Math 321 – Chapter 5 – Confidence Intervals
(draft version 2019/06/20-13:30:42)

Contents

1 Introduction

The situation we find ourselves in is that we have some population or process that we will collect data from. We will analyze our dataset in order to learn about the underlying population or process that generated the dataset.

We assume that there is a probability distribution with certain parameters that each data point comes from independently. We want to estimate the parameters of this underlying population probability distribution.

Generally we have a population parameter θ and we use sample data to calculate an estimate of this parameter. what we calculate form our sample data will be called an *estimator*, and is usually denoted with a 'hat' over it, $\hat{\theta}$.

This estimator $\hat{\theta}$ is called a *point estimate* of the parameter θ.

We will mostly be concerned with estimating a population mean and variance. Recall the sample mean

$$\overline{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$$

and the sample variance

$$S_n^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X}_n)^2.$$

These can be used as point estimates for the population parameters, $\hat{\mu} = \overline{x}$ and $\hat{\sigma}^2 = s^2$.

Example: Resistance is normally distributed with mean μ and standard deviation σ, we collect a dataset of 11 resistors and find the sample mean resistance to be $\overline{x} = 8.89$ ohms and sample standard deviation $s = 0.38$ ohms. We would like to think that the true mean and standard deviation are close to our sample values: $\mu \approx 8.89$ and $\sigma \approx 0.38$.

Instead of a point estimate, it may be more desirable to give a range of values where the population parameter might be. This makes sense given that there will always be underlying uncertainty and randomness. No point estimate will ever be exact, generally, or at least we can never be absolutely certain it is exact.

So we construct *interval estimates.* An interval estimate will generally be of the form

$$\hat{\theta} \pm Q \cdot SE \ \text{ which means } \ (\hat{\theta} - Q \cdot SE, \hat{\theta} + Q \cdot SE)$$

or sometimes of the form

$$(\hat{\theta}Q_L, \hat{\theta}Q_U)$$

where $\hat{\theta}$ is our point estimate as describe above, and Q, Q_L, Q_U are quantiles or percentiles from some probability distribution, and SE is a standard error. the quantile Q and standard error SE are calculated according to know rules and theorems, such as the central limit theorem, and the properties that we know about the random variables and probability functions that we have studied.

We will study confidence intervals for: mean, variance, proportion, and differences between two means, variances, and proportions.

2 Confidence interval for mean μ

First we discuss how to estimate a population mean, μ.

2.1 Known variance

Here are two situations:

Under the assumptions that we know σ, and one of the following applies:

- $X_i \sim N(\mu, \sigma^2)$, or
- the data are not normal, but the sample size is large (usually $n \geq 30$ is an acceptable rule of thumb)

then we can use the following formula for a $(1-\alpha)100\%$ confidence interval for μ:

$$\overline{X}_n \pm z_{1-\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}$$

where $z_{1-\frac{\alpha}{2}}$ is the $(1-\frac{\alpha}{2})100^{th}$-percentile for the standard normal distribution given in R as

$$z_{1-\frac{\alpha}{2}} = \texttt{qnorm(1-}\alpha\texttt{/2)}.$$

This confidence interval is *exact* when the data are normal and *approximate* otherwise. What the term *exact* means is that the true confidence is exactly $1-\alpha$. When the data are not normal, then the true confidence may be different than $1-\alpha$. It could be higher or lower, but generally it is lower.

Another was to understand this is to think about it in terms of sampling distributions. If the data are normal, then we can exactly calculate probabilities on what the sample mean will be using the CLT. If the data are not normal, then we can still use the CLT, but the resulting probabilities about what the sample mean are will be approximate probabilities.

In R:

```
mean(x)+c(-1,1)*qnorm(1-α/2)*σ/sqrt(length(x))
```

or

```
> x = c(x1,x2,...,xn)
  xbar = mean(x)
  n = length(x)
  xbar + c(-1,1) * qnorm(1-α/2) * σ/sqrt(n)
```

2.2 Unknown variance, large sample

If we do not know the population standard deviation, σ, then we can approximate it by the sample standard deviation, s_n.

If we have a large sample size ($n \geq 30$ is an acceptable rule of thumb), then an *approximate* $(1-\alpha)100\%$ confidence interval for μ is given by:

$$\overline{X}_n \pm z_{1-\frac{\alpha}{2}} \frac{s_n}{\sqrt{n}}$$

where $z_{1-\alpha/2}$ is the $(1-\frac{\alpha}{2})100^{th}$-percentile for the standard normal distribution.

This does not rely on any assumption about the underlying population distribution, however if it is badly skewed or has very high probability of outliers, the approximation may be quite poor. Translation: if you want a 95% confidence interval, but the data has many outliers, your true confidence may actually be much less than 95%, maybe even as low as $50-70\%$, or even less. You can think of this as being because we have less control over the probabilities of the random sample's mean (the CLT approximation isn't as good in these cases for n too small). No matter the properties of the underlying population, though, we can always choose an n large enough to make this approximation as good as we want. In the 'bad' cases, it just may require a sample size in the thousands or even millions.

We know that the sample mean statistic standardized with the sample variance follows a Student's t-distribution with $n-1$ degrees of freedom.

$$\frac{\overline{X}_n - \mu}{\frac{s_n}{\sqrt{n}}} = t \sim T(n-1)$$

We also know that as the sample size gets large, the t-distribution converges to the standard normal. That is why this approximation works for large sample sizes.

In R:

```
mean(x)+c(-1,1)*qnorm(1-α/2)*sd(x)/sqrt(length(x))
```

or

```
> x = c(x1,x2,...,xn)
  xbar = mean(x)
  s = sd(x)
  n = length(x)
  xbar + c(-1,1) * qnorm(1-α/2) * s/sqrt(n)
```

2.3 Unknown variance, small sample

As stated above, we know that the sample mean statistic standardized with the sample variance follows a Student's t-distribution with $n-1$ degrees of freedom.

$$\frac{\overline{X}_n - \mu}{\frac{s_n}{\sqrt{n}}} = t \sim T(n-1)$$

So we can use this to construct a confidence interval:

$$\overline{X}_n \pm t_{1-\frac{\alpha}{2},n-1} \frac{s_n}{\sqrt{n}}$$

where $t_{1-\frac{\alpha}{2},n-1}$ is the $(1-\frac{\alpha}{2})100^{th}$ percentile of the t-distribution with $n-1$ degrees of freedom. In R this is

$$t_{1-\frac{\alpha}{2},n-1} = \texttt{qt(1-}\alpha\texttt{/2,df=}n\texttt{-1).}$$

Under the assumption that the data is normal, this is an exact CI. If the data are not normal, this formula can be used, but just know that the true confidence may be very much below what is desired. You can somewhat compensate for this by decreasing α, i.e. trying to construct a 99.9% CI instead of a 95% one in the hope that the true confidence may still be 95% or above.

In R, if we have a dataset, we can generally construct a confidence interval as follows:

```
> x = c(x1,x2,...,xn)
  xbar = mean(x)
  s = sd(x)
  n = length(x)
  a = α
  xbar + c(-1,1) * qt(1-a/2, df =n-1) * s/sqrt(n)
```

2.4 One-sided confidence bounds

In may cases, we only are interested in an upper or lower bound on the population mean. In this case, we only need the $(1-\alpha)100^{th}$ percentile instead of the $(1-\frac{\alpha}{2})100^{th}$ as in the two-sided case.

2.4.1 Known variance

Lower bound on μ

$$\overline{X}_n - z_{1-\alpha} \frac{\sigma}{\sqrt{n}}$$

Upper bound on μ

$$\overline{X}_n + z_{1-\alpha} \frac{\sigma}{\sqrt{n}}$$

2.4.2 Unknown variance, large sample

Lower bound on μ

$$\overline{X}_n - z_{1-\alpha}\frac{s_n}{\sqrt{n}}$$

Upper bound on μ

$$\overline{X}_n + z_{1-\alpha}\frac{s_n}{\sqrt{n}}$$

2.4.3 Unknown variance, small sample

Lower bound on μ

$$\overline{X}_n - t_{1-\alpha,n-1}\frac{s_n}{\sqrt{n}}$$

Upper bound on μ

$$\overline{X}_n + t_{1-\alpha,n-1}\frac{s_n}{\sqrt{n}}$$

2.5 Bootstrap CI for μ

Bootstrapping is a very nice technique for generating confidence intervals. It is a *non-parametric* technique. This means that it makes no underlying assumptions about what the population distribution is, e.g. that the data are normal or approximately normal. The only assumption is that the sample is independent and identically distributed.

For a sample of size n, bootstrapping will re-draw from this original sample a sample of size n with replacement. Each new resample of size n can then be used to calculate a sample mean. These resampled means will generally be different from the original sample mean since sampling with replacement will mean that some original data points are left out while others are drawn more than once. This resampling is done many times, often 100 or 1,000 is sufficient and then lower and upper quantiles are taken from the set of resampled means. These quantiles give a confidence interval for the underlying population mean.

Here is how the procedure works. We start with our original dataset $D = \{x_1, x_2, ..., x_n\}$. We randomly draw, with replacement, from this dataset to get a new dataset of the same size n, $Y = \{y_1, y_2, ..., y_n\}$ where for each i we have that $y_i = x_j$ for some j. Some of the y's might be the same x-value and some of the x-values may not appear in the resampled Y list. Now we do this many times to generate many resampled lists: $Y_1, Y_2, ..., Y_N$ usually $N = 100$ or 1,000 is sufficient. From each resampled list Y_i we calculate a mean $\overline{y}_i$. This gives us a list of resampled means $\overline{y}_1, \overline{y}_2, ..., \overline{y}_N$. We then get the $\alpha/2$ and $1-\alpha/2$ quantiles, $\tilde{\overline{y}}_{\alpha/2}$ and $\tilde{\overline{y}}_{1-\alpha/2}$, from this list of resampled means and our $(1-\alpha)100\%$ confidence interval for μ is $(\tilde{\overline{y}}_{\alpha/2}, \tilde{\overline{y}}_{1-\alpha/2})$.

Here is R code for the bootstrapping procedure for a confidence interval for the mean. You only need to make sure your data is stored in `x` and set your desired confidence level.

```
> x = c(x1,x2,...,xn) # input dataset
  a = 0.05 # set confidence level (1-a)
  N = 100 # number of resamples
  ybars = vector(mode="numeric", length=N)
  for (i in 1:N){
      bsamp = sample(x, size=length(x), replace=TRUE)
      ybars[i] = mean(bsamp)
  }
  ci = quantile(ybars,c(a/2,1-a/2)) # confidence interval
```

3 Confidence interval for proportion p

A $(1-\alpha)100\%$ confidence interval for a binomial proportion p:

$$\hat{p} \pm z_{1-\frac{\alpha}{2}} \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

The requirement is that $n\hat{p} > 5$ and $n(1-\hat{p}) > 5$.

Example: A machine produces parts and we wish to estimate the maximum possible true defective rate with 99.99% confidence. Out of 10,000 parts produced in a day, 120 were defective.

Note that we are doing an one-sided upper bound CI here. It makes sense that we are only interested in a one-sided interval here.

$n = 10000$ and $\hat{p} = 120/10000 = 0.012$ so we do satisfy $n\hat{p} > 5$ and $n(1-\hat{p}) > 5$.

$\alpha = 0.00001$ and `qnorm(0.9999)`$=3.719016$

So the 99.99% confidence upper bound on the true defective proportion is

$$0.012 + 3.719016 \cdot \sqrt{\frac{0.012 \cdot 0.988}{10000}} = 0.01604946$$

Thus we are 99.99% confident that the machine will produce no more than 1.6% defective parts.

4 Confidence interval for variance σ^2

A $(1-\alpha)100\%$ confidence interval for a normal variance σ^2:

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

where $\chi^2_{n-1,p}$ is the $p(100)\%$ quantile of the χ^2 distribution with $n-1$ degrees of freedom. Note that the quantiles might seem like they are 'switched' since they are in denominators.

In R: $\chi^2_{n-1,p} =$ `qchisq(`p`,df=n-1)`

Example: Construct a 95% CI for the variance of a normally distributed population from a sample of size 25 with sample variance $s^2 = 10$.

$\alpha = 0.05$ so teh quantile used for the lower bound is `qchisq(0.975,df=24)=` 39.36408 and the quantile used for the upper bound is `qchisq(0.025,df=24)=` 12.40115. And the interval is:

$$\left(\frac{24 \cdot 10}{39.4}, \frac{24 \cdot 10}{12.4}\right) = (6.1, 19.4)$$

That seems like quite a wide range of variances!

4.1 CI for standard deviation, σ

To get a CI for standard deviation, we just construct one for varaince and then take the square roots:

$$\left(\sqrt{\frac{(n-1)s^2}{\chi^2_{n-1,1-\alpha/2}}}, \sqrt{\frac{(n-1)s^2}{\chi^2_{n-1,\alpha/2}}}\right)$$

5 Confidence interval for difference in means $\mu_1 - \mu_2$

5.1 Welch's two-sample interval (unequal variances)

A $(1-\alpha)100\%$ confidence interval for the difference between the means $\mu_1 - \mu_2$:

$$(\overline{x}_1 - \overline{x}_2) \pm t_{\nu,1-\alpha/2} \cdot \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

where the degrees of freedom are

$$\nu = \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\frac{s_1^4}{n_1^2(n_1-1)} + \frac{s_2^4}{n_2^2(n_2-1)}}$$

Round ν down.

This interval assumes the underlying data are normally distributed. It is still often used whether or not we know the underlying data distributions, but if we have reason to suspect the data deviate far from normal, then only use this interval when both sample sizes are 30 or larger.

5.2 Equal variances

Here we assume that we have two samples X_i and Y_i from normal populations with identical variances.

A $(1-\alpha)100\%$ confidence interval for the difference between the means $\mu_1 - \mu_2$:

$$(\overline{x}_! - \overline{x}_2) \pm t_{n_1+n_2-2,1-\alpha/2} \cdot S_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$

where S_p is the pooled sample variance:

$$S_p = \sqrt{\frac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{n_1+n_2-2}}$$

Notice that the degrees of freedom for the t-distribution here is $n_1 + n_2 - 2$.

If the sample sizes are large ($n_1 \geq 30$ and $n_2 \geq 30$), then we can use a standard normal quantile instead:

$$(\overline{x}_1 - \overline{x}_2) \pm z_{1-\alpha/2} \cdot S_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$

Again, if the data are not normal, then only use these interval formulas when the sample sizes are large enough. Even for smaller sample sizes, though, these confidence intervals can be reasonably accurate for non-normal data as long as there are not too many outliers.

Generally, Welch's interval will be a better choice for a confidence interval with two samples. In either of these cases, one could also opt to use the minimum of $n_1 - 1$ and $n_2 - 1$ as the degrees of freedom. This is a *conservative* approach, which means in this context, that we are just making our interval wider to make sure we are sufficiently close to our desired confidence level. Smaller degrees of freedom will make the interval wider.

6 CI for difference in proportions $p_1 - p_2$

A $(1-\alpha)100\%$ confidence interval for the difference between binomial proportions $p_1 - p_2$:

$$(\hat{p}_1 - \hat{p}_1) \pm z_{1-\frac{\alpha}{2}}\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}$$

The requirement is that $n_1\hat{p}_1 > 5$, $n_1(1-\hat{p}_1) > 5$, $n_2\hat{p}_2 > 5$, and $n_2(1-\hat{p}_2) > 5$. This means that both samples each have more than 5 successes and more than 5 failures.

Example: Suppose 310 out of 500 surveyed Democrats support a particular congressional bill, and 220 out of 400 surveyed Republicans support the bill. Estimate with 95% confidence the true difference in support among Democrats and Republicans.

$n_1 = 500$, $\hat{p}_1 = 310/500 = 0.62$, $n_2 = 400$, $\hat{p}_1 = 220/400 = 0.55$, so we think the difference is 7%. This is our point estimate of the difference. The interval estimate of the difference is:

$$0.62 - 0.55 \pm 1.96 \cdot \sqrt{\frac{(0.62)(0.38)}{500} + \frac{(0.55)(0.45)}{400}} = (0.0053, 0.1347)$$

So even though we think there is a 7% difference, it may be as little as 0.5% or as high at 13%!

Changing our confidence level to 99% shows that Republicans might even support the bill at a higher proportion since the interval becomes $(-0.015, 0.155)$.

7 Confidence interval for ratio of variances σ_1^2/σ_2^2

A $(1-\alpha)100\%$ confidence interval for the rtio of normal variances σ_1^2/σ_2^2:

$$\left(\frac{s_1^2/s_2^2}{F_{1-\frac{\alpha}{2},n_1-1,n_2-1}}, \frac{s_1^2/s_2^2}{F_{\frac{\alpha}{2},n_1-1,n_2-1}}\right)$$

where F_{p,n_1-1,n_2-1} is the $p(100)^{th}$ percentile from the F-distribution with $n_1 - 1$ numerator degrees of freedom and $n_2 - 1$ denominator degrees of freedom.

In R: $F_{p,n_1-1,n_2-1} =$ `qf(`p`,`n_1`-1,`n_2`-1)`.

8 Summary

R commands:

```
mean(x)+c(-1,1)*qnorm(1-a/2)*sigma/sqrt(n)
mean(x)+c(-1,1)*qt(1-a/2,n-1)*sd(x)/sqrt(n)
```

Notation and formulas:

$\overline{x} \pm z_{1-a/2} \frac{\sigma}{\sqrt{n}}$

$\overline{x} \pm t_{1-a/2,n-1} \frac{s}{\sqrt{n}}$

$\hat{p} \pm z_{1-\frac{\alpha}{2}} \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$

$(\overline{x}_1 - \overline{x}_2) \pm t_{\nu,1-\alpha/2} \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$

$\nu = \frac{\left(\frac{s_1^2}{n_1}+\frac{s_2^2}{n_2}\right)^2}{\frac{s_1^4}{n_1^2(n_1-1)}+\frac{s_2^4}{n_2^2(n_2-1)}}$

$(\hat{p}_1 - \hat{p}_1) \pm z_{1-\frac{\alpha}{2}} \sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}$