PHP 2510 Principles of Biostatistics & Data Analysis

Week 6: Sampling Distributions

OUTCOMES

After this week's classes, along with the required readings (CHIHARA Chapter 4; SPEEGLE Chapter 5, Sections 5.4+), you should be able to:

- Determine how well-defined functions of RVs behave
- Define the concepts of a sampling distribution and the standard error of a statistic
- Explain the Central Limit Theorem and its importance for making inferences
- Differentiate between the standard deviation of a population and the standard error of the mean

This Week's Plan

1 — Functions of RVs

- Motivation
- Convolution & Order Statistics

2 — Sampling Distributions

- Sample Mean

3 — Inference

- Paradigm
- Central Limit Theorem
- Sample Mean (again)
- Terminology

4 — In Action

- Simulations
- Practice Problems

Population

Has some data generation process
Has some (unknown) distribution for the outcome of interest

Sampling or Reporting Process

Data (in hand!)

Transformation process

S Summary Statistic

To use S to make claims about the population, we need to understand each of these steps

Examples

1. I want to know if a new medicine is effective. I try it on a few [SAMPLING PROCESS] patients and report their average outcome [TRANSFORMATION PROCESS]. For example, BMI change after a new weight loss drug.

2. I create a study that asks participants [SAMPLING PROCESS] to write down how many migraines they experience in a week. After 6 months, I want to study the total number of migraines [TRANSFORMATION PROCESS]

3. Police officers provide breathalyzer tests to drivers suspected [SAMPLING PROCESS] of operating under the influence. Their protocol is to take 3 readings. They report the average to control for measurement error [REPORTING PROCESS]. How often do they detect illegal activity [TRANSFORMATION PROCESS]?

Mixtures of Random Variables are Random Variables themselves

- Sum of n independent bernoulli (p) → binomial(n,p)
  - By definition
  - Convince yourself with the formulae
- Normal + Normal (if independent OR jointly Normal) → Normal
  - Let's convince ourselves with R simulations
  - How might this relate to the drug treatment example?
  - Another special fact about Normal: if X~N then aX+b is also Normal
- Poisson + Poisson (if independent) → Poisson
  - Let's prove it!* (then confirm in R)
  - How does this relate to the migraine example?

This is sometimes called convolution

*for this proof, we will need to use the binomial theorem

Part 1

Order Statistics

A police officer measures the blood alcohol content of a person who is suspected to be driving under the influence 3 successive times. You are worried some police officers report only the minimum, while others report only the maximum. For each approach:

1. Draw a picture of what this reporting process does to our data
2. Derive the cdf generally
3. In R, simulate the pdf when
   a. BAC ~ U[0,1]
   b. BAC ~ beta[alpha = 2, beta = 30]
4. Can you derive the distribution of 3a by hand?

The sampling/reporting process can really influence our data!

How does the mean of a simple random sample behave?

Let's prove theorem A.7 in CHIHARA (pg 498)

For iid samples from X, a RV with mean μ, var σ^2

$\bar{X}$ has mean μ, variance σ^2/n

Relating this back to the original graphic:

- What assumptions did we place on the population DGP?
- What assumptions are we making about the sampling process?
- What assumptions are we making about the transformation process?

The sample mean is special. Do you agree? Explain

Population

Has some data generation process

Has some (unknown) distribution for the outcome of interest

Sampling or Reporting Process

Data (in hand!)

Transformation process

S Summary Statistic

To use S to make claims about the population, we need to understand each of these steps

BIG PICTURE

A statistic is a random variable, with some distribution, that depends on unknown population parameters

Using our data to learn about the likely values of the population parameters is ***statistical inference***

But we must understand our sampling and transformation process, and make/justify some assumptions

Inference Paradigm

1. Assume something about the true population
2. Assume something about your sampling/reporting process
3. Postulate a statistic; prove it is close to the population parameter you care about
4. Make claims about the population parameter based on the value we actually observe in our dataset

The Sample Mean (for iid samples)

We proved that it’s an unbiased and consistent estimator of the (true) population mean

- What does this mean?
- Why is this valuable to us?

→ For any n, we know its expectation and variance, but not it’s (full) distribution

What happens with n is large? CENTRAL LIMIT THEOREM

Part 3

CLT Intuition

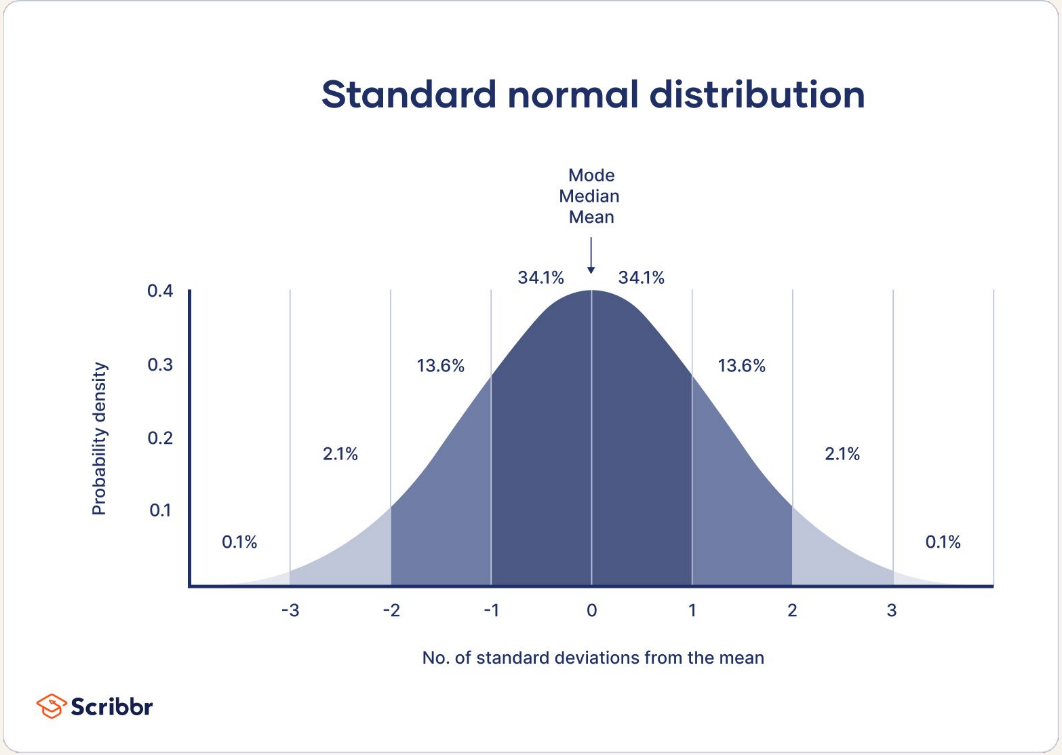
Inference Paradigm - the sample mean

1. Population has some true mean μ and variance σ^2
   - not necessary independent parameters
2. Take an iid sample and calculate the sample mean ($\bar{X}$)
3. Per CLT (for large n), $\bar{X}$ follows a $N(\mu, \sigma^2/n)$
   - What does this mean?
4. If the value of $\bar{X}$ *that we see in our one dataset* is very rare for a certain μ, that μ is unlikely to be true

Let's draw a picture of what is happening

DATA IS EVIDENCE

Normal Distribution Reminder



If $X \sim N(\mu, \sigma^2)$

then

$aX + b \sim N(\mu+b, a^2\sigma^2)$

… therefore:

$(X-\mu)/\sigma \sim N(0,1)$

Practice the Paradigm

X ~ f(x) with mean 2 and sd 3
X_i (i = 1, 2, …, n) are iid draws from X, and n is large

For what values of a, b is $(\bar{X} - a)/b \sim N(0,1)$?

Is $\bar{X} = 3$ a reasonable observation if n = 40?

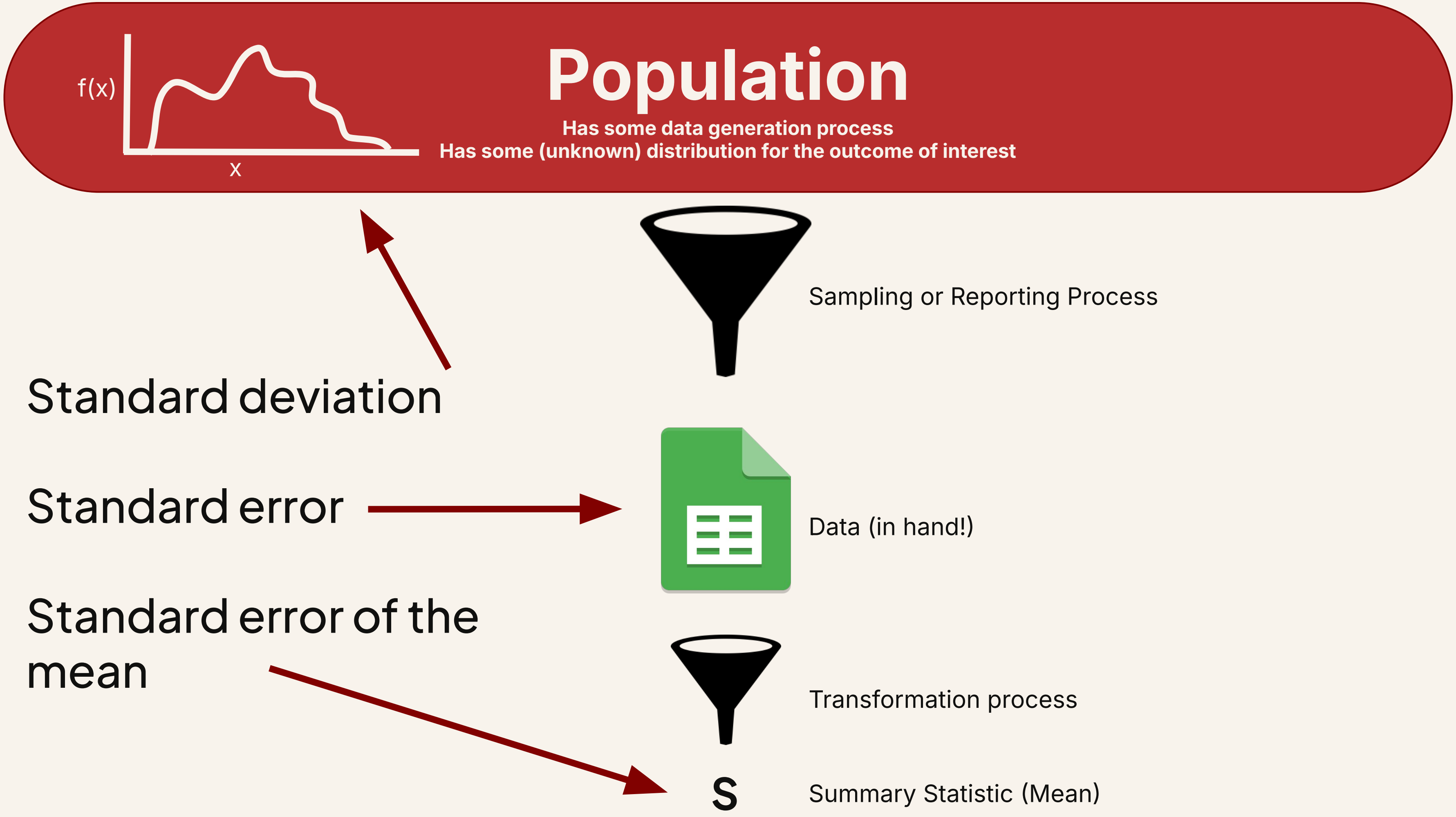
Repeat for X ~ bin(5, .8). Why is the result so different?

Terminology

1. Standard deviation
2. Standard error
3. Standard error of the mean

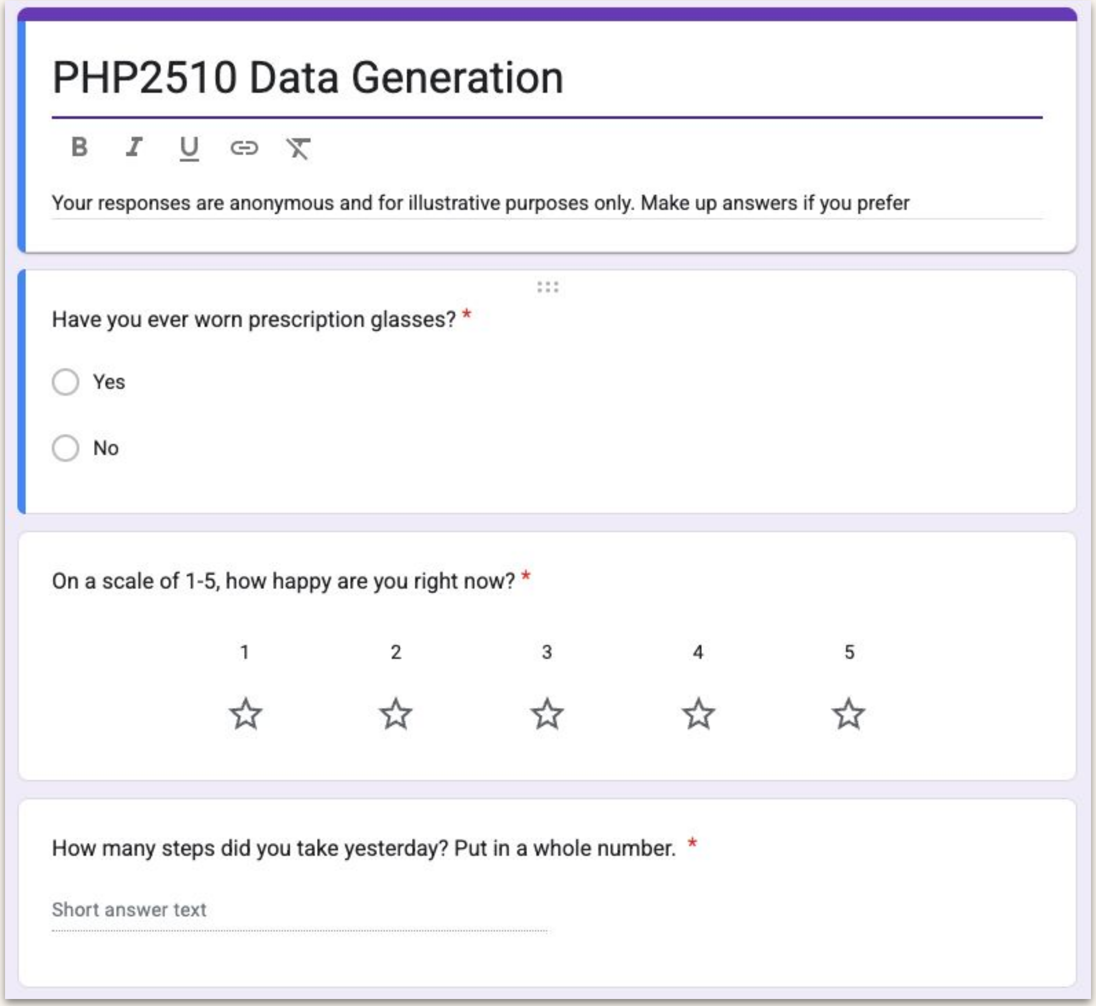
Terminology

1. Standard deviation ← what we learned in week 4 (RV theory)
2. Standard error ← an estimate of (1), from our data
3. Standard error of the mean ← an estimate of the standard deviation of the mean, a function of (1) and n


Population
Has some data generation process
Has some (unknown) distribution for the outcome of interest
f(x)
x
Sampling or Reporting Process
Standard deviation
Standard error
Data (in hand!)
Standard error of the mean
Transformation process
S
Summary Statistic (Mean)

Part 4

CLT Simulations



PHP2510 Data Generation

Your responses are anonymous and for illustrative purposes only. Make up answers if you prefer

Have you ever worn prescription glasses? *

- Yes
- No

On a scale of 1-5, how happy are you right now? *

1 2 3 4 5

How many steps did you take yesterday? Put in a whole number. *

Short answer text

Practice Problems

CHIHARA 4.16 (pg 96) - a and c only

Maria claims she has drawn a random sample of size 30 from the exponential distribution with lambda = 1/10. The mean of her sample is 12.

What is the expected value of a sample mean?

Is her sample mean unusual?

Practice Problems

CHIHARA 4.19 (pg 97)

$X\_1, \ldots, X\_10 \sim \text{iid } N(20, 8^2)$
$Y\_1, \ldots, Y\_15 \sim \text{iid } N(16, 7^2)$

Let $W = \overline{X} + \overline{Y}$

What is the **exact** distribution of W?
Simulate in R and confirm
Calculate $P(W < 40)$ via simulations and exactly

Practice Problems

CHIHARA 4.37 (pg 102) - a only

A random sample of size $n=100$ is drawn from a distribution with:
$f(x) = 3(1-x)^2$ for $0<x<1$

Use the CLT to approximate $P(\bar{X} < 0.27)$

Note: the pdf above comes from taking the minimum of 3 draws from a U[0,1] distribution, like in the BAC example

Reminder

Exam #1 on Tuesday, covering:

- Probability (joint, conditional, marginal)
- Diagnostic testing and 2×2 tables
- Random variables, including expectations & variance
- Foundational R coding
- Sampling distributions & CLT

Logistics

- No external resources, computers, calculators, etc - just pencil & paper
- Don’t memorize distributional formula - it will be given if needed
- 45 minutes; Second half of class we will discuss hypothesis testing

OUTCOMES

Q&A

After this week's classes, along with the required readings (CHIHARA Chapter 4; SPEEGLE Chapter 5, Sections 5.4+), you should be able to:

- Determine how well-defined functions of RVs behave
- Define the concepts of a sampling distribution and the standard error of a statistic
- Explain the Central Limit Theorem and its importance for making inferences
- Differentiate between the standard deviation of a population and the standard error of the mean

Next Week

Hypothesis Testing

Thank you!