Caltech Department of Mathematics

Ma 3/103
Introduction to Probability and Statistics

KC Border
Winter 2017

Lecture 19: Estimation II

Relevant textbook passages:

Larsen–Marx [1]: Sections 5.2–5.7

19.1 The method of moments

Larsen–Marx [1]: pp. 293–296

Let $X_1, \dots, X_n$ be independent and identically distributed with density $f(x; \theta_1, \dots, \theta_m)$. Then the k^{th} sample moment is

$$\frac{\sum_{i=1}^n x_i^k}{n}.$$

The distribution's k^{th} moment is

$$\int x^k f(x; \theta_1, \dots, \theta_m)\, dx.$$

Solving for the $(\hat\theta_1, \dots, \hat\theta_m)$ that equates the first m moments is called the **method of moments**.

$$\int x^k f(x; \hat\theta_1, \dots, \hat\theta_m)\, dx = \frac{\sum_{i=1}^n x_i^k}{n} \qquad (k = 1, \dots, m).$$

Larsen–Marx [1]: Example 5.2.5, pp. 294–295

19.1.1 Example (Method of moments and the Gamma distribution) Recall that the Gamma(r, λ) distribution ($r > 0$, $\lambda > 0$) has density given by

$$f(t) = \frac{\lambda^r}{\Gamma(r)} t^{r-1} e^{-\lambda t} \qquad (t > 0).$$

The parameter r is the **shape parameter**, and λ is the **scale parameter**. The mean and variance of a Gamma(r, λ) random variable are given by

$$\boldsymbol{E}\, X = \frac{r}{\lambda}, \qquad \boldsymbol{Var}\, X = \frac{r}{\lambda^2}.$$

It is difficult to derive closed form expressions for the MLE of a Gamma, because the gamma function $\Gamma(r)$ does not have a closed form expression. But it is straightforward to derive the method of moments estimators for r and λ.

Using the fact that for any random variable X, we have $E(X^2) = \boldsymbol{Var}\, X + (\boldsymbol{E}\, X)^2$ (see Section 6.10), given a sample $x_1, \dots, x_n$ of n independent draws from a Gamma, we just need to solve the two equations

$$\bar x = \frac{\sum_i x_i}{n} = \frac{r}{\lambda}, \qquad \frac{\sum_i x_i^2}{n} = \frac{r}{\lambda^2} + \left(\frac{r}{\lambda}\right)^2 = \frac{r(r+1)}{\lambda^2}.$$

The solution is gotten by solving the first for $r = \lambda \bar x$ and substituting that into the second to get

$$\begin{aligned} \frac{\sum_i x_i^2}{n} &= \frac{(\lambda \bar x)(1 + \lambda \bar x)}{\lambda^2} \\ &= \frac{\lambda \bar x + \lambda^2 \bar x^2}{\lambda^2} \\ &= \frac{\bar x}{\lambda} + \bar x^2 \end{aligned}$$

so

$$\hat\lambda = \frac{\bar x}{\sum_i x_i^2/n - \bar x^2},$$

and

$$\hat r = \hat\lambda \bar x.$$

□

19.1.2 Example (Method of moments and the Normal distribution) The Normal(μ, σ^2) has mean μ and variance σ^2, so the method of moments estimators solve

$$\hat\mu = \bar x, \qquad , \sum_i x_i^2/n = \widehat{\sigma^2} + \hat\mu^2.$$

Solving gives

$$\hat\mu = \bar x, \qquad \widehat{\sigma^2} = \sum_i x_i^2/n - \bar x^2 = \sum_i (x_i - \bar x)^2/n.$$

The moments estimators for μ and σ^2 are the same as the maximum likelihood estimators. □

19.2 Other ways to generate estimators

Most other general methods for finding estimators involve some sort of maximization or minimization. For instance, there are minimum χ^2 estimators, that frequently have nice properties. Mosteller's [3] analysis of the World Series considers minimum χ^2 estimation in addition to MLE. I'll describe this kind of estimation later on, when we discuss χ^2 tests.

Most general methods for generating estimators involve choosing a method of either similarity or distance between the observed data and the data that might have been generated by the dgp with a given parameter. There are deep reasons why such estimators have good properties, but that's a topic for a more advanced course.

19.3 Digression: The quantiles z_α

Larsen–Marx [1]: p. 307

Statisticians have adopted the following special notation. Let Z be a Standard Normal random variable, with cumulative distribution function denoted Φ. For $0 < \alpha < 1$, define z_α by

$$P(Z > z_\alpha) = \alpha$$

or equivalently

$$P(Z \leqslant z_\alpha) = 1 - \alpha.$$

Then

$$\boxed{z_\alpha = \Phi^{-1}(1 - \alpha)}$$

This is something you can look up with R or Mathematica's built-in quantile functions. (Remember the **quantile function** is Φ^{-1}.) By symmetry,

$$P(Z < -z_\alpha) = \alpha \qquad \text{and} \qquad P(|Z| > z_\alpha) = 2\alpha$$

so

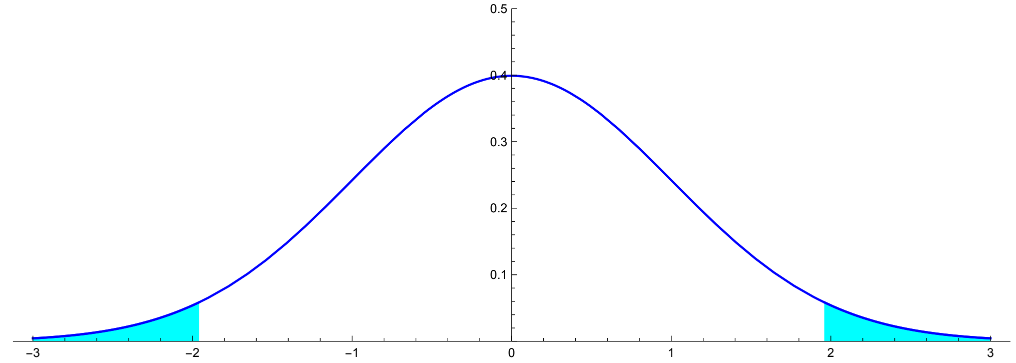
$$P\,(-z_\alpha \leqslant Z \leqslant z_\alpha) = 1 - 2\alpha.$$

The last inequality is often expressed as

$$P\left(-z_{\alpha/2} \leqslant Z \leqslant z_{\alpha/2}\right) = 1 - \alpha.$$

Here are some commonly used values of α and the corresponding z_α to two decimal places.

α	z_α	$1-2\alpha$
0.1	1.28	0.80
0.05	1.64	0.90
0.025	1.96	0.95
0.01	2.33	0.98
0.005	2.58	0.99



This shaded area is the probability of the event $(|Z| > 1.96)$, which is equal to 0.05. Values outside the interval $(-1.96, 1.96)$ are often regarded as unlikely to have occurred "by chance."

19.4 Confidence intervals for Normal means if σ is known

So far we have looked at **point estimates**, and barely made a dent in the subject. (Erich L. Lehmann's classic *Theory of Point Estimation* [2] runs to about 500 pages.) But it is time to move on.

Interval estimates are closely related to hypothesis testing (coming up soon) and are sometime more useful than point estimates.

Go back to the Normal estimation case. The maximum likelihood estimator $\hat\mu_{\text{MLE}}$ of the mean μ is just the sample mean $\bar{x} = \sum_i x_i/n$, but how "good" is that estimate? If $X_1, \dots, X_n$ are independent and identically distributed $N(\mu, \sigma^2)$, then

$$\hat\mu_{\text{MLE}} = \frac{X_1 + \cdots + X_n}{n} \sim N(\mu, \sigma^2/n),$$

so by standardizing $\hat\mu$ we have

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

We have just seen that

$$z_{0.025} = 1.96.$$

Therefore

$$P\left(-1.96 \leqslant \frac{\hat{\mu}-\mu}{\sigma/\sqrt{n}} \leqslant 1.96 \right) = 0.95$$

But this event is also equal to the event

$$\left(\frac{\hat{\mu}-1.96\sigma}{\sqrt{n}} \leqslant \mu \leqslant \frac{\hat{\mu}+1.96\sigma}{\sqrt{n}} \right).$$

So another way to interpret this is

$$P\left(\mu \in [\hat{\mu}-1.96\sigma/\sqrt{n},\ \hat{\mu}+1.96\sigma/\sqrt{n}]\right) = 95\%$$

even though μ is not random. The interval

$$I = [\hat{\mu}-1.96\sigma/\sqrt{n},\ \hat{\mu}+1.96\sigma/\sqrt{n}]$$

is called a 95% **confidence interval** for μ. More generally we have the following

To get a $1-\alpha$ **confidence interval** for μ when σ is known, set

$$I = \left[\hat{\mu} - \frac{z_{\alpha/2}\sigma}{\sqrt{n}},\ \hat{\mu} + \frac{z_{\alpha/2}\sigma}{\sqrt{n}}\right]. \tag{1}$$

Then

$$P\left(\mu \in I\right) = 1-\alpha.$$

19.4.1 Interpreting confidence intervals

Remember that μ is not random, rather the interval $I(X) = [\hat{\mu}-1.96\sigma/\sqrt{n},\ \hat{\mu}+1.96\sigma/\sqrt{n}]$ is random, since it is based on the random $\hat{\mu}$. But once I calculate I, μ either belongs to I or it doesn't, so what am I to make of the 95% probability? I think the way to think about it is this:

No matter what the values of μ and σ are, following the procedure "draw a sample X from the distribution $N(\mu,\sigma^2)$, and use (1) to calculate the interval $I(X)$," the interval $I(X)$ will then have a 95% probability of containing μ.

This is not the same as saying, I used (1) to calculate the interval I, so no matter what the values of μ and σ are, the interval I has a 95% probability of containing μ.

It is the *procedure*, not the interval per se, that gives us the confidence.

Figure 19.1 shows the result of using this procedure 100 times to construct a symmetric 95% confidence interval for μ, based on (pseudo-)random samples of size 5 drawn from a standard normal distribution. Note that in this instance, 5 of the 100 intervals missed the true mean 0.

19.4.2 Hold on

But wait! The confidence interval given by (1) depends on σ. What if we don't know σ? We can use $\hat{\sigma}$ to estimate σ to get a confidence interval. The catch is that $\frac{\hat{\mu}-\mu}{\hat{\sigma}/\sqrt{n}}$ is *not* a Standard Normal random variable. Instead it has a "Student t" distribution. We will discuss this later in Lecture 21, sections 21.6 and 21.7.

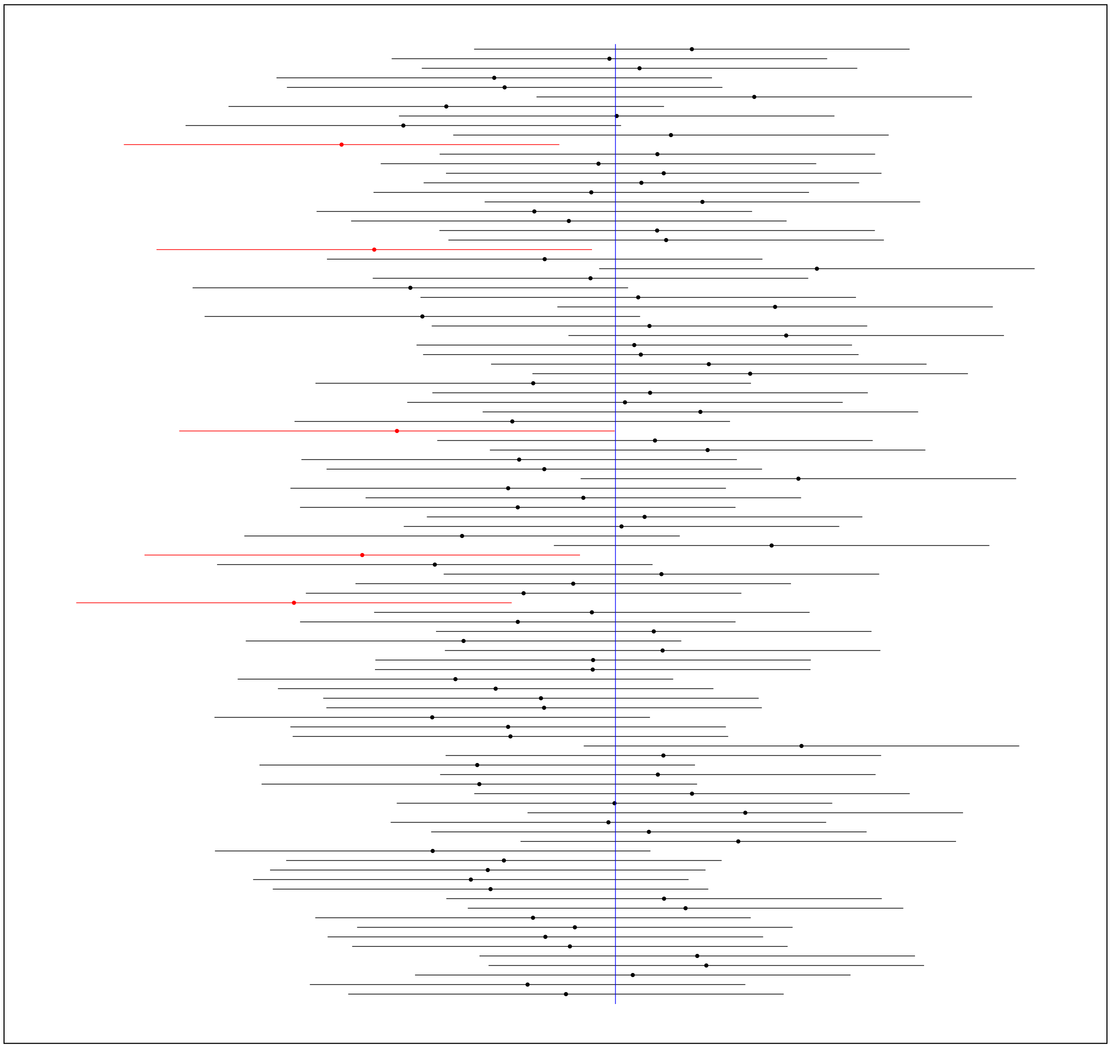


Figure 19.1. Here are one hundred 95% confidence intervals for the mean from a Monte Carlo simulation of a sample of size 5 independent standard normals. The intervals that do not include the true mean 0 are shown in red.

You might ask, when might I know σ, but not know μ? Maybe in a case like this: I can imagine the variance in a measurement of weight using a balance beam scale depends on the friction in the balance bearing. I can also imagine that the mean measurement of a sample's mass depends on the sample's actual mass. I might have a lot of experience with this particular of scale, so that I know the variance σ, but the mean of the measurement depends on which sample I am weighing. To get a good estimate of the weight, I might make several measurements,[1] and I could then use this procedure to generate a confidence interval. (I just made this up, and it sounds plausible, but do any of you chemists or engineers have any real information on such scales?)

19.5 Considerations in constructing confidence intervals

There are two more points worth noting.

- Suppose we know μ, and we want to choose an interval I so that the standard normal random variable $Z = \frac{\hat{\mu}-\mu}{\sigma/\sqrt{n}}$ lies in I with probability $1-\alpha$. Any interval $[a,b]$ satisfying $\int_a^b \frac{1}{2\pi} e^{-z^2/2}\,dz = 1-\alpha$ has this property.

Because of the symmetry of the normal distribution, the symmetric interval $[-\frac{z_{\alpha/2}\sigma}{\sqrt{n}}, \frac{z_{\alpha/2}\sigma}{\sqrt{n}}]$ is the *shortest* such interval.

- Because of the properties of the standard normal distribution, the length of the interval $\left[\hat{\mu} - \frac{z_{\alpha/2}\sigma}{\sqrt{n}}, \hat{\mu} + \frac{z_{\alpha/2}\sigma}{\sqrt{n}}\right]$ does not depend on μ.

- For distributions that are not symmetric, you may want to construct asymmetric confidence intervals. I can think of at least two principles you could use.

1. Choose the *shortest* interval $[a,b]$ containing your point MLE $\hat{\theta}$ that has $P_{\hat{\theta}}\big([a,b]\big) = 1-\alpha$. This would be the interval where the likelihood (= density) is highest. Since $\hat{\theta}$ maximizes the likelihood, we know it will be in the interval.

Oops. How do we know that an interval is the shortest set? Maybe we would be better off taking two short intervals instead one long one. For unimodal (single-peaked) densities, this won't happen.

2. The other principle you might consider is to choose an interval $[a,b]$ so that $P(\theta < a) = P(\theta > b) = \alpha/2$, bearing in mind the above interpretation of the probability.

In the normal case, these two principles are not in conflict and procedure for constructing the interval described above is consistent with both.

Bibliography

[1] R. J. Larsen and M. L. Marx. 2012. *An introduction to mathematical statistics and its applications*, fifth ed. Boston: Prentice Hall.

[2] E. L. Lehmann. 1983. *Theory of point estimation*. Wiley Series in Probability and Mathematical Statistics. New York: John Wiley and Sons.

[3] F. Mosteller. 1952. The world series competition. *Journal of the American Statistical Association* 47(259):355–380. http://www.jstor.org/stable/2281309

[1] My grandfather was a carpenter, so I am quite familiar with the old saw, "Measure twice, cut once." (Sorry, I couldn't help myself.)