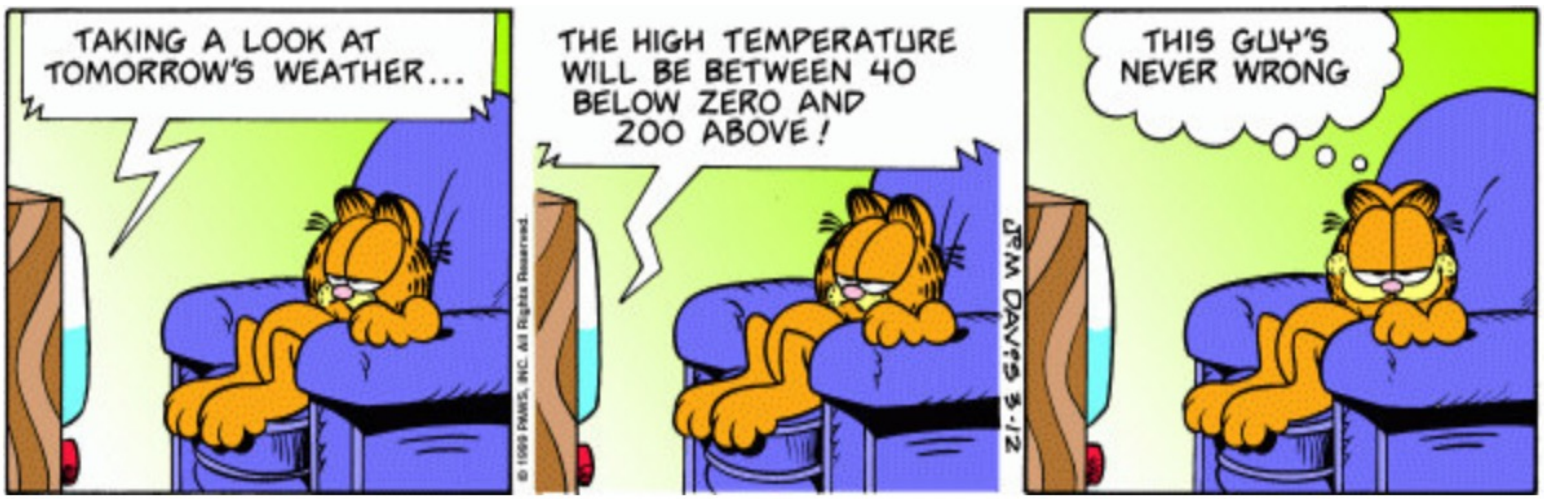
Stat 88: Prob. & Mathematical Statistics in Data Science



Lecture 22 : 4/11/2022

Section 9.3, 9.4

Confidence intervals

Goal: Estimating a parameter

- Say we have a population whose average, μ, we want to estimate
- How would we do it? We could draw one data point X_1 and use it to estimate μ. Do you think this is a good method of estimation? If not, why not?

No! One data point does not make a good sample

- What about if we draw a sample of size 2: X_1, X_2 where each of the X_i have expectation μ? Is this better? Can we use the average of these two?

Not really.

$X_1, X_2, \ldots X_n$ s.t. $E(X_k)=\mu$, $Var(X_k)=\sigma^2$

- We generally use a larger sample, say n is a large number and we draw an iid sample $X_1, X_2, \ldots, X_n$. Why is this a better idea? The expectation of each of the X_i is μ, so the expectation of the sample mean is also μ. But this was true even for $n = 2$. Why use larger n?

We need a large n to either apply the CLT or get a representative sample to use the bootstrap method.

1. Suppose I draw a sample of size 5 from a population & compute $\bar{X}$ (A_5)

2. Repeat step 1 one thousand times & plot the histogram of these $\bar{X}$'s (A_5)

Q Will this histogram be bell shaped b/c of the C.L.T?

Using $\bar{X}$ to estimate μ

- $\bar{X}$ is an unbiased estimator of μ (what does that mean?) $\mathbb{E}(\bar{X}) = \mu$
- If we also know that each of the X_k had SD σ, what can we say about $SD(\bar{X})$?

$SD(\bar{X}) = \sigma/\sqrt{n}$

- What does the Central Limit theorem say about the sample mean?

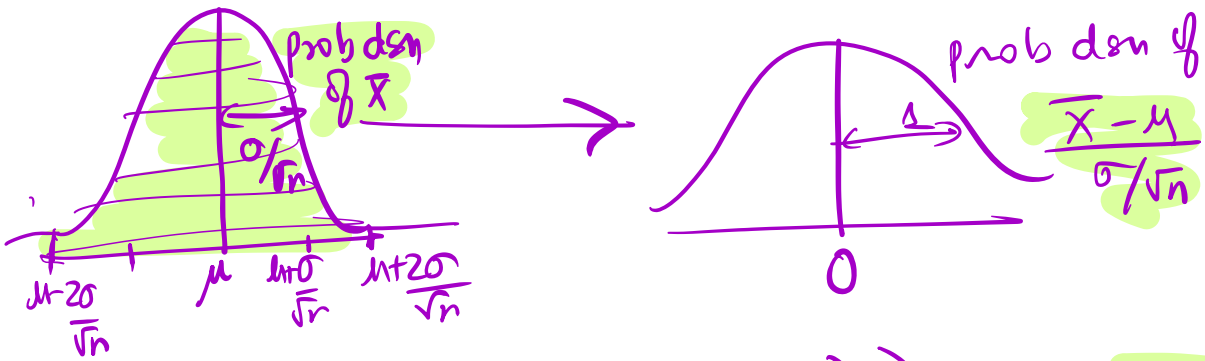
as long as n is large enough, the dsn of $\bar{X}$ will be approx bell shaped

- We will use the CLT and the sample mean to define a random interval (why is it random?) that will *cover* the true mean with a specified probability, say 95%
- Based on data from a *random sample*, we will construct an interval of estimates for some unknown (but fixed) population parameter.

Interval is random b/c it depends on the sample mean $\bar{X}$ which is a r.v.

$\bar{X} \overset{\text{approximately}}{\approx} \text{Normal}\left(\mu, \frac{\sigma^2}{n}\right)$, $Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \approx N(0,1)$

Goal: Estimate μ w/ a $\pm$ #.



$$P\left(\bar{X} \in \left(\mu - 2\frac{\sigma}{\sqrt{n}},\ \mu + 2\frac{\sigma}{\sqrt{n}}\right)\right) \approx 0.95$$

($\bar{X} \sim N(\mu, \sigma/\sqrt{n})$ (68%-95%-99.7% rule applied to $\bar{X}$, approx Normal))

$$\Rightarrow P\left(\mu - 2\frac{\sigma}{\sqrt{n}} < \bar{X} < \mu + 2\frac{\sigma}{\sqrt{n}}\right) \approx 0.95$$

$$\Rightarrow P\left(-2\frac{\sigma}{\sqrt{n}} < \bar{X} - \mu < +2\frac{\sigma}{\sqrt{n}}\right) \approx 0.95$$

$$\Rightarrow P\left(-\bar{X} - 2\frac{\sigma}{\sqrt{n}} < -\mu < -\bar{X} + 2\frac{\sigma}{\sqrt{n}}\right) \approx 0.95$$

multiply by -1

$$P\left(\bar{X} + 2\frac{\sigma}{\sqrt{n}} > \mu > \bar{X} - 2\frac{\sigma}{\sqrt{n}}\right) \approx 0.95$$

$$P\left(\underset{\text{rdm}}{\bar{X}} - \underset{\text{const}}{2\frac{\sigma}{\sqrt{n}}} < \underset{\text{const}}{\mu} < \bar{X} + \underset{\text{const}}{2\frac{\sigma}{\sqrt{n}}}\right) \approx 0.95$$

Interval with RANDOM endpoints $\rightarrow$ RandomInterval.

Confidence intervals

- In the previous slide, we derived an ***approximate 95% Confidence Interval for the population mean μ***

The C.I is our "net" until we plug in an observed value for $\bar{x}$,

- Why is the interval random?

the interval is random b/c the endpoints are random $\bar{X} \pm 2\frac{\sigma}{\sqrt{n}}$

- 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.

$C \cdot \sigma/\sqrt{n}$

- 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.
- 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 $1-\alpha$

Example

$\mathbb{E}(X_k) = \mu \qquad X_1, X_2, \ldots X_{64}$

- A population distribution is known to have an SD of 20. The average of an iid sample of 64 observations is 55. What is your 95% confidence interval for the population mean?

$SD(X_k) = 20$

$\bar{X} = A_{64}$, $\mathbb{E}(\bar{X}) = \mu$ (unknown), $SD(\bar{X}) = \frac{20}{\sqrt{64}} = \frac{20}{8}$

$\bar{x}$ = observed value of $\bar{X}$ is 55

95% Confidence interval is given by $\bar{x} \pm 2 \cdot \frac{\sigma}{\sqrt{n}} = 55 \pm 2 \cdot \frac{20}{8} = 55 \pm 5$

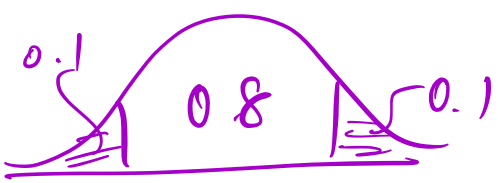
Observed C.I $= (50, 60)$: 95% C.I for μ.

Confidence levels

- 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.

- What would we do differently if we wanted a 68% CI? 99.7% CI?

- What about an 80% CI? 99% CI?



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.

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

- 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.

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 is many smaller than N, so can use the CLT. Let p = percentage of unemployed people in the town

$\hat{p}$: sample %, $X_1, X_2 \cdots X_{400} \sim$ Bernoulli(p)

$\hat{p} = \frac{S_{400}}{400}$ = sample average

95% C.I = $\hat{p} \pm 2 \cdot \frac{\sigma}{\sqrt{n}}$

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

Answer 0.22 ± 0.0414

Example

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