ESTIMATION, CONFIDENCE INTERVALS, AND TESTS OF HYPOTHESES

Notes by

DONALD RICHARDS

Department of Statistics
Center for Astrostatistics
Penn State University

notes revised and lectures delivered
by
B. V. Rao

1. A problem.

Van den Bergh (1985, ApJ 297, p. 361) considered the luminosity function (LF) for globular clusters in various galaxies.

V-d-B's conclusion: The LF for clusters in the Milky Way is adequately described by a normal distribution.

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left[-\frac{(x-\mu)^2}{2\sigma^2}\right].$$

$\mu(\equiv M_0)$ is the mean visual absolute magnitude and σ is the standard deviation of visual absolute magnitude. Magnitudes are log variables (a log-normal distribution). This appears to be one of the few normal distributions in astronomy.

Statistical Problems:

1. On the basis of collected data, estimate the numbers μ and σ. Also, derive a plausible range of values for each of them; etc.

2. V-d-B concludes that the LF is "adequately described" by a normal distribution. How can we quantify the plausibility of this conclusion?

2. Some terminology.

Population: This term is used in two different contexts. If you are studying the luminosity function of globular clusters, then the globular clusters constitute the population and you select a sample from this population and make measurements to draw conclusions. Second context, and this is how we use, is the following. You want to study a particular attribute X, like the luminosity. You make a probabilistic model for this attribute. For example you may want to say that the possible values of X follow a particular density, $f(x)$. Then this model is called the population. Thus, normal population means that the attribute under study obeys density

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left[-\frac{(x-\mu)^2}{2\sigma^2}\right],$$

for some numbers μ and $\sigma > 0$.

Remember, this means that the chances of your observation falling in the interval, say, $(4, 10)$ is the area under the above curve between these two limits. In practice, this means the proportion of observations that lie in this interval equals, approximately, this area. Only when we prescribe the values of μ and σ, the model is completely specified. Otherwise, it is a class of models for the attribute.

The function $f(x)$ is called the **probability density function (p.d.f.)** of X. A **statistical model** is a choice of p.d.f. for X. We wish to choose a model which "adequately describes" data collected on X. A **parameter** is a number that appears in the choice of the density, which is to be determined from observations. For example, μ and σ are parameters for the p.d.f. of the LF for Galactic globulars. **parameter space** is the set of permissible values of the parameters. In the above normal model, the parameter space is $\Omega = \{(\mu, \sigma) : -\infty < \mu < \infty, \sigma > 0\}$. Thus, for example, σ can not be negative.

A **random sample** means mutually independent random variables X_1, $\ldots$, X_n; which all have the same distribution as X. Here n is called the size of the sample. In practice, this amounts to data values $X_1, \ldots, X_n$ which are *fully representative* of the population. In general, Roman letters are used to represent data, and Greek letters are used to represent parameters. For example θ, μ, σ are parameters where as $X_1, \cdots, X_5$, are data. A **statistic** is a number computed from the observations, that is, from the random sample $X_1, \ldots, X_n$. Here are two examples. Sample mean defined as $\bar{X} = \frac{1}{n}\sum\limits_{i=1}^{n} X_i$ and sample variance defined as $S^2 = \frac{1}{n-1}\sum\limits_{i=1}^{n}(X_i - \bar{X})^2$ are statistics. In general, a statistic could be any function $Y = u(X_1, \ldots, X_n)$ of the observations. The **sampling distribution** of a statistic is the probability distribution of the statistic. For example, if $X_1, \cdots, X_n$ is sample from the normal population described above, and the sample mean $\bar{X}$ is the statistic, then the sampling distribution of this statistic is also normal, but with parameters μ and $\sigma/\sqrt{n}$.

An **estimator** or **estimate** for a parameter is a statistic. Thus estimator is nothing but a number computed from the sample. But when you say estimator, you should also say the parameter for which this is proposed as an estimator. For example $\bar{X}$ is an estimator for μ and S^2 is an estimator for σ^2. These are also called **point estimators**.

3. Estimation.

As mentioned earlier, unless the parameters are explained, the model is not fully specified. Having proposed a class of models for the attribute under study, how do we estimate the parameters, to fully specify the model. For example, in modeling the LF, the proposal was that a normal model fits the data. But which normal model?

How do we construct estimates and how do we know a good estimate from a bad one. There are several methods for constructing estimates for the parameters. Judicious guessing, the method of Maximum Likelihood, the method of Moments, method of Minimum χ^2, Bayesian methods, Decision-theoretic methods etc. There are several criteria proposed for estimators. Unbiased estimator, Consistent estimator, Efficient estimator, etc. Keep in mind that an estimator is a random variable, because it depends on the observations and the observations are, in turn, random variables.

An estimator Y for a parameter θ is **unbiased** if $E(Y) = \theta$. Intuitively,

Y is unbiased if its long-term average value is equal to θ. In the above normal population model, $\bar{X}$ is an unbiased estimator of μ. This is because, $E(X_i) = \mu$ for each i. Also S^2 is an unbiased estimator of σ^2. On the other hand, if you put Y as the largest of the observations, then Y is NOT an unbiased estimator. It appears that, after all, this maximum is one of the observations and each observation has expected value μ, so Y must have the same property. But it is not so. Indeed, if the sample size is at least two, then $E(Y) > \mu$.

An estimator is consistent if it gets closer and closer to the parameter value as the sample size increases. One way of stating this is to say that the chances of it differing from the parameter, by a preassigned quantity, become smaller and smaller, no matter what the preassigned quantity is. More precisely, an estimator Y, for a parameter θ, is **consistent** if for any $\epsilon > 0$, we have $P(|Y - \theta| \geq \epsilon) \longrightarrow 0$ as $n \to \infty$. You should remember that the estimator Y depends on the sample size n. Actually, we should have written Y_n for the estimator based on a sample of size n. In the above normal population model $\bar{X}$ is a consistent estimator of μ. This is because, given any $\epsilon > 0$,

$$P(|\bar{X} - \mu| \geq \epsilon) \leq \frac{var(\bar{X})}{\epsilon^2} = \frac{\sigma^2}{\epsilon^2}\frac{1}{n} \longrightarrow 0.$$

Here we have used the Chebyshev's inequality.

This argument shows that for any model, with μ denoting the population mean, $\bar{X}$ is a consistent estimator of μ. This depends only on the fact that the variance of $\bar{X}$ is σ^2/n where σ^2 is population variance, assumed to be finite. In fact the same argument gives us a general fact: If Y_n is an unbiased estimator (based on a sample of size n) of θ and if $var(Y_n) \longrightarrow 0$, then Y_n is a consistent estimator of θ.

If Y is unbiased estimator of θ, then, of course, $E(Y - \theta)^2$ is nothing but the variance of Y. However, if Y is not unbiased, then this is no longer the variance of Y. This quantity $E(Y - \theta)^2$ is called the **Mean Square Error (MSE)**. An estimator is said to have **minimum mean square error** if this quantity is the least possible. When the estimator is unbiased, then mean square error being its variance, an unbiased estimator with minimum mean square error is called **Minimum Variance Unbiased Estimator (MVUE)**. $\bar{X}$ has minimum variance among all estimators which are linear combinations of $X_1, \ldots, X_n$.

4. Confidence intervals.

Point estimators are not always perfect. We wish to quantify the accuracy of the estimator. One way to measure the accuracy is to see its variance. The smaller the variance, the better it is. But there is a fundamentally different method of looking at the problem of estimation. Instead of saying that a number is an estimator of the parameter μ, why not prescribe an interval and quantify by saying that the parameter lies in this interval with a certain probability which is high. This leads to the notion of confidence intervals.

Let us start with our normal example for LF. We know that $\bar{X}$ is an unbiased estimator of μ, its variance is σ^2/n and in fact $\bar{X} \sim N(\mu, \frac{\sigma^2}{n})$. As a result,

$$\frac{\bar{X}-\mu}{\sigma/\sqrt{n}} \sim N(0,1).$$

If $Z \sim N(0,1)$, then $P(-1.96 < Z < 1.96) = 0.95$, so that

$$P\Big(-1.96 < \frac{\bar{X}-\mu}{\sigma/\sqrt{n}} < 1.96\Big) = 0.95.$$

The above inequality can be restated as

$$P\Big(\bar{X} - 1.96\frac{\sigma}{\sqrt{n}} < \mu < \bar{X} + 1.96\frac{\sigma}{\sqrt{n}}\Big) = 0.95.$$

The probability that the interval

$$\Big(\bar{X} - 1.96\frac{\sigma}{\sqrt{n}}\ , \bar{X} + 1.96\frac{\sigma}{\sqrt{n}}\Big)$$

"captures" μ is 0.95. This interval is called a 95% **confidence interval** for μ. It is a plausible range of values for μ together with a quantifiable measure of its plausibility.

A confidence interval is a *random* interval; it changes as the collected data changes. This explains why we say "**a** 95% confidence interval" rather than "the 95% confidence interval". We chose the "cutoff limits" ±1.96 symmetrically around 0 to minimize the length of the confidence interval. "cutoff limits" are also called "percentage points".

Example (devised from van den Bergh, 1985): $n = 148$ Galactic globular clusters. $\bar{x} = -7.1$ mag. We assume that $\sigma = 1.2$ mag. Let M_0 be the population mean visual absolute magnitude. A 95% confidence interval for M_0 is

$$\Big(\bar{x} - 1.96\frac{\sigma}{\sqrt{n}}\ , \bar{x} + 1.96\frac{\sigma}{\sqrt{n}}\Big) = \Big(-7.1 - 1.96\frac{1.2}{\sqrt{148}}, -7.1 + 1.96\frac{1.2}{\sqrt{148}}\Big).$$

Thus, (-7.1 ∓ 0.193) is a plausible range of values for M_0.

Warning: Don't bet your life that your 95% confidence interval has captured μ! There is a chance (5%) of it not capturing. Should we derive intervals with higher levels of confidence, 96%, 98%, 99%? Return to the tables of the $N(0,1)$ distribution and observe that $P(-2.33 < Z < 2.33) = 0.98$. Repeat the earlier arguments. Assuming that σ is known,

$$P\Big(-2.33 < \frac{\bar{X}-\mu}{\sigma/\sqrt{n}} < 2.33\Big) = 0.98.$$

leading to a 98% confidence interval,

$$\Big(\bar{X} - 2.33\frac{\sigma}{\sqrt{n}}\ , \bar{X} + 2.33\frac{\sigma}{\sqrt{n}}\Big).$$

If σ is unknown then the method outlined above for getting confidence intervals does not work. A basic principle in statistics is: *Replace any unknown parameter with a good estimator.* Consider the LF data problem. We have a random sample $X_1, \dots, X_n$ drawn from $N(\mu, \sigma^2)$. We want to construct confidence interval for μ using the statistic $\frac{\bar{X}-\mu}{S/\sqrt{n}}$. To repeat the above method,we need the sampling distribution of this statistic. It is not normally distributed.

The t-distribution: If $X_1, \dots, X_n$ is a random sample drawn from $N(\mu, \sigma^2)$ then the statistic

$$T = \frac{\bar{X} - \mu}{S/\sqrt{n}}$$

has a t-distribution with $n - 1$ degrees of freedom. Once this is granted, we construct confidence intervals as before. Suppose that $n = 16$, then see the tables of the t-distribution with 15 degrees of freedom.

$$P(-2.131 < T_{15} < 2.131) = 0.95.$$

Therefore

$$P\Big(- 2.131 < \frac{\bar{X} - \mu}{S/\sqrt{n}} < 2.131\Big) = 0.95.$$

Thus, a 95% confidence interval for μ is

$$\Big(\bar{X} - 2.131\frac{S}{\sqrt{n}}, \bar{X} + 2.131\frac{S}{\sqrt{n}}\Big).$$

For example, with $n = 16$, $\bar{x} = -7.1$ mag, $s = 1.1$ mag, a 95% confidence interval for μ is -7.1 ∓ 0.586. If you are curious about the t-density, here it is for p degrees of freedom.

$$f(t) = \frac{\Gamma((p+1)/2)}{\Gamma(p/2)}\ \frac{1}{\sqrt{p\pi}}\ \left(1 + \frac{t^2}{p}\right)^{-(p+1)/2} \qquad - \infty < t < \infty.$$

The χ^2-distribution: So far we have been considering confidence intervals for μ. Let us now discuss confidence intervals for σ based on a random sample $X_1, \dots, X_n$, under normal model. We know S^2 is an unbiased and consistent estimator of σ^2. What is the sampling distribution of S^2? The statistic $(n-1)S^2/\sigma^2$ has a *chi-squared* χ^2 distribution with $n - 1$ degrees of freedom. We now construct confidence intervals as before. Consult the tables of the χ^2 distribution. Find the percentage points, and solve the various inequalities for σ^2. Denote the percentage points by a and b.

$$P(a < \chi^2_{n-1} < b) = 0.95.$$

We find a, b using tables of the χ^2 distribution. Usually, this is done by choosing a so that $P(\chi^2_{n-1} < a) = .025$ and $P(\chi^2_{n-1} > b) = .025$. Solve for σ^2 the inequalities: $a < \frac{(n-1)S^2}{\sigma^2} < b$. A 95% confidence interval for σ^2 is

$$\Big(\frac{(n-1)S^2}{b}, \frac{(n-1)S^2}{a}\Big)$$

For example, if $n = 16$, $s = 1.2$ mag, percentage points from the χ^2 tables (with 15 degrees of freedom) are 6.262 and 27.49. Hence a 95% confidence interval for σ^2 is

$$\Big(\frac{15 \times (1.2)^2}{27.49}, \frac{15 \times (1.2)^2}{6.262}\Big) = (0.786, 3.449).$$

If you are curious about the χ^2 density, here it is for p degrees of freedom.

$$f(x) = \frac{2^{-p/2}}{\Gamma(p/2)}\, e^{-x/2}\, x^{\frac{p}{2}-1} \qquad x > 0.$$

If we want a greater the level of confidence, the confidence interval will, in general, be longer. The larger the sample size, the shorter will be the confidence interval. How do we choose n? In our 95% confidence intervals for μ, the term $1.96\sigma/\sqrt{n}$ is called the **margin of error**. We choose n to have a desired margin of error. To have a margin of error of 0.01 mag, we choose n so that

$$\frac{1.96\sigma}{\sqrt{n}} = 0.01, \quad \text{that is,} \quad n = \Big(\frac{1.96\sigma}{0.01}\Big)^2.$$

A very interesting question arises now. Could we get the above confidence interval for μ only because we assumed a normal model? On the face of it this seems so, because we used the fact that a certain statistic is normal. There is indeed more to this construction. Here is a **modified Central Limit Theorem** that will help us. Let $X_1, \ldots, X_n$ be a random sample; μ be the population mean; $\bar{X}$ be the sample mean and S be the sample standard deviation. If n is large, then

$$\frac{\bar{X} - \mu}{S/\sqrt{n}} \approx N(0, 1).$$

In other words, for large values of n, the probability that the above statistic lies between a and b is same as the corresponding area under the standard normal curve. The conclusion does not depend on the population probability distribution of the X_i. As long as the population mean and variance are finite, this will hold. Instead of the exact sampling distribution used earlier, we can use this approximate distribution to construct confidence intervals. The resulting confidence intervals for μ also do not depend on the population probability distribution. Several papers on LF for globular clusters have large sample sizes like 300, 1000, etc.

5. Testing of Hypotheses.

A LF researcher believes that $M_0 = -7.7$ mag for the M31 globular clusters. The researcher collects data — using a *random sample* — from M31. A natural question: "Are the data strongly in support of the claim that $M_0 = -7.7$ mag?"

A **statistical hypothesis** is a statement about the parameters of the population. A **statistical test of significance** is a procedure for comparing observed data with a hypothesis whose plausibility is to be assessed. The **null hypothesis** is the statement being tested, usually denoted by H_0. The **alternative hypothesis** is a competing statement, usually denoted by H_a. In general, the alternative hypothesis is chosen as the statement for which there is likely to be supporting evidence. In the case of our M31 LF researcher, the null hypothesis is H_0: $M_0 = -7.7$. An alternative hypothesis is H_a: $M_0 \neq -7.7$. This is an example of a **two-sided** alternative hypothesis. If we have reasons to believe that M_0 can not be above -7.7, then we should make the alternative hypothesis **one sided**, namely, H_a: $M_0 < -7.7$.

The basic idea in devising a test is the following. Based on the observations, we calculate a specially chosen informative statistic. See which of the hypotheses makes the observed value of this chosen statistic more plausible. First, some terminology is needed. A **test statistic** is a statistic that will be calculated from the observed data. This will measure the compatibility of H_0 with the observed data. It will have a sampling distribution free of unknown parameters (under the null hypothesis). A **rejection rule** is a rule which specifies the values of the test statistic for which we reject H_0. Here is an illustration.

Example: A random sample of 64 measurements has mean $\bar{x} = 5.2$ and standard deviation $s = 1.1$. Test the null hypothesis $H_0 : \mu = 4.9$ against the alternative hypothesis $H_a : \mu \neq 4.9$

1. The null and alternative hypotheses are $H_0 : \mu = 4.9, \quad H_a : \mu \neq 4.9$.

2. The test statistic is $T = \frac{\bar{X}-4.9}{S/\sqrt{n}}$.

3. The distribution of the test statistic T, under the assumption that H_0 is valid, is $\approx N(0,1)$.

4. The rejection rule: Reject H_0 if $|T| > 1.96$, the upper 95 percentage point in the tables of the standard normal distribution. Otherwise, we *fail to reject* H_0.

This cutoff point 1.96 is also called a **critical value**. This choice of critical value results in a 5% **level of significance** of the test of hypotheses. This mean that that there is a 5% chance of our rejecting hypothesis H_0, when it is actually true.

5. Calculate the value of the test statistic. It is

$$\frac{\bar{x}-4.9}{s/\sqrt{n}} = \frac{5.2-4.9}{1.1/\sqrt{64}} = 2.18$$

6. Decision: We reject H_0; the calculated value of the test statistic exceeds the critical value, 1.96.

We report that the statistic is **significant**. There is a **statistically significant** difference between the population mean and the hypothesized value of 4.9.

7. The P-value of the test is the smallest significance level at which the statistic is significant.

6. Return to χ^2.

We briefly encountered χ^2 in discussing confidence intervals for σ^2. We now discuss a little more of this. This arises in both testing **goodness of fit**, and also in estimation.

We first start with testing problem. This is best explained with a discrete model. Suppose that you have a random variable X that takes r values $a_1, \cdots, a_r$. Someone proposes a hypothesis that for each i the chance of value a_i is p_i. Here $p_i > 0$ for all i and $\sum p_i = 1$. How do we test this? Make n independent observations of X and suppose that in your data the value a_i appears n_i times. Of course $\sum n_i = n$. If the hypothesis is correct, we *expect* to see the value a_i approximately np_i many times. So the discrepancy relative to our expectation is $(n_i - np_i)^2/(np_i)$ and the total discrepancy is

$$\sum_1^r \frac{(n_i - np_i)^2}{np_i}$$

and this is named as the χ^2 value for the data. This statistic is called χ^2 statistic. It can be shown that for large n, this statistic indeed has a χ^2 distribution with $(r-1)$ degrees of freedom. This fact can be used to test whether the proposed hypothesis is plausible — large values of this statistic being not in favour of the hypothesis.

Now We turn to an important method of estimation. As in the earlier para, assume that X takes r values $a_1, \cdots, a_r$. Let $P(X = a_i) = p_i(\theta)$, that is, the probability depends on a parameter θ. Once θ is found out, the value p_i is known. How do we estimate θ? Here is a way to do it, choose that value of θ which minimizes the discrepancy. In other words, choose that value of θ for which

$$\chi^2(\theta) = \sum_1^r \frac{(n_i - np_i)^2}{np_i}$$

is minimum. Note that this is **not** a statistic, it depends on the parameter θ. You use calculus, differentiate w.r.t. θ, remember p_i are functions of θ. You end up solving

$$\sum_1^r \left(\frac{n_i - np_i(\theta)}{p_i(\theta)} + \frac{(n_i - np_i(\theta))^2}{2np_i^2(\theta)} \right) \frac{dp_i(\theta)}{d\theta} = 0.$$

This is called the **minimum χ^2 method** of estimation. Unfortunately, the presence of θ in the denominator makes things messy. So one uses the **modified minimum χ^2 method** where, one ignores the second term in the above equation.

If the model is continuous and not discrete, one groups the observations and proceeds. We shall not go into the details.

7. Truncation.

Sometimes we need to use truncated distributions for modeling. As an example, suppose that in the LF study, we believe that there is an absolute magnitude limit. Say, we believe that the magnitude can not be above M^*. Then we should not model the LF data with normal distribution. We should use the truncated normal.

$$f(x;\mu,\sigma^2) = \begin{cases} \frac{C}{\sqrt{2\pi}\sigma} \exp\left[-\frac{(x-\mu)^2}{2\sigma^2}\right], & \text{if} \quad x \le M^* \\ 0, & \text{if} \quad x > M^* \end{cases}$$

where the constant C is so chosen as to make the area under the curve unity. See Garcia-Munoz, et al. "The relative abundances of the elements silicon through nickel in the low energy galactic cosmic rays," In: Proc. Int'l. Cosmic Ray Conference, Plovdiv, Bulgaria, 1977, Conference Papers. Volume 1. Sofia, B'lgarska Akademiia na Naukite, 1978, p. 224-229.

As another example, consider, Protheroe, et al. "Interpretation of cosmic ray composition - The path length distribution," ApJ., 247 1981. If our instruments can not detect rays with path length below a certain value, then our observations will not be a random sample from the exponential population. Rather, they would only be a sample from the truncated exponential, namely,

$$f(x;\theta_1,\theta_2) = \begin{cases} \theta_1^{-1} \exp[-(x-\theta_2)/\theta_1], & \text{if} \quad x \ge \theta_2 \\ 0, & \text{if} \quad x < \theta_2 \end{cases}$$

Here we have two parameters $\theta_1 > 0$ and $\theta_2 > 0$.