STAT 210
Applied Statistics and Data Analysis
Review of Inference II: Interval Estimation

Joaquín Ortega
KAUST

Fall 2020

Interval estimation

Introduction

We studied previously how to get pointwise estimates for the mean of a distribution, using the sample mean.

The result of this procedure is a number, and this is not very satisfactory, because we have no idea of how accurate this estimation is.

If we have pointwise estimates coming from two different samples, one of size 1000 and another of size 10, we have no way of comparing the uncertainty associated to each one, if we only consider the single values we obtain in each case.

Introduction

As we saw before, the variance associated with the first estimate will be 100 times smaller than the variance associated with the second, and this means that we should expect the first estimated to have less ‘uncertainty’.

One way to quantify this uncertainty is through an interval estimate, or confidence interval, that gives a range of values for the unknown parameter, and has an associated confidence level.

Confidence intervals

Let α be a number in the interval $(0, 1)$, and let $Z \sim N(0, 1)$. Let z_α be the real number defined by the relation

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

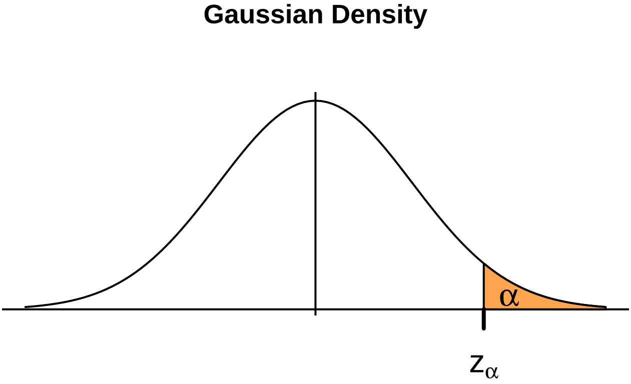

Figure 1: Definition of z_α for a Gaussian distribution

Confidence intervals

Figure 1 shows the relation between α and z_α.

Observe that the value of z_α increases as α decreases: To make the area smaller, we need to move z_α further to the right.

Since the normal distribution function is continuous and strictly increasing, this number is unique.

By the symmetry of the normal distribution,

$$P(Z < -z_\alpha) = \alpha$$

Confidence intervals

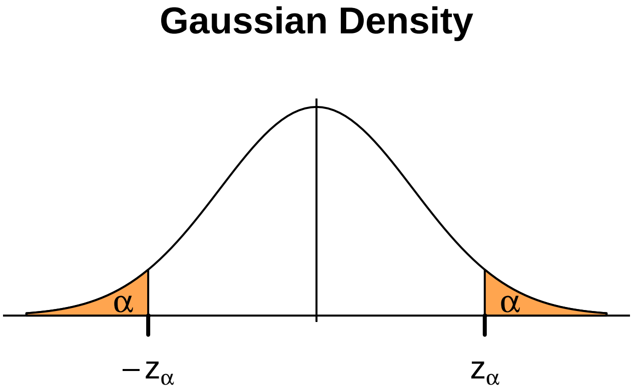

Figure 2: Definition of z_α and $-z_\alpha$ for a Gaussian distribution

Confidence intervals

This says that the probability that Z belongs to the interval $[-z_\alpha, z_\alpha]$ is $1 - 2\alpha$, or equivalently, using $\alpha/2$ instead of α,

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

In what follows, we will assume that the sample comes from a normal population.

If this is not true, but the sample size is large, the distribution of the sample mean is approximately normal, and the confidence intervals we get are also approximate.

Confidence intervals

Recall now that $\hat{\mu} \sim N(\mu, \sigma^2/n)$. Therefore

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

and this implies that $\sqrt{n}(\hat{\mu} - \mu)/\sigma$ has the same distribution as Z.

Therefore, replacing Z by $\sqrt{n}(\hat{\mu} - \mu)/\sigma$ in 1,

$$P\Big(\Big|\frac{\sqrt{n}(\hat{\mu} - \mu)}{\sigma}\Big| \leq z_{\alpha/2}\Big) = 1 - \alpha \tag{2}$$

Confidence intervals

After some manipulation of the inequality in the expression above we get that

$$P\Big(\hat{\mu} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \leq \mu \leq \hat{\mu} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\Big) = 1 - \alpha. \tag{3}$$

This equation says that, with probability $1 - \alpha$, the interval

$$I(\alpha; \sigma, n) = \Big[\hat{\mu} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}}, \hat{\mu} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\Big]$$

contains the actual value of the parameter μ.

We say that $(1 - \alpha)$ is the **confidence level** of the interval, frequently expressed as a percentage $100(1 - \alpha)\%$. The smaller α, the higher the confidence level.

Confidence intervals

Observe that the extremes of the interval are random. Therefore the probability statement we just made applies to them and not to the parameter, which has a fixed (but unknown) value.

The interval $I(\alpha; \sigma, n)$ is centered at $\hat{\mu}$ and depends on α, σ and n.

The standard deviation σ is a parameter of the population and is usually unknown, but fixed.

We will see later on how we can deal with the situation where σ is unknown, and we have to estimate it, adding uncertainty to the confidence interval. For the moment, we consider it known.

The width of the interval is

$$\left| I(\alpha; \sigma, n) \right| = 2 z_{\alpha/2} \frac{\sigma}{\sqrt{n}} \tag{4}$$

and the smaller the width, the sharper our estimate is.

Confidence intervals

We saw before that making α smaller makes z_α bigger.

Therefore, if we want to increase the confidence level of the interval, we need to reduce α, and this will increase z_α and make the confidence interval wider.

Higher confidence levels imply wider confidence intervals.

In other words, for fixed sample size, the precision (width) of the interval, and the confidence level go in opposite directions. If we want to reduce one of them, we have to increase the other.

Confidence intervals

On the other hand, if we increase sample size, we reduce the width of the interval and therefore increase its precision.

For instance, if we want to reduce the width by $1/2$, we need to increase the sample size by 4, since we have a square root in the denominator.

Increasing sample size is usually linked to costs, and therefore is not always feasible.

It is possible to use equation (4) to determine the sample size required to have a given width and confidence level, as long as we know the standard deviation σ or express the desired width in terms of σ.

Example 2

Suppose that we want the confidence interval for the mean of a normally distributed variable to be less than or equal to one half of the (unknown) standard deviation, with a confidence level of 98%.

This means that $\alpha = 0.02$. The interval width is given by expression

$$\left| I(\alpha; \sigma, n) \right| = 2 z_{\alpha/2} \frac{\sigma}{\sqrt{n}} \tag{4}$$

and we want this to be equal to $\sigma/2$.

This gives

$$\sqrt{n} = 4 z_{0.01}$$

We need to calculate $z_{0.01}$.

Example 2

Observe from figure 2 that $-z_\alpha$ is the α quantile of the standard normal distribution, and therefore we can use the function `qnorm` to calculate its value.

```
(alp <- qnorm(0.01))
```

```
## [1] -2.326348
```

Thus, $z_{0.01} = 2.326348$ and the sample size is

```
(4*abs(alp))^2
```

```
## [1] 86.59031
```

So, a sample of size 87 would be large enough to satisfy the conditions of the example.

One-sided confidence intervals

In some situations, we are only interested in one-sided confidence intervals.

For example, in a chemical reaction, it may be important that the temperature in the reaction does not exceed a specific critical value or that a sanitizing agent is at least as effective as the standard required by a particular control agency.

From relations (1) and (2), the one-sided confidence intervals are

$$\Big(\hat{\mu} - z_\alpha \frac{\sigma}{\sqrt{n}}, \infty\Big) \quad \text{and} \quad \Big(-\infty, \hat{\mu} + z_\alpha \frac{\sigma}{\sqrt{n}}\Big).$$

Confidence intervals when the variance is unknown

Confidence intervals when the variance is unknown

In the previous section, we assumed that the variance was known, but this is not likely. Almost always, both parameters of the distribution are unknown, so we also have to estimate the variance, and this will add uncertainty to the confidence interval.

In fact, the sampling distribution for the mean changes if we do not know the variance and have to estimate it from the sample.

We saw that if $\hat{\mu}_n = \frac{1}{n}\sum_1^n X_i$ and the X_i are iid with $N(\mu, \sigma^2)$ distribution then

$$\hat{\mu}_n \sim N(\mu, \sigma^2/n).$$

We estimate the variance by the empirical variance s_n^2:

$$s_n^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \hat{\mu}_n)^2.$$

Confidence intervals when the variance is unknown

The t distribution arises as the sampling distribution of the (empirical) mean $\hat{\mu}_n$ when the data come from a normal distribution with unknown variance.

It was shown in 1908 by W.S. Gosset in a paper in Biometrika, published under the pseudonym *Student*, that

$$\frac{\hat{\mu}_n - \mu}{s_n/\sqrt{n}} \sim t_{n-1} \tag{5}$$

where t_{n-1} denotes a t distribution with $n-1$ degrees of freedom.

For $n \geq 30$, the t distribution is very similar to the normal distribution, but for n small, there are substantial differences.

The t distribution has ‘heavier’ tails than the normal, which means that large values are more probable.

Confidence intervals when the variance is unknown

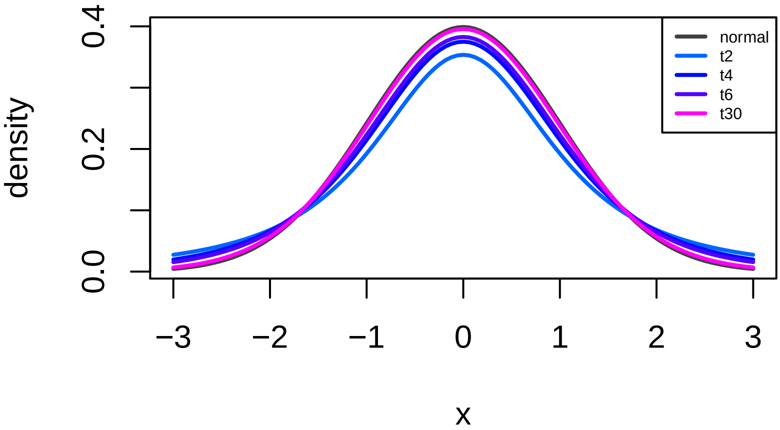

Figure 3: Density of the t distribution for different values of the degrees of freedom parameter. The normal standard density (black) is included as reference.

Confidence intervals when the variance is unknown

Now that we know the true distribution for the sample mean with unknown variance (under the hypothesis that the data are normal), we can obtain confidence intervals for the mean without assuming that the variance is known.

Let T_n be a random variable with t distribution with n degrees of freedom. We define $t_{\alpha,n}$ to be the real number that satisfies

$$P(T_n > t_{\alpha,n}) = \alpha.$$

By symmetry we have that $P(T_n < -t_{\alpha,n}) = \alpha$.

Confidence intervals when the variance is unknown

Following a similar argument as before, from equation (5) we get that

$$I^*(\alpha; n) = \left[\hat{\mu} - t_{\alpha/2,n-1}\frac{s_n}{\sqrt{n}}, \hat{\mu} + t_{\alpha/2,n-1}\frac{s_n}{\sqrt{n}}\right]$$

is a confidence interval of level $1 - \alpha$ for the mean. One-sided confidence intervals can be obtained similarly.

Example 1 revisited

Recall that we have a sample of size ten and the estimated value for the mean is

```
mean.smpl1
```

```
## [1] 4.311172
```

We know that the sample was simulated from a normal distribution with mean 4.5 and variance 4. Since, in this case, we know the exact value for the parameters, we can build different confidence intervals and compare them. We will consider

- interval built with the true value of the variance
- interval built with the estimated value of the variance but using the normal as the sampling distribution.
- interval built with the estimated value of the variance but using the t_{n-1} as the sampling distribution.

Example 1 revisited

```
I10 <- matrix(numeric(6), ncol=3)
(zz <- abs(qnorm(0.025)))
```

```
## [1] 1.959964
```

Interval using known variance

```
(I10[,1] <- c(mean.smpl1 - (zz*2/sqrt(10)),mean.smpl1 + (zz*2/sqrt(10))))
```

```
## [1] 3.071582 5.550762
```

Interval using estimated variance and normal sampling distribution

```
(I10[,2] <- c(mean.smpl1 - (zz*sqrt(var.smpl1)/sqrt(10)),
  mean.smpl1 + (zz*sqrt(var.smpl1)/sqrt(10))))
```

```
## [1] 3.116586 5.505758
```

Intervals using estimated variance and *t* sampling distribution

```
tt9 <- abs(qt(0.025,9))
(I10[,3] <- c(mean.smpl1 - (tt9*sqrt(var.smpl1)/sqrt(10)),
  mean.smpl1 + (tt9*sqrt(var.smpl1)/sqrt(10))))
```

```
## [1] 2.932401 5.689942
```

Example 1 revisited

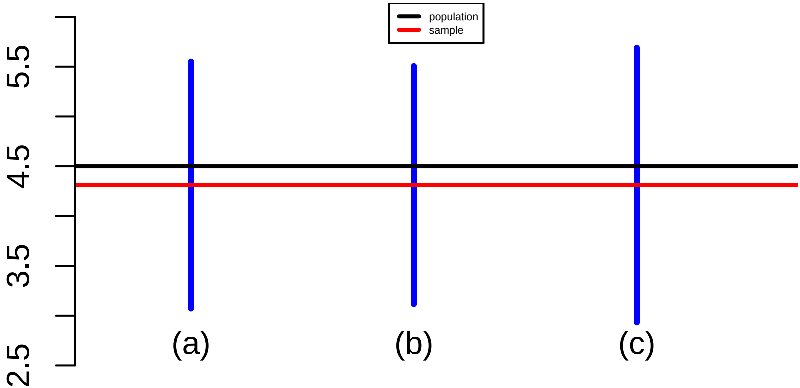

Figure 4: Confidence intervals for the mean in a sample of size 10 from the normal distribution N(4.5,4), (a) Interval using known variance, (b) Interval using estimated variance and normal sampling distribution, (c) Interval using estimated variance and t sampling distribution

Example 1 revisited

Figure 4 shows the three intervals, from left to right, interval using the true value for the standard deviation, interval using an estimated value for the standard deviation and a normal sampling distribution, and interval using an estimated value for the standard deviation and a t_9 sampling distribution.

The horizontal lines represent the true (black) and estimated (red) means.

Observe that in this case, the second interval is shorter than the first. The reason for this is that for the calculation of these intervals, the only difference is the value of the standard deviation, which in the first case is the true value, 2, while in the second is the estimated value, 1.927388.