ($P_{\theta\in\Theta_0}(\mathbf{X} \in R)$))	✓ $(1 - P_{\theta\in\Theta_0}(\mathbf{X} \in R))$
$\theta \in \Theta_0^c$	✓ ($P_{\theta\in\Theta_0^c}(\mathbf{X} \in R)$))	Type II Error $(1 - P_{\theta\in\Theta_0^c}(\mathbf{X} \in R))$

Define power function:

$$\beta(\theta) = P_\theta(\mathbf{X} \in R) = \begin{cases} P(\text{Type I Error}) & \text{if } \theta \in \Theta_0 \\ 1 - P(\text{Type II Error}) & \text{if } \theta \in \Theta_0^c \end{cases}$$

Ideally, the power function takes 0 for all $\theta \in \Theta_0$ and 1 for all $\theta \in \Theta_0^c$.

Type I or Type II Error, which is a more serious crime? Der Heilige im Walde has already given you a great answer: Type I Error means harm to the whole society (unsound "scientific" conclusion), but Type II Error means harm to only one person - our unfortunate researcher cannot publish her paper. Therefore we normally care more about controlling Type I Error, that is, we first ensure that Type I Error is below some threshold (must!), then we minimize Type II Error (just do your best), i.e., $\max_R P(x \in R|\theta \in \Theta_0^c)\ s.t.\ P(x \in R|\theta \in \Theta_0) \leq \alpha$. A test with $\beta(\theta)$ is a level α test if $\sup_{\theta\in\Theta_0} \beta(\theta) \leq \alpha$.

Well, aside from these 2 errors there may be some other good qualities you want your power function to possess. For one, you hope that you are more likely to reject H_0 if $\theta \in \Theta_0^c$ than if $\theta \in \Theta_0$. In talking about level α tests you were only putting restrictions on the worst type I error, but now you also hope (1−worst type II error) ≥ (worst type I error always). Formally, a test is unbiased if $\beta(\theta') \geq \beta(\theta'')$ for every $\theta' \in \Theta_0^c$ and $\theta'' \in \Theta_0$.

You never settle for "better." If a way to the better there be, it exacts a full look at the best (Thomas Hardy winces). How can a test be the best? Similar to the step-by-step logic (kinda like dynamic programming & principle of optimality but not quite) in finding a best unbiased estimator, you only consider level α tests, then among them, you want the test that guarantees a uniformly smallest Type II Error. Formally, let C be a class of (all level α) test. A test in class C, with power function $\beta(\theta)$, is a uniformly most powerful (UMP) class C test if $\beta(\theta) \geq \beta'(\theta)$, $\forall\theta \in \Theta_0^c$, $\forall\beta'(\theta) \in C$.

Then just like how you paddled through the swamp of Cramer-Rao Lower Bound, Rao-Blackwell, etc., to find UMVUE, you are now in search of a systematic way to see whether a UMP exists, and if it does, what is it. Or to say the least you want to be able to recognize a UMP if you see one on the street. The Neyman-Pearson Lemma delineates the UMP level α tests when both H_0, H_1 are simple hypotheses. Formally,

$H_0 : \theta = \theta_0, H_1 : \theta = \theta_1 . f(\mathbf{x}|\theta_i), i = 0, 1$. Using a test with R that satisfies:
1) $\mathbf{x} \in R\ if\ f(\mathbf{x}|\theta_1) > kf(\mathbf{x}|\theta_0)$ & $\mathbf{x} \in R^c\ if\ f(\mathbf{x}|\theta_1) < kf(\mathbf{x}|\theta_0)$, for some $k \geq 0$.
$\Leftrightarrow \{\mathbf{x} : \frac{f(\mathbf{x}|\theta_1)}{f(\mathbf{x}|\theta_0)} > k\} \subset \{\mathbf{x} \in R\} \subset \{\mathbf{x} : \frac{f(\mathbf{x}|\theta_1)}{f(\mathbf{x}|\theta_0)} \geq k\}$ Relative likelihood by division, a more intuitive form.
and
2) $\alpha = P_{\theta_0}(\mathbf{X} \in R)$ Then
a. (Sufficiency) Any test that satisfies 1) and 2) is a UMP level α test.
b. (Necessity) If ∃ a test satisfying 1) and 2) with $k > 0$, then every UMP level α test is a size α test (2)), and every UMP level α test satisifies (1)) except perhaps on a set A satisfying $P_{\theta_0}(\mathbf{X} \in A) = P_{\theta_1}(\mathbf{X} \in A) = 0$.

As usual, you can use sufficient statistic $T(X)$ to replace X and get an equivalent result. Up to this point you might have already gotten the feeling that sufficient statistics are loyal proxies for the data.

However, tackling simple hypotheses doesn't satisfy you. The good news is that the Neyman-Pearson Lemma can also be used to find UMP tests with composite hypotheses, with some help from a concept called monotone likelihood ratio (MLR). $\{g(t|\theta) : \theta \in \Theta\}$ has a monotone likelihood ratio (MLR) if, $\forall \theta_2 > \theta_1$, $\frac{g(t|\theta_2)}{g(t|\theta_1)})$ is a (weakly) monotone function of t on $\{t : g(t|\theta_2) > 0 \ or \ g(t|\theta_1) > 0\}$. Say $H_0 : \theta \leq \theta_0, H_1 : \theta > \theta_0$, if sufficient statistic T has MLR, and $P(t > t_0|\theta = \theta_0) = \alpha$, then $R : \{t : t > t_0\}$ is UMP level α test (Karlin-Rubin Theorem). Intuitively, if MLR= $\frac{g(t|\theta_2)}{g(t|\theta_1)}$ increases in t, then larger t renders the larger θ more *and more* plausibility (I emphasize the "and more" here because "more" alone is already guaranteed by large t, largeR t says more). Imagine a rabbit-turtle race with a normal rabbit, and how the comparative advantage increases as time elapses.

Karlin-Rubin Theorem basically specifies a condition (MLR) under which a one-sided hypothesis test can be reduced to a simple-hypothesis test, thereby extending the applicability of the Neyman-Pearson Lemma.

Here you might wonder, like me, why $g(t|\theta)$ is called likelihood but not probability. Think about the variable we are interested in here.

A long time ago I wrote a short passage trying to compare the critical value approach and the p-value approach in hypothesis testing based on what pathetically little I had learned - just like what I'm doing now with this bedtime story series. Back then I felt quite good about p-values, naively betrayed by a certain prettiness. Now is a good time to reexamine this issue.

How has α been interpreted thus far? It is the probability of mistakenly rejecting H_0 when it is true; it is the worst Type I Error we allow; it is the "significance level," the smaller the α, the more significantly correct the H_1. What about p-value? It is the the probability of getting a test statistic at least as extreme as the one represented by the sample data when H_0 is true; it is a measure of extremeness of the observed sample given H_0 is true. An obvious advantage of p-values, as was mentioned in my naive passage, is that it reports the test result on a continuous scale and is no longer confined to the dichotomous "reject" or "accept." However, p-value ignores Type II Error completely, and the definition of extremeness can easily be arbitrary and subject to manipulation. According to der Heiliger, p-value is one of the most abused statistical techniques. We say a p-value is valid if $P_\theta(p(\mathbf{X}) \leq \alpha) \leq \alpha$, $\forall \theta \in \Theta_0, \forall 0 \leq \alpha \leq 1$. A valid p-value can guarantee that $R : \{\mathbf{x} : p(\mathbf{x}) \leq \alpha\}$ is a level α test.

5 Interval Estimation: Methods of Finding and Evaluating Interval Estimators

Remember those parks you used to take post-dinner strolls in with grandma holding your hand when you were little? (Alternatively, think about an amusement park.) Point estimation is like balloon-shooting, and interval estimation is like ring-tossing. Instead of guessing a value, you guess an interval, thus trading precision for confidence about your assertion.

An interval estimator (your ring) looks like this: $[L(\mathbf{X}), U(\mathbf{X})]$. Coverage probability $P_\theta(\theta \in [L(\mathbf{X}), U(\mathbf{X})])$ is the probability that your interval covers a given θ (gotcha!), and confidence coefficient $\inf_{\theta\in\Theta} P_\theta(\theta \in [L(\mathbf{X}), U(\mathbf{X})])$ is the infimum of coverage probabilities - worst case analysis again.

An essential feature of interval estimation is that it is a dual problem of hypothesis testing. The intuition given by Cinderella and her glass slippers still applies, it is only that before, the spotlight was on x, and now it is on θ_0.

The equivalence can be formalized as follows: Let $A(\theta_0)$ be the acceptance region of a level α test of $H_0 : \theta = \theta_0$. Then $C(\mathbf{x}) = \{\theta : \mathbf{x} \in A(\theta_0)\} \Rightarrow C(\mathbf{X})$ is a $1 - \alpha$ confidence set. Conversely, let $C(\mathbf{X})$ be a $1 - \alpha$ confidence set. Then $A(\theta_0) = \{\mathbf{x} : \theta_0 \in C(\mathbf{x})\} \Rightarrow A(\theta_0)$ is the acceptance region of a level α test of $H_0 : \theta = \theta_0$.

After finding a reservoir of interval estimators that make sense, you face the same ole problem of selecting the best among them, or, more fundamentally, forming nice KPIs for this selection.

How is the reward mechanism designed for a ring-tossing game? First you have to capture *something* to get something. Then the smaller your ring, the higher prize you get. These correspond to coverage probability and size, respectively. Size is usually measured by length (of interval, or volume of a multidimentional sets) and coverage probability by confidence coefficient, although there are certainly other ways of measure. By now you should be competent in formalizing your problem:

$$\min \text{size of confidence interval} = E_x(U(x) - L(x))$$

$$s.t. \quad \inf_{\theta\in\Theta} P_\theta(L(x) \leq \theta \leq U(x)) \leq 1 - \alpha$$

We always seem to put bias (accuracy) before variance (precision) don't we, putting the former in the constraint and the latter in the objective function. If we can construct a pivot $Q(x, \theta)$ and WLOG assume it is decreasing, then our problem can be simplified to

$$\min_{a,b} E(Q^{-1}(a,x) - Q^{-1}(b,x))$$

$$s.t. P(a \leq Q(x,\theta) \leq b) = 1 - \alpha$$

My job here is done. The road (and its shortcuts) shall be left to the real traveler in the dark wood.