The Likelihood Principle. The Likelihood Principle is the claim that the likelihood function contains within it all of the ’useful’ information about the parameters that is to be found in the data set. One place it shows up is in analyzing psychometric function data from staircase methods and method of constant stimuli. The levels of the stimulus tested in a staircase method are themselves random; in MOCS, they are pre-defined and only the order of testing is randomized. And yet we fit the two kinds of data in the same way (if you haven’t worked with psychometric functions, you may not be aware of this common practice). The Likelihood Principle is the justfication for doing this, and we will return to it in discussing Bayesian decision theory.

The first optimal property is the *invariance property* of MLEs. Suppose we have a sample $X_1,....,X_n$ from a normal distribution with mean μ and variance σ^2. We can estimate the MLE estimates $\hat{\mu}$ and $\hat{\sigma}^2$ as above. But above we estimated not σspu2 but σ. Does it matter whether we estimate σ^2 by MLE directly, or whether we estimate σ and then square the *estimate* $(\hat{\sigma})^2$?

For MLE's, the answer is no. In fact,

Fact 1: If $f(x;\theta)$ is a pdf, and $\hat{\theta}$ is a MLE estimator of θ, then for any increasing or decreasing function $g(.)$, $g(\hat{\theta})$ is the MLE of $g(\theta)$.

So if we perversely decided to reparameterize the Normal distribution Normal(μ,σ^2) by Normal(ν,σ) where $\nu = \mu^3$, then we'd just have to take the MLE estimates of μ and σ^2 and transform them similarly to get MLE estimates of ν and σ.

I'll prove the fact above in class (it's easy but requires a little arm-waving). What's remarkable is that, if a distribution has several parameters (say μ, σ) and we transform these parameters by any function $g(\mu,\sigma)$, then the MLE of this new quantity is just $g(\hat{\mu},\hat{\sigma})$. So the 'increasing or decreasing' assumption of the previous theorem is unnecessary, and we need not confine ourselves to transformations of one parameter at a time.

If we suspect that our choice of parameterization is utterly arbitrary, then, we can take comfort from the fact that, if we are using MLE, it doesn't matter. A second estimation method that we'll discuss later in the course, Bayesian decision theory (BDT), does not have this invariance property. If we change parameterization and use BDT, we can get a seriously different picture of the world from our estimates. This should make Bayesians uncomfortable but, hey, *nothingFP makes Bayesians uncomfortable.*

A second property: the MLE may be biased, but it is asymptotically unbiased as were the Method of Moments Estimates.

A third property: Asymptotically, the MLE is distributed as a Normal random variable, and no other estimator that is asymptotically unbiased and Normally distributed (and satisfies a few technical conditions) can have a lower asymptotic variance. In jargon, the MLE is Best-Asymptotically-Normal (BAN). So eventually, for large enough samples, the MLE is the least variable estimator.

A more complex (and more common) summary assumes that the p_i are determined by a *psychometric curve* which has a limited number of parameters (typically 2). The Weibull-Quick Psychometric function assumes that $p_i = 1 - e^{-(\frac{x_i}{\alpha})^\beta}$? There are only two α and β. Once these parameters are chosen the p_i and the observer's performance are characterized. If we assumed that the Quick-Weibull described the observer, we would want to estimate α and β from the data. To do so by MLE is fairly easy.

First we compute the likelihood,

$$L(\alpha,\beta;DATA) = \prod_{i=1}^{n} p_i^{c_i}(1-p_i)^{n_i - c_i}.$$

The $p_i = p_i(\alpha,\beta)$ are really functions of α and β, but for simplicity, I've not written the full form of the equation above. The log-likelihood is even simpler.

$$l(\alpha,\beta;DATA) = \sum_{i=1}^{n}[c_i \log(p_i) + n_i - c_i \log(1-p_i)].$$

We would like to maximize this as a function of α and β. To do so by calculus, turns out to be basically impossible. However, there are many computer methods to maximize this equation. For many MLE applications this is the only way to solve for the estimates. The use of the Weibull-Quick together with MLE is due to A. B. Watson.

12. Optimal Properties of the Maximum Likelihood Estimator

MLE was first employed by Daniel Bernoulli. Leonhard Euler (for whom 'e' is ntotated) replied to Bernoulli's modest attempt to use MLE with a devastating critique. Why, after all, should we assume that the most parameters that render the data most likely are anywhere near the true parameters? Fig. 3 contains a troubling counterexample. As a consequence, perhaps, of Euler's onslaught, MLE was ignored for over a century by Ronald Fisher, the greatest statistician of all time. He defended its use not by intuition but by showing that MLE had many desirable properties. So, if you buy into MLE, your estimator has these properties. We've already seen that Method of Moments has one such property, asymptotic unibasedness. Fisher showed that MLE had this property and several others as well. More importantly, he set the terms of discourse for comparing estimators (Bayesian, MLE, whatever) as a comparison of desirable properties of each kind of estimator.

11. Maximum Likelihood Estimation of the Weibull Psychometric Function

Suppose, however, you wanted to fit something other than a line to your data. Or it's unrealistic to assume that your error terms are iid Normal variables. You may find a computational formula somewhere, but then again you may not. In many cases, the Method of Maximum Likelihood may be used to create a new method for fitting data.

Suppose you wanted to fit psychometric curve to observer's data. To describe the experimental situation: the experimenter presents stimuli of varying intensities, $I_1, I_2, \cdots, I_n$. He presents stimuli of intensity I_i n_i times. The subject responds YES [I detect the stimulus] or NO [I don't detect the stimulus after every stimulus presentation. The data for each subject is the number of times the subject said YES (denoted c_i) at each intensity I_i. The following table summarizes one series of trials:

Ii	ni	ci
7.0	20	2
7.1	20	5
7.2	20	10
7.3	20	13
7.4	19	15
7.5	20	18

These numbers could be a measure of the intensity of any stimulus. For some reason, the subject only attempted 19 trials at intensity 7.4, but this poses no problem for the fitting method. The number of trials at each intensity need not be the same. There just needs to be enough of them overall.

If the presentation order of all 119 stimuli was randomized, we would call this the *method of constant stimuli*. There are other methods, but they are less often used. Now, let's assume that, at each fixrd intensity p_i, the subject has p_i probaiblity of saying YES, and that his performance across trials produces independent random variables. With these assumptions, the c_i at each intensity are Binomial(n_i,p_i) random variables. The probability p_i are unknown to the experimenter, and we could estimate them by $p_i = \frac{c_i}{n_i}$, the MLE for p for the binomial. The p_i, $i = 1,n$ then summarize the observer's performance.

$$\sum_{i=1}^{n}\overline{x}((Y_i-\overline{Y})-\hat{\alpha}(x_i-\overline{x}))$$

However, if you look closely at this equation, you'll see that we are adding up all the deviations from the mean of the Y_i and all the deviations from the mean of the x_i (multiplied by various terms). But the sum of the deviations about the mean is 0! So the term in the equation above is 0, and we can just add it the Equation (a''):

$$\sum_{i=1}^{n}(x_i-\overline{x})((Y_i-\overline{Y})-\hat{\alpha}(x_i-\overline{x})) = 0. \tag{a'''}$$

We begin to see variances and covariances of the sample emerging. Recall that the covariance and variance of the samples are the covariance and variance of the discrete r.v.s whose cdf's are the empirical cdf's corresponding to the sample. For the record, these are, Define

$$\mathrm{Cov}[X_{discrete},Y_{discrete}] = \frac{\sum_{i=1}^{n}(x_i-\overline{x})(Y_i-\overline{Y})}{n}$$

$$\mathrm{Cov}[X_{discrete},Y_{discrete}] = \frac{\sum_{i=1}^{n}(x_i-\overline{x})^2}{n}$$

Note that the x's are not r.v.s! But we can still define the empirical cdf corresponding to the x_i' and then its covariance with Y and its variance are well defined. Se we can rewrite Equation (a''')

$$n\mathrm{Cov}[Xdiscrete,Ydiscrete]-\hat{\alpha}n\mathrm{Var}[Xdiscrete] = 0. \tag{a''''}$$

and, finally, we have the MLE's for linear regression:

$$\hat{\alpha} = \frac{\mathrm{Cov}[Xdiscrete,Ydiscrete]}{\mathrm{Var}[Xdiscrete]} \tag{A}$$

$$\hat{\beta} = \frac{\sum_{i=1}^{n}(Y_i-\hat{\alpha}x_i)}{n} = \overline{Y}-\hat{\alpha}\overline{x} \tag{B}$$

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^{n}(Y_i-\hat{\alpha}x_i-\hat{\beta})^2}{n}. \tag{C}$$

So that's how we fit lines to data by MLE when the data is perturbed by equal-variance Normal errors. These are the standard computational formulas (or equivalent to the standard computational formulas) that you may have encountered before.

$$l(\alpha,\beta,\sigma) = -n\log(\sigma) - \sum_{i=1}^{n} \frac{(\Delta_i)^2}{2\sigma^2}.$$

Substituting for Δ_i, we get,

$$l(\alpha,\beta,\sigma) = -n\log(\sigma) - \sum_{i=1}^{n} \frac{(Y_i - \alpha x_i - \beta)^2}{2\sigma^2}.$$

Taking the *three* partial derivative with respect to α, β and σ we get,

$$D_\alpha l(\alpha,\beta,\sigma) = -2\sum_{i=1}^{n} \frac{x_i(Y_i - \alpha x_i - \beta)}{2\sigma^2}. \quad \text{(a)}$$

$$D_\beta l(\alpha,\beta,\sigma) = -2\sum_{i=1}^{n} \frac{(Y_i - \alpha x_i - \beta)}{2\sigma^2}. \quad \text{(b)}$$

$$D_\sigma l(\alpha,\beta,\sigma) = -\frac{n}{\sigma} + \sum_{i=1}^{n} \frac{(Y_i - \alpha x_i - \beta)^2}{\sigma^3}. \quad \text{(c)}$$

These aren't as bad as they look! We set these three partial derivatives to 0. These aren't as bad as they look! In particular, the {sigma sup 2} and the constant can be dropped from the first two equations. They don't affect the solution.

$$\sum_{i=1}^{n} x_i(Y_i - \hat{\alpha} x_i - \hat{\beta}) = 0. \quad \text{(a')}$$

$$\sum_{i=1}^{n} (Y_i - \hat{\alpha} x_i - \hat{\beta}) = 0. \quad \text{(b')}$$

The third equation is easily solved for σ^2 in terms of $\hat{\alpha}$ and $\hat{\beta}$:

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^{n} (Y_i - \hat{\alpha} x_i - \hat{\beta})^2}{n}. \quad \text{(c')}$$

So we have two equations in two unknowns $\hat{\alpha}$ and $\hat{\beta}$. First we solve Equation (b') for $\hat{\beta}$:

$$\hat{\beta} = \frac{\sum_{i=1}^{n} (Y_i - \hat{\alpha} x_i)}{n} = \overline{Y} - \hat{\alpha}\overline{x} \quad \text{(b'')}$$

where $\overline{Y}$ is the mean of the Y_i, $\overline{x}$ is the mean of the x_i (the x_i aren't r.v.s, but we can still average them.) Then we substitute $\hat{\beta}$ into Equation (a'):

$$\sum_{i=1}^{n} x_i((Y_i - \overline{Y}) - \hat{\alpha}(x_i - \overline{x})) = 0. \quad \text{(a'')}$$

Now, look at that first x_i. We're going to turn that into x_i–xbar. That means that we have to add and subtract:

$$l(\mu,\lambda;X) = n\log(\lambda) - \lambda\sum_{i=1}^{n}|X_i - \mu|.$$

The partial derivatives with respect to λ is,

$$D_\lambda l(\mu,\lambda;X) = \frac{n}{\lambda} - \sum_{i=1}^{n}|X_i - \mu|.$$

This gives us,

$$\hat{\lambda} = \frac{n}{\sum_{i=1}^{n}|X_i - \hat{\mu}|}.$$

which will be the estimate of λ when we figure out the estimate of μ. But we have a problem with the partial derivative D_μ. It's not defined when μ equals one of the X_i. And in between, it's a bit peculiar. A graph (Figure 5) shows the story for the case $n=5$, the sample being 1.2, 3, 4.5, 6, 10. The negative of the log likelihood function is plooted versus. (I plotted the negative so that the maximum would become a minimum and be near the x-axis and thereby easier to read.

We'll explore the behavior of this estimator in class. Can you guess what the estimator is for n odd? Can you prove it? [Hint: consider a continuous function whose slope is not defined at a finite number of points. Suppose the slope goes from negative to positive across a region. Does there have to be a local minimum of the function in that region?] What about when n is even?

10. Maximum Likelihood Estimation and Linear Regression

The MLE Method now we are going to derive linear regression. The model is this: We pick values $x_1, \cdots, x_n$. These aren't random. The we assume that at each x_i we observe Y_i determined by

$$Y_i = \alpha x_i + \beta + \varepsilon_i$$

where the ε_i are iid Normal with mean 0 and variance σ^2. We want to estimate α and β, the slope and intercept of the line relating x_i to Y_i.

Now the key step is to rewrite the equation as

$$\Delta_i(\alpha,\beta,\sigma) = Y_i - \alpha x_i - \beta = \varepsilon_i$$

and see that, the $\Delta_i(\alpha,\beta,\sigma)$ are iid normal with mean 0 and variance σ^2. The Log Likelihood Function is (leaving out the constant terms),

$$D_\sigma l(\mu,\sigma;X) \;=\; = \frac{-n}{\hat{\sigma}} + \frac{1}{\hat{\sigma}^3}\sum_{i=1}^{n}(X_i - \hat{\mu})^2 = 0$$

And from the second,

$$\sigma = \sqrt{\frac{\sum_{i=1}^{n}(X_i - \overline{X})^2}{n}}.$$

So the MLE Estimates are the same as the Method-of-Moments estimators.

8. Maximum Likelihood Estimation and the Uniform Distribution

The MLE Estimate and the Method of Moments Estimate need not always be the same. In fact, in doing the problem from last time, you found that the Method of Moments can give parameter estimates that are *impossible* given the data. The distribution was the Uniform distribution $\frac{1}{a}I[0,a]$ and the estimate of a was *less than* one of the sample values, an impossibility. The MLE estimate, though, will always provide estimates that are consistent with the data, for otherwise the likelihood function would be 0. If it were 0 for all possible parameter values, the sample really is impossible! Otherwise, it's positive for some other value of the parameter(s) and the zero-likelihood value would never be chosen.

So, let's work out the MLE of the Uniform above. The trick is to renumber the sample so that X_1 is the smallest and X_n the largest. Now if, $a<X_n$, the likelihood is 0. Whenever $a \geq X_n$, the likelihood function $L(a;X)$ is $\frac{1}{a^n}$. A plot of this function shows that it is decreasing with a. The maximum is therefore at X_n. Remembering that we reordered the X_i, we get

$$\hat{a} = \max(X_1, \cdots, X_n).$$

The MLE is very different from the Method of Moments Estimator. We might describe it whimsically as 'shrink wrap the data'.

9. Maximum Likelihood Estimation and the Laplacean Distribution

The Laplacean distribution has the pdf $\frac{\lambda}{2}e^{-\lambda|x-\mu|}$. I've plotted it in Figure 1. It's two half-height exponentials back-to-back but with allocation parameter μ added to the natural parameter of the exponential λ. Given a sample XSub1, $\cdots, X_n$, the log-likelihood function is (with a constant left out),

define likelihood.

Let $f(x,\Theta)$ be the pdf (with unknown parameters Θ). The likelihood of the sample is,

$$L(\Theta;X) = f_{X_1 \cdots X_n}(X_1 \cdots X_n;\Theta)$$

where the pdf on the right hand side is the multivariate pdf of the sample. Since the r.v.s in a sample are independent, this becomes,

$$L(\Theta;X) = f(X_1;\Theta)f(X_2;\Theta) \cdots f(X_n;\Theta).$$

Likelihood is well-defined for a sample and is also well-defined even when the X_i are not independent or not identically-distributed. In that case we must use the first equation with the multivariate pdf. Note: Likelihood will show up again and again in statistics.

7. Maximum Likelihood Estimation and the Normal Distribution

For now, let's estimate the mean μ and variance σ^2 of a sample of iid Normal r.v.'s. For this derivation, I'm going to use σ as the variable rather than σ^2. The reason? Didactic, as you'll see further on. So the sample is $X_1, \cdots ,X_n$ and

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

We can take the logarithm of each $f(X_i;\mu,\sigma)$ and then ADD these quantities together, and maximize the resulting log-likelihood function. So let compute $\log(f(X_i;\mu,\sigma))$:

$$\log(f(X_i;\mu,\sigma)) = -\log(\sqrt{2\pi}) - \log(\sigma) - \frac{(X_i-\mu)^2}{2\sigma^2}$$

and, to begin with we can drop the first term as it is constant and can't affect the maximum. Adding the terms, we get,

$$l(\mu,\sigma;X) = -n\log(\sigma) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(X_i-\mu)^2$$

and this is what we wish to maximize as a function of the two variables μ and σ. To do that we take the partial derivative with respect to μ and the partial derivative with respect to σ: We set both to 0. This give us two equations to solve in two unknowns.

$$D_\mu l(\mu,\sigma;X) \; = \; \frac{1}{2\hat{\sigma}^2}\sum_{i=1}^{n}2(X_i-\hat{\mu}) = 0$$

We could maximize $L(p;X)$ graphically (in class demonstration). It seems plausible that 0.63 is our estimate, in agreement with the Method of Moments.

We can maximize the function $L(p;X)$ by calculus. We differentiate with respect to p and set the derivative to 0. Then we solve for the points where the derivative is 0. There are various ways to decide which of these points is a true maximum of which graphing is acceptable.

To directly differentiate $L()$ in this case and solve for the maximum is hard work. Here's a trick. The logarithm is a striclty increasing function on the positive numbers where it is defined, and as a consequence, the maximum of $L()$ and $\log L()$ is the same. But the *log-likelihood function*, usually denoted $l()$ is often simpler and easier to differentiate as it is here:

$$l(p;X) = 63\log(p)+37\log(1-p)$$

and differentiating,

$$D_p l(p;X) = \frac{63}{p} - 37 over 1-p$$

and, setting this to 0 and solving for p gives $63/100$. If you go back through this derivation with the number of heads set to h instead of 63, and the size of the sample set to n, you will end up with h/n which is the same as $\overline{X}$ for this sample.

The Maximum Likelihood Estimate [MLE] and the Method of Moments Estimate will often coincide. After all, they're both trying to do a good job of estimating the same thing. When they agree the estimator has all the properties of both kinds of estimates. In this case, for example, we know that the MLE is asymptotically unbiased since it coincides with the Method of Moments estiamtor which is asymptotically unbiased. (Actually, it turns out that this particular MLE is unbiased. We don't need the 'asymptotically'.). We'll see below that MLE's have several good properties some of which I'll describe later on in this lecture. We discussed some properties of the Method of Moments Estimate above.

For a discrete r.v., the Likelihood Function is just the probability of the observed outcome, i.e. the sample. For a continuous r.v., though, we have the same old problem we always have. the probability of any one outcome (e.g. X=9.7334521...) is always precisely 0. The probability of the sample is also 0. So what do we do? We invented the pdf so that we could compute expectations for continuous variables just to get around this 'any specific outcome never happens' problem. It turns out that we will use the pdf again to

(You should get 1.30 approximately).

Example 3: 3. You estimated that α is 1.30. Does it bother you that the sample includes 2? Why? Comment: the Method of Moments can produce estimates that are not just bad, but impossible.

Example 4: 4. Suppose that $\alpha > 0$ is a scale parameter. That is, we're working with a whole family of cdf's that are of the form $F(\frac{x}{\alpha})$ for various choice of α.

Let $G(x) = F(\frac{x}{\alpha})$ be the cdf resulting from fixing a particular value of α. $F(x)$ is of course the 'base' cdf that gives rise to the α 'family'. Generate some examples to show that a QQ-plot of data from $G(x)$ against the theoretical quantile function of $F(x)$ is a straight line of slope α. [Let F() be some specific cdf like the exponential or Gaussian.] Can you prove it? [very hard. Show that the points $(G^{-1}(t), F^{-1}(t))$ fall on a straight line. Then use the Glivenko-Cantelli Theorem to show that, for large enough samples, the inverse of the empirical cdf approaches $G(x)$, the theoretical cdf.]

6. Maximum Likelihood Estimation

The second method of estimation we will go through is Maximum Likelihood Estimation. Let's start with an example, a Bernoulli(p) random variable with atoms 0 and 1 and where the probability of a 1 is p. We toss the coin n=100 times and get the sample $X_1, X_2, \cdots, X_n$ where each X_i is 0 or 1. Let's say there are 63 1's and 37 0's. How should we estimate p? The Method-of-Moments estimator is easily shown to be $\overline{X} = 0.63$. A second approach is to examine the probability of the outcome. The probability of the particular outcome we got is a long product of p's and 1-p's with 63 p's and 37 1-p's.

$$L(p; X_1, \cdots, X_n) = p^{63}(1-p)^{37}.$$

The notation on the left-hand side needs explaining. Here the variable is the parameter p and the X_i are fixed: the sample has been taken, they are all either 0 or 1, and there is nothing uncertain about them. Often I'll abbreviate $L(p; X_1, \cdots, X_n)$ to just $L(p; X)$. The notation is similar to that of a parameterized distribution $f(x;\theta)$ but the role of parameters and variables have been reversed. The function $L()$ is an example of a *likelihood function*. (*It is not a probability density function*. In particular, it need not integrate to 1. We'll come back to this point when we examine Bayesian decision theory.) We are going to estimate p by maximizing this function.

$$S^2 = \frac{n}{n-1}\hat{\sigma}^2$$

as

$$E[S^2] = \frac{n}{n-1}E[\hat{\sigma}^2] = \frac{n}{n-1}\frac{n-1}{n}\sigma^2 = \sigma^2.$$

and we see that S^2, the estimate of the variance you likely learned in undergraduate statistics, is unbiased.

5. Conclusion

We've seen one method for generating statistics and learned about two desireable properties that a statistic may have: unbiasedness and asymptotic unbiasedness. Not all methods of moments estimators are unbiased, but they are all asymptotically unbiased. The Glivenko-Cantelli theorem provides the basis for showing that the moments of the sample converge *in probability* to the moments of the underlying theoretical distribution and that, therefore, all methods of moments estimators are asymptotically unbiased. Convergence in probability is the sort of convergence we saw for the Glivenko-Cantelli theorem itself. If you name a probability p less than 1 and a precision $\varepsilon>0$, then I can find a sample size N_0 such that the method of moments estimator for a given parameter will be within ε of the true value of the parameter with probability p or more.

So, if we use method of moments estimation, and collect lots of data, we can expect that our estimates will typically be close to what we're trying to estimate.

That the good part about the Method of Moments. The bad part is that there's a certain arbitrariness in the method. We could have, for example, in estimating λ in the exponential, used the second raw moment in place of the first, or for that matter, any of the raw moments. We get estimators in each case, and it's not obvious which to choose. We'll see below that there are better methods that guarantee more desireable properties.

Example 1: 1. Consider the uniform pdf $\frac{1}{\alpha}I[0,\alpha]$ (the r.v. is drawn uniformly from the interval $[0,\alpha]$), which has one unknown parameter $\alpha > 0$. Derive the Method of Moments estimate of α.

Example 2: 2. Apply your estimator $\hat{\alpha}$ to the sample of size 8:

2.000, 0.063, 0.911, 0.147, 0.269, 0.924, 0.713, 0.162

If we compute the expected value of $\hat{\sigma}^2$, we get

$$E[\hat{\sigma}^2 \; = \; E[\sum_{i=1}^{n} \frac{(X_i - \overline{X})^2}{n} \; = \; \sum_{i=1}^{n} \frac{E[(X_i - \overline{X})^2]}{n}.$$

Now, $E[(X_i - \overline{X})^2]$ is the same for every $i = 1,2, \cdots ,n$, and the expression above simplifies to,

$$E[\hat{\sigma}^2 \; = \; E[(X_1 - \overline{X})^2].$$

(I picked X_1 arbitrarily. Any of the X_i would do.) Now, at first glance this *does* look like the variance $E[(X_1 - \mu)^2]$, but it's not since we have $\overline{X}$ in place of μ. So we need to keep computing. The trick that makes things work out is the following:

$$E[(X_1 - \overline{X})^2]. \; = \; E[((X_1 - \mu) + (\mu - \overline{X}))^2] \; = \; E[((X_1 - \mu) - (\overline{X} - \mu))^2].$$

Now, we are going to expand the square term,

$$E[((X_1 - \mu) - (\overline{X} - \mu))^2] \; = \; E[(X_1 - \mu)^2] - E[2(X_1 - \mu)(\overline{X} - \mu)] + E[(\overline{X} - \mu)^2]$$

Now, the first term is σ^2, and the third term is (since $E[\overline{X}] = \mu$),

$$Cov[\overline{X},\overline{X}] = \frac{i}{n^2} Var[X_1 + \cdots + X_n] = \frac{1}{n^2} n\sigma^2 \; = \; \frac{1}{n}\sigma^2.$$

Now, we must compute the middle term, but it is just,

$$-2Cov[X_1,\overline{X}] = -\frac{2}{n} Cov[X_1, X_1 + X_2 + \cdots + X_n] = -\frac{2}{n}\sigma^2.$$

(as the X_i are independent).

Adding all three terms together, we get

$$E[\hat{\sigma}^2 \; = \; (1 - \frac{1}{n})\sigma^2 = \frac{n-1}{n}\sigma^2.$$

So $\hat{\sigma}^2$ is a biased estimator with bias $\frac{\sigma^2}{n}$. We see that Method of Moments estimators can be biased.

However, it is true that as $n \rightarrow \infty$, the bias present in any Method of Moments estimator shrinks to 0. In the above, the bias $\frac{\sigma^2}{n}$ shrinks to 0 as n increases. We say that Method of Moments estimators are *asymptotically unbiased*. The term *asymtotically* will recur and it refers to something that becomes true as we take larger and larger samples.

Since unbiasedness is a desireable property in a statistic, we might try to modify the Method of Moments estimator to remove bias. Multiplying by $\frac{n-1}{n}$ does the trick,

$$\hat{\mu} = \overline{X}$$

$$S^2 = \hat{\sigma}^2 + \hat{\mu}^2.$$

Solving,

$$\hat{\mu} = \overline{X}$$

$$\hat{\sigma}^2 = S^2 - \overline{X}^2.$$

If we work with this last formula, we get

$$\hat{\sigma}^2 = \sum_{i=1}^{n} \frac{(x_i - \overline{X})^2}{n},$$

which is almost the formula you learned in undergraduate statistics. Here we divide by n, not n-1. We'll see why in a moment.

so Method of Moments estimators are not so hard, and I think you now have a better idea of where the formulas you learned in undergraduate statistics come.

4. Bias and Unbiased Estimators

Now the methods we used to design statistics (estimators) usually guarantee that the estimators will have certain desireable properties. For instance, suppose that α is an unknown parameter of a distribution of interest, and, suppose that $\hat{\alpha}$ is an estimator of α. Now, $\hat{\alpha}$ is a function of the random variables in the sample, and despite its Greek heritage, is itself a random variable. We should write it explicitly as a function of the sample: $\hat{\alpha}(X_1, \cdots, X_n)$. And so $E[\hat{\alpha}]$ is the perfectly well-defined expected value of $\hat{\alpha}$.

For the Method-of-Moments estimators $\hat{\mu}$ and $\hat{\sigma}^2$ that we derived for the Gaussian parameters μ and σ^2, we have

$$E[\hat{\mu}] = E\left[\frac{\sum_{i=1}^{n} X_i}{n}\right] = \frac{1}{n}\sum_{i=1}^{n} E[X_i]$$

using the linearity of expectation. But $E[X_1] = \cdots = E[X_n] = \mu$ as the X_i are iid Gaussian. So, $E[\hat{\mu}] = \mu$. The expected value of the estimator is the thing estimated. We say that the estimator $\hat{\mu}$ is *unbiased*. This property holds for any size sample (for any n).

The *bias* of a statistic is $\hat{\alpha}$ is

$$\text{Bias}[\hat{\alpha}] = E[\hat{\alpha}] - \alpha$$

and an *unbiased statistic* is one where $\text{Bias}[\hat{\alpha}] = 0$.

cumulative distribution function. What we really want is the first raw moment of Y, corresponding to the first raw moment of the distribution computed above. Let the sample be $X_1, \cdots, X_n$:

$$E[Y] = \frac{\sum_{i=1}^{n} X_i}{n} = \overline{X}$$

where $\overline{X}$ ('X bar') is just the usual notation for the mean of the sample.

Now, the last step in the Method of Moments is to equate the corresponding moments of the empirical cdf and the moments of the distribution expressed in terms of its parameters. Here we have one of each and we get

$$\overline{X} = \frac{1}{\lambda}.$$

This equation isn't really true. We'd have to be impossibly lucky for $\overline{X}$ to be exactly $\frac{1}{\lambda}$. What we really mean is that $\overline{X}$ is a good estimate for $\frac{1}{\lambda}$. We solve the notation problem by re-introducing the 'hat' notation: $\hat{\lambda}$. Now we solve for the parameter in terms of the data,

$$\hat{\lambda} = \frac{1}{\overline{X}}.$$

This is our Methods-of-Moments estimator for λ denoted $\hat{\lambda}$).

Now, in general, when we have more than one unknown parameter, we'll need as many equations as there are unknowns to solve for the unknown parameters. Said above that we compute as many raw moments of both the theoretical distribution and the empirical cdf of the sample as there are unknown parameters, and that's not quite true. Instead, we will compute as many raw moments as we need to in order to be able to solve for the parameters of the theoretical distribution in terms of the raw moments. This is almost always 'as many raw moments as there are prameters' but it needn't be. What's important is to be able to solve the equations for the parameters in terms of the moments.

Let's try the Gaussian distribution next:

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

The first two raw moments of the theoretical distribution are, μ and $\sigma^2+\mu^2$. The first two raw moments of the empirical cdf are $\overline{X}$ and $S^2 = \sum_{i=1}^{n} X_i^2$. So, we have the simultaneous equations to solve,

A small 'hat' over a parameter means that the quantity represented is an estimate of the corresponding parameter. So our estimate of standard deviation is $\hat{\sigma}$ and our estimate of variance is this quantity squared: $\hat{\sigma}^2 = (\hat{\sigma})^2$.

Now, the graphical estimation procedures we've examined work pretty well. And they have the advantage that, if your assumptions about the distribution are way off, you'll find out rather quickly. There'll be no way to shape the pdf to the histogram, the theoretical QQ plot won't be a straight line, etc. But there are two big problems (and a few little ones) with these procedures. The first is that you need a *lot* of data. The second is that the estimating procedure is *ad hoc*.

What we are going to do next is to learn some *methods* of creating good (or even optimal) statistics. Another word for 'estimator' is 'statistic', so you could think of this as *meta-statistics*: they tell us how to design statistics. The term 'mathematical statistics' (contrasted with just 'statistics') also refers to methods for creating and evaluating estimators.

I'm going to describe two methods in detail and sketch a third: The Method of Moments, The Maximum Likelihood Method, and Bayesian Estimation. We'll begin with the Method of Moments.

3. The Method of Moments

The Method of Moments is the oldest method and is more or less superseded by more modern methods. Let's apply it to the exponential problem above. We begin by calculating the *raw moments* of the distribution. They'll be formulas containing the parameter λ. We calculate at least as many moments as we have unknown parameters, i.e. one in this case. Let's calculate the first raw moment. If X is an exponential rv with parameter λ

$$E[X] = \int_0^\infty x\lambda x e^{-\lambda x} dx = \frac{1}{\lambda}$$

by Statistician's Integration.

Next, we calculate the corresponding moments of the sample. The moments of the sample are just the moments are of the *empirical cdf* of the sample, the discrete rv that has the sample values as atoms and assigns each of the values the probability $\frac{1}{n}$ (if the sample values are all distinct. If some value occurs m times in the sample, it gets the probability $\frac{m}{n}$). So let Y be the discrete rv corresponding to the empirical

In brief, we have *partial knowledge* of f(x), but one or parameters remain unknown. We could call this situation a *black box*. A black box, like a white box is a source of iid random variables, and it has knobs on the side (parameters) controlling its behavior. But now we don't know the settings of the knobs, and we would like to estimate one or more of these settings.

Often we will write the pdf in a suggestive way: $f(x;\lambda)$ for the exponential, for example. This reminds us that λ is unknown and that $f(x;\lambda)$ really represents any of an infinite number of possible exponential distributions. We would write the Gaussian pdf (with unknown mean and variance) as $f(x;\mu,\sigma^2)$. Note the use of semicolon and comma. If we're talking about a generic distribution (as when proving theorems or making definitions, we'll denote it $f(x;\theta_1, \cdots \theta_n)$ using the Greek letter θ ('theta'). Since the above notation is a little clumsy, we'll often just use a single Greek capital letter in place of all the θ_i: $f(x;\Theta)$. Remember that this Θ stands for $(\theta_1, \cdots \theta_n)$. Later on we'll even subscript Θ when we're deciding which of two setting of the parameters is more plausible (Θ_0 or Θ_1) in hypothesis testing.

For now, though, let's return to our sample and assume that it is iid exponential with unknown parameter λ. We wish to estimate λ as best we can. How do we do it?

If the sample is very large, we could histogram it, and then draw the pdf of the exponential for various values of λ until it looks about the same as the histogram. It's not as crazy as it sounds!

A cleverer way is to remember what a QQ Plot does when you change λ. If you have the correct distribution, QQ Plot of the data versus the distribution produces a straight line, and the slope of that line will be the scale parameter in the pdf, the intercept will be the location parameter. For the exponential, λ isn't a scale parameter, but $\frac{1}{\lambda}$ is. So we could certainly estimate λ by fitting a line to the theoretical QQ plot.

For a two-parameter distribution, we could still take the same approach. Fitting the histogram is tedious, but possible. If the two parameters are location and scale parameters, respectively, the theoretical QQ plot would work well.

So consider estimating the mean and variance of a normal random variable. The pdf is denoted $f(x;\mu,\sigma^{2)}$, and while variance is not a scale parameter, σ, the standard deviation is. So perhaps we should estimate σ via a theoretical QQ plot and then square the estimated σ to get the estimated σ^2.

$\mathrm{Var}[\overline{X}] = \frac{\sigma^2}{n}$.

The proof is easy. We make a joint n-dimensional r.v. of the X_i. Since they are independent, we just multiply their pdf's together to get the joint pdf. Then

$$\mathrm{Var}[\overline{X}] = \mathrm{Cov}[\overline{X},\overline{X}] = \frac{1}{n^2}\mathrm{Cov}[X_1 + \cdots X_n, X_1, X_n] =$$

$$= \frac{1}{n^2}(\mathrm{Var}[X_1] + \cdots + \mathrm{Var}[X_n]) = \frac{\sigma^2}{n}.$$

2. Introduction to Estimation

So far in this course we've examined the tools needed to build, analyze, and simulate stochastic models. You can now implement a model you read about in a journal article, simulate it and look at what it does. You can alter the assumptions of the model, and see what effect your alterations have. If the simulations suggest that something interesting is going on, you might be able to capture it analytically by working through a bit of calculus. If you were to assign a single generic label to all these activities, it would likely be 'probability.' Now we are going to examine 'statistics' proper.

A typical problem in statistics is the following: you have a *sample*: $X_1, \cdots, X_n$. Calling these numbers a sample implies that (1) we believe they are drawn from a single distribution (let's assume it's a continuous distribution and denote its pdf by $f(x)$), and (2) the numbers are independent. The phrase 'independent, identically-distributed' random variables (abbreviated iid) is often used. You will hear 'iid Gaussian variables with mean 0 and variance σ^2)', or 'iid random variables drawn from a distribution with mean 0'. In the latter case, we may not know the distribution, just that the mean of the distribution (*not* the mean of the sample is 0).

Now, often we're willing to assume that we know the kind of distribution (exponential, Gaussian) that the sample was drawn from, but we don't know one or more of the parameters associated with the distribution. We might know that $f(x)$ is some exponential:

$$f(x) = \lambda e^{-\lambda x} I[0,\infty)$$

but we don't know what λ is. Or we might assume that $f(x)$ is Gaussian but we don't know either the location parameter (the mean μ) or its scale parameter (the standard deviation σ).

Notice that 2 does not assume independence.

Fact 3: Let C be a constant. Then $E[C]=C$.

Let $\mu_X = E[X]$ and $\mu_Y = E[Y]$. Then

Fact 4: $E[X-\mu_X] = 0$.

The *covariance* of two random variables, defined above, is denoted, $\mathrm{Cov}[X,Y] = E[(X-\mu_X)(Y-\mu_Y)]$. Clearly, Cov[X,Y] = Cov[Y,X].

Fact 5: If X and Y are independent, then $\mathrm{Cov}[X,Y] = 0$.

Fact 6: For any constant a, $\mathrm{Cov}[aX,Y] = \mathrm{Cov}[X,aY] = a\,\mathrm{Cov}[X,Y]$.

The proof of the last fact amounts to little more than writing out the expression in integral form and factoring out the constant.

Fact 7: For any r.v.'s X, Y and Z, $\mathrm{Cov}[X+Y,Z] = \mathrm{Cov}[Z,X+Y] = \mathrm{Cov}[X,Y] + \mathrm{Cov}[X,Z]$.

This last fact moves us up to *three* dimensions (three r.v.'s at a time), but the expectation notation hides all this machinery from us. The proof amounts to writing out the expression in integral form and using the linearity of the integral. This is not a short proof, but it is easy.

Fact 8: Let X and Y be joint r.v.s. For any r.v. X $\mathrm{Cov}[X,X] = \mathrm{Var}[X]$.

This one is notationally tricky. The Covariance is with respect to the joint pdf f(x,y), the variance is with respect to the distribution of X. Let's do it out in full.

$$E_{XY}[(X-\mu_X)(X-\mu_X)] = \iint (x-\mu_X)^2 f(x,y)\,dx\,dy$$

$$= \int (x-\mu_X)^2 \left[\int f(x,y)\,dy\right] dx$$

and, as the quantity in brackets in the marginal pdf of X,

$$= \int (x-\mu_X)^2 f_X(x)\,dx.$$

The last expression is the variance of X treated as a univariate variable.

Fact 9: Let X and Y be joint r.v.s. For any r.v. X $\mathrm{Cov}[X+Y,X+Y] = \mathrm{Var}[X+Y] = \mathrm{Var}[X] + \mathrm{Var}[Y] + 2\mathrm{Cov}[X,Y]$.

Comparing **Fact 5** and this fact, we see that if the two variables are independent, then variances add as before. The covariance term above is what messes up additivity for non-independent variables.

Fact 10: Let $X_1, \cdots, X_n$ be independent r.v.s. each with variance σ^2. Let $\overline{X} = \frac{\sum_{i=1}^{n} X_i}{n}$. Then

Lecture 6: Outline

The road to be travelled over ... is full of interest of its own. It familiarizes us with the measurement of variability, and with curious laws of chance that apply to a vast diversity of social subjects. This part of the inquiry may be said to run along a road on a high level, that affords wide views in unexpected directions, and from which easy descents may be made to totally different goals to those we have now to reach. I have a great subject to write upon.

Sir Francis Galton, *Natural Inheritance*, 1889.

1. Covariance and Variance

Last time we looked at the central and raw moments of univariate random variables. We define moments for multivariate random variables by a straighforward analogy. If, for example, (X,Y) has the bivariate pdf $f_{XY}(x,y)$, then the (m,n)'th raw moment of (X,Y) is,

$$\mu'_{mn} = \iint x^m y^n f_{XY}(x,y)\,dx\,dy.$$

It is very easy to show that $\mu_m 0$ is just the m'th raw moment of the univariate marginal random variable X and μ_{0n} is just the n'th raw moment of the univariate marginal random variable Y. Note that $E(X) = \mu_X = \mu_{10}$, where the first two terms are defined in terms of the marginal random variable X, the last in terms of the bivariate random variable (X,Y). That is, the raw moments of the univariate marginal variables X and Y are included in the bivariate moments of (X,Y). The central moments are defined by

$$\mu_{mn} = \iint (x-\mu_{10})^m (y-\mu_{01})^n f_{XY}(x,y)\,dx\,dy.$$

The central moments of the univariate marginal variables X and Y are also included in the bivariate moments of (X,Y). The above definitions readily expand to the N-variate case for $N>2$.

The bivariate central moment $\mu_{11} = E[(X-\mu_X)(Y-\mu_Y)]$ is far more frequently encountered than other multivariate moments and it has a special name: *covariance*.

Now we are going to list a series of simple facts about expectation, especially expectations involivng two or more variables.

Fact 1: If X and Y are independent, then $E[XY] = E[X]E[Y]$.

The Expectation notation is hiding a lot of needless complexity. For example, the first expectation is a 2D integral, the others are 1D.

Fact 2: $E[X+Y] = E[X] + E[Y]$

testing 123