Stat 3701 Lecture Notes: Statistical Models, Part II

Charles J. Geyer

November 27, 2022

1 License

2 R

- The version of R used to make this document is 4.2.1.
- The version of the `rmarkdown` package used to make this document is 2.17.
- The version of the `alabama` package used to make this document is 2022.4.1.
- The version of the `numDeriv` package used to make this document is 2016.8.1.1.

3 General Statistical Models

The time has come to learn some theory. This is a preview of STAT 5101–5102. We don't need to learn much theory. We will proceed with the general strategy of all introductory statistics: don't derive anything, just tell you stuff.

3.1 Probability Models

3.1.1 Kinds of Probability Theory

There are two kinds of probability theory. There is the kind you will learn in STAT 5101-5102 (or 4101–4102, which is more or less the same except for leaving out the multivariable stuff). The material covered goes back hundreds of years, some of it discovered in the early 1600's. And there is the kind you would learn in MATH 8651–8652 if you could take it, which no undergraduate does (it is a very hard Ph. D. level math course). The material covered goes back to 1933. It is all "new math". These two kinds of probability can be called *classical* and *measure-theoretic*, respectively.

The old kind (classical) is still very useful. For every 1000 people who know a lot of probability theory, 999 know only the classical theory. This includes a lot of working scientists. And it is *a lot* easier than the new kind (measure-theoretic), which is why we still teach it at the undergraduate and master's level and Ph. D. level scientists in all fields. (Only math and stat Ph. D. students take the measure-theoretic probability course.) And Minnesota is not different from any university in this respect.

3.1.2 Classical Probability Models

In classical probability theory there are two kinds of probability models (also called probability *distributions*). They are called *discrete* and *continuous*. The fact that there are two kinds means everything has to be done twice, once for discrete, once for continuous.

3.1.2.1 Discrete Probability Models

A discrete probability model is specified by a finite set S called the *sample space* and a real-valued function f on the sample space called the *probability mass function* (PMF) of the model. A PMF satisfies two properties

$$f(x) \ge 0, \qquad x \in S$$
$$\sum_{x \in S} f(x) = 1$$

We say that $f(x)$ is the *probability* of the *outcome* x. So the two properties say that probabilities are nonnegative and sum to one.

That should sound familiar, just like what they told you probability was in your intro statistics course. The only difference is that, now that you know calculus, S can be an infinite set so the summation here can be an infinite sum.

In principle, the sample space can be any set, but all discrete probability models that are well known, have names, and are used in applications have sample spaces that are subsets of the integers. Here are a few examples.

3.1.2.1.1 The Binomial Distribution

The binomial distribution describes the number of successes in n stochastically independent and identically distributed (IID) random process that can only have two outcomes, conventionally called *success* and *failure*, although they could be anything, the important point is that there are only two possible outcomes.

If p is the probability of success in any single trial then the probability of x successes in n trials is

$$f(x) = \binom{n}{x} p^x (1-p)^{n-x}, \qquad x = 0, 1, 2, \ldots, n.$$

where

$$\binom{n}{x} = \frac{n!}{x!\,(n-x)!}$$

is called a *binomial coefficient* and gives the distribution its name. And this is the PMF of the *binomial distribution*.

The fact that probabilities sum to one is a special case of the *binomial theorem*

$$\sum_{x=0}^{n} \binom{n}{x} a^x b^{n-x} = (a+b)^n.$$

Special cases are

$$(a+b)^2 = a^2 + 2ab + b^2$$
$$(a+b)^3 = a^3 + 3a^2 b + 3ab^2 + b^3$$
$$(a+b)^4 = a^4 + 4a^3 b + 6a^2 b^2 + 4ab^3 + b^4$$

and so for, which you may remember from high school algebra.

The binomial distribution is very important. It arises any time there are data with only two outcomes: yes or no, for or against, vanilla or chocolate, whatever.

There is a generalization called the *multinomial distribution* that allows for any (finite) number of outcomes. But we won't bother with that. (It is the basis of STAT 5421, categorical data analysis.)

3.1.2.1.2 The Poisson Distribution

The *Poisson* distribution (named after a man named *Poisson*, it's not about fish) describes the number of things in any part of a stochastic process where the locations of things are stochastically independent (one are affected by any of the others). Examples would be the number of winners of a lottery, the number of raisins in a slice of carrot cake, the number of red blood cells in a drop of blood, the number of visible stars in a region of the sky, the number of traffic accidents in Minneapolis today. It doesn't matter what is counted, so long as the thingummies counted have nothing to do with each other, you get the Poisson distribution.

Its PMF is

$$f(x) = \frac{\mu^x e^{-\mu}}{x!}, \qquad x = 0, 1, 2, \ldots,$$

where μ can be any positive real number (more on this later).

The fact that probabilities sum to one is a special case of the Maclaurin series (Taylor series around zero) of the exponential function

$$e^\mu = \sum_{x=0}^{\infty} \frac{\mu^x}{x!}$$

Here the sample space is infinite, so the fact that probabilities sum to one involves an infinite series.

The Poisson distribution was initially derived from the binomial distribution. It is what you get when you let p go to zero in the binomial PMF in such a way so that $np \to \mu$. So the Poisson distribution is an approximation for the binomial distribution when n is very large p is very small, and np is moderate sized. This illustrates how one probability distribution can be derived from another.

3.1.2.1.3 The Zero-Truncated Poisson Distribution

We already met the zero-truncated Poisson distribution. This arises when you have a Poisson distribution except for zero counts. There may be other reasons why zero occurs other than Poisson variation; the chef may have forgotten the raisins in the recipe rather than your slice of carrot cake has no raisins for no other reason other than chance variation — that's just the way things came out in the mixing of the batter and slicing of the cake.

The zero-truncated Poisson distribution is widely used in aster models, and we used it as an example of a function that requires extreme care if you want to calculate it accurately using computer arithmetic (supplementary notes).

The exact definition is that the zero-truncated Poisson distribution is what you get when you take Poisson data and throw out all the zero counts. So its PMF is the PMF of the Poisson distribution with zero removed from the sample space and all of the probabilities re-adjusted to sum to one.

For the Poisson distribution $f(0) = e^{-\mu}$ so the probability of nonzero is $1 - e^{-\mu}$ so the zero-truncated Poisson distribution has PMF

$$f(x) = \frac{\mu^x e^{-\mu}}{x! \, (1 - e^{-\mu})}, \qquad x = 1, 2, 3, \ldots.$$

This is another illustration of how one probability distribution can be derived from another.

3.1.2.2 Univariate Continuous Probability Models

A univariate continuous probability model is specified by a a real-valued function f of one real variable called the *probability density function* (PDF) of the model. A PDF satisfies two properties

$$f(x) \geq 0, \qquad -\infty < x < \infty$$

$$\int_{-\infty}^{\infty} f(x) \, dx = 1$$

We say that $f(x)\,dx$ is the *probability* of an outcome in the interval from x to $x + dx$ when dx is very small. For this to be exactly correct dx has to be infinitesimal. To get the probability for a finite interval, one has to integrate

$$\Pr(a < X < b) = \int_a^b f(x)\,dx.$$

That should sound familiar, just like what they told you probability was in your intro statistics course. Integrals are area under a curve. Probability is area under a curve (for continuous distributions).

3.1.2.2.1 The Normal Distribution

The normal distribution arises whenever one averages a large number of IID random variables (with one proviso, which we will discuss later). This is called the *central limit theorem* (CLT).

Its PDF is

$$f(x) = \frac{1}{\sqrt{2\pi}\,\sigma} e^{-(x-\mu)^2/(2\sigma^2)}, \qquad -\infty < x < \infty.$$

The fact that this integrates to one is something they didn't teach you in calculus of one variable (because the trick of doing it involves multivariable calculus, in particular, polar coordinates).

The special case when $\mu = 0$ and $\sigma = 1$ is called the *standard* normal distribution. Its PDF is

$$f(z) = \frac{1}{\sqrt{2\pi}} e^{-z^2/2}, \qquad -\infty < z < \infty.$$

3.1.2.2.2 The Cauchy Distribution

The Cauchy distribution arises in no applications I know of. It is a mathematical curiosity mostly useful as a counterexample. Many things that are true of other distributions don't hold for Cauchy. For example, the average of IID Cauchy random variables does not obey the CLT (more on this later). If X and Y are independent standard normal random variables, then X/Y has a Cauchy distribution, but this is an operation that does not seem to arise in applications.

Its PDF is

$$f(x) = \frac{1}{\pi\sigma \left[1 + \left(\frac{x-\mu}{\sigma}\right)^2\right]}, \qquad -\infty < x < \infty.$$

The special case when $\mu = 0$ and $\sigma = 1$ is called the *standard* Cauchy distribution. Its PDF is

$$f(z) = \frac{1}{\pi(1 + z^2)}, \qquad -\infty < z < \infty.$$

The fact that these integrate to one involves, firstly, change-of-variable, the substitution $z = (x - \mu)/\sigma$ establishing that

$$\frac{1}{\pi\sigma} \int_{-\infty}^{\infty} \frac{1}{1 + \left(\frac{x-\mu}{\sigma}\right)^2}\,dx = \frac{1}{\pi} \int_{-\infty}^{\infty} \frac{1}{1 + z^2}\,dz,$$

and, secondly,

$$\int \frac{1}{1 + z^2}\,dz = \operatorname{atan}(z) + \text{a constant},$$

where atan denotes the arctangent function, and, lastly,

$$\lim_{z \to \infty} \operatorname{atan}(z) = \frac{\pi}{2}$$
$$\lim_{z \to -\infty} \operatorname{atan}(z) = -\frac{\pi}{2}$$

3.1.2.3 Multivariate Continuous Probability Models

A distribution for two or more continuous random variables is the same except this is multivariable calculus. For example, a probability distribution for three variables x, y, and z has a PDF that is nonnegative and integrates to one, but now this involves a triple integral

$$f(x, y, z) \ge 0, \qquad -\infty < x < \infty, -\infty < y < \infty, -\infty < z < \infty$$
$$\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f(x, y, z)\, dx\, dy\, dz = 1$$

As in the case of univariate models, the PDF does not give probability but rather (as the name says) probability *density:* $f(x, y, z)\, dx\, dy\, dz$ is the probability of the box $(x, x + dz) \times (y, y + dy) \times (z, z + dz)$ if dx, dy, and dz are infinitesimal, but for finite regions, you have to do a triple integral

$$\Pr\{(X, Y, Z) \in A\} = \iiint_A f(x, y, z)\, dx\, dy\, dz$$

and that is something only learned in multivariable calculus.

3.1.2.4 Stochastic Independence

We say random variables are *stochastically independent* or *statistically independent* or *independent* (with no qualifying adjective) if

- the values of any of them have nothing to do with the values of the others (this is the concept we use for applications), or
- the PMF or PDF factors into a product of functions of one variable

  $$f(x_1, \ldots, x_n) = \prod_{i=1}^{n} f_i(x_i)$$

  (this is the theoretical concept). This equation is so important that it has its own terminology: the phrase the joint distribution is the product of the marginal distributions, or even shorter, the joint is the product of the marginals, means the *joint distribution* of all the variables (the left-hand side of this equation) is equal to the product (on the right-hand side of this equation) of the *marginal distributions*, meaning $f_i(x_i)$ is the PDF or PMF, as the case may be, of X_i.

So we have two concepts of independence, one applied (that we use to tell use what applications can use this concept) and one theoretical (that we use to tell us how this concept affects the mathematics).

In statistics, we should never use *independent* with any other meaning to avoid confusion with any other notion of independence. In particular, in regression models we *never* say dependent and independent variables, but always say predictor and response variable instead.

The theoretical concept implies the applied concept because the PDF or PMF factoring implies that probability calculations will also factor: in the continuous case

$$\Pr(a_i < X_i < b_i,\ i = 1, \ldots, n) = \prod_{i=1}^{n} \int_{a_i}^{b_i} f_i(x_i)\, dx_i$$

and in the discrete case

$$\Pr(a_i < X_i < b_i,\ i = 1, \ldots, n) = \prod_{i=1}^{n} \sum_{\substack{x_i \in S_i \\ a_i < x_i < b_i}} f_i(x_i)$$

(and we see that in the last case the sample space also has to factor as a Cartesian product $S_1 \times S_2 \times \cdots \times S_n$ so that the values of each variable that are possible have nothing to do with the values of the other — this was automatic in the continuous case because we took the sample space to be the Cartesian product R^n, the n-fold product of the real line with itself, the set of all n-tuples of real numbers, just like $S_1 \times S_2 \times \cdots \times S_n$ denotes the set of all n-tuples $(x_1, x_2, \ldots, x_n)$ with $x_i \in S_i$ for each i). And independence gets us back (almost) to univariate calculus. We have integrals or sums involving only one variable.

3.1.2.5 Independent and Identically Distributed (IID)

The phrase in the section title, so important that it gets its own TLA (three-letter acronym) is just the special case of independence where all the random variables have the same distribution, so the theoretical concept is

$$f(x_1, \ldots, x_n) = \prod_{i=1}^{n} f(x_i)$$

(all the marginals on the right-hand side are the same distribution (here we have f where the analogous equation above had f_i).

3.1.2.6 Introductory Statistics versus Theoretical Statistics

In intro stats the only statistical model discussed is *finite population sampling*, there are N individuals, which are taken to be fixed not random, a specified population. For example, the population could be the students registered at the University of Minnesota *today* at 8:00 a. m. (which students the university has changes over time). Then we take a *simple random sample* (SRS) of this population, which is a special case of IID. The random variables X_i are measurements on each individual (quantitative or qualitative) selected for the sample. And SRS means the same as IID: whether one individual is selected for the sample has nothing to do with which other individuals are selected. This means that X_i and X_j can be the *same* individual: which individual X_i is a measurement on has nothing to do with which individual X_j is a measurement on, and this means, in particular, that we cannot require that these individuals cannot be the same individual (that would make X_i have something to do with X_j). For those who have heard that terminology, we are talking about so-called *sampling with replacement* as opposed to *sampling without replacement*.

To be theoretically astute, you have to move out of finite population sampling and replace SRS with IID. In SRS we are sampling from a finite set (the population) so every variable is discrete whether we think of it that way or not. In IID we can have continuous random variables. But then the SRS story breaks down. Sampling from in infinite population doesn't make sense.

Strangely statistics teachers and applied statisticians often use the terminology of SRS (the sample and the population) even when they are talking about IID (where those terms don't make any sense — so they are making an imprecise analogy with finite population sampling).

3.1.2.7 Random Variables and Expectation

Applicationally, a random variable is any measurement on a random process. Theoretically, a random variable a function on the sample space. Either of these definitions make any function of a random variable or variables another random variable.

If X is the original variable (taking values in the sample space) and $Y = g(X)$, then the *expectation* or *expected value* or *mean* or *mean value* (all these terms mean the same thing) is

$$E(Y) = E\{g(X)\} = \int_{-\infty}^{\infty} g(x) f(x)\, dx$$

in the continuous case and

$$E(Y) = E\{g(X)\} = \sum_{x \in S} g(x) f(x)$$

and analogous formulas for multivariable cases, which we will try to avoid.

So to calculate the expectation (a. k. a. mean) of a random variable, you multiply the values of the random variable (here $g(x)$) by the corresponding probabilities or probability density (here $f(x)$) and sum or integrate, as the case may be.

3.1.2.8 Mean, Variance, and Standard Deviation

We have already said that the expectation of the variable itself is called the *mean*

$$\mu = E(X)$$

or sometimes when there is more that one variable under discussion we decorate the μ

$$\mu_Y = E(Y).$$

The expected squared deviation from the mean is another important quantity called the *variance*

$$\sigma^2 = \operatorname{var}(X) = E\{(X - \mu)^2\}$$

or with decoration

$$\sigma_Y^2 = \operatorname{var}(Y) = E\{(Y - \mu_Y)^2\}$$

The *standard deviation* is the square root of the variance, and, conversely, the variance is the square of the standard deviation, always

$$\begin{aligned} \operatorname{sd}(Y) &= \sqrt{\operatorname{var}(Y)} \\ \operatorname{var}(Y) &= \operatorname{sd}(Y)^2 \end{aligned}$$

Why to such closely related concepts? In applications the standard deviation is more useful because it has the same units as the variable. If Y is measured in feet then μ_Y also has units feet, but $(Y - \mu_Y)^2$ and $\operatorname{var}(Y)$ have units square feet (ft^2) so $\operatorname{sd}(Y)$ is back to units feet (ft). But theoretically, the square root is a nuisance that just makes many formulas a lot messier than they need to be.

Here's an example. The expectation of a sum is the sum of the expectations, always,

$$E\left(\sum_{i=1}^n X_i\right) = \sum_{i=1}^n E(X_i)$$

and the variance of a sum is the sum of the variances, not always, but when the variables are *independent*,

$$\operatorname{var}\left(\sum_{i=1}^n X_i\right) = \sum_{i=1}^n \operatorname{var}(X_i)$$

The latter is a lot messier when one tries to state it in terms of standard deviations

$$\operatorname{sd}\left(\sum_{i=1}^n X_i\right) = \sqrt{\sum_{i=1}^n \operatorname{sd}(X_i)^2}$$

and standard deviation doesn't look so simple any more.

3.1.2.9 Operations and Expectations

It is FALSE that a general operation can be taken outside of an expectation. For an arbitrary function g

$$E\{g(X)\} \neq g\big(E(X)\big)$$

However this is true for some special situations. One can always take addition or subtraction out

$$\begin{aligned} E(X + Y) &= E(X) + E(Y) \\ E(X - Y) &= E(X) - E(Y) \end{aligned}$$

One can always take constants out

$$E(aX) = aE(X)$$

where a is constant (non-random).

One can always take linear functions out

$$\begin{aligned} E(a + bX) &= a + bE(X) \\ \operatorname{var}(a + bX) &= b^2 \operatorname{var}(X) \end{aligned}$$

(the second of these doesn't fit the pattern we are talking about here, but is very important and used a lot).

In the special case of *independent* random variables, one can take out multiplication and division

$$\begin{aligned} E(XY) &= E(X)E(Y) \\ E\left(\frac{X}{Y}\right) &= \frac{E(X)}{E(Y)} \end{aligned}$$

but only if X and Y are *independent* random variables.

3.1.2.10 Mean and Variance of the Sample Mean

We can use these operations to prove the formulas for the mean and variance of the sample mean in the IID case

$$\begin{aligned} \overline{X}_n &= \frac{1}{n} \sum_{i=1}^{n} X_i \\ E(\overline{X}_n) &= \frac{1}{n} \sum_{i=1}^{n} E(X_i) \\ &= \frac{n\mu}{n} \\ &= \mu \\ \operatorname{var}(\overline{X}_n) &= \frac{1}{n^2} \sum_{i=1}^{n} \operatorname{var}(X_i) \\ &= \frac{n\sigma^2}{n^2} \\ &= \frac{\sigma^2}{n} \end{aligned}$$

3.1.2.11 Expectations Need Not Exist

The Cauchy distribution does not have a mean or variance.

This is related to a concept that we should perhaps already fussed about in the preceding two sections. In probability theory we do not allow improper integrals or infinite sums (where the value depends on how you take the limits). We only allow absolute integrability or summation. That is, we say $E(X)$ exists only when $E(|X|)$ also exists. (It is a theorem of advanced calculus that then the integral or sum does not depend on how you take the limits.)

So for the standard Cauchy distribution

$$E(|X|) = \frac{1}{\pi} \int_{-\infty}^{\infty} \frac{|x| \, dx}{1 + x^2} = \frac{2}{\pi} \int_{0}^{\infty} \frac{x \, dx}{1 + x^2}$$

and we claim these integrals do not exist (the area under the curve is infinite). Even part of the area under the curve is infinite. For $a > 0$

$$\frac{x}{1 + x^2} = \frac{1}{\frac{1}{x} + x} \geq \frac{1}{\frac{1}{a} + x}$$

and

$$\int \frac{dx}{\frac{1}{a} + x} = \log\left(\frac{1}{a} + x\right) + \text{a constant}$$

and the right-hand side goes to infinity as $x \to \infty$ so

$$\int_0^\infty \frac{x\,dx}{1 + x^2} \geq \int_a^\infty \frac{x\,dx}{1 + x^2} = \infty$$

Then the variance doesn't exist because

$$\text{var}(X) = \frac{1}{\pi} \int_{-\infty}^\infty \frac{x^2\,dx}{1 + x^2} = \frac{2}{\pi} \int_0^\infty \frac{x^2\,dx}{1 + x^2} \geq \frac{2}{\pi} \int_0^\infty \frac{x\,dx}{1 + x^2} = E(X) = \infty$$

3.1.2.12 The Law of Large Numbers (LLN)

For any sequence of IID random variables X_1, X_2, ... that have a mean μ, the sample mean for sample size n

$$\overline{X}_n = \frac{1}{n} \sum_{i=1}^n X_i$$

converges to μ as n goes to infinity (in senses of convergence that we won't even bother to explain, suffice it to say that with very high probability $\overline{X}_n$ will be very close to μ as when n is very large).

This is a theorem of mathematical probability theory that is not only beyond this course but also beyond STAT 5101–5102 (although they do prove the LLN under the additional assumption that the X_i also have finite variance, which is a much easier proof).

If the mean does not exist, for example when the X_i are IID Cauchy, then the LLN does not hold.

3.1.2.13 The Central Limit Theorem (CLT)

For any sequence of IID random variables X_1, X_2, ... that have a mean μ and a nonzero (and finite) variance σ^2, with the sample mean $\overline{X}_n$ as defined in the preceding section, the probability distribution of

$$Z_n = \frac{\overline{X}_n - \mu}{\sigma / \sqrt{n}}$$

converges to the standard normal distribution as n goes to infinity in the sense that $\Pr(a < Z_n < b)$ converges to $\Pr(a < Z < b)$, where Z is a standard normal random variable for any real numbers $a < b$.

This is a theorem of mathematical probability theory that is not only beyond this course but also beyond STAT 5101–5102.

Both the LLN and the CLT are proved in MATH 8651–8652.

If the variance does not exist, for example when the X_i are IID Cauchy, then the CLT does not hold.

3.1.3 Measure-Theoretic Probability Models

Measure theory allows all of the above to be combined and streamlined. The discrete and continuous cases become special cases of one kind of model and the sums and integrals become special cases of one kind of integral (a vast generalization of the integral taught in calculus). And there are probability models that are neither discrete nor continuous.

But in practical applications, measure theory does nothing. All the expectations one could calculate in a practical example are done by calculus. The measure theory only supplies mathematical elegance and generality and simplification of proofs (which, even simplified, are still very difficult).

So we aren't missing anything important by not understanding this.

3.2 Statistical Models

A *statistical model* is a family of probability models. They come in two kinds, parametric and nonparametric.

3.2.1 Probability Theory versus Theoretical Statistics

In probability theory we give you a probability model and your job is to calculate something about it, for example, the expectation of some random variable in the model.

In theoretical statistics we give you a statistical model (a family of probability models) and some data that purportedly obey one probability model in this family, called the *truth* or the *true unknown model* or the *true unknown distribution* (all mean the same thing) and your job is to say something about which single probability model it was that the data obey. Obviously you cannot say anything with certainty. Any data at all are compatible with any continuous distribution, but they are not all equally probable. So turning this around, we get inexact *statistical inference*, which may take any of the following familiar forms:

- frequentist inference:
  - point estimators, which are statistics that don't even to claim to exactly estimate whatever they are trying to estimate they only claim to get close,
  - confidence intervals and confidence regions, which are functions of statistics that do claim to contain whatever they are trying to estimate but not certainly but only with a specified probability,
  - hypothesis tests, which are accept-reject decisions about whether the true unknown probability model is in some subset of the statistical model or another, and they don't claim to do this correctly with certainty but only with a specified error rate (the same goes for hypothesis tests reported in terms of P-values rather than decisions because the two views of hypothesis tests are interconvertible: the P-value is the smallest α level for which the hypothesis rejects the null hypothesis),
- and Bayesian inference, more on this in another handout.

We assume you already know enough about point estimators, confidence intervals, and hypothesis tests from STAT 3011 and 3032. Confidence regions were illustrated in Section 8.2 of the course notes on optimization and equation solving. Bayesian inference will be the subject of other course notes.

3.2.2 Parametric Statistical Models

In a *parametric statistical model* the PMF or PDF of the distributions in the family are given by a formula, like those in Section 3.1.2 above that has adjustable constants called the *parameters*. Each different set of parameter values gives a different probability model.

Hence to be pedantically correct we should say that

- Section 3.1.2.1.1 above does not describe (as the section title says) *the* binomial distribution but rather the binomial *family* of distributions (a parametric statistical model, the parameter being p),
- Section 3.1.2.1.2 above does not describe (as the section title says) *the* Poisson distribution but rather the Poisson *family* of distributions (a parametric statistical model, the parameter being μ),
- Section 3.1.2.1.3 above does not describe (as the section title says) *the* zero-truncated Poisson distribution but rather the zero-truncated Poisson *family* of distributions (a parametric statistical model, the parameter being μ),
- Section 3.1.2.2.1 above does not describe (as the section title says) *the* normal distribution but rather the normal *family* of distributions (a parametric statistical model, the parameters being μ and σ^2 or μ and σ, whichever you prefer), and
- Section 3.1.2.2.2 above does not describe (as the section title says) *the* Cauchy distribution but rather the Cauchy *family* of distributions (a parametric statistical model, the parameters being μ and σ).

As you know from STAT 3011 and 3032, parameters are always denoted by Greek letters, like those above, except for the exceptions, again like the exception above, which is the parameter p of the binomial distribution. Although there are a few textbooks that use π for this parameter, so many people consider this a frozen letter than cannot mean anything but the area of the unit circle, $3.141592653589793238462\ldots$ and so cannot be the parameter of the binomial distribution.

As you also know from STAT 3011 and 3032, estimators of parameters are often denoted by putting a hat on the Greek letter: the statistic $\hat{\mu}$ estimates the parameter μ, the statistic $\hat{\theta}$ estimates the parameter θ, or on the non-Greek letter if that's what we have: the statistic $\hat{p}$ estimates the parameter p. Of course, we don't have to use this convention if we don't want to: the statistic $\bar{x}$ estimates the parameter μ.

3.2.3 Nonparametric Statistical Models

A *nonparametric statistical model* the PMF or PDF of the distributions in the family cannot be given by one formula with a finite set of parameters. Here are some nonparametric statistical models:

a. all probability distributions on R (this one requires measure-theoretic probability to describe the distributions that are neither discrete nor continuous),

b. all continuous probability distribution on R (those having PDF),

c. all continuous probability distributions on R that are symmetric about zero (those whose PDF f satisfies $f(x) = f(-x)$),

d. all continuous, symmetric about zero, and unimodal probability distributions on R (those whose PDF f satisfies $f(x) = f(-x)$ and f is an increasing function on $(-\infty, 0]$ and a decreasing function on $[0, \infty)$).

e. all continuous probability distributions on R that are symmetric about some point μ (those whose PDF f satisfies $f(\mu + h) = f(\mu - h)$),

These go with various statistical procedures.

- For statistical model (b) the usual estimator of location is the sample median, which estimates the population median, and the usual hypothesis test about location is the so-called *sign test* which comes with an associated confidence interval.
- For statistical model (e) the usual estimator of location is the median of Walsh averages, which was discussed in Section 6.5.4 of the course notes on matrices, arrays, and data frames, also called the *Hodges-Lehmann estimator*, which estimates the center of symmetry μ, which is also the population median and the population mean, if the mean exists, and the usual hypothesis test about location is the so-called *Wilcoxon signed rank test* which comes with an associated confidence interval.
- Statistical model (c) is the error model assumed for the nonparametric regression procedures `lqs` and `rlm`, which were discussed in Section 3.4.1 of Part I of the course notes about statistical models (the notes you are reading now being Part II).

We are not going to do any more about nonparametrics because there is a whole course STAT 5601 about this stuff and because there may have been a little bit about it in STAT 3011 and will probably be something about it in STAT 5102.

3.2.4 General Asymptotics

Here "general" means more general than the LLN and the CLT and, as usual, "asymptotics" refers to approximate, large sample size, n goes to ∞ results.

There are a large number of tools taught in theoretical probability and statistics (including STAT 5101–5102) for figuring out the asymptotic distribution of estimators other than the sample mean.

3.2.4.1 The Sample Median

The *median* of the distribution of a random variable X is any point μ such that

$$\Pr(X \le \mu) \ge \tfrac{1}{2} \quad \text{and} \quad \Pr(X \ge \mu) \ge \tfrac{1}{2}$$

(the reason these probabilities do not necessarily add to one is because we are double counting the point $x = \mu$). Such a point always exists but need not be unique. It will be unique if the distribution is continuous and everywhere positive or even positive in a neighborhood of the median. When we are using the bad analogy with finite population sampling (Section 3.1.2.6 above) we call μ the population median.

The *sample median* of a sample X_1, X_2, ..., X_n is the middle value in sorted order if n is odd and the average of the two middle values in sorted order if n is even, what the R function `median` calculates.

For statistical model (b) under the additional assumption that the PDF f is nonzero and continuous at the population median μ (saying X is continuous does not mean its PDF f is continuous, rather it just means it has a PDF, not necessarily continuous; here we are assuming that f is continuous at μ, that is $x_n \to \mu$ implies $f(x_n) \to f(\mu)$ and also assuming $f(\mu) > 0$), the asymptotic distribution of the sample median is known: if $\widetilde{X}_n$ denotes the sample median, then

$$Z_n = (\widetilde{X}_n - \mu) \cdot 2 f(\mu) \sqrt{n}$$

converges to the standard normal distribution as n goes to infinity in the same sense as in the CLT discussed in Section 3.1.2.13 above.

3.2.4.2 Sample Mean versus Sample Median

We know the standard deviation of the sample mean is $\sigma/\sqrt{n}$, where σ is the "population" standard deviation (with "population" in scare quotes because this is a general result valid for any distribution for which the variance exists; it has nothing to do with finite population sampling).

Now we have learned that the asymptotic standard deviation of the sample median is $1/(2f(\mu)\sqrt{n})$ where f is the PDF and μ is the "population" median (again this is a general result having nothing to do with finite population sampling) provided $f(\mu) > 0$ and f is continuous at μ.

How these compare depends on the distribution. For a *symmetric* distribution the mean and the median are the same and equal to the center of symmetry (if the mean exists). So the sample mean and sample median both estimate the same quantity. We could use either. It depends on the distribution.

3.2.4.2.1 Normal Distribution

If we plug $x = \mu$ into the formula for the PDF of the normal distribution Section 3.1.2.2.1 above, we get

$$f(\mu) = \frac{1}{\sqrt{2\pi}\,\sigma}$$

so the asymptotic standard deviation of the sample median is

$$\frac{1}{2 \cdot \frac{1}{\sqrt{2\pi}\,\sigma} \cdot \sqrt{n}} = \sqrt{\frac{\pi}{2}} \cdot \frac{\sigma}{\sqrt{n}}$$

and this is $\sqrt{\pi/2} \approx 1.2533141$ times the standard deviation of the sample mean.

3.2.4.2.2 Cauchy Distribution

If we plug $x = \mu$ into the formula for the PDF of the Cauchy distribution Section 3.1.2.2.2 above, we get

$$f(\mu) = \frac{1}{\pi\sigma}$$

so the asymptotic standard deviation of the sample median is

$$\frac{1}{2 \cdot \frac{1}{\pi\sigma} \cdot \sqrt{n}} = \frac{\pi}{2} \cdot \frac{\sigma}{\sqrt{n}}$$

Here σ is not the "population" *standard deviation* but just one of the two parameters (μ and σ) of the Cauchy distribution. Like μ and σ for the normal distribution, μ measures what theoretical statistics calls *location* and intro stats calls *center* and σ measures what theoretical statistics calls *scale* and intro stats calls *spread* (I have no idea why intro stats uses different terminology than the rest of statistics but if you do `?Cauchy` in R you will find that these parameters are called `location` and `scale`).

For the Cauchy distribution, μ is the center of symmetry and also the median, but it is not the mean because the mean does not exist for the Cauchy distribution (Section 3.1.2.11 above).

For the Cauchy distribution, σ happens to be twice the interquartile range (IQR), a robust measure of scale appropriate for symmetric distributions. It is also the median absolute deviation from the median (MAD), another robust measure of scale appropriate for symmetric distributions. For symmetric distributions IQR $= 2 \cdot$ MAD.

```
mu <- rnorm(10)
sigma <- runif(10)
unique(pcauchy(mu - sigma, location = mu, scale = sigma))
```

```
## [1] 0.25 0.25
```

```
unique(pcauchy(mu, location = mu, scale = sigma))
```

```
## [1] 0.5
```

```
unique(pcauchy(mu + sigma, location = mu, scale = sigma))
```

```
## [1] 0.75
```

R agrees that the lower quartile, median, and upper quartile are $\mu - \sigma$, μ, and $\mu + \sigma$. But σ is not the standard deviation because the variance (and hence standard deviation) does not exist for the Cauchy distribution (Section 3.1.2.11 above).

Because the "population" standard deviation does not exist, the standard deviation of the sample mean does not exist. In fact, it can be shown, and is shown in STAT 5101–5102, that the distribution of the sample mean for the Cauchy distribution is exactly the same Cauchy distribution that is the "population" distribution, that is, if $X_1, \ldots, X_n$ are IID Cauchy, the $\overline{X}_n$ has the *same* distribution as the X_i. The sample mean is (for Cauchy) no better an estimator of location than any one of data points it is the average of! So the sample mean (for Cauchy) does not obey the LLN, does not obey the CLT, and does not obey what some "statistics for poets" courses (like STAT 1001) call the *square root law* (statistical precision varies as the square root of sample size). We saw above that the sample mean and median for normal data obey the square root law. We saw above that the sample median for Cauchy data obeys the square root law. But the sample mean for Cauchy does not. It does not get closer and closer to μ as n increases; its distribution does not change at all as n increases.

3.2.4.2.3 Other Distributions

Some estimators are better than others. Which estimator is better depends on the "population" distribution. We could do a lot more examples, but that is not appropriate for this course (they are lecture examples and homework and test questions for STAT 5102).

3.2.4.3 Asymptotic Relative Efficiency

The ratio of standard deviations is *not* the appropriate measure of goodness of estimators. The ratio of *variances* or *asymptotic variances* is. Why?

Suppose you have two consistent and asymptotically normal (CAN) estimators of the same parameter, and their asymptotic standard deviations are $\tau_1/\sqrt{n}$ and $\tau_2/\sqrt{n}$ (these τ_i having nothing necessarily to do with the "population" standard deviation, consider the asymptotics for the sample median), so both obey the "square root law", and the corresponding asymptotic (large sample, approximate) confidence intervals are

$$\text{estimator} \pm 1.96 \cdot \frac{\tau_i}{\sqrt{n}}$$

Now suppose we want two equally accurate confidence intervals, which we can always achieve by using a larger sample size for the two estimators. That is, we want

$$\frac{\tau_1}{\sqrt{n_1}} = \frac{\tau_2}{\sqrt{n_2}}$$

or

$$\frac{\tau_1^2}{n_1} = \frac{\tau_2^2}{n_2}$$

or

$$\frac{\tau_1^2}{\tau_2^2} = \frac{n_1}{n_2}$$

If we make the reasonable assumption that the cost of obtaining data is proportional to sample size, then we see that the ratio of asymptotic variances (the left-hand side in the last equation above) is the appropriate criterion. It is called *asymptotic relative efficiency* (ARE).

Since it is not always clear which estimator you are calling estimator 1 and which estimator 2, it is always a good idea to say which is best in addition to giving the ARE.

- for the normal distribution the sample mean is a better estimator of location than the sample median and the ARE is $\pi/2 \approx 1.5707963$.
- for the Cauchy distribution the sample median is a better estimator of location than the sample mean and the ARE is ∞.

3.2.4.4 Maximum Likelihood Estimation

Is there always a best estimator? Not always-always, regardless of how weird the "population" distribution is. But almost always, including every practical example I know.

This is called the *maximum likelihood estimator* (MLE) or the *maximum likelihood estimate* (MLE), depending on whether you are taking about the procedure (estimator) or its result (estimate). Actually, I cannot be bothered to make this distinction most of the time. MLE for either. MLE can also stand for *maximum likelihood estimation* (another name for the procedure).

3.2.4.4.1 Likelihood

The *likelihood* is just the PMF or PDF *considered as a function of the parameter or parameters (data being held fixed)* rather than the other way around. We can symbolize this by the equation

$$L_x(\theta) = f_\theta(x)$$

The likelihood is a function of the parameter θ. The PMF/PDF is a function of the data x. The PMF/PDF gives probabilities or probability density; $f_\theta(x)$ is the probability of x (for discrete) or the probability density at x (for continuous). But likelihood does not give probabilities; $L_x(\theta)$ is not the probability of θ or probability density at θ. In fact, θ is not a random variable (if you are a frequentist, more on this in another notes).

It turns out that for most of the theory it does not matter whether the data (here x) is scalar or vector, nor does it matter whether the parameter (here θ) is scalar or vector. So here x can also stand for a vector of data, and θ can also stand for a vector of parameters.

It also turns out that it makes no difference to any use of likelihood (anywhere in statistics, not just for MLE) that it makes no difference if we drop multiplicative terms from the likelihood that do not contain the parameter or parameters.

For example, we can write

- $L_x(p) = p^x(1-p)^{n-x}$ for the likelihood for the binomial distribution, dropping the term $\binom{n}{x}$ from the PMF, or
- $L_x(\mu, \sigma) = (1/\sigma) \exp(-(x-\mu)^2/(2\sigma^2))$ for the likelihood of the normal distribution, dropping the term $1/\sqrt{2\pi}$ from the PDF.

3.2.4.4.2 Log Likelihood

In MLE, we usually use log likelihood rather than likelihood.

$$l_x(\theta) = \log L_x(\theta)$$

and now we are even farther away from probability. What is log probability? Nothing!

The rule that *multiplicative* terms that do not contain the parameter or parameters can be dropped from the *likelihood* implies The rule that *additive* terms that do not contain the parameter or parameters can be dropped from the *log likelihood* (because log of a product is sum of the logs).

3.2.4.4.3 Likelihood, Log Likelihood, and IID

For IID data, the joint is the product of the marginals

$$f_\theta(x_1, \ldots, x_n) = \prod_{i=1}^n f_\theta(x_i)$$

(Section 3.1.2.5 above). So the likelihood is also a product

$$L_{x_1,\ldots,x_n}(\theta) = \prod_{i=1}^n f_\theta(x_i)$$

except that possibly we can drop multiplicative terms that do not contain the parameter or parameters, and the likelihood is a sum

$$l_{x_1,\ldots,x_n}(\theta) = \sum_{i=1}^n \log f_\theta(x_i)$$

except that possibly we can drop additive terms that do not contain the parameter or parameters (because the log of a product is the sum of the logs).

Since the notation on the left-hand side of these two equations is pretty horrible, we usually leave out the data and just keep the sample size

$$L_n(\theta) = \prod_{i=1}^n f_\theta(x_i)$$

$$l_n(\theta) = \sum_{i=1}^n \log f_\theta(x_i)$$

(possibly with multiplicative or additive terms, as appropriate, dropped). The likelihood or log likelihood is still a function of random data, hence a random function, whether or not this is indicated explicitly.

For example, for the normal distribution, we write

$$\begin{aligned} L_n(\mu, \sigma) &= \prod_{i=1}^{n} \left(\frac{1}{\sigma} e^{-\frac{(x_i - \mu)^2}{2\sigma^2}} \right) \\ l_n(\mu, \sigma) &= \sum_{i=1}^{n} \left(-\log(\sigma) - \frac{(x_i - \mu)^2}{2\sigma^2} \right) \\ &= -n \log(\sigma) - \sum_{i=1}^{n} \frac{(x_i - \mu)^2}{2\sigma^2} \end{aligned}$$

3.2.4.4.4 Estimation

The MLE is the *right local maximum* of either the likelihood or the log likelihood (and we have to explain what "right" means here). It does not matter whether we maximize likelihood or log likelihood because log is a monotone function. Almost always log likelihood is simpler.

When there are multiple local maxima of the log likelihood, only one will be a good estimator, sometimes called the *efficient likelihood estimator* (ELE). One can find this local maximum by starting sufficiently close (at any preliminary estimator that obeys the square root law) and then applying an optimization method that is guaranteed to go uphill in every step. All of the optimization procedures we considered in the course notes on optimization make this guarantee (except simulated annealing).

The ELE need not be the global maximizer, and the global maximizer need not exist (the supremum of the log likelihood can be infinite but the ELE exists and can be found by the procedure in the preceding paragraph in most applications).

Thus to find best estimates (MLE, a. k. a. ELE), one needs some *good* (but not best) estimates to start the optimization.

3.2.4.4.5 Asymptotics

Fisher Information

In "nice" problems (technical conditions for this to happen are given in theory courses)

$$\operatorname{var}_\theta\{l_n'(\theta)\} = -E_\theta\{l_n''(\theta)\}$$

In words, the variance of the first derivative of the log likelihood is minus the expectation of its second derivative.

Either side is called *Fisher information* for the parameter θ for sample size n, denoted $I_n(\theta)$, that is,

$$\begin{aligned} I_n(\theta) &= \operatorname{var}_\theta\{l_n'(\theta)\} \\ &= -E_\theta\{l_n''(\theta)\} \end{aligned}$$

Because the log likelihood is the sum of IID terms (Section 3.2.4.4.3 above), so is its first and second derivative. Because the expectation of a sum is the sum of the expectations and similarly for variance (for IID) (Section 3.1.2.8 above),

$$I_n(\theta) = nI_1(\theta)$$

In words, Fisher information for sample size n is n times Fisher information for sample size one.

Observed Information

Because the log likelihood is the sum of IID terms (Section 3.2.4.4.3 above), the LLN says minus the second derivative of the log likelihood is close to its expectation, which is Fisher information, for large n.

Thus

$$J_n(\theta) = -l_n''(\theta)$$

is close to $I_n(\theta)$ for large n, and we can use this as an estimate when it is inconvenient or impossible to do the sum or integral (as the case may be) in the calculation of Fisher information. This quantity is called *observed information*.

Some people call Fisher information *expected Fisher information* and call observed Fisher information *observed Fisher information*. I do this myself sometimes.

Asymptotics of the ELE

If $\hat{\theta}_n$ denotes the efficient likelihood estimator, then in "nice" problems (technical conditions for this to happen are discussed in theory courses)

$$\hat{\theta}_n \approx \text{Normal}\left(\theta, I_n(\theta)^{-1}\right)$$

The asymptotic distribution of the ELE (a. k. a., MLE) is normal centered at the true unknown parameter value θ and the asymptotic variance is inverse Fisher information for sample size n. Of course, we don't know θ so we have to plug in our estimate, which does not change the asymptotics. We have two good estimates, either expected or observed Fisher information with the ELE plugged in. Hence

$$\hat{\theta}_n \approx \text{Normal}\left(\theta, I_n(\hat{\theta}_n)^{-1}\right)$$
$$\hat{\theta}_n \approx \text{Normal}\left(\theta, J_n(\hat{\theta}_n)^{-1}\right)$$

All of the above are valid in "nice" problems for large n.

It is not obvious that this obeys the square root law, but it does

$$I_n(\theta)^{-1} = \frac{1}{I_n(\theta)} = \frac{1}{nI_1(\theta)} = \frac{I_1(\theta)^{-1}}{n}$$

so the asymptotic standard deviation is $I_1(\theta)^{-1/2}/\sqrt{n}$ (an instance of the square root law).

But we state the asymptotics the way we have with out an explicit $\sqrt{n}$ because experience shows that when students try to put the $\sqrt{n}$ explicitly they often get it in the wrong place. It is safer to just use Fisher information (expected or observed) for the actual sample size n.

3.2.4.4.6 Efficiency

In "nice" problems, no estimator can be better than the ELE for all θ (and since θ is the unknown true parameter value, what good would it be to only estimate well for some θ and not others). That is why the ELE is called the ELE. It is most efficient as measured by ARE.

So that is why the method of maximum likelihood is so important. It gives the best estimators (at least asymptotically best).

3.2.4.4.7 Hypothesis Tests

Wald

According to the asymptotic theory of the MLE

$$\frac{\hat{\theta}_n - \theta_0}{I_n(\hat{\theta}_n)^{-1/2}} = (\hat{\theta}_n - \theta_0)\sqrt{I_n(\hat{\theta}_n)}$$

or

$$\frac{\hat{\theta}_n - \theta_0}{J_n(\hat{\theta}_n)^{-1/2}} = (\hat{\theta}_n - \theta_0)\sqrt{J_n(\hat{\theta}_n)}$$

is asymptotically standard normal under the null hypothesis that the true unknown parameter value is equal to θ_0 and can be used as a test statistic for one or two tailed tests concerning this hypothesis.

These are by far the most commonly used hypothesis tests. They are, for example, the ones used to obtain the P-values reported by `summary.lm` and `summary.glm`. These were illustrated with LM in Section 3.2 of Part I of the course notes about statistical models (the notes you are reading now being Part II) and with GLM in Section 3.3 of Part I.

Wilks (a. k. a. Likelihood Ratio)

We could also use the method illustrated in Section 8.3 of the course notes on optimization. This says

$$T_n = 2\big[l_n(\hat{\theta}_n) - l_n(\theta_0)\big]$$

is asymptotically chi-square distributed with degrees of freedom that are the number of parameters in the model (here just one) under the null hypothesis that the true unknown parameter value is equal to θ_0. So this can also be used as a test statistic but only for two-tailed tests.

These are also commonly used hypothesis tests. They are, for example, the ones used to obtain the P-values reported by `anova.glm` when the options `test = "Chisq"` or`test` = "LRT" (which do exactly the same thing) are specified. These were illustrated with GLM in Section 3.3.3 of Part I of the course notes about statistical models.

One can convert this into a procedure for one-tailed tests using the so-called *signed likelihood ratio* by basing the test on both the likelihood ratio test statistic T_n and the sign of $\hat{\theta}_n - \theta_0$. Since we know that $\hat{\theta}_n - \theta_0$ is asymptotically normal centered at zero under the null hypothesis $\theta = \theta_0$, it will be positive with probability 1/2 and negative with probability 1/2 under this null hypothesis.

Thus to test, for example

$$H_0 : \theta = \theta_0$$
$$H_1 : \theta > \theta_0$$

We only reject H_0 at level α if $\hat{\theta}_n > \theta_0$ and T_n is greater than the upper chi-square critical value for $\alpha/2$

```
qchisq(alpha / 2, df = p, lower.tail = FALSE)
```

where `alpha` is the significance level and `p` is the number of parameters in the model.

Equivalently, just take the P-value from the two-tailed test and cut it in half, but only report it when $\hat{\theta}_n > \theta_0$. (Otherwise report $P = 1$. Do not reject H_0 at any significance level when the sign of $\hat{\theta}_n - \theta_0$ is not what is specified by the chosen alternative hypothesis. This is the same as any one-tailed test works.)

Rao (a. k. a. Score)

The *Rao test*, also called the *score test* because Fisher called $l_n'(\theta)$ the "score", and also called the *Lagrange multiplier test* for somewhat obscure reason (but only so-called by economists because the name was coined by some economists), is based on the fact that the first derivative of the likelihood is also asymptotically normal centered at zero. Under the null hypothesis $\theta = \theta_0$

$$\frac{l_n'(\theta_0)}{\sqrt{I_n(\theta_0)}}$$

is asymptotically standard normal. So one can use that as a test statistic for one or two tailed tests.

One-tailed tests make sense because the sign of $\hat{\theta}_n - \theta_0$ is always the same as the sign of $l_n'(\theta_0)$. If $l_n'(\theta_0) > 0$, then the $l_n(\theta)$ is a increasing function in a neighborhood of θ_0 so any method of finding a local maximum of the log likelihood that uses steps that always increase the log likelihood must find a local maximum $\hat{\theta}_n$ that is greater than θ_0. Conversely, if $l_n'(\theta_0) < 0$, then the $l_n(\theta)$ is a decreasing function in a neighborhood of θ_0 so any such method will find $\hat{\theta}_n < \theta_0$.

These are less commonly used hypothesis tests, but might be used more often than they are. They are, for example, the ones used to obtain the P-values reported by `anova.glm` when the option `test = "Rao"` is specified. These were illustrated with GLM also in Section 3.3.3 of Part I of the course notes about statistical

models. (I think of this as new in R, because this option to `anova.glm` and `add1.glm` and `drop1.glm` only appeared in R-2.14.0, released in October 2011.)

Summary

That is a lot of ways to do hypothesis tests and only the Wald tests use standardized point estimates, which is the only kind of tests explained in STAT 3011 except for chi-square for contingency tables (which are a special case of Rao tests).

All of these procedures are equivalent in asymptopia where n has already gone to infinity. When n is only very large, they are "asymptotically equivalent" which means they will report nearly the same P-values for the same data. So it doesn't matter which one you use. You don't need to know all of them, just one of them. (And the dogma of hypothesis testing says it is invalid to try all of them and pick the one (if any) that says $P < 0.05$.)

When n is not large, they can be quite different, but asymptotic theory cannot tell us which is better. Some are obviously better than others in some particular situations. But there is no general theory that says one is always preferred.

3.2.4.4.8 Confidence Intervals

Wald

According to the asymptotic theory of the MLE

$$\hat{\theta}_n \pm 1.96 \cdot I_n(\hat{\theta}_n)^{-1/2}$$

or

$$\hat{\theta}_n \pm 1.96 \cdot J_n(\hat{\theta}_n)^{-1/2}$$

is an asymptotic (large sample, approximate) 95% confidence interval for the true unknown θ. To get other confidence levels, change 1.96 to the appropriate quantile of the standard normal distribution.

Level Sets of the Likelihood

We could also use the method illustrated in Section 8.3 of the course notes on optimization. This says the random set

$$\left\{ \theta \in R^p : 2[l_n(\hat{\theta}_n) - l_n(\theta)] > c \right\}$$

is a 95% confidence region for the true unknown parameter value when c is the $1 - \alpha$ quantile of the chi-square distribution with p degrees of freedom, where p is the number of parameters.

When $p = 1$, the region is an interval. As in the section of the optimization and zero-finding notes linked above, this requires nonlinear optimization to find the boundary of the confidence region or zero-finding to find the endpoints of the confidence interval.

Rao

One can also invert the Rao test to obtain confidence intervals, for example, the confidence intervals produced by `prop.test` with optional argument `correct = FALSE` are Rao (a. k. a., score) intervals for the binomial distribution.

But for once we won't elaborate how to do this in general.

3.2.5 Summary of General Asymptotics

There are many CAN estimators. The best is the ELE.

That's all the theory for a while. Let's try some examples.

4 Estimating Location in a Normal Location Model

By "normal location model" we mean that we are pretending that only the location parameter μ is unknown and that we know the scale parameter σ. This is bogus for real data, but (as in intro stats) we do this because, as of right now, we only know how to deal with one-parameter problems (we cover multi-parameter problems below).

The usual estimator of μ is the sample mean. What is the ELE?

Differentiating the log likelihood for the normal distribution given in Section 3.2.4.4.3 above with respect to μ gives

$$\begin{aligned} l_n'(\mu) &= -\sum_{i=1}^n \frac{\partial}{\partial \mu} \frac{(x_i - \mu)^2}{2\sigma^2} \\ &= \sum_{i=1}^n \frac{x_i - \mu}{\sigma^2} \\ &= \frac{n(\bar{x}_n - \mu)}{\sigma^2} \end{aligned}$$

where $\bar{x}_n$ is the sample mean for sample size n

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

Setting equal to zero and solving for μ gives $\mu = \bar{x}_n$. So the MLE (a. k. a., ELE) is the usual estimator, the sample mean.

But this is only for the normal distribution.

Let's check that the asymptotic theory of the ELE agrees with what we know from intro stats and Section 3.1.2.10 above.

$$\begin{aligned} I_n(\mu) &= \operatorname{var}\{l_n'(\mu)\} \\ &= \operatorname{var}\left(\frac{n(\bar{x}_n - \mu)}{\sigma^2}\right) \\ &= \frac{n^2}{\sigma^4} \operatorname{var}(\bar{x}_n - \mu) \\ &= \frac{n^2}{\sigma^4} \cdot \frac{\sigma^2}{n} \\ &= \frac{n}{\sigma^2} \\ &= -E\{l_n''(\mu)\} \\ &= -\frac{\partial}{\partial \mu} \frac{n(\bar{x}_n - \mu)}{\sigma^2} \\ &= \frac{n}{\sigma^2} \end{aligned}$$

So using either formula for Fisher information we get inverse Fisher information, which is the asymptotic variance of the ELE, to be σ^2/n, which is what we already know from either the formula for the variance of the sample mean (Section 3.1.2.10 above) or from the CLT (Section 3.1.2.13 above).

Although we used asymptotic theory of the MLE here, we know this sampling variance is actually exact not asymptotic for this particular estimator (Section 3.1.2.10 above) and asymptotic normality holds for any distribution having a variance not just this particular distribution (Section 3.1.2.13 above).

Because the log likelihood for the normal distribution is exactly quadratic, Wald, Wilks, and Rao procedures (tests and confidence intervals) do exactly the same thing. You won't see any differences among them when normality is assumed (this includes LM in general).

But this is a very special case. In general, the asymptotic variance will not be exact and will only hold for the particular statistical model we are assuming.

5 Estimating Location in a Cauchy Location Model

5.1 Introduction

Since the Cauchy distribution is symmetric we have two obvious estimators of location (the sample mean and the sample median).

But we already know the sample mean is a terrible estimator that does not even obey the LLN. It is not even *consistent*.

The sample median is CAN. But the ELE is even better.

5.2 Log Likelihood and Its Derivatives

Rather than use calculus ourselves let us be lazy and let R do it. From Section 3.1.2.2.2 above and from Section 3.2.4.4.3 above the likelihood for the Cauchy distribution for an IID sample of size n is

$$L_n(\mu) = \prod_{i=1}^{n} \frac{1}{1 + \left(\frac{x_i - \mu}{\sigma}\right)^2}$$

(we can drop the multiplicative constants π and σ that do not contain the parameter μ; we are assuming σ is known here so we have a one-parameter model to analyze). And the log likelihood is

$$\begin{aligned} l_n(\mu) &= \sum_{i=1}^{n} \log\left(\frac{1}{1 + \left(\frac{x_i - \mu}{\sigma}\right)^2}\right) \\ &= -\sum_{i=1}^{n} \log\left(1 + \left(\frac{x_i - \mu}{\sigma}\right)^2\right) \end{aligned}$$

R cannot do derivatives involving `sum` so we have to apply rule that the derivative of a sum is the sum of the derivatives ourselves. R can calculate the derivatives of one term of the sum. Then we have to do the rest.

```
foo <- expression(- log(1 + ((x - mu) / sigma)^2))
foo.deriv <- D(foo, "mu")
foo.deriv.deriv <- D(foo.deriv, "mu")
foo.deriv
```

```
## 2 * (1/sigma * ((x - mu)/sigma))/(1 + ((x - mu)/sigma)^2)
```

```
foo.deriv.deriv
```

```
## -(2 * (1/sigma * (1/sigma))/(1 + ((x - mu)/sigma)^2) - 2 * (1/sigma *
##     ((x - mu)/sigma)) * (2 * (1/sigma * ((x - mu)/sigma)))/(1 +
##     ((x - mu)/sigma)^2)^2)
```

Pretty messy. Here are some Cauchy data simulated from a distribution with $\sigma = 1$.

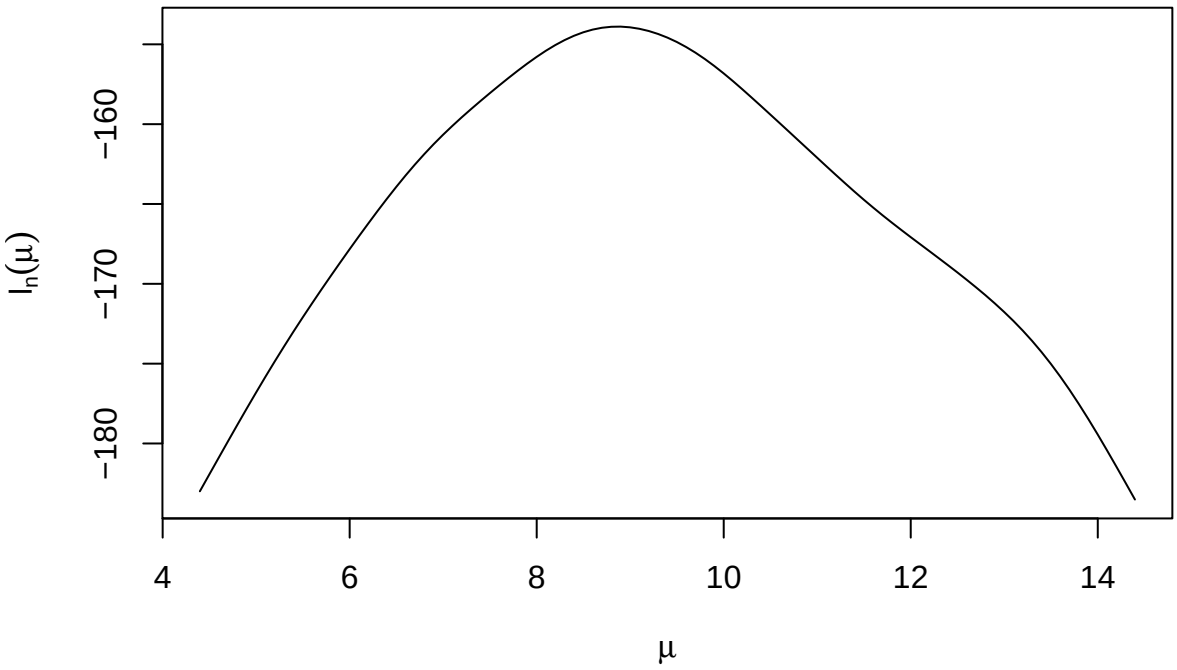
```
x <- read.csv(url("http://www.stat.umn.edu/geyer/5102/data/prob7-1.txt"))$x
```

We can code up the log likelihood using global variables as

```
sigma <- 1
logl <- function(mu) sum(eval(foo))
```

and plot it

```
fred <- Vectorize(logl)
curve(fred, from = median(x) - 5, to = median(x) + 5,
    xlab = expression(mu), ylab = expression(l[n](mu)))
```



5.3 Optimization

There is no closed-form expression for the MLE. Setting the derivative equal to zero and trying to find a solution won't work. It is not just beyond your current math skills. No one could do it. No such formula exists.

So we have to use a computer to find the MLE. As said above (Section 3.2.4.4.4) we need to start the optimization at what is already a good estimator, and we happen to have one, the sample median.

5.3.1 Without Derivatives

Since `nlm` only does minimization, we need to hand it minus the log likelihood.

```
mlogl <- function(mu) sum(- eval(foo))
nout <- nlm(mlogl, median(x))
nout$code %in% c(1, 2)
```

```
## [1] TRUE
```

```
nout$estimate
```

```
## [1] 8.864535
```

```
median(x)
```

```
## [1] 9.3977
```

Our two estimators are different but not that different.

5.3.2 With Derivatives

We seem to have gotten good answers, but we should supply analytic derivatives, if we can, and we can here.

```
mlogl <- function(mu) {
    result <- sum(- eval(foo))
    attr(result, "gradient") <- sum(- eval(foo.deriv))
    return(result)
}
nout.with <- nlm(mlogl, median(x))
nout.with$code %in% c(1, 2)
```

```
## [1] TRUE
```

```
nout$estimate - nout.with$estimate
```

```
## [1] -4.431772e-06
```

5.3.3 Without Derivatives Again

```
mlogl.r <- function(mu) sum(- dcauchy(x, mu, log = TRUE))
nout.again <- nlm(mlogl.r, median(x))
nout.again$code %in% c(1, 2)
```

```
## [1] TRUE
```

```
nout$estimate - nout.again$estimate
```

```
## [1] 6.228316e-10
```

5.4 Confidence Intervals

5.4.1 Using Sample Median

We already know the asymptotic standard deviation of the sample median is

$$\frac{\pi}{2} \cdot \frac{\sigma}{\sqrt{n}}$$

(Section 3.2.4.2.2 above). So here is a confidence interval based on the sample median.

```
conf.level <- 0.95
crit <- qnorm((1 + conf.level) / 2)
crit
```

```
## [1] 1.959964
```

```
median(x) + c(-1, 1) * crit * pi * sigma / (2 * sqrt(length(x)))
```

```
## [1] 8.938754 9.856646
```

5.4.2 Using MLE and Observed Fisher Information

We can ask `nlm` to calculate the second derivative of minus the log likelihood (which is observed information) with an optional argument

```
nout <- nlm(mlogl, nout$estimate, hessian = TRUE)
nout$code %in% c(1, 2)
```

```
## [1] TRUE
```

```
nout$estimate
```

```
## [1] 8.864539
```

```
nout$hessian
```

```
##          [,1]
## [1,] 5.212839
```

so this large-sample approximate 95% confidence interval is

```
nout$estimate + c(-1, 1) * crit * sqrt(1 / as.vector(nout$hessian))
```

```
## [1] 8.006097 9.722981
```

The reason for the `as.vector` above is because R gives a warning otherwise.

5.4.3 Using MLE and Expected Fisher Information

It is a bit worrying that the confidence interval based on the MLE is wider than the confidence interval based on the sample median.

Perhaps we should have used expected Fisher information because the sample size

```
length(x)
```

```
## [1] 45
```

is not that large. There is an R function `integrate` that does one-dimensional integrals in the R core.

```
mu <- nout$estimate
sigma <- 1
integrand <- function(x) - eval(foo.deriv.deriv) * dcauchy(x, mu, sigma)
exp.fish <- integrate(integrand, lower = -Inf, upper = Inf)
exp.fish
```

```
## 0.5 with absolute error < 6.4e-05
```

We had to evaluate at a specific value of μ because it is doing numerical integration. Since we want to evaluate expected Fisher information at the MLE, that is the value we should choose.

What is this thing? According to `?integrate` it is a list and the result is

```
exp.fish$value
```

```
## [1] 0.5
```

The fact that the second derivative is a function of $x - \mu$ tells us that the integral does not, in fact, depend on μ. It does however depend on σ. The fact that the second derivative is a function of $(x - \mu)/\sigma$ except for the $1/\sigma^2$ term tells us that

$$I_1(\mu) = \frac{1}{2\sigma^2}$$

(and this agrees with the integral done in Mathematica).

$$I_n(\mu) = \frac{n}{2\sigma^2}$$

and the asymptotic standard deviation of the MLE is

$$I_n(\mu)^{-1/2} = \sqrt{2} \cdot \frac{\sigma}{\sqrt{n}}$$

So here is another asymptotically valid confidence interval for μ

```
nout$estimate + c(-1, 1) * crit * sqrt(1 / exp.fish$value) / sqrt(length(x))
```

```
## [1] 8.451343 9.277736
```

and this is shorter than the confidence interval based on the sample median.

We divided by $\sqrt{n}$ here because the R object `exp.fish` is expected Fisher information for sample size 1.

5.4.4 Likelihood Level Set

We could also use the method illustrated in Section 8.3 of the course notes on optimization with theory specified in Section 3.2.4.4.8.2 above.

Since we are trying to find end points of an interval, zero-finding is called for.

```
crit.chisq <- qchisq(conf.level, df = 1)
crit.chisq
```

```
## [1] 3.841459
```

```
mu.hat <- nout$estimate
fred <- function(mu) 2 * (logl(mu.hat) - logl(mu)) - crit.chisq
uout.low <- uniroot(fred, interval = c(mu.hat - 5, mu.hat),
    extendInt = "downX")$root
uout.hig <- uniroot(fred, interval = c(mu.hat, mu.hat + 5),
    extendInt = "upX")$root
c(uout.low, uout.hig)
```

```
## [1] 7.993667 9.771138
```

5.4.5 Summary

estimator	standard error	interval
median	$1/2f(\hat{\mu})$	$(8.94, 9.86)$
MLE	$I_n(\hat{\mu})^{-1/2}$	$(8.45, 9.28)$
MLE	$J_n(\hat{\mu})^{-1/2}$	$(8.01, 9.72)$
likelihood	NA	$(7.99, 9.77)$

None of these are exact. All are only asymptotically valid (have approximately the nominal coverage for large n).

Also TIMTOWTDI.

6 Multivariable and Multiparameter

Some things change when we go from random variables to random vectors or from one parameter (estimates of which are random variables) to a vector of parameters (estimates of which are random vectors).

6.1 Probability Theory

6.1.1 Mean Vectors

If X is a random vector having components X_i, its *mean* is the (non-random) vector μ having components μ_i, where

$$\mu_i = E(X_i), \qquad i = 1, \ldots, d,$$

where d is the dimension of X and μ.

This vector notation allows us to write

$$\mu = E(X)$$

rather than the more complicated equation above. Simple.

6.1.2 Covariance

The *covariance* of two random variables X and Y is

$$\operatorname{cov}(X, Y) = E\{(X - \mu_X)(Y - \mu_Y)\},$$

where μ_X and μ_Y are the means of X and Y, respectively.

6.1.3 Correlation

The *correlation* of two non-constant random variables X and Y is

$$\operatorname{cor}(X, Y) = \frac{\operatorname{cov}(X, Y)}{\operatorname{sd}(X) \operatorname{sd}(Y)}$$

The *non-constant* is there to assure that the standard deviations are not zero so we do not have divide by zero.

So correlation is standardized covariance.

6.1.4 Variance Matrices

If X is a random vector having components X_i, its *variance* is the (non-random) matrix Σ having components σ_{ij}, where

$$\sigma_{ij} = \operatorname{cov}(X_i, X_j), \qquad i = 1, \ldots, d, \ j = 1, \ldots, d,$$

where d is the dimension of X, so Σ is a $d \times d$ matrix.

This vector-matrix notation allows us to write

$$\Sigma = \operatorname{var}(X)$$

rather than the more complicated equation above. Simple.

Or maybe not so simple.

- The variance of a random vector is a (non-random) matrix.
- It is common to use the Greek letter Σ for this matrix, possibly decorated: Σ_Y is the variance matrix of the the random vector Y.
- This is kind of like using σ^2 for the variance of a random variable. But not really because σ is the standard deviation of the random variable and Σ is the variance matrix.
- So one might think we should use the Greek letter corresponding to the English v (for variance) but there is no such Greek letter that has the same sound at the English v.

- This is the same Greek letter used for the summation sign, but you can tell them apart because the summation sign is bigger

$$\Sigma = \mathrm{var}(X)$$
$$\overline{X}_n = \frac{1}{n} \sum_{i=1}^{n} X_i$$

- There are many names for this thingummy, and I have chosen one of the least popular.
- Many theory books call this the *covariance* matrix of X because its components are covariances rather than variances. But that is really bad terminology because what then does one call the covariance matrix of two random vectors?
- Notice that $\mathrm{cov}(X, X) = \mathrm{var}(X)$, so the diagonal components of the variance matrix are variances, but the off-diagonal components are covariances that are not variances. This leads some to call $\mathrm{var}(X)$ the *variance-covariance* matrix of X. And this terminology is unambiguous and does not use a term that is needed for other things.
- But variance-covariance is wordy, so someone invented the term *dispersion* matrix.
- Your humble author always uses *variance matrix* on the grounds that
  - this terminology is also unambiguous — there isn't anything else it could mean — and
  - the variance matrix of a random vector plays the same theoretical role as the variance of a random variable.
- But you can use variance-covariance matrix or dispersion matrix or even covariance matrix (bad terminology though this one is) if you like.

6.1.5 Multivariate Linear Transformations

If X is a random variable, then

$$Y = a + bX,$$

where a and b are (non-random) constants, is another random variable that is a linear function of X.

If X is a random vector, then

$$Y = a + BX,$$

where a is a (non-random) constant vector and B is a (non-random) constant matrix, is another random vector that is a linear function of X.

Here the dimensions of a and B must be such that the equation makes sense.

- BX is a matrix multiplication (X is a column vector, as we always have by default if we don't say otherwise). In order for this to make sense, the column dimension of B must match the row dimension of X, which is just the dimension of X. For more on this see Section 4.1 of the course notes on matrices.
- The addition in $a + BX$ is vector addition, so a must have the same dimension of as BX, which is the row dimension of B (assuming BX makes sense).
- Putting both criteria together,
  - the dimension of a is the same as the row dimension of B, and
  - the dimension of X is the same as the column dimension of B, and
  - the dimension of Y is the same as the row dimension of B.

Thus, when B is not a square matrix (both dimensions the same), this linear transformation maps from vectors of one dimension to vectors of another dimension.

6.1.6 Multivariate Linear Transformations, Mean Vectors, and Variance Matrices

We learned in Section 3.1.2.9 above the important facts about means and variances of linear transformations of random variables. We repeat them here

$$E(a + bX) = a + bE(X)$$
$$\operatorname{var}(a + bX) = b^2 \operatorname{var}(X)$$

The analogous formulas for linear transformations of random vectors are

$$E(a + BX) = a + BE(X)$$
$$\operatorname{var}(a + BX) = B \operatorname{var}(X) B^T$$

almost the same but a little bit different. Recall that B^T is the transpose of B Section 6.1 of the course notes on matrices. The reason that the last formula does not have B^2 is because B^2 doesn't even make sense when B is not square, and the reason that B and B^T are not together is because matrix multiplication is not commutative (Section 4.1.2 of the course notes on matrices).

We can also write these equations as

$$\mu_Y = a + B\mu_X$$
$$\Sigma_Y = B\Sigma_X B^T$$

where we are still using the notation

$$Y = a + BX.$$

6.1.7 The Multivariate Normal Distribution

A random vector whose components are IID standard normal random variables is said to be a *multivariate standard normal random vector*. Its mean vector is the zero vector (all components zero). Its variance matrix is the identity matrix (all components diagonal components one, all off-diagonal components zero).

A random vector is said to be a (general) *multivariate normal random vector*. If $Y = a + BZ$, where Z is a standard normal random vector, then $E(Y) = a$ and $\operatorname{var}(Y) = BB^T$.

We won't say any more about the multivariate normal distribution (theory courses say a lot more).

6.1.8 The Multivariate LLN

If X_1, X_2, ... is a sequence of IID random vectors (note that the subscripts here do not indicate components of vectors but rather elements of a sequence — each X_i is a random vector, not a random scalar) having mean vector μ (this means the expectations of the components of X_i exist), and

$$\overline{X}_n = \frac{1}{n} \sum_{i=1}^{n} X_i$$

is the sample mean vector (the summation sign here is vector addition, the multiplication by $1/n$ is scalar multiplication, the result has the same dimension as the dimension of X_i), then $\overline{X}_n$ converges to μ in a sense that we won't even bother to explain here any more than we explained the scalar LLN in Section 3.1.2.12 above.

6.1.9 The Multivariate CLT

If X_1, X_2, ... is a sequence of IID random vectors having mean vector μ and variance matrix Σ, then the sample mean vector $\overline{X}_n$ defined in the preceding section has an approximate multivariate normal distribution with mean vector μ and variance matrix Σ/n for very large n.

We won't bother to explain the exact details. Suffice it to say, using the formulas for linear transformations above, that if

$$\overline{Y}_n = a + B\overline{X}_n$$

makes sense (the dimensions of the vector a and the matrix B are such that this makes sense), then this has an approximate multivariate normal distribution with mean vector $a + B\mu$ and variance matrix $B\Sigma B^T/n$ for very large n. mean vector μ and variance matrix Σ/n for very large n.

In particular, this applies when B has only one row, so we can write it as a row vector b^T and then a must also have only one row, so it is a scalar. Then the equations become

$$\begin{aligned}
\overline{Y}_n &= a + b^T\overline{X}_n \\
\mu_{\overline{Y}_n} &= a + b^T\mu \\
\sigma^2_{\overline{Y}_n} &= \frac{b^T\Sigma b}{n}
\end{aligned}$$

6.2 Multivariable Calculus: Differentiation

6.2.1 Partial Derivatives

Partial derivatives are derivatives with respect to one variable holding the other variables fixed. They are written $\partial y/\partial x$ rather than the "total" derivatives of single variable calculus written dy/dx.

Higher derivatives are more complicated because they can be mixed. If y is a function of x and z, we have four second derivatives

$$\frac{\partial^2 y}{\partial x^2} \qquad \frac{\partial^2 y}{\partial x \partial z} \qquad \frac{\partial^2 y}{\partial z \partial x} \qquad \frac{\partial^2 y}{\partial z^2}$$

In "nice" situations, when these are continuous functions of the variables involved, the middle two are equal (it makes no difference which order one does the differentiation.

For example, take

$$y = \sin(x^2 + z^3)$$

so

$$\begin{aligned}
\frac{\partial y}{\partial x} &= \cos(x^2 + z^3) \cdot 2x \\
\frac{\partial y}{\partial z} &= \cos(x^2 + z^3) \cdot 3z^2 \\
\frac{\partial^2 y}{\partial x^2} &= \cos(x^2 + z^3) \cdot 2 - \sin(x^2 + z^3) \cdot (2x)^2 \\
\frac{\partial^2 y}{\partial z^2} &= \cos(x^2 + z^3) \cdot 6z - \sin(x^2 + z^3) \cdot (3z^2)^2 \\
\frac{\partial^2 y}{\partial x \partial z} = \frac{\partial^2 y}{\partial z \partial x} &= -\sin(x^2 + z^3) \cdot (2x)(3z^2)
\end{aligned}$$

```
D(D(expression(sin(x^2 + z^3)), "x"), "x")
```

```
## cos(x^2 + z^3) * 2 - sin(x^2 + z^3) * (2 * x) * (2 * x)
```

```
D(D(expression(sin(x^2 + z^3)), "z"), "x")
```

```
## -(sin(x^2 + z^3) * (2 * x) * (3 * z^2))
```

```
D(D(expression(sin(x^2 + z^3)), "x"), "z")
```

```
## -(sin(x^2 + z^3) * (3 * z^2) * (2 * x))
```

```
D(D(expression(sin(x^2 + z^3)), "z"), "z")
```

```
## cos(x^2 + z^3) * (3 * (2 * z)) - sin(x^2 + z^3) * (3 * z^2) *
##     (3 * z^2)
```

6.2.2 First Derivatives

A general (not necessarily linear) function $y = f(x)$ that maps vectors to vectors, not necessarily of the same dimension, has component functions $Y_i = f_i(X)$ that give the components of the result.

The *derivative* of f at a point (vector) x is a matrix, sometimes called the *Jacobian matrix* which has components $\partial f_i(x)/\partial x_j$.

This is the Jacobian matrix that the R function `jacobian` in the CRAN package `numDeriv` calculates

```
library(numDeriv)
f <- function(x) {
    u <- x[1]
    v <- x[2]
    c(sin(u) + v^2, cos(u) + v^3, tan(u) + v^4)
}
jacobian(f, c(0, 1))
```

```
##      [,1] [,2]
## [1,]    1    2
## [2,]    0    3
## [3,]    1    4
```

Since `f` maps two-dimensional vectors to three-dimensional vectors, the Jacobian matrix is 3 by 2. (Exercise for the reader: check that `jacobian` got the right answer.)

It is also the Jacobian matrix needed for the argument `hin.jac` of the R function `auglag` in the CRAN package `alabama` that was illustrated in Section 8.3 of the course notes on optimization.

In case f is a scalar-valued function of one vector variable, the first derivative is a vector with components $\partial f(x)/\partial x_i$. Then is is often called the *gradient vector*.

The difference:

- when f is vector-to-vector, so the first derivative is a matrix, it is called the *Jacobian matrix* or just the *Jacobian*, and
- when f is vector-to-scalar, so the first derivative is a vector, it is called the *gradient vector* or just the *gradient*.

The CRAN package `numDeriv` also has an R function `grad` to do this job.

```
f <- function(x) {
    u <- x[1]
    v <- x[2]
    sin(u) + v^2
}
jacobian(f, c(0, 1))
```

```
##      [,1] [,2]
## [1,]    1    2
```

```
grad(f, c(0, 1))
```

```
## [1] 1 2
```

(The only difference is that `jacobian` always produces a matrix and `grad` always produces a vector.)

In mathematics, the gradient vector is denoted

$$g(x) = \nabla f(x).$$

This means $g(x)$ is the gradient vector of f at the point x.

6.2.3 Second Derivatives

The *second derivative* of a vector-to-vector function f at a point (vector) x is a three-index thingummy $\partial f_i(x)/\partial x_j \partial x_k$. This could be put into an R array but cannot be considered a matrix (hence it has no name other than second derivative of f).

The *second derivative* of a vector-to-scalar function f at a point (vector) x can be considered a matrix having components $\partial f(x)/\partial x_i \partial x_j$. This called the *Hessian matrix*.

In writing, it is always spelled Hessian with a capital H because it is named after a person. In R code, it is usually `hessian` with a lower-case h because that is the R way. For example in the `hessian` argument to the R function `nlm` and in the `hessian` component of the result of this function.

In mathematics, the Hessian matrix is denoted

$$h(x) = \nabla^2 f(x).$$

This means $h(x)$ is the Hessian matrix of f at the point x.

The CRAN package `numDeriv` also has an R function `hessian`

```
hessian(f, c(0, 1))
```

```
##               [,1]          [,2]
## [1,]  9.436974e-07 -1.868117e-13
## [2,] -1.868117e-13  2.000000e+00
```

The R function `deriv` in the R core will do gradients and Hessians but not Jacobians (that is, it only does vector-to-scalar functions)

```
foo <- expression(sin(u) + v^2)
deriv3(foo, namevec = c("u", "v"), function.arg = TRUE)
```

```
## function (u, v)
## {
##     .expr1 <- sin(u)
##     .value <- .expr1 + v^2
##     .grad <- array(0, c(length(.value), 2L), list(NULL, c("u",
##         "v")))
##     .hessian <- array(0, c(length(.value), 2L, 2L), list(NULL,
##         c("u", "v"), c("u", "v")))
##     .grad[, "u"] <- cos(u)
##     .hessian[, "u", "u"] <- -.expr1
##     .hessian[, "u", "v"] <- .hessian[, "v", "u"] <- 0
##     .grad[, "v"] <- 2 * v
##     .hessian[, "v", "v"] <- 2
##     attr(.value, "gradient") <- .grad
##     attr(.value, "hessian") <- .hessian
##     .value
## }
```

```
# Wow!  That's hard to read.
fred <- deriv3(foo, namevec = c("u", "v"), function.arg = TRUE)
fred(0, 1)
```

```
## [1] 1
## attr(,"gradient")
##      u v
## [1,] 1 2
## attr(,"hessian")
## , , u
##
##      u v
## [1,] 0 0
##
## , , v
##
##      u v
## [1,] 0 2
```

6.3 Multiparameter Maximum Likelihood Estimation

When θ is a vector parameter, the MLE $\hat{\theta}_n$ is a random vector. As in the one-parameter case, we do not know its exact sampling distribution but do know its asymptotic (large sample size, approximate) distribution.

6.3.1 Log Likelihood Derivatives

When there is more than one parameter, the log likelihood is a vector-to-scalar function so we write

- its first derivative as
$$\nabla l_n(\theta)$$
(its *gradient vector*)
- and its second derivative as
$$\nabla^2 l_n(\theta)$$
(its *Hessian matrix*).

6.3.2 Fisher Information Matrix

For "nice" models (we won't try to explain exactly what this mean any more than we did in the one-parameter case)

$$\operatorname{var}_\theta\{\nabla l_n(\theta)\} = -E_\theta\{\nabla^2 l_n(\theta)\}$$

(on the left-hand side we have the variance matrix of the random vector $\nabla l_n(\theta)$, on the right-hand side we have the expectation of the random matrix $\nabla^2 l_n(\theta)$, which is the non-random matrix whose components are the expectations of the components of $\nabla^2 l_n(\theta)$). Perhaps this is clearer if written out componentwise

$$E_\theta\left\{\frac{\partial l_n(\theta)}{\partial\theta_i}\cdot\frac{\partial l_n(\theta)}{\partial\theta_j}\right\} = E_\theta\left\{\frac{\partial^2 l_n(\theta)}{\partial\theta_i\partial\theta_j}\right\}, \qquad \text{for all } i \text{ and } j$$

The left-hand side is a covariance because

$$E_\theta\left\{\frac{\partial l_n(\theta)}{\partial\theta_i}\right\} = 0, \qquad \text{for all } i$$

But, as we are just about to see, we must think of Fisher information as a matrix. Its components do no work separately.

Either side of this identity is called Fisher information

$$I_n(\theta) = \operatorname{var}_\theta\{\nabla l_n(\theta)\} = -E_\theta\{\nabla^2 l_n(\theta)\}$$

6.3.3 Observed Information Matrix

As in the one-parameter case, the Hessian of the log likelihood will be close to its expectation when sample size is large by the (multivariate) LLN. So we define the *observed Fisher information matrix*

$$J_n(\theta) = -\nabla^2 l_n(\theta)$$

it can be used instead of $I_n(\theta)$ when n is large.

As in the one-parameter case we sometimes say

- *expected Fisher information* instead of *Fisher information*, and
- *observed Fisher information* instead of *observed information*.

6.3.4 Asymptotics of the MLE

For large n

$$\hat{\theta}_n \approx \text{Normal}\left(\theta, I_n(\hat{\theta}_n)^{-1}\right) \approx \text{Normal}\left(\theta, J_n(\hat{\theta}_n)^{-1}\right)$$

where here inverse Fisher information must mean *matrix inverse* (Section 6.6.1 of the course notes on matrices) and "Normal" indicates multivariate normal distribution having mean vector the true unknown parameter vector θ and variance matrix inverse Fisher information (either expected or observed).

7 Multivariate Statistical Theory for MLE

7.1 Hypothesis Tests

Suppose, as in Section 8.3 of the course notes about optimization, the parameter vector θ has the form (γ, δ), where γ and δ are both possibly vectors, called the *parameter of interest* and the *nuisance parameter*, respectively. (Or one could say the components of γ are called the *parameters of interest* and the components of δ are called the *nuisance parameters*).

Although we write (γ, δ) so all the parameters of interest come before all the nuisance parameters in the (combined) parameter vector, this is just for notational convenience. Obviously nothing depends on the order in which the parameters appear in the parameter vector, so long as you keep straight which is which.

Suppose we want to do a hypothesis test with null and alternative hypotheses

$$\begin{aligned} H_0 &: \gamma = \gamma_0 \\ H_1 &: \gamma \neq \gamma_0 \end{aligned}$$

where γ_0 is the value of the nuisance parameter vector specified under the null hypothesis.

7.1.1 Wilks

The Wilks test, also called the likelihood ratio test (LRT) uses the test statistic

$$T_n = 2\left[\left(\sup_{\substack{\gamma \in R^p \\ \delta \in R^q}} l_n\big((\gamma, \delta)\big)\right) - \left(\sup_{\delta \in R^q} l_n\big((\gamma_0, \delta)\big)\right)\right]$$

where p and q are the dimensions of γ and δ, respectively. Under the null hypothesis, the asymptotic distribution of this test statistic is chi-square with p degrees of freedom.

If we want to state the test statistic in terms of the MLE, let

- $(\hat{\gamma}_n, \hat{\delta}_n)$ denote the MLE for the alternative hypothesis and
- $(\gamma_0, \tilde{\delta}_n)$ denote the MLE for the null hypothesis.

then

$$T_n = 2 \left[l_n\big((\hat{\gamma}_n, \hat{\delta}_n)\big) - l_n\big((\gamma_0, \tilde{\delta}_n)\big) \right]$$

Note that in maximizing the likelihood over the alternative hypothesis where both parameters of interest and nuisance parameters need to be estimated we get MLE of both $\hat{\gamma}_n$ and $\hat{\delta}_n$, but in maximizing the likelihood over the null hypothesis, where the parameters of interest are specified ($\gamma = \gamma_0$) so only the nuisance parameters need to be estimated, we get MLE of only the nuisance parameters $\tilde{\delta}_n$. Also note that in general $\hat{\delta}_n \neq \tilde{\delta}_n$. The MLE of δ is different depending on whether γ is also estimated or is held fixed.

7.1.2 Wald

The Wilks test seems to require finding the MLE for both the null and alternative hypotheses, or at least finding the optimal value in two optimization problems. The Wald test requires only the MLE for the alternative. Its test statistic is

$$W_n = (\hat{\gamma}_n - \gamma_0)^T \left[I_n\big((\hat{\gamma}_n, \hat{\delta}_n)\big)^{-1}_{\gamma\gamma} \right]^{-1} (\hat{\gamma}_n - \gamma_0)$$

where

$$I_n\big((\hat{\gamma}_n, \hat{\delta}_n)\big)^{-1}_{\gamma\gamma}$$

denotes the γ, γ block of the inverse Fisher information matrix (the variance matrix of the asymptotic distribution of $\hat{\gamma}_n$).

Note that there are two matrix inverses in the formula for the Wald test statistic. First we invert the full Fisher information matrix. That is the asymptotic variance matrix of the MLE of all the parameters. Then we take the block of this matrix that is the asymptotic variance matrix of the MLE for the parameters of interest. Then we invert that block.

If we used observed instead of expected Fisher information the test statistic would have the same asymptotic distribution. The asymptotic distribution of the test statistic is the same as for the Wilks test: chi-square with p degrees of freedom.

In fact, the Wald and Wilks tests are asymptotically equivalent, meaning the difference between these test statistics will be negligible for very large n. These tests will give almost the same P-values for the same data.

7.1.3 Rao

The Wilks test seems to require finding the MLE for both the null and alternative hypotheses, or at least finding the optimal value in two optimization problems. The Wald test requires only the MLE for the alternative. The Rao test, also called the score test and the Lagrange multiplier test, requires only the MLE for the null. Its test statistic is

$$R_n = \big[\nabla l_n\big((\gamma_0, \tilde{\delta}_n)\big)\big]^T \big[I_n\big((\gamma_0, \tilde{\delta}_n)\big)\big]^{-1} \big[\nabla l_n\big((\gamma_0, \tilde{\delta}_n)\big)\big]$$

Note that

$$\nabla l_n\big((\gamma_0, \tilde{\delta}_n)\big)$$

is the first derivative of the log likelihood for the big model (alternative hypothesis) evaluated at the MLE for the little model (null hypothesis). So it is not zero. The components corresponding to nuisance parameters are zero because we have maximized over those parameters, but the components corresponding to parameters of interest are not zero because we did not maximize over those parameters.

If we used observed instead of expected Fisher information the test statistic would have the same asymptotic distribution. The asymptotic distribution of the test statistic is the same as for the Wilks test: chi-square with p degrees of freedom.

In fact, the Rao, Wald and Wilks tests are asymptotically equivalent, meaning the difference between these test statistics will be negligible for very large n. These tests will give almost the same P-values for the same data.

7.2 Confidence Regions and Intervals

The confidence region for any of these tests is simply the parameter values that are not rejected by the test. More specifically, a confidence region for the vector parameter of interest γ having coverage probability $1 - \alpha$ is the set of parameter values γ_0 such that test does not reject the null hypothesis $H_0 : \gamma = \gamma_0$ at level α.

If one wants a confidence region for all the parameters, just consider all the parameters to be parameters of interest (so δ has length zero).

If one wants a confidence interval for one parameter, just consider it the only parameter of interest.

8 MLE for Two-Parameter Normal

The log likelihood for the normal distribution is given in Section 3.2.4.4.3 above. We repeat it here

$$l_n(\mu, \sigma) = -n \log(\sigma) - \sum_{i=1}^{n} \frac{(x_i - \mu)^2}{2\sigma^2}$$

Since this is theory, we want to estimate the variance, so we change parameters to $\nu = \sigma^2$. Then the log likelihood is

$$l_n(\mu, \nu) = -\frac{n}{2} \log(\nu) - \frac{1}{2\nu} \sum_{i=1}^{n} (x_i - \mu)^2$$

Then

$$\begin{aligned} \frac{\partial l_n(\mu, \nu)}{\partial \mu} &= \frac{1}{\nu} \sum_{i=1}^{n} (x_i - \mu) \\ \frac{\partial l_n(\mu, \nu)}{\partial \nu} &= -\frac{n}{2\nu} + \frac{1}{2\nu^2} \sum_{i=1}^{n} (x_i - \mu)^2 \end{aligned}$$

Setting the first equal to zero and solving gives $\mu = \bar{x}_n$, so the MLE of μ is $\hat{\mu}_n = \bar{x}_n$.

So plugging this into the second and solving gives

$$\hat{\nu}^2 = \frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x}_n)^2$$

which we could call the "sample variance" if intro stats books had not stolen this term to apply to the same thing with n replaced by $n - 1$ (of course, for large n this makes very little difference).

Now

$$\begin{aligned} \frac{\partial^2 l_n(\mu, \nu)}{\partial \mu^2} &= -\frac{n}{\nu} \\ \frac{\partial^2 l_n(\mu, \nu)}{\partial \mu \partial \nu} &= -\frac{1}{\nu^2} \sum_{i=1}^{n} (x_i - \mu) \\ \frac{\partial^2 l_n(\mu, \nu)}{\partial \nu^2} &= \frac{n}{2\nu^2} - \frac{1}{\nu^3} \sum_{i=1}^{n} (x_i - \mu)^2 \end{aligned}$$

In taking expectations we know that constants come out and sums come out and

$$\begin{aligned} E(X_i - \mu) &= 0 \\ \operatorname{var}(X_i - \mu) &= \nu \end{aligned}$$

so

$$\begin{aligned} -E_{\mu,\nu}\left\{\frac{\partial^2 l_n(\mu,\nu)}{\partial\mu^2}\right\} &= \frac{n}{\nu} \\ -E_{\mu,\nu}\left\{\frac{\partial^2 l_n(\mu,\nu)}{\partial\mu\partial\nu}\right\} &= 0 \\ -E_{\mu,\nu}\left\{\frac{\partial^2 l_n(\mu,\nu)}{\partial\nu^2}\right\} &= -\frac{n}{2\nu^2} + \frac{1}{\nu^3}\cdot n\nu \\ &= \frac{n}{2\nu^2} \end{aligned}$$

Since $E(\overline{X}_n) = \mu$ (Section 3.1.2.10 above), the expectation of the mixed partial derivative is zero and we get the Fisher information matrix

$$I_n(\mu,\nu) = \begin{pmatrix} n/\nu & 0 \\ 0 & n/(2\nu^2) \end{pmatrix}$$

It is very easy to invert a diagonal matrix; just invert each diagonal component.

```
foo <- diag(1:4)
foo
```

```
##      [,1] [,2] [,3] [,4]
## [1,]    1    0    0    0
## [2,]    0    2    0    0
## [3,]    0    0    3    0
## [4,]    0    0    0    4
```

```
solve(foo)
```

```
##      [,1] [,2]      [,3] [,4]
## [1,]    1  0.0 0.0000000 0.00
## [2,]    0  0.5 0.0000000 0.00
## [3,]    0  0.0 0.3333333 0.00
## [4,]    0  0.0 0.0000000 0.25
```

So the inverse Fisher information matrix is

$$I_n(\mu,\nu)^{-1} = \begin{pmatrix} \nu/n & 0 \\ 0 & 2\nu^2/n \end{pmatrix}$$

And we have found out something we didn't know. We already knew the asymptotic variance of the MLE of μ, which we found to be the sample mean, is $\nu/n = \sigma^2/n$. But we didn't know that the asymptotic variance of the MLE of ν, which we found to be the sample variance except for a factor of $n/(n-1)$, which is negligible for large n, is $2\nu^2/n = 2\sigma^4/n$.

9 MLE for Two-Parameter Cauchy

9.1 Log Likelihood and Its Derivatives

Unlike in Section 5.2 above when we treat both μ and σ as parameters, we cannot drop terms containing σ. So we have to redo the log likelihood calculation.

$$\begin{aligned} l_n(\mu, \sigma) &= \log\left(\prod_{i=1}^n \frac{1}{\pi\sigma\left[1 + \left(\frac{x_i - \mu}{\sigma}\right)^2\right]}\right) \\ &= \sum_{i=1}^n \log\left(\frac{1}{\pi\sigma\left[1 + \left(\frac{x_i - \mu}{\sigma}\right)^2\right]}\right) \\ &= -n\log(\pi) - n\log(\sigma) - \sum_{i=1}^n \log\left(1 + \left(\frac{x_i - \mu}{\sigma}\right)^2\right) \end{aligned}$$

and we can drop the additive term that does not contain the parameters obtaining

$$l_n(\mu, \sigma) = -n\log(\sigma) - \sum_{i=1}^n \log\left(1 + \left(\frac{x_i - \mu}{\sigma}\right)^2\right)$$

Again because we want R to do the derivatives, we look at the log likelihood for sample size one

$$l_1(\mu, \sigma) = -\log(\sigma) - \log\left(1 + \left(\frac{x_1 - \mu}{\sigma}\right)^2\right)$$

and have R differentiate that

```
foo <- expression(- log(sigma) - log(1 + ((x - mu) / sigma)^2))
foo.mu <- D(foo, "mu")
foo.sigma <- D(foo, "sigma")
foo.mu.mu <- D(foo.mu, "mu")
foo.mu.sigma <- D(foo.mu, "sigma")
foo.sigma.sigma <- D(foo.sigma, "sigma")
foo.mu
```

```
## 2 * (1/sigma * ((x - mu)/sigma))/(1 + ((x - mu)/sigma)^2)
```

```
foo.sigma
```

```
## -(1/sigma - 2 * ((x - mu)/sigma^2 * ((x - mu)/sigma))/(1 + ((x -
##     mu)/sigma)^2))
```

```
foo.mu.mu
```

```
## -(2 * (1/sigma * (1/sigma))/(1 + ((x - mu)/sigma)^2) - 2 * (1/sigma *
##     ((x - mu)/sigma)) * (2 * (1/sigma * ((x - mu)/sigma)))/(1 +
##     ((x - mu)/sigma)^2)^2)
```

```
foo.mu.sigma
```

```
## -(2 * (1/sigma * ((x - mu)/sigma^2) + 1/sigma^2 * ((x - mu)/sigma))/(1 +
##     ((x - mu)/sigma)^2) - 2 * (1/sigma * ((x - mu)/sigma)) *
##     (2 * ((x - mu)/sigma^2 * ((x - mu)/sigma)))/(1 + ((x - mu)/sigma)^2)^2)
```

```
foo.sigma.sigma
```

```
## 1/sigma^2 - (2 * ((x - mu)/sigma^2 * ((x - mu)/sigma^2) + (x -
```

```
##     mu) * (2 * sigma)/(sigma^2)^2 * ((x - mu)/sigma))/(1 + ((x -
##     mu)/sigma)^2) - 2 * ((x - mu)/sigma^2 * ((x - mu)/sigma)) *
##     (2 * ((x - mu)/sigma^2 * ((x - mu)/sigma)))/(1 + ((x - mu)/sigma)^2)^2)
```

9.2 Optimization

Since the derivatives are so messy, let's just do the no-derivative form.

We need good estimates of the parameters to use for starting values.

- As in the one-parameter case, the sample median is a good estimator of μ.
- Since we know the quartiles are $-\sigma$ and σ, half the interquartile range is a good estimator of σ.

```
mlogl <- function(theta) {
    stopifnot(is.numeric(theta))
    stopifnot(is.finite(theta))
    stopifnot(length(theta) == 2)
    stopifnot(theta[2] > 0)
    mu <- theta[1]
    sigma <- theta[2]
    sum(- dcauchy(x, mu, sigma, log = TRUE))
}
theta.twiddle <- c(median(x), IQR(x) / 2)
nout <- nlm(mlogl, theta.twiddle)
nout$code %in% c(1, 2)
```

```
## [1] TRUE
```

```
nout$estimate
```

```
## [1] 9.266664 4.362070
```

```
theta.twiddle
```

```
## [1] 9.3977 3.9765
```

It seems the assumption $\sigma = 1$ that we used in Section 5 above was quite wrong. The scale parameter is more like 4.

9.3 Confidence Intervals

9.3.1 Using MLE and Observed Fisher Information

```
conf.level <- 0.95
crit <- qnorm((1 + conf.level) / 2)
crit
```

```
## [1] 1.959964
```

```
nout <- nlm(mlogl, nout$estimate, hessian = TRUE)
nout$code %in% c(1, 2)
```

```
## [1] TRUE
```

```
nout$estimate
```

```
## [1] 9.266664 4.362070
```

```
nout$hessian
```

```
##            [,1]       [,2]
## [1,]  1.2045727 -0.1179716
## [2,] -0.1179716  1.1599769
```

```
# inverse observed Fisher information
obsinfo.inverse <- solve(nout$hessian)
# confidence intervals
for (i in 1:2)
   print(nout$estimate[i] + c(-1, 1) * crit * sqrt(obsinfo.inverse[i, i]))
```

```
## [1]  7.471908 11.061419
## [1] 2.53314 6.19100
```

9.3.2 Using MLE and Expected Fisher Information

Now that Fisher information is a matrix, we are going to have to do integrals component by component.

```
mu <- nout$estimate[1]
sigma <- nout$estimate[2]
integrand <- function(x) - eval(foo.mu.mu) * dcauchy(x, mu, sigma)
exp.fish.mu.mu <- integrate(integrand, lower = -Inf, upper = Inf)
integrand <- function(x) - eval(foo.mu.sigma) * dcauchy(x, mu, sigma)
exp.fish.mu.sigma <- integrate(integrand, lower = -Inf, upper = Inf)
integrand <- function(x) - eval(foo.sigma.sigma) * dcauchy(x, mu, sigma)
exp.fish.sigma.sigma <- integrate(integrand, lower = -Inf, upper = Inf)
exp.fish.mu.mu
```

```
## 0.02627755 with absolute error < 6.6e-05
```

```
exp.fish.mu.sigma
```

```
## 8.776661e-12 with absolute error < 4.6e-05
```

```
exp.fish.sigma.sigma
```

```
## 0.02627755 with absolute error < 6.7e-05
```

```
exp.fish <- matrix(c(exp.fish.mu.mu$value,
                     exp.fish.mu.sigma$value,
                     exp.fish.mu.sigma$value,
                     exp.fish.sigma.sigma$value), nrow = 2)
exp.fish
```

```
##              [,1]         [,2]
## [1,] 2.627755e-02 8.776661e-12
## [2,] 8.776661e-12 2.627755e-02
```

```
exp.fish.inverse <- solve(exp.fish)
```

We see that the off-diagonal component is nearly zero, so the two estimators $\hat{\mu}_n$ and $\hat{\sigma}_n$ are asymptotically nearly uncorrelated. Actually, this is just error in the numerical integration. Exact symbolic integration in Mathematica proves the off-diagonal component is exactly zero. It also says the two diagonal components are equal, despite being about dissimilar parameters. Both are $1/(2\sigma^2)$. Let's check that.

```
all.equal(1 / (2 * sigma^2), exp.fish[1, 1])
```

```
## [1] TRUE
```

```
all.equal(1 / (2 * sigma^2), exp.fish[2, 2])
```

```
## [1] TRUE
```

So the corresponding confidence intervals are

```
for (i in 1:2)
   print(nout$estimate[i] + c(-1, 1) * crit *
       sqrt(exp.fish.inverse[i, i]) / sqrt(length(x)))
```

```
## [1]  7.464271 11.069057
## [1] 2.559677 6.164462
```

9.3.3 Profile Likelihood Level Set

The term *profile likelihood* means the likelihood as a function of the parameter of interest after the nuisance parameter has been maximized over. If γ is the parameter of interest and δ is the nuisance parameter, then the profile likelihood useful for inference about γ is

$$\tilde{l}_n(\gamma) = \sup_{\delta \in R^q} l_n(\gamma, \delta).$$

Conceptually, a confidence region for the parameter of interest is a level set of the profile likelihood. But it is very inefficient to actually compute the profile likelihood. The computationally efficient method is to follow Section 8.3 of the course notes about optimization.

```
crit.chisq <- qchisq(conf.level, df = 1)

theta.hat <- nout$estimate
names(theta.hat) <- c("mu", "sigma")
theta.hat
```

```
##       mu    sigma
## 9.266664 4.362070
```

```
library(alabama)

fred <- function(theta) crit.chisq - 2 * (mlogl(theta) - mlogl(theta.hat))
foo <- matrix(NaN, 2, 2)
for (i in 1:2) {
    sally <- function(theta) theta[i]
    aout <- auglag(theta.hat, fn = sally, hin = fred,
        control.outer = list(trace = FALSE))
    # important to check for convergence
    stopifnot(aout$convergence == 0)
    sally <- function(theta) - theta[i]
    aout.too <- auglag(theta.hat, fn = sally, hin = fred,
        control.outer = list(trace = FALSE))
    # important to check for convergence
    stopifnot(aout.too$convergence == 0)
    foo[i, 1] <- aout$value
    foo[i, 2] <- (- aout.too$value)
}
rownames(foo) <- names(theta.hat)
colnames(foo) <- c("lower", "upper")
foo
```

```
##          lower     upper
## mu    7.480974 11.190893
## sigma 2.861286  6.645003
```

9.3.4 Summary

Table 2: 95% Confidence Intervals for Location Parameter

type	interval
Wald with observed Fisher information	(7.47, 11.06)
Wald with expected Fisher information	(7.46, 11.07)
Wilks, a. k. a., profile likelihood	(7.48, 11.19)

None of these are exact. All are only asymptotically valid (have approximately the nominal coverage for large n).

Table 3: 95% Confidence Intervals for Scale Parameter

type	interval
Wald with observed Fisher information	(2.53, 6.19)
Wald with expected Fisher information	(2.56, 6.16)
Wilks, a. k. a., profile likelihood	(2.86, 6.65)

None of these are exact. All are only asymptotically valid (have approximately the nominal coverage for large n).

9.3.5 Confidence Regions

9.3.5.1 Wilks

We can make confidence regions that are level sets of the log likelihood following Section 8.2 of the course notes about optimization.

```
phi <- seq(0, 2 * pi, length = 501)
crit.chisq <- qchisq(conf.level, df = 2)

r.wilks <- double(length(phi))
for (i in seq(along = phi)) {
    fred <- function(r) {
        theta <- theta.hat + r * c(cos(phi[i]), sin(phi[i]))
        crit.chisq - 2 * (mlogl(theta) - mlogl(theta.hat))
    }
    r.wilks[i] <- uniroot(fred, interval = c(0, 1), extendInt = "upX")$root
}
xx <- theta.hat[1] + r.wilks * cos(phi)
yy <- theta.hat[2] + r.wilks * sin(phi)
xx <- c(xx, xx[1])
yy <- c(yy, yy[1])
plot(xx, yy, type = "l", xlab = expression(mu), ylab = expression(sigma))
points(theta.hat[1], theta.hat[2])
```

9.3.5.2 Wald

Here, where we want a confidence interval for all the parameters so there are no nuisance parameters, the formula for the Wald test statistic in Section 7.1.2 above reduces to

$$W_n = (\hat{\gamma}_n - \gamma_0)^T I_n(\hat{\gamma}_n)(\hat{\gamma}_n - \gamma_0)$$

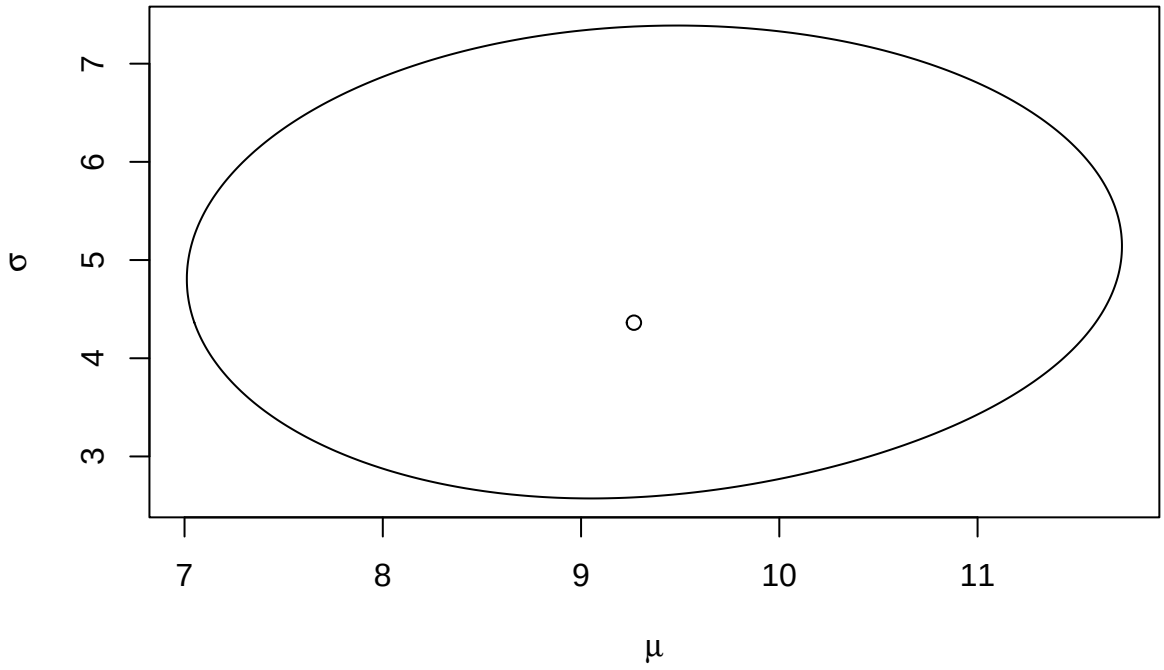


Figure 1: Likelihood Confidence Region. Dot is MLE

so we are just trying to find an ellipse that is a level set of this thought of as a function of γ_0. There are faster ways to figure out an ellipse than trying to mimic the preceding section, but we won't bother with speed here.

```
r.wald <- double(length(phi))
for (i in seq(along = phi)) {
    fred <- function(r) {
        theta <- theta.hat + r * c(cos(phi[i]), sin(phi[i]))
        crit.chisq - length(x) *
            drop(t(theta - theta.hat) %*% exp.fish %*% (theta - theta.hat))
    }
    r.wald[i] <- uniroot(fred, interval = c(0, 1), extendInt = "upX")$root
}
xx <- theta.hat[1] + r.wald * cos(phi)
yy <- theta.hat[2] + r.wald * sin(phi)
xx <- c(xx, xx[1])
yy <- c(yy, yy[1])
plot(xx, yy, type = "l", xlab = expression(mu), ylab = expression(sigma))
points(theta.hat[1], theta.hat[2])
```

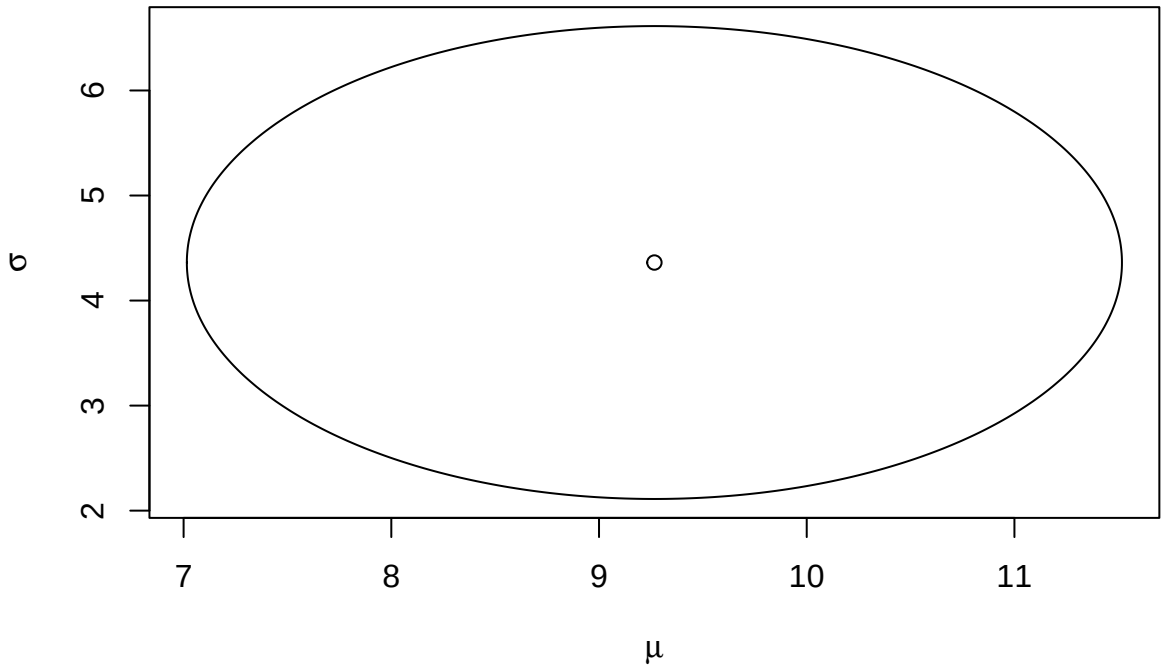


Figure 2: Wald Confidence Region. Dot is MLE

We need `length(x)` in the formula for the Wald statistic because `exp.fish` is expected Fisher information for sample size one.

9.3.5.3 Rao

Here, where we want a confidence interval for all the parameters so there are no nuisance parameters, the

formula for the Rao test statistic in Section 7.1.3 above reduces to

$$R_n = \left[\nabla l_n(\gamma_0)\right]^T \left[I_n(\gamma_0)\right]^{-1} \left[\nabla l_n(\gamma_0)\right]$$

so we are trying to find a level set of this thought of as a function of γ_0.

This looks hard. We would have to do several integrals to calculate expected Fisher information at each point γ_0 we test to figure out whether it is in or out of the region. Even if we replaced expected Fisher information with observed Fisher information, it would still be messy. Let's not bother.

9.3.5.4 Summary

Both regions on the same plot.

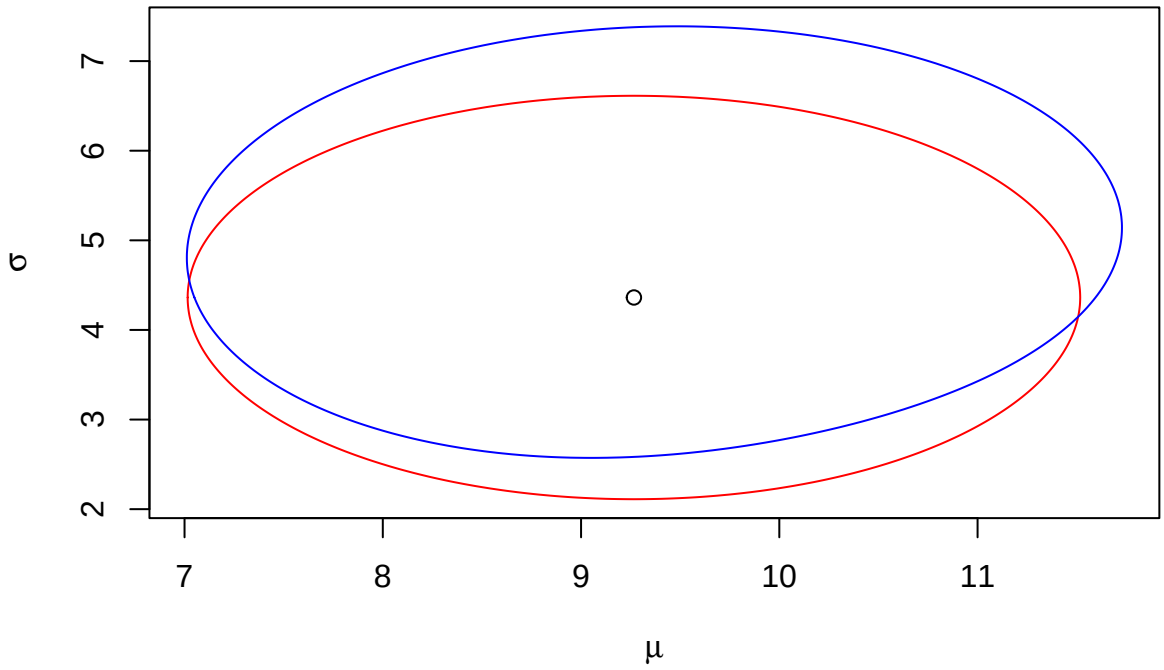


Figure 3: Wilks and Wald Confidence Regions. Dot is MLE. Wilks boundary blue. Wald boundary red.

The fact that the regions don't agree exactly means we aren't in asymptopia. The fact that the regions don't agree nearly means we aren't nearly in asymptopia. The asymptotic approximation is somewhat imperfect.