STA 580 — Spring 2009 — Dr. Charnigo

Lecture 4

Introduction. Lecture 1 focused on numerical and graphical techniques for summarizing sample values $x_1, \ldots, x_n$. Lectures 2 and 3 then developed a probabilistic framework, central to which was the concept of a random variable. I stated in Lecture 3 that sample values could be viewed as realizations of random variables $X_1, \ldots, X_n$. In particular, I claimed that understanding the behavior of random quantities like $\bar{X}$ could help us to make inferences about a population. That is our agenda for today — and the rest of the semester!

Population parameters and normality

Scenario. Suppose that we have identified a population and a specific attribute of interest; further, assume that the attribute can be described numerically. For instance, we may be interested in the population of all hospital employees and the attribute of serum cholesterol reduction after one month on a vegetarian diet (Cf. Table 2.13, which presents results for a sample but not the full population). Or we may be interested in the population of all HIV-positive hemophiliacs and the attribute of latency time to AIDS (Cf. Table 6.12).

Population mean. The population mean is denoted μ and represents the average value for the attribute within the population. The population mean μ should be contrasted with the sample mean $\bar{x}$, which represents the average value for the attribute within a sample.

If the numerical values $x_1, \ldots, x_n$ have arisen through a sampling process in which any group of n people in the population had the same chance of being selected (i.e., we have a "simple random sample"), then the random analogues $X_1, \ldots, X_n$ satisfy $E[X_1] = \cdots = E[X_n] = \mu$. That is, $X_1, \ldots, X_n$ inherit their mean from the population.

Population variance. The population variance is denoted σ^2 and represents the average squared deviation from the mean within the population. The population variance σ^2 should be contrasted with the sample variance s^2, which represents the average squared deviation from the mean within a sample, apart from the presence of $(n-1)$ in the denominator rather than n.

If we have a simple random sample, then the random analogues $X_1, \ldots, X_n$ satisfy $Var[X_1] = \cdots = Var[X_n] = \sigma^2$. That is, $X_1, \ldots, X_n$ also inherit their variance from the population.

A normal population. We say that a population with mean μ and variance σ^2 is normal if the fraction of measurements in the population less than x equals $\Phi\left(\frac{x-\mu}{\sigma}\right)$ for any number x.

If we have a simple random sample from a normal population, then the random analogues $X_1, \ldots, X_n$ have a normal distribution. In fact, some people take this as a working definition: a population is normal if the random analogues of sample values are normally distributed.

A key idea. We do not know what the population mean μ and the population variance σ^2 are. We wish to make inferences about them based on the sample data. In particular, we can use the sample mean $\bar{x}$ and the sample variance s^2 as guesses for the population mean μ and the population variance σ^2, but we also need to know how "good" those guesses are.

Estimating a population mean

Point estimation. Let $X_1, \ldots, X_n$ be the random analogues of numerical values from a simple random sample. I noted in Lecture 3 that $E[\bar{X}] = \mu$, $Var[\bar{X}] = \sigma^2/n$, and $\bar{X}$ is approximately normally distributed for large n. Together, these facts imply that the random quantity $\bar{X}$ is likely to be close to μ if n is large. Hence, once we have selected the sample and obtained the numerical values $x_1, \ldots, x_n$, we can be rather confident that the number $\bar{x}$ is close to μ if n is large.

Note my use of the word "confident" above rather than the word "likely". Although this may seem to be an issue of semantics, we must remember that nothing is random about $\bar{x}$. The randomness disappeared at the moment n specific people were selected for the sample; in effect, the random quantity $\bar{X}$ became the number $\bar{x}$.

We refer to $\bar{x}$ as a "point estimate" or, more simply, an "estimate" of μ. While $\bar{x}$ is not the only possible estimate, it is an estimate with two desirable properties. First, its random analogue $\bar{X}$ has expected value equal to the population parameter that we are trying to estimate, so there is no systematic tendency towards overestimation or underestimation across repeated sampling. Second, in many situations $\bar{X}$ has a smaller variance than the random analogue of any competing estimate, so $\bar{X}$ is more likely to be close to μ.

Interval estimation, large sample. Besides furnishing a point estimate for μ, we can construct an "interval estimate" or a "confidence interval". A confidence interval is a range of plausible values for μ based on the sample data. In essence, a confidence interval conveys how "good" the point estimate is by revealing how far away you can move from the point estimate before arriving at an implausible value for μ.

Assume that n is large. Then, even if the population is not normal, $\bar{X}$ is approximately normal with mean μ and standard deviation $\sigma/\sqrt{n}$, so that

$$Z := \frac{\bar{X} - \mu}{\sigma/\sqrt{n}}$$

is approximately standard normal (Equation 6.4). Hence, we have

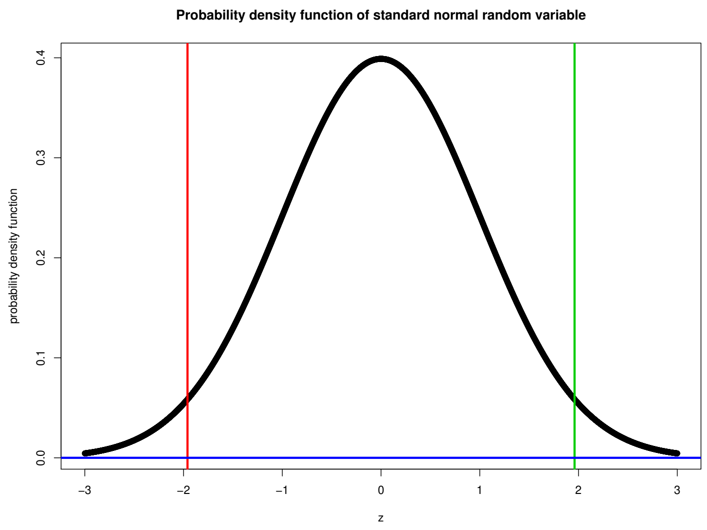
$$P(-z_{1-\alpha/2} \leq Z \leq z_{1-\alpha/2}) \approx 1 - \alpha$$

for any small positive α. For example, with $\alpha = 0.05$ we have

$$P(-z_{0.975} \leq Z \leq z_{0.975}) = P(-1.96 \leq Z \leq 1.96) \approx 0.95.$$

This is illustrated in Figure 1 below.

Figure 1:



The area bounded above by the standard normal probability density function, bounded below by the blue horizontal line at 0, bounded on the left by the red vertical line at $a = -1.96$, and bounded on the right by the green vertical line at $b = 1.96$, is $P(-1.96 < Z \leq 1.96) \approx 0.95$.

Algebraic rearrangement of $-z_{1-\alpha/2} \leq Z \leq z_{1-\alpha/2}$ leads to $\bar{X} - z_{1-\alpha/2}\, \sigma/\sqrt{n} \leq \mu \leq \bar{X} + z_{1-\alpha/2}\, \sigma/\sqrt{n}$. Hence,

$$P(\bar{X} - z_{1-\alpha/2}\, \sigma/\sqrt{n} \leq \mu \leq \bar{X} + z_{1-\alpha/2}\, \sigma/\sqrt{n}) \approx 1 - \alpha.$$

So, for example,

$$P(\bar{X} - 1.96\, \sigma/\sqrt{n} \leq \mu \leq \bar{X} + 1.96\, \sigma/\sqrt{n}) \approx 0.95.$$

We refer to

$$\bar{x} - z_{1-\alpha/2}\, \sigma/\sqrt{n} \ \text{ to } \ \bar{x} + z_{1-\alpha/2}\, \sigma/\sqrt{n}$$

as a $100(1-\alpha)\%$ confidence interval for μ. Thus,

$$\bar{x} - 1.96\, \sigma/\sqrt{n} \ \text{ to } \ \bar{x} + 1.96\, \sigma/\sqrt{n}$$

is a 95% confidence interval for μ.

Unfortunately, the above confidence interval depends on σ^2, which is typically unknown. However, we have available an estimate of σ^2. This estimate is s^2. With large n, we anticipate that s^2 is close to σ^2. So we simply replace σ^2 by s^2. This yields (Equation 6.7)

$$\bar{x} - z_{1-\alpha/2}\, s/\sqrt{n} \ \text{ to } \ \bar{x} + z_{1-\alpha/2}\, s/\sqrt{n},$$

which we still refer to as a $100(1-\alpha)\%$ confidence interval.

As an aside, I can now tell you why we divide by $(n-1)$ instead of n when computing the sample variance. The random analogue to the sample variance, $S^2 := \sum_{i=1}^{n}(X_i - \bar{X})^2/(n-1)$, has expected value σ^2. Hence, the division by $(n-1)$ ensures that we avoid systematic underestimation of σ^2 across repeated sampling (Equation 6.10).

How large is large? Your textbook author feels that Equation 6.7 should be used only when $n > 200$. Others feel that Equation 6.7 can be used when $n \geq 30$. The issue is not so much whether the Central Limit Theorem can be invoked when $30 \leq n \leq 200$ (most people feel that it can) but whether s^2 is a satisfactory substitute for σ^2 when $30 \leq n \leq 200$ (this is open to debate). My opinions are as follows: If $30 \leq n \leq 200$ and you have a normal population, then to be safe just use the small-sample procedure (described later). However, if $30 \leq n \leq 200$ and you do not have a normal population, then using Equation 6.7 is more easily justified than using the small-sample procedure.

Example (interval estimation, large sample). Refer to "Infectious Disease" on page 222. Let μ denote the average time to onset of AIDS following seroconversion. We have $n = 287$, $\bar{x} = 5.24$, $s^2 = 3.60$, and $s = \sqrt{3.60} = 1.90$. Also, with $\alpha = 0.05$, we have $z_{1-\alpha/2} = z_{0.975} = 1.96$. Thus, a 95% confidence interval for μ is

$$5.24 - 1.96 \times 1.90/\sqrt{287} \ \text{ to } \ 5.24 + 1.96 \times 1.90/\sqrt{287}.$$

Numerically this simplifies to $[5.02, 5.46]$.

Such computations can also be performed with SAS. The first four entries on page 1 of {TABLE612.pdf} show the sample size, lower boundary of a 95% confidence interval for μ, point estimate for μ, and upper boundary of a 95% confidence interval for μ. Actually, SAS employs the small-sample procedure regardless of n, but the difference in results is negligible when n is as large as 287.

Interval estimation, small sample. If n is small, we cannot simply substitute s^2 for σ^2 because we do not have enough data to estimate σ^2 reliably. In particular, we must make an additional adjustment to Equation 6.7. Instead of using $z_{1-\alpha/2}$ to construct the confidence interval, we must use $t_{n-1,1-\alpha/2}$.

What is $t_{n-1,1-\alpha/2}$? Just as $z_{1-\alpha/2}$ is the $1-\alpha/2$ quantile of a standard normal distribution, which can be recovered from Table 3 in the back of your book, $t_{n-1,1-\alpha/2}$ is the $1-\alpha/2$ quantile of another probability distribution called the T distribution on $(n-1)$ degrees of freedom. For small n and selected α, you can recover $t_{n-1,1-\alpha/2}$ from Table 5 in the back of your book. I have also prepared SAS code that will provide $t_{n-1,1-\alpha/2}$ for any n and α.

So, we now have (Equation 6.6)

$$\bar{x} - t_{n-1,1-\alpha/2}\, s/\sqrt{n} \quad \text{to} \quad \bar{x} + t_{n-1,1-\alpha/2}\, s/\sqrt{n}$$

as a $100(1-\alpha)\%$ confidence interval for μ.

Unfortunately, there is a catch. While Equation 6.6 is valid for any $n > 1$, it does require that the population be normal.

Is the population normal? Recall from Lecture 1 that we discussed not only numerical summaries of sample data but also graphical summaries. The graphical summaries can be useful in determining whether a population is normal (or, more accurately, whether we will be comfortable proceeding as if the population were normal). For instance, suppose that we construct a histogram and find that the bars form a bell-shaped pattern. Since the sample distribution is anticipated to reflect the population distribution, we will be comfortable proceeding as if the population were normal.

Example (interval estimation, small sample). Consider the "Before - after" serum cholesterol data in Table 2.13 on page 39. Assuming normality, let us find a 95% confidence interval for μ. Previously we found that $n = 24$, $\bar{x} = 19.54$, $s^2 = 282.4$, and $s = \sqrt{282.4} = 16.80$. Using Table 5 or SAS, we see that $t_{23,0.975} = 2.069$. The 95% confidence interval is

$$19.54 - 2.069 \times 16.80/\sqrt{24} \ \text{ to } \ 19.54 + 2.069 \times 16.80/\sqrt{24},$$

which simplifies to $[12.44, 26.64]$.

Remarks. A subtle but common mistake is to assert that μ is inside a specific confidence interval with probability $1 - \alpha$. This is impossible because neither μ nor the confidence interval is random. The correct interpretation (Equation 6.8) is that if we repeatedly take samples and construct confidence intervals, then about $100(1 - \alpha)\%$ of the intervals will contain μ.

A confidence interval for μ tends to become narrower as n becomes larger (Equation 6.9). Intuitively, we have more information about the population when n is large, so the range of plausible values for the population mean diminishes: after all, only one value can be correct!

Estimating a population variance

Point and interval estimation. A point estimate for σ^2 is s^2. If the population is normal, then the quantity $(n - 1)S^2/\sigma^2$ follows a special probability distribution called the chi-square distribution on $(n - 1)$ degrees of freedom (Equations 6.14 and 6.13). Let $\chi^2_{n-1,\alpha/2}$ and $\chi^2_{n-1,1-\alpha/2}$ denote the $\alpha/2$ and $1 - \alpha/2$ quantiles of this distribution, which can be recovered (for selected α and n) from Table 6 in the back of your book. Since

$$\chi^2_{n-1,\alpha/2} \leq (n - 1)S^2/\sigma^2 \leq \chi^2_{n-1,1-\alpha/2}$$

is algebraically equivalent to

$$(n-1)S^2/\chi^2_{n-1,1-\alpha/2} \le \sigma^2 \le (n-1)S^2/\chi^2_{n-1,\alpha/2},$$

we have

$$P((n-1)S^2/\chi^2_{n-1,1-\alpha/2} \le \sigma^2 \le (n-1)S^2/\chi^2_{n-1,\alpha/2}) = 1-\alpha.$$

Hence, we refer to (Equation 6.15)

$$\frac{(n-1)s^2}{\chi^2_{n-1,1-\alpha/2}} \text{ to } \frac{(n-1)s^2}{\chi^2_{n-1,\alpha/2}}$$

as a $100(1-\alpha)\%$ confidence interval for σ^2.

Example (point and interval estimation). Consider the "Before - after" serum cholesterol data in Table 2.13 on page 39. Assuming normality, let us find a 95% confidence interval for σ^2. We already know that $n = 24$ and $s^2 = 282.4$. Using Table 6 or SAS, $\chi^2_{23,0.975} = 38.08$ and $\chi^2_{23,0.025} = 11.69$. Thus, a 95% confidence interval for σ^2 is

$$\frac{23 \times 282.4}{38.08} \text{ to } \frac{23 \times 282.4}{11.69},$$

which simplifies to $[170.6, 555.6]$.

Remarks. A confidence interval for σ^2 is obtained by multiplicatively adjusting s^2 to obtain lower and upper bounds for the range of plausible values. In particular, a confidence interval for σ^2 is not symmetric about s^2. This is unlike a confidence interval for μ, which is obtained by additively adjusting $\bar{x}$ and which is symmetric about $\bar{x}$.

Estimating a population proportion

Scenario. Now consider a somewhat different situation. We want to estimate the proportion of people in the population for whom a certain statement is true. Call this proportion p. For instance, what proportion of HIV-positive hemophiliacs develop AIDS more than 5 years after seroconversion?

Point and interval estimation. An intuitive point estimate is $\hat{p}$ (Equation 6.16), the proportion of people in the sample for whom the statement is true. Let $\hat{P}$ denote the random analogue to $\hat{p}$. If n is large, then by the argument given in Lecture 3, $\hat{P}$ $(= Y/n$ in the notation of Lecture 3) is approximately normally distributed with mean p and variance $p(1-p)/n$. Hence,

$$Z := (\hat{P} - p)/\sqrt{p(1-p)/n}$$

is approximately standard normal. Since $-z_{1-\alpha/2} \leq Z \leq z_{1-\alpha/2}$ is algebraically equivalent to

$$\hat{P} - z_{1-\alpha/2}\sqrt{p(1-p)/n} \leq p \leq \hat{P} + z_{1-\alpha/2}\sqrt{p(1-p)/n},$$

we have

$$P\left(\hat{P} - z_{1-\alpha/2}\sqrt{p(1-p)/n} \leq p \leq \hat{P} + z_{1-\alpha/2}\sqrt{p(1-p)/n}\right) \approx 1-\alpha.$$

This suggests a $100(1-\alpha)\%$ confidence interval of

$$\hat{p} - z_{1-\alpha/2}\sqrt{p(1-p)/n} \ \text{ to } \ \hat{p} + z_{1-\alpha/2}\sqrt{p(1-p)/n}.$$

But there is a problem. We can't use p, which is unknown, to help construct a confidence interval for itself! If n is large, however, $\hat{p}$ is a good substitute for p and we obtain (Equation 6.19)

$$\hat{p} - z_{1-\alpha/2}\sqrt{\hat{p}(1-\hat{p})/n} \ \text{ to } \ \hat{p} + z_{1-\alpha/2}\sqrt{\hat{p}(1-\hat{p})/n}.$$

Example (point and interval estimation). Let us find a 95% confidence interval for the proportion of HIV-positive hemophiliacs who develop AIDS more than 5 years after seroconversion. We have $\hat{p} = 122/287 = 0.425$, $n = 287$, and $z_{1-\alpha/2} = z_{0.975} = 1.96$. The confidence interval is

$$0.425 - 1.96\sqrt{0.425(0.575)/287} \ \text{ to } \ 0.425 + 1.96\sqrt{0.425(0.575)/287},$$

which simplifies to $[0.368, 0.482]$.

We can also perform this calculation in SAS, as shown on page 2 of {TABLE612.pdf}. The first four entries in the box "Binomial Proportion..." are the point estimate $\hat{p}$, the "asymptotic standard error" $\sqrt{\hat{p}(1-\hat{p})/n}$, and the endpoints of a 95% confidence interval.

Remarks. When estimating proportions, we regard a large sample as one for which $n\hat{p}(1-\hat{p}) \geq 5$ (preferably ≥ 10) instead of using a fixed demarcation such as 30 or 200. This is because the quality of the normal approximation underlying Equation 6.19 depends heavily on p.