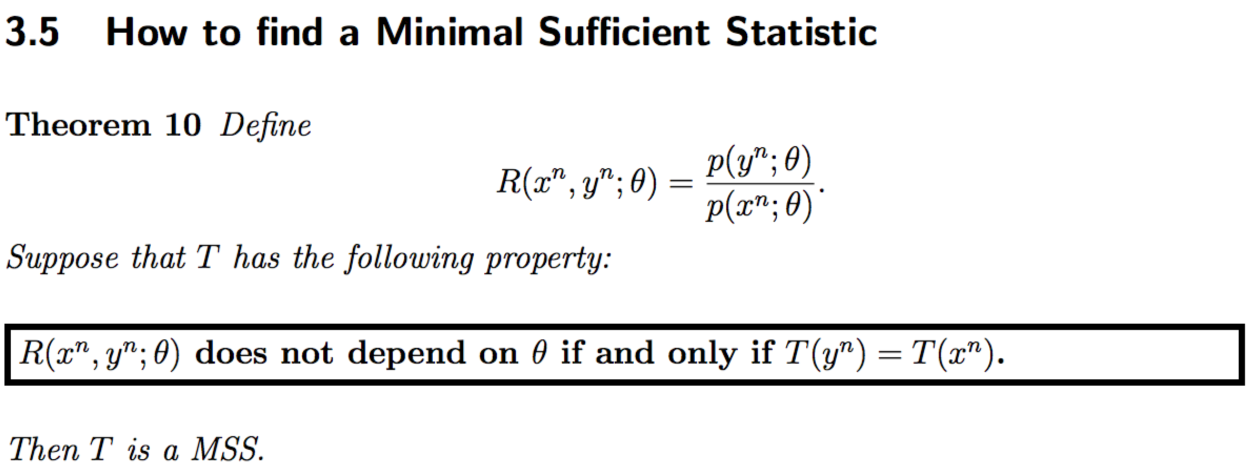
sufficiency, likelihood, point estimates

- the difference between probability and statistics: up till now, we have mainly focused on the question: here's a distribution, what can i say about that data that comes out of this. that is the central question probability focuses on. statistics focused on: "here is some data. what can i say about the underlying distribution that generated the data/where it comes from/what information can we derive from this data".
- statistical model is a set of probability distributions (or a collection of probability densities). given some data, we might have a model (a set of distributions, one of which generated that data). we don't know which one it is but we'd like to estimate (find a probability distribution of) which distribution may have generated that data.
- in a simplified view of things, we can say there are parametric and non-parametric models. parametric models are models in a finite dimensional parameter space. given a parameter vector, you specify the distribution. eg. the set of normal distributions is a parametric model specified by μ and σ. non parametric models are those that cannot be specified like this.
- really, non-parametric models are infinitely parametric models. (eg. the set of all distributions is a nonparametric model).
- parametric models make assumptions, non-parametric models don't really. it is an art to put the right set of assumptions. there is a lot of gray area between parametric and nonparametric models. to make things worse, there is this set of things called semi parametric models.
- a typical example of a non-parametric model is, given some data $X_1 ... X_n$ that are iid with some cdf F, estimate F. I've told you nothing about F. So we will have to assume a nonparametric model.
- first we are going to do the easy thing: parametric models.
- remember, a statistic is a function of the given data. by this definition, $\overline{X} - \mu$ is not a statistic as it depends not only on data but also knowing the underlying mean, which we usually don't in statistics (we did in probability)
- To indicate that we are working with a parametric model, we will use $p(x; \theta)$ to indicate that we are dealing with a set of distributions parametrized by θ.
- sufficiency deals with the question of is there some statistic that summarizes the data? this was a lot more important in the olden days until computing happened, but is definitely still useful even today.
- a trivial statistic that is sufficient is: $g(X_1, X_2 ... X_n) = (X_1, X_2 ... X_n)$ → the statistic that outputs the data itself. which is great. but we are really interested in minimal sufficiency. preserving information with the least amount of information. this is called data reduction.
- suppose $X_1 ... X_n$ ~ $p(x; \theta)$. Statistic T is sufficient for θ if the conditional distribution of $X_1 ... X_n | T$ does not depend on θ. in other words, $p(X_1, ... X_n | t, \theta) = p(X_1, ... X_n | t)$ where $t = T(X_1 ... X_n)$.
- notice that a statistic splits up the set of all possible assignments of $X_1, X_2 ... X_n$. A value of the statistic $g(X_1 ... X_n)$ can correspond to multiple assignments of $X_1, ... X_n$. Each such assignment is a partition.
- Statistics are equivalent if they cause the same partition. the value of the statistics itself is irrelevant. we will see more about this in the example below. this is because the $p(X^n | t)$ is the same acroos these (except the value of t itself is different since the statistic is a different function, but think of t as more an identifier for the partition). Note that X^n is just shorthand for $X_1, X_2 ... X_n$.

- so remember, any statistic that drops the dependence on θ's given the statistic, is sufficient. that is, when $p(X_1, ...X_n|t, \theta) = p(X_1, ...X_n|t)$.
- factorization theorem and proof: if a statistic is sufficient, we will be able to write $p(X^n|\theta)$ as $h(X^n)g(t, \theta)$. also, (more usefully) if we are able to write $p(X^n|\theta)$ as $h(X^n)g(t, \theta)$ then the statistic is sufficient. notice that the proof above proves both directions in the order mentioned here.
- T is a minimally sufficient statistic if:
  - T is sufficient
  - If U is any other sufficient statistic, then $T = g(U)$ for some function g.
- Things Larry states but doesn't prove (kind of gives intuition but definitely doesn't prove):
  - A minimal sufficient T generates the coarsest sufficient partition.
  - The minimal sufficient statistic is not unique. But, the minimal sufficient partition is unique.
  - test for minimally sufficient statistic:



3.5 How to find a Minimal Sufficient Statistic

Theorem 10 *Define*

$$R(x^n, y^n; \theta) = \frac{p(y^n; \theta)}{p(x^n; \theta)}.$$

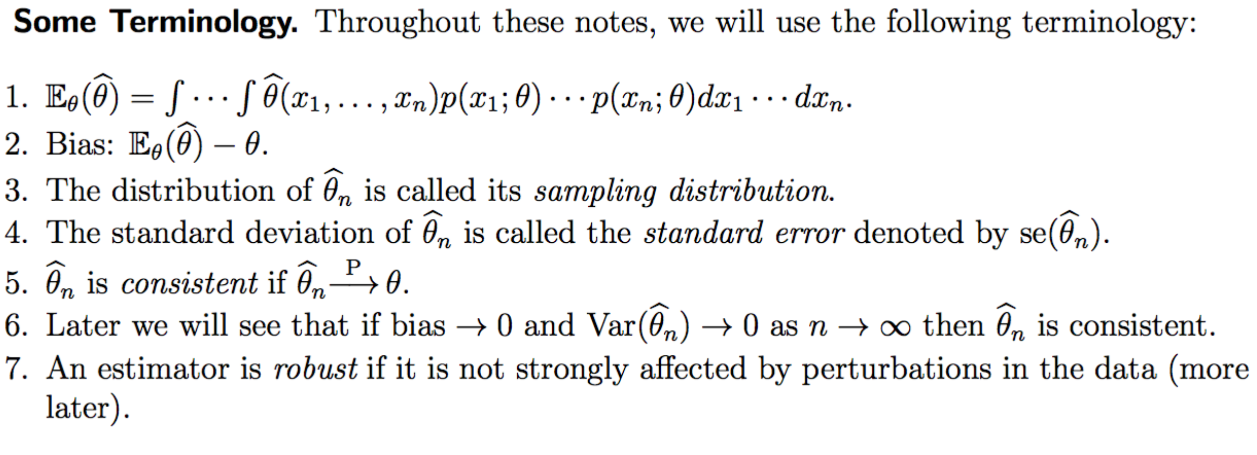
Suppose that T has the following property:

$$\boxed{R(x^n, y^n; \theta) \textbf{ does not depend on } \theta \textbf{ if and only if } T(y^n) = T(x^n).}$$

Then T is a MSS.

- basically, larry hasn't given us a convincing (with proof) technique to test for minimal sufficient statistic
- assuming the above theorem, we can get a hint on how to find a minimal sufficient statistic. we have to find a function that so R does not depend on θ iff $T(y^n) = T(x^n)$.
- point estimates: given some data, how do we estimate θ? the goal is to get to this question. but before that, likelihood functions.
- why study likelihood functions?
  - to do point estimates
  - to do bayesian inference (which we'll talk about)
  - its also related to sufficiency
- a likelihood function $L(\theta) = p(x_1, ...x_n|\theta)$ → (if iid) → $\prod p(x|\theta)$. Note that the likelihood function is *not* a probability distribution, does not integrate to 1, etc. (yes, $p(\theta|x)$ is a pmf/pdf).
- log likelihood function is the log of the likelihood function. because the likelihood is a product for iid's, the log likelihood is a bunch of sums.
- if we have two likelihood functions of μ that are constant multiples of each other, then they are the same likelihood function. why? (my stab at it: if you think of a likelihood function as a function where you plug in the $\theta = k$ and it outputs the probability of data given $\theta = k$, then in that sense it is a pdf/pmf. multiplying by a constant means nothing cuz we need to make it sum upto 1 anyway)
- If we take the set of all possible datasets (all possible X^n's) and partition them according to the values of a sufficient statistic T, then all datasets within the same parition have the same likelihood function $p(x|\theta)$. (this kind of follows from the definition of sufficiency for a statistic)

- Given a sufficient statistic, we can compute the likelihood function. why? because the statistic is sufficient, $p(x|t,\theta) = p(x|t)$. the likelihood function is $p(x|\theta)$, but since we have x, we can compute t and $p(x|t)$ gives us $p(x|t,\theta)$ but note that x (and hence t are given) so this is really the likelihood function.
- so really, the minimal sufficient statistic has all the information we need to compute the likelihood function. the notion of sufficiency is that its "enough to compute the likelihood function".
- now lets try to answer the question: can we guess θ, given a dataset? this is called point estimation. its so called because the θ we estimate is really a point.
- later we will do other related things like confidence intervals, which is to get an interval instead of a single point, hypothesis testing, which is how to test some hypothesis about given data but we will start with the best guess of θ, which we will call $\hat{\theta}$.
- the hat sign means its an estimate. remember, that the thus estimated $\hat{\theta}$ is a function of the data. the difference between the parameter and the estimated parameter is that θ is a fixed, unknown number. $\hat{\theta}$ is really a random variable (backed by a distribution). here we are interested in the value of the random variable that has the highest pdf/pmf.
- if our model is correct, then one of the distributions in the model is the true distribution (the one that generated the data).
- there is a lot of ways of coming up with point estimators. we will cover the 3 most common ones: method of moments, maximum likelihood estimator and bayes estimators
- we will talk about the methods of these estimators and about evaluating these estimators → which one to use, which is best, what is best etc
- we will evaluate point estimators based on:
  - consistency: given more and more data, do they converge to the true value?
  - bias, variance
  - mean squared error (MSE)
  - minimax (a way to formally compare estimators)
  - robustness → such and such an estimator is good under these circumstances. what if the assumptions/circumstances get violated by the data?
- We will complete this section by introducing some terminology that will be used in the next section on point estimation. just note this terminology. these will be discussed and proved in the coming sections.



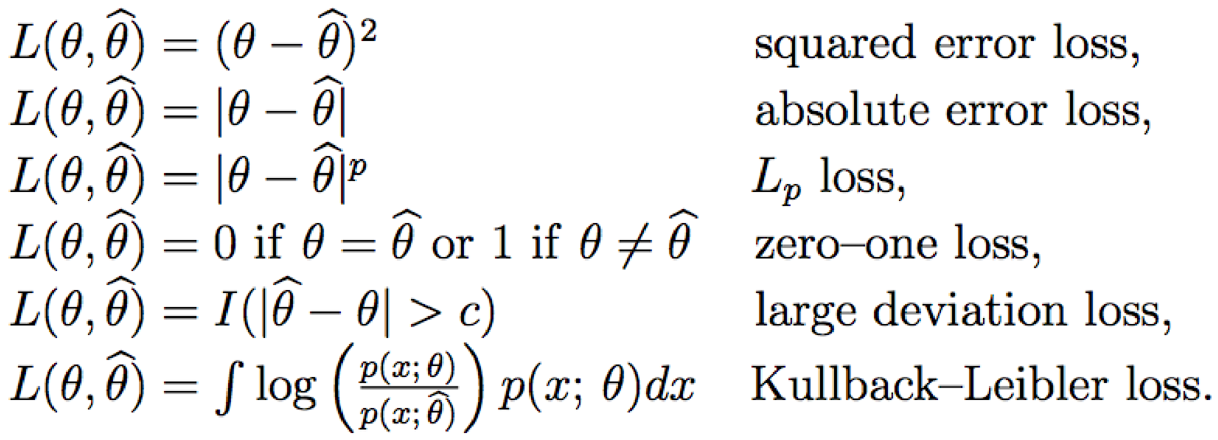
Some Terminology. Throughout these notes, we will use the following terminology:

1. $\mathbb{E}_\theta(\widehat{\theta}) = \int \cdots \int \widehat{\theta}(x_1, \ldots, x_n) p(x_1;\theta) \cdots p(x_n;\theta) dx_1 \cdots dx_n$.
2. Bias: $\mathbb{E}_\theta(\widehat{\theta}) - \theta$.
3. The distribution of $\widehat{\theta}_n$ is called its *sampling distribution.*
4. The standard deviation of $\widehat{\theta}_n$ is called the *standard error* denoted by $\mathrm{se}(\widehat{\theta}_n)$.
5. $\widehat{\theta}_n$ is *consistent* if $\widehat{\theta}_n \xrightarrow{\mathrm{P}} \theta$.
6. Later we will see that if bias $\to 0$ and $\mathrm{Var}(\widehat{\theta}_n) \to 0$ as $n \to \infty$ then $\widehat{\theta}_n$ is consistent.
7. An estimator is *robust* if it is not strongly affected by perturbations in the data (more later).

- if you think back to our law of large numbers proof, you can see that the proof (and the result) applies not only to the first moment but also for higher moments. the method of moments uses that result to do a point estimate. Lets say we want to come with up $\hat{\theta}$ (the point estimate). $\hat{\theta}$ is, in general, a vector (because a model, which is just a distribution space, can be parametrized by multiple parameters (mean and variance in case of normal, form example)). to find this vector, we compute the various moments from the given dataset and the theoretical moments (in terms of the parameters). this way if we have k parameters, we get k equations in k unknowns. If these equations are nice and solvable, the solution gives us our estimate. Note that this relies on the weak law of large numbers which says we converge in probability as we get more and more data.

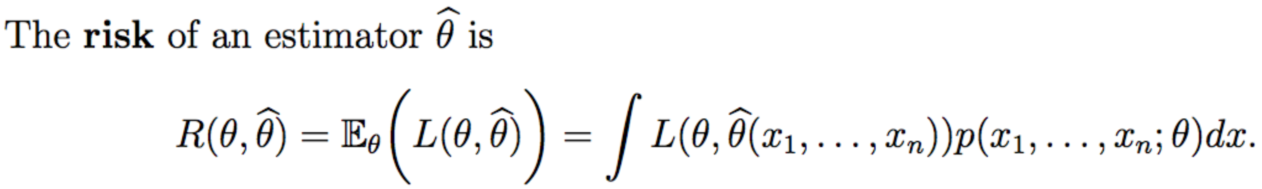
- Maximum likelihood estimate (MLE): Basically find the $\hat{\theta}$ that maximizes $L(\theta)$, the likelihood function. this is the same as maximizing the log likelihood function since the log is monotonic and this is often easier than maximing the likelihood function directly, although the answer doesn't change at all. MLE $= argmax\ p(X^n|\theta)$. there is a whole field of maximizing functions whose theory can be applied here.
- If θ has multiple parameters, then we have to find an assignment of values to those θ_i's that maximizes L. We can either think of this as a global optimization across the whole space of θ or as a process of finding the max for each value that some θ_k can take (lets call this a profile max) and then finding the max across these "profile max"s. Both search the whole space of θ's (just in different order) and end up finding the same max answer.
- A related idea is a re-indexing of the distribution space. Instead of using θ to index the space, we use $g(\theta)$. If $\hat{\theta}$ gives us the max L in the original indexing, then $g(\hat{\theta})$ will in the new indexing.
- If g is many to one function, then we must choose g so that it maps to the L value that corresponds to the max of all the θ values that have the same g mapping. This is similar to the idea of profiling.
- From the above, we see that with MLE, we have an estimator not just for a parameter but for any function of that parameter! this is called the equivariance property.
- In the above discussion the likelihood function for $g(\theta)$ is some $L^*(g(\theta))$ which is, obviously, not equal to L.
- Bayes' estimator: For this, we will need a prior distribution $p(\theta)$ that is our belief before we've seen the dataset. From Bayes' theorem, $p(\theta|X^n) = \frac{p(X^n|\theta)p(\theta))}{p(X^n)} \propto p(X^n|\theta)p(\theta)$ (since the marginal $p(X^n)$ is a constant) $p(X^n|\theta)p(\theta) = L(\theta)p(\theta)$. Now we are basically treating θ as a random variable and have obtained its distribution $p(\theta|X^n)$. Next we just compute its expected value and that's our estimate. There is arguments to use median, mode etc. but for now we will just use the expected values as the bayes' estimate once we've obtained the distribution of θ. when mode is used it is called the maximum a posteriori (MAP) estimator.
- notice that the above approach involves us coming up with a prior for θ. which prior we choose very much influences the bayes' estimate.
- go back and look at the discussion on conjugate priors in the probability course. that applies here.
- given some data there are infinitely many bayes estimates because there are infinitely many priors we can choose.
- how do we evaluate estimators? we will start with a method called mean squared error. mean squared error is $E((\hat{\theta} - \theta)^2)$ where θ is the true, unknown parameter and $\hat{\theta}$ is the point estimate.
- one can easily show that $mse = bias^2 + variance$
- If we do MLE on a data with a normal model, we will see that MSE goes to 0 as $n \to \infty$. this seems like a good thing. mean squared error of the estimate is going to 0.
- In general, the MSE is a function of the true mean, variance or other parameters! So MSE is a function of the thing we are trying to estimate. So MSE gives us some visibility, but doesn't make it obvious which estimator is better, given the MSEs of two estimators.
- Unbiased estimators are those where $E(\hat{\theta}) = \theta$. (which leads to 0 bias)
- Minimax theory provides a framework for answering the question: what is the best possible estimator. you decide the cost/loss function. we'll provide a framework to measure it. some examples of loss

functions:



$$
\begin{array}{ll}
L(\theta,\widehat{\theta}) = (\theta-\widehat{\theta})^2 & \text{squared error loss,}\\
L(\theta,\widehat{\theta}) = |\theta-\widehat{\theta}| & \text{absolute error loss,}\\
L(\theta,\widehat{\theta}) = |\theta-\widehat{\theta}|^p & L_p \text{ loss,}\\
L(\theta,\widehat{\theta}) = 0 \text{ if } \theta=\widehat{\theta} \text{ or } 1 \text{ if } \theta\neq\widehat{\theta} & \text{zero–one loss,}\\
L(\theta,\widehat{\theta}) = I(|\widehat{\theta}-\theta|>c) & \text{large deviation loss,}\\
L(\theta,\widehat{\theta}) = \int \log\left(\frac{p(x;\theta)}{p(x;\widehat{\theta})}\right) p(x;\,\theta)dx & \text{Kullback–Leibler loss.}
\end{array}
$$

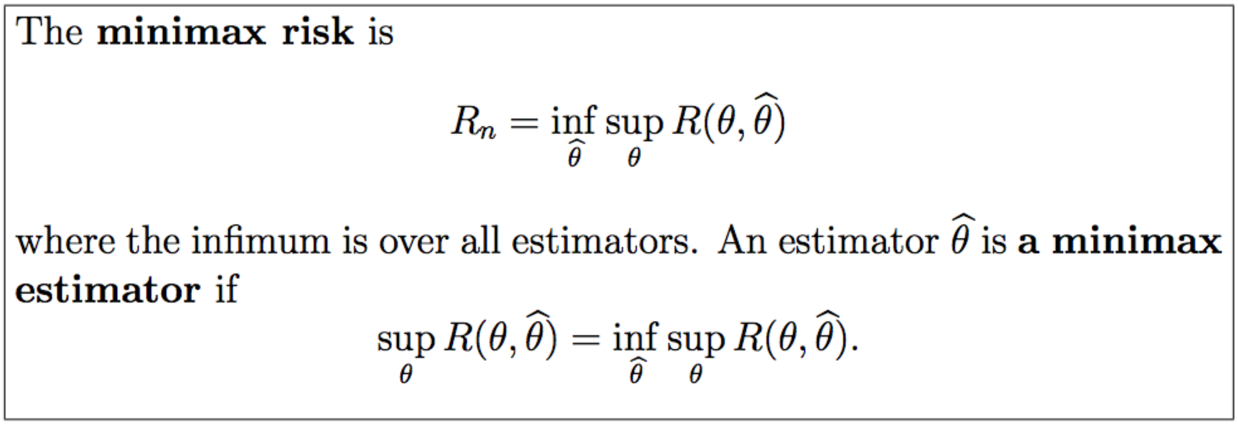
- The risk is defined as the expected loss. The estimation is with respect to the probability distribution of the data.



The **risk** of an estimator $\widehat{\theta}$ is

$$R(\theta,\widehat{\theta}) = \mathbb{E}_\theta\Big(L(\theta,\widehat{\theta})\Big) = \int L(\theta,\widehat{\theta}(x_1,\ldots,x_n))p(x_1,\ldots,x_n;\theta)dx.$$

- So MSE is just L_2 loss. Note this still has the same problem as MSE (that it is a function of the unknowns parameters we are trying to estimate)
- so what do we do? lets see what the worst case is. one idea is to look at the maximum of the risk function and minimize this maximum risk. this is a way to protect ourselves from the worst case. minimizing the max loss (hence called minimax).



The **minimax risk** is

$$R_n = \inf_{\widehat{\theta}} \sup_{\theta} R(\theta,\widehat{\theta})$$

where the infimum is over all estimators. An estimator $\widehat{\theta}$ is **a minimax estimator** if

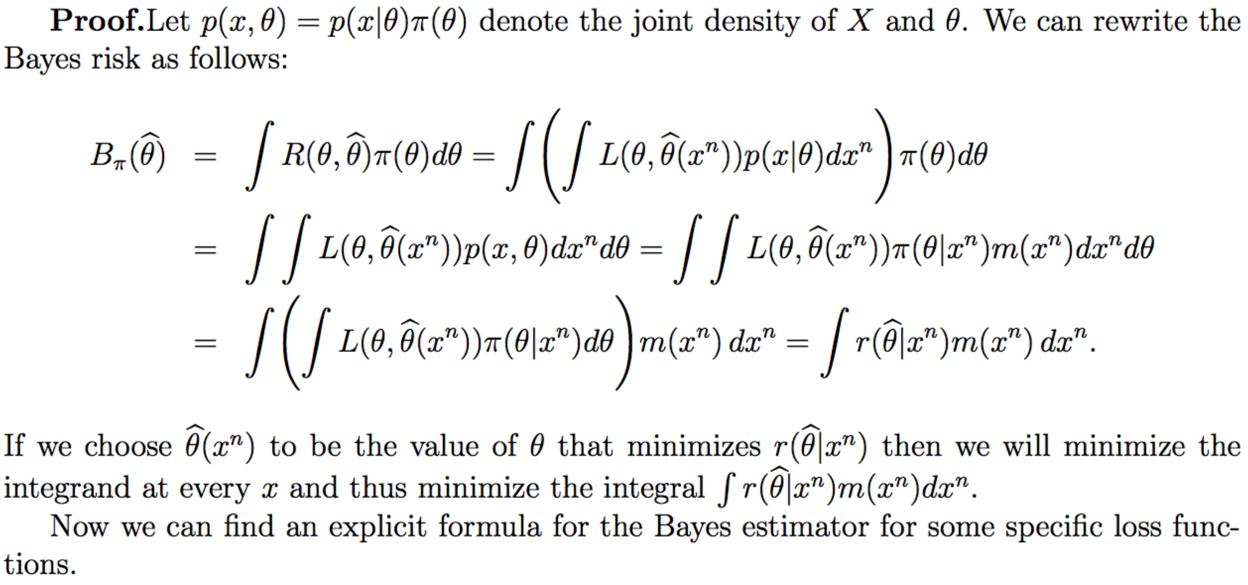
$$\sup_{\theta} R(\theta,\widehat{\theta}) = \inf_{\widehat{\theta}} \sup_{\theta} R(\theta,\widehat{\theta}).$$

minimax means choosing the estimate that has the least loss

- An alternative to minimax is to define a prior over the θ's and find the mean risk over this prior. Then we choose the estimator that has the least mean risk over the prior. This is called bayes' risk

$B_\pi(\hat{\theta}) = \int R(\theta, \hat{\theta})\pi(\theta)d\theta$. The subscript π in B_π denotes that the bayes risk is computed with the prior π.

- An estimator that minimizes the Bayes risk is called a Bayes estimator. Yes, we discussed a different bayes' estimator earlier. we will later see they are basically the same thing. for now, remember that we've seen two different "bayes' estimators"s.
- To find the risk, we integrated the loss function over the sampling distribution. instead we will try a different thing now. part of the process of computing the first kind of bayes estimator was computing a posterior $p(\theta|X^n) \propto L(\theta)\pi(\theta)$. what if we integrate the loss function over this posterior? it won't be function of θ anymore, it will just be a function of the data. (note that this is different from the risk → the risk is a function of the unknown, true θ, because we integrated over the data distribution).
- so once we have it as a function of the data, we can integrate over the data (for which we also have a marginal probability since we know have a prior for θ ($\pi(\theta)$)
- In both bayes risk and in the discussion right above ^, the claim is we get the same result. intuition: in bayes risk, we started with loss function, integrated over data given θ and then over the prior for θ. here we start with the loss, integrate over the posterior for the θ given data and then integrate over the data given the prior for θ.
- proof:
  - the proof is pretty straightforward. start with the first definition of bayes risk we saw, write out the full integral in gory detail, and switch around terms and you'll see that we get to what we derived in the second method.



Proof. Let $p(x,\theta) = p(x|\theta)\pi(\theta)$ denote the joint density of X and θ. We can rewrite the Bayes risk as follows:

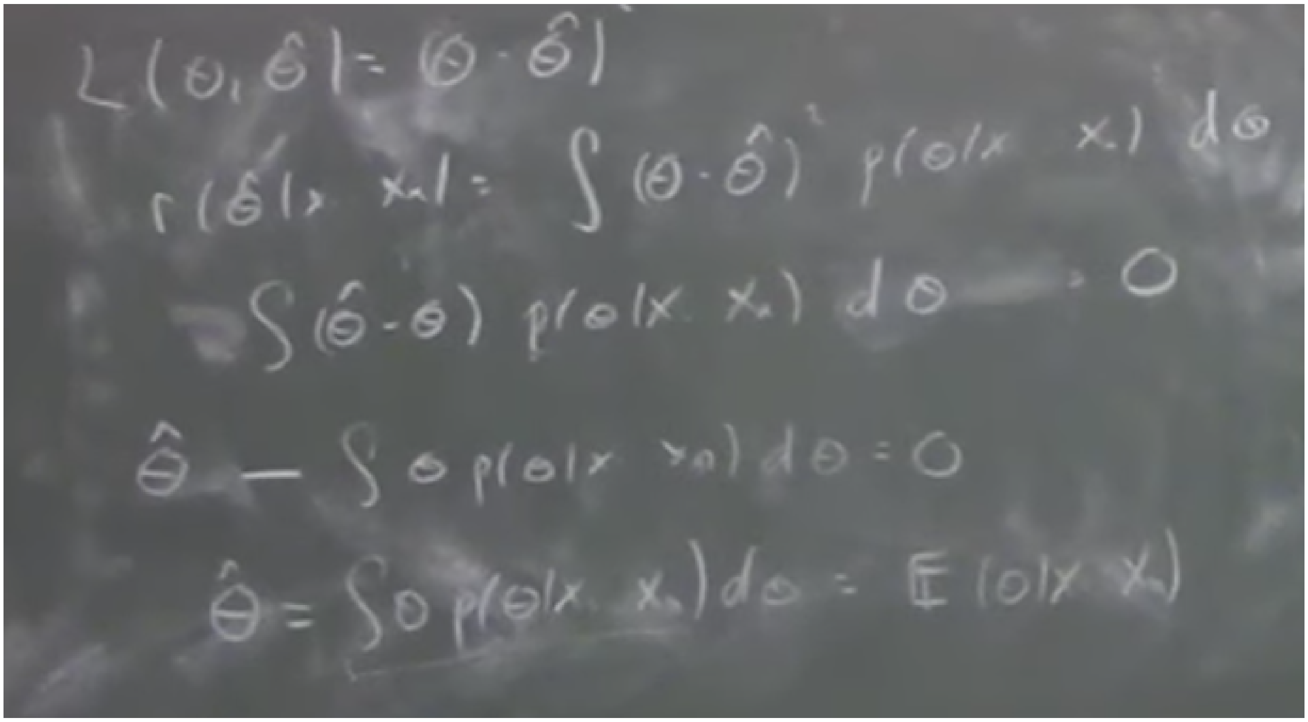
$$\begin{aligned} B_\pi(\widehat{\theta}) &= \int R(\theta,\widehat{\theta})\pi(\theta)d\theta = \int\left(\int L(\theta,\widehat{\theta}(x^n))p(x|\theta)dx^n\right)\pi(\theta)d\theta \\ &= \int\int L(\theta,\widehat{\theta}(x^n))p(x,\theta)dx^n d\theta = \int\int L(\theta,\widehat{\theta}(x^n))\pi(\theta|x^n)m(x^n)dx^n d\theta \\ &= \int\left(\int L(\theta,\widehat{\theta}(x^n))\pi(\theta|x^n)d\theta\right)m(x^n)\,dx^n = \int r(\widehat{\theta}|x^n)m(x^n)\,dx^n. \end{aligned}$$

If we choose $\widehat{\theta}(x^n)$ to be the value of θ that minimizes $r(\widehat{\theta}|x^n)$ then we will minimize the integrand at every x and thus minimize the integral $\int r(\widehat{\theta}|x^n)m(x^n)dx^n$.

Now we can find an explicit formula for the Bayes estimator for some specific loss functions.

- the reason we do this second method is because it often makes it a lot easier to compute the estimator that minimizes bayes risk this way instead of doing it from the definition.

how do we find an explicit formula for bayes estimator for some specific loss function? see below. note that this uses the idea of taking the derivative and setting it to 0 which works well for convex/concave functions:



$$L(\theta,\hat\theta) = (\theta-\hat\theta)$$

$$r(\hat\theta|x_1 \ldots x_n) = \int (\theta-\hat\theta)^2 p(\theta|x_1 \ldots x_n)\, d\theta$$

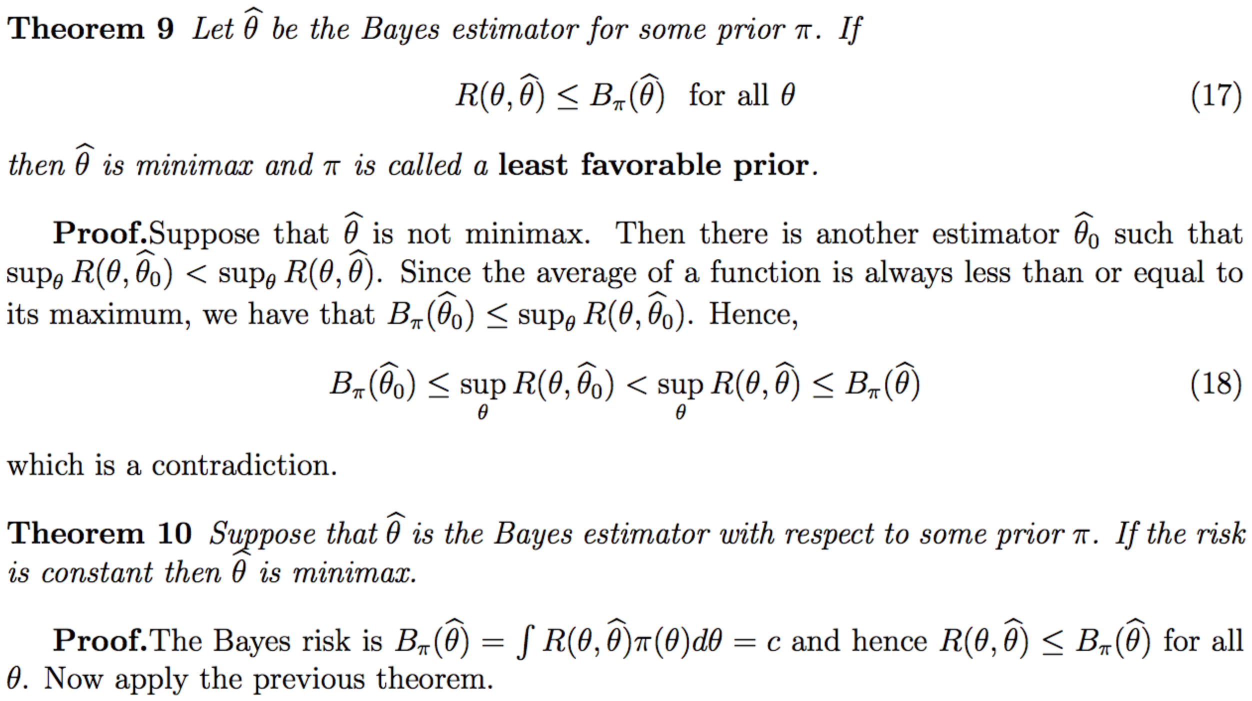
$$\int (\hat\theta-\theta)\, p(\theta|x_1 \ldots x_n)\, d\theta = 0$$

$$\hat\theta - \int \theta\, p(\theta|x_1 \ldots x_n)\, d\theta = 0$$

$$\hat\theta = \int \theta\, p(\theta|x_1 \ldots x_n)\, d\theta = \mathbb{E}(\theta|x_1 \ldots x_n)$$

We see the result turns out to be exactly the same as the bayes estimate we saw earlier! hence all the confusing names → they are all the same thing

- Key result: Bayes' estimators with a constant risk function are minimax. Proof is done well in the lecture notes:



Theorem 9 *Let $\widehat{\theta}$ be the Bayes estimator for some prior π. If*

$$R(\theta,\widehat{\theta}) \leq B_\pi(\widehat{\theta}) \text{ for all } \theta \tag{17}$$

then $\widehat{\theta}$ is minimax and π is called a **least favorable prior**.

Proof. Suppose that $\widehat{\theta}$ is not minimax. Then there is another estimator $\widehat{\theta}_0$ such that $\sup_\theta R(\theta,\widehat{\theta}_0) < \sup_\theta R(\theta,\widehat{\theta})$. Since the average of a function is always less than or equal to its maximum, we have that $B_\pi(\widehat{\theta}_0) \leq \sup_\theta R(\theta,\widehat{\theta}_0)$. Hence,

$$B_\pi(\widehat{\theta}_0) \leq \sup_\theta R(\theta,\widehat{\theta}_0) < \sup_\theta R(\theta,\widehat{\theta}) \leq B_\pi(\widehat{\theta}) \tag{18}$$

which is a contradiction.

Theorem 10 *Suppose that $\widehat{\theta}$ is the Bayes estimator with respect to some prior π. If the risk is constant then $\widehat{\theta}$ is minimax.*

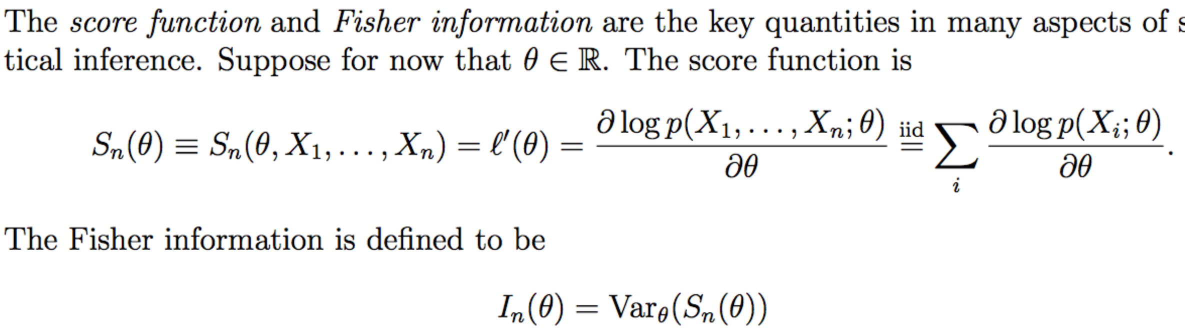
Proof. The Bayes risk is $B_\pi(\widehat{\theta}) = \int R(\theta,\widehat{\theta})\pi(\theta)d\theta = c$ and hence $R(\theta,\widehat{\theta}) \leq B_\pi(\widehat{\theta})$ for all θ. Now apply the previous theorem.

So if we make up the estimator $\hat\theta$ so that $\int L(\theta, \hat\theta(x_1...x_n))p(x_1...x_n|\theta)dx$ is a constant, that is minimax estimator. we just found a powerful way to compute a minimax estimator!

- let $X_1...X_n$ ~ $N(\theta, 1)$. With L_2 loss, we will show that $\hat\theta = \overline{X}$ is minimax.
  proof: note that the risk of the minimax estimator is the risk such that $R_m = inf_{\hat\theta}\ sup_\theta\ R(\theta, \hat\theta)$.

For any other estimator $\hat{\theta}_{other}$, we can write $R_m \leq sup_\theta \, R(\theta, \hat{\theta}_{other})$.
Now, $R_m = \; inf_{\hat{\theta}} \; sup_\theta \; R(\theta, \hat{\theta}) \geq inf_{\hat{\theta}} \int R(\theta, \hat{\theta})\pi(\theta)d\theta = inf_{\hat{\theta}}B_\pi(\hat{\theta})$ → the $\hat{\theta}$ that minizmies the bayes risk is the bayes estimator.
so now we can say pick an arbitrary bayes estimator, compute its bayes risk and we have a lower bound for R_m. Then pick any other estimator, compute its max risk and we have an upper bound for R_m.
For the normal, if we choose $\hat{\theta} \; = \overline{X}$, the $sup_\theta \; R(\theta, \hat{\theta}_{other}) =$ MSE $= \frac{1}{n}$. This is the upper bound. For the lower bound, we don't show the full details here. but basically we take the bayes estimate using the likelihood and prior, then compute the L_2 risk, then compute the bayes risk. doing the math, we see that as n → ∞, the lower bound → $\frac{1}{n}$. That means as n → ∞, R_m → $\frac{1}{n}$. Hence $\overline{X}$ is minimax for large n.

- Remember that Kullback Leibler distance was written as $\int p \, log(p/q)$
- Similarly, Hellinger distance is defined as $\sqrt{\int(\sqrt{p} - \sqrt{q})^2}$
- Now we will turn our attention to the question, given an estimator, what can we say about its behavior as we accumulate more and more data?
- Under some conditions (called regularity conditions), we can show that the MLE is consistent. the conditions:
    - the dimensionality of the parameter space is fixed (just means a fixed number of θ_i's in the model)
    - $p(x|\theta)$ is a smooth function of θ for each x.
    - the parameter is uniquely identifiable. that is given a θ, it only refers to 1 distribution
- Score function and Fisher information:

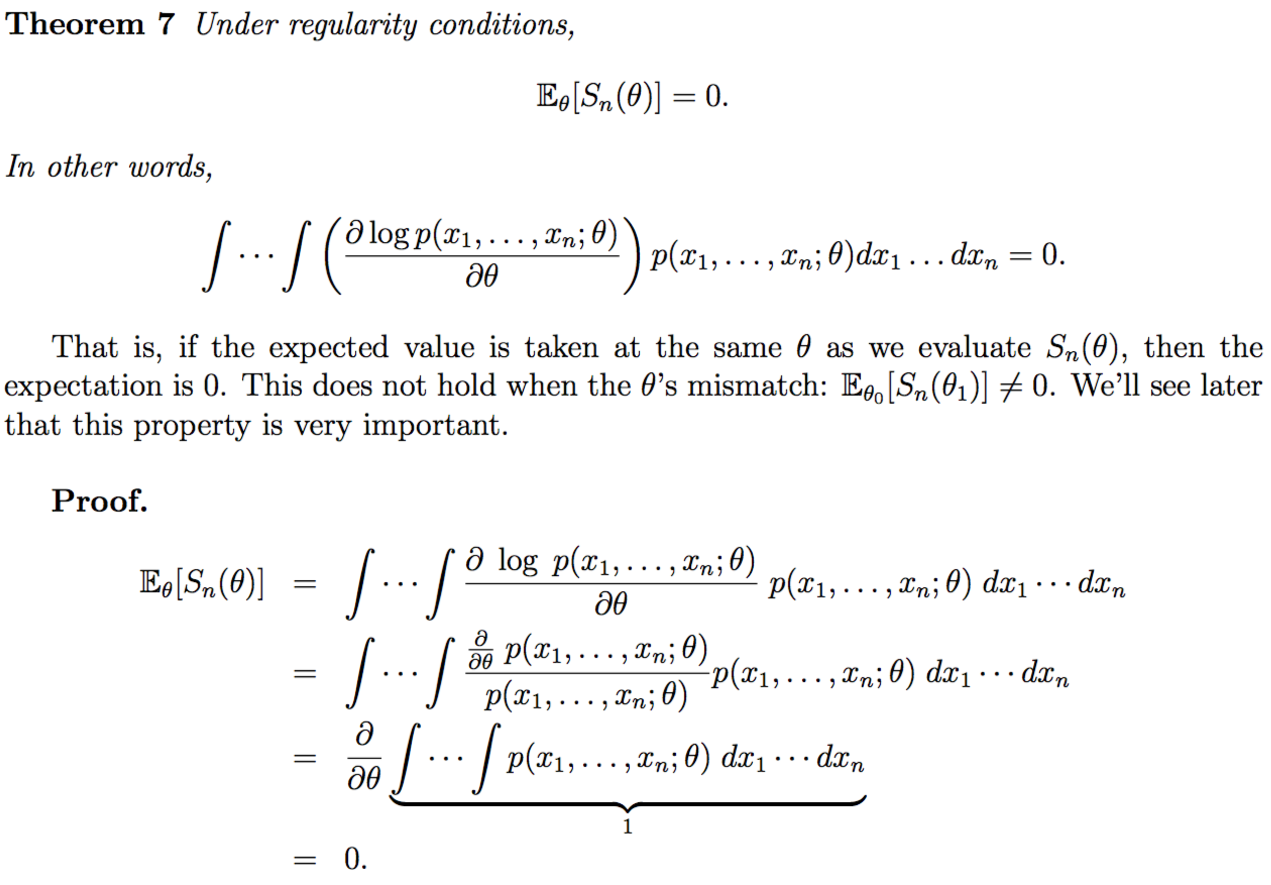


The *score function* and *Fisher information* are the key quantities in many aspects of statistical inference. Suppose for now that $\theta \in \mathbb{R}$. The score function is

$$S_n(\theta) \equiv S_n(\theta, X_1, \ldots, X_n) = \ell'(\theta) = \frac{\partial \log p(X_1, \ldots, X_n; \theta)}{\partial \theta} \overset{\text{iid}}{=} \sum_i \frac{\partial \log p(X_i; \theta)}{\partial \theta}.$$

The Fisher information is defined to be

$$I_n(\theta) = \text{Var}_\theta(S_n(\theta))$$



Theorem 7 *Under regularity conditions,*

$$\mathbb{E}_\theta[S_n(\theta)] = 0.$$

In other words,

$$\int \cdots \int \left(\frac{\partial \log p(x_1, \ldots, x_n; \theta)}{\partial \theta} \right) p(x_1, \ldots, x_n; \theta) dx_1 \ldots dx_n = 0.$$

That is, if the expected value is taken at the same θ as we evaluate $S_n(\theta)$, then the expectation is 0. This does not hold when the θ's mismatch: $\mathbb{E}_{\theta_0}[S_n(\theta_1)] \neq 0$. We'll see later that this property is very important.

Proof.

$$\begin{aligned}
\mathbb{E}_\theta[S_n(\theta)] &= \int \cdots \int \frac{\partial \ \log \ p(x_1, \ldots, x_n; \theta)}{\partial \theta} \ p(x_1, \ldots, x_n; \theta) \ dx_1 \cdots dx_n \\
&= \int \cdots \int \frac{\frac{\partial}{\partial \theta} \ p(x_1, \ldots, x_n; \theta)}{p(x_1, \ldots, x_n; \theta)} p(x_1, \ldots, x_n; \theta) \ dx_1 \cdots dx_n \\
&= \frac{\partial}{\partial \theta} \underbrace{\int \cdots \int p(x_1, \ldots, x_n; \theta) \ dx_1 \cdots dx_n}_{1} \\
&= 0.
\end{aligned}$$

The last step in the above proof involves two assumptions (hence the dependence on the regularity conditions:

  - smoothness of $p(x|\theta)$
  - the set over which the likelihood function is smooth does not depend on θ. (this is not true for the uniform distribution).

the above conditions let us switch the integral and the differential.

- From wikipedia: Intuitively, the **Fisher information** (sometimes simply called **information**[1]) is a way of measuring the amount of information that an observable random variable *X* carries about an unknown parameter *θ* of a distribution that models *X*.
- Note that in the mean and variance (fisher information) calculation, the expectation is over the data and θ is fixed.
- Since we have mean 0, the variance is $E(S_n^2)$
- For iid, note that the fisher information $I_n(\theta)$ is $nI(\theta)$ just because log likelihoods add up and fisher information is basically a variance.
- We can show easily that the fisher information $I_n(\theta) = -E_\theta[\frac{\partial}{\partial \theta^2} l_n(\theta)]$. the proof of this is pretty straightforward, just algebra, we won't prove it here, it is in the lecture.
- Note that given some dataset, score function is random (depends on θ). The fisher information is the variance, so it is not random, just a function of the parameter.
- Under some regularity conditions, the MLE is asymptotically normal. That means as n gets large, it converges in distribution to normal. That is, $\sqrt{n}(\hat{\theta} - \theta)$ ~ $N(0, \frac{1}{I(\theta)})$. proof is on page 8. we use the definitions and proofs about fisher information and the score function we developed earlier.
- The above means that $\hat{\theta} \approx N(\theta, \frac{1}{nI(\theta)})$. Note that the standard error (standard deviation) goes to 0 as n goes higher.

- In summary, remember that this is saying, given a real life physical situation, if the model is good (one of them is the true distribution) and if the regularity conditions are satisfied, then MLE produces this result.
- Note that lower variance of the estimator is great, this means a tighter estimate with higher probability. Asymptotic relative efficiency is a way to measure this. It's the ratio of the asymptotic variances of the estimators.
- There is a proof that Larry does not do that says the MLE has the best asymptotic relative efficiency. Assuming this is true, one reason not to choose MLE in a real world situation is if you're not sure the model is correct. eg. you might use a poisson model and compute MLE but your friend might have an estimator with lesser efficiency, but a broader model, then why use your friend's? Well, if your model is wrong then your friend's might work better.
- Larry Wasserman: "Everything I've told you up until now assumes the parametric model is correct. The parametric model is never correct. I can't think of a single example in real life where anybody would believe the parametric model is correct."
- Larry says that when the model is not perfect, the median can be a more robust estimator than the mean if we are assuming a normal model. One reason is if we in our sample some very very large numbers that skews the mean, but not the median.