009 - Confidence Intervals

EPIB 607 - FALL 2020

Sahir Rai Bhatnagar
Department of Epidemiology, Biostatistics, and Occupational Health
McGill University

sahir.bhatnagar@mcgill.ca

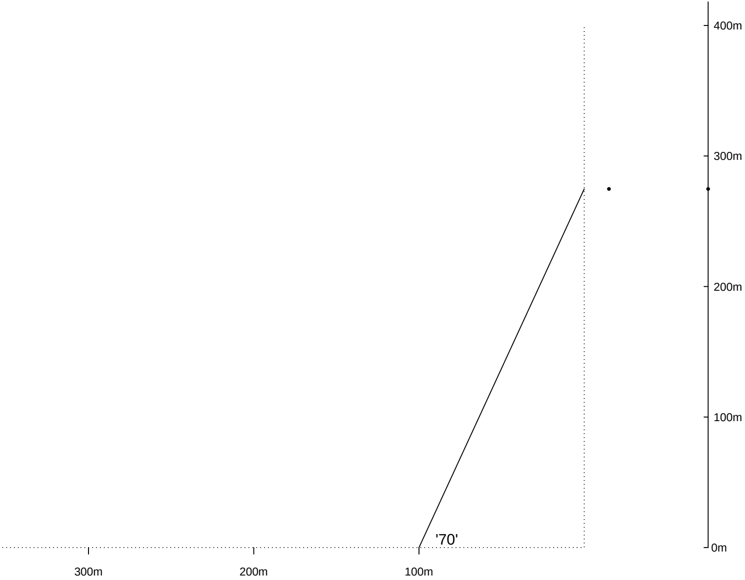
slides compiled on September 25, 2020

Key takeaways and next steps

1. We’ve been exclusively talking about point estimates
2. How confident are we about these point estimates?
3. Thought experiment: Estimate the average temperature in Montreal in August over the past 100 years. How much money would you be willing to bet on it?
4. We’re going into stat territory now.

Example 1: Height of a building

- Consider the height of a particular building θ.
- We use tools of mathematical science (trignometry) to measure it.
- Suppose you measure the height of this building by standing a known horizontal distance (e.g. 100 metres) from the bottom of the building and using an instrument to measure the angle between the horizontal and the top of the building.


400m
300m
200m
100m
0m
'70'
300m
200m
100m

Example 1: Height of a building

- After calculating this, you learn that the measuring instrument only displays the angle is to the nearest 10 degrees. This means that the true angle is somewhere between 65 and 75 degrees.
- So you **cannot say** that the true height is **exactly** 275 metres. What **can** you say? And with what certainty?
- You can put **limits** on the true height by asking **what are the minimum and maximum heights that could have produced the observed reading of 70 degrees?**

Example 1: Height of a building

- **Lower limit:** What is the **minimum** angle that could have given the (rounded) readout of 70 degrees ?

- **Upper limit:** What is the **maximum** angle that could have given the (rounded) readout of 70 degrees ?

Example 1: Height of a building

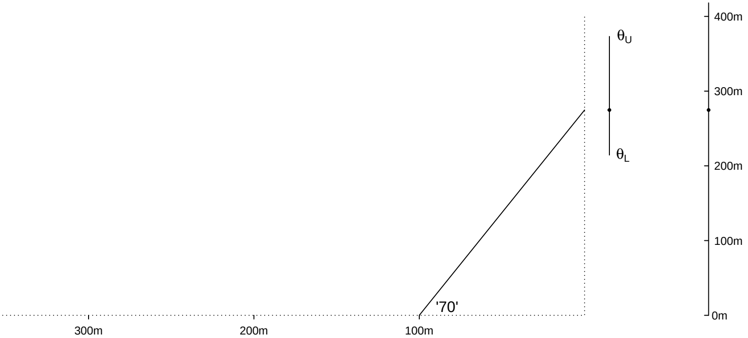


Figure: Estimating the height of an building by measuring subtended angles. The '70' signifies that the real angle was somewhere between 65 and 75 degrees; thus the real height lies between the L and U limits of 214 and 373 metres.

Example 2: Estimating calendar age[1]

A Royal Australian Navy officer stands on the coach house of a boat carrying suspected illegal immigrants near Ashmore Reef, about 850 km west of Darwin. Defence department photograph, April 16th, 2009. REUTERS/Australian Department

[1] People smugglers, statistics and bone age by Tim Cole (2012). Significance Magazine. https://doi.org/10.1111/j.1740-9713.2012.00568.x

Example 2: Estimating calendar age[2]

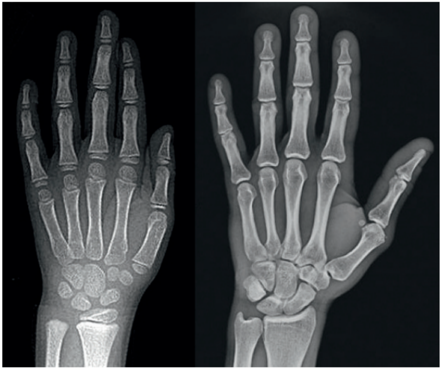


Figure 1. Composite X-rays of, left, an immature hand, and right, an adult hand. The growth plates at the end of the long bones can be clearly seen in the left-hand image. In the right-hand image the growth plates have fused with the long bones. It is impossible to tell the owner's age from this mature X-ray

[2] People smugglers, statistics and bone age by Tim Cole (2012). Significance Magazine. https://doi.org/10.1111/j.1740-9713.2012.00568.x

Example 2: Estimating calendar age

- The person's correct chronological age is a particularistic parameter, one that had nothing to do with science, or universal laws of Nature. But it can be estimated by using the laws of mathematics and statistics.
- Consider first a single indirect measurement of chronological age, that yielded a value of 17.6 years.
- Given what you know about the sizes of the possible errors, you **cannot say** that the true age is **exactly** 17.6 years What **can** you say? And with what certainty?
- You can put **limits** on the true age by asking **what are the minimum and maximum ages that could have produced the observed reading of** 17.6 years.

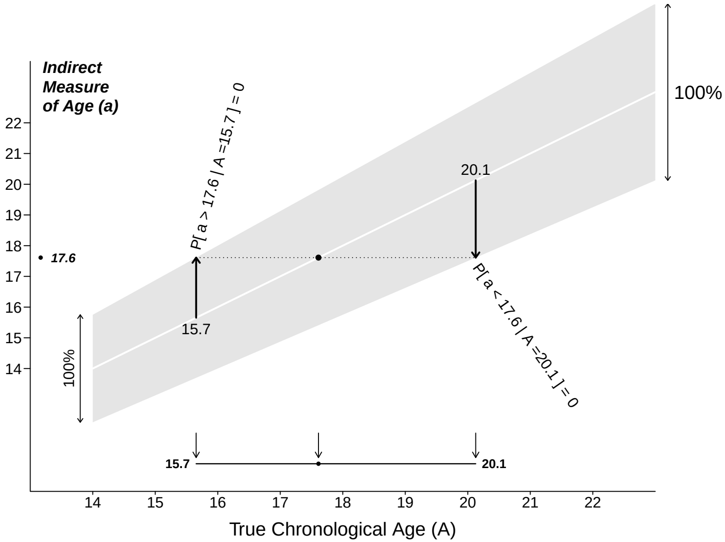


Figure: 100% Confidence Intervals for a person's chronological age when error distributions (that in this example are wider at the older ages) are 100% confined within the shaded ranges.

Confidence Interval

Definition (Confidence Interval)

A level C confidence interval for a parameter has two parts:

1. *An interval calculated from the data, usually of the form*

   $$\textit{estimate} \pm \textit{margin of error}$$

   where the estimate is a sample statistic and the margin of error represents the accuracy of our guess for the parameter.
2. *A confidence level C, which gives the probability that the interval will capture the true parameter value in different possible samples. That is, the confidence level is the success rate for the method*

Confidence Interval: A simulation study

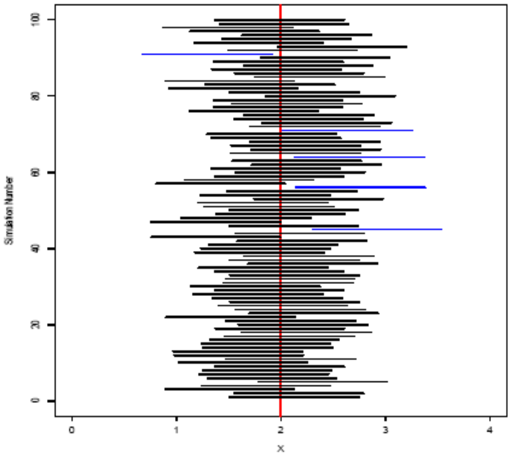


Figure: True parameter value is 2 (red line). Each horizontal black line represents a 95% CI from a sample and contains the true parameter value. The blue CIs do not contain the true parameter value. 95% of all samples give an interval that contains the population parameter.

Confidence Intervals: we only get one shot

- In practice, we don't take many simple random samples ("repeated" samples) to estimate the population parameter θ.

- Because the method has a 95% success rate, all we need is one simple random sample to compute one CI.

Interpreting a frequentist confidence interval

- The confidence level is the success rate of the method that produces the interval.
- We don't know whether the 95% confidence interval from a particular sample is one of the 95% that capture θ (the unknown population parameter), or one of the unlucky 5% that miss.
- To say that we are 95% confident that the unknown value of θ lies between U and L is shorthand for "We got these numbers using a method that gives correct results 95% of the time."

More about a frequentist confidence interval

- The confidence level of 95% has to say something about the sampling procedure:
  - The confidence interval depends on the sample. If the sample had come out differently, the confidence interval would have been different.
  - With some samples, the interval 'estimate $\pm$ margin of error' does trap the population parameter (the word statisticians use is cover). But with other samples, the interval fails to cover.
- It's like buying a used car. Sometimes you get a lemon – a confidence interval which doesn't cover the parameter.

```
  covers               a lemon                 another lemon

|--P--------|       P  |-----------|          |-----------| P
```

Figure: 3 confidence intervals 'chasing' (taking a shot at) the population parameter P

More about a frequentist confidence interval

- In the frequentist approach, θ is regarded as a fixed (but unknowable) constant, such as the exact speed of light to an infinite number of digits, or the exact mean depth of the ocean at a given point in time.

- It doesn't "fall" or "vary around" any particular values; in contrast you can think of the statistic $\hat{\theta}$ "falling" or "varying around" the fixed (but unknowable) value of θ

Polling companies

- Polling companies who say “polls of this size are accurate to within so many percentage points 19 times out of 20” are being statistically correct $\rightarrow$ they emphasize the **procedure** rather than what has happened in this specific instance.

- Polling companies (or reporters) who say “this poll is accurate .. 19 times out of 20” are talking statistical nonsense – this specific poll is either right or wrong. On average 19 polls out of 20 are “correct”. But this poll cannot be right on average 19 times out of 20.

Example: Inference for a single population mean

We begin with the (unrealistic) assumption that the population variance is known.

- Then the true variance of the sample mean is known!
- We can use `mosaic::xpnorm(q = c(-1.96, 1.96))` to find that there is a 95% chance that a $\mathcal{N}(0,1)$ random variable lies within 1.96 standard errors of the population mean of the distribution. So then:

$$P\left(-1.96 \leq \frac{\bar{y} - \mu}{\sigma/\sqrt{n}} \leq 1.96\right) = 0.95$$

What does allow us to learn about μ?

Example: Inference for a single population mean

We can use this probability statement about the standardized version of $\bar{y}$ to place bounds on where we think the true mean lies by examining the probability that $\bar{y}$ is within $1.96\frac{\sigma}{\sqrt{n}}$ of μ.

$$
\begin{aligned}
P\left(-1.96 \leq \tfrac{\bar{y}-\mu}{\sigma/\sqrt{n}} \leq 1.96\right)
&= P\left(-1.96\frac{\sigma}{\sqrt{n}} \leq \bar{y}-\mu \leq +1.96\frac{\sigma}{\sqrt{n}}\right) \\
&= P\left(-\bar{y}-1.96\frac{\sigma}{\sqrt{n}} \leq -\mu \leq -\bar{y}+1.96\frac{\sigma}{\sqrt{n}}\right) \\
&= P\left(\bar{y}+1.96\frac{\sigma}{\sqrt{n}} \geq \mu \geq \bar{y}-1.96\frac{\sigma}{\sqrt{n}}\right) \\
&= P\left(\bar{y}-1.96\frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{y}+1.96\frac{\sigma}{\sqrt{n}}\right) \\
&= 0.95
\end{aligned}
$$

We call the interval $\left(\bar{y}-1.96\frac{\sigma}{\sqrt{n}}, \bar{y}+1.96\frac{\sigma}{\sqrt{n}}\right)$ a **95% confidence interval** for μ.

Example: Inference for a single population mean

So what does the CI allow us to learn about μ??

- In classical (frequentist) statistics, we assume that the population mean, μ is a **fixed** but unknown value.
- With this view, it doesn't make sense to think of μ as having a distribution. Therefore we can't make probability statements about μ.
- What about the CI? It is made up of the sample mean and other fixed numbers (1.96, the square root of the known sample size n, and the known standard deviation, σ).
- **The CI is a random quantity.**
- Remember: a random quantity is one in which the outcome is not known ahead of time. We don't know the lower and upper limits of the CI before the sample has been collected since we don't yet know the value of the random quantity $\bar{x}$.

Example: Inference for a single population mean

So what does the CI allow us to learn about μ??

- It tells us that if we repeated this procedure again and again (collecting a sample mean, and constructing a 95% CI), 95% of the time, the CI would *cover* μ.

- That is, with 95% probability, the *procedure* will include the true value of μ. Note that we are making a probability statement about the CI, not about the parameter.

- Unfortunately, we do not know whether the true value of μ is contained in the CI in the particular experiment that we have performed.

Interactive visualization of CIs

http://rpsychologist.com/d3/CI/

Exercise: How deep is the ocean?

1. For your samples of $n = 5$ and $n = 20$ of depths of the ocean, calculate the
   1.1 sample mean ($\bar{y}$)
   1.2 standard error of the sample mean ($SE_{\bar{y}}$)
2. Calculate the 68%, 95% and 99% confidence intervals (CI) for both samples of $n = 5$ and $n = 20$.
3. Enter your results in the Google sheet
4. Plot the CIs for each student using the following code:

```
plot(dt$Mean.5.depths, 1:nrow(dt), pch=20,
xlim=range(pretty(c(dt$lower.mean.5.66, dt$upper.mean.5.66))),
xlab='Depth of ocean (m)', ylab='Student (sample)',
las=1, cex.axis=0.8, cex=1.5)
abline(v = 3700, lty = 2, col = "red", lwd = 2)
segments(x0 = dt$lower.mean.5.66, x1=dt$upper.mean.5.66,
y0 = 1:nrow(dt), lend=1)
```