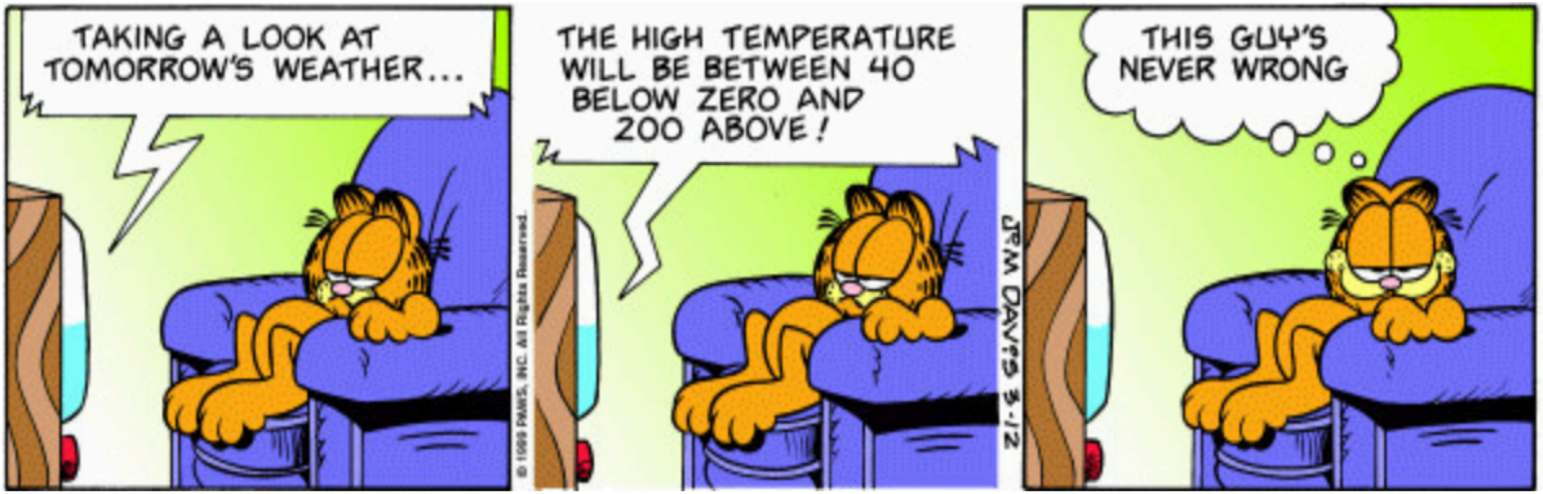
Stat 88: Probability & Mathematical Statistics in Data Science



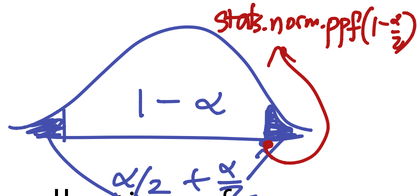
Lecture 31 : 4/9/2021

Section 9.4

Confidence Intervals

Confidence intervals

$(1-\alpha)100\%$ C.I.



- 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.

- 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.

r.v. does not have randomness ($z_{\frac{\alpha}{2}} \cdot SD(\bar{X})$) margin of error

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

sample mean is the center of your C.I.

- 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 $1-\alpha$

$$P\Big(\underset{\text{constant}}{\mu} \in \big(\bar{X} - (z_{\frac{\alpha}{2}} \cdot SD(\bar{X})),\ \bar{X} + (z_{\frac{\alpha}{2}} \cdot SD(\bar{X}))\big)\Big) = 1-\alpha$$

- 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.

Dealing with proportions

- 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.

$N = 10^6$, $n = 400$, observed value of $\bar{X} = \hat{p} = 0.22$

$\hat{p}$ is what we often write instead of $\bar{X}$ if the original X_k's are 0 or 1

Even though we have a SRS, $n << N$, so pretend that we are sampling w/ replacement so that we can use the C.L.T.

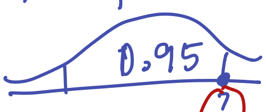
$X_1, X_2, \ldots, X_{400} \sim$ Bernoulli(p). Need to estimate p.

$SD(X_k) = \sigma = \sqrt{p(1-p)}$

$E(\bar{X}) = p$, $SD(\bar{X}) = \dfrac{\sqrt{p(1-p)}}{\sqrt{n}}$, 95% C.I margin of error

m.e. $= z_{\frac{\alpha}{2}} \cdot SD(\bar{X})$

m.e $= 2 * \dfrac{\sqrt{(0.22)(0.78)}}{\sqrt{\cdots}} = 0.0414$



$\sqrt{400}$

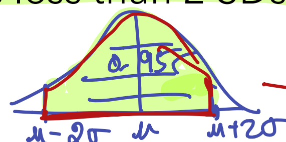
$95\% \cdot C.I = \bar{X} \pm 2 \cdot SD(\bar{X}) \longrightarrow 95\%\ C.I = 0.22 \pm 0.0414$

$= (0.1786, 0.2614)$

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

Section 9.4: Interpretation

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

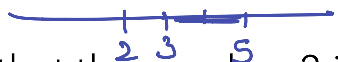


$E(\bar{X}) = \mu$ true

by the CLT, $P(\mu - 2\sigma < \bar{X} < \mu + 2\sigma) \sim 0.95$

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

- Which object is random in each of these sentences?



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

No

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

No

(rounding 17.86 & 26.14)

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%?

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

Confidence intervals: What is random?

- 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 ($\bar{X}$) to compute it.

  observed value

- 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.

X = # of successful 95% C.I for a true mean. out of 100 C.I.

$X \sim \text{Binomial}(100, 0.95)$

Confidence Intervals

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

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

higher / lower
wider / narrower

3. The higher the confidence level, the wider the interval

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

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

0 or 1

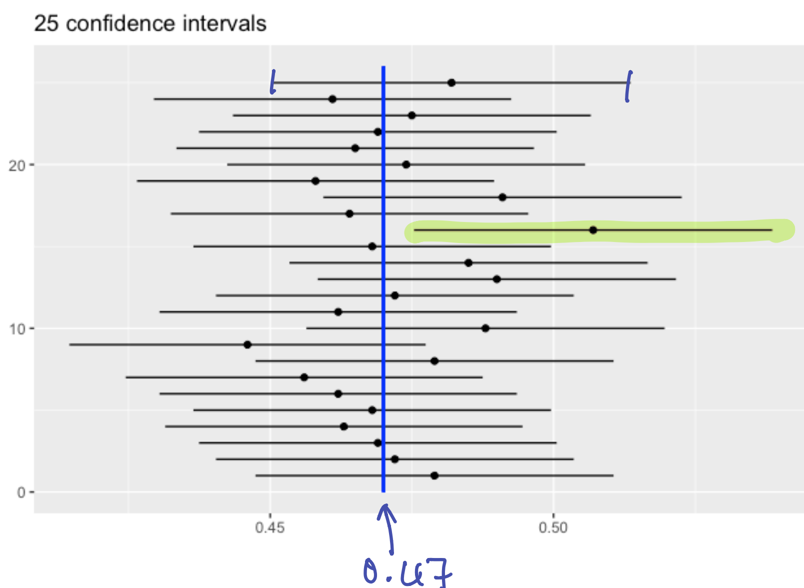
Probability of coverage

- We draw 25 samples (sample size 100) from a Bernoulli distribution with p=0.47.
- Construct a 95% CI from each sample. 25 . 95% C.I
- How many intervals covered the blue line? How many did you expect? 23.75 $= 25 * 0.95$
- 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?

$2 * z_{\frac{\alpha}{2}} \cdot SD(\bar{X})$

$\frac{\sqrt{p(1-p)}}{\sqrt{25}}$

$= \frac{\sqrt{(0.47)(0.53)}}{\sqrt{25}}$



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

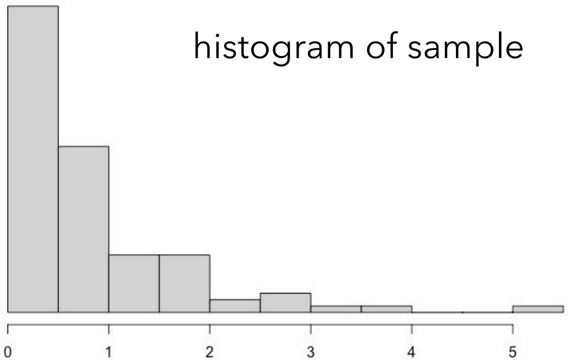
B. Decrease width of CI by increasing *n* by 4 times

C. Decrease width of CI by increasing *n* by 16 times

Comparison with bootstrap CI

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

histogram of sample



Bootstrapped histogram
of sample mean
(10,000 reps)