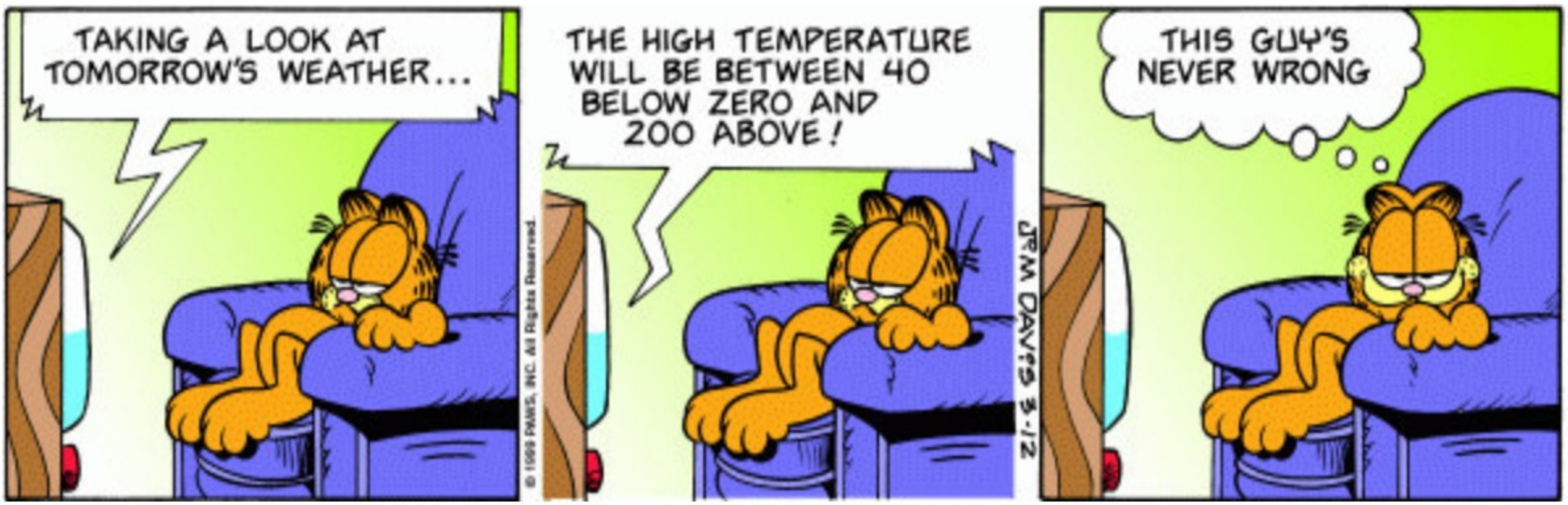
Stat 88: Prob. & Mathematical Statistics in Data Science



Lecture 31: 4/10/2024

Section 9.1, 9.2

Confidence intervals

Goal: Estimate a population parameter.

Using $\bar{X}$ to estimate μ

$X_1, X_2, \ldots X_n$ iid: μ, σ^2

- $\bar{X}$ is an unbiased estimator of μ (what does that mean?) $E(X_k) = \mu$
- 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?

For n large enough $\bar{X} \approx$

- 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.

$\approx$ "approximately distributed as"

$X \sim Bin(n=100, p=0.5)$ $\quad \sigma = \sqrt{npq} = \sqrt{100 \times \frac{1}{2} \times \frac{1}{2}} = 5$

$\boxed{X \approx N(\mu=50, \sigma=5)}$ ← example

Given an iid sample $X_1, X_2, \ldots X_n$

By CLT, $\bar{X} = \frac{X_1 + X_2 + \cdots + X_n}{n}$ is

approx normal with $E(\bar{X}) = \mu$, $Var(\bar{X}) = \frac{\sigma^2}{n}$

$$Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \approx N(0,1) \quad \text{(by CLT)}$$

$$P(-2 \le Z \le 2) \approx 0.95$$

$$P(|Z| \le 2) = 0.95$$

↑ dist b/w Z & $0 \le 2$

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

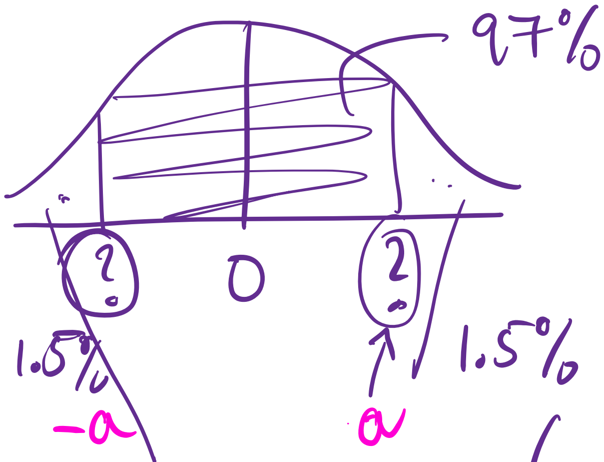
$$P(-2 \le Z \le 2) \approx 0.95$$

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

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

$\ge \qquad \ge$

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



$\Phi^{-1}(0.015) = -a$

stats.norm.ppf(0.015)

$P\left(\bar{X} - a\frac{\sigma}{\sqrt{n}} \le \mu \le \bar{X} + a\frac{\sigma}{\sqrt{n}}\right)$

$\approx 0.97 = 97\%$

$\left(\bar{X} - 2\frac{\sigma}{\sqrt{n}}, \bar{X} + 2\frac{\sigma}{\sqrt{n}}\right)$ is called 95% CI

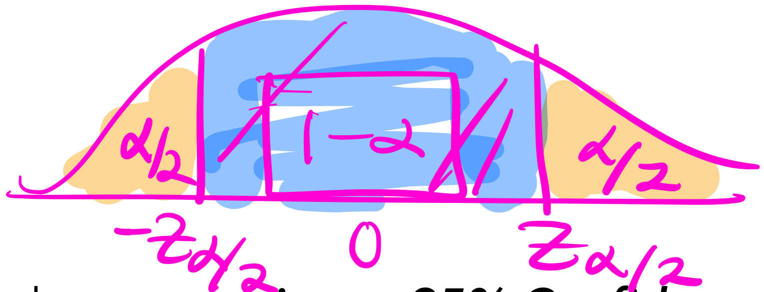
Once I plug in my observed value for $\bar{X}$ (call this $\bar{x}$)

$\left(\bar{x} - 2\frac{\sigma}{\sqrt{n}}, \bar{x} + 2\frac{\sigma}{\sqrt{n}}\right)$ 95% CI for μ.

Example Say the ht of randomly selected Data 88 student is 67", σ = 5"

Take a sample & construct a 95% CI to get (60", 70")

Confidence intervals



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

100(1−α)% conf. Int for μ

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

- Why is the interval random?

- ***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.

width of interval = 2 * 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.

- 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

$\sigma = 20 \quad \mu = ?$

- 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?

$n = 64 \quad \bar{x} = 55$

$\boxed{1.96}$ = z-score for 95% CI

$$I = \left(55 - 2 * \frac{20}{8},\ 55 + 2 * \frac{20}{8}\right) = (55-5, 55+5) = (50, 60)$$

What is the prob that this interval I contains μ. (0 or 1)

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?

Exercise

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.

$X_1, X_2 \ldots X_{400}$ pretend Bin.

When the X_k's are 0 or 1, $\bar{X}$ is a proportion & $\bar{X}$ is just called $\hat{p}$

$X \sim \text{Bin}(400, p)$?

$\mathbb{E}(\bar{X})$ $\text{Var}(\bar{X})$? ?

$\hat{p} = 0.22$ $\quad E(X) = np \quad Var(X) = npq$

$X = X_1 + X_2 + \cdots + X_{400}$

$E(X_k) = p = ? \quad Var(X_k) = p(1-p) \approx \hat{p}(1-\hat{p})$

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

Exercise Complete this

$$\hat{p} \pm 2 * \frac{\sigma}{\sqrt{n}}$$

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

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.