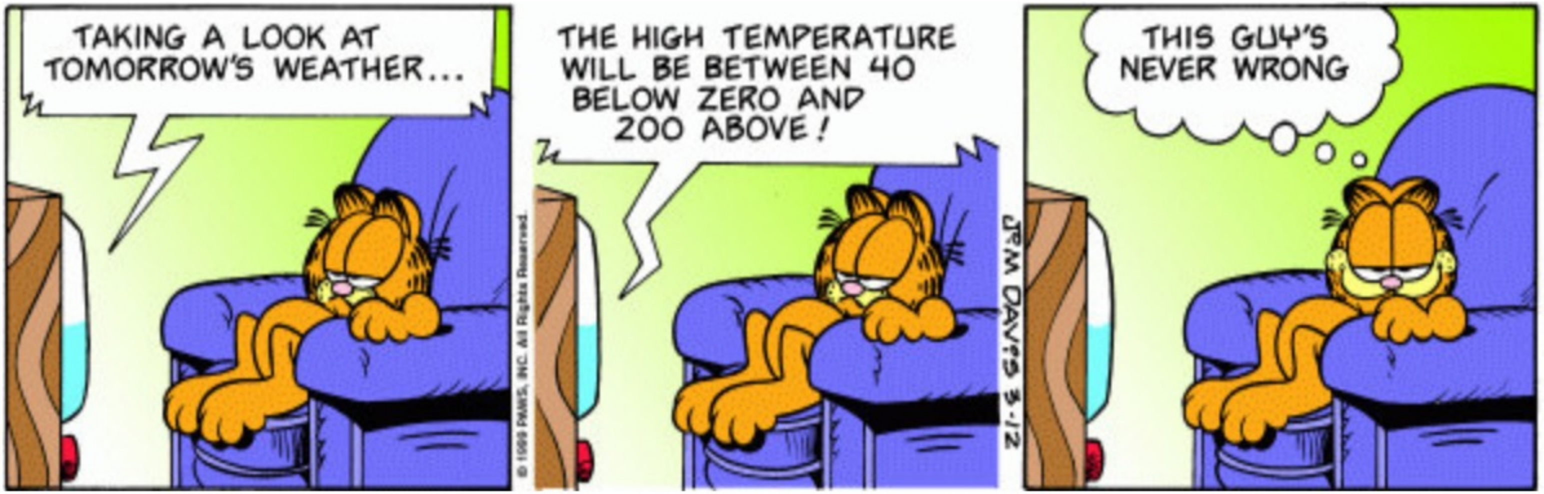
Stat 88: Prob. & Mathematical Statistics in Data Science



Lecture 23 Part 1: 4/14/2022

Section 9.4

Interpreting confidence intervals

Confidence intervals for the population mean: recap

- A *confidence interval* is an interval on the real line, that is, a collection of values, that are plausible estimates for the true mean μ.

(Interval estimate for μ)

Can write down CI using SD or $\bar{X}$

95% CI for μ : $\boxed{\bar{X} \pm (1.96) \times SD(\bar{X})}$

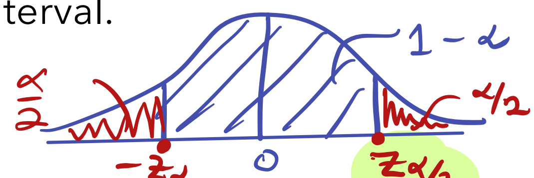
- Using the CLT, we can estimate the chance that this interval contains the true mean. If we want the chance to be higher, we make the interval bigger. The interval is like a net. We are trying to catch the true mean in our net.

- The CLT takes the form: $\boxed{\bar{X} \pm \textit{margin of error}}$, where the margin of error tells us how big our interval is, and depends on the SD of the sample mean.

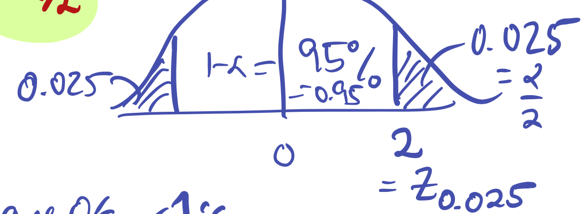
- The margin of error $= z_{\alpha/2} \times SD(\bar{X})$, where $z_{\alpha/2}$ is the quantile we need to have an area of $1-\alpha$ in the middle, that is, a ***coverage probability*** of $1-\alpha$

- The probability with which our ***random*** interval will cover the mean is called the confidence level.

- In reality (vs theory), we will have just one ***realization*** (observed value) of the sample mean (from our data sample), and we use that value to write down the **realization** of our random interval.



$N(0,1)$



Dealing with proportions

$\hat{p} = \bar{X}$ when $X_1, X_2, \ldots X_n$ are 0's or 1's

- A sample proportion is just the sample mean of a special population of 0's and 1's.
- This kind of population is so common since many of our problems deal with *classifying* and *counting*.
- We have a population of 1 million in a town. We take a SRS of size 400 and find that 22% of the sample is unemployed. Estimate the percentage of unemployed people in the town.

$N = 10^6$, $n = 400$ $\longrightarrow$ even though sample is a SRS, can pretend $X_1, X_2 \ldots X_{400}$ are indep. $\sim$ Bernoulli (p)

$\mathbb{E}(X_k) = p$, $SD(X_k) = \sqrt{pq}$, $q = 1-p$, $SD(X_k) = \sqrt{p(1-p)}$

$X_k = \begin{cases} 1 & \text{w.p. } p \\ 0 & \text{w.p. } 1-p \end{cases}$

$\boxed{p \text{ unknown}}$

$\hat{p}$ = sample value of $p = 0.22$

$\boxed{\hat{p} \text{ a.k.a. } \bar{X} \text{ a.k.a } A_{400}}$

$\mathbb{E}(\hat{p}) = \mu = \mathbb{E}(\bar{X}) = \mathbb{E}(X_k) = p$

$SD(\bar{X}) = SD(\hat{p}) = \frac{\sigma}{\sqrt{n}} = \frac{\sqrt{p(1-p)}}{\sqrt{400}}$

By the CLT $\hat{p}$ is approx $N\left(\mu, \frac{\sigma}{\sqrt{n}}\right)$

$$SD(\hat{p}) = \frac{\sqrt{p(1-p)}}{20}$$. Don't know p, use $\hat{p}$ to approximate the SD.

("bootstrapping not the way you do in data 8, though")

$$\text{approx } SD(\hat{p}) = \frac{\sqrt{(0.22)(0.78)}}{20} \approx 0.0207$$

95% C.I for p is given by

$$0.22 \pm 2 \times 0.0207 = 0.22 \pm 0.0414$$

$$22\% \pm 4.14\% = (22-4.14)\%, 22+4.14\%)$$

$$= (17.86\%, 26.14\%)$$

Random interval

$$\hat{p} \pm z_{\frac{\alpha}{2}} \cdot \underbrace{SD(\hat{p})}_{\frac{\sigma}{\sqrt{n}}}$$ ← before we plug in observed values, this is a random interval with a prob. of $1-\alpha$ of "covering" the true value of p.

Once we plug in the observed value of $\hat{p}$ & use it to approximate $\frac{\sigma}{\sqrt{n}} = \frac{\sqrt{p(1-p)}}{\sqrt{n}}$,

we have only particular numbers & no randomness. We cannot talk about probabilities any more and say that the resulting interval is a $(1-\alpha)100\%$ CONFIDENCE INTERVAL for the true p. In our example, the confidence level of the interval is 95%.

$(1-\alpha = 0.95)$

Example

$X_1, X_2, \ldots X_{400}$ $(n=400)$ $\bar{X}$: r.v. $\bar{x}$: observed value

- In a simple random sample of 400 voters in a state, 23% are undecided about which way they will vote. Find a 95% CI for the proportion of undecided voters in the state.

μ expected value of $\bar{X}$

- In the above problem, find 99.7% confidence interval.

$n=400$, $1-\alpha = 95\% = 0.95$, $z_{\frac{\alpha}{2}} = 2$ (or 1.96)

$\hat{p} = 0.23$, $SD(\hat{p}) = \frac{\sqrt{p(1-p)}}{\sqrt{400}} \approx \frac{\sqrt{(0.23)(0.77)}}{20}$

$0.23 \pm 2 \times \frac{\sqrt{(0.23)(0.77)}}{20}$

$X_k \sim \text{Bernoulli}(p)$

$\hat{p} = \frac{X_1 + X_2 + \cdots + X_n}{n}$, $E(\hat{p}) = \frac{nE(X_k)}{n} = p$

$SD(\hat{p}) = \sqrt{Var(\hat{p})}$, $Var(\hat{p}) = \frac{Var(X_k)}{n}$

For a 99.7% C.I : $0.23 \pm 3 \times \sqrt{\frac{(0.23)(0.77)}{400}}$

margin of error (m.e)

to halve the m.e, need to quadruple n.

Section 9.4: Interpretation

$= P(|\bar{X}-\mu| < 2\sigma) \approx 0.95$

- Chance that sample mean is less than 2 SDs away from population mean is about 0.95

$P(\mu-2\sigma < \bar{X} < \mu+2\sigma)$

(by CLT & empirical rule)

- Therefore the chance that population mean is less than 2 SDs away from sample mean is about 0.95

$P(|\mu-\bar{X}| < 2\sigma) = 0.95$

- Which object is random in each of these sentences?

ONLY THE SAMPLE MEAN. NOT μ.

- Does it make sense to say "The probability that the number 2 is between 3 and 5 is 0.95" ?

- Does it make sense to say "The probability that the population mean is between 18 and 26 is 0.95"?

Interpretation

- Let’s think about tossing coins. *Before* we toss a coin some number of times, we can say that the number of heads is random, since we *don’t know* how many heads we will get.

- Suppose we have tossed the coin (say 100 times) and we see 53 heads, can we say 53 is a random number and the chance that 53 lies between 40 and 50 is 95%?

observed value

- 53 is our ***realization*** of the random ”number of heads” in this ***particular*** instance of 100 tosses.

Confidence intervals: What is random?

$\bar{X}$ $\bar{x}$ $z_{\frac{\alpha}{2}}$

- Note that if we use the sample mean and extend one or two SDs in either direction, we *may* or *may not* cover the true population percentage.
- The *interval* is random, since ~~we use a realization of the random variable~~ until we plug in the actual value of the realization ($\bar{X}$) to compute it.

$\bar{X} \pm z_{\frac{\alpha}{2}} \cdot SD(\bar{X})$ is rdm, but $\bar{x} \pm z_{\frac{\alpha}{2}} \cdot SD(\bar{X})$ is not

- What fraction of such intervals (each interval computed from a random sample of data) will cover the true value μ?

- This *coverage probability* **(before we actually collect the data)** is called the ***confidence level*** of the confidence interval.

If Y is the # of successful 95% C.I, in 100 independently computed C.I.

$Y \sim Bin(100, 0.95)$

95% CI, $1-\alpha = 0.95$
$\alpha = 0.05$

Confidence Intervals

1. Which would be wider : a 99% CI or a 95% CI?

2. What about a 90% CI? 68%?

3. The _______ the confidence level, the _______ the interval

   higher / (lower)   wider / (narrower)

4. This does not make sense! Why are we using a normal distribution when the sample consists of Bernoulli random variables?

   We are applying the CLT

5. What is the chance that the population %, **p**, is in the interval (18%, 26%)?

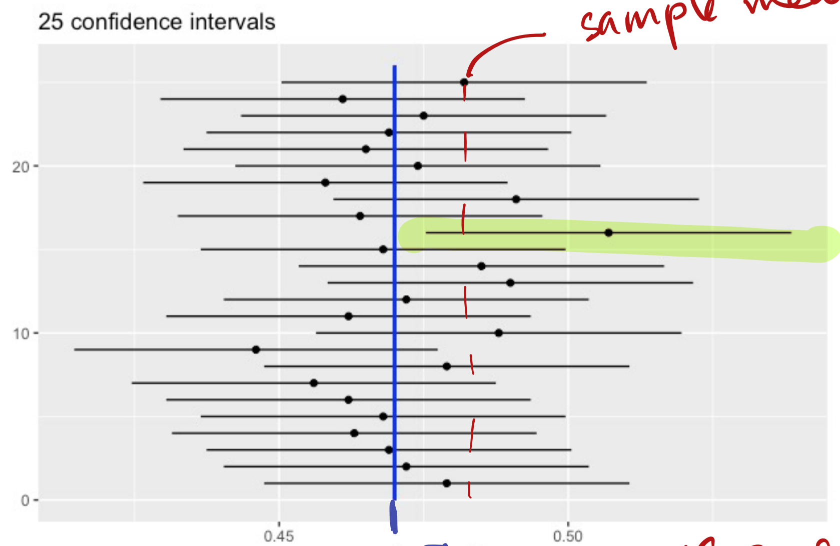
Probability of coverage

$X_1, X_2, \ldots X_{25}$

$E(X_k) = p$, $SD(X_k) = p(1-p)$

- We draw 25 samples (sample size 100) from a Bernoulli distribution with p=0.47.
- Construct a 95% CI from each sample.
- How many intervals covered the blue line? How many did you expect? 24 $= (0.95)(25) = 23.75 \approx 24$
- What is the *chance* that each CI will cover the true p (before you plug in #s)? 0.95
- If X=number of successful intervals, what is the distribution of X? $= Bin(25, 0.95)$
- Why are the centers different? Are the widths the same?

sample means ($\hat{p}$'s) observed values



$\vdash \bar{X} \dashv$

Width

$= (1.96) \times SD(\hat{p})$

$= (1.96) \times \frac{\sqrt{p(1-p)}}{\sqrt{100}}$

If $p = 0.47$ was used, widths are the same.

If $\hat{p}$ was used, widths are not quite the same (to approx p)

Margin of error

- We have a confidence interval. Now we want to keep the **same confidence level**, but want to improve our accuracy. For example, say our *margin of error* is 4 percentage points, and we want it to be 1 percentage point. What should we do?

A. increase width of CI 4 times by increasing SD

B. Decrease width of CI by increasing *n* by 4 times

C. Decrease width of CI by increasing *n* by 16 times ← square root law

Exercise work the math out!!

$$\text{old m.e.} = z_{\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n_{old}}} = 4$$

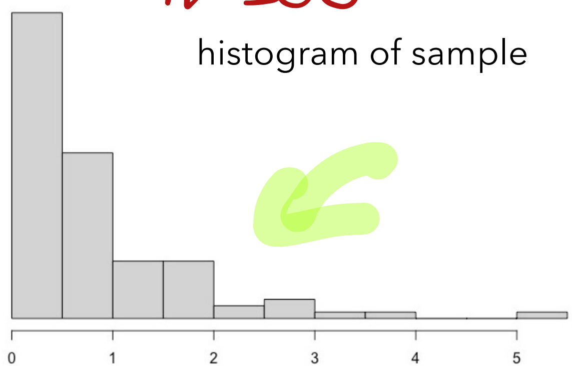
$$\text{new m.e.} = z_{\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n_{new}}} = 1$$

Comparison with bootstrap CI

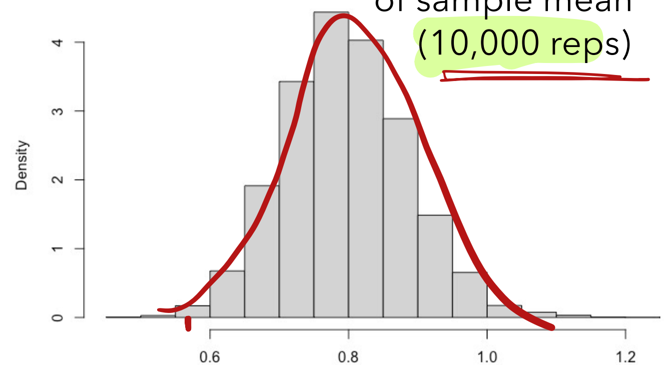
- How do you create a bootstrap CI for the population mean?

$n = 100$

histogram of sample



Bootstrapped histogram of sample mean (10,000 reps)



250^{th} sorted # 9750^{th} sorted #

Take B (large #) "resamples" of size n and compute the sample mean each time

2.5th percentile   97.5th percentile.