Statistics 4109: Probability and Statistics
Liam Paninski, liam@stat.columbia.edu
Fall 2010

The course covers the material of two other courses, STAT4105 and STAT4107, in a single semester. The pace, therefore, is fast, and not all students will be able to keep up. Furthermore, the material is cumulative, that is, almost every lecture builds on previously discussed concepts, and students unable to keep up will find themselves in a very uncomfortable position. Students who doubt their preparation or who are concerned that they will not be able to consistently devote time to the course would be well advised to consider taking STAT4105 this semester followed by STAT4107 the next. However, if you're thinking of taking 4105 and 4107 in the same semester, I strongly recommend you take 4109 instead; 4109 offers the big advantage of covering the material in the proper sequence.

Also see

Probability and Statistics, 3rd Ed., by DeGroot and Schervish (ISBN 0-201-52488-0
A First Course in Probability (by S. Ross)
Statistical Inference (by Casella and Berger).

Table of Contents

Book I:

Introduction, sample spaces, probability axioms
Conditional probability, Bayes rule. Independent events.
continuous r.v.'s; multivariate distributions
Moment-generating functions; Covariance and correlation; sample means
Special continuous distributions

Book II

Inequalities, LLN, SpecialDiscrete, Central limit theorem

Estimation Theory: Order statistics; basic simulation theory: Monte Carlo integration, importance sampling
Decision theory: admissibility; minimax and Bayes decision rules; Bias/variance of estimators
Cramer-Rao bound;

Hypothesis Testing: Simple hypothesis testing; likelihood ratio tests; Neyman-Pearson lemma

Probability inequalities[11]

There is an adage in probability that says that behind every limit theorem lies a probability inequality (i.e., a bound on the probability of some undesired event happening). Since a large part of probability theory is about proving limit theorems, people have developed a bewildering number of inequalities.

Luckily, we'll only need a few key inequalities. Even better, three of them are really just versions of one another. **Exercise 29**: Can you think of example distributions for which each of the following inequalities are tight (that is, the inequalities may be replaced by equalities)? Are these "extremal" distributions unique?

Markov's inequality

Figure 5: Markov.

For a nonnegative r.v. X,

$$P(X > u) \leq \frac{E(X)}{u}.$$

So if $E(X)$ is small and we know $X \geq 0$, then X must be near zero with high probability. (Note that the inequality is *not* true if X can be negative.)

[11] HMC 1.10

The proof is really simple:

$$uP(X > u) \leq \int_u^\infty t p_X(t) dt \leq \int_0^\infty t p_X(t) dt = E(X). \quad \square$$

Chebyshev's inequality

Figure 6: Chebyshev.

$$P(|X - E(X)| > u) \leq \frac{V(X)}{u^2},$$

aka

$$P\left(\frac{|X - E(X)|}{\sigma(X)} > u\right) \leq \frac{1}{u^2}.$$

Proof: just look at the (nonnegative) r.v. $(X - E(X))^2$, and apply Markov.

So if the variance of X is really small, X is close to its mean with high probability.

Chernoff's inequality

$$P(X > u) = P(e^{sX} > e^{su}) \leq e^{-su} M(s).$$

So the mgf controls the size of the tail of the distribution — yet another surprising application of the mgf idea.

Figure 7: Chernoff.

The really nice thing about this bound is that it is easy to deal with sums of independent r.v.'s (recall our discussion above of mgf's for sums of independent r.v.'s). **Exercise 30**: Derive Chernoff's bound for sums of independent r.v.'s (i.e., derive an upper bound for the probability that $\sum_i X_i$ is greater than u, in terms of $M(X_i)$).

The other nice thing is that the bound is exponentially decreasing in u, which is much stronger than Chebyshev. (On the other hand, since not all r.v.'s have mgf's, Chernoff's bound can be applied less generally than Chebyshev.)

The other other nice thing is that the bound holds for all s simultaneously, so if we need as tight a bound as possible, we can use

$$P(X > u) \leq \inf_s e^{-su} M(s),$$

i.e., we can minimize over s.

Jensen's inequality

This one is more geometric. Think about a function $g(u)$ which is curved upward, that is, $g''(u) \geq 0$, for all u. Such a $g(u)$ is called "convex." (Downward-curving functions are called "concave." More generally, a convex

Figure 8: Jensen.

function is bounded above by its chords:

$$g(tx + (1-t)y) \leq tg(x) + (1-t)g(y),$$

while a concave function is bounded below.)

Then if you draw yourself a picture, it's easy to see that

$$E(g(X)) \geq g(E(X)).$$

That is, the average of $g(X)$ is always greater than or equal to g evaluated at the average of X. **Exercise 31**: Prove this. (Hint: try subtracting off $f(X)$, where f is a linear function of X such that $g(X) - f(X)$ reaches a minimum at $E(X)$.)

Exercise 32: What does this inequality tell you about the means of $1/X$? of $-X \log X$? About $E_i(X)$ vs. $E_j(X)$, where $i > j$?

Cauchy-Schwarz inequality

$$|C(X,Y)| \leq \sigma(X)\sigma(Y),$$

that is, the correlation coefficient is bounded between -1 (X and Y are anti-correlated) and 1 (correlated).

The proof of this one is based on our rules for adding variance:

$$C(X,Y) = \frac{1}{2}[V(X+Y) - V(X) - V(Y)]$$

(assuming $E(X) = E(Y) = 0$). **Exercise 33**: Complete the proof. (Hint: try looking at X and $-X$, using the fact that $C(-X,Y) = -C(X,Y)$.)

Exercise for the people who have taken linear algebra: interpret the Cauchy-Schwarz inequality in terms of the angle between the vectors X and Y (where we think of functions — that is, r.v.'s — as vectors, and define the dot product as $E(XY)$ and the length of a vector as $\sqrt{E(X^2)}$). Thus this inequality is really geometric in nature.

Limit theorems[12]

An approximate answer to the right question is worth a great deal more than a precise answer to the wrong question.

The first golden rule of applied mathematics, sometimes attributed to John Tukey

(Weak) law of large numbers

Chebyshev's simple inequality is enough to prove perhaps the fundamental result in probability theory: the law of averages. This says that if we take the sample average of a bunch of i.i.d. r.v.'s, the sample average will be close to the true average. More precisely, under the assumption that $V(X) < \infty$, then

$$P\left(|E(X) - \frac{1}{N}\sum_{i=1}^{N} X_i| > \epsilon\right) \to 0$$

as $N \to \infty$, no matter how small ϵ is.

The proof:

$$E\left(\frac{1}{N}\sum_{i=1}^{N} X_i\right) = E(X),$$

by the linearity of the expectation.

$$V\left(\frac{1}{N}\sum_{i=1}^{N} X_i\right) = \frac{V(X)}{N},$$

by the rules for adding variance and the fact that X_i are independent.

Now just look at Chebyshev. □

Remember, the LLN does not hold for all r.v.'s: remember what happened when you took averages of i.i.d. Cauchy r.v.'s? **Exercise 34**: What goes wrong in the Cauchy case?

Stochastic convergence concepts

In the above, we say that the sample mean $\frac{1}{N}\sum_{i=1}^{N} X_i$ "converges in probability" to the true mean. More generally, we say r.v.'s Z_N converge to Z in

[12] HMC chapter 4

probability, $Z_N \to_P Z$, if

$$P(|Z_N - Z| > \epsilon) \to 0$$

as $N \to \infty$. (The weak LLN is called "weak" because it asserts convergence in probability, which turns out to be a somewhat "weak" sense of stochastic convergence, in the mathematical sense that there are "stronger" forms of convergence — that is, it's possible to find sequences of r.v.'s which converge in probability but not in these stronger senses. In addition, it's possible to prove the LLN without assuming that the variance exists; existence of the mean turns out to be sufficient. But discussing these stronger concepts of convergence would take us too far afield[13]; convergence in probability will be plenty strong enough for our purposes.)

We discussed convergence of r.v.'s above; it's often also useful to think about convergence of distributions. We say a sequence of r.v.'s with cdf's $F_N(u)$ "converge in distribution" if

$$\lim_{N\to\infty} F_N(u) \to F(u)$$

for all u such that F is continuous at u (here F is itself a cdf). **Exercise 35**: Explain why do we need to restrict our attention to continuity points of F. (Hint: think of the following sequence of distributions: $F_N(u) = 1(u < 1/N)$, where the "indicator" function of a set A is one if $x \in A$ and zero otherwise.)

It's worth emphasizing that convergence in distribution — because it only looks at the cdf — is in fact weaker than convergence in probability. For example, if p_X is symmetric, then the sequence $X, -X, X, -X, ...$ trivially converges in distribution to X, but obviously doesn't converge in probability. **Exercise 36**: Prove that convergence in probability actually is stronger, that is, implies convergence in distribution.

Central limit theorem

The second fundamental result in probability theory, after the LLN, is the CLT: if X_i are i.i.d. with mean zero and variance 1, then

$$\frac{1}{\sqrt{N}} \sum_{i=1}^{N} X_i \to_D \mathcal{N}(0,1),$$

[13] Again, see e.g. Breiman '68 for more information.


Figure 9: De Moivre and Laplace.

where $\mathcal{N}(0,1)$ is the standard normal density. More generally, the usual rescalings tell us that

$$\frac{1}{\sigma(X)\sqrt{N}} \sum_{i=1}^{N} (X_i - E(X)) \rightarrow_D \mathcal{N}(0,1).$$

Thus we know not only that (from the LLN) the distribution of the sample mean approaches the degenerate distribution on $E(X)$, but moreover (from the CLT) we know exactly what this distribution looks like, asymptotically, if we take out our magnifying glass and zoom in on $E(X)$, to a scale of $N^{-1/2}$. In this sense the CLT is a stronger result than the WLLN: it gives more details about what the asymptotic distribution actually looks like.

One thing worth noting: keep in mind that the CLT really only tells us what's going on in the local neighborhood $(E(X) - N^{-1/2}c, E(X) + N^{-1/2}c)$ — think of this as the mean plus or minus a few standard deviations. But this does *not* imply that, say,

$$P(\frac{1}{N}\sum_{i=1}^{N} X_i \leq -\epsilon) \sim \int_{\infty}^{-\epsilon} \mathcal{N}(0, \frac{1}{N})(x)dx = \int_{\infty}^{-\sqrt{N}\epsilon} \mathcal{N}(0,1)(x)dx \quad \textbf{not true};$$

a different asymptotic approximation typically holds for the "large devia-

tions," the tails of the sample mean distribution[14].

More on stochastic convergence

So, as emphasized above, convergence in distribution can drastically simplify our lives, if we can find a simple approximate (limit) distribution to substitute for our original complicated distribution. The CLT is the canonical example of this; the Poisson theorem is another. What are some general methods to prove convergence in distribution?

Delta method

The first thing to note is that if X_N converge in distribution or probability to a constant c, then $g(X_N) \to_D g(c)$ for any continuous function $g(.)$. **Exercise 37**: Prove this, using the definition of continuity of a function: a function $g(u)$ is continuous at u if for any possible fixed $\epsilon > 0$, there is some (possibly very small) δ such that $|g(u+v) - g(u)| < \epsilon$, for all v such that $-\delta < v < \delta$. (If you're having trouble, just try proving this for convergence in probability.)

So the LLN for sample means immediately implies an LLN for a bunch of functions of the sample mean, e.g., if X_i are i.i.d. with $V(X) < \infty$, then

$$\left(\prod_{i=1}^{N} e^{X_i} \right)^{1/N} = e^{\frac{1}{N}\sum_{i=1}^{N} X_i} \to_P e^{E(X)},$$

(which of course should not be confused with $E(e^X)$; in fact, **Exercise 38**: Which is greater, $E(e^X)$ or $e^{E(X)}$? Give an example where one of $E(e^X)$ or $e^{E(X)}$ is infinite, but the other is finite).

We can also "zoom in" to look at the asymptotic distribution (not just the limit point) of $g(Z)$, whenever g is sufficiently smooth. For example, let's say $g(.)$ has a Taylor expansion at u,

$$g(z) = g(u) + g'(u)(z-u) + o(|z-u|), \ |z-u| \to 0,$$

where $|g'(u)| > 0$ and $z = o(y)$ means $z/y \to 0$. Then if

$$a_N(z_N - u) \to_D q,$$

[14]See e.g., Large deviations techniques and applications, Dembo and Zeitouni '93, for more information.

for some limit distribution q and a sequence of constants $a_N \to \infty$ (think $a_N = N^{1/2}$, if Z_N is the sample mean), then

$$a_N \frac{g(Z_N) - g(u)}{g'(u)} \to_D q,$$

since

$$a_N \frac{g(Z_N) - g(u)}{g'(u)} = a_N(Z_N - u) + o\left(a_N \frac{|Z_N - u|}{g'(u)}\right);$$

the first term converges in distribution to q (by our assumption) and the second one converges to zero in probability (**Exercise 39**: Prove this; i.e., prove that the remainder term

$$a_N \frac{g(Z_N) - g(u)}{g'(u)} - a_N(Z_N - u)$$

converges to zero in probability, by using the Taylor expansion formula). In other words, limit distributions are passed through functions in a pretty simple way. This is called the "delta method" (I suppose because of the deltas and epsilons involved in this kind of limiting argument), and we'll be using it a lot. The main application is when we've already proven a CLT for Z_N,

$$\sqrt{N} \frac{Z_N - \mu}{\sigma} \to_D N(0, 1),$$

in which case

$$\sqrt{N}(g(Z_N) - g(\mu)) \to_D N(0, \sigma^2 (g'(\mu))^2).$$

Exercise 40: Assume $N^{1/2} Z_N \to_D \mathcal{N}(0, 1)$. Then what is the asymptotic distribution of 1) $g(Z_N) = (Z_N - 1)^2$? 2) what about $g(Z_N) = Z_N^2$? Does anything go wrong when applying the delta method in this case? Can you fix this problem?

Mgf method

What if the r.v. we're interested in, Y_N, can't be written as $g(X_N)$, i.e., a nice function of an r.v. we already know converges? Are there methods to prove limit theorems directly?

Here we turn to our old friend the mgf. It turns out that the following generalization of the mgf invertibility theorem we quoted above is true:

Theorem 2. *The distribution functions F_N converge to F if:*

- *the corresponding mgf's $M_{X_N}(s)$ and $M_X(s)$ exist (and are finite) for all $s \in (-z, z)$, for all N, for some positive constant z.*
- *$M_{X_N}(s) \to M_X(s)$ for all $s \in (-z, z)$.*

So, once again, if we have a good handle on the mgf's of X_N, we can learn a lot about the limit distribution. In fact, this idea provides the simplest way to prove the CLT.

Proof: assume X_i has mean zero and unit variance; the general case follows easily, by the usual rescalings.

Now let's look at $M_N(s)$, the mgf of $\frac{1}{\sqrt{N}} \sum_{i=1}^{N} X_i$. If X_i has mgf $M(s)$, then $\frac{1}{\sqrt{N}} \sum_{i=1}^{N} X_i$ has mgf

$$M(s/\sqrt{N})^N.$$

Now let's make a Taylor expansion. We know that $M(0) = 1, M'(0) = 0$, and $M''(0) = 1$. (Why?) So we can write

$$M(s) = 1 + s^2/2 + o(s^2).$$

Now we just note that $M_N(s)$ converges to $e^{s^2/2}$, recall the mgf of a standard normal r.v., and then appeal to our general convergence-in-distribution theorem for mgf's. □

Part III
Estimation theory

We've established some solid foundations; now we can get to what is really the heart of statistics.

Point estimation

"Point estimation" refers to the decision problem we were talking about last class: we observe data X_i drawn i.i.d. from $p_\theta(x)$[16], and our goal is to *estimate* the parameter $\theta \in \Theta$ from the data. An "estimator" is any decision rule, that is, any function from the data space $\mathcal{X}^N$ into the parameter space Θ. E.g. the sample mean, in the Gaussian case. Or the function that assigns "2" to every possible observed data sample.

Bias and variance[17]

There are two important functions associated with any estimator $\hat{\theta}$ that are useful as a thumbnail sketch of how well the estimator is doing: the "bias" $B_{\hat{\theta}}(\theta) =$

$$E_\theta(\hat{\theta}-\theta) = \int_{-\infty}^{\infty} ... \int_{-\infty}^{\infty} \prod_{i=1}^{N} p_\theta(X_i) \left(\hat{\theta}(\{X_1, X_2, ...X_N\}) - \theta \right) \prod_{i=1}^{N} dX_i = E_\theta(\hat{\theta}) - \theta$$

and the variance

$$V_{\hat{\theta}}(\theta) = V_\theta(\hat{\theta})$$

There is a very useful relationship between the bias, variance, and the mean-square error (MSE) of any estimator. Using the usual rules for expectations of squares, it's easy to show that the square error decomposes into a bias and variance term:

$$E_\theta\left((\theta - \hat{\theta})^2 \right) = B_{\hat{\theta}}(\theta)^2 + V_{\hat{\theta}}(\theta)$$

So the MSE of an estimator can be simply described as a sum of a term measuring how far off the estimator is "on average" (not average square) and a term measuring the variability of the estimator.

[16]Note that this is a *huge* assumption, or rather set of assumptions. We assume that the data is mutually independent — that is, seeing one data point doesn't affect the other data points at all — and even more strongly, that the true underlying distribution of the data happens, rather too conveniently, to some easy-to-analyze family of distributions $p_\theta(x)$, where θ is some simple parameter that tells us everything we need to know about the data. The message is to take all of the following with a grain of salt: this i.i.d. story is a simple, tractable *model* of the data — while it's a very helpful model, as we'll see, it's important to remember that in 99% of cases it's something of a caricature.

[17]HMC 4.1

Note that both the bias and variance are functions of θ (and are therefore usually unknown, although we will see some exceptions to this below); the bias could be positive for some parameters but negative for others, for example.

Here's an example: let x_i be i.i.d. coin flips from some coin that has p probability of coming up heads. We want to estimate p. Then it's easy to compute the bias if we take our estimator to be the sample mean number of heads: we just need to compute

$$E_{B_{N,p}}(n/N) = p.$$

Therefore the bias is zero, no matter what p is. The variance is also easy to compute, if we recall the binomial variance and use our scaling rule for variance:

$$V_{B_{N,p}}(n/N) = \frac{1}{N}p(1-p).$$

Here the variance of our estimator depends on the parameter p.

Unbiased, minimum-variance estimators

This bias-variance decomposition leads to another possible way to choose between all possible admissible estimators. (Recall that we discussed two such principles before, the Bayes — minimum average error — principle and the minimax — minimum worst-case error.)

Here's a third: choose the best *unbiased* estimator. That is, choose an estimator $\hat{\theta}$ such that

$$B_{\hat{\theta}}(\theta) = 0 \quad \forall \theta,$$

and such that the variance is minimized over the class of all such unbiased estimators. Unbiased estimators are right "on average," which is not a bad thing to aim for (although the condition of exactly zero bias turns out to be pretty strong, ruling out a lot of otherwise good estimators, so it's much more questionable whether we should exclude all estimators with any bias whatsoever).

Exercise 52: Is the sample mean unbiased for Poisson data?

Exercise 53: Provide an unbiased estimator for b if the data is $U([0, b])$.

Exercise 54: Provide an unbiased estimator for σ^2, the variance parameter of the Gaussian distribution $\mathcal{N}(\mu, \sigma^2)$: 1) in the case that μ is known; 2) in the case that μ is unknown.

Note that this unbiasedness condition rules out trivial estimators such as $\hat{\theta}(D) \equiv 2 \ \forall D$, which is nice. In fact, in some situations we'll see that a "uniformly minimum-variance unbiased" estimator (UMVUE) exists: such an estimator satisfies

$$V_{\hat{\theta}_{UMVU}}(\theta) \leq V_{\hat{\theta}}(\theta) \quad \forall \theta,$$

for any unbiased estimator $\hat{\theta}$; therefore an UMVUE dominates all other unbiased estimators under the squared-error cost function. In this case, it obviously makes sense to use the UMVUE.

One last interesting thing to note: when a UMVUE does exist, it is automatically unique. Suppose U_1 and U_2 are both UMVUE, with variance $V(\theta)$, then the average $U = (U_1 + U_2)/2$ is also an unbiased estimator. Now let's look at the new function $W = (U_1 - U_2)/2$. Now

$$V_U + V_W = \frac{V_{U_1} + V_{U_2}}{2} = V_{U_1} \leq V_U.$$

This means $V_W = 0$, i.e., $U_1 = U_2$ with probability 1.

Exercise 55: Is the sample mean from a Gaussian with known variance (say, $\mathcal{N}(\mu, 1)$) a UMVUE for the mean parameter μ? If not, can you think of one, or prove that none exists? (Hint: try the case $N = 1$, then $N = 2$, first.)

Exercise 56: Is the sample maximum from a uniform distribution $U(a, b)$ a UMVUE for the maximum parameter b? If not, can you think of one, or prove that none exists? (Hint: try the case $N = 1$, then $N = 2$, first.)

Reparameterization

One very important thing to note about the unbiasedness property is that it is not invariant with respect to reparameterizations. That is, if we relabel the parameters, the estimator might not remain unbiased. This is, in fact, one of the strongest arguments that has been raised against the idea of restricting our attention to unbiased estimators. **Exercise 57**: Is the sample mean an unbiased estimator of $\sqrt{p}$, where p is the parameter of a binomial distribution? What about the square root of the sample mean: is $\hat{p} = \sqrt{n/N}$ an unbiased estimator of $\sqrt{p}$?

The other big drawback, as mentioned above, is that unbiased estimators don't necessarily exist. **Exercise 58**: Does an unbiased estimator exist for $\log p$, where p is the parameter of a binomial distribution? If so, supply one; if not, prove it.

Maximum likelihood estimators (MLE)

Figure 11: Fisher.

The idea of maximum likelihood is perhaps the most important concept in parametric estimation. It is straightforward to describe (if not always to implement), it can be applied to any parametric family of distributions (that is, any class of distributions that is indexed by a finite set of parameters $\theta_1, \theta_2, ..., \theta_d$), and it turns out to be optimal in an asymptotic sense we will develop more fully in a couple lectures.

The basic idea is to choose the parameter $\hat{\theta}_{ML}$ under which the observed data, D, was "most likely." I.e., choose the $\hat{\theta}_{ML}$ which maximizes the likelihood, $p_\theta(D)$. Let's think about this in terms of Bayes' rule: if we start with some prior on the parameters, $p(\theta)$, then the posterior on the parameters given the data is

$$p(\theta|D) = \frac{1}{Z} p(\theta) p(D|\theta) = \frac{1}{Z} p(\theta) p_\theta(D),$$

where $Z = p(D) = \int p(\theta) p_\theta(D) d\theta$ is a constant ensuring the normalization of the posterior. So if we ignore the prior $p(\theta)$ on the right hand side (or assume $p(\theta)$ is roughly constant in θ), then maximizing the likelihood is roughly the same as maximizing the posterior probability density of θ, given the data.

It's clear that this can be applied fairly generally. But how do we actually compute the maximum? Well, we can ask a computer to do it, using a

numerical optimization scheme. Sometimes we can find the optimum analytically. For example, assume (as usual) that we have i.i.d. data X_i. This means that

$$p_\theta(D) = \prod_i p_\theta(X_i).$$

This suggests that we maximize the *log-likelihood* instead, since sums are easier to work with than products:

$$\hat{\theta}_{ML} = \arg\max_\theta p_\theta(D) = \arg\max_\theta \prod_i p_\theta(X_i) = \arg\max_\theta \sum_i \log p_\theta(X_i).$$

Often we can take the gradient of

$$L(\theta) \equiv \sum_i \log p_\theta(X_i)$$

and set it to zero to obtain a local optimum; then we can sometimes additionally argue that the optimum is unique. **Exercise 59**: Give a simple condition, in terms of the second derivative of the log-likelihood $d^2 \log p_\theta(x)/d\theta^2$, ensuring that the likelihood has a unique global maximum as a function of $\theta \in \Theta$.

Here's an example: let x_i be exponential. Then

$$p_\theta(\{x_i\}) = \prod_i \theta e^{-\theta x_i},$$

so if we set the derivative of the loglikelihood equal to zero, we get

$$\begin{aligned} 0 &= \frac{\partial}{\partial\theta}\left(N\log\theta - \sum_{i=1}^N \theta x_i\right)\Big|_{\theta=\hat{\theta}_{MLE}} \\ &= \frac{N}{\hat{\theta}_{MLE}} - \sum_i x_i, \end{aligned}$$

so we have that

$$\hat{\theta}_{MLE} = \left(\frac{1}{N}\sum_i x_i\right)^{-1};$$

the MLE for θ is just the inverse of the sample mean of x_i. This makes sense: if we see that the sample mean is very close to zero, then it seems likely that θ is large.

Exercise 60: Find the MLE for (μ, σ^2) for Gaussian data, given N i.i.d. data samples x_i. Is the MLE biased? If so, compute the bias.

Exercise 61: Find the MLE for (a, b) for uniform $U(a, b)$ data, given N i.i.d. data samples x_i. Is the MLE biased? If so, compute the bias.

Invariance

One more important point about the MLE is that it is *invariant* with respect to reparameterizations. That is, if $\hat{\theta}_{ML}$ is an MLE for θ, then $g(\hat{\theta}_{ML})$ is an MLE for $g(\theta)$ whenever $g(.)$ is invertible. For example, the MLE for σ is just $\sqrt{\hat{\sigma}^2_{ML}}$. **Exercise 62**: Prove this.

Regression[18]

One important application of ML estimation is to regression analysis. Unfortunately we don't have time to go very deeply into this very important topic; check out W4315 for more information.

The basic model for regression is as follows: we see *paired* data $\{X_i, Y_i\}$, and we have reason to believe X_i and Y_i are related. In fact, we can hypothesize a model:

$$Y_i = aX_i + e_i,$$

where e_i is some (unobserved) i.i.d. noise source; i.e., Y is given by aX, a linearly-scaled version of X, but contaminated by additive noise. Let's assume that $e_i \sim \mathcal{N}(0, \sigma^2)$. What is the MLE for the parameters (a, σ^2)?

Well, first we write down the loglikelihood.

$$\begin{aligned} L(a, \sigma^2) &= \sum_i \log \mathcal{N}(0, \sigma^2)(Y_i - aX_i) \\ &= \sum_i -\log(\sigma\sqrt{2\pi}) - \frac{(Y_i - aX_i)^2}{2\sigma^2}. \end{aligned}$$

Now if we take the gradient and set it to zero, we get two equations (one for a and one for σ^2:

$$\sum_i X_i(Y_i - \hat{a}_{ML}X_i) = 0,$$

[18] HMC 12.3

and

$$\sum_i -\frac{1}{\sigma^2} + \frac{(Y_i - aX_i)^2}{(\sigma^2)^2} = 0.$$

So

$$\hat{a}_{ML} = \frac{\sum_i X_i Y_i}{\sum_i X_i X_i}$$

and

$$\hat{\sigma}^2_{ML} = \frac{\sum_{i=1}^N (Y_i - aX_i)^2}{N},$$

both of which have a fairly intuitive interpretation.

Again, check out W4315 to learn more about what to do when X is multidimensional, when more complicated (nonlinear) relationships hold between X and Y, when e_i is not normal or even i.i.d., etc.

Robustness

One very important point is that the MLE depends strongly on the parametric family chosen. For example, if your data is actually Cauchy, but you apply the MLE assuming Gaussian data, then you're not going to do very well. (**Exercise 63**: Why?) This is an extreme case, but a lot of work on the "robustness" of the MLE indicates that things can go fairly badly wrong even when your data is "mostly" Gaussian (e.g., when data are drawn from a "mixture" distribution

$$\sum a_i p_i(x),$$

where the mixture weights a_i are positive and sum to one; think e.g. of a Gaussian distribution for p_1 mixed with some occasional "outliers," $a_2 < a_1$, with p_2 having heavier tails than p_1). Since, of course, we don't know a priori what distribution our data is drawn from, this is a bit of a problem. If there's time at the end of the semester after developing the basic theory, we'll return to this robustness question. If not, of course, feel free to look up this topic on your own; see 12.1-12.2 in HMC for a start.

Sufficiency[19]

Let's look more closely at this likelihood idea. One thing that we saw was that not every bit of the data really mattered to our estimate. For example, if we have i.i.d. data, it doesn't really matter what order the data appeared in. So we can throw out the order of the data and do inference given the unordered data just as well.

Similarly, se saw that we didn't need to remember everything about Gaussian data, just the sample mean and sample variance (or equivalently, the sample mean and sample mean square, since we can derive one from the other), since the likelihood depends on the data only through these statistics:

$$\begin{aligned} L(\mu,\sigma^2) &= \sum_i \log \mathcal{N}(\mu,\sigma^2)(X_i) \\ &= \sum_i -\log(\sigma\sqrt{2\pi}) - \frac{(X_i-\mu)^2}{2\sigma^2} \\ &= -N\log(\sigma\sqrt{2\pi}) - \frac{1}{2\sigma^2}\left(\sum_i X_i^2 - 2\mu\sum_i X_i + N\mu^2\right). \end{aligned}$$

This is quite a savings: we've compressed N data points into just two.

This turns out to be a pretty common phenomenon, if we think about it. We're led to a definition: Any function $T(D)$ of the data D (and only of the data D) is called a "statistic." A statistic is called *sufficient* for the parameter θ if we can split the data into two parts: 1) the sufficient statistic, and 2) the aspects of the data that have nothing to do with estimating θ.

More mathematically,

$$p_\theta(D) = F(\theta, T(D))G(D),$$

for some functions $F(.)$ and $G(.)$. This is equivalent (although we'll skip the proof) to saying that the conditional distribution of the data D, given $T(D)$, does not depend on θ at all. That is, the function

$$p_\theta(D|T) = \frac{p_\theta(T(D)|D)p_\theta(D)}{p_\theta(T(D))} = \frac{p_\theta(D)}{p_\theta(T(D))}$$

[19] HMC 7.

does not depend on θ.

(Yet another way of saying this: θ — T — D is a "Markov chain": D is conditionally independent of θ given T, for any prior distribution on θ, i.e.

$$p(\theta, D|T) = p(\theta|T)p(D|T).)$$

Here are some more examples:

- Binomial data: if $x_i = 1$ or 0 depending on if the i-th i.i.d. coin flip came up heads or tails, then

  $$p_\theta(\{x_i\}) = \binom{N}{\sum_i x_i} p^{\sum_i x_i}(1-p)^{N-\sum_i x_i};$$

  from this we can easily see that $n = \sum_i x_i$ is sufficient.

- If we have N i.i.d. Poisson observations, then

  $$p_\theta(\{x_i\}) = \prod_i e^{-\theta}\frac{\theta^{x_i}}{x_i!} = e^{-N\theta}\frac{\theta^{\sum_i x_i}}{\prod_i (x_i!)},$$

  and once again $\sum_i x_i$ is sufficient.

- For uniform $U[(0,\theta])$ data,

  $$p_\theta(\{x_i\}) = \prod_i 1_{[\theta \geq x_i]}\frac{1}{\theta} = \frac{1}{\theta^N}1_{[\theta \geq \max_i x_i]},$$

  i.e., $\max_i x_i$ is sufficient.

- For uniform $U[(\theta, \theta+1])$ data,

  $$p_\theta(\{x_i\}) = 1_{[\theta \leq \min_i x_i]}1_{[\theta+1 \geq \max_i x_i]}$$

  i.e., the *pair* $(\min_i x_i, \max_i x_i)$ is sufficient (even though there is only one parameter θ).

Exercise 64: What is a sufficient statistic for exponential data, $x_i \sim \exp(\lambda)$?

Exercise 65: What is a sufficient statistic for uniform data, $x_i \sim U([a,b])$?

Exercise 66: What is a sufficient statistic for Gaussian data, $x_i \sim \mathcal{N}(0,\theta)$? (I.e., the mean is known but the variance is not.)

A couple things to note:

- The MLE can only depend on the data through sufficient statistics, since

$$\arg\max_\theta p_\theta(D) = \arg\max_\theta F(\theta, T(D))G(D) = \arg\max_\theta F(\theta, T(D)).$$

  Exercise 67: Prove the following statement (if true), or (if false) give a counterexample and salvage the statement if possible. "If the MLE is unique, it must necessarily be a function of any sufficient statistic."

- For similar reasons, Bayes estimators only depend on the data through sufficient statistics. **Exercise 68**: Show this using Bayes' rule.

- Sufficiency is only defined in the context of a parametric family. That is, a statistic may be sufficient for one parametric family but not for another one. **Exercise 69**: Give an example of this.

- any invertible function (relabeling) of a sufficient statistic is itself sufficient. (Hence sufficient statistics are very nonunique.) **Exercise 70**: Prove this using the factorization definition of sufficiency.

Minimal sufficiency

This last point leads to another important concept. A sufficient statistic is *minimal* if it can be written as a function of every other conceivable sufficient statistic. In a sense, minimal statistics have all the redundancy compressed out — there's nothing irrelevant left to throw out.

In a sense, anything that doesn't change the likelihood can be thrown out, as the following simplification of a theorem (which we won't prove) by Lehmann-Scheffé shows:

Theorem 3. *T is minimal sufficient if and only if the following two statements are equivalent:*

1. *For any data samples D and D',*

$$\frac{p_\theta(D)}{p_\theta(D')} \text{ is constant in } \theta$$

2. *$T(D) = T(D')$.*

This condition is often much easier to check than whether a given statistic is a function of every other sufficient statistic.

Example: $x_i \sim \exp(\theta)$. Then

$$p_\theta(D) = \prod_{i=1}^{N} \theta \exp(-\theta x_i) = \theta^N \exp(-\theta \sum_{i=1}^{N} x_i)$$

for $x_i > 0 \; \forall i$. Clearly $\sum x_i$ is sufficient; is it minimal? Let's apply the above theorem: choose another arbitrary data sample, $D' = \{x'_1, x'_2, ..., x'_N\}$. Now we can see that

$$\frac{p_\theta(D)}{p_\theta(D')} = \exp\left[\theta(\sum x'_i - \sum x_i)\right]$$

is constant as a function of θ if and only if $\sum x_i = \sum x'_i$. Thus $\sum x_i$ is a minimal sufficient statistic; this was much easier to check than whether $\sum x_i$ was a function of all other possible sufficient stats! Conversely, let's look at a sufficient statistic which is not minimal, namely the full data D. Here it's clear that $2 \implies 1$ but 1 does not imply 2; hence, the full data D is not minimal.

Here's another one. Recall that $(\min_i x_i, \max_i x_i)$ was sufficient for uniform $U[(\theta, \theta + 1])$ data. It's clear from this theorem that this statistic is also minimal.

We'll see more examples in a moment, when we discuss exponential families.

Exercise 71: Let $x \sim \mathcal{N}(0, \sigma^2)$ (i.e., the mean is known but the variance is not). Is $|x|$ sufficient for σ^2? If so, is it minimal?

Exercise 72: Let x_i be drawn i.i.d. from a density in a "location family"; that is, we know $f(x)$ and we know $p_\theta(x) = f(x - \theta)$, we just want to know θ. Can you come up with a minimal sufficient statistic for θ (and, of course, prove that this statistic is minimal sufficient)? (Hint: the order statistics might be useful here, as it's intuitively clear that we don't need to remember what order the data actually came in.)

Rao-Blackwell theorem[20]

Not only does restricting our attention to sufficient statistics make life easier; it also improves (or at least can't hurt) our estimation accuracy:

[20] HMC 7.3


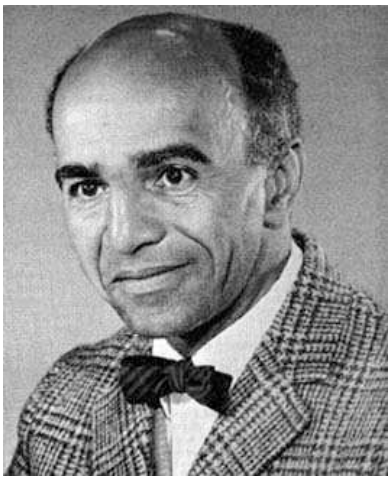

Figure 12: Rao and Blackwell.

Theorem 4 (Rao-Blackwell)**.**

$$E[g(E(\hat{\theta}|T) - \theta)] \leq E[g(\hat{\theta} - \theta)]$$

for any estimator $\hat{\theta}$, convex error function g, and sufficient statistic T. In particular,

$$E[(E(\hat{\theta}|T) - \theta)^2] \leq E[(\hat{\theta} - \theta)^2]$$

In words, take any estimator $\hat{\theta}$. Then form the estimator $E(\hat{\theta}|T)$ (note that $E(\hat{\theta}|T)$ is a bona fide statistic — that is, doesn't depend on θ — if T is sufficient). The theorem says that the risk of $E(\hat{\theta}|T)$ is never worse (and in practice, often better) than that of the original estimator $\hat{\theta}$, as long as the loss function g is convex. Also note that $E(\hat{\theta}|T)$ is unbiased whenever $\hat{\theta}$ is (**Exercise 73**: Why?).

We'll prove the special ($g(u) = u^2$) case to give a sense of what's going on here. Just write out $E[(\hat{\theta} - \theta)^2]$:

$$\begin{aligned} E[(\hat{\theta} - \theta)^2] &= E_T\Big(E[(\hat{\theta} - \theta)^2|T]\Big) \\ &= E_T[(E(\hat{\theta}|T) - \theta)^2] + E_T\Big(E[(\hat{\theta} - E(\hat{\theta}|T))^2]\Big) \\ &\geq E_T[(E(\hat{\theta}|T) - \theta)^2]. \end{aligned}$$

Exercise 74: Prove the general case using Jensen's inequality and the rules for combining conditional expectations.

It's worth noting that a very similar proof establishes the important fact we've used a couple times now, that the optimal Bayesian estimator under square loss is the conditional expectation of θ given the data. I'll leave the proof as an exercise.

Also, the proof doesn't seem to rely directly on sufficiency — the inequalities above hold for any statistic T, not just for sufficient T. The point to remember, again, is that if T is not sufficient then $E(\hat{\theta}|T)$ is not guaranteed to even be a valid statistic.

Exponential families[21]

Let's talk about a class of statistical families whose sufficient statistics are very easy to describe. We say a parametric family is an "exponential family" if

$$p_\theta(x) = \begin{cases} \exp\left(f(\theta)k(x) + s(x) + g(\theta)\right) & \text{if } a < x < b \\ 0 & \text{otherwise,} \end{cases}$$

for some $-\infty \leq a, b \leq \infty$ (note that a and b *don't* depend on θ).

An example: $x \sim \exp(\theta)$.

$$p_\theta(x) = \exp([-\theta x] + [0] + [\log \theta]),$$

from which we can read off $f(\theta), k(x), s(x)$, and $g(\theta)$.

Another example: $\mathcal{N}(\theta, \sigma^2)$ (σ^2 known). After a little manipulation, we can write

$$p_\theta(x) = \exp\left([\frac{-\theta}{\sigma^2}x] + [\frac{-x^2}{2\sigma^2} - \frac{1}{2}\log(2\pi\sigma^2)] + [\frac{-\theta^2}{2\sigma^2}]\right).$$

It's pretty easy to define a minimal sufficient statistic here: if we write

$$p_\theta(x) \sim \exp\left(f(\theta)k(x)\right)\exp\left(s(x)\right)$$

and recall the sufficient statistic factorization, then $k(x)$ is a good candidate.

[21] HMC 7.5.

Now for the really useful part: if we look at multiple i.i.d. samples x_i, then it's easy to see that $\sum_i k(x_i)$ is minimal sufficient for the full data $\{x_i\}$. (**Exercise 75**: Prove this.) This saves a whole lot of work — to come up with a minimal s.s. (and therefore come up with an improved estimator, according to Rao-Blackwell), all we need to do is manipulate $p_\theta(x)$ into the above form. We'll get some more practice with this in a moment.

More generally, sometimes we need more than one statistic to adequately describe the data. In this case, we can define a k-dimensional exponential family as a parametric family satisfying

$$p_\theta(x) = \begin{cases} \exp\left(\sum_{j=1}^k f_j(\theta) k_j(x) + s(x) + g(\theta) \right) & \text{if } a < x < b \\ 0 & \text{otherwise.} \end{cases}$$

Here, $\{k_j(x)\}_{1 \le j \le k}$ are minimal sufficient together (but not alone).

These concepts give us a canonical parameterization of our parametric family: $f_j(\theta)$ (we call $\{f_j(\theta)\}_{1 \le j \le k}$ the "canonical parameter"), and the natural parameter space is the set of all θ for which the above form makes sense, that is, the set of θ such that $g(\theta)$, as defined above, is finite.

Exercise 76: Write out $f(\theta), k(x), s(x)$, and $g(\theta)$ for 1) $\mathcal{N}(\mu, \theta)$ (μ known), 2) $B(N, \theta)$, and 3) $Poiss(\theta)$. Write out $f_j(\theta), k_j(x), s(x)$, and $g(\theta)$ for the normal family in the case that both μ and σ^2 are unknown.

Now to make life simpler assume we're dealing with the canonical parameterization, i.e. $f(\theta) = \theta$. Let's look more closely at $g(\theta)$. First, there is some redundancy here: we know, since $\int p_\theta(x) dx = 1$, that

$$g(\theta) = -\log\left(\int \exp[\theta k(x) + s(x)] dx \right).$$

We can go a little further if we remember some of our facts about moment-generating functions and recognize that an mgf is hiding in the above definitions. Now, remember that taking derivatives of mgf's kicks out moments (hence the name). In this case, we have

$$\frac{\partial g}{\partial \theta} = -E_\theta k(x).$$

This is because

$$\begin{aligned} -\frac{\partial g}{\partial \theta} &= \frac{\partial}{\partial \theta} \log \left(\int \exp[\theta k(x) + s(x)] dx \right) \\ &= \frac{\int k(x) \exp[\theta k(x) + s(x)] dx}{\exp[-g(\theta)]} \\ &= \int p_\theta(x) k(x) dx = E_\theta k(x). \end{aligned}$$

Exercise 77: Derive a relationship between $g(\theta)$, $E_\theta k(x)$, and the MLE by taking the derivative of the loglikelihood and setting it to zero.

We can use similar techniques to show that

$$\frac{\partial^2 g}{\partial \theta^2} = -V_\theta k(x).$$

This, in turn, proves that the log-likelihood $\log p_\theta(x)$ is a concave function of θ whenever θ is the canonical parameter of an exponential family, which you'll recall is quite handy in the context of ML estimation. **Exercise 78**: Prove the above formula, and use this to establish the concavity of the log-likelihood in the canonical exponential family setting.

Of course, exponential families are a special case; life isn't always so easy. **Exercise 79**: Try writing $U(a, b)$ in the exponential form. What goes wrong? (Hint: don't forget to keep track of the support of $U(a, b)$.)

It's interesting to note (though we won't pursue this) that exponential families are the only ones for which a finite-dimensional sufficient statistic exists for all sample sizes N. This is called the "Koopman - Darmois" theorem, if you want to read more about it.

Exercise 80: Give a minimal sufficient statistic for Cauchy data.

Completeness and uniqueness (time permitting)[22]

"Completeness" of a statistic, in the context of a given probability family, is a property that guarantees the uniqueness of the unbiased estimator which may be written as a function of a sufficient statistic; this estimator is then automatically the UMVUE.

[22] HMC 7.4.

We call the statistic $U(x)$ "complete" if[23]

$$E_\theta(g(U)) = 0 \ \ \forall\theta \implies g(U) = 0.$$

Exercise 81: Prove that the completeness of a sufficient statistic U, as defined above, guarantees that if $\phi(U)$ is an unbiased estimator for θ, then $\phi(U)$ is the UMVUE. (Hint: think about what the completeness condition says about the difference between $\phi(U)$ and any other unbiased estimator that is a function of U. Then think about Rao-Blackwell, and the uniqueness of UMVUEs.)

Exercise 82: Is the natural sufficient statistic in an exponential family complete?

[23]The term "complete" is inherited from functional analysis (or, in the case of discrete data, linear algebra): $U(x)$ is complete if $p_\theta(U)$ is complete in $\mathrm{L}_2(\mathcal{U})$, the space of square-integrable functions on the range of $U(x)$, $\mathcal{U}$. If you've taken linear algebra, just think of functions of U as vectors — you can add them and multiply them by scalars to get new functions of U, just like ordinary vectors. Now the completeness condition just says that $p_\theta(U)$ span the set of all functions of U: if any function is orthogonal to all $p_\theta(U)$ (where we interpret $E_\theta g(U) = \int p_\theta(u)g(u)du$ as a dot product), then the function must be zero.

Asymptotic ideas

Now we turn to the asymptotic properties of our estimators. We'll discuss two questions in particular:

1. Does our estimator work properly asymptotically? That is, does it provide us with the correct answer if we give it enough data?
2. How asymptotically efficient is our estimator? For example, can we come up with an estimator which is at least as good as any other estimator, in some asymptotic sense?

Consistency[24]

We say an estimator is *consistent* (in probability) if it provides us with the correct answer asymptotically. That is,

$$\hat{\theta} \rightarrow_P \theta.$$

(More precisely, we're talking about convergence of a *sequence* of estimators, one for each N, i.e.,

$$\hat{\theta}_N \rightarrow_P \theta.$$

But usually we'll suppress this extra notation.)

How can we establish that an estimator is consistent? Well, the easiest thing to do is to establish that the estimator is asymptotically unbiased,

$$B(\theta, N) \rightarrow_{N\rightarrow\infty} 0,$$

and that the variance goes to zero,

$$V(\theta, N) \rightarrow_{N\rightarrow\infty} 0;$$

then we can just apply our bias-variance decomposition and Chebysheff's inequality, and we're done.

Exercise 83: Use this method to develop a simple consistent estimator of the binomial parameter p.

It's worth noting that it's possible to come up with examples in which the estimator is consistent but either the bias or the variance does not tend

[24]HMC p. 206.

to zero; in other words, for consistency it is sufficient but not necessary that the bias and variance both tend to 0. For example, an estimator might have very fat tails, such that the variance is infinite for any N, but nonetheless most of the mass of $p(\hat{\theta}_N)$ becomes concentrated around the true θ. I'll leave it to you to work out the details of such an example.

Method of moments

One way to generate consistent estimators is to find a function $U(x)$ of a single observation X_i such that $E_\theta(U(x)) = f(\theta)$, where $f(\theta)$ is chosen to be a one-to-one function with a continuous inverse. Then we can use our results about convergence in probability of continuous functions to see that the estimator

$$\hat{\theta}_{MM} = f^{-1}\left(\frac{1}{N}\sum_{i=1}^{N} U(x_i)\right)$$

is consistent for θ. (To see this, note that

$$\frac{1}{N}\sum_{i=1}^{N} U(x_i) \rightarrow_P E_\theta U(x) = f(\theta),$$

by the law of large numbers and the definition of $f(\theta)$, then apply f^{-1} to both sides and use what we've learned about continuous mappings and convergence in probability.)

In the case of multiple parameters, we would solve this equation simultaneously for several U_j. When U_j are taken to be the first j moments of X (i.e., we choose $\hat{\theta}$ to match the observed moments to the true moments as a function of θ), this estimation technique is known as the "method of moments."

Here's an example. Let $x_i \sim exp(\theta)$. Now let $U(x) = x$. Then

$$E_\theta(U(x)) = f(\theta) = 1/\theta.$$

Therefore

$$\hat{\theta}_{MM} = \left(\frac{1}{N}\sum_{i=1}^{N} U(x_i)\right)^{-1} = \left(\frac{1}{N}\sum_{i=1}^{N} x_i\right)^{-1}$$

is consistent for θ.

We saw another example of this kind of moment-matching estimator recently in the homework: in an exponential family with the canonical parameterization, we saw that

$$E_{\hat{\theta}_{MLE}}(U(x)) = \frac{1}{N}\sum_{i=1}^{N} U(x_i),$$

where $U(x) = k(x)$ is the minimal sufficient statistic of the exponential family. Thus in this special case (but not in general!) the MLE is exactly the method of moments estimator.

Exercise 84: Develop the "method of moments" estimator for λ, the parameter of the Poisson distribution, using a) $U(x) = x$ and b) $U(x) = x^2$. Are these estimators consistent? How are these estimators related to the MLE?

Exercise 85: Develop the "method of moments" estimator for (μ, σ^2), the parameters of the Gaussian distribution. Is this estimator consistent? How is this estimator related to the MLE?

Exercise 86: Assume $U(x)$ has some finite variance, $V_\theta(U(x))$, and that f^{-1} is continuously differentiable, with strictly nonzero derivative. Use the central limit theorem and the delta method to derive the asymptotic distribution of $\hat{\theta}_{MM}$.

This type of estimator, constructed by finding the solutions of some equations that the parameter estimate must satisfy, is often called a "Z-estimator," because of the special case of the MLE, when we set the gradient of the likelihood equal to zero (hence the "Z") and solve the resulting equations. We'll look at some more examples in the next section.

Convergence rates and asymptotic normality

Just as we discussed the CLT as a "finer" result than the LLN, we'd like to know more about an estimator than just the fact that it converges in probability. For example, we'd like to know the convergence *rate* — how quickly it converges to θ, for example the $N^{-1/2}$ rate we saw when adding together i.i.d. r.v.'s — and once the rate is established, what the asymptotic distribution is on the scale defined by the convergence rate. For example, can we prove asymptotic normality on the $N^{-1/2}$ scale,

$$N^{1/2}(\hat{\theta} - \theta) \rightarrow_D \mathcal{N}(0, \sigma^2)?$$

And finally, what is the variance σ^2 of this asymptotic rescaled distribution? We'll address these questions for the MLE in the next section.

Confidence intervals[25]

Before we get too deeply into the question of how to prove these kinds of results, let's step back a moment and think about what we're going to do with this kind of asymptotic approximation. Probably the most important application of this idea is in the construction of *confidence intervals* — "error bars."

Let's imagine we know that

$$\sqrt{N}(\hat{\theta} - \theta) \rightarrow_D \mathcal{N}(0, \sigma^2(\theta)),$$

for some estimator $\hat{\theta}$. That is, $\hat{\theta}$ is asymptotically unbiased and normal about the true parameter θ (for now assume we know the asymptotic variance coefficient $\sigma^2(\theta)$, even though we don't know θ). How can we use this information? Well, by definition of asymptotic normality, we know that

$$P\left(-2 < \frac{\sqrt{N}(\hat{\theta} - \theta)}{\sigma(\theta)} < 2\right) \approx 0.95;$$

thus

$$P\left(\hat{\theta} - 2\frac{\sigma(\theta)}{\sqrt{N}} < \theta < \hat{\theta} + 2\frac{\sigma(\theta)}{\sqrt{N}}\right) \approx 0.95.$$

So we've gone from just making a guess about θ to something stronger: we've bracketed θ in a set, $(\hat{\theta} - 2\frac{\sigma(\theta)}{\sqrt{N}}, \hat{\theta} + 2\frac{\sigma(\theta)}{\sqrt{N}})$, into which θ falls with about 95% probability, assuming N is sufficiently large (it's essential to remember — and unfortunately easy to forget — that this argument only holds asymptotically as $N \rightarrow \infty$). In other words, we've given an approximate "95% confidence interval" for θ.

We left one little problem: how do we get $\sigma(\theta)$ without knowing θ? Well, the coefficient $\sigma(\theta)$ is a function of the parameter θ. If we can estimate θ, then we can also estimate a function of θ. So we estimate $\sigma(\theta)$: we know from the continuity properties of stochastic convergence theory that if an estimator $\hat{\sigma}$ is consistent for $\sigma(\theta)$, and recall we're already assuming that

$$\sqrt{N}(\hat{\theta} - \theta) \rightarrow_D \mathcal{N}(0, \sigma^2(\theta)),$$

[25] HMC 5.4

then[26]

$$\sqrt{N}\left(\frac{\hat{\theta}-\theta}{\hat{\sigma}}\right) \rightarrow_D \mathcal{N}(0,1).$$

Note, importantly, that the (unknown) parameter θ no longer appears on the right hand side. Now, applying the same logic as above, $(\hat{\theta} - 2\hat{\sigma}/\sqrt{N}, \hat{\theta} + 2\hat{\sigma}/\sqrt{N})$ is an approximate 95% confidence interval for θ, and importantly, we don't need to know θ to construct this interval.

Exercise 87: Generalize this analysis to get a 99% confidence interval (instead of 95%). What about the $(1-\alpha)\cdot 100\%$ confidence interval, where $0 < \alpha < 1$ is some arbitrary (fixed) error tolerance?

Let's look at a simple example. Let $n \sim Bin(N,p)$. Then $\hat{p}_{MLE} = n/N$. We know from the standard CLT and the formula for the variance of n that

$$\sqrt{N}\left(\hat{p}_{MLE} - p\right) \rightarrow_D \mathcal{N}(0, \sigma^2(p)),$$

where

$$\sigma^2(p) = p(1-p).$$

A simple estimator for σ is

$$\hat{\sigma} = \sqrt{\sigma^2(\hat{p}_{MLE})} = \sqrt{\hat{p}_{MLE}(1-\hat{p}_{MLE})};$$

we can prove that this estimator is consistent by our usual delta method arguments. Thus

$$\sqrt{N}\left(\frac{\hat{p}_{MLE}-p}{\hat{\sigma}}\right) \rightarrow_D \mathcal{N}(0,1),$$

and $(\hat{p}_{MLE} - 2\hat{\sigma}/\sqrt{N}, \hat{p}_{MLE} + 2\hat{\sigma}/\sqrt{N})$ is an approximate 95% confidence interval for p.

[26] If you're reading along in HMC, be careful — the corresponding equation on page 255 (the equation just before eq. 5.4.4) is incorrect. Do you see why?

MLE asymptotics[27]

Finally we get to the discussion of the asymptotic properties of the maximum likelihood estimator. As we said long ago, when we first introduced the idea of ML, really the best justification for the MLE is in its asymptotic properties: it turns out to be asymptotically optimal in a sense we will define below.

Consistency: identifiability and the Kullback-Leibler divergence

Before we talk about the asymptotic optimality of the MLE, though, let's ask a more basic question: is the MLE even consistent in general?

The answer is (generally speaking) yes, if the parameters are "identifiable," that is, if the distribution of data under any parameter θ in our parameter space Θ differs from the distribution of the data under any other parameter, θ'. That is,

$$p_\theta(D) \neq p_{\theta'}(D), \ \forall \ D.$$

This condition makes intuitive sense — if two parameters, say θ_1 and θ_2 — were not identifiable, of course we wouldn't be able to distinguish them based on their likelihoods (because the likelihoods would be equal), and so the MLE is doomed to be inconsistent.

The interesting thing is that this simple identifiability condition is enough to guarantee consistency in most cases, if the data are i.i.d. Let's write out the likelihood and try to manipulate it into a form we can deal with.

$$\log p_\theta(x_1, x_2, ...x_N) = \sum_{i=1}^{N} \log p_\theta(x_i).$$

Let's say the true parameter is θ_0. We don't know θ_0, of course (otherwise we wouldn't need to estimate it), but subtracting off its (unknown) loglikelihood and dividing by N won't change the location of the MLE:

$$\hat{\theta}_{ML} = \arg\max_\theta \sum_{i=1}^{N} \log p_\theta(x_i) = \arg\max_\theta \frac{1}{N} \sum_{i=1}^{N} \left(\log p_\theta(x_i) - \log p_{\theta_0}(x_i) \right);$$

[27] HMC 6.1-6.2

remember, $\log p_{\theta_0}(x_i)$ is constant in θ, so subtracting it off doesn't perturb the MLE at all.

Now we're left with something that looks a little more familiar: the *log-likelihood ratio*

$$\frac{1}{N}\sum_{i=1}^{N}\left(\log p_\theta(x_i) - \log p_{\theta_0}(x_i)\right) = \frac{1}{N}\sum_{i=1}^{N}\log\frac{p_\theta(x_i)}{p_{\theta_0}(x_i)}$$

— the log of the ratio of the likelihood of the data under θ_0 and under θ — is a sample average of i.i.d. r.v.'s! So, by the LLN,

$$\frac{1}{N}\sum_{i=1}^{N}\log\frac{p_\theta(x_i)}{p_{\theta_0}(x_i)} \to_P E_{\theta_0}\log\frac{p_\theta(x)}{p_{\theta_0}(x)} = -E_{\theta_0}\log\frac{p_{\theta_0}(x)}{p_\theta(x)};$$

the expectation is taken under θ_0 because, remember, that is the true parameter (i.e., our data are drawn i.i.d. from $p_{\theta_0}(x)$).

Now, this last term has a name you might have encountered before; $E_{\theta_0}\log\frac{p_{\theta_0}(x)}{p_\theta(x)}$ is called the "Kullback-Leibler divergence" between the distributions $p_{\theta_0}(x)$ and $p_\theta(x)$. This is called a "divergence" because it measures the distance between $p_{\theta_0}(x)$ and $p_\theta(x)$, in the sense that

$$D_{KL}(p_{\theta_1};p_{\theta_2}) = E_{\theta_1}\log\frac{p_{\theta_1}(x)}{p_{\theta_2}(x)} \geq 0,$$

with equality only when $p_{\theta_1}(x) = p_{\theta_2}(x)$ with p_{θ_1}-probability one. **Exercise 88**: Prove this using Jensen's inequality.

So what have we learned? We now know that, for any fixed θ, the normalized log-likelihood ratio, $\frac{1}{N}\sum_{i=1}^{N}\log\frac{p_\theta(x_i)}{p_{\theta_0}(x_i)}$, tends to a function, $-D_{KL}(p_{\theta_0};p_\theta)$, which has a unique maximum at θ_0. (Why is the maximum unique? Identifiability.) So we can argue that the MLE asymptotically picks out the argmax of $-D_{KL}(p_{\theta_0};p_\theta)$, i.e., is consistent. (Actually, completing this consistency argument rigorously does require a couple technical conditions — e.g., it is enough that $p_\theta(x)$ is continuous in θ with probability one, and

$$E_{\theta_0}\left(\max_{\theta\in\Theta}|\log p_\theta(x)|\right) < \infty$$

— but we'll ignore these technical details. The basic logic — LLN + definition of K-L divergence + Jensen — should be clear.)

Asymptotic normality and the Fisher information

OK, that takes care of consistency: now we know that $\hat{\theta}_{MLE} \to_P \theta_0$. But just as in the LLN case, we want to know more. How fast does the MLE converge? What is the limiting (rescaled) distribution?

It turns out to be useful and informative to step back and look at the behavior of the posterior density. We know from the above that

$$\begin{aligned} p(\theta|x_1, x_2, ...x_N) = \frac{1}{Z} p(\theta) p(x_1, x_2, ...x_N|\theta) &= \frac{1}{Z} p(\theta) \prod_{i=1}^{N} p(x_i|\theta) \\ &\approx \frac{1}{Z} \exp(-N D_{KL}(p(x|\theta_0); p(x|\theta))). \end{aligned}$$

Note immediately that we're ignoring the prior $p(\theta)$ asymptotically; as N becomes large the likelihood term dominates the shape of the posterior, because the likelihood term is growing linearly with N, whereas the prior term is fixed as a function of N.

Next, remember that $-D_{KL}(\theta_0; \theta)$ has a unique maximum at θ_0; this implies that

$$\frac{\partial}{\partial \theta} D_{KL}(\theta_0; \theta)\Big|_{\theta=\theta_0} = 0.$$

Now if we make a second-order expansion around θ_0,

$$\begin{aligned} \log p(\theta|x_1, x_2, ...x_N) &\sim -N D_{KL}(\theta_0; \theta) \\ &= -N(0 + 0 + \frac{1}{2}(\theta - \theta_0) I(\theta_0)(\theta - \theta_0) + ...), \end{aligned}$$

i.e., the posterior likelihood is well-approximated by a Gaussian with mean θ_0 and variance

$$\frac{1}{N} I(\theta_0)^{-1},$$

where we have abbreviated the curvature of the D_{KL} function at θ as

$$I(\theta_0) = \frac{\partial^2}{\partial \theta^2} D_{KL}(\theta_0; \theta)\Big|_{\theta=\theta_0}.$$

This simple geometric quantity $I(\theta_0)$ is called the "Fisher information" at θ_0; it's called "information" because the larger I is, the smaller the asymptotic variance of the posterior is — thus, in a sense, large values of I indicate that

the data tell us a lot about the true underlying θ, and vice versa. (This number is named after Fisher, the statistician who first developed a great deal of the estimation theory we've been talking about in this section.)

What about the mean of this Gaussian? We know it's close to θ_0, because the posterior decays exponentially everywhere else. (This is another way of saying that $\hat{\theta}_{MLE}$ is consistent.) But how close exactly? We took a relatively crude approach above: we took the random log-likelihood function $\log p(x_i|\theta)$ and substituted its average, $D_{KL}(\theta_0;\theta)$. What if we try expanding the loglikelihood directly: We look at $\sum_i \log p(x_i|\theta)$:

$$\begin{aligned}\sum_i \log p(x_i|\theta_0) &+ \sum_i \frac{\partial}{\partial\theta}\log p_\theta(x_i)\Big|_{\theta_0}(\theta-\theta_0) + \frac{1}{2}\sum_i \frac{\partial^2 \log p(x_i|\theta)}{\partial\theta^2}\Big|_{\theta_0}(\theta-\theta_0)^2 \\ &\approx K_N + \sum_i \frac{\partial}{\partial\theta}\log p_\theta(x_i)\Big|_{\theta_0}(\theta-\theta_0) - \frac{1}{2}NI(\theta_0)(\theta-\theta_0)^2\end{aligned}$$

Look at

$$\frac{\partial}{\partial\theta}\log p_\theta(x)\Big|_{\theta_0}.$$

This is an important random variable — albeit one with a somewhat complicated definition — known as the "score." **Exercise 89**: Prove that this r.v. has mean zero and variance $I(\theta_0)$ (a nice coincidence!). (Caveat: to prove this, you'll need to interchange an integral and a derivative, which isn't always legal. We'll mostly ignore this mathematical delicacy here, but note that it does lead to problems in some cases, e.g. in the case that $p_\theta(x)$ is uniform $U(0,\theta)$.)

So we can apply the CLT: if we abbreviate

$$G_N = \sum_i \frac{\partial}{\partial\theta}\log p_\theta(x_i)\Big|_{\theta_0},$$

then G_N is asymptotically Gaussian, with mean zero and variance $NI(\theta_0)$.

So $\log p(D|\theta)$ looks like a random upside-down bowl-shaped function:

$$\sum_i \log p(x_i|\theta) \sim G_N(\theta-\theta_0) - \frac{1}{2}NI(\theta_0)(\theta-\theta_0)^2.$$

The curvature of this bowl is $-NI(\theta_0)$. The top of the bowl (i.e., the MLE) is random, on the other hand, asymptotically Gaussian with mean θ_0 and

variance $(NI(\theta_0))^{-1}$. **Exercise 90**: Prove this, using what you know about the peak of an upside-down quadratic, what you already know about the mean and variance of G_N, and the usual rules for multiplication of variances and the fact that Gaussians are preserved under addition and multiplication by scalars.

To sum up, our main result: the posterior likelihood is well-approximated by a Gaussian shape, with variance $(NI(\theta_0))^{-1}$ and mean

$$\sqrt{N}\left(\hat{\theta}_{MLE} - \theta_0\right) \rightarrow_D \mathcal{N}(0, I(\theta_0)^{-1}).$$

Note that, as we've seen with our concrete examples, the variance asymptotically depends on the underlying true parameter θ_0, because the Fisher information $I(\theta_0)$ depends on θ_0.

Exercise 91: We've seen some examples where computing the asymptotic variance of the MLE is easy by direct methods. Use direct methods to compute the asymptotic variance of the MLE, then use the above formula to derive the Fisher information $I(\theta_0)$, for a) the Gaussian with unknown mean and known variance; b) the binomial; c) the Poisson.

Exercise 92: Use the delta method to compute the asymptotic variance of the MLE for exponential data. Now compute the Fisher information $I(\theta_0)$, and from this derive the asymptotic variance. Do your answers agree?

Exercise 93: Compute the score and Fisher information in an exponential family with the canonical parameterization.

Exercise 94: Compute the MLE for double-exponential data,

$$p_\theta(x) = \frac{1}{2}\exp(-|x-\theta|).$$

Now compute the asymptotic variance of the median under double-exponential data.

Exercise 95: Compute the Fisher information in N i.i.d. observations. More generally, if x and y are conditionally independent given θ (i.e., $p(x, y|\theta) = p(x|\theta)p(y|\theta)$, what is the information in the joint sample (x, y)? (Hint: write out the score, and take the variance.)

Exercise 96: It might be helpful to step back and look at all this from a more general viewpoint. We are estimating θ by maximizing a function of the form

$$M_N(\theta) = \sum_{i=1}^{N} m(x_i, \theta);$$

here $m(x_i, \theta)$ is just the log-likelihood of one sample point, $\log p_\theta(x_i)$ (and as usual, x_i are i.i.d. from $p_{\theta_0}(x)$). What can you say about the asymptotic distribution of the "M-estimator" (where "M" stands for "maximization")

$$\hat{\theta}_N = \arg\max_{\theta \in \Theta} M_N(\theta),$$

if you know that:

1. $E_{p_{\theta_0}(x)}[m(x, \theta)]$ has a unique maximum at θ_0;
2. $$\frac{\partial^2}{\partial \theta^2} E_{p_{\theta_0}(x)}[m(x, \theta)]\Big|_{\theta=\theta_0} = -A, \ A > 0;$$
3. $V_{p_{\theta_0}(x)}[\frac{\partial m(x,\theta)}{\partial \theta}] = B, \ 0 < B < \infty$.

Now how does this result fit in with what we just proved about the MLE (e.g., what form do A and B take in the MLE case)?

Multiparameter case

A similar asymptotic analysis can be performed when we need to estimate more than one parameter simultaneously. We'll leave the details to you, but the basic result is that if $(\hat{\theta}_{MLE,1}, \hat{\theta}_{MLE,2})$ is the MLE for (θ_1, θ_2), then we can construct the Fisher information matrix

$$I_{ij} = E \frac{\partial}{\partial \theta_i} \frac{\partial}{\partial \theta_j} \log_{\theta_1, \theta_2}(x),$$

and the asymptotic covariance matrix of the MLE is given exactly by $(NI)^{-1}$, where here $(.)^{-1}$ is interpreted as a matrix inverse.

Exercise 97: What is the asymptotic variance of $\hat{\theta}_{MLE,1}$ if θ_2 is known? What if θ_2 is unknown?

Exercise 98: Compute the asymptotic covariance of $\hat{\mu}_{MLE}$ and $\hat{\sigma}^2_{MLE}$ under Gaussian data, in two ways: 1) directly, and 2) using the multiparameter Fisher information.

Cramer-Rao bound, efficiency, asymptotic efficiency

We just established that the asymptotic variance of the MLE looks like $(NI(\theta))^{-1}$. It turns out that this is asymptotically optimal, as the following bound shows:

Theorem 5 (Cramer-Rao lower bound on variance)**.** *Let $\hat{\theta}$ be unbiased. Then*

$$V(\hat{\theta}) \geq (I(\theta))^{-1}.$$

More generally, for any estimator $\hat{\theta}$,

$$V(\hat{\theta}) \geq \frac{\left[dE_\theta(\hat{\theta})/d\theta\right]^2}{I(\theta)}.$$

Proof. For the general case, compute the covariance of $\hat{\theta}$ and the score; then apply Cauchy-Schwartz.

For the special unbiased case, just plug in $E_\theta(\hat{\theta}) = \theta$. □

Exercise 99: Fill in the gaps in the above proof.

Estimators for which the Cramer-Rao bound is achieved exactly are called "efficient." However, such estimators are the exception rather than the rule, as the following exercise demonstrates. Nonetheless, efficiency is still an extremely useful concept, if applied in an asymptotic sense: a sequence of estimators $\hat{\theta}_N$ is "asymptotically efficient" if it asymptotically meets the C-R bound, that is,

$$\lim_{N\to\infty} [NV(\hat{\theta}_N)] = I(\theta)^{-1}.$$

Exercise 100: Look at the derivation of the Cramer-Rao bound more closely. What can you say about the case that the bound is met exactly (i.e., equality holds in the bound)? More precisely: if the bound is met precisely, what does this imply about the parametric family $p_\theta(x)$?

Exercise 101: If $\hat{\theta}_1$ and $\hat{\theta}_2$ are two asymptotically efficient estimators, what can you say about their (rescaled) difference, $\sqrt{N}(\hat{\theta}_1 - \hat{\theta}_2)$? Does this imply anything about the "asymptotic uniqueness" of the MLE as an asymptotically efficient estimator?

Sufficiency and information loss

Exercise 102: Compute the Fisher information for a sufficient statistic and compare it to the Fisher information in the full data. Is it necessarily true that a sufficient statistic preserves all the information in the full sample? How about the converse: if $I_{T(x)}(\theta) = I_x(\theta)$, then is $T(x)$ automatically sufficient? Can we ever have $I_{T(x)}(\theta) > I_x(\theta)$? If yes, give an example; if no, prove it.

One-step estimators

It's often a hassle to exactly solve the likelihood equation

$$\frac{\partial L(\theta)}{\partial \theta} = 0.$$

However, in some cases we can come up with a decent estimator $\hat{\theta}_1$ that at least gets us close: say, $\hat{\theta}_1$ is $\sqrt{N}$-consistent.

Now, we have established that the loglikelihood surface is asymptotically (as $N \to \infty$) well-approximated (on a $N^{-1/2}$ scale) by an upside-down quadratic. So a natural idea is to use the estimator $\hat{\theta}_2$ derived by applying one step of "Newton's algorithm"[28] for finding the local maximum of a function which looks like an upside-down quadratic:

$$\hat{\theta}_2 = \hat{\theta}_1 - \frac{l'(\hat{\theta}_1)}{l''(\hat{\theta}_1)}$$

(where l' and l'' are the first and second derivative of the loglikelihood with respect to θ, respectively).

Now the very interesting result is that this "one-step" estimator — which can be computed analytically whenever $\hat{\theta}_1$ can — is asymptotically efficient, that is, asymptotically just as good as the MLE, even though the MLE might be a lot harder to compute exactly. **Exercise 103**: Prove this; that is, establish the asymptotic efficiency of the one-step estimator. (Hint: the most direct way to do this uses basically the same logic we used to establish the optimality and asymptotic distribution of the MLE. So try mimicking that proof.)

[28] Recall the logic of Newton's algorithm: to maximize the function $f(x)$ given an initial guess x_1, approximate $f(x)$ with a second-order Taylor expansion about x_1 and then (analytically) solve for the maximum of this quadratic approximation to obtain x_2. Draw a picture and do the relevant algebra to remind yourself of what's going on here.

Part IV
Hypothesis testing

Do not put your faith in what statistics say until you have carefully considered what they do not say.

William W. Watt

A caricature of one recipe might read: Apply a significance test to each result, believe the result implicitly if the conventional level of significance is reached, believe the null hypothesis otherwise. Such a complete flight from reality and its uncertainties is fortunately rare, but periodically considering its extremism may help us keep our balance.

F. Mosteller & J. W. Tukey, 1977, p 25.

...no one believes an hypothesis except its originator but everyone believes an experiment except the experimenter.

W. I. B. Beveridge, 1950, p 65.

Simple hypotheses[29]

The simplest version of the hypothesis testing problem is as follows: we have two possible models, $p_0(D)$ and $p_1(D)$, and based on the observed data have to make a choice between them. (This is called, appropriately enough, a "simple" hypothesis test; later we'll consider testing between more than just two hypotheses at a time.) The example to keep in mind: we draw samples x_i i.i.d. from a Gaussian distribution. We know the Gaussian's variance is 1 and we know that the mean is either -1 or 1. I.e., we may take

$$p_0(D) = \prod_{i=1}^{N} \frac{1}{\sqrt{2\pi}} e^{-(x_i+1)^2/2}$$

and

$$p_1(D) = \prod_{i=1}^{N} \frac{1}{\sqrt{2\pi}} e^{-(x_i-1)^2/2}.$$

How do we distinguish between these two hypotheses (models of the world)?

Well, given that we've just spent a month or so talking about likelihood-based methods, one obvious approach would be to use maximum likelihood. That is, choose the hypothesis under which the likelihood of the observed data is largest. In other words, we look at the likelihood ratio

$$\frac{p_0(D)}{p_1(D)} = \exp\left(\frac{1}{2}\sum_{i=1}^{N}(x_i-1)^2 - (x_i+1)^2\right) = \exp\left(-2\sum_{i=1}^{N} x_i\right);$$

if this ratio is larger than 1, then $p_0(D) > p_1(D)$ and we decide that the true mean was -1, or otherwise choose 1. This is a straightforward and intuitive thing to do, and we'll see in just a moment that in many cases this approach is in fact optimal.

But first let's look a little more closely at our decision rule here. If we simplify the above likelihood ratio, we see that our decision really comes down to: if $\sum_i x_i > 0$, choose 1, and otherwise choose -1. (Of course in theory $\sum_i x_i$ could equal zero, in which case we could just flip a coin; but this exact-tie case happens with probability zero here, so we'll ignore it for now.) If we recall, $\sum x_i$ was a minimal sufficient statistic for the Gaussian

[29] HMC 8.1.

family with known variance (write the Gaussian as an exponential family to remind yourself of this fact if you've forgotten).

And of course this can be generalized: tests between two hypotheses based on likelihood ratios only depend on the data through sufficient statistics. **Exercise 104**: Prove this, using the product decomposition for sufficiency.

Exercise 105: Let's say we observe N i.i.d. observations x_i from an exponential distribution whose mean is known to be either 1 or 2. Write down the ML test between these two alternatives, in as simple a form as possible. What role does the minimal sufficient statistic of this exponential family play?

Exercise 106: Let's say we observe N i.i.d. observations x_i from a Gaussian distribution whose mean is known and whose variance is known to be either 1 or 2. Write down the ML test between these two alternatives, in as simple a form as possible. What role does the minimal sufficient statistic of this exponential family play?

Decision-theoretic approach

What if we take a more general decision-theoretic point of view? I.e., we have some prior information and a cost function. Now what is the optimal decision rule? Let's write down our expected loss function:

$$E[C(truth, guess(D))] = \sum_{truth=i\in\{-1,1\}} p(i) \sum_D p_i(D) C(i, guess(D)).$$

Here $C(.,.)$ is some cost function (just a two-by-two table of numbers, in this case) and $p(i)$ is the prior probability of model i. Now as usual we want to choose $guess(D)$ in such a way as to minimize the expected cost. It's clear that all we need to do, for each possible data observation D, is to choose $guess(D)$ such that

$$\sum_{truth=i\in\{-1,1\}} p(i) p_i(D) C(i, guess(D))$$

is as small as possible.

Now let's simplify things a little: assume $C(i,i) = 0$, that is, there's no cost associated with choosing the model correctly. Now the optimal decision rule is as follows:

$$guess(D) = \begin{cases} 1 & \text{if } p(-1)p_{-1}(D)C(-1,1) < p(1)p_1(D)C(1,-1) \\ -1 & \text{otherwise.} \end{cases}$$

Thus we see that if the cost of errors is symmetric $C(-1,1) = C(1,-1)$, and each hypothesis as equally probable *a priori*, $p(-1) = p(1)$, then our decision rule is exactly the ML rule described above. So our intuitive ML approach is actually a special case of the decision-theoretic optimal rule. More generally, the optimum rule says that if our prior belief is that -1 is more likely than 1, then it makes sense to "lean" towards -1 in the sense that we will choose -1 even if the likelihood ratio is slightly weighted towards 1.

Exercise 107: Repeat the last two homework problems (the exponential and Gaussian ML hypothesis tests) in this more general decision-theoretic context. What is the decision-theoretically optimal test if the costs of a mistake are $C(1,2) = a$ and $C(2,1) = b$, as a function of the prior distribution? (Assume again that $C(i,i) = 0$.)

Null and alternate hypotheses

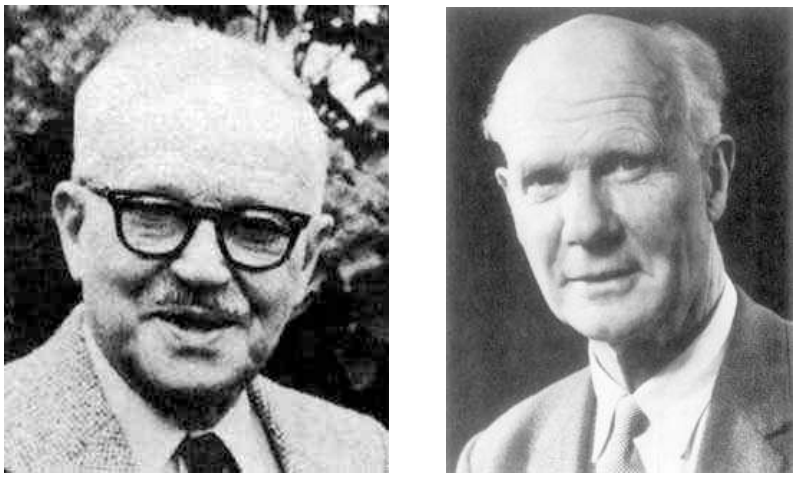

Figure 13: Neyman and Pearson.

In the above discussion we treated both hypotheses equally. In some situations, though, it makes more sense to distinguish between the two hypotheses. For example, if we are testing the fairness of a coin, it might be reasonable to think of the hypothesis that the coin is fair as the "null" hypothesis $p_0(D)$, and the hypothesis that the coin is biased (towards heads, say) as the "alternate" hypothesis, $p_1(D)$. Some specialized terminology has developed in this case:

- the "critical region" A of a test is the region of data space such that if the data D falls in A, we "reject" the null hypothesis; that is, we choose the alternate hypothesis instead.
- the probability $\alpha = \int_{D \in A} p_0(D)$ of incorrectly rejecting the null is called the "size," or "significance level"; this kind of error is called a "type I" error.
- a "type II" error is when we incorrectly accept the null hypothesis.
- the probability $\int_{D \in A} p_1(D)$ of correctly rejecting the null is called the "power."

Now, clearly, we want to make the power of our test as large as possible, while making the size as small as possible. These are contrary goals, of course: making A smaller decreases the size but also decreases the power. So one approach is to hold the size fixed at a given level, say $\alpha = 0.05$, and then try to maximize the power over all possible tests with size less than or equal to 0.05.

It turns out this is not hard to do. Conveniently, this optimal test is of exactly the likelihood ratio form we dealt with above.

Theorem 6 (Neyman-Pearson lemma)**.** *The likelihood ratio test*

$$A = \{D : \frac{p_1(D)}{p_0(D)} \geq T_\alpha\},$$

with the threshold T_α chosen so that the size of the test is equal to α, is the most powerful test of size α.

Proof. Let A_1 be the critical region of any other test of size α. We need to prove that A is at least as powerful as A_1. We have

$$\begin{aligned}
\int_{D\in A} p_1(D) - \int_{D\in A_1} p_1(D) &= \int_{D\in A\cap A_1^c} p_1(D) - \int_{D\in A_1\cap A^c} p_1(D) \\
&\geq T\int_{D\in A\cap A_1^c} p_0(D) - T\int_{D\in A_1\cap A^c} p_0(D) \\
&= T\int_{D\in A} p_0(D) - T\int_{D\in A_1} p_0(D) \\
&= T\alpha - T\alpha = 0.
\end{aligned}$$

□

To return to our Gaussian example above, we have that the most powerful test of size α is to choose 1 whenever $\sum_{i=1}^N x_i > T_\alpha$, where T_α is chosen such that

$$\alpha = \int_{T_\alpha}^\infty \frac{1}{\sqrt{2N\pi}} e^{-(u+1)^2/2N} du.$$

So, to sum up, all three of the approaches we've looked at — ML, decision-theoretic, and Neyman-Pearson — all say basically the same thing: for simple hypothesis testing, the optimal thing to do is to use a test based on the likelihood ratio.

Exercise 108: What is the most powerful test between two exponential distributions with mean 1 and 2, at some fixed size α? What is the power of this test as a function of the number of samples N?

Exercise 109: What is the most powerful test between two uniform distributions, $U([0,1])$ and $U([0,2])$, at some fixed size α? What is the power of this test as a function of the number of samples N?

A side note: it is possible to use many of the same tricks we developed earlier to describe the asymptotic power as N becomes large. We won't go into the details, but if we look at the log-likelihood ratio

$$\sum_{i=1}^{N} \log \frac{p_0(x_i)}{p_1(x_i)},$$

it's clear that we may apply the LLN, CLT, etc. to elucidate the asymptotic behavior here; again we find that the Kullback-Leibler divergence plays a key role in determining the asymptotic behavior of the error (the main difference here is that the hypothesis testing error goes to zero *exponentially* in N, while we saw in the last section that the estimation error, at least in the mean-square setting, goes to zero like $1/N$). We leave the details to the interested reader.

Compound alternate hypotheses[30]

More generally, we're interested in *compound* hypothesis tests: were the data generated by a model $\theta \in H_0$ or $\theta \in H_A$, where H_0 and H_A are two disjoint *sets* of models, the null and alternate sets respectively.

We'll start with the simplest case: the null hypothesis is simple (that is, H_0 consists of just one model, $p_0(D)$), but the alternate is compound. In this case it's reasonable to ask if there is a test with a given size which maximizes the power over every single alternate hypothesis $p_1(D) \in H_A$. (Remember, the size of a test only depends on p_0, so the size will be the same for all p_1 here.) Such a test is called a "uniformly most powerful test," or UMP test for short.

From the Neyman-Pearson theory, we already know that a UMP test has to be based on likelihood ratio tests. This makes it clear that UMP tests often simply don't exist: **Exercise 110**: Prove that no UMP test exists when we are drawing data from a Gaussian of variance 1, if the null hypothesis is $\mu = 0$ and the alternate hypothesis is $\mu \neq 0$. (Hint: break the alternate hypothesis into two sets, $\mu > 0$ and $\mu < 0$, and look at the likelihood ratio tests for each of these individual alternate hypotheses. If these tests are not the same, then argue that no UMP test can exist.)

Is there a simple way to guarantee that a UMP test exists? Well, by the same logic we used above, it's enough to establish that the LR test is independent of $\theta \in H_A$. Here's an example: look at our old friend the exponential family in canonical form. Let's write down the loglikelihood ratio given i.i.d. observations:

$$\begin{aligned} \log \frac{p_0(D)}{p_{\theta \in H_A}(D)} &= \log \frac{\exp[\theta_0 \sum_i k(x_i) + \sum_i s(x_i) + N g(\theta_0)]}{\exp[\theta \sum_i k(x_i) + \sum_i s(x_i) + N g(\theta)]} \\ &= [\theta_0 - \theta] \sum_i k(x_i) + N[g(\theta_0) - g(\theta)]. \end{aligned}$$

It's not too hard to see that a test of the form $T(D) = \sum k(x_i) > c$, for some constant c, leads to a UMP test of the hypothesis that $\theta_0 > \theta$ here. **Exercise 111**: Complete this argument.

How do we choose the test when no UMP test exists? Well, this puts us back in the situation we've encountered before, when e.g. no uniformly

[30] HMC 8.2.

optimal decision rule or UMVUE was available: we have to look at other optimality criteria, e.g. minimax or Bayesian criteria. For example, a Bayesian would choose a hypothesis according to its posterior probability ratio, namely

$$\frac{P(H_0|D)}{P(H_A|D)} = \frac{p(H_0)p_0(D)}{\int_{\theta \in H_A} p(\theta)p_\theta(D)}.$$

More generally, we can always choose some test statistic, $T(D)$, compute its sampling distribution under the null hypothesis $p_0(T)$, and then ask if the observed $T(D)$ is "significantly different" than what we would have expected under the null hypothesis. For example, if we see that $T(D)$ falls outside of the interval defined by the 1st and 99th quantile of $p_0(T)$, then it is reasonable to suspect that D was not, in fact drawn from $p_0(D)$, but rather from some other distribution, and we would reject the null hypothesis here. (This test would not be guaranteed to be optimal in any sense — and importantly, depending on your choice of your test statistic T, your test might have much more power against some alternatives than others — but nonetheless this idea often leads to useful tests in the real world.)

This leads naturally to the concept of a "power curve": namely, for a given test (at some given size), we plot the power $\int_A p_\theta(x)dx$ as a function of $\theta \in H_A$. This function plays a similar role to the risk function played in our decision theory section: basically, we want to make the power as large as possible over the "relevant" part of the parameter space (where "relevant" here depends in some sense on which θ are allowed, or most probable, etc.), just as we wanted to make the risk function as small as possible over the relevant part of the parameter space.

We can also make use of the large sample estimation theory we learned not so long ago: recall that if we have a consistent estimator $\hat{\theta}$ for θ and know that

$$\sqrt{N}\left(\hat{\theta}_N - \theta\right) \rightarrow_D \mathcal{N}(0, V(\theta)),$$

and a consistent estimator of $\sqrt{V(\theta)}$, $\hat{\sigma}$, then as we discussed earlier we may build tests for θ_0 versus $\theta > \theta_0$ based on $\sqrt{N}(\hat{\theta} - \theta)/\hat{\sigma}$ with an asymptotically correct significance level.

Compound null and alternate hypotheses

In this last section we talk about the general case: both null and alternate hypotheses are compound. To find a test of size α here, we would look for some critical region A such that

$$\int_A p_\theta = \alpha \ \forall \ \theta \in H_0.$$

Sometimes this isn't even possible, and we have to relax our standards and look for tests such that

$$\int_A p_\theta \leq \alpha \ \forall \ \theta \in H_0.$$

Sometimes it is possible to find a test whose size is the same for all $\theta \in H_0$, though: for example, if we can find a test statistic T whose distribution is the same for all $\theta \in H_0$, then clearly a test constructed on this statistic will have size independent of $\theta \in H_0$ as well.

Here's an example: Gaussian data of unknown variance; we want to test the null $\mu = 0$ versus the alternate $\mu > 0$. Thus H_0 is the set of all Gaussians with mean 0, and H_A is the set of all Gaussians with mean greater than zero. Before we used the sample mean as our test statistic, but this won't work here because its distribution clearly depends on the variance (and therefore the size of any test based on the sample mean will depend on the underlying unknown true variance). But we know that the sample mean and sample variance are sufficient for this family; why don't we look at the standardized sample mean,

$$T(D) = \frac{\bar{x}}{\bar{\sigma}},$$

with

$$\bar{x} = \frac{1}{N} \sum_{i=1}^{N} x_i$$

and

$$\bar{\sigma}^2 = \frac{1}{N} \sum_{i=1}^{N} (x_i - \bar{x})^2.$$

It's not hard to show that the distribution of $\bar{x}/\bar{\sigma}$ is independent of the variance under the null $\mu = 0$ (**Exercise 112**: prove this); this statistic (or rather, the normalized statistic $\sqrt{N-1}\bar{x}/\bar{\sigma}$) is called a "t-statistic," and its

distribution was originally derived by a statistician working for the Guinness brewery who disguised his identity when publishing his work (to avoid getting in trouble for revealing trade secrets) under the pseudonym "Student." Thus the t-statistic is also called "Student's t," and the commonsense process of dividing by the sample standard deviation is known as "studentizing." **Exercise 113**: Derive the distribution of Student's t.

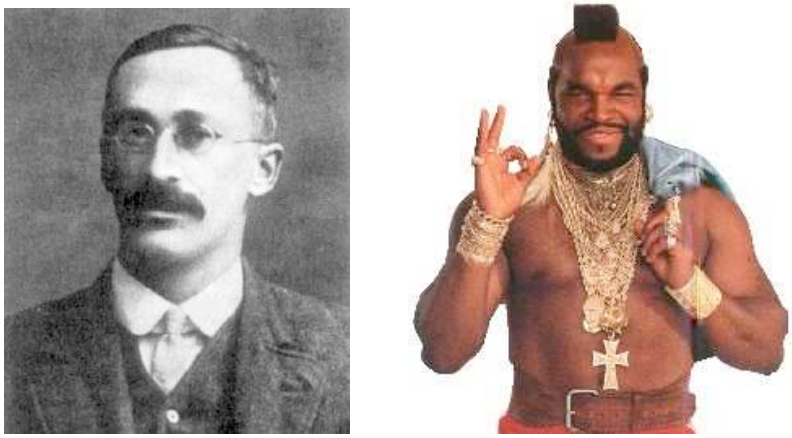

Figure 14: Gosset (aka "Student").

Another example: Gaussian again, but with unknown mean and we'd like to test $\sigma^2 = 1$ versus $\sigma^2 > 1$. Clearly the sample variance is independent of the mean, so we can use this as our test statistic; the null distribution of $N\bar{\sigma}^2$ is a chi-square with $N-1$ degrees of freedom, and thus we can use a test of the form $\bar{\sigma}^2 > c$, for some c chosen such that the size of the test is α. **Exercise 114**: Prove that $N\bar{\sigma}^2$ is a chi-square with $N-1$ degrees of freedom. (Hint: start by proving that the sample variance $\bar{\sigma}^2$ and the sample mean $\bar{x}$ are independent if x_i are i.i.d. Gaussian.)

Outside of the Gaussian family it is a little harder (though not impossible) to find test statistics which are independent of $\theta \in H_0$; nonetheless, again, we may use our large-sample theory to take advantage of this nice property of the Gaussian distribution.

More generally we can always turn back to our Bayesian methods: a Bayesian would choose a hypothesis according to its posterior probability

ratio, namely

$$\frac{P(H_0|D)}{P(H_A|D)} = \frac{\int_{\theta \in H_0} p(\theta) p_\theta(D)}{\int_{\theta \in H_A} p(\theta) p_\theta(D)}$$

in general. This is often simplified into the maximum likelihood ratio test, based on

$$\lambda = \frac{\max_{\theta \in H_0} p_\theta(D)}{\max_{\theta \in H_A} p_\theta(D)} :$$

we reject if λ is sufficiently small. (Again, we often resort to large sample theory to determine exactly how small is "sufficiently small.") This maximal likelihood ratio test is sometimes easier to compute numerically than the integral-based Bayesian test, and the two tests turn out to behave similarly asymptotically (this may be shown, again, using expansions of the loglikelihood similar to those we used in establishing the asymptotic behavior of the MLE).