Notes: Confidence Intervals

CS 3130 / ECE 3530: Probability and Statistics for Engineers

November 14, 2017

Confidence Intervals. So far, we've talked about estimates for parameters that give a single number as the best guess for that parameter given the data we have. Sometimes it is useful to instead estimate an interval for the possible values of the parameter and put a probability on how confident we are that the true parameter value falls inside this interval.

Consider a random sample $X_1, X_2, \ldots, X_n$ that comes from some distribution F with parameter θ.
A **100(1 - α)% confidence interval** for θ is a pair of statistics L_n, U_n, such that

$$P(L_n < \theta < U_n) = 1 - \alpha.$$

A common choice is $\alpha = 0.05$, which results in a 95% confidence interval.

Confidence Intervals for the Mean (Known Variance). The most common case for confidence intervals is when we have a Gaussian random sample, that is, $X_i \sim N(\mu, \sigma^2)$, and we want to estimate the mean. Remember when we discussed the Central Limit Theorem that we defined a normalized sample mean statistic:

$$Z_n = \frac{\bar{X}_n - \mu}{\sigma / \sqrt{n}}$$

Because the X_i are iid Gaussian, we know that $\bar{X}_n \sim N(\mu, \sigma^2/n)$. Therefore, the normalized sample mean has a standard normal distribution, i.e., $Z_n \sim N(0, 1)$. We want to find a **critical value** $z_{\alpha/2}$ that satisfies

$$P(-z_{\alpha/2} < Z_n < z_{\alpha/2}) = 1 - \alpha,$$

where the probability is computed using the standard normal cdf. With a little algebraic manipulation (see book), we can convert this to

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

This means that our confidence interval is given by

$$L_n = \bar{X}_n - z_{\alpha/2}\frac{\sigma}{\sqrt{n}}, \qquad U_n = \bar{X}_n + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}.$$

Using R: To calculate the critical value $z_{\alpha/2}$, notice that this is just another word for the $1 - \alpha/2$ quantile. The R command to get this value is:

```
qnorm(1 - 0.5 * alpha)
```

For the most common choice of $\alpha = 0.05$ (again, a 95% confidence interval), we get that $z_{0.025} \approx 1.96$, for a 99% confidence interval, we would get the critical value $z_{0.005} \approx 2.58$.

Problem: You want to estimate the average snowfall in the Wasatch front this year, and you take snowfall measurements at 40 different locations along the front. Experience from previous years indicates that there is a variance of 36 inches between measurements at these locations. You compute the average snowfall for the year is 620 inches. What is a 95% confidence interval for the average snowfall?

Answer: Try this in R. You should get the 95% confidence interval to be: $(618.1406, 621.8594)$.

Example: Margin of error for polls. You may notice that when you see the results of a poll, there is often a statement such as "the margin of error for this poll is $\pm 3\%$". What does this mean, and how do they come up with this number? If we are asking people about a choice between two candidates, then we can model their answers as a Bernoulli distribution. The goal of the poll is to estimate the parameter p, which is the proportion of people that will vote for candidate "1" over candidate "0".

Let $X_1, X_2, \ldots, X_n$ be iid Bernolli random variables representing people's responses. Recall that the sample mean, $\bar{X}_n$, is equivalent to the proportional of people that answer "1", and that this is an unbiased estimator of p. Also, remember that the *sum* of the X_i is a Binomial random variable, that is, $\sum X_i \sim Bin(n, p)$. We learned earlier that the Binomial distribution can be closely approximated by a Gaussian, $N(np, np(1-p))$. The sample mean statistic is just the sum scaled by the factor $1/n$, and therefore, the distribution of $\bar{X}_n$ can be approximated by the Gaussian $N(p, p(1-p)/n)$. So, our 100(1-α)% confidence interval will be

$$L_n = \bar{X}_n - z_{\alpha/2}\sqrt{\frac{p(1-p)}{n}}, \qquad U_n = \bar{X}_n + z_{\alpha/2}\sqrt{\frac{p(1-p)}{n}}$$

The **margin of error** (MOE) is the interval about our estimate for p, that is, the MOE equals $\pm z_{\alpha/2}\sqrt{\frac{p(1-p)}{n}}$.

Example: If I poll 100 people, what would my margin of error be (using $\alpha = 0.05$)?
Since we don't know the true p, we assume the worst-case scenario of $p = 0.5$ (remember this gives us the maximal variance $p(1-p) = 0.25$). We get that

$$z_{\alpha/2}\sqrt{\frac{p(1-p)}{n}} = 1.96\frac{0.5}{10} \approx 0.098$$

Example: How many people do I need to poll to get a 3% margin of error (again with $\alpha = 0.05$)?
We want $z_{\alpha/2}\sqrt{\frac{p(1-p)}{n}} = 0.03$. Again, assume worst-case $p = 0.5$ and plug in $\alpha = 0.05$ to get $z_{\alpha/2} = 1.96$. Now, we just need to solve for the unknown n.

$$n = \left(z_{\alpha/2}\frac{\sqrt{p(1-p)}}{0.03}\right)^2 = \left(1.96\frac{0.5}{0.03}\right)^2 \approx 1067$$

Confidence Intervals for the Mean (Unknown Variance). The problem with the above estimation of confidence intervals is that we have to know the true value of the variance. Typically, the true value of σ^2 is not known, and the best we can do is estimate it using the sample variance S_n^2. What happens if we replace the standard deviation σ with the sample statistic $S_n = \sqrt{S_n^2}$? Instead of Z_n, we get the following random variable:

$$T_n = \frac{\bar{X}_n - \mu}{S_n/\sqrt{n}}.$$

This variable will no longer have a $N(0,1)$ distribution. However, a fellow named William Gossett figured out the formula for this distribution in 1908. (Actually he had a slightly different statistic – the modern form of the t-statistic and its distribution were discovered by Ronald Fisher.) It is called Student's t-distribution because Gossett used the pseudonym "Student" so that he wouldn't get in trouble with his employer (Guinness brewery) for publishing his work. The pdf for the t-distribution is very similar to the Gaussian (see the book, Wikipedia, or you can plot it in R). It is centered at zero and a symmetric "hill" shape, but it does have "heavier tails", meaning that it goes to zero slower than the Gaussian. As n gets large, the t-distribution converges in the limit to a standard normal $N(0,1)$. The t-distribution has one parameter, the degrees of freedom m. The random variable T_n above has a t-distribution with degrees of freedom equal to $m = n-1$. In terms of notation, this is written $T_n \sim t(n-1)$.

Back to confidence intervals, we can also get critical values of the t-distribution, denoted $t_{\alpha/2}$, which satisfy

$$P(-t_{\alpha/2} < T_n < t_{\alpha/2}) = 1 - \alpha$$

Equivalent to the Gaussian case above, we construct our confidence interval as

$$L_n = \bar{X}_n - t_{\alpha/2}\frac{S_n}{\sqrt{n}}, \qquad U_n = \bar{X}_n + t_{\alpha/2}\frac{S_n}{\sqrt{n}}$$

Using R: The commands in R for the t-distribution are `dt` (gives the pdf), `pt` (gives the cdf), and `qt` (gives the quantiles). Each command takes a required parameter `df` for the degrees of freedom. To calculate the critical value $t_{\alpha/2}$, remember that this is just another word for the $1 - \alpha/2$ quantile. The R command to get this value is:

```
qt(1 - 0.5 * alpha, df = n - 1)
```

where `n` is your sample size used to compute the mean statistic. As an example, if the sample size was 10 and $\alpha = 0.05$ (again, a 95% confidence interval), we would call

```
qt(1 - 0.025, df = 9)
```

which returns the result 2.26. Notice that this gives a *wider* confidence interval than the Gaussian case above did. If n were to grow larger, our confidence interval would shrink and eventually match the 1.96 value of the Gaussian.

Problem: Repeat the Wasatch snowfall analysis above, but this time you do not rely on previous estimates of the snowfall variance. You compute the variance in your measurements to be $S_n^2 = 34$ inches. How did the confidence interval change?

Answer: Now you should get the 95% confidence interval to be: $(618.1352, 621.8648)$.