

Interval estimation and statistical inference

We have looked at statistics before that are *estimates*: best guesses of parameter values. For example, we estimate μ , the population mean, with \bar{X} , the sample mean; we estimate σ^2 , the population variance, with s^2 , the sample variance. These are **point estimates** — a single-valued guess of the parametric value. It is simple, but doesn't give a sense of the *variability* of the estimate as a guess for the parameter. That is, how close to our guess do we think the parameter really is?

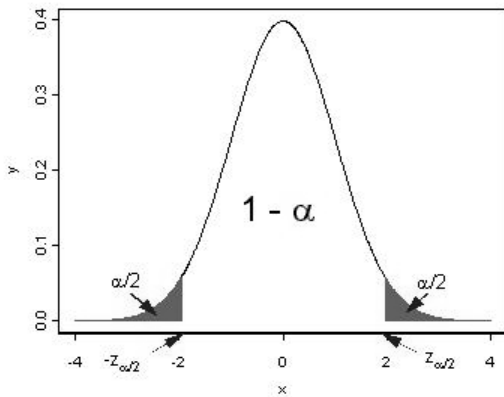
Interval estimates provide a range of values for a parameter value, within which we have a stated degree of confidence that the parameter lies. How can we construct such estimates? The key idea is to turn around what we already know from things like the Central Limit Theorem: **if a high percentage of the time in repeated samples values of \bar{X} are within a certain distance of μ , then μ is within the same distance of \bar{X} a high percentage of the time.** Thus, to get an interval estimate for μ , we simply invert an interval using \bar{X} based on the Central Limit Theorem. Recall that the CLT says that $\bar{X} \sim N(\mu, \sigma^2/n)$:

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

By definition,

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

where the following diagram defines $\pm z_{\alpha/2}$:



Oft-used values of $z_{\alpha/2}$ are

$$\alpha = .10 \Rightarrow z_{.05} = 1.645$$

$$\alpha = .05 \Rightarrow z_{.025} = 1.96$$

$$\alpha = .01 \Rightarrow z_{.005} = 2.575$$

Substituting in the definition of Z gives

$$P\left(-z_{\alpha/2} \leq \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \leq z_{\alpha/2}\right) = 1 - \alpha;$$

rearranging terms finally gives

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

This defines a $100 \times (1 - \alpha)\%$ **confidence interval** to be

$$\bar{X} \pm z_{\alpha/2}\sigma/\sqrt{n}.$$

Example. Say $\bar{X} = 102$ in the contract example (recall that $n = 50$ and $\sigma = 10$). What is a 95% confidence interval for μ ?

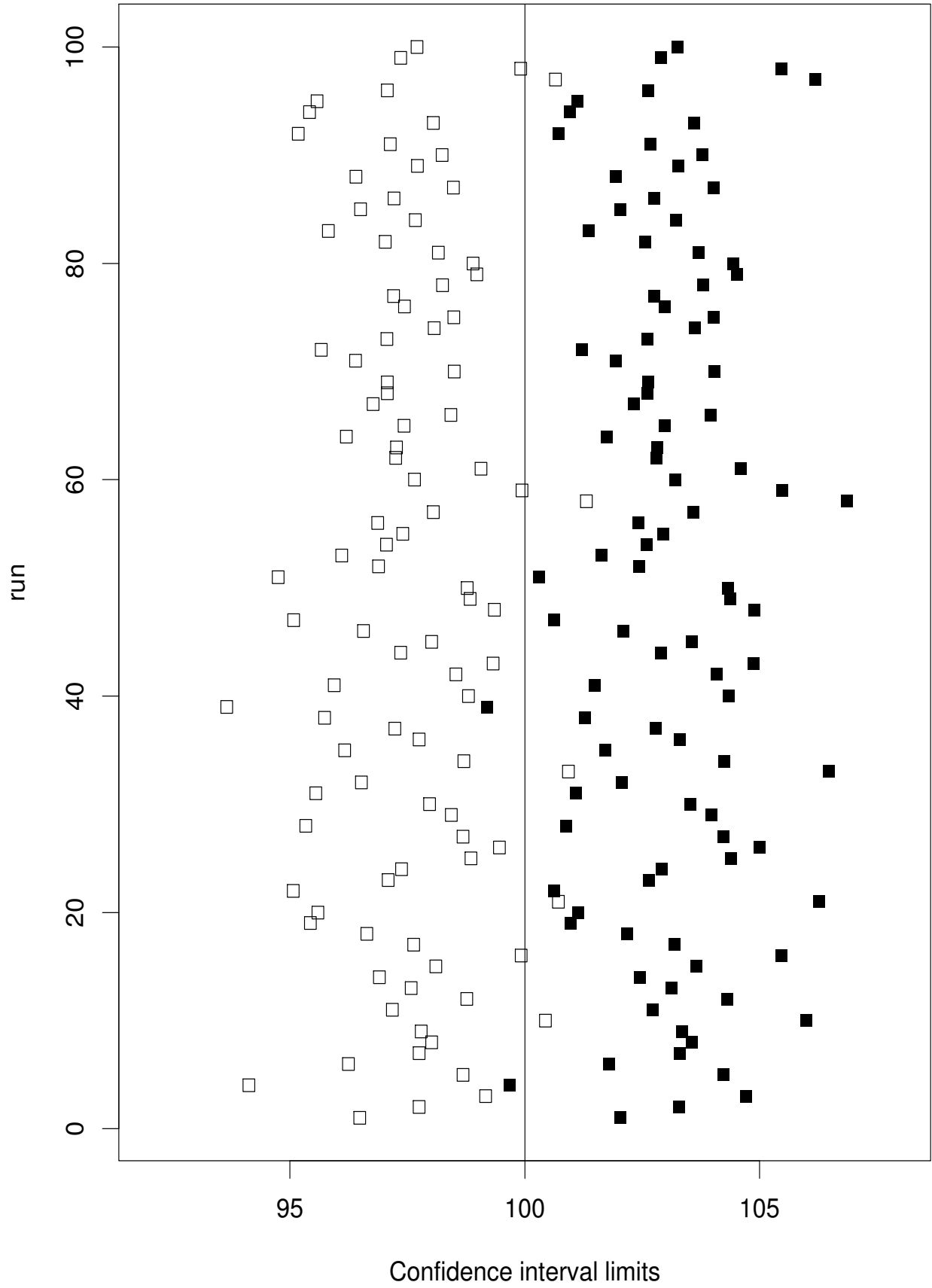
$$102 \pm (1.96)(10/\sqrt{50}) = 102 \pm 2.77,$$

or $(99.23, 104.77)$. Note that the interval includes 100, which implies that this value is not inconsistent with the observed data.

It is very tempting now to say that

$$P(99.23 \leq \mu \leq 104.77) = .95.$$

This is **NOT** true. The parameter μ either *is* in the interval, or *is not* in the interval; there is no probability attached to it (remember, μ is fixed (but unknown), not random). What **is** true is that if we took many random samples of size 50 from this population, and for each sample constructed a 95% confidence interval using the formula $\bar{X} \pm z_{\alpha/2}\sigma/\sqrt{n}$, roughly 95% of the intervals so constructed would actually contain μ . The following picture illustrates what is going on:



Note the important fact that the data analyst can never know if they are in the 95% of the time that the interval does contain μ , or the 5% of the time that it does not! All over the world, people are constructing 95% confidence intervals, taking comfort from the fact that the odds are 19 to 1 in their favor that the interval *does* contain the parameter of interest, and 5% of them are wrong! This strange concept of probability led to the invention of the term *confidence* for it, first coined by Jerzy Neyman in 1934. The number $100 \times (1 - \alpha)\%$ (e.g., 95%) is called the *confidence level* or *coverage* of the interval.

What if those 19 to 1 odds aren't high enough for you? That is, what if you want more confidence that μ is in your interval? No problem — just increase the confidence level to say 99%. There is a price to pay, however — the interval will be wider, since $z_{.005} = 2.575$ (while $z_{.025} = 1.96$). Note also that there's no such thing as a 100% confidence interval, unless you count the vacuous interval $(-\infty, \infty)$, so we can never be absolutely sure that our confidence interval contains the true value, for any confidence level.

There is an obvious flaw in this confidence interval construction: it requires knowledge of σ , which is not typically known. The obvious solution is to just substitute s , the sample standard deviation, for σ . This is exactly what people did until 1908, when William Gosset (“Student”) showed that this is not correct. The problem is that s doesn't exactly equal σ , and this additional source of variability has to be taken into account (especially for small samples). The Central Limit Theorem says that for large enough samples,

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

Gosset showed that if we are willing to assume the population is itself normally distributed,

$$\frac{\bar{X} - \mu}{s/\sqrt{n}} \sim t_{n-1},$$

where t_{n-1} refers to a (Student's) t distribution on $n - 1$ degrees of freedom. Then, by using the same arguments as were used earlier, a $100 \times (1 - \alpha)\%$ confidence interval for μ is

$$\bar{X} \pm t_{\alpha/2}^{(n-1)} s/\sqrt{n}.$$

For large sample sizes $t_{\alpha/2}^{(n-1)}$ and $z_{\alpha/2}$ become very close to each other, so the t -based interval looks like the z -based interval. In fact, for large enough samples, the Central Limit Theorem takes over and the z -based interval is appropriate, even if the population is not normally distributed.

Example. A few years ago I gave out a survey in the core course, asking (among other things) the annual salary of the jobs that the students had before enrolling at Stern. Here is a subset ($n = 10$) of those responses (in thousands of dollars):

20, 34, 52, 21, 26, 29, 71, 41, 23, 67

These data are quite long right-tailed; here is a stem-and-leaf display:

Character Stem-and-Leaf Display

Stem-and-leaf of Income N = 10
Leaf Unit = 1.0

```
(6)  2 01369
      4  3 4
      4  4 1
      3  5 2
      2  6 7
      1  7 1
```

This violates the assumption of the t -based confidence interval. Here $\bar{X} = 38.4$ and $s = 18.9$, so a 95% confidence interval for the true average income for incoming full-time MBA students is

$$\bar{X} \pm t_{.025}^{(9)} s / \sqrt{n} = (24.88, 51.91).$$

This interval is not very trustworthy, since \bar{X} is not very close to being normally distributed for such a small sample.

Let's say that we were satisfied with the confidence interval above. Does this interval imply that if we wanted to predict the income of a **particular** student, we would use that interval? The answer is resoundingly NO, since the confidence interval is designed to be an estimate for the **average income of all students**. If we had a very large sample, we would know μ almost exactly, and the confidence interval would be very narrow, but we still wouldn't be able to predict an individual's income with near perfect accuracy. The interval we want, which reflects a region where we feel a new observation will fall, is called a **prediction interval**, and it has the form

$$\bar{X} \pm t_{\alpha/2}^{(n-1)} s \sqrt{1 + \frac{1}{n}}.$$

This is similar to the confidence interval, except that it includes an extra “1”, which come from the inherent variability in the population. We are assuming here that the data values are a random sample from a roughly Gaussian distribution, just as was true for the confidence interval.

Unlike a confidence interval, a prediction interval reflects genuine probability. A 95% prediction interval, for example, is an interval within which the probability is .95 that a new observation will fall. Equivalently, we expect about 95% of the population values to fall inside the 95% prediction interval. Prediction intervals are useful in forecasting (predicting) future values, of course; they are also useful in quality control. In the latter context, “good data” are used to construct a prediction interval; then, if in future operation of a process the values move outside the interval, the process might be “out of control.” In this context a relatively wide interval is typically used, corresponding to at least $\alpha = .99$.

Example. Consider again the salary data given earlier. A 95% prediction interval for a randomly chosen salary is

$$38.4 \pm (2.262)(18.9)\sqrt{1.1} = (-6.44, 82.24)!$$

This is obviously ridiculous, as it includes impossible negative values (and recall that the minimal observed value is 20!). The problem is that the data are severely non-Gaussian, and this prediction interval is useless.

In this case, however, we can improve things using a log transformation to correct the long-tailedness. Here is a stem-and-leaf display of the logged incomes:

Character Stem-and-Leaf Display

Stem-and-leaf of Loginc N = 10
 Leaf Unit = 0.010

```

  4   13 026
(2)  14 16
  4   15 3
  4   16 1
  3   17 1
  2   18 25

```

A 95% prediction interval for logged income can be converted back to one for income by just using the antilog (exponentiating). Here's how it works. For the logged incomes, $\bar{X} = 1.54$ and $s = .2036$, so a 95% prediction interval for a randomly chosen logged income is

$$1.54 \pm (2.262)(.2036)\sqrt{1.1} = (1.057, 2.023).$$

That is, we expect that 95% of the logged population income values are between 1.057 and 2.023. But the logarithm is a one-to-one transformation, so 95% of the income values are between

$$10^{1.057} = 11.40$$

and

$$10^{2.023} = 105.44.$$

That is, a 95% prediction interval is (\$11400, \$105440). Note this interval is not symmetric (it is “centered” at the geometric mean of the incomes, $10^{1.54} = \$34674$), as it reflects the inherent long right tail in the incomes themselves. The interval is wide, but reflects the actual income distribution far more effectively than the one using the unlogged incomes.

We need to emphasize a fundamental difference here between confidence and prediction intervals. Since prediction intervals are designed to reflect the actual underlying distribution of the population, if a transformation can make that distribution look more Gaussian, a prediction interval built in that scale and then back-transformed can be quite effective. **Thus, if the sample comes from long right-tailed distribution, logging the values, constructing a prediction interval, and then anti-logging (exponentiating) the endpoints to get a prediction interval in the original scale is perfectly appropriate.** This is **not** the case for confidence intervals. A confidence interval for the mean of a logged distribution cannot be anti-logged to give a confidence interval for the mean of the original distribution, because the expected value of the logged distribution is not the same as the log of the expected value of the original distribution. **Thus, if the sample comes from long right-tailed distribution, logging the values, constructing a confidence interval, and then anti-logging (exponentiating) the endpoints to get a confidence interval in the original scale is not appropriate, and should not be attempted.**

On the other hand, confidence intervals do have an advantage over prediction intervals when data come from a non-Gaussian distribution that cannot be easily transformed to normality (say, long left-tailed or multimodal). If the sample size is big enough, the

Central Limit Theorem takes over, and a confidence interval constructed in the usual way on the original data will be a reasonable confidence interval for the mean of the original distribution. This is not the case for a prediction interval, since it is directly determined by the distribution in the original scale. Thus, if the sample cannot be transformed to a form that is at least roughly Gaussian, the method discussed here cannot be used to construct a prediction interval, no matter how large the sample is.

There's an important philosophical point that should be mentioned here. The confidence interval construction uses *sampling variability* to get the distributions of \bar{X} and s , and hence the z - and t -based intervals. In the terminology of W. Edwards Deming, the famous quality control expert, these are *enumerative studies*, where we wish to describe the parameters of the current population. If we were able to take a sample equal to the population, there would be no variability left, and the confidence interval would have width zero. An alternative situation is the *analytical study*, where we wish to make predictions about a future process. In this latter context, our "sample" can in fact be the entire population (a census of per capita incomes of the 50 states, for example), but confidence intervals only make sense if we are willing to assume that the underlying process doesn't change over time, so that we can view the current values as a "sample" from a stable process over time (that is, a "snapshot" in time).

Minitab commands

A z -based confidence interval is obtained by clicking **Stat** → **Basic Statistics** → **1-Sample z**. Enter the variable of interest under **Variables:**, and the true population standard deviation in the box next to **Sigma:**. The coverage (confidence level) of the interval can be changed by changing the number in the box next to **Level:**.

A t -based confidence interval is obtained by clicking **Stat** → **Basic Statistics** → **1-Sample t**. Enter the variable of interest under **Variables:**. The coverage level of the interval can be changed by changing the number in the box next to **Level:**.

“For better or worse, statistical inference has provided an entirely new style of reasoning. The quiet statisticians have changed our world — not by discovering new facts or technical developments but by changing the way we reason, experiment and form our opinion about it.”

— Ian Hacking

“Uncertain knowledge + Knowledge about the amount of uncertainty in it = Usable knowledge.”

— C. Radhakrishna Rao