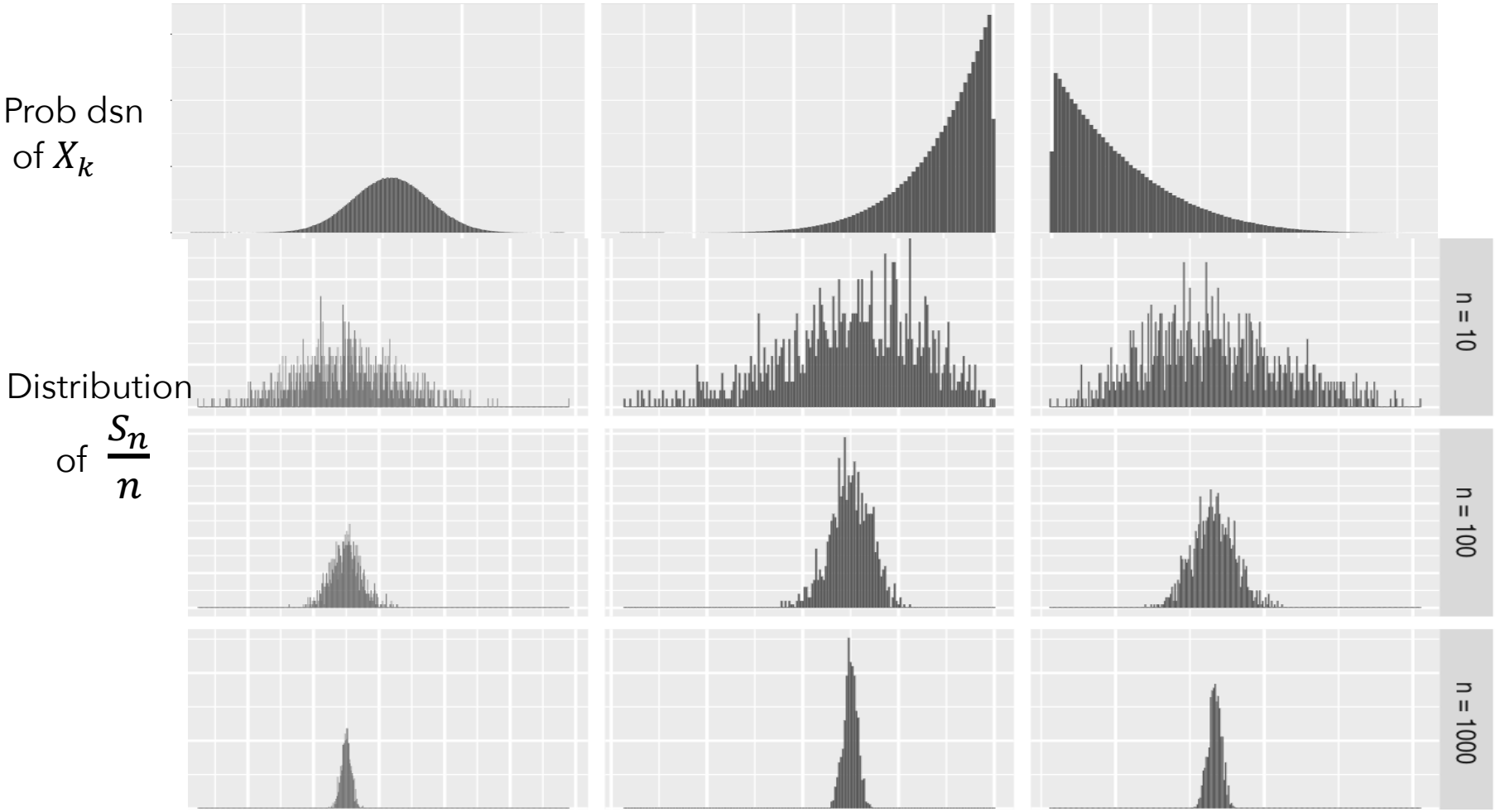
Stat 88: Prob. & Mathematical Statistics in Data Science



Lecture 30 PART 1: 4/8/2024

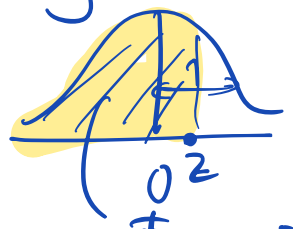
Finishing up chapter 8 and the Central Limit Theorem

Standard normal cdf, symmetries, percentiles

If you think of the standard normal curve as a probability histogram, then it is natural to think of areas under the curve as probabilities. In particular, the function Φ defined by:

If the dsn of Z is given by the N(0,1)

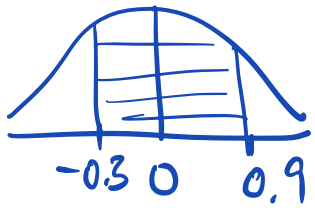
(Go to Jupyter)



$\Phi(z) = P(Z \le z)$

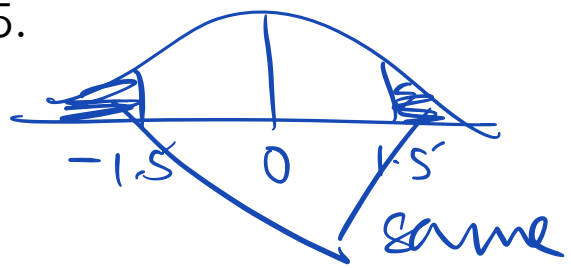
(a) Verify empirical rule using stats.norm.cdf()

(b) Find the probability that Z is between -0.3 and 0.9



$\Phi(0.9) - \Phi(-0.3)$

(c) Find the probability that Z is outside -1.5 and 1.5.

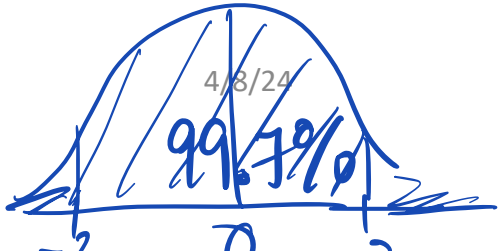


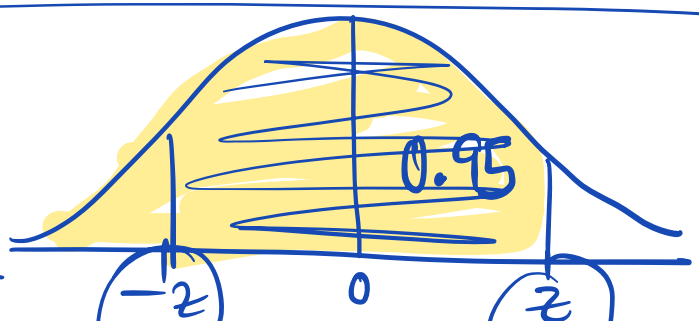
$2 * \Phi(-1.5)$

(d) Find z such that $\Phi(z) = 0.95$

stats.norm.ppf(0.95)

(e) Find z so that the area in the **middle** is 0.95 ($\Phi(z) - \Phi(-z) = 0.95$)





stats.norm.ppf(0.975)

stats.norm.ppf(0.025)

Example

- Suppose that each of 300 patients has a one in three chance of being helped by a treatment, independently of its effect on other patients. Find the probability that at least half the patients are helped by the treatment.

Approximate

Let X = # of patients helped by the treatment

$X \sim Bin(300, \frac{1}{3})$ $\quad X \approx N(100, Var(X))$

$\boxed{P(X \geq 150) \approx 0}$

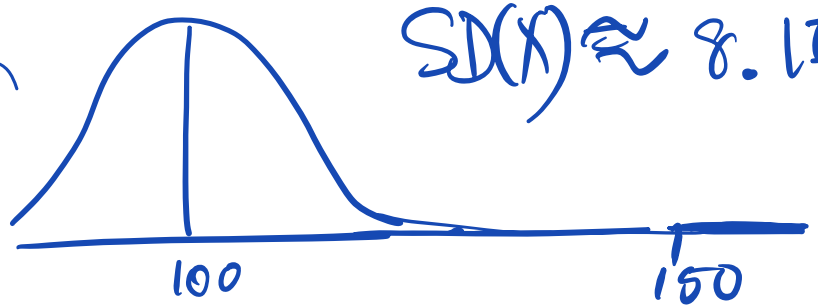
$Var(X) = npq$

$= 300 \times \frac{1}{3} \times \frac{2}{3} \approx 66.67$

$X \approx N(100, (8.2)^2)$

$SD(X) \approx 8.17 \approx 8.2$

$\rightarrow Z = \frac{X - 100}{8.2}$



$Z = \frac{150 - 100}{8} \geq 6$

$P(X \geq 150) = P(X \geq 100 + 50)$

The Central Limit Theorem

- Suppose that $X_1, X_2, \ldots, X_n$ are iid with mean μ and SD σ
- Let $S_n = X_1 + \cdots + X_n$ be the sample sum, and $A_n = \frac{S_n}{n}$ be the sample mean

- Then the distribution of S_n (and A_n) is ***approximately normal*** for large enough n.

- For S_n, the distribution is approximately normal (bell-shaped), centered at $\boldsymbol{E(S_n) = n\mu}$ and with spread given by by $\boldsymbol{SD(S_n) = \sqrt{n}\,\sigma}$.

- For A_n, the distribution is centered at $\boldsymbol{E(A_n) = \mu}$ with spread $\boldsymbol{SD(A_n) = \sigma/\sqrt{n}}$

Normal approximations : standard units

- Let X be any random variable, with expectation μ and SD σ, consider a new random variable that is a linear function of X, created by shifting X to be centered at 0, and dividing by the SD. If we call this new rv X^* then X^* has expectation __0__ and SD __1__.
- $E(X^*) =$ 0
- $SD(X^*) =$ 1

$$Z = \frac{X - E(X)}{SD(X)}$$ in Standard Units

- This new rv does not have units since it measures how far above or below the average a value is, in SD's. Now we can compare things that we may not have been able to compare.

- Because we can convert anything to standard units, ***every normal curve is the same.***

How to decide if a distribution could be normal

- Need enough SDs on both sides of the mean.

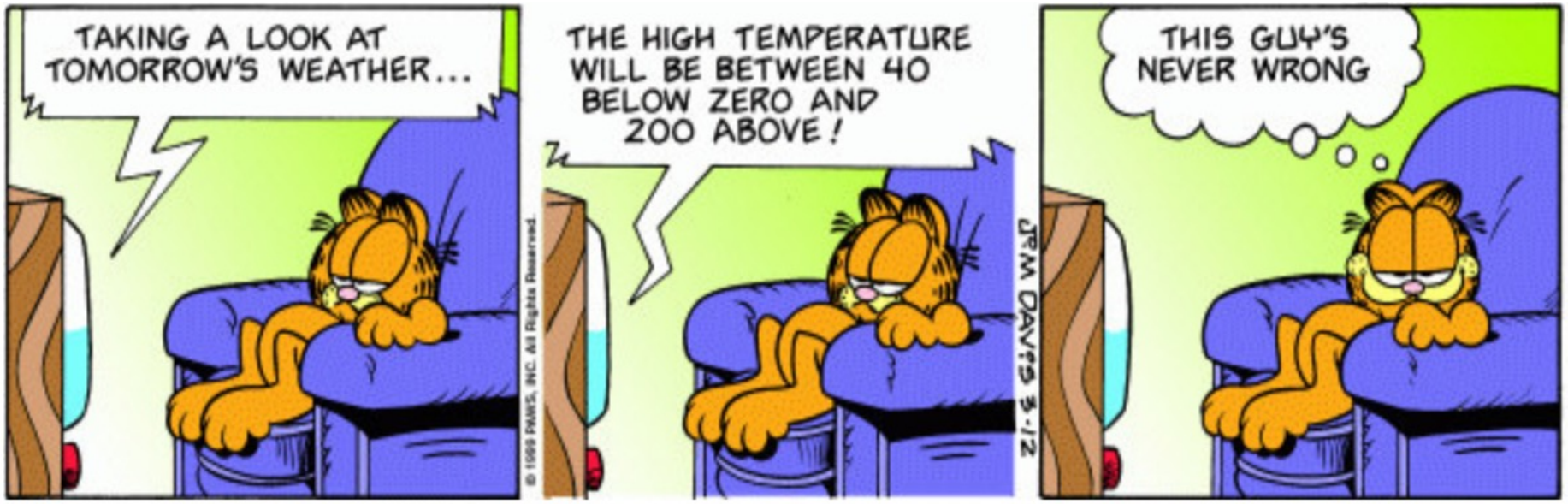
- In 2005, the murder rates (per 100,000 residents) for 50 states and D.C. had a mean of 8.7 and an SD of 10.7. Do these data follow a normal curve?

- If you have indicators, then you are approximating binomial probabilities. In this case, if n is very large, but p is small, so that np is close to 0, then you can't have many sds on the left of the mean. So need to increase n, stretching out the distribution and the n the normal curve begins to appear.

- If you are not dealing with indicators, then might bootstrap the distribution of the sample mean and see if it looks approximately normal.

Read last section in Ch. 8

Stat 88: Prob. & Mathematical Statistics in Data Science



Lecture 22 PART 2 : 4/8/2024

Section 9.1, 9.2

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. b/c too much variability

- 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?

- 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?

$X_1, \ldots, X_n \sim \mu, \sigma^2$

$\mathbb{E}(\bar{X})$ or $\mathbb{E}(A_n) = \mu,$

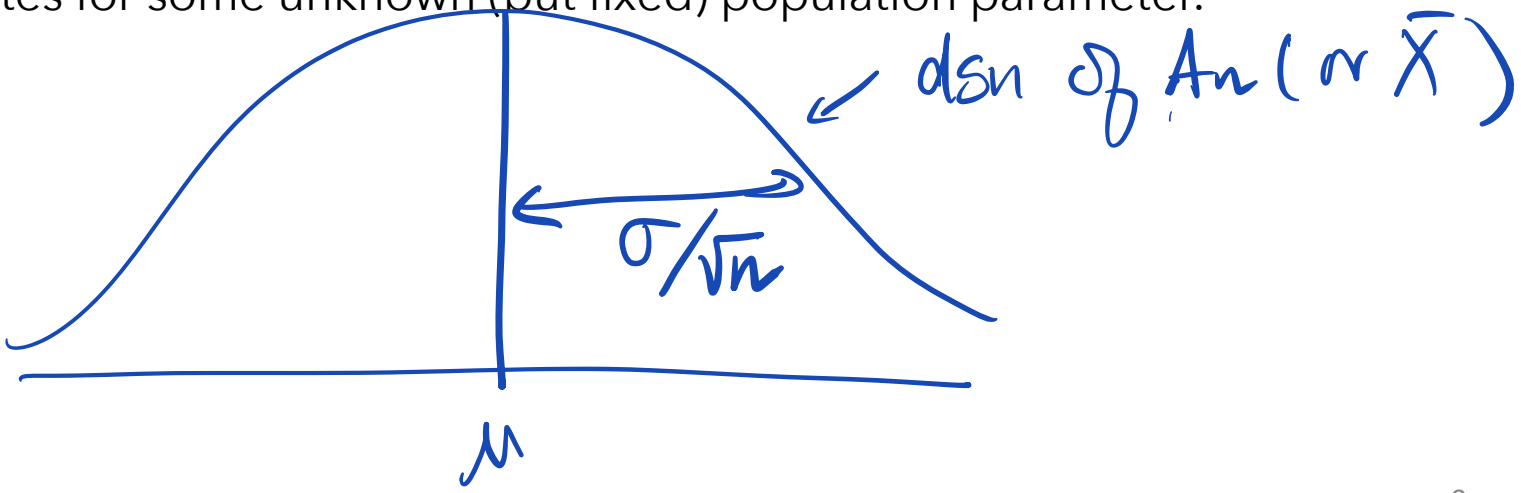
Using $\bar{X}$ to estimate μ

$X_1, \ldots, X_n$ iid

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

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

- 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.



Instead of A_n, lets consider

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

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

$$P\left(-2 \le \frac{A_n - \mu}{\sigma/\sqrt{n}} \le 2\right) = 0.95$$

$$P\left(-2 * \frac{\sigma}{\sqrt{n}} \le A_n - \mu \le 2 * \frac{\sigma}{\sqrt{n}}\right) = 0.95$$

$$P\left(-2 * \frac{\sigma}{\sqrt{n}} - A_n \le -\mu \le 2 * \frac{\sigma}{\sqrt{n}} - A_n\right)$$

$$\Rightarrow P\left(2 * \frac{\sigma}{\sqrt{n}} + A_n \ge \mu \ge -2\frac{\sigma}{\sqrt{n}} + A_n\right)$$

$$= 0.95$$

$$\Rightarrow P\left(A_n - 2\frac{\sigma}{\sqrt{n}} \le \mu \le A_n + 2\frac{\sigma}{\sqrt{n}}\right)$$

$$= 0.95$$

What is random?

Confidence intervals

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

- 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.

- 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 $1 - \alpha$