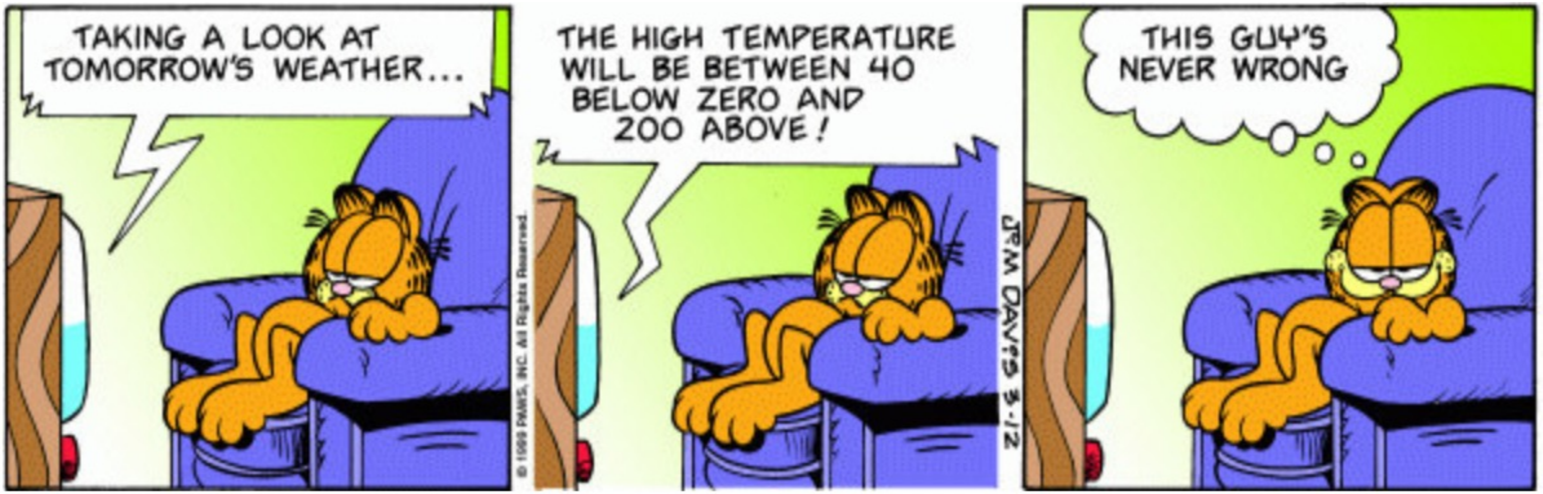
Stat 88: Probability & Mathematical Statistics in Data Science



Lecture 30 : 4/7/2021

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. It's lousy

  $n = 1$ too small.

- 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?

  Still not good enough.

- 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? ← need large n to apply CLT

1. Say we draw a sample of size 5_n from some population & compute the $\bar{X}$
2. Repeat step 1 one thousand times & plot the histogram of the sample means. Will this histogram be bellshaped by CLT?

(NO! sample size = 5, so no guarantee that histogram will be bell-shaped).

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 $SD(\bar{X})$?

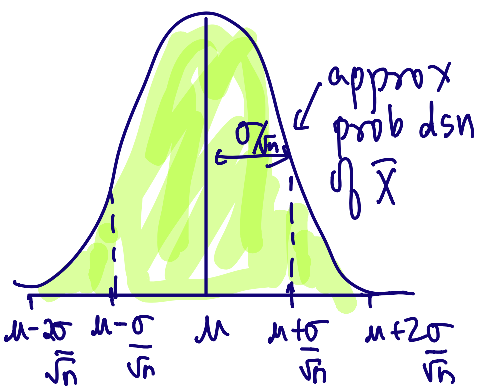
  $X_1, X_2, \ldots X_n$, $E(X_k) = \mu$, $SD(X_k) = \sigma$

- What does the Central Limit theorem say about the sample mean? $SD(\bar{X}) = \frac{\sigma}{\sqrt{n}}$

  The dsn of the sample mean $\bar{X}$ will be approximately (n is large) normal (bell shaped). $Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}}$ is approx std normal.

- 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%

$\bar{X}$ is approx. normal with mean μ, SD $\frac{\sigma}{\sqrt{n}}$



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

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

$\bar{X}$ ↑ random

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

subtract $\bar{X}$

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

multiply everything by -1

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

Constant but unknown

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

constant, perhaps known, or estimated from data.

random variable

$$\left(\bar{X} - \frac{2\sigma}{\sqrt{n}}, \bar{X} + \frac{2\sigma}{\sqrt{n}}\right) \longleftarrow$$

random interval (because the endpoints interval) that has a prob. of 0.95 of capturing the true mean μ.

(approximate) 95% CONFIDENCE INTERVAL

Confidence intervals

- In the previous slide, we derived an ***approximate 95% Confidence Interval for the population mean μ*** net is C.I.

- Why is the interval random? Because endpoints are functions of sample mean which is a r.v.

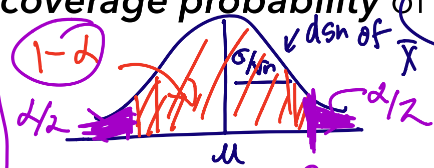
- A *confidence interval* is an interval on the real line, that is, a collection of values, that are plausible estimates for the true mean μ.

- 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: $\boxed{\bar{X} \pm \textit{margin of error,}}$ 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 $1-\alpha$

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



$?? \quad z_{\frac{\alpha}{2}}$ area to RIGHT is $\frac{\alpha}{2}$

Example

$\sigma = 20$

- 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?

$n = 64, \sqrt{n} = 8$

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

$P(\mu \in \mathcal{I}) = 0.95$

approximate 95% C.I. for μ $\left(\bar{X} - \frac{2\sigma}{\sqrt{n}}, \bar{X} + \frac{2\sigma}{\sqrt{n}}\right)$

Plug in observed value of $\bar{X}$

Observed C.I (realized): $\left(55 - 2 \cdot \frac{20}{8}, 55 + 2 \cdot \frac{20}{8}\right)$

$= (50, 60)$

0   50   60

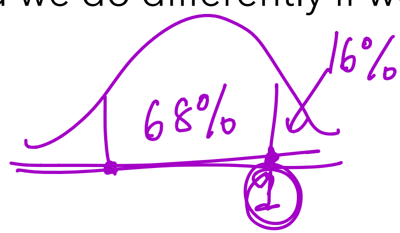
Confidence levels

- 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.

stats.norm.ppf$(1-\frac{\alpha}{2})$

$\bar{x} \pm z_{\frac{\alpha}{2}} \cdot SD(\bar{X}) = \bar{x} \pm z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}$

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



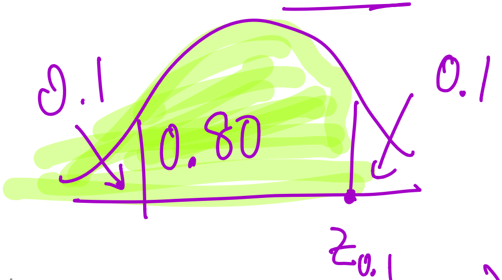
$\bar{x} \pm 1 \cdot \frac{\sigma}{\sqrt{n}}$

$\bar{x} \pm 3 \cdot \frac{\sigma}{\sqrt{n}}$

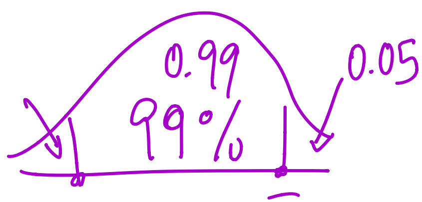
$\bar{x} = 55,\ \sigma = 20,\ n = 64$

80% C.I: $55 \pm (1.28)(\frac{20}{8})$

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



stats.norm.ppf(0.9) ≈ 1.28

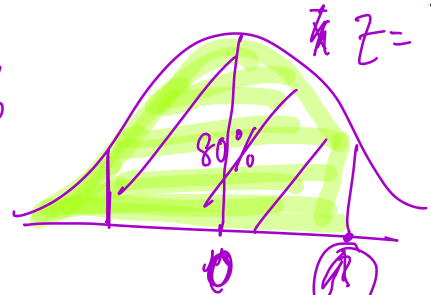


stats.norm.ppf(0.995)

We want an 80% C.I

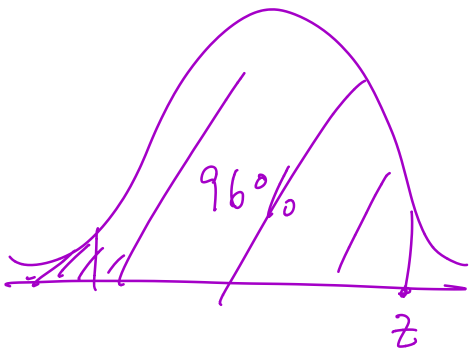
$P(\mu \in \underbrace{C.I.}_{\text{random}}) = 80\%$

$z = \frac{\bar{X}-\mu}{\sigma/\sqrt{n}}$



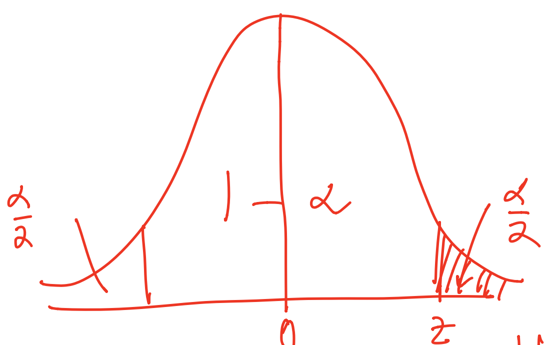
$\bar{X} \pm z \cdot \frac{\sigma}{\sqrt{n}}$

stats.norm.ppf(0.9)



96% C.I

stats.norm.ppf(0.98)



Want z so that (1−α) prob or area is in middle

by the symmetry of normal curve, $\frac{\alpha}{2}$ is in the tails.

So to use python function, I need area to left of this z.

Area to left = $(1-\alpha) + \frac{\alpha}{2} = 1 - \frac{\alpha}{2}$